Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015 doi: 10.1111/2041-210X.12332

Generating realistic assemblages with a joint species distribution model

David J. Harris*

Center for Population Biology, 1 Shields Avenue, Davis, CA 95616, USA

Summary

1. Species distribution models (SDMs) represent important analytical and predictive tools for ecologists. Until now, these models have either assumed (i) that species' occurrence probabilities are uncorrelated or (ii) that species respond linearly to preselected environmental variables. These two assumptions currently prevent ecologists from modelling assemblages with realistic co-occurrence and species richness properties.

2. This paper introduces a stochastic feedforward neural network, called 'mistnet', which makes neither assumption. Thus, unlike most SDMs, mistnet can account for non-independent co-occurrence patterns driven by unobserved environmental heterogeneity. And unlike several recently proposed joint SDMs, the model can also learn nonlinear functions relating species' occurrence probabilities to environmental predictors.

3. Mistnet makes more accurate predictions about the North American bird communities found along Breeding Bird Survey transects than several alternative methods tested. In particular, typical assemblages held out of sample for validation were each tens of thousands of times more likely under the mistnet model than under independent combinations of single-species predictions.

4. Apart from improved accuracy, mistnet shows two other important benefits for ecological research and management. First: by analysing co-occurrence data, mistnet can identify unmeasured – and perhaps unanticipated – environmental variables that drive species turnover. For example, the model identified a strong grassland/forest gradient, even though only temperature and precipitation were given as model inputs. Second: mistnet is able to take advantage of outside information to guide its predictions towards more realistic assemblages. For example, mistnet automatically adjusts its expectations to include more forest-associated species in response to a stray observation of a forest-dwelling warbler.

Key-words: birds, breeding bird survey, neural network, probabilistic graphical model, species assemblages, species distribution model

Introduction

A major goal of community ecology is to understand the processes, such as environmental filtering and species interactions, that determine where species could occur and which species can occur together (Chase 2003). Traditional multivariate methods for studying these issues in community ecology – such as ordination techniques for summarizing a data matrix's multivariate geometry – will not always provide the best approach to these questions, as they typically do not specify a data-generating mechanism or make predictions about new assemblages (but see Walker & Jackson 2011). More recent approaches, such as generalized linear models (Jackson *et al.* 2012; Wang *et al.* 2012; Jamil & ter Braak 2013) and species distribution models (SDMs; Elith *et al.* 2006), can make specific predictions. Just as importantly, these predictions can be evaluated quantitatively based on their likelihoods.

Modern SDMs need not assume that species respond to environmental variation in a pre-specified way (e.g. linearly or quadratically); relaxing this assumption has improved our ability to make predictions about individual species (Elith *et al.* 2006). For many community-level questions, however, species-level predictions may be of limited use. While SDMs can be combined ('stacked') to generate assemblage-level predictions (Pellissier *et al.* 2013), doing so implies that species' occurrence probabilities are uncorrelated (Clark *et al.* 2013; Calabrese *et al.* 2014). Ignoring the (potentially unobserved) factors driving these correlations can lead stacked models to generate incoherent jumbles of species rather than realistic assemblages (Clark *et al.* 2013). Given that most models only use climate variables as predictors (Austin & Van Niel 2011), the set of unobserved factors will usually include *all of ecology* apart from climatic influences. SDMs' failure to include other ecological processes is thus widely considered to be a major omission from statistical ecology's toolbox (Austin & Van Niel 2011; Guisan & Rahbek 2011; Kissling *et al.* 2012; Clark *et al.* 2013; Wisz *et al.* 2013).

In the last few years, several mixed models have been proposed to help explain the co-occurrence patterns that stacked SDMs ignore (Latimer *et al.* 2009; Ovaskainen, Hottola &

*Correspondence author. davharris@ucdavis.edu

Siitonen 2010; Clark *et al*. 2013; Golding 2013; Pollock *et al*. 2014). These *joint* species distribution models (JSDMs) can produce mixtures of possible species assemblages (points in Fig. 1a), rather than relying on a small number of environmental measurements to fully describe each species' probability of occurrence (which would collapse the distribution in Fig. 1a to a single point). In JSDMs (as in nature), a given set of climate estimates could be consistent with a number of different sets of co-occurring species, depending on factors that ecologists have not necessarily measured or even identified as important. JSDMs represent these unobserved (latent) factors as random variables whose true values are unknown, but whose existence would still help explain discrepancies between the data and the stacked SDMs' predictions (Fig. 1). While JSDMs represent a major advance in community-level modelling (Clark *et al*. 2013; Pollock *et al*. 2014), existing implementations have all assumed that species' responses to the environment are linear (in the sense of a generalized linear model), limiting their accuracy and utility.

Here, I present a new R package for assemblage-level modelling – called *mistnet* – that does not rely on independence (as stacks of single-species models do) or linearity (as previous JSDMs have). Mistnet models are stochastic feed-forward neural networks (Neal 1992; Tang & Salakhutdinov 2013) that combine the flexibility of modern nonlinear models with the latent variables found in previous JSDMs. To demonstrate the value of this approach, I compared mistnet's predictive likelihood with that of several existing models, using observational data from thousands of North American Breeding Bird Survey transects (BBS; Sauer *et al*. 2011). A high predictive likelihood indicates that the model correctly expects to see the kinds of assemblages that were actually found out of sample, while a very low likelihood means that the model has effectively ruled those assemblages out due to overfitting or underfitting.

An accurate JSDM would open up new possibilities for research and effective management. For example, although most models only have access to climate data (Austin & Van Niel 2011), a successful model of community structure should also be able to identify the major axes of non-climate variation that drive species turnover based on the species' observed co-occurrence patterns. Moreover, a successful assemblage-level model would be able to take advantage of the presence (or absence) of indicator species to inform its predictions about the rest of the assemblage. This ability to transfer information from easily detected, well-documented taxa to more cryptic or rare species would prove valuable for community ecologists and conservationists alike.

The mistnet source code can be freely downloaded from https://github.com/davharris/mistnet/releases.

Materials and methods

Methods are presented in five main sections:

- Data sets used
- Introduction to stochastic neural networks
- The specific model used here
- Summary of existing methods used for comparison
- Criteria for model evaluation

DATA

Field observations were obtained from the 2011 Breeding Bird Survey (BBS; Sauer *et al*. 2011). The BBS data consist of thousands of transects ('routes'), which served as the main unit for the analysis. Each route includes 50 stops, about 0·8 km apart. At each stop, all the birds observed in a 3-min period are recorded, using a standardized procedure. Following BBS recommendations, I omitted non-standard routes and data collected on days with bad weather.

To evaluate the SDMs' predictive capabilities, I split the routes into a 'training' data set consisting of 1559 routes and a 'test' data set consisting of 280 routes (Fig. 2; Appendix S1). The two data sets were separated by a 150-km buffer to ensure that models could not rely on spatial autocorrelation to make accurate predictions about the test set (cf. Bahn & McGill 2007; Appendix S1). Each model was fit to the same training set, and then its performance was evaluated out of sample on the test set.

Observational data for each species were reduced to 'presence' or 'absence' at the route level, ignoring the possibility of observation error for the purposes of this analysis. Three hundred and sixty-eight species

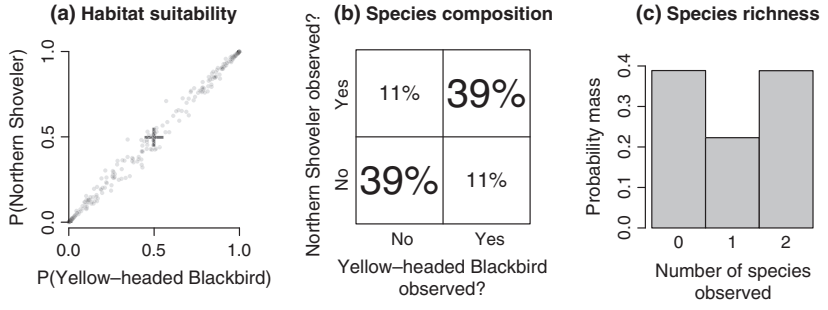

Fig. 1. Unobserved environmental heterogeneity can induce correlations between species; ignoring this heterogeneity can produce misleading results. (a) Based on climate, a pair of single-species models might predict 50% occurrence probabilities for each of two wetland species (black cross). However, a site's suitability for these species cannot be fully determined without information about the availability of wetland habitat. Most real habitats will be suitable for both species (dense cloud of points in upper-right corner) or neither (lower-left corner), depending on this unmeasured variable. (b) This correlation among species substantially alters the set of assemblages one would expect to observe. (Under independence, all four possibilities would be equally probable.) (c) Positive correlations among species can even induce a bimodal distribution of species richness values.

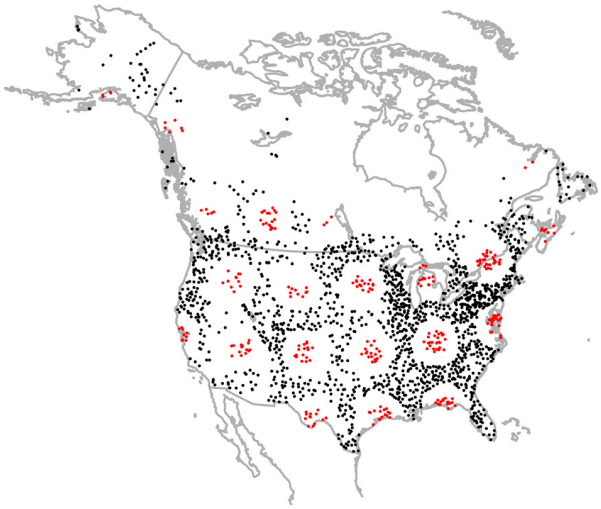

Fig. 2. Map of the BBS routes used in this analysis. Black points are training routes; red ones are test routes. The training and test routes are separated by a 150-km buffer to minimize spatial autocorrelation across the two partitions.

were chosen for analysis according to a procedure described in Appendix S1.

I extracted the 19 Bioclim climate variables for each route from Worldclim (version 1.4; Hijmans *et al.* 2005) for use as environmental predictors. After removing predictors that were nearly collinear, eight climate-based predictors remained for the analyses (Appendix S1). Since most SDMs do not use land cover data (Austin & Van Niel 2011) and one of mistnet's goals is to make inferences about unobserved environmental variation, no other variables were included in this analysis.

Finally, I obtained habitat classifications for each species from the Cornell Lab of Ornithology's 'All About Birds' website (AAB; www.allaboutbirds.org) using an R script written by K. E. Dybala.

INTRODUCTION TO STOCHASTIC NEURAL NETWORKS

This section discusses the stochastic networks in general terms; the specific model used for avian communities is discussed next. Ecologists have generally not had much success using neural networks for SDM (e.g. Dormann *et al.* 2008), but their recent success in other machine learning contexts (including contexts with latent random variables; Murphy 2012; Bengio 2013) makes them worth a second look. While one can build stochastic versions of other nonlinear regression methods as well (e.g. Hutchinson, Liu & Dietterich 2011), the relative simplicity of the backpropagation algorithm for training neural networks (Murphy 2012) makes them very appealing for exploratory research.

A neural net is a statistical model that makes its predictions by applying a series of nonlinear transformations to one or more predictor variables such as environmental measurements (Fig. 3; Appendix S2). After a suitable transformation of the environmental data, a final operation performs logistic regressions in the transformed space to make predictions about each species' occurrence probability (cf. Leathwick *et al.* 2005). Training a neural network entails simultaneously adjusting the parameters associated with these transformations to optimize the overall likelihood or posterior density (Appendix S3).

Most neural networks' predictions are deterministic functions of their inputs. Applied to SDMs, this would mean that each species' occurrence probability would be fully specified by the small number of variables that ecologists happen to measure. Mistnet's neural networks, in contrast, are *stochastic* (Neal 1992; Tang & Salakhutdinov 2013; Appendix S2), meaning that they allow species' occurrence probabilities to depend on unobserved environmental factors as well. The true values of these unobserved factors are (by definition) not known, but one can still represent their *possible* values using samples from a probability distribution. In the absence of any information about what these variables should represent, mistnet defaults to sampling them from standard normal distributions. Depending on which values are sampled (i.e. on the possible states of the environment), the model could expect to see radically different kinds of species assemblages (Figs 1 and 3).

Inference can also proceed backward through a stochastic network: the presence (or absence) of one species provides information about the local environment, which can then be used to make better predictions about other species. For example, suppose that a researcher has more data about the local distribution of waterfowl – which are of special interest to hunters and conservation groups – than about other species. If waterfowl species are known to be present along a given route, then a mistnet model could infer that suitable habitat must have been available to them. The model could then infer that the same habitat must have been available to other species, such as grebes and rails, with similar requirements. These species' predicted occurrence probabilities should thus increase automatically wherever waterfowl have been detected. Notably, the required correlations are automatically inferred from species' co-occurrence patterns, so the accuracy of these updated predictions does not depend closely on the user's ecological intuition about species' environmental tolerances.

As with most neural networks, a mistnet model's coefficients are initialized randomly, and then an optimization procedure iteratively adjusts the coefficients towards values with higher likelihoods via gradient-based hill climbing. In mistnet models, these adjustments are calculated with a variant of the backpropagation algorithm suggested by Tang & Salakhutdinov (2013; Appendix S3). This variant uses a Monte Carlo expectation-maximization procedure (Wei & Tanner 1990; also called generalized expectation-maximization, Neal & Hinton 1998), which alternates between inferring the states of the latent variables that produced the observed assemblages (via importance sampling) and updating the model's coefficients to make better predictions (via weighted backpropagation). While most stochastic networks rely on severely autocorrelated Markov chains for Monte Carlo sampling (Bengio 2013), mistnet's importance sampling-based approach generates independent samples very efficiently (Tang & Salakhutdinov 2013). By iteratively improving the model's estimates of the latent environmental factors and of the parameters governing species' responses to them, mistnet's generalized expectation-maximization procedure will eventually bring the model – with probability one – to a local maximum likelihood estimate. Importantly, this procedure works even if each step includes substantial Monte Carlo error (Neal & Hinton 1998; Tang & Salakhutdinov 2013).

Most successful neural networks are regularized to avoid overfitting, meaning that they operate on a *modified* likelihood surface that favours reduced model complexity (Murphy 2012). In the mistnet package, regularization is formulated as prior distributions favouring smaller magnitude parameter values over larger ones (Appendix S2). In Bayesian terms, this means that mistnet maximizes the model's posterior probability rather than the likelihood (maximum a posteriori estimation); in mathematically equivalent frequentist terms (Murphy 2012), mistnet optimizes a constrained or penalized likelihood.

The ill-conditioned and highly multimodal likelihood functions associated with most neural networks make it difficult to quantify the uncertainty associated with parameter estimates. For this reason, mistnet – like most neural network software – only provides point estimates

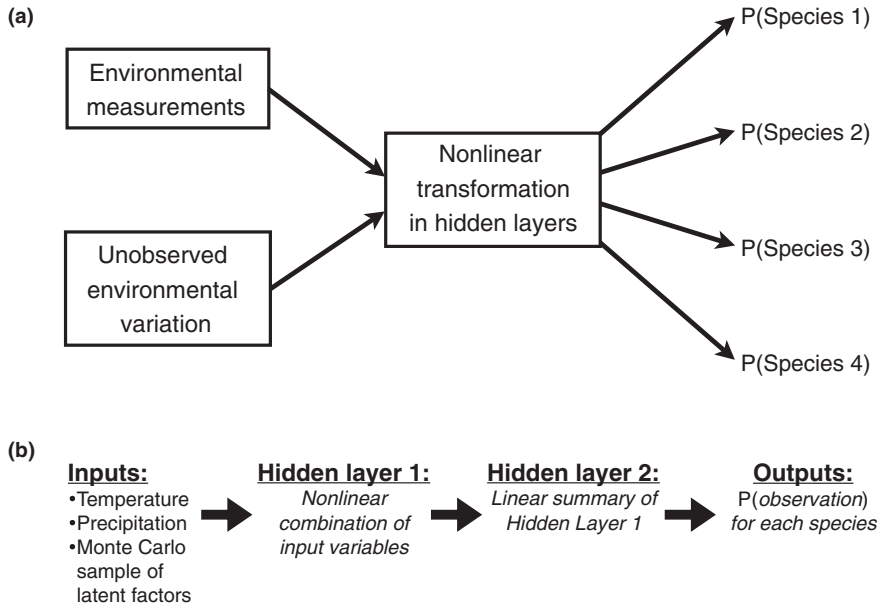

Fig. 3. (a) A generalized diagram for stochastic feed-forward neural networks that transform environmental variables into occurrence probabilities for multiple species. The network's hidden layers perform a nonlinear transformation of the observed and unobserved ('latent') environmental variables; each species' occurrence probability then depends on the state of the final hidden layer in a generalized linear fashion. (b) The specific network used in this paper, with two hidden layers. The inputs include Worldclim variables involving temperature and precipitation, as well as random draws from each of the latent environmental factors. These inputs are multiplied by a coefficient matrix and then nonlinearly transformed in the first hidden layer. The second hidden layer uses a different coefficient matrix to linearly reduce the dimensionality of the model's representation. A third matrix of coefficients links each species' occurrence probability to each of the variables in this linear summary (like one instance of logistic regression for each species). The coefficients are all learned using a variant of the backpropagation algorithm.

for model parameters, though its Monte Carlo samples do provide a measure of uncertainty regarding occurrence probabilities.

A MISTNET MODEL FOR BIRD ASSEMBLAGES

Mistnet models can take different forms, depending on the statistical or biological problems of interest. The model used here (Fig. 3b) uses two hidden layers that transform the environmental data into a form that is suitable for a linear classifier; the final layer essentially performs logistic regression in this transformed space. As discussed below, this structure is designed to improve the interpretability of the model, relative to other nonlinear SDMs.

Each hidden unit ('neuron') in the first layer is sensitive to a different linear combination of environmental features (e.g. hot-and-forested or wet-and-open). The hidden units' responses to the environment are nonlinear (Appendix S2), expressing the possibility that – for example – species might be more sensitive to a one-degree change in temperature from 25 to 26°C than to a change of the same magnitude from 19 to 20°C.

The second hidden layer collapses first layer's description of the environment down to a smaller number of values (e.g. 15 in this analysis; Appendix S4), using a linear transformation. This aspect of the network's structure ensures that the number of coefficients determining each species' response to the environment is manageable (e.g. one for each of the 15 transformed environmental variables, plus an intercept term). These few coefficients each have a consistent interpretation across species, essentially describing species' responses to the leading principal components of environmental variation (cf. Vincent *et al.* 2010). For comparison, the boosted regression tree SDMs used below (Elith, Leathwick & Hastie 2008) have tens of thousands of coefficients per species, and each species' coefficients have different interpretations.

Apart from limiting the number of coefficients per species, two additional factors constrained the model's capacity for overfitting. First, the coefficients in each layer were constrained using weak Gaussian priors, preventing any one variable from dominating the network. Second, a very weak Beta (1.000001, 1.000001) prior was used to reduce the prevalence of overconfident predictions ($|\text{odds ratio}| > 10^6$). The size of each layer and optimization details were chosen by cross-validation (see Appendix S4 for the settings that were evaluated, along with their cross-validated likelihoods).

EXISTING MODELS USED FOR COMPARISON

I compared mistnet's predictive performance with two machine learning techniques and with a linear JSDM. Each technique is described briefly below; see Appendix S4 for each model's settings.

The first machine learning method that I used for comparison, boosted regression trees (BRT), is among the most powerful techniques available for single-species SDM (Elith *et al.* 2006; Elith, Leathwick & Hastie 2008). I trained one BRT model for each species using the `gbm` package (Ridgeway 2013) and stacked them following the recommendations in Calabrese *et al.* (2014).

I also used a deterministic neural network from the `nnet` package (Venables & Ripley 2002) as a baseline to assess the importance of mistnet's latent random variables. This network shares some information among species (i.e. all species' occurrence probabilities depend on the same hidden layer), but like most other multispecies SDMs (Leathwick *et al.* 2005; Ferrier *et al.* 2007), it is not a JSDM and does not explicitly model co-occurrence (Clark *et al.* 2013).

Finally, I trained a linear JSDM using the BayesComm package (Golding 2013; Golding & Harris 2014) to assess the importance of mistnet's nonlinearities compared to a linear alternative that also models co-occurrence explicitly.

EVALUATING MODEL PREDICTIONS ALONG TEST ROUTES

I evaluated mistnet's predictions both qualitatively and quantitatively. Qualitative assessments involved looking for patterns in the model's predictions and comparing them with ornithological knowledge (e.g. the AAB habitat classifications).

Each model was evaluated quantitatively on the test routes (red points in Fig. 2) to assess its predictive accuracy out of sample. Models were scored according to their predictive likelihoods, i.e. the probabilities they assigned to various scenarios observed in the test data. Models with high likelihoods tend to produce realistic co-occurrence patterns and should yield more biologically relevant insights about the processes underlying those patterns. Models that overfit or underfit will have lower out-of-sample likelihoods, and drawing scientific conclusions from them could be unwise. I tested each model's ability to make several kinds of predictions, ranging from the species level to predictions about the richness and composition of entire assemblages. Models that assumed species were uncorrelated should see an exponential decay in their likelihoods as the number of species increases (since the probability of making correct predictions for a set of uncorrelated species equals the product of their individual probabilities), while BayesComm and mistnet should be able to simplify the problem for larger assemblages by using correlational information. As mistnet models' likelihoods involve intractable integrals over the latent variables, I approximated their likelihoods using importance sampling. Appendix S5 demonstrates that the sampling error introduced by this approximation is probably negligible.

In addition to assessing the models' overall likelihoods, I also focused on predictions about species richness by comparing the range of possible richness values they expected along each test route with what was actually observed. For each model, I used the Poisson binomial distribution (Hong 2013) to find confidence intervals for species richness, as described in Calabrese *et al.* (2014). The Poisson binomial distribution (not to be confused with the better-known Poisson distribution for counting rare events) represents each species' occurrence as an independent Bernoulli trial with its own probability of success; the total number of successes determines the overall richness. For the two JSDMs, I calculated the confidence intervals for the appropriate mixtures of Poisson binomial distributions (as estimated from 1000 independent Monte Carlo samples).

Results and discussion

MISTNET'S VIEW OF NORTH AMERICAN BIRD ASSEMBLAGES

I began by decomposing the variance in mistnet's species-level predictions into variance among routes (which varied in their climate values) and residual (within-route) variance (Appendix S6). On average, the residuals accounted for 30% of the variance in mistnet's predictions, suggesting that non-climate factors play a substantial role in habitat filtering.

If the non-climate factors mistnet identified were biologically meaningful, then there should be a strong correspondence between the 15 coefficients assigned to each species by mistnet and the AAB habitat classifications. A linear discriminant analysis (LDA; Venables & Ripley 2002) demonstrated such a correspondence (Fig. 4). Mistnet's coefficients cleanly distinguished several groups of species by habitat association (e.g. 'Grassland' species vs. 'Forest' species), though the model largely failed to distinguish 'Marsh' species from 'Lake/Pond' species and 'Scrub' species from 'Open Woodland' species. These results indicate that the model has identified the broad differences among communities, but that it lacks some fine-scale resolution for distinguishing among types of wetlands and among types of partially wooded areas. Alternatively, perhaps these finer distinctions are not as salient at the scale of a 40-km transect or require more than two dimensions to represent.

While one might be able to produce a similar-looking scatterplot using ordination methods such as non-metric multidimensional scaling (NMDS; McCune, Grace & Urban 2002), the interpretation would be very different. Species' positions in ordination plots are chosen to preserve the multivariate geometry of the data and do not usually connect to any data-generating process or to a predictive model. In Fig. 4, by contrast, each species' x–y coordinates describe the predicted slopes of its responses to two axes of environmental variation; these slopes could be used to make specific predictions about occurrence probabilities at new sites. Likewise, deviations from these predictions could be used to falsify the underlying model, without the need for expensive permutation tests or comparison with a null model. The close connection between model and

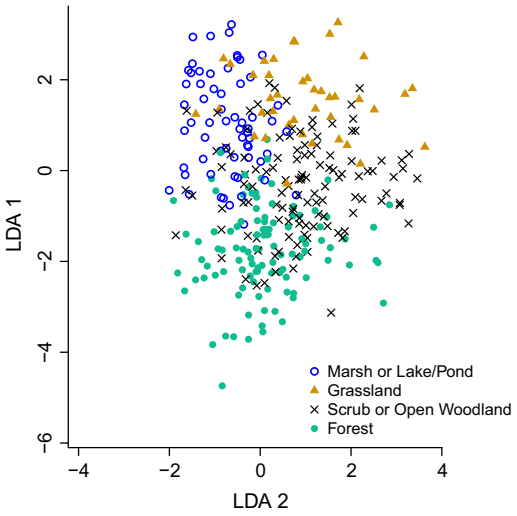

Fig. 4. Each species' mistnet coefficients, projected into a two-dimensional space by linear discriminant analysis (LDA) to maximize the spread between the six habitat types assigned to species by AAB. Mistnet cleanly separates 'Grassland' species from 'Forest' species, with 'Scrub' and 'Open Woodland' species representing intermediates along this axis of variation. 'Marsh' and 'Lake/Pond' species cluster together in the upper left. Habitat classes with fewer than 15 species were omitted from this analysis.

visualization demonstrated in Fig. 4 may prove especially useful in contexts where prediction and understanding are both important.

The environmental gradients identified in Fig. 4 are explored further in Fig. 5. Figure 5a shows how the forest/grassland gradient identified by mistnet affects the model's predictions for a pair of species with opposite responses to forest cover. The model cannot tell *which* of these two species will be observed (since it was only provided with climate data), but the model has learned enough about these two species to tell that the probability of observing *both* along the same 40-km transect is much lower than would be expected if the species were uncorrelated.

Figure 5a reflects a great deal of uncertainty, which is appropriate considering that the model has no information about a crucial environmental variable (forest cover). Often, however, additional information is available that could help resolve this uncertainty, and the mistnet package includes a built-in way to do so, as indicated in Fig. 5b,c. These panels show how the model is able to use a chance observation of a forest-associated Nashville Warbler (*Oreothlypis ruficapilla*) to indicate that a whole suite of other forest-dwelling species are likely to occur nearby and that a variety of species that prefer open fields and wetlands should be absent. Similarly, Fig. 5d shows how the presence of a Redhead duck (*Aythya americana*) can inform the model that a route likely contains suitable wetland habitat for waterfowl, marsh-breeding blackbirds, shorebirds and rails (along with the European Starling and Bobolink, whose true wetland associations are somewhat weaker). None of these inferences would be possible from a stack of disconnected single-species SDMs, nor would traditional ordination methods have been able to quantify the changes in occurrence probabilities.

MODEL COMPARISON: SPECIES RICHNESS

Environmental heterogeneity plays an especially important role in determining species richness, which is often overdispersed relative to models' expectations (O'Hara 2005). Figure 6 shows that mistnet's predictions respect the heterogeneity one might find in nature: areas with a given climate could plausibly be either very unsuitable for most waterfowl (Anatid richness <2 species) or much more suitable (Anatid richness >10 species). Under the independence assumption used for stacking SDMs, however, both of these scenarios would be ruled out (Fig. 6a).

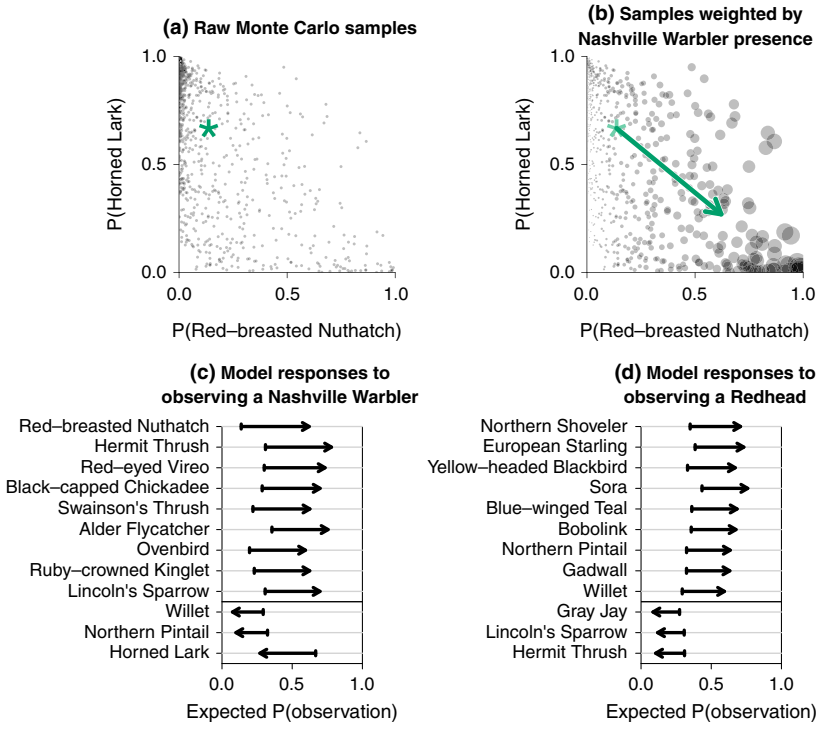

Fig. 5. (a) The mistnet model has learned that Red-breasted Nuthatches (*Sitta canadensis*) and Horned Larks (*Eremophila alpestris*) have opposite responses to an unobserved environmental factor. Based on these two species' biology, an ornithologist could infer that this unobserved variable is related to forest cover, with the nuthatch favouring more forested areas and the lark favouring more open areas. The green asterisk marks the marginal probabilities of observing the two species. (b) The presence of a forest-dwelling Nashville Warbler (*Oreothlypis ruficapilla*) provides the model with a strong indication that the area is forested, increasing the weight assigned to Monte Carlo samples that are suitable for the nuthatch and decreasing the weight assigned to samples that are suitable for the lark. The model's updated expectations can be found at the head of the green arrow. (c) The Warbler's presence similarly suggests increased occurrence probabilities for a variety of other forest species (top portion of panel) and decreased probabilities for species associated with open habitat (bottom portion). (d) If a Redhead (*Aythya americana*) had been observed instead, the model would correctly expect to see more water-associated birds and fewer forest dwellers.

Stacking leads to even larger errors when predicting richness for larger groups, such as the complete set of birds studied here. Models that stacked independent predictions consistently underestimated the range of biologically possible outcomes (Fig. 6b), frequently putting million-to-one or even billion-to-one odds against species richness values that were actually observed. These models' 95% confidence intervals were so narrow that half of the observed species richness values fell outside the predicted range. The overconfidence associated with stacked models could have serious consequences in both management and research contexts if we fail to prepare for species richness values outside such unreasonably narrow bounds (e.g. expecting a reserve to protect 40–50 species even though it only supports 15). Mistnet, on the other hand, was able to explore the range of possible non-climate environments to avoid these missteps: 90% of the test routes fell within mistnet's 95% confidence intervals, and the log-likelihood ratio decisively favoured it over stacked alternatives.

MODEL COMPARISON: SINGLE SPECIES

Figure 7a compares the models' ability to make predictions for a single species across all the test routes (shown as the exponentiated expected log-likelihood). While there was substantial variation among species, the two neural network models' predictions averaged more than an order of magnitude better than BRT's. Moreover, these models' advantage over BRT was largest for low-prevalence species (linear regression of log-likelihood ratio vs. log-prevalence; $P = 3*10^{-4}$), which will often be of the greatest concern to conservationists. The most likely reason for this improvement was a reduction in overfitting: while the overall model included complex nonlinear transformations, the number of degrees of freedom associated with any given species in the final logistic regression layer was modest (15 weights plus an intercept term).

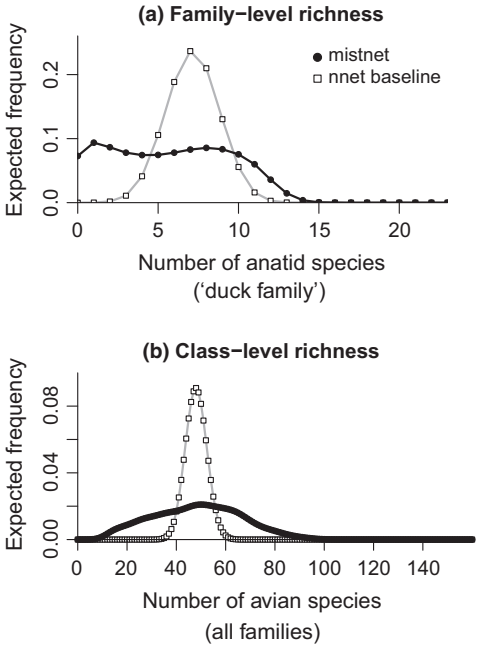

Fig. 6. The predicted distribution of species richness values one would expect to find based on predictions from mistnet and from the deterministic neural network baseline. (a) Anatid (waterfowl) species richness. (b) Total species richness. BRT's predictions (not shown) are similar to the baseline network, since neither one accounts for the effects of unmeasured environmental heterogeneity. In general, both networks' mean predictions are equally distant from the observed values, but only mistnet represents its uncertainty adequately.

BayesComm's predictions were substantially worse than any of the machine learning methods tested, which I attribute mostly to its inability to learn nonlinear responses to the environment (Elith *et al.* 2006). Adding quadratic terms or interaction terms (cf. Austin 1985; Jamil & ter Braak 2013) would have led to severe overfitting for many rare species. Even if one added a regularizer to the software to mitigate this problem, these extra pre-specified terms may still not provide enough flexibility to compete with modern nonlinear techniques.

Applying BayesComm to a large data set also highlighted one other area where mistnet appears to outperform existing JSDMs. Despite its assumed linearity, the BayesComm model required 70 000 parameters, most of which served to identify a distinct correlation coefficient between a single pair of species. Tracing this many parameters through hundreds of Markov chain iterations routinely caused BayesComm to run out of memory and crash, even after the code was modified to reduce its memory footprint. Sampling long Markov chains over a dense, full-rank covariance matrix (as has apparently been done in all other JSDMs to date) thus appears to be a costly strategy with large assemblages.

MODEL COMPARISON: COMMUNITY COMPOSITION

While making predictions about individual species is fairly straightforward with this data set (since most species have relatively narrow breeding ranges), community ecology is more concerned with co-occurrence and related patterns involving community composition (Chase 2003). Mistnet was able to use the correlation structure of the data to reduce the number of independent bits of information needed to make an accurate prediction. As a result, mistnet's route-level likelihood averaged 430 times higher than the baseline neural network's and 45 000 times higher than BRT's (Fig. 7b). BayesComm demonstrated a similar effect, but not strongly enough to overcome the low quality of its species-level predictions.

Conclusion

The large discrepancy between the performance of linear and nonlinear methods shown in Fig. 7a confirms previous results: accuracy in SDM applications requires the flexibility to learn about the functional form of species' environmental responses from the data (Elith *et al.* 2006). Likewise, mistnet's large improvement over stacked models (Figs 6 and 7b) provides strong evidence that accurate assemblage-level predictions require accounting for unmeasured environmental heterogeneity – especially when reasonable confidence intervals are

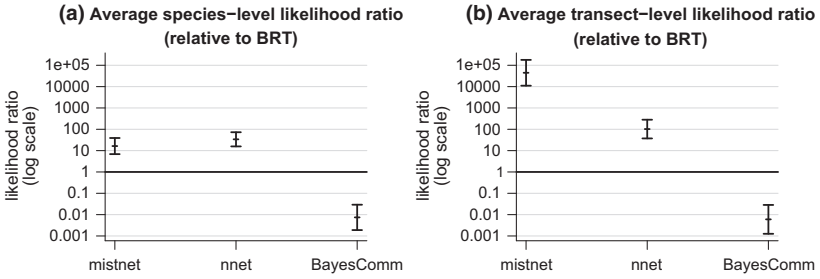

Fig. 7. Relative predictive performance of the evaluated methods, as compared to BRT (mean ±95% CI, calculated from paired *t*-tests on the log-likelihood scale). (a) Expected likelihood ratio for predictions about one species across 280 test-set routes. (b) Expected likelihood ratio when predicting species composition of a test route.

required. Currently, mistnet appears to be the only software package that meets both of these criteria, providing both nonlinear responses to the environment and a method for dealing with assemblage-level responses to unobserved environmental heterogeneity.

Mistnet can also identify some of the same similarities among species that a skilled biologist would expect to find. For taxa on the frontier of our knowledge, a model like mistnet could help guide the biologists to ask the best questions and organize their understanding by suggesting which species have similar habitat requirements – even when the factors controlling their occurrence are still unknown (cf. indirect gradient analysis).

Unlike with stacked methods, one can read this information directly from mistnet's coefficient tables with no more difficulty than interpreting a principal components analysis. Also, where most ordination techniques merely describe the multivariate geometry of an existing data matrix, mistnet's coefficients are directly tied to quantitative – and falsifiable – predictions about community structure in unobserved locations. Nonlinear JSDMs should thus be able to take on a variety of roles in ecologists' toolboxes, providing a unified framework for summarizing community structure, developing forecasts and evaluating hypotheses about community structure.

Future research should look for ways to use other forms of ecological knowledge about species to impose some structure on models, coefficients and nudge the models towards more biologically reasonable predictions (Kearney & Porter 2009; Lankau *et al*. 2011; Kissling *et al*. 2012). Such a research programme could also be useful in other areas of predictive ecology (Pearse *et al*. 2013). JSDMs' ability to use asymmetrical or low-quality data sources to improve their predictions should also increase the value of low-effort data collection procedures such as short transects – especially since these data sources can be incorporated without the need for fitting a new model.

Finally, while it would be tempting to attribute JSDMs' correlation structure to species interactions, this approach may not be as fruitful as some authors have hoped. The correlations are all driven indirectly via shared dependencies on latent variables, rather than the direct response of one species to another implied by species interactions. Pollock *et al*. (2014)'s covariance decomposition allows for some progress towards inferring interactions from JSDMs, but it would be much more straightforward to use a different approach [e.g. Markov random fields (Azaele *et al*. 2010) or ensembles of classifier chains (Yu *et al*. 2011)] whose coefficients describe direct pairwise interactions much more explicitly. Latent variable models are more appropriate for studies like this one at large spatial scales where direct species interactions will tend to be weaker and most of the variation is driven by environmental filtering and range limits.

Mistnet's accuracy, interpretability and flexibility to work with opportunistic samples indicate that nonlinear JSDMs will be important in a variety of basic and applied contexts, from forecasting, to quantifying differences among species, to developing new insights about community structure. Ecologists' models for these tasks need not be neural nets, but these analyses suggest that the most comprehensive and useful models will have many of the same features, such as latent random variables, nonlinearity and low rank.

Acknowledgements

This work benefitted greatly from discussions with A. Sih and his laboratory meeting group, M. L. Baskett, R. J. Hijmans, R. McElreath, J. H. Thorne, M. W. Schwartz, B. M. Bolker, R. E. Snyder, A. C. Perry and C. S. Tysor, as well as comments from S. C. Walker and an anonymous reviewer. It was funded by a Graduate Research Fellowship from the National Science Foundation, the UC Davis Center for Population Biology, and the California Department of Water Resources. I gratefully acknowledge the field biologists that collected the BBS data, as well as the US Geological Survey, Cornell Lab of Ornithology, and Worldclim for making their data publicly available.

Data accessibility

- All data sets used here are freely downloadable from their original sources.
- The mistnet source code is at https://github.com/davharris/mistnet and can be installed with the `install_github` function from the `devtools` package.

References

Austin, M.P. (1985) Continuum concept, ordination methods, and niche theory. *Annual Review of Ecology and Systematics*, **16**, 39–61.

Austin, M.P. & Van Niel, K.P. (2011) Improving species distribution models for climate change studies: variable selection and scale. *Journal of Biogeography*, **38**, 1–8.

Azaele, S., Muneepeerakul, R., Rinaldo, A. & Rodriguez-Iturbe, I. (2010) Inferring plant ecosystem organization from species occurrences. *Journal of Theoretical Biology*, **262**, 323–329.

Bahn, V. & McGill, B.J. (2007) Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, **16**, 733–742.

Bengio, Y. (2013) Deep learning of representations: looking forward. *Statistical Language and Speech Processing* (eds A.-H. Dediu, C. Martín-Vide, R. Mitkov & B. Truthe), pp. 1–37. Springer, Berlin, Heidelberg.

Calabrese, J.M., Certain, G., Kraan, C. & Dormann, C.F. (2014) Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, **23**, 99–112.

Chase, J.M. (2003) Community assembly: when should history matter? *Oecologia*, **136**, 489–498.
Clark, J.S., Gelfand, A.E., Woodall, C.W. & Zhu, K. (2013) More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, **24**, 990–999.
Dormann, C.F., Purschke, O., Márquez, J.R.G., Lautenbach, S. & Schröder, B. (2008) Components of uncertainty in species distribution analysis: a case study of the great grey shrike. *Ecology*, **89**, 3371–3386.
Elith, J., Graham*, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.
Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.
Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2007) Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, **13**, 252–264.
Golding, N. (2013) *Mapping and understanding the distributions of potential vector mosquitoes in the UK: New methods and applications*. PhD thesis
Golding, N. & Harris, D.J. (2014) *BayesComm: Bayesian Community Ecology Analysis*.
Guisan, A. & Rahbek, C. (2011) SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, **38**, 1433–1444.
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.
Hong, Y. (2013) *Poibin: The Poisson Binomial Distribution*.
Hutchinson, R.A., Liu, L.-P. & Dietterich, T.G. (2011) Incorporating boosted regression trees into ecological latent variable models. *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 1343–1348. Association for the Advancement of Artificial Intelligence. Available at: https://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3711.
Jackson, M.M., Turner, M.G., Pearson, S.M. & Ives, A.R. (2012) Seeing the forest and the trees: multilevel models reveal both species and community patterns. *Ecosphere*, **3**, 79.
Jamil, T. & ter Braak, C.J. (2013) Generalized linear mixed models can detect unimodal species-environment relationships. *PeerJ*, **1**, 95.
Kearney, M. & Porter, W. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, **12**, 334–350.
Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J. *et al.* (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–2178.
Lankau, R., Jørgensen, P.S., Harris, D.J. & Sih, A. (2011) Incorporating evolutionary principles into environmental management and policy. *Evolutionary Applications*, **4**, 315–325.
Latimer, A.M., Banerjee, S., Sang Jr H., Mosher, E.S. & Silander Jr J.A. (2009) Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. *Ecology Letters*, **12**, 144–154.
Leathwick, J.R., Rowe, D., Richardson, J., Elith, J. & Hastie, T. (2005) Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology*, **50**, 2034–2052.
McCune, B., Grace, J.B. & Urban, D.L. (2002) *Analysis of Ecological Communities*. MjM Software Design, Gleneden Beach, Oregon, USA.
Murphy, K.P. (2012) *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, Massachusetts, USA.
Neal, R.M. (1992) Connectionist learning of belief networks. *Artificial Intelligence*, **56**, 71–113.
Neal, R.M. & Hinton, G.E. (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models* (ed. M.I. Jordan), pp. 355–368. Kluwer, Boston, Massachusetts, USA.
O'Hara, R.B. (2005) Species richness estimators: how many species can dance on the head of a pin? *Journal of Animal Ecology*, **74,** 375–386.
Ovaskainen, O., Hottola, J. & Siitonen, J. (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, **91**, 2514–2521.
Pearse, I.S., Harris, D.J., Karban, R. & Sih, A. (2013) Predicting novel herbivore–plant interactions. *Oikos*, **122**, 1554–1564.
Pellissier, L., Espíndola, A., Pradervand, J.-N., Dubuis, A., Pottier, J., Ferrier, S. & Guisan, A. (2013) A probabilistic approach to niche-based community models for spatial forecasts of assemblage properties and their uncertainties. *Journal of Biogeography*, **40**, 1939–1946.
Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., Vesk, P.A. & McCarthy, M.A. (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, **5**, 397-406.
Ridgeway, G. (2013) *Gbm: Generalized Boosted Regression Models*.
Sauer, J.R., Hines, J.E., Fallon, J., Pardieck, K.L., Ziolkowski Jr D.J. & Link, W.A. (2011) The North American breeding bird survey, results and analysis 1966–2011. *Version 2011.0*.
Tang, Y. & Salakhutdinov, R. (2013) Learning stochastic feedforward neural networks. *Advances in Neural Information Processing Systems 26* (eds & trans C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K.Q. Weinberger), pp. 530–538. Neural Information Processing Systems Foundation, San Diego, California, USA.
Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics With S*. Springer, New York, New York, USA.
Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P.-A. (2010) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, **9999**, 3371–3408.
Walker, S.C. & Jackson, D.A. (2011) Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. *Ecological Monographs*, **81**, 635–663.
Wang, Y., Naumann, U., Wright, S.T. & Warton, D.I. (2012) mvabund–an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, **3**, 471–474.
Wei, G.C. & Tanner, M.A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, **85**, 699–704.
Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F. *et al.* (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews*, **88**, 15–30.
Yu, J., Wong, W.-K., Dietterich, T., Jones, J., Betts, M., Frey, S., Shirley, S., Miller, J. & White, M. (2011) Multi-label classification for multi-species distribution modeling. *Proceedings of the 28th International Conference on Machine Learning*.

Received 6 April 2014; accepted 9 December 2014
Handling Editor: David Warton

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Data.

Appendix S2. Neural network structure.

Appendix S3. Model fitting.

Appendix S4. Model configuration.

Appendix S5. Sampling error in marginal likelihood estimates.

Appendix S6. Variance decomposition.