A robust-design formulation of the incidence function model of metapopulation dynamics applied to two species of rails

Benjamin B. Risk,[1] Perry de Valpine, and Steven R. Beissinger[2]

*Department of Environmental Science, Policy, and Management, University of California, Berkeley, 137 Mulford Hall #3114,
Berkeley, California 94720-3114 USA*

Abstract. The incidence function model (IFM) uses area and connectivity to predict metapopulation dynamics. However, false absences and missing data can lead to underestimates of the number of sites contributing to connectivity, resulting in overestimates of dispersal ability and turnovers (extinctions plus colonizations). We extend estimation methods for the IFM by using a hierarchical Bayesian model to account both for false absences due to imperfect detection and for missing data due to sites not surveyed in some years. We compare parameter estimates, measures of metapopulation dynamics, and forecasts using stochastic patch occupancy models (SPOMs) among three IFM models: (1) a Bayesian formulation assuming no false absences and omitting site–year combinations with missing data; (2) a hierarchical Bayesian formulation assuming no false absences but incorporating missing data; and (3) a hierarchical Bayesian formulation allowing for imperfect detection and incorporating missing data. We fit the models to multiyear data sets of occupancy for two bird species that differ in body size and presumed dispersal ability but inhabit the same network of sites: the small Black Rail (*Laterallus jamaicensis*) and the medium-sized Virginia Rail (*Rallus limicola*). Incorporating missing data affected colonization parameters and led to lower estimates of dispersal ability for the Black Rail. Detection rates were high for the Black Rail in most years but moderate for the Virginia Rail. Incorporating imperfect detection resulted in higher occupancy and lower turnover rates for both species, with largest effects for the Virginia Rail. Forecasts using SPOMs were sensitive to both missing data and false absences; persistence in models assuming no false absences was more optimistic than from robust models. Our results suggest that incorporating false absences and missing data into the IFM can improve (1) estimates of dispersal ability and the effect of connectivity on colonization, (2) the scaling of extinction risk with patch area, and (3) forecasts of occupancy and turnover rates.

Key words: area; connectivity; false absences; hierarchical Bayesian model; incidence function model;
Laterallus jamaicensis coturniculus *and* Rallus limicola; *metapopulation; missing data; robust design;*
stochastic patch occupancy models.

INTRODUCTION

Spatially realistic metapopulation models are used to understand local extinction and colonization dynamics of populations in habitat patches connected by occasional dispersal events (Hanski 1998). The most common of these, the incidence function model (IFM), only requires data on occupancy, patch size, and interpatch distances (Hanski 1999). The colonization probability of a focal site is a function of the occupancy status of the other sites, their distance to the focal site, and their area. Extinction probability is primarily related to patch area. The IFM can be used to estimate a species' average dispersal distance and its sensitivity to

[1] Present address: Department of Statistical Science, Cornell University, 301 Malott Hall, Ithaca, New York 14853 USA.

[2] Corresponding author. E-mail: beis@berkeley.edu

patch area (Moilanen 2004). It is also used to create stochastic patch occupancy models (SPOMs) that forecast occupancy dynamics and metapopulation persistence (e.g., O'Hara et al. 2002, Bulman et al. 2007). The IFM has been applied to diverse taxa including amphibians (ter Braak and Etienne 2003), mammals (Moilanen et al. 1998, Ozgul et al. 2006), birds (Hanski 1998), and insects (Hanski 1999).

Moilanen (2002) identified three data quality issues that adversely affected the parameterization of the IFM: (1) poorly estimated site areas, which led to mischaracterizations of the scaling of extinction risk; (2) missing data, which led to overestimates of dispersal ability; and (3) false absences, which affected all model components. Moilanen (2002) advocated the use of GIS methods to estimate site areas and suggested using average occupancy in multiyear data sets in place of missing data in calculations of connectivity, which could be adequate if fewer than 10% of values are missing (Moilanen 2004). Alternatively, the joint distribution of occupancy for

missing data can be estimated directly using a hierarchical approach, where missing data are treated as unobserved state variables (ter Braak and Etienne 2003).

False absences are the most problematic of the factors affecting IFM parameterization (Moilanen 2002) and occur when surveyors fail to detect the presence of a species. They lead to overestimates of dispersal ability and the number of turnovers (local extinctions plus colonizations). Moilanen (2002) formulated a method that used an independent estimate of the probability of a false absence at a site in a year, but it did not allow for multiple visits to a site within a year and assumed the probability of detection was constant across time. More recently, the robust design from capture–recapture studies (Pollock 1982) has been applied to occupancy studies. Occupancy at a site is assumed to be closed within a season and detection rates are estimated from repeated visits to sites within a season (MacKenzie et al. 2003, 2006). By incorporating imperfect detection rates, these hierarchical occupancy models account for the probability of false absences. The effect of covariates for the detection model (e.g., visit number or year) and the occupancy model (e.g., area or connectivity) can be estimated using maximum likelihood (e.g., MacKenzie et al. 2003, 2006) or a Bayesian approach (e.g., Royle and Kery 2007, Royle and Dorazio 2008). Covariates can include occupancy-based connectivity metrics, i.e., metrics calculated from the occupancy status of nearby sites. MacKenzie et al. (2003) noted that metapopulation studies should incorporate false absences into connectivity. However, incorporating false absences in models with connectivity leads to a more complicated hierarchical model because the transition probabilities of every site depend on the imperfectly known state of every other site. False absences could potentially be incorporated into connectivity using classical methods, such as Monte Carlo kernel likelihood (de Valpine 2004) or data cloning (Lele et al. 2007), or with a Bayesian analysis.

In this study, we present a hierarchical Bayesian formulation of the IFM that incorporates missing data and imperfect detection and then fit the models to detection data on two species of secretive rails in wetland patches in the Sierra Nevada foothills. The Black Rail (*Laterallus jamaicensis*) is the smallest (about 30 g) North American rail (Eddleman et al. 1994) and requires relatively small territories (Legare and Eddleman 2001). It may be a poor disperser because it is small, has a limited geographic range, and performs very short, low flights on the rare occasions when it leaves dense vegetation (Taylor 1998). The Virginia Rail (*Rallus limicola*) is three times larger than the Black Rail, may require larger territories, and has a wide geographic range, all of which suggest it may be a better disperser than the Black Rail (Conway and Eddleman 1995). Both species occupy small wetlands amid a matrix predominantly composed of upland pasture that may present a barrier to dispersal. Additionally, turnovers (extinctions

and colonizations) are frequently observed. Thus, the network of local populations for each species may compose a metapopulation.

We compare parameter estimates, measures of metapopulation dynamics, and SPOM outcomes among three models: (1) a Bayesian formulation of the IFM assuming no false absences and omitting sites for particular years in which data were missing (IFM naive); (2) a hierarchical Bayesian formulation of the IFM assuming no false absences but incorporating missing data (IFM missing); and (3) a hierarchical Bayesian formulation of the IFM allowing for imperfect detection and incorporating missing data (IFM robust). To assess the impact of missing data, we compare whether estimates of dispersal ability differ between the IFM naive and IFM missing. We make a distinction between sites where data have been recorded in some years but not others and unknown sites where no data have been recorded; we do not address the issue of unknown sites in this study. To examine the impact of false absences, we compare the IFM missing to the IFM robust. For both species, we used each variant of the IFM to estimate (1) dispersal ability and the scaling of extinction risk with area, (2) occupancy and turnover rates, and (3) occupancy dynamics projected for 100 years.

Bayesian Formulations of the IFM

The IFM

We summarize the IFM as presented in Moilanen (2002). Our notation is described in Table 1 (see also Appendix A). The colonization probability at the ith site is defined as

$$C_{i,t} = \frac{S_{i,t}^2}{S_{i,t}^2 + \delta^2} \tag{1}$$

where $S_{i,t}^2$ is the square of the connectivity metric and δ^2 is a parameter creating a sigmoidally shaped colonization probability function. $S_{i,t}$ is a proxy for the number of migrants that survive from all other sites j in the network to site i and is

$$S_{i,t} = \sum_{j \neq i}^{m} z_{j,t-1} A_j^\beta \exp(-\alpha d_{ij}). \tag{2}$$

The quantity $z_{j,t-1} A_j^\beta$ is a proxy for population size, where $z_{j,t-1}$ is the occupancy (0 or 1) at site j in the previous year, and A_j is its area. The distribution of dispersal distances is assumed to be inversely related to the distance between the natal patch j and the focal patch i by the equation $\exp(-\alpha d_{ij})$, and α^{-1} is the "average dispersal distance" (Moilanen 2004), which we refer to as the dispersal index. Consistent with O'Hara et al. (2002), we assumed that connectivity was conditionally independent of the occupancy of sites in the current year given occupancy in the previous year.

TABLE 1. Notation in the incidence function model (IFM) for IFM naive, IFM missing, and IFM robust.

Symbol	Description
A_j	Area of site j.
α	Scales the effect of interpatch distance in the dispersal kernel.
β	Scales the effect of area on the number of emigrants from a patch.
$C_{i,t}$	Probability of colonization at site i in year t.
χ	Scales the effect of area on extinction risk.
δ	Larger values decrease colonization. Used to convert connectivity to a probability.
$d_{i,j}$	The distance between the centroids of sites i and j.
E_i	Intrinsic extinction rate at site i (time invariant).
$E_{i,t}$	Extinction rate at site i in year t with rescue effect.
$I_{i,t}$	Indicator variable equal to 0 if data are missing at site i in year t; 1 otherwise.
m	The total number of sites.
μ	Larger values increase extinction. In numerator of intrinsic extinction rate.
n	The number of years of data.
Ψ_1	Probability of occupancy in the first year of data.
$\psi_{i,t}$	Probability of occupancy at site i in year t.
p_t	Probability of detection given occupancy in year t.
$\mathbf{p}$	Vector of detection probabilities of length n.
$q_{i,t}$	The number of visits to site i in year t.
$s_{i,t}$	The number of visits with detections at site i in year t.
$S_{i,t}$	Connectivity at site i in year t.
Θ	Shorthand for the vector of IFM parameters, α, β, δ, μ, χ, and Ψ_1 when applicable.
$y_{i,t,r}$	Detection record at site i in year t for replicate (visit) r.
$\mathbf{y}$	The three-dimensional array of all detection records (for IFM robust only).
$z_{i,t}$	Occupancy state of site i in year t.
$\mathbf{z}_t$	Occupancy states from year t.
$\mathbf{z}$	Matrix of occupancy states for all sites and all years.

Intrinsic extinction rate (E_i) for the ith site describes the extinction probability in the absence of immigration and is independent of time:

$$E_i = \min\left(1, \frac{\mu}{A_i^\chi}\right) \qquad (3)$$

(see Table 1). The minimum patch area, A_0, is the threshold at or below which the intrinsic extinction rate is 1 and is equal to $\mu^{1/\chi}$ (Moilanen 1999). The rescue effect occurs when extinction risk is reduced by immigration from other sites; that is, the distances between a focal patch and occupied patches affect the probability of extinction in the focal patch. It is incorporated into the IFM by multiplying intrinsic extinction rates by the probability of no recolonization, $(1 - C_{i,t})$ (Hanski 1994, 1999):

$$E_{i,t} = (1 - C_{i,t})\left[\min\left(1, \frac{\mu}{A_i^\chi}\right)\right]. \qquad (4)$$

Eq. 4 is appropriate in systems where small sites with high intrinsic extinction rates can be occupied due to immigration from nearby sites (Hanski 1994). We detected rails in small sites that were proximate to larger sites. Let $\mathbf{z}$ denote all occupancy states and let $z_{i,t-1}$ be an indicator variable for occupancy at the ith site in year $t - 1$. The probability of occupancy is $\psi_{i,t}$:

$$\psi_{i,t} = z_{i,t-1}(1 - E_{i,t}) + (1 - z_{i,t-1})C_{i,t}. \qquad (5)$$

This formula is used in likelihood calculations, as described in the next section.

The IFM assuming perfect detection and omitting site-years with no data (IFM naive)

In the IFM naive model, an element $z_{i,t-1}$ of $\mathbf{z}$ was equal to one if there was a detection in any of the visits to site i in year $t - 1$ and 0 otherwise (Eq. 5; see Table 1). Here and throughout we make the standard assumption that occupancy of a site does not change between the visits within one year's survey season (but may change between years). The likelihood is

$$P(\mathbf{z} \mid \Theta) = \prod_{t=2}^{n} \prod_{i=1}^{m} \left(\psi_{i,t}^{z_{i,t}}(1 - \psi_{i,t})^{(1-z_{i,t})}\right)^{I_{i,t}} \qquad (6)$$

where $I_{i,t}$ is an indicator variable equal to 0 if site i was not sampled in both years t and $t - 1$; $\Theta = (\delta, \beta, \alpha, \mu, \chi)$; and there are m sites and n years. Sites with data in one year were included in the likelihood when they had data in the previous year, as it is only possible to calculate $\psi_{i,t}$ when $z_{i,t-1}$ is available (Eq. 5). For example, if site 2 was surveyed in all years except year 3, then the transitions from years 2 to 3 and from years 3 to 4 are not incorporated in parameter estimation. In effect, this approach assumed the network changed across years to the extent that the sites surveyed in pairs of years differed. The same sites in the likelihood for year t were used in calculating $S_{i,t}$ (i.e., sites with data in year t and $t - 1$).

The joint posterior distribution of IFM parameters was estimated via Bayes' Theorem, $P(\Theta \mid \mathbf{z}) = P(\mathbf{z} \mid \Theta)P(\Theta)/\int P(\mathbf{z} \mid \Theta)P(\Theta)d\Theta$, where $P(\mathbf{z} \mid \Theta)$ is the likelihood (Eq. 6) and $P(\Theta)$ are the priors. We used Markov chain Monte

Carlo (MCMC) and the Metropolis-Hastings (MH) algorithm with component-wise updating to sample parameters from their joint conditional distribution (see Appendix A; e.g., Gilks et al. 1996, Gelman et al. 2004). Priors are described in *Priors, model fitting, and model fit*.

The IFM incorporating missing data (IFM missing)

The second model dealt with the unsurveyed site–year combinations differently by substituting unobserved random variables for missing data. The unobserved random variables represent latent states of the population. We use the joint posterior distribution of these states in estimates of connectivity, occupancy, and turnover rates. We included a model for the first year of data, where occupancy at a site followed a Bernoulli distribution with probability equal to Ψ_1, although the formulation could be extended to include covariates. Our approach differed from O'Hara et al. (2002), who conditioned all inference on occupancy in the first year, and from ter Braak and Etienne (2003), who estimated the posterior distribution of first year data by assuming quasi-stationarity.

In the IFM missing model, $z_{i,t}$ was the same as for the IFM naive model for sites with data, but $z_{i,t}$ could be 0 or 1 for sites with missing data according to the Bayesian posterior probabilities. It is helpful to write the likelihood of occupancy states by factoring occupancy in different years and canceling conditionally independent terms:

$$P(\mathbf{z} \mid \Theta) = P(\mathbf{z}_n \mid \mathbf{z}_{n-1}, \Theta)P(\mathbf{z}_{n-1} \mid \mathbf{z}_{n-2}, \Theta)\cdots$$
$$P(\mathbf{z}_2 \mid \mathbf{z}_1, \Theta)P(\mathbf{z}_1 \mid \Theta) \qquad (7)$$

where now $\Theta = (\delta, \beta, \alpha, \mu, \chi, \Psi_1)$. Then, $P(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \Theta)$ for $t \neq 1$ is equal to

$$\prod_{i=1}^{m} \psi_{i,t}^{z_{i,t}}(1 - \psi_{i,t})^{(1-z_{i,t})} \qquad (8)$$

where $\psi_{i,t}$ is calculated using Eq. 5, and $P(\mathbf{z}_1 \mid \Theta)$ is calculated with $\psi_{i,1}$ equal to Ψ_1. Note that the term $S_{i,t+1}$ is calculated over all states $\mathbf{z}_t$ (i.e., including missing data) (Eq. 2) and then used in colonization and extinction probabilities (Eqs. 1 and 4). The missing data are included in the MCMC using block sampling (Appendix A: Section 3). Priors are described in *Priors, model fitting, and model fit*.

Robust-design formulation of the IFM (IFM robust)

The IFM robust is a hierarchical state-space model that includes an observation model (the detection model) and a process model (the occupancy model). Data used in the IFM robust are generated from multiple surveys at each site within a season, where the patches are assumed to be closed to changes in occupancy within a season (MacKenzie et al. 2003, 2006). In the observation model, we used year-specific detection probabilities, $p_1, \ldots, p_n$, where $\mathbf{p}$ is the vector

of these detection rates, and $\mathbf{y}$ denotes all detection data from all years, including the multiple visits to each site. This approach is a relatively straightforward way to account for potential year-to-year variation in surveyors and/or survey methodology, which characterized our study system (see *Study area and sampling methods*).

A state $z_{i,t}$ is assumed to be known perfectly if at least one detection was recorded. We write the likelihood of the data as

$$P(\mathbf{y} \mid \Theta, \mathbf{p}) = \sum_{\text{all } z=\{0,1\}} P(\mathbf{y}, \mathbf{z} \mid \Theta, \mathbf{p})$$
$$= \sum_{\text{all } z=\{0,1\}} P(\mathbf{y} \mid \mathbf{z}, \mathbf{p})P(\mathbf{z} \mid \Theta) \qquad (9)$$

where $\Theta = (\delta, \beta, \alpha, \mu, \chi, \Psi_1)$ as in the IFM missing. The term $P(\mathbf{y} \mid \mathbf{z}, \mathbf{p})$ is the observation model, and the term $P(\mathbf{z} \mid \Theta)$ is the process model. The summation is over all possible combinations of 0 and 1 for all $z_{i,t}$ states; this is not directly calculated but is reflected in the MCMC algorithm. When there are multiple visits to an occupied site, the detections in the observation model follow a binomial distribution with the probability of a success equal to the probability of detection, p. Let $q_{i,t}$ and $s_{i,t}$ denote the number of visits and detections (i.e., trials and successes), respectively, at site i in year t

$$P(\mathbf{y} \mid \mathbf{z}, \mathbf{p})$$
$$= \prod_{t=1}^{n}\prod_{i=1}^{m}\left[z_{i,t}\binom{q_{i,t}}{s_{i,t}}p_t^{s_{i,t}}(1-p_t)^{q_{i,t}-s_{i,t}} + (1-z_{i,t})\right]^{I_{i,t}} \qquad (10)$$

(notation described in Table 1). In this case, $I_{i,t}$ is 1 if site i was visited in year t, 0 otherwise. Next, the process model is equal to the likelihood described in Eq. 7. The latent variables (non-detections and missing data) are updated in the same manner as the IFM missing using block sampling. Details are provided in Appendix A: Section 4.

Priors, model fitting, and model fit

IFM parameters were restricted to non-negative values, where negative parameter values are biologically unreasonable (α, β, μ, χ) or redundant (δ). In our preliminary model estimation, the posterior of some IFM parameters included zero. We estimated the posteriors for the log of IFM parameters to allow better mixing in the region near 0. Initially, we used uniform priors from -10 to 10 on the logs of α, β, δ, μ, and χ and from 0 to 1 for Ψ_1 and p_t. In these preliminary models, $\log \beta$ was highly correlated with $\log \delta$ at values of $\log \beta$ greater than 0 (Pearson's $\rho > 0.95$ in both the IFM naive and IFM missing in both species). Consequently, we used an informative uniform prior from -10 to 0 on $\log \beta$. Values of β greater than one indicate that population size increases at an increasing rate with area (see Eq. 2); thus, restricting β to less than 1 has a reasonable biological interpretation. Furthermore, previous studies of the IFM often use fixed values for β, with β

sometimes set to 0.5 (e.g., ter Braak and Etienne 2003, Bulman et al. 2007). We believe our approach balances the potential bias introduced assuming a known value for β with the issues of model overspecification.

Two chains were estimated for each model for each data set, and convergence was diagnosed using multiple methods: visual examination of mixing and cumulative quantile plots (0.025, 0.5, and 0.975), Geweke's diagnostic, and Heidelberger and Welch's diagnostic (Appendix B; Cowles and Carlin 1996). We executed 255,000 iterations per chain in all models, discarded the first 5000 as the burn-in period, and retained one iteration for each 100 iterations (i.e., thinning = 100). Our analysis was done using the software R (R Development Core Team 2010) and execution times ranged from 5 hours (Virginia Rail IFM naive) to 55 hours (Black Rail IFM robust) on a 2-GHz AMD Opteron 246 processor (Penguin Computing, Fremont, California, USA). Simulated data and MCMC code are available in the Supplement. WinBUGS can potentially be used, but we found it to be computationally impracticable when large distance matrices were used in connectivity calculations.

We assessed model fit by constructing a Bayesian p value from the deviance of the posterior predictive distribution and the deviance of the data, where p values near 0 or 1 indicate a substantive failure of the model (Gelman et al. 2004). Details of these calculations are provided in Appendix B.

Assessing model performance with simulated data sets

To demonstrate the importance of missing data and the observation model, we fit the IFM models to a simulated data set where the true states of sites were perfectly known. A rigorous simulation study with hundreds of data sets is beyond the scope of the present investigation. Instead, we focused on one simulated data set, rather arbitrarily chosen, in order to compare the models to the true states. We created a network of sites based on parameters from our field data. Then we created a 10-year data set based on known IFM parameter values, and we set 10% of the data to missing and annual detection rates ranging from 0.2 to 0.9. The three IFM models were fit to the data, and the resulting posteriors were used to simulate occupancy for nine years, starting from the occupancy states in the first year of the true data set. Using this approach, differences between the three models were due to differences in parameter estimates and not initial conditions. The predicted number of occupied sites and number of turnovers were then compared to the true states. See Appendix B for details of the simulation data set and parameters.

Forward projections using stochastic patch occupancy models

Stochastic patch occupancy models use transition probabilities defined by a metapopulation model to create a Markov chain of occupancy patterns (Moilanen 1999). SPOMs were simulated for a network of 228 sites (see *Study area and sampling methods*). For the IFM naive, a set of parameter values was randomly sampled from the joint posterior, and initial occupancy was equal to the most recently observed occupancy state. For the IFM missing and IFM robust, a set of parameter values and occupancy in the last year of the study was randomly sampled from the joint posterior. Transition probabilities for year $t + 1$ were calculated from occupancy in year t using the IFM parameters (Eq. 5). For each model variant, a total of 5000 samples from the joint posterior were used to simulate occupancy dynamics for 100 years for each sample. We calculated year-specific occupancy, number of turnovers (extinctions plus colonizations), and the proportion of simulations where global extinction occurred.

Predicted effects of incorporating missing data and false absences

Missing data if ignored in the analysis may cause an overestimate of dispersal ability, and false absences can affect all parameters and lead to inflated turnover probabilities (Moilanen 2002). We considered how these data issues would affect α, μ, and χ. We predicted α would be smallest in the IFM naive (and the dispersal index, 1/α, would be largest), intermediate in the IFM missing (dispersal index intermediate), and largest in the IFM robust (dispersal index smallest). We did not have a priori expectations for whether missing data would affect the parameters determining intrinsic extinction rate, i.e., μ and χ. However, we predicted μ would be smaller and χ larger in the IFM robust than the other models, resulting in lower intrinsic extinction rates in the IFM robust. We predicted turnover rates would be higher in models that did not incorporate imperfect detection and that average occupancy would be highest in the IFM robust. We predicted that SPOMs based on the IFM naive and IFM missing would have more variable dynamics and higher rates of global extinction, as, all else equal, higher turnover rates could increase the probability of simultaneous local extinctions.

STUDY AREA AND SAMPLING METHODS

The field study was conducted from 2002 to 2008 at 228 sites in Yuba, Butte, and Nevada counties at the eastern edge of the Central Valley and in the foothills of the Sierra Nevada Mountains in California (33–790 m). See Richmond et al. (2008, 2010a) for a map of the study area and other details. Sites are palustrine emergent persistent wetlands (Cowardin et al. 1979) separated by unsuitable pasture or oak savanna. The median site area was 0.48 ha, and the range was 0.01–13.9 ha. The median nearest neighbor distance between site centroids was equal to 0.5 km, and the maximum distance between any two sites was 49 km.

We attempted to survey all sites in the study area but were unable to obtain permission from landowners for

some sites. Sites that exist but were never surveyed are not accounted for in our models, and thus represent a potential source of error in our estimation of connectivity. Sites were delineated with a Trimble GPS unit (Trimble, Sunnyvale, California, USA) by walking the edge of a wetland, which was defined as having greater than 50% wetland plants (Cowardin et al. 1979). Areas were calculated using ArcGIS 9.2 (ESRI, Redlands, California, USA). Areas of open water were excluded as rails avoid them. The distance between patches was measured as the Euclidean distance between the centroids of sites.

Black Rails were surveyed from May to early August in 2002–2008 and Virginia Rails were surveyed in 2004–2008. For Black Rails, we used playback surveys with stations every 50 m as described in Richmond et al. (2008). Sites were visited up to three times following a removal design for Black Rail detections (MacKenzie et al. 2006). In 2004, separate playbacks were not conducted for Virginia Rails, but occupancy was noted when Virginia Rails spontaneously called or responded to Black Rail playbacks. In 2005, Virginia Rail playbacks were conducted at two stations anywhere in a wetland during the first visit but were rarely done during the second and third visits. From 2006–2008, Virginia Rail playbacks were conducted whenever a site was visited at points 100 m apart until a detection was recorded or the entire wetland was traversed; for details, see Richmond et al. (2010b). In 2008, the visit protocol was modified slightly, where sites were visited at least twice for the Virginia Rail. Surveys were conducted from 30 minutes before sunrise to two hours afterwards and two hours before to 30 minutes after sunset. For the Black Rail, there were 109 (49% of sites), 121 (53%), 136 (60%), 169 (74%), 195 (86%), 201 (88%), and 203 (89%) sites surveyed from 2002 to 2008, respectively. For the Virginia Rail, there were 107 (47%), 168 (73%), 193 (85%), 201 (88%), and 203 (89%) sites surveyed from 2004 to 2008, respectively. Note that the number of sites not surveyed exceeded 10%, suggesting that missing data should be treated in a more exhaustive way than using the site averages from years with data (i.e., the rule of thumb from Moilanen 2004). The increasing number of sites was due to the expansion of the study and motivated direct treatment of missing data.

RESULTS

Assessing model performance with simulated data

When fit to the simulated data set with false absences and missing data, the actual occupancy and turnover fractions that occurred in the simulation were accurately predicted by the IFM robust, while the IFM missing performed less well, and the IFM naive performed poorly. The IFM naive tended to overpredict both occupancy and turnovers (Fig. 1), which may have been due to overestimates of dispersal ability resulting from the missing data. The IFM missing tended to underpredict occupancy and overpredict turnovers. Compared

to the IFM naive, it predicted occupancy better but performed only slightly better with respect to turnovers. The true occupancy and turnover rates were well within the 95% credible intervals for the IFM robust. The IFM robust tended to underpredict more often than overpredict occupancy, although the true numbers of sites occupied were well within the 95% credible intervals, and we believe this tendency was negligible.

Detection, occupancy, and turnover rates for rails

Detection probabilities in most years were high for Black Rails and moderate for Virginia Rails (Fig. 2A, B). The median probability of detecting a Black Rail per visit exceeded 0.74 for all years except 2007 (0.36). As a result, the probability of detecting a Black Rail at a site with up to three visits exceeded 0.98 for all years except 2007 (0.74). Single visit detection probabilities for Virginia Rails varied considerably by year; the median value was lowest in 2004 (0.13) when playback surveys were used sporadically but ranged from 0.49 to 0.73 in other years. The probability of detecting a Virginia Rail at a site with up to three visits ranged from 0.87 to 0.98 except for 2004 (0.34). Thus, false absences should be expected to have a greater effect on model estimates for Virginia compared to Black Rails.

The effects of IFM model structure on occupancy and turnovers depended largely on detectability. For the highly detectable Black Rail, occupancy (Fig. 2C) and turnover rates (Fig. 2E) from the three models were generally similar in all years except 2007, although turnover rates tended to be somewhat higher in the IFM missing and intermediate in the IFM robust. In 2007 when detection rates were lowest, occupancy was substantially higher and turnover was lower in the IFM robust compared to the other models. For the less detectable Virginia Rail, in all years occupancy was higher (Fig. 2D) and turnover was lower (Fig. 2F) in the IFM robust compared to the IFM naive and IFM missing.

Effects of model structure on IFM parameters for rails

Colonization parameters appeared to be more strongly affected by incorporating missing data into the IFM than false absences in the highly detectable Black Rail. The parameters α and δ substantially increased and decreased, respectively, when missing data were incorporated but changed little when false absences were included, while β differed markedly among the three models (Fig. 3A–C). Notably, the high value of β in the IFM no missing suggested that the number of dispersers leaving a patch increases strongly with patch area, while the credible interval of β in the IFM missing indicated highly variable effects, and the narrower credible interval and median near zero in the IFM robust indicated weak effects. The median and credible intervals for δ were larger in the IFM naive than in the other models, while the quantiles for δ were similar in the IFM missing and IFM robust. The dispersal index

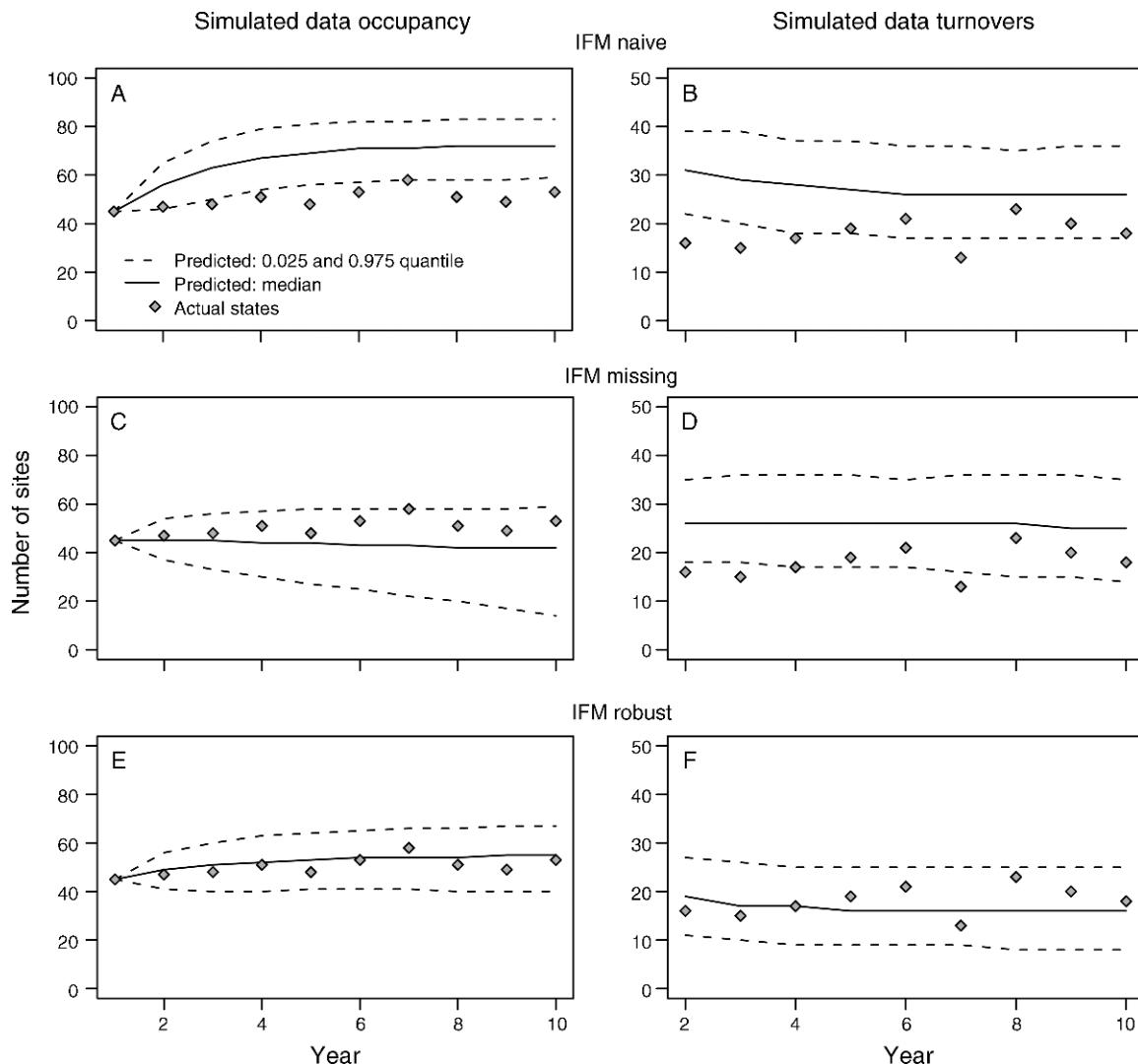## Simulated data occupancy / Simulated data turnovers

FIG. 1. Predicted occupancy and turnovers vs. actual occupancy and turnovers from a simulated data set based on a network with 100 sites, 10% missing data, and year-specific detection rates varying from 0.2 to 0.9. See Appendix B for additional details.

(i.e., $1/\alpha$) in the IFM naive indicated moderate to high dispersal distance (median 22.4 km, 95% CI 9.6–2729.1) relative to the scale of our study area, where the maximum distance between sites was 49 km. However, it decreased dramatically in the IFM missing and IFM robust, indicating restricted dispersal ability (median 8.9 km, 95% CI 5.4–18.6; median 8.1 km, 95% CI 5.1–16.6, respectively; Fig. 3D). The decrease in estimates of the dispersal index is consistent with predictions.

The effects of missing data and false absences on colonization parameters in the Virginia Rail were more variable. The parameter α differed little among the three models but β varied greatly, and δ increased strongly when false absences were included (Fig. 3A–C). Notably, the high values of β in the IFM naive indicated a strong effect of area on the number of dispersers, whereas β declined to near zero in the IFM missing.

Interestingly, the 95% credible interval of β for the IFM missing was narrow, while the IFM robust credible interval was broad, and its posterior for β differed little from the prior (Fig. 3B). The parameter δ was largest in the IFM robust, which also had the largest credible intervals. The dispersal index for the Virginia Rail was much larger than the Black Rail and differed little among models (Fig. 3D).

Extinction parameters changed little when missing data were incorporated into IFM models, whereas incorporating false absences had a larger influence on parameter estimates for the less-detectable Virginia Rail than for the Black Rail. Credible intervals for both μ and χ overlapped for the three Black Rail models (Fig. 3E, F); μ tended to be lower in the IFM robust, suggesting some effect of false absences. In the Virginia Rail, μ and χ were relatively unaffected by missing data,
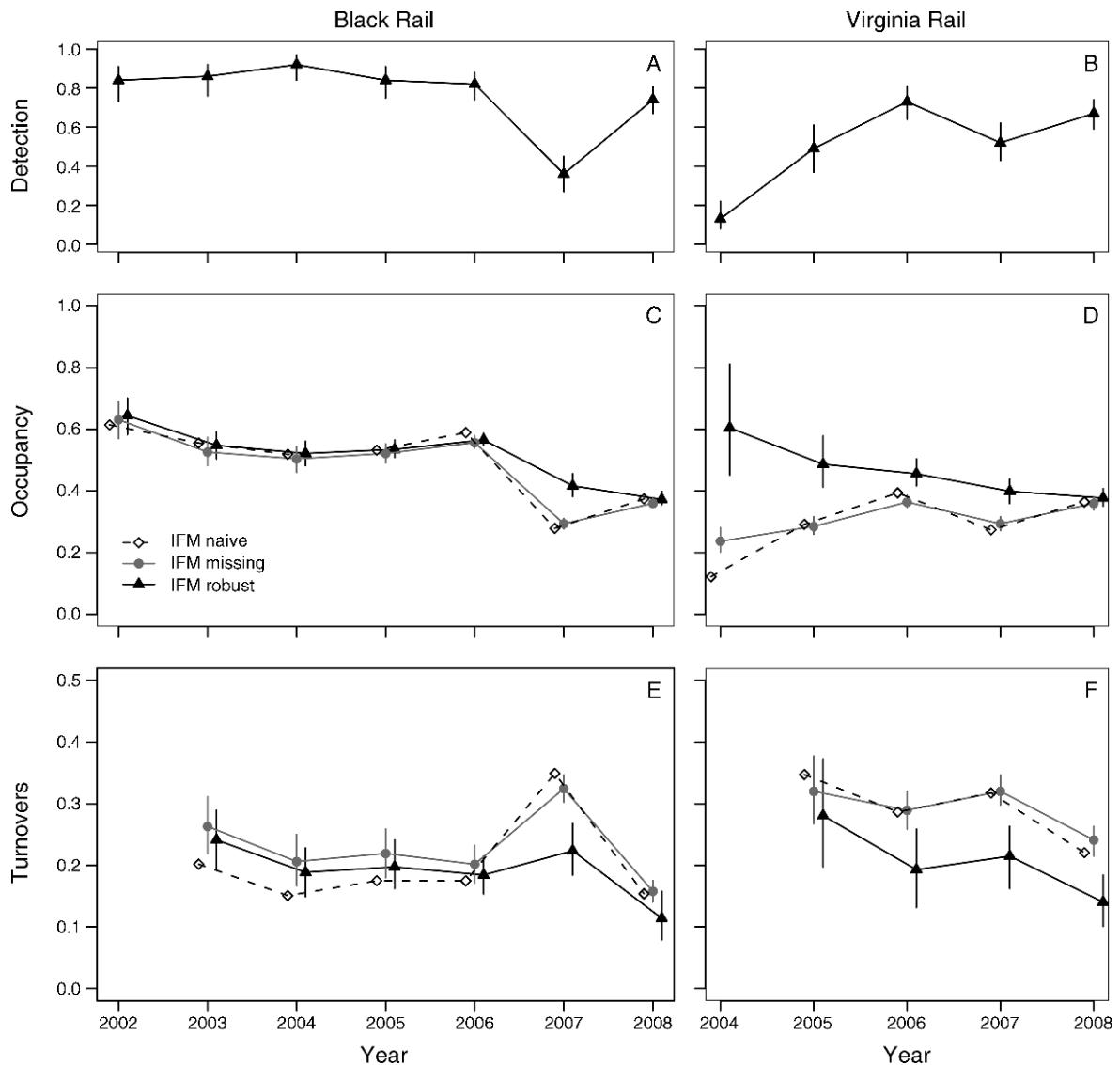
FIG. 2. Single-visit detection probabilities by year for the (A) Black Rail and (B) Virginia Rail estimated using the IFM robust. (C, D) Black and Virginia Rail average occupancy and (E, F) turnover rates estimated using the IFM naive, IFM missing, and IFM robust. Medians are plotted, and vertical lines represent the 95% credible intervals (0.025–0.975 quantiles).

but μ was much lower and χ was somewhat higher when false absences were incorporated in the IFM robust (Fig. 3E, F); this is consistent with predictions. For both species, minimum patch area (A_0) and extinction risk of median and maximum patch area tended to be lower in the IFM robust, but differences among models were much greater for the Virginia Rail (Fig. 3G–I). Minimum patch area tended to be larger in Virginia than Black Rails, although the 95% credible intervals of their respective IFM robust models overlapped. Intrinsic extinction rates were 1.5–3 times greater for the Virginia than Black Rail in models assuming perfect detection (nonoverlapping at median area), but they were similar when accounting for false absences (Fig. 3H, I).

Based on the posterior predictive P values from the deviance, model fit of the IFM naive performed poorly for both the Black and Virginia Rail data sets ($P < 0.01$ for both). The IFM missing performed adequately for the Black Rail ($P = 0.31$), while less well but still acceptably for the Virginia Rail ($P = 0.13$). The IFM robust performed well for both species ($P = 0.74$ and 0.45, respectively).

Projected occupancy, turnover rates, and persistence

Results from SPOMs for the Black Rail metapopulation were highly variable, and a large proportion of simulations went extinct in all model variants (Fig. 4A–C; Fig. 5). Median occupancy approached zero more
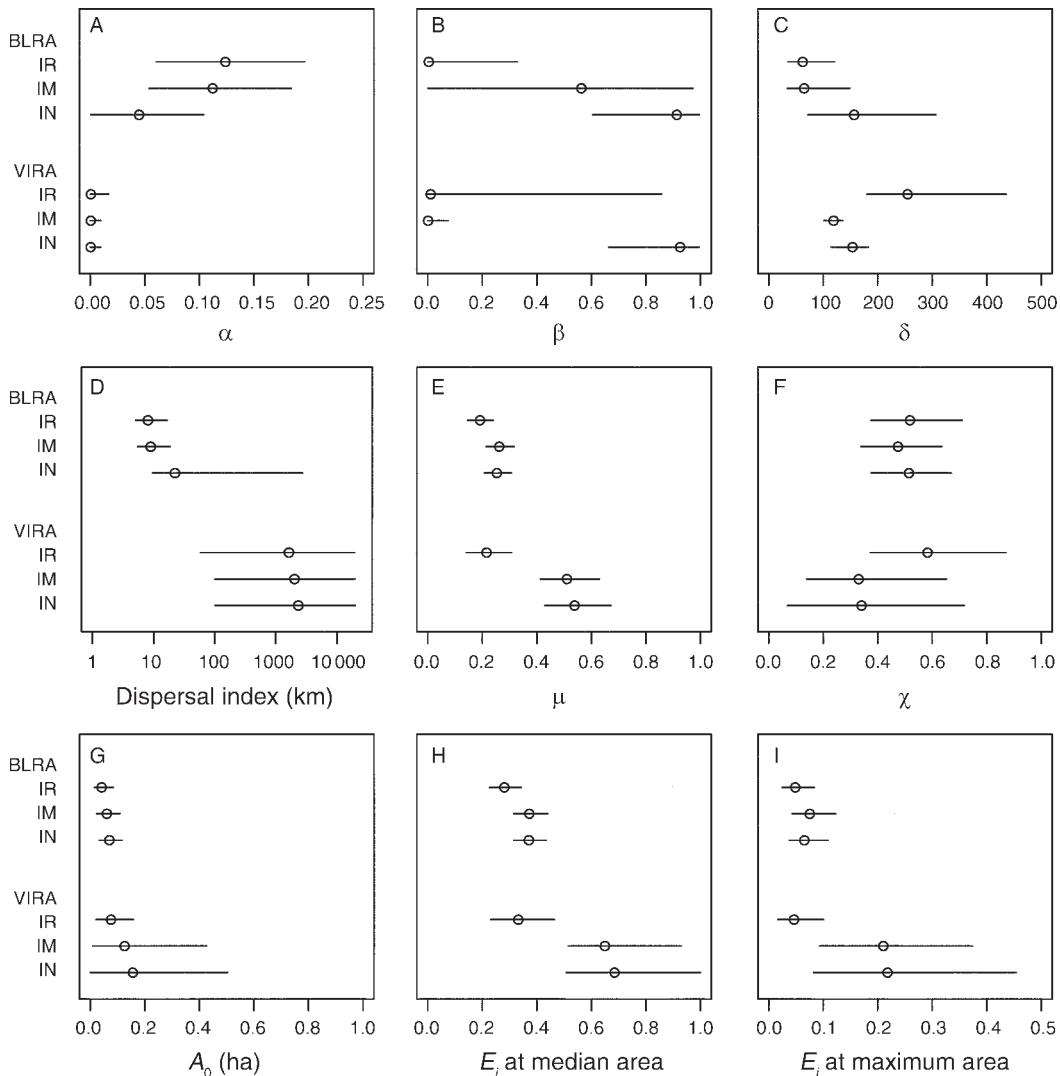
FIG. 3. IFM parameters and derived measures of metapopulation dynamics. Dots indicate medians, and lines indicate 95% credible intervals (0.025–0.975 quantiles). Logs of the IFM parameters were estimated, and back-transformed quantiles are plotted here. Key to abbreviations: BLRA, Black Rail; VIRA, Virginia Rail; IR, IFM robust; IM, IFM missing; IN, IFM naive. The dispersal index (D) is equal to $1/\alpha$; note the log scale. (G) A_0 is the minimum patch area and is calculated as $\mu^{1/\chi}$. E_i (panels H and I) is the intrinsic extinction rate and is equal to (μ/A_i^χ). Other variables are defined in Table 1. Note that α, β, δ, μ, and χ are unitless parameters, and E_i is a probability.

slowly in the IFM naive than the IFM missing and IFM robust (Fig. 4A–C). The 95% credible intervals reveal considerable variation in the number of sites occupied. Lower bounds suggest imminent global extinction in all models. Upper bounds suggest the metapopulation may have reached quasi-stationary equilibrium in the IFM naive and IFM missing, while occupancy was still decreasing in year 100 in the IFM robust. Turnover rates were highest in the IFM naive and only somewhat higher in the IFM missing than in the IFM robust (Fig. 4D–F). Contrary to predictions, higher turnover rates did not lead to a higher proportion of simulations where global extinction occurred, and in fact, the IFM naive had fewer global extinctions (about 50%) than the

IFM missing and IFM robust (roughly 80% in both; Fig. 5A).

SPOMs for the Virginia Rail were highly variable in the IFM naive but less so in the IFM missing and IFM robust. Median occupancy was similar in the IFM naive (nearly 150 sites occupied at equilibrium; Fig. 4G) and IFM missing (approximately 170; Fig. 4H) but declined quickly to zero in the IFM robust (Fig. 4I). Occupancy quickly declined to zero at the 0.025 quantile in the IFM naive and IFM robust, yet was stable at approximately 140 sites in the IFM missing (Fig. 4G–I). Turnover rates in the IFM naive and IFM missing were higher than the IFM robust (Fig. 4J–L), which is consistent with the presence of false absences. However, these higher rates
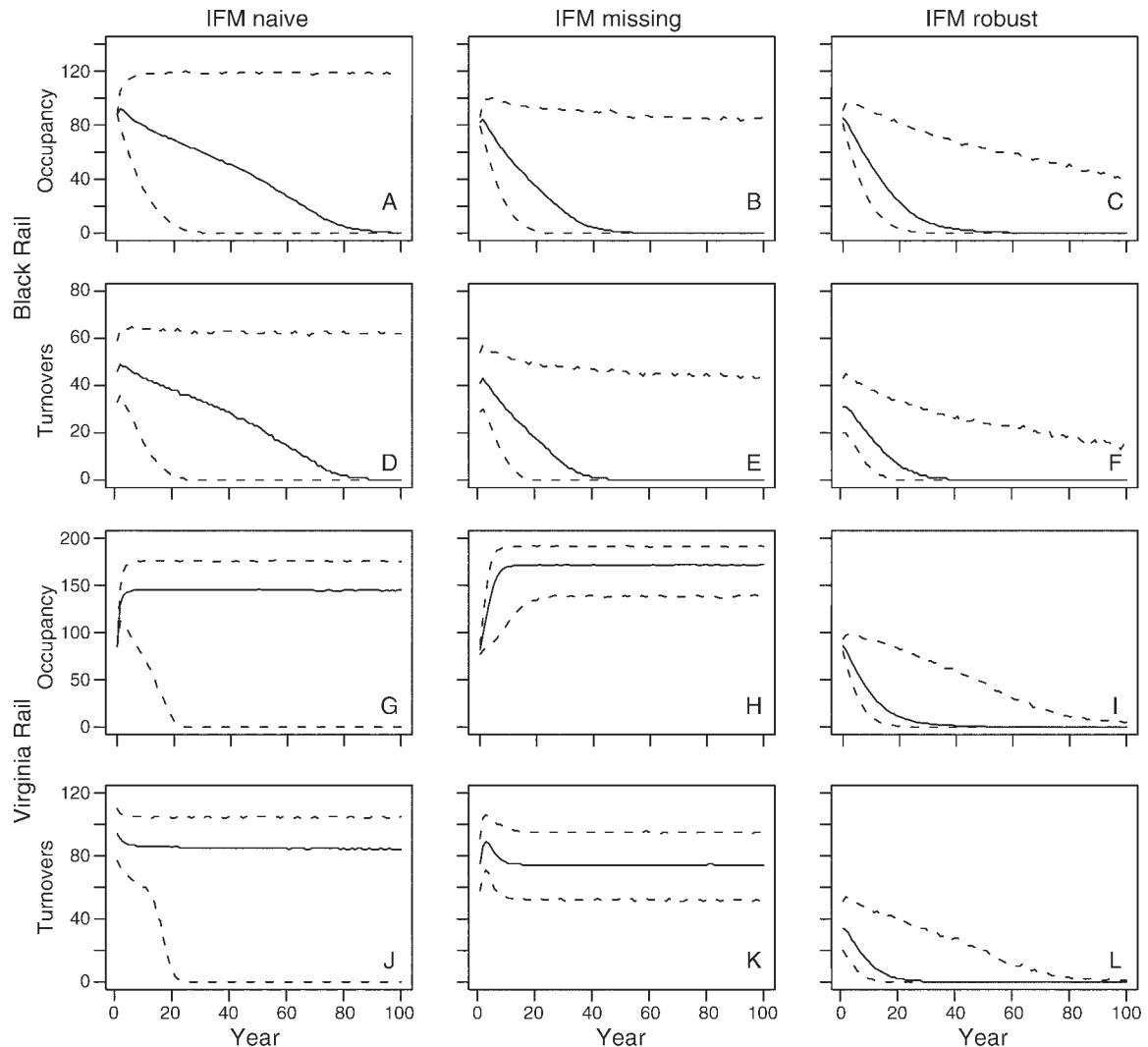
Fig. 4. Stochastic patch occupancy models for the Black and Virginia Rail. The median number of sites (solid lines) and 0.025 and 0.975 quantiles (dashed lines) occupied by the Black Rail (A–C) and Virginia Rail (G–I) estimated using the joint posterior of parameters for each model variant. Median number of turnovers (solid lines) and 0.025 and 0.975 quantiles (dashed lines) for the Black Rail (D–F) and Virginia Rail (J–L).

of turnover did not result in a higher proportion of global extinctions (Fig. 5B). At year 100, roughly 15% of simulations went extinct in the IFM naive, 0% in the IFM missing, and over 85% in the IFM robust.

DISCUSSION

The IFM has been a useful tool in understanding the dynamics of species existing in discrete habitat patches with barriers to movement (Hanski 1999). It uses relatively few data to make powerful predictions about the persistence of local populations and inform conservation plans (Drechsler et al. 2003, Bulman et al. 2007). However, applying the IFM to studies with imperfect detection and missing data can produce biased results (Moilanen 2002). Models accounting for false absences in occupancy were recently developed (MacKenzie et al.

2003), and many occupancy studies now incorporate imperfect detection (e.g., Pellet et al. 2007, Moritz et al. 2008, Nichols et al. 2008, Rizkalla et al. 2009, Kéry and Royle 2010). Yet, the robust design had not been applied to the IFM. In this study, we presented a robust formulation of the IFM that also estimates the full joint posterior for missing data, and we fit the model to two data sets.

In years with low detection rates, occupancy and turnover rates differed between the IFM robust and other models. Notably, detection of Black Rails was substantially lower in 2007 than other years, and this decrease in detection coincided with considerably lower occupancy (Fig. 2A, C). The reduction in detection rate may partly have been due to variation among observers, although there was some overlap in observers from 2006
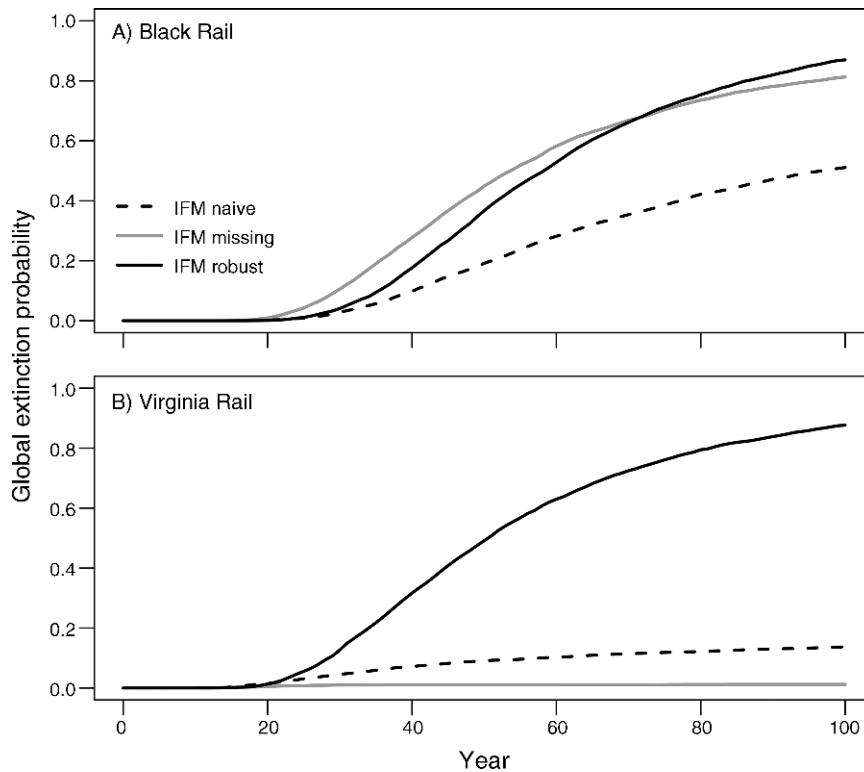
Fɪɢ. 5. Proportion of simulations with global extinction (0 sites occupied) estimated using the IFM naive (dashed line), IFM missing (solid gray line), and IFM robust (solid black line) for the (A) Black Rail and (B) Virginia Rail.

to 2008 and all received extensive training. It could also have been related to a depressed population size, where fewer individuals at a site would lead to a lower probability of detection. West Nile virus (WNV) antibodies were detected in blood samples collected from this population of Black Rails in the summer of 2007 (unpublished results) after the disease swept through the region, suggesting WNV may have caused the population decline. If the decrease in average occupancy were attributable to an ecological process not modeled by our formulation of the IFM, then the IFM robust could potentially underestimate turnovers in 2007 and 2008 and overestimate false absences in 2007. However, given the increase in detections in 2008, the average occupancy estimated by the IFM robust seems plausible. Virginia Rail detectability was relatively low in most years (Fig. 2B), resulting in large differences in occupancy and turnover rates between the IFM robust and other models (Fig. 2D, F). Detection rate also declined in 2007, as in Black Rails, but was particularly low in 2004 due to a difference in survey methodology (see *Study area and sampling methods*).

Both missing data and false absences affected IFM estimates of colonization parameters, and accounting for these forms of error can strengthen inference about dispersal ability. Notably, the median dispersal index decreased by 50% for the Black Rail when missing data and false absences were incorporated, whereas it was unaffected and large in the Virginia Rail, even after accounting for these sources of error (Fig. 3A, D). Moreover, estimates of the effect of patch area on the number of dispersers (β) were strongly affected by both missing data and false absences in both species (Fig. 3B). These results demonstrate that the effects of missing data and false absences can be difficult to predict, as evidenced by the differing responses of the two species. They also suggest that the Virginia Rail can disperse much greater distances than the Black Rail, which is consistent with differences in body mass and flight behavior of the two species (Paradis et al. 1998, Taylor 1998).

Estimates of the parameters of intrinsic extinction rate and their derived statistics were also affected by false absences but were not greatly influenced by missing data (Fig. 3E–I). The lower detection rates for the Virginia Rail indicated the presence of false absences (Fig. 2), and, consequently, IFM models that did not account for them overestimated extinction parameters (Fig. 3E, F). The Virginia Rail appeared more sensitive to area than the Black Rail in the IFM naive and IFM missing, but sensitivity was more similar when false absences and missing data were accounted for in the IFM robust (Fig. 3H, I). Although the credible intervals overlapped, minimum patch area was larger for the Virginia than for the Black Rail in the IFM robust, which is consistent

with the general pattern that larger species require larger territories (Schoener 1968).

Forecasts of occupancy dynamics from SPOMs were also sensitive to missing data and false absences. Projections that did not account for these sources of error (IFM naive or IFM missing) tended to be more optimistic than forecasts that did (IFM robust), despite the fact that the former models recorded more turnovers (Fig. 4). Projected persistence differed dramatically when missing data were incorporated for the Black Rail (Fig. 5A) and when false absences were estimated for the Virginia Rail (Fig. 5B).

Although there is increasing recognition that false absences bias occupancy models, there is little discussion of the effects of missing data. In the framework of MacKenzie et al. (2003, 2006), missing data are excluded from the model likelihood, which is justified on the grounds that missing data do not contribute to knowledge of the system. However, this is problematic for studies examining the effect of occupancy-based measures of connectivity. Missing data in the IFM cause the number of sites that contribute dispersers to be underestimated. In long-term studies where the amount of missing data is minimal, the average observed occupancy estimated from years with data may be a sufficient correction for missing data (Moilanen 2004). In the Black Rail data set, 12–51% of the sites were missing data (i.e., were not surveyed each year), which caused dispersal distance to be overestimated in the IFM naive (Fig. 3D). Notably, the IFM missing and robust still did not account for all sites, as on a few occasions we were unable to gain permission to visit a site from private landowners. Future studies could extend the hierarchical models to missing sites that have never been visited but for which information on area and location are available, e.g., by using satellite imagery. Simulation studies could examine the ability of the models to impute missing values for sites that have never been visited to determine whether the data on covariates provides sufficient information for estimating the probability of occupancy.

This study focused on the incidence function model, but the issues of missing data and false detections in explanatory variables could also be incorporated into other types of occupancy models. Inferences from studies using connectivity metrics may be especially sensitive to false absences and missing data. A meta-analysis of 1015 population networks found that a landscape connectivity measure (distance to nearest neighbor) was a poorer predictor of occupancy than area, but that an occupancy-based connectivity measure (distance to nearest occupied neighbor) and area had similar predictive power, although the majority of variation in occupancy was still unexplained (Prugh et al. 2008). Incorporating a robust measure of connectivity could explain a larger amount of the variation in occupancy. Using a hierarchical Bayesian framework, we calculated a connectivity metric from the full joint posterior of the latent states to more accurately estimate connectivity. These methods could be extended to logistic regression models (e.g., Program Mark, Program Presence). Hierarchical models that incorporate false absences and missing data in connectivity metrics may better characterize connectivity and ultimately improve our understanding of the factors affecting species distributions in patchy landscapes.

LITERATURE CITED

Bulman, C. R., R. J. Wilson, A. R. Holt, L. G. Bravo, R. I. Early, M. S. Warren, and C. D. Thomas. 2007. Minimum viable metapopulation size, extinction debt, and the conservation of a declining species. Ecological Applications 17:1460–1473.

Conway, C. J., and W. R. Eddleman. 1995. Virginia Rail. The birds of North America. Volume 173. Cornell Lab of Ornithology, Ithaca, New York, USA.

Cowardin, L. M., V. Carter, F. C. Golet, and E. T. LaRoe. 1979. Classification of wetlands and deepwater habitats of the United States. U.S. Department of the Interior, Fish and Wildlife Service, Washington, D.C., USA.

Cowles, M. K., and B. P. Carlin. 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association 91:883–904.

de Valpine, P. 2004. Monte Carlo state-space likelihoods by weighted posterior kernel density estimation. Journal of the American Statistical Association 99:523–536.

Drechsler, M., K. Frank, I. Hanski, R. B. O'Hara, and C. Wissel. 2003. Ranking metapopulation extinction risk: from patterns in data to conservation management decisions. Ecological Applications 13:990–998.

Eddleman, W. R., R. E. Flores, and M. L. Legare. 1994. Black Rail. The Birds of North America. Volume 123. Cornell Lab of Ornithology, Ithaca, New York, USA.

Gelman, A., J. Carlin, H. Stern, and D. Rubin. 2004. Bayesian data analysis. Second edition. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter, editors. 1996. Markov chain Monte Carlo in practice. Chapman and Hall, Washington, D.C., USA.

Hanski, I. 1994. A practical model of metapopulation dynamics. Journal of Animal Ecology 63:151–162.

Hanski, I. 1998. Metapopulation dynamics. Nature 396:41–49.

Hanski, I. 1999. Metapopulation ecology. Oxford University Press, New York, New York, USA.

Kéry, M., and J. A. Royle. 2010. Hierarchical modelling and estimation of abundance and population trends in metapopulation designs. Journal of Animal Ecology 79:453–461.

Legare, M. L., and W. R. Eddleman. 2001. Home range size, nest-site selection and nesting success of Black Rails in Florida. Journal of Field Ornithology 72:170–177.

Lele, S., B. Dennis, and F. Lutscher. 2007. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecology Letters 10:551–563.

MacKenzie, D. I., J. D. Nichols, J. E. Hines, M. G. Knutson, and A. B. Franklin. 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology 84:2200–2207.

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines. 2006. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Academic Press, Burlington, Massachusetts, USA.

Moilanen, A. 1999. Patch occupancy models of metapopulation dynamics: efficient parameter estimation using implicit statistical inference. Ecology 80:1031–1043.

Moilanen, A. 2002. Implications of empirical data quality to metapopulation model parameter estimation and application. Oikos 96:516.

Moilanen, A. 2004. SPOMSIM: software for stochastic patch occupancy models of metapopulation dynamics. Ecological Modelling 179:533–550.

Moilanen, A., I. Hanski, and A. T. Smith. 1998. Long-term dynamics in a metapopulation of the American pika. American Naturalist 152:530–542.

Moritz, C., J. L. Patton, C. J. Conroy, J. L. Parra, G. C. White, and S. R. Beissinger. 2008. Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. Science 322:261–264.

Nichols, J. D., L. L. Bailey, A. F. O'Connell, Jr., N. W. Talancy, E. H. Campbell Grant, A. T. Gilbert, E. M. Annand, T. P. Husband, and J. E. Hines. 2008. Multi-scale occupancy estimation and modelling using multiple detection methods. Journal of Applied Ecology 45:1321–1329.

O'Hara, R. B., E. Arjas, H. Toivonen, and I. Hanski. 2002. Bayesian analysis of metapopulation data. Ecology 83:2408–2415.

Ozgul, A., K. B. Armitage, D. T. Blumstein, D. H. Vanvuren, and M. K. Oli. 2006. Effects of patch quality and network structure on patch occupancy dynamics of a yellow-bellied marmot metapopulation. Journal of Animal Ecology 75:191–202.

Paradis, E., S. R. Baillie, W. J. Sutherland, and R. D. Gregory. 1998. Patterns of natal and breeding dispersal in birds. Journal of Animal Ecology 67:518–536.

Pellet, J., E. Fleishman, D. S. Dobkin, A. Gander, and D. D. Murphy. 2007. An empirical evaluation of the area and isolation paradigm of metapopulation dynamics. Biological Conservation 136:483–495.

Pollock, K. H. 1982. A capture-recapture design robust to unequal probability of capture. Journal of Wildlife Management 46:752–757.

Prugh, L. R., K. E. Hodges, A. R. E. Sinclair, and J. S. Brashares. 2008. Effect of habitat area and isolation on fragmented animal populations. Proceedings of the National Academy of Sciences USA 105:20770–20775.

R Development Core Team. 2010. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ⟨www.r-project.org⟩

Richmond, O. M., S. K. Chen, B. B. Risk, J. Tecklin, and S. R. Beissinger. 2010a. California black rails depend on irrigation-fed wetlands in the Sierra Nevada foothills. California Agriculture 64:85–93.

Richmond, O. M., J. E. Hines, and S. R. Beissinger. 2010b. Two species occupancy models: a new parameterization applied to co-occurrence of secretive rails. Ecological Applications 20:xx–xx.

Richmond, O. M., J. Tecklin, and S. R. Beissinger. 2008. Distribution of California Black Rails in the Sierra Nevada foothills. Journal of Field Ornithology 79:381–390.

Rizkalla, C. E., J. E. Moore, and R. K. Swihart. 2009. Modeling patch occupancy: relative performance of ecologically scaled landscape indices. Landscape Ecology 24:77–88.

Royle, J. A., and R. M. Dorazio. 2008. Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations, and communities. Academic Press, Amsterdam, The Netherlands.

Royle, J. A., and M. Kery. 2007. A Bayesian state-space formulation of dynamic occupancy models. Ecology 88:1813–1823.

Schoener, T. W. 1968. Sizes of feeding territories among birds. Ecology 49:123–141.

Taylor, B. 1998. Rails: a guide to the rails, crakes, gallinules and coots of the world. Yale University Press, New Haven, Connecticut, USA.

ter Braak, C. J. F., and R. S. Etienne. 2003. Improved Bayesian analysis of metapopulation data with an application to a tree frog metapopulation. Ecology 84:231–241.

APPENDIX A

Metropolis-Hastings sampling for the incidence function models (*Ecological Archives* E092-040-A1).

APPENDIX B

Posterior densities and model diagnostics for the incidence function models (*Ecological Archives* E092-040-A2).

SUPPLEMENT

R-code used to simulate occupancy data and to fit the IFM naive, IFM missing, and IFM robust models (*Ecological Archives* E092-040-S1).