ECOLOGY LETTERS

Ecology Letters, (2020) **23**: 734–747 doi: 10.1111/ele.13462

METHODS

Neural hierarchical models of ecological populations

Maxwell B. Joseph*

Earth Lab, Cooperative Institute for Research in Environmental Sciences University of Colorado Boulder Boulder, CO 80303, USA

**Correspondence: E-mail: maxwell.b.joseph@colorado.edu*

The peer review history for this article is available at https://publons.com/publon/10.1111/ele.13462

Abstract

Neural networks are increasingly being used in science to infer hidden dynamics of natural systems from noisy observations, a task typically handled by hierarchical models in ecology. This article describes a class of hierarchical models parameterised by neural networks – neural hierarchical models. The derivation of such models analogises the relationship between regression and neural networks. A case study is developed for a neural dynamic occupancy model of North American bird populations, trained on millions of detection/non-detection time series for hundreds of species, providing insights into colonisation and extinction at a continental scale. Flexible models are increasingly needed that scale to large data and represent ecological processes. Neural hierarchical models satisfy this need, providing a bridge between deep learning and ecological modelling that combines the function representation power of neural networks with the inferential capacity of hierarchical models.

Keywords

Deep learning, hierarchical model, neural network, occupancy.

Ecology Letters (2020) **23**: 734–747

INTRODUCTION

Deep neural networks have proved useful in myriad tasks due their ability to represent complex functions over structured domains (LeCun *et al*. 2015). While ecologists are beginning to use such approaches, for example to identify plants and animals in images (Norouzzadeh *et al*. 2018; Fricker *et al*. 2019), there has been relatively little integration of deep neural networks with ecological models.

Ecological processes are difficult to observe. Inference often proceeds by modelling the relationship between imperfect data and latent quantities or processes of interest with hierarchical models (Wikle 2003). For example, occupancy models estimate the presence or absence of a species using imperfect detection data (MacKenzie *et al*. 2002), and 'dynamic' occupancy models estimate the extinction and colonisation dynamics (MacKenzie *et al*. 2003). Population growth provides another example, motivating hierarchical models that link noisy observations to mechanistic models (De Valpine & Hastings 2002). In such models, it is often desirable to account for heterogeneity among sample units (e.g. differences among habitats and survey conditions), to better understand ecological dynamics.

Many hierarchical models in ecology account for heterogeneity among sample units using a linear combination of explanatory variables, despite there often being reasons to expect non-linearity (Lek *et al*. 1996; Austin 2002; Oksanen & Minchin 2002). A variety of solutions exist to account for non-linearity. For example, Gaussian processes have been used for species distribution models (Latimer *et al*. 2009; Golding & Purse 2016), in animal movement models (Johnson *et al*. 2008), and in point process models for distance sampling data (Johnson *et al*. 2010; Yuan *et al*. 2017). Generalised additive models also have been used to account for spatial autocorrelation (Miller *et al*. 2013; Webb *et al*. 2014), nonlinear responses to habitat characteristics (Knapp *et al*. 2003; Bled *et al*. 2013) and differential catchability in capture–recapture studies (Zwane & Van der Heijden 2004).

Machine learning provides additional tools for approximating nonlinear functions in hierarchical models. For example, Hutchinson *et al*. (2011) combined a site-occupancy model with an ensemble of decision trees to predict bird occurrence, combining a structured observation and process model from ecology with a flexible random forest model. Similar hybrid approaches could be developed for other classes of ecological models and/or machine learning methods. Neural networks seem particularly worthy of attention, given their success in other domains (LeCun *et al*. 2015).

Neural networks have been used in ecology, but to date have not been integrated into hierarchical models that explicitly distinguish between ecological processes and imperfect observations. For example, neural networks have modelled observed abundances of aquatic organisms, but without distinguishing observed from true abundance (Chon *et al*. 2001; Jeong *et al*. 2001, 2008; Malek *et al*. 2012). Neural networks also have been used to model stock-recruitment and apparent presence/absence data (Manel *et al*. 1999; Chen & Hare 2006; Özesmi *et al*. 2006; Harris 2015; Chen et al. 2016), but have not yet been extended to account for imperfect detection, despite increasing recognition of its importance (Guillera-Arroita 2017; Tobler *et al*. 2019). Notably, standard neural networks that classify or predict data are of limited use for ecological applications where imperfect data are used to learn about dynamics of hard to observe systems. As a solution, this article describes neural hierarchical models that combine the function representation capacity of neural networks with hierarchical models that represent ecological processes.

RELATED-WORK

Neural networks

Neural networks are function approximators. Linear regression is a special case, where an input vector $\mathbf{x}$ is mapped to a predicted value y:

$$y = \mathbf{w}^T\mathbf{x},$$

where $\mathbf{w}$ is a parameter vector (Fig. 1a). Predicted values are linear combinations of the inputs $\mathbf{x}$ due to the product $\mathbf{w}^T\mathbf{x}$, which restricts the complexity of the function mapping $\mathbf{x}$ to y.

Instead of modelling outputs as linear functions of $\mathbf{x}$, a model can be developed that is a linear function of a nonlinear transformation of $\mathbf{x}$. Nonlinear transformations can be specified via polynomial terms, splines or another basis expansion (Hefley *et al.* 2017), but neural networks parameterise the transformation via a set of sequential 'hidden layers' (Goodfellow *et al.* 2016).

In a neural network with one hidden layer, the first hidden layer maps the length D input $\mathbf{x}$ to a length $D^{(1)}$ vector of 'activations' $\mathbf{a}^{(1)} = \mathbf{W}^{(1)}\mathbf{x}$, where $\mathbf{W}^{(1)}$ is a $D^{(1)} \times D$ parameter matrix.

The activations are passed to a differentiable nonlinear activation function g to obtain the 'hidden units' of the first layer $\mathbf{h}^{(1)} = g(\mathbf{a}^{(1)})$ (Fig. 1b).

The final layer of a neural network maps the hidden layer to an output. For a neural network with one hidden layer, if the output variable $\mathbf{y}$ is a K dimensional vector, the output unit activations are given by $\mathbf{a}^{(2)} = \mathbf{W}^{(2)}\mathbf{h}^{(1)}$, where $\mathbf{W}^{(2)}$ is a $K \times D^{(1)}$ parameter matrix.

The output activation can be written as a composition of functions that transform the inputs x:

$$\mathbf{a}^{(2)} = \mathbf{W}^{(2)}(g(\mathbf{W}^{(1)}\mathbf{x})).$$

Similar to link functions in generalised linear models, outputs can be transformed by an output activation function. For example, if $\mathbf{y}$ is unbounded, the identity function can act as an output activation so that $\mathbf{y} = \mathbf{a}^{(2)}$. Neural networks that predict probabilities typically use a sigmoid (inverse logit) activation function.

Neural networks usually are trained using stochastic gradient-based optimisation to minimise a loss function, for example the negative log likelihood of a Gaussian distribution for a regression task, a Bernoulli distribution for binary classification or a Poisson distribution for a count model. Partial derivatives of the model parameters are computed with respect to the loss via backpropagation, and parameters are updated to reduce the loss. In practice, these partial derivatives are often computed via automatic differentiation over a 'mini-batch' of samples, which provides a noisy estimate of the gradient (Ruder 2016).

Neural networks are popular both because of their practical successes in a wide variety of applications, and because they possess some desirable theoretical properties. A neural

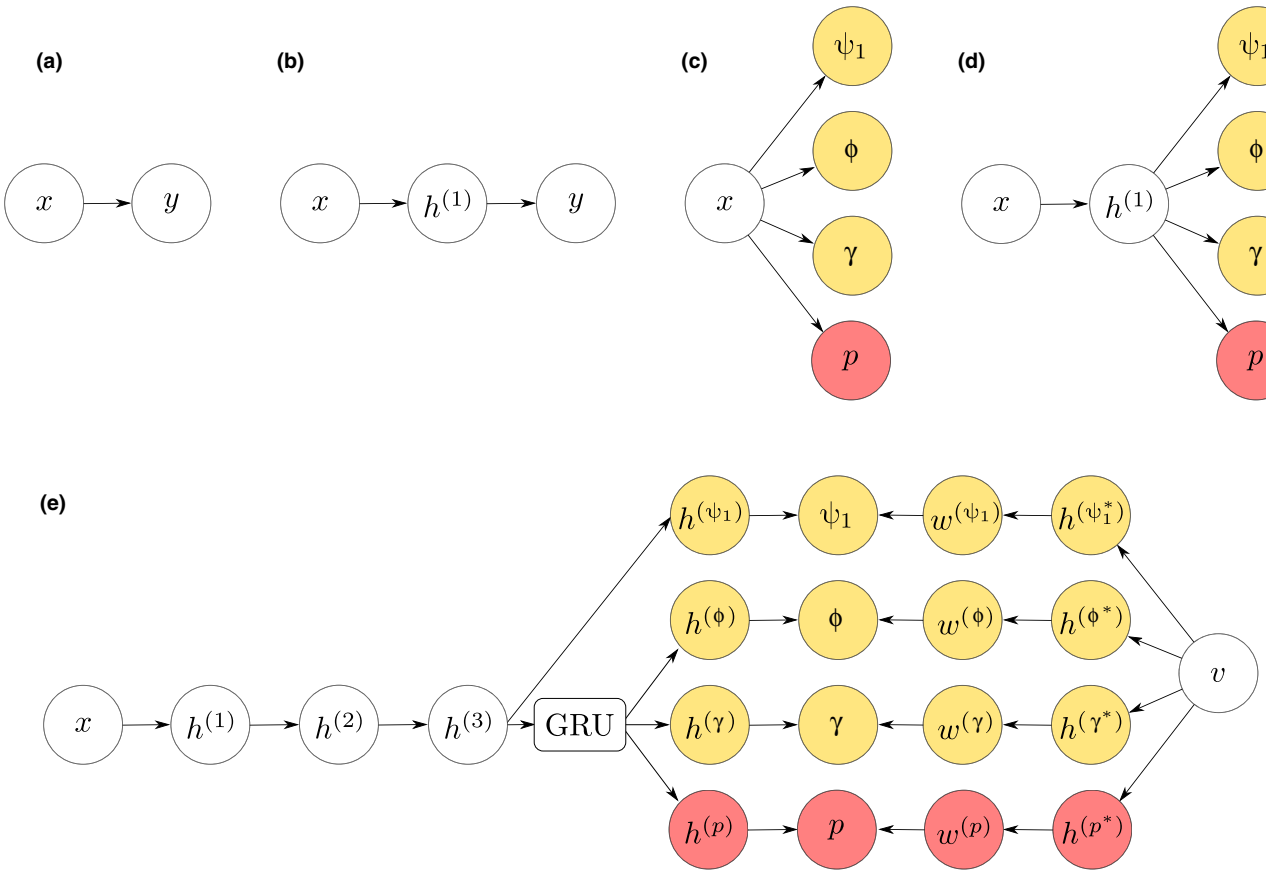

Figure 1 Computation graphs for (a) linear regression, (b) a neural network with one hidden layer, (c) a dynamic occupancy model, (d) a single-species neural dynamic occupancy model and (e) a deep multi-species neural dynamic occupancy model. Yellow and red indicate quantities specific to process and observation components respectively. Inputs are represented by x, and predicted values in panels (a) and (b) by y. Hidden layers are represented by h, with layer-specific superscripts. Outputs include initial occupancy (ψ_1), persistence (φ), colonisation (γ) and detection (p) probabilities. Latent species embedding vectors are represented by v, and GRU indicates a gated recurrent unit.

network with suitable activation functions and a single hidden layer containing a finite number of neurons can approximate nearly any continuous function on a compact domain (Cybenko 1989; Hornik 1991). 'Deep' neural networks with many sequential hidden layers also act as function approximators (Lu et al. 2017). The field of deep learning, which applies such networks, provides a variety of network architectures to account for temporal structure (Hochreiter & Schmidhuber 1997), spatial structure on regular grids (Long *et al*. 2015) or graphs (Niepert *et al*. 2016), sets of unordered irregular points (Li et al. 2018) and spatiotemporal data on grids or graphs (Xingjian et al. 2015; Jain *et al*. 2016).

The potential for deep learning has been recognised in Earth science (Reichstein *et al*. 2019), the natural sciences (Ching *et al*. 2018; Gazestani & Lewis 2019; Roscher et al. 2019), physical sciences (Carleo et al. 2019), chemical sciences (Butler *et al*. 2018) and ecology (Christin *et al*. 2018; Desjardins-Proulx et al. 2019). For example, models of lake temperature that combine neural networks with loss functions consistent with known physical mechanisms perform better than physical models and neural networks applied alone (Karpatne et al. 2017). Similarly, generative adversarial networks with loss functions that encourage mass balance have expedited electromagnetic calorimeter data generation from the Large Hadron Collider (Paganini *et al*. 2018; Radovic *et al*. 2018). Convolutional neural networks also have been successfully deployed in population genetics to make inferences about introgression, recombination, selection and population sizes (Flagel *et al*. 2018). There are various ways to combine science knowledge with deep learning. Ba et al. (2019) provide a useful taxonomy in the context of physics-based deep learning. In ecology, hierarchical models present an opportunity to build upon existing approaches to derive science-based deep learning methods.

Hierarchical models

Hierarchical models combine a data model, a process model, and a parameter model (Berliner 1996; Wikle 2003). Data models represent the probability distribution of observations conditioned on a process and some parameters, for example the probability of capturing a marked animal, given its true state (alive or dead). Process models represent states and their dynamics, conditioned on some parameters. State variables are often incompletely observed, for example whether an individual animal is alive or whether a site is occupied. Parameter models represent probability distributions for unknown parameters – priors in a Bayesian framework. In a non-Bayesian setting, parameters are treated as unknowns and estimated from the data, but parameter uncertainty is not represented using probability distributions (Cressie *et al*. 2009).

NEURAL HIERARCHICAL MODELS

Neural hierarchical models are hierarchical models in which the observation, process or parameter model is parameterised by a neural network. These models are hierarchical (sensu Berliner (1996)) if they distinguish between a modelled process and available data, for example between partial differential equations and noisy observations of their solutions (Raissi 2018). 'Deep Markov models' – hidden Markov models parameterised by neural networks – provide an example, with successful applications in polyphonic music structure discovery, patient state reconstruction in medical data and time series forecasting (Krishnan *et al*. 2017; Rangapuram et al. 2018). State-space neural networks that use recurrent architectures provide another example dating back two decades (Zamarreño & Vega 1998; Van Lint *et al*. 2002). This class of models inherits the flexibility and scalability of neural networks, along with the inferential power of hierarchical models, but applications in ecology and environmental science are just beginning to emerge (Wikle 2019).

Construction of such models from existing hierarchical models is straightforward. For example, one can propose neural variants of occupancy models (MacKenzie *et al*. 2002), dynamic occupancy models (MacKenzie *et al*. 2003; Royle & Kéry 2007), N-mixture models (Royle 2004), mark-recapture models (Jolly 1965; Calvert *et al*. 2009) and other hidden Markov models (Patterson *et al*. 2009, 2017; Langrock *et al*. 2012). Output activation functions can be determined from inverse link functions, for example sigmoid (inverse logit) activations for probabilities, and loss functions can be constructed from the negative log likelihoods (see Appendix S1 in Supporting Information for example model specifications). Specialised neural network architectures that operate on structured data can be readily integrated into such models. For example, Appendix S2 in Supporting Information provides a simulated animal movement case study where aerial imagery is mapped to state transition probabilities of a hidden Markov model with a convolutional neural network. To provide an empirical use case, a neural dynamic occupancy model is developed for extinction and colonisation dynamics of North American bird communities.

CASE STUDY

The North American Breeding Bird Survey (BBS) is a large-scale annual survey aimed at characterising trends in roadside bird populations (Link & Sauer 1998; Sauer & Link 2011; Sauer *et al*. 2013; Pardieck *et al*. 2019). Thousands of routes are surveyed once a year during the breeding season. Surveys consist of volunteer observers that stop 50 times at points 800 m apart on a transect, recording all birds detected within 400 m for 3 min. Species can be present, but not detected, and may go locally extinct or colonise new routes from year to year, motivating the development of dynamic occupancy models which use imperfect detection data to estimate the latent presence or absence states (MacKenzie *et al*. 2003; Royle & Kéry 2007).

This case study uses BBS data from 1997 to 2018 excluding unidentified or hybrid species, restricting the analysis to surveys meeting the official BBS criteria (Pardieck *et al*. 2019). The resulting data consists of 647 species sampled at 4540 routes, for a total of 59 384 surveys (not every route is surveyed in each year), 2 937 380 observation history time series and 38 421 448 detection/non-detection observations.

Process model

A multi-species dynamic occupancy model for spatially referenced routes $s = 1, ..., S$, surveyed in years $t = 1, ...T$, for species $j = 1, ..., J$ aims to estimate colonisation and extinction dynamics through time. The true occupancy state $z_{t,s,j} = 1$ if species j is present, and $z_{t,s,j} = 0$ if species j is absent. The model represents $z_{t,s,j}$ as a Bernoulli distributed random variable, where $\Pr(z_{t,s,j} = 1) = \psi_{t,s,j}$. The probability of occurrence on the first timestep is $\psi_{t=1,s,j}$. Subsequent dynamics are determined by probabilities of persistence from time t to $t + 1$ denoted $\phi_{t,s,j}$, and probabilities of colonisation from time t to $t + 1$ denoted $\gamma_{t,s,j}$, so that the probability of occurrence in timesteps $t = 2, ..., T$ is (MacKenzie *et al.* 2003):

$$\psi_{t,s,j} = z_{t-1,s,j}\phi_{t-1,s,j} + (1 - z_{t-1,s,j})\gamma_{t-1,s,j}.$$

Observation model

Let $y_{t,s,j}$ represent the number of stops where species j was detected in year t on route s. Conditional on a species being present, it is detected at each stop with probability $p_{t,s,j}$. Assume that there are no false-positive detections, so that if a species is absent, it cannot be detected (but see Royle & Link 2006). With $k = 50$ replicate stops on each transect, the observations can be modelled using a Binomial likelihood for one species-route-year combination:

$$[y_{t,s,j}|z_{t,s,j}, p_{t,s,j}] = \text{Binomial}(y_{t,s,j}|z_{t,s,j}p_{t,s,j}, k),$$

where square brackets denote the probability function (e.g. in this case, $[y_{t,s,j}|z_{t,s,j}, p_{t,s,j}] = \Pr(Y_{t,s,j} = y_{t,s,j}|z_{t,s,j}, p_{t,s,j})$ where $Y_{t,s,j}$ is a discrete random variable and $y_{t,s,j}$ is a particular value).

The joint likelihood corresponds to the product of these terms for all years, routes, and species (Dorazio *et al.* 2010).

Parameter models

Heterogeneity in parameter values among routes, years, species, and surveys was modelled using three different approaches.

(1) **Single-species baseline models** mapped input features to occupancy parameters with a linear combination on the logit scale (MacKenzie *et al.* 2003; Fig. 1c).
(2) **Single-species neural hierarchical models** mapped inputs to occupancy parameters using a neural network with one hidden layer (Fig. 1d).
(3) **A multi-species deep neural hierarchical model** was developed to model occupancy dynamics of all species simultaneously (Fig. 1e).

Input features included Environmental Protection Agency (EPA) level 1 ecoregions, the first eight principal components of the standard 19 WorldClim Bioclimatic variables averaged across the years 1970–2000 (Fick & Hijmans 2017), BBS route spatial coordinates, distance from the coast, elevation and road density within a 10 km buffer of each route (Meijer *et al.* 2018). The model of detection probabilities additionally included survey-level features including temperature, duration, wind and air conditions.

The neural hierarchical models were motivated by joint species distribution models in which species load onto a shared set of latent factors (Thorson *et al.* 2015, 2016; Warton *et al.* 2015; Ovaskainen *et al.* 2016; Tikhonov *et al.* 2017), and by recent work on deep neural basis expansions (McDermott & Wikle 2019; Wikle 2019).

Analogously, the neural networks combined inputs into a latent vector for each route (the hidden layers), which act as latent factors that are mapped to parameters (Fig. 1d). Because the single-species neural hierarchical models were fit separately for each species, latent factors were species-specific. In contrast, latent factors were shared among species in the multi-species model.

The multi-species neural dynamic occupancy model additionally built upon previous work on deep multi-species embedding. Deep multi-species embedding uses vector-valued 'entity embeddings' to represent each species (Chen et al. 2016; Guo & Berkhahn 2016). These entity embeddings are mappings from categorical data (e.g. a species identity) to continuous numeric vector representations. Embeddings are used extensively in language models such as word2vec, which maps words to vector spaces (Mikolov et al. 2013).

Further, the multi-species model combined species embeddings with encoder–decoder components to estimate occupancy parameters. Encoder–decoder neural networks are used in sequence-to-sequence translation (Sutskever et al. 2014). They encode inputs (e.g. a sequence of words) into a latent vector space, then decode that vector representation to generate another sequence using a neural network. Similarly, the multi-species model encoded route-level features into vector representations. These route vectors were decoded by a recurrent neural network to generate a multivariate time series of latent vectors associated with colonisation, persistence and detection (Chung et al. 2014). Finally, the multi-species model combined these route-level latent vectors with species-level embeddings to compute colonisation, persistence and detection probabilities (Fig. 1e). For additional details, see Appendix S3 in Supporting Information.

Model comparisons

To compare the performance of the three modelling approaches, the data were partitioned into a training, validation and test set at the EPA level 2 ecoregion (Roberts *et al.* 2017). All routes within an ecoregion were assigned to the same partition. This resulted in 2154 training routes, 948 validation routes and 1438 test routes. K-fold cross-validation would also be possible, though it requires retraining of each model K times (Roberts *et al.* 2017). Because of the size of the BBS data and the computational resources required to train these models, a simpler train/validation/test split was used.

For each of the three modelling approaches, routes in the training set were used for parameter estimation. Single-species models were fit separately for each species (modelling approaches 1 and 2), and one multi-species model (approach 3) was fit using all of the training data. Then, using the trained models, the mean predictive log-density of the validation data was evaluated to identify the best performing model

(Gelman *et al*. 2014). This step indicated which model fit best to the withheld validation data. Finally, the best performing model was retrained using the training and validation data, and its predictive performance was evaluated on the withheld test set (Russell & Norvig 2016).

Final model evaluation

The final model's performance was evaluated quantitatively and qualitatively. Quantitative predictive performance was evaluated in two ways. First, 95% prediction interval coverage was computed for test set counts. Prediction intervals were constructed using the quantiles of the binomial distribution, marginalising over latent occupancy states. Second, the area under the receiver operator characteristic curve (AUC) was computed for binarised test set counts. The AUC analysis also marginalised over latent states to derive predicted probabilities, for example $\Pr(y_{t,s,j} = 0|p_{t,s,j}, \psi_{t,s,j}) = 1 - \psi_{t,s,j} + \psi_{t,s,j}(1 - p_{t,s,j})^{50}$, where the parameters p and ψ are estimated by the model. These two approaches provide information on how well the final model could predict the number of stops with detections at each route, and whether any detections would occur on each route respectively.

Qualitative analyses were based on predicted occupancy states from the final model. The most likely occupancy states were computed at all BBS routes for each year and species using the Viterbi algorithm (Viterbi 1967). The estimated occupancy states were then used to compute finite sample population growth rates for each species (Royle & Kéry 2007). Occupancy state estimates were also used to compute annual spatial centroids for each species, by taking the spatial centroids of occupied route coordinates. These are hereafter referred to as 'BBS range centroids' to differentiate from the actual centroid of a species entire range.

To evaluate whether the results were qualitatively consistent with previous findings about colonisation and extinction gradients over species ranges, correspondence between BBS range centroids and both colonisation and persistence probabilities were assessed using linear regression (Mehlman 1997; Doherty et al. 2003; Royle & Kéry 2007). For this analysis, the response variable was colonisation or persistence averaged over time, the predictor was distance from BBS range centroid. Survey routes were the sample units, and separate analyses were conducted for each species. To avoid bias associated with recently added BBS routes and variance associated with rare species, these analyses only used data from BBS routes surveyed in every year, species observed in every year, and species that occurred in 100 or more routes (Sauer *et al*. 2017).

Finally, visualisations were developed to graphically interpret the final model. First, route-level features were visualised using t-distributed stochastic neighbour embedding, which maps high dimensional vectors to low dimensional representations (Maaten & Hinton 2008). In this low dimensional space, routes with similar embeddings are close together, and routes with dissimilar embeddings appear distant (Rauber *et al*. 2016). Second, species loading vectors were compared in terms of cosine similarity, which measures the orientation of loading vectors in latent space. Species occupancy should be positively related if loading vectors are oriented similarly.

This expectation was checked graphically by comparing estimated occupancy time series of species that had the most similar and most different loading vectors.

RESULTS

The multi-species neural hierarchical model performed best on the withheld validation routes. The difference in mean validation set negative log likelihood was 5.015 relative to the baseline model and 2.108 relative to the single-species neural hierarchical model. For the final model, 95% prediction interval coverage for observed counts at withheld test set routes was 93.6%, with a standard deviation of 1.5%, a minimum of 86.5% and a maximum of 97.6%. The model also predicted whether species would be detected on test set routes fairly well, with a mean test set AUC of 0.953, an among-route standard deviation of 0.032, a minimum of 0.667 and a maximum of 0.993 (Fig. 2).

Qualitative results related to range shifts and population growth rates were plausible given previous work. The model identified the invasive Eurasian Collared-Dove (*Streptopelia decaocto*) as having the greatest range centroid displacement from 1997 to 2018 and the highest finite sample population growth rate (Fig. 3), consistent with its invasion of North America from Florida following its introduction in the 1980s (Bled *et al*. 2011). Increasing trends also have previously been reported for the species with the next three highest population growth rates: Bald Eagle (*Haliaeetus leucocephalus*), Wild Turkey (*Meleagris gallopavo*) and Osprey (*Pandion haliaetus*; Sauer *et al*. 2013).

The majority (77%) of common species were less likely to persist at routes that were distant from their estimated BBS range centroids. Similarly, 98% of common species were less likely to colonise routes distant from their BBS range centroids. There were examples of species with positive and negative distance coefficients for persistence and colonisation (Fig. 4a). Negative relationships were most apparent for common species that occupied a large fraction of BBS routes (Fig. 4b).

Results for representative species are displayed in Fig. 4c–d. Route vectors combined information from the categorical and continuous route-level features. Unsurprisingly (because ecoregion was an input feature), routes in the same ecoregions clustered together (Fig. 5a). Route embeddings also revealed relationships among ecoregions. For example, Marine West Coast Forest ecoregion routes were similar to Northwestern Forested Mountains routes and most different from Northern Forests routes (Fig. 5a). Variation within clusters also related to continuous route-level features (Fig. 5b).

The model predicted nonlinear dependence among species that is interpretable in terms of the cosine similarity among species-specific parameter vectors. For example, parameters for Mourning Dove (*Zenaida macroura*) were closest to those of Barn Swallow (*Hirundo rustica*), and most different from (Myrtle Warbler) Yellow-rumped Warbler (*Setophaga coronata coronata*; Fig. 6a). On BBS routes where Mourning Doves are likely to occur, Barn Swallows are also likely to

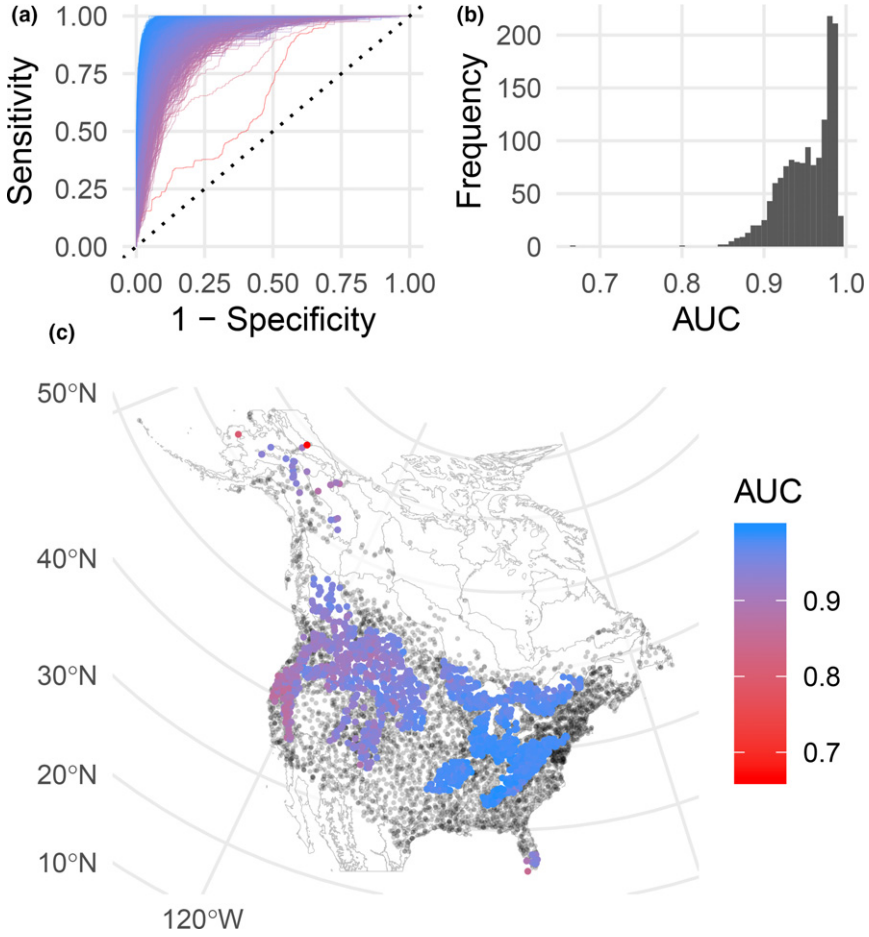

Figure 2 (a) Receiver operator characteristic curves of binarised detection/non-detection data for each survey route in the test set, coloured by the area under the curve (AUC). Here, the x-axis is the complement of specificity: the ratio of the number of false positives (incorrectly predicted detections, with no observed detections) to the sum of false positives and true negatives (correctly predicted non-detections with observed non-detections). The y-axis is true positive rate: the ratio of the number of true positives (correctly predicted detections with observed detections) to the sum of true positives and false negatives (incorrectly predicted non-detections with observed detections). The overall distribution of AUC values is shown in (b), and (c) shows the locations of test set routes coloured by AUC values, where black dots represent routes that were used to train the final model.

occur, and (Myrtle Warbler) Yellow-rumped Warblers are unlikely to occur. Species pairs of the most similar and most dissimilar loadings are also provided for Eurasian Collared-Dove and Bald Eagle (Fig. 6b and c).

DISCUSSION

Neural hierarchical models provide a bridge between hierarchical models built from scientific knowledge and neural networks that approximate functions over structured domains. This framework integrates research in science-based deep learning and ecological modelling, and can use existing hierarchical models as a starting point. The case study provides a proof of concept example of constructing a scalable and performant neural hierarchical model based on a multi-species dynamic occupancy model, providing insights about colonisation and extinction dynamics of North American bird assemblages at a continental scale.

The breeding bird survey case study indicates that neural hierarchical models can outperform simpler models. Notably, the final model was performant both quantitatively and qualitatively, detecting population increases and range expansions that are consistent with prior work. The case study extends previous analyses of persistence probabilities at range edges (Royle & Kéry 2007), indicating that common species tend to have higher persistence and colonisation probabilities at routes close to their BBS range centroid. Yet, interpreting this result is complicated by irregular range geometry which can lead to centroids that near the edge (or even outside) of a species range, and divergence between actual and estimated ranges due to incomplete sampling (Sagarin & Gaines 2002; Fortin *et al.* 2005; Dallas *et al.* 2017; Knouft 2018). Additional complexity is apparent in Fig. 4c, which indicates that gradients in colonisation and persistence may not be isotropic (the same in every direction). With those caveats, the results are broadly consistent with a theoretical expectation that range boundaries can arise from gradients in local extinction and colonisation rates (Holt & Keitt 2000).

The case study used a gated recurrent neural network architecture to handle temporal structure (Chung et al. 2014), and architectures designed for other data structures present additional opportunities for ecological applications.

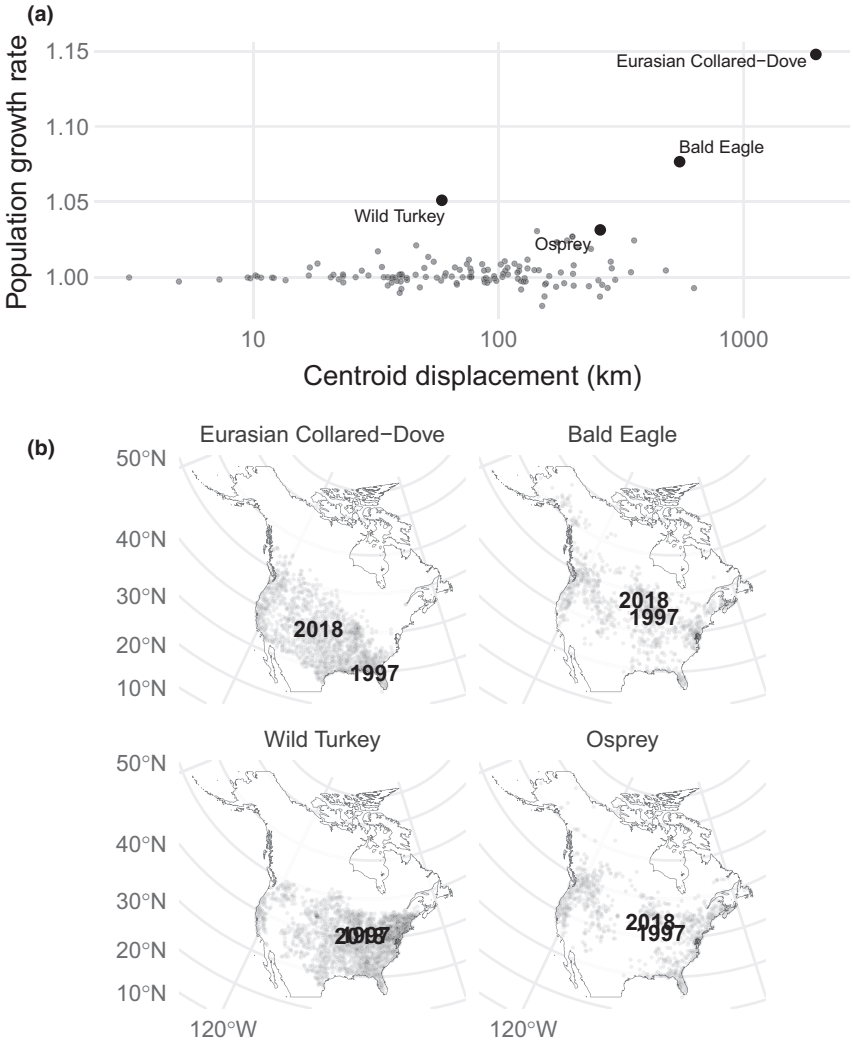

Figure 3 (a) Centroid displacement of bird species (x-axis) from 1997 to 2018 in kilometers vs. finite-sample population growth rate (y-axis), where a growth rate of 1 is stable, values less than 1 are decreasing, and values greater than 1 are increasing. Species with the highest population growth rates are highlighted. Panel (b) shows the locations of breeding bird survey range centroids in 1997 and 2018 for each highlighted species, along with grey points that represent survey routes where the species was estimated to be present in at least 1 year.

For example, neural hierarchical models could be used to couple a convolutional neural network observation model for camera trap images (Norouzzadeh *et al*. 2018; Tabak *et al*. 2019) with an ecological process model that describes animal density, movement, or community composition (Burton *et al*. 2015). Gridded data such as modelled climate data, remotely sensed Earth observations, and even 96 well microplates also might use convolutional neural network architectures (Rawat & Wang 2017). Appendix S2 provides a simulated example where gridded data (aerial imagery) are mapped to a state transition matrix of a hidden Markov model. In addition, many ecological data sets exhibit graph structure related to phylogenies, social networks or network-like spatial structure. While it is possible to adapt convolutional neural networks to operate on distance matrices computed from graphs (Fioravanti *et al*. 2018), graph representation learning can also provide embeddings for nodes that encode network structure and node attributes (Hamilton et al. 2017; Cai *et al*. 2018).

Neural hierarchical models can scale to data that are too large to fit in computer memory. Indeed, memory limitations precluded comparison of a fully Bayesian multi-species dynamic occupancy model against the multi-species neural hierarchical model. The current state of the art multi-species occupancy models use approximately one-tenth of the number of species, at about half the number of sites, and are static in the sense that colonisation and extinction dynamics through time are not represented (Tobler *et al*. 2019). Species distribution models are increasingly scalable due to advances in approximate Gaussian process models (Tikhonov *et al*. 2019), but multi-species dynamic occupancy models have not previously been reported at this scale. This is particularly relevant for extensions of the BBS case study, given the volume of (imperfect) bird data accumulating through citizen science programs (Sullivan *et al*. 2009). The key strategy providing scalability in the case study is stochastic optimisation that uses mini-batches – small subsets of a larger data set – to generate noisy estimates of model performance and partial

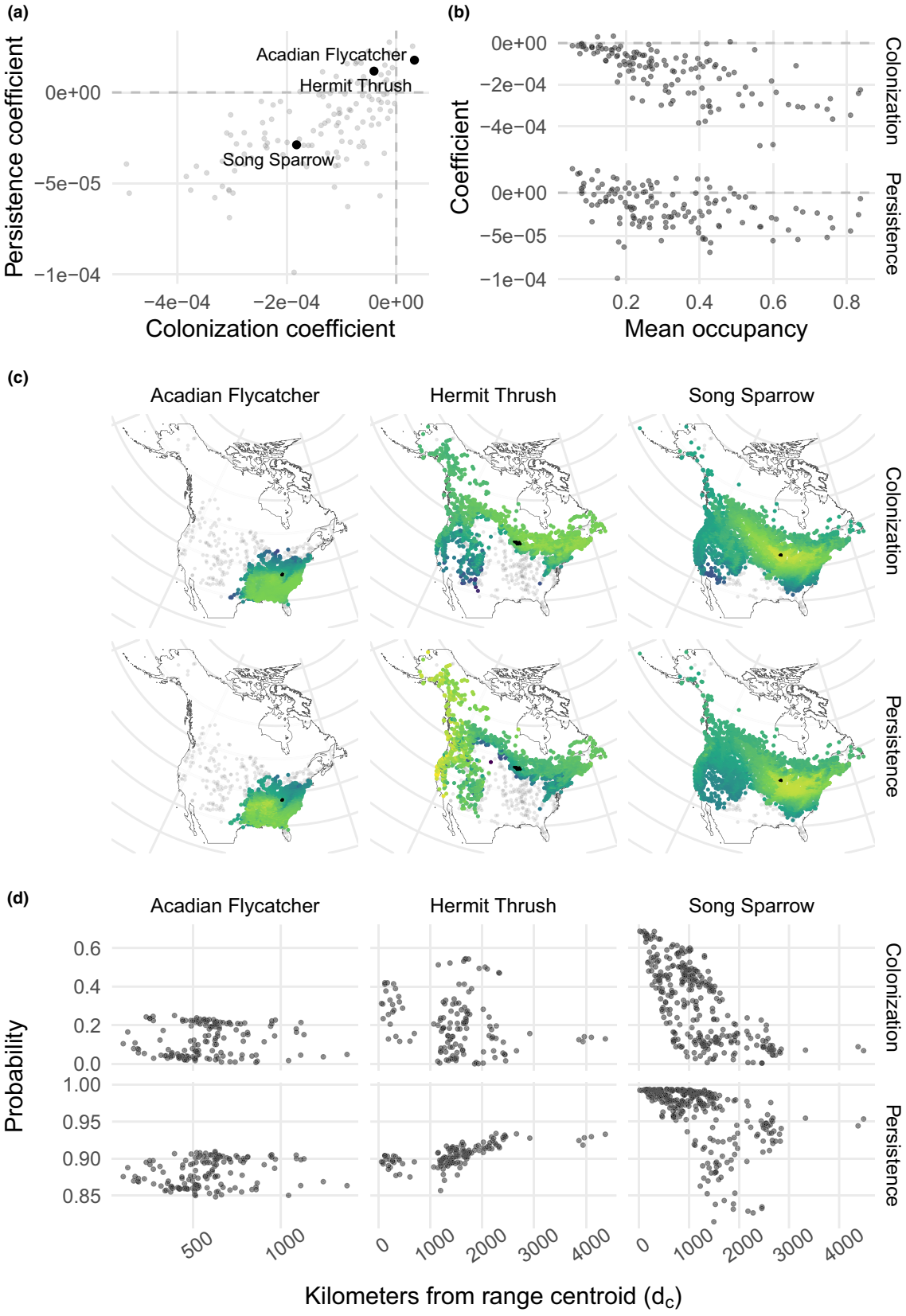

Figure 4 (a) Scatter plot relating species-specific distance decay coefficients for colonisation and persistence, with focal species from each quadrant highlighted. (b) Mean finite-sample occupancy (x-axis) vs. coefficients relating distance from range centroid and the probability of colonisation (y-axis, top row), and persistence (y-axis, bottom row). (c) Maps of colonisation and persistence probabilities at each route where focal species were likely absent (in grey) and present (in colour, normalised to increase the visibility of gradients). Centroids for each year are shown in solid black. (d) Colonisation and persistence probabilities (y-axis) as a function of distance from range centroid, averaged among years.

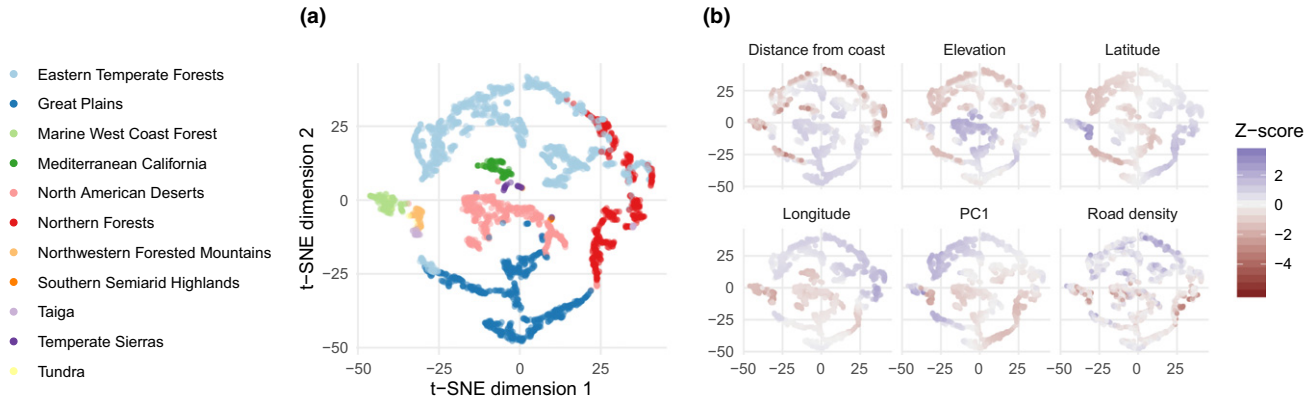

Figure 5 Clustering of North American Breeding Bird Survey routes by ecoregion and route-level features. All panels show a two dimensional plot of route vectors, computed using t-distributed stochastic neighbour embedding (t-SNE) on the latent vectors for initial occupancy, persistence, colonisation and detection probabilities. Each route is shown as a point. Colour indicates (a) EPA level 1 ecoregion and (b) z-standardised continuous route-level features. PC1 refers to the first principal component of WorldClim climate data.

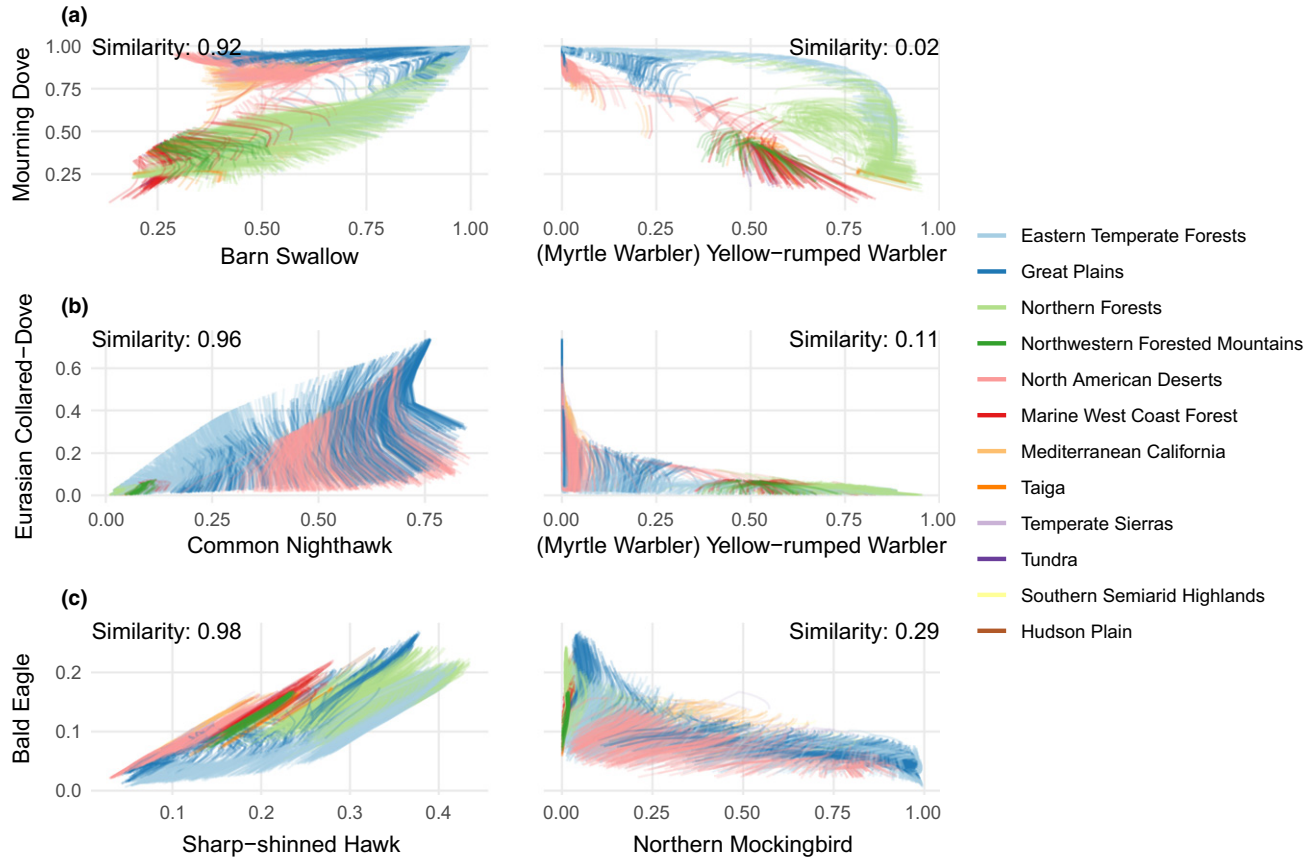

Figure 6 Relationships among species estimated occupancy probabilities through time, by location for species pairs that are nearest neighbours in parameter cosine distance (left column), and most cosine dissimilar (right column). Colour indicates EPA level 1 ecoregions. Cosine similarity values for species pairs are printed in the upper corners of each panel. Each route is a line segment that connects estimates of occupancy probabilities for each year from 1997 to 2018. Species pairs with high cosine similarity tend to have positively related occupancy trajectories.

derivatives that can be used during training (Ruder 2016). With this approach, the entire data set does not need to be loaded into memory at the same time. This could be useful as a way to scale models that integrate BBS, eBird and other data (Ngiam et al. 2011; Pacifici *et al.* 2017; Zipkin *et al.* 2017; Lin *et al.* 2019).

Although neural hierarchical models can perform well in the large data regime, they might be overparameterised in a small data setting. For this reason, performance comparisons with simpler baseline models are useful. Appendix S2 provides a simulated example, where simple baselines work better with small data sets, and more complex neural hierarchical models

work better with large data sets. Approaches familiar to ecologists such as k-fold cross-validation and (to a lesser extent) information criteria such as the Akaike information criterion (AIC) have been used with neural networks, but such approaches are less common than cross-validation with training, validation and test partitions of the data (Anderson & Burnham 2004; Jiang & Chen 2016; Ran & Hu 2017). With large data sets, considerable computational resources are required for training, so that retraining the same model many times might make k-fold cross-validation infeasible. As a practical matter, even if a neural hierarchical model demonstrates superior predictive performance, a simpler model might be preferred if interpretability and/or explainability is a high priority, as it might be in a decision-making context (Rudin 2018).

Neural networks have a reputation for being 'black-box' models. However, the interpretation and explanation of such models is an active area of research (Roscher et al. 2019). Interpretations map abstract concepts to domains that humans can make sense of, for example mapping neural network parameters to species identity or spatial location (Montavon *et al*. 2018). Explanations are collections of features in such a domain that contributed to a prediction (Montavon *et al*. 2018), for example sensitivity of model output to perturbations of the input (Olden & Jackson 2002; Gevrey *et al*. 2003).

Looking ahead, in a forecasting or decision-making setting, it would be important to estimate both aleatoric uncertainty (arising from noise in an observation process) and epistemic uncertainty (arising from uncertainty in a model and its parameters; Clark *et al*. 2001; Kendall & Gal 2017). Recent advances in accounting for uncertainty in neural network parameter estimates and architectures could be applied in future work. Approaches include 'Bayes by backpropagation' (Blundell et al. 2015), normalising flows for variational approximations (Kingma et al. 2016), adversarial training (Lakshminarayanan et al. 2017), methods that use dropout or its continuous relaxation (Gal et al. 2017), and ensemble approaches (McDermott & Wikle 2017). These methods can also help explain neural networks, for example to probabilistically estimate sensitivity to model inputs to random masking (Chang *et al*. 2017), or to decompose predictive uncertainty into component parts (Thiagarajan *et al*. 2019).

The potential for combining neural networks with mechanistic models was recognised more than 30 years ago (Psichogios & Ungar 1992; Meade & Fernandez 1994; Lagaris *et al*. 1998). This potential is more easily realised today due to methodological spillover from deep learning into the natural sciences, but also increases in computing power, availability of ecological data, and the proliferation of educational content for quantitative ecology and machine learning. Furthermore, modern deep learning frameworks provide abstractions that allow users to focus on model construction rather than the details of implementation, increasing accessibility in the same way that WinBUGS, JAGS, OpenBUGS and Stan have done (Lunn *et al*. 2000; Plummer & others 2003; Spiegelhalter et al. 2005; Carpenter *et al*. 2017).

Although deep learning and ecological modelling may seem to be separate activities, neural hierarchical models bridge these disciplines. Given the increasing availability of massive ecological data, scalable and flexible science-based models are increasingly needed. Neural hierarchical models satisfy this need, and can provide a framework that links imperfect observational data to ecological processes and mechanisms by construction.

ACKNOWLEDGEMENTS

The authors thank David Zonana and Roland Knapp for discussing on the potential of neural hierarchical models and Susie Ellis for providing feedback on a draft of the manuscript. This manuscript was also greatly improved by three anonymous reviewers. The authors also thank Brandon Edwards for developing the bbsBayes R package, which was used to acquire and parse the North American Breeding Bird Survey data. This work was motivated by a workshop on machine learning in Earth science, hosted by Earth Lab and organised by the Federation of Earth Science Information Partners and the National Aeronatics and Space Administration Advanced Information Systems Technology program. This work was made possible by the CU Boulder Grand Challenge initiative and the Cooperative Institute for Research in the Environmental Sciences through their investment in Earth Lab. The author extends their thanks to all the participants of US and Canada who annually perform and coordinate the North American Breeding Bird Survey.

AUTHORSHIP

MJ designed and performed the research, and wrote the manuscript.

DATA AVAILABILITY STATEMENT

The data are available from the United States Geological Survey Patuxent Wildlife Research Center: https://www.pwrc.usgs.gov/bbs/ (Pardieck et al. 2019).

REFERENCES

Anderson, D. & Burnham, K. (2004). *Model Selection and Multi-Model Inference*. Springer-Verlag, 2nd. NY, p. 63.

Austin, M. (2002). Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. *Ecol. Model.*, 157, 101–118.

Ba, Y., Zhao, G. & Kadambi, A. (2019). Blending diverse physical priors with neural networks. arXiv preprint arXiv:1910.00201.

Berliner, L.M. (1996). Hierarchical bayesian time series models. In: *Maximum Entropy and Bayesian Methods* (eds Hanson K.M. & Silver R.N.). Springer, New York, pp. 15–22.

Bled, F., Royle, J.A. & Cam, E. (2011). Hierarchical modeling of an invasive spread: The eurasian collared-dove streptopelia decaocto in the united states. *Ecol. Appl.*, 21, 290–302.

Bled, F., Nichols, J.D. & Altwegg, R. (2013). Dynamic occupancy models for analyzing species' range dynamics across large geographic scales. *Ecol. Evol.*, 3, 4896–4909.

Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. (2015). Weight uncertainty in neural networks. arXiv preprint arXiv:1505.05424.

Burton, A.C., Neilson, E., Moreira, D., Ladle, A., Steenweg, R., Fisher, J.T. *et al.* (2015). Wildlife camera trapping: A review and recommendations for linking surveys to ecological processes. *J. Appl. Ecol.*, 52, 675–685.

Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O. & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, 559, 547.

Cai, H., Zheng, V.W. & Chang, K.C.-C. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.*, 30, 1616–1637.

Calvert, A.M., Bonner, S.J., Jonsen, I.D., Flemming, J.M., Walde, S.J. & Taylor, P.D. (2009). A hierarchical bayesian approach to multi-state mark–recapture: Simulations and applications. *J. Appl. Ecol.*, 46, 610–620.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., et al. (2019). Machine learning and the physical sciences. arXiv preprint arXiv:1903.10563.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M. *et al.* (2017). Stan: A probabilistic programming language. *J. Stat. Software*, 76.

Chang, C.-H., Creager, E., Goldenberg, A. & Duvenaud, D. (2017). Interpreting neural network classifications with variational dropout saliency maps. *Proc. NIPS*, 6.

Chen, D.-G. & Hare, S.R. (2006). Neural network and fuzzy logic models for pacific halibut recruitment analysis. *Ecol. Mod.*, 195, 11–19.

Chen, D., Xue, Y., Chen, S., Fink, D. & Gomes, C. (2016). Deep multi-species embedding. arXiv preprint arXiv:1609.09353.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P. *et al.* (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface*, 15, 20170387.

Chon, T.-S., Kwak, I.-S., Park, Y.-S., Kim, T.-H. & Kim, Y. (2001). Patterning and short-term predictions of benthic macroinvertebrate community dynamics by using a recurrent artificial neural network. *Ecol. Model.*, 146, 181–193.

Christin, S., Hervet, E. & Lecomte, N. (2018). Applications for deep learning in ecology, bioRxiv, 334854.

Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Clark, J.S., Carpenter, S.R., Barber, M., Collins, S., Dobson, A., Foley, J.A. *et al.* (2001). Ecological forecasts: An emerging imperative. *Science*, 293, 657–660.

Cressie, N., Calder, C.A., Clark, J.S., Hoef, J.M.V. & Wikle, C.K. (2009). Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling. *Ecol. Appl.*, 19, 553–570.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, 2, 303–314.

Dallas, T., Decker, R.R. & Hastings, A. (2017). Species are not most abundant in the centre of their geographic range or climatic niche. *Ecol. Lett.*, 20, 1526–1533.

Desjardins-Proulx, P., Poisot, T. & Gravel, D. (2019). Artificial intelligence for ecological and evolutionary synthesis. *Front. Ecol. Evol.*, 7. https://doi.org/10.3389/fevo.2019.00402

Doherty, P.F. Jr, Boulinier, T. & Nichols, J.D. (2003). Local extinction and turnover rates at the edge and interior of species' ranges. In: Annales zoologici fennici. JSTOR, 145–153.

Dorazio, R.M., Kery, M., Royle, J.A. & Plattner, M. (2010). Models for inference in dynamic metacommunity systems. *Ecology*, 91, 2466–2475.

Fick, S.E. & Hijmans, R.J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.*, 37, 4302–4315.

Fioravanti, D., Giarratano, Y., Maggio, V., Agostinelli, C., Chierici, M., Jurman, G. *et al.* (2018). Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinformatics*, 19, 49.

Flagel, L., Brandvain, Y. & Schrider, D.R. (2018). The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Mol. Biol. Evol.*, 36, 220–238.

Fortin, M.-J., Keitt, T.H., Maurer, B., Taper, M., Kaufman, D.M. & Blackburn, T. (2005). Species' geographic ranges and distributional limits: Pattern analysis and statistical issues. *Oikos*, 108, 7–17.

Fricker, G.A., Ventura, J.D., Wolf, J.A., North, M.P., Davis, F.W. & Franklin, J. (2019). A convolutional neural network classifier identifies tree species in mixed-conifer forest from hyperspectral imagery. *Remote Sens.*, 11, 2326.

Gal, Y., Hron, J. & Kendall, A. (2017). Concrete dropout. In: *Advances in Neural Information Processing Systems*. pp. 3581–3590.

Gazestani, V.H. & Lewis, N.E. (2019). From genotype to phenotype: Augmenting deep learning with networks and systems biology. Current Opinion. *Systems Biol.*, 15, 68–73.

Gelman, A., Hwang, J. & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Stat. Comput.*, 24, 997–1016.

Gevrey, M., Dimopoulos, I. & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol. Model.*, 160, 249–264.

Golding, N. & Purse, B.V. (2016). Fast and flexible bayesian species distribution modelling using gaussian processes. *Methods Ecol. Evol.*, 7, 598–608.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. MIT press, Cambridge.

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: Advances, challenges and opportunities. *Ecography*, 40, 281–295.

Guo, C. & Berkhahn, F. (2016). Entity embeddings of categorical variables. arXiv preprint arXiv:1604.06737.

Hamilton, W., Ying, Z. & Leskovec, J. (2017). Inductive representation learning on large graphs. In: *Advances in Neural Information Processing Systems*. pp. 1024–1034.

Harris, D.J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods Ecol. Evol.*, 6, 465–473.

Hefley, T.J., Broms, K.M., Brost, B.M., Buderman, F.E., Kay, S.L., Scharf, H.R. *et al.* (2017). The basis function approach for modeling autocorrelation in ecological data. *Ecology*, 98, 632–646.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9, 1735–1780.

Holt, R.D. & Keitt, T.H. (2000). Alternative causes for range limits: A metapopulation perspective. *Ecol. Lett.*, 3, 41–47.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4, 251–257.

Hutchinson, R.A., Liu, L.-P. & Dietterich, T.G. (2011). Incorporating boosted regression trees into ecological latent variable models. In: Twenty-fifth aaai conference on artificial intelligence

Jain, A., Zamir, A.R., Savarese, S. & Saxena, A. (2016). Structural-rnn: Deep learning on spatio-temporal graphs. In: *Proceedings of the ieee conference on computer vision and pattern recognition*. pp. 5308-5317

Jeong, K.-S., Joo, G.-J., Kim, H.-W., Ha, K. & Recknagel, F. (2001). Prediction and elucidation of phytoplankton dynamics in the nakdong river (korea) by means of a recurrent artificial neural network. *Ecol. Model.*, 146, 115–129.

Jeong, K.-S., Kim, D.-K., Jung, J.-M., Kim, M.-C. & Joo, G.-J. (2008). Non-linear autoregressive modelling by Temporal Recurrent Neural Networks for the prediction of freshwater phytoplankton dynamics. *Ecological Modelling*, 211, 292–300.

Jiang, P. & Chen, J. (2016). Displacement prediction of landslide based on generalized regression neural networks with k-fold cross-validation. *Neurocomputing*, 198, 40–47.

Johnson, D.S., Thomas, D.L., Ver Hoef, J.M. & Christ, A. (2008). A general framework for the analysis of animal resource selection from telemetry data. *Biometrics*, 64, 968–976.

Johnson, D.S., Laake, J.L. & Ver Hoef, J.M. (2010). A model-based approach for making ecological inference from distance sampling data. *Biometrics*, 66, 310–318.

Jolly, G.M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52, 225–247.

Karpatne, A., Watkins, W., Read, J. & Kumar, V. (2017). Physics-guided neural networks (pgnn): An application in lake temperature modeling. arXiv preprint arXiv:1710.11431.

Kendall, A. & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In: *Advances in neural information processing systems*. pp. 5574–5584.

Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I. & Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In: *Advances in neural information processing systems*. pp. 4743–4751.

Knapp, R.A., Matthews, K.R., Preisler, H.K. & Jellison, R. (2003). Developing probabilistic models to predict amphibian site occupancy in a patchy landscape. *Ecol. Appl.*, 13, 1069–1082.

Knouft, J.H. (2018). Appropriate application of information from biodiversity databases is critical when investigating species distributions and diversity: A comment on dallas et al. (). *Ecol. Letters*, 21, 1119–1120.

Krishnan, R.G., Shalit, U. & Sontag, D. (2017). Structured inference networks for nonlinear state space models. In: *Thirty-first aaai conference on artificial intelligence*. pp. 2101–2109.

Lagaris, I.E., Likas, A. & Fotiadis, D.I. (1998). Artificial neural networks for solving ordinary and partial differential equations. *IEEE Trans. Neural Networks*, 9, 987–1000.

Lakshminarayanan, B., Pritzel, A. & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Advances in neural information processing systems*. pp. 6402–6413.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D. & Morales, J.M. (2012). Flexible and practical modeling of animal telemetry data: Hidden markov models and extensions. *Ecology*, 93, 2336–2342.

Latimer, A., Banerjee, S., Sang, H. Jr, Mosher, E. & Silander, J. Jr (2009). Hierarchical models facilitate spatial analysis of large data sets: A case study on invasive plant species in the northeastern united states. *Ecol. Lett.*, 12, 144–154.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, 521 (7553), 436–444.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J. & Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. *Ecol. Model.*, 90, 39–52.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X. & Chen, B. (2018). PointCNN: Convolution on x-transformed points. In: *Advances in Neural Information Processing Systems*. pp. 820–830.

Lin, T.-Y., Winner, K., Bernstein, G., Mittal, A., Dokter, A.M., Horton, K.G. *et al.* (2019). MistNet: Measuring historical bird migration in the US using archived weather radar data and convolutional neural networks. *Methods Ecol. Evol.*, 10, 1908–1922.

Link, W.A. & Sauer, J.R. (1998). Estimating population change from count data: Application to the north american breeding bird survey. *Ecol. Appl.*, 8, 258–268.

Van Lint, J., Hoogendoorn, S. & van Zuylen, H.J. (2002). Freeway travel time prediction with state-space neural networks: Modeling state-space dynamics with recurrent neural networks. *Transp. Res. Rec.*, 1811, 30–39.

Long, J., Shelhamer, E. & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In: *Proceedings of the ieee conference on computer vision and pattern recognition*. pp. 3431-3440

Lu, Z., Pu, H., Wang, F., Hu, Z. & Wang, L. (2017). The expressive power of neural networks: A view from the width. In: *Advances in Neural Information Processing Systems*. pp. 6231–6239.

Lunn, D.J., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS-a bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput.*, 10, 325–337.

van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learning Res.*, 9, 2579–2605.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J. & Langtimm, C.A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83, 2248–2255.

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

Malek, S., Salleh, A., Milow, P., Baba, M.S. & Sharifah, S. (2012). Applying artificial neural network theory to exploring diatom abundance at tropical Putrajaya lake, Malaysia. *J. Freshwater Ecol.*, 27, 211–227.

Manel, S., Dias, J.-M. & Ormerod, S.J. (1999). Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: A case study with a himalayan river bird. *Ecol. Model.*, 120, 337–347.

McDermott, P.L. & Wikle, C.K. (2017). An ensemble quadratic echo state network for non-linear spatio-temporal forecasting. *Stat*, 6, 315–330.

McDermott, P.L. & Wikle, C.K. (2019). Deep echo state networks with uncertainty quantification for spatio-temporal forecasting. *Environmetrics*, 30, e2553.

Meade, A.J. Jr & Fernandez, A.A. (1994). The numerical solution of linear ordinary differential equations by feedforward neural networks. *Math. Comp. Mod.*, 19, 1–25.

Mehlman, D.W. (1997). Change in avian abundance across the geographic range in response to environmental change. *Ecol. Appl.*, 7, 614–624.

Meijer, J.R., Huijbregts, M.A., Schotten, K.C. & Schipper, A.M. (2018). Global patterns of current and future road infrastructure. *Environ. Res. Lett.*, 13, 064006.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Miller, D.L., Burt, M.L., Rexstad, E.A. & Thomas, L. (2013). Spatial models for distance sampling data: Recent developments and future directions. *Methods Ecol. Evol.*, 4, 1001–1010.

Montavon, G., Samek, W. & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1–15.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. & Ng, A.Y. (2011). Multimodal deep learning. In: *Proceedings of the 28th international conference on machine learning (icml-11)*. pp. 689–696.

Niepert, M., Ahmed, M. & Kutzkov, K. (2016). Learning convolutional neural networks for graphs. In: *International conference on machine learning*. pp. 2014-2023.

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C. *et al.* (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proc. Natl Acad. Sci.*, 115, E5716–E5725.

Oksanen, J. & Minchin, P.R. (2002). Continuum theory revisited: What shape are species responses along ecological gradients? *Ecol. Model.*, 157, 119–129.

Olden, J.D. & Jackson, D.A. (2002). Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks. *Ecol. Model.*, 154, 135–150.

Ovaskainen, O., Roy, D.B., Fox, R. & Anderson, B.J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods Ecol. Evol.*, 7, 428–436.

Özesmi, S.L., Tan, C.O. & Özesmi, U. (2006). Methodological issues in building, training, and testing artificial neural networks in ecological applications. *Ecol. Model.*, 195, 83–93.

Pacifici, K., Reich, B.J., Miller, D.A., Gardner, B., Stauffer, G., Singh, S. *et al.* (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, 98, 840–850.

Paganini, M., de Oliveira, L. & Nachman, B. (2018). Accelerating science with generative adversarial networks: An application to 3D particle showers in multilayer calorimeters. *Phys. Rev. Lett.*, 120, 042003.

Pardieck, L. K., Ziolkowski, D. J. Jr., Lutmerding, M., Aponte, V., & Hudson, M-A. R., (2019). North american breeding bird survey

dataset, 1966–2018. version 2018.0. U.S. Geological Survey, Patuxent Wildlife Research Center. https://doi.org/10.5066/P9HE8XYJ
Patterson, T.A., Basson, M., Bravington, M.V. & Gunn, J.S. (2009). Classifying movement behaviour in relation to environmental conditions using hidden markov models. *J. Anim. Ecol*., 78, 1113–1123.
Patterson, T.A., Parton, A., Langrock, R., Blackwell, P.G., Thomas, L. & King, R. (2017). Statistical modelling of individual animal movement: An overview of key methods and a discussion of practical challenges. *AStA Advances Stat. Anal*., 101, 399–438.
Plummer, M. (2003). JAGS: A program for analysis of bayesian graphical models using gibbs sampling. In: *Proceedings of the 3rd international workshop on distributed statistical computing*. Vienna, Austria.
Psichogios, D.C. & Ungar, L.H. (1992). A hybrid neural network-first principles approach to process modeling. *AIChE J*., 38, 1499–1511.
Radovic, A., Williams, M., Rousseau, D., Kagan, M., Bonacorsi, D., Himmel, A. *et al*. (2018). Machine learning at the energy and intensity frontiers of particle physics. *Nature*, 560, 41.
Raissi, M. (2018). Deep hidden physics models: Deep learning of nonlinear partial differential equations. *J. Mach. Learning Res*., 19, 932–955.
Ran, Z.-Y. & Hu, B.-G. (2017). Parameter identifiability in statistical machine learning: A review. *Neural Comput*., 29, 1151–1203.
Rangapuram, S.S., Seeger, M.W., Gasthaus, J., Stella, L., Wang, Y. & Januschowski, T. (2018). Deep state space models for time series forecasting. In: *Advances in neural information processing systems*. pp. 7785–7794.
Rauber, P.E., Fadel, S.G., Falcao, A.X. & Telea, A.C. (2016). Visualizing the hidden activity of artificial neural networks. *IEEE Trans. Visual Comput. Graphics*, 23, 101–110.
Rawat, W. & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Comput*., 29, 2352–2449.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. *et al*. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, 566, 195.
Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G. *et al*. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40, 913–929.
Roscher, R., Bohn, B., Duarte, M.F. & Garcke, J. (2019). Explainable machine learning for scientific insights and discoveries. arXiv preprint arXiv:1905.08883.
Royle, J.A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60, 108–115.
Royle, J.A. & Kéry, M. (2007). A bayesian state-space formulation of dynamic occupancy models. *Ecology*, 88, 1813–1823.
Royle, J.A. & Link, W.A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87, 835–841.
Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.
Rudin, C. (2018). Please stop explaining black box models for high stakes decisions. arXiv preprint arXiv:1811.10154.
Russell, S.J. & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson Education Limited, Malaysia.
Sagarin, R.D. & Gaines, S.D. (2002). The 'abundant centre'distribution: To what extent is it a biogeographical rule? *Ecol. Lett*., 5, 137–147.
Sauer, J.R. & Link, W.A. (2011). Analysis of the north american breeding bird survey using hierarchical models. *Auk*, 128, 87–98.
Sauer, J.R., Link, W.A., Fallon, J.E., Pardieck, K.L. & Ziolkowski, D.J. Jr (2013). The north american breeding bird survey 1966–2011: Summary analysis and species accounts. *North American Fauna*, 79, 1–32.
Sauer, J.R., Niven, D.K., Pardieck, K.L., Ziolkowski, D.J. Jr & Link, W.A. (2017). Expanding the north american breeding bird survey analysis to include additional species and regions. *Journal of Fish and Wildlife Management*, 8, 154–172.
Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2005). OpenBUGS version 2.10, user manual. MRC Biostatistics Unit, Cambridge, United Kingdom.
Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D. & Kelling, S. (2009). EBird: A citizen-based bird observation network in the biological sciences. *Biol. Cons*., 142, 2282–2292.
Sutskever, I., Vinyals, O. & Le, Q.V. (2014). Sequence to sequence learning with neural networks. In: *Advances in neural information processing systems*. pp. 3104–3112.
Tabak, M.A., Norouzzadeh, M.S., Wolfson, D.W., Sweeney, S.J., VerCauteren, K.C., Snow, N.P. *et al*. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods Ecol. Evol*., 10, 585–590.
Thiagarajan, J.J., Kim, I., Anirudh, R. & Bremer, P.-T. (2019). Understanding deep neural networks through input uncertainties. In: *ICASSP 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, pp. 2812-2816.
Thorson, J.T., Scheuerell, M.D., Shelton, A.O., See, K.E., Skaug, H.J. & Kristensen, K. (2015). Spatial factor analysis: A new tool for estimating joint species distributions and correlations in species range. *Methods Ecol. Evol*., 6, 627–637.
Thorson, J.T., Ianelli, J.N., Larsen, E.A., Ries, L., Scheuerell, M.D., Szuwalski, C. *et al*. (2016). Joint dynamic species distribution models: A tool for community ordination and spatio-temporal monitoring. *Glob. Ecol. Biogeogr*., 25, 1144–1158.
Tikhonov, G., Abrego, N., Dunson, D. & Ovaskainen, O. (2017). Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods Ecol. Evol*., 8, 443–452.
Tikhonov, G., Duan, L., Abrego, N., Newell, G., White, M., Dunson, D. *et al*. (2019). Computationally efficient joint species distribution modeling of big spatial data. *Ecology*, e02929.
Tobler, M.W., Kéry, M., Hui, F.K., Guillera-Arroita, G., Knaus, P. & Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, 100, e02754.
De Valpine, P. & Hastings, A. (2002). Fitting population models incorporating process noise and observation error. *Ecol. Monogr*., 72, 57–76.
Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, 13, 260–269.
Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C. *et al*. (2015). So many variables: Joint modeling in community ecology. *Trends Ecol. Evol*., 30, 766–779.
Webb, M.H., Wotherspoon, S., Stojanovic, D., Heinsohn, R., Cunningham, R., Bell, P. *et al*. (2014). Location matters: Using spatially explicit occupancy models to predict the distribution of the highly mobile, endangered swift parrot. *Biol. Cons*., 176, 99–108.
Wikle, C.K. (2003). Hierarchical models in environmental science. *Inter. Stat. Rev*., 71, 181–199.
Wikle, C.K. (2019). Comparison of deep neural networks and deep hierarchical models for spatio-temporal data. *J. Agri. Biol. Environ. Stat*., 24, 175–203.
Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K. & Woo, W.-C. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: *Advances in Neural Information Processing Systems*. pp. 802–810.
Yuan, Y., Bachl, F.E., Lindgren, F., Borchers, D.L., Illian, J.B., Buckland, S.T. *et al*. (2017). Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *Ann. App. Statist*., 11, 2270–2297.
Zamarreño, J.M. & Vega, P. (1998). State space neural network. Properties and application. *Neural Networks*, 11, 1099–1112.
Zipkin, E.F., Rossman, S., Yackulic, C.B., Wiens, J.D., Thorson, J.T., Davis, R.J. *et al*. (2017). Integrating count and detection–nondetection data to model population dynamics. *Ecology*, 98, 1640–1650.

Zwane, E. & Van der Heijden, P. (2004). Semiparametric models for capture–recapture studies with covariates. *Comput. Stat. Data Anal.*, 47, 729–743.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Editor, Carl Boettiger
Manuscript received 16 September 2019
First decision made 17 October 2019
Manuscript accepted 23 December 2019