Bayesian data assimilation provides rapid decision support for vector-borne diseases

Chris P. Jewell1 and Richard G. Brown2

1CHICAS, Lancaster University, Bailrigg, Lancaster LA1 4YG, UK
2Institute of Fundamental Sciences, Massey University, Private Bag 11222, Palmerston North 4442, New Zealand

Predicting the spread of vector-borne diseases in response to incursions requires knowledge of both host and vector demographics in advance of an outbreak. Although host population data are typically available, for novel disease introductions there is a high chance of the pathogen using a vector for which data are unavailable. This presents a barrier to estimating the parameters of dynamical models representing host–vector–pathogen interaction, and hence limits their ability to provide quantitative risk forecasts. The Theileria orientalis (Ikeda) outbreak in New Zealand cattle demonstrates this problem: even though the vector has received extensive laboratory study, a high degree of uncertainty persists over its national demographic distribution. Addressing this, we develop a Bayesian data assimilation approach whereby indirect observations of vector activity inform a seasonal spatio-temporal risk surface within a stochastic epidemic model. We provide quantitative predictions for the future spread of the epidemic, quantifying uncertainty in the model parameters, case infection times and the disease status of undetected infections. Importantly, we demonstrate how our model learns sequentially as the epidemic unfolds and provide evidence for changing epidemic dynamics through time. Our approach therefore provides a significant advance in rapid decision support for novel vector-borne disease outbreaks.

1. Introduction

During outbreaks of infectious diseases, effective decision-making is key to implementing efficient control measures. Quantitative methods are now commonplace for supporting such decisions, and in particular mathematical models are now central to informing control strategy [1]. Before a disease outbreak, models may be used offline to investigate disease dynamics within a particular population, and hence inform policies such as childhood immunization [2] and livestock disease containment [3]. During an outbreak, models have the potential to be used for forecasting future disease spread, provided the difficult task of adequate design and calibration is performed appropriately [4]. In an era characterized by rapid climatic and sociological change, the increasing emergence of new diseases therefore requires a modelling approach that not only adapts quickly to the current outbreak, but is also flexible to the availability of data [5].

For contagious diseases such as SARS and foot and mouth disease, recent methodology has enabled models to be employed in real-time outbreak situations [6,7]. At any particular time point during the epidemic, a forecast of ongoing disease spread may be calculated given knowledge of the underlying population and case data collected so far. Such a forecast may be used not only for monitoring the overall extent of the outbreak, but also for optimal targeting of control strategies, such as surveillance and quarantine, to high-risk individuals [8]. However, mathematical models of vector-borne diseases do not appear to have been adopted for real-time forecasting purposes, despite their long history (e.g. [9–12]). A possible reason for this is that while high-quality host demographic data are often available from government databases, less is known about national distributions of vector populations. Indeed, keeping current ecological records on vector populations is subject to prioritization, such that high-quality information relevant to all possible vector-borne diseases is not guaranteed [13]. Furthermore, even where such records do exist, the...
effect of climate change on vector populations is likely to increase the rate at which existing ecological vector studies become irrelevant to new outbreaks [14–16]. Thus, a particular challenge for forecasting is to construct a model that is able to adapt quickly to the spatial and seasonal characteristics of the ecology of a new vector, without having detailed ecological data [17,18].

An important aspect of forecasting is the capacity to draw information from all available sources of data. Fitting models to data in this way allows inference about unknown model parameters, quantifies uncertainty about missing data and aids choice between competing model structures [19]. As previously discussed by Jewell et al. [7], a Bayesian approach to data assimilation and forecasting offers a number of advantages for prediction over classical approaches. Firstly, the probabilistic likelihood-based framework allows a highly flexible approach for assimilating a wide variety of available information in a way that allows the model to adapt to the availability of data [20,21]. Secondly, the ability to use data augmentation Markov-chain Monte Carlo (MCMC) methods to sample from a posterior distribution provides the opportunity to treat missing data, such as unobserved infection times, as latent variables even in large populations [22]. Finally, the forecast itself is represented by the Bayesian predictive distribution—the probability distribution of the future epidemic, conditional on what has been observed to date as well as the epidemic model structure. Importantly for decision-making, the predictive distribution rigorously quantifies uncertainty both in terms of the stochasticity of the epidemic process and parameter estimation [23].

In this paper, we present for the first time a fully Bayesian approach to inference and forecasting for vector-borne diseases in the absence of detailed information on vector ecology. Motivated by an outbreak of a novel tick-borne pathogen in New Zealand (NZ) cattle, we construct an epidemic model that represents heterogeneity in both the host and vector populations. Most significantly, we capture spatio-temporal variation in disease transmission due to vector abundance using a seasonal discrete-space latent risk surface, informed by indirectly collected serosurveillance data, heuristic expert opinion, and the case time series itself. A trans-dimensional MCMC algorithm is used to fit the model to the available data at various time points throughout the outbreak, from which we present forecasts of ongoing disease spread.

2. Motivating example

This study was motivated by a recent epidemic of theileriosis in NZ cattle, caused by the vector-borne pathogen *Theileria orientalis* (Ikeda) [24]. While many *T. orientalis* genotypes are endemic in NZ, and cause only rare cases of clinical disease, the Ikeda genotype appears to be associated with high morbidity and mortality haemolytic anaemia [25,26]. The tick *Haemaphysalis longicornis* is a putative vector for *T. orientalis* spp. in NZ and is known to be endemic throughout the North Island [27]. This vector is sensitive to climatic conditions and, therefore, has a spatially varying distribution as well as a seasonal pattern of activity [28]. Alternative vectors have been hypothesized, in particular the *Stomoxys calcitrans* stablefly, which is active in the same regions as the tick [29]. Beyond laboratory studies, however, little is known about the quantitative relationship between climatic factors and vector abundance in the environment. The likely determinants of disease spread are therefore environmental factors related to vector presence, distance from infected herds and animal movements.

2.1. Case report data

The first case of bovine theileriosis in NZ due to *T. orientalis* (Ikeda) was detected on the 12 August 2012 [24,30]. By 1 August 2014, the outbreak had grown to 633 infected cattle herds, concentrated largely in the Waikato region of the North Island. Figure 1 shows the spatial distribution of cases as well as the log cumulative case detections time series. The latter demonstrates a pronounced seasonality, with periods of low case incidence corresponding to winter and summer, and higher incidence in autumn and spring. This is probably due to the seasonality of the putative vector populations, though increased physiological stress experienced by cattle post-calving in late winter (August–September) may also have an effect.

The case data are curated by AsureQuality’s Agribase™ database. Each case is assigned a unique identifier and a detection time. Cases are joined to the Farms Online demographic data (see below) using spatial queries based on farm geospatial polygon records.

2.2. Demographic data

Demographic data characterizing the NZ cattle population were obtained from Farms Online (FOL), the NZ government-owned database of rural properties, as well as from the National Animal Identification and Tracking (NAIT) cattle movement database.

The FOL data comprised a list of 220,668 premises each with a unique identifier (FOL ID), geographical centroid using the New Zealand Transverse Mercator projection (EPSG:2936) and the number of beef and dairy cattle. To identify those farms owning cattle, we limited our working dataset to included herds with IDs present in the NAIT movement records (see below), or being listed as having at least one beef or dairy animal, or having been detected as a *T. orientalis* (Ikeda) case. The resultant working dataset contained 100,288 herds.

Cattle movement data were supplied from NAIT in the form of 611,230 animal movement records spanning the 557 day period from 1 January 2012 to 31 July 2013. Each record represents a cohort of cattle moved and includes the date, source and destination identifier for the respective persons in charge of animals (PICA), and number of animals moved. The NAIT data are represented by a geolocated dynamic network, with nodes corresponding to the 100,288 herds in the FOL dataset, and directed edges weighted by the frequency of animal movements (see the electronic supplementary material). In total, 520,940 non-zero directed edges were identified, representing 0.005% of the maximum possible edges in the network (i.e. 100,288²).

2.3. Spatio-temporal vector distribution

In the absence of direct observations, data on spatial vector abundance on a national scale were available in the form of expert entomological opinion and indirect observations of
herds exposed to any *T. orientalis* genotype. These were supplied as areal aggregations for 72 Territorial Land Authority (TLA) regions across NZ.

Expert opinion on the spatial distribution of the putative vector *H. longicornis* was obtained from Dr Allen Heath, encoded by classifying TLAs into high, medium and low tick risk (figure 1). Laboratory studies of ixodid tick species indicate threshold effects of both humidity and temperature on development and activity. For example, Stafford [31] demonstrates greatly increased mortality below approximately 93% relative humidity, and Ogden et al. [32] conclude a sharp-shouldered power-law curve for reproductive activity in response to temperature. These findings are consistent with both Heath [28] and Knüll & Rudolph [33], suggesting steeply increasing and decreasing tick activity in response to seasonal climatic variation.

After the discovery of the *T. orientalis* Ikeda outbreak, stored blood samples that had been collected from NZ cattle herds in conjunction with routine BVD surveillance were PCR tested for the presence of *T. orientalis* subspecies [34,35]. These data provide the number of farms tested and the number of farms returning positive for *T. orientalis* in each TLA region. These data allow us to calculate the apparent prevalence of endemic *T. orientalis* strains which serves as

---

**Figure 1.** Summary of the NZ *T. orientalis* Ikeda epidemic as of 1 August 2014. Top row: spatial distribution of infected herds, with the cumulative number of case detections throughout the epidemic. Bottom row: median prior and posterior tick occurrence probabilities for NZ TLAs.
an indirect measure of vector activity, without having to specifically identify the vector species.

3. Modelling

3.1. Epidemic model

The occurrence of cases of *T. orientalis* (Ikeda) in the NZ cattle population is represented by a continuous-time SID (susceptible–infected–detected) model, which assumes that herds progress from susceptible to infected, and are subsequently detected according to a time inhomogeneous Poisson process [7]. In contrast to other epidemic scenarios, no 'removed' status is used, since we assume that once *T. orientalis* Ikeda is circulating within the herd, the herd remains infectious. In this sense, our basic model set-up resembles an SI model, with infection times being indirectly observed via detection events.

The effects of spatial location, cattle breed, NAIT-recorded animal movements, tick density and seasonality are captured by specifying a model for the pairwise disease transmission rate. We assume that at time *t*, a susceptible herd *j* ∈ *S(t)* experiences infectious pressure at rate

$$
\lambda_j(t) = \sum_{i \in I(t)} \beta_{ij}(t),
$$

(3.1)

where *S(t)* and *I(t)* are the sets of susceptible and infected herds at time *t*, respectively. Further, we assume that infection is transmitted between an infected herd *i* and susceptible *j* at time *t* at rate

$$
\beta_{ij}(t) = h(j; t; \Phi)[\beta_j K(i, j; \delta) + \beta_2 c_{ij}],
$$

(3.2)

where *h(j; t; \Phi) ≥ 0* represents susceptibility of farm *j* at time *t*. *K(i, j; \delta)* is a function describing the decay of transmissibility with distance between herds for non-network infections, with \(\beta_1\) the baseline rate of disease transmission. The daily frequency of animal movements between farms *i* and *j* is denoted *c*\(_{ij}\) with parameter \(\beta_2\) interpreted as the maximal probability of an animal movement resulting in infection.

We define

$$
K(i, j; \delta) = \frac{\delta}{(\delta^2 + ||x_i - x_j||^2)^{3/2}},
$$

(3.3)

a Cauchy-type decay kernel with distance \(||x_i - x_j||\) between locations *x*\(_i\) and *x*\(_j\), and decay parameter \(\delta\). \(\omega = 1.2\) is chosen to optimize statistical identifiability between \(\beta_1\) and \(\delta\).

Given the paucity of quantitative data relating vector activity to measurable spatial and temporal climatic data, as well as uncertainty in breed susceptibility to *T. orientalis* Ikeda, we model seasonal infection risk as a separable spatio-temporal latent process

$$
h(j; t; \Phi) = s(t; \alpha, \nu) \zeta_{t0} p_{i(j)},
$$

(3.3)

where \(\Phi = (\alpha, \nu, \zeta, \pi)\), \(s(t; \alpha, \nu)\) represents the seasonal transmission risk at time *t*, and \(\zeta\) represents the susceptibility of dairy farms relative to non-dairy; \(\kappa_j = 1\) if *j* is a dairy farm and 0 otherwise. The probability parameter \(p_{i(j)}\) is a proxy for vector occurrence on farm *j*, a member of TLA region *k*, and allows us to connect the epidemic model to independent *T. orientalis* surveillance.

The biannual pattern of theileriosis incidence indicates the need for a flexible seasonal function capable of capturing differences in transmission peaks and troughs throughout the year. The observed threshold effects of humidity and temperature on tick activity suggests that a steep-shouldered function to be more appropriate [18]. To address this, we first adopted a piecewise cubic spline function as an analytically tractable approximation to a trigonometric function (see the electronic supplementary material). Surprisingly, on the basis of in-sample predictive performance (figure 2), this function was rejected in favour of a periodic square wave function with changepoints fixed at quarter-year epochs:

$$
s(t; \alpha, \nu) = \begin{cases} 
1 & \text{if } 0 \leq t^* < 0.25 \\
\alpha_1 & \text{if } 0.25 \leq t^* < 0.5 \\
\alpha_2 & \text{if } 0.5 \leq t^* < 0.75 \\
\alpha_3 & \text{if } 0.75 \leq t^* < 1 
\end{cases} 
= t + v - \lfloor t + v \rfloor,
$$

(3.4)

with *t* in years. Setting the height of the autumn peak to 1 enables identifiability between the \(\alpha\) and \(\beta = (\beta_1, \beta_2)\), and

Figure 2. Six-month in-sample prediction obtained by simulating forward from 1 February 2014, using parameter estimates generated as of 1 August 2014. Piecewise cubic spline (left) and square wave (right) seasonal functions. Red smudge: the probability distribution over the predicted epidemic trajectory with 95% quantiles marked by the dotted lines. Blue line: observed cumulative cases.
we allow $0 \leq v \leq 0.5$ to allow fine-tuning of the phase of the seasonal function.

We define the infectious period $d$ to be the time between an infection and detection. We assume that for each individual $i$, the infectious period is conditionally independent given the infection times, and distributed according to

$$d_i \sim \text{Gamma}(\alpha, \beta)$$

with $\alpha = 4$ based on previous infectious disease analyses (e.g. [7]), and $\beta$ an unknown scale parameter.

### 3.2. Surveillance model

The specification of $p_k$ as the occurrence probability for ticks in TLA region $k$ above presents the opportunity to model an independent disease testing process alongside the epidemic. From samples obtained from BVD surveillance, we have for each TLA region $k$ the number of herds tested, $n_k$, and number testing positive for $T. orientalis$ species, $x_k$. We assume a Binomial model such that

$$x_k \sim \text{Binomial}(n_k, p_k),$$

allowing us to make inference on $p_k$ for each TLA region. We remark that $p_k$ is only a proxy for vector occurrence, since the link between the number of cases testing positive and vector activity is complicated by many factors including the exposure of the host to the vector, host genetics and test sensitivity. As $T. orientalis$ requires a vector for transmission between cattle hosts, $p_k$ may then be thought of as a measure of the risk that infection will spread through a herd, given that it is introduced.

### 4. Data assimilation and model fitting

In this section, we describe in outline how the epidemic and surveillance models may be used together to estimate the joint posterior distribution of the model parameters, infection times and occult infections. This then facilitates the calculation of the predictive distribution of the epidemic. The implementation for model fitting and simulation is available as an R package at https://github.com/chrismdwk/infer/releases/tag/nztheliera-v1.0.

We proceed by assuming the epidemic process and BVD surveillance programme to be independent. This allows us to multiply the statistical likelihood functions for the epidemic and surveillance models—$L_E(\theta, I|D)$ and $L_s(p|X)$, respectively—to obtain a joint likelihood function for parameters $\theta = \{\alpha, \beta_1, \beta_2, \gamma, p, b\}$ and unobserved (and undetected) infection times $I$, given the case detections $D$ and surveillance data $X$ (see the electronic supplementary material for full details).

The joint posterior distribution function $\pi(\theta, I|D, X)$ is proportional to the product of the joint likelihood function and prior distributions $f_\theta(\theta)$ for each parameter

$$\pi(\theta, I|D, X) \propto L_E(\theta, I|D)L_s(p|X)\prod_\theta f_\theta(\theta),$$

which allows inference on tick occurrence probabilities $p$ to be informed by both the epidemic data and the BVD sampling data. $L_E(\theta, I|D)$ takes the form of a continuous-time inhomogeneous Poisson process likelihood, where individuals become infected according to an exponential distribution with rate given by the infectious pressure $\lambda_i(t)$ from equation (3.1). $L_s(p|X)$ then takes a binomial likelihood function for independent observations from each TLA region.

Prior probability distributions are chosen for each parameter, informed by expert opinion and heuristic expectation of the resultant epidemic (see the electronic supplementary material). The latter was obtained by simulation exploration of the behaviour of the model without consideration of the case detection time series to date. To incorporate expert opinion on spatial tick distribution, independent Beta($a_k, b_k$) prior distributions are chosen to reflect ‘high’, ‘medium’ and ‘low’ risk and are applied to $p_k$ for each region corresponding to the expert-classified TLA regions. The properties of the Beta distributions chosen are shown in the electronic supplementary material.

The Bayesian model was fitted to the observed case data using a modification of the adaptive reversible jump MCMC algorithm presented in [7]. This algorithm performs inference by drawing samples from the joint posterior distribution over the model parameters, infection times and occult infections. In our particular implementation, we use ASIS methodology to enable the algorithm to work efficiently in the face of strong a priori dependence between the marginal posterior distributions for the infection times and infectious period scale parameter $b$ [36]. Convergence of the algorithm was confirmed by running four parallel independent chains starting at randomly chosen values, as shown in the electronic supplementary material.

Having estimated the joint posterior distribution, we employ a continuous-time Doob–Gillespie simulation algorithm to construct the posterior predictive distribution of the future epidemic, with retrospective sampling used to account for the seasonal function [7]. While being less computationally efficient than a discrete time algorithm, this approach avoids discretization error which might bias the resulting disease forecast. The predictive distribution $f_Y(Y|D, X)$ of the future epidemic $Y$ conditional on the observed case detections and surveillance data are calculated by simulating over the joint posterior distribution such that

$$f_Y(Y|D, X) = \int_{\Theta} f_Y(Y|D, X, \theta)\pi(\theta, I|D, X)d\theta.$$  

### 5. Results

The analysis of the NZ $T. orientalis$ (Ikeda) outbreak began on the 1 November 2013, once it became clear that the epidemic had established. Ongoing predictive analyses were made on a monthly basis as updates to the case detection dataset were obtained. We summarize these results at quarterly intervals—1 November 2013, 1 February 2014, 1 May 2014 and 1 August 2014—indicating how the predictions adapt in the face of learning from increasing case data. Here, we focus on results relevant to forecasting, though further results relevant to now-casting may be found in the electronic supplementary material.

Early in the analyses, it became apparent that the cubic spline seasonal function was inadequate to give sufficient posterior prediction power both for in-sample and out-of-sample predictions. Figure 2 presents the in-sample prediction for the August analysis, simulated from 1 February. This clearly shows that the cubic seasonal function was inadequate to give sufficient posterior prediction power both for in-sample and out-of-sample predictions. However, the discontinuities in the posterior distribution introduced by the square wave presents significant difficulties in achieving MCMC convergence for the
period parameter \( v \). We therefore assumed \( v = 19/365 \) from expert opinion [37]. The square wave was used for all subsequent results and is represented in figure 3. The model suggests that most infection occurs in the third epoch, corresponding to mid-June to mid-September for peak disease transmissibility.

The marginal prior and posterior distributions for the tick occurrence vector \( p \) as of 1 August 2014 are summarized as median values in figure 1. These posterior median values allow for population density due to the spatial kernel and, therefore, provide information on farm density adjusted regional transmission risk. On a national scale, the posterior medians show a similar tick distribution to the prior, as expected from the distribution of epidemic cases and expert opinion. However, marked regional heterogeneity is present in the North Island compared to the prior, reflecting a synthesis of apparent tick prevalence from sampling and regional differences in disease transmission. We note that while the blood sampling data are the main influence on posterior tick occurrence, the effect of joint modelling with the epidemic data have a moderating effect on individual TLA regions (electronic supplementary material, figure S1).

To assess the effect of time in our sequential analyses, figure 3 summarizes the marginal posterior distributions for key parameters in equation (3.2). A decrease in \( \beta_2 \) during 2014 indicates a decreased importance of the cattle movement network in transmitting disease, concomitantly with a marked decrease in environmental transmission rate after February 2014 as seen by the median of the posterior \( \beta_1 K(i, j; d) \) spatial function. The posterior distributions for \( z \), the susceptibility of dairy herds versus non-dairy shows two populations of distributions. Here, dairy farms in the analyses prior to March 2014 are estimated to be approximately eight times as susceptible as non-dairy farms, whereas for the May and August 2014 this drops to approximately five times. We note that these graphs are both consistent with the flattening of the logged cumulative case curve in figure 1. Interestingly, whereas the model consistently estimates autumn and spring transmission to be negligible, an increasing trend is seen for the height of the spring seasonal function. This indicates that although the overall apparent transmission rate is tending to decrease with time, there is increased evidence for disease spread being concentrated during the winter, given the

Figure 3. Prior and posterior distributions at the four analysis times for (a) the susceptibility of a dairy farm compared to non-dairy \( \zeta \) (top left); (b) the probability of a NAIT movement transmitting an infection \( \beta_2 \) (top right); (c) the median posterior seasonal step function shown for each analysis time—violin plots of the marginal posterior densities for each analysis are shown for \( \alpha_1 \), \( \alpha_2 \) and \( \alpha_3 \), corresponding to epochs 2, 3 and 4, respectively; (d) the median environmental transmission with distance \( \beta_1 K(i, j; d) \) (bottom right).
acquisition of increasing amounts of data. The infectious period (the time between a herd’s infection and detection events) is estimated consistently at 73 days (see the electronic supplementary material), consistent with the lag between seasonal transmission during the winter, and the marked increase in case detection rate observed in the spring.

A critical quantity in determining policy for a given disease outbreak is the predicted size and extent of the epidemic. Figure 4 presents six-month ahead predictions of cumulative numbers of cases detected based on the three analyses prior to August. Increases in the rate of case detections are predicted for the autumn and spring periods. This is due to the phase of the seasonal function increasing the transmission rate during the summer and winter periods in combination with the 73 day mean infection to detection time. In a sequential setting, we evaluate out-of-sample predictive ability by comparing the predictive distributions with the subsequent cumulative case detections curve. An over-prediction is initially seen for both the November and February analyses. For the May analysis, this is much less apparent, with the true number of case detections by 1 August lying on the 0.01 percentile of the predictive distribution. The improvement in this prediction relates to the decreasing transmission rates and concentration of the infection risk into the winter period as previously discussed.

The predicted spatial extent of the epidemic is represented by the probability of individual herds becoming infected by six months ahead of the analysis times, shown in figure 5. These maps reflect the subsequent spatial pattern of the epidemic as of 1 August (figure 1) and therefore are indicative of the likely extent of the epidemic being confined to the north of the North Island in the medium term. The large number of farms further South in the North Island account for the significant number of subsequent cases occurring outside this area (figure 1) event though individual infection probabilities are low. We note, however, that predictions at the individual herd level are likely to be inaccurate due to local model inadequacy and population data inaccuracies.

6. Discussion

The motivation behind this epidemic analysis was to provide rapid predictions in response to a sudden incursion of the novel strain of vector-borne protozoan *T. orientalis* (Ikeda) in NZ cattle. Little is known about the national level spatio-temporal dynamics of the tick vector, and we have shown that a parsimonious approach which models the vector population as a seasonal discrete-space latent risk surface is a viable option for serial prediction purposes. The results indicate an epidemic determined by both vector presence and environmental transmission, with movements recorded in the NAIT database having a low risk of propagating infection.

The main feature of our forecasting approach is to jointly model independent disease surveillance results with epidemic data. The *a posteriori* estimates of the parameter
vector $p$ therefore reflect a synthesis of static sample-based data and the dynamic epidemic data. A more accurate interpretation of $p$ is therefore a proxy for vector-driven disease transmission in each TLA region. Of concern, however, is the level of spatial discretization used for the BVD sampling data, though this was necessary to obtain anonymized data for the sampled farms. We do not expect tick activity to be constant across TLA regions, nor do we expect a step change in tick activity across region borders. Future research will therefore focus on integrating the recently characterized class of log Gaussian Cox processes for inference on continuous space risk surfaces given point data in preference to areal aggregations [38].

In our analyses, we have compared our models against subsequently observed data using both in-sample and out-of-sample comparisons. Early in the analysis, in-sample assessment of predictive performance quickly identified a strong preference for the square wave seasonal function over the cubic spline, consistent with studies of the effect of humidity and temperature on the activity of ixodid ticks [16,31–33,39]. We conclude therefore that in terms of disease transmission, these threshold effects are mimicked well by the square wave. In statistical terms, the disadvantage of the square wave is its effect on the mixing quality of the MCMC, which is caused by the discontinuous nature of the posterior distribution with respect to changes in $v$. A more elaborate cubic spline function, designed as a continuous approximation to the square wave, may well alleviate this particular difficulty albeit with the introduction of more parameters. However, given that this dataset provides observations for only two replicates of an annual seasonal pattern, it is likely that a more complex model would exhibit a loss of statistical identifiability, again interfering with efficient model fitting. We note also that non-identifiability between the phase ($v$) and infectious period ($b$) parameters is inherent to any periodic function, as the majority of infection times are dictated by the season in which most disease transmission occurs. Thus, MCMC mixing issues are still apparent even for smooth functions. While further research is required to identify alternative seasonal functions, a promising approach to resolving this problem is to incorporate climatic covariates into the analysis. Quantities such as vapour pressure, relative humidity and temperature at noon are all known to affect tick activity, and estimating their effects as a hierarchical component within the $h(j,t;\psi)$ function (equation (3.3)) represents a straightforward extension to our model.

The out-of-sample predictive accuracy of our results changes markedly throughout the epidemic as the spatio-temporal case detection data increase in volume. In terms of the number of cases over time, the November and February analyses over predict the number of future case detections by a large margin, with early exponential growth dominating the rate of new case discovery far into 2014. However, the subsequent observed cumulative case time series (figure 1) shows a slowing of case detection rate in

Figure 5. Six-month predicted individual infection probability, conditional on the observed epidemic as of 1 November 2013, 1 February 2014, 1 May 2014 and 1 August 2014.
comparison to exponential growth. This is only captured at the May analysis which predicts the subsequent three month period far better with the acquisition of 126 new cases since the February analysis. As such, our results are consistent with the tendency of epidemic models to over predict numbers of cases, as is common due to unidentified heterogeneity in the population (e.g. [40,41]). Features such as the timing of epidemic peaks, seasonal effects and spatial extent are however generally well identified. Additionally, it is likely that a downward reporting bias is occurring as well as a genuine slowing of the transmission rate as the cattle industry adapts to the outbreak.

A striking difference between our results for NZ theileriosis and previous results from directly communicable diseases in animal populations is the decay rate of the environmental transmission with distance. Previous studies in foot and mouth disease, and avian and equine influenza indicate that the majority of herd to herd transmission occurs within 5 km [8,42–44]. By contrast, our results suggest that environmental transmission of T. orientalis (Ikeda) occurs over much greater distances. We propose two possible explanations for this finding. Firstly, though ticks are relatively short ranging arthropods (in comparison to flying insect vectors), wildlife hosts may be capable of translocating infected ticks over long distances [45]. Haemaphysalis longicornis has three life cycle stages—larva, nymph and adult—with each stage feeding on a host [28]. In principle, then, it may be possible for an adult, which has ingested T. orientalis as a larva, to transmit the infection to a host in a remote location, after translocation at the nymphaal stage. We note that this may not require that the nymphal stage host be competent for T. orientalis. Secondly, a more plausible explanation lies in the accuracy of NAIT-recorded movements with respect to actual cattle movements around NZ, and also the accuracy of joining Agribase™ case identifiers to FOL and NAIT records. We note that since the sparsity of the NAIT network is high, inaccuracies in georeferencing animal movements will have a marked downward bias on β2. Additionally, NAIT is a nascent movement recording system and we believe that, even though cattle movement recording is mandatory, compliance may be low. In the absence of a national movement ban, it is therefore highly likely that the apparent long-range spatial transmission observed here is a result of reporting bias: the spatial transmission kernel compensates for unrecorded animal movements, with a corresponding downward bias on β2. That T. orientalis (Ikeda) can be transmitted by the movement of tick-infested cattle is supported both by common sense and anecdotal evidence. For example, Islam et al. [26] provide strong evidence for such a mode of infection through genetic typing of the pathogen. However, even via this mode of transmission, infection of a naive herd depends on local environmental conditions conducive to tick survival [25]. We therefore conclude that risk of infection via incoming animal movements is determined by the geographical regions and time periods within which the vector population is active.

The availability of demographic and ecological data will always be the limiting factor for detailed dynamical models of disease. While maintaining databases on livestock industry demographics is commonly carried out at the national level by government bodies, keeping pace with all possible vector populations in the face of changing climate and habitats is economically unfeasible. Bayesian data assimilation and inference therefore provides a robust and rigorous solution for quantitative decision support in disease response situations.

Data accessibility. The C++ code implementing the simulation and MCMC algorithms used in this study may be found at https://github.com/christomdvwk/infer/releases/tag/tztheileria-v1.0, released under the GNU Public Licence v. 3. Access to the underlying data is protected by confidentiality agreement between Massey University and the Ministry for Primary Industries, NZ. Requests for these data should be directed to the Ministry for Primary Industries, NZ.

Authors’ contributions. C.P.J. was responsible for obtaining the data, formulation of the epidemic model, coding of the Gillespie and MCMC algorithms and writing the manuscript. R.G.B. provided the cubic spline formulation of the seasonal component of the model, and helped with formatting of the composite figures and copy-editing of the manuscript.

Competing interests. We declare we have no competing interests.

Funding. We received no funding for this study.

Acknowledgements. The authors gratefully thank Andy McFadden, Mary van Andel, Daan Vink and Kevin Lawrence at The Ministry for Primary Industries, and Robertanson at Assure Quality for supplying all data used in this study. We thank Dr Allen Heath for his helpful discussions on arthropod vectors for Theileria, and for supplying the prior information summarized in figure 1.

References
