Supporting Information

Parry et al. 10.1073/pnas.1310997111

SI Text

Data. The spread of Huanglongbing (HLB) during the four sweeps is shown in Fig. S1. The epidemic first becomes apparent in young trees in the east and then spreads both westward and to older trees. Removal of symptomatic trees usually occurred 1–3 mo after detection.

Modeling Exposure to Infection. Because the regions of interest are of suborchard scale and the timescale of the observations allows the vector to move a considerable distance, for the purposes of the model, we assume that the total psyllid density is roughly constant in space. Letting $\tilde{\rho}$ denote the total psyllid density, κ denote, the fraction that are infected, and Λ denote the intrinsic infection rate per psyllid, the rate at which a susceptible tree becomes exposed will be $\Lambda\kappa\tilde{\rho}$. It is convenient for modeling purposes, however, to distinguish between vectors coming from outside and inside the observed region. Let $\tilde{\rho}_0$ denote the density of external psyllids (assumed constant over the region of interest) and $\tilde{\rho}_{ij}$ denote the density of psyllids at tree i arriving from tree j. The density fraction $\tilde{\rho}_{ij}$ can be further modeled via a dispersal kernel, namely, $\tilde{\rho}_{ij} = RK(\mathbf{x}_i, \mathbf{x}_j)$, where R is a scaling constant and $\mathbf{x}_i$ denotes the position of tree i, etc. The dispersal kernels we consider are isotropic, $K(\mathbf{x}_i, \mathbf{x}_j) = k(r_{ij}/\alpha)$, where r_{ij} is the Euclidean distance between tree i and tree j and α is a length scale to be determined.

To account for psyllid control in Southern Gardens, this was carried out in the region under observation and surrounding areas, it is necessary to let $\tilde{\rho} = \tilde{\rho}(t)$. We assume that the same function of time controls the time dependence of both $\tilde{\rho}_0(t)$ and $\tilde{\rho}_{ij}(t)$, namely $\rho(t) := \tilde{\rho}_0(t)/\tilde{\rho}_0(t_0) = \tilde{\rho}_{ij}(t)/\tilde{\rho}_{ij}(t_0)$, and call it the relative psyllid density. We estimate the relative psyllid density as a piecewise linear function

$$\rho(t) = \begin{cases} 1, & t \le 1 \\ 2 - t, & 1 < t \le 2 \\ 0, & t > 2 \end{cases}, \qquad \textbf{[S1]}$$

where t is measured in years and $t = 0$ corresponds to the start of 2005. Note that $t_0 < 1$ is implied.

Defining $\epsilon = \Lambda\kappa\tilde{\rho}_0(t_0)$ and $\beta = \Lambda\kappa R$, the instantaneous rate of infection at time t for susceptible tree i can be written as

$$\dot{\Phi}_i(t) = \rho(t)\left[\varepsilon + \beta \sum_j k\left(\frac{r_{ij}}{\alpha}\right) \mathbb{1}\left(t_j^I < t \le t_j^R\right)\right], \qquad \textbf{[S2]}$$

where $\mathbb{1}$ is the indicator function that returns 1 when its argument is true and 0 otherwise. We identify ε as the primary or external rate of infection and β as the secondary rate of infection; both are model parameters to be estimated.

Although the form of Eq. **S2** is standard, the physical and biological reasoning underpinning the model of the exposure process leads to the expressions for ε and β given above. As a direct consequence, the ratio of primary to secondary transmission rates is the same constant for all spatial regions of interest

$$\frac{\epsilon}{\beta} = \frac{\tilde{\rho}_0(t_0)}{R} = \text{constant}. \qquad \textbf{[S3]}$$

Note that had we introduced the normalized secondary rate of infection per tree, $\beta' = \beta \cdot Z(\alpha)$, where $Z(\alpha) = (1/N)\sum_{i \ne j} k(r_{ij}/\alpha)$ and N is the number of trees, this result would not have been apparent: the ratio ϵ/β' depends on α and the details of the observed region.

Latent Period Parameters. The parameters of the cyclic, exponential, and cyclic Weibull models are estimated from in-orchard and in-nursery observations. Expert opinion, based on in-orchard observations, for the upper bound on the combined latent and cryptic periods was interpreted as the time for 95% of trees to move from exposed to symptomatic. (Our results do not depend sensitively on the assumption of 95%.) In-nursery observations of the cryptic period then allow us to estimate the $T_{0.95}$, the time for 95% of trees to move from exposed to infectious. By itself, $T_{0.95}$ is not sufficient to fix the parameters of the cyclic, exponential, and cyclic Weibull models, because these models describe a latent period that depends on t^E. Fortunately, the dependence on t^E becomes small for latent periods longer than the seasonal period. The in-orchard data are given in Table S1.

On other hand, the gamma model is a two-parameter model. Its parameters are fixed by $T_{0.95}$ and by expert opinion that the lower bound on the combined latent and cryptic periods is at least 6 mo and usually more than 1 y. The lower bound was interpreted as the time for 5% of trees to move from exposed to symptomatic. Recently, however, it has been realized that the latent period can be as a little as 1 mo. This shortness of the latent period is in line with the conclusions of our study, being entirely consistent with the cyclic, exponential, and cyclic Weibull models.

It is also important to point out that seasonal effects were introduced to account for the increase in symptomatic counts during the second and third sweeps and the reduced counts in the fourth sweep. Having seasonality in the latent period was forced on us by a combination of the assumption of a compartmental model for the epidemic, the early appearance of symptomatic trees in Southern Gardens, the introduction of efficient psyllid control measures, and the relatively long incubation period for HLB.

Cryptic Period Parameters. The parameters are fixed by in-nursery observations showing that 5% (95%) of trees become symptomatic within 2 (3) mo of becoming infectious.

Note that both the latent period and cryptic period parameters could also have been estimated as part of the Bayesian inference. Because the in-orchard and in-nursery observations lead to rather strong priors on these parameters, it was the felt the extra computational complexity would have yielded limited additional information.

Parameter Estimation. We adopt a Bayesian approach to parameter estimation. We use Markov chain Monte Carlo (MCMC) techniques, with uninformative exponential priors, to obtain, after a burn-in period, a joint posterior density for the parameters, conveniently represented as a vector $(\alpha, \beta, \epsilon, t_0)$ in the parameter space Θ_0.

Suppose there are initially N susceptible trees and that T is the final observation time. Let $\mathcal{S}$ be the index set of trees never exposed, $\mathcal{E}$ be those exposed but never infectious, $\mathcal{I}$ be those infectious but never symptomatic, and $\mathcal{D}$ be those that were symptomatic, up to time T. It follows that $\mathcal{S}$, $\mathcal{E}$, $\mathcal{I}$, and $\mathcal{D}$ are disjoint sets and that $\mathcal{S} \cup \mathcal{E} \cup \mathcal{I} \cup \mathcal{D} = \{1, \ldots, N\}$. Then, taking the removal times as known, the theoretical joint probability density for the epidemic times is

$$p\left(\{t_i^E, t_i^I, t_i^D\}\middle|\alpha,\beta,\epsilon,t_0\right) = \prod_{i\in\mathcal{S}} \mathrm{e}^{-\Phi_i(T)} \prod_{i\in\mathcal{E}} \dot{\Phi}_i\left(t_i^E\right)\mathrm{e}^{-\Phi_i\left(t_i^E\right)}\overline{F}\left(T|t_i^E\right)$$
$$\times \prod_{i\in\mathcal{I}} \dot{\Phi}_i\left(t_i^E\right)\mathrm{e}^{-\Phi_i\left(t_i^E\right)} f\left(t_i^I|t_i^E\right)\overline{G}\left(T|t_i^I\right)$$
$$\times \prod_{i\in\mathcal{D}} \dot{\Phi}_i\left(t_i^E\right)\mathrm{e}^{-\Phi_i\left(t_i^E\right)} f\left(t_i^I|t_i^E\right) g\left(t_i^D|t_i^I\right), \qquad \textbf{[S4]}$$

where $f(t^I|t^E)$ is the probability density function for the latent period, $\overline{F}(T|t^E) := \int_T^\infty \mathrm{d}t\, f(t|t^E)$ is its (right) tail distribution, $g(t^D|t^I)$ is the probability density function for the cryptic period, and $\overline{G}(T|t^I) := \int_T^\infty \mathrm{d}t\, g(t|t^I)$ is its tail distribution.

In practice, however, the times t_i^E and t_i^I are unobserved, and we observe only that $t_i^D \in \Delta_i$, where Δ_i is some time interval. Indeed, we cannot even specify the index sets $\mathcal{S}$, $\mathcal{E}$, or $\mathcal{I}$. Ordinarily, we would integrate the joint density with respect to the latent variables and over the censoring intervals to obtain the probability $\mathbb{P}(\{t_i^D \in \Delta_i : i \in \mathcal{D}\}|\alpha,\beta,\epsilon,t_0)$, but such an approach is not tractable here. The reason is that $\Phi_i(t)$ depends on all infectious times $t_j^I < t$, making it impossible to find a closed form expression for the integral of Eq. **S4** over the unobserved times. To solve this problem, we use data augmentation and reversible-jump MCMC methods.

Reversible-Jump MCMC with Data Augmentation. Let δ denote a partition of the set of trees that were never symptomatic into sets of trees that were not exposed, exposed but not infectious, and infectious: $\mathcal{S}^\delta \cup \mathcal{E}^\delta \cup \mathcal{I}^\delta = \mathcal{D}'$. MCMC with data augmentation (1–3) extends the parameter space from Θ_0 to Θ^δ, where Θ^δ contains Θ_0 and all times consistent with the data and the partition δ. Given a partition δ, MCMC will generate an empirical joint posterior density on Θ^δ. The marginal density on Θ_0 is then the desired posterior.

The reversible-jump technique (4) enables MCMC to explore Θ^δ for different δ, a requirement that is clearly necessary because the actual partition is not known. Reversible-jump MCMC generates an empirical joint posterior density on the parameter space $\Theta := \bigcup_\delta\{\delta\}\times\Theta^\delta$. Because $\Theta_0 \subset \Theta^\delta$ for all δ, we marginalize over all times and partitions to obtain the posterior for the parameter vector.

Age Dependence in the Rates of Infection. The age dependence in the rates of infection is show in Fig. S2.

Model Checking and Model Comparison. For subregion 13a, we compare eight different models, a model being specified by a choice of dispersal kernel and latent period model: (*i*) exponential kernel plus cyclic model with yearly oscillations; (*ii*) exponential kernel plus exponential latent period; (*iii*) Cauchy kernel (r^{-2} power law kernel) plus yearly cyclic model; (*iv*) exponential kernel plus gamma model; (*v*) r^{-4} power law kernel plus yearly cyclic model; (*vi*) exponential kernel plus cyclic Weibull model; (*vii*) r^{-8} power law kernel plus yearly cyclic model; and (*viii*) exponential kernel plus twice-yearly cyclic model.

The labeling has been chosen so that in Figs. S4 and S5, the first column shows the results of different dispersal kernels and the second column shows the results of different latent period models. We fit each model using reversible-jump MCMC with data augmentation. From the resulting joint posterior for Θ_0, we randomly drew 100 parameter sets and simulated epidemics using the Selke algorithm. We then compared the temporal and spatial structure of simulated outcomes with those of the actual outcome. Specifically, we considered the counts of symptomatic trees in each of the four sweeps, and the two-point spatial correlation of all symptomatic trees observed up to and including the final sweep.

Dispersal kernel length scales. The estimated length scale α varies with choice of dispersal kernel and latent period model. Table S2 gives the explicit form of the kernel $k(r/\alpha)$ in each of the eight models above and the posterior mean and 95% credible region for α measured in units of meters. Except for model c, the marginal posterior densities for α were roughly bell-shaped. For the Cauchy kernel in model c, the marginal posterior peaks at $\alpha = 0$. The interpretation when α is vanishingly small is that external sources of infection are entirely responsible for driving the epidemic.

Counts of symptomatic trees. The counts of symptomatic trees are indicated in Fig. S3. The red line is the actual observed count; the frequency histogram of counts from the 100 simulations is in blue. Only in models a and c are all actual counts well within the distribution of simulated counts. Model d, the only model that favors a long latent period, is completely ruled out.

Two-point correlation function. The two-point correlation function we use is a modified Moran's I statistic for presence-absence data. Letting $s_i = 1$, if tree i is found to be symptomatic before or at the final observation time and $s_i = 0$ otherwise, we define

$$\xi_2(r_1,r_2) = \frac{\left[\sum_{ij} w_{ij}(r_1,r_2)s_is_j\right] - \bar{s}^2}{\bar{s}(1-\bar{s})\sum_{ij} w_{ij}(r_1,r_2)}, \qquad \textbf{(5)}$$

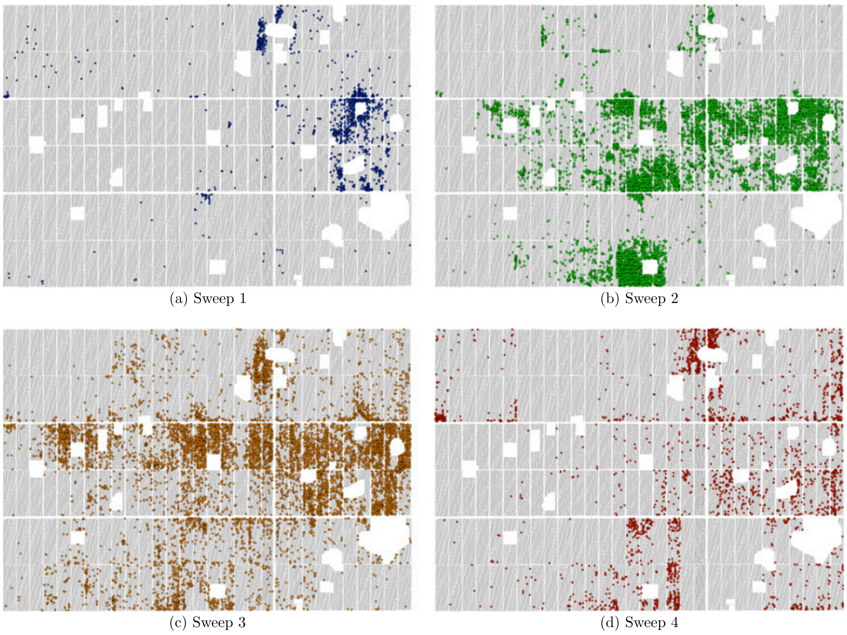
where $\bar{s} = (1/N)\sum_i s_i$ and $w_{ij}(r_1,r_2) = \mathbb{1}(r_1 \le |\mathbf{x}_i - \mathbf{x}_j| < r_2)$ is a ring weighting. Note that in the limit $r \to 0$, we have $\xi_2(0,r) = 1$.

The two-point correlation functions are shown in Fig. S4. We chose a sequence of ring radii, $r_1, r_2, \ldots$, and plotted $\xi(r_i) := \xi_2(r_i, r_{i+1})$. The red line is the actual observed two-point correlation; the two-point correlations from the 100 simulations for each model are in green. Each model gives correlation functions that are consistent with the vanishing of the observed correlation function for r > 80 m. Conversely, only model a accommodates the observed correlation function for r > 80 m. Models e and g fit reasonably well except in the range of 30–40 m. Once again, model d appears to be completely ruled out.

Model predictions. In Fig. S5, the cumulative counts of symptomatic trees are shown for 100 simulations of each model that have been run more than 3 y past the final observation time. Note that simulations are run from the randomly drawn initial times $t_0 \in \Theta_0$ and incorporate the removal of symptomatic trees. Although the simulations are not constrained by the actual detections at the four observational sweeps, they are in good agreement with the data, as evidenced by the red lines that meet at the total cumulative detection count at the final observation time. Alternatively, we could have sampled from the full posterior for Θ, which, for each sample, would have allowed us to infer a complete epidemiological state of subregion 13a at the final observation time. We could then have run each simulation forward from this time.

Model limitations. Our model for HLB spread is not uniformly applicable. In subregions with very few symptomatic trees, a very large length scale α can be favored. This outcome reflects an underlying identifiability problem because in this case the dispersal kernel is essentially constant and mimics an external infection. Such a situation occurred in the large subregion in the southeast corner of Southern Gardens. Additionally, subregions with trees of differing ages can be problematic because the epidemic will proceed very differently in different parts of the subregion. Because the parameters are phenomenological quantities combining effects due to host and vector, dispersal parameters can only be understood in an averaged sense.

1. Gibson GJ (1997) Investigating mechanisms of spatiotemporal epidemic spread using stochastic models. *Phytopathology* 87(2):139–146.
2. Gibson GJ (1997) Markov chain Monte Carlo methods for fitting spatio-temporal stochastic models in plant epidemiology. *J R Stat Soc [Ser C]* 46(2):215–233.
3. O'Neill PD, Roberts GO (1999) Bayesian inference for partially observed stochastic epidemics. *J R Stat Soc [Ser A]* 162(1):121–129.
4. Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82(4):711–732.

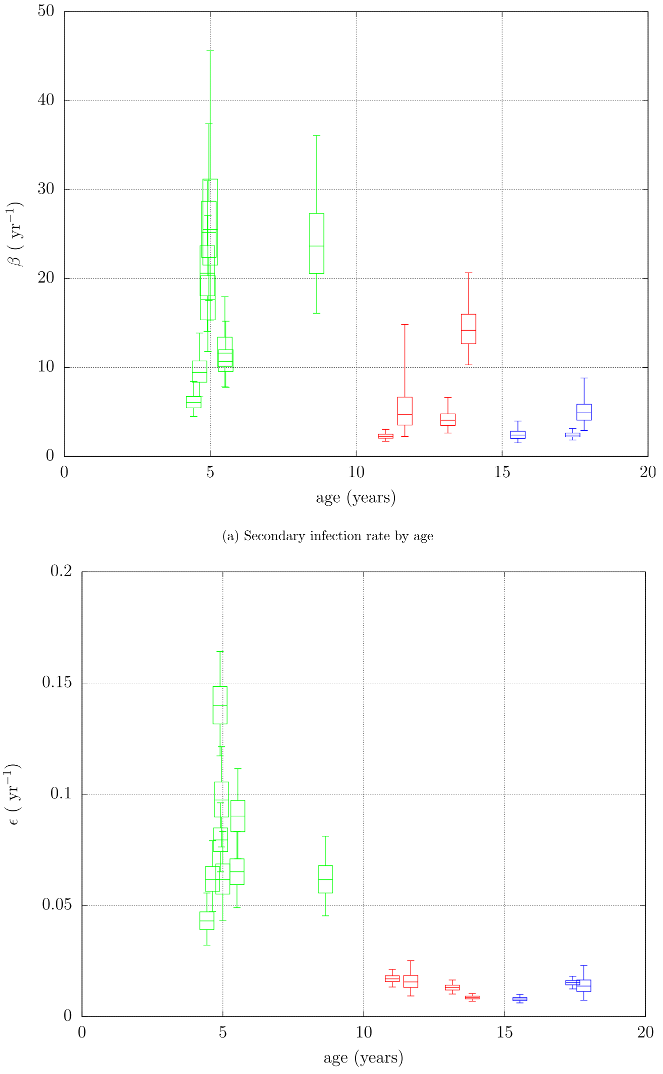

(a) Sweep 1

(b) Sweep 2

(c) Sweep 3

(d) Sweep 4

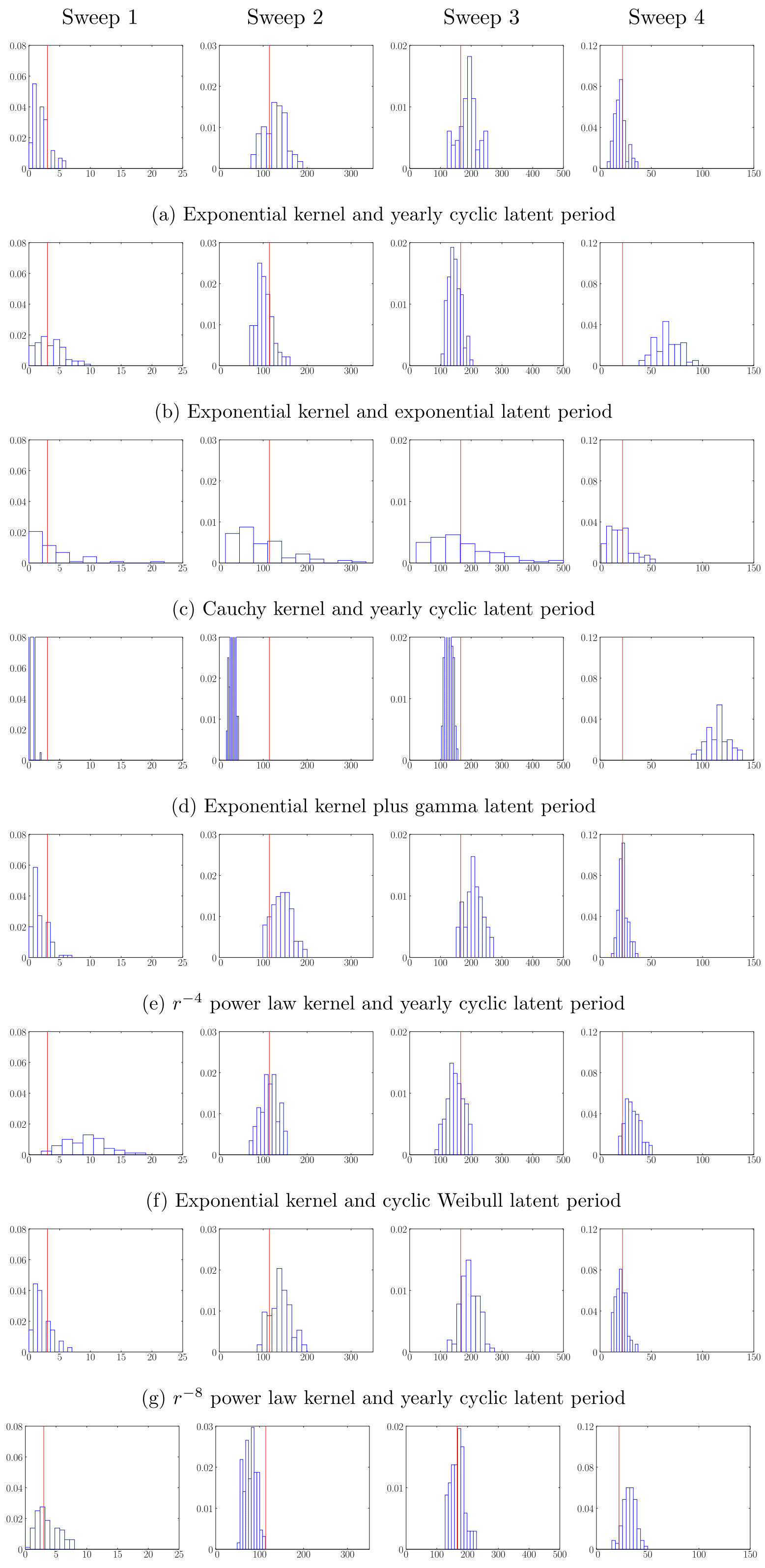
Fig. S1. (*A*) Sweep 1: November 2005 to March 2006, 1,748 detections, shown in blue. (*B*) Sweep 2: September 2006 to November 2006, 12,962 detections, shown in green. (*C*) Sweep 3: January 2007 to April 2007, 10,591 detections, shown in orange. (*D*) Sweep 4: May 2007 to July 2007, 2,626 detections, shown in red.



(a) Secondary infection rate by age

(b) Primary infection rate by age

Fig. S2. Boxplots of the transmission parameter posteriors for 16 subregions; the subregions are color-coded by age: young (green), old (blue), and mixed (red). (*A*) Secondary rate of infection, β, by average age of subregion at estimated epidemic start time. (*B*) Primary rate of infection, ε, by average age of subregion at estimated epidemic start time.



(a) Exponential kernel and yearly cyclic latent period

(b) Exponential kernel and exponential latent period

(c) Cauchy kernel and yearly cyclic latent period

(d) Exponential kernel plus gamma latent period

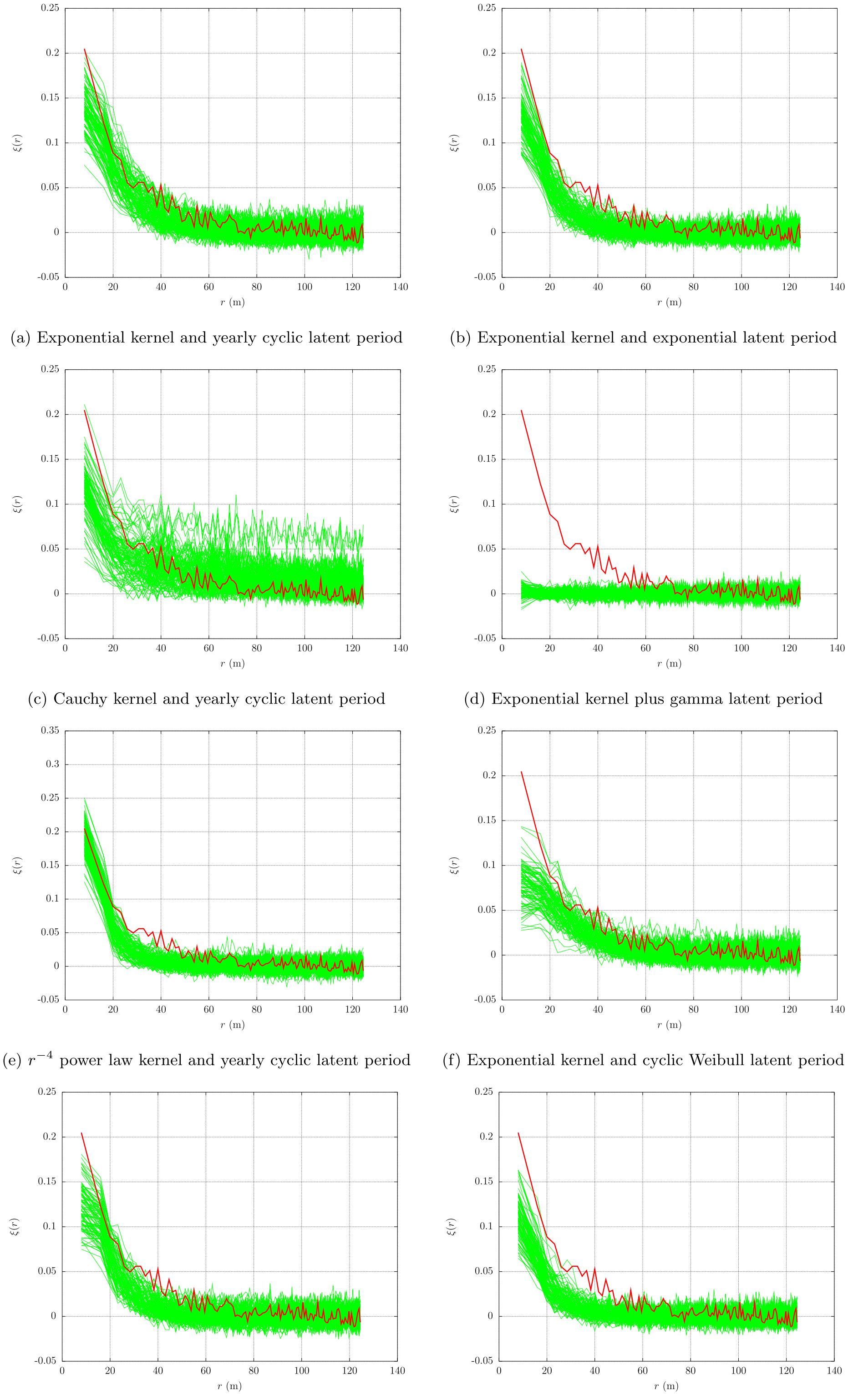
(e) r^{-4} power law kernel and yearly cyclic latent period

(f) Exponential kernel and cyclic Weibull latent period

(g) r^{-8} power law kernel and yearly cyclic latent period

(h) Exponential kernel and twice-yearly cyclic latent period

Fig. S3. Comparing models by observed detections in each of the four successive sweeps. The red line is the actual observed count; the frequency histogram (scaled to have unit area) of counts from 100 simulations is in blue. The vertical and horizontal scales are the same for each sweep.



(a) Exponential kernel and yearly cyclic latent period

(b) Exponential kernel and exponential latent period

(c) Cauchy kernel and yearly cyclic latent period

(d) Exponential kernel plus gamma latent period

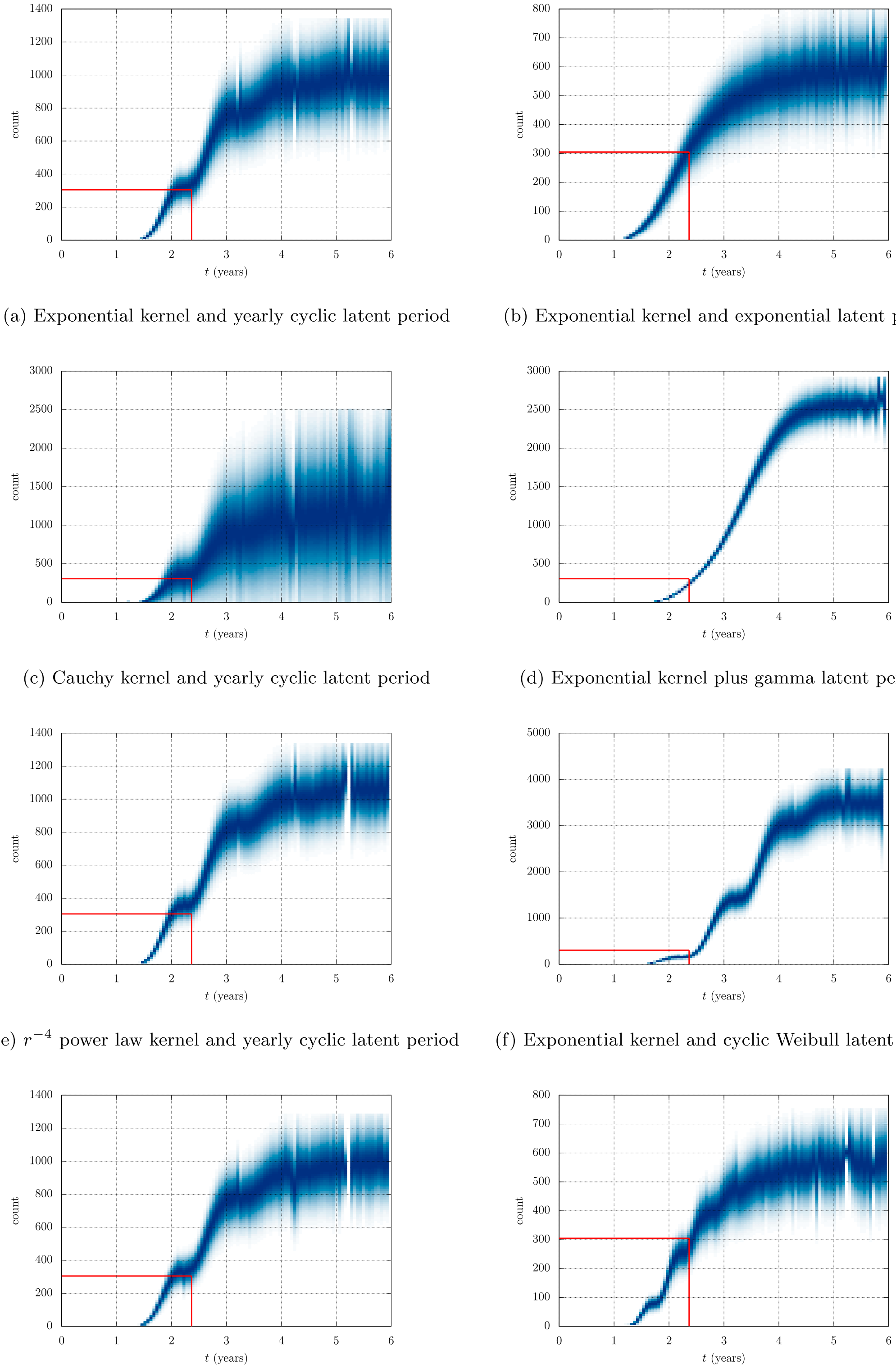
(e) r^{-4} power law kernel and yearly cyclic latent period

(f) Exponential kernel and cyclic Weibull latent period

(g) r^{-8} power law kernel and yearly cyclic latent period

(h) Exponential kernel and twice-yearly cyclic latent period

Fig. S4. Comparing models by the two-point spatial correlation of all detections. The red line is the actual observed two-point correlation; the two-point correlations from 100 simulations are in green. Note that the actual correlation function goes to zero at about 80 m.



(a) Exponential kernel and yearly cyclic latent period

(b) Exponential kernel and exponential latent period

(c) Cauchy kernel and yearly cyclic latent period

(d) Exponential kernel plus gamma latent period

(e) r^{-4} power law kernel and yearly cyclic latent period

(f) Exponential kernel and cyclic Weibull latent period

(g) r^{-8} power law kernel and yearly cyclic latent period

(h) Exponential kernel and twice-yearly cyclic latent period

Fig. S5. Comparing models by predicted future detection counts. The red lines meet at the actual observed count at the final observation time; the counts from 100 simulations are in blue. White stripes in the plots are a result of the latent period models with cyclic hazard rates.

Table S1. Expert opinion of upper bound on combined latent and cryptic periods for trees of different ages

Age (y)	Upper combined latent and cryptic period (y)
$1 < a \leq 3$	1.5
$3 < a \leq 10$	3
$a > 10$	4

Table S2. Posterior mean and 95% credible region for α measured in units of meters

Model	Dispersal kernel $k(u)$	Length scale α (m)
a	$\exp(-u)$	$6.9^{8.7}_{5.3}$
b	$\exp(-u)$	$6.0^{7.8}_{4.4}$
c	$(1+u^2)^{-1}$	$0.6^{1.6}_{0.02}$
d	$\exp(-u)$	$4.0^{5.7}_{2.8}$
e	$(1+u^4)^{-1}$	$7.2^{11.1}_{5.0}$
f	$\exp(-u)$	$6.9^{8.9}_{5.3}$
g	$(1+u^8)^{-1}$	$16.4^{13.5}_{20.1}$
h	$\exp(-u)$	$6.0^{7.8}_{4.4}$