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

Chris P Jewell[1] and Richard G Brown[2]

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

Summary

Predicting the spread of vector-borne diseases in response to incursions requires knowledge of both host and vector demographics in advance of an outbreak. Whereas host population data is typically available, for novel disease introductions there is a high chance of the pathogen utilising a vector for which data is 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 provides evidence for changing epidemic dynamics through time. Our approach therefore provides a significant advance in rapid decision support for novel vector-borne disease outbreaks.

Keywords vector-borne disease | seasonal epidemic | Bayesian inference | risk forecasting | MCMC

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 immunisation [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 characterised 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, 3]. 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 (see for example [9, 10, 11, 12]). A possible reason for this is that whilst 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 prioritisation, 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, 15, 16]. Thus a particular challenge for forecasting is to construct a model which is able to adapt quickly to the spatial and seasonal characteristics of a new vector ecology, 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 in Jewell et al.[3], 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, 1]. Secondly, the ability to use data augmentation Markov-chain Monte Carlo 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 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 timeseries itself. A trans-dimensional Markov-chain Monte Carlo algorithm is used to fit the model to the available data at various timepoints 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 New Zealand cattle, caused by the vector-borne pathogen *Theileria orientalis* (Ikeda) [24]. Whilst 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 New Zealand, 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 hypothesised, 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 12th August 2012 [30, 24]. By 1st 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 timeseries. 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 likely 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 is 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 characterising the NZ cattle population were obtained from Farms OnLine (FOL), the New Zealand 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 220668 premises each with a unique identifier (FOL ID), geographic 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 100288 herds.

Cattle movement data was supplied from NAIT in the form of 611230 animal movement records spanning the 557 day period from 1st January 2012 to 31st 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 is represented by a geolocated dynamic network, with nodes corresponding to the 100288 herds in the FOL dataset, and directed edges weighted by the frequency of animal movements (see supporting information). In total, 520940 non-zero directed edges were identified, representing 0.005% of the maximum possible edges in the network (i.e. 100288^2).

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 *Haemaphysallis 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, [31] demonstrate greatly increased mortality below approximately 93% relative humidity, and [32] conclude a sharp-shouldered power law curve for reproductive activity in response to temperature. These findings are consistent with both [28] and [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 a 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 model, which assumes that herds progress from Susceptible to Infected, and are subsequently Detected according to a time inhomogeneous Poisson process [3]. 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 setup 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 $j \in \mathcal{S}(t)$ experiences infectious pressure at rate

$$\lambda_j(t) = \sum_{i \in \mathcal{I}(t)} \beta_{ij}(t) \tag{1}$$

where $\mathcal{S}(t)$ and $\mathcal{I}(t)$ 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

$$\beta_{ij}(t) = h(j, t; \boldsymbol{\psi}) \left[\beta_1 K(i, j; \delta) + \beta_2 c_{ij}\right] \tag{2}$$

where $h(j, t; \boldsymbol{\psi}) \geq 0$ 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 β_1 the baseline rate of disease transmission. The daily frequency of animal movements between farms i and j is denoted c_{ij}, with parameter β_2 interpreted as the maximal probability of an animal movement resulting in infection.

We define

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

a Cauchy-type decay kernel with distance $||x_i - x_j||$ between locations x_i and x_j, and decay parameter δ. $\omega = 1.2$ is chosen to optimise statistical identifiability between β_1 and δ.

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 spatiotemporal latent process

$$h(j, t; \boldsymbol{\psi}) = s(t; \boldsymbol{\alpha}, \nu) \zeta^{\kappa_j} p_{k(j)} \tag{3}$$

where $\boldsymbol{\psi} = \{\boldsymbol{\alpha}, \nu, \zeta, \boldsymbol{p}\}$. $s(t; \boldsymbol{\alpha}, \nu)$ represents the seasonal transmission risk at time t, and ζ represents the susceptibility of dairy farms relative to non-dairy; κ_j is 1 if j is a dairy farm and 0 otherwise. The probability parameter $p_{k(j)}$ 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 approximated by a square wave might be appropriate. However, the combined effect of humidity and temperature on the vector, the degree of uncertainty surrounding the identity of other vector species, and the fact that this model is capturing the occurrence of new cases rather than the vector itself, might suggest a sinusoidal 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 supporting information). 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:

$$s(t;\boldsymbol{\alpha},\nu) = \begin{cases} 1 & \text{if } 0 \leq t^\star < 0.25 \\ \alpha_1 & \text{if } 0.25 \leq t^\star < 0.5 \\ \alpha_2 & \text{if } 0.5 \leq t^\star < 0.75 \\ \alpha_3 & \text{if } 0.75 \leq t^\star < 1 \end{cases}, t^\star = t + \nu - \lfloor t + \nu \rfloor \tag{4}$$

with t in years. Setting the height of the autumn peak to 1 enables identifiability between the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta} = \{\beta_1, \beta_2\}$, and we allow $0 \leq \nu \leq 0.5$ 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

$$d_i \sim \text{Gamma}(a, b)$$

with $a = 4$ based on previous infectious disease analyses (see for example [3]), and b 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

$$x_k \sim \text{Binomial}(n_k, p_k) \tag{5}$$

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 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. Since *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 of detected herds, and the presence of undetected 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/chrism0dwk/infer/releases/tag/nztheileria-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 – $L_E(\boldsymbol{\theta}, \boldsymbol{I}|\boldsymbol{D})$ and $L_S(\boldsymbol{p}|\boldsymbol{X})$ respectively – to obtain a joint likelihood function for parameters $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \beta_1, \beta_2, \nu, \zeta, \boldsymbol{p}, b\}$ and unobserved (and undetected) infection times $\boldsymbol{I}$, given the case detections $\boldsymbol{D}$ and surveillance data $\boldsymbol{X}$ (see supporting information for full details).

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

$$\pi(\boldsymbol{\theta}, \boldsymbol{I}|\boldsymbol{D}, \boldsymbol{X}) \propto L_E(\boldsymbol{\theta}, \boldsymbol{I}|\boldsymbol{D}) L_S(\boldsymbol{p}|\boldsymbol{X}) \prod_{\boldsymbol{\theta}} f_\theta(\theta) \tag{6}$$

which allows inference on tick occurrence probabilities $\boldsymbol{p}$ to be informed by both the epidemic data *and* the BVD sampling data. $L_E(\boldsymbol{\theta}|\boldsymbol{I},\boldsymbol{D})$ 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 $\lambda_j(t)$ from Equation 1. $L_S(\boldsymbol{p}|\boldsymbol{X})$ 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 supporting information). The latter was obtained by simulation exploration of the behaviour of the model without consideration of the case detection timeseries 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 supporting information.

The Bayesian model was fitted to the observed case data using a modification of the adaptive reversible jump Markov chain Monte Carlo algorithm presented in [3]. 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 4 parallel independent chains starting at randomly chosen values, as shown in the supporting information.

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 [3]. Whilst being less computationally efficient than a discrete time algorithm, this approach avoids discretisation error which might bias the resulting disease forecast. The predictive distribution $f_Y(Y|\boldsymbol{D},\boldsymbol{X})$ of the future epidemic Y conditional on the observed case detections and surveillance data is calculated by simulating over the joint posterior distribution such that

$$f_Y(Y|\boldsymbol{I},\boldsymbol{D},\boldsymbol{X}) = \int_{\boldsymbol{\Theta},\boldsymbol{I}} f_Y(Y|\boldsymbol{D},\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{I})\pi(\boldsymbol{\theta},\boldsymbol{I}|\boldsymbol{D},\boldsymbol{X})d\boldsymbol{\theta}d\boldsymbol{I} \tag{7}$$

5 Results

The analysis of the New Zealand *Theileria orientalis* (Ikeda) outbreak began on the 1st 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 summarise these results at quarterly intervals – 1st November 2013, 1st February 2014, 1st May 2014, and 1st 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 supporting information.

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 1st February. This clearly shows that the cubic spline leads to a marked over-prediction in the size of the epidemic, whereas the square wave leads to a greatly superior in-sample prediction. However, the discontinuities in the posterior distribution introduced by the square wave presents significant difficulties in achieving MCMC convergence for the period parameter ν. We therefore assumed $\nu = 19/365$ 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 $\boldsymbol{p}$ as of 1st August 2014 are summarised 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 whilst the blood sampling data is the main influence on posterior tick occurrence, the effect of joint modelling with the epidemic data has a moderating effect on individual TLA regions (Figure S1 in supporting information).

To assess the effect of time in our sequential analyses, Figure 3 summarises the marginal posterior distributions for key parameters in Equation 2. A decrease in β_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 $\beta_1 K(i,j;\delta)$ spatial function. The posterior distributions for ζ, 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 8 times as susceptible as non-dairy farms, whereas for the May and August 2014 this drops to approximately 5 times. We notes 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 negibigble, 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 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 supporting information), 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 6-month ahead predictions of cumulative numbers of cases detected based on the 3 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 August 1st 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 6 months ahead of the analysis times, shown in Figure 5. These maps reflect the subsequent spatial pattern of the epidemic as of 1st 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 $\boldsymbol{p}$ therefore reflect a synthesis of static sample-based data and the dynamic epidemic data. A more accurate interpretation of $\boldsymbol{p}$ is therefore as a proxy for vector-driven disease transmission in each TLA region. Of concern, however, is the level of spatial discretisation used for the BVD sampling data, though this was necessary to obtain anonymised 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 characterised 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 [39, 33, 31, 32, 16]. 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 ν. 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 (ν) 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. Whilst 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 $h(j, t; \boldsymbol{\psi})$ function (Equation 3) represents a straightforward extension to our model.

The out-of-sample predictive accuracy of our results changes markedly throughout the epidemic as the spatiotemporal 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 timeseries (Figure 1) shows a slowing of case detection rate in comparison to exponential growth. This is only captured at the May analysis which predicts the subsequent 3 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 [40, 41, for example]. 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 5km [42, 43, 8, 44]. In 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]. *H. longicornis* has three lifecycle stages – larva, nymph, 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 nymphal 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 β_2. 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 β_2. 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, [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 naïve 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. Whilst 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.

Acknowledgments

The authors gratefully thank Andy McFadden, Mary van Andel, Daan Vink, and Kevin Lawrence at The Ministry for Primary Industries, and Robert Sanson at Asure 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 summarised in Figure 1.

References

[1] Roche S, Garner M, Sanson R, Cook C, Birch C, Backer J, et al. Evaluating vaccination strategies to control foot-and-mouth disease: a model comparison study. Epidemiology and infection. 2014;p. 1–20.

[2] Grenfell B, Bjørnstad O, Kappey J. Travelling waves and spatial hierarchies in measles epidemics. Nature. 2001;414(6865):716–723.

[3] Tildesley M, Savill N, Shaw D, Deardon R, Brooks S, Woolhouse M, et al. Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. Nature. 2006;440(7080):83–86.

[4] Wearing HJ, Rohani P, Keeling MJ. Appropriate models for the management of infectious diseases. PLoS medicine. 2005;2(7):e174.

[5] Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. Nature. 2008;451(7181):990–993.

[6] Cauchemez S, Boëlle PY, Donnelly C, Ferguson N, Thomas G, Leung G, et al. Real-time estimates in early detection of SARS. Emerging Infect Dis. 2006;12:110–113.

[7] Jewell C, Kypraios T, Neal P, Roberts G. Bayesian Analysis for Emerging Infectious Diseases. Bayes Anal. 2009;4(3):465–496.

[8] Jewell CP, Keeling MJ, Roberts GO. Predicting undetected infections during the 2007 foot-and-mouth disease outbreak. J R Soc Interface. 2009 Dec;6(41):1145–1151.

[9] Porco T. A mathematical model of the ecology of Lyme disease. IMA J Math Appl Med Biol. 1999 Sep;16(3):261–296.

[10] Mandal S, Sarkar R, Sinha S. Mathematical models of malaria – a review. Malar J. 2011;10:202–220.

[11] Sutton A, Karagenc T, Bakirci S, Sarali H, Pekel G, Medley G. Modelling the transmission dynamics of Theileria annulata: model structure and validation for the Turkish context. Parasitology. 2012 Apr;139(4):441–453.

[12] Lourenço J, Recker M. The 2012 Madeira Dengue Outbreak: Epidemiological Determinants and Future Epidemic Potential. PLoS Negl Trop Dis. 2014;8(8):e3083.

[13] Braks M, van der Giessen J, Kretzschmar M, van Pelt W, Scholte EJ, Reusken C, et al. Towards an integrated approach in surveillance of vector-borne diseases in Europe. Parasit Vectors. 2011;4:192.

[14] Purse BV, Mellor PS, Rogers DJ, Samuel AR, Mertens PPC, Baylis M. Climate change and the recent emergence of bluetongue in Europe. Nat Rev Microbiol. 2005 Feb;3(2):171–181.

[15] Bouzid M, Colón-González F, Lung T, Lake I, Hunter P. Climate change and the emergence of vector-borne diseases in Europe: case study of dengue fever. BMC Public Health. 2014;14(1):781–792.

[16] Ostfeld RS, Brunner JL. Climate change and Ixodes tick-borne diseases of humans. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2015;370(1665):20140051.

[17] Graesboll K, Sumner T, Enoe C, Christiansen L, Gubbins S. A Comparison of Dynamics in Two Models for the Spread of a Vector-Borne Disease. Transbound Emerg Dis. 2014 Jul;.

[18] Parry M, Gibson GJ, Parnell S, Gottwald TR, Irey MS, Gast TC, et al. Bayesian inference for an emerging arboreal epidemic in the presence of control. Proc Natl Acad Sci U S A. 2014 Apr;111(17):6258–6262.

[19] Luo Y, Ogle K, Tucker C, Fei S, Gao C, LaDeau S, et al. Ecological forecasting and data assimilation in a data-rich era. Ecol Appl. 2011;21(5):1429–1442.

[20] Presanis A, Gill O, Chadborn T, Hill C, Hope V, Logan L, et al. Insights into the rise in HIV infections in England and Wales, 2001 to 2008: a Bayesian synthesis of prevalence evidence. AIDS. 2010;24(18):2849–2858.

[21] Jewell CP, Roberts GO. Enhancing Bayesian risk prediction for epidemics using contact tracing. Biostatistics. 2012 Sep;13(4):567–579.

[22] Neal P, Roberts G. Statistical inference and model selection for the 1861 Hagelloch measles epidemic. Biostatistics. 2004;5(2):249–261.

[23] Dawid AP. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. Journal of the Royal Statistical Society Series A (General). 1984;p. 278–292.

[24] Lawrence K, McFadden A, D P. *Theileria orientalis* (Ikeda) associated bovine anaemia: The epidemic to date. Vetscript. 2013;November:11–13.

[25] McFadden AMJ, Rawdon TG, Meyer J, Makin J, Morley CM, Clough RR, et al. An outbreak of haemolytic anaemia associated with infection of Theileria orientalis in naive cattle. N Z Vet J. 2011 Mar;59(2):79–85.

[26] Islam MK, Jabbar A, Campbell BE, Cantacessi C, Gasser RB. Bovine theileriosis–an emerging problem in south-eastern Australia? Infect Genet Evol. 2011 Dec;11(8):2095–2097.

[27] James M, Saunders B, Guy L, Brookbanks E, Charleston W, Uilenberg G. *Theileria orientalis*, a blood parasite of cattle. First report in New Zealand. NZ Vet J. 1984;32(9):154–156.

[28] Heath A. The temperature and humidity preferences of *Haemaphysalis longicornis*, *Ixodes holocyclus*, and *Rhipicephalus sanguineus* (Ixodidae): studies on engorged larvae. Int J Parasitol. 1981;11(2):169–175.

[29] Heath A. Distribution, seaseasonal and relative abundance of *Stomoxys calcitrans* (stablefly) (Diptera: Muscidae) in New Zealand. NZ Vet J. 2002;50(3):93–98.

[30] McFadden A, Pulford D, Lawrence K, Frazer J, van Andel M, Donald J, et al. Epidemiology of *Theileria orientalis* in cattle in New Zealand. In: Proceedings of the Dairy Cattle Veterinarians of the NZVJ Annual Conference; 2013. p. 207–217.

[31] Stafford KC. Survival of immature Ixodes scapularis (Acari: Ixodidae) at different relative humidities. Journal of medical entomology. 1994;31(2):310–314.

[32] Ogden N, Lindsay L, Beauchamp G, Charron D, Maarouf A, O'Callaghan C, et al. Investigation of relationships between temperature and developmental rates of tick Ixodes scapularis (Acari: Ixodidae) in the laboratory and field. Journal of medical entomology. 2004;41(4):622–633.

[33] Knülle W, Rudoph D. Humidity relationships and water balance of ticks. In: Obenchain F, Galun R, editors. Physiology of Ticks. Pergamon Press, Oxford.; 1982. p. 43–70.

[34] Perera PK, Gasser RB, Pulford DJ, Stevenson MA, Firestone SM, McFadden AM, et al. Comparison of the performance of three PCR assays for the detection and differentiation of Theileria orientalis genotypes. Parasites & Vectors. 2015;8(1):1–11.

[35] Pulford D, Gias E, Bueno I, McFadden A. Developing high throughput diagnostic molecular tests for typing *Theileria orientalis* from bovine blood samples. N Z Vet J. 2015;In Press.

[36] Yu Y, Meng XL. To center or not to center: That is not the question – an Ancillarity-Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency. Journal of Computational and Graphical Statistics. 2011;20(3):531–570.

[37] Heath A. Ticks (*Haemaphysalis longicornis*). In: Charleston W, editor. Ectoparasites of sheep in New Zealand and their control. Sheep and Beef Cattle Society of the New Zealand Veterinary Association; 1985. p. 23–29.

[38] Diggle P, Moraga P, Rowlingson B, Taylor B. Spatial and Spatio-temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm. Statist Sci. 2013;28(4):452–563.

[39] Shirashi S, Yoshino K, Uchida T. Studies on seasonal fluctuations of population and overwintering in the cattle tick, *Haemaphysalis longicornis*. J Fac Agric Kyushu Univ. 1989;34:43–52.

[40] Srinivasa Rao ASR, Chen MH, Pham BZ, Tricco AC, Gilca V, Duval B, et al. Cohort effects in dynamic models and their impact on vaccination programmes: an example from hepatitis A. BMC Infect Dis. 2006;6:174.

[41] Ong JBS, Chen MIC, Cook AR, Lee HC, Lee VJ, Lin RTP, et al. Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. PLoS One. 2010;5(4):e10036.

[42] Chis Ster I, Singh B, Ferguson N. Epidemiological inference for partially observed epidemics: the example of the 2001 foot and mouth epidemic in Great Britain. Epidemics. 2009;1(1):21–34.

[43] Cowled B, Ward M, Hamilton S, Garner G. The equine influenza epidemic in Australia: spatial and temporal descriptive analyses of a large propagating epidemic. Prev Vet Med. 2009 Nov;92(1-2):60–70.

[44] Minh P, Stevenson M, Jewell C, French N, Schauer B. Spatio-temporal analyses of highly pathogenic avian influenza H5N1 outbreaks in the Mekong River Delta, Vietnam, 2009. Spatial and Spatiotemporal Epidemiology. 2011;2(1):49–57.

[45] Tenquist J, Charleston W. A revision of the annotated checklist of ectoparasites of terrestrial mammals in New Zealand. J Roy Soc NZ. 2001;31(3):481–542.

Figure 1: Summary of the New Zealand *T. orientalis* Ikeda epidemic as of 1st August 2014. Top row: Spatial distribution of infected herds, with the log cumulative number of case detections throughout the epidemic. Bottom row: Median prior and posterior tick occurrence probabilities for NZ TLAs.

Figure 2: 6-month in-sample prediction obtained by simulating forward from 1st February 2014, using parameter estimates generated as of 1st 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.

Figure 3: Prior and posterior distributions at the four analysis times for a) the probability of a NAIT movement transmitting an infection β_2 (top left); b) the susceptibility of a dairy farm compared to non-dairy ζ (top left); 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 α_1, α_2, and α_3, corresponding to epochs 2, 3, and 4 respectively; d) the median environmental transmission with distance $\beta_1 K(i,j;\delta)$ (bottom right).

Figure 4: 6-month out-of-sample prediction of cumulative case detections as of 1st November 2013 (top left), 1st February 2014 (top right), 1st May 2014 (bottom left) and 1st August 2014 (bottom right). Red smudge: the probability distribution over the predicted epidemic trajectory with 95% quantiles marked by the dotted lines. Blue curve: The observed cumulative cases up to 1st August 2014.

Figure 5: 6-month predicted individual infection probability, conditional on the observed epidemic as of 1st November 2013, 1st February 2014, 1st May 2014, and 1st August 2014.

Appendix A Preparation of NAIT movement network

Demographic data for the New Zealand cattle herd population currently exist in a number of component databases. These require joining of records prior to analysis. For our study, we have used Farms OnLine, NAIT, and Agribase.

Farms OnLine (FOL) is the government-owned agricultural property database. Each record contains a unique identifier, owner contact information, geographic land parcel polygons (from which centroids are calculated), and presence/absence information for different animal types.

The National Animal Identification and Tracking (NAIT) database records individual cattle and deer movements within New Zealand. This is a mandatory system where each record represents an animal moving on a given date, between source and destination "Persons In Charge of Animals" (PICAs). Furthermore, NAIT links each PICAs to FOL identifiers providing the opportunity to georeference animal movements for the purposes of joint network-spatial modelling.

To begin, we represent the NAIT data as a dynamic network A of cattle movements where nodes represent PICAs. Let $N_{sr}(t)$ be a counting process describing the number of cattle moved between nodes s and r up to time t. We estimate the mean directed edge frequency

$$\hat{a}_{sr} = \frac{N_{sr}(t_1) - N_{sr}(t_0)}{t_1 - t_0}$$

with $t_1 - t_0$ the time interval represented by our NAIT extract.

To georeference this network, we map PICAs to FOL entities to obtain a new network C with nodes represented by FOL entities. However, a many-to-many relationship exists between PICAs and FOL entities. Movements between PICAs linked to more than one FOL entity were therefore assumed to have occurred between a single pair of source/destination FOL locations with equal probability. The movement was therefore "distributed" between all possible pairs with equal weighting. Furthermore, for consistency with the SID model we explicitly disallow loops. Thus the edge frequency for C between FOL locations i and j is estimated as

$$\hat{c}_{ij} = \begin{cases} \sum_{s\sim i} \sum_{r\sim j} \frac{\hat{a}_{sr}}{m_s m_r} & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

where $s \sim i$ denotes PICA s being linked to FOL entity i and $m_s = \sum_i 1[s \sim i]$ denotes the number of FOL entities associated with s. Similarly for r and j.

We note that no attempt to reflect uncertainty in the pair weighting is made in our study, though in principle this could be done either through a Dirichlet model, or via trans-dimensional MCMC as in [1].

Appendix A.1 Seasonal variation modelling

For rapid likelihood-based inference and prediction, the seasonal function $s(t; \boldsymbol{\alpha}, \nu)$ in Equation 3 (main paper) requires the following characteristics:

1. As part of the parameter estimation procedure, the definite integral of the seasonal variation component $\int_{t_0}^{t} s(t; \boldsymbol{\alpha}, \nu)\, dt$ is computed many thousands of times, for different parameter and t values. As such it is important that the integral is analytically tractable to allow direct evaluation rather than costly quadrature computations.
2. The relative heights and depths of the spring/autumn and summer/winter peaks and troughs, respectively, should be allowed to differ.

3. It is desirable to set at least one of the seasonal peaks to a fixed value. This aids statistical identifiability between $\boldsymbol{\alpha}$ and the baseline transmission parameters β_1 and β_2 (Equation 2, main paper) because the majority of the statistical information is concentrated into periods of high case incidence, corresponding to seasonal peaks.

Whilst at first a trigonometric function appears suitable for capturing seasonality, requirements 2 and 3 inevitably mean that 1 is not satisfied. To this end, we initially modelled seasonal variation $s(t; \boldsymbol{\alpha}, \nu)$ using the following piecewise cubic formulation.

$$s(t; \boldsymbol{\alpha}, \nu) = \begin{cases} f_{\text{spline}}(t^\star; 0, 0.25, 1, \alpha_1) & \text{if } 0 \leq t^\star < 0.25 \\ f_{\text{spline}}(t^\star; 0.25, 0.5, \alpha_1, \alpha_2) & \text{if } 0.25 \leq t^\star < 0.5 \\ f_{\text{spline}}(t^\star; 0.5, 0.75, \alpha_2, \alpha_3) & \text{if } 0.5 \leq t^\star < 0.75 \\ f_{\text{spline}}(t^\star; 0.75, 1, \alpha_3, 1) & \text{if } 0.75 \leq t^\star < 1 \end{cases},$$

where $t^\star = t + \nu - \lfloor t + \nu \rfloor$, with t in years, and where $f_{\text{spline}}(t; t_0, t_1, s_0, s_1)$ is the unique cubic passing through (t_0, s_0) and (t_1, s_1) with $f'_{\text{spline}}(t_0) = f'_{\text{spline}}(t_1) = 0$. The parameter $0 \leq \nu \leq 0.5$ allows for phase adjustment of the seasonality. Note that to avoid identifiability issues, the spring peak ($t^\star = 0$) is fixed at 1 as the overall scale is set by the $\boldsymbol{\beta}$ parameters (Equation 2, main paper).

This function is, however, unable to specify neither prolonged periods of tick activity or inactivity nor rapid threshold-like transitions between regimes, and was subsequently rejected in favour of the square wave function, (Equation 4, main paper) which performed significantly better on the data. See the discussion for more details. Figure 6 shows example square-wave and cubic spline seasonality functions.

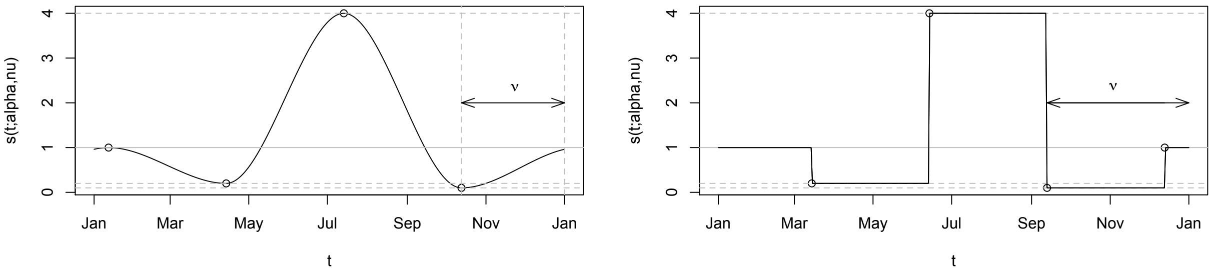

Figure 6: Sample spline (left) and square wave (right) seasonality functions.

Appendix B Statistical likelihood function

We assume the epidemic and BVD sampling processes to be independent conditional on the underlying regional tick occurrence vector $\boldsymbol{p}$ as defined in the main text. Our likelihood function falls into two parts: $L_E(\boldsymbol{\theta}, \boldsymbol{I}|\boldsymbol{D})$ is the likelihood for the model parameters (including $\boldsymbol{p}$) and infection times conditional on the detection times, and $L_S(\boldsymbol{p}|\boldsymbol{X})$ is the likelihood for $\boldsymbol{p}$ given the sampling data $\boldsymbol{X}$.

To model the SID epidemic we use a continuous time inhomogeneous Poisson process setup. The derivation of this likelihood function has been previously described (see for example [2]).

Let T_{obs} be the analysis time of the epidemic, a time at which the epidemic may still be in progress. At this time, we will have observed the times and identities of case detections represented by the m_I dimensional vector $\boldsymbol{D}$. The model parameters $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \beta_1, \beta_2, \nu, \zeta, \boldsymbol{p}, b\}$ are unknown. Furthermore, we do not observe the M_I dimensional vector of infection times $\boldsymbol{I}$, which may include infection times corresponding to elements of $\boldsymbol{D}$ as well as those corresponding to undetected infections.

Since $\boldsymbol{I}$ is not observed, we cannot write an explicit likelihood function for $\boldsymbol{\theta}$. However, recalling the Gamma distributed infection to detection time (main text), we may write a conditional likelihood

$$\begin{aligned} L_E(\boldsymbol{\theta}|\boldsymbol{I}, \boldsymbol{D}) &\propto \prod_{j=1} M_I \left[\lambda_j(I_j^-)\right] \exp\left[\int_{I_\kappa}^{T_{obs}} \sum_{i \in \mathcal{P}} \lambda_j(t) dt\right] \\ &\times \prod_{j=1}^{m_I} f_D(D_j - I_j) \\ &\times \prod_{j=m_I+1}^{M_I} (1 - F_D(D_j - I_j)) \end{aligned}$$

where $\mathcal{I}$ is the set of herds that have been infected up to the analysis time T_{obs}, I_j and D_j are the infection and detection times of the jth herd, and $\mathcal{P}$ is the set of individual herds comprising the entire population. $\lambda_j(I_j^-)$ represents the infectious pressure on j immediately before its infection, as described in the main text. MCMC data augmentation methodology is then used to integrate over $\boldsymbol{I}$ in terms of both infection times and presence of undetected infections [3].

To model the BVD sampling process, we use independent Binomial distributions for each TLA region $k = 1, \ldots, 72$. In many discrete-space statistical models, spatial dependency is explicitly included. However, since the TLA regions are in general large, and the epidemic process explicitly accounts for spatial dependency, we choose not to incorporate this level of complexity.

Given sampling data $\boldsymbol{X} = \{\boldsymbol{n}, \boldsymbol{x}\}$ where $\boldsymbol{n}$ is the number of samples collected and $\boldsymbol{x}$ is the number of positive samples for *Theileria orientalis* species, we have

$$L_S(\boldsymbol{p}|\boldsymbol{X}) \propto \prod_{k=1}^{72} p_k^{x_k}(1-p)^{n_k - x_k}$$

The joint (conditional) likelihood for the epidemic and BVD sampling process is therefore

$$L(\boldsymbol{\theta}|\boldsymbol{I}, \boldsymbol{D}) \propto L_E(\boldsymbol{\theta}|\boldsymbol{I}, \boldsymbol{D}) L_S(\boldsymbol{p}|\boldsymbol{X})$$

In terms of inference, since $\boldsymbol{p}$ appears in both likelihood components (through $\lambda_j(t)$ for the epidemic), we allow its value to be determined not only by the BVD sampling, but also by the epidemic process. $\boldsymbol{p}$ therefore represents a measure of tick-associated transmissibility in each TLA region.

Appendix C MCMC Convergence diagnostics

MCMC convergence was assessed by running 4 parallel chains starting at difference values for the parameters. Superimposed timeseries plots were visually inspected. The marginal traceplots for four parameters from the August analysis are shown in Figure 7. These plots show the overall mixing quality of the algorithm to be satisfactory, though not perfect. Importantly, the plots show that the assumption of the chains converging to the same limiting distribution is acceptable, confirmed by a Gelman-Rubin statistic of 1.001 [4].

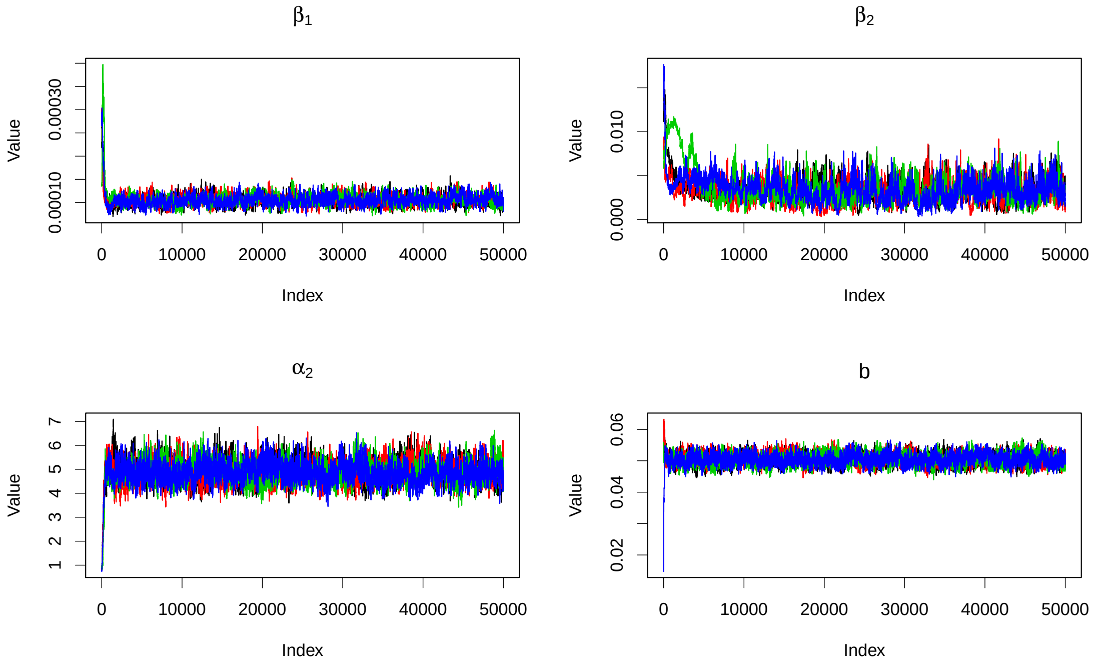

Figure 7: Superimposed traceplots for 4 parallel MCMC runs on the August dataset. Graphs for β_1, β_2, α_2, and b are shown.

Appendix D Priors

Functional forms for the prior distributions used in our analysis were chosen to match the support of each parameter. Typically, Gamma distribution are chosen for parameters describing rates (transmission rate, detection rate, etc), and Beta distributions are used for probability parameters (i.e. tick occurrence). An exception to this is for parameters α_1 and α_3 which were given Beta distributions to ensure their definition as "troughs" in the seasonal function. Prior distributions for $\boldsymbol{\alpha}$, β_1, β_2, ζ, δ, and b are shown in Table 1. For $\boldsymbol{p}$, TLA regions are aggregated into "high", "medium", and "low" regions as shown in Figure 1 of the main text. These are shown in Table 2.

Table 1: Prior distributions for parameters of the epidemiological model

Parameter	**Distribution**
α_1	Beta(1,50)
α_2	Gamma(32, 8)
α_3	Beta(1,50)
β_1	Gamma(4, 16000)
β_2	Beta(2, 2)
ζ	Gamma(5, 2)
δ	Gamma(1, 1)
b	Gamma(2.5, 50)

Table 2: (Hyper)Parameters of the Beta(a, b) distributions, together with the median and 2.5% and 97.5% quantiles, used to encode information about spatial vector risk for the three prior tick risk levels.

Prior vector risk	a	b	Median	2.5% quantile	97.5% quantile
"high"	51	1	0.97	0.93	1.0
"medium"	20	20	0.50	0.35	0.65
"low"	1	50	0.014	0.00051	0.071

Appendix E Now-casting results

This section contains further results on the current state of the epidemic, taken from the inference algorithm. The maps in Figure 8 provide a spatial representation of occult probability – the probability that each presumed-susceptible farm is in fact an undetected infection. Herds with a high probability of infection are confined to the high herd density region of northern Waikato and Auckland, reflecting the predominantly spatial nature of this epidemic. These maps may be used to inform targeted surveillance by ranking farms in order of the most likely to be infected [1].

Figure 9 shows the posterior distributions for infection to detection time, a convolution of the infection to detection time model and the marginal posterior for b. The results show a stability in the infectious period throughout the epidemic, with a reduction in posterior variance compared to the prior. We note that for our data, the infectious period is essentially dictated by the phase of seasonal function.

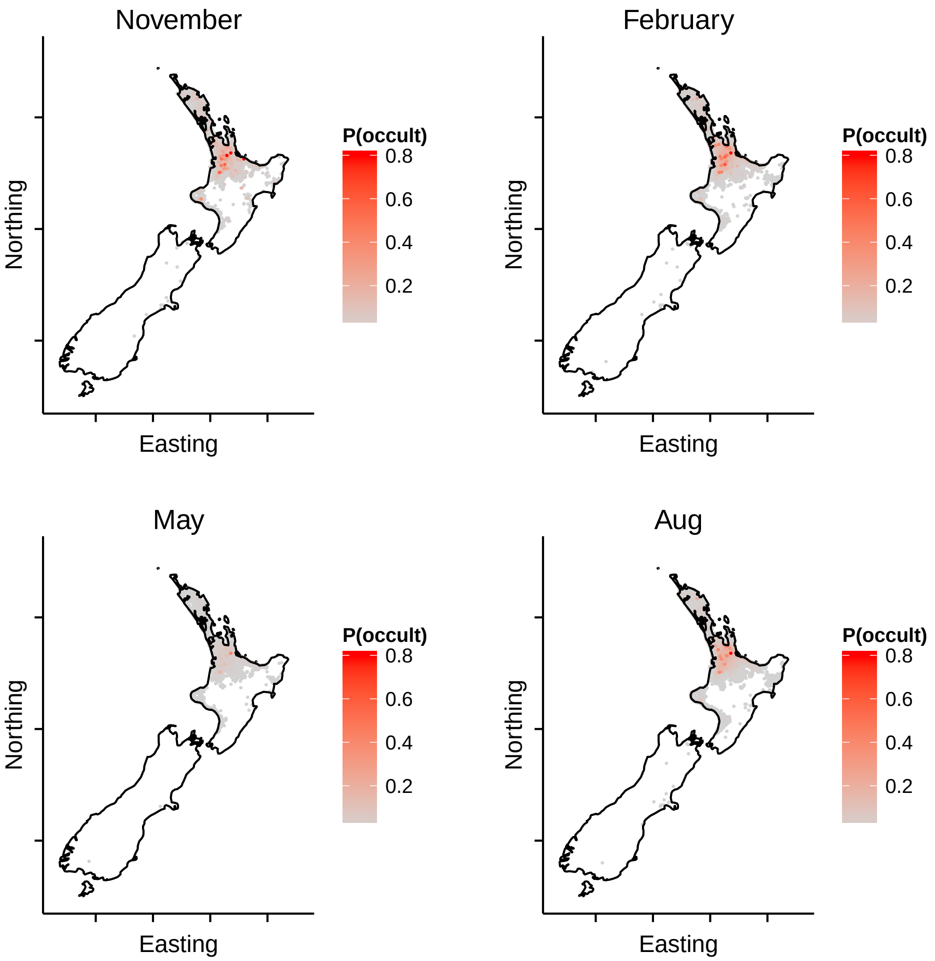

Figure 8: Spatial distribution of occult probabilities at each analysis timepoint.

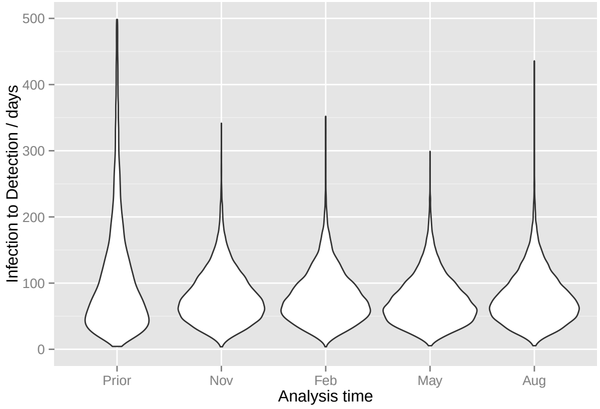

Figure 9: Prior and posterior infectious period distributions for the 4 analysis times.

Appendix F Effect of joint modelling on posterior tick occurrence

Figure 10 compares the posterior tick occurrence surface using just BVD sampling data, compared to that using the the joint sampling/epidemic model. The overall pattern of tick activity is similar between the two models. However, differences between the two maps in the high case density region of the North Island demonstrates how the innovation in joint modelling is required beyond marginal modelling of tick activity based on the sampling data alone.

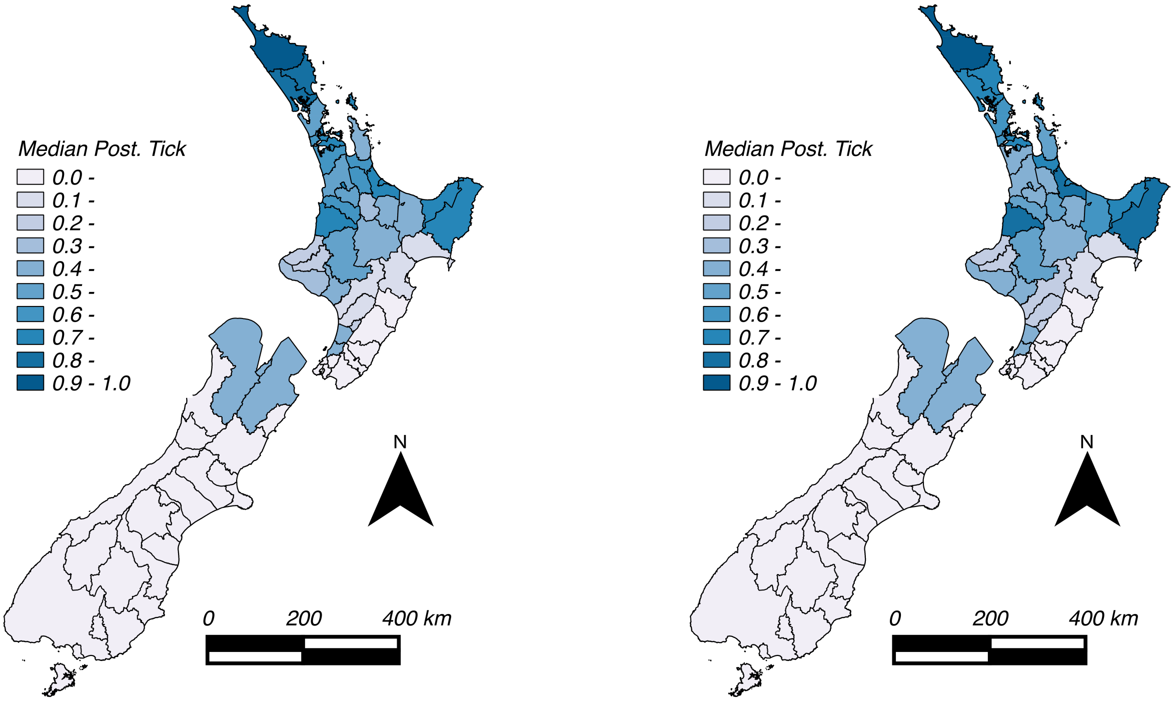

Figure 10: A comparison of posterior TLA region tick occurrence for BVD sample data only (left) and joint sample-epidemic data (right).

References

[1] Jewell CP, Roberts GO (2012) Enhancing bayesian risk prediction for epidemics using contact tracing. *Biostatistics* 13(4):567–579.

[2] Andersson H, Britton T (2000) Stochastic Epidemic Models and Their Statistical Analysis. *Lecture Notes in Statistics* (Springer, Heidelberg).

[3] Jewell CP, Kypraios T, Neal P, Roberts GO (2009) Bayesian analysis for emerging infectious diseases. *Bayes. Anal.* 4(3):465–496.

[4] Brooks SP, Gelman A (1998) Alternative Methods for Monitoring Convergence of Iterative Simulations. *J. Comput. Graph. Stat.* 7:434-455.