Stochastic mechanistic models and Bayesian inference for direct and environmental disease transmission: applications in aquaculture

Lee Benson

A thesis presented for the degree of
Doctor of Philosophy

Computing Science and Mathematics
University of Stirling
United Kingdom

March 2021

Stochastic mechanistic models and Bayesian inference for direct and environmental disease transmission: applications in aquaculture

Lee Benson

Abstract

Stochastic dynamic epidemiological models are key to data-driven understanding of infectious disease. This thesis aims to expand, develop and provide deeper insights into applications of direct (DT) and environmental transmission (ET) models as mechanistic descriptions of environmentally transmitted infections like cholera and diseases affecting aquaculture production.

We explore timescale separation between host and environmental pathogens as the factor determining whether ET may be adequately described by a DT model. When fitting DT models to data, graphical posterior-prediction demonstrates robustness to departures from DT.

Rates of environmental transmission and pathogen decay of white spot disease among penaeid shrimp are estimated from published data and used to formulate the DT, susceptible-exposed-infected-removed (SEIR), and ET susceptible-exposed-infected-removed-pathogen (SEIR-P) models. Investigation of regular partial removal of dead and diseased shrimp reveals that host-disease dynamics of both models are almost indistinguishable with 24-hourly removals. However, for 6-hourly removals the SEIR model under-estimates average reduction in outbreak size and over-estimates duration compared with the SEIR-P model, demonstrating limitations of DT models.

Novel methods of Bayesian inference are shown to reliably estimate key model parameters using data from routine immersion challenge experiments (ICE). How design of such experiments can be aided using simulation and ideas from information theory is shown. Using published data and particle-marginal Markov chain Monte Carlo (PM-MCMC), we estimate the transmission rate and mean duration of latent infectiousness for a novel strain of Piscine orthoreovirus affecting rainbow trout.

Finally, using PM-MCMC, we show that rates of environmental transmission, pathogen emission and loss of pathogen viability of an ET model that distinguishes between viable and non-viable pathogen may be estimated from routine ICE data with additional digital polymerase chain reaction measurements of waterborne pathogen load. This work opens up application of ET models to develop understanding of pathogen dynamics beyond the confining assumption of direct transmission.

Supervisors: Dr Andy Hoyle, Dr Darren Green, Prof. Glenn Marion, Dr Ross Davidson & Prof. Mike Hutchings

Acknowledgements

I would like to give my deepest thanks to my supervisors Glenn Marion, Darren Green, Andy Hoyle, Ross Davidson and Mike Hutchings for their support and sage guidance throughout. These last few years have been happy and productive ones, thanks to their wisdom, experience and good humour. It takes a village to write a PhD thesis, and so thanks also to all the staff and students at BioSS, the University of Stirling and SRUC. Hopefully "Biscuit Wednesday" will soon be back! A big thank you to David Nutter for all the help with software and IT.

Finally, to Ajesh Patalay, for the loving support right from the off, thank you.

This work is dedicated to my parents, Lynn and Roy Benson, and grandparents, Josephine and Eric Wales.

Contents

Declaration

I hereby declare that this dissertation is the result of my own work and includes nothing that is the outcome of work done in collaboration, except where specifically indicated in the text and bibliography.

I also declare that this dissertation (or any significant part of this dissertation) is not substantially the same as any that I have submitted, or that is being currently submitted, for a degree or diploma or other qualification at the University of Stirling or similar insitution.

This disseration is a record of the work carried out at the University of Stirling between 2016 and 2021, under the supervision of Dr Andrew Hoyle and Dr Darren Green and of Professor Glenn Marion at BioSS and Professor Michael Hutchings and Dr Ross Davidson at SRUC.

Lee Benson

March 2021

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Modern epidemiology's basis in the germ theory of disease, which posits that pathogenic microorganisms, such as bacteria, viruses, fungi and macroparasites, are the cause of transmissible illness, has its roots in empirical study by scientists in the nineteenth century. Most notable among these are John Snow, whose investigations into the spread of cholera [Sno49] offered an evidence-based refutation of the, then prevailing, miasma theory of disease, as well as Louis Pasteur, Robert Koch and Joseph Lister [Gay11, BB10, Wor13, McG85]. In a continual arms race [WSB+21], these pathogens exhibit multifarious ways of invading their hosts' defences in order to replicate themselves. The distances and durations that pathogens can travel from one host to the next vary from immediate, short-range transmission due to contact with bodily fluids or close proximity, so called *direct* routes of transmission, to environmental routes involving disperal of aerosols and contamination of water and food sources.

The aim of this thesis is to improve the understanding of epidemiology through the representation of complex processes of disease transmission and progression by mechanistic models. Especially, we will address the quantification of such models using data that is becoming increasingly available with the advancement of a wealth of technologies. Data-informed mechanistic models of disease spread will bring predictive power to formation of disease control and surveillance strategies. Some of the practical challenges in modelling and inference that this raises will be addressed by the work here. Direct transmission models (DTM) are more widely used for quantitative prediction, because the data requirements are reduced, compared with environmental transmission models (ETM) and partly because modern statistical methods for fitting ETMs are less developed than for their DTM counterparts. This thesis aims to develop tools and insights that enable expanded use of both types of transmission models, and in particular to narrow this gap between them.

1.1.1 Disease surveillance, monitoring and prediction in the Information Age

Data and mechanistic models together provide the key to understanding the spread of infectious disease among humans, livestock and wild animal populations. The first instance of systematic collection and reporting of disease cases came in the nineteenth century, coinciding with early public health interventions, such as the introduction of the smallpox vaccine [SGOV16] and the Broad Street water pump investigation of 1854, which by plotting local incidences of cholera on a streetmap of London, John Snow was able to pinpoint a water source as the source of infection [fI21] - see Figure 1.1. Reporting developed at the turn of the twentieth century to include the coding of cause of death on death certificates, and with the advent of computers in the twentieth century, health reporting went digital [SGOV16]. The leap from an empirical, observational understanding of disease spread (e.g. studying case incidences) to a mechanistic one that captures disease outbreaks as a process came in 1927 with Kermack & McKendrick's Contribution to the Mathematical Theory of Epidemics [KM27]. Mechanistic disease models (which we will return to below) enable the close study of the complex patterns of outbreaks, which themselves are difficult, if not impossible, to observe fully. More importantly, they allow predictions to be made about future patterns of disease outbreaks. There has since been an explosion in the field of mathematical epidemiolgy, with particular focus on how these models are quantified using observational data.

In 2012, Forbes' magazine heralded the Age of Big Data [Pet12]. Nearly a decade on, we live in an increasingly data rich world. This is transforming how we understand and respond to outbreaks of infectious disease. Internet search patterns, digitised health records and social media now augment classical disease surveillance methods. Between 2008 and 2013, Google operated Flu Trends, which mapped influenza and influenza-like illnesses in real time using search engine query data. This was highly successful up until the 2012-2013 winter influenza epidemic [GMP$^+$09]. Currently, internet search data are being used to map the clinical course of COVID-19 disease progression [LR21]. GPS tracking and proximity detection technologies, such as bluetooth, are providing online data flows detailing human movements and interactions (thanks to smartphones, which by the end of 2021 almost half of the world's 7.9 billion people are expected to own [O'D20]). GPS data are also assisting the study of diseases in wild animals, e.g. bovine TB in badgers [GKM$^+$18] and livestock [BTKT18]. The use of wearable fitness sensors is projected to exceed 1.1 billion by 2022 [Hol20] and more medically oriented devices integrating a wide range of technologies, including optical, temperature, movement and chemical sensing [YMHÜ$^+$18] will provide streams of physiological data. Novel pathogen detection technologies, such as digital polymerase chain reaction (dPCR) [SJ13], are enabling real-time monitoring for airborne pathogens in public spaces [FY15, LSJ$^+$20]. Similarly, low cost technologies can provide autonomous early detection of waterborne pathogens, such as cholera, which is transmitted to humans via contaminated water sources [WHO21b, ZGH$^+$19] and the many viral and bacterial pathogens that affect aquaculture production. Fifth generation (5G) networks, allowing up to a million devices per square kilometre to connect to the internet [ARM21], and distributed, mobile edge computing and machine learning will bring us closer to the mass adoption of "telemedicine" and automatic disease diagnosis [ZLZ$^+$20].

An extremely useful dataset describing human movements and interactions was obtained from a citizen science experiment conducted by the BBC to mark the centenary of the 1918 Spanish flu pandemic [oC18]. In two experiments, the movements of 28,947 [KKG18] volunteers from across the UK were tracked. The first experiment tracked people for three days in the southern English town of Haslemere, and the second experiment tracked a number of participants for 24 hours nationwide. Participants were then asked to report any contacts they had had in the course of their day, along with some estimated demographic information about those they came into contact with. The dataset formed from the nationwide experiment was used by a team of researchers from the University of Cambridge and the London School of Hygiene and Tropical Medicine to put together a model simulating UK-wide spread of an influenza-like illness that first emerges in Haslemere. The team noted interesting interactions between patterns of movement and the participants' age, gender and whether they live in an urban or rural setting, with rural dwellers tending to spend more time far from home [KKG18]. By mapping the origins and destinations of journeys of more than 200 km, and noting that these tended to coincide predominantly with areas of high population density, such as towns and cities, the data allowed the researchers to calibrate that part of their model that governs how the infection jumps between distant areas of higher population density. Rather than relying on an assumption of how infections jump over long distances, the data informed a much more specific mechanism [KKG18] and therefore, potentially, a more realistic picture of disease spread. The BBC Contagion! dataset, now in the public domain, has since been used to model localised COVID-19 control strategies and how such interventions interact with complex human social behaviours [FHK+20].

In addition to these data flows there are *challenge studies*, i.e. observations of the effects of infectious disease and the dynamics of its progression through a population, under controlled conditions. These shed light on how various factors, such as the environment and genetics, interact with the epidemiology of the infection, such as the likely severity of illness and the ease with which it is transmitted among the population. How genetics can confer greater or lesser resistance to a disease, influence a host's infectiousness and the likely severity of resulting illness are currently the subject of many challenge studies in animals, e.g. a study by [RBM+14] suggests that resistance to bacterial infection by *Aeromonas hydrophila* in rohu carp (*Labeo rohita*) is linked to genotype and a study by [DAS+21] suggests that breed composition may influence the progression of bovine spongiform encephalopathy in steers. The effects of environmental factors on disease are also the subject of numerous studies, including the effect of water temperature on expression of nervous necrosis virus in the Australian bass species *Macquaria novemaculeata* [JFWH19] the effects of acidity and salt concentration on the mortality of Vietnamese striped catfish, *Pangasianodon hypophthalmus* due to bacterial infection with Edwardsiella ictaluri [PRC20]. There have been recent developments in the use of models to enable inference of genetic effects from challenge studies, e.g. [PMB+20].

Challenge studies are not, however, limited to fish, livestock and wild animals. In October 2020, a human challenge study of SARS-CoV-2 in the UK was the first of its kind to have received ethical approval [Kir20]. Alongside testing the effectiveness of vaccines, an aim of this study is to ascertain the minimum viral dose required to induce infection in young healthy people. Such information on minimal dose will help to inform a *dose-response* relationship,

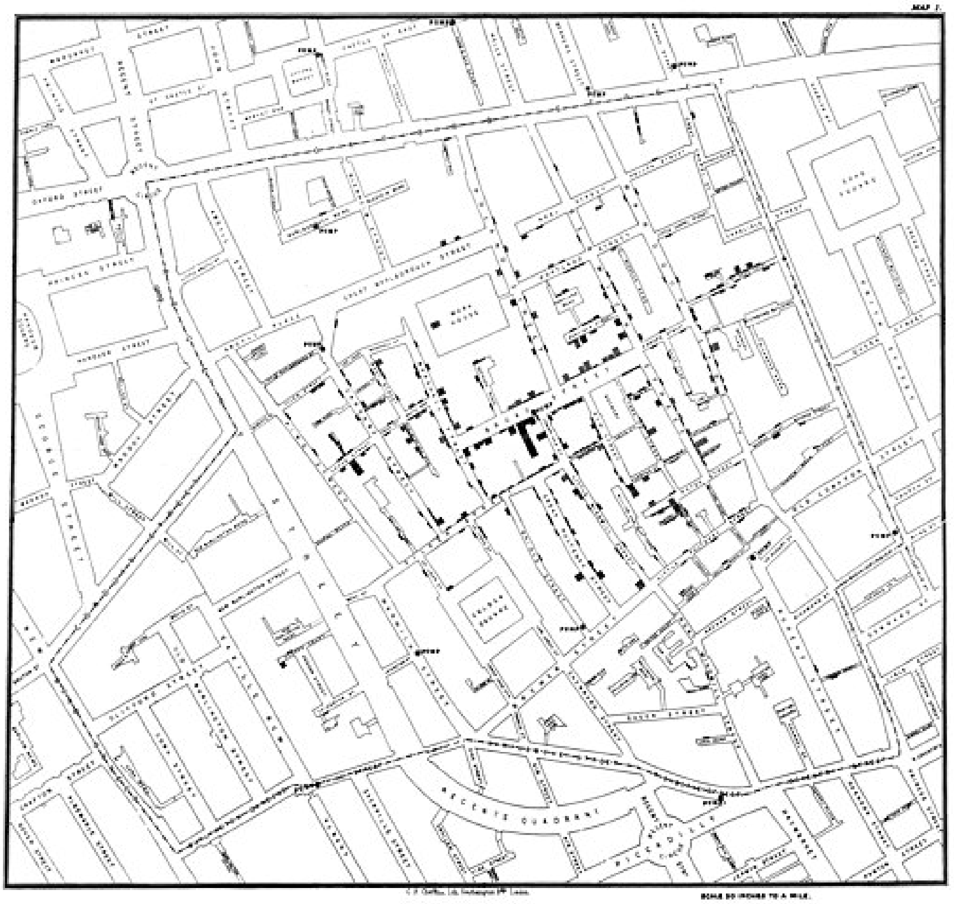


Figure 1.1: **John Snow's map** showing cholera cases (indicated by stacked rectangles) in the London epidemic of 1854. Clear clustering around Broad Street can be seen from the pattern of cases and is the location of the contaminated water source. Source `https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak#/media/File:Snow-cholera-map-1.jpg`.

.

relating the likelihood that an individual is infected upon exposure to the virus to the amount of pathogen that they are exposed to. More broadly, such information can, in turn, be linked into knowledge of persistence of pathogens in the environment and the dynamics of air flows or water currents in order to predict the spatial spread of infections. An example of this having been done is a reproduction, in simulation, of a large community outbreak of SARS in the Amoy Gardens residential complex in Hong Kong in 2003 [YLW$^+$04]. Using computational fluid dynamics, researchers were able to reproduce the temporal and spatial pattern of cases observed to occur around 24-29 March 2003 in the vicinity of an air shaft thought to be the origin of a virus-laden plume of air. The directional bias in the spatial distribution of cases toward "the north, west, and southwest of Amoy Gardens" was linked to the prevailing wind direction at the time [YQTW14]. This and similar studies helped to confirm the airborne spread of SARS and now work is underway to mitigate the risks of airborne transmission of SARS-CoV-2 in public buildings and transport infrastructure [CSBM20, MC20, PPS20].

1.1.2 Waterborne pathogens and the global aquaculture industry

A particular focus of application for the work of this thesis are diseases that affect aquaculture production. Disease poses the most significant threat to aquaculture production, to the tune of 6 billion US dollars annually, according to the World Bank [Ban14]. The Food and Agriculture Organization of the United Nations (FAO) estimate that aquaculture, i.e. the farming and harvesting of fish (i.e. "fish, crustaceans, molluscs and other aquatic animals, but exclud[ing] aquatic mammals, reptiles, seaweeds and other aquatic plants" [FAO20]) accounted for 46% of the 179 million tonnes (401 billion USD in value) of fish produced in 2018 and 52% of fish consumed by humans [FAO20]. Overall growth in production in the Americas and Europe has reversed in recent decades [FAO20] - they contributed 14% and 10% respectively to global production in 2018. However, Asia and, in particular, China have seen significant growth, both in overall production, and in the share of this contributed by inland and marine aquaculture [FAO20]. China has seen rapid development of production capacity in the last 30 years [GTL+18, Chapter 1.5] and in 2018 accounted for 35% of global fish production [FAO20]. One key driver for this growth is the low requirement for external feed inputs for many stocks meaning aquaculture is an efficient means of protein production [GTL+18, Chapter 1.5].

Aquaculture practices vary with geography and environmental conditions. For example, in Scotland the rearing of Atlantic salmon and other salmonid species, such as rainbow trout, in addition to shellfish, including Pacific and native oysters, scallops and mussels, accounts for the bulk Scottish of aquaculture [Sco21]. Atlantic salmon is a major export industry in Scotland, with over 162 thousand tonnes produced in 2016 [Sco21]. In Southeast Asia production focusses on freshwater species, such as carp, catfish and tilapia, as well as numerous species of shrimp in brackish waters and molluscs in marine environments [HBRY09]. Shrimp represent a major export market for this region [HBRY09].

While overall growth in capture fisheries has stalled since around the late 1980s [FAO20], due to increased aquaculture, growth in fish production has outpaced all other food production sectors over the last 40 years [BBS+15]. In the context of the United Nations' call for the doubling of food production by 2050 to remain in line with projected population growth [Nat09], fish consumption is expected to play a critical role in prevention of malnutrition [LMWH18]. The aquaculture industry is therefore vital to global food security.

The culture of salmonid species, including Atlantic salmon and rainbow trout, alone is threatened by an array of viral pathogens that includes infectious pancreatic necrosis virus (IPNV), which causes severe mortality [LSB19, SJBPP03] and piscine orthoreovirus (PRV), the cause of heart and skeletal muscle inflammation (HSMI) [PMSG19]. Many such pathogens have previously existed endemically in wild populations, only to become a problem when "environmental conditions become stressful to populations, compromising immunity and thereby the capacity to cope with pathogens, resulting in epidemics" [VLD+20]. The incursion of pathogens into populations without natural defences can also result in disease outbreaks [VLD+20].

The outbreak of the viral disease *infectious salmon anaemia* (ISA) among farmed salmon populations in Chile beginning around 2007 shows the potential scale of economic impact that disease outbreaks can have upon a single country's aquaculture production. Prior to this, none of the pathogens affecting salmon farming in Northern Europe, including ISA, had yet reached this part of the world and Chile was considered to be a "nearly disease-free environment" for Atlantic salmon farming [FCY+19]. The outbreak is estimated to have cost from 350-400 thousand tons of fish, 2 billion USD and 20,000 jobs [Ban14]. The plot in Figure 1.2 shows estimated annual outputs of Atlantic Salmon for Chile for the years 2000 to 2018; the size of the slump in output around the year 2010 is clear. The World Bank cites rapid, unregulated growth in the salmon farming industry, and insufficient controls to prevent disease introduction and spread as causal factors for the outbreak. Measures taken in response to this outbreak included the establishment of mandatory surveillance and a closer partnership between industry and government [Ban14].

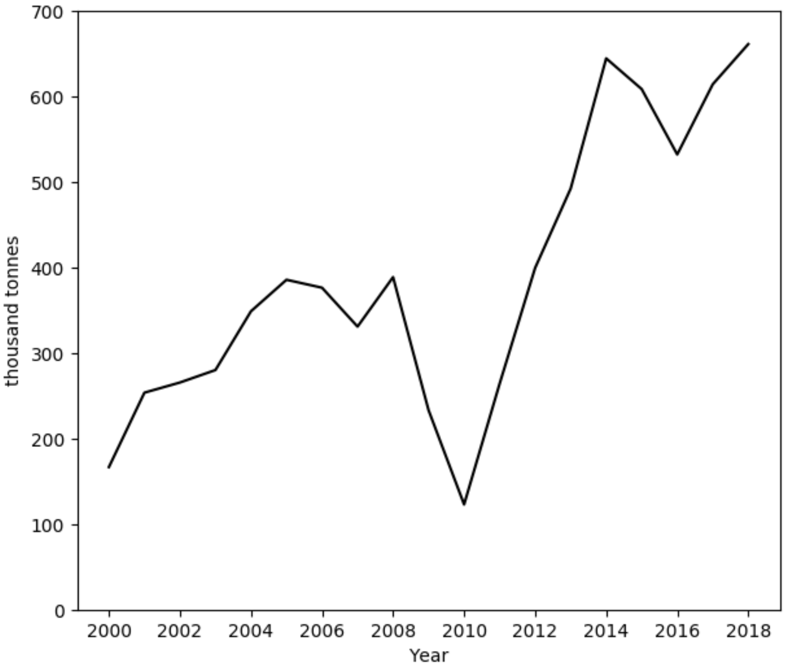


Figure 1.2: **Estimated annual output of Atlantic Salmon from Chile 2000-18**. The fall in output in 2010 is due to an outbreak of ISA among farmed salmon populations. Source: FAO - Fisheries and Aquaculture Information and Statistics Branch `http://www.fao.org/fishery/statistics/global-aquaculture-production/query/en`. Query terms: Country, Chile; Fishing area, All; Environment, All; Species, Atlantic salmon; Years, 2000-2018. Query date: 24/02/2021.

White spot disease (WSD) in penaeids, caused by the white spot syndrome virus (WSSV), is another example of a disease affecting aquaculture production with a signficant economic impact. WSD first emerged in Taiwan in 1992 and spread throughout Southeast Asia and the Americas by 1999 [SP10, Ban14]. Once the virus has entered a shrimp pond, 100% mortality can occur within 3 to 10 days [Ban14]. It is known to be long-persistent in numerous substrates [SAR+13] and affects, or is carried by, a wide range of species [SP10, Ban14, ELEBCH+09]. The particular susceptibility of the giant tiger prawn *Penaeus monodon* to the virus led to the precipitous fall in its global share of shrimp production and to its subsequent replacement with more resistant species, including the whiteleg shrimp *Penaeus vannamei* (see Figure 1.3). We present a detailed case study of WSD in Chapter 2.

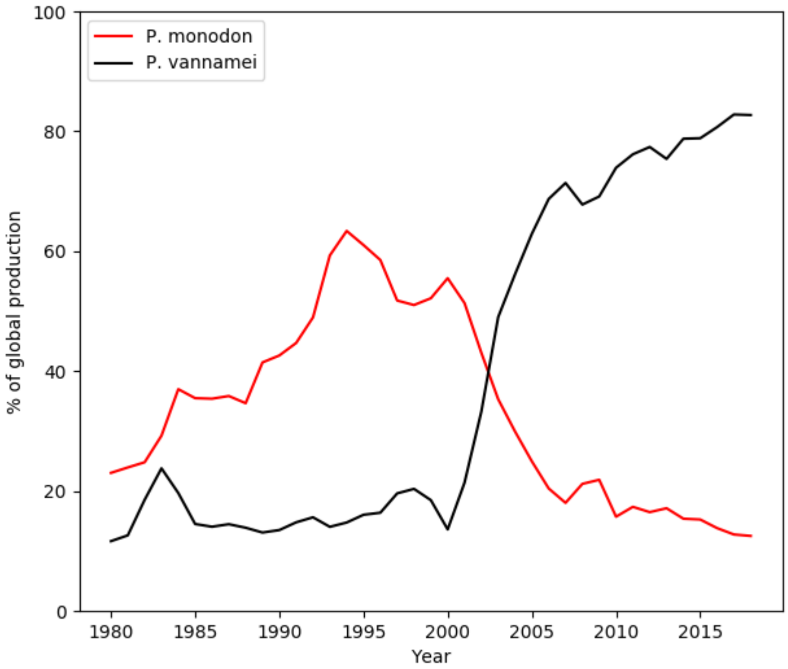


Figure 1.3: **Estimated percentage share, by weight, of global annual output of shrimp represented by giant tiger prawn**, *P. monodon*, **and whiteleg shrimp**, *P. vannamei*, **1980-2018**. The World Bank cites the particular vulnerability of *P. monodon* to WSSV as the cause of its virtual replacement as key penaeid production species [Ban14]. Source: - Fisheries and Aquaculture Information and Statistics Branch `http://www.fao.org/fishery/statistics/global-aquaculture-production/query/en`. Query terms: Country, All; Fishing area, All; Environment, All; Species, giant tiger prawn & whiteleg shrimp; Years, 1980-2018. Query date: 24/02/2021.

According to [BOM17], the epidemiological challenges that specifically confront disease surveillance in aquaculture include sampling confidence; the authors ask "does the absence of disease presence really mean the presence of disease absence?" Surveillance of diseases, similar to WSD, that can wipe out an entire stock within days of introduction are therefore particularly problematic for classical surveillance methods that involve periodic sampling of animals. The same authors additionally cite the hierarchical, interconnected nature of aquaculture; ponds are connected within farms, and farms are connected by common water courses. Pathogens can be carried by water currents to infect neighbouring sea cages [ODW^{+}18]. Modern, data-driven disease surveillance in the aquaculture sector would therefore benefit greatly from a quantification of the risks associated with disease outbreaks in relation to the presence and dynamics of environmental pathogens.

1.2 Disease modelling

In terms of understanding outbreaks of an infectious disease, often we are looking to answer a range of questions. For example, what is the likelihood that an outbreak occurs following the entry of a single infected individual, *the index case*, into a disease-free population? And supposing that an outbreak does happen, how rapidly will it spread, and how much of the population will be affected? We might be concerned with the progression of disease; how many will get seriously unwell and how long will it take to recover? Or we may wish to understand the effect that a range of environmental, genetic, social and physiological factors have on the patterns of disease outbreaks, or indeed the effects of various interventions, such as vaccination, social distancing and face masks, or water flow-through in aquaculture production.

Disease outbreaks are often difficult phenomena to observe in their entirety, however. For example, mild or asymptomatic cases will often go unreported, at least in the early stages, e.g. the asymptomatic transmission of COVID-19 [PL20]. Data on disease outbreaks are often sparse and the times of critical events, such as when individuals first became infectious, are typically unavailable [Bec89]. *Dynamic*, or *mechanistic, epidemiological models* (disease models) are mathematical abstractions of the "the mechanisms of disease transmission and pathogenesis" [LEH$^+$15]. By providing an approximate, idealised version of these mechanisms, disease models fill in the gaps in observation due to unobservable events and missing data, enabling prediction and the testing of hypotheses [LEH$^+$15].

Models vary in terms of the extent to which they present a simplified picture of real-world complexity. A very common simplifcation in the field of disease modelling is to subdivide the host population into a small number of definite, clearly defined, disease states. Disease progression is then described soley in terms of movements of individuals between these disease states. All disease models used in this thesis are, so-called, *compartmental models.* The seminal *Susceptible-Infected-Recovered* (SIR, see Fig. 1.4) model of Kermack & McKendrick (1927), which we discuss below, is the cannonical example. As the name suggests, the SIR model splits up the target population into those who risk catching the disease, those who currently have it and pose a risk to others, and those who, having had it, are at no further risk of infection. Using this SIR model, the authors addressed the question of whether disease outbreaks only come to a halt when either there are no more individuals susceptible to the disease or when the "virulence of the causative organism has gradually decreased" [KM27]. Through analysis of the model, the authors suggested that a combination of factors concerning population density and "various factors of infectivity, recovery and mortality" determine the magnitude of an outbreak. Many variations of the SIR model have subsequently been employed, e.g. inclusion of additional disease states to reflect more complex patterns of disease progression (such as latently infected, or vaccinated states), or demographic structuring of the host population to reflect interactions with factors such as age and occuption, on an individual's susceptibility to disease.

When an outbreak of disease spreads sufficiently rapidly that we do not have to take into account demographic movements, such as births, non disease-related deaths and immigration, etc., then outbreaks tend to follow a common, general trajectory. There is an initial phase when cases grow exponentially, followed by a decline in the number of new cases as the number of susceptible members of the population is depleted. It is commonplace to think in terms of chance, however, when considering the complex patterns of who is infected when, and by whom. Indeed, we often ask how *likely* is an outbreak to occur under certain circumstances or what are our *chances* of catching the disease. *Stochastic* disease models incorporate these chance effects into their mechanistic representation of the disease system. In small populations, and more generally at the early stages of an outbreak when case numbers are still low, stochastic models "are preferable" [Bri10] to wholly determinstic ones. Since we will predominantly be thinking about disease spread among small popualtions, in this thesis we consider only stochastic compartmental models.

1.2.1 Direct transmission models

One use of stochasticity in models is the representation of the mechanisms that drive transmission among the members of the population. We now describe a common class of stochastic compartmental disease models used to describe infections passed by either physical contact or close proximity, so called *direct transmission models* (DTM). The SIR model (Figure 1.4) will serve as our example.

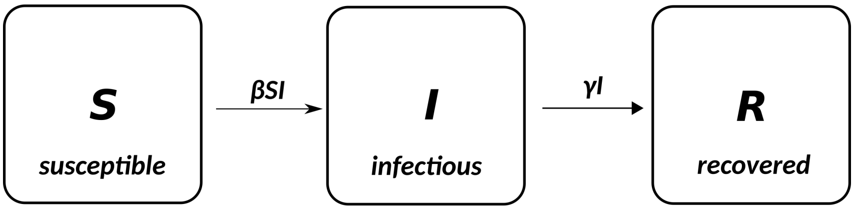


Figure 1.4: **Susceptible-infectious-recovered model diagram**. The boxes represent the three disease states into which the population is divided up and the arrows are the allowed transitions between states. The parameters β and γ are, respectively, the direct transition and the recovery rate, assuming exponentially distributed infectious lifetimes.

Disease transmission, according to a DTM, occurs only at random sequences of events that represent close contacts between pairs of members of the population at which there is a risk that the infection can be passed from one individual to the other. Depending on the infection, this can describe direct physical contact, as well as close-range transmission via respiratory droplets and even faecal-oral transmission. In fact, many infectious human and non-human diseases have been described with DTMs, as discussed in the introduction to Chapter 2. Following Diekmann, Heesterbeek & Britton (2012), we use the term *infectious contact* to refer to these idealised events. Whenever, an infectious contact occurs between two individuals, one of whom is susceptible and the other infectious, transmission will occur.

For the DTMs that we employ in this thesis we make two very significant assumptions. The first of these is that each member of the population's pattern of infectious contacts follows a *homogeneous Poisson process* (PP). A PP is obtained by drawing a sequence of independent random *inter-event times*, $T_0, T_1, T_2, \ldots$, with exponential distributions with a fixed rate, known as the rate of the PP. T_0 is the first contact time, and for $t < T_0$, no infectious contacts have yet happened; for $T_0 \leq t < T_0 + T_1$, there has been exactly one, and so on. More precisely, the number of infectious contacts that an individual experiences up to and including time t is

$$C(t) = \min\{n \geq 0 : t < T_0 + \cdots + T_n\}. \tag{1.1}$$

An important feature of PPs is that they are examples of *Markov processes*. The Markov property implies that, knowing the number of contacts that have occurred by time t, the parts of the process occurring before and after t - the "past" and the "future" - are independent of each other. This property is often expressed as *memorylessness*, since the process "forgets" its history if its present state is fully known. In fact, if we have observed c infectious contacts up to and including time t, then $C'(s) = C(s) - c$ for $s > t$, the pattern of contacts going forward, is yet another PP that is unaffected by what has occurred prior to time t. A typical initial pattern of contacts for one individual is shown in Figure 1.5.

For our models we always assume a fixed population of size N, and that the rate of PP describing each member of the population's sequence of infectious contacts is proportional to this population size, so that the rate is βN; this is called *density-dependent* transmission, since if the population size is doubled (while keeping the area or volume in which the population lives the same), the rate of infectious contact is doubled also. The quantity β is the *direct transmission rate* appearing above the first arrow in Figure 1.4. Another common relationship between the rate of infectious contact and population size or density is described by *frequency-dependent* transmission, where the contact rate is independent of the population size or density and commonly describes sexual transmission, for example [LSGW04]. We do not discuss frequency-dependent transmission in this thesis.

This Markovian PP formulation of the contact and infection process, though widely adopted (see, e.g. [NTS$^+$16, DGB$^+$19, CRW$^+$20, Mon20]), is not the end of the story. The model may be equivalently reformulated in terms of accumulation of infection pressure by each susceptible individual, each of whom posseses an "infection threshold" drawn randomly from an exponential distribution [Sel83] . As soon as the susceptible individual's accumulated infection pressure reaches their threshold, they become infected. Streftaris and Gibson (2012) propose an extension whereby these infection thresholds are distributed more generally in order to account better for effects such as previous immunity due to vaccination and variations in susceptibility due to genetic factors, etc. [SG12].

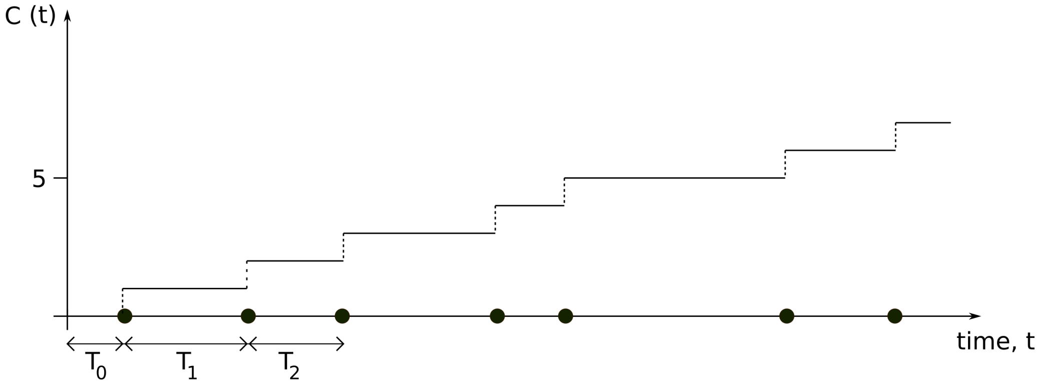


Figure 1.5: **An illustration of a Poisson process** describing the temporal pattern of infectious contacts, $C(t)$, experienced by a singe host. The times between contacts are independent of the times of previous contacts and are exponentially-distributed with some given, fixed rate. The *value* of the process at time t is the number of contacts that have occurred up to this time. Throughout this thesis we will assume that the rate of this process is βN, where β is the rate of direct transmission.

The second assumption that we make is that the host population is *mixing freely.* This means that each host comes into contact with all other hosts with equal likelihood, with no preference for certain members of the population over others, and in particular, no avoidance of infected individuals. As a consequence, when there are I members of the population currently infected at the time of some susceptible individual's infectious contact, the probablility that contact is with an infectious individual is $\frac{I}{N}$.

From these two assumptions we can easily calculate the force of infection (FoI) from the current state of the system. This is a susceptible individual's immediate risk, or *hazard*, of getting the infection at time t when there are currently I infected. Because the process is Markovian, for some susceptible individual at time t, the probability that their next infectious contact occurs in the interval $(t, t + \delta t)$ is the same as the probability that $T_0' < \delta t$, where $T_0' \sim \exp(\beta N)$ is the first contact time of their contact process set back to zero and started over at time t. Therefore, the next contact occurs in the interval $(t, t + \delta t)$ with probability

$$\mathbb{P}(T_0' < \delta_t) = \int_0^{\delta t} \beta N e^{-\beta N s}\, ds = 1 - e^{-\beta N \delta t} \tag{1.2}$$

and the probability that such a contact occurs in the interval $(t, t + \delta t)$ with an infectious individual is

$$(1 - e^{-\beta N \delta t})\frac{I}{N}. \tag{1.3}$$

Dividing the above by δt and taking the limit as $\delta t \to 0$ gives a susceptible's immediate risk of infection at time t under this model

$$\text{FoI} = \lim_{\delta t \to 0} \frac{(1 - e^{-\beta N \delta t})\frac{I}{N}}{\delta t} = \beta N \cdot \frac{I}{N} = \beta I. \tag{1.4}$$

Adding up the βI for the S members of the population who are currently susceptible gives the overall rate of transmission βSI, which appears above the leftmost arrow in Figure 1.4. Newly infected individuals remain so for a period of time, known as their *infectious lifetime*, which is a random quantity that is typically (but not necessarily) exponentially distributed. Once the infectious lifetime has passed, the infected individual moves into the R compartment (this is the second arrow in the Figure).

In the case that the infectious lifetimes are exponentially-distributed (as in Figure 1.4, with rate γ), the SIR model is an example of a continuous-time Markov chain with transition rates listed in Table 1.1.

SIR model outbreaks following the introduction of a single infectious individual into a population of size $N - 1$ can be efficiently simulated using the *Doob-Gillespie* algorithm as follows:

1. set the initial compartment sizes, $(S, I, R) = (N - 1, 1, 0)$, and initial time $t = 0$

Transition		rate	description
$(S, I, R) \to$	$(S-1, I+1, R)$	βSI	infection
	$(S, I-1, R+1)$	γI	recovery

Table 1.1: **Markov transition rates** for the SIR continuous-time Markov chain, with exponentially-distributed infectious lifetimes. The rates are interpreted as, for example, $\lim_{h\to 0}\frac{1}{h}\{\mathbb{P}(S_{t+h} = s-1, I_{t+h} = i+1 | S_t = s, I_t = i) - \beta si\} = 0$.

2. set $t \to t + \Delta t$, where $\Delta t \sim \exp(\beta SI + \gamma I)$

3. draw $U \sim \mathrm{U}(0, 1)$. If
$$U < \frac{\beta SI}{\beta SI + \gamma I} \tag{1.5}$$
an infection has occurred, so set $S \to S-1, I \to I+1$. Otherwise, a recovery has occurred, so set $I \to I-1, R \to R+1$

4. repeat steps 2 and 3 until $I = 0$.

1.2.2 Environmental transmission models

As explained above, a consequence of modelling disease transmission via series of infectious contacts between the members of the host population is that the force of infection is directly proportional to the number of infectious individuals currently present. This may be too restrictive for some purposes, such as when considering transmission via environmental pools of pathogens that have been shed by infectious hosts. We now briefly describe *environmental transmission models* (ETM), a second class of stochastic compartmental disease model that we consider in this thesis.

ETMs describe two populations, the hosts and the environmental pathogens. In addition to density-dependent transmission between hosts, ETMs allow infection to be transmitted indirectly due to contact between susceptible hosts and the pathogens that have been shed by infectious hosts. An example is the susceptible-infected-recovered-pathogen (SIR-P) model, summarised in Figure 1.6. This model has three additional parameters, the environmental transmission rate, α, and the pathogen emission and decay rates, ϵ and ρ.

In addition to the free mixing of the host population, as required for the SIR model, we also assume that the pathogen is also evenly mixed throughout the space occupied by the hosts. its two routes of transmission, the FoI under the SIR-P model is

$$\mathrm{FoI} = \alpha P + \beta I \tag{1.6}$$

where P is the size of the pathogen load.

As for the SIR model, in the case of exponentially-distributed infectious lifetimes, SIR-P trajectories can be simulated using the Doob-Gillespie algorithm, this time with transition rates in Table 1.2.

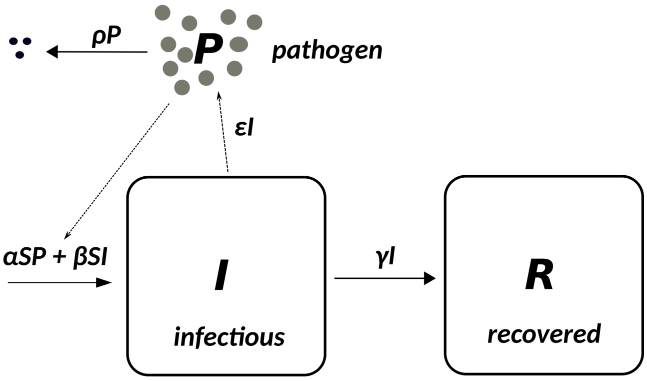


Figure 1.6: **Susceptible-infectious-recovered-pathogen model diagram**. The boxes represent the three disease states into which the population is divided up and the arrows are the allowed transitions between states. The parameters β and γ are the direct transmission and host recovery rates, as for the SIR model. Additionally, there is the environmental transmission rate, α, that determines the risk of infection to susceptible hosts, relative to the magnitude of the pathogen load. The behaviour of the pathogen load itself is determined by the pathogen emission rate, ϵ, which is the amount of pathogen infectious hosts emit, per unit of time, and the rate that pathogen becomes uninfectious in the environment, ρ. Solid arrows represent the possible changes within the host and pathogen populations and the dashed arrows show how both the host and the pathogen populations affect the dynamics of the other.

Transition		rate	description
$(S, I, R, P) \rightarrow$	$(S-1, I+1, R, P)$	$\alpha SP + \beta SI$	infection
	$(S, I-1, R+1, P)$	γI	recovery
	$(S, I, R, P+1)$	ϵI	pathogen emission
	$(S, I, R, P-1)$	ρP	pathogen decay

Table 1.2: **Markov transition rates** for the SIR-P continuous-time Markov chain, with exponentially-distributed infectious lifetimes.

In the following chapters various generalisations of the SIR and SIR-P models will be presented. For example, models with additional host and pathogen states will be considered, as well as models with non-exponentially-distributed infectious lifetimes. In Chapter 4 an ETM will be discussed in detail where the pathogen load is modelled as a continuous, rather than discrete process.

1.3 Bayesian inference

In order for our understanding of the spread of disease to be data-driven, we have to inform our models from observation of outbreaks. Model-based Bayesian statistical inference allows us to express our knowledge about the unknown parameters and unobserved events of a model in terms of "probablility statements" [GCS+13] and a framework for updating these as more information about the system is obtained.

Throughout, we will use the generic symbols $p(\ldots)$ and $p(\ldots \mid \ldots)$ to represent distribtions or densities. Letting θ, x and y stand respecitively for the model's unknown parameters, the unobserved event sequence, such as infection times, and observed data, then their joint probability distribution can be factorised

$$p(y, x, \theta) = p(y \mid x) p(x \mid \theta) p(\theta). \tag{1.7}$$

The first of the three factors on the right hand side of Eq. 1.7 is the *observation model* that describes the probabilistic dependence of the observed data y upon the disease model's unobserved event sequence and the second factor is the *model likelihood*, which expresses the distribution of the model's event sequence in terms of fixed parameter values.

The third factor on the right hand side of Eq. 1.7 is the *prior.* This summarises our present knowledge about the model parameters. There exists a quantity, *Shannon entropy* [CAFD09, HK07], that measures the uncertainty of a distribution, which for the prior, $p(\theta)$, is

$$H(\theta) = -\int_{-\infty}^{\infty} p(\theta) \log p(\theta) \, d\theta. \tag{1.8}$$

The Shannon entropy is a key concept in a measure of expected information gained from an experimental outcome, the *mutual information*, discussed in Chapter 3. Among priors on a finite interval, (a, b), the uniform distribution on (a, b) has maximal Shannon entropy [CAFD09]. Generalising this to general distributions, this expresses the idea that "flatter" priors convey *less* information about the quantity of interest.

By conditioning $p(y, x, \theta)$ on the result of an observation, y, we can use Bayes' rule to obtain the *posterior distribution*

$$p(x, \theta \mid y) = \frac{p(y, x, \theta)}{\int \int p(y, x, \theta) \, dx \, d\theta}. \tag{1.9}$$

For a Bayesian, the posterior is a complete description of the model parameters and event sequence after observing the data y. For this reason, we sometimes say that we have "updated" our prior with the data y. The denominator in Eq. 1.9 is called the *model evidence* and measures the overall likelihood of the observation y, averaging over the parameters and unobserved events.

Having updated our prior, $p(\theta)$ with data, y, how do we get predictions from our model? The *posterior-predictive* distribution, $p(\tilde{y} \mid y)$, can be obtained by averaging the likelihood over the posterior

$$\begin{aligned} p(\tilde{y} \mid y) &= \int p(\tilde{y} \,|\theta, y) \, p(\theta \mid y) \, d\theta \\ &= \int p(\tilde{y} \,|\theta) \, p(\theta \mid y) \, d\theta \end{aligned} \tag{1.10}$$

since the observations y and $\tilde{y}$ are conditionally independent, given the parameters, θ. The posterior-predictive distribution is a key constituent of graphical goodness-of-fit checks for models, known as *graphical posterior-predictive checks* (see, e.g. [GCS$^+$13]).

1.3.1 An example from Neal and Roberts (2005)

As a concrete example, we consider the problem presented in [NR05] of obtaining the posterior distribution, $p(\mathbf{I}, \beta, \gamma \,|\, \mathbf{R})$, where β and γ are the parameters of an SIR model of disease with exponentially distributed infectious lifetimes. In the case under study the infection times, $\mathbf{I} = (I_1, \ldots, I_m)$, $I_1 < I_2 < \cdots < I_m$, are unobserved and $\mathbf{R} = (R_1, \ldots R_m)$ are the observed recovery times, forming the data y, using the above notation. In this case, the data $y = \mathbf{R}$ are just a subset of the full SIR event sequence, $x = \mathbf{I}, \mathbf{R}$. As in [NR05], we can write down the model likelihood, $p(\mathbf{I}, \mathbf{R} \,|\, \beta, \gamma)$, by working through the sequence of probabilistic events comprising the Doob-Gillespie algorithm described at the end of the previous Section, and re-arranging to obtain

$$p(\mathbf{I}, \mathbf{R} \,|\, \beta, \gamma) = \prod_{i=2}^{m} \{\beta S_{I_i-} I_{I_i-}\} e^{-\beta \int_0^{R_{\max}} S_t I_t \, dt} \prod_{j=1}^{m} \gamma e^{-\gamma \left(R_j - I_j\right)} \tag{1.11}$$

(see Chapter 2 for more detail). In the above equation, S_{I_i-} and I_{I_i-} are the numbers of susceptibles and infectives just before the i^{th} infection and $R_{\max}$ is the last recovery time. Examining the functional form of Eq. 1.11, as a function of β only (holding all other quantities fixed), the right hand side of Eq. 1.11 is proportional to

$$\beta^{m-1} e^{-\beta \int_0^{R_{\max}} S_t I_t \, dt} \tag{1.12}$$

and likewise for γ, the right hand side of 1.11 is proportional to

$$\gamma^m e^{-\gamma \sum_{j=1}^{m} (R_j - I_j)}. \tag{1.13}$$

This suggests that the family of gamma-distributions are a convenient choice from which to select priors

$$\begin{aligned} p(\beta) &\propto \beta^{\nu_\beta} e^{-\lambda_\beta \beta} \\ p(\gamma) &\propto \gamma^{\nu_\gamma} e^{-\lambda_\gamma \gamma} \\ p(\beta, \gamma) &= p(\beta) p(\gamma) \end{aligned}$$

Diffuse priors can be obtained by setting the “shape” parameters, ν_β and ν_γ, to unity and the “rate” parameters, λ_β and λ_γ to a value such as 0.001, as suggested by Neal & Roberts (2005).

Equation 1.9 will give us the posterior, $p(\beta, \gamma, \mathbf{I} \,|\, \mathbf{R})$ as soon as we calculate the model evidence for the removal times $\mathbf{R}$

$$p(\mathbf{R}) = \int p(\mathbf{I}, \mathbf{R} \,|\, \beta, \gamma) p(\beta, \gamma) \, d\beta \, d\gamma \, dI_1 \ldots dI_m. \tag{1.14}$$

However, this problem is analytically intractable due to the integral of the likelihood with respect to $I_1 \ldots I_m$. Rather than seeking a closed analytical form for the posterior, we would instead draw on a large and growing array of approximate sampling techniques, known collectively as *Markov chain Monte Carlo* (MCMC). These involve using computer simulation to draw chains of dependent samples, whose marginal distributions eventually converge to the posterior. Once this convergence has been judged to have been achieved (typically by inspection of trace plots or numerical diagnostics), we can begin to collect samples. From these samples we can estimate various quantities relating to the posterior, such as the posterior mean and variance, as well as to obtain density estimates etc. MCMC in practice is part-theory-part-art and there are significant practical challenges relating primarily to the speed at which the marginal distributions of the samples converge to their target. Selection of sampling routines appropriate to the problem at hand, and the tuning of these routines, are not trivial tasks. However, in this thesis we present several methods in application to various problems.

1.4 Thesis outline

The goal of this thesis is to advance an undertanding of the spread of infectious disease through mechanistic models and Bayesian inference. This comes at a time when technologies ranging from communication to pathogen detection promise rich disease surveillance and monitoring data. Because of the enormity of the impact that disease has on the aquaculture sector and its crucial role in future global food supply, our focus in applications will be environmental spread of waterborne pathogens that affect these production stocks in particular.

In Chapter 2 we will discuss the wide application of direct transmission models (DTM), such as the susceptible-infectious-recovered (SIR) model, including in cases of environmental transmission. As described above, the form usually taken by the SIR and similar models can be derived from the assumptions that the population is mixing freely and disease is transmitted between members of the population only at discrete points in time described as a Poisson process. We see in the literature the wide application of these models to cases of disease where the second of these assumptions might not apply, such as when disease is transmitted by contact with long-persistent pools of environmental pathogens. Using simulations, we will show that DTMs offer a faithful picture of the host-disease dynamics in many of these cases and we will explain why this is the case. We will then fit DTMs to simulated data describing environmentally transmitted disease and assess their fit by comparing the data with the model's posterior-predictive distribution. This way we will see that we can fit DTMs to data from a wide array of environmentally-transmitted diseases and get sensible parameter estimates. This is a great benefit to researchers since DTMs do not require observation of the dynamics of the pathogen load in order to fit them. The second half of Chapter 2 is an examination of a case study of white spot disease among penaeid shrimp. We estimate the parameters of an environmental transmission model (ETM) of the disease with both direct and environmental routes of transmission. We then use this model to compare the effectiveness of two disease control strategies in production settings, exploring some of the ideas introduced earlier in the chapter.

There are many descriptions of challenge experiments in the recent scientific literature examining the spread of waterborne infections under controlled conditions. Many of these experiments are used to form empirical judgements about e.g. the effectiveness of treatments and vaccines. However, the cases where these data are used to estimate the parameters of mechanistic models are relatively few. This presents numerous wasted opportunities. Even a very simple model with parameters estimated from observation of the system of interest would allow researchers to make a range of predictions and test varied hypotheses about the impacts of interventions on disease spread. Simulations of mechanistic models with parameters estimated from experiment are a low cost way to meet the legal and ethical obligations to reduce the number of animals used in such studies (e.g. the 3R principle in the UK [ASP14]). A fitted model can also offer a guide as to where the gaps in our knowledge are and so guide further data collection [LEH+15]. Chapter 3 is an overview of model estimation using Bayesian inference and data obtained by immersion challenge experiment. We present a case study in which we estimate the parameters of a susceptible-exposed (i.e. latently infected)-infectious (SEI) model of a novel *Piscine orthoreovirus* among rainbow trout from real data gathered in a trial by [HVT+17]. We then discuss the elements of the design of similar experiments and how this is aided by computer simulation and ideas from information theory.

Although, as we will discuss in Chapter 2, the mechanisms of the spread of environmentally transmitted disease are well-described by DTMs, more predictive power can be gained from understanding how the dynamics of these infections relate to those of the environmental pathogen load. For example, the risk that a localised viral outbreak poses to clusters of nearby salmon sea cages could be quantified from models of prevailing water currents and an understanding of how various environmental conditions affect the longevity of the virus in the water body, along the lines of the analysis of the Amoy Gardens SARS outbreak of 2003 [YLW+04, YQTW14]. As a first step, an ETM of the disease in a closed population relates the transmission dynamics between the hosts to the behaviour of the pathogen load. Chapter 4 tackles the problems of the design of a challenge study and the estimation of the ETM model parameters. We show that designs of real experiments described in the scientific literature, augmented by a small number of measurements of waterborne pathogen concentration using digitial polymerase chain reaction (dPCR) technology, are sufficient to quantify an ETM of the disease. We describe the application of a novel MCMC method of Bayesian inference to the fitting of ETMs.

Finally, in Chapter 5, we pull together some final observations and suggest directions for future work. The novel contributions of this thesis include the identification of a stochastic compartmental DTM as the timescale separation limit of certain stochastic compartmental ETMs in Chapter 2 and the fitting to data of stochastic compartmental ETMs that account for the discrepancy between detectable pathogen concentration levels using technologies such as dPCR and the concentration of viable, infectious pathogen in Chapter 4. In addition, the estimation of the parameters of an ETM describing white spot disease in penaeid shrimp has not been attempted elsewhere and the discussion of stochastic modelling and Bayesian inference in the context of study of waterborne infection via immersion challenge experiment data seeks to bring greater appreciation of these tools to a wider audience.

Chapter 2

When and why direct transmission models can be used for environmentally persistent pathogens

2.1 Introduction

The famed Susceptible-Infectious-Recovered (SIR) compartmental model framework of Kermack and McKendrick [KM27], and its many subsequent extensions (see [KHM19], for example), stand as prominent examples of what can be gained from simple models of complex systems. In addition to the assumption that the host population can be divided into a finite number of discrete states, transmission of infection within such models is characterised by a force of infection that depends linearly upon the numbers of infectious individuals. Throughout this paper, we call such a model a direct transmission model (DTM) (as in [McC01]). The proportionality constant, known as the transmission rate, is often interpreted as the rate at which individuals within the population come into contact with each other, times the probability that such a contact leads to the transmission of infection (termed an infectious contact in [Die00]), times the probability of successful transmission. However, this simplified representation has been extended, for example, to account for non-uniform frequency of contact among the population and levels of infectiousness varying across individuals and over time.

Such approaches have been pivotal in gaining valuable insights into the dynamics and patterns of the spread of disease throughout many varied populations. Recently, for example, as the basis for understanding drivers of spatial spread of Ebola virus [MAF$^+$15], the likely effectiveness of scaling up certain vaccination, treatment and testing regimes in the fight to control hepatitis B [NTS$^+$16] and the importance of targeting household transmission of MRSA as a preventative strategy [DGB$^+$19]. Incorporation of a spatial element into the DTM framework enables the observed spatial-temporal trajectory of the 2001 foot and mouth outbreak in the UK to be closely replicated and provides insight for control [Kee01, Kee05]. DTMs have also been recently drafted into the effort to understand and predict the dynamics of SARS-CoV-2 [CRW$^+$20, Mon20].

Despite these sucesses, DTMs may not always be appropriate, e.g. when members of the host population are in contact with environmental sources of infection, such as pools of pathogens residing on surfaces or in water bodies. The focus of this paper is to critique the use of DTMs in describing the spread of disease in the presence of such environmental pools of persistent pathogens. Examples of relevant disease systems include Cholera [VPHC10, AMT14], avian influenza [SSKZ90, HCR^{+}17] and even respiratory infections, such as SARS-CoV-2 [ZLZ^{+}20]. Our case study (Section 2.4) focusses on infectious disease spread in aquaculture systems which are likely to feature a large degree of environmental transmission. In scenarios such as these it is prolonged exposure to these sources, in addition to direct infectious contact between individuals, that gives rise to new infections, in the most general case.

A critical step in applications is the fitting of DTMs to data on disease outbreaks. Using Bayesian model fitting methods and graphical posterior predictive checks (GPPCs) that target observable characteristics of an outbreak, such as its final size and when it peaks, we show that DTMs fit very well to simulated host-disease event times that would occur when the infection is transmitted environmentally but the rates of pathogen emission and removal are high (Section 2.3). Using a combination of GPPCs each targetting different observable aspects of an outbreak, increases the likelihood of spotting deficiencies in the fitted model and consequently provides a more reliable test of the model. Gibson (2018) suggests that the choice of outbreak characteristic to target with GPPCs should depend on the intended use of the fitted model [GST18]. E.g. if we are primarily interested in estimating attack rates, then our model's predictions of final outbreak size should chiefly interest us.

We show that the fit of DTMs to ET data is explained by the force of infection within an *environmental transmission model* (ETM, Section 2.2) behaving increasingly like that of a DTM when the timescale of the pathogen population is shorter than that of the host population (Section 2.3). We find that, in terms of predicting outbreak size and duration, when fitted to outbreak data, DTMs are robust even in the case of long-lived pathogens with low rates of emission, and it is only the rate at which an outbreak grows toward its peak, and the timing of that peak that are poorly predicted.

These issues are further highlighted in a case study illustrating the use of DTMs as approximate descriptions of outbreaks of disease due to an environmentally persistent pathogen (Section 2.4). Using parameter values estimated from published data on white spot disease (WSD - see Section 2.4) infection in penaeid shrimp, we explore how imperfect interventions that aim to remove dead and diseased hosts at regular intervals impact outbreak control in closed populations of this aquaculture disease system. With removal attempts spaced at 24 hourly intervals, average outbreak trajectories, final outbreak size and outbreak duration are accurately captured using a DTM without needing to model the pathogen load. When the frequency of the removal events are increased to every 6 h, we begin to see divergence between the two models, so that e.g. the DTM predicts slightly larger outbreaks of shorter duration than the ETM. This case study illustrates the potential practical consequences of ignoring issues of timescale separation when applying DTMs to environmentally transmitted pathogens. Control and other processes in such disease systems may be accurately captured by DTMs at one timescale but are poorly represented at others; in this case under estimating the benefit of high frequency removal.

2.2 Materials and methods

2.2.1 Models for direct and environmental transmission of disease

2.2.1.1 A direct transmission model: Susceptible-Exposed-Infectious-Removed (SEIR)

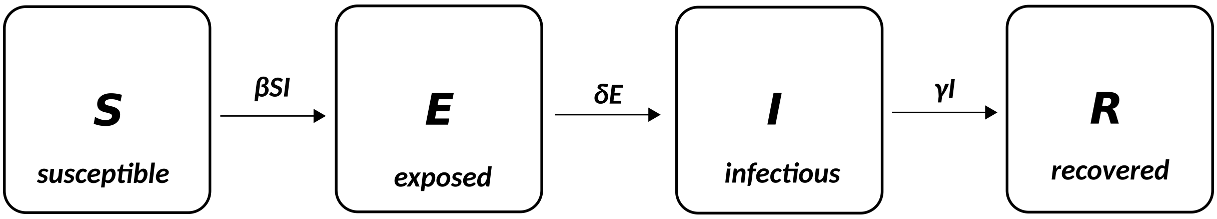


Figure 2.1: **Susceptible-exposed-infectious-removed (SEIR) direct transmission model diagram** illustrating the four compartments of the model and the rates of the possible movements between them. Parameter β is the *direct transmission rate* while, in the Markovian case, δ and γ are the rates that hosts move from the E to the I and the I to the R states, respectively.

The stochastic, compartmental SEIR (see Figure 2.1) model referred to throughout this paper is a DTM that treats the focal population (the *hosts*) as compartmentalised into four sub-populations: hosts that are susceptible to disease (S), have been exposed to the disease but not yet infectious (E), infectious (I) and recovered or removed from the population (R). Hosts in the R compartment play no further role in the spread of disease. Here we consider outbreaks started by the introduction of a single host to a wholly susceptible population of size N (this is known as the *index exposure*), however, our results are generalisable to greater numbers of initial exposures. Additionally, we assume a closed population, so that there is no immigration, births or non disease-induced mortality. Hosts remain in the E and I compartments for periods of time determined by the *exposed* and *infectious lifetime distributions*, i.e. random variables with continuous, positive distributions. When exponentially-distributed lifetime distributions are assumed, the resulting process is a continuous time Markov chain and the current state of the system is fully determined by the number of hosts in the four compartments. In such a case we use the symbols δ and γ to respectively represent the rates that hosts move from the E to I and from the I to the R compartments. The SEIR model may be specialised by stipulating that hosts spend a period of zero duration in the E compartment (i.e. $\delta = \infty$), resulting in the familiar SIR model. Alternative gamma and Weibull-distributed lifetimes can also be assumed. These more general distributions model non-constant hazard of becoming infectious after exposure and recovering (or dying) from the disease after onset of infectiousness. In Section 2.3.2 we fit a SEIR model with exponentially-distributed lifetimes to simulated SEIR-P data where the corresponding lifetimes are gamma-distributed. We find that the fitted model at least correctly produces the expected durations in the E and I states.

In Section 2.4 we modify this model by additionally allowing, at regularly spaced intervals, each host in the E and the I compartments to go directly to the R compartment with probabilities π_E and π_I. This simulates regular attempts, with error, to remove all exposed and infectious hosts from the system.

The direct transmission assumption means that secondary cases are generated at a rate dependent on the number of infected individuals. Here we adopt the standard approach to modelling this. The probability that each susceptible host at time t becomes exposed to disease over the short time interval $(t, t+h]$, is $\beta I_t h$ to first order, where β is the *direct transmission rate* and I_t is the number of infectious hosts present at time t. The force of infection, i.e. the rate of secondary infections, per susceptible host, or the imminent risk of infection that is faced by each susceptible host, at time t is therefore expressed as

$$\lambda(t) = \beta I_t. \tag{2.1}$$

This means that the force of infection is proportional (by the factor β) to the number of infected hosts currently present. When the number of infected hosts increases or decreases, there is a corresponding instantaneous change in the force of infection. This is a key feature of this and other DTMs and represents an important modelling assumption.

2.2.1.2 An environmental transmission model: Susceptible-Exposed-Infectious-Removed-Pathogen (SEIR-P)

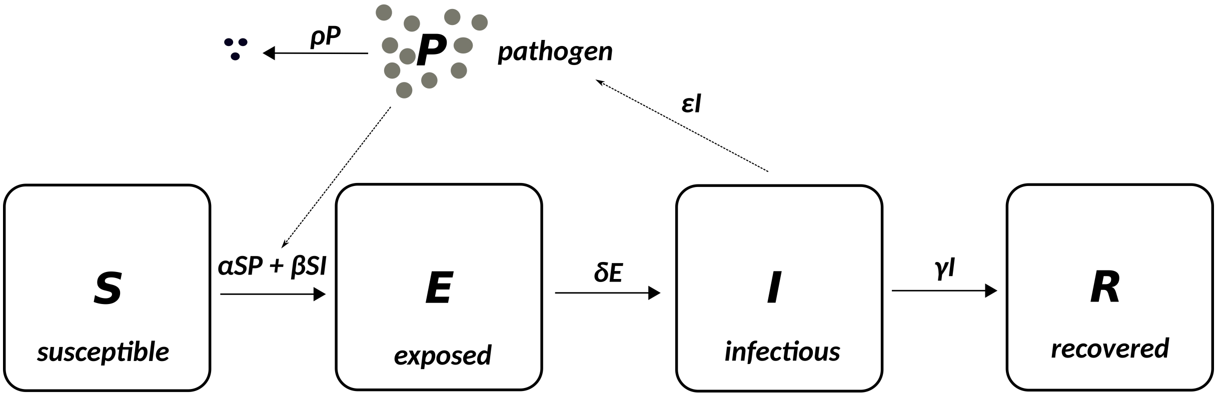


Figure 2.2: **Susceptible-exposed-infectious-removed-pathogen (SEIR-P) environmental transmission model diagram** illustrating the four host compartments and single pathogen compartment of the model. The solid arrows indicate the movement of hosts between host compartments and the loss of viable pathogen from the system. The dotted arrows indicate how the host and pathogen parts of the model influence each other. In addition to the parameters β, δ and γ, which are the same as for SEIR (for the Markovian case, see above), α is the *environmental transmission rate* and ϵ and ρ are the *rates of pathogen emission and removal.*

Here we define a class of models that represent both direct (DT) and environmental transmission (ET) of disease (see, for example, [SR13]) where the latter occurs via interaction of susceptible hosts with environmental pools of infectious pathogen. SEIR-P describes the time evolution of two populations: the hosts, divided into S, E, I and R sub-populations (as in the DTM SEIR, described above) and the pathogen population (P) in the environment, external to the hosts. When hosts enter state I they begin to emit pathogen at the fixed rate ϵ, i.e. they begin to contribute to an increase in the environmental pathogen load. The pathogen population decays exponentially at rate ρ.

Each susceptible host at time t now becomes exposed to disease over the short time interval $(t, t+h]$ with probability $(\alpha P_t + \beta I_t)h$, to first order, where β and I_t represent direct transmission, as in the SEIR model, and α and P_t represent the *indirect* or *environmental transmission rate* and size of the pathogen population at time t, respectively. The force of infection is now

$$\lambda(t) = \alpha P_t + \beta I_t, \tag{2.2}$$

which depends linearly on both the size of the environmental pathogen load and the number of infectious hosts. A change in the number of infected hosts now produces a delayed response in the force of infection due to the pathogen load taking time to either build up or decay.

Setting α to zero and restricting attention to the four host population compartments recovers the SEIR model described above. Similarly, $\beta = 0$ produces a purely ET model, and finally, setting $\delta = \infty$, so that hosts pass instantaneously from the E to the I compartment, yields a SIR-P model with both DT and ET.

2.2.2 Bayesian model fitting and inference

2.2.2.1 SEIR posterior, likelihood and prior densities

In Section 2.3 we fit the SEIR model with exponential exposed and infectious lifetimes to simulated data generated by both the SEIR and SEIR-P processes. Since in practice, the times of host exposure, $\mathbf{E}$, are very rarely observed directly, these are treated as *missing data.* It is often the case that that the times of onset of infectivity, $\mathbf{I}$, are also unobservable. However, we will be using the SEIR and SEIR-P models to describe WSD in penaeid shrimp (see Section 2.4), for which onset of infectiousness coincides roughly with the death of the shrimp. Therefore, in this particular case, the times of entry into the I state, corresponding to death (and consequent onset of infectiousness), and the R state, corresponding to removal from the system (either due to complete decay or physical removal from the waterbody) are feasibly observable. In cases where we cannot observe either the onset of infectivity or removal directly then model assessment is impacted since our data are less informative.

Bayesian inference regarding the parameters, β, δ and γ of the SEIR model is based entirely on the *posterior density*

$$p(\,\beta, \delta, \gamma, \mathbf{E} \,|\, \mathbf{I}, \mathbf{R}\,) \propto p(\,\mathbf{E}, \mathbf{I}, \mathbf{R} \,|\, \beta, \delta, \gamma\,)\, p(\,\beta, \delta, \gamma\,) \tag{2.3}$$

where $\mathbf{I}$ and $\mathbf{R}$ are the observed times of onset of infectivity and removal and $\mathbf{E}$ are the unobserved times of exposure. The two factors on the right hand side are respectively the *likelihood* and the *prior* densities. The prior density summarises our knowledge and uncertainty about the parameters *prior* to observing the data. Throughout, we will assume that the three parameters are *a priori* independent, i.e.

$$p(\,\beta,\delta,\gamma\,) = p(\,\beta\,)p(\,\delta\,)p(\,\gamma\,) \tag{2.4}$$

and that

$$\begin{aligned} p(\,\beta\,) &\sim \mathrm{Exp}(\lambda_\beta) \\ p(\,\gamma\,) &\sim \mathrm{Exp}(\lambda_\gamma) \\ p(\,\delta\,) &\sim \mathrm{Exp}(\lambda_\delta) \text{ or } \mathrm{U}(0, 10) \end{aligned} \tag{2.5}$$

where $\lambda_\beta = \lambda_\gamma = \lambda_\delta = 0.001$. Exponentially-distributed priors are chosen since they are conjugate with the functional form of the likelihood so that the full conditional distributions of the parameters are easily obtained, and such a value for the hyperparameters $\lambda_\beta, \lambda_\gamma$ and λ_δ means that there is a high prior variance, reflecting a lack of prior knowledge.

In the case of long-lived pathogen, SEIR is far from the "correct" model for the data (e.g. see [Ber66, Wal13] for accounts of fitting mis-specified models) and this can present issues with the convergence of the MCMC chains since there are no strong candidates among parameter values that simultaneously explain the data. In order to aid convergence, therefore, the alternative, uniformly-distributed, prior for δ is used in the long-lived pathogen case only (Section 2.3.2). By placing an upper bound (in this case, 10.0) on the support of δ, we prevent the chains from exploring parts of the sample space that represent exposed lifetimes shrinking to zero.

The likelihood, $p(\,\mathbf{E},\mathbf{I},\mathbf{R}\,|\,\beta,\delta,\gamma\,)$, describes how the data depend on the parameters of the model, which, in the case of exponentially distributed times in the E and I states is

$$p(\,\mathbf{E},\mathbf{I},\mathbf{R}\,|\,\beta,\gamma,\delta\,) \propto \prod_{i\neq\kappa}\{\beta I_{E_i}\}\, e^{-\beta\int_{E_\kappa}^{\infty} S_t I_t\,dt} \prod_{j=1}^{m} \delta e^{-\delta(I_j - E_j)} \prod_{j=1}^{m} \gamma e^{-\gamma(R_j - I_j)}. \tag{2.6}$$

$E_1,\ldots,E_m$ with integer subscripts denote the times of exposure events, whereas E_t, with a real-valued subscript, is the number of hosts that are exposed at time t. The same applies to I_i, I_t, etc. For example, the quantity I_{E_i} in (2.6) means the number of infected hosts at time E_i - the i$^{\text{th}}$ chronological time of exposure. The products over $j = 1,\ldots m$ in (2.6) are the contributions to the likelihood from the exponentially-distributed times spent in states E and I. The product over $i \neq \kappa$ are the contributions from the exposure times (which each have hazard proportional to βI_t).

The notation used here is similar to that used by Neal and Roberts in [NR05], so that E_κ represents the (perhaps unobserved) first exposure time. The reason why we write them this way, as opposed to assuming that the initial infection or exposure occurs at some known time, is that in Section 2.3 we will be fitting these models in scenarios where the exposure times have not been observed and so consequently we do not know when the outbreak began.

For details of model fitting, see Appendix A.

2.2.2.2 Model checking using graphical posterior-predictive checks (GPPC)

Graphical posterior predictive checks (GPPC) [GCSR95, GS13] are used here to test for departure from DT model assumptions. These are a standard model checking tool, offering a visual comparison of quantities derived from the observed data, $h(\mathbf{I},\mathbf{R})$, with $h(\mathbf{I}',\mathbf{R}')$, where $\mathbf{I}',\mathbf{R}'$ are simulated from the posterior-predictive distribution of the fitted model, with density

$$\begin{aligned} p(\,\mathbf{I}',\mathbf{R}' \,|\, \mathbf{I},\mathbf{R}\,) &= \int p(\,\mathbf{I}',\mathbf{R}' \,|\, \beta,\delta,\gamma,\mathbf{E},\mathbf{I},\mathbf{R}\,)\, p(\,\beta,\delta,\gamma,\mathbf{E} \,|\, \mathbf{I},\mathbf{R}\,)\, d\beta\, d\delta\, d\gamma\, d\mathbf{E} \\ &= \int p(\,\mathbf{I}',\mathbf{R}' \,|\, \beta,\delta,\gamma\,)\, p(\,\beta,\delta,\gamma,\mathbf{E} \,|\, \mathbf{I},\mathbf{R}\,)\, d\beta\, d\delta\, d\gamma\, d\mathbf{E} \qquad (2.7) \end{aligned}$$

Uncertainty regarding parameter values is expressed by drawing them from their posterior distribution. The idea is to check that the fitted model replicates the original data with reasonable probability, with no systematic disagreements between the data and model predictions.

Among the salient features of a disease outbreak are its size, at both peak and completion, and characteristic timescales, e.g. time from index exposure to peak of outbreak and total duration. Such statistics are of interest in their own right and there are known formulas in the deterministic case for peak and final outbreak sizes and initial exponential growth rates for SIR, SEIR and similar models [Fen07, Mil12, Ma20]. For the GPPCs here we obtain a probabilistic picture of similar quantities with the timing of the outbreak's peak standing as proxy for the initial growth rate. The following four quantities are considered

1. final outbreak size, m
2. outbreak duration, $R_{\max}-I_{\min}$, where $I_{\min} = \min\{I_1,\ldots,I_m\}$ and $R_{\max} = \max\{R_1,\ldots,R_m\}$
3. time of outbreak peak, t_{peak}
4. size of outbreak at peak, $I_{t_{\text{peak}}}$.

Since the simulated data contains times of onset of infectivity and host removal, the above quantities are indeed directly calculable. Additionally, the outbreak size trajectory, I_t, over the course of an outbreak will be examined. Due to likely correlations between m and $R_{\max} - I_{\min}$ and between t_{peak} and $I_{t_{\text{peak}}}$, these are plotted bivariately.

The time of outbreak peak, t_{peak}, is interpolated between the first and last times that the outbreak size is within the range $\kappa I_{\max} \leq I_t \leq I_{\max}$, where $0 < \kappa < 1$, i.e.

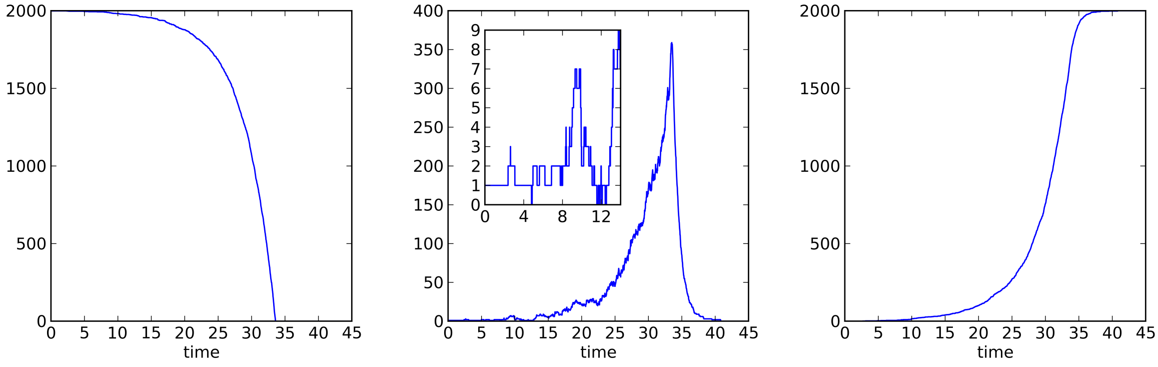
$$t_{\text{peak}} = \frac{\max\{t : I_t \geq \kappa I_{\max}\} - \min\{t : I_t \geq \kappa I_{\max}\}}{2}. \qquad (2.8)$$

Calculating t_{peak} this way, rather than simply taking it to be the time that the outbreak size first reaches its maximum, avoids the complication of the outbreak hitting its maximum size more than once and, more importantly, aims to reduce the variance of its posterior-predictive distribution, and therefore produce a sharper test for model departure. Here κ is chosen to be 0.3, since the GPPC outputs were not found to be sensitive to the particular value chosen.

2.3 Investigating direct transmission model performance

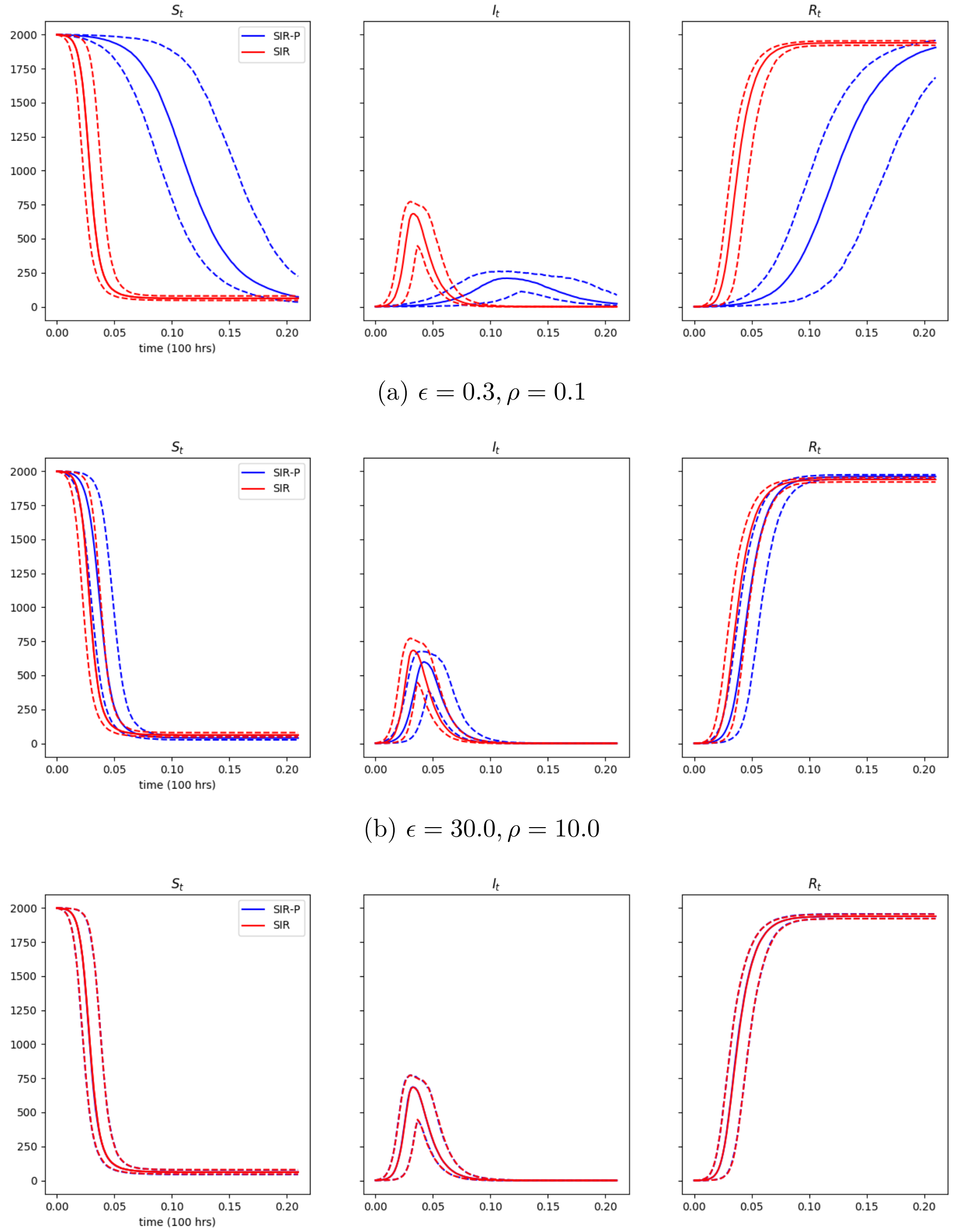
2.3.1 The direct transmission approximation as timescale limit of SEIR-P process

The two populations described by an ETM, the hosts and environmental pathogen, each have a timescale characterising their evolution, and the extent to which these differ has a qualtitative effect upon the behaviour of the model. For example, ETMs with shed pathogen retaining infectivity for long durations compared with the typical host infectious lifetime can exhibit outbreaks that appear to have died out, in terms of infected individuals, only to restart. Figure 2.3 shows output from a single simulation of the SIR-P model with exponentially-distributed exposed and infectious lifetimes, host population size $N = 2000$ and $(\alpha, \beta, \gamma, \epsilon, \rho) = (0.25, 0.0, 1.0, 0.32, 0.01)$. An emission rate of 3.0 part h^{-1} and host mortality rate of 1.0 h^{-1} means that infected hosts shed on average 3 pathogen particles during their period of infectiousness. Nonetheless, this shed pathogen, with a large indirect transmission rate, α, and low pathogen decay rate, ρ, eventually causes a large outbreak after a considerable period of time in which the outbreak appears to be small and contained. SIR-P and SEIR-P models such as this one are able to exhibit greater variation in the sizes and durations of outbreaks than their SIR and SEIR direct transmission counterparts due to their dependence upon the stochastic environmental pathogen load. On the other hand, ETMs with short-lived pathogens produce host-disease dynamics that are reproducible with a host-only DTM, as we now demonstrate.



(a) S_t (b) I_t (c) R_t

Figure 2.3: **Susceptible, infectious and removed host output from a single simulation of SIR-P model** with $N = 2000$ and $(\alpha, \beta, \gamma, \epsilon, \rho) = (0.25, 0.0, 1.0, 0.32, 0.01)$. Notice the delayed onset of the exponential growth phase of the outbreak, with several instances of the number of infectious hosts reaching zero (shown in the inset in the second plot) before the outbreak eventually takes off.



(a) $\epsilon = 0.3, \rho = 0.1$

(b) $\epsilon = 30.0, \rho = 10.0$

(c) $\epsilon = 3.0 \times 10^4, \rho = 1.0 \times 10^4$

Figure 2.4: **Susceptible, infectious and removed host sub-population sizes** of SIR-P (blue) process averaged over 2000 simulations for fixed $\alpha = 6.0 \times 10^{-4}$ and $\gamma = 1.0$. The rates of pathogen emission, ϵ, and pathogen removal, ρ, are increased while keeping their ratio fixed, $\frac{\epsilon}{\rho} = 3.0$. For comparison, the same sub-population sizes for the DTA SIR process with fixed direct transmission rate $\beta = \alpha\frac{\epsilon}{\rho} = 1.8 \times 10^{-3}$ and $\gamma' = 1.0$ are plotted in (red). Median population sizes indicated by bold lines, dashed lines indicate 5th and 95th percentiles. In sub-figure 2.4a the two processes are visibly distinct in their output with the SIR-P outbreaks tending to peak later and to be smaller in magnitude. With a hundred-fold increase in the pathogen decay rate, a closer alignment between the two sets of trajectories can be seen in sub-figure 2.4b. In the last case (sub-figure 2.4c) there is no difference on the scale of the plots between the SIR-P and SIR model outputs.

Recalling Section 2.2, the probability that a susceptible host at time t becomes exposed within the short interval of time $(t, t+h]$ is $(\alpha P_t + \beta I_t)h$, to first order. The relationship between the qualitative behaviour of the SEIR-P model and pathogen timescale comes down to the appearance of P_t in this expression. How does P_t behave? At time t, each infectious host is emitting pathogen at the rate ϵ so, overall, new pathogen is entering the population at the rate ϵI_t. At the same time, the pathogen decays exponentially at rate ρ, or equivalently, each pathogen element remains viable on average for a duration $1/\rho$ (in appropriate units). This amounts to P_t being an immigration-death process (IDP) with inhomogenous immigration rate ϵI_t and death rate ρ. As a consequence of elementary theory of Markov chains (see, e.g. [And91]), if we momentarily regard the number of infectives as fixed, $I_t = i$ for $t \geq 0$, then the distribution of P_t tends to Poisson($\frac{\epsilon i}{\rho}$) and

$$\mathbf{E}(P_t) \to \frac{\epsilon i}{\rho} \quad \text{as } t \to \infty. \tag{2.9}$$

Increasing the values of the parameters ϵ and ρ, while keeping their ratio constant, increases the rate of convergence but has no other effect upon this limiting behaviour, with the limiting distribution unchanged. Returning now to variable I_t, in the limiting case, P_t's behaviour can be characterised approximately as being in equilibrium, i.e. $P_t \sim$ Poisson($\frac{\epsilon i}{\rho}$) during intervals of constant $I_t = i$, and jumping without transition to a new equilibrium when I_t changes.

Consequently, for large ϵ and ρ, we may approximate the probability that a susceptible host becomes exposed over the interval $(t, t+h]$ as

$$(\alpha \mathbf{E}(P_t) + \beta I_t)h = (\alpha \frac{\epsilon}{\rho} + \beta) I_t h = \hat{\beta} I_t h. \tag{2.10}$$

The part of the SEIR-P model that describes the host disease dynamics is approximately a direct transmission SEIR model, with an *effective direct transmission rate* $\hat{\beta}$ and E and I lifetime distributions unchanged. This SEIR model is the *direct transmission approximation* (DTA) of the ETM.

This approximation of the sub-process, P_t, is the stochastic analogue of the "quasi-steady state approximation" of the pathogen concentration discussed by Tien and Earn in their susceptible-infected-water reservoir-removed (SIWR) ordinary differential equation model of cholera outbreaks among humans [TE10], in which the pathogen concentration in water sources is restricted to the "critical manifold" of fixed points of the flow.

That ET via short-lived pathogen can be approximated as DT is because of the vanishing lag between changes in the number of shedding, infectious hosts, and the resulting response in the force of infection. Systems with ET that results from the accumulation of relatively long-lived pathogen retain a *memory* of the size of the infectious host sub-population since hosts that have since been removed, or have ceased to shed, may still be the cause of new exposures via the pathogen that they had previously emitted. This delay between changes in the number of infectious hosts and their effect on the system dynamics violates the implicit assumption of DTMs that the force of infection is directly related to the number of infectious hosts, or the number of hosts who are shedding pathogen. As the pathogen lifetime decreases, so does this memory effect, and we see an increasing similarity with the dynamics produced by direct transmission.

Figure 2.4 demonstrates this behaviour by showing susceptible, infectious and removed subpopulation sizes averaged over a number of independent simulations from three SIR-P models with fixed α and γ and increasing ϵ and δ, with $\frac{\epsilon}{\rho}$ held constant. These are plotted in each case against similar outputs from a SIR model with the same γ and β chosen to equal $\alpha\frac{\epsilon}{\rho}$, so that the SIR model is the DTA of the SIR-P model. As shown in the figure, increasing rates of pathogen emission and decay rate lead to increasing similarity between the SIR-P and SIR model outputs, and in the case of short-lived pathogen, the SIR-P and DTA SIR outputs are indistinguishable on the scale of the plots.

2.3.2 Estimating DTA parameters from outbreak data

The presence of environmental transmission violates an assumption of DTMs: that the force of infection is directly related to the number of infectious hosts currently in the system. Here we fit the DT SEIR model, with exponentially-distributed exposed and infectious lifetimes, to data from simulated SEIR-P outbreaks with varied rates of pathogen emission and decay and assess the goodness of fit of the model using GPPCs. The simulated data consists of times of onset of infectivity and host removal, i.e. the times of entry of hosts into the I and the R states, the observation of which is feasible in cases where onset of infectivity coincides with the death of the host (see Section 2.4).

The data set specifications are summarised in Table 2.1 and the MCMC methods of model fitting and a description of the GPPCs are found in Section 2.2 and Appendix A. The first three data sets are generated using a SEIR-P process in which there is both environmental transmission due to pathogen as well as direct transmission from host to host. The rates of pathogen emission, ϵ, and decay, ρ, are increased with $\frac{\epsilon}{\rho}$ kept constant. The fourth data set has no environmental transmission rate, and so is the output of a DTM. We additionally test the fitted SEIR model's sensitivity to slight deviations from exponentially-distributed exposed and infectious lifetimes by using gamma-distributions with shape parameter greater than unity to simulate these.

2.3.2.1 MCMC convergence and posterior coverage

Convergence of the MCMC sampling chains is checked by running two separate chains for each fitted model with separated initial values and observing that they converge to a common stationary distribution. Figure B1 in Appendix B contains trace plots for the two parameters that were MCMC sampled, β and δ.

Figure 2.5 is a graphical summary of the samples obtained from the posterior distribution of the parameters and of the SEIR basic reproduction ratio (the expected number of new cases of infection that result from addition of a single infected host into a large, wholly susecptible population), $R_0 = \frac{\beta N}{\gamma}$, for each of the fitted models. In order to show that fitting the Markovian SEIR model to non-Markovian SEIR-P-generated host event times results in meaningful parameter estimates, these are plotted against certain reference values:

- $p(\beta \,|\, \mathbf{I}, \mathbf{R})$ is compared with $\hat{\beta} = \alpha\frac{\epsilon}{\rho} + \beta$ (α, ϵ, ρ and β are given in Table 2.1) - $\hat{\beta}$ is the effective direct transmission rate for the SEIR-P model, defined in Section 2.3.1
- $p(\delta \,|\, \mathbf{I}, \mathbf{R})$ with $\hat{\delta} = (\mathbb{E}(I_j - E_j))^{-1}$ and $p(\gamma \,|\, \mathbf{I}, \mathbf{R})$ with $\hat{\gamma} = (\mathbb{E}(R_j - I_j))^{-1}$. This way we assess the fitted Markovian SEIR model's ability to recover the mean, gamma-distributed, SEIR-P exposed and infectious lifetimes
- $p(R_0 \,|\, \mathbf{I}, \mathbf{R})$ ($R_0 = \frac{\beta N}{\gamma}$, the basic reproductive ratio for the SEIR-P process, according to the survival function formulation (see, e.g. [HD96, HSW05]) with $\hat{R_0} = \frac{\hat{\beta} N}{\hat{\gamma}}$.

The position of $\hat{\beta}, \hat{\delta}$ and $\hat{\gamma}$ close to the centre of the parameter posterior distribution in the short-lived pathogen case (Figure 2.5c) means that a SEIR process with direct transmission rate $\hat{\beta} = \alpha\frac{\epsilon}{\rho} + \beta$ and (exponentially-distributed) E and I lifetimes with means that match those of the underlying process is the most likely model from among the class of SEIR models with exponential lifetimes. As a result, the fitted model produces a very good estimate of R_0 in relation to the underlying SEIR-P process.

As the pathogen lifetime increases in Figures 2.5a and 2.5d, leading to a lesser degree of timescale separation, the fitted models both underestimate R_0 and $\hat{\beta}, \hat{\delta}$ and $\hat{\gamma}$ lie further from the centre of the parameter posterior distribution (in the tails in the long-lived pathogen case).

2.3.2.2 Assessing DTA model fit against ET outbreak

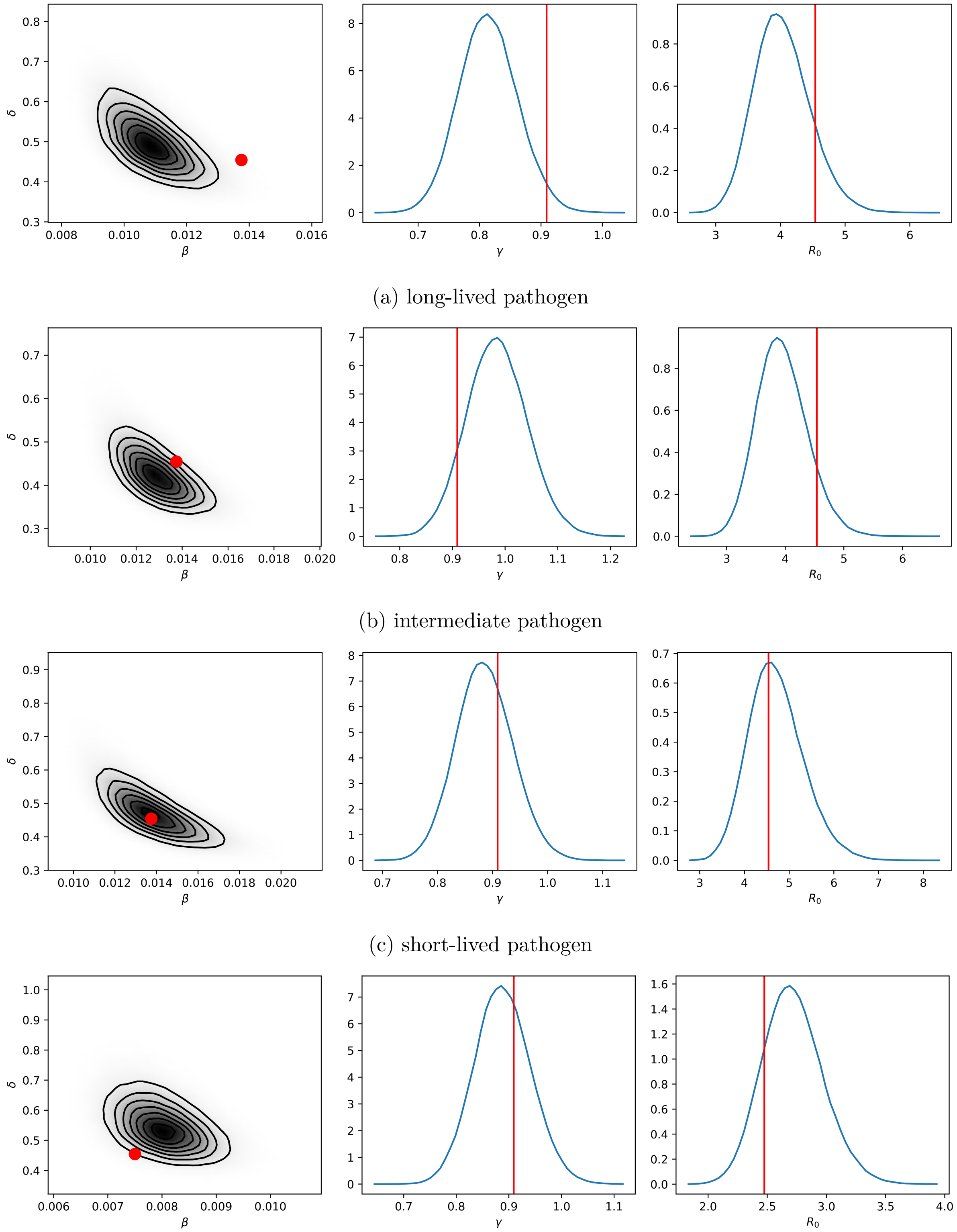
First, we compare the outbreak size trajectories predicted by the fitted model with the trajectory obtained from the data. These are obtained by simulating SEIR event times with parameter values drawn from the posterior samples, as discussed in Section 2.2. If the model fits well then we should expect these posterior-predicted trajectories to look similar to the observed trajectory [GCSR95, GS13]. Figure 2.6 graphically compares the posterior-predictive outbreak size trajectories for the four cases described above with the underlying observed outbreak trajectory. The outbreak size trajectory from the data is indicated by the solid red line and superposed on this is a graphical summary of posterior predictive outbreak size trajectories. The solid blue line indicates the median predicted number of infectious hosts, while the blue dashed lines are the 5th, 25th, 75th and 95th percentiles. In the cases of intermediate and short-lived pathogen, the predicted model output appears to agree well with the data, as is the case with DT only. However, for the long-lived pathogen, the model appears to predict outbreaks that reach their peak and begin to recede sooner than was actually observed in the data. Nonetheless, the fitted model does agree with the data in terms of peak outbreak size.

Figure 2.7 compares graphically the final outbreak size and the outbreak duration associated with each of the four sets of onset of infectivity and removal times with the same quantities drawn from their respective posterior-predictive distributions. Figure 2.8 is similar, comparing size and time of outbreak peak (i.e. the outbreak at its largest) (see Section 2.2 for details).

As is indicated clearly in the plots, for the long-lived pathogen case, the model makes a poor prediction of when the outbreak peaks and the how long the outbreak persists before finally dying out. Better agreement is evident for the intermediate and short-lived pathogen, more so, in fact, than is evident in the DT only case.

Data generating process	parameter values	N	m
long-lived pathogen	$\alpha = 0.001$, $\beta = 0.007$ $I_j - E_j \sim \text{Gamma}(1.10, 0.5)$ $R_j - I_j \sim \text{Gamma}(1.10, 1.0)$ $\epsilon = 5.4$, $\rho = 0.8$	300	295
intermediate pathogen	$\alpha = 0.001$, $\beta = 0.007$ $I_j - E_j \sim \text{Gamma}(1.10, 0.5)$ $R_j - I_j \sim \text{Gamma}(1.10, 1.0)$ $\epsilon = 54.0$, $\rho = 8.0$	300	294
short-lived pathogen	$\alpha = 0.001$, $\beta = 0.007$ $I_j - E_j \sim \text{Gamma}(1.10, 0.5)$ $R_j - I_j \sim \text{Gamma}(1.10, 1.0)$ $\epsilon = 5.4 \times 10^4$, $\rho = 0.8 \times 10^4$	300	297
direct transmission only	$\alpha = 0.0$, $\beta = 0.0075$ $I_j - E_j \sim \text{Gamma}(1.10, 0.5)$ $R_j - I_j \sim \text{Gamma}(1.10, 1.0)$	300	272

Table 2.1: Scenarios for simulation study. In order to test the sensitivity of the Markovian SEIR model when fitted to non exponentially-distributed exposed and infectious lifetimes, these here are simulated with gamma-distributions. Total host population size = N. Final outbreak size = m.



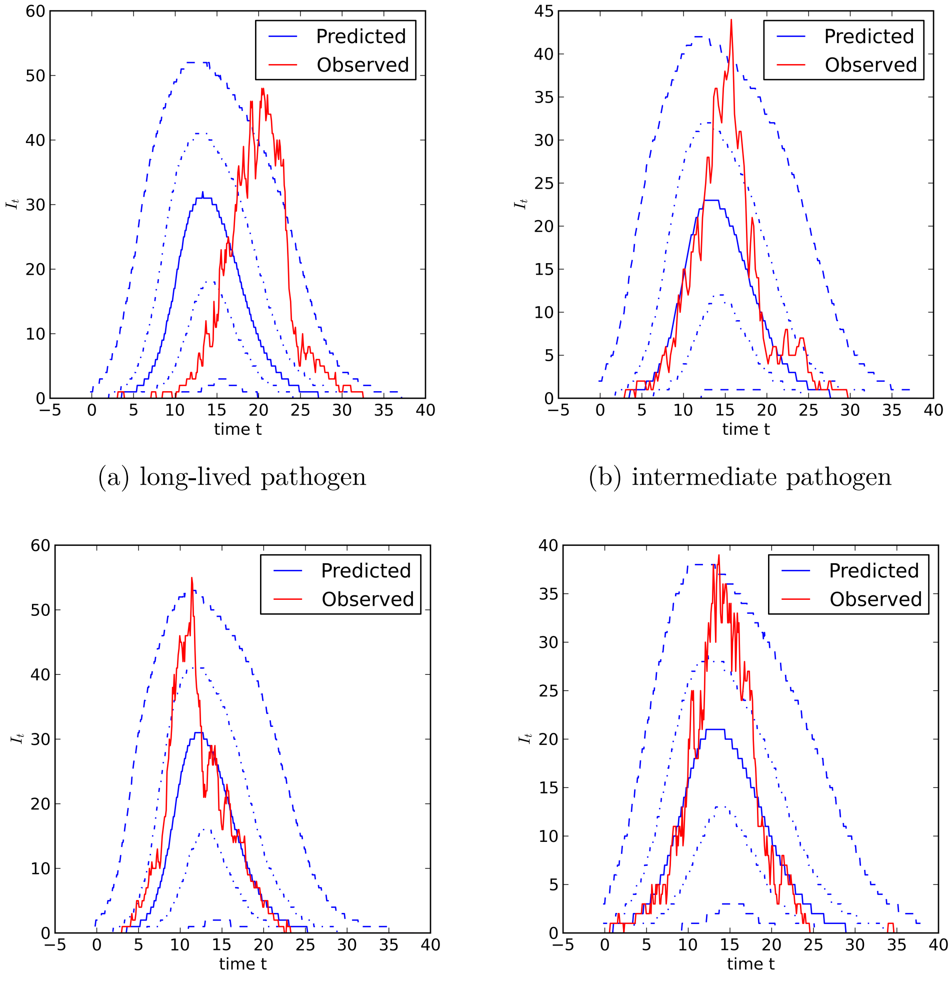
(a) long-lived pathogen

(b) intermediate pathogen

(c) short-lived pathogen

(d) direct transmission only

Figure 2.5: **Density estimates of SEIR parameter posterior distributions and R_0.** The red dot in the leftmost panels indicates $(\hat{\beta}, \hat{\delta})$, where $\hat{\beta} = \alpha\epsilon/\rho + \beta$ and $\hat{\delta} = (\mathbb{E}(I_j - E_j))^{-1}$. The red vertical lines in the central and rightmost panels indicate $\hat{\gamma} = (\mathbb{E}(R_j - I_j))^{-1}$ and $\hat{R}_0 = \frac{\hat{\beta}N}{\hat{\gamma}}$, respectively (see main text for more detail). The marginal posterior distributions γ and (β, δ) are conditionally independent, given the data, and so are plotted separately.



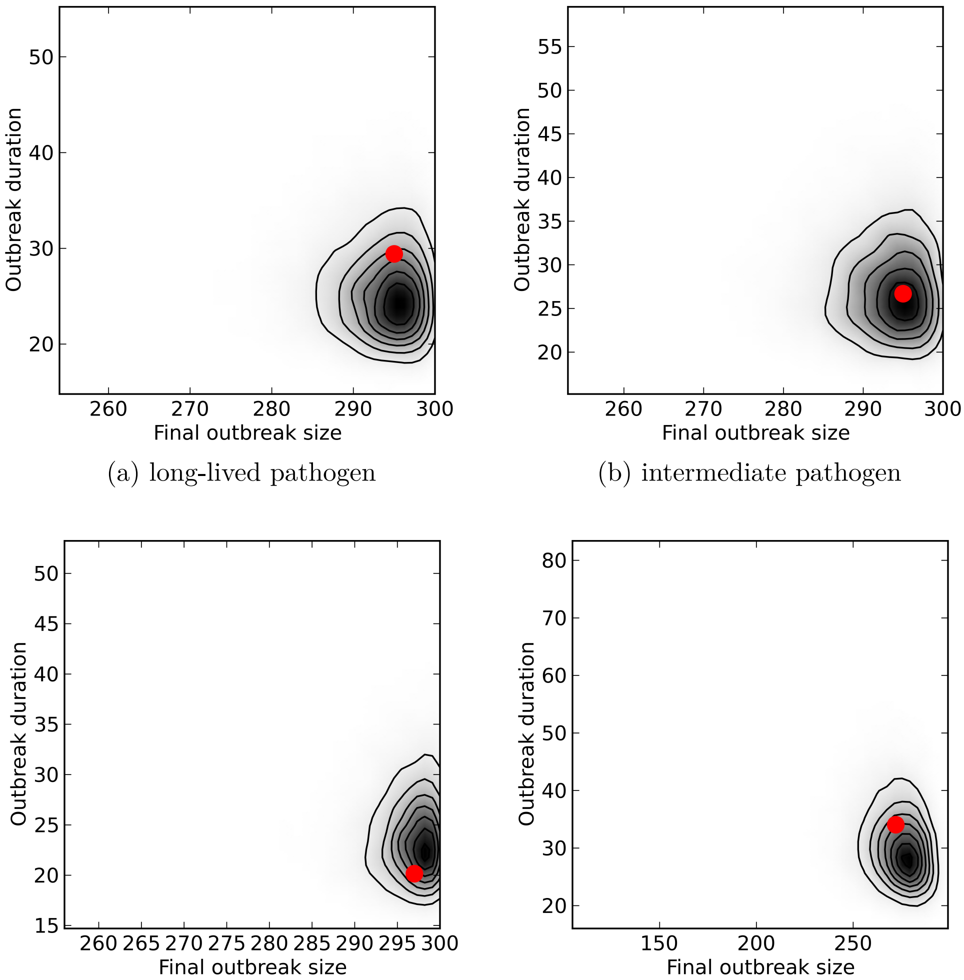
(a) long-lived pathogen

(b) intermediate pathogen

(c) short-lived pathogen

(d) direct transmission only

Figure 2.6: **Observed outbreak size trajectories,** I_t (solid red line): long-lived (a), intermediate (b) and short-lived pathogen (c) and direct transmission only (d). These are compared with the trajectories obtained from SEIR model with 15 000 parameters values taken uniformly from the MCMC sample chains obtained while fitting the SEIR model, with *small* outbreaks (≤ 50) discarded. The time axis was discretised (400 points) and 5$^{\text{th}}$, 25$^{\text{th}}$, 50$^{\text{th}}$ (median), 75$^{\text{th}}$ and 95$^{\text{th}}$ percentiles of the SEIR-predicted outbreak size were estimated at each discrete time point. The solid blue line indicates the median outbreak size while the dashed blue lines are the other percentiles. In the short-lived and direct transmission only cases, the shape of the predicted outbreak size trajectories (as indicated by the blue lines) mirrors that of the observed outbreak size, with the solid red and blue lines aligning at the initial exponential growth phase, as well as at the end of the outbreak when the number of infective hosts dies out. This is not the case for the long-lived pathogen case, for which the model predicts earlier onset of growth of the outbreak, peaking somewhat earlier than was observed.



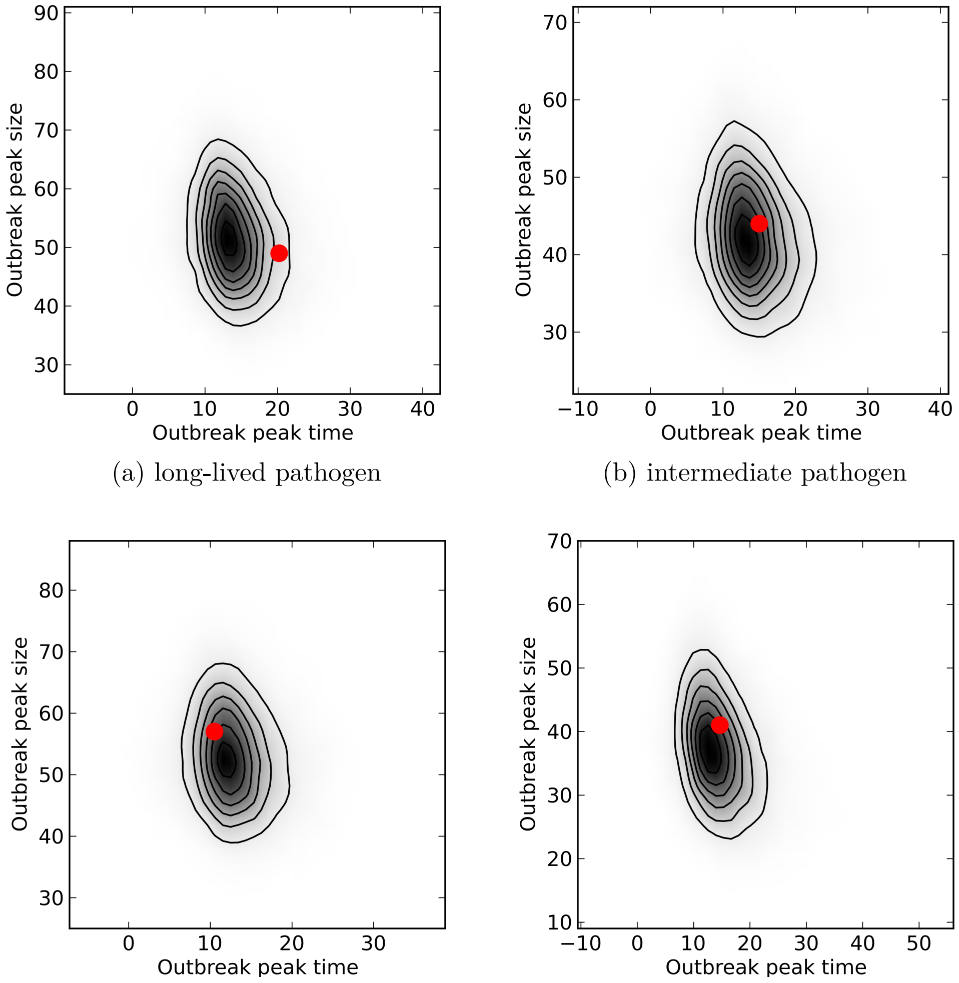
(a) long-lived pathogen

(b) intermediate pathogen

(c) short-lived pathogen

(d) direct transmission only

Figure 2.7: **Graphical comparisons of final outbreak size (total hosts infected during outbreak) vs. outbreak duration (latest removal time minus time of first onset of infectivity) with their posterior predictive distributions**. The red dot indicates observed value of statistics from one outbreak: long-lived (a), intermediate (b) and short-lived pathogen (c) and direct transmission only (d). The shading and contours were obtained from a kernel density estimate after simulating 15000 SEIR outbreak trajectories with parameter values taken at uniform intervals from the MCMC chains obtained while fitting the SEIR model, with *small* outbreaks (≤ 50) discarded. In the case of long-lived pathogen, the fitted model tends to predict shorter duration outbreaks but otherwise agrees with the data in terms of final outbreak size. This is indicated by the red dot aligning horizontally with the darkest part of the density estimate but being shifted vertically. Better agreement between the data and fitted model is evident in the short-lived and intermediate pathogen and DT-only cases.



(a) long-lived pathogen

(b) intermediate pathogen

(c) short-lived pathogen

(d) direct transmission only

Figure 2.8: **Graphical comparisons of size of outbreak peak, i.e. the size of I_t at its largest, and time of outbreak peak, as defined in main body of text with their posterior predictive distributions**. The red dot indicates observed value of statistics from one outbreak: long-lived (a), intermediate (b) and short-lived pathogen (c) and direct transmission only (d). The shading and contours were obtained from a kernel density estimate after simulating 15000 SEIR outbreak trajectories with parameter values taken uniformly from the MCMC sample chains obtained while fitting the SEIR model, with *small* outbreaks (≤ 50) discarded. For long-lived pathogen, the fitted SEIR model predicts that outbreaks peak, on average, at the size observed in the data. However, the model predicts outbreaks that peak *earlier*. This is also evident in Figure 2.6a, where the predicted outbreak size trajectories clearly peak earlier than the observed outbreak size trajectory indicated by the solid red line. Better agreement between data and model predictions are visible in Figures 2.8c and 2.8d.

2.4 Case study: white spot disease in penaeid shrimp

In this section we focus on white spot disease (WSD) in penaeid shrimp, since this is a key example of an infectious disease transmitted via both DT and ET due to a pathogen known to be long-lived in the environment, under the right conditions. We formulate a SEIR-P model of the WSD-penaeid system in which attempts are made at regular intervals to remove dead and diseased hosts from the system. We estimate its parameters from published data and then, using simulations, we compare the effect upon the outbreak trajectories resulting from stepping up the frequency of removals from every 24 h to every 6 h.

This increase in the frequency of removals is an example of how relative shortening of the SEIR-P timescales might come about in practice. We see that with removals every 24 h the SEIR-P host-disease dynamics are closely replicated by its direct transmission approximation (DTA), whereas SEIR-P and DTA visibly diverge in their averaged outputs with removals every 6 h.

2.4.1 Background: WSD and its impact

WSD is a viral disease caused by white spot syndrome virus (WSSV) that affects penaeid shrimp, including the species Asian tiger shrimp, *Penaeus monodon* (PM) and whiteleg shrimp, *Litopenaeus vannamei* (LV). These and similar species are farmed extensively in South East Asia and Central America [Rön01] and consumed worldwide. By time of publication, the World Bank Group estimates in its report of 2014 [Ban14] that WSD had caused economic losses exceeding 8 billion US Dollars globally to the commercial culture of penaeid shrimp since the first documented emergence of the disease in 1992 in Taiwan. Within three years, the disease had spread throughout Southeast Asia, reaching South Texas by 1995 and central America by 1999 [SP10, Ban14]. The rapidity and extent to which WSD infected PM, the chief source of brood stock, was a factor in the move to the culture of alternative species of penaeid, including LV [Ban14]. WSD is spread between ponds and canals by human activity, by movement of wild species [SP10] and, potentially, even the action of seabirds [VNL04]. The virus' wide host range is one key factor in the extent of the devastation caused by the disease, with hosts and carriers of the disease known to include 93 species of arthropod, various species of phytoplankton and zooplankton (ranging from 0.2 to 200 micrometers), micro-algae, polychaetes and rotifers [SP10, Ban14, ELEBCH$^+$09]. WSSV is transmitted directly as a result of ingestion of infected material and indirectly due to waterborne virus within algae and plankton or that is free-floating [ELEBCH$^+$09]). There is evidence of long-term persistence of the virus in ponds where WSD outbreaks have occurred [QHDA09].

Under laboratory conditions, WSSV has been shown to retain its infectivity in sea water for up 12 days and in sun dried and water-logged pond sediment for up to 19 and 35 days, respectively [SAR$^+$13]. Through periodic sampling of seawater from abandoned shrimp culture ponds and surrounding canals in Vietnam, where previously an outbreak of WSD had led to 100% mortality of cultivated shrimp, the authors of [QHDA09] found that WSSV remained detectable for up to 20 months. The detection rates declined throughout the duration of the study with the steepest declines observed between July and December of both 2001 and 2002. The authors suggest this is linked to decreased plankton biomass during that period, which in turn suggests that WSSV is able to replicate within certain plankton species. Esparza-Leal, et.al. [ELEBCH+09] suggest that free-floating WSSV virions have the potential to infect shrimp in pond water at around 27°C, whereas pond water temperatures in range of 30-33°C prohibit infection. In [ELEBCH+09] it was noted that detectability of WSSV varied among pond water samples taken simultaneously from the same pond, leading the authors to note a degree of stochasticity in relation to the waterborne pathogen load.

2.4.2 Modelling WSD and estimation of SEIR-P parameters

Estimates of rates of WSD transmission via ingestion and cohabitation have been made by Lotz and Soto [SL01] for LV and by Tuyen, et al [TVVdJ14] for both LV and PM. Each of these estimates rely on the assumption that the force of infection is proportional to the number of infected shrimp currently present in the tank (I_t) and therefore responds immediately to changes in this number.

Tuyen, et al, decompose the overall rate of transmission into two components relating to ingestion and cohabitation, $\beta = \beta_{\text{ingest}} + \beta_{\text{cohab}}$ (our notation), so that the force of infection at time t is $\frac{\beta I_t}{N}$ (where they assume that transmission is frequency dependent). They estimate the two components of β using regression analysis of data obtained via an immersion challenge experiment where the relative amounts of exposure via the two routes are controlled. This is done for LV and PM independently and for both species combined.

Lotz and Soto expose LV to WSSV exclusively via either ingestion or cohabitation for a set duration and estimate $\tilde{\beta}_{\text{ingest}}$ and $\tilde{\beta}_{\text{cohab}}$ from the numbers of shrimp that later developed the disease. Using a Reed-Frost model of epidemics (e.g. see [Abb52]), the latter two quantities are probabilities of disease transmission per distinct susceptible-infected shrimp pair during a time interval of duration Δt. The force of infection here is approximately $\frac{\tilde{\beta} I_t}{\Delta t}$ (see Appendix C), where $\tilde{\beta}$ is either $\tilde{\beta}_{\text{ingest}}$ or $\tilde{\beta}_{\text{cohab}}$. Lotz and Soto found $\tilde{\beta}_{\text{cohab}}$ to be not significantly different from zero in a first experiment and to be over an order of magnitude smaller than $\tilde{\beta}_{\text{ingest}}$ when the experiment was repeated ($\tilde{\beta}_{\text{ingest}} = 0.56, \tilde{\beta}_{\text{cohab}} = 0.02$). Such a relatively low rate of transmission due to cohabitation led the authors to omit this from their model of WSD in LM described in [LS02]. Tuyen, et al, found a similar result in the case of PM ($\beta_{\text{ingest}} = 0.22\,\text{h}^{-1}, \beta_{\text{cohab}} = 0.0026\,\text{h}^{-1}$) but for LV they in fact found that the reverse was true, in that the rate of transmission via cohabitation was greater than via ingestion ($\beta_{\text{ingest}} = 0.0038\,\text{h}^{-1}, \beta_{\text{cohab}} = 0.018\,\text{h}^{-1}$).

Underlying the estimates of β_{cohab} and $\tilde{\beta}_{\text{cohab}}$ above is the assumption that the force of infection responds without delay to a change in the number of infected shrimp, I_t. This assumption is indeed valid across a wide variety of cases in which the size of the environmental pathogen load responds more or less rapidly to changes in I_t, as discussed in the previous section. However, given the slow rate of decay of infectivity of WSSV and its persistence in water bodies long after outbreaks have occurred, it is perhaps fruitful to consider the relationship between environmental pathogen load and the rate of environmental transmission as described, for example, by the SEIR-P model described in Section 2.2. For example, Wang, et al. (2012) [WJL+12] find that an environmental transmission model similar to SEIR-P of avian flu among duck populations was able to account for the complex periodic outbreak patterns of the disease over long time periods.

As far as we know, there is no published estimate of the environmental transmission rate, α, for WSD among penaeids. Here, however, we obtain a lower estimate of $\alpha = 10^{-4}\,\mathrm{ml\,part^{-1}\,h^{-1}}$ for PM, along with an upper estimate of the pathogen decay rate, $\rho = 0.005\,\mathrm{h}^{-1}$, from the results of the WSSV viability in seawater experiment by Kumar, et al, ([SAR+13], details in Appendix D). Lotz and Soto ensure use a shrimp density of 12 animals per square metre of water surface in their experiment in order to mimic densities of wild populations [SL01]. In the simulation study, described below, we adopt a nominal water volume of $46.2\,\mathrm{m}^3$ and water surface area of $77\,\mathrm{m}^2$ to obtain a similar host density with 1000 shrimps initially in the system. Since the estimates of α and β both have dimensions $\texttt{volume} \times \texttt{time}^{-1}$, we scale them by this nominal volume before carrying out the simulations.

The pathogen emission rate, ϵ, is chosen from within a range of known shedding rates for waterborne viruses (e.g. see [WSKK17, KF12]). Since each dead shrimp contributes $\frac{\epsilon}{\rho}$ to the environmental pathogen load, at equilibrium, and therefore $\alpha\frac{\epsilon}{\rho}$ to the force of infection via environmental transmission, we choose a direct transmission rate $\beta = 10 \times \alpha\,\frac{\epsilon}{\rho}$. This is in accord with the relative sizes of Lotz and Soto's estimates of $\tilde{\beta}_{\text{ingest}} \approx 10 \times \tilde{\beta}_{\text{cohab}}$.

The S, E, I and R compartments of the SEIR-P model (summarised in Figure 2.10) are, respectively, shrimp that are susceptible (S), have been exposed, but still alive (E), dead, and now causing new infections either via shedding virus into the water body due to decay or being scavenged upon (I) and physically removed from the system (R). We assume for simplicity that there is no viral shedding or infectivity during the E stage and that times from exposure to mortality are gamma-distributed with shape and rate parameters ν_δ and δ such that the mean time from exposure to mortality ($\frac{\nu_\delta}{\delta}$) agrees with the estimate given by [TVVdJ14].

There are two processes by which shrimp are removed from the system. Firstly, there is the long process of decay characterised by gamma-distributed times in the I compartment, with shape and rate parameters ν_γ and γ with mean $\frac{\nu_\gamma}{\gamma} = 333.3\,\mathrm{h} \approx 14\,\mathrm{d}$. Secondly, removals occur probabilistically from both the E and I compartments at regularly spaced time points with probabilities of success of $\pi_E = 0.05$ and $\pi_I = 0.95$, respectively, so that $\mathrm{bin}(E_t, \pi_E)$ and $\mathrm{bin}(I_t, \pi_I)$ shrimp are removed at each removal point, t. Shrimp that are dead are therefore removed at the first removal time, post-mortem, with probability 0.95 and at the second with probability 0.9975. This means that it is highly unlikely that shrimp are removed from the system due to natural decay in this scenario. We assume that no living, disease-free shrimp are accidentally removed in this process. All of these model quantities are summarised together in Table 2.2.

2.4.3 Impact of removal frequency on WSD outbreaks among penaeids

Using simulations, we study outbreak patterns under 24 and 6-hourly removals under the SEIR-P model described above. Alongside these we also look at those of the DTA of this model, where $\hat{\beta} = \alpha\frac{\epsilon}{\rho} + \beta$, for comparison. Density plots for the final outbreak size and outbreak duration

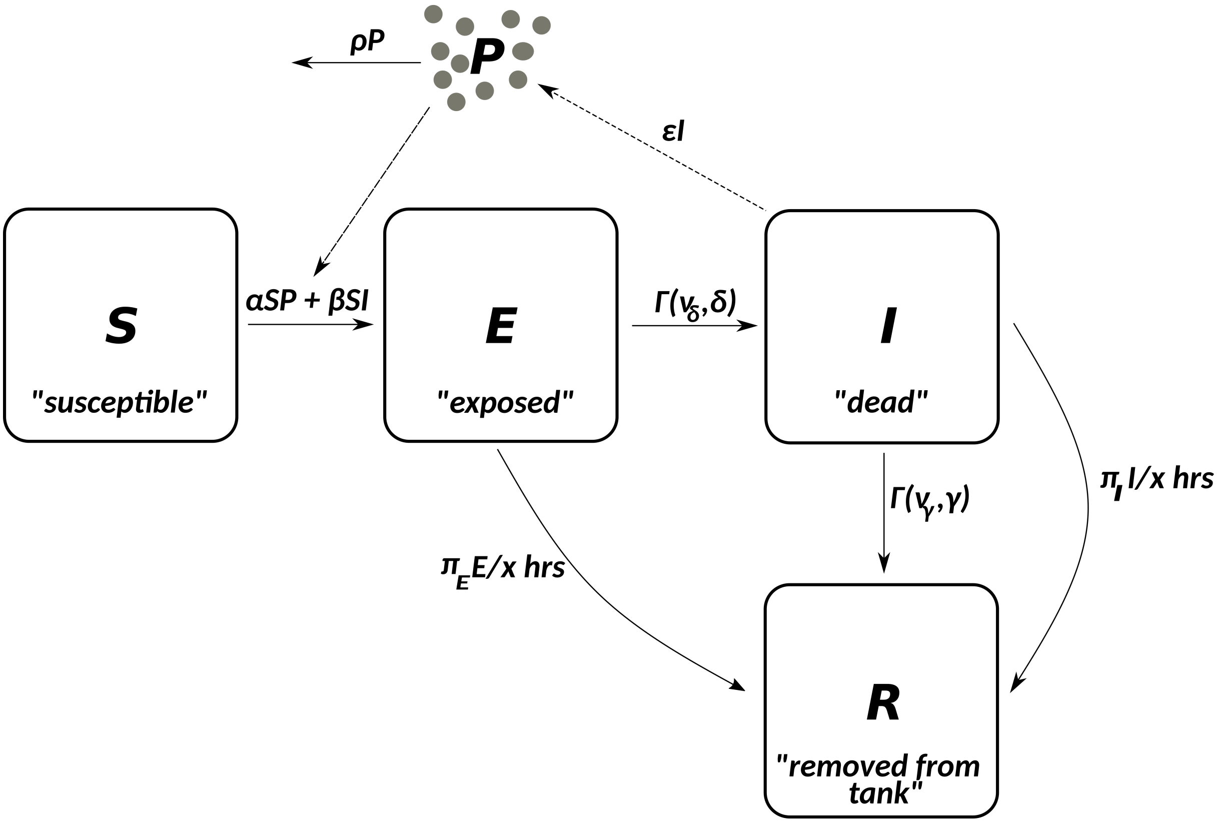


Figure 2.9: **Summary of SEIR-P model** of WSD among penaeids. Parameter values are listed in Table 2.2. The arrow from I to R labelled $\Gamma(\nu_\gamma, \gamma)$ represents removal of dead hosts after a gamma-distributed time to full natural decay. The curved arrow from I to R represents removal at one of the x-hourly removal attempts, with probability π_I, similarly for the curved arrow from E to I.

$(R_{\max} - E_0)$ of the SEIR-P model and the DTA are displayed in Figures 2.11 and 2.12 for 24 and 6 hourly removals, respectively. Figures 2.13 and 2.14 show the simulated outbreak trajectories for the four host compartments of the SEIR-P model and DTA. The top row in these two figures are typical individual outbreak trajectories while the bottom row are trajectories averaged over 3×10^4 independent simulations, with "small" outbreaks of fewer than 10 infections.

Figures 2.11 and 2.13 show that attempting to remove the dead shrimp from system at 24 h intervals, even with a success rate of 95%, is not sufficient to prevent large outbreaks of WSD, with outbreaks overwhelmingly affecting more than 90% of the shrimp population and lasting more than 600 h from index exposure to final removal. Increasing the intensity of surveillance, however, by removing dead and diseased shrimp every 6 h, eliminates such large outbreaks, limiting the final outbreak size to about 60% of the population. Additionally, outbreaks tend to be eradicted sooner at around 200 h, although a sizeable proportion continue for longer (see Figure 2.12).

It is interesting to compare the SEIR-P and DTA trajectories when going from 24 to 6-hourly removals, since the time that infectious shrimp are in the system is reduced by about a quarter, on average. This is an example of how two distinct degrees of host and pathogen timescale separation may be observed for the same host-disease system. The plots in Figures 2.11 and 2.13 suggest very close alignment between the SEIR-P and DTA models in their host compartment

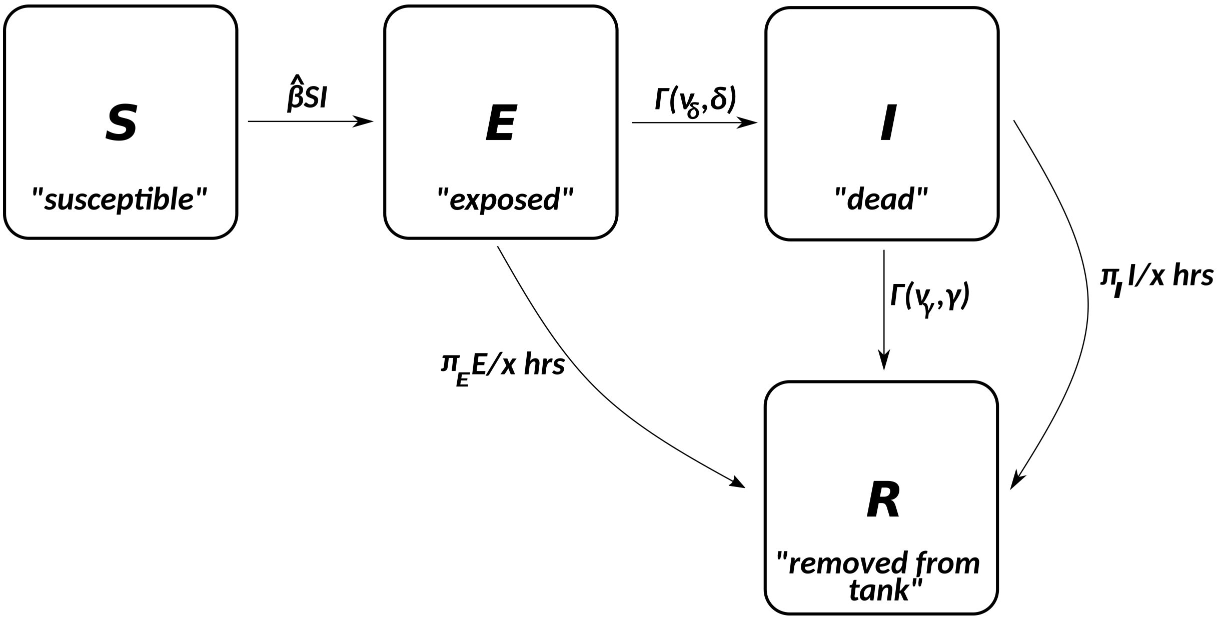


Figure 2.10: **Summary of DTA model** of WSD corresponding to the SEIR-P model in Figure 2.10 (see Section 2.3). This is the direct transmission, SEIR approximation, with direct transmission rate $\hat{\beta} = \alpha\frac{\epsilon}{\rho} + \beta$. All other aspects are as for the SEIR-P model.

dynamics, final outbreak sizes and outbreak durations, meaning that we can faithfully reproduce the environmental transmission without needing to model the pathogen load. Under such a scenario, most hosts remain infectious for less than 24 h. Nonetheless, we see that the resulting outbreak patterns are captured as equally well by the DTM as by the full SEIR-P model. The shortening of the host timescale by removing every 6 h is sufficient, however, to begin to observe divergence between the SEIR-P and the DTA, most noticeably, perhaps, in the distributions of the final outbreak size and outbreak duration (Figure 2.12). Indeed, the DTA underestimates, on average, the reduction in final outbreak size and overestimates the reduction in outbreak duration. The DTA outbreaks at 6-hourly removals tend to grow slightly faster than the SEIR-P outbreaks (see Figure 2.14).

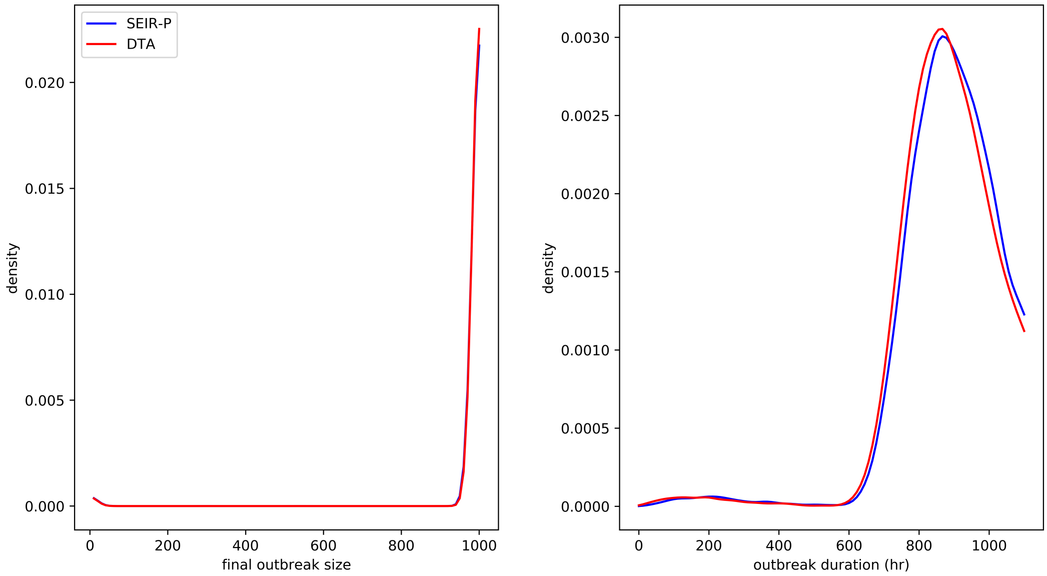


Figure 2.11: **Estimated density plots** of final outbreak size (left) and outbreak duration (right) for the SEIR-P (blue) and DTA (red) models of WSD with **24-hourly removals**. Both quantities are distributed very similarly under the two models.

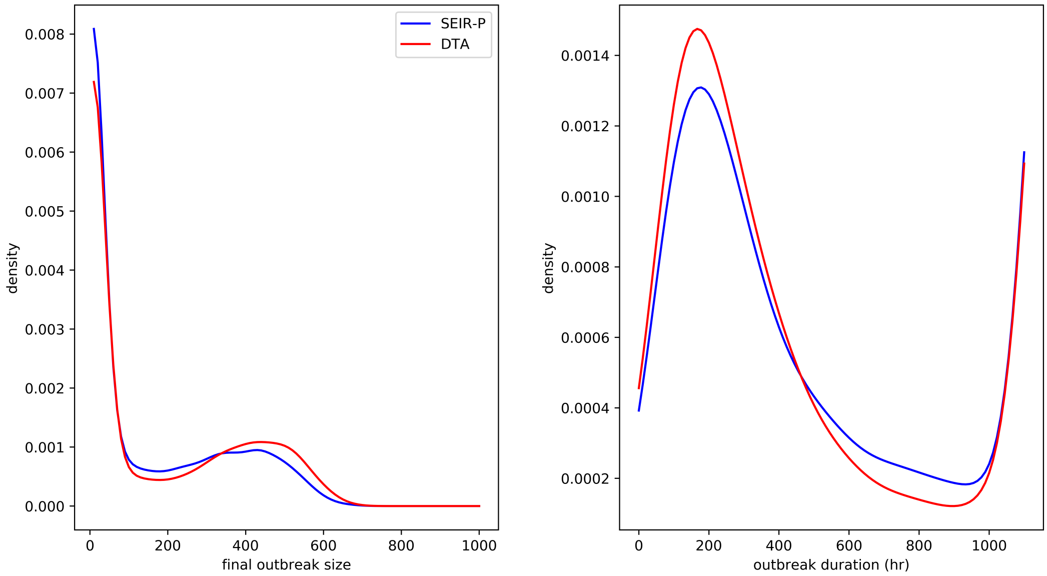
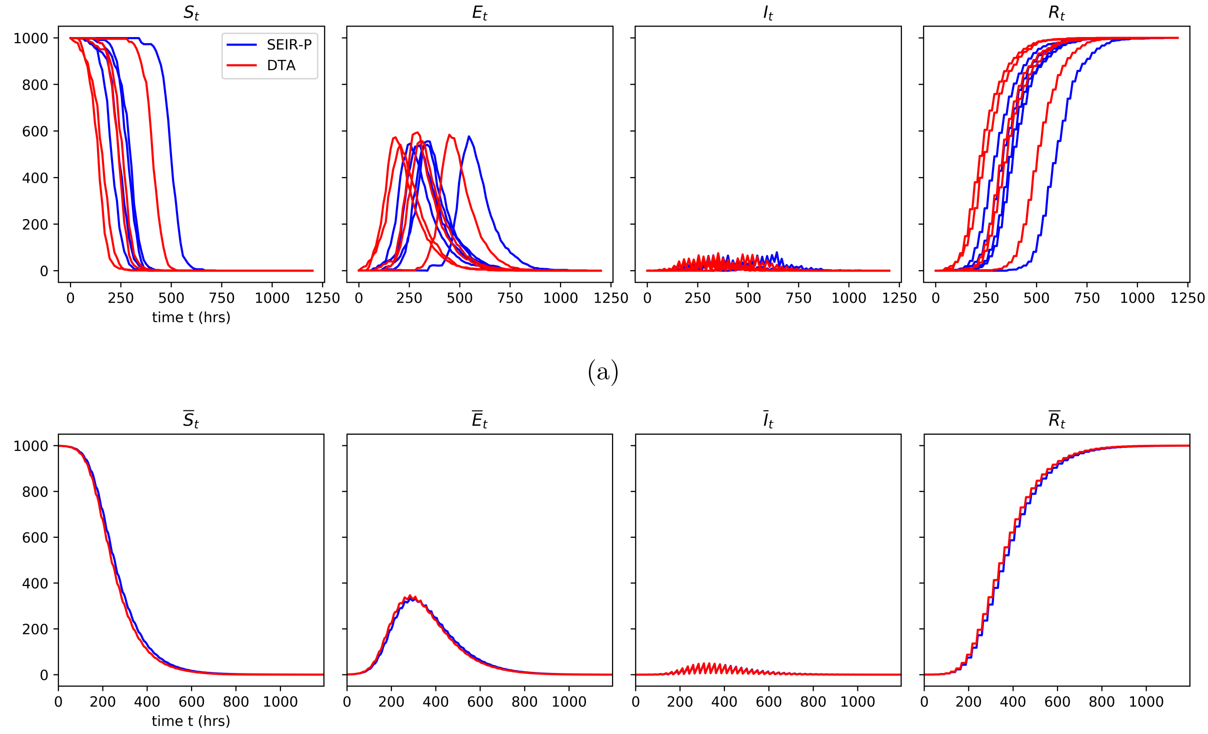


Figure 2.12: **Estimated density plots** of final outbreak size (left) and outbreak duration (right) for the SEIR-P (blue) and DTA (red) models of WSD with **6-hourly removals**. The benefit of increasing the removal frequency, in terms of reduction in mean final outbreak size, is underestimated slightly by the DTA and the reduction in outbreak duration is over-estimated.

	description		source / comment
α	transmission (cohabitation)	$10^{-4}\,\mathrm{ml\,part^{-1}\,h^{-1}}$ $2.16 \times 10^{-12}\,\mathrm{part^{-1}\,h^{-1}}$	estimated from [SAR+13] (Appendix D) scaled by $46.2\,\mathrm{m}^3$
β	transmission (ingestion)	$8.64 \times 10^{-4}\,\mathrm{shrimp^{-1}\,h^{-1}}$	$10 \times \alpha\frac{\epsilon}{\rho}$ (see [SL01] and above discussion)
$\hat{\beta}$	direct transmission (DTA)	$9.5 \times 10^{-4}\,\mathrm{shrimp^{-1}\,h^{-1}}$	$\alpha\frac{\epsilon}{\rho} + \beta$
ν_δ	mortality (shape)	1.5	
δ	mortality (rate)	$0.0112\,\mathrm{h^{-1}}$	[TVVdJ14]
ν_γ	removal (decay) (shape)	2.0	
γ	removal (decay) (rate)	$0.006\,\mathrm{h^{-1}}$	
π_E	success of removal (from E)	0.05	
π_I	success of removal (from I)	0.95	
ϵ	WSSV shedding	$2 \times 10^5\,\mathrm{part\,shrimp^{-1}\,h^{-1}}$	(see e.g. [WSKK17, KF12])
ρ	loss of WSSV infectivity	$0.005\,\mathrm{h^{-1}}$	estimated from [SAR+13] (see App. D)

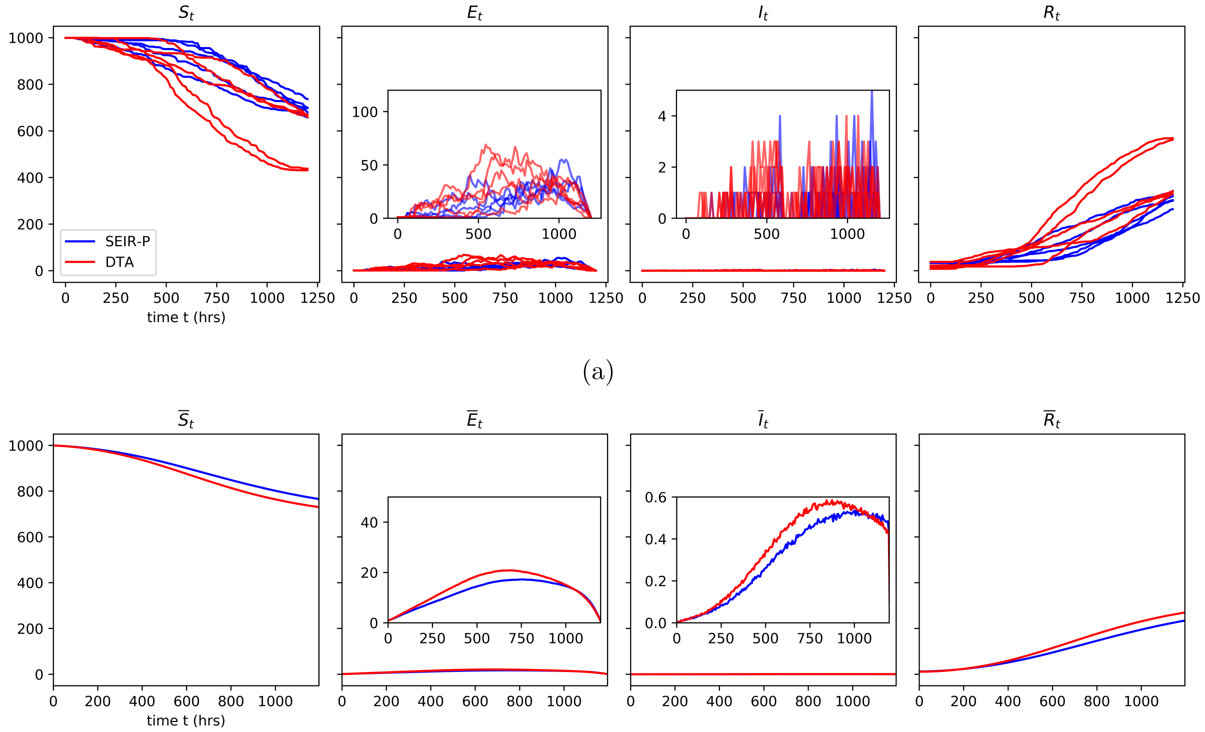
Table 2.2: **Parameter estimates and sources** for SEIR-P model of WSSV in penaeids.



(a)

(b)

Figure 2.13: **Simulations** of the SEIR-P (blue) and DTA (red) models of WSD in penaeid shrimp with removals of exposed (E) and dead (I) hosts at 24-hourly intervals, with probabilities of success 0.05 and 0.95, respectively. Single outbreak trajectories (2.13a) and averages over 30 000 independent simulations with small outbreaks (fewer than 10 infections) excluded (2.13b). The zig-zag pattern in the 3$^{\text{rd}}$ panel of (2.13b) is due to the periodic removals. The averaged model outputs show a high degree of similarity between SEIR-P and DTA, meaning that at these timescales the environmental transmission of WSD can be well approximated with direct transmission among the hosts.



(a)

(b)

Figure 2.14: **Simulations** of the SEIR-P (blue) and DTA (red), as in Figure 2.13, with removals at 6-hourly intervals. Although the outbreaks of single SEIR-P and DTA trajectories appear similar, a small but definite divergence between the two models appears when studying their averaged outputs.

2.5 Discussion

We have seen in Section 2.3 that the SIR and SEIR models approximate the host-disease dynamics arising from a combination of direct and environmental transmission, as modelled by SIR-P or SEIR-P, that this approximation improves with increasing rates of pathogen shedding and decay and that when fitting these models to data, using Bayesian inference and data augmentation, they are highly robust to violations of the assumption of direct transmission. For example, these results suggest that the direct transmission approximation will be suitable for modelling transmission of SARS-CoV-2 within a closed environment, such as a hospital, since it has a half life of about 1 h in aerosols and 1 h, 3.5 h, 5.75 h and 7 h on copper, cardboard, stainless steel and plastic surfaces [vDBM+20] but the mean infectious period is considerably longer: 5 d to 11 d for asymptomatic cases, up to 4 d for presymptomatic cases [BMC+20] and about 7 d for symptomatic cases [WCG+20].

Tien & Earn, in their investigation of multiple transmission routes of cholera among humans [TE10], cite the rate of pathogen decay in the waterbody as the important factor in determining whether one should consider modelling the environmental and direct routes of transmission separately, or combined as one direct route. As suggested by the simulation study in Section 2.4, a viral lifetime of 200 h versus a much shorter host mean infectious lifetime of around 24 h also results in disease dynamics closely reproducible with a DTM, in spite of the low rate of pathogen decay. In this case the high rate of pathogen shedding produces sufficient host-pathogen timescale separation in order that DT provides a good approximate description of the transmission via both direct and environmental routes. When the rates of shedding and pathogen decay are both low, as in Figure 2.4a and the long-lived pathogen of Section 2.3.2, then we do not expect the DT approximation to work. Macro-parasite infections are one class of disease system within this grouping and our results indicate why models describing the complex host-parasite interaction, similar to those of similar to that of Anderson & May [AM78, MA78], are often used for these systems (e.g. see [MNH+12]).

While individual outbreak trajectories appear very similar, statistical comparison over many runs reveals a strong and practically important divergence between environmental and direct transmission models of WSD among penaeids under more effective disease control (i.e. more frequent removals). The model fitting and checking in Section 2.3 was done under the rather special scenario that both the times of onset of infectivity and host removal are known so as to construct the outbreak size trajectories displayed in Figure 2.6, which provides the clearest indication of lack of model fit in the long-lived pathogen case. However, DTMs can and have been fitted to a wide range of partial epidemic data [NR05, GMP17] and our conclusions will hold in such scenarios. Nonetheless GPPCs in general may not be the sharpest way to detect departure from direct transmission when the data from an outbreak is less complete, as is often the case. *Exposure time residuals* (ETRs) (Lau, et al [LMSG14]) could potentially yield a numerical measure of model fit. ETRs are defined, relative to some putative model, as joint functions of the data, latent variables and parameters and their joint posterior predictive density should approximate an independent uniform sample when the parameters are close to the mode of their posterior. However, in the case of there being latent variables, such as when the host event times are not fully observed, analysis of their high-dimensional posterior distribution is not straightforward. A second numerical method of model checking is the use of tail-area probabilities or posterior-predictive p-values [GCS^{+}13, GST18], which offer a measure of the probability of observing certain aspects of an outbreak (such as final outbreak size) that are at least as extreme as what was observed. However, as noted by Gibson (2018), use of low-dimensional summaries of complex high-dimensional processes may impact upon the power of model tests based upon poster-predictive distributions to correctly identify model deficiencies but such impacts can be mitigated by using several low-dimensional summaries, as we did here.

Common methods of model comparison, such as model evidence [PM18] and the Bayesian and the deviance information criteria (BIC & DIC) would require an alternative ETM fitted to the same data in order to make a comparison. It is an open question whether ETMs can be fitted to host-disease events without measurement of the environmental pathogen load or strong prior information about the pathogen shedding and decay rates. Tien and Earn [TE10] comment that even when pathogen dynamics are slow, parameters quantifying rates of environmental and direct transmission (α and β) are still unidentifiable from disease incidence data alone. Methods that measure pathogen density in the waterbody, such as polymerase chain reaction [RCLMJ^{+}15, AWS98] are therefore required in order to quantify environmental transmission from data.

Statistical modelling and prediction are tools, and most prominent among these are direct transmission compartmental models within the SIR framework, are becoming increasingly important in the control and understanding of infectious disease, with their simple, yet powerful, picture of disease transmission. It is a common wisdom that a simple model is often preferable over a more complex one. Nonetheless, all models should be applied critically, in order that the conclusions and predictions we draw from them are sound. This work should reassure that direct transmission models retain their validity even as their field of application widens to include environmental tranmission via long-living pathogens. However, even this highly successful and widely applied assumption should be assessed critically as its validity depends not just on the host-pathogen system in question but also on the management regimes imposed upon it.

Chapter 3

Bayesian inference for direct transmission models as a tool to analyse and design challenge studies

3.1 Introduction

The purpose of this chapter is to demonstrate that analysis of data obtained via *infection challenge experiment* (ICE) using dynamic stochastic disease models enables estimation of key characteristics of disease dynamics that are not typically considered in such studies and can be used to improve their design. ICE are the means whereby infectious disease is observed under controlled conditions and among their numerous and varied objectives are the greater understanding of modes and mechanisms of transmission, such as the transmission of African swine fever by carrier pigs [EHW$^+$19] and transmission dynamics of Rift Valley fever virus among various types of livestock [KBP$^+$20]. Additionally, the purpose of ICE can be to quantify the minimum infectious dose of a pathogen associated with a transmission route, e.g. transmisison of foot and mouth disease to pigs via aerosols [ABD02] as well as to study behavioural factors in disease transmission, such as faecal avoidance in grazing cattle and its impact upon infection risk [SWMH09].

Efffectiveness of vaccines are typically first demonstrated using ICE, such as the study of the protection against tuberculosis (TB) in Eurasian badgers following intramuscular injection of BCG [LPGS$^+$11] and such studies often accompany wider field studies investigating the impacts of the same interventions on prevalence and transmission in the wild, such as [CCR$^+$12]. At the time of writing, the first human challenge study of SARS-CoV-2 in humans has received ethics approval [GOV, Kir20] and will examine the transmission mechanisms of the virus and test various candidate vaccines.

Because of the magnitude of the economic impact of disease upon aquaculture production (6 billion USD per annum, as estimated by the World Bank [Ban14]), ICE are widely used for research on diseases in fish and other aquatic animals. To state some examples published in the first two months of 2021, susceptibility to tenacibaculosis in conger eels [IA21] and the transmissibility of pilchard orthomyxovirus among Atlantic salmon via seawater [SRT+21] have been established using ICEs. Challenge protocols have been developed for *Streptococcus agalactiae* in Nile tilapia [HLLL21] and Piscirickettsia salmonis in various salmon species [LGJ21]. For ameobic gill disease in Atlantic salmon, the effectiveness of immersion in sodium percarbonate as a treatment [TSC+21] and presence of Nolandella sp. and Pseudoparamoeba sp. as aggravating factors [EBA+21] have been investigated using ICE. Recent examples of trials of vaccination regimes include immersion with live nervous necrosis virus (NNV) of sevenband grouper [KKO21] and a DNA vaccine against a salmonid alphavirus in Atlantic salmon [TWI+21]. The time variation of the rate of NNV viral shedding and the within-host viral load among sevenband grouper have been measured [KQKO21].

None of these studies made use of stochastic dynamic disease models in their analysis of data that were gathered, in spite of the ongoing development in the fields of dynamic modelling and statistical inference and their application to the problems of quantification of disease transmission, prediction and experimental design. Such models provide a dynamic, probabilistic, picture of the focal population passing between a small number of disease states [CHBCC09, BCCF19] and are the cornerstone of statistical analysis of disease spread. The continous time Susceptible-Infectious-Removed (SIR) model of Kermack & McKendrick [KM27] and the discrete time Reed-Frost model [Abb52] being canonical examples. For a homogeneously mixing population - that is where each member of the population comes into contact with every other member with equal likelihood (a reasonable assumption in the context of an ICE) - SIR-like models offer a faithful picture of population-level disease dynamics, despite their simplicity. Even when there is local heterogeneity of mixing for subpopulations in a particular trial, SIR-like behaviour is often recovered at the large scale, as discussed by [THY+20] concerning COVID-19 transmission in US cities.

With computing power getting ever less expensive [Nor07], powerful techniques are available for the fast and efficient simulation of disease outbreaks described by these models. Algorithms based upon those originally employed to simulate chemical reactions, such as the Doob-Gillespie algorithm from Markov chain theory [PP13, vK07], allow us to sample disease outbreak trajectories over sets of parameter values, initial conditions and population structures, cheaply assessing, for example, likely effects of disease prevention and management interventions, comparative efficacies of different vaccination regimes, effects upon transmission and mortality of various environmental factors etc.

It is commonplace that data from disease outbreaks, and indeed ICE, provide only a partial picture of how a disease progresses through a population. Whereas previously the estimation of stochastic dynamic model parameters, in the absence of complete data, was an intractable problem, the increased availability and speed of computation mean that powerful modern statistical techniques, including a widening array of Markov chain Monte Carlo (MCMC) methods supporting Bayesian inference, may now be brought to bear on such problems. For example, particle filtering techniques [ADH10] simultaneously estimate unknown parameters of the model and any unobserved or missing data by building up sequentially sets of “particles”, i.e. samples from the unobserved parts of the dynamic process. Data augmentation [TW10] alternately samples the parameters and missing data, often using Gibbs or Metropolis-within-Gibss sampling, and has variants such as partial non-centering [NR05] and model-based proposals [PBM15], which aim to improve efficiency. To assess model structure, besides the various information criteria ([Sch78, Mey16]), we also have at our disposal various means of model checking and comparison, such as latent residuals [LMSG14], which can be used to test the specification of each component part (e.g. whether durations of infectiousness are exponentially distributed) of the dynamic model against data. In addition, there are graphical posterior-predictive checks (GP-PCs) ([GS13], described below) and ways to incorporate estimation of the model evidence (i.e. the likelihood of the data averaged over the entire parameter prior distribution) into MCMC sampling routines [PM18]. Finally, it is possible to produce numerical information-theoretic comparisons of hypothetical experimental designs using Bayesian inference and computer simulation (e.g. [VTHvR12, LFKS13]) so that prior to carrying out expensive ICEs we can first assess various aspects of their design.

To give a brief outline of the Chapter, in Section 3.3 we go through two cycles of model fitting and checking in search of a model of transmission and progression of a novel strain of Piscine orthoreovirus in rainbow trout using data reported from a trial conducted by [HVT$^+$17]. We arrive at a stochastic dynamic model of this disease with a susceptible, an exposed (but not yet infectious) and an infectious disease state and two parameters determining the rate of transmission and the duration that fish are latently infected (in the E state). Such quantities were not estimated by the authors of the trial, but we show here that this indeed can be done. In Section 3.4 we explore how trial design influences precision of parameter estimates when fitting models. As suggested by [LFKS13], we use the *mutual information* (MI) between the parameter prior and experimental outcome to compare several variants of the Hauge study design in terms of their expected information gain. Building on the development of [LFKS13], we give a detailed account of how we may estimate the Monte Carlo error in the MI estimate.

3.2 Materials and methods

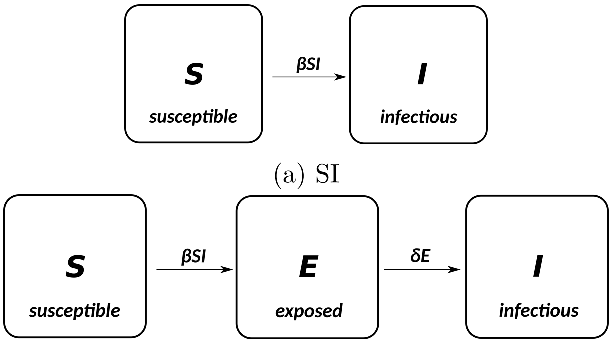
3.2.1 Hauge trial 1A - description and data

The following trial is described in [HVT$^+$17] and was conducted to establish a challenge protocol for a novel strain of piscine ortheovirus, (PRV-Om), affecting farmed rainbow trout in Norway. The trial also confirmed the transmissability of the virus due to cohabitation.

The trial begins with $N_1 = 20$ innoculated ("Group 1") and $N_2 = 20$ disease-naive ("Group 2") hosts cohabited in a tank of volume $V_\mathrm{t} = 150\,\mathrm{l}$. Therefore $S_0 = N_2, I_0 = N_1$. At $t_i = 1$, 2, 4, 6 & 8 wpc (weeks post-challenge) d_i^1 hosts from Group 1 and d_i^2 hosts from Group 2 are selected at random, euthanised and their blood is tested using RT-qPCR for presence of the virus. The numbers of fish testing positive is reproduced in Table 3.1 below.

Sample time t (wpc)	d^1, d^2	r
1	4, 4	0
2	4, 4	0
4	4, 3	1
6	4, 4	2
8	3, 5	5

Table 3.1: **Group 1 and 2 sample sizes,** d^1, d^2 **and positive-testing Group 2 fish,** r from Trial 1A of [HVT+17]. All Group 1 fish tested positive when sampled.



(a) SI

(b) SEI

Figure 3.1: **Summaries of SI and SEI models**. The arrows show the possible transitions between disease states. The parameter $\beta\,\text{wpc}^{-1}$ is the direct transmission rate, which determines the force of infection in relation to the number of infectious hosts present in the system. The time that hosts spend in the exposed (E) state is determined by the rate parameter $\delta\,\text{wpc}^{-1}$.

3.2.2 The models

3.2.2.1 Susceptible-Infectious (SI)

The first of two dynamic disease models used in this chapter is the Susceptible-Infected (SI) stochastic compartmental model, which describes direct transmission of disease among a closed population of hosts that are in one of two disease states: those that are susceptible to disease (S) and those that are infected (I). There is a single parameter β, the *direct transmission rate*, that determines the force of infection (the rate of secondary infections, per susceptible host). The probability that some susceptible host becomes infected during the interval $(t, t + \delta t)$ is

$$\beta S_t I_t \delta t + o(\delta t) \tag{3.1}$$

where S_t and I_t are the numbers of susceptible and infected hosts at time t and $o(\delta t)$ is a generic symbol for a function of δt for which $\frac{o(\delta t)}{\delta t} \to 0,$ as $\delta t \to 0.$ Figure 3.1b summarises the SI model; the solid arrow represents the only possible transition between disease states.

3.2.2.2 Susceptible-Exposed-Infectious (SEI)

The second model (summarised in Fig. 3.1b) has an additional disease state describing hosts that are exposed, but not yet infectious (E). Equation 3.1 also describes the probability of a susceptible host entering the E state during the interval $(t, t + \delta t)$. The SEI model has an additional parameter, δ, that determines the rate that hosts move from the E to the I state, at which point they become infectious. More precisely, upon entering the exposed state, hosts will move into the infectious state after a random duration, τ, where

$$\tau \sim \text{Exponential}(\delta) \tag{3.2}$$

so that hosts are in the E state for a duration of $\frac{1}{\delta}$, on average. Transitions from S to E and from E to I are the only ones possible.

3.2.2.3 Sampling model

The state of the SI or SEI systems at any time t is not directly observed. Instead, a series of samples of sizes $\mathbf{d}$ are taken at times $\mathbf{t}$ and tested for disease. We assume that this test is perfectly reliable, in terms of sensitivity (no false negatives) and specificity (no false positives), so that, with probability one, infected fish test positive and susceptible fish test negative, when sampled. Sampling, considered here, is destructive, i.e. fish are not returned to the tank.

For the SI model considered here, if there are S susceptible and I infectious Group 2 hosts at time of sampling, then on taking a sample of size d, the number of positive hosts in the sample, r, follows a hypergeometric distribution:

$$\mathbb{P}(r = r') = \frac{\binom{I}{r'}\binom{S}{d-r'}}{\binom{S+I}{d}} \tag{3.3}$$

where e.g. $\binom{S+I}{d}$ is the number of different ways of choosing a sample of size d from the $S + I$ hosts in the tank.

For the SEI model, with S, E and susceptible, exposed and infectious Group 2 hosts

$$\mathbb{P}(r = r') = \frac{\binom{E+I}{r'}\binom{S}{d-r'}}{\binom{S+E+I}{d}}. \tag{3.4}$$

3.2.3 Simulation

Appendix F.2 contains sample Python2 code used to simulate counts of Group 2 positive-testing at sample points similar to the Hauge trial, with underlying model SEI.

3.2.4 Inference

We now describe how we fit the SEI model to Hauge trial data, employing Bayesian inference. Throughout the rest of this chapter the generic symbol $p(x)$ denotes the density or probability mass function (pmf) for the random quantity x and $p(x \,|\, y)$ denotes the conditional density or pmf of x, given y. The posterior density summarises everything we wish to know about the unknown parameter values, having observed the data

$$p(\beta, \delta \,|\, \mathbf{r}). \tag{3.5}$$

Since the form of this density is analytically intractable, we use techniques associated with Markov chain Monte Carlo (MCMC) to approximately draw dependent samples from (3.5), as we now outline. By Bayes' rule, Eq. 3.5 is proportional to

$$p(\mathbf{r} \,|\, \beta, \delta)\, p(\beta, \delta) \tag{3.6}$$

The first factor of Eq. 3.6 is the likelihood and second is the prior for the two unknown parameters β and δ, where we assumed these to be independently exponentially distributed with rate $\lambda = 1.0 \times 10^{-3}\,\mathrm{wpc}^{-1}$, i.e.

$$p(\beta, \delta) \propto e^{-\lambda(\beta+\delta)}. \tag{3.7}$$

Such a diffuse prior, suggested by [NR05], reflects an initial lack of knowledge about the unknown parameters. Since β and δ take values in the positive real line only, it is convenient to sample them on the natural logarithmic scale in what follows, so that instead we to draw samples from

$$p(\log\beta, \log\delta \,|\, \mathbf{r}) \propto p(\mathbf{r} \,|\, \beta, \delta)\, p(\beta, \delta)\, \beta\, \delta, \tag{3.8}$$

where the additional factor $\beta\delta$ is the required Jacobian term due to the fact that we have log-transformed the parameters.

We use a sequential Monte Carlo (SMC) step with bootstrap particle filter ([ADH10, GC01]) to approximate

$$\hat{p}(\mathbf{r} \,|\, \beta, \delta) \approx p(\mathbf{r} \,|\, \beta, \delta) \tag{3.9}$$

which, having sampled a set of parameters, proceeds by forward simulation of a number of "particles", consisting of the sequences of unobserved events of the SEI model between measurement times. At the measurement times, the particles are weighted according to their agreement with the counts of positive-testing fish, $\mathbf{r}$, and then resampled. This way, those particles with underlying states that agree well with the data are more likely to be propagated. Since sampling is lethal, we must adjust the S, E and I compartment sizes for each particle before the next set of forward simulations; we remove the r_i positive-testing Group 2 hosts randomly from the E and I compartments, d_i^1 Group 1 hosts from the I compartment and $d_i^2 - r_i$ negative-testing Group 2 hosts from the S compartment. Any particles with insufficient hosts in the three compartments receive a weighting of zero and are not propagated further. See Appendix E for a detailed description of the routine, implemented using `C++`, used here. See the introduction to Chapter 4 for a discussion of some of the alternatives to this approach.

The sampling routine for $i = 1, 2, 3, \dots, N$ (with initial N_{burn} iterations used for burn in and adaptation) is as follows:

1. Initialise:
   a) Set initial values for the log-transformed parameters, $\log \beta_0, \log \delta_0$
   b) Initialise the variance-covariance matrix (variances chosen to scale suitably, relative to initial $\log \beta_0$, etc.)

$$\Sigma = \begin{pmatrix} \left(\frac{\log \beta_0}{1.96}\right)^2 & 0 \\ 0 & \left(\frac{\log \delta_0}{1.96}\right)^2 \end{pmatrix} \tag{3.10}$$

2. Propose:

$$(\log \tilde{\beta}, \log \tilde{\delta}) = (\log \beta_{i-1}, \log \delta_{i-1}) + v_i$$

   where $v_i \sim \text{MVN}(\mathbf{0}, \Sigma)$ is drawn from a multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix Σ. To optimise sampling efficiency, before we can begin to use or collect useful samples from the parameter posterior distribution we must first tune Σ - we follow the method used by [PBM15].

3. Estimate:

$$\hat{p}(\mathbf{r} \mid \tilde{\beta}, \tilde{\delta}) \approx p(\mathbf{r} \mid \tilde{\beta}, \tilde{\delta}) \quad \text{(Eq. 3.9 and App. } E\text{)}$$

4. Accept-reject:

$$\begin{aligned} &\text{set} \quad \log \beta_i, \log \delta_i = \log \tilde{\beta}, \log \tilde{\delta} \\ &\text{with probability} \quad \frac{\hat{p}(\mathbf{r} \mid \tilde{\beta}, \tilde{\delta})\, p(\tilde{\beta}, \tilde{\delta})\, \tilde{\beta}\, \tilde{\delta}}{\hat{p}(\mathbf{r} \mid \beta_{i-1}, \delta_{i-1})\, p(\beta_{i-1}, \delta_{i-1})\, \beta_{i-1}\, \delta_{i-1}} \\ &\text{otherwise,} \quad \log \beta_i, \log \delta_i = \log \beta_{i-1}, \log \delta_{i-1}. \end{aligned}$$

   If $i \leq N_{\text{burn}}$, then if accepted set

$$\Sigma = 1.002 \times \Sigma \tag{3.11}$$

   and if rejected set

$$\Sigma = 0.999 \times \Sigma. \tag{3.12}$$

   In addition, while $i \leq N_{\text{burn}}$, at every N_{rescale} iterations (usually between 750 and 1000), set the elements of Σ to the variances and covariances of the sequences of the log-transformed parameter samples. This step adapts the proposal density to information we have gained so far about the shape of the joint posterior, improving the acceptance rate of the MCMC chains.

3.2.5 Mutual information

Throughout this subsection, the symbols θ, Y and X denote the unknown model parameters, the experimental outcome and unobserved data. Densities are again denoted generically by $p(\ldots)$.

In experimental outcomes there are two sources of randomness. The first comes from our initial uncertainty about the parameters, θ, of the model, as expressed by the prior distribution. The *prior entropy*

$$H(\theta) = -\int p(\theta) \log p(\theta) \, d\theta$$

is a measure of the uncertainty implied by the prior and is maximised when θ's distribution is highly dispersed and minimised when its density is concentrated in a small region. The second source of randomness is the outcome of the experiment, Y (which we model via the likelihood $p(y \mid \theta)$). The *posterior entropy*

$$H(\theta|Y) = -\int\int p(\theta|Y) \log p(\theta|Y) \, d\theta dY$$

measures the uncertainty in the posterior density, averaged over all experimental outcomes. The difference between these latter two quantities is the *mutual information* (see e.g. [HK07])

$$I(\theta, Y) = H(\theta) - H(\theta|Y). \tag{3.13}$$

The mutual information can be described as the expected loss of uncertainty, or gain of information, when updating the prior $p(\theta)$ with an observation of Y. [LFKS13] suggest this as a measure with which to compare the average information gain of different experimental designs in biology; those designs with a higher mutual information will produce, in the long run, the greatest redution in uncertainty when passing from the prior to the posterior.

As in [LFKS13], Eq. (3.13) can be rewritten

$$I(\theta, Y) = \int\int p(Y \mid \theta) p(\theta) \log \frac{p(Y|\theta)}{p(Y)} \, d\theta dY = \mathbb{E}_{Y,\theta}\left[\log \frac{p(Y|\theta)}{p(Y)}\right] \tag{3.14}$$

suggesting how we may obtain a Monte Carlo estimate of $I(\theta, Y)$ by first drawing $\theta_i \sim p(\theta)$ and then $Y_i \sim p(Y|\theta_i)$, for $i = 1, \ldots, N_1$ and calculating

$$\hat{I} = \frac{1}{N_1} \sum_{i=1}^{N_1} \log \frac{p(Y_i|\theta_i)}{p(Y_i)} \tag{3.15}$$

which is straightforward, if we can write down the likelihood and evidence terms, in terms of Y_i and θ_i. In the current application, however, we cannot calculate these directly since they are both integrals over the unobserved data likelihood $p(X|\theta)$ and / or the prior, i.e.

$$
\begin{aligned}
p(Y_i|\theta_i) &= \int p(Y_i \mid X, \theta_i) p(X \mid \theta_i)\, dX \\
p(Y_i) &= \int \int p(Y_i \mid X, \theta) p(X \mid \theta) p(\theta)\, dX d\theta
\end{aligned}
\tag{3.16}
$$

and so these too must be estimated. For each Y_i, we estimate the evidences $p(Y_i)$ by first sampling $\theta_j \sim p(\theta)$ and then sampling $X_j \sim p(X \mid \theta_j)$ for $j = 1, \dots, N_2$ and putting

$$
\hat{p}(Y_i) = \frac{1}{N_2} \sum_j^{N_2} p(Y_i|X_j, \theta_j). \tag{3.17}
$$

which converges to $p(Y_i)$ by the law of large numbers. Similarly, we estimate the likelihoods, $p(Y_i|\theta_i)$, by sampling $X_k \sim p(X \mid \theta_i)$ for $k = 1, \dots, N_3$ and putting

$$
\hat{p}(Y_i|\theta_i) = \frac{1}{N_3} \sum_k^{N_3} p(Y_i|X_k, \theta_i) \tag{3.18}
$$

and we replace $p(Y_i)$ and $p(Y_i|\theta_i)$ with $\hat{p}(Y_i)$ and $\hat{p}(Y_i|\theta_i)$ in Eq. 3.15.

In order to construct approximate confidence intervals for the estimate, $\hat{I}$, we first need to estimate its standard error

$$
\text{std. err}(\hat{I})^2 = \frac{1}{N_1} \text{var}\Big(\log \frac{\hat{p}(Y|\theta)}{\hat{p}(Y)} \Big) \tag{3.19}
$$

accounting for the error arising both from sampling $Y, \theta \sim p(Y|\theta)$ and from estimating the likelihoods and evidences in Eq. 3.15. We can partition the variance of $\log \frac{\hat{p}(Y|\theta)}{\hat{p}(Y)}$ into contributions from each of these two sources according to the law of total variance,

$$
\text{var}\Big(\log \frac{\hat{p}(Y|\theta)}{\hat{p}(Y)} \Big) = \text{var}\left(\mathbb{E}\Big[\log \frac{\hat{p}(Y|\theta)}{\hat{p}(Y)} \Big| Y, \theta \Big] \right) + \mathbb{E}\left[\text{var}\Big(\log \frac{\hat{p}(Y|\theta)}{\hat{p}(Y)} \Big| Y, \theta \Big) \right]. \tag{3.20}
$$

The first term on the right hand size of Eq. 3.20 is the variance of the *conditional expectation* of $\log \frac{\hat{p}(Y|\theta)}{\hat{p}(Y)}$, given Y and θ. To estimate this term we first note that by expanding $\log \frac{\hat{p}(Y|\theta)}{\hat{p}(Y)}$ to first order around $p(Y_i|\theta_i)$ and $p(Y_i)$, and then taking conditional expectations

$$
\begin{aligned}
&\mathbb{E}\Big[\log \frac{\hat{p}(Y|\theta)}{\hat{p}(Y)} \Big| Y = Y_i, \theta = \theta_i \Big] \\
&\approx \mathbb{E}\Big[\log \frac{p(Y_i|\theta_i)}{p(Y_i)} - \frac{1}{p(Y_i)} (\hat{p}(Y) - p(Y_i)) + \frac{1}{p(Y_i|\theta_i)} (\hat{p}(Y|\theta) - p(Y_i|\theta_i)) \Big| Y = Y_i, \theta = \theta_i \Big] \\
&= \log \frac{p(Y_i|\theta_i)}{p(Y_i)}
\end{aligned}
\tag{3.21}
$$

and then take the sample variance

$$\operatorname{var}\left(\mathbb{E}\Big[\log\frac{\hat{p}(Y|\theta)}{\hat{p}(Y)}\Big|\,Y,\theta\Big]\right)\approx\widehat{\operatorname{var}}\Big\{\log\frac{p(Y_i|\theta_i)}{p(Y_i)}:i=1,\dots,N_1\Big\}. \tag{3.22}$$

The second term on the right of Eq. 3.20 is the mean of the *conditional variance* of $\log\frac{\hat{p}(Y|\theta)}{\hat{p}(Y)}$, which given Y and θ is defined as

$$\operatorname{var}\Big(\log\frac{\hat{p}(Y|\theta)}{\hat{p}(Y)}\Big|\,Y,\theta\Big)=\mathbb{E}\Big[\Big(\log\frac{\hat{p}(Y|\theta)}{\hat{p}(Y)}-\mathbb{E}\big[\log\frac{\hat{p}(Y|\theta)}{\hat{p}(Y)}\big|\,Y,\theta\big]\Big)^2\Big|\,Y,\theta\Big] \tag{3.23}$$

We can however estimate the conditional variance appearing in second term on the right hand side of Eq. 3.20 by once more expanding to first order and then taking conditional variances,

$$\begin{aligned}\operatorname{var}\Big(\log\frac{\hat{p}(Y|\theta)}{\hat{p}(Y)}\Big|\,Y,\theta\Big)&\approx\operatorname{var}\Big(\log\frac{p(Y|\theta)}{p(Y)}-\frac{1}{p(Y)}\left(\hat{p}(Y)-p(Y)\right)+\frac{1}{p(Y|\theta)}\left(\hat{p}(Y|\theta)-p(Y|\theta)\right)\Big|\,Y_i,\theta_i\Big)\\&=\frac{1}{p(Y)^2}\operatorname{var}\Big(\hat{p}(Y)\Big|\,Y,\theta\Big)+\frac{1}{p(Y|\theta)^2}\operatorname{var}\Big(\hat{p}(Y|\theta)\Big|\,Y,\theta\Big).\end{aligned} \tag{3.24}$$

There is no covariance term in the last line since the Monte Carlo estimates $\hat{p}(Y_i)$ and $\hat{p}(Y_i|\theta_i)$ for each $i=1,\dots,N_1$ are respectively generated from the independent sequences

$$\begin{aligned}&\{X_j,\theta_j\sim p(X,\theta),j=1,\dots,N_2\}\\\text{and}\quad&\{X_k\sim p(X|\theta_i),k=1,\dots,N_3\}\end{aligned} \tag{3.25}$$

and so are conditionally independent given Y_i and θ_i.

Finally, letting V_i^{e} and V_i^{l} denote the sample variances of the sets $\{p(Y_i|X_j,\theta_j)\ j=1,\dots,N_2\}$ and $\{p(Y_i|X_k,\theta_k)\ k=1,\dots,N_3\}$ we used to estimate $p(Y_i)$ and $p(Y_i|\theta_i)$ for each $i=1,\dots,N_1$, then (replacing the p's with $\hat{p}$'s)

$$\begin{aligned}\mathbb{E}\Big[\operatorname{var}\Big(\log\frac{\hat{p}(Y|\theta)}{\hat{p}(Y)}\Big|\,Y,\theta\Big)\Big]&\approx\frac{1}{N_1}\sum_{i=1}^{N_1}\operatorname{var}\Big(\log\frac{\hat{p}(Y_i|\theta_i)}{\hat{p}(Y_i)}\Big|\,Y_i,\theta_i\Big)\\&\approx\frac{1}{N_1}\sum_{i=1}^{N_1}\Big\{\frac{1}{\hat{p}(Y_i)^2}\frac{1}{N_2}V_i^{\mathrm{e}}+\frac{1}{\hat{p}(Y_i|\theta_i)^2}\frac{1}{N_3}V_i^{\mathrm{l}}\Big\}\end{aligned} \tag{3.26}$$

Putting all of the above together, we approximate $\mathrm{var}(\hat{I})$ in terms of all the various means and sample variances

$$\mathrm{var}(\hat{I}) \approx \frac{1}{N_1}\left\{V^0 + \frac{1}{N_1}\sum_{i=1}^{N_1}\left\{\frac{1}{\hat{p}(Y_i)^2}\frac{1}{N_2}V_i^{\mathrm{e}} + \frac{1}{\hat{p}(Y_i|\theta_i)^2}\frac{1}{N_3}V_i^{1}\right\}\right\}. \tag{3.27}$$

and a 95% confidence interval for the estimate $\hat{I}$, for example, is

$$\left(\hat{I} - 1.96 \times \sqrt{\mathrm{var}(\hat{I})}, \hat{I} + 1.96 \times \sqrt{\mathrm{var}(\hat{I})}\right). \tag{3.28}$$

3.2.6 Model checking

We use graphical posterior predictive checks (GPPC) [GCSR95, GS13] to graphically compare the data (e.g., as in Figure 3.3) with corresponding posterior-predictive distributions of similar quantities. These being in marked disagreement suggests we need to look again at the model.

For example, the posterior-predictive distribution of r_1, the number of positive-testing fish found in the first sample, for the SI model is

$$\begin{aligned} p(\hat{r}_1 \,|\, r_1) &= \int p(\hat{r}_1, S_1, I_1 \,|\, r_1)\, dS_1 dI_1 \\ &= \int\int p(\hat{r}_1 \,|\, S_1, I_1) p(S_1, I_1 \,|\, \beta, r_1) p(\beta \,|\, r_1)\, dS_1 dI_1 d\beta. \end{aligned} \tag{3.29}$$

where S_1 and I_1 are the numbers of susceptible and infectious hosts at the first sample time. We sample from $p(\hat{r}_1 \,|\, r_1)$ by first sampling β_i, $i = 1, \ldots N$ independently from the posterior (or a Guassian density estimate of the posterior estimated from the MCMC posterior sampled using SciPy's `guassian_kde` function [VGO$^+$20]) and then using forward simulation of the SI process with $\beta = \beta_i$ to sample the S_1^i, I_1^i. Finally, we approximate

$$P(\hat{r} = j \,|\, r_1) \approx \frac{1}{N}\sum_{i=1}^{N} P(\hat{r}_1 = j \,|\, S_1^i, I_1^i). \tag{3.30}$$

3.3 Case study: novel piscine ortheovirus (PRV-Om) in rainbow trout [HVT$^+$17]

We present here a case study where we fit a stochastic dynamic disease model to data gathered from an ICE described by [HVT$^+$17] demonstrating how useful quantities that characterise this and other infections can be estimated from these kinds of data using the theory and techniques described above. We show how the straightforward model checking technique of graphical posterior predictive checks (GPPC) can alert us to problems with our fitted model.

In the trial, rainbow trout *Onchorhyncus mykiss* were intraperitoneally injected with a challenge innoculum of a novel variant of piscine orthoreovirus (PRV-*Om*), typically associated with heart and skeletal muscle inflammation in Atlantic salmon. This new variant was observed to be causing disease in farmed rainbow trout in Norway [HVT+17] and the particular experiment that we study here was a small-scale preliminary trial whose purpose was to establish that innoculation via intraperitoneal injection was effective in initiating infection in healthy fish and that infection via cohabitation was one of the possible routes of transmission. The innoculated fish were then cohabited with a group of disease-naive fish and samples of the innoculates and initially disease naive fish were taken at 1, 2, 4, 6 and 8 weeks post-challenge (wpc). The trial lasted eight weeks in total. Blood from all of the sampled fish was tested using real-time polymerase chain reaction (RT-qPCR) to detect presence of viral RNA. In addition tests for organ pathologies were also carried out for a sub-sample, although we do not use those results here. All of the innoculate group tested positive for virus throughout the trial with blood viral levels observed to rise to a peak at 4 wpc. The numbers of positive testing fish in each sample comprise the data that we use here to fit the model. Due to the results of this trial, the researchers were able to conclude that the virus replicates in the blood of rainbow trout and the disease could be transmitted environmentally.

See Section 3.2 for a more detailed description of the trial and the resulting data. We first attempt to fit the SI model, and so, initially, our assumptions are

1. the underlying dynamic model is SI, i.e. fish are either susceptible or infectious
2. Group 1 (innoculates) occupy the I state throughout the trial
3. Group 2 (naive) are initially in the S state
4. a positive RT-qPCR test on blood implies infectiousness and a negative test implies susceptibility, i.e. fish in the S state test negative and fish in the I state test positive.

Assumption 4 in the above list is somewhat reasonable in this case since the organs of a subsample of sampled fish are examined post-mortem for signs of disease, providing validation of the blood RT-qPCR test. When we cannot assume either 100% sensitivity or specificity (or both) then we must adjust our observation model to account for the possibility of false positives or negatives.

Figure 3.2a contains a plot of the estimated posterior density and trace plot for the single parameter of the above model - the transmission rate, β (see caption below figure for detail). The trace plot is shown to demonstrate that both of the MCMC chains converged to a common stationary distribution, which the associated theory tells us must be the posterior. Having started off with a very diffuse exponential prior, with rate 1×10^{-3} giving a prior mean for β of $1000\,\mathrm{wpc}^{-1}$, we see that the data are informative about β, which has a posterior mode at approximately $0.03\,\mathrm{wpc}^{-1}$ and posterior 2.5 and 97.5 percentiles at about $0.01\,\mathrm{wpc}^{-1}$ and $0.04\,\mathrm{wpc}^{-1}$, so that there is 95% of β's posterior mass lying between these two values.

However, we must still check that our SI model fits the data. A graphical posterior-predictive check (GPPC - see Section 3.2) compares the data with the numbers of positive-testing Group

2 fish, $\mathbf{r}$, predicted by the model at the sampling times 1 wpc, 2 wpc, 4, 6 and 8 wpc. We do this by drawing values of β from its posterior distribution and then peforming a simulated run of the experimental set-up. After binning up the results, we obtain an approximation of the posterior-predicted positive sample numbers, which may then be compared graphically with the data, as in Figure 3.3a. What jumps out immediately from the Figure is the over-prediction by the model of positive samples at 1 wpc, 2 wpc, 4 and 6, or in other words, the data lie in a low probability region of the posterior-predictive distribution, $p(\,\hat{\mathbf{r}}|\,\mathbf{r}\,)$. In fact, by looking at the overall number of positive tests from the five samples combined, the model makes an over-prediction with probability 0.97. The values for the parameter β that are *likely*, as expressed by the posterior, would lead to too many infections, implying that the one parameter SI model is an overly-simplistic picture of the mechanisms of transmission behind the trial data. It should also be noted that by comparing the model's prediction of the number of positive testing fish at each sampling event with the numbers obtained by experiment we have a more sensitive test of model fit than if we based our test of model fit on low-dimensional summaries of the same data, as discussed by Gibson (2018) [GST18].

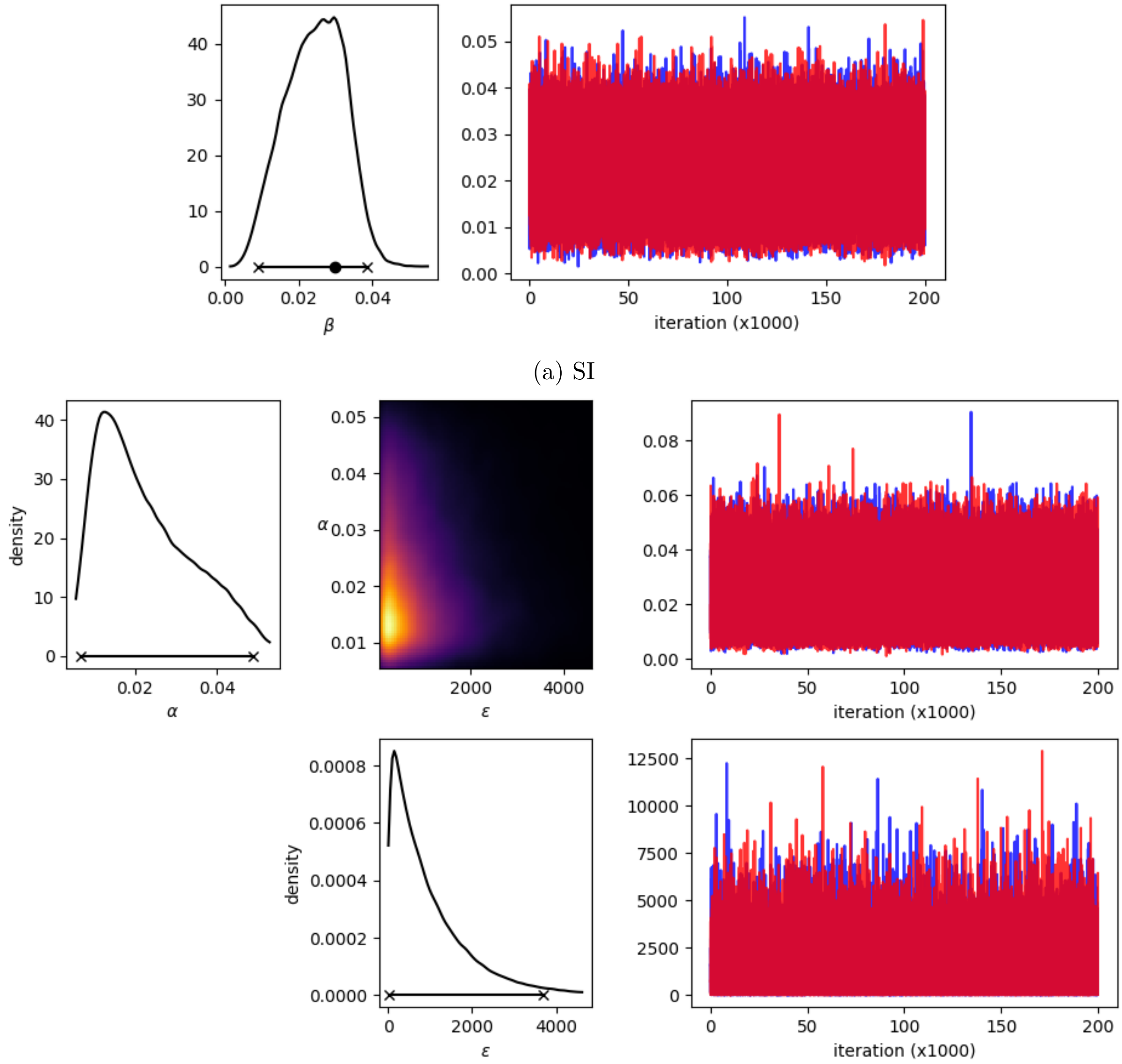
Bearing in mind the assumptions above, a possible step towards improving the model is to consider the possibility that some of the fish that test positive throughout the trial, while they have contracted the disease (as evidenced by presence of the virus in their blood) are not yet infectious to others. This can be accounted for by simply expanding the disease model to include an *exposed* (E) state, so that our underlying model is SEI (Section 3.2). We now also need to link this new state into our observation model by assuming that fish in the E state test positive. The updated assumptions are

1. the underlying dynamic model is SEI
2. Group 1 occupy the I state throughout the trial and Group 2 are initially in the S state
3. fish in the S state test negative and fish in the E and I states test positive for the disease.

A summary of the posterior distribution for the two parameters of the SEI model, β, δ is shown in Figure 3.2b and Figure 3.3b shows that an improved fit, with the predicted number of positive samples much more resembling what was seen in the experiment, with the probability of over-prediction reduced to 0.78. The SEI transmission rate, β, has a marginal posterior mode of about $0.013\,\text{wpc}^{-1}$ and δ, the E-I transition rate, has a mode at approximately $166\,\text{wpc}^{-1}$ and 2.5 and 97.5 percentiles at approximately 25 and $3698\,\text{wpc}^{-1}$. Equivalently, our model tells us that the duration of the latently infected state (i.e. $\frac{1}{\delta}$) is in the range 0.05 h to 6.72 h, with a probability of 0.95.

3.4 Computer simulation to improve design of experiments

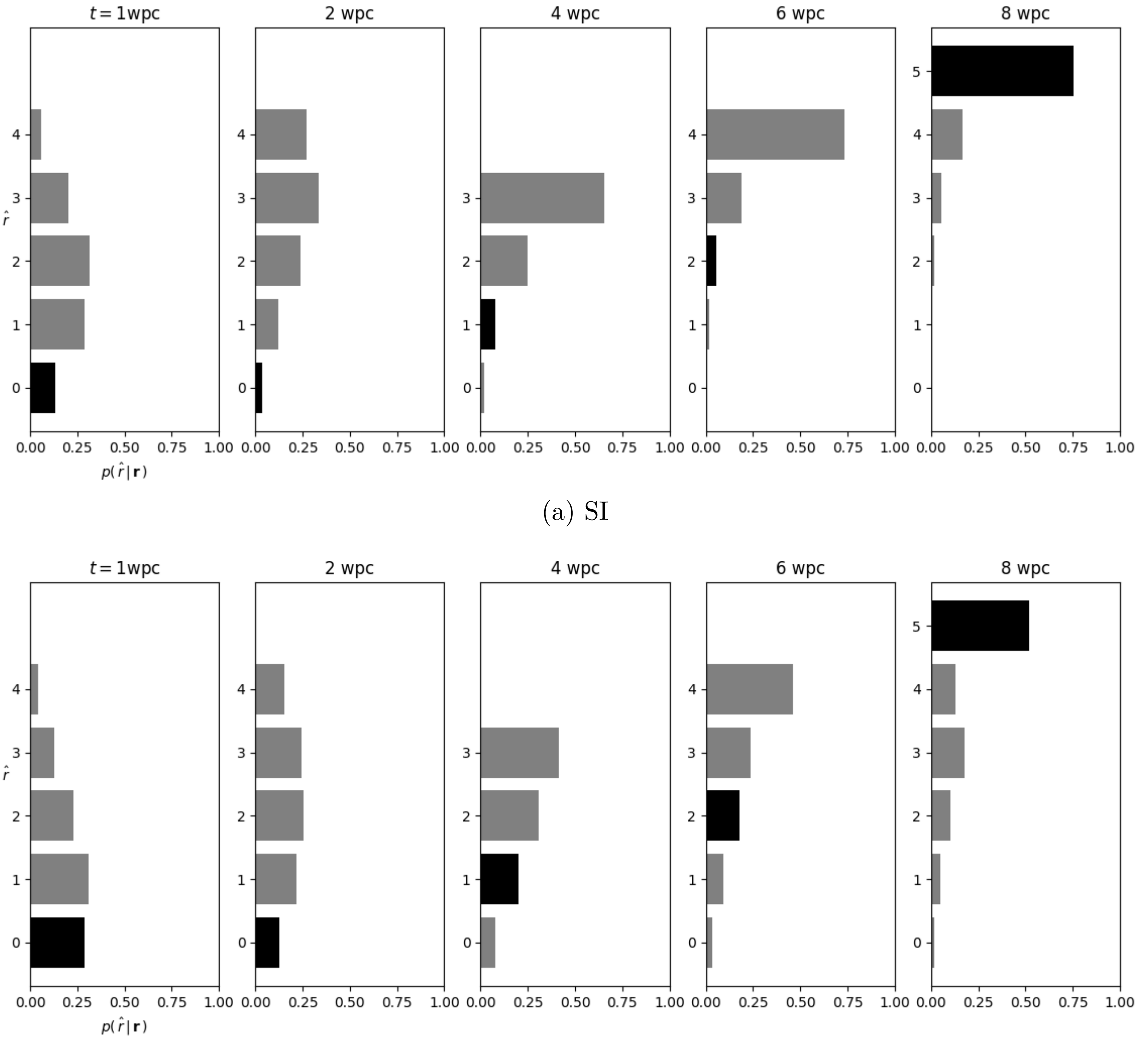
There are many variable aspects in the design of a ICE. To list some of these, there are environmental factors that can be controlled, such as tank volume, water flow-though rate, water temperature, pH, salinity, etc. and the characteristics of the tests used to monitor for infection,



(a) SI

(b) SEI

Figure 3.2: **Estimated posterior density and trace plots** for the parameters of the SI (a) and SEI (b) models fitted to the data from Trial 1A of [HVT+17]. In the leftmost plots, the 2.5 and 97.5 posterior percentiles and posteriors mode of the marginal parameter densities are indicated at the bottom. Exponentially-distributed marginal priors (indepently in the case of (b)) are assumed, reflecting an initial lack of knowledge about the rate of transmission. Particle-marginal MCMC was used to obtain the samples (see Sec. 3.2), with each iteration consisting of first a Metropolis-Hastings step for the parameter β followed by an SMC step to sample the unobserved infection event times. A total of 1200 independent particles were used for each chain with weighting and resampling at each of the five sampling events at 1 wpc, 2 wpc, 4 wpc, 6 wpc and 8 wpc. The trace plots are shown as a graphical indication of convergence to stationarity of the two indepedent MCMC chains started with different initial values, however, Gelman and Rubin convergence statistics, via the R coda package [R C18, PBCV06], were also inspected as a numerical confirmation of convergence. 2×10^5 iterations are sampled after discarding 2×10^4 iterations for burn-in and adaptation. Estimated densities were obtained using SciPy's `guassian_kde` function [VGO+20].



(a) SI

(b) SEI

Figure 3.3: **Estimated posterior-predictive distribution** of numbers of positive-testing Group 2 fish from Trial 1A of [HVT+17] for the SI (a) and the SEI (b) models, where 4, 4, 3, 4 and 5 Group 2 fish were sampled at 1 wpc, 2 wpc, 4 wpc, 6 wpc and 8 wpc, respectively. The black bars indicate the values $\mathbf{r} = r_1, \ldots, r_5$ observed in the trial. Both models are likely to over-predict the number of positive tests, in comparison with the data: the posterior-predictive probability of there being *more* positive tests in total than were observed in the trial, $P(\sum \hat{r_i} > \sum r_i \,|\, \mathbf{r})$, is 0.97 for (a) and 0.78 for (b). Based on this summary statistic, the SEI model therefore, with its one extra disease state and parameter, provides a better fit to the trial data. Posterior-predictive samples were generated by taking 5×10^4 evenly spaced values of the parameters from the MCMC-samples and simulating a run of the trial.

i.e. whether they are lethal, as in the Hauge trial, or non-lethal, such as swabbing of the gills and skin mucus (see e.g. [SRT+21]) and their reliability (sensitivity, specificity). We may also vary the size of the trial, both in terms of the number of replicate tanks as well as the number of fish per tank and the sizes and timings of samples taken. How the trials are initiated is another factor, e.g. innoculated and naive groups of 20 fish, in the cae of Hauge, or prepartion of tank water to a recorded pathogen density.

Experiments have costs associated with them as well as the ethical requirements to reduce animal suffering as well as the number of live animals sacrificed in the study. We demonstrate in this

Section how computer simulation and ideas from information theory may help us explore how certain design factors will likely influence how much information we can expect to extract from an experimental outcome.

Figure 3.4 shows posterior summaries for the parameter β after fitting an SI model to six simulated runs each of a hypothetical infection under two different sampling designs (see caption beneath Figure for details). The plot shows that, of the two designs, Design 1, where four sets of 20 samples taken at 24 h-ly intervals, gives uniformly narrower parameter estimates than does Design 2, where eight sets of 10 samples are taken every 12 h. Although the 1st run of Design 1 produces a poorer estimate than the others, overall, Design 1 gives better results than Design 2.

Returning to the case study of Section 3.3, suppose now that we intend to perform a follow-up challenge trial of PRV-*Om* in rainbow trout with the purpose of refining the parameter estimates that we obtained previous for the SEI model. We wish to select a trial design that best achives this from among a list of candidates with varying sample sizes and times, summarised in the first three columns of Table 3.2. Our present knowledge about the two parameters, β and δ, is now described by their posterior distribution, which is summarised in Figure 3.2b. We take this posterior to be our new prior, which we intend to update once again with data from this follow-up trial. In light of Figure 3.4, we want the design that, on average, results in the greatest reduction in remaining parameter uncertainty. If $\mathbf{r}_Z$ denotes the experimental outcome of trial design Z (the counts of positive tests at each of the sample times) then the *mutual information* (MI), $I(\mathbf{r}_Z, \beta, \delta)$ (see Section 3.2) offers a numerical comparison of this reduction in parameter uncertainty from our current prior (our old posterior) to the new posterior, as described by [LFKS13]. The measure summarises the reduction in parameter uncertainty, or information gain, averaged over both the prior and all experimental outcomes, as modelled by the likelihood.

Since, the likelihood $p(\mathbf{r}_Z|\beta, \delta)$ is an integral over the numbers of susceptible, infectious and exposed hosts at each of the sample times, weighted by their respective likelihoods according to the SEI model, it is intractable and cannot be written down as a closed functional form. Therefore, for each sampled β and δ the likelihoods themselves must be Monte Carlo estimated along with the evidences, that are used to estimate the MI. This additional source of error must be taken into account when constructing confidence intervals for $\hat{I} \approx I(\mathbf{r}_Z, \beta, \delta)$. See Section 3.2 for detailed discussion.

The final column of Table 3.2 contains Monte Carlo estimates of $\hat{I}$ for each of the six designs, along with approximate 95% confidence intervals. These are also plotted in Figure 3.5. Taking into account the margin of error, Design F turns out to be the clear winner with the highest estimated MI, while Designs C and D, where for both of these, four sets of 5 samples were taken, are expected to perform very poorly, in comparison. Designs A, B seem to offer comparable performance, where two lots of 10 fish are sampled. Based on this exercise, however, it would be advisable for researchers to simply repeat the trial without altering the design.

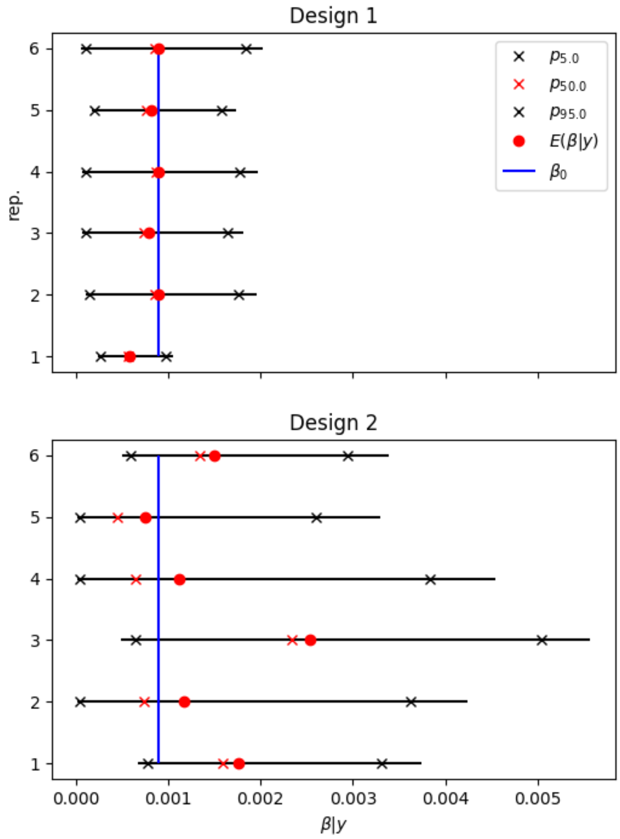


Figure 3.4: **Summary of posterior distributions** of parameter β of SI model fitted to six simulated runs each of two experimental designs of an SI disease system with $\beta_0 = 0.0009$ with evenly spaced lethal samples, starting with a high variance exponential prior for β with mean 1×10^3 h. All trials are initiated when 1 infectious fish is cohabited with 80 susceptible fish. In **Design 1**, 4 samples of size 20 are taken 24 h apart, starting at $t = 24$ h while in **Design 2** 8 samples of size 10 are taken at every 12 h starting at $t = 12$ h. Marked on the plot are the posterior mean ($E(\beta|y)$, medians ($p_{50.0}$) and 5 and 95 posterior percentiles ($p_{5.0}, p_{95.0}$)for each run. We see a clear difference between the posteriors obtained from the two designs; Design 1 produces the posterior means and medians close to β_0 in five out of the size runs, meaning these are reliable point summaries. The posteriors we get from Design 2 are more variable in two ways: they are more *dispersed*, meaning less precise estimates of β, and they are more variable *between runs*, so that some experimental runs will produce more precise estimates than others.

3.5 Discussion

We have seen how it is possible to extract useful epidemiological information from ICE data. Indeed, from data gathered in Trial 1A of [HVT+17] we are able to estimate the rate of disease transmission of the novel Piscine orthoreovirus PRV-*Om* among rainbow trout, as well as the average duration of latent infection. We then explored how computer simulation can be used to compare the efficiency of a number of experimental designs, via the mutual information, $I(\theta, Y)$. Adopting the posterior distribution associated with an SEI model fitted to the Hauge data as our new prior, we compared a number of trial designs with varying sample times of different sizes and found, in fact, that the orginal design performed the best.

Design	$\mathbf{d}^1, \mathbf{d}^2$	$\mathbf{t}$ (wpc)	$\hat{I}$ (nat) (95% CI)
A	10, 10	3, 5	0.814 (0.812, 0.816)
B	10, 10	3, 6	0.816 (0.814, 0.818)
C	5, 5, 5, 5	3, 5, 7, 9	0.764 (0.762, 0.766)
D	5, 5, 5, 5	3, 6, 9, 12	0.711 (0.709, 0.712)
E	4, 4, 4, 4, 4	1, 2, 3, 4, 5	0.804 (0.803, 0.805)
F	4, 4, 4, 4, 4	1, 2, 4, 6, 8	0.826 (0.824, 0.827)

Table 3.2: **Six candidate trial designs** for a follow up trial of PRV-Om in rainbow trout. In all cases, the trial is initiated with 20 innoculates and 20 disease-naive fish. $\mathbf{d}^1 = d_1^1, \ldots$ innoculates and $\mathbf{d}^2 = d_2^1, \ldots$ of the initially disease-free fish are removed at times $\mathbf{t} = t_1, \ldots$ for testing. The prior distribution, $p(\beta, \delta)$, is a Gaussian density (obtained using SciPy's `guassian_kde` function [VGO+20]) estimate of the sampled posterior distribution for these parameters on fitting the SEI model to the Hauge trial data, described in the previous Section. The third column contains Monte Carlo estimates of the mutual information (see Section 3.2) along with approximate 95% confidence intervals summarising the accuracy of these estimates. For Designs A to D, 5000 samples were drawn from $p(\beta, \delta)$, and then for each of these samples, 5000 further samples were drawn to estimate the likelihood and 5000 to estimate the evidence. These were sufficient to obtain sufficiently accurate estimates. To see the effect upon increasing the number of samples on the error, for Designs E and F 20×10^3 samples were drawn to estimate both the likelihood and the prior, for each of 5000 parameter values sampled from the prior. This increase in samples reduced the width of the confidence interval from about 0.004 to 0.002.

Obtaining a fitted dynamic model of some disease system should not be the end of the story, however. Returning to the over-prediction of positive tests in the posterior-predictive analysis summarised in Figure 3.3b, we see that more could be done in terms of forming a mechanistic picture of the transmission of PRV-*Om* in rainbow trout. As [LEH+15] suggest, a fitted model helps to identify where gaps remain in our understanding of the system under study, prompting ideas about what further data need to be collected. Returning to the stated assumptions underlying the SEI model, we did not allow for the possibility that some of the positive testing fish had ceased to be infectious, although they still had detectable blood viral loads. A goal of a future study could be to better understand the within-host virology of this new viral strain by using a non-lethal method of testing for shedding of virus, such as taking gill swabs, and allowing for the possibility that some of the fish cease to shed during the trial. This would be described by expanding the SEI model to include a recovered (R) state. Another assumption made when fitting the SEI model was that our test was perfectly reliable both in terms of correctly detecting cases (i.e. the test is 100% sensitive) without reporting false positives (100% specific). A better fitting model may be found where these assumptions are tweaked to allow a degree of unreliability of one or both of these kinds.

While here we estimated the mutual information for six arbitrarily chosen variations of the Hauge design, a more systematic search of, at least part of, the design space should be straightforward to implement. Cost is an important factor when considering the design of new trials and there are also legal and ethical obligations to minimise animal suffering, e.g. the "3R" principles of "replacement, reduction and refinement" in terms of animals used in research [ASP14]. Simulation of models in order to test hypotheses and for purposes of power analysis [LEH+15], prior to conducting the trial is a low-cost way to meet these requirements. While the estimation of measures such as the mutual information is computationally intensive, this is still less expensive than

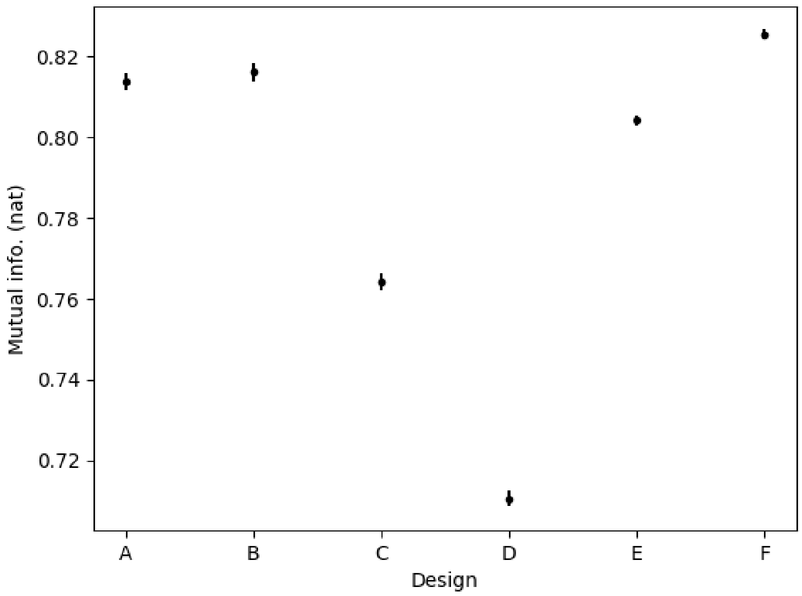


Figure 3.5: **Plots of mutual information and 95% confidence interval error bars** for the size candidate designs for a follow-up trial of PRV-Om in rainbow trout. The narrowed error bars for Designs E and F came at the cost of a fourfold increase in the number of independent samples used to estimate the likelihoods and evidences (see Sec. 3.2).

a poorly designed experiment. An alternative means of experimental design comparison could be offered by expanding the kind of analysis presented in Figure 3.4. Here posterior location and dispersion statistics were obtained from fitting models under various experimental designs to simulated data with fixed parameter values. In this case a "good" experimental design is one where the posterior distribution has low dispersion and consistently has the true parameter values in its support. This could work especially well when there is only a single parameter value to infer.

Chapter 4

Novel statistical tools for environmental transmission models and digital-PCR pathogen load data

4.1 Introduction

The kinetics of pathogen shedding and the viability of pathogen in the environment are two important factors in the indirect, environmental transmission of diseases including cholera [WHO21a] in humans, paratuberculosis in cattle [DMW$^+$12, PMH$^+$17, Fec18] and many aquatic infectious diseases [ODW$^+$18]. The quantification of these processes is therefore important if we are to understand and make predictions about the spread of environmentally-transmitted diseases. Stochastic compartmental environmental transmission models (ETM) extend the framework exemplified by the susceptible-infectious-removed (SIR) model of Kermack and McKendrick [KM27] to allow infection to be transmitted indirectly via an external pathogen pool, as well as directly from host to host. In this Chapter we advance the tools for ETMs in two ways. Firstly, we develop methods for estimating the parameters from data that can be obtained from suitably designed challenge trials or observational studies. Secondly, we introduce a modelling approach that explicitly accounts for loss of viability of environmental pathogens and show how this can be estimated within our inferential framework. This novel inference framework builds on Bayesian techniques related to particle filtering and sequential Monte Carlo [ADH10, GC01] to allow a wider application of ETMs in the study of environmentally-transmitted diseases, especially those affecting aquaculture production. ETMs offer a natural way to relate the rate of transmission to environmental conditions that affect the viability of environmental pathogens [ODW$^+$18, PMPG$^+$16, MÁC$^+$12] and here we develop the methods to fit these models and show what extra data are required.

As discussed in Chapter 2, although direct transmission models (DTM) are good descriptions of single outbreaks of environmentally transmitted disease within and between closed populations, there are circumstances where it is necessary to quantify environmental transmission in terms of its dependence upon the size or density of the environmental pathogen load. For example, survival of pathogen external to the host is identified by Grassly and Fraser as one of the factors that drives seasonal change in incidence of disease [GF06]. In [WJL+12] the authors show how an environmental transmission model (ETM) of avian influenza reproduces the complex, multi-year, patterns of major outbreaks observed among wild duck populations. These authors show that the frequency of outbreaks is related to intensity of environmental transmission, which is itself a function of pathogen density in water bodies. Oidtmann, et al., (2018), observe that whereas "live fish movements are one of the main routes of long-distance spread of viral pathogens, waterborne spread of virus is the most relevant route for regional spread" and therefore, in order to piece together a global picture of the spread of waterborne disease, we need local models of transmission and its dependence upon a spatially and temporally varying pathogen density.

Experimental challenge allows researchers to study the dynamics of infectious disease and the efficacies of treatment and prevention strategies under controlled conditions. Some recent work concerning diseases that impact upon commercial aquaculture production include development of vaccines against a salmonid alphavirus in Atlantic salmon [TWI+21] and infectious pancreatic necrosis virus in rainbow trout [LHL+20]. The rate of pathogen shedding as a function of time has also been studied for various hosts and diseases. Relevant to the work here, are two interesting studies that track the rate of viral shedding from hosts over the course of an infection. Wargo et al., (2017) find, from measuring viral RNA copies shed per hour using quantitative polymerase chain reaction (qPCR), that there is a peak in viral shedding from juvenile rainbow trout around 2 d after innoculation with infectious hematopoietic necrosis virus, with secondary and even tertiary peaks thereafter. We show that, despite such complex shedding kinetics at the individual host level, models that describe a fixed uniform rate of pathogen shedding during each host's period of infectiousness provide a good description of disease spread at the population level.

The study of sexual transmission of Zika virus among immunodeficient AG129 mice by [DRP+17] goes further in that they measure both the viral RNA copy numbers, using qPCR, and infectious virion numbers (in plaque-forming units, PFU), via plaque assay, in ejaculate samples taken from mice at intervals, post-innoculation. Here they find that the measurements of live virions and of RNA copy numbers are not correlated, with viral material remaining detectable in samples for up to 40 d past the point at which mice cease to shed live virus. This presents a critical challenge when attempting to quantify the rate of secondary infections in terms of environmental viral concentration as measured by qPCR, since not all viral material that is present is able to cause new infections. Widders, Broom and Broom term this the "viral shedding vs infectivity dilemma" [WBB20], in relation to SARS-CoV-2. The inability to account for such differences between qPCR measurements and viable viral load could render fitted ETM models highly misleading. In response to this, we propose a stochastic compartmental ETM with one compartment each describing the amounts of live and dead pathogen in the system.

When fitting stochastic compartmental disease models to data, there is the issue of missing data and unobservable events, e.g. we often cannot observe directly when members of the population first become exposed to, or "catch", an infection. Many techniques in the field of Bayesian inference and Monte Carlo sampling have been brought to bear on this problem. For example, *data augmentation* [TW87, TW10], alternately samples parameter values and values for the unobserved events from their respective full conditional distributions, treating all unknown quantities on an equal footing. Gibson and Renshaw (1998) and O'Neill and Roberts (1999) further developed these ideas, introducing data-augmented MCMC to infer pathogen transmission and disease dynamics alongside model parameters from partial observed epidemics, using continuous time stochastic processes. Notable developments of these ideas, developed to improve computational tractability, include partial non-centering (PNC) by Neal and Roberts, applied in their 2005 paper to fitting the SIR model to a set of host-removal times [NR05], as well as the model-based proposal (MBP) [PBM15]. PNC seeks to improve the speed of convergence of the MCMC chains produced by data augmentation by seeking an optimal re-parameterisation of the likelihood that, in some way, lies between a "centred" parameterisation (where the parameters and the unobserved events are conditionally dependent, given the data) and a "non-centred" one (where the parameters and the unobserved events are marginally *in*dependent). The MBP, unlike initial data augmentation approaches and PNC, attempts to update both parameters and unobserved events simultaneously by exploiting knowledge of the model being fitted.

When there are a large number of unknown parameters or unobserved events then we run up against the so-called "curse of dimensionality" [Bel61] where the volume of the sample space and the required number of samples grow rapidly with the number of quantities we are required to estimate. For this reason, it is not appropriate to treat every minute fluctuation of the pathogen load as an unobserved event. A more appropriate class of methods for the problem-at-hand are those associated with sequential Monte Carlo (SMC) [DFG01] and particle filtering, especially particle-marginal MCMC (PM-MCMC) [ADH10]. In our application, each iteration of the PM-MCMC sampler first chooses a new set of parameter values and then samples a number of "particles", which are forward simulations of the pathogen load and unobserved host event times between subsequent pairs of measurement events. This is why such methods are termed "sequential". At each measurement event, the particles are "filtered" probabilistically; those that do not agree well with the current observed measurements are pruned while others that are more in agreement with the data are allowed to propogate, continue and be tested against the next set of measurements in the sequence. The key to PM-MCMC's applicability here is that it reduces the dimensionality of the problem, i.e. the number of quantities to be estimated, to the number of unknown parameters, plus the number of hosts in each disease state, plus the sizes of the pathogen load compartments at each of the measurement event times. Another significant advantage of particle methods in general is that the forward simulations of the particles can be run in parallel batches, greatly speeding up the process on a multi-core machine. In the previous chapter, PM-MCMC was used to infer parameter values and missing event times for the problem of fitting SI and SEI models to data describing transmission and progression of a viral disease in rainbow trout, as described by Hauge, et al. (2017). For that particular application, since the number of missing event times is modest in comparison, data augmentation or the MBP could possibly work just as well.

This chapter addresses the problems of estimating the parameters of an ETM from data as follows. We begin by describing, in Section 4.2.1, the SI-PQ model, which is a two-host (susceptible & infectious), two-pathogen compartmental (live & dead) ETM, where the pathogen load is described by a pair of coupled diffusion processes. ETMs do not typically subdivide the pathogen population in this way (e.g. [BPK$^+$16, WJL$^+$12, BDSR09]). We then describe an observation model that relates the underlying disease states to results of tests of samples of hosts and digital polymerase chain reaction (dPCR - [SJ13]) measurements, a sensitive and powerful technology that yields absolute quantification of viral RNA copy numbers, before describing two hypothetical experimental designs (Sec. 4.2.2); the first is an expansion of the one employed by Hauge et al., (2017) in their study of a novel Piscine orthoreovirus in rainbow trout (*Oncorhynchus mykiss*) and Atlantic salmon (*Salmo salar L.*); and the second is similar to that of Kumar et al., (2013) in their study of the viability of white spot syndrome virus (WSSV) in sterile seawater. Using PM-MCMC we fit the SI-PQ model (Sec. 4.2.3) to a simulated outcome of the trial with the first of these hypothetical designs (Sec. 4.2.5). The simulation incorporates some of the characteristics of data collected in real trials and observational studies, such as time-varying rates of pathogen shedding. We show that as long as external pathogen loses its viability at a slow rate relative to the known rate of water filtration, we may obtain good estimates of the rates of environmental transmission and pathogen shedding (defined below). However, combining data from both designs we are able, in addition, to estimate the rate of pathogen decay, as well as get improved estimates of the environmental transmission and pathogen shedding rates (Sec. 4.3).

4.2 Materials and methods

Table 4.2 contains a summary of all symbols used in the current chapter.

4.2.1 Environmental transmission and disease progression model

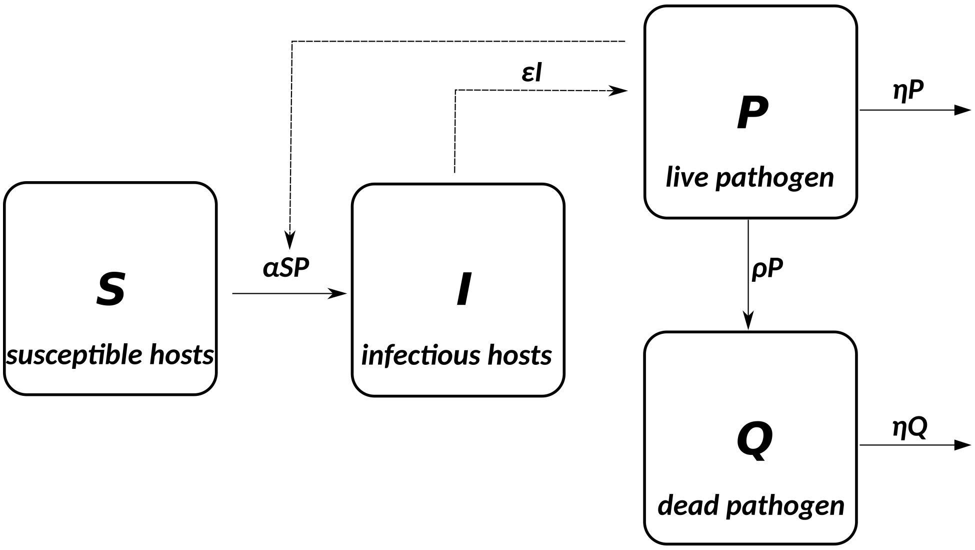


Figure 4.1: **SI-PQ model diagram.** Susceptible and infectious hosts in the system are represented by **S** and **I**, respectively. Live pathogen is represented by **P** while **Q** represents dead, or unviable, pathogen (incapable of causing new infections) still present in the tank. The behaviour of **P** and **Q** over time is described by the coupled diffusions, Eq. (4.2). The parameters α and ϵ are the rates of environmental transmission and pathogen emission. Pathogen leaves the tank by water filtration at rate η and loses its viability at rate ρ.

In order to capture the probabilistic nature of disease transmission among the host population in relation to the environmental dynamics of the pathogen, we propose a stochastic compartmental environmental transmission model (ETM) (which we term SI-PQ and describe below) that subdivides the host and the pathogen populations each into two sub-populations. In light of the shedding vs. infectivity dilemma [WBB20] mentioned in the Introduction, we wish to allow for the possibility that only a subset of detectable pathogen in the tank is able to cause new infections. The SI-PQ model describes the dynamics of the processes by which pathogens decay from viable to unviable and through which both types are lost from the system due to water filtration.

The two host sub-populations of the SI-PQ model are those hosts that are *susceptible* to disease (S) and those that, having been exposed, are *infectious* (I). The pathogen in the tank is assumed either to be viable, and able to infect susceptible hosts (P), or unviable (Q). The model is summarised in Figure 4.1 where the boxes represent the four sub-populations and the possible transitions between them are represented by the solid arrows. The transitions shown *do not include* any removals of hosts or pathogen from the tank when samples are taken, but these are accounted for in the inference procedure described below. Letting S_t, I_t, P_t and Q_t denote the sizes of the susceptible and infectious host and the viable and unviable pathogen subpopulations, then conditional upon S_t and P_t, the probability that an infection occurs during the interval $(t, t + \delta t)$ is

$$\alpha S_t P_t \delta t + o(\delta t) \tag{4.1}$$

with α the *environmental transmission rate* ($o(\delta t)$ is a generic symbol for a function of δt for which $\frac{o(\delta t)}{\delta t} \to 0,$ as $\delta t \to 0$). There is no direct route of transmission in this model.

Throughout, we express the sizes of the two parts of the pathogen load, P_t, Q_t, in terms of RNA copy numbers, since the concentration of these is measured by digital polymerase chain reaction (dPCR, see below). An alternative pathogen quantification (not used here) is the *quantum of infection*. This is the amount of pathogen required to infect a susceptible member of the population with probability $1 - e^{-1} = 0.632$ [BNS+03, SC10].

Equations (4.2) describe the time evolution of P_t and Q_t in relation to I_t, the number of infectious hosts. The pathogen emission rate, ϵ, is the rate that each infected host emits pathogen into the water, η is the rate at which pathogen is removed from the water via filtration and the pathogen decay rate, ρ, is the rate that pathogen becomes unviable. The four $dB_{\lambda t}$ terms are the sources of randomness for each of the sub-processes of pathogen emission (dB^E), decay (dB^D) and removal by water filtration from the P and Q subcompartments (dB^{F_P} and dB^{F_Q}). These stochastic infinitesimals of Brownian motion, “sped up” by the factor λ, and have normal distributions with mean zero and variance λdt.

$$
\begin{aligned}
dP_t &= (\epsilon I_t - \rho P_t - \eta P_t)\, dt + \left[\sqrt{\epsilon I_t}\, dB^E_{\lambda t} - \sqrt{\rho P_t}\, dB^D_{\lambda t} - \sqrt{\eta P_t}\, dB^{F_P}_{\lambda t}\right] \\
dQ_t &= (\rho P_t - \eta Q_t)\, dt + \left[\sqrt{\rho P_t}\, dB^D_{\lambda t} - \sqrt{\eta Q_t}\, dB^{F_Q}_{\lambda t}\right]
\end{aligned}
\tag{4.2}
$$

For a small time increment, Δt, the corresponding increments in P_t and Q_t are approximately

$$
\begin{aligned}
\Delta P_t &\approx (\epsilon I_t - \rho P_t - \eta P_t)\, \Delta t + \left[\sqrt{\epsilon I_t}\, \mathrm{N}^E(0, \lambda \Delta t) - \sqrt{\rho P_t}\, \mathrm{N}^D(0, \lambda \Delta t) - \sqrt{\eta P_t}\, \mathrm{N}^{F_P}(0, \lambda \Delta t)\right] \\
\Delta Q_t &\approx (\rho P_t - \eta Q_t)\, \Delta t + \left[\sqrt{\rho P_t}\, \mathrm{N}^D(0, \lambda \Delta t) - \sqrt{\eta Q_t}\, \mathrm{N}^{F_Q}(0, \lambda \Delta t)\right]
\end{aligned}
\tag{4.3}
$$

where the $\mathrm{N}^D(0, \lambda\Delta t)$ etc. are independent and normally distributed. Eqs. 4.3 are the key to how in Section 4.2.5 P_t and Q_t are simulated using the Euler-Maruyama method [Hig01]. A well defined solution to Eqs. 4.2 exists [RW00, Ch. 5] at least when when $P_0 > 0$, $Q_0 > 0$. When simulating care should be taken that P_t and Q_t remain non-negative, for example, by setting

$$
\begin{aligned}
P_{t+\Delta t} &= \max\{0, P_t + \Delta P_t\} \\
Q_{t+\Delta t} &= \max\{0, Q_t + \Delta Q_t\}.
\end{aligned}
\tag{4.4}
$$

as in Eq. 4.28.

We require that the form of Eqs. 4.2 continues to hold under a rescaling of the pathogen load, e.g. from RNA copy numbers to quanta of infection. This is ensured by the parameter λ, since for a positive scaling factor c, $\frac{1}{\sqrt{c}} dB_{c\lambda t} \sim dB_{\lambda t}$ and $\tilde{P}_t = c^{-1} P_t, \tilde{Q}_t = c^{-1} Q_t, \tilde{I}_t = I_t$ and $\tilde{\epsilon} = c^{-1}\epsilon, \tilde{\rho} = \rho, \tilde{\eta} = \eta, \tilde{\lambda} = c^{-1}\lambda$. If we remove the terms in square brackets from (4.2) then what remains is a pair of ordinary differential equations, whose solution, for fixed I_t, is the *determinstic part* of P_t and Q_t's behaviour. The parameter λ can be used to adjust the sizes of the stochastic fluctuations of P_t and Q_t around the values determined by the deterministic part of (4.2) - small values of λ produce near deterministic pathogen load behaviour, for fixed I_t, while larger values produce noisier pathogen load behaviour. However, we will not attempt to estimate λ in what follows, instead treating it as a nuisance parameter. By fitting two SI-PQ models with distinct, fixed values for λ we will show that inferences about the other parameters are not affected by the value we fix for λ.

As outlined in [RW00, Ch. 5], Eqs. 4.2 with fixed $I_t = I$ and $\lambda = 1$, can be derived as an approximating diffusion process

$$
\begin{pmatrix} dP_t \\ dQ_t \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} dt + \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \begin{pmatrix} dB^1_t \\ dB^2_t \end{pmatrix}
\tag{4.5}
$$

to the discrete state space, continuous-time Markov chain (DS-CTMC) $(\tilde{P}_t, \tilde{Q}_t)$ with transition rates in Table 4.1 below, choosing $\binom{b_1}{b_2}$ and $\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$ in order to match up the first and second moments of $\binom{dP_t}{dQ_t}$ and $\binom{\tilde{P}_{t+\Delta t} - \tilde{P}_t}{\tilde{Q}_{t+\Delta t} - \tilde{Q}_t}$ to first order. However, the resulting equations do not have the required scaling property mentioned above.

Transition		rate	description
$(\tilde{P}, \tilde{Q}) \to$	$(\tilde{P}+1, \tilde{Q})$	ϵI	pathogen emission
	$(\tilde{P}-1, \tilde{Q}+1)$	$\rho \tilde{P}$	pathogen decay
	$(\tilde{P}-1, \tilde{Q})$	$\eta \tilde{P}$	water filtration from P
	$(\tilde{P}, \tilde{Q}-1)$	$\eta \tilde{Q}$	water filtration from Q

Table 4.1: **Markov transition rates** for discrete state space $\tilde{P}_t, \tilde{Q}_t$ process.

4.2.1.1 dPCR measurements of pathogen

The state of the SI-PQ system at any time, t, can only be inferred indirectly via two kinds of sampling. The first is a dPCR measurement of the total waterborne viral material, $P_t + Q_t$, at time t, in RNA copy numbers. The dPCR technology biomechanically separates a water sample into C separate reaction chambers, each of volume V_c, and tests for the presence of *at least one* target molecule in each chamber. The number of "hits", H_t, constitutes a measurement. Dube, Qin and Ramakrishnan (2008) show that the probability, θ_t, of a hit is

$$\theta_t = 1 - \exp\left(-\frac{V_\mathrm{c}}{V_\mathrm{t}}(P_t + Q_t)\right), \tag{4.6}$$

where V_t is the total volume of the tank or vessel from which the sample is taken. By way of explanation of Eq. 4.6, the number of target molecules turning up in a given reaction chamber may be modelled as a Poisson random variable with mean $\mu = \frac{V_\mathrm{c}}{V_\mathrm{t}}(P_t + Q_t)$ and so the probability of there being no molecule in the chamber is $e^{-\mu}$ and the probability of there being one or more (a hit) is $1 - e^{-\mu}$.

The number of hits, H_t, is therefore assumed to follow the binomial distribution

$$\begin{aligned} H_t|\theta_t &\sim \text{binomial}(C, \theta_t) \\ \mathbb{P}(H_t = h|\theta_t) &= \binom{C}{h} \theta_t^h (1-\theta_t)^{C-h}. \end{aligned} \tag{4.7}$$

Dube, Qin and Ramakrishnan discuss a dPCR device with 12 panels each containing 765 chambers (9180 in total) each of volume 6 nl. We assume the same specification throughout this Chapter. Since

$$\begin{aligned}\mathbb{E}\Big(\frac{H_t}{C}\Big) &= \theta_t \\ \mathrm{var}\Big(\frac{H_t}{C}\Big) &= \frac{1}{C}\theta_t(1-\theta_t)\end{aligned}$$

$\frac{H_t}{C}$ provides an unbiased estimate of θ_t, and this estimate is improved by increasing the number of chambers, C.

4.2.1.2 Testing for host disease

The second kind of sampling is the testing of hosts for signs of disease, which we assume to be perfectly reliable in terms of sensitivity (no false negatives) and specificity (no false positives), so that, with probability one, infected fish test positive and susceptible fish test negative, when sampled. This is somewhat reasonable here since in the original trial by Hauge et al., (2017) sampled fish are euthanised and subjected to thorough post-mortem.

If, at the time of sampling, there are S susceptible and I infectious hosts, and we take a sample of size d, the number of positive hosts in the sample, r, follows a hypergeometric distribution:

$$\mathbb{P}(r = r') = \frac{\binom{I}{r'}\binom{S}{d-r'}}{\binom{S+I}{d}} \tag{4.8}$$

where e.g. $\binom{S+I}{d}$ is the number of different ways of choosing a sample of size d from the $S+I$ hosts in the tank.

4.2.2 Experimental designs

4.2.2.1 Experiment A: Immersion challenge

What data are required to fit the SI-PQ model? We consider first an experimental design “Experiment A” based on that by Hauge et al., (2017), used to establish the possibility of environmental transmission of a novel strain of PRV among both rainbow trout *Onchorhyncus mykiss* and Atlantic salmon *Salmo salar*. We make a hypothetical modifcation to the trial design by simulating the taking of dPCR measurements of the viral concentration of the tank water at the same times that fish are sampled.

1. The trial begins with N_1 innoculated (“Group 1”) and N_2 disease-naive (“Group 2”) hosts cohabited in a tank of volume V_t, so that $S_0 = N_2, I_0^1 = N_1, I_0^2 = 0$. Fish from one of the two groups are marked so that they can be distinguished from the those in the other group.

2. Viral matter is removed from the tank by the water filtration system at the rate η.

3. At pre-determined times $0 < t_1 < \cdots < t_n$:
   a) $d_1^1, \ldots, d_n^1$ hosts from Group 1 are selected at random and removed. Samples are taken from Group 1 only to conform with the original trial design of [HVT+17]
   b) $d_1^2, \ldots, d_n^2$ hosts from Group 2 are sampled at random and tested for disease, with the number of these testing positive, $r_1, \ldots, r_n$, recorded and
   c) a dPCR measurement, $H_1, \ldots, H_n$, with C chambers, each of volume V_c, is taken.

All aspects of the design: sampling sizes d_i^1 and d_i^2, number of dPCR chambers, C, sampling times t_i, tank volume V_t and water filtration rate μ are fixed in advance. Let $\mathbf{r} = (r_1, \ldots, r_n)$ and $\mathbf{H} = (H_1, \ldots, H_n)$ denote the vectors of Group 2 positive sample sizes and dPCR readings.

Fixing $I_t = I$ and setting $dP_t = dQ_t = 0$ in the determinstic part of Eq. 4.2 we can solve for P_t and Q_t at temporary equilibrium, P^I and Q^I

$$\begin{aligned} P^I &= \frac{\epsilon I}{\rho + \eta} \\ Q^I &= \frac{\rho}{\eta} P^I \\ \frac{P^I}{P^I + Q^I} &= \frac{1}{1 + \frac{\rho}{\eta}}. \end{aligned} \tag{4.9}$$

The water filtration rate, η, is fixed in the design of the experiment, and so if the unknown pathogen decay rate, ρ, is signifcantly smaller than η then

$$\frac{P^I}{P^I + Q^I} \approx 1 \tag{4.10}$$

i.e. nearly all of the pathogen in the system is viable. We will show below that when fitting the SI-PQ model, if we can assume that $\rho = 0$, or equivalently that all pathogen is viable until it is removed by water filtration, data from this trial design are sufficient for the estimation of both the rates of environmental transmission, α, and pathogen shedding ϵ.

Rather than assuming that $\rho = 0$, if instead we attempt to estimate the rate of pathogen decay, alongisde the rates of environmental transmission, α, and pathogen emission, ϵ, then in the absence of an informative prior, we run into a problem of parameter identifiability. This is due to the fact that the total pathogen load, $P_t + Q_t$, and the force of infection, αP_t, are measured directly (by dPCR and counting positive-testing hosts, respectively), but neither α nor the proportion of viable pathogen P_t can be identified.

4.2.2.2 Experiment B: Viral viability

The second trial design we consider, “Experiment B”, allows the estimation of the rates of environmental transmission, α, and pathogen decay, ρ and is based on the investigation of the viability of white spot syndrome virus in sterile seawater by [SAR+13]:

1. k containers of identical volume are prepared with the same predetermined pathogen load, p_0.

2. The filtration mesh is large enough to avoid filtering out viral particles from the tank water.

3. At pre-determined times, $0 = t_0 < t_1 < \cdots < t_k$, one host is introduced into an unoccupied container (chosen at random).

4. At time $T > t_k$ all hosts are tested for disease (again assumed to be perfectly sensitive and specific).

The number of containers, k, trial duration, T, introduction times, t_i and initial pathogen load p_0 are fixed. The trial is repeated, yielding $\mathbf{y} = (y_0, y_1, \ldots, y_k)$ counts of positive-testing fish in each of the containers, out of a total of m repetitions. As we shall see below, these counts together with the numbers of positive tests, $\mathbf{r}$, and dPCR readings, $\mathbf{H}$ from Experiment A enable the estimation of the rate of viral decay, ρ, and to improve our estimates of the rates of environmental transmission, α, and viral emission, ϵ.

We will show that data from Experiments A and B combined allow estimation of the parameters α, ϵ and ρ simultaneously.

4.2.3 Inference methodology

We outline here the process of fitting the SI-PQ model to data from Experiments A and B combined, where the data from A consist of counts of positive tests, $\mathbf{r} = (r_1, \ldots, r_n)$ and dPCR readings $\mathbf{H} = (H_1, \ldots, H_n)$ taken at the n sampling times $t_1, \ldots, t_n$ and data from B are the counts of positive testing fish in each of the k containers $\mathbf{y} = (y_1, \ldots, y_k)$ out of m repetitions. Table 4.2 contains a summary of all symbols employed in the current Chapter. It should also be clear from the following how to fit the model to data from Experiment A only. The methods described here, such as sequential Monte Carlo (SMC), are generic and can, in principle, be easily adapted to account for specifics of disease systems of interest by, for example, adapting the process model to have more compartments, e.g. SEIR etc, or by adjusting the observation model.

Throughout, we employ Bayesian inference with densities represented by the generic symbols $p(\ldots)$ and $p(\ldots \mid \ldots)$. Everything we wish to know about the unknown parameter values, α, ϵ and ρ, having observed the data, is encapsulated by the posterior density

$$p(\alpha, \epsilon, \rho \,|\, \mathbf{H}, \mathbf{r}, \mathbf{y}). \tag{4.11}$$

Symbol	**description**
SI-PQ state variables:	
S_t, I_t	susceptible / infectious hosts at time t
P_t, Q_t	viable / unviable environmental pathogen load at time t
N_1, N_2	Initial sizes of Groups 1 and 2
SI-PQ model parameters:	
α	environmental transmission (h^{-1})
ϵ	pathogen shedding (h^{-1})
ρ	pathogen decay (h^{-1})
η	water filtration (h^{-1})
λ	relative size of stochastic fluctuations of P_t, Q_t (dimensionless)
Experiment A:	
d_i^1, d_i^2	i^{th} Group 1 and 2 samples sizes
r_i	Number of positive tests in i^{th} Group 2 sample
H_t	dPCR reading at time t (RNA copy numbers)
C	total dPCR chambers
θ_t	probability of a single dPCR hit
$V_{\text{t}}, V_{\text{c}}$	tank (l) and dPCR chamber (nl volumes
dB_t	infinitesimal increment of Brownian motion
a, b, m	parameters of variable shedding rate (simulation only)
$t_1, \ldots, t_n$	sampling and dPCR measurement times (wpc)
Experiment B:	
y_i, k	positive-testing hosts in i^{th} container, out of k repetitions
p_0	initial pathogen load (part ml^{-1})
$t_0, \ldots, t_k$	introduction times (d)
T	trial duration (h)

Table 4.2: **Table of symbols used in current Chapter**. Occasionally we add the indices ij to symbols assoicatied with Experiment B - i refers to container and j denotes repetition. Bold face $\mathbf{r}, \mathbf{y}, \mathbf{H}$, denote corresponding vectors.

Since such densities are analytically intractable, we use techniques associated with Markov chain Monte Carlo (MCMC) to approximately draw dependent samples from (4.11), as we now outline. By Bayes' rule, (4.11) is proportional to

$$p(\mathbf{H}, \mathbf{r} \mid \alpha, \epsilon, \rho)\, p(\mathbf{y} \mid \alpha, \rho)\, p(\alpha, \epsilon, \rho). \tag{4.12}$$

The first and second factors of (4.12) are respectively the likelihoods for the outcomes of Experiments A and B. These describe how the data depend upon the parameter values. Conditional on the parameters, $\mathbf{H}, \mathbf{r}$ are indepedent of $\mathbf{y}$. The third factor is the prior density for the three parameters and encapsulates our knowledge and beliefs about the parameters *prior* to performing the experiments and obtaining the data. We will assume flat (improper) priors, i.e.

$$p(\alpha, \epsilon, \rho) \propto 1 \tag{4.13}$$

which means roughly that, before carrying out the experiments, any set of values for these parameters is, for us, as likely as any other to be the "true values". However, more precise knowledge of these parameters can be accounted for when this is available. The three parameters take values in the positive real line only, and so it is more convenient to instead sample them on the natural logarithmic scale in what follows, so that instead we are to draw samples from

$$p(\log\alpha, \log\epsilon, \log\rho \,|\, \mathbf{H}, \mathbf{r}, \mathbf{y}) \propto p(\mathbf{y} \,|\, \alpha, \rho)\, p(\mathbf{H}, \mathbf{r} \,|\, \alpha, \epsilon, \rho)\, p(\alpha, \epsilon, \rho)\alpha\epsilon\rho, \tag{4.14}$$

where the additional factor $\alpha\epsilon\rho$ is the required Jacobian term due to the fact that we have transformed the parameters.

Having obtained the data $\mathbf{H}, \mathbf{r}$ and $\mathbf{y}$, we will need to estimate $p(\mathbf{H}, \mathbf{r} \,|\, \alpha, \epsilon, \rho)$ and $p(\mathbf{y} \,|\, \alpha, \rho)$ for given α, ϵ and ρ. To do so we must relate the state of the simulated system to the observations as now described.

Concerning Experiment B, letting κ_i be the probability that the host in container i *not* is infected at time T after introduction to the bucket at time t_i with initial pathogen load p_0

$$p(\mathbf{y} \,|\, \alpha, \rho) = \binom{m}{y_i} \prod_i \kappa_i^{y_i} (1 - \kappa_i)^{m - y_i} \tag{4.15}$$

where, conditional on P_t^i, the pathogen load in container i,

$$\kappa_i = 1 - \exp\left(-\alpha \int_{t_i}^{T} P_t^i \, dt\right) \tag{4.16}$$

since the host in container i becomes infected with hazard αP_t, between the times $t = t_i$ and T.

To simplify, we replace P_t^i with a deterministic process of exponential decay at rate ρ, the right hand side of Eq. (4.16) is approximately

$$\begin{aligned} \hat{\kappa}_i = 1 - & \exp\left(-\alpha \int_{t_i}^{T} p_0 e^{-\rho t} \, dt\right) \\ = & 1 - \exp\left(-\frac{\alpha p_0}{\rho}\left(e^{-\rho t_i} - e^{-\rho T}\right)\right) \end{aligned} \tag{4.17}$$

and

$$\tilde{p}(\mathbf{y} \,|\, \alpha, \rho) = \binom{m}{y_i} \prod_i \hat{\kappa}_i^{y_i} (1 - \hat{\kappa}_i)^{m - y_i} \approx p(\mathbf{y} \,|\, \alpha, \rho). \tag{4.18}$$

We use a sequential Monte Carlo (SMC) step ([ADH10, GC01]) to numerically estimate

$$\tilde{p}(\mathbf{H}, \mathbf{r} \,|\, \alpha, \epsilon, \rho) \approx p(\mathbf{H}, \mathbf{r} \,|\, \alpha, \epsilon, \rho) \tag{4.19}$$

which proceeds by forward simulation between measurement times of a number of "particles", each representing a simulation of the system based in the model with (potentially) different parameter values. Once the measurement times are reached, the particles are reweighted according to Eqs. 4.8 and 4.7, so that those particles with underlying states that agree well with the data are more likely to be propagated. Since sampling in Experiment A is lethal, we must adjust the S and I compartment sizes for each particle after sampling and reweighting; we remove the r_i positive-testing Group 2 hosts from the I compartment, d_i^1 Group 1 hosts from the I compartment and $d_i^2 - r_i$ negative-testing Group 2 hosts from the S compartment. Any particles with insufficient hosts in either compartment receive a weighting of zero and are not propagated further. See Appendix E for a detailed description of the routine used here, however, the upshot is that we obtain an unbiased estimate of the likelihood (Eq. 4.19) that can be used within the following algorithm

The sampling routine for $i = 1, 2, 3, \ldots, N$ (with initial N_{burn} iterations used for burn in and adaptation) is as follows:

1. Initialise:

   a) Set initial values for the log-transformed parameters, $\log \alpha_0, \log \epsilon_0, \log \rho_0$

   b) Initialise the variance-covariance matrix (variances chosen to scale suitably, relative to initial $\log \alpha_0$, etc.)

$$\Sigma = \begin{pmatrix} \left(\frac{\log \alpha_0}{1.96}\right)^2 & 0 & 0 \\ 0 & \left(\frac{\log \epsilon_0}{1.96}\right)^2 & 0 \\ 0 & 0 & \left(\frac{\log \rho_0}{1.96}\right)^2 \end{pmatrix} \tag{4.20}$$

2. Propose:

$$(\log \tilde{\alpha}, \log \tilde{\epsilon}, \log \tilde{\rho}) = (\log \alpha_{i-1}, \log \epsilon_{i-1}, \log \rho_{i-1}) + v_i$$

   where $v_i \sim \text{MVN}(\mathbf{0}, \Sigma)$ is drawn from a multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix Σ. To optimise sampling efficiency, before we can begin to use or collect useful samples from the parameter posterior distribution we must first tune Σ - we follow the method used by [PBM15].

3. Estimate:

$$\hat{p}(\mathbf{y} \mid \tilde{\alpha}, \tilde{\rho}) \approx p(\mathbf{y} \mid \tilde{\alpha}, \tilde{\rho})$$
$$\hat{p}(\mathbf{H}, \mathbf{r} \mid \tilde{\alpha}, \tilde{\epsilon}, \tilde{\rho}) \approx p(\mathbf{H}, \mathbf{r} \mid \tilde{\alpha}, \tilde{\epsilon}, \tilde{\rho}) \qquad \text{(Eqs. 4.18, 4.19 \& App. } E\text{)}$$

4. Accept-reject:

   set $(\log \alpha_i, \log \epsilon_i, \log \rho_i) = (\log \tilde{\alpha}, \log \tilde{\epsilon}, \log \tilde{\rho})$

   with probability $\dfrac{\hat{p}(\mathbf{y} \mid \tilde{\alpha}, \tilde{\rho})\, \hat{p}(\mathbf{H}, \mathbf{r} \mid \tilde{\alpha}, \tilde{\epsilon}, \tilde{\rho})\, p(\tilde{\alpha}, \tilde{\epsilon}, \tilde{\rho})\tilde{\alpha}\tilde{\epsilon}\tilde{\rho}}{\hat{p}(\mathbf{y} \mid \alpha_{i-1}, \rho_{i-1})\, \hat{p}(\mathbf{H}, \mathbf{r} \mid \alpha_{i-1}, \epsilon_{i-1}, \rho_{i-1})\, p(\alpha_{i-1}, \epsilon_{i-1}, \rho_{i-1})\alpha_{i-1}\epsilon_{i-1}\rho_{i-1}}$

   otherwise, $(\log \alpha_i, \log \epsilon_i, \log \rho_i) = (\log \alpha_{i-1}, \log \epsilon_{i-1}, \log \rho_{i-1})$.

If $i \leq N_{\text{burn}}$, then if accepted set

$$\Sigma = 1.002 \times \Sigma \tag{4.21}$$

and if rejected set

$$\Sigma = 0.999 \times \Sigma. \tag{4.22}$$

In addition, while $i \leq N_{\text{burn}}$, at every N_{rescale} iterations (usually between 750 and 1000), set the elements of Σ to the variances and covariances of the sequences of the log-transformed parameter samples. This step adapts the proposal density to information we have gained so far about the shape of the joint posterior, improving the acceptance rate of the MCMC chains.

4.2.4 Model checking

The least we ought to expect from a fitted model is that the data are not viewed as unlikely with regards to the posterior distribution. Graphical posterior predictive checks (GPPC) [GCSR95, GS13] are a standard model checking tool we use to compare a fitted model's predictions with the data it is fitted to.

Given $(\alpha_i, \epsilon_i, \rho_i), i = 1, 2, 3, \ldots$ sampled approximately from $p(\alpha, \epsilon, \rho \,|\, \mathbf{H}, \mathbf{r}, \mathbf{y})$, as described above, we simulate new data $\hat{\mathbf{H}}_i, \hat{\mathbf{r}}_i, \hat{\mathbf{y}}_i$ and compare these graphically with $\mathbf{H}, \mathbf{r}, \mathbf{y}$.

4.2.5 Data simulation

We describe below how the simulated data sets for Experiments A and B are obtained. The fitted SI-PQ model assumes a constant rate of pathogen shedding during each host's period of infectiousness, this rate being uniform across the population. However, in order to show that the model is robust to the kinds of complexities in shedding kinetics similar to those observed by [DRP+17], we also allow each infectious host's rate of shedding to vary over time. In our simulation of results from Experiment B we introduce a small amount of variability into the simulation parameters for each container and repetition.

4.2.5.1 Simulation routine for Data Set A

Table 4.4 summarises the simulated data set for Experiment A (Data Set A), for a trial commencing at $t = 0$ with $N_1 = 20$ inncoulated hosts in Group 1 and $N_2 = 20$ disease-naive hosts in Group 2. Five samples of four hosts from each of Groups 1 and 2 are taken at 24 hourly intervals, starting at $t = 24\,\text{h}$. Additionally a dPCR reading is taken with $C = 12 \times 765 = 9180$ chambers, each of volume 6 nl, as in [DQR08]. Samples are taken from Group 1 only to conform with the original trial design of [HVT+17]; we are not concerned with the disease status of hosts in these samples.

Parameter(s)	**description (experiment)**	**value(s)**
α	env. transmission	$5.0 \times 10^{-10}\,\mathrm{h}^{-1}$
a, b, σ	time-varying pathogen shedding rate	$482.36\,\mathrm{h}^{-1}$, 3.4, 2×10^{3}
ρ	pathogen decay	$0.005\,\mathrm{h}^{-1}$
η	water filtration (A)	$0.1\,\mathrm{h}^{-1}$
λ	stochastic fluctuation of pathogen load	1.0×10^{3}
N_1, N_2	initial size Group 1 & 2 (A)	20, 20
V_t	tank volume (A)	150 l
V_c	dPCR chamber volume (A)	6 nl
C	number of dPCR chambers (A)	9180
p_0	mean initial pathogen load per container (B)	1×10^{7}
T	trial duration (B)	432 h

Table 4.3: **Summary of values used to simulate Data Sets A & B.**

Time (hrs)	r (/4)	H (/9180)
24.0	0	64
48.0	3	65
72.0	2	57
96.0	2	40
120.0	3	30

Table 4.4: **Simulated data set for Experiment A - "Data Set A".** Initially, there are $N_1 = 20$ hosts in the innoculate group (Group 1) and $N_2 = 20$ in the disease-naive group (Group 2). Simulation parameters are in Table 4.3. At the times indicated, 4 hosts are sampled at random from Group 1, and 4 are sampled from Group 2. Those from Group 2 are tested for disease and the number of positive tests, r, is record. At the same time, a dPCR measurement, H, of pathogen load in the tank water is taken, with $C = 9180, V_\mathrm{c} = 6\,\mathrm{nl}$.

For the purposes of simulating data to test our inference approach we allow a time-dependent rate of shedding,

$$\epsilon(t) = \frac{a}{\Gamma(b)\sigma^b} t^{b-1} e^{-\frac{t}{\sigma}} \tag{4.23}$$

for each infected host, where t is time since infection. Our choice of functional form for $\epsilon(t)$, and the values for the parameters a, b and σ, allow for an initial increase in the rate of shedding up to a peak at 48.0 h post-infection, at the rate of $1.0 \times 10^6\,\mathrm{part\,h}^{-1}$ before falling away slowly - see Figure 4.2). Studies quantifying the shedding kinetics of various pathogens are rare (but see [WSKK17] and [DRP+17]) but a general pattern of shedding initially increasing with time, post-infection, and then falling back away, is reasonable.

Rather than initially innoculating all Group 1 hosts precisely at $t = 0$, for each host i in Group 1, we draw τ_i from an exponential distribution of mean $0.5\,\mathrm{h}^{-1}$ and set host i's shedding rate from the beginning of the trial until they are sampled to be

$$\epsilon_i(t) = \epsilon(t + \tau_i) \quad t \geq 0 \tag{4.24}$$

which is equivalent to innoculating each host in Group 1, on average, half an hour prior to

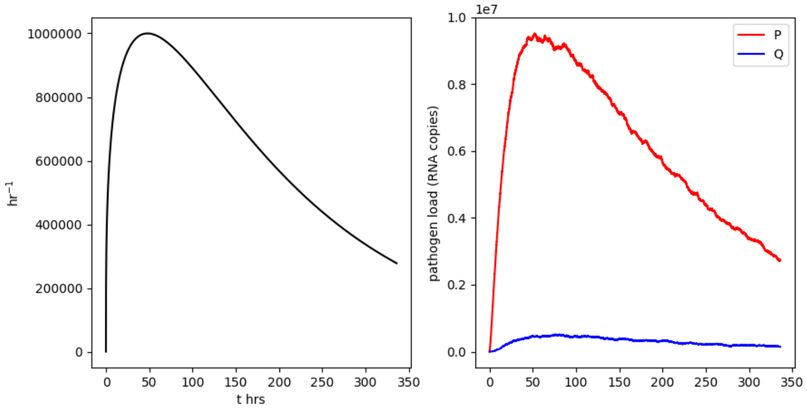


Figure 4.2: **Time-dependent rate of pathogen shedding**, $\epsilon(t)$, (left panel) in simulation of Data Set A. In the right panel is a corresponding typical P-Q pathogen trajectory of one host with the parameter values selected above ($\rho = 0.005, \eta = 0.1$).

the start of the trial. Values chosen for the model parameters for the simulations are $\alpha = 5.0 \times 10^{-10}, \rho = 0.005, \eta = 0.1$ and $\lambda = 1.0 \times 10^{-3}$ and the experimental tank volume is $V_\mathrm{t} = 150\,\mathrm{l}$.

After intialisation, simulation of Data Set A proceeds by alternately forward-simulating the underlying SI-PQ process (with time-varying total rate of shedding) *between* each subsequent pair of sampling times, and then when a sampling time is reached, simulating the random sample and testing, as well as the dPCR measurement before making the appropriate adjustments to the underlying state, to account for destructive sampling. We now describe the process in detail.

1. **Initialise**:

   a) Fix sampling times $0 < t_1 < \cdots < t_n$ and Group 1 and 2 sample sizes $(d_1^1, \ldots, d_n^1)$ and $(d_1^2, \ldots, d_n^2)$

   b) choose a stepping distance h - this should be several orders of magnitude smaller than the inter-sampling times $t_i - t_{i-1}$

   c) set state variables $S_0 = N_2, I_0^1 = N_1, I_0^2 = 0$ (S are susceptibles, I^1 and I^2 correspond to infectious hosts in Groups 1 and 2)

   d) Initialise shedding rates for Group 1. For $i = 1, \ldots, N_1$, draw

   $$\tau_i \sim \text{Exponential}(2.0) \tag{4.25}$$

   The shedding rate for Group 1 host number i, until it is removed, is $\epsilon_i(t) = \epsilon(t + \tau_i)$. This models the fact that Group 1 are not innoculated precisely at $t = 0$, but on average half an hour prior to then (due to the choice of rate in Eq. 4.25).

2. **Simulate between sampling times,** t_{i-1} **and** t_i. Divide the interval $[t_{i-1}, t_i]$ into small time steps $t_{i-1} = s_0 < s_1 < \cdots < s_n = t_i$ of equal duration h. For each time step

   a) sum all the current shedding rates for all infected hosts and set

   $$\epsilon = \sum \epsilon_i(s_j) \tag{4.26}$$

b) sample $P_{s_{j+1}}$ and $Q_{s_{j+1}}$ using an Euler-Maruyama step:

i. Draw independently

$$\Delta^E, \Delta^D, \Delta^{F_P} \text{ and } \Delta^{F_Q} \sim \text{Normal}(0, \lambda h) \tag{4.27}$$

ii. Set

$$\begin{aligned}
\Delta P_{s_j} &= (\epsilon(I^1_{s_j} + I^2_{s_j}) - \rho P_{s_j} - \eta P_{s_j})h + \sqrt{\epsilon(I^1_{s_j} + I^2_{s_j})}\Delta^E - \sqrt{\rho P_{s_j}}\Delta^D - \sqrt{\eta P_{s_j}}\Delta^{F_P} \\
\Delta Q_{s_j} &= (\rho P_{s_j} - \eta Q_{s_j})h + \sqrt{\rho P_{s_j}}\Delta^D - \sqrt{\eta Q_{s_j}}\Delta^{F_Q} \\
P_{s_{j+1}} &= \max\{0, P_{s_j} + \Delta P_{s_j}\} \\
Q_{s_{j+1}} &= \max\{0, Q_{s_j} + \Delta Q_{s_j}\}
\end{aligned} \tag{4.28}$$

c) an infection occurs in the interval $(s_j, s_{j+1}]$ with approximate hazard $\alpha P_{s_j} S_{s_j}$, therefore

i. draw $w \sim \text{Exponential}(\alpha P_{s_j} S_{s_j})$

ii. if $s_j + w < s_{j+1}$ then an infection in Group 2 has occurred. Set

$$\begin{aligned}
S_{s_{j+1}} &= S_{s_j} - 1 \\
I^2_{s_{j+1}} &= I^2_{s_j} + 1
\end{aligned} \tag{4.29}$$

otherwise,

$$\begin{aligned}
S_{s_{j+1}} &= S_{s_j} \\
I^2_{s_{j+1}} &= I^2_{s_j}.
\end{aligned} \tag{4.30}$$

3. **Simulate measurements at sample time t_i**

a) Take dPCR measurement:

$$\begin{aligned}
&\text{Draw } H_i \sim \text{Binomial}(C, \theta_i) \\
&\text{where } \theta_i = 1 - \exp\left(-\frac{V_\text{c}}{V_\text{t}}(P_{t_i} + Q_{t_i})\right)
\end{aligned} \tag{4.31}$$

b) Remove d^1_i hosts from Group 1 (all infected)

$$I^1_{t_i} = I^1_{t_i} - d^1_i \tag{4.32}$$

and choose d^2_i hosts from Group 2 at random. With r_i infectious out of sample of size d^2_i, adjust the state variables

$$\begin{aligned}
S_{t_i} &= S_{t_i} - (d^2_i - r_i) \\
I^2_{t_i} &= I^2_{t_i} - r_i - d^1_i
\end{aligned} \tag{4.33}$$

Time (hrs)	y (/10)
0.0	6
48.0	7
96.0	6
144.0	3
192.0	4
240.0	1
288.0	1
336.0	1
384.0	0

Table 4.5: **Simulated data set for Experiment B - "Data Set B"**. Ten repetitions of the Experiment B are performed with nine containers each. Repetition j of container i is prepared with an initial viral concentration of $p_{0,ij}$ part ml^{-1}. Table lists the times that hosts are added to the containers and the counts (out of 10) of infected hosts found therein at the end of the trial at $T = 432$ h.

4.2.5.2 Simulation routine for Data Set B

Table 4.5 contains Data Set B, which are the counts of infected hosts from $k = 10$ repetitions of Experiment B. Simulation of Data Set is much more straightforward. At each repetition of the trail, hosts are introduced into one of nine containers starting at $t = 0$ and every 48 h thereafter. The trial then continues until $T = 432.0$ h.

Each container within each repetition has its own values for the parameters α and ρ, as well as the initial pathogen load, p_0, drawn as

$$\begin{aligned} \log \alpha_{ij} &\sim \mathrm{N}\left(\log(5.0 \times 10^{-10}), 0.05\right) \\ \log \rho_{ij} &\sim \mathrm{N}\left(\log(0.005), 0.05\right) \\ \log_{10} p_{0,ij} &\sim \mathrm{N}\left(\log_{10}(1 \times 10^{7}), 0.05\right) \end{aligned} \tag{4.34}$$

The means of the normal distributions above match the values used to simulate Dataset A, and the variances are chosen to introduce a small amount of noise into the simulation.

Whether container i contains an infected host at repetition j is simply a draw from a Bernoulli trial with probability θ_{ij}, where

$$\theta_{ij} = 1 - \exp\left\{-\frac{\alpha_{ij} p_{0,ij}}{\rho_{ij}}\left(e^{-\rho_{ij} t_{0,i}} - e^{-\rho_{ij} T}\right)\right\} \tag{4.35}$$

(see Eq. 4.16).

4.3 Results

Our aim in this chapter is to show that, using particle-marginal MCMC (PM-MCMC), we can estimate the SI-PQ model rates of environmental transmission, α, pathogen shedding, ϵ, and decay, ρ, from the kinds of data that are routinely collected in challenge studies of waterborne infectious disease, augmented with dPCR.

Figure 4.3 summarises the samples drawn from the posterior distribution resulting from updating the flat prior $p(\alpha, \epsilon) \sim 1$ with simulated data from Experiment A, under the assumption that $\rho = 0$. This is equivalent to assuming that all of the pathogen in the tank is viable, which is a reasonable approximation for systems in which only a small amount of the measured pathogen load corresponds to dead pathogen, as is the case when the rate of removal of pathogen due to water filtration, η, is sufficiently greater than ρ (see Eq. 4.9). In order to assess the effect of the particular value we choose for the scale parameter λ, we perform two repetitions of this exercise with (a) $\lambda = 1 \times 10^3$ and (b) $\lambda = 1 \times 10^4$. Experiment A is based on the trial conducted by [HVT$^+$17] with the addition of five dPCR measurments accompanying the taking of host samples. We see Data Set A is very informative about the two quantities α and ϵ since we get posterior estimates with bounded support that are at the correct orders of magnitude in relation to the known values chosen for simulation, as listed in Table 4.3, and that inferences for α and ϵ are not affected by the value of λ. Figure 4.3 also contains trace plots demonstrating the rapid convergence of two independent sampling chains given different initial values. Such rapid convergence is highly desirable, since "slow mixing" can be a bar to practical usage of certain MCMC sampling routines. Visual inspection of trace plots is the primary means of checking convergence is achieved, however, the Gelman and Rubin convergence diagnostic [R C18, PBCV06, GR92] (not reported), showed no evidence of lack of convergence in any of the cases of model fitting.

In Figure 4.4 a graphical posterior-predictive check (GPPC) comparing this fitted model's predictions of the postive sample numbers and dPCR readings with those of Data Set A is presented, drawing parameter values from the first of the two sets of posterior samples drawn, with $\lambda = 1 \times 10^3$. From this we see that the predicted dPCR readings agree well with the data. However, the host samples are somewhat under-predicted from $t = 48\,\text{h}$ onwards, where a better-fitting model would correctly predict these quantities.

Figures 4.5 and 4.6 similarly show samples from the posterior of all three parameters after updating $p(\alpha, \epsilon, \rho) \sim 1$ this time with data additionally from Data Set B, which is a simulation of a second hypothetical trial similar in design to that of [SAR$^+$13]. Again, we perform this exercise twice with $\lambda = 1 \times 10^3$ (Fig. 4.5) and $\lambda = 1 \times 10^4$ (Fig. 4.6). Table 4.6 lists sample statistics for all marginal parameter posteriors. In addition to an estimate of the rate of pathogen decay, which was not available to us with data from Experiment A alone, we have narrower, more precise, estimates of α and ϵ, with the difference between the 5 and 95 percentiles reducing from $1.8 \times 10^{-10}\,\text{h}$ to $1.4 \times 10^{-10}\,\text{h}$ in the case of α and from $4.1 \times 10^5\,\text{h}$ to $3.6 \times 10^5\,\text{h}$ for ϵ. What is more, the GPPC, presented in Figures 4.7 and 4.8, with samples drawn from the first set of posterior samples, with $\lambda = 1 \times 10^3$ shows an improvement in terms of predicted positive sample sizes.

The crosses on the univariate and bivariate density subplots in Figures 4.3, 4.5 and 4.6 indicate the values of α and ρ (where appropriate) used to simulate the data sets (see Table 4.3) as well as the *peak* rate of pathogen emission of $1 \times 10^6\,\mathrm{h}^{-1}$. Since in our simulations, the pathogen emission rate was assumed to be variable, rising rapidly to a peak of $1 \times 10^6\,\mathrm{h}^{-1}$ and then declining, one would expect the inferred value for a *steady* rate of pathogen emission (as implied by the fitted SIR-PQ model) to lie somewhere below this peak value, which is the case here.

Another interesting feature of the fitted posteriors in Figures 4.5 and 4.6 is the marked correlation between the parameters α and ρ. Such a correlation suggests that appropriately increasing or decreasing the environmental transmission and pathogen decay rates together produces a range of models consistent with the data. This correlation is also an artefact of the of lack model identifiability when attempting to infer values for the three parameters α, ϵ and ρ from Data Set A alone. Addition of a second source of information, such as Data Set B, allows us to learn something about ρ, independently of α, and narrows down the plausible ranges of values for these two parameters, albeit with some residual correlation.

		mean	**5%-ile**	**median**	**95%-ile**
(i) A	$p(\alpha \,\vert\, \mathbf{H}, \mathbf{r})$	1.3×10^{-10}	0.4×10^{-10}	1.2×10^{-10}	2.2×10^{-10}
	$p(\epsilon \,\vert\, \mathbf{H}, \mathbf{r})$	9.5×10^{5}	7.6×10^{5}	9.4×10^{5}	11.7×10^{5}
(ii) A & B	$p(\alpha \,\vert\, \mathbf{H}, \mathbf{r}, \mathbf{y})$	2.1×10^{-10}	1.5×10^{-10}	2.1×10^{-10}	2.9×10^{-10}
	$p(\epsilon \,\vert\, \mathbf{H}, \mathbf{r}, \mathbf{y})$	9.3×10^{5}	7.5×10^{5}	9.2×10^{5}	11.1×10^{5}
	$p(\rho \,\vert\, \mathbf{H}, \mathbf{r}, \mathbf{y})$	1.0×10^{-3}	0.1×10^{-3}	1.0×10^{-3}	3.0×10^{-3}

Table 4.6: **Summary statistics** (mean, median, 5 and 95 percentiles) **of marginal posterior distributions** resulting from fitting SI-PQ model to (i) Data Set A and (ii) Data Sets A and B with $\lambda = 1 \times 10^3$. All units are h^{-1}.

4.4 Discussion

In this chapter we have developed and implemented a Bayesian methodology that enables the inference of the parameters of an ETM, as well as shown what kinds of data are required. While our focus has been on the SI-PQ model, the methods used here are generic and very versatile. The PM-MCMC routine can be implemented in `C++` in a generic way, allowing, for example, extra compartments to the SI-PQ model to be added without significant recoding. The forward simulations of particles can be run in batches within independent threads of execution, significantly speeding up the PM-MCMC routine on a multi-core machine. The required data can be collected using experimental protocols already in use in aquaculture studies.

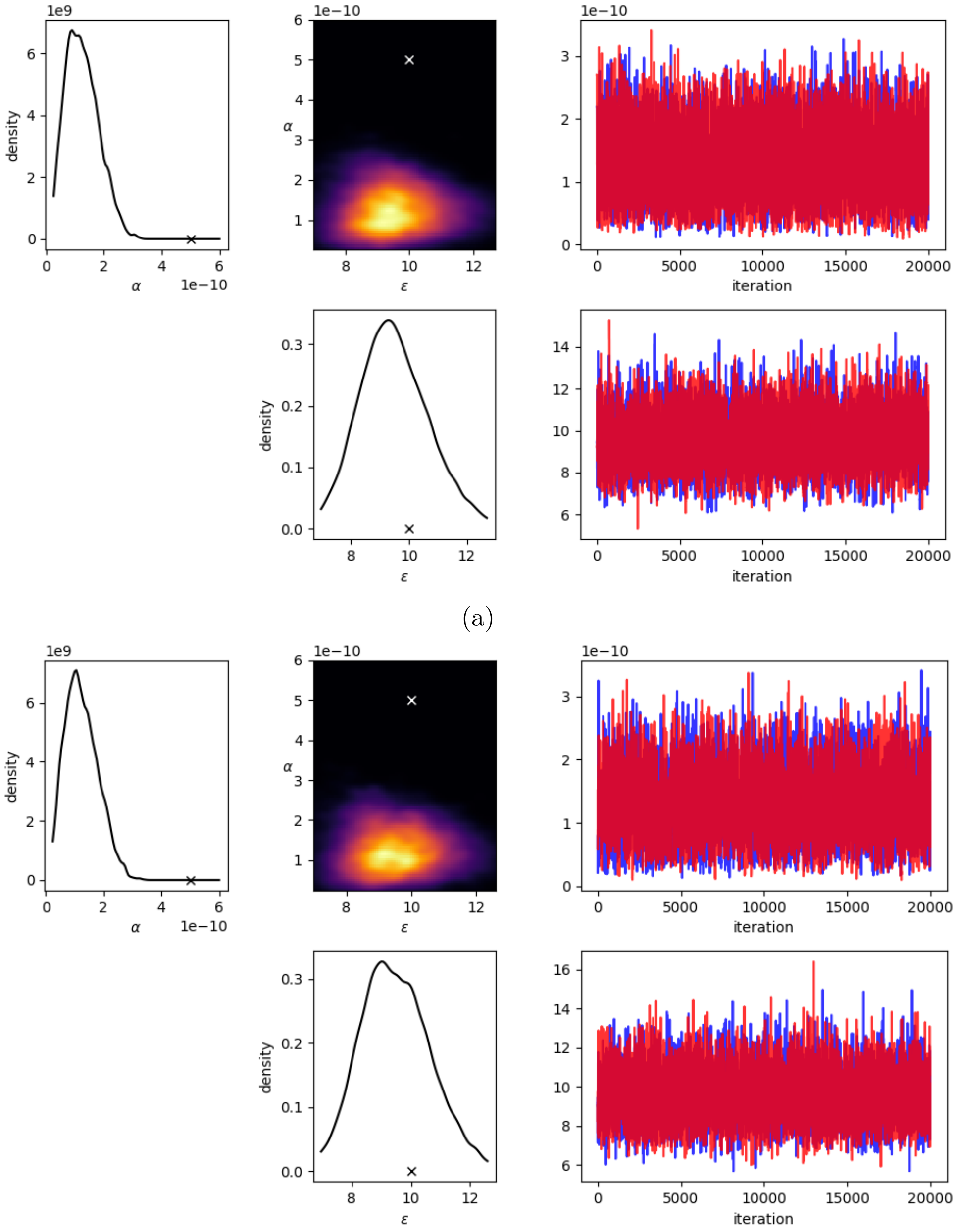
By making appropriate changes to the forward simulations of the particles as part of the PM-MCMC routine (see Appendix E) we may easily include a latently infected (E) state within a SEI-PQ model, corresponding to a per-host shedding rate that remains low for an initial period before growing to a peak. If the shedding rate falls away quickly after reaching a peak then we may also include a recovered (R) state in a SIR-PQ, or even a SEIR-PQ model.

By making appropriate changes to how we adjust host states we can also expand upon the observation model presented here. For example, a method of non-lethal testing (e.g. gill and skin swabs) would be modelled in the same way but without reductions to the numbers of hosts following the measurement, as we do in the case studied here. We can also allow for the possibilities of false positives or false negatives in testing by making adjustments to the host states, after sampling and reweighting, subject to a probabilistic rule determined by the test's *sensitivity* and *specificity*.

What should be considered carefully, however, is whether the data collected and the observation model allow all of the parameters to be estimated. For example, the pathogen decay rate cannot be estimated using data from Experiment A alone and although in some circumstances loss of pathogen viability can be ignored, where it cannot transmission studies such as Experiment A can be supplemented with relatively straightforward experiments that enable the rate of viability loss to be estimated, such as Experiment B. Fitting models to simulations of trials (as we have done in this Chapter and in Chapter 3) is a low-cost way of detecting any model identifiability problems ahead of committing expensive resources.

The SI-PQ parameter λ governs the expected size of the stochastic fluctuations of both parts of the pathogen load around the solution to the deterministic part of Eq. 4.2. This was introduced so that these equations hold after rescaling the size of P_t and Q_t, e.g. to reflect different scales on which pathogen load can be measured. If we set $\lambda = 0$, then conditional upon I_t, P_t and Q_t become deterministic functions of t that responds to changes in the number of shedding hosts. Conversely, as λ grows, the pathogen load gets increasingly "noisy" and the fluctuations of P_t and Q_t dominate the pathogen dynamics. We have not attempted here to estimate this quantity from data, treating λ instead as a nuisance parameter and fixing its value when fitting the model. The value at which λ is fixed does not affect the inferences about the other parameters. A suggested strategy when model fitting is first to set λ either to zero or to some small value, and then to increase it if problems are encountered either in the assessment of the model fit or in the model fitting prodedure itself, e.g. slow mixing of the MCMC chains.

ETMs, informed by data, offer a deeper understanding of the relationship the dynamics of environmental pathogen levels and the risk of outbreaks of environmentally transmitted diseases, such as cholera, paratuberculosis and the many aquatic infectious diseases affecting aquaculture production. Real-time waterborne pathogen detection technologies will boost surveillance for aquatic infectious diseases in aquaculture production settings. The tools we have begun to develop here enable such impacts to be quantified and opens up the environmental transmission of disease to a more granular analysis of its constituent sub-processes, e.g. shedding kinetics and pathogen viability in the environment.



(a)

(b)

Figure 4.3: **Summary of posterior distribution of environmental transmission and pathogen emission rates, $p(\alpha, \epsilon \,|\, \mathbf{H}, \mathbf{r})$, and MCMC trace plots** for SI-PQ model fitted to Data Set A with (a) $\lambda = 1 \times 10^3$ and (b) $\lambda = 1 \times 10^4$. Axes are rescaled by 1×10^5 for ϵ. Univariate marginal posterior density estimates are shown on the diagonal along with full bivariate density estimate. Data Set A was simulated with $\alpha = 5 \times 10^{-10}\,\mathrm{h}^{-1}$, $\rho = 0.005\,\mathrm{h}^{-1}$ and variable pathogen emission peaking at $1 \times 10^6\,\mathrm{h}^{-1}$. Crosses are used to indicate these values on the plots for comparison and we see that posterior estimates for these parameters are at the correct order of magnitude. In order to estimate these two parameters from Data Set A alone, we made the approximating assumption that $\rho = 0$ (for the simulation we set $\rho = 0.005\,\mathrm{h}, \eta = 0.1\,\mathrm{h}$, i.e. $\rho \ll \eta$). The fitted model, however, still provides good estimates of the parameters α and ϵ. The trace plots indicate that two chains run independently with different initial values have converged to a common stationary distribution after discarding 5000 iterations from each chain for burn-in and adaptation of the proposal density.

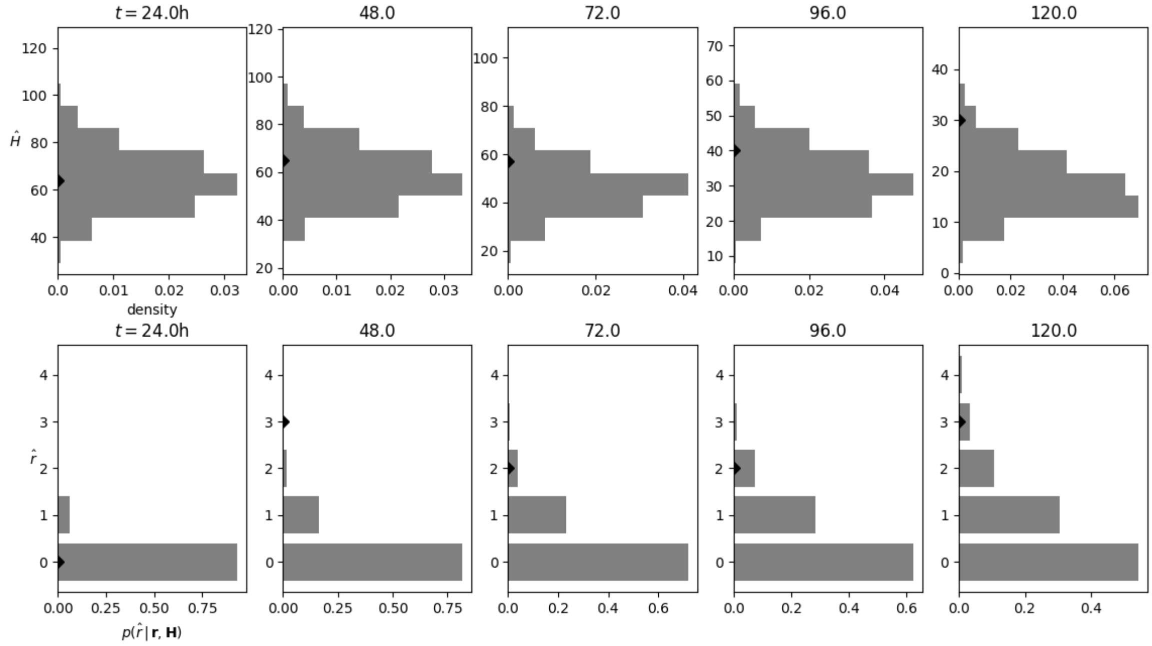


Figure 4.4: **Marginal posterior predictive distributions for dPCR readings H and positive test results r,** $p(\hat{H}_i \,|\, \mathbf{H}, \mathbf{r})$ **and** $p(\hat{r}_i \,|\, \mathbf{H}, \mathbf{r})$ corresponding to samples drawn from posterior summarised in Fig. 4.3a above with $\lambda = 1 \times 10^3$. The numbers of positive-testing hosts in each sample and dPCR readings comprising Data Set A are indicated by ⧫ on the plot. We see that the posterior-predictive distribution of the dPCR counts are roughly centered upon the values in Data Set A (apart from the reading at $t = 120\,\text{h}$), however, fewer positive samples from $t = 48\,\text{h}$ onwards are predicted by the model than were actually observed. This suggests that another model might provide a better fit. To sample from the posterior-predictive distributions, 5×10^4 evenly spaced values were taken from the MCMC parameter samples.

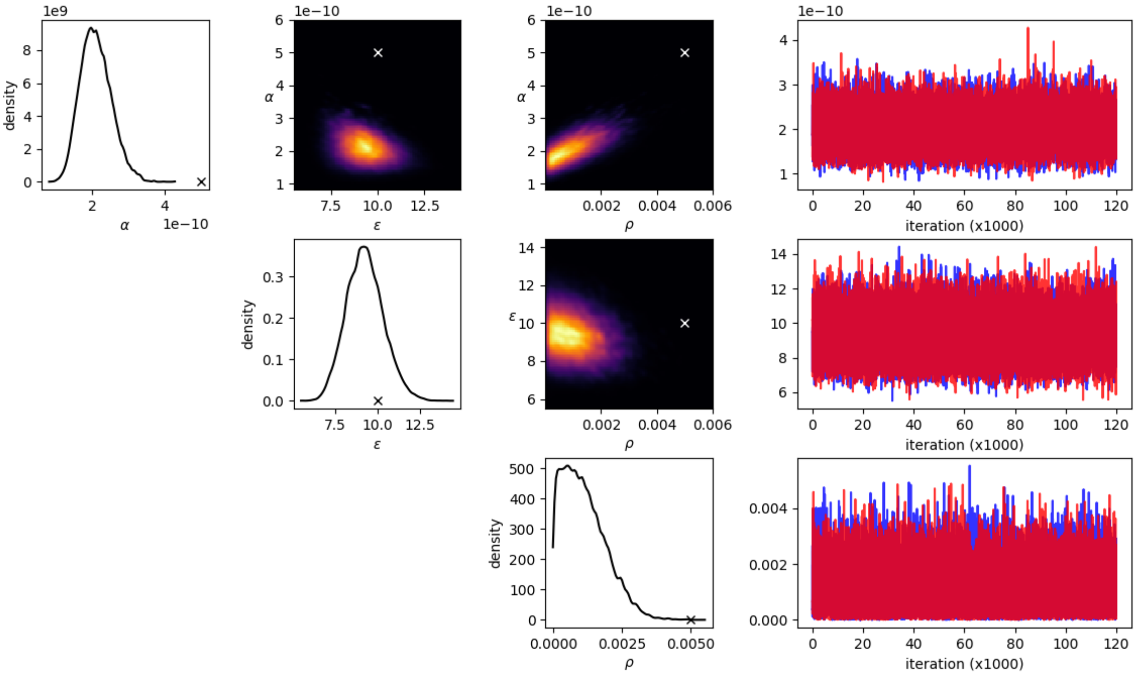


Figure 4.5: **Summary of posterior distribution of environmental transmission, pathogen decay and emission rates, $p(\alpha, \epsilon, \rho \,|\, \mathbf{H}, \mathbf{r}, \mathbf{y})$, and MCMC trace plots** for SI-PQ model fitted to Data Sets A and B with $\lambda = 1 \times 10^3$. Axes are rescaled by 1×10^5 for ϵ. Having begun with flat priors for all parameters, $p(\alpha, \epsilon, \rho) \propto 1$ (see Sec. 4.2.3), we see that Data Sets A and B together are informative, and lead to estimates of all three unknown parameters of the model. Data Set A and B were simulated with $\alpha = 5 \times 10^{-10}\,\mathrm{h}^{-1}$, $\rho = 0.005\,\mathrm{h}^{-1}$ and variable pathogen emission peaking at $1 \times 10^6\,\mathrm{h}^{-1}$. Crosses are used to indicate these values on the plots for comparison and we see that posterior estimates for these parameters are at the correct order of magnitude. As before, the trace plots are shown to indicate the rapid mixing of the two MCMC chains (started separately and run independently). Between 10 and 15 thousand iterations were discarded before calculating each of the above density estimates.

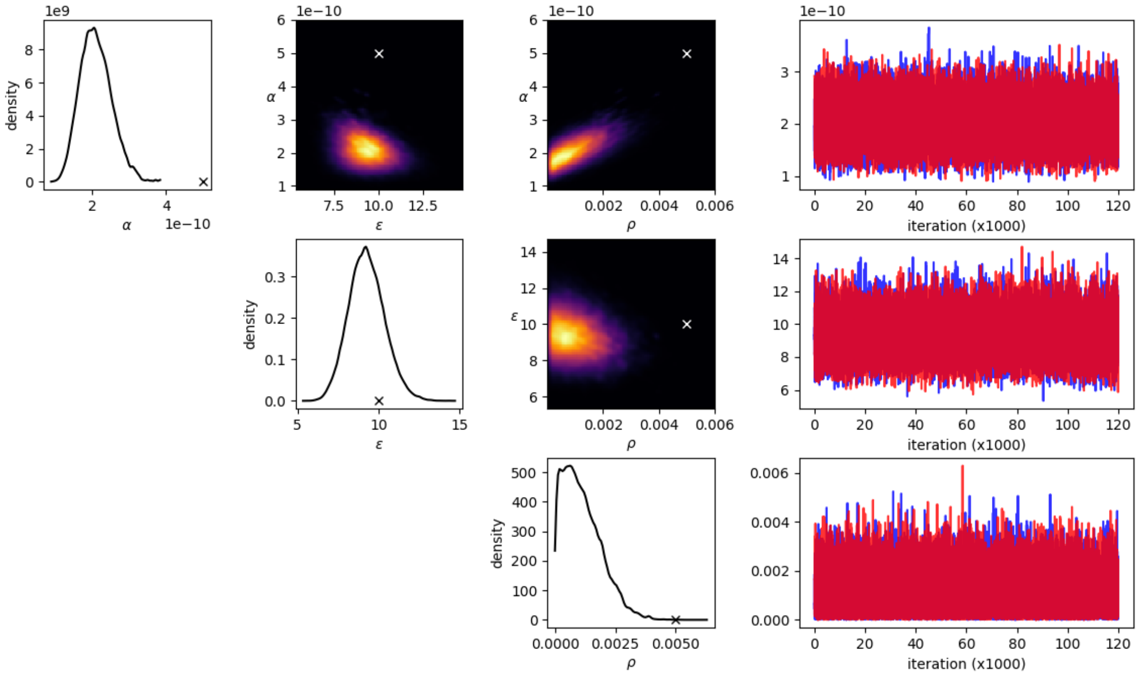


Figure 4.6: **Summary of posterior distribution of environmental transmission, pathogen decay and emission rates, $p(\alpha, \epsilon, \rho \,|\, \mathbf{H}, \mathbf{r}, \mathbf{y})$, and MCMC trace plots** for SI-PQ model fitted to Data Sets A and B with $\lambda = 1 \times 10^4$. Axes are rescaled by 1×10^5 for ϵ. Having begun with flat priors for all parameters, $p(\alpha, \epsilon, \rho) \propto 1$ (see Sec. 4.2.3), we see that Data Sets A and B together are informative, and lead to estimates of all three unknown parameters of the model. Data Set A and B were simulated with $\alpha = 5 \times 10^{-10}\,\mathrm{h}^{-1}$, $\rho = 0.005\,\mathrm{h}^{-1}$ and variable pathogen emission peaking at $1 \times 10^6\,\mathrm{h}^{-1}$. Crosses are used to indicate these values on the plots for comparison and we see that posterior estimates for these parameters are at the correct order of magnitude. As before, the trace plots are shown to indicate the rapid mixing of the two MCMC chains (started separately and run independently). Between 10 and 15 thousand iterations were discarded before calculating each of the above density estimates.

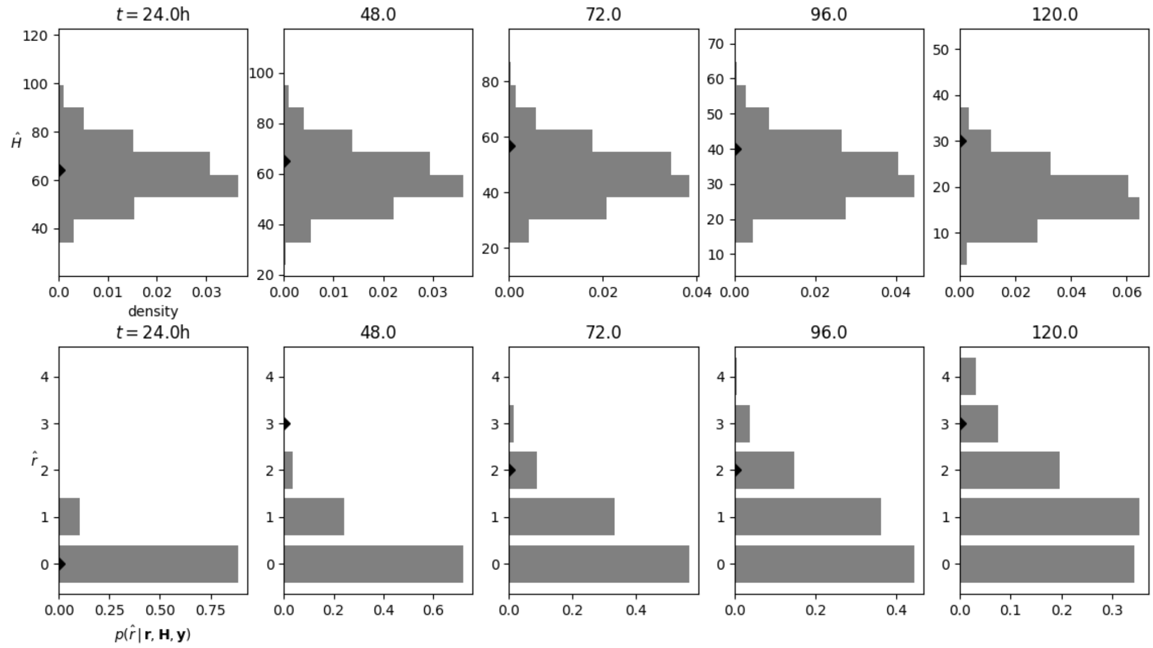


Figure 4.7: **Marginal posterior predictive distributions for dPCR readings H and positive test results** $p(\hat{H}_i \,|\, \mathbf{H}, \mathbf{r}, \mathbf{y})$ **and** $p(\hat{r}_i \,|\, \mathbf{H}, \mathbf{r}, \mathbf{y})$ corresponding to samples drawn from posterior summarised in Fig. 4.5 above, with $\lambda = 1 \times 10^3$. The numbers of positive-testing hosts in each sample and dPCR readings comprising Data Set A are indicated by ⧫ on the plot. As before, the model's predictions of the dPCR readings are in good agreement with those of Data Set A. However, we see an improvement in the model's predictions of the numers of positive-testing fish in the samples at $t = 72, 96$ and $120\,$h.

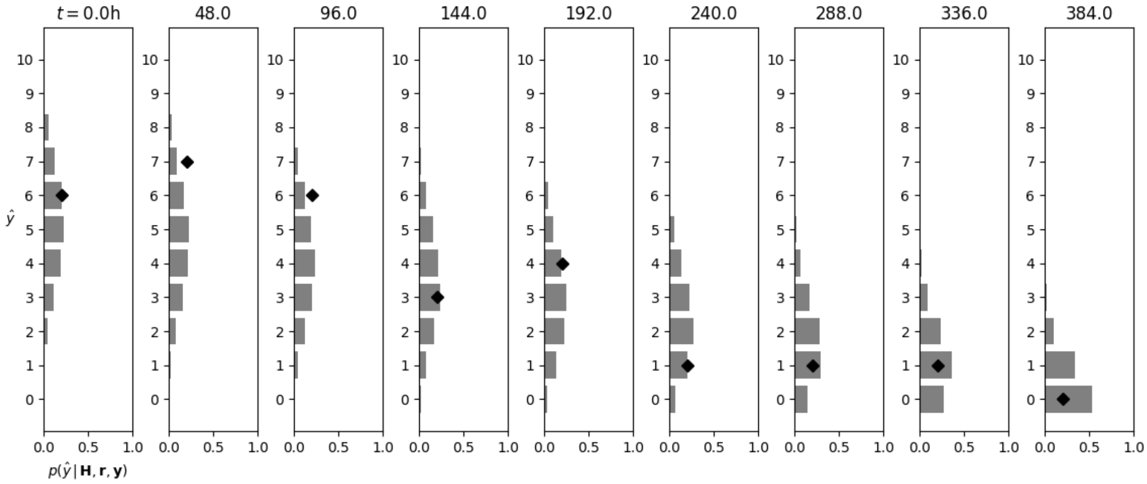


Figure 4.8: **Marginal posterior predictive distributions for positive bucket counts, y,** $p(\hat{y}_i \,|\, \mathbf{H}, \mathbf{r})$ corresponding to corresponding to samples drawn from posterior summarised in Fig. 4.5 above, with $\lambda = 1 \times 10^3$. The numbers of positive testing hosts in each container are indicated by ⧫ on the plots, where we see in each case that the fitted model's predicted counts are similar to those seen.

Chapter 5

Discussion

The aim of this thesis has been to expand and develop the tools employed in stochastic epidemiological modelling and computationally intensive Bayesian inference. A greater understanding of disease spread through mechanistic models, informed by observation, offers greater predictive power, more accurate quantification of the risk of outbreaks and better assessment of the effects of potential treatments and interventions [LC16].

In Chapter 2 we took a critical look at direct transmission (DT) models [McC01] and provided insight into their application to infections with environmental transmission (ET) routes (e.g. [BDSR09, ZLZ^{+}20, PKF^{+}21]). We identified timescale separation between the hosts and pathogens as the factor that determines whether simple DT is an accurate model description, even when the infection is also transmitted via contact with environmental pathogens. The demonstrated robustness of DT models to departure from the DT assumption, and their reduced data requirement (since they model the host-disease dynamics without describing the pathogen load) means these models are a sound and parsimonious choice as a dynamic description of a wide range of host-pathogen interactions.

In Chapter 3 we showed how the large and growing body of observational data obtained from immersion challenge experiments (ICE) can be used to build DT models of aquatic infectious diseases and how computer simulation and an information-theoretic measure of the expected information gain [VTHvR12, LFKS13] can help in the design of such experiments. Furthermore we demonstrated that, using data-informed simulations, researchers can maximise the benefits of their experiments by testing hypotheses about treatments and interventions not included in the original design. Fitted models can help to identify remaining gaps in knowledge, motivating further data collection [LEH^{+}15] and efficient design of future experiments.

Finally, in Chapter 4, by novel application of particle filtering Markov chain Monte Carlo methods, we tackled the problem of estimating the parameters of an ET model. We showed that adding digital polymerase chain reaction (PCR) measurements [DQR08, SJ13] of waterborne pathogen load to common ICE designs yields estimates of important quantities that characterise ET: the rates of environmental transmission, pathogen emission and loss of pathogen viability. By proposing an ET model that differentiates between infectious pathogens and non-infectious pathogenic material (which is also detected by PCR), we addressed the "shedding versus infectivity dilemma" discussed by [WBB20] in relation to SARS-CoV-2. Although, as discussed Chapter 2, in many cases DT models accurately describe ET, there are benefits to considering how pathogen levels affect the likelihood and severity of outbreaks. For example, Wang, et al. (2012) found that a simple, stochastic ET model of avian influenza could explain complex phenomena relating to multi-year patterns of outbreaks in the wild [WJL+12].

There are many promising directions for future work. These include a measure of the departure from DT in the system under study, which would complement the graphical posterior-predictive checks used throughout this thesis. The first place to look would be the *exposure time residuals* (ETR), which are a subclass of more general *latent residuals* (LR) [LMSG14]. The posterior distribution of a sequence of LRs, which are a function of the data, unobserved events and parameters, can be examined in order to assess whether the assumptions made about particular parts of the fitted model are valid. However, the posterior distribution of the ETRs is often high dimensional (one dimension per disease exposure), and requires graphical inspection.

Supporting real-time surveillance for environmental pathogens that have become endemic within the wider host population requires longer-term modelling that includes demographic movements, which in food production can include trading and restocking [BJK+18]. The models considered in this thesis assume that the timescales over which outbreaks occur are sufficiently short that these demographic movements can be neglected. However, by adding extra routes of entry into and exit out of the host and pathogen population compartments, a closed-population ET model with parameter estimates using ICE data may be adapted to include demographic movements, as well as incursions of pathogens into the local environment from outside.

By providing predictions about outbreaks based upon measurements of pathogens in the local water body, ET models can inform real-time disease surveillance and management operations in aquaculture production systems [Sub05]. Models that better describe the relationship between background pathogen levels and the risks of local outbreaks would complement surveillance methods that focus on the detection of cases. By simulating a number of scenarios covering different disease management and surveillance strategies, as in Chapter 2, ET models will help to find optimal trade-offs between outbreak risk reduction and the costs associated with disease surveillance and any consequent control actions undertaken. The development of approaches that assimilate data in real time and provide information to support decision making for disease control, or even provide automated decision making and action, could be an important step in realizing the potential benefits of digital farming [FotUN21].

ET models could additionally be combined with computational fluid dynamics (CFD) models of water currents, such as the one used to study the airborne spread of SARS at the Amoy Gardens housing complex [YLW+04, YQTW14], so that the risks posed from disease outbreaks at neighbouring farms can be modelled and quantified [DD21].

Understanding how environmental conditions affect the epidemiology of ET diseases would significantly improve predictions about disease outbreaks and inform prevention and mitigation strategies. An additional refinement to ET models would be to take into account how epidemiological parameters are affected by environmental factors, such as water temperature (observed to affect the virulence of white spot syndrome virus [SP10]). This would possibly take the form of a simple generalised linear model regression model relating e.g. the pathogen decay rate parameter to explanatory variables. The regression parameters could be quantified using ICE and then plugged into an ET model.

However, the work of this thesis should already serve to encourage epidemiological researchers in aquaculture, and other fields, on two fronts. Firstly, to apply well-known DT models critically, but with confidence in their effectiveness. And secondly, to take full advantage of the increasing richness of observational disease data in order to quantify both DT and ET models using inferential methods, such as particle-marginal MCMC, described here. The result of this will be future disease surveillance and prevention systems that are smarter, more proactive and will ultimately save lives.

Bibliography

[Abb52] Helen Abbey. An Examination of the Reed-Frost Theory of Epidemics. *Human Biology*, 24(3):201, 1952.

[ABD02] S. ALEXANDERSEN, I. BROTHERHOOD, and A. I. DONALDSON. Natural aerosol transmission of foot-and-mouth disease virus to pigs: minimal infectious dose for strain O 1 Lausanne. *Epidemiology and Infection*, 128(2):301–312, apr 2002.

[ADH10] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, jun 2010.

[ADHR04] G. Altekar, S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20(3):407–415, feb 2004.

[AM78] Roy M. Anderson and Robert M. May. Regulation and Stability of Host-Parasite Population Interactions: I. Regulatory Processes. *The Journal of Animal Ecology*, 47(1):219, feb 1978.

[AMT14] Salvador Almagro-Moreno and Ronald K. Taylor. Cholera: Environmental Reservoirs and Impact on Disease Transmission. In *One Health*, volume 1, pages 149–165. ASM Press, Washington, DC, USA, apr 2014.

[And91] William J. Anderson. *Continuous-Time Markov Chains*. Springer Series in Statistics. Springer New York, New York, NY, 1991.

[ARM21] ARM. *Unparalleled Performance for 5G Network Infrastructure*, 2021. url: `https://www.arm.com/solutions/5g/infrastructure`. Last accessed: 2021-03-04.

[ASP14] Consolidated version of ASPA 1986 - GOV.UK. (January):1–40, 2014.

[AWS98] Nectar Aintablian, Pramila Walpita, and Mark H. Sawyer. Detection of Bordetella pertussis and Respiratory Syncytial Virus in Air Samples from Hospital Rooms. *Infection Control and Hospital Epidemiology*, 19(12):918–923, dec 1998.

[Ban14] T. W. Bank. Reducing disease risk in aquaculture (English). Technical Report 88257, World Bank Group, 2014.

[BB10] Steve M. Blevins and Michael S. Bronze. Robert Koch and the ‘golden age’ of bacteriology. *International Journal of Infectious Diseases*, 14(9):e744–e751, sep 2010.

[BBS+15] Christophe Béné, Manuel Barange, Rohana Subasinghe, Per Pinstrup-Andersen, Gorka Merino, Gro-Ingunn Hemre, and Meryl Williams. Feeding 9 billion by 2050 – Putting fish back on the menu. *Food Security*, 7(2):261–274, apr 2015.

[BCCF19] Fred Brauer, Carlos Castillo-Chavez, and Zhilan Feng. *Mathematical Models in Epidemiology*, volume 69 of *Texts in Applied Mathematics*. Springer New York, New York, NY, 2019.

[BDSR09] Romulus Breban, John M. Drake, David E. Stallknecht, and Pejman Rohani. The Role of Environmental Transmission in Recurrent Avian Influenza Epidemics. *PLoS Computational Biology*, 5(4):e1000346, apr 2009.

[Bec89] Niels G. Becker. *Analysis of Infectious Disease Data.* Chapman and Hall/CRC, 1989.

[Bel61] Richard E. Bellman. *Adaptive Control Processes.* Princeton University Press, jan 1961.

[Ber66] Robert H. Berk. Limiting Behavior of Posterior Distributions when the Model is Incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

[BJK+18] Andrew M. Bate, Glyn Jones, Adam Kleczkowski, Rebecca Naylor, Jon Timmis, Piran C. L. White, and Julia Touza. Livestock Disease Management for Trading Across Different Regulatory Regimes. *EcoHealth*, 15(2):302–316, jun 2018.

[BMC+20] Andrew William Byrne, David McEvoy, Aine B. Collins, Kevin Hunt, Miriam Casey, Ann Barber, Francis Butler, John Griffin, Elizabeth A. Lane, Conor McAloon, Kirsty O’Brien, Patrick Wall, Kieran A. Walsh, and Simon J. More. Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ open*, 10(8):e039856, 2020.

[BNS+03] Clive B. Beggs, C. J. Noakes, P. A. Sleigh, L. A. Fletcher, and K. Siddiqi. The transmission of tuberculosis in confined spaces: an analytical review of alternative epidemiological models. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*, 7(11):1015–26, nov 2003.

[BOM17] Cecile Brugere, Dennis Mark Onuigbo, and Kenton Ll Morgan. People matter in animal disease surveillance: Challenges and opportunities for the aquaculture sector. *Aquaculture*, 467:158–169, jan 2017.

[BPK$^+$16] Gorka Bidegain, Eric N. Powell, John M. Klinck, Tal Ben-Horin, and Eileen E. Hofmann. Marine infectious disease dynamics and outbreak thresholds: contact transmission, pandemic infection, and the potential role of filter feeders. *Ecosphere*, 7(4):e01286, apr 2016.

[Bri10] Tom Britton. Stochastic epidemic models: A survey. *Mathematical Biosciences*, 225(1):24–35, may 2010.

[BTKT18] Derek W. Bailey, Mark G. Trotter, Colt W. Knight, and Milt G. Thomas. Use of GPS tracking collars and accelerometers for rangeland livestock production research1. *Translational Animal Science*, 2(1):81–88, apr 2018.

[CAFD09] C. F. Chicco, C. B. Ammerman, J. P. Feaster, and B. G. Dunavant. A Mathematical Theory of Communication. In *Claude E. Shannon*, volume 36, pages 986–993. IEEE, may 2009.

[CCR$^+$12] Stephen P. Carter, Mark A. Chambers, Stephen P. Rushton, Mark D. F. Shirley, Pia Schuchert, Stéphane Pietravalle, Alistair Murray, Fiona Rogers, George Gettinby, Graham C. Smith, Richard J. Delahay, R. Glyn Hewinson, and Robbie A. McDonald. BCG Vaccination Reduces Risk of Tuberculosis Infection in Vaccinated Badgers and Unvaccinated Badger Cubs. *PLoS ONE*, 7(12):e49833, dec 2012.

[CHBCC09] Gerardo Chowell, James M. Hyman, Luís M. A. Bettencourt, and Carlos Castillo-Chavez, editors. *Mathematical and Statistical Estimation Approaches in Epidemiology.* Springer Netherlands, Dordrecht, 2009.

[CRW$^+$20] Tian Mu Chen, Jia Rui, Qiu Peng Wang, Ze Yu Zhao, Jing An Cui, and Ling Yin. A mathematical model for simulating the phase-based transmissibility of a novel coronavirus. *Infectious Diseases of Poverty*, 9(1):1–8, 2020.

[CSBM20] Francesco Chirico, Angelo Sacco, Nicola Luigi Bragazzi, and Nicola Magnavita. Can Air-Conditioning Systems Contribute to the Spread of SARS/MERS/COVID-19 Infection? Insights from a Rapid Review of the Literature. *International Journal of Environmental Research and Public Health*, 17(17):6052, aug 2020.

[DAS$^+$21] Sandor Dudas, Renee Anderson, Antanas Staskevicus, Gordon Mitchell, James C. Cross, and Stefanie Czub. Exploration of genetic factors resulting in abnormal disease in cattle experimentally challenged with bovine spongiform encephalopathy. *Prion*, 15(1):1–11, jan 2021.

[DD21] Talib Dbouk and Dimitris Drikakis. Fluid dynamics and epidemiology: Seasonality and transmission dynamics. *Physics of Fluids*, 33(2):021901, feb 2021.

[DFG01] Arnaud Doucet, Nando Freitas, and Neil Gordon, editors. *Sequential Monte Carlo Methods in Practice.* Springer New York, New York, NY, 2001.

[DGB$^+$19] Francesco Di Ruscio, Giorgio Guzzetta, Jørgen Vildershøj Bjørnholt, Truls Michael Leegaard, Aina Elisabeth Fossum Moen, Stefano Merler, and Birgitte Freiesleben de Blasio. Quantifying the transmission dynamics of MRSA in the community and healthcare settings in a low-prevalence country. *Proceedings of the National Academy of Sciences*, 116(29):14599–14605, jul 2019.

[DHB12] Odo Diekmann, Hans Heesterbeek, and Tom Britton. *Mathematical Tools for Understanding Infectious Disease Dynamics.* Princeton University Press, nov 2012.

[Die00] O Diekmann. *Mathematical epidemiology of infectious diseases : model building, analysis, and interpretation.* Wiley series in mathematical and computational biology. Wiley, Chichester, 2000.

[DMW$^+$12] Ross S. Davidson, Iain J. McKendrick, Joanna C. Wood, Glenn Marion, Alistair Greig, Karen Stevenson, Michael Sharp, and Michael R. Hutchings. Accounting for uncertainty in model-based prevalence estimation: paratuberculosis control in dairy herds. *BMC Veterinary Research*, 8(1):159, 2012.

[Dow16] Allen B. Downey. *The Little Book of Semaphores.* Green Tea Press, 2016.

[DQR08] Simant Dube, Jian Qin, and Ramesh Ramakrishnan. Mathematical Analysis of Copy Number Variation in a DNA Sample Using Digital PCR on a Nanofluidic Device. *PLoS ONE*, 3(8):e2876, aug 2008.

[DRP$^+$17] Nisha K. Duggal, Jana M. Ritter, Samuel E. Pestorius, Sherif R. Zaki, Brent S. Davis, Gwong-Jen J. Chang, Richard A. Bowen, and Aaron C. Brault. Frequent Zika Virus Sexual Transmission and Prolonged Viral RNA Shedding in an Immunodeficient Mouse Model. *Cell Reports*, 18(7):1751–1760, feb 2017.

[EBA$^+$21] Chloe J English, Natasha A Botwright, Mark B Adams, Andrew C Barnes, James W Wynne, Paula C Lima, and Mathew T Cook. Immersion challenge of naïve Atlantic salmon with cultured Nolandella sp. and Pseudoparamoeba sp. did not increase the severity of Neoparamoeba perurans -induced amoebic gill disease (AGD). *Journal of Fish Diseases*, 44(2):149–160, feb 2021.

[EHW$^+$19] P.L. Eblé, T.J. Hagenaars, E. Weesendorp, S. Quak, H.W. Moonen-Leusen, and W.L.A. Loeffen. Transmission of African Swine Fever Virus via carrier (survivor) pigs does occur. *Veterinary Microbiology*, 237(June):108345, oct 2019.

[ELEBCH$^+$09] Héctor M. Esparza-Leal, César M. Escobedo-Bonilla, Ramón Casillas-Hernández, Píndaro Álvarez-Ruíz, Guillermo Portillo-Clark, Roberto C. Valerio-García, Jorge Hernández-López, Jesús Méndez-Lozano, Norberto Vibanco-Pérez, and Francisco J. Magallón-Barajas. Detection of white spot syndrome virus in filtered shrimp-farm water fractions and experimental evaluation of its infectivity in Penaeus (Litopenaeus) vannamei. *Aquaculture*, 292(1-2):16–22, jul 2009.

[FAO20] *The State of World Fisheries and Aquaculture 2020.* FAO, 2020.

[FCY^{+}19] Jaime Figueroa, Juan Cárcamo, Alejandro Yañez, Victor Olavarria, Pamela Ruiz, René Manríquez, Claudio Muñoz, Alex Romero, and Ruben Avendaño-Herrera. Addressing viral and bacterial threats to salmon farming in Chile: historical contexts and perspectives for management and control. *Reviews in Aquaculture*, 11(2):299–324, may 2019.

[Fec18] Marie-Eve Fecteau. Paratuberculosis in Cattle. *Veterinary Clinics of North America: Food Animal Practice*, 34(1):209–222, mar 2018.

[Fen07] Zhilan Feng. Final and peak epidemic sizes for SEIR models with quarantine and isolation. *Mathematical Biosciences and Engineering*, 4(4):675–686, 2007.

[FHK^{+}20] Josh A. Firth, Joel Hellewell, Petra Klepac, Stephen Kissler, Adam J. Kucharski, and Lewis G. Spurgin. Using a real-world network to model localized COVID-19 control strategies. *Nature Medicine*, 26(10):1616–1622, oct 2020.

[fI21] British Society for Immunology. *John Snow's pump (1854)*, 2021. url: `https://www.immunology.org/john-snows-pump-1854`. Last accessed: 2021-03-21.

[FotUN21] Food and Agriculture Organization of the United Nations. *Digital Agriculture*, 2021. url: `http://www.fao.org/digital-agriculture/en/`. Last accessed: 2021-03-31.

[FY15] Christopher F. Fronczek and Jeong-Yeol Yoon. Biosensors for Monitoring Airborne Pathogens. *Journal of Laboratory Automation*, 20(4):390–410, aug 2015.

[Gay11] R Gaynes. Louis Pasteur and the Germ Theory of Disease. In *Germ Theory*, pages 143–171. American Society of Microbiology, jan 2011.

[GC01] Simon Godsill and Tim Clapp. Improvement Strategies for Monte Carlo Particle Filters. In *Sequential Monte Carlo Methods in Practice*, pages 139–158. Springer New York, New York, NY, 2001.

[GCS^{+}13] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman and Hall/CRC, nov 2013.

[GCSR95] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman and Hall/CRC, jun 1995.

[GF06] Nicholas C. Grassly and Christophe Fraser. Seasonal infectious disease epidemiology. *Proceedings of the Royal Society B: Biological Sciences*, 273(1600):2541–2550, oct 2006.

[GKM^{+}18] Aoibheann Gaughran, David J. Kelly, Teresa MacWhite, Enda Mullen, Peter Maher, Margaret Good, and Nicola M. Marples. Super-ranging. A new ranging strategy in European badgers. *PLOS ONE*, 13(2):e0191818, feb 2018.

[GMP^{+}09] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, feb 2009.

[GMP17] Kokouvi Gamado, Glenn Marion, and Thibaud Porphyre. Data-Driven Risk Assessment from Small Scale Epidemics: Estimation and Model Choice for Spatio-Temporal Data with Application to a Classical Swine Fever Outbreak. *Frontiers in Veterinary Science*, 4(February):1–14, feb 2017.

[GOV] GOV.UK. Press release: World's first coronavirus Human Challenge study receives ethics approval in the UK itle.

[GR92] Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, nov 1992.

[GR98] G. Gibson and E. Renshaw. Estimating parameters in stochastic compartmental models using Markov chain methods. *Mathematical Medicine and Biology*, 15(1):19–40, mar 1998.

[GS13] Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38, feb 2013.

[GST18] Gavin J. Gibson, George Streftaris, and David Thong. Comparison and Assessment of Epidemic Models. *Statistical Science*, 33(1), feb 2018.

[GTL$^+$18] Jian-Fang Gui, Qisheng Tang, Zhongjie Li, Jiashou Liu, and Sena S. De Silva, editors. *Aquaculture in China.* John Wiley & Sons Ltd, Chichester, UK, jun 2018.

[HBRY09] N (FAO) Hishamunda, P (NACA) Buena, N (University of New Brunswick) Ridler, and W.G. (Aquaculture-Based countryside Development Enterprises Foundation Inc.) Yap. Analysis of aquaculture development in Southeast Asia. Technical report, Food and Agriculture Organization of the United Nations, Rome, 2009.

[HCR$^+$17] R. Hauck, B. Crossley, D. Rejmanek, H. Zhou, and R. A. Gallardo. Persistence of Highly Pathogenic and Low Pathogenic Avian Influenza Viruses in Footbaths and Poultry Manure. *Avian Diseases*, 61(1):64–69, mar 2017.

[HD96] J. A. P. Heesterbeek and K. Dietz. The concept of R0 in epidemic theory. *Statistica Neerlandica*, 50(1):89–110, mar 1996.

[Hig01] Desmond J. Higham. An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations. *SIAM Review*, 43(3):525–546, jan 2001.

[HK07] Te Han and Kingo Kobayashi. *Mathematics of Information and Coding*, volume 203 of *Translations of Mathematical Monographs.* American Mathematical Society, Providence, Rhode Island, mar 2007.

[HLLL21] R.Z. He, Z.C. Li, S.Y. Li, and A.X. Li. Development of an immersion challenge model for Streptococcus agalactiae in Nile tilapia (Oreochromis niloticus). *Aquaculture*, 531:735877, jan 2021.

[Hol20] A (Statista) Holst. *Wearable technology - Statistics & Facts*, March 2020. url: `https://www.statista.com/topics/1556/wearable-technology/#dossierSummary__chapter2`. Last accessed: 2021-03-10.

[HSW05] J.M Heffernan, R.J Smith, and L.M Wahl. Perspectives on the basic reproductive ratio. *Journal of The Royal Society Interface*, 2(4):281–293, sep 2005.

[HVT$^+$17] Helena Hauge, Niccolo Vendramin, Torunn Taksdal, Anne Berit Olsen, Øystein Wessel, Susie Sommer Mikkelsen, Anna Luiza Farias Alencar, Niels Jørgen Olesen, and Maria Krudtaa Dahle. Infection experiments with novel Piscine orthoreovirus from rainbow trout (Oncorhynchus mykiss) in salmonids. *PLOS ONE*, 12(7):e0180293, jul 2017.

[IA21] Rute Irgang and Ruben Avendaño-Herrera. Experimental tenacibaculosis infection in adult conger eel (Genypterus chilensis , Guichenot 1948) by immersion challenge with Tenacibaculum dicentrarchi. *Journal of Fish Diseases*, 44(2):211–216, feb 2021.

[JFWH19] Diana Jaramillo, Stewart Fielder, Richard J Whittington, and Paul Hick. Host, agent and environment interactions affecting Nervous necrosis virus infection in Australian bass Macquaria novemaculeata. *JOURNAL OF FISH DISEASES*, 42(2):167–180, feb 2019.

[KBP$^+$20] Andrea Louise Kroeker, Shawn Babiuk, Bradley S. Pickering, Juergen A. Richt, and William C. Wilson. Livestock Challenge Models of Rift Valley Fever for Agricultural Vaccine Testing. *Frontiers in Veterinary Science*, 7(May):1–13, may 2020.

[Kee01] M. J. Keeling. Dynamics of the 2001 UK Foot and Mouth Epidemic: Stochastic Dispersal in a Heterogeneous Landscape. *Science*, 294(5543):813–817, oct 2001.

[Kee05] Matt J. Keeling. Models of foot-and-mouth disease. *Proceedings of the Royal Society B: Biological Sciences*, 272(1569):1195–1202, jun 2005.

[KF12] Robert K. Kim and Mohamed Faisal. Shedding of viral hemorrhagic septicemia virus (Genotype IVb) by experimentally infected muskellunge (Esox masquinongy). *The Journal of Microbiology*, 50(2):278–284, apr 2012.

[KHM19] Adam Kleczkowski, Andy Hoyle, and Paul McMenemy. One model to rule them all? Modelling approaches across OneHealth for human, animal and plant epidemics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1775):20180255, jun 2019.

[Kir20] Tony Kirby. COVID-19 human challenge studies in the UK. *The Lancet Respiratory Medicine*, 8(12):e96, dec 2020.

[KKG18] Petra Klepac, Stephen Kissler, and Julia Gog. Contagion! The BBC Four Pandemic – The model behind the documentary. *Epidemics*, 24(March):49–59, sep 2018.

[KKO21] Si-Woo Kim, Soo-Jin Kim, and Myung-Joo Oh. Efficacy of live NNV immersion vaccine immunized at low temperature in sevenband grouper, Epinephelus septemfasciatus. *Virus Research*, 292:198227, jan 2021.

[KM27] W O Kermack and A G McKendrick. 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, aug 1927.

[KQKO21] Rahul Krishnan, Syed Shariq Nazir Qadiri, Jae-Ok Kim, and Myung-Joo Oh. Infection dynamics and shedding kinetics of nervous necrosis virus in juvenile seven band grouper using an intraperitoneal infection-cohabitation model. *Aquaculture*, 530:735957, jan 2021.

[LC98] Jun S. Liu and Rong Chen. Sequential Monte Carlo Methods for Dynamic Systems. *Journal of the American Statistical Association*, 93(443):1032–1044, sep 1998.

[LC16] Justin Lessler and Derek A. T. Cummings. Mechanistic Models of Infectious Disease and Their Impact on Public Health. *American Journal of Epidemiology*, 183(5):415–422, mar 2016.

[LEH^{+}15] J. Lessler, W.J. Edmunds, M.E. Halloran, T.D. Hollingsworth, and A.L. Lloyd. Seven challenges for model-driven data collection in experimental and observational studies. *Epidemics*, 10:78–82, mar 2015.

[LFKS13] Juliane Liepe, Sarah Filippi, Michał Komorowski, and Michael P. H. Stumpf. Maximizing the Information Content of Experiments in Systems Biology. *PLoS Computational Biology*, 9(1):e1002888, jan 2013.

[LGJ21] Amy Long, Aidan Goodall, and Simon R.M. Jones. Development of a Piscirickettsia salmonis immersion challenge model to investigate the comparative susceptibility of three salmon species. *Journal of Fish Diseases*, 44(1):1–9, jan 2021.

[LHL^{+}20] Shouhu Li, Yonghao Hu, Xuerui Li, Shengyi Han, Bo Zhang, Zunqiang Yan, Ruiling Xue, Qiang Gao, Jintang Wu, Xingxu Zhao, and Jixing Liu. Development of a live vector vaccine against infectious pancreatic necrosis virus in rainbow trout. *AQUACULTURE*, 524, jul 2020.

[LMSG14] Max S Y Lau, Glenn Marion, George Streftaris, and Gavin J Gibson. New model diagnostics for spatio-temporal systems in epidemiology and ecology. *Journal of The Royal Society Interface*, 11(93):20131093–20131093, feb 2014.

[LMWH18] Owen R. Liu, Renato Molina, Margaret Wilson, and Benjamin S. Halpern. Global opportunities for mariculture development to promote human nutrition. *PeerJ*, 6(5):e4733, may 2018.

[LPGS^{+}11] Sandrine Lesellier, Si Palmer, Sonya Gowtage-Sequiera, Roland Ashford, Deanna Dalley, Dipesh Davé, Ute Weyer, F. Javier Salguero, Alejandro Nunez, Timothy

Crawshaw, Leigh A.L. Corner, R. Glyn Hewinson, and Mark A. Chambers. Protection of Eurasian badgers (Meles meles) from tuberculosis after intra-muscular vaccination with different doses of BCG. *Vaccine*, 29(21):3782–3790, may 2011.

[LR21] Tina Lu and Ben Y. Reis. Internet search patterns reveal clinical course of COVID-19 disease progression and pandemic spread across 32 countries. *npj Digital Medicine*, 4(1):22, dec 2021.

[LS02] Jeffrey M. Lotz and M. Andres Soto. Model of white spot syndrome virus (WSSV) epidemics in Litopenaeus vannamei. *Diseases of Aquatic Organisms*, 50(3):199–209, 2002.

[LSB19] Bestha Lakshmi, Shameer Syed, and Viswanath Buddolla. Current Advances in the Protection of Viral Diseases in Aquaculture With Special Reference to Vaccination. In *Recent Developments in Applied Microbiology and Biochemistry*, pages 127–146. Elsevier, 2019.

[LSGW04] James O. Lloyd-Smith, Wayne M. Getz, and Hans V. Westerhoff. Frequency–dependent incidence in models of sexually transmitted diseases: portrayal of pair–based transmission and effects of illness on contact behaviour. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1539):625–634, mar 2004.

[LSJ$^+$20] Inae Lee, Youngung Seok, Huijin Jung, Byungjin Yang, Jiho Lee, Jaeyoung Kim, Heesoo Pyo, Chang-Seon Song, Won Choi, Min-Gon Kim, and Joonseok Lee. Integrated Bioaerosol Sampling/Monitoring Platform: Field-Deployable and Rapid Detection of Airborne Viruses. *ACS Sensors*, 5(12):3915–3922, dec 2020.

[MA78] Robert M. May and Roy M. Anderson. Regulation and Stability of Host-Parasite Population Interactions: II. Destabilizing Processes. *The Journal of Animal Ecology*, 47(1):249, feb 1978.

[Ma20] Junling Ma. Estimating epidemic exponential growth rate and basic reproduction number. *Infectious Disease Modelling*, 5:129–141, 2020.

[MÁC$^+$12] Juliana R. Moser, Diego A. Galván Álvarez, Fernando Mendoza Cano, Trinidad Encinas Garcia, Daniel E. Coronado Molina, Guillermo Portillo Clark, Maria Risoleta F. Marques, Francisco J. Magallón Barajas, and Jorge Hernández López. Water temperature influences viral load and detection of White Spot Syndrome Virus (WSSV) in Litopenaeus vannamei and wild crustaceans. *Aquaculture*, 326-329:9–14, jan 2012.

[MAF$^+$15] Stefano Merler, Marco Ajelli, Laura Fumanelli, Marcelo F C Gomes, Ana Pastore y. Piontti, Luca Rossi, Dennis L. Chao, Ira M. Longini, M. Elizabeth Halloran, and Alessandro Vespignani. Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. *The Lancet Infectious Diseases*, 15(2):204–211, feb 2015.

[MC20] Lidia Morawska and Junji Cao. Airborne transmission of SARS-CoV-2: The world should face the reality. *Environment International*, 139:105730, jun 2020.

[McC01] H McCallum. How should pathogen transmission be modelled? *Trends in Ecology & Evolution*, 16(6):295–300, jun 2001.

[McG85] Roderick E. McGrew. *Encyclopedia of Medical History*. Palgrave Macmillan UK, London, 1985.

[Mey16] Renate Meyer. Deviance Information Criterion (DIC). In *Wiley StatsRef: Statistics Reference Online*, pages 1–6. John Wiley & Sons, Ltd, Chichester, UK, aug 2016.

[Mil12] Joel C. Miller. A Note on the Derivation of Epidemic Final Sizes. *Bulletin of Mathematical Biology*, 74(9):2125–2141, sep 2012.

[MNH+12] N.J. McPherson, R.A. Norman, A.S. Hoyle, J.E. Bron, and N.G.H. Taylor. Stocking methods and parasite-induced reductions in capture: Modelling Argulus foliaceus in trout fisheries. *Journal of Theoretical Biology*, 312:22–33, nov 2012.

[Mon20] L.H.A. Monteiro. Short Communication. *Ecological Complexity*, page 100836, may 2020.

[Nat09] United Nations. *Food Production Must Double by 2050 to Meet Demand from World's Growing Population, Innovative Strategies Needed to Combat Hunger, Experts Tell Second Committee*, October 2009. url: `https://www.un.org/press/en/2009/gaef3242.doc.htm`. Last accessed: 2021-03-21.

[Nor07] William D. Nordhaus. Two Centuries of Productivity Growth in Computing. *The Journal of Economic History*, 67(1):128–159, mar 2007.

[NR05] Peter Neal and Gareth Roberts. A case study in non-centering for data augmentation: Stochastic epidemics. *Statistics and Computing*, 15(4):315–327, oct 2005.

[NTS+16] Shevanthi Nayagam, Mark Thursz, Elisa Sicuri, Lesong Conteh, Stefan Wiktor, Daniel Low-Beer, and Timothy B. Hallett. Requirements for global elimination of hepatitis B: a modelling study. *The Lancet Infectious Diseases*, 16(12):1399–1408, dec 2016.

[oC18] University of Cambridge. *Citizen science experiment predicts massive toll of flu pandemic on the UK*, May 2018. url: `https://www.cam.ac.uk/research/news/citizen-science-experiment-predicts-massive-toll-of-flu-pandemic-on-the-uk`. Last accessed: 2021-03-01.

[O'D20] S (Statista) O'Dea. *Smartphone users worldwide 2016-2021*, December 2020. url: `https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/`. Last accessed: 2021-03-10.

[ODW+18] Birgit Oidtmann, Peter Dixon, Keith Way, Claire Joiner, and Amanda E. Bayley. Risk of waterborne virus spread - review of survival of relevant fish and crustacean viruses in the aquatic environment and implications for control measures. *Reviews in Aquaculture*, 10(3):641–669, aug 2018.

[OR99] Philip D. O'Neill and Gareth O. Roberts. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):121–129, feb 1999.

[PBCV06] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11, 2006.

[PBM15] C M Pooley, S C Bishop, and G Marion. Using model-based proposals for fast parameter inference on discrete state space, continuous-time Markov processes. *Journal of the Royal Society Interface*, 12(107), 2015.

[Pet12] Brad (Forbes) Peters. *The Age of Big Data*, 2012. url: `https://www.forbes.com/sites/bradpeters/2012/07/12/the-age-of-big-data/`. Last accessed: 2021-03-01.

[PKF+21] Mira L. Pöhlker, Ovid O. Krüger, Jan-David Förster, Wolfgang Elbert, Janine Fröhlich-Nowoisky, Ulrich Pöschl, Christopher Pöhlker, Gholamhossein Bagheri, Eberhard Bodenschatz, J. Alex Huffman, Simone Scheithauer, and Eugene Mikhailov. Respiratory aerosols and droplets in the transmission of infectious diseases. 2021.

[PL20] Allyson M Pollock and James Lancaster. Asymptomatic transmission of covid-19. *BMJ*, page m4851, dec 2020.

[PM18] C M Pooley and G Marion. Bayesian model evidence as a practical alternative to deviance information criterion. *Royal Society Open Science*, 5(3):171519, mar 2018.

[PMB+20] Christopher M. Pooley, Glenn Marion, Stephen C. Bishop, Richard I. Bailey, and Andrea B. Doeschl-Wilson. Estimating individuals' genetic and non-genetic effects underlying infectious disease transmission from temporal epidemic data. *PLOS Computational Biology*, 16(12):e1008447, dec 2020.

[PMH+17] Jamie C. Prentice, Glenn Marion, Michael R. Hutchings, Tom N. McNeilly, and Louise Matthews. Complex responses to movement-based disease control: when livestock trading helps. *Journal of The Royal Society Interface*, 14(126):20160531, jan 2017.

[PMPG+16] M. K. Purcell, C. L. McKibben, S. Pearman-Gillman, D. G. Elliott, and J. R. Winton. Effects of temperature on Renibacterium salmoninarum infection and transmission potential in Chinook salmon, Oncorhynchus tshawytscha (Walbaum). *Journal of fish diseases*, 39(7):787–798, 2016.

[PMSG19] Mark P. Polinski, Gary D. Marty, Heindrich N. Snyman, and Kyle A. Garver. Piscine orthoreovirus demonstrates high infectivity but low virulence in Atlantic salmon of Pacific Canada. *Scientific Reports*, 9(1):3297, dec 2019.

[PP13] Alida Palmisano and Corrado Priami. Stochastic Simulation Algorithm. In *Encyclopedia of Systems Biology*, pages 2009–2010. Springer New York, New York, NY, 2013.

[PPS20] Ilenia Pierantoni, Mariano Pierantozzi, and Massimo Sargolini. COVID 19—A Qualitative Review for the Reorganization of Human Living Environments. *Applied Sciences*, 10(16):5576, aug 2020.

[PRC20] Nguyen Ngoc Phuoc, Randolph Richards, and Margaret Crumlish. Environmental conditions influence susceptibility of striped catfish Pangasianodon hypophthalmus (Sauvage) to Edwardsiella ictaluri. *AQUACULTURE*, 523, jun 2020.

[QHDA09] Nguyen Duc Quang, Phan Thi Phuong Hoa, Tran Thanh Da, and Phan Hoai Anh. Persistence of white spot syndrome virus in shrimp ponds and surrounding areas after an outbreak. *Environmental Monitoring and Assessment*, 156(1-4):69–72, sep 2009.

[R C18] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.

[RBM$^+$14] Nicholas Robinson, Matthew Baranski, Kanta Das Mahapatra, Jatindra Nath Saha, Sweta Das, Jashobanta Mishra, Paramananda Das, Matthew Kent, Mariann Arnyasi, and Pramoda Kumar Sahoo. A linkage map of transcribed single nucleotide polymorphisms in rohu (Labeo rohita) and QTL associated with resistance to Aeromonas hydrophila. *BMC genomics*, 15(1):541, 2014.

[RCLMJ$^+$15] Flor Ramírez-Castillo, Abraham Loera-Muro, Mario Jacques, Philippe Garneau, Francisco Avelar-González, Josée Harel, and Alma Guerrero-Barrera. Waterborne Pathogens: Detection Methods and Challenges. *Pathogens*, 4(2):307–334, may 2015.

[Rön01] P Rönnbäck. Shrimp aquaculture - State of the art. Technical report, Swedish University of Agricultural Sciences (SLU), Uppsala, 2001.

[RW00] L. C. G. Rogers and David Williams. *Diffusions, Markov Processes, and Martingales.* Cambridge University Press, apr 2000.

[SAR$^+$13] S. Satheesh Kumar, R. Ananda Bharathi, J. J.S. Rajan, S. V. Alavandi, M. Poornima, C. P. Balasubramanian, and A. G. Ponniah. Viability of white spot syndrome virus (WSSV) in sediment during sun-drying (drainable pond) and under non-drainable pond conditions indicated by infectivity to shrimp. *Aquaculture*, 402-403:119–126, 2013.

[SC10] G. N. Sze To and C. Y. H. Chao. Review and comparison between the Wells-Riley and dose-response approaches to risk assessment of infectious respiratory diseases. *Indoor Air*, 20(1):2–16, feb 2010.

[Sch78] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, mar 1978.

[Sco21] Natural Scotland. *Scotland's Aquaculture*, 2021. url: `http://aquaculture.scotland.gov.uk/our_aquaculture/our_aquaculture.aspx`. Last accessed: 2021-03-08.

[Sel83] Thomas Sellke. On the asymptotic distribution of the size of a stochastic epidemic. *Journal of Applied Probability*, 20(02):390–394, jun 1983.

[SG12] George Streftaris and Gavin J Gibson. Non-exponential tolerance to infection in epidemic systems–modeling, inference, and assessment. *Biostatistics*, 13(4):580–593, sep 2012.

[SGOV16] Lone Simonsen, Julia R. Gog, Don Olson, and Cécile Viboud. Infectious Disease Surveillance in the Big Data Era: Towards Faster and Locally Relevant Systems. *Journal of Infectious Diseases*, 214(suppl 4):S380–S385, dec 2016.

[SJ13] Ruth Hall Sedlak and Keith R. Jerome. Viral diagnostics in the era of digital polymerase chain reaction. *Diagnostic Microbiology and Infectious Disease*, 75(1):1–4, jan 2013.

[SJBPP03] Sylvia Rodriguez Saint-Jean, Juan J Borrego, and Sara I Perez-Prieto. Infectious Pancreatic Necrosis Virus: Biology, Pathogenesis, and Diagnostic Methods. pages 113–165. 2003.

[SL01] M.Andres Soto and Jeffrey M Lotz. Epidemiological Parameters of White Spot Syndrome Virus Infections in Litopenaeus vannamei and L. setiferus. *Journal of Invertebrate Pathology*, 78(1):9–15, jul 2001.

[Sno49] John Snow. On the mode of communication of cholera. 1849.

[SP10] Arturo Sánchez-Paz. White spot syndrome virus: an overview on an emergent concern. *Veterinary Research*, 41(6):43, nov 2010.

[SR13] NKG Salama and B. Rabe. Developing models for investigating the environmental transmission of disease-causing agents within open-cage salmon aquaculture. *Aquaculture Environment Interactions*, 4(2):91–115, jul 2013.

[SRT$^+$21] Francisca Samsing, Megan Rigby, Hedda K Tengesdal, Richard S Taylor, Daniela Farias, Richard N Morrison, Scott Godwin, Carla Giles, Jeremy Carson, Chloe J English, Roger Chong, and James W Wynne. Seawater transmission and infection dynamics of pilchard orthomyxovirus (POMV) in Atlantic salmon (Salmo salar). *Journal of Fish Diseases*, 44(1):73–88, jan 2021.

[SSKZ90] D. E. Stallknecht, S. M. Shane, M. T. Kearney, and P. J. Zwank. Persistence of Avian Influenza Viruses in Water. *Avian Diseases*, 34(2):406, apr 1990.

[Sub05] Rohana P. Subasinghe. Epidemiological approach to aquatic animal health management: opportunities and challenges for developing countries to increase aquatic production through aquaculture. *Preventive Veterinary Medicine*, 67(2-3):117–124, feb 2005.

[SWMH09] Lesley A. Smith, Piran C.L. White, Glenn Marion, and Michael R. Hutchings. Livestock grazing behavior and inter- versus intraspecific disease risk via the fecal-oral route. *Behavioral Ecology*, 20(2):426–432, 2009.

[TE10] Joseph H. Tien and David J. D. Earn. Multiple Transmission Pathways and Disease Dynamics in a Waterborne Pathogen Model. *Bulletin of Mathematical Biology*, 72(6):1506–1533, aug 2010.

[THY$^+$20] Loring J Thomas, Peng Huang, Fan Yin, Xiaoshuang Iris Luo, Zack W Almquist, John R Hipp, and Carter T Butts. Spatial heterogeneity can lead to substantial local variations in COVID-19 timing and severity. *Proceedings of the National Academy of Sciences*, 117(39):24180–24187, sep 2020.

[TSC$^+$21] Richard S Taylor, Joel Slinger, Paula Camargo Lima, Chloe J English, Ben T Maynard, Francisca Samsing, Russell McCulloch, Petra R. Quezada-Rodriguez, and James W Wynne. Evaluation of sodium percarbonate as a bath treatment for amoebic gill disease in Atlantic salmon. *Aquaculture Research*, 52(1):117–129, jan 2021.

[TVVdJ14] N. X. Tuyen, J. Verreth, J. M. Vlak, and M. C M de Jong. Horizontal transmission dynamics of White spot syndrome virus by cohabitation trials in juvenile Penaeus monodon and P. vannamei. *Preventive Veterinary Medicine*, 117(1):286–294, 2014.

[TW87] Martin A. Tanner and Wing Hung Wong. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398):528, jun 1987.

[TW10] Martin A. Tanner and Wing H. Wong. From EM to Data Augmentation: The Emergence of MCMC Bayesian Computation in the 1980s. *Statistical Science*, 25(4):506–516, 2010.

[TWI$^+$21] Ragnar Thorarinsson, Jeffrey C Wolf, Makoto Inami, Lisa Phillips, Ginny Jones, Alicia M Macdonald, Jose F Rodriguez, Hilde Sindre, Eystein Skjerve, Espen Rimstad, and Oystein Evensen. Effect of a novel DNA vaccine against pancreas disease caused by salmonid alphavirus subtype 3 in Atlantic salmon (Salmo salar). *FISH & SHELLFISH IMMUNOLOGY*, 108:116–126, jan 2021.

[vDBM$^+$20] Neeltje van Doremalen, Trenton Bushmaker, Dylan H. Morris, Myndi G. Holbrook, Amandine Gamble, Brandi N. Williamson, Azaibi Tamin, Jennifer L. Harcourt, Natalie J. Thornburg, Susan I. Gerber, James O. Lloyd-Smith, Emmie de Wit, and Vincent J. Munster. Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. *New England Journal of Medicine*, 382(16):1564–1567, apr 2020.

[VGO+20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[vK07] N.G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, 2007.

[VLD+20] Knut Wiik Vollset, Robert J Lennox, Jan Grimsrud Davidsen, Sindre Håvarstein Eldøy, Trond E Isaksen, Abdullah Madhun, Sten Karlsson, and Kristina M Miller. Wild salmonids are running the gauntlet of pathogens and climate as fish farms expand northwards. *ICES Journal of Marine Science*, oct 2020.

[VNL04] Kristie A. Vanpatten, Linda M. Nunan, and Donald V. Lightner. Seabirds as potential vectors of penaeid shrimp viruses and the development of a surrogate laboratory model utilizing domestic chickens. *Aquaculture*, 241(1-4):31–46, nov 2004.

[VPHC10] Luigi Vezzulli, Carla Pruzzo, Anwar Huq, and Rita R. Colwell. Environmental reservoirs of Vibrio cholerae and their role in cholera. *Environmental Microbiology Reports*, 2(1):27–33, jan 2010.

[VTHvR12] J. Vanlier, C. A. Tiemann, P. A. J. Hilbers, and N. A. W. van Riel. A Bayesian approach to targeted experiment design. *Bioinformatics*, 28(8):1136–1142, apr 2012.

[Wal13] Stephen G Walker. Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, 143(10):1621–1633, oct 2013.

[WBB20] Arabella Widders, Alex Broom, and Jennifer Broom. SARS-CoV-2: The viral shedding vs infectivity dilemma. *Infection, Disease & Health*, (xxxx), may 2020.

[WCG+20] Roman Wölfel, Victor M. Corman, Wolfgang Guggemos, Michael Seilmaier, Sabine Zange, Marcel A. Müller, Daniela Niemeyer, Terry C. Jones, Patrick Vollmar, Camilla Rothe, Michael Hoelscher, Tobias Bleicker, Sebastian Brünink, Julia Schneider, Rosina Ehmann, Katrin Zwirglmaier, Christian Drosten, and Clemens Wendtner. Virological assessment of hospitalized patients with COVID-2019. *Nature*, 581(7809):465–469, may 2020.

[WHO21a] WHO. *Fact sheets / Detail / Cholera*, February 2021. url: `https://www.who.int/news-room/fact-sheets/detail/cholera`. Last accessed: 2021-03-10.

[WHO21b] WHO. *Water sanitation hygiene - Water-related Diseases*, 2021. url: `https://www.who.int/water_sanitation_health/diseases-risks/diseases/cholera/en/`. Last accessed: 2021-03-03.

[WJL$^+$12] Rong-Hua Wang, Zhen Jin, Quan-Xing Liu, Johan van de Koppel, and David Alonso. A Simple Stochastic Model with Environmental Transmission Explains Multi-Year Periodicity in Outbreaks of Avian Flu. *PLoS ONE*, 7(2):e28873, feb 2012.

[Wor13] Michael Worboys. Joseph Lister and the performance of antiseptic surgery. *Notes and Records of the Royal Society*, 67(3):199–209, sep 2013.

[WSB$^+$21] Qi Wang, Nadia Shakoor, Adam Boyher, Kira M. Veley, Jeffrey C. Berry, Todd C. Mockler, and Rebecca S. Bart. Escalation in the host-pathogen arms race: A host resistance response corresponds to a heightened bacterial virulence response. *PLOS Pathogens*, 17(1):e1009175, jan 2021.

[WSKK17] Andrew R. Wargo, Robert J. Scott, Benjamin Kerr, and Gael Kurath. Replication and shedding kinetics of infectious hematopoietic necrosis virus in juvenile rainbow trout. *Virus Research*, 227:200–211, jan 2017.

[YLW$^+$04] Ignatius T.S. Yu, Yuguo Li, Tze Wai Wong, Wilson Tam, Andy T. Chan, Joseph H.W. Lee, Dennis Y.C. Leung, and Tommy Ho. Evidence of Airborne Transmission of the Severe Acute Respiratory Syndrome Virus. *New England Journal of Medicine*, 350(17):1731–1739, apr 2004.

[YMHÜ$^+$18] Ali K. Yetisen, Juan Leonardo Martinez-Hurtado, Barış Ünal, Ali Khademhosseini, and Haider Butt. Wearables in Medicine. *Advanced Materials*, 30(33):1706910, aug 2018.

[YQTW14] I. T.-S. Yu, Hong Qiu, Lap Ah Tse, and Tze Wai Wong. Severe Acute Respiratory Syndrome Beyond Amoy Gardens: Completing the Incomplete Legacy. *Clinical Infectious Diseases*, 58(5):683–686, mar 2014.

[ZGH$^+$19] Shufang Zhou, Tong Gou, Jiumei Hu, Wenshuai Wu, Xiong Ding, Weibo Fang, Zhenming Hu, and Ying Mu. A highly integrated real-time digital PCR device for accurate DNA quantitative analysis. *Biosensors and Bioelectronics*, 128:151–158, mar 2019.

[ZLZ$^+$20] Renyi Zhang, Yixin Li, Annie L. Zhang, Yuan Wang, and Mario J. Molina. Identifying airborne transmission as the dominant route for the spread of COVID-19. *Proceedings of the National Academy of Sciences*, 117(26):14857–14863, jun 2020.

Appendix **A**

Appendices

A Metropolis-cooled MCMC routine for Chapter 2

A.1 SEIR likelihood

As stated in Section 2.2 the SEIR likelihood is,

$$p(\,\mathbf{E},\mathbf{I},\mathbf{R}\,|\,\beta,\gamma,\delta\,) \propto \beta^{m-1}B(\mathbf{E},\mathbf{I},\mathbf{R})e^{-\beta A(\mathbf{E},\mathbf{I},\mathbf{R})}\prod_{j=1}^{m}\delta e^{-\delta(I_j-E_j)}\prod_{j=1}^{m}\gamma e^{-\gamma(R_j-I_j)}$$

where m is the final outbreak size (i.e. number of exposure, onset of infectivity and removal events) and E_κ (below) is the unobserved time of the index exposure

$$\begin{aligned} A &= \int_{E_\kappa}^{\infty} S_t I_t \, dt \\ B &= \prod_{i\neq\kappa} \{I_{E_i}\} \end{aligned}$$

The integral $\int_{E_\kappa}^{\infty} S_t I_t \, dt$ in (2.6) has an easily computable form, found in [NR05]

$$A = \sum_{j=1}^{m}\sum_{i=1}^{N}\{\min{(R_j, I_i)} - \min{(I_j, I_i)}\} \tag{A.1}$$

where j sums over all exposure events and i sums over the entire host population. This quantity represents the total susceptible-infectious host contact time throughout the outbreak.

To explain the form of the likelihood, each exposed host, j, contributes terms $\delta e^{-(I_j-E_j)}$ and $\gamma e^{-(R_j-I_j)}$, since they spend Exp(δ) and Exp(γ) times in the E and I states, respectively. After exposure, all onsets of infectiousness and recoveries are observed. Ordering the exposure events, $E_\kappa = E_1 < \cdots < E_m$, each exposure except the first also contributes a term

$$\beta S_{E_i} I_{E_i} e^{-\beta \int_{E_{i-1}}^{E_i} S_t I_t \, dt} \qquad i > 1$$

since the i$^{\text{th}}$ exposure event has hazard $\beta S_t I_t$ beginning at time E_{i-1}. There is a final factor of

$$e^{-\beta \int_{E_i}^{\infty} S_t I_t \, dt} = e^{-\beta \int_{E_i}^{R_{\max}} S_t I_t \, dt}$$

Since no further exposures occurred. Since $S_{E_1} S_{E_2} \dots S_{E_m} = (N-1)(N-2)\dots(N-m) =$ constant, these factors are omitted from the likelihood.

A.2 Markov chain Monte Carlo (MCMC)

Since we can write down the posterior density for γ, and within the MCMC routine described in Section 2.2, samples of δ and $\mathbf{E}$ do not depend on previous samples of β, the main routine need only consist in sampling δ and $\mathbf{E}$ and recording these values along with A, B and E_κ. The parameter β can then be sampled from its full conditional distribution, which given an exponential prior $p(\beta) \sim \text{Exp}(\lambda_\beta)$ is

$$p(\beta | \delta, \gamma, \mathbf{E}, \mathbf{I}, \mathbf{R}) \sim \text{Gamma}(m, A + \lambda_\beta)$$

Since the full posterior distribution of parameters and exposure times, $p(\, \beta, \delta, \gamma, \mathbf{E} \,|\, \mathbf{I}, \mathbf{R}\,)$ does not have a closed form, we use MCMC methods to sample from it. This is done by setting initial values for the parameters and exposure times, $\beta_0, \delta_0, \gamma_0, \mathbf{E}_0$ and then, for some $T \geq 1.0$ (the *temperature* - see below), iterating through the following steps for $i = 1, \dots, n$:

1. Update δ via Metropolis-Hastings, i.e., propose $\delta' \sim N(\delta^{i-1}, \sigma_\delta^2)$ and with probability α_1 set $\delta^i = \delta'$, otherwise $\delta^i = \delta^{i-1}$, where

   $$\alpha_1 = \min\left(1, \frac{f_T(\delta', \mathbf{E} | \mathbf{I}, \mathbf{R})}{f_T(\delta^{i-1}, \mathbf{E} | \mathbf{I}, \mathbf{R})}\right)$$

   f_T is the marginal conditional density for $\delta, \mathbf{E}$ at temperature T (see below). The parameter σ_δ is tuned during an number of iterations in order to get an acceptance rate of between 20% and 40%.

2. Choose an exposure time to update (with index j) uniformly at random. Propose $\mathbf{E}'$, where $E_k' = E_k^{i-1}$ for $k \neq j$ and $I_j - E_j' \sim \text{Exp}(\delta^i)$ and with probability α_2 set $E_j^i = E_j'$, otherwise set $E_j^i = E_j^{i-1}$ (all other exposure times are unaltered), where

   $$\alpha_2 = \min\left(1, \frac{f_T(\delta^i, \mathbf{E}' | \mathbf{I}, \mathbf{R})}{f_T(\delta^i, \mathbf{E}^{i-1} | \mathbf{I}, \mathbf{R})} e^{-\delta^i (E_j' - E_j^i)}\right)$$

3. Update β by the Gibbs sampler, i.e. sample from the full conditional distribution $\beta^i \sim \Gamma(m, A + \lambda_\beta)$.

The above follows a fully-centred parameterisation, as discussed by Neal and Roberts for the SIR model ([NR05]). Note that there is no need to sample γ as part of the above routine, since having assumed an exponentially-distributed prior $p(\gamma) = \lambda_\gamma e^{-\lambda_\gamma \gamma}$, its posterior density

$$p(\gamma|\mathbf{E}, \mathbf{I}, \mathbf{R}) \propto \gamma^m e^{-\gamma \sum_{j=1}^m (R_j - I_j)} \times e^{-\lambda_\gamma \gamma}$$

and therefore $p(\gamma|\mathbf{E}, \mathbf{I}, \mathbf{R}) \sim \text{Gamma}(m+1, \sum_{j=1}^m (R_j - I_j) + \lambda_\gamma)$.

Due to the high dimensionality of the sample space and the likelihood function having local maxima, the sampling chains can sometimes be slow to converge to stationarity and even become stuck at certain parameter values, with an acceptance rate going towards zero. *Metropolis coupled* MCMC, or $(\text{MC})^3$ [ADHR04] is the strategy adopted here to alleviate poor mixing and is summarised as follows: several of the above chains are run with several closely spaced temperatures $1.0 = T_1 < T_2 < \cdots < T_r$. The first chain, with temperature 1.0, is termed the *cold chain* and is the only chain from which we obtain samples. The other chains are known as the *heated* chains. After performing a fixed number of iterations for each chain in parallel, two are selected uniformly at random, with temperatures T' and T'' and current states $X' = \beta', \delta', \mathbf{E}'$ and $X'' = \beta'', \delta'', \mathbf{E}''$. The states are then exchanged with probability α_3, where

$$\alpha_3 = \min\left(1, \frac{f_{T_1}(X''|\mathbf{I}, \mathbf{R})}{f_{T_1}(X'|\mathbf{I}, \mathbf{R})} \frac{f_{T_2}(X'|\mathbf{I}, \mathbf{R})}{f_{T_2}(X''|\mathbf{I}, \mathbf{R})}\right)$$

Although the samples from the heated chains are ultimately discarded, the method has the advantage that is easily parallelised on a multi-core machine. For this work, this was achieved using Python's multiprocessing module. Six chains were run in parallel with the temperatures $T = 1.00, 1.02, 1.04, 1.06, 1.08, 1.10$ and exchanges of state were attempted every 400 iterations.

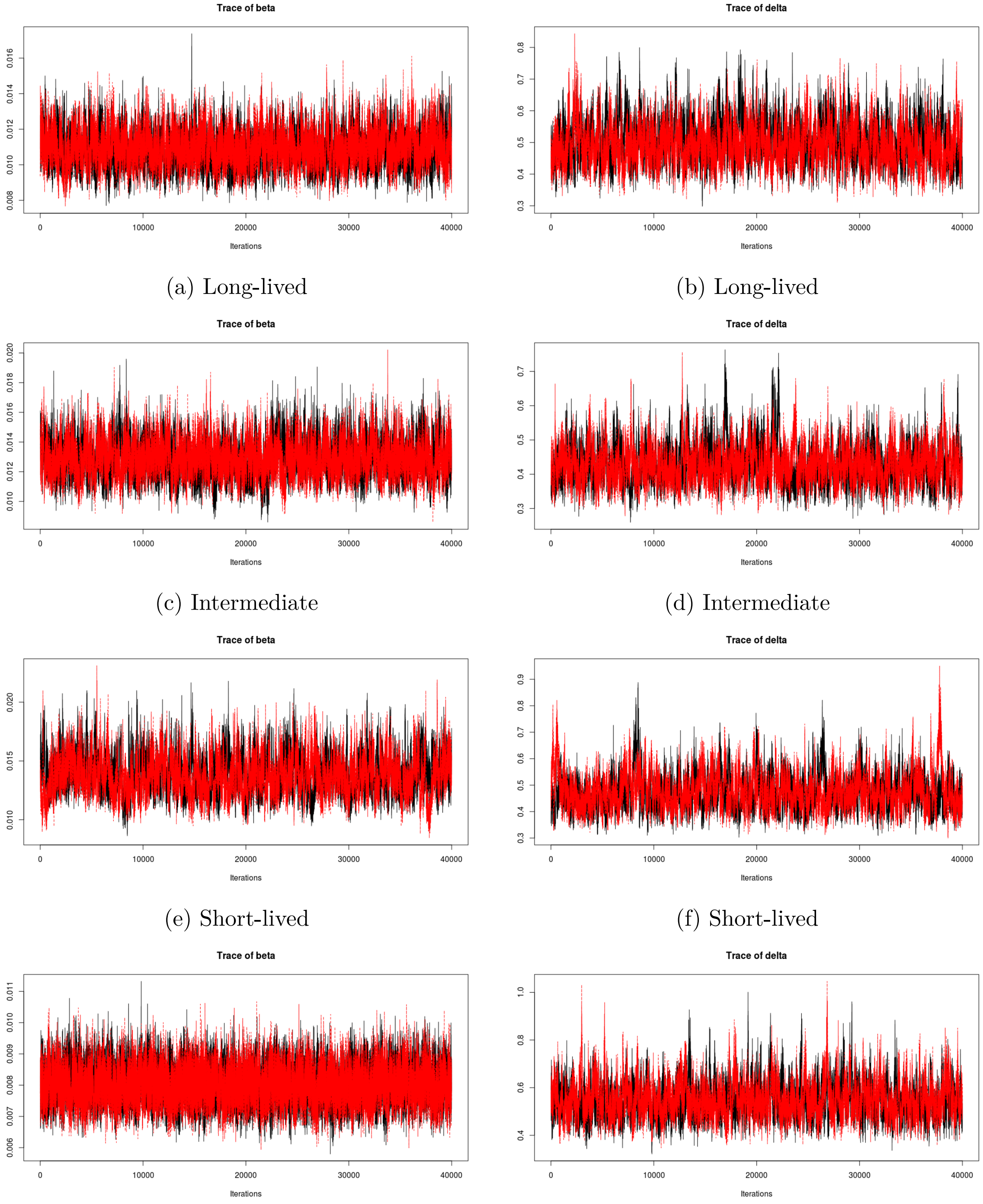
We can therefore marginalise the posterior density at temperature T, obtaining f_T which is, for $p(\delta) \sim \text{Exp}(\lambda_\delta)$

$$\begin{aligned} f_T(\delta, \mathbf{E}|\mathbf{I}, \mathbf{R}) &\propto p(\,\delta, \mathbf{E} \,|\, \mathbf{I}, \mathbf{R}\,)^{\frac{1}{T}} \\ &\propto \left\{ \int \int p(\,\mathbf{E}, \mathbf{I}, \mathbf{R} \,|\, \beta, \delta, \gamma\,) p(\beta) p(\delta) p(\gamma) d\beta \, d\gamma \right\}^{\frac{1}{T}} \\ &\propto \left\{ \prod_{i \neq \kappa} \{I_{E_i}\} \left(A + \lambda_\beta\right)^{-m} \left(\sum_{j=1}^m (R_j - I_j) + \lambda_\gamma \right)^{m+1} \delta^m e^{-\delta(\sum_{j=1}^m (I_j - E_j) + \lambda_\delta)} \right\}^{\frac{1}{T}} \end{aligned}$$

and for $p(\delta) \sim \text{U}(0, 10)$

$$f_T(\delta, \mathbf{E}|\mathbf{I}, \mathbf{R}) \propto \left\{ \prod_{i \neq \kappa} \{I_{E_i}\} \left(A + \lambda_\beta\right)^{-m} \left(\sum_{j=1}^m (R_j - I_j) + \lambda_\gamma \right)^{m+1} \delta^m e^{-\delta \sum_{j=1}^m (I_j - E_j)} \mathbf{1}_{\delta \in (0,10)} \right\}^{\frac{1}{T}}$$

B MCMC summary plots for Chapter 2



(a) Long-lived

(b) Long-lived

(c) Intermediate

(d) Intermediate

(e) Short-lived

(f) Short-lived

(g) Direct transmission only

(h) Direct transmission only

Figure B1: Trace plots using R `coda` package. In the case of *long-lived pathogen* (b), a restricted prior was adopted for δ: $p(\delta) \sim \mathrm{U}(0, 10.0)$ in order to aid mixing. In all other cases $p(\beta), p(\delta) \sim \exp(0.001)$ independently. Trace plots show all iterations, including those later discarded for burn in, in order to demonstrate convergence of two independent chains to stationarity.

C Force of infection in Reed-Frost epidemic model

Lotz and Soto in [SL01] adopt a Reed-Frost model in order to describe transmission of WSD among shrimp in a controlled experiment in which transmission may either be exclusively direct, via ingestion of dead infected shrimp, or exclusively environmental, via cohabitation with a live, infected shrimp. Reed-Frost (see e.g. [Abb52]) is a discrete time model of the numbers of susceptible (S_t), infected (I_t) and removed (R_t) individuals at each of a series of closely spaced time points, separated by a duration Δt. The transmission parameter β is the probability of transmission from some particular infected individual to some other particular susceptible individual during one time step, so that the probability of some susceptible individual not being infected during one time step is

$$(1-\beta)^{I_t} \tag{A.2}$$

and the probability that there is transmission from *at least one* infected individual to this specified susceptible is therefore

$$1-(1-\beta)^{I_t} \tag{A.3}$$

and the expected number of new infections occuring is

$$S_t(1-(1-\beta)^{I_t}). \tag{A.4}$$

The force of infection is therefore

$$\frac{1-(1-\beta)^{I_t}}{\Delta t} \approx \frac{\beta I_t}{\Delta t}. \tag{A.5}$$

D Estimation of α, ϵ and ρ for SEIR-P model of WSD in shrimp

D.1 Estimation of pathogen decay rate, ρ

By a challenge experiment in which *P. monodon* were immersed in sterile seawater that had been spiked with WSSV a variable number of days prior to immersion, Kumar et.al. were able to estimate how long a known quantity of WSSV remains viable in seawater under laboratory conditions. Ten experimental and one control bucket were filled with 10l of sterile seawater. To the experimental buckets pure WSSV was added to a final concentration of 1000 part ml^{-1}, meaning that around 10^7 particles were present in each bucket. On days 0 up to 18, 10 juvenile *P. monodon* were added to one of the unoccupied buckets and all shrimp were monitored at 8h intervals for mortality or signs of WSSV infection. Dead shrimp were removed from the buckets and proportions of living shrimp were recorded daily for each bucket. The authors found that under the conditions of the experiment seawater-borne WSSV remains infective for up to 12 days.

The plots given in [SAR+13, Figure 2] for buckets 0 to 8 indicate similar rates of mortality across these buckets, suggesting that the WSSV lost little of its infectivity during the first eight days in seawater. Total mortality occurred at around the four day mark following introduction to the 0 to 8-day buckets. A reduction in mortality rates is then noticeable following immersion in the 10 day and 12 day buckets, where 100% mortality was observed each at the 7 day mark (i.e 7 days after immersing the shrimp). No mortalities occurred following immersion in the 14, 16 and 18 day buckets, at least during the experimental period, suggesting that the amount of viable WSSV had decayed significantly by 12-14 days in seawater. Since at 8 days there was still a sufficient quantity of WSSV to inoculate all 10 of the shrimp, we expect that the mean infectious lifetime a WSSV particle in seawater to be no less than 8 days, giving pathogen decay rate, ρ, no greater than 0.005 h^{-1}.

D.2 Estimation of environmental transmission rate, α

The second column of Table D1 contains the time in hours, t_{100}, from immersion to 100% mortality of shrimp in each of the buckets, labelled by the number of days after the introduction of WSSV to the bucket that the shrimp were immersed. A lower, order of magnitude, estimate for α (indirect rate of WSD transmission) can be obtained from the same data by assuming that mortality comes immediately upon infection (thus underestimating the infectivity) and assuming that the pathogen density remained at its initial level of $1000\,\mathrm{part\,ml}^{-1}$, at least for buckets 0 to 8, where the rates of mortality were similar.

The transmission rate α should be no more than $10/(10^3 \times t_{100})$, since 10 exposures occurred in t_{100} hours, and the rate of new exposures is $\alpha P = 10^3\alpha$. These values are in Column 3 of Table D1. Similarly, upper estimates for each bucket by supposing that 100% of the exposures had happened at t_{100} - 48, i.e. fixing a period of 2 days between exposure and death for each shrimp since no deaths occurred before 2 days, post-immersion. These are in Column 4. Taking medians, we can say that α is of the order of $10^{-4}\,\mathrm{ml\,part}^{-1}\,\mathrm{h}^{-1}$.

Bucket	t_{100} h	$\alpha\,\mathrm{ml\,part}^{-1}\,\mathrm{h}^{-1}$	
0	96	1.0×10^{-04}	2.1×10^{-4}
2	120	8.3×10^{-05}	1.4×10^{-4}
4	96	1.0×10^{-04}	2.1×10^{-4}
6	72	1.4×10^{-04}	4.2×10^{-4}
8	96	1.0×10^{-04}	2.1×10^{-4}

Table D1: **From graphical plots by Kumar, et.al.** ([SAR+13, Fig 2]). Buckets are labelled by time in days between addition of WSSV to bucket and immersion of shrimp. t_{100} - the time in hours to 100% mortality. Upper and lower estimates of α are given to 1 decimal place and calculations are described in main body of text.

E Bootstrap sequential Monte Carlo routine for particle-marginal MCMC

E.1 General outline

Here we describe in detail the bootstrap SMC routine used to obtain the approximations $\hat{p}(\mathbf{r} \,|\, \beta, \delta) \approx p(\mathbf{r} \,|\, \beta, \delta)$ and $\hat{p}(\mathbf{H}, \mathbf{r} \,|\, \alpha, \epsilon, \rho) \approx p(\mathbf{H}, \mathbf{r} \,|\, \alpha, \epsilon, \rho)$ used to fit the SI, SEI and SI-PQ models of Chapters 3 and 4. Such a routine is a key step in the Particle-Marginal MCMC routine of [ADH10]. Key references also include [GC01, DFG01, LC98]. We shall follow closely the notation and development in [ADH10, GC01]. Let θ be a vector of unknown parameters of some stochastic mechanistic disease model, $\mathbf{Y}_{1:m} = (Y_1, \ldots, Y_m)$ a sequence of observations of the system at the observation times $0 = t_0 < t_1 < \cdots < t_m$ and let $\mathbf{X}_{1:m} = (X_1, \ldots, X_m)$ denote event sequences, X_j, of the underlying disease model between the times t_{j-1} and t_j. For a Markovian disease model, $\mathbf{X}_{1:m}$ can equivalently be the state of the system at the observation times $t_1, \ldots, t_m$. For example, $X_i = (S_{t_i}, I_{t_i}, P_{t_i}, Q_{t_i})$ is the state of the Markovian SI-PQ model of Chapter 4 at time $t = t_i$ and $Y_i = (H_{t_i}, r_{t_i})$ are the results of a dPCR reading (measuring $P_{t_i} + Q_{t_i}$), taken at time t_i, and the number of positive-testing hosts found in a sample taken from the entire host population at $t = t_i$, of size d_i. SMC allows us to obtain, for fixed θ, a weighted set of N *particles* $\mathbf{X}^k_{1:m}$ importance sampled sequentially from the conditional distributions

$$\begin{aligned} X_1^k &\sim p(X_1 \,|\, Y_1, \theta) \\ \mathbf{X}^k_{1:2} &\sim p(\mathbf{X}_{1:2} \,|\, \mathbf{Y}_{1:2}, \theta) \\ &\cdots \\ \mathbf{X}^k_{1:m} &\sim p(\mathbf{X}_{1:m} \,|\, \mathbf{Y}_{1:m}, \theta). \end{aligned} \tag{A.6}$$

Each completed particle $\mathbf{X}^k_{1:m}$ comes with a normalised, W^k_m, and an unnormalised weight, w^k_m, such that

$$\hat{p}(\mathbf{X}_{1:m} \,|\, \mathbf{Y}_{1:m}, \theta) = \sum_{k=1}^{N} W^k_m \delta_{\mathbf{X}^k_{1:m}} \approx p(\mathbf{X}_{1:m} \,|\, \mathbf{Y}_{1:m}, \theta) \tag{A.7}$$

(where the $\delta_{X^k_1}$ are probability measures degenerate at X^k_1 - i.e. for some measureable set B, $\delta_{X^k_1}(B)$ equals 1 if $X^k_1 \in B$ and 0 otherwise), and taking the mean of the unnormalised weights we get, by the law of large numbers

$$\frac{1}{N} \sum_{k=1}^{N} w^k_m \approx p(\mathbf{Y}_{1:m} \,|\, \theta). \tag{A.8}$$

At each generation, r, we use importance sampling to update the particles from the previous generation and their weights. To avoid numerical instability arising from a subset of the weights underflowing to zero, we then perform a re-sampling step from the set of particles. The purpose of this is to cut out particles with low weights and increase the representation in the sample of those with higher weights. This is the essence of the *bootstrap particle filter* [GC01]. Since we are sampling by forward simulation, working from $t = 0$ onwards, no account is taken of the information contained in future observations, $Y_{r+1}, Y_{r+2}, \dots, Y_m$. Some particles may currently have low weights due to outliers occuring early in the observation sequence. To avoid too many of these getting prematurely killed off, we follow a recommendation of Liu, et al. [DFG01, Chapter 11] and smooth the weights, as described below. We now describe the routine in detail.

Generation 1:

To begin, we forward simulate the underlying disease model from $t = 0$ to $t = t_1$ with initial condition X_0 and parameter values θ in order to obtain samples from the importance density

$$X_1^1, \dots, X_1^N \overset{\text{iid}}{\sim} p(X_1 \,|\, \theta) \tag{A.9}$$

which, since θ is fixed, receive the weights

$$w_1^k = w_1(X_1^k) = p(Y_1 \,|\, X_1^k) \tag{A.10}$$

The weights are normalised

$$W_1^k = \frac{w_1^k}{\sum_{k=1}^N w_1^k}. \tag{A.11}$$

and we have the approximate distribution

$$\hat{p}(X_1 \,|\, Y_1, \theta) = \sum_{k=1}^N W_1^k \delta_{X_1^k} \approx p(X_1 \,|\, Y_1, \theta). \tag{A.12}$$

By the law of large numbers

$$\begin{aligned} \frac{1}{N} \sum_{k=1}^N w_1^k &\approx \int p(Y_1 \,|\, X_1) p(X_1 \,|\, \theta) \, dX_1 \\ &= p(Y_1 \,|\, \theta). \end{aligned} \tag{A.13}$$

Generation r:

Suppose that we have so far sampled the particles $\mathbf{X}_{1:r-1}^1, \dots, \mathbf{X}_{1:r-1}^N$ with normalised weights $W_{r-1}^1, \dots, W_{r-1}^N$, such that

$$\hat{p}(\mathbf{X}_{1:r-1} \,|\, \mathbf{Y}_{1:r-1}, \theta) = \sum_{k=1}^N W_{r-1}^k \delta_{\mathbf{X}_{1:r-1}^k} \approx p(\mathbf{X}_{1:r-1} \,|\, \mathbf{Y}_{1:r-1}, \theta). \tag{A.14}$$

For chosen $v_{r-1}^1, \ldots, v_{r-1}^N$ with $\sum_{k=1}^N v_{r-1}^k = 1$ (see below), we form the importance density

$$q(\mathbf{X}_{1:r} \,|\, \theta) = \sum_{k=1}^N v_{r-1}^k p(X_r | \mathbf{X}_{1:r-1}^k, \theta) \tag{A.15}$$

which we use to sample the r^{th} particle generation. Drawing from $q(\mathbf{X}_{1:r} \,|\, \theta)$ is equivalent to first selecting, for $k = 1, \ldots, N$, particle k's parent with index A_{r-1}^k from the previous generation, with probabilities $v_{r-1}^1, \ldots, v_{r-1}^N$, and then forward simulating the disease model between the times $t = t_{r-1}$ and $t = t_r$, conditional on θ and the parent particle's history $\mathbf{X}_{1:r-1}^{A_{r-1}^k}$. When the disease model is Markovian, the state of the parent particle $X_{r-1}^{A_{r-1}^k}$ at time $t = t_{r-1}$ determines the probabilistic behaviour of its offspring, independently of its history prior to t_{r-1}. As in [ADH10], the v_{r-1}^k are typically the normalised weights W_{r-1}^k from the first particle generation. However, it is often beneficial to *smooth* the weights, for example by setting

$$\begin{aligned} &v_{r-1}^k \propto \sqrt{W_{r-1}^k} \\ &\sum_{k=1}^N v_{r-1}^k = 1. \end{aligned} \tag{A.16}$$

This pushes those normalised weights that equal neither 0 nor 1 toward the interior of the interval $[0, 1]$ so that a greater diversity of particles is propagated into the next generation. There are numerous options for how weights are smoothed, Liu, et al. [DFG01, Chapter 11] discuss an adaptive method based on calculating the *effective sample size* at each generation. However, taking the square root, as above, before normalising, again as suggested by Liu, et al., works sufficiently well for our purposes. In all applications of SMC discussed in this thesis, the weights are smoothed in this way. Since by Eq. A.14, $W_{r-1}^{A_{r-1}^k} \approx p(\mathbf{X}_{1:r-1}^{A_{r-1}^k} \,|\, \mathbf{Y}_{1:r-1}, \theta)$,

$$\begin{aligned} \frac{p(\mathbf{X}_{1:r}^k \,|\, \mathbf{Y}_{1:r}, \theta)}{q(\mathbf{X}_{1:r} \,|\, \theta)} &= \frac{p(X_r^k, \mathbf{X}_{1:r-1}^{A_{r-1}^k} \,|\, \mathbf{Y}_{1:r}, \theta)}{q(\mathbf{X}_{1:r} \,|\, \theta)} \\ &= \frac{1}{p(Y_r \,|\, \mathbf{Y}_{1:r-1}, \theta)} \times \frac{p(Y_r \,|\, X_r^k) p(X_r^k \,|\, \mathbf{X}_{1:r-1}^{A_{r-1}^k}, \theta) p(\mathbf{X}_{1:r-1}^{A_{r-1}^k} \,|\, \mathbf{Y}_{1:r-1}, \theta)}{q(\mathbf{X}_{1:r} \,|\, \theta)} \\ &= \frac{1}{p(Y_r \,|\, \mathbf{Y}_{1:r-1}, \theta)} \times \frac{p(Y_r \,|\, X_r^k) p(X_r^k \,|\, \mathbf{X}_{1:r-1}^{A_{r-1}^k}, \theta) p(\mathbf{X}_{1:r-1}^{A_{r-1}^k} \,|\, \mathbf{Y}_{1:r-1}, \theta)}{p(X_r^k \,|\, \mathbf{X}_{1:r-1}^{A_{r-1}^k}, \theta) v_{r-1}^{A_{r-1}^k}} \\ &\approx \frac{1}{p(Y_r \,|\, \mathbf{Y}_{1:r-1}, \theta)} \times \frac{W_{r-1}^{A_{r-1}^k}}{v_{r-1}^{A_{r-1}^k}} p(Y_r \,|\, X_r^k) \end{aligned} \tag{A.17}$$

if we set

$$w_r^k = w_r(\mathbf{X}_{1:r}^k) = \frac{W_{r-1}^{A_{r-1}^k}}{v_{r-1}^{A_{r-1}^k}} p(Y_r \mid X_r^k) \tag{A.18}$$

then by the law of large numbers

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N} w_r^k &\approx \int p(Y_r \mid \mathbf{Y}_{1:r-1}, \theta) \frac{p(\mathbf{X}_{1:r}^k \mid \mathbf{Y}_{1:r}, \theta)}{q(\mathbf{X}_{1:r} \mid \theta)} q(\mathbf{X}_{1:r} \mid \theta)\, d\mathbf{X}_{1:r} \\ &= p(Y_r \mid \mathbf{Y}_{1:r-1}, \theta). \end{aligned} \tag{A.19}$$

We normalise the weights

$$W_r^k = \frac{w_r^k}{\sum_{k=1}^N w_r^k} \tag{A.20}$$

and obtain the approximate distribution

$$\hat{p}(\mathbf{X}_{1:r} \mid \mathbf{Y}_{1:r}, \theta) = \sum_{k=1}^{N} W_r^k \delta_{\mathbf{X}_{1:r}^k} \approx p(\mathbf{X}_{1:r} \mid \mathbf{Y}_{1:r}, \theta). \tag{A.21}$$

Finally, after m generations

$$\begin{aligned}\prod_{i=1}^{m} \Big(\frac{1}{N}\sum_{k=1}^{N} w_i^k\Big) &\approx p(Y_m|\mathbf{Y}_{1:m-1}, \theta) \cdots p(Y_2|Y_1, \theta)p(Y_1|\theta) \\ &= p(\mathbf{Y}_{1:m}|\theta). \end{aligned} \tag{A.22}$$

E.2 Implementation

The above routine is implemented in `C++`. Because the implementation makes use of abstract classes and polymorphism, the implementation can be adapted easily for different disease models. Associated with each disease model is a namespace, under which is defined the `mint` struct that stores the states and histories of each of N particles. For example, for the SI-PQ model of Chapter 4

```
namespace SIPQ
{
struct mint
{
  int parent_id;
  int S, I1, I2;
  double P, Q;

  mint(){};
  mint(int pid, int S, int I1, int I2, double P, double Q)
  :parent_id{pid},S{S},I1{I1},I2{I2},P{P},Q{Q}{}
  void print(std::ostream&);
};
}
```

For Markovian disease models, the evolution of the next generation of particles depends only on the final compartment sizes. To allow for non-Markovian models, such as non exponentially-distributed infectious liftimes, an additional data structure, such as a `priority queue`, can be added to the `mint` struct to hold, e.g. future recovery times.

The struct `job`

```
struct job
{
  job(){};
  virtual ~job(){}

  virtual void run()=0;
  virtual void run(std::mt19937_64&)=0;
};
```

is an abstract handle by which various computation tasks related to a particular `mint`, e.g. performing a forward simulation of a particle between two measurement times or calculating the likelihood of a particular measurement, such as that of a dPCR reading, given the amount of live and dead pathogen currently in the system, $p(H_t \,|\, P_t, Q_t)$, can be called and manipulated polymorphously.

At each generation, pointers to `jobs` are loaded onto a generic job queue and a number of worker threads, running in a loop, take `jobs` off the queue and executing them. Once a thread finds the queue empty, it breaks out of the loop and exits. This is how the implementation leverages multiple cores to break down the computations and perform them in parallel batches, for example

```
void do_jobs(std::deque<job*>& jobq)
{
  job* local_jobptr;

  while(true)
    {
      job_mutex.lock();
```

```
   if( jobq.empty() )
{
 job_mutex.unlock();
 break;
}

   local_jobptr = jobq.front();
   jobq.pop_front();
   job_mutex.unlock();
   local_jobptr->run();

   delete local_jobptr;
  }
}
```

Generally, between 1000 and 1500 particles operating in 4 or 5 independent threads are found to be sufficient for the work in this thesis.

The function reshuffle sets up the correct number mints and their current weights as parents for the next generation of particles. It does this in-place by making two passes through the array of mints

```
void reshuffle(
    std::vector< std::tuple< int, int > >& uniq_cnts
    ,mint marray[num_experiments][num_particles]
    ,double* lv
    ,double* lw
    )
{
 std::vector< std::tuple< int, int > >::iterator u{uniq_cnts.begin()};

 while( std::get<1>(*u) == 0 ) ++u;

 for(int n{0}; n < num_particles; ++n)
  {
   //is particle n in the list?...
   if( any_of(uniq_cnts.begin(), uniq_cnts.end()
 ,[n](std::tuple<int, int> x){return (std::get<0>(x)==n); }))
//...yes.  Do nothing...
for( int j{0}; j < num_experiments; ++j )
 marray[j][n].parent_id = n; //...apart from set parent_id

   else //Otherwise, produce an offspring...
{
  for( int j{0}; j < num_experiments; ++j )
    {
      marray[j][n] = marray[j][std::get<0>(*u)];
      marray[j][n].parent_id = std::get<0>(*u);
    }

  lv[n] = lv[std::get<0>(*u)];
  lw[n] = lw[std::get<0>(*u)];
  --std::get<1>(*u);//...and decrement.
}
      //skip past parents with no more offspring to produce
      while( std::get<1>(*u) == 0 ) ++u;
    }

}
```

The helper function `count_uniques` takes as input a vector `a` of integer values and populates the vector `uniq_cnts` with distinct values and their counts. `a` is a vector of sampled indices from a multinomial distribution determining the parents of the next generation of particles. `count_uniques` is called by the function `reshuffle`

```
void count_uniques(
   std::vector< std::tuple< int, int > >& uniq_cnts
   ,const std::vector< int >& a
   )
{
    /*
      Assumptions:
      a is sorted

      Returns:
      vector of tuples {a,m} - particle a will have m+1 offspring
    */
#ifdef DEBUG
  assert( std::is_sorted( a.begin(), a.end() ));
#endif
  uniq_cnts.clear();

  uniq_cnts.push_back( std::tuple<int, int>{a[0], 0} );

  for( int n{1};  n < num_particles; ++n )
    {
      if( a[n] == std::get<0>( uniq_cnts.back() ))
++std::get<1>(uniq_cnts.back());

      else uniq_cnts.push_back( std::tuple<int, int>{a[n], 0} );
    }
}
```

F Sample code

F.1 Python code to implement metropolis-cooled MCMC routine from Chapter 2

The following Python2 code uses the `multiprocessing` module in the standard library to implement several parallel Metropolis-Hastings samplers comprising the Metropolis-cooled MCMC routine of Altekar, et al. (2004) [ADHR04]. The details of the sampling for each chain are contained within the `seir_sampler` member of class `SamplerThread`. The concepts of Barrier and Reusable barrier are from The Little Book of Semaphores, Allen B. Downey [Dow16].

```
import multiprocessing as m_proc
import numpy as np
import time

'''
Little Book of Semaphores
'''
class Barrier:
    def __init__(self, n):
        self.n = n
        self.count = m_proc.Value('i',0)
        self.mutex = m_proc.Semaphore(1)
        self.barrier = m_proc.Semaphore(0)

    def wait(self):
        self.mutex.acquire()
        self.count.value += 1
        self.mutex.release()

        if self.count.value == self.n:
            self.barrier.release()

        self.barrier.acquire()
        self.barrier.release()
```

```
class ReusableBarrier:
    def __init__(self, n):
        self.n = n
        self.count = m_proc.Value('i',0)
        self.mutex = m_proc.Semaphore(1)
        self.barrier_one = m_proc.Semaphore(0)
        self.barrier_two = m_proc.Semaphore(1)

    def wait_phase_one(self):
        self.mutex.acquire()
        self.count.value += 1
        if self.count.value == self.n:
            self.barrier_two.acquire() #lock second barrier
            self.barrier_one.release() #unlock first barrier
        self.mutex.release()

        self.barrier_one.acquire()
        self.barrier_one.release()

    def wait_phase_two(self):
        self.mutex.acquire()
        self.count.value -= 1
        if self.count.value == 0:
            self.barrier_one.acquire() #lock first barrier
            self.barrier_two.release() #unlock second barrier
        self.mutex.release()

        self.barrier_two.acquire()
        self.barrier_two.release()



class SamplerThread(m_proc.Process):
    def __init__(self,
                 index,
                 sampler,
                 reusable_barrier,
                 mutex,
                 connection,
                 block_size,
                 num_blocks,
                 blocks_to_burn,
                 thin,
                 file_handles,
                 sh_temperatures,
                 sh_cross_temperatures,
                 sh_likelihoods,
                 sh_cross_likelihoods,
                 sh_sigma_delta,
                 sh_write):
        m_proc.Process.__init__(self)
        self.index = index
        self.reusable_barrier = reusable_barrier
        self.mutex = mutex
        self.seir_sampler = sampler
        self.connection = connection
        self.block_size = block_size
        self.num_blocks = num_blocks
        self.blocks_to_burn = blocks_to_burn
        self.thin = thin
        self.file_handles = file_handles
        self.sh_temperatures = sh_temperatures
        self.sh_cross_temperatures = sh_cross_temperatures
        self.sh_likelihoods = sh_likelihoods
        self.sh_cross_likelihoods = sh_cross_likelihoods
        self.sh_sigma_delta = sh_sigma_delta
        self.sh_write = sh_write

    def run(self):
        '''
        burn-in phase.
        Iterate <block_size> number of times and then report current sigma_delta
        '''
        self.seir_sampler.T = self.sh_temperatures[self.index]
        self.seir_sampler.sigma_delta = self.sh_sigma_delta[self.index]
        self.seir_sampler.write = self.sh_write[self.index]

        for i in range(self.blocks_to_burn):
            self.seir_sampler.run( self.block_size, self.thin, self.file_handles, True )
            self.sh_sigma_delta[self.index] = self.seir_sampler.sigma_delta

            self.reusable_barrier.wait_phase_one()
            self.reusable_barrier.wait_phase_two()

        '''
        sampling phase.  Iterate <block_size> number of times
        '''
        for i in range(self.num_blocks):
```

```
'''
1.
main thread:
draw random transposition (k,l)
update sh_cross_temperatures (kth and lth positions transposed)
'''
self.reusable_barrier.wait_phase_one()

'''
2.
main thread waits
'''
self.reusable_barrier.wait_phase_two()



self.seir_sampler.T = self.sh_temperatures[self.index]
self.seir_sampler.sigma_delta = self.sh_sigma_delta[self.index]
self.seir_sampler.write = self.sh_write[self.index]

self.seir_sampler.run( self.block_size, self.thin, self.file_handles, False )
self.sh_likelihoods[self.index]
     = self.seir_sampler.log_likelihood_partial_integrated(
            self.seir_sampler.T)
self.sh_cross_likelihoods[self.index]
     = self.seir_sampler.log_likelihood_partial_integrated(
                                    self.sh_cross_temperatures[self.index] )



'''
3.
main thread:
calculate acceptance probability
accept / reject
if accept:
transpose sh_<>
'''
self.reusable_barrier.wait_phase_one()
self.reusable_barrier.wait_phase_two()
```

F.2 Python code to simulate results of trial by Hauge, et al. (2017) from Chapter 3

```
import time
import sys
import numpy as np

seed = int(time.time())
print 'seed = ', seed
ranD = np.random.RandomState(seed)

N1 = 20
N2 = 20

d1_v = [4, 4, 4, 4, 3]
d2_v = [4, 4, 3, 4, 5]
remtimes = [2.0, 4.0, 6.0, 12.0, 16.]

bet = 0.02
delta = 0.2

def E_sampler( E_params ):
    delta, ranD = E_params
    return ranD.exponential( scale = 1/delta )

rout = []

Hauge_SEI_Sim( 0.0, bet, E_sampler, (delta, ranD), d1_v, d2_v,
            remtimes, N1, N2, rout, ranD )


def Hauge_SEI_Sim( t0, bet, E_sampler, E_params, d1_v, d2_v, remTimes,
            N1, N2, rout, ranD ):
    ev_queue = []

    S = N2; E = 0; I1 = N1; I2 = 0
    t = t0

    for d1, d2, rt in zip(d1_v, d2_v, remTimes):
        #print t0, '-', rt, 'd1 = ', d1, 'd2 = ', d2

        while True:
            if S == 0:
                while not ev_queue == [] and ev_queue[0] < rt:
                    E -= 1
                    I2 += 1
                    heapq.heappop(ev_queue)
                break
            else:
                t += ranD.exponential( scale = 1/( bet * S * (I1 + I2) ) )

            if t >= rt:
                break
            elif not ev_queue == [] and t >= ev_queue[0]:
                t = ev_queue[0]
                E -= 1
                I2 += 1
                heapq.heappop(ev_queue)
            else:
                S -= 1
                E += 1
                heapq.heappush(ev_queue, t + E_sampler( E_params ))

            #print ’S =’, S, ’E =’, E, ’I1 =’, I1, ’I2 =’, I2, ’t =’, t, ’  ’, ev_queue

        t = rt


        #print ’============================\n’
        #print ’Removal time, t = ’, rt, ’hit’, ’(1)’, d1, ’(2)’, d2
        #print ’(before) S =’, S, ’E =’, E, ’I1 =’, I1, ’I2 =’, I2, ’t =’, t


        I1 -= d1

        rout.append(0)
        ranD.shuffle( ev_queue )
        indices = []

        for i in range(S):
            indices.append(’s’)
        for i in range(E):
            indices.append(’e’)
        for i in range(I2):
            indices.append(’i’)

        ranD.shuffle( indices )
        #print indices
```

```
for i in range(d2):
    j = indices.pop()
    if j == ’s’:
        S -= 1
    elif j == ’e’:
        E -= 1
        ev_queue.pop()
    elif j == ’i’:
        I2 -= 1
        rout[-1] += 1
    else:
        print ’Error’
        sys.exit()

heapq.heapify( ev_queue )


#print ’(after) S =’, S, ’E =’, E, ’I1 =’, I1, ’I2 =’, I2, ’t =’, t
#print ev_queue
#print ’=============================\n’
t0 = rt
#for loop ends
}
```

Colophon

This thesis is based on a template developed by Matthew Townson and Andrew Reeves. It was typeset with LaTeX 2$_\varepsilon$. It was created using the *memoir* package, maintained by Lars Madsen, with the *madsen* chapter style. The font used is Latin Modern, derived from fonts designed by Donald E. Kunith.