BirdFlow: Learning Seasonal Bird Movements from Citizen Science Data

Miguel Fuentes*[1], Benjamin M. Van Doren[2], Daniel Fink[2], and Daniel Sheldon[1]

[1]Manning College of Information and Computer Sciences, University of Massachusetts Amherst, 140 Governors Drive, Amherst, MA 01003, USA

[2]Cornell Lab of Ornithology, Cornell University, Ithaca, NY 14850, USA

*Corresponding Author: Email: mmfuentes@cs.umass.edu

Abstract

Large-scale monitoring of seasonal animal movement is integral to science, conservation, and outreach. However, gathering representative movement data across entire species ranges is frequently intractable. Citizen science databases collect millions of animal observations throughout the year, but it is challenging to infer individual movement behavior solely from observational data. We present BirdFlow, a probabilistic modeling framework that draws on citizen science data from the eBird database to model the population flows of migratory birds. We apply the model to 11 species of North American birds, using GPS and satellite tracking data to tune and evaluate model performance. We show that BirdFlow models can accurately infer individual seasonal movement behavior directly from eBird relative abundance estimates. Supplementing the model with a sample of tracking data from wild birds improves performance. Researchers can extract a number of behavioral inferences from model results, including migration routes, timing, connectivty, and forecasts. The BirdFlow framework has the potential to advance migration ecology research, boost insights gained from direct tracking studies, and serve a number of applied functions in conservation, disease surveillance, aviation, and public outreach.

Key words: bird migration, movement ecology, graphical models, big data, species distributions, forecasting

1 Introduction

The movements of animals span the globe, and movement is integral to behavior, survival, and reproduction. Monitoring movement is particularly important in the face of climate and landscape change, forces that shape how animals interact with their environments (Bauer et al., 2019; Dunn & Møller, 2019). Capturing movement patterns is critical for effective conservation actions, which may hinge on accurate knowledge of animals' locations and how geographic and environmental interactions change over time (Fraser et al., 2018; Katzner & Arlettaz, 2020). For these reasons, incomplete movement information frequently impedes progress in science and conservation (Fraser et al., 2018; Katzner & Arlettaz, 2020). Often, these challenges arise from constraints on the number of animals that can be monitored, captured, or re-captured in the field; the weight and shape of

tracking devices; the number of tracking devices that can be deployed; and the geographic areas that can be adequately covered.

Migratory birds exemplify the challenges facing movement researchers, as well as the urgent need for additional movement information to inform science and conservation. Migratory birds are important indicators of ecosystem health that connect peoples and places in ways few phenomena can. Migrants rely on a predictable series of seasonally and regionally varying resources which, unfortunately, makes them susceptible to rapid global change (Bairlein, 2016; Rosenberg et al., 2019; Sanderson et al., 2006). In North America alone, an estimated three billion birds have been lost in the last half-century, representing nearly a third of the continent's avifauna (Rosenberg et al., 2019). To conserve migratory birds and study their responses to global change, data and methods are needed that can capture their movements at population scales. For example, a better understanding of the migratory connectivity of different populations of bird species is crucial (Schuster et al., 2019; Webster & Marra, 2005), but detailed connectivity information is lacking for most species. Unfortunately, wireless tracking devices are too heavy for most bird species, limiting the information that scientists can gather on their movements (McKinnon & Love, 2018). Other sources of direct movement data, such as Doppler weather radars, provide no information on species identities or individual behavior (Bauer et al., 2019; Dokter et al., 2018; Van Doren & Horton, 2018).

Citizen and community science projects provide a source of data on animal occurrence and abundance across the globe. In particular, the eBird (Sullivan et al., 2014) database comprises over one billion global bird observations and has been used highly successfully for population distribution modeling (Fink, Auer, et al., 2020; Fink, Auer, et al., 2020; Fink et al., 2014; Fink et al., 2013; Fink et al., 2010; Johnston et al., 2015). Although these projects are collecting increasing volumes of data across a variety of taxa (e.g. iNaturalist, camera trapping projects, etc.), most of these data only provide snapshots of occurrence across a population. Without tracking the movements of individuals, it is difficult to infer movement from these datasets. Methods that accurately infer movement behavior from large-scale observational data would unlock troves of citizen science data for use by movement researchers and conservation practitioners.

Previous studies have approached modeling movement from observational data by first extensively cleaning the data to correct for variability from the observation process, and then investigating

specific quantities of interest like centroid movement or estimated movement speed (Supp et al., 2021). Other promising approaches include deterministic models based on the concept of global energy efficiency, in which simulated birds are distributed to optimize both resource acquisition and energy expenditure (Somveille et al., 2021). However, it has proven challenging to accurately infer individual-level behavior across large spatial scales while accounting for the stochasticity inherent in the movement behavior of individuals.

Here, we present BirdFlow, a probabilistic modeling framework that uses relative abundance data from citizen science repositories to infer movement behavior across the geographic range of a species. Our method builds on previous work on collective graphical models, which reason about individual behavior from aggregate information about a population (Sheldon & Dietterich, 2011; Sheldon et al., 2013; Sun et al., 2015), and on a related modeling framework from private data analysis in human populations (McKenna et al., 2019). Inputs to BirdFlow are weekly high-resolution relative abundance models produced by the eBird Status & Trends project (Fink, Auer, et al., 2020). Outputs are weekly spatial transition matrices that can be interrogated for biological insight, including estimates of migratory paths, timing, connectivity, and forecasting. BirdFlow models can be trained on any species, even those not tracked by eBird, as long as relative abundance models are available. Direct tracking methods are not required but, in the event that direct tracking data are available, these data can be used to fine tune model hyperparameters in order to improve performance. In this paper, we investigate the performance of BirdFlow models on several bird species. We train models from eBird relative abundance estimates and use GPS and satellite-tracking data from wild birds to validate and evaluate model performance. We evaluate the sensitivity of the model to hyperparameter selection, asking whether trained models perform well under general settings or if species-specific tuning is required. Finally, we demonstrate how these probabilistic models can produce a range of high-resolution and temporally explicit biological inferences across species' entire ranges.

2 Methods and Materials

BirdFlow models reason about the distribution of tracks of birds of one species over discrete time steps. A track of one individual is modeled as a sequence of random variables $X_1, \ldots, X_T$, where

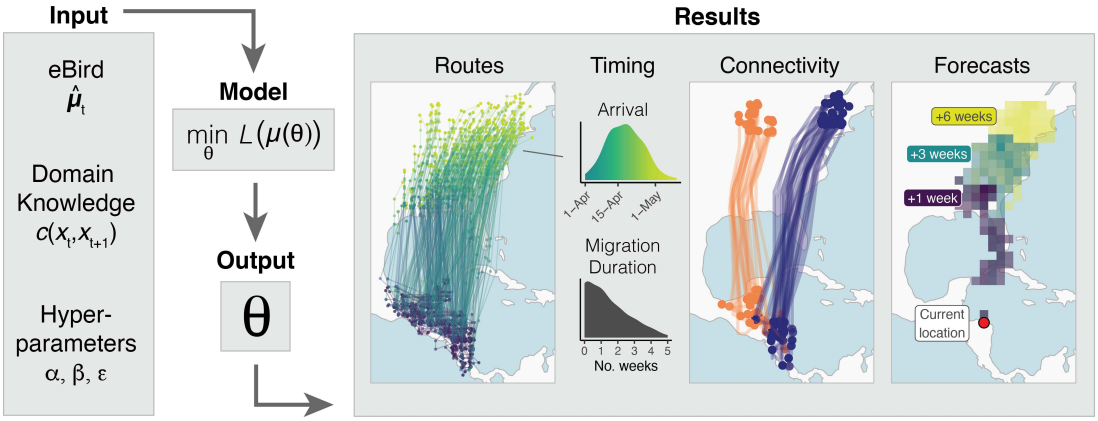

Figure 1: Methodology Outline.

$X_t \in \mathcal{X}$ represents the location at time t, from a discrete set $\mathcal{X}$ of locations (e.g., map grid cells). For the rest of the paper, we will use a weekly time step with week index t ranging from 1 to 52 to match the temporal resolution of eBird data. The randomness represents variability in tracks of individuals drawn from the population. The goal of BirdFlow is to estimate the population track distribution $p(x_1, \ldots, x_T) = \Pr(X_1 = x_1, \ldots, X_T = x_T)$, which can be conceptualized as a vector $\mathbf{p}$ with $|\mathcal{X}|^T$ entries, one for each possible track.

A key challenge in animal movement modeling is obtaining a broadly representative sample of individual movement tracks. To address this, we use weekly relative abundance estimates produced by the eBird Status & Trends project (Fink, Auer, et al., 2020). These eBird-based estimates provide direct evidence about marginal distributions of $\mathbf{p}$, the probability distribution averaged over the population of all individual tracks at local spatial scales. These estimates are released at a weekly time scale so our model will infer movement on that same time scale. Specifically, the normalized relative abundance estimates across a species range at week t corresponds to a *single-time-step marginal*, a vector $\boldsymbol{\mu}_t$ representing the distribution of the population over locations at week t. This vector has entries $\mu_t(x_t) = \Pr(X_t = x_t)$.

2.1 eBird Data

The eBird database (Sullivan et al., 2014) currently includes over 1 billion bird observations. eBird observers report information on observing effort and counts of all birds they observe during birding

trips in the form of species checklists. Over 77 million complete checklists currently provide presence-absence data for almost every bird species in the world. These data have seen broad applications advancing the field of 'big data' ornithology (La Sorte et al., 2018) and have been used to estimate full annual cycle relative abundance for almost every migratory species breeding in North America (Fink, Auer, et al., 2020; Fink, Auer, et al., 2020; Fink et al., 2014; Fink et al., 2013; Johnston et al., 2015). The eBird Status & Trends project[1] estimates the relative abundance of over 600 species at a spatial resolution of 3km x 3km and a *weekly* temporal resolution (Fink, Auer, et al., 2020; Fink, Auer, et al., 2020), providing spatial and temporal detail on the seasonally changing population-level abundance patterns of migratory species. These estimates of relative abundance at fine spatial and temporal scale were first completed in January 2020 and thus provide a unique and timely opportunity to estimate patterns of population movement across the full extent of their annual western hemispheric distributions. We used Status & Trends version 2020, which uses eBird data from 2006–2020 and produces estimates that are broadly representative of that time period.

2.1.1 Processing eBird Distribution Data

We downloaded relative abundance estimates for 11 bird species that also had available GPS or satellite tracking data (see Table 1 for list of species) as raster files from eBird Status & Trends project using the *ebirdst* R package (Auer et al., 2020). These estimates are provided at a spatial resolution of 3km x 3km and a *weekly* temporal resolution for 52 weeks. We chose to use the eBird-based relative abundance estimates instead of the eBird observations directly because (1) the estimates provide a spatiotemporally complete data set by filling spatiotemporal gaps based on modeled relationships with remotely sensed environmental data (Fink et al., 2014; Fink et al., 2013; Johnston et al., 2015), and (2) the estimates remove bias by accounting for systematic patterns of variation inherent in citizen-science observations (Fink, Auer, et al., 2020). We loaded rasters at 27 km resolution, re-projected to the Mollweide equal-area projection and further aggregated them to obtain an approximate grid resolution of 100-250 km, depending on the total size of the species' distribution. For species with larger distributions, we used coarser grids to keep total computational memory usage withing the limitations of our compute environment; specifically, our GPU memory was limited to grids with about 4000 or fewer cells for a 52-week modeling period. We used a

[1] https://ebird.org/science/status-and-trends

110-m resolution shapefile of global coastlines from Natural Earth (naturalearthdata.com) to mask open water, restricting our modeled area to terrestrial environments. For each weekly grid, we standardized relative abundance values by dividing each cell value by the total summed abundance so that the cells sum to one. This gave us weekly "ground truth" estimates $\hat{\boldsymbol{\mu}}_t$ of the single-time-step marginals, where $\hat{\boldsymbol{\mu}}_t(x_t)$ is the fraction of the population in grid cell x_t in week t as estimated by eBird Status & Trends. (Auer et al., 2020).

2.2 The BirdFlow Model

BirdFlow seeks to estimate a track distribution that has single-time-step marginals that approximately match distribution estimates from eBird Status & Trends. However, this alone will not ensure realistic *movement trajectories*. To ensure that modeled movements are reasonable, BirdFlow incorporates additional biological knowledge to approximately minimize the movement cost of individuals. Mathematically, this is done through pairwise marginals of the track distribution: the *pairwise marginal* at week t is a matrix $\boldsymbol{\mu}_{t,t+1}$ with entries $\boldsymbol{\mu}_{t,t+1}(x_t, x_{t+1}) = \Pr(X_t = x_t, X_{t+1} = x_{t+1})$, giving the probability an individual is in location x_t at week t and moves to location x_{t+1} at week $t+1$.

For any track distribution $\mathbf{p}$, let $\boldsymbol{\mu}$ be the vector consisting of all of its single-time-step and pairwise marginals. Because each marginal probability is obtained by summing certain entries of $\mathbf{p}$, there is a matrix A such that $\boldsymbol{\mu} = A\mathbf{p}$; the matrix A is the "marginalization operator". BirdFlow estimates a distribution by solving the following optimization problem:

$$\min_{\mathbf{p}} L_{\text{loc}}(A\mathbf{p}, \hat{\boldsymbol{\mu}}) + \alpha L_{\text{mov}}(A\mathbf{p}). \tag{1}$$

This problem searches over all probability distributions, but the objective only depends on the distribution $\mathbf{p}$ through its marginals $\boldsymbol{\mu} = A\mathbf{p}$. The function $L_{\text{loc}}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ is a location loss function that encourages the single-time-step marginals to match the eBird estimates $\hat{\boldsymbol{\mu}}$. The function $L_{\text{mov}}(\boldsymbol{\mu})$ is a movement loss function to encourage biologically appropriate movements. The scalar α is a non-negative hyperparameter to control the relative weight of the two loss functions.

2.2.1 Loss Functions

For the location loss function, we use the mean squared error between the model marginals and the eBird marginals:

$$L_{\text{loc}}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \frac{1}{T|\mathcal{X}|} \sum_{t=1}^{T} \|\boldsymbol{\mu}_t - \hat{\boldsymbol{\mu}}_t\|_2^2. \tag{2}$$

This is a natural choice because it is a differentiable metric for the distance between the marginals.

The movement loss is a proxy for energetic and fitness costs. A very general movement loss function is:

$$L_{\text{mov}}(\boldsymbol{\mu}) = \sum_{t=1}^{T-1} \sum_{x_t \in \mathcal{X}} \sum_{x_{t+1} \in \mathcal{X}} \boldsymbol{\mu}_{t,t+1}(x_t, x_{t+1}) c(x_t, x_{t+1}), \tag{3}$$

where $c(x_t, x_{t+1})$ is any user-defined cost for transitioning from x_t to x_{t+1}. It is straightforward to see that $L_{\text{mov}}(\boldsymbol{\mu})$ is equivalent to the population mean of the track cost $c(X_1, X_2) + c(X_2, X_3) + \ldots + c(X_{T-1}, X_T)$. One proxy for the energy required for movement this is represented by $c(x_t, x_{t+1}) = d(x_t, x_{t+1})$, the distance between locations x_t and x_{t+1}, in which case $L_{\text{mov}}(\boldsymbol{\mu})$ gives the average total distance moved by an individual. Minimizing this will ensure that the birds will try to minimize the distance they have to fly in order to arrive at their migratory destination. However, we will see later that performance is improved by using $c(x_t, x_{t+1}) = \big(d(x_t, x_{t+1})\big)^{\epsilon}$ for $\epsilon < 1.0$. This transition cost penalizes small distances more than large distances and therefore promotes a model where birds are likely to make fewer large movements instead of many small movements. This behavior is observed in many bird species, so this loss function is motivated by biological knowledge (Newton, 2008).

2.2.2 Optimization over Markov Chains

It is important to notice that our main loss functions L_{loc} and L_{mov} depend only on the marginals of the full model distribution $\mathbf{p}$. This implies that the optimization problem could be converted to one that searches over the space of valid marginals instead of full distributions. However, for some optimal marginals $\boldsymbol{\mu}$, there are arbitrarily many distributions $\mathbf{p}$ which share those marginals, so the problem is under-determined. We follow the principle of maximum entropy to determine what form $\mathbf{p}$ should take. By well known results in the theory of graphical models, the maximum entropy distribution with a certain set of marginals is a graphical model with a dependence graph in which two variables are connected if and only if they co-occur in one of the specified marginal

distributions (Wainwright & Jordan, 2008). With single-time-step marginals and pairwise marginals for adjacent time steps, which are the only marginals required for L_{loc} and L_{mov}, the graph structure is a chain or path on the variables X_1 to X_T, which means the maximum entropy distribution is a Markov chain. For any set of marginals $\boldsymbol{\mu} > 0$, there is a *unique* Markov chain with those marginals. This means we can instead optimize our loss function over the space of non-stationary Markov chains. Specifically, we parameterize an arbitrary Markov chain via parameters $\boldsymbol{\theta}$, introduce a differentiable mapping $\boldsymbol{\mu}(\boldsymbol{\theta})$ from the Markov chain parameters to its marginals, and then minimize the loss function with respect to the Markov chain parameters:

$$\min_{\boldsymbol{\theta}} L_{\text{loc}}\big(\boldsymbol{\mu}(\boldsymbol{\theta}), \hat{\boldsymbol{\mu}}\big) + \alpha L_{\text{mov}}\big(\boldsymbol{\mu}(\boldsymbol{\theta})\big). \tag{4}$$

We emphasize that after solving the problem in Equation (4) to obtain the optimal parameters $\boldsymbol{\theta}$, the resulting Markov chain $\mathbf{p}_{\boldsymbol{\theta}}$ is a global minimizer of the original problem in Equation (1), and has maximum entropy among all minimizers of that problem.

2.2.3 Entropy Regularization

We expect real bird movements to be more variable than those obtained by solving the optimization problems we have introduced for two reasons: (1) our movement cost function only approximates true energy and fitness costs, and (2) a real population is not expected to exactly minimize energy and fitness costs, instead showing substantial individual variation in behavior. To account for these facts, we use an entropy-based regularization term $J(\boldsymbol{\mu}) = -H(\boldsymbol{\mu})$, where H is the Shannon entropy of the distribution with marginals $\boldsymbol{\mu}$, to encourage optimal solutions to have higher entropy. This calculation is generally computationally intractable, but for reasons that are mathematically subtle but well established (Wainwright & Jordan, 2008), the negative entropy of a Markov chain can be written as a function of only the marginals as

$$J(\boldsymbol{\mu}) = \sum_{t=1}^{T} H(\boldsymbol{\mu}_t) - \sum_{t=1}^{T-1} H(\boldsymbol{\mu}_{t,t+1}), \tag{5}$$

where $H(\boldsymbol{\mu}_t)$ and $H(\boldsymbol{\mu}_{t,t+1})$ are Shannon entropies of corresponding marginal distributions, specifically:

$$H(\boldsymbol{\mu}_t) = -\sum_{x_t \in \mathcal{X}} \boldsymbol{\mu}_t(x_t) \log \boldsymbol{\mu}_t(x_t),$$
$$H(\boldsymbol{\mu}_t, \boldsymbol{\mu}_{t+1}) = -\sum_{x_t \in \mathcal{X}} \sum_{x_{t+1} \in \mathcal{X}} \boldsymbol{\mu}_{t,t+1}(x_t, x_{t+1}) \log \boldsymbol{\mu}_{t,t+1}(x_t, x_{t+1}).$$

Since $J(\boldsymbol{\mu})$ also only depends on the single time step and pairwise marginals, we can introduce it to our loss term while maintaining a computationally tractable and well defined optimization problem. The new problem will have the form

$$\min_{\boldsymbol{\theta}} L_{\text{loc}}\big(\boldsymbol{\mu}(\boldsymbol{\theta}), \hat{\boldsymbol{\mu}}\big) + \alpha L_{\text{mov}}\big(\boldsymbol{\mu}(\boldsymbol{\theta})\big) + \beta J\big(\boldsymbol{\mu}(\boldsymbol{\theta})\big), \tag{6}$$

where β is another non-negative hyperparameter.

2.2.4 Optimization Scheme

We now describe the remaining optimization details, including our Markov chain parameterization, the mapping from parameters to marginals, and the optimization algorithm. Let $n = |\mathcal{X}|$ be the number of grid cells. We will make use of the softmax function σ, which operates on a vector $\mathbf{u}$ of n real numbers and produces a normalized probability distribution with ith entry

$$\sigma_i(\mathbf{u}) = \frac{\exp(u_i)}{\sum_{j=1}^{n} \exp(u_j)}. \tag{7}$$

For an $n \times n$ matrix U, we will also write $\sigma(U)$ to indicate the mapping that applies the softmax function separately to each row of U to produce a new $n \times n$ matrix with rows that are non-negative and sum to one.

We parameterize a Markov chain by the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(1,2)}, \boldsymbol{\theta}^{(2,3)}, \ldots, \boldsymbol{\theta}^{(T-1,T)})$, where $\boldsymbol{\theta}^{(1)} \in \mathbb{R}^n$ determines the initial distribution of X_1, and, for each t, the matrix $\boldsymbol{\theta}^{(t,t+1)} \in \mathbb{R}^{n \times n}$ determines the conditional distribution of X_{t+1} given X_t. The total number of parameters in $\boldsymbol{\theta}$ is $N = n + n^2(T-1)$. We use the softmax function to transform from unconstrained parameters to probability distributions: the inital parameters $\boldsymbol{\theta}^{(1)}$ are mapped to the initial marginal distri-

bution $\boldsymbol{\mu}_1 = \sigma(\boldsymbol{\theta}^{(1)})$, and the transition parameters $\boldsymbol{\theta}^{(t,t+1)}$ for all t are mapped to the transition distributions $\mathbf{T}_{t,t+1}(i,j) = P(X_{t+1} = j | X_t = i) = \left(\sigma(\boldsymbol{\theta}^{(t,t+1)})\right)_{i,j}$.

The mapping $\boldsymbol{\mu}(\boldsymbol{\theta})$ to obtain marginals from parameters uses these probability distributions together with additional Markov chain calculations, and is given in Algorithm 1.

Algorithm 1: Differentiable mapping from parameters $\boldsymbol{\theta}$ to marginals $\boldsymbol{\mu}$

Data: $\boldsymbol{\theta}, T$
Result: $\boldsymbol{\mu}$
$\boldsymbol{\mu}_1 \leftarrow \sigma(\boldsymbol{\theta}^{(1)})$
for $t = 1$ *to* $T - 1$ **do**
  $\mathbf{T} \leftarrow \sigma(\boldsymbol{\theta}^{(t,t+1)})$
  $\boldsymbol{\mu}_{t,t+1}(i,j) \leftarrow \boldsymbol{\mu}_t(i)\mathbf{T}(i,j)$ for all $i, j \in \mathcal{X}$
  $\boldsymbol{\mu}_{t+1}(j) \leftarrow \sum_{i \in \mathcal{X}} \boldsymbol{\mu}_{t,t+1}(i,j)$ for all $j \in \mathcal{X}$
end
return $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_T, \boldsymbol{\mu}_{1,2}, \ldots, \boldsymbol{\mu}_{T-1,T})$

Because the parameters $\boldsymbol{\theta}$ are unconstrained and the mapping $\boldsymbol{\mu}(\boldsymbol{\theta})$ of Algorithm 1 is differentiable, we can solve the problem in Equation (6) by gradient descent over $\boldsymbol{\theta} \in \mathbb{R}^N$. There are other methods to solve Problem (1), for example the proximal algorithm of (McKenna et al., 2019); we selected this approach because it is simple, practical, and compatible with current deep learning tool boxes.

2.3 Validation

To validate BirdFlow models and tune hyperparameters, we obtained tracking data for 11 different bird species from the MoveBank repository (Kranstauber et al., 2011) and other data sources (Table 1). All tracks were obtained with high-precision GPS or satellite tracking devices to ensure minimal uncertainty in location estimates. For Argos data, we retained locations with a location class of 1, 2, or 3, indicating estimated error of <1500 m. For each tracking dataset, we subsampled observations to weekly resolution to match the temporal resolution of eBird relative abundance estimates. To do this, we picked the tracking observation closest in time to the date of relative abundance distribution, as long as the observation was within 4 days of the distribution date. We then matched all tracking observations to the corresponding cell of the distribution raster. When tracking data spanned multiple calendar years, we considered the data from each calendar year as a separate track.

2.3.1 Average Log Likelihood

Once the track data were processed, the primary metric we used to evaluate our model is average log-likelihood (ALL). For an observed track $x = (x_1, \ldots, x_T)$ and parameters $\boldsymbol{\theta}$, the log-likelihood is $\log p_{\boldsymbol{\theta}}(x_1, ..., x_T) = \log p_{\boldsymbol{\theta}}(X_1 = x_1) + \sum_{t=1}^{T-1} \log p_{\boldsymbol{\theta}}(X_{t+1} = x_{t+1} | X_t = x_t)$. In practice, many of the tracks span shorter time periods than an entire year and some species have many more tracks than other species. Therefore, to more easily compare results across different species with different numbers of observations, we used the average log-likelihood of bird movements over the total number of observed transitions for that species. Specifically, each track is split into a collection of weekly movements (t, x, x') where t is the starting week, x is the bird's observed location in week t, and x' is the bird's location in week $t+1$, for each week t for which consecutive observations were available. These movements are combined to form the validation dataset $\mathcal{D}$. Then, the average log likelihood is given by

$$\mathrm{ALL}(\mathcal{D}, \boldsymbol{\theta}) = \frac{1}{|\mathcal{D}|} \sum_{(t,x,x') \in \mathcal{D}} \log p_{\boldsymbol{\theta}}(X_{t+1} = x' | X_t = x). \tag{8}$$

This captures how well the model predicts the movement of the the observed birds and it is comparable for tracks of different lengths and species with different numbers of tracks. Because of this, the average log likelihood is a crucial indicator of model quality. To further contextualize this metric, we constructed a baseline from the eBird relative abundance estimates. The baseline approach ignores the initial position x and considers only the log probability of the destination position x' according to the eBird marginal $\hat{\boldsymbol{\mu}}_{t+1}$

$$\mathrm{ALL}_{\mathrm{Baseline}}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(t,x,x') \in \mathcal{D}} \log \hat{\boldsymbol{\mu}}_{t+1}(x'). \tag{9}$$

This corresponds to a model where each bird selects a location at random from the population marginal distribution in each time step, without regard to its location in the previous time step. This random redistribution baseline is not biologically realistic, but it captures the information included in the ground truth marginals alone and can be used to demonstrate how much improvement can be gained by incorporating the biologically-inspired information about pairwise marginals. The values of this baseline for the 11 species we evaluate can be seen in Table 1. Note that an ALL improvement of three nats (the unit for log likelihood) over this baseline means that the average weekly movement

is about 20 times ($e^3 \approx 20$) more likely under our model than under the baseline and the average 52 week track is about 1040 times more likely under our model than under the baseline.

2.3.2 Model Calibration

An important capability of BirdFlow is the ability to make probabilistic forecasts, such as forecasting the distribution of a bird's location at week $t + 4$ given that it was in a certain location in week t. When making forecasts, it is important to understand the model's *calibration*, or the extent to which the variability of the forecasted distributions matches the observed variability of true outcomes (i.e., a tracked bird's locations in the future). To measure calibration, we used the *probability integral transform* (PIT) (Gneiting et al., 2007). This transformation uses the cumulative distribution function (CDF) F of the forecasted distribution for an eventually observed outcome variable z, where z is a scalar. If z is actually distributed according to the forecasted distribution, then $F(z)$ will be a uniform random variable; otherwise, the distribution of $F(z)$ can reveal specific types of miscalibration, such as forecasts being over- or under-dispersed. The distribution of $F(z)$ is assessed by constructing histograms over many pairs of forecasts and observed values.

We were particularly interested in geographic calibration, that is, the calibration of forecasts of a bird's location in future weeks given its current location. Since PIT diagnostics apply to scalar quantities, we assessed calibration of forecasts for north-south positions and east-west positions separately. For example, for any grid cell $x \in \mathcal{X}$, let $u(x)$ be its east-west position, and let $U_t = u(X_t)$ be the random variable for the east-west position of a bird at time t. Conditioned on the bird's location x at time t, the CDF of the forecast distribution for U_{t+1} is

$$F_t(u|x) = \Pr_{\boldsymbol{\theta}}(U_{t+1} \leq u | X_t = x). \tag{10}$$

The PIT transform computes the values $F_t(u_{t+1}|x_t)$ for all observed triples of the form (t, x_t, u_{t+1}) where t is a time index, x_t is the bird's grid cell at time t, and u_{t+1} is the east-west position at time $t + 1$.

However, since our map in discrete, we must modify this procedure to correctly account for the probability assigned to discrete outcomes, specifically, the nonzero probability that $U_t = u$ in

Equation (10). For discrete variables it is common practice to use the *randomized PIT* transform

$$F_t(u|x) = \Pr_{\boldsymbol{\theta}}(U_{t+1} < u | X_t = x) + \nu \Pr_{\boldsymbol{\theta}}(U_{t+1} = u | X_t = x), \tag{11}$$

where ν is a random variable chosen uniformly in $[0, 1]$. This randomized PIT is evaluated in the same way as the standard PIT.

Since each observed $F_t(u_{t+1}|x_t)$ should be uniformly distributed, we can make a histogram of these values and check for uniformity. We followed the same procedure to evaluate north-south calibration, the only difference is that we use the north-south position $V_t = v(X_t)$ of the grid cell instead of the east-west position $U_t = u(X_t)$.

2.4 Experiments

We conducted experiments to assess BirdFlow's predictive performance, comparisons to baseline models, and sensitivity to hyperparameters.

2.4.1 Hyperparameter Grid Search

We addressed several questions by performing a grid search of model hyperparameters and evaluating the resulting models. The three hyperparameters we are interested in are α, β and ϵ (the weights on the movement loss L_{mov}, the entropy regularization term J, and the distance exponent applied to the cost function c, respectively). Initial experiments showed that the model is less sensitive to the choice of α than other hyperparameters and that a value of $\alpha \ll 1$ consistently performed well. So, to reduce the search space, we fixed $\alpha = 0.005$ and trained models with different values of β and ϵ. Conceptually, this places a very high relative weight on the location loss function, which means that BirdFlow weekly distributions will closely match the eBird estimates; then, subject to that "constraint", the model will minimize the movement costs and entropy costs. We trained the model using every combination of values for $\beta \in \{0.0, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006\}$ and $\epsilon \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. We believe this range captures most reasonable values for these hyperparameters because none of the models perform best with the extremal values and the performance seems to vary smoothly as the hyperparameters change. We compared the average log-likelihoods of the resulting models to determine which settings of the hyperparameters led to

models that best explain the observed tracks and to understand how hyperparameters affect model quality.

The first question we investigated with the grid search results is the effect of the entropy regularization term and the distance exponent on model quality. We performed an ablation study that compares four model configurations for each species. We compared models with no entropy regularization to models with entropy regularization and models with distance power equal to one to models with distance power less than one. This lets us evaluate how impactful those components are for model quality in isolation and also together.

The second question we investigated with the grid search results was the sensitivity of the model to the choice of hyperparameters. We examined model performance across two methods of hyperparameter selection. First, we tuned each species model by determining hyperparameter values that gave the best average log likelihood for that species; we refer to these as "tuned" model settings. Second, we examined how well each species model performed using hyperparameters chosen based on performance on all *other* species, excluding the focal species. These "leave one out" (LOO) parameters for a species are the hyperparameter values from the grid search results that give the best average log likelihood across all other bird species. We then compared performance using both methods of hyperparameter selection. In particular, wanted to know whether the LOO settings performed well, or if species-specific tuning was required for acceptable performance.

2.4.2 Entropy Calibration

We investigated the effect of the entropy regularization term on the calibration of model predictions. Intuitively, we would expect that if we increase the weight of the entropy regularization term, the joint marginals will become more diffuse. In order to evaluate this, we computed the PIT score for each of the transitions for the American Woodcock (*Scolopax minor*) under several versions of the model and plotted the score in a histogram. A convex histogram indicates under-dispersion, and a concave histogram indicates over-dispersion. A uniform (flat) histogram indicates optimal dispersion and a well-calibrated model.

2.4.3 k-Week Forecasting

We also investigated model performance for the task of k-week ahead forecasting for $k > 1$ to understand how prediction accuracy decreases with time horizon. The procedure for computing the average log-likelihood was slightly modified to compute the average log-likelihood for forecasts k weeks into the future. Instead of splitting the tracks into bird movements in consecutive weeks, tracks were split into positions of a single bird k weeks apart, that is, we created a data set $\mathcal{D}_k$ with triples of the form (t, x, x') where x was the bird's position at time t and x' was it's position at time $t + k$. Then, the model and baseline were evaluated on how well they predicted these positions. These modified average log likelihoods were computed as follows

$$\mathrm{ALL}(\mathcal{D}_k, \boldsymbol{\theta}) = \frac{1}{|\mathcal{D}_k|} \sum_{(t,x,x') \in \mathcal{D}_k} \log p_{\boldsymbol{\theta}}(X_{t+k} = x' | X_t = x), \tag{12}$$

$$\mathrm{ALL}_{k,\mathrm{Baseline}}(\mathcal{D}_k) = \frac{1}{|\mathcal{D}_k|} \sum_{(t,x,x') \in \mathcal{D}_k} \log \hat{\boldsymbol{\mu}}_{t+k}(x'). \tag{13}$$

2.5 Demonstration

To demonstrate the inferences one can draw from BirdFlow models, we generated and evaluated model outputs for American Woodcock. We chose this species because we had high-quality validation data from GPS-tracked birds (Table 1), and because it represents a bird species of approximately average body size from among our sample of tracked species. In order to select the hyperparameters, we performed a finer grid search around the best parameters from the original coarser grid search. We selected the model from the finer grid search with the best average log likelihood. From the trained woodcock flow model, we simulated 5000 migration trajectories, representing plausible routes of individual woodcocks through the year. From these simulated trajectories, we calculated three measures of the spring migration: (1) the distribution of migration departure timing, (2) the distribution of migration arrival timing, and (3) the migratory connectivity of breeding populations. We calculated the distributions of spring migration departure and arrival dates using the *alongTrackDistance* function in the `geosphere` R package (Hijmans, 2017), assessing when each simulated bird moved at least 100 km from its starting location and arrived within 100 km of its ending location. To infer migratory connectivity, we used simulated tracks from the fall migration. We subselected

trajectories that began in the northwest and northeast sectors of the woodcock breeding range to compare the modeled connectivity of populations originating from different parts of the breeding range. Then, we compared the modeled non-breeding destinations of individuals from these two groups, asking whether the model inferred different wintering areas for these two subpopulations.

We generated visual representations of modeled tracks alongside actual GPS-tracked individuals to compare modeled trajectories to observed migration routes. For each observed track, we generated 2500 simulated trajectories originating at the same location as the GPS-tracked bird and continuing for the same duration. Then, we plotted observed and simulated routes together.

Finally, we produced visual representations of short-term forecasts. For observed GPS-tracked birds, we extracted the future probability distribution of a bird at a given location and time at 3, 6, or 12 weeks into the future. Then, we compared the predicted movement forecast to observed movements.

Table 1: Summary of tracking data used. A "track" is defined as the path of an individual during a calendar year. Individuals monitored for multiple years will therefore have multiple tracks.

Species	# Indiv.	# Tracks	Mean weeks/track	Baseline ALL	References
American Woodcock	67	107	16.8	-6.52	[a]
Black-bellied Plover	15	38	28.3	-5.53	[b]
Broad-winged Hawk	20	35	18.1	-5.93	[c]
Blue-winged Teal	42	51	21.8	-5.43	[d]
Long-billed Curlew	91	240	34.0	-4.86	[e]
Osprey	230	415	21.6	-5.74	[f]
Swainson's Hawk	43	76	17.0	-4.96	[g]
Tundra Swan	50	176	34.2	-5.42	[h]
Turkey Vulture	19	76	35.6	-6.95	[i]
Whimbrel	32	62	28.3	-5.66	[j]
Wood Thrush	20	37	12.0	-5.93	[k]

[a] Moore et al., 2021a, 2021b.
[b] Harrison, 2022.
[c] R. McCabe and Goodrich, 2022; R. A. McCabe et al., 2020.
[d] Ramey et al., 2019.
[e] Carlisle, 2022.
[f] Bierregaard, 2019; Jensen, 2018; Martell and Douglas, 2019; Martell et al., 2001.
[g] Kochert, 1998; Kochert et al., 2011.
[h] Ely et al., 2020.
[i] Bildstein et al., 2014; Dodge et al., 2014.
[j] Tibbitts et al., 2018.
[k] Stanley et al., 2021.

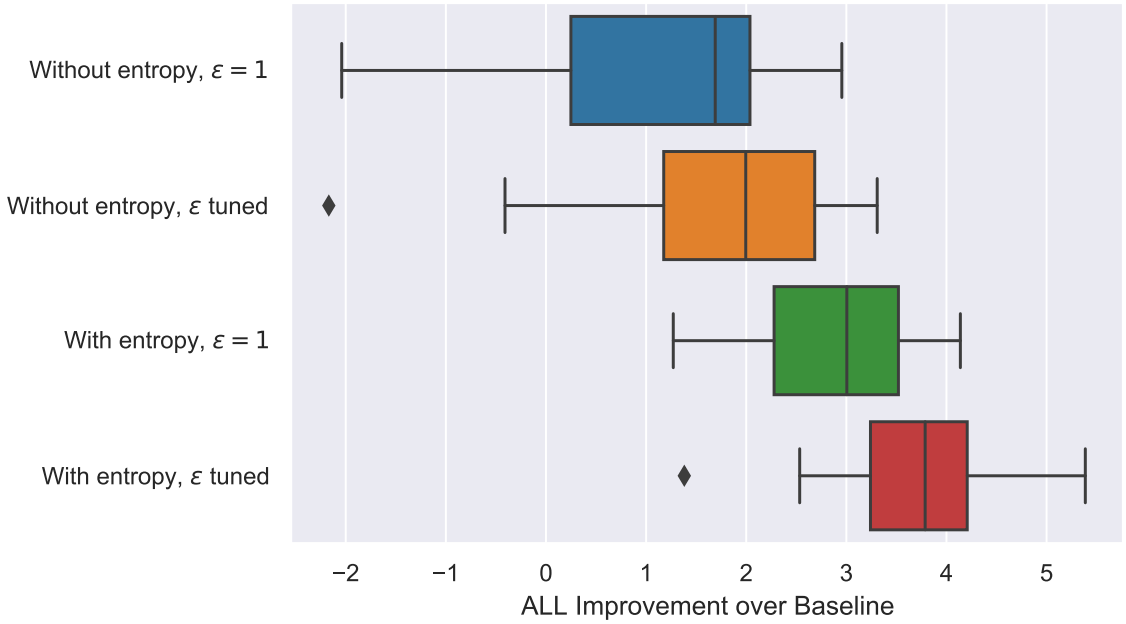

Figure 2: Model type ablation study. For each model version, the distribution of performance (average-log likelihood) improvement over the baseline model for 11 species is displayed as a box and whisker plot. Whisker length is at most 1.5 times the interquartile range, with outliers shown as diamonds. The model with non-zero entropy and tuned distance exponent (ϵ) achieves higher log-likelihoods than other model versions.

3 Results

We now present results of our model validation experiments and demonstration of model outputs for the American Woodcock.

3.1 Validation

Figure 2 shows the results of the ablation study comparing the performance of different model configurations on tracked wild birds. All BirdFlow model types performed better than a baseline model that incorporated only weekly species relative abundance. Models with non-zero entropy regularization and tuned distance penalty exponent (ϵ) performed best overall, followed by models with entropy regularization and $\epsilon = 1$.

Figure 3 assesses sensitivity to hyperparameters. For most species, the "leave one out" (LOO) parameters, which were selected using only the validation tracks from *other* species, performed nearly as well as models tuned using tracking data from that species. The difference in average

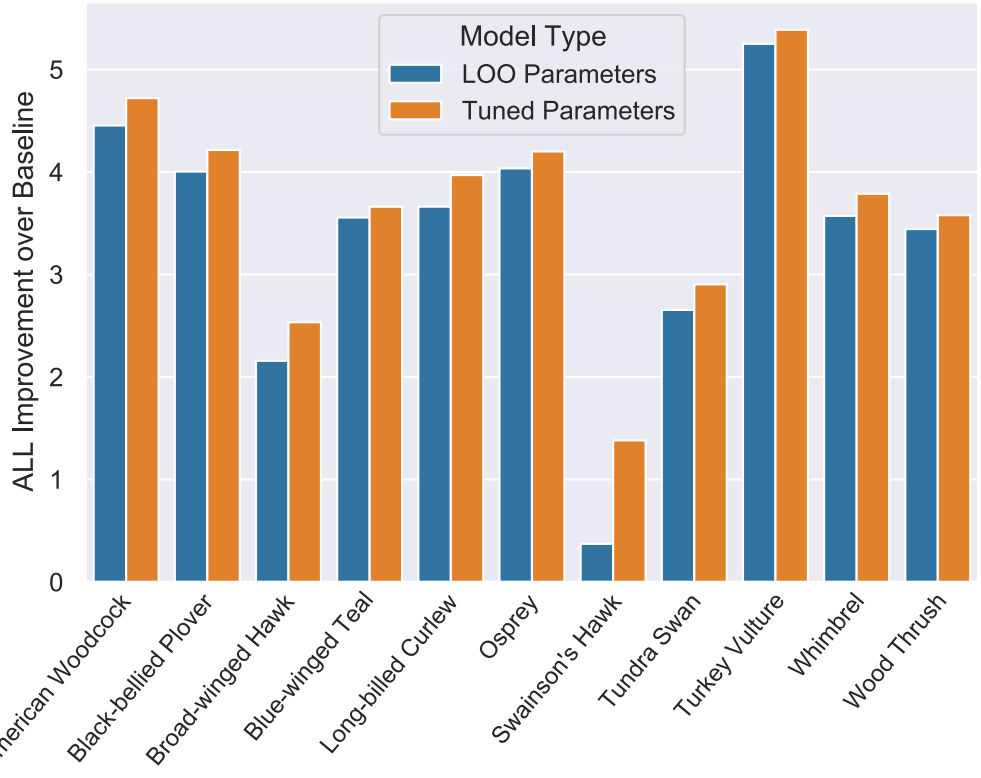

Figure 3: Parameter sensitivity. Bars show improvement over the baseline model for BirdFlow models with "leave one out" parameters (selected using validation data from other species) vs "tuned" parameters (selected using validation data from the target species). Performance is measured as average log-likelihood of one-week transitions.

log-likelihood between the LOO parameters and the tuned parameters is small compared to the difference between either setting and the baseline. The most notable exception is Swainson's Hawk, where the LOO parameters perform much worse than the tuned parameters.

Figure 4 shows the effect of entropy regularization on model calibration, which was substantial. PIT histograms for four versions of the American Woodcock model are shown, with distance exponent (ϵ) fixed to 0.3 and varying entropy regularization weights. The PIT histograms are closest to uniform for entropy weights of 0.0005 and 0.001, which indicates the best model calibration. Entropy weights that are higher or lower strongly negatively impact calibration. With zero entropy, too many observations occur at the extremes of the forecast distribution, which indicates underdispersed forecasts. With high entropy, too few observations occur at the extremes of the forecast distribution, which indicates overdispersed forecasts.

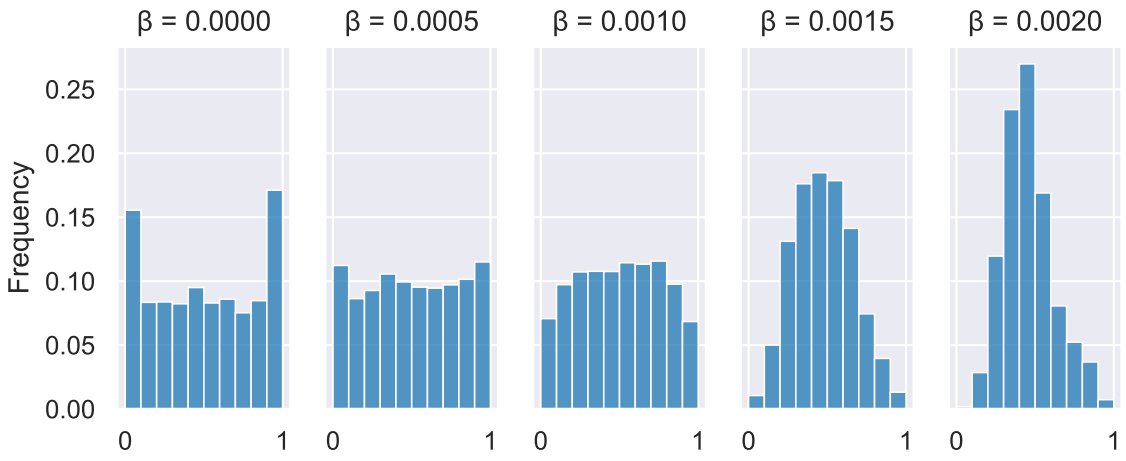

Figure 4: Effect of entropy regularization on model calibration. Randomized PIT histograms are shown for 1-week ahead forecasts of American Woodcock east-west positions for models trained with different entropy weights. Histograms that are nearly uniform indicate well calibrated models.

Figures 5 and 6 show model performance relative to forecast horizon (in weeks). We identified the best-performing model from the hyperparameter grid search (using average log-likelihood) for every species and evaluated the improvement over the baseline for k-week-ahead average log-likelihood for all forecast horizons k from 1 to 17. Figure 5 displays those results for each species. For every species, the improvement over the baseline decreases with k. However, there is substantial variation: some species continue to perform substantially better than the baseline up to a forecast horizon of 17 weeks, while others approach the performance of the baseline. We also compared the tuned woodcock parameters to the LOO woodcock parameters and the baseline in an absolute sense (Figure 6). The gap between the tuned parameters and the LOO parameters is small at first, but increases with forecast horizon, which indicates that the tuned model performs better relative to the LOO model at larger horizons. Both models performed better than baseline model at all prediction horizons tested.

3.2 Demonstration

We demonstrated example model outputs from our trained model of American Woodcock movements. Simulated spring migration trajectories (Figure 7a) allowed us to estimate the distributions of migration departure and departure timing (Figure 7b,c). Simulated woodcocks left their wintering

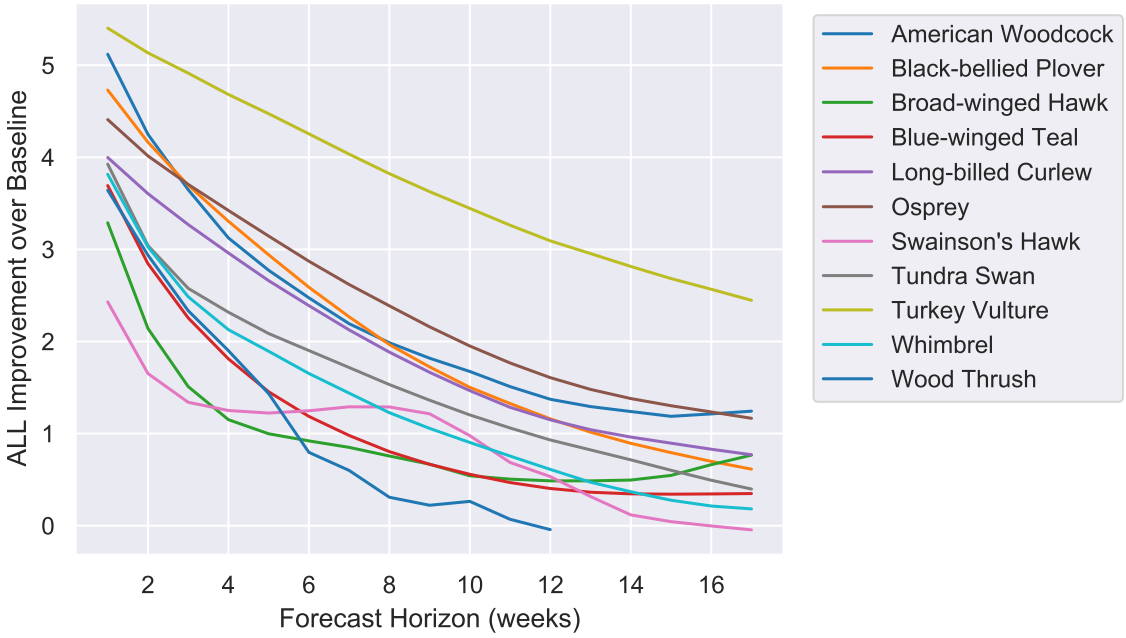

Figure 5: Performance by forecast horizon. The plot shows the improvement in log-likelihood over the baseline model vs. forecast horizon for each species. BirdFlow hyperparameters for each species are selected using 1-week-ahead average log-likelihood for that species.

grounds between mid-January and early March, arriving largely between early March and early May. Our model inferred meaningful differences in migratory connectivity between woodcocks breeding in the northeast US and in the midwest (Figure 7d). The model inferred that woodcocks breeding in the northeast primarily spend the winter in the mid-Atlantic and southeast. In contrast, the model inferred that woodcocks breeding in the midwest winter primarily along the western Gulf Coast.

We generated simulated migration trajectories alongside observed routes of GPS-tracked birds. The observed routes were generally well-represented among simulated trajectories (Figures 7e,f,g and 9). Similarly, short-term conditional forecast distributions also successfully captured observed movements (Figures 7h,i,j and 10). More of the plots containing simulated trajectories and short-term forecasts can be seen in the appendix Figures 9 and 10.

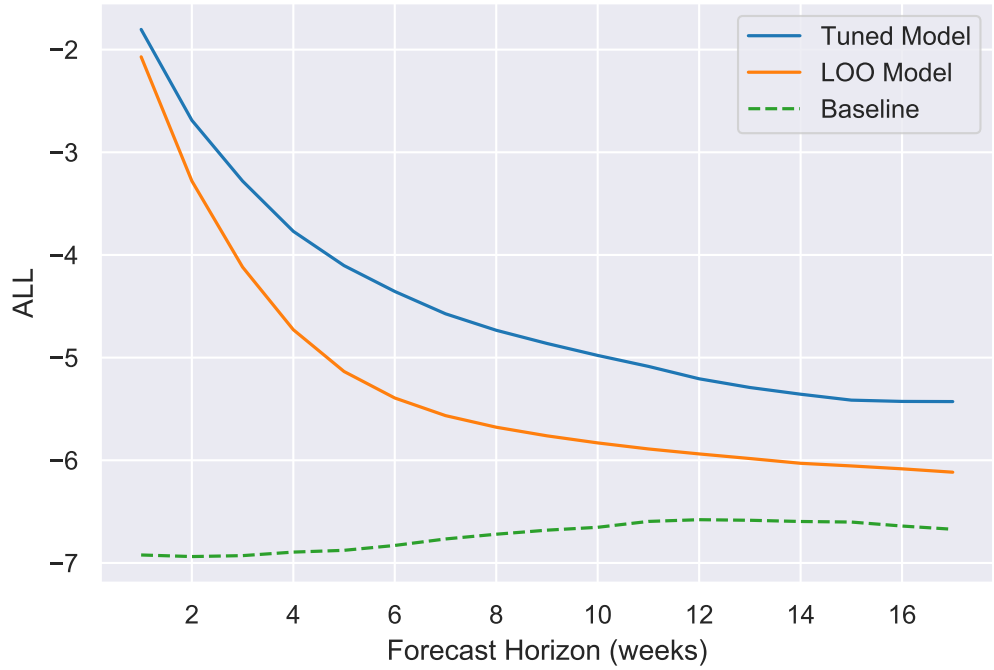

Figure 6: Performance by forecast horizon for American Woodcock. The plot shows the log-likelihood at various prediction horizons of the best performing American Woodcock model, the LOO American Woodcock model, and the baseline model. The baseline model changes with k because the number of observed transitions spanning k weeks for tracked birds decreases with k.

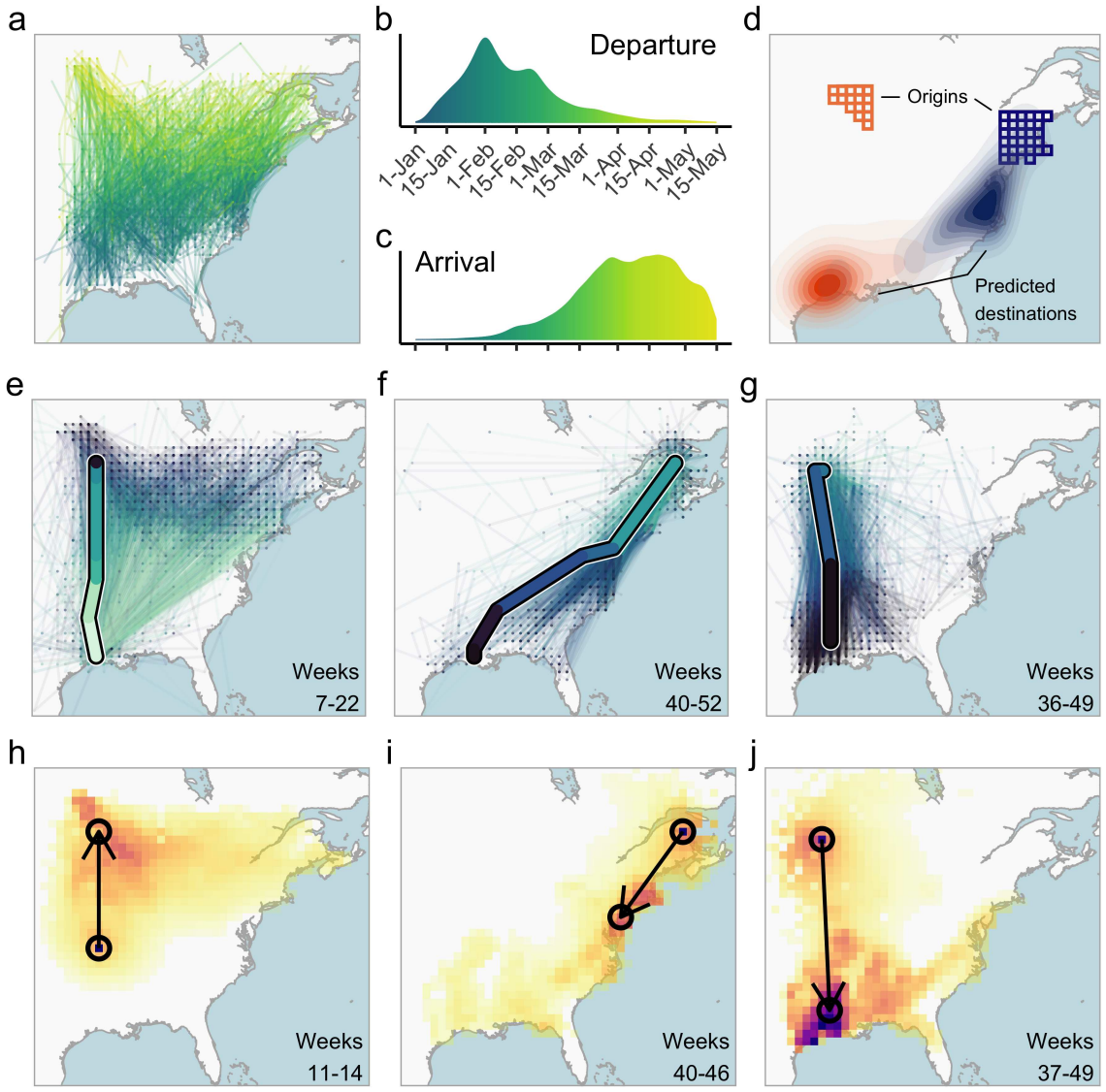

Figure 7: Demonstration of model inferences. Shown are derived model outputs from American Woodcock *Scolopax minor*. (a) Simulated spring migration trajectories (n=1000). (b) Timing of spring migration departure and (c) arrival derived from simulated trajectories. (d) Migratory connectivity: square cells show breeding origins of individuals in the northwest (orange) and northeast (blue) parts of the breeding range. Filled density contours show the predicted wintering distributions of individuals breeding in those respective regions. (e-g) Observed movements of GPS-tracked woodcocks (single thick path) and simulated trajectories (thin paths) for 2500 simulated birds originating at the same starting location as observed birds. (h-j) Conditional forecast distributions: each heatmap shows the predicted movement distribution of a GPS-tracked individual originating within the circle at the base of the arrow. Darker colors indicate a higher predicted likelihood of movement to that area. The point of the arrow shows the observed ending location. Shown are examples of 3-week (h), 6-week (i), and 12-week (j) conditional forecasts.

4 Discussion

Our probabilistic BirdFlow models accurately inferred individual movement behavior using weekly relative abundance estimates from citizen science data. For all species studied, our movement model predicted the movements of GPS- and satellite-tracked birds substantially better than a baseline model that included only the weekly species distribution maps. A set of general (LOO) model parameters performed well across nearly all species, suggesting that BirdFlow could be used to accurately infer movements without any tracking data inputs in many species. Models fine-tuned with tracking data were most accurate, but the difference between LOO models and tuned models was small compared to the improvements over the baseline model. Overall, our results show that by combining relative abundance estimates derived from citizen science observations with models of movement costs, it is possible to infer individual movement behavior in a way that is substantially more accurate than baseline models.

Impacts of model hyperparameters Addition of an entropy regularization term was crucial for proper model calibration, and using a distance exponent less than one in the movement cost term was important for producing realistic movement patterns. When these components were removed (labeled "Without entropy, $\epsilon = 1$" in Figure 2), several species under-performed the baseline. The entropy regularization term seems to be particularly important, because its inclusion alone ensures that the model outperforms the baseline for every single species. However, inclusion of both components resulted in the best performance.

Is model tuning required? One of the important advantages of our modeling approach is that track data are not explicitly needed for training, although track data proved useful for validating the model and tuning the hyperparameters. Our sensitivity experiment shows that the difference between the LOO parameters and the tuned parameters was usually small compared to the difference between either setting of the parameters and the baseline. However, the results from the Swainson's Hawk indicate that hyperparameter settings will not translate equally well from species to species. Of the species evaluated, Swainson's Hawk migrates the longest distances, with many individual traveling from northern North America to southern South America. This may be the reason why the hyperparameters did not transfer as well. Further work is needed in order to fully determine

under what conditions hyperparameter settings will transfer well and how to select hyperparameters when no tracks are available. We hypothesize that hyperparameters that work well for other ultra-long-distance migrants may transfer better to Swainson's Hawk.

Importance of proper calibration Average log-likelihood is not the only metric by which we measured model performance; we are also interested in model calibration and how the calibration of the model can be modified. Our results show a direct relationship between the entropy regularization term and the dispersion of model predictions. Insufficient entropy will result in an over-confident model, while excess entropy will lead to biologically implausible movement patterns. In choosing an entropy regularization weight, a user could use a set of observed tracks, as we did in this study. If no observed tracks are available, our results suggest that substituting hyperparameters from a similar species or group of species may suffice at a starting point. Users can also determine based on their application if they would prefer to err on the side of over-dispersion or under-dispersion and choose an entropy weight based on that preference.

Short-term movement forecasting One capability of the BirdFlow model is to predict the likely position of a bird several weeks into the future, given a starting time and location. The further into the future the prediction is made, the more uncertainty about the birds position accumulates. It is therefore encouraging that the k-week forecasting experiment showed that the model performs consistently better than the baseline even many weeks into the future.

Data quality and loss functions A crucial component for the performance of BirdFlow is the match between the marginals encouraged by the loss function and the true marginals of the target population. In order to ensure a good match, the ground truth marginals used for training must accurately reflect the actual distribution of the species in question. These ground truth marginals could be derived from raw observational data but we would expect spacial and temporal gaps and noise to lead to low quality ground truth marginals. Similarly, ground truth marginals could be derived from occurrence models but, the probability of occurrence does not directly encode the proportion of the population at a location so we would not expect this to match the true marginals well. The other terms in the loss function should reflect the biological properties of the target

population as accurately as possible. In our case, the movement loss reflects the energy cost of moving and the different values of the distance exponent encodes how much a species will tend to make few large movements compared to many small movements. The entropy regularization term encodes that a real population is not expected to exactly minimize energy and fitness costs, instead showing substantial individual variation in behavior.

Limitations and open questions There are several limitations and open questions that should guide short-term applications and future method development for BirdFlow. While BirdFlow shows promise for broad-scale application to many species, including those without tracking data, the extent to BirdFlow will generalize to the thousands of other migratory species is unknown, and practitioners should exercise caution. It is best practice, when possible, to validate BirdFlow results using tracking data, either from the target species or a closely related species. In the short-term, we expect it will be beneficial to have a human expert vet models and select parameters based on visual examination of model outputs. Over time, the use of BirdFlow is likely to lead to a set of best practices and better understanding of its generalization capabilities. Even when tracking data are available, we found that selecting a model based only on log-likelihood did not always lead to synthetic routes that were the most consistent with biological knowledge. In particular, there can be a difficult tradeoff where models trained with low entropy learn distributions that are far too narrow but models trained with a higher entropy learn distributions that send birds in unrealistic directions (see Figure 8). Choosing models via their average log-likelihood sometimes favors an entropy weight that produces routes that are more variable than expected. This suggests that there may be a better way to encode biological knowledge about variability in migration paths: intuitively, a high entropy distribution will be very uniform and lead to variability in all directions; there may be some other loss function which could encourage variability only in desirable directions. Designing these sorts of loss terms which better encode our biological knowledge is an interesting direction for future work.

The loss functions employed by BirdFlow lead to a Markovian movement model, which has several known limitations. Because the distribution of future locations depends only on a bird's current location, the model treats all birds in the same location at the same time identically: their future routes may diverge, but only due to randomness of transitions, and not due to long-term "memory". This means, for example, that the current implementation of BirdFlow cannot model year-to-

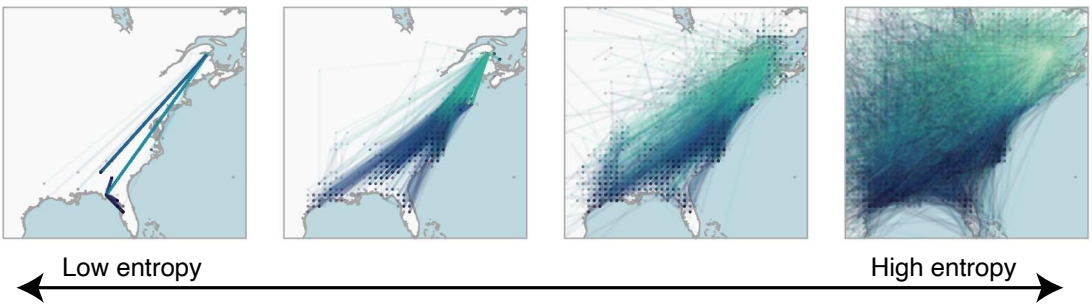

Figure 8: The effect of entropy levels on American Woodcock model samples. The models were trained with the distance exponent (ϵ) fixed to 0.3 and with the entropy weights (0.00, 0.01, 0.02, 0.04). The plot displays 2500 tracks sampled from each model.

year site fidelity. That is, simulated full-year routes are unlikely to return to the same location one year later. For this reason, we currently recommend applying BirdFlow for single migration seasons. For the same reasons, BirdFlow cannot differentiate between individuals of different subpopulations that have different migration strategies but coincide both spatially and temporally. For example, BirdFlow could not correctly model two distinct subpopulations that cross through the same location at the same time. We believe this limitation is minor in practice, because populations with different migration strategies are often separated either spatially or temporally. Future methodological research could incorporate site fidelity and other considerations into the BirdFlow model. Conceptually, site fidelity could be modeled by adding loss functions that depend on the marginal distribution of a bird's location at a given time together with its location one year later. Other phenomena could be modeled with loss functions on other marginal distributions—based on either biological knowledge or additional data sources such as banding data. However, it is known that such loss terms will increase the computational difficulty of solving the BirdFlow optimization problem, so computational research will be a key part of this future work. BirdFlow could also be applied to study inter-annual variation with the use of several relative abundance estimates which each pertain to different years or groups of years.

Related work BirdFlow builds on prior methods for learning a probability distribution from evidence about its marginal distributions. Notably, we previously developed *collective graphical models* (CGMs) (Sheldon & Dietterich, 2011), which are a general formalism for learning the parameters

of a probabilistic graphical model from noisy aggregate observations. CGMs were inspired by bird migration modeling (Sheldon et al., 2008), and later used to model human population flows (Akagi et al., 2018; Iwata et al., 2017). Inference and estimation in CGMs is computationally challenging (Sheldon et al., 2013), but many approximations have been proposed (Sheldon et al., 2013; Singh et al., 2020; Sun et al., 2015; Vilnis et al., 2015; Yasunori et al., 2020).

A similar problem setting arises in privacy-preserving data analysis, where noisy aggregate population statistics are released by a central agency such as a census bureau to provide information about population demographics while ensuring privacy of individuals (Dwork et al., 2006). From these noisy, aggregate statistics, an analyst wishes to estimate a full distribution over demographic variables. Private-PGM (McKenna et al., 2019) is a recent algorithmic framework we developed for this setting, which has been successful as a key component of winning entries in privacy competitions (www.nist.gov, 2018, 2020) and of mechanisms for releasing private synthetic data (Cai et al., 2021).

BirdFlow builds on the conceptual underpinnings of Private-PGM, rather than CGMs, to estimate bird movement models. One key difference compared to CGMs is that BirdFlow and Private-PGM ignore sampling variability due to the population being drawn from an underlying distribution. This is appropriate for large populations, where sampling error is smaller in magnitude than measurement noise, and leads to simpler estimation algorithms. A second key difference is that in BirdFlow the model output is a probabilistic model (a Markov chain), while in CGMs the model output is a reconstruction of population flows. While this difference is minor mathematically (one object can be converted to the other), it is a significant practical and conceptual advance to treat the model output as a probabilsitic model from which we can construct synthetic routes and create forecasts and many other products. Finally, although CGMs were motivated by bird migration modeling, the current study is the first in-depth examination of the capabilities of any of these methods to accurately model bird migration at this scope, including many species, validation using real tracks, and tuning of of key parameters such as entropy regularization and distance exponent to obtain biologically realisitic model outputs.

Recently, Somveille et al. (2021) developed a closely related model for inferring migratory connectivity from breeding and non-breeding distributions and cost-based estimation. Two key differences

are that: (1) BirdFlow models the entire track $X_1, \ldots, X_T$ instead of just the starting and ending locations, (2) BirdFlow incorporates entropy regularization to combat problems that arise with exact cost minimization, including too little variability in inferred routes (cf. Figures 4 and 8).

Future applications We show that it is possible to accurately model animal movement solely from aggregate data—in this case, from citizen science observations. We demonstrate how one can extract a range of behavioral inferences from BirdFlow models, including migratory routes, timing, connectivity, and forecasts. This modeling framework has the potential to advance migration ecology research in a variety of ways, for example through inferences of population migratory connectivity (i.e. where a given breeding population spends the non-breeding period), stopover behavior, and responses to global change. In addition, movement researchers with access to even a small amount of tracking data could use our model to infer individual behavior across the species' entire range—in essence, combining insights from citizen science data with direct tracks to achieve a more complete understanding of animal movements than either approach can alone. Applications exist well beyond ecology, and include movement forecasting to inform disease surveillance (e.g. for avian influenza) and ensure safer aviation. Finally, BirdFlow can raise public awareness about biodiversity and ecosystem health by providing a tool for outreach to engage scientists, bird-watchers, policy-makers, and the general public.

5 Acknowledgements

We are grateful to the eBird Status & Trends team. We thank Tom Auer and Adriaan Dokter for assistance and feedback on our work, and Rob Bierregaard, Autumn-Lynn Harrison, and Michael N. Kochert for permission to use tracking data in this study. This material is based upon work supported by the National Science Foundation under Grant Nos. 1522054 and 1661259. The work of BMVD was supported by a Cornell Presidential Postdoctoral Fellowship. We thank the Leon Levy Foundation; The Wolf Creek Charitable Foundation; NSF DBI-1939187. Computing support was provided by the NSF CNS-1059284 and CCF-1522054, and the Extreme Science and Engineering Discovery Environment (XSEDE) NSF ACI-1548562, through allocation TG-DEB200010 run on Bridges at the Pittsburgh Supercomputing Center. Additional computing efforts were performed with equipment

obtained under a grant from the Collaborative R&D Fund managed by the Massachusetts Technology Collaborative.

References

Akagi, Y., Nishimura, T., Kurashima, T., & Toda, H. (2018). A fast and accurate method for estimating people flow from spatiotemporal population data. *IJCAI*, 3293–3300.

Auer, T. [Tom], Fink, D., & Strimas-Mackey, M. (2020). *Ebirdst: Tools for loading, plotting, mapping and analysis of ebird status and trends data products* [R package version 0.2.0]. https://cornelllabofornithology.github.io/ebirdst/

Bairlein, F. (2016). Migratory birds under threat. *Science*, *354*(6312), 547–548. https://doi.org/10.1126/science.aah6647

Bauer, S., Shamoun-Baranes, J., Nilsson, C., Farnsworth, A., Kelly, J. F., Reynolds, D. R., Dokter, A. M., Krauel, J. F., Petterson, L. B., Horton, K. G., & Chapman, J. W. (2019). The grand challenges of migration ecology that radar aeroecology can help answer. *Ecography*, *42*(5), 861–875. https://doi.org/https://doi.org/10.1111/ecog.04083

Bierregaard, R. (2019). *Movebank: Osprey bierregaard north and south america*. Retrieved February 16, 2022, from https://www.movebank.org/cms/webapp?gwt_fragment=page=studies,path=study8868155

Bildstein, K., Barber, D., & Bechard, M. J. (2014). *Data from: Environmental drivers of variability in the movement ecology of turkey vultures (cathartes aura) in north and south america*. http://doi.org/10.5441/001/1.46ft1k05

Cai, K., Lei, X., Wei, J., & Xiao, X. (2021). Data synthesis via differentially private Markov random fields. *Proceedings of the VLDB Endowment*, *14*(11), 2190–2202.

Carlisle, J. (2022). *Movebank: Long-billed curlew migration from the intermountain west*. Retrieved February 16, 2022, from https://www.movebank.org/cms/webapp?gwt_fragment=page=studies,path=study42451582

Dodge, S., Bohrer, G., Bildstein, K., Davidson, S. C., Weinzierl, R., Bechard, M. J., Barber, D., Kays, R., Brandes, D., Han, J., & Wikelski, M. (2014). Environmental drivers of variability in the movement ecology of turkey vultures (cathartes aura) in north and south america [Publisher: Royal Society]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1643), 20130195. https://doi.org/10.1098/rstb.2013.0195

Dokter, A. M., Farnsworth, A., Fink, D., Ruiz-Gutierrez, V., Hochachka, W. M., La Sorte, F. A., Robinson, O. J., Rosenberg, K. V., & Kelling, S. (2018). Seasonal abundance and survival of north america's migratory avifauna determined by weather radar [Number: 10 Publisher: Nature Publishing Group]. *Nature Ecology & Evolution*, *2*(10), 1603–1609. https://doi.org/10.1038/s41559-018-0666-4

Dunn, P. O., & Møller, A. P. (Eds.). (2019). *Effects of climate change on birds (2nd edition)*. Oxford University Press.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Third Theory of Cryptography Conference*.

Ely, C. R., Terenzi, J., Tibbitts, L., & Douglas, D. C. (2020). Tracking data for tundra swan (cygnus columbianus) [Medium: csv,zip Type: dataset]. https://doi.org/10.5066/P9KBR79C

Fink, D., Auer, T., Johnston, A., Strimas-Mackey, M., Robinson, O., Ligocki, W., Hochachka, W. M., Wood, C., Davies, I., Iliff, M. J., & Seitz, L. (2020). *eBird status and trends, data version: 2019; released: 2020*. Cornell Lab of Ornithology. Ithaca, New York. https://doi.org/10.2173/ebirdst.2019

Fink, D., Auer, T. [Tom], Johnston, A., Ruiz-Gutierrez, V., Hochachka, W. M., & Kelling, S. (2020). Modeling avian full annual cycle distribution and population trends with citizen science data [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/eap.2056]. *Ecological Applications*, *30*(3), e02056. https://doi.org/https://doi.org/10.1002/eap.2056

Fink, D., Damoulas, T., Bruns, N. E., Sorte, F. A. L., Hochachka, W. M., Gomes, C. P., & Kelling, S. (2014). Crowdsourcing meets ecology: Hemisphere-wide spatiotemporal species distribution models [Number: 2]. *AI Magazine*, *35*(2), 19–30. https://doi.org/10.1609/aimag.v35i2.2533

Fink, D., Damoulas, T., & Dave, J. (2013). Adaptive spatio-temporal exploratory models: Hemisphere-wide species distributions from massively crowdsourced eBird data. *AAAI*.

Fink, D., Hochachka, W. M., Zuckerberg, B., Winkler, D. W., Shaby, B., Munson, M. A., Hooker, G., Riedewald, M., Sheldon, D., & Kelling, S. (2010). Spatiotemporal exploratory models for broad-scale survey data [tex.publisher: Ecological Society of America]. *Ecological Applications*, *20*(8), 2131–2147.

Fraser, K. C., Davies, K. T. A., Davy, C. M., Ford, A. T., Flockhart, D. T. T., & Martins, E. G. (2018). Tracking the conservation promise of movement ecology. *Frontiers in Ecology and Evolution*, *6*. Retrieved February 15, 2022, from https://www.frontiersin.org/article/10.3389/fevo.2018.00150

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 243–268.

Harrison, A. (2022). *Movebank: MCP black-bellied plover alaska*. Retrieved February 16, 2022, from https://www.movebank.org/cms/webapp?gwt_fragment=page=studies,path=study77248725

Hijmans, R. J. (2017). *Geosphere: Spherical trigonometry*. R Package. https://CRAN.R-project.org/package=geosphere

Iwata, T., Shimizu, H., Naya, F., & Ueda, N. (2017). Estimating people flow from spatiotemporal population data via collective graphical mixture models. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, *3*(1), 1–18.

Jensen, B. (2018). *Movebank: Pandion haliaetus osprey - SouthEast michigan*. Retrieved February 16, 2022, from https://www.movebank.org/cms/webapp?gwt_fragment=page=studies,path=study10204361

Johnston, A., Fink, D., Reynolds, M. D., Hochachka, W. M., Sullivan, B. L., Bruns, N. E., Hallstein, E., Merrifield, M. S., Matsumoto, S., & Kelling, S. (2015). Abundance models improve spatial and temporal prioritization of conservation resources [_eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf 1826.1]. *Ecological Applications*, *25*(7), 1749–1756. https://doi.org/10.1890/14-1826.1

Katzner, T. E., & Arlettaz, R. (2020). Evaluating contributions of recent tracking-based animal movement ecology to conservation management. *Frontiers in Ecology and Evolution*, *7*. Retrieved February 15, 2022, from https://www.frontiersin.org/article/10.3389/fevo.2019.00519

Kochert, M. N. (1998). *Movebank: Swainson's hawks*. Retrieved February 16, 2022, from https://www.movebank.org/cms/webapp?gwt_fragment=page=studies,path=study204253

Kochert, M. N., Fuller, M. R., Schueck, L. S., Bond, L., Bechard, M. J., Woodbridge, B., Holroyd, G. L., Martell, M. S., & Banasch, U. (2011). Migration patterns, use of stopover areas, and austral summer movements of swainson's hawks. *The Condor*, *113*(1), 89–106. https://doi.org/10.1525/cond.2011.090243

Kranstauber, B., Cameron, A., Weinzerl, R., Fountain, T., Tilak, S., Wikelski, M., & Kays, R. (2011). The movebank data model for animal tracking. *Environmental Modelling & Software*, *26*(6), 834–835. https://doi.org/10.1016/j.envsoft.2010.12.005

La Sorte, F. A., Lepczyk, C. A., Burnett, J. L., Hurlbert, A. H., Tingley, M. W., & Zuckerberg, B. (2018). Opportunities and challenges for big data ornithology. *The Condor*, *120*(2), 414–426. https://doi.org/10.1650/CONDOR-17-206.1

Martell, M. S., & Douglas, D. (2019). *Data from: Fall migration routes, timing, and wintering sites of north american ospreys as determined by satellite telemetry*. http://doi.org/10.5441/001/1.sv6335t3

Martell, M. S., Henny, C. J., Nye, P. E., & Solensky, M. J. (2001). Fall migration routes, timing, and wintering sites of north american ospreys as determined by satellite telemetry. *The Condor*, *103*(4), 715–724. https://doi.org/10.1093/condor/103.4.715

McCabe, R., & Goodrich, L. (2022). *Movebank: Broad-winged hawk habitat use, range, and movement ecology*. Retrieved February 16, 2022, from https://www.movebank.org/cms/webapp?gwt_fragment=page=studies,path=study28691134

McCabe, R. A., Goodrich, L. J., Barber, D. R., Master, T. L., Watson, J. L., Bayne, E. M., Harrison, A., Marra, P. P., & Bildstein, K. L. (2020). Satellite tracking reveals age and origin differences in migration ecology of two populations of broad-winged hawks (buteo platypterus) [Publisher: The Wilson Ornithological Society]. *The Wilson Journal of Ornithology*, *132*(1), 1–14. https://doi.org/10.1676/1559-4491-132.1.1

McKenna, R., Sheldon, D., & Miklau, G. (2019). Graphical-model based estimation and inference for differential privacy. *International Conference on Machine Learning*, 4435–4444.

McKinnon, E. A., & Love, O. P. (2018). Ten years tracking the migrations of small landbirds: Lessons learned in the golden age of bio-logging. *The Auk*, *135*(4), 834–856. https://doi.org/10.1642/AUK-17-202.1

Moore, J. D., Andersen, D. E., Cooper, T. R., Duguay, J. P., Oldenburger, S. L., Stewart, C. A., & Krementz, D. G. (2021a). *Data from: Migration phenology and patterns of american woodcock in central north america derived using satellite telemetry*. http://doi.org/10.5441/001/1.8764q39q

Moore, J. D., Andersen, D. E., Cooper, T. R., Duguay, J. P., Oldenburger, S. L., Stewart, C. A., & Krementz, D. G. (2021b). Migration phenology and patterns of american woodcock in central north america derived using satellite telemetry [Publisher: Nordic Board for Wildlife Research]. *Wildlife Biology*, *2021*(1), wlb.00816. https://doi.org/10.2981/wlb.00816

Newton, I. (2008). *The migration ecology of birds* [Google-Books-ID: BndIbshDWTgC]. Academic Press.

Ramey, A. M., Soos, C., Link, P., Walther, P., Tibbitts, L., & Douglas, D. C. (2019). Tracking data for blue-winged teal (anas discors) [Medium: zip,csv Type: dataset]. https://doi.org/10.5066/P9Z9BA9F

Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., Stanton, J. C., Panjabi, A., Helft, L., Parr, M., & Marra, P. P. (2019). Decline of the north american avifauna [Publisher: American Association for the Advancement of Science Section: Report]. *Science*, *366*(6461), 120–124. https://doi.org/10.1126/science.aaw1313

Sanderson, F. J., Donald, P. F., Pain, D. J., Burfield, I. J., & van Bommel, F. P. J. (2006). Long-term population declines in afro-palearctic migrant birds. *Biological Conservation*, *131*(1), 93–105. https://doi.org/10.1016/j.biocon.2006.02.008

Schuster, R., Wilson, S., Rodewald, A. D., Arcese, P., Fink, D., Auer, T., & Bennett, J. R. (2019). Optimizing the conservation of migratory species over their full annual cycle [Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Animal migration;Conservation biology;Decision making;Sustainability Subject_term_id: animal-migration;conservation;decision-making;sustainability]. *Nature Communications*, *10*(1), 1754. https://doi.org/10.1038/s41467-019-09723-8

Sheldon, D., & Dietterich, T. (2011). Collective graphical models. *Advances in neural information processing systems (NIPS)*, 1161–1169.

Sheldon, D., Elmohamed, M. A. S., & Kozen, D. (2008). Collective inference on Markov models for modeling bird migration. *Advances in neural information processing systems (NIPS)*, 1321–1328.

Sheldon, D., Sun, T., Kumar, A., & Dietterich, T. G. (2013). Approximate inference in collective graphical models. *Proceedings of the 30th international conference on machine learning (ICML)*, 1004–1012.

Singh, R., Haasler, I., Zhang, Q., Karlsson, J., & Chen, Y. (2020). Inference with aggregate data: An optimal transport approach. *arXiv preprint arXiv:2003.13933*.

Somveille, M., Bay, R. A., Smith, T. B., Marra, P. P., & Ruegg, K. C. (2021). A general theory of avian migratory connectivity [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ele.13817]. *Ecology Letters*. https://doi.org/10.1111/ele.13817

Stanley, C. Q., Dudash, M. R., Ryder, T. B., Shriver, W. G., Serno, K., Adalsteinsson, S., & Marra, P. P. (2021). Seasonal variation in habitat selection for a neotropical migratory songbird using high-resolution GPS tracking [_eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecs2.3421 *Ecosphere*, *12*(3), e03421. https://doi.org/https://doi.org/10.1002/ecs2.3421

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W. M., Iliff, M. J., Lagoze, C., La Sorte, F. A., ... Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Sun, T., Sheldon, D., & Kumar, A. (2015). Message passing for collective graphical models. *Proceedings of the 32nd international conference on machine learning (ICML)*, 853–861.

Supp, S. R., Bohrer, G., Fieberg, J., & La Sorte, F. A. (2021). Estimating the movements of terrestrial animal populations using broad-scale occurrence data. *Movement Ecology*, *9*(1), 1–19.

Tibbitts, T., Ruthrauff, D. R., Gill, R. E., & Douglas, D. C. (2018). Tracking data for whimbrels (numenius phaeopus) [Medium: csv,zip Type: dataset]. https://doi.org/10.5066/P978PX2X

Van Doren, B. M., & Horton, K. G. (2018). A continental system for forecasting bird migration. *Science*, *361*(6407), 1115–1118. https://doi.org/10.1126/science.aat7526

Vilnis, L., Belanger, D., Sheldon, D., & McCallum, A. (2015). Bethe projections for non-local inference. *Proceedings of the 29th conference on uncertainty in artificial intelligence (UAI)*, 892–901.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, *1*(1-2), 1–305.

Webster, M. S., & Marra, P. P. (2005). The importance of understanding migratory connectivity and seasonal interactions. In R. Greenberg & P. P. Marra (Eds.), *Birds of two worlds: The ecology and evolution of migration*. Johns Hopkins University Press.

www.nist.gov. (2018). 2018 differential privacy synthetic data challenge. https://www.nist.gov/communications-technology-laboratory/pscr/funding-opportunities/open-innovation-prize-challenges-1

www.nist.gov. (2020). 2020 differential privacy temporal map challenge. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/current-and-upcoming-prize-challenges/2020-differential

Yasunori, A., Nishimura, T., Tanaka, Y., Kurashima, T., & Toda, H. (2020). Exact and efficient inference for collective flow diffusion model via minimum convex cost flow algorithm. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), 3163–3170.

A Additional Figures

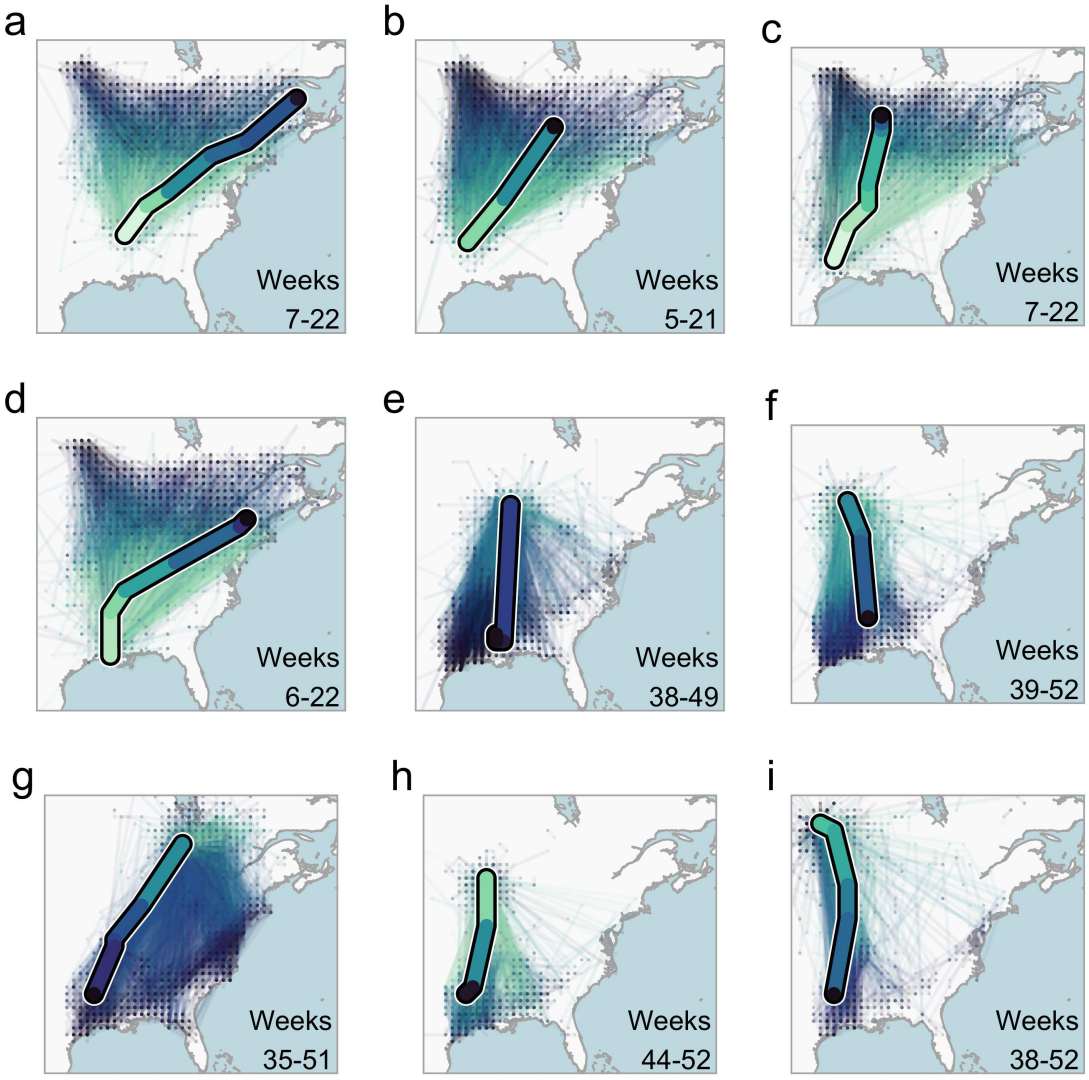

Figure 9: Model-simulated trajectories for GPS-tracked American Woodcocks *Scolopax minor*. Observed movements of GPS-tracked woodcocks (single thick path) and simulated trajectories (thin paths) for 2500 simulated birds originating at the same starting location as observed birds.

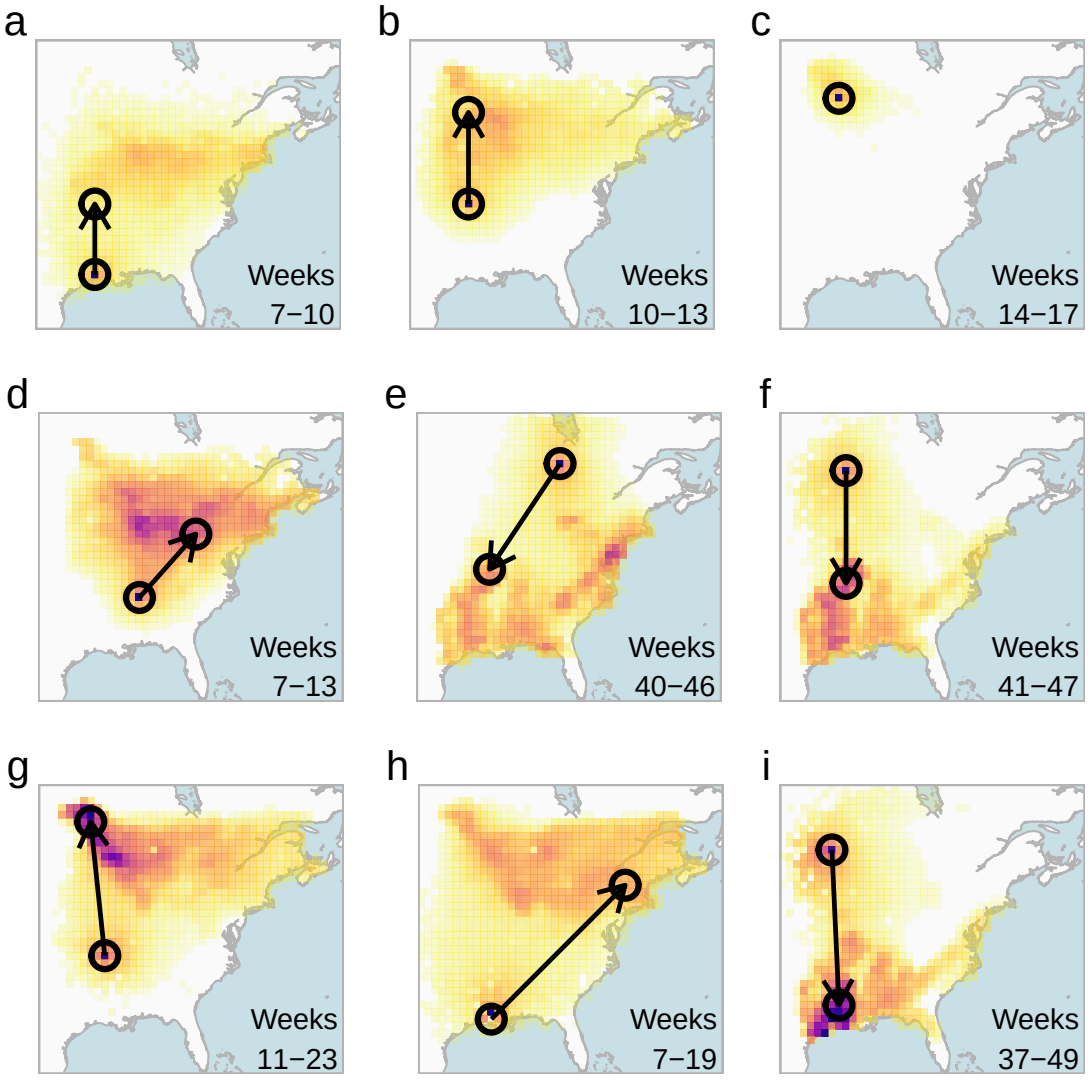

Figure 10: Conditional forecast distributions for GPS-tracked American Woodcocks *Scolopax minor*. Each heatmap shows the predicted movement distribution of a GPS-tracked individual originating within the circle at the base of the arrow. Darker colors indicate a higher predicted likelihood of movement to that area. The point of the arrow shows the observed ending location. Shown are examples of 3-week (a-c, same individual), 6-week (d-f, different individuals), and 12-week (g-i, different individuals) conditional forecasts.