



HAL
open science

FlywayNet: A hidden semi-Markov model for inferring the structure of migratory bird networks from count data

Sam Nicol, Marie-Josée Cros, Nathalie Peyrard, Régis Sabbadin, Ronan Trépos, Richard A Fuller, Bradley K Woodworth

► **To cite this version:**

Sam Nicol, Marie-Josée Cros, Nathalie Peyrard, Régis Sabbadin, Ronan Trépos, et al.. FlywayNet: A hidden semi-Markov model for inferring the structure of migratory bird networks from count data. 2022. hal-03800886

HAL Id: hal-03800886

<https://hal.inrae.fr/hal-03800886>

Preprint submitted on 6 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FlywayNet: A hidden semi-Markov model for inferring the structure of migratory bird networks from count data

Sam Nicol ^{☆a,1,*}, Marie-Josée Cros ^{☆b,2}, Nathalie Peyrard ^{☆b,3}, Régis Sabbadin ^{☆b,4}, Ronan Trépos ^{☆b,5}, Richard A. Fuller^{c,7}, Bradley K. Woodworth ^{c,6}

^a*CSIRO Land and Water, Dutton Park QLD, Australia*

^b*INRAE, UR MIAT, F-31320 Castanet-Tolosan, France*

^c*School of Biological Sciences, The University of Queensland, St. Lucia, QLD, Australia*

Abstract

1. Every year, millions of birds migrate between breeding and non-breeding habitat, but the relative numbers of animals moving between sites is difficult to observe directly.
2. Here we propose FlywayNet, a discrete network model based on observed count data, to determine the most likely migration links between regions using statistical modelling and efficient inference tools. Our approach advances on previous studies by accounting for noisy observations and flexible stopover durations by modelling using interacting hidden semi-Markov Models. In FlywayNet, individual birds sojourn in stopover nodes for a period of time before moving to other nodes with an unknown probability that we aim to estimate. Exact estimation using existing approaches is not possible, so we designed customised versions of the Monte Carlo

[☆]These authors contributed equally to this study

*Corresponding author

Email address: sam.nicol@csiro.au (Sam Nicol [☆])

¹ORCID: <https://orcid.org/0000-0002-1160-7444>

²ORCID: <https://orcid.org/0000-0002-6395-5563>

³ORCID: <https://orcid.org/0000-0002-0356-1255>

⁴ORCID: <https://orcid.org/0000-0002-6286-1821>

⁵ORCID: <https://orcid.org/0000-0002-3338-9337>

⁶ORCID: <https://orcid.org/0000-0002-4528-8250>

⁷ORCID: <https://orcid.org/0000-0001-9468-9678>

Expectation-Maximisation and Approximate Bayesian Computation algorithms for our model. We compare the efficiency and quality of estimation of these approaches on synthetic data and an applied case study.

3. Our algorithms performed well on benchmark problems, with low absolute error and strong correlation between estimated and known parameters. On a case study using citizen science count data of the Far Eastern Curlew (*Numenius madagascariensis*), an endangered shorebird from the East Asian-Australasian Flyway, the ABC and MCEM algorithms generated contrasting recommendations due to a difference in optimisation criteria and noise in the data. For ABC, we recovered key features of population level movements predicted by experts despite the challenges of noisy unstructured data.
4. Understanding connectivity places local conservation efforts and threat mitigation in the global context, yet it has proven difficult to rigorously quantify connectivity at the population level. Our approach provides a flexible framework to infer the structure of migratory networks in birds and other organisms.

Résumé

1. Chaque année, des millions d'oiseaux migrent vers et depuis leurs sites de reproduction en empruntant différentes routes de migration. Il est difficile d'estimer comment les oiseaux se répartissent entre ces différentes routes.
2. Nous proposons FlywayNet, un modèle de réseau de migration basé sur des données de comptage. Le modèle permet de reconstruire les routes de migration les plus probables en utilisant la modélisation statistique et des outils d'inférence efficaces. Notre approche améliore les travaux existants en permettant la prise en compte d'observations bruitées et de durées de séjour flexibles dans les sites-étape de la migration, à l'aide de modèles semi-Markoviens cachés, couplés. Dans FlywayNet, chaque oiseau séjourne dans des sites-étape pendant un certain temps, avant de s'envoler vers un nouveau site parmi plusieurs sites possibles. Nous util-

isons les comptage observés afin d’estimer les lois de probabilité des temps de séjour d’un oiseau dans chaque site et la probabilité de rejoindre chacun des sites suivants. Une estimation exacte par des méthodes classiques est trop complexe, aussi nous avons construit des versions adaptées des algorithmes Monte-Carlo Expectation-Maximization (MCEM) et Approximate Bayesian Computation (ABC) permettant une estimation approchée des paramètres du modèle FlywayNet. Nous proposons une comparaison empirique de l’efficacité et de la qualité d’estimation de ces deux algorithmes, sur des données synthétiques et des données issues d’un cas d’étude.

3. Les deux algorithmes montrent de bonnes performances sur les données synthétiques, avec une faible erreur d’estimation et une forte corrélation entre les paramètres estimés et réels des modèles. Nous avons également considéré un cas d’étude utilisant des données d’observations citoyennes concernant le Courlis de Sibérie (*Numenius madagascariensis*), espèce de rivage menacée migrant le long des rivages entre la Sibérie et l’Océanie. Sur ce cas d’étude, les approches ABC et MCEM donnent des résultats contrastés, liés à la rareté des données d’observations et aux critères optimisés par les deux approches. La méthode ABC a toutefois permis de retrouver les mouvements d’oiseaux prédits par les experts, malgré le défi posé par le manque de données et le bruit les entachant.
4. La compréhension des connectivités entre sites-étape permet de replacer les efforts de conservation et de réduction des menaces locales dans un contexte global de conservation. Néanmoins, il est difficile d’estimer les liens de connectivité entre sites pour des populations entières d’oiseaux. Notre approche fournit un moyen flexible (basé sur des données participatives d’observations) d’inférer la structure d’un réseau de migration à l’échelle globale, pour des espèces aviaires, ou autres...

Keywords: Approximate Bayesian Computation, connectivity, East Asian-Australasian flyway, eBird, Hidden Semi Markov Model, Monte Carlo

1 **1. Introduction**

2 The seasonal migration of animals around our planet is one of Earth's great
3 natural spectacles. Apart from the inspiration humans draw from the endurance
4 of migrants who undertake such arduous journeys across our world, migration
5 is critical for connecting ecosystem processes and services across vast distances
6 (Semmens et al. 2011, Wilcove & Wikelski 2008). Sadly, the phenomenon of mi-
7 gration is threatened and many formerly abundant species are declining globally
8 (Clemens et al. 2016, Rappole & McDonald 1994, Robbins et al. 1989, Wilcove
9 & Wikelski 2008).

10 As well as understanding the drivers of decline, arresting declines in migra-
11 tory species requires knowing the degree of geographic linkage between different
12 stages of a species' annual range due to the movement trajectories of individ-
13 uals as they complete their migration, referred to as migratory connectivity
14 (Marra et al. 2019, Webster et al. 2002). Connectivity determines how changes
15 in habitat at one part of a migratory network may influence others. For exam-
16 ple, connectivity can explain how poor non-breeding habitat quality will impact
17 the breeding population (Silllett et al. 2000), the disproportionate impacts of the
18 loss of migratory structure on population size (Iwamura et al. 2013, Runge et al.
19 2014), or how disease is likely to spread through a migration network (Webster
20 et al. 2002). Understanding connectivity places local conservation efforts and
21 threat mitigation in the global context. For example, if we know the main routes
22 travelled by populations, we can prioritise management of threats in the parts of
23 the flyway that are critical habitat for the largest number of migrants. Connec-
24 tivity should also be the basis for informed reserve design for migratory species,
25 yet recent analysis suggests that existing reserve networks rarely account for
26 connectivity of migratory bird species across their annual cycle (Runge et al.
27 2015).

28 Measuring migratory connectivity is challenging due to the wide geographic

29 areas and vast numbers of individuals involved (Webster et al. 2002). Great
30 progress has been made in recent decades, with sophisticated advances in tradi-
31 tional mark-recapture/ banding studies (Cohen et al. 2014), improved satellite
32 tracking and geolocator technology, stable isotope analysis and genetic tech-
33 niques all providing alternative ways to learn more about where individuals
34 move (Webster et al. 2002). Accompanying these advances is an extensive liter-
35 ature on movement ecology, including models for analysing tracking data, which
36 we do not attempt to review here. Of particular relevance to our study are the
37 works of Joo et al. (2013), which used a hidden semi-Markov Model and track-
38 ing data to model foraging behaviour, and Kölzsch et al. (2018) which used
39 tracking data to infer a migratory network structure. Despite these powerful
40 methods and the increasingly clever ways that they are being combined, the
41 expense and low scalability of tracking individuals (Webster et al. 2002) means
42 that in most species our understanding of migratory routes is still drawn from
43 a tiny subsample of the population, often just a few individuals of any given
44 species.

45 A complementary approach to tracking individuals is to infer connectivity
46 from count data at known aggregation sites. Count data, particularly for birds
47 using the citizen science database eBird (Sullivan et al. 2009), have been used to
48 complement and boost inferences about connectivity from other methods such
49 as geolocator data (Hallworth et al. 2015) and stable isotope analysis (Fournier
50 et al. 2017). Unlike most tracking data and banding data, count data has the
51 tremendous advantage of being widely and freely available, at least for birds,
52 but increasingly for other organisms (e.g. Tonachella et al. 2012). There is an
53 opportunity to develop more powerful estimation techniques that make greater
54 use of count data in its own right.

55 Models of migratory networks have been posed to explore theoretical prop-
56 erties of migratory structure using assumed parameters (Taylor & Norris 2010),
57 and others have used tracking data to infer structure from a few individuals
58 (Kölzsch et al. 2018). However, until recently there have been relatively few at-
59 tempts to infer migratory network structure using count data, but a handful of

60 studies have attempted to solve the problem for migratory bird networks. The
61 most relevant to our study, (Sheldon et al. 2007, Jain & Dilkina 2015), infer
62 Markov transition probabilities for a migratory network from eBird data but
63 make the unrealistic assumption of perfect observations. Our study advances
64 previous attempts because it incorporates imperfect detection (i.e. it allows for
65 error in the observed counts) and explicitly estimates the time spent at stopover
66 locations.

67 As in previous models of migratory networks (e.g. Kölzsch et al. (2018), Jain
68 & Dilkina (2015), Taylor & Norris (2010)), we model a migratory system as a
69 network consisting of nodes (breeding nodes, non-breeding nodes and stopover
70 nodes) connected by edges. Individuals sojourn in stopover nodes for a period of
71 time before moving to other nodes with an unknown probability that we aim to
72 estimate. From the set of estimated transition probabilities we can reconstruct
73 a weighted network which represents connections between stopovers and their
74 relative strength. The model also enables us to estimate the mean duration of
75 a bird’s sojourn at each stopover (hereafter ‘sojourn time’).

76 Since animals are difficult to count precisely, to estimate the characteris-
77 tics of the network, we introduce a hidden semi-Markov Modelling (HSMM, Yu
78 (2010), Joo et al. (2013)) approach to model imperfectly detected count data.
79 The hidden part of the model is the position of each bird at each time step. The
80 HSMM is an extension of the well-known hidden Markov Model (HMM). The
81 HMM assumes that the sojourn time in a given hidden state follows a geometric
82 distribution; extending the HMM to the HSMM relaxes this assumption and
83 allows explicit modelling of sojourn times. The geometric distribution assumes
84 that the most probable sojourn time is always 1 time unit, which is a limit-
85 ing assumption for birds that may spend a few weeks resting at sites before
86 continuing their migration. To circumvent this limitation we use a HSMM.

87 Due to the dimension of the hidden variables, exact estimation of the model
88 parameters using classical approaches is not feasible for even a small number of
89 nodes. To overcome this, we present two dedicated estimation algorithms for our
90 model: Monte Carlo Expectation-Maximisation (MCEM, Wei & Tanner (1990))

91 and Approximate Bayesian Computation (ABC, Csilléry et al. (2010)). We
92 present and compare the efficiency and quality of estimation of these approaches
93 on synthetic data before applying them to a case study of a migratory shorebird
94 in the East Asian-Australasian flyway.

95 **2. Methods**

96 *2.1. The model of the migratory system*

97 We assume that we are following the migration of a population of N birds
98 over a set of I distinct sites (i.e. breeding, non-breeding or stopover locations)
99 over time. Sites are connected via migration links ('edges' in the following) for
100 which we have some a priori knowledge, however we do not know the strength
101 of the connections, and our goal is to learn the most likely structure from count
102 data. We name our model 'FlywayNet'.

103 *2.1.1. A priori knowledge of the migratory network*

104 We introduce some *a priori* knowledge on the presence or absence of an edge
105 between two sites. First, since migration is a directed movement (from North
106 to South or from South to North, depending on the season) we assume that
107 birds do not fly backward. Although it is known that some birds do terminate
108 migration and return to their place of origin (e.g. Driscoll & Ueta (2002)), the
109 number of birds returning to their origin sites is very small compared to the
110 number completing their migration. So we assume an ordering of the I sites
111 such that if $i < j$ then a bird cannot fly from site j to site i . The set of all
112 potential connections is given by the set of oriented edges from i to j for every
113 $i < j$. This assumption ensures that the graph is acyclic, simplifying the model
114 estimation.

115 *2.1.2. Semi-Markov model of bird migration*

116 We consider that each bird trajectory is modeled as a semi-Markov model
117 over a finite discrete time horizon $H = \{0, 1, 2, \dots, T\}$, and that the N bird
118 trajectories are independent. The state of a trajectory at a given time can be

119 one of the I sites, or the state ‘death’ which corresponds to a bird who dies before
 120 time T . Rigorously, to have a semi-Markov model, one should add the states
 121 corresponding to a bird flying towards a given site. Since flight durations are
 122 known and fixed, for sake of simplicity, we do not burden the model description
 123 with these extra states.

For bird n ($1 \leq n \leq N$), the trajectory π_n can be summarized by the sequence of visited states and the time of arrival in the state:

$$\pi_n = \left((i_0^n, t_0^n), (i_1^n, t_1^n), \dots, (i_{F_n}^n, t_{F_n}^n) \right). \quad (1)$$

124 In expression (1), trajectory π_n has F_n stages, bird n starts in site i_0^n at time
 125 $t_0^n = 0$, and t_k^n is the date of arrival of bird n at site i_k^n , for every $1 \leq k \leq F_n$.
 126 By convention, $i_{F_n}^n$ is the last state occupied by bird n , i.e. the bird entered
 127 state $i_{F_n}^n$ at time $t_{F_n}^n \leq T$ and is still in this state at time T .

128 If for some bird n , we have $i_k^n = \text{“death”}$, then $t_k^n < T$ represents the date
 129 of death of bird n . In this case, π_n is stopped at t_k^n .

130 We assume that for every pair of sites (i, j) such that $i < j \leq I$, the flight
 131 duration between i and j , $f_{i,j} \in \{1, 2, \dots\}$ is known. Under these hypotheses,
 132 expression (1) defines a unique bird trajectory.

133 Then, the semi-Markov model for a bird’s trajectory is defined as follows:

- 134 • Transition probabilities between states. We define the $I \times I$ matrix R
 135 of transition probabilities between states that are sites. The probability
 136 that any bird leaving site i at any given time goes to site j is $R(i, j)$. If
 137 $i \geq j$ then $R(i, j) = 0$, so R is an upper triangular matrix. Note that,
 138 accounting for mortality, we may have, for any $i < I$,

$$\sum_{j \in 1 \dots I} R(i, j) = \sum_{j=i+1}^I R(i, j) < 1.$$

139 The value $\mu_i = 1 - \sum_{j \in 1 \dots I} R(i, j)$, for $i < I$ is the mortality probability in
 140 site i which is assumed known. For a bird leaving site $i < I$, the destina-
 141 tion is thus selected according to a categorical distribution of parameters

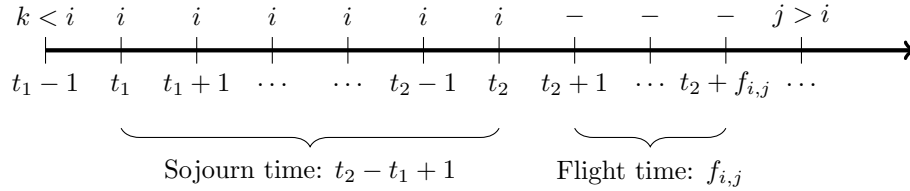


Figure 1: Sojourn time definition. The arrow represents time evolution, and the discrete times are indicated below the arrow. The state of the bird trajectory is indicated above the arrow, with k, i and j being three distinct sites and ‘-’ coding for a bird flying. In this example, the duration of sojourn in state i is $t_2 - t_1 + 1$.

142 $(R(i, i + 1), \dots, R(i, I), \mu_i)$. For the breeding site I , we assume that when
 143 a bird ‘leaves’ site I , it necessarily moves to death so $\mu_I = 1$. This as-
 144 sumption has no influence on the estimation, since the breeding node is
 145 the terminal node and we do not estimate sojourn time or transitions from
 146 it.

- 147 • Sojourn time. We assume that the sojourn time distribution in state $i \leq I$
 148 is a shifted Poisson distribution of parameter λ_i . The shift is equal to one
 149 to ensure that sojourn time is larger than 0 (as is done in the R package
 150 MHSM, O’Connell & Hojsgaard (2011)). Thus the probability that the
 151 sojourn time τ_i in site i is equal to d is:

$$P_{\lambda_i}(\tau_i = d) = \frac{(\lambda_i)^{d-1}}{(d-1)!} e^{-\lambda_i}, \forall d = 1, 2, \dots$$

152 The sojourn time in state ‘death’ is infinite (‘death’ is an absorbing state).
 153 The definition for a sojourn of duration d is the following: if the state of
 154 bird n ’s trajectory at time t is i for $t = t_1, t_1 + 1, \dots, t_2$ and is not state
 155 i at $t_1 - 1$ and at $t_2 + 1$, then $d = t_2 - t_1 + 1$ (see Figure 1). Sojourn
 156 time distributions depend on the site, but are the same for each bird.
 157 Furthermore, we assume that sojourn times of two sites are independent.

- 158 • Initial distribution. γ_i^0 is the distribution of the number of birds at site i at
 159 time zero. Rigorously, in the semi-Markov model framework, if we wanted

160 to track the position of each individual bird, we would need to define
 161 a separate probability distribution for the initial position of each bird.
 162 However, since our model does not distinguish between birds, it is sufficient
 163 to summarise these individual distributions with a single distribution for
 164 site i , γ_i^0 . In this study we assume that the initial distribution is known.

165 These notions are formally defined in Supporting information S1.

166 2.1.3. Observation model

167 Given the set of trajectories $\Pi = \{\pi_1, \dots, \pi_n\}$, we can determine N_i^t , the
 168 number of birds located in site i at time t (this variable is formally defined in
 169 Supporting Information S1).

170 Observations are observed counts O_i^t for a set $\Omega \subseteq \{1 \dots I\} \times \{1 \dots T\}$ of
 171 observed site-times. We will consider that, conditional on the birds' trajectories,
 172 these counts are independent. Furthermore the distribution of O_i^t (for $(i, t) \in \Omega$)
 173 conditional on Π is equal to the distribution of O_i^t conditional on N_i^t . We
 174 model it as a negative binomial distribution with parameters r_i^t and p where
 175 $r_i^t = \delta N_i^t p / (1 - p)$ and δ is the probability to report a bird (see Supporting
 176 Information S2 for details and also for other choices of observation model).

177 Since the observed counts may not be discrete (e.g. where they are averaged
 178 across several observers), observations are rounded to the nearest integer value.

179 The joint distribution of all the observations $O = \{O_i^t\}_{(i,t) \in \Omega}$ given the tra-
 180 jectories is $P(O|\Pi) = \prod_{(i,t) \in \Omega} \mathcal{NB}_{(r_i^t, p)}(O_i^t)$, where $\mathcal{NB}_{(r_i^t, p)}(\cdot)$ is the negative
 181 binomial distribution.

182 2.2. Parameter interpretation and estimation

183 Let us denote $\Lambda = (R, \{\lambda_i\}_{1 \leq i \leq I}, \delta)$ the set of parameters of the HSMM
 184 model that we would like to estimate. For the negative binomial observation
 185 model, we also estimate parameter p from data, but prior to the joint estimation
 186 of Λ , see Supplementary Information S5.

187 Since we want to infer the most likely network of migration links between
 188 nodes, a parameter of interest is the matrix R of probabilities of transitions

189 between sites. From this matrix, we can build a weighted migratory network
190 where there is an edge from site i toward site j if $R(i, j) > 0$. The weight
191 of the edge is $R(i, j)$. For a fixed i , the non-zero $R(i, j)$'s provide the relative
192 importance of the routes $i \rightarrow j$. The parameter λ_i indicates the expected
193 duration that a bird stays at site i .

194 Estimating the model parameters is difficult for several reasons. First, this
195 is a model with hidden data: neither the individual bird trajectories nor the
196 real counts N_i^t are observed. Second, conditional on the observed counts, the
197 N bird trajectories are no longer independent.

198 Direct optimization of the likelihood is intractable, yet realisations of the
199 O_i^t from the model are easy to simulate. Indeed, given parameters Λ we can
200 first simulate each bird's trajectory, then compute the N_i^t and finally simulate
201 each O_i^t . We designed two simulation-based methods to estimate the param-
202 eters, based on the Monte Carlo Expectation-Maximisation method (MCEM,
203 Andrieu et al. (2003)) and the Approximate Bayesian Computation method
204 (ABC, Csilléry et al. (2010)) respectively. With MCEM, we obtain a point-
205 wise estimate for each model parameter (frequentist approach) while with ABC
206 we obtain an approximation of the posterior distribution of each parameter
207 (Bayesian approach). The reason to design two estimation algorithms, from dif-
208 ferent approaches (i.e. frequentist and Bayesian) and with different optimisation
209 criteria, is to help to diagnose the confidence we can have in the estimated pa-
210 rameters, i.e. if the algorithms find different parameter values then this should
211 prompt further investigation to understand the cause of the difference.

212 *2.2.1. Monte Carlo Expectation Maximisation*

213 The Expectation-Maximisation (EM) algorithm is an iterative algorithm
214 that maximises the likelihood of the observed data when hidden variables pre-
215 clude the use of direct maximisation of the likelihood. For our model, for a
216 current value of the estimated parameters, Λ_{old} , in the E-step, the conditional
217 probabilities $P_{\Lambda_{old}}(\Pi|O)$ are computed for all possible sets of trajectories Π .
218 Then in the M-step, the parameter estimator is updated to Λ_{new} :

$$\Lambda_{new} = \arg \max_{\Lambda} \sum_{\Pi} \log (P_{\Lambda}(\Pi, O)) P_{\Lambda_{old}}(\Pi|O),$$

219 where the sum is taken over every possible sets Π of N independent trajecto-
 220 ries. So computing Λ_{new} requires the evaluation of the distribution $P_{\Lambda_{old}}(\Pi|O)$
 221 which is too complex (E-step). It is possible to approximate the updating for-
 222 mulas using Monte-Carlo techniques (Andrieu et al. 2003, Levine & Casella
 223 2001) by drawing many samples from $P_{\Lambda_{old}}(\Pi|O)$. The corresponding updating
 224 formulas are equations 6 and 7 in Supporting information S2.

225 The challenging part of the Monte-Carlo Expectation-Maximisation (MCEM)
 226 approach is therefore to draw samples from $P_{\Lambda_{old}}(\Pi|O)$. To do this, we used
 227 a Metropolis-Hastings algorithm (Hastings 1970). This approach, as well as a
 228 more complete presentation of the MCEM algorithm, is described in detail in
 229 Supporting Information S2.

230 2.2.2. Approximate Bayesian Computation

231 The idea of an ABC algorithm (Csilléry et al. 2010, Jabot et al. 2013) is to
 232 generate parameter values Λ from proposed prior distributions (or in our case,
 233 a particle filter, since we use a more complex version of ABC; see Supporting
 234 Information S3), then to generate observations O_{Λ} for these values of the model
 235 parameters. If the simulated observations O_{Λ} are close to the true observation
 236 O then the parameter values Λ are accepted. The procedure is repeated a large
 237 number of times. The histogram of the set of accepted values is then used as
 238 an approximation of the true posterior distribution $P(\Lambda|O)$.

239 We used the Lenormand sequential sampling method of the EASYABC pack-
 240 age in R (Jabot et al. 2015) to obtain the posterior distribution of every param-
 241 eter of the model. We selected the set of all observed counts, O , as the summary
 242 statistics. Further details of the ABC algorithm, including the particle filtering
 243 algorithm for drawing candidate parameter values, are included in Supporting
 244 Information S3.

245 *2.3. Benchmarking*

246 The performance of the MCEM and ABC algorithms were assessed by esti-
247 mating the model parameters from data simulated from the HSMM model for
248 several networks with known values of Λ . In these experiments the parameter
249 p of the negative binomial distribution was not estimated but was fixed to its
250 true value. Because the parameters of these benchmark networks are known a
251 priori, we can test the performance of MCEM and ABC by comparing how well
252 they recover the transition probabilities, the sojourn times and the reporting
253 probability given different numbers of nodes and network structures.

254 The network structure tested during these experiments varied depending on
255 the number of nodes (4-10 nodes) and the maximum number of outgoing edges
256 per node (2-4 outgoing edges/node, where outgoing edges are departure routes
257 from a node). We used a range of [1, 3] for generating sojourn times. Five sets of
258 parameter values (transition probabilities and sojourn times) were generated for
259 each structure. The total population of birds was 10000 and it was distributed
260 equally over the set of nodes that have no incoming edge (source nodes). The
261 number of parameters to estimate ranged from 6-20 depending on the problem
262 structure. The total number of generated problems in this benchmark was
263 300. As well as varying the network structure, we tested the effect of missing
264 observation data on parameter estimation by removing varying proportions of
265 the observed data and re-estimating parameter values.

266 Performance was assessed quantitatively by comparing the log likelihood
267 of estimated parameters and the mean absolute error (*meanAE*) of estimated
268 parameters rescaled into [0, 1]. Computation of log likelihood as well as the
269 *meanAE* are detailed in the Supporting information S4. ABC provides an esti-
270 mate of the posterior distribution of each parameter, so to compute *meanAE* we
271 extracted point estimates from these distributions. Point estimates were repre-
272 sented using the mode of each distribution (the mean being less representative,
273 in particular for non symmetric distributions). We compared several methods to
274 estimate the mode of a distribution. Among them, the Lientz function (Lientz
275 (1972)) and the Venter method (Venter (1967)) (both with bandwidth 0.2) re-

276 turned similar results and led to the lowest *meanAE* values. We selected the
277 Venter method because the bandwidth parameter is easier to interpret.

278 Wilcoxon tests (Wilcoxon 1945) were performed on the *meanAE* and log
279 likelihoods obtained for each of the benchmark networks. The tests compared
280 the differences between the results of ABC and MCEM algorithms. The Wilcoxon
281 method was used because we did not want to make any assumption on the dis-
282 tribution of the differences and the pairing option was chosen when we could
283 focus on differences within benchmark problems. Common notations were used
284 when displaying the p values computed by the test, i.e.: *ns* if $p > 0.05$, * if
285 $p \leq 0.05$, ** if $p \leq 0.01$, *** if $p \leq 0.001$ and **** if $p \leq 0.0001$.

286 2.4. Case Study: Eastern Curlews in the East Asian-Australasian flyway

287 We applied our model to infer the northward migration of the Eastern Curlew
288 (*Numenius madagascariensis*) population in the East Asian-Australasian Fly-
289 way (EAAF). Eastern Curlews are the largest migratory shorebirds in the
290 world, making an annual migration from their breeding grounds in Siberia and
291 Kamchatka through east Asia to their predominantly Australian non-breeding
292 grounds, before returning to breed. Approximately 80% of the population is es-
293 timated to utilise the Yellow Sea during the northern migration Department of
294 the Environment (2015), making the Yellow Sea a critically important stopover
295 site for the species.

296 The global population of Eastern Curlews was estimated to be 32000 birds
297 in 2021 (Wetlands International 2021). The population is declining at a rate of
298 81% over three generations, leading to the species being listed as Endangered
299 globally (BirdLife International 2017) and critically endangered in Australia
300 (Department of the Environment 2015). An identified priority information need
301 is to better quantify the dependence of the species on key migratory staging
302 sites (Garnett et al. 2011).

303 Individual Eastern Curlews are known to follow different routes on their
304 northward and southward migrations (Minton, Jessop, Collings & Standen 2011).
305 To demonstrate our approach, we focus on the northward migration, which is

Node name	Description
SAUS	Southern Australia
SEAUS	Southeastern Australia
NWAUS	Northwestern Australia
NAUS	Northern Australia
MSIA-IND	Malaysia, Indonesia
JPN-SK	Japan, South Korea
YS-NK	Yellow Sea, North Korea
BREED	Breeding

Table 1: Description of defined migration network nodes

306 better understood (Minton, Jessop, Collings & Standen 2011). Birds depart
307 the Australian nonbreeding grounds in late February and March, with more
308 southerly birds departing and arriving at their destinations earlier. Most birds
309 make a non-stop, long-distance flight to the southern parts of Japan, Korea and
310 the Yellow Sea, in 3-4 weeks, arriving in late March or early April. Birds depart
311 the Yellow Sea and Korean peninsula and arrive on their breeding ground dur-
312 ing April and early May. The 10000km journey from the southerly Victorian
313 non-breeding grounds to the breeding grounds is completed in 6-8 weeks, while
314 the shorter trip from the South-Eastern (8000km) and North-Western (7500km)
315 Australian non-breeding areas to the breeding grounds takes roughly 5-6 weeks
316 (Minton, Jessop, Collings & Standen 2011). For our case study of northward
317 migration, we model the first 26 weeks of the year (i.e. 1 January- late June)
318 with a weekly timestep.

319 Although the major migration linkages are known from recaptures and re-
320 sightings, as well as some satellite tracking and count data, little is known about
321 the timing or the duration that curlews spend at stopover sites (Minton, Jessop,
322 Collings & Standen 2011) or how the individual sightings data can be extrapo-
323 lated to the population level.

324 In our case study we model the migration network using 8 nodes representing

325 the major known stopover regions for Eastern Curlews (see table 1), connected
326 by 12 edges (Figure 6). Nodes and edges are based on observed sightings and
327 descriptions (Minton, Jessop, Collings & Standen 2011, Minton, Wahl, Gibbs,
328 Jessop, Hassell & Boyle 2011), an expert-derived network (Iwamura et al. 2013),
329 distribution maps (BirdLife International 2017) and eBird observations for East-
330 ern Curlew. eBird sample data are incomplete and spatially biased (Strimas-
331 Mackey et al. 2020), so to obtain estimates of total observed count (O_i^t) within
332 each node i at time t , we completed the following steps:

- 333 1. Drew approximate node boundaries based on the expert-derived network
334 (Iwamura et al. 2013). The geographical extent of the approximate nodes
335 were defined to capture the Internationally Important Sites designated on
336 the basis of Eastern Curlew numbers (Bamford et al. 2008) and include
337 as many eBird checklists as practicable.
- 338 2. Clipped the geographical extent to the intersection of the approximate
339 node boundaries and the Birdlife species distribution maps (BirdLife In-
340 ternational 2017) to give a refined node area.
- 341 3. Overlaid these intersected areas with a hexagonal grid (cell size 100km^2 ,
342 which roughly coincides with the $10 \times 10\text{km}$ grid cell sizes used to extrap-
343 olate populations in Hansen et al. (2016)).
- 344 4. Within each hexagon, computed the mean count observed in each hexagon
345 in each week. This step aimed to reduce the impacts of double counting
346 and spatially variable survey effort.
- 347 5. Obtained an extrapolated count estimate for the node by assigning the
348 mean count per hexagon (from hexagons with observed records) to all
349 hexagons. The sum of counts over all hexagons was assumed to be an
350 estimate of O_i^t .

351 Weekly extrapolated count estimates were extracted for the first 26 weeks of
352 2018 and 2019. Initial node counts were assigned based on expert estimates
353 from Iwamura et al. (2013). See Supporting Information S5 for additional details
354 regarding node and edge definition as well as how eBird count data was assigned

355 to the nodes.

356 Flight durations between nodes (Supporting Information S5) were estimated
357 using the distance between key aggregation sites in nodes and assuming a mi-
358 gration flight speed of 50km/h (ground speed) consistent with Driscoll & Ueta
359 (2002) (estimated flight speed of 50km/h); and Minton, Jessop, Collings &
360 Standen (2011), Minton et al. (2013) (median tracked speed 50.2km/h). Flight
361 durations were rounded to the nearest week with a minimum assumed travel
362 time of 1 week.

363 For our case study, we assume that mortality during the migration is zero.
364 In the absence of mortality estimates during migration for Eastern Curlew, and
365 most other species, it remains an open question whether the impacts of loss
366 of staging habitat impact species directly during the migration, or indirectly
367 through reduced breeding success or survival while at breeding or non-breeding
368 sites. However if future analyses are able to determine mortality during migra-
369 tion, it would be simple to include this estimate in our analysis.

370 Parameter p of the negative binomial distribution was estimated directly
371 from the data prior to the estimation of the other model parameters with MCEM
372 or ABC (see Supplementary Information S5).

373 Our objective is to estimate the edge strength between nodes ($R(i, j)$), pro-
374 viding estimates of population connectivity during migration, and the sojourn
375 durations at each node. The uncertainties tested in the case study are the routes
376 taken by birds migrating from Southern, Southeast and Northern Australia–
377 specifically the proportions of internal migration along the eastern and north-
378 ern Australian nodes and the relative proportions of birds using stopovers in
379 the Yellow Sea compared to those using South Korea and Japan (Figure 6).

380 **3. Results**

381 *3.1. Benchmarking the MCEM and ABC algorithms*

382 Both algorithms performed well on the benchmark experiments. Across
383 all benchmark problems, MCEM and ABC estimates of the parameters were

384 associated with mean absolute errors of 0.07 and 0.13 respectively. Parameter
 385 estimates were well correlated with the true parameters, with a correlation of
 386 0.84 and 0.65 respectively ($p < 2.2e^{-16}$ for the Pearson’s correlation tests).
 387 MCEM statistically performed better than ABC. Increasing the number of nodes
 388 (Figure 2) and the maximum number of outgoing edges (Figure 3-a) increased
 389 the error of estimation, however mean absolute errors remained reasonably low
 390 across all benchmark problems.

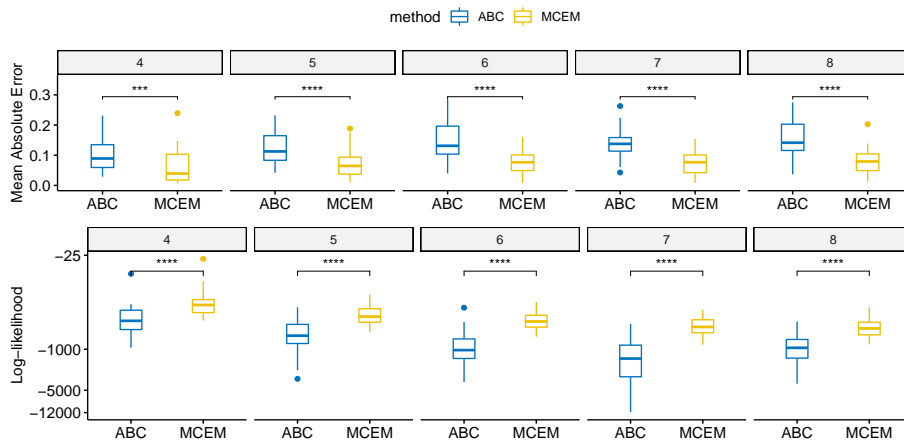


Figure 2: Mean absolute error (above), and log likelihood (below) of estimated parameters using ABC and MCEM, according to the number of sites. Mean absolute error is computed using the average error over all sojourn, transition and observation parameters. The statistical tests performed were paired Wilcoxon tests since the differences are based on the same benchmark problems. Significance test symbols * * * and * * * * refer to a p -value less than 0.001 and 0.0001 respectively.

391 Missing observations did not substantially impact the quality of estimation
 392 (Figure 3-b). Transition probabilities were estimated with lower error than
 393 sojourn mean times and appeared less sensitive to the number of sites (Figure 4).
 394 Error on the estimation of the observation parameter δ was close to 0 regardless
 395 of the number of sites (Figure 4). An additional illustration of the estimation
 396 results on a benchmark problem is provided in Supporting information S6.

397 In terms of number of simulations, MCEM required a median number of

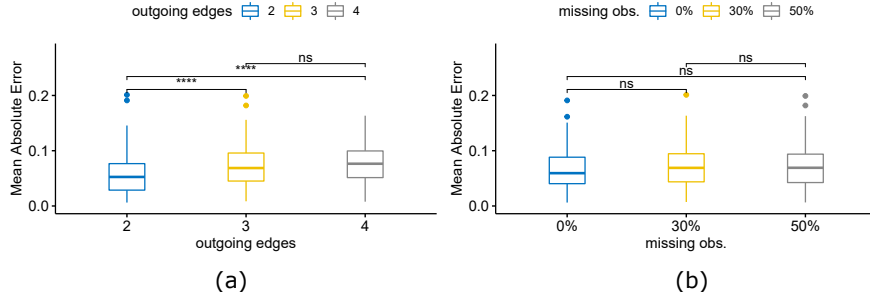


Figure 3: Mean absolute error according to the maximal number of outgoing edges (a) and Mean absolute error according to the percentage of missing observations (b). Mean absolute error was computed using the combined parameter estimates from the ABC and MCEM algorithms. The statistical tests performed were unpaired Wilcoxon tests since the differences are based on different benchmark problems. Significance test symbols *ns* and ***** refer to a *p*-value greater than 0.05 and less than 0.0001 respectively.

398 106650 simulations and a median time of 0.5 hours to reach convergence per
 399 problem, taking into account the 5 optimizations using different initial values
 400 of the parameters. ABC was much more expensive with a median number of
 401 245000 simulations and a median time of 4 hours. ABC is expensive due to a
 402 high rejection rate of simulated observations. Rejections occur because there is
 403 a low probability that a sampled parameter set generates observations that are
 404 close to the true observations.

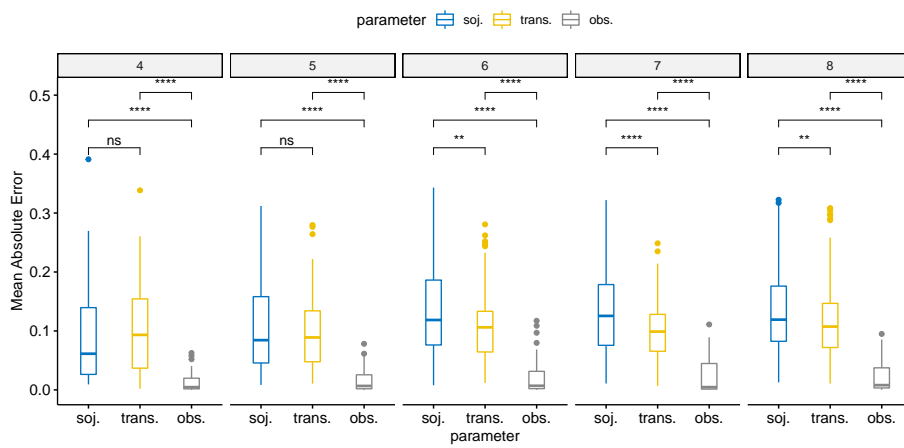


Figure 4: Mean absolute error of estimated sojourn mean time compared to mean absolute error of estimated transition probabilities and the estimated observation parameter δ according to the number of sites. Mean absolute error was computed using the combined parameter estimates from the ABC and MCEM algorithms. The statistical tests performed were paired Wilcoxon tests since the differences are based on the same benchmark problems. Significance test symbols *ns*, **** and ****** refer to a *p*-value greater than 0.05, less than 0.01 and less than 0.0001 respectively.

405 *3.2. Case Study: The Eastern Curlew*

406 MCEM and ABC estimates of parameters for the migratory network sup-
407 ported different hypotheses about the routes taken by Eastern Curlews (Figure
408 5; Figure 6; Venter mode parameter estimates are included in Supporting Infor-
409 mation S7). ABC results estimated strong reliance on the Yellow Sea in both
410 2018 and 2019, with only small proportions visiting the Japan-South Korean
411 node. Although the majority of birds flew directly to north Asia from their
412 origin, ABC estimated that many birds staged in a more northerly Australian
413 node before undertaking their migration, particularly in 2018.

414 In 2018, MCEM estimated that Australian birds flew to the Yellow Sea, but
415 in 2019 most birds instead flew to the South Korean-Japanese node. In MCEM,
416 the amount of staging in a more northerly Australian node was much stronger
417 than in ABC. Unlike in ABC, where many birds flew directly to the Yellow Sea
418 from their origin node, in MCEM large majorities of birds (70-90%) “hopped”
419 north to the Northern Australian node before undertaking their long flight to
420 the Yellow Sea or Japan-South Korea.

421 For both algorithms, sojourn time parameter estimates were stable between
422 years. MCEM estimated sojourn times at Southern (13 weeks) and North-
423 western Australia (13 weeks) that were longer than expected, suggesting a later
424 start to migration than expected for the species. In contrast, ABC estimates for
425 these two nodes were approximately 8 weeks for both nodes, matching the late
426 February departure expected from observations. MCEM also predicted unusu-
427 ally short sojourn times (~ 0 weeks) for both the Yellow Sea and Japan-South
428 Korean nodes (ABC predicted 2 weeks for both nodes).

429 To further highlight the differences between the two algorithms we simu-
430 lated trajectories using FlywayNet with either MCEM or ABC estimators and
431 computed the number N_i^t of birds at site i and time t from these trajectories
432 (Figure 8). Reasonable departure times were estimated from the ABC simulated
433 trajectories (approximately week 8; February/March). For MCEM, birds began
434 departing Southeastern Australia slightly early (mean departure week 5), and
435 left Southern Australia later than expected (mean departure week 12). Since

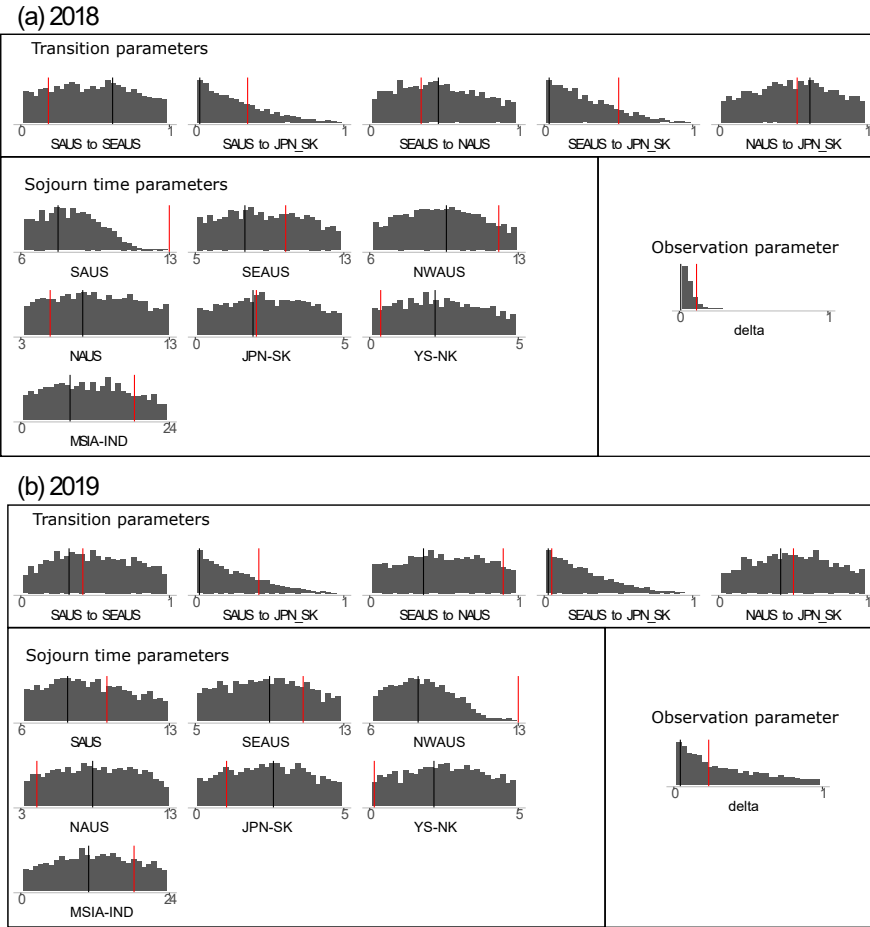


Figure 5: The marginals of the posterior distribution for each parameter estimated with ABC for the Curlew problem using the (a) 2018 and (b) 2019 datasets; the black lines represent the Venter mode of the marginals of the posterior distribution. Red vertical lines represent the estimated parameters computed by MCEM.

436 the majority of birds originate in Southeastern Australia, this had the effect of
 437 causing the migration to be shifted earlier for MCEM. Consequently, combined
 438 with very short estimated sojourn times at the Yellow Sea and South Korean-
 439 Japanese nodes (mean sojourn durations for both nodes < 1 week), MCEM
 440 estimated trajectories had very early arrivals at the breeding node (first ar-
 441 rivals late January). With ABC estimators, trajectories were closer to observed

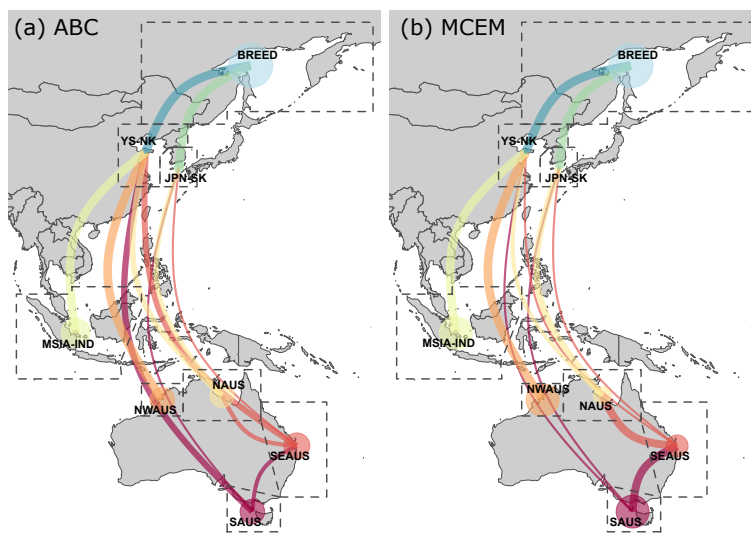


Figure 6: Estimated Eastern Curlew networks using 2019 eBird records and (a) ABC and (b) MCEM algorithms. Edge widths depict the relative transition probabilities between nodes; node sizes represent relative sojourn time lengths. Dotted lines depict the node boundaries. Colours depict edges from the same origin.

442 Curlew behaviour (Figure 7): there was a peak departure for both Northwest-
 443 ern and Southern Australian nodes near the end of February and a peak in bird
 444 numbers at the Yellow Sea in April. The mean migration time for ABC was
 445 7.0 weeks in 2018 and 6.2 weeks in 2019, close to the 6-8 weeks estimated in
 446 the literature Minton, Jessop, Collings & Standen (2011); for MCEM it was 5.1
 447 weeks and 5.7 weeks for 2018-19 respectively.

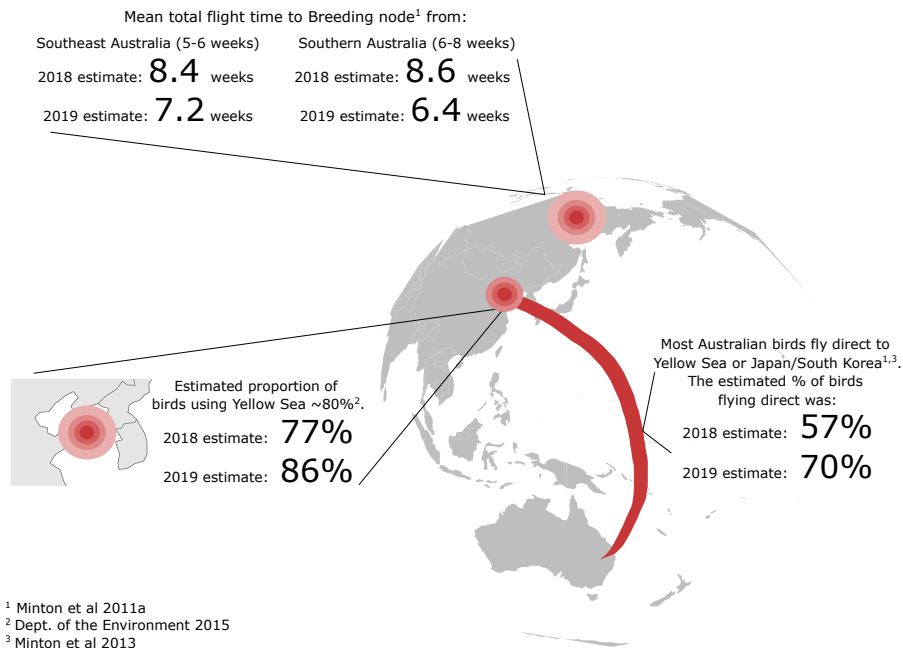


Figure 7: Comparison of estimated ABC results with values derived from the literature

448 4. Discussion

449 When tested on benchmark problems, both the ABC and MCEM algorithms
 450 performed well and recovered parameter values with good accuracy across vari-
 451 ous sized networks and numbers of connected neighbours. This suggests that, if
 452 our HSMM model assumptions hold and sufficiently accurate counts are avail-
 453 able, our algorithms provide a powerful way to recover network structure (i.e.

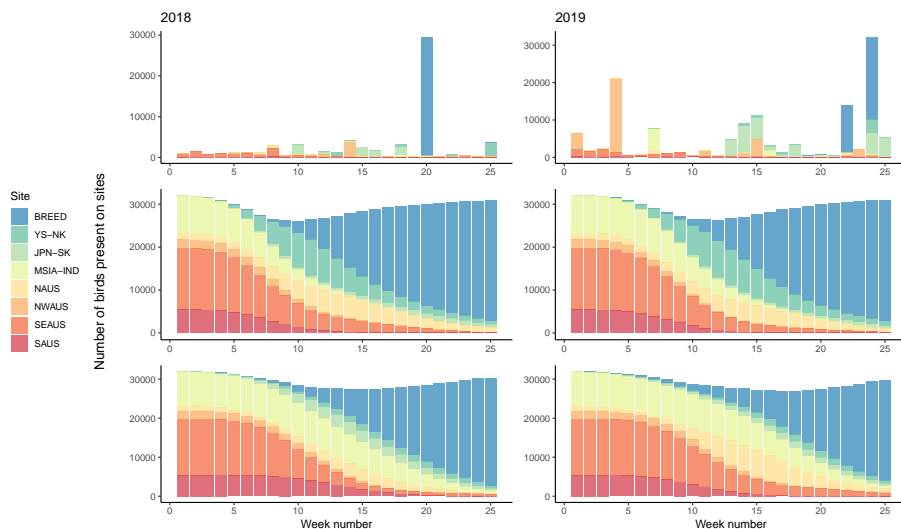


Figure 8: Simulated weekly trajectories for the Eastern Curlew in 2018 (left) and 2019 (right). Plots show (top) observed counts, (middle) trajectories simulated using ABC estimators and (bottom) trajectories simulated using MCEM estimators.

454 the relative importance of edges and the durations of stopovers). Unlike previ-
 455 ous approaches (Kölzsch et al. 2018), using this approach enables us to account
 456 for error on count data and provides a flexible framework that can make infer-
 457 ence with incomplete spatiotemporal count data. In contrast to previous studies
 458 which require high-resolution data which is difficult to obtain for species that
 459 are not included in formal monitoring programs (Kölzsch et al. 2018, Jain &
 460 Dilkina 2015), our approach requires only a basic network structure and count
 461 information, which is more widely available than individual trajectories (e.g.
 462 from satellite tagging of birds), providing a useful complementary source of in-
 463 ference to traditional bird tracking studies. Count data is the default method
 464 of data collection for bird watchers globally, so harnessing this data source is a
 465 powerful way to make best use of a global citizen science network.

466 Although our algorithms worked well for connectivity estimation on bench-
 467 mark problems, we obtained contrasting results when we attempted to estimate
 468 real Eastern Curlew networks from eBird data. In contrast to our benchmark

469 testing results where MCEM had lower estimation error, the ABC results for
470 the case study appeared to better match existing knowledge about curlew move-
471 ments, particularly in terms of reproducing the dependence on the Yellow Sea
472 rather than South Korea-Japan. This could be partly due to how the node
473 boundaries are selected, however the different estimates make it hard to con-
474 clude which algorithm best estimates the true migratory behaviour of Eastern
475 Curlews.

476 We have two main hypotheses that could explain the differences between
477 the algorithm estimations for Eastern Curlews. Firstly, they have different
478 objective functions: MCEM maximises the likelihood while ABC optimises a
479 customised set of statistics (here we minimise a weighted sum squared error
480 between observed and simulated data). We investigated this hypothesis by
481 changing the ABC acceptance criterion to better match the MCEM likelihood
482 and obtained results that were closer to the MCEM estimates (see Supporting
483 Information S8). This suggested that some of the difference between algorithms
484 may be due to different optimisation criteria, however it does not suggest which
485 results are closer to the real bird network dynamics.

486 Secondly, the eBird count data could be too noisy for the algorithms to
487 reach a common estimation. Given that our benchmark performances were
488 similar despite their different optimisation criteria, we believe that this is the
489 most probable explanation for the discrepancy between algorithms. Although
490 some nodes had substantial observed count data and we used the best available
491 estimates of suitable range to develop weekly count estimates, geographic and
492 temporal coverage is variable in all nodes and our node abundance estimates
493 had high week-to-week variability for all nodes (Figure 8, top panel). The vari-
494 ability was evident in the posterior distributions of the ABC results, which were
495 relatively flat for several nodes (Figure 5). Further, our model assumes that the
496 initial population size at each node is known, but we drew this information from
497 expert knowledge rather than count data. We pursued data from the Interna-
498 tional Waterbird Census (IWC, Delany, S 2005), which is a highly promising
499 dataset since it contains systematic count data recorded at the same time of year

500 (January; which roughly corresponds to the beginning of our simulation period
501 in our migration model). However, the IWC data has incomplete spatial cover-
502 age and variable survey effort in different areas, so additional research is needed
503 to use IWC data to generate node abundance estimates that could be used in
504 this study. Further research to better estimate node abundance from count data
505 would likely greatly improve our estimation ability. Potentially useful methods
506 may include smoothing the weekly observed data to infer observations at miss-
507 ing locations (e.g. Sheldon et al. 2007) or clustering to test the locations of the
508 node boundaries (Jain & Dilkina 2015). Although we deliberately tested how
509 well we could estimate using very minimal data from eBird, another promising
510 approach may be to incorporate additional datasets, such as tracking and/or
511 banding data, to guide the simulated trajectories. Formally incorporating envi-
512 ronmental covariates such as habitat type, temperature or the results of species
513 distribution modelling may provide additional information to improve abun-
514 dance estimates. Including covariates would also be useful for predicting other
515 parameters, most notably sojourn times, since birds make stopover decisions
516 based on environmental conditions such as the time of year and wind conditions
517 (Klzsch et al. 2016). It may also be useful to try the method on other well-
518 documented animal migratory data, particularly on species which are easier to
519 track and count, such as ungulate migration (Sawyer et al. 2009, Convention
520 on Migratory Species Secretariat 2021), where observation errors may be lower
521 than for shorebirds.

522 A key question for application is how to diagnose when the algorithms are
523 performing well. Clearly, if the two algorithms estimate very different parameter
524 values, users should seek to diagnose the cause of the difference. However,
525 there may also be other useful indicators of reliability. For example, in the
526 curlew problem, MCEM tended to seek the boundaries of its domain, suggesting
527 that the likelihood surface is increasing monotonically (and therefore that it is
528 unlikely that parameter estimates will be reliable). The posterior distributions
529 estimated by ABC for some parameters were relatively flat, which also acts as a
530 simple check to test for parameters that are difficult to estimate. Where users

531 find that MCEM estimates a majority of parameter values on the edge of their
532 domain and ABC finds numerous flat posterior distributions, we recommend
533 reviewing the data quality rather than accepting parameter estimates.

534 For our curlew case study, the difference in the results between algorithms
535 means that the findings should be interpreted with reference to the general
536 movement patterns of the species. Expert knowledge, such as that used here
537 (e.g. Figure 7), should be used to verify the predictions of the algorithms. The
538 general migratory behaviour of most bird species is known, and if an algorithm
539 does not reproduce this behaviour then it can be said to be performing poorly.
540 Where the algorithm results do align with known movement behaviour, the tests
541 outlined in the previous paragraph are a useful guide. In particular, for our
542 curlew case study, we suggest that although the ABC results are encouraging,
543 some parameters have flat marginal posterior distributions (Figure 5) and these
544 should be further scrutinised before being used in applied conservation.

545 The migratory movement patterns of some bird species are poorly known.
546 Our model accommodates this by requiring minimal input information, specif-
547 ically: count data, the location boundaries of the nodes, and the suspected
548 connections between nodes. Count data is readily available for any species via
549 eBird, so this should not be a limiting factor except where there are few observa-
550 tions of the species recorded. The locations of the nodes and the hypothesized
551 connections between nodes can be obtained from a combination of eBird list
552 locations, published studies and expert knowledge (See Supporting Information
553 S5). For poorly-known species where studies and expert knowledge are lacking,
554 eBird data alone could be used to set the hypothesized network structure, al-
555 though further research would be valuable to determine the most robust way to
556 set node boundaries (e.g. clustering techniques).

557 Our model makes some assumptions that could be improved in future iter-
558 ations. First, we assumed that the time spent in a node is the same regardless
559 of the origin and destination (i.e. sojourn times are independent of the origin
560 and destination). Where nodes are both a nonbreeding origin node as well as
561 a stopover node (e.g. SEAUS in our model), this may confound sojourn times

562 that occur between week zero and when the migration starts with sojourn times
563 due to true stopovers during migration. If these sojourn times are considerably
564 different, then sojourn estimation may be affected. Modelling conditional so-
565 journ times is possible within our HSMM framework, however it would increase
566 the number of parameters at each node and increase the difficulty of estimation.

567 Second, we assumed that sojourn times at different sites are independent.
568 Strictly, since we have an idea of the total duration of the migration, the sum
569 of the sojourn times along a trajectory should not substantially exceed the ex-
570 pected total duration, so the independence assumption may not hold. We expect
571 that in practice the observed data will minimise the impacts of this assumption
572 by enforcing average movement between nodes at reasonable times (i.e. that
573 match the observed movement times), even without explicitly modelling depen-
574 dence between nodes. There could be some trajectories that are overly long
575 or short due to independent sojourn times, but these should be minimised by
576 fitting to observed movements.

577 Third, we assumed that migratory birds progress in one direction (north-
578 ward migration only) and that migration time between nodes was constant. In
579 practice, some birds are known to abort migrations and return to their origin
580 (Driscoll & Ueta (2002)), but for modelling reasons we assumed that this pro-
581 portion was small. We also only modelled the northward migration; it would be
582 theoretically trivial to model the full annual cycle of migration however doing
583 so would increase the number of parameters that need to be estimated. It may
584 be more practical to model northward and southward migrations separately as
585 we have done here.

586 Fourth, eBird observations (and count data in general) will tend to be un-
587 derestimates of the true population, since at any time it is unlikely that an
588 observer will see and record all the birds in an area. This creates the possibility
589 of systematic bias in the count data, which is not explicitly captured in our neg-
590 ative binomial observation model. However, there are other sources of error in
591 the node abundance estimates, most notably the extrapolation process used to
592 estimate counts in areas of the node where no lists have been recorded. For the

593 curlew study, the extrapolation process was likely to dwarf the errors caused by
594 underestimation due to the area of extrapolation required, minimizing the effects
595 of systematic bias. However, for other studies where minimal extrapolation is
596 required, further attention (and perhaps alternative observation distributions)
597 may be required to deal with systematic under-counting.

598 In a situation where we are sure that the variable used for modeling the
599 observed count is an underestimation of the real count, we should use a Binomial
600 distribution, $B(N_i, p_i)$, with N_i the true count, and p_i the probability to see a
601 bird (this formulation is included in Supporting Information S1). However, the
602 drawback of the Binomial formulation is that we don't avoid overdispersion with
603 this distribution. Since we are not certain that the node count estimates are
604 underestimates, we use the negative Binomial to manage overdispersion.

605 The ABC algorithm estimates marginal modes for each parameter, but
606 strictly speaking, the multivariate posterior mode is most comparable to the
607 MCEM estimate. The multivariate mode was not used because it is more com-
608 putationally expensive to generate (requires estimation of a multivariate kernel
609 density function and optimization of the density function) than the straight-
610 forward computation required to generate the Venter mode of the marginals.
611 Other methods have also used the mode of the marginals to represent the ABC
612 posterior (e.g. Nunes & Prangle (2015)). We are also confident that posterior
613 is nice enough to be summarized by the mode of marginals (see e.g. Figure 2 in
614 Appendix S6), at least for the benchmark experiments.

615 Both algorithms became time-consuming to run as the networks became
616 complex, particularly for the curlew case study. Runtime may limit performance
617 on large networks, so it may be beneficial to investigate alternative methods to
618 estimate the network connectivity. Variational EM in a frequentist approach
619 (VEM Neal & Hinton 2000) or Bayes Expectation-Maximisation (VBEM Beal
620 2003) in a Bayesian approach may be interesting solutions for a trade-off be-
621 tween runtime and the quality of estimators. Instead of relying on simulations,
622 variational approaches perform estimation by replacing the complex distribution
623 (here the HSMM model) by a closer model in a family of tractable distributions.

624 We are currently investigating whether VEM or VBEM can be used to solve our
625 migratory networks problem.

626 **5. Conclusion**

627 Understanding how migratory populations move is crucial because it allows
628 us to design conservation measures accordingly. Here we have developed a new
629 way to estimate the connectivity of migratory populations based only on limited
630 count data at discrete locations. The method accounts for observation error and
631 predicts both migratory structure and sojourn times. Although information
632 about migratory connectivity can be inferred from individual tracking studies,
633 few studies have attempted to extrapolate individual behaviours to population-
634 level movements. Our study complements existing tracking work by providing
635 a statistical model to exploit the most commonly collected form of bird data.
636 Although questions remain about how best to estimate node abundance, our
637 approach has tremendous promise because of the explosion in availability of
638 citizen science count data through platforms such as eBird. As these datasets
639 grow, existing geographic and temporal gaps in the datasets will be filled. As
640 this happens, there will be increasing demand for algorithms that are sufficiently
641 flexible to draw inference from unstructured data.

642 **6. Author Contributions Statement**

643 Sam Nicol (SN), Marie-Josée Cros (MJC), Nathalie Peyrard (NP), Régis
644 Sabbadin (RS) and Ronan Trépos (RT) contributed equally to this manuscript.
645 Specifically, NP, RS and SN conceived and designed the HSMM model. RT
646 implemented the MCEM and ABC algorithms and designed benchmarks. MJC
647 implemented the case study. SN extracted the data for the curlew case study and
648 wrote the draft manuscript. Richard Fuller and Bradley Woodworth provided
649 advice on the curlew case study. All authors edited the manuscript and provided
650 critical comments.

651 **7. Data availability**

652 The FlywayNet package, including installation instructions and a vignette
653 demonstrating a minimal example, is available from the INRAE git repository:
654 <https://forgemia.inra.fr/birdnet/FlywayNet>.

655 Results and additional materials to run the study are available on figshare:
656 <https://doi.org/10.6084/m9.figshare.16658185.v5>.

657 Specifically, the figshare contains two files:

- 658 • “Flywaynet_experiments.zip” contains the results of the benchmarking
659 and curlew experiments.
- 660 • “CurlewCaseStudy_GetWeeklyObservations.tar.gz” contains the scripts and
661 data used to convert the raw eBird eastern curlew counts into weekly ob-
662 servation data, as outlined in Supporting Information S5.

663 Further information about each of the files and how to reproduce the results is
664 contained in the readme files contained with the downloads.

665 Scripts to re-run the benchmarking experiments and the curlew case study
666 are available from: https://forgemia.inra.fr/birdnet/FlywayNet_experiments. Note
667 that re-running the scripts is time consuming without substantial computing
668 resources— if users want to run their own example or view the results, we instead
669 recommend using the FlywayNet R package or viewing the figshare repository
670 respectively.

671 **8. Acknowledgements**

672 The authors acknowledge the computational resources provided by RECORD
673 platform (Bergez et al. 2013). SN was supported by a CSIRO Julius Career
674 Award. NP, RS, MJC and RT acknowledge the support of the French Agence
675 Nationale de la Recherche (ANR), under grant ANR-21-CE40-005 (project HSMM-
676 INCA).

677 **9. Conflict of Interest Statement**

678 The authors declare no conflict of interest.

679 Andrieu, C., Freitas, N. D., Doucet, A. & Jordan, M. (2003), ‘An introduction
680 to MCMC for machine learning’, Machine Learning **50**, 5–43.

681 Bamford, M., Watkins, D., Bancroft, W., Tischler, G. & Wahl, J. (2008), Migra-
682 tory shorebirds of the East Asian-Australasian flyway: population estimates
683 and internationally important sites, Technical report, Wetlands International-
684 Oceania.

685 Beal, M. (2003), Variational Algorithms for Approximate Bayesian Inference,
686 Ph.d. thesis, Gatsby Computational Neuroscience Unit, University College
687 London.

688 Bergez, J.-E., Chabrier, P., Gary, C., Jeuffroy, M., Makowski, D., Quesnel,
689 G., Ramat, E., Raynal, H., Rouse, N., Wallach, D., Debaeke, P., Durand, P.,
690 Duru, M., Dury, J., Faverdin, P., Gascuel-Oudou, C. & Garcia, F. (2013), ‘An
691 open platform to build, evaluate and simulate integrated models of farming
692 and agro-ecosystems’, Environmental Modelling and Software **39**(1), 39–49.

693 BirdLife International (2017), *Numenius madagascariensis* (amended ver-
694 sion of 2016 assessment). the IUCN Red List of Threatened Species
695 2017: e.t22693199a118601473. [https://dx.doi.org/10.2305/iucn.uk.2017-
696 3.rlts.t22693199a118601473.en](https://dx.doi.org/10.2305/iucn.uk.2017-3.rlts.t22693199a118601473.en), Technical report.

697 Clemens, R., Rogers, D. I., Hansen, B. D., Gosbell, K., Minton, C. D. T., Straw,
698 P., Bamford, M., Woehler, E. J., Milton, D. A., Weston, M. A., Venables,
699 B., Wellet, D., Hassell, C., Rutherford, B., Onton, K., Herrod, A., Studds,
700 C. E., Choi, C.-Y., Dhanjal-Adams, K. L., Murray, N. J., Skilleter, G. A. &
701 Fuller, R. A. (2016), ‘Continental-scale decreases in shorebird populations in
702 Australia’, Emu - Austral Ornithology **116**(2), 119–135.

- 703 Cohen, E. B., Hostetler, J. A., Royle, J. A. & Marra, P. P. (2014), ‘Estimating
704 migratory connectivity of birds when re-encounter probabilities are heteroge-
705 neous’, Ecology and Evolution **4**(9), 1659–1670.
- 706 Convention on Migratory Species Secretariat (2021), ‘Global initiative on un-
707 gulate migration’.
708 **URL:** <https://www.cms.int/en/gium>
- 709 Csilléry, K., Blum, M. G. B., Gaggiotti, O. E. & François, O. (2010), ‘Ap-
710 proximate Bayesian Computation (ABC) in practice’, Trends in Ecology and
711 Evolution **25**(7), 410–418.
- 712 Delany, S (2005), Guidelines for participants in the international waterbird cen-
713 sus (iwc), Technical report, Wetlands International, Wageningen, The Nether-
714 lands.
- 715 Department of the Environment (2015), Conservation advice *Numenius mada-*
716 *gascariensis* eastern curlew., Technical report, Department of the Environ-
717 ment.
- 718 Driscoll, P. V. & Ueta, M. (2002), ‘The migration route and behaviour of eastern
719 curlews *Numenius madagascariensis*’, Ibis **144**(3), E119–E130.
- 720 Fournier, A. M. V., Sullivan, A. R., Bump, J. K., Perkins, M., Shieldcastle,
721 M. C. & King, S. L. (2017), ‘Combining citizen science species distribution
722 models and stable isotopes reveals migratory connectivity in the secretive
723 Virginia rail’, Journal of Applied Ecology **54**(2), 618–627.
- 724 Garnett, S., Szabo, J. & Dutson, G. (2011), The Action Plan for Australian
725 Birds 2010, Birds Australia. CSIRO Publishing, Melbourne, Australia.
- 726 Hallworth, M. T., Sillett, T. S., Van Wilgenburg, S. L., Hobson, K. A. & Marra,
727 P. P. (2015), ‘Migratory connectivity of a neotropical migratory songbird re-
728 vealed by archival light-level geolocators’, Ecological Applications **25**(2), 336–
729 347.

- 730 Hansen, B., Fuller, R., Watkins, D., Rogers, D., Clemens, R., Newman, M.,
731 Woehler, E. & Weller, D. (2016), Revision of the East Asian-Australasia
732 flyway population estimate for 37 listed migratory shorebird species. un-
733 published report for the Department of the Environment, Technical report,
734 BirdLife Australia.
- 735 Hastings, W. (1970), ‘Monte Carlo sampling methods using Markov chains and
736 their applications’, Biometrika **57**(1), 97–109.
- 737 Iwamura, T., Possingham, H. P., Chadés, I., Minton, C., Murray, N. J., Rogers,
738 D. I., Treml, E. A. & Fuller, R. A. (2013), ‘Migratory connectivity magnifies
739 the consequences of habitat loss from sea-level rise for shorebird populations’,
740 Proceedings of the Royal Society B: Biological Sciences **280**(1761), 20130325.
- 741 Jabot, F., Faure, T. & Dumoulin, N. (2013), ‘Easyabc: performing efficient
742 approximate Bayesian computation sampling schemes using R’, Methods in
743 Ecology and Evolution **4**(7), 684–687.
- 744 Jabot, F., Faure, T., Dumoulin, N. & Albert, C. (2015), EasyABC: Efficient
745 Approximate Bayesian Computation Sampling Schemes. R package version
746 1.5.
747 **URL:** <https://CRAN.R-project.org/package=EasyABC>
- 748 Jain, N. & Dilkina, B. (2015), Coarse models for bird migrations using cluster-
749 ing and non-stationary Markov chains, in ‘AAAI Workshop: Computational
750 Sustainability’, pp. 63–68.
751 **URL:** <http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10209>
- 752 Joo, R., Bertrand, S., Tam, J. & Fablet, R. (2013), ‘Hidden Markov models:
753 The best models for forager movements?’, PLoS One **8**(8), e71246.
- 754 Klzsch, A., Mskens, G., Kruckenberg, H., Glazov, P., Weinzierl, R., Nolet, B.
755 & Wikelski, M. (2016), ‘Towards a new understanding of migration timing:
756 slower spring than autumn migration in geese reflects different decision rules
757 for stopover use and departure’, Oikos **125**(10), 1496–1507.

- 758 Kölzsch, A., Kleyheeg, E., Kruckenberg, H., Kaatz, M. & Blasius, B. (2018), ‘A
759 periodic Markov model to formalize animal migration on a network’, Royal
760 Society Open Science **5**(6), 180438.
- 761 Levine, R. A. & Casella, G. (2001), ‘Implementations of the Monte Carlo EM
762 algorithm’, Journal of Computational and Graphical Statistics **10**(3), 422–
763 439.
- 764 Lientz, B. (1972), ‘Properties of modal intervals’, SIAM Journal on Applied
765 Mathematics **23**, 1–5.
- 766 Marra, P. P., Cohen, E., Harrison, A.-L., Studds, C. E. & Webster, M. S. (2019),
767 Migratory Connectivity, 2nd edn, Academic Press, Oxford, pp. 455–461.
- 768 Minton, C., Gosbell, K., Johns, P., Christie, M., Klaassen, M., Hassell, C.,
769 Boyle, A., Jessop, R. & Fox, J. (2013), ‘New insights from geolocators de-
770 ployed on waders in Australia’, Wader Study Group Bulletin **120**, 37–46.
- 771 Minton, C., Jessop, R., Collings, P. & Standen, R. (2011), ‘The migration
772 of eastern curlew *Numenius Madagascariensis* to and from Australia’, Stilt
773 **59**, 6–16.
- 774 Minton, C., Wahl, J., Gibbs, H., Jessop, R., Hassell, C. & Boyle, A. (2011),
775 ‘Recoveries and flag sightings of waders which spend the non-breeding season
776 in Australia’, Stilt **59**, 17–43.
- 777 Neal, R. & Hinton, G. (2000), ‘A view of the EM algorithm that justifies incre-
778 mental, sparse, and other variants’, Learning in graphical models **89**, 355–368.
- 779 Nunes, M. A. & Prangle, D. (2015), ‘abctools: An R Package for Tuning Ap-
780 proximate Bayesian Computation Analyses’, The R Journal **7**(2), 189–205.
- 781 O’Connell, J. & Hojsgaard, S. (2011), ‘Hidden semi Markov models for multiple
782 observation sequences: The mhsmm package for R’, Journal of Statistical
783 Software **39**(4).

- 784 Rappole, J. H. & McDonald, M. V. (1994), ‘Cause and effect in population
785 declines of migratory birds’, The Auk **111**(3), 652–660.
- 786 Robbins, C. S., Sauer, J. R., Greenberg, R. S. & Droege, S. (1989), ‘Population
787 declines in North American birds that migrate to the neotropics’, Proceedings
788 of the National Academy of Sciences **86**(19), 7658.
- 789 Runge, C. A., Martin, T. G., Possingham, H. P., Willis, S. G. & Fuller, R. A.
790 (2014), ‘Conserving mobile species’, Frontiers in Ecology and the Environment
791 **12**(7), 395–402.
- 792 Runge, C. A., Watson, J. E. M., Butchart, S. H. M., Hanson, J. O., Possingham,
793 H. P. & Fuller, R. A. (2015), ‘Protected areas and global conservation of
794 migratory birds’, Science **350**(6265), 1255.
- 795 Sawyer, H., Kauffman, M. J., Nielson, R. M. & Horne, J. S. (2009), ‘Identifying
796 and prioritizing ungulate migration routes for landscape-level conservation’,
797 Ecological Applications **19**(8), 2016–2025.
798 **URL:** [https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/08-](https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/08-2034.1)
799 [2034.1](https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/08-2034.1)
- 800 Semmens, D. J., Diffendorfer, J. E., López-Hoffman, L. & Shapiro, C. D. (2011),
801 ‘Accounting for the ecosystem services of migratory species: Quantifying mi-
802 gration support and spatial subsidies’, Ecological Economics **70**(12), 2236–
803 2242.
- 804 Sheldon, D., Saleh Elmohamed, M. & Kozen, D. (2007), Collective inference
805 on Markov models for modeling bird migration, in J. C. Platt, D. Koller,
806 Y. Singer & S. T. Roweis, eds, ‘Advances in Neural Information Processing
807 Systems 20, Proceedings of the Twenty-First Annual Conference on Neural
808 Information Processing Systems, Vancouver, British Columbia, Canada, De-
809 cember 3-6, 2007’, Curran Associates, Inc., pp. 1321–1328.
- 810 Sillett, T. S., Holmes, R. T. & Sherry, T. W. (2000), ‘Impacts of a global

811 climate cycle on population dynamics of a migratory songbird’, Science
812 **288**(5473), 2040.

813 Strimas-Mackey, M., Hochachka, W., Ruiz-Gutierrez, V., Robin-
814 son, O., Miller, E., Auer, T., Kelling, S., Fink, D. & John-
815 ston, A. (2020), Best Practices for Using eBird Data. Version 1.0
816 <https://cornelllabofornithology.github.io/ebird-best-practices/>, Cornell
817 Lab of Ornithology, Ithaca, New York.

818 Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D. & Kelling,
819 S. (2009), ‘ebird: A citizen-based bird observation network in the biological
820 sciences’, Biological Conservation **142**(10), 2282–2292.

821 Taylor, C. M. & Norris, D. R. (2010), ‘Population dynamics in migratory net-
822 works’, Theoretical Ecology **2**, 65–73.

823 Tonachella, N., Nastasi, A., Kaufman, G., Maldini, D. & Rankin, R. W. (2012),
824 ‘Predicting trends in humpback whale (*Megaptera novaeangliae*) abundance
825 using citizen science’, Pacific Conservation Biology **18**(4), 297–309.
826 **URL:** <https://www.publish.csiro.au/paper/PC120297>

827 Venter, J. H. (1967), ‘On estimation of the mode’, The Annals of Mathematical
828 Statistics **38**(5), 1446–1455.

829 Webster, M. S., Marra, P. P., Haig, S. M., Bensch, S. & Holmes, R. T. (2002),
830 ‘Links between worlds: unraveling migratory connectivity’, Trends in Ecology
831 and Evolution **17**(2), 76–83.

832 Wei, G. C. G. & Tanner, M. A. (1990), ‘A Monte Carlo implementation of the
833 EM algorithm and the poor man’s data augmentation algorithms’, Journal of
834 the American Statistical Association **85**(411), 699–704.

835 Wetlands International (2021), ‘Waterbird population estimates’.
836 **URL:** wpe.wetlands.org

- ⁸³⁷ Wilcove, D. S. & Wikelski, M. (2008), ‘Going, going, gone: Is animal migration
⁸³⁸ disappearing’, PLOS Biology **6**(7), e188.
- ⁸³⁹ Wilcoxon, F. (1945), ‘Individual Comparisons by Ranking Methods’, Biometrics
⁸⁴⁰ Bulletin **1**(6), 80–83.
- ⁸⁴¹ Yu, S.-Z. (2010), ‘Hidden semi-Markov models’, Artificial Intelligence .