`embarcadero`: Species distribution modelling with Bayesian additive regression trees in R

Colin J. Carlson[1,†]

[1] *Department of Biology, Georgetown University, Washington, D.C. 20057, USA.*
[†] *Correspondence should be directed to cjc322@georgetown.edu.*

Submitted to *Methods in Ecology and Evolution* on September 18, 2019

Abstract

1. Classification and regression tree methods, like random forests (RF) or boosted regression trees (BRT), are one of the most popular methods of mapping species distributions.

2. Bayesian additive regression trees (BARTs) are a relatively new alternative to other popular regression tree approaches. Whereas BRT iteratively fits an ensemble of trees each explaining smaller fractions of the total variance, BART starts by fitting a sum-of-trees model and then uses Bayesian backfitting with an MCMC algorithm to create a posterior draw. So far, BARTs have yet to be applied to species distribution modeling.

3. `embarcadero` is an `R` package of convenience tools for researchers interested in species distribution modeling with BARTs. It includes functionality for spatial prediction, an automated variable selection and importance procedure, and other functionality for rapid implementation and data visualization.

4. To show how `embarcadero` can be used by ecologists, we re-map the distribution of Crimean-Congo haemorrhagic fever and a likely vector, *Hyalomma truncatum*, in Africa.

Keywords: Bayesian additive regression trees, species distribution modeling, ecological niche modeling, Crimean Congo haemorrhagic fever

1 Introduction

In the last two decades, over a dozen statistical and machine learning methods have been proposed for species distribution modeling (SDM). Over time, a handful of methods have risen to predominance due to ease of implementation, computational speed, and strong predictive performance in rigorous cross-validation. Some methods are especially popular for specific applications, mostly because of disciplinary tradition. For example, maximum entropy (MaxEnt) models are widely popular for studies of global ecological responses to climate change. (VanDerWal *et al.*, 2013; Warren *et al.*, 2013) In disease ecology, boosted regression trees (BRTs) have become the dominant tool for mapping vectors, reservoirs, and transmission risk of infectious zoonoses and vector-borne diseases (Carlson *et al.*, 2019; Pigott *et al.*, 2014; Messina *et al.*, 2016), largely due to an influential 2013 paper on dengue virus. (Bhatt *et al.*, 2013)

Boosted regression trees are easily implemented as an out-of-the-box machine learning method, are powerful for ecological inference, and consistently perform well in rigorous tests of SDM performance. The classification tree approach is also more intuitive than the complex fitting procedures "under the hood" of MaxEnt or Maxlike methods. However, BRTs also have downsides: they can be prone to overfitting, and fitting procedures are largely handed down as anecdotal best practices, with most studies choosing learning rates and tree depth based on the default settings recommended by early work (Elith *et al.*, 2008), with very few studies selecting parameters from formal cross-validation with packages like `caret`. Furthermore, uncertainty is usually measured by generating an unweighted ensemble of BRT submodels over subsetted training data, generating a confidence interval from data permutations rather than formal assumptions about model uncertainty.

In this paper, we propose the use of a Bayesian alternative to boosted regression trees, called *Bayesian additive regression trees* (BARTs), for the problem of species distribution modeling. In computer science, BARTs are used for everything from medical diagnostics to self-driving car algorithms (Sparapani *et al.*, 2018; Tan *et al.*, 2018); however, they have yet to find any widespread application in ecology. A study from 2011 used BARTs as a tool to examine habitat selection data on birds (Yen *et al.*, 2011); a 2017 study used BARTs to evaluate performance data of other species distribution modeling methods. (Farley, 2017) But so far, they have not been used for the purpose of predicting species distributions. We introduce an `R` package, `embarcadero`, as a convenience tool for running SDMs with BARTs.

2 Bayesian additive regression trees

Whereas BRTs fit an ensemble of trees each explaining smaller fractions of unexplained variance, BART starts by fitting an initial sum-of-trees model, and then uses Bayesian backfitting with an MCMC algorithm to create a posterior draw. The Bayesian component of BART is designed to have minimal user input, making it an unusually out-of-the-box algorithm as Bayesian SDM methods go. The priors set for covariates and for the distribution of splitting rules are both uniform by default. (Chipman *et al.*, 2010) The remaining priors, which govern tree depth and the error distribution, are given a set of default hyperparameters that are robust enough for the developers of the method to have recommended automated tuning. The number of trees is the main free parameter; a default of 200 is recommended, though cross-validation can be used to tune this very easily.

3 SDMs with BARTs

3.1 Using `dbarts` and `embarcadero`

At least four `R` packages currently exist that can implement BARTs: `BayesTree` (Chipman & McCulloch, 2016), `bartMachine` (Kapelner & Bleich, 2013), `BART` (McCulloch *et al.*, 2018), and `dbarts` (Chipman *et al.*, 2014). Their functionality differs in important ways, and not all of them are currently capable of important features like partial dependence plots that are important for SDMs. Our package is an SDM-oriented workflow wrapper for `dbarts`, which includes most of the basic functionality needed for species distribution modeling: BART models with binary response variables, multithread model training, easy generation of partial dependence surfaces (including two-predictor plots and spatial projections), and simple prediction including full posteriors. Compared to other packages, `dbarts` only has one major limitation. In future versions, we hope to make `embarcadero` flexible for work with any underlying engine.

3.2 Variable selection

Variable importance can be measured in BART models by counting the number of times a given variable is used by a tree split across the full posterior draw of trees. (This is similar to variable importance in BRTs, which is calculated from

the number of tree splits and the corresponding improvement they cause in the model.) Because the trees are weak learners, in models with higher numbers of trees, the difference in variable importance becomes less pronounced, and less informative variables become more commonly represented. Consequently, variable selection can be performed by observing variable importance in models with progressively fewer trees, and selecting variables that have improved performance in progressively smaller ensembles. (Chipman *et al.*, 2010) In our package, variable importance can be calculated with `varimp`, and variable selection diagnostic plots can be generated with `varimp.plot`.

As a way to standardize inclusion rules across workflows, we implemented an automatic variable elimination procedure in `variable.step`, which (1) fits a full model with all predictors and a small tree ensemble (default m = 10), a fixed number of times (n = 50); (2) eliminates the least informative variable across all 50 runs; (3) re-runs the models again (n = 50), recording the root mean square error; (4) repeats steps 2 and 3 until there are only three covariates left; and (5) finally selects the model with the lowest average RMSE is selected. The variable sets recommended by this procedure are almost always nearly identical to the sets generated from a subjective reading of the diagnostic plots.

We recommend careful analysis of all diagnostic information, but include a full automated variable selection pipeline in `bart.var`, which (a) produces the initial multi-m diagnostic plot, (b) runs automated variable selection, (c) returns a model trained with the optimal variable set, (d) plots variable importance in the final model, and (e) calculates the AUC of the final model. Despite automation, this procedure is not a fail-safe against the inclusion of uninformative predictors, or false inference on them; this is true of almost all methods, and predictors should always be chosen based on at least some expert opinion about biological plausibility. (Fourcade *et al.*, 2018) Similarly, validation of partial depencence curves against biological knowledge should be treated as an additional level of model validation, potentially more informative than measuring predictive accuracy. (Warren *et al.*, 2019)

3.3 Visualizing model results

BART model predictions can be visualized several ways using `embarcadero`. Unfortunately `dbarts::bart.predict` cannot handle spatial data in its native format. For spatial prediction, we provide `bart.map` as a wrapper, which can also

pull the 5% and 95% credible interval layers from the posterior distribution. These raster layers can be exported to other spatial data packages in `R`, or external software like ArcGIS or QGIS, for more professional visualization.

We include several methods for generating partial dependence plots. The function `partial` is written as a wrapper for `dbarts::pdbart`, and can be used to generate partial dependence plots with a customizable aesthetic, including multiple ways of visualizing uncertainty. (As with overall predictions, credible intervals on partial plots are true Bayesian credible intervals.) Two-dimensional partial dependence plots (interactions among two predictor variables) can also be generated using `dbarts::pd2bart`. Finally, we designed a new visualization called *spatial partial dependence plots*, which reclassify predictor rasters based on their partial dependence plots, and show the relative suitability of different regions for an individual covariate. The `spartial` function can be used to generate these maps, and answer questions like "What desert regions are too arid, even in their wettest month, for spadefoot toads?" or "Where are the soils with the best pH for redwood growth?" We illustrate some of these visualization options in Figure 1.

3.4 Notes on model performance

The strength of a given SDM method is very rarely resolved in a single study, and is challenging to understand in the context of "real world" datasets. Simulation studies have become common practice as a way to unpack model performance relative to confounding factors, such as sample size and bias, pseudoabsence design, and colinearity among predictors; each of these is usually worth testing in isolation. Moreover, the definition of "performance" is subjective, and SDMs are used for several (sometimes conflicting) purposes in ecology (Guillera-Arroita *et al.*, 2015); for example, methods that usually perform well in overall accuracy may not necessarily handle variable importance well. (Smith & Santos, 2019) Ultimately, the most comprehensive "bake-off" studies, with dozens of methods tested by dozens of authors, have usually found that more recent and popular methods like MaxEnt or BRTs are more sensitive to poor calibration than they are discrepant with each other. (Norberg *et al.*, 2019)

In this descriptor, we did not compare the performance of BARTs to other methods using AUC or comparable accuracy metrics, because our aim was not to produce a method that universally outperforms `gbm` or more distantly related methods. Instead, we aimed to produce an accessible environment for explor-

ing BARTs, and to propose their use for SDMs as an approach with different conceptual strengths than existing methods. BART is a nice compromise between Bayesian prediction—which works well in SDM, but has yet to become widely popular—and the conceptual familiarity and strengths of regression tree approaches like BRTs. Some other Bayesian methods for SDMs have been proposed, but are not widely adopted (Golding & Purse, 2016; Redding *et al.*, 2017), possibly due to the dominance of more familiar model families. Most importantly, BARTs handle uncertainty with formal Bayesian logic, and generate a posterior from a single model implementation. We argue this is more coherent than how uncertainty is usually generated for BRTs, with several hundred BRTs trained as a usually-unweighted ensemble on randomly-subsetted training data. This inherently reduces the training sample size and amplifies within-sample biases (especially geographic biases) fed into every submodel, and produces a "confidence interval" that lacks a formal statistical definition as such. We believe the structure of BART makes it easier to include uncertainty in applied tasks using SDMs, like estimating the population at risk from a given pathogen.

4 Example: Crimean-Congo Haemorrhagic Fever

Crimean-Congo haemorrhagic fever virus (CCHFV) is a tick-borne Bunyavirus that causes extremely severe, and often fatal, illness in humans. Very little is known about CCHFV, compared to other cosmopolitan tick-borne illnesses like Lyme disease or tularemia. The definitive reservoir of CCHFV is unknown but likely ungulates (Babayan *et al.*, 2018); outbreaks frequently affect sheep and other domestic ruminants, and other members of the Nairoviridae infect similar hosts. The vectors of CCHFV are better known, and are presumed to almost always be *Hyalomma* ticks, which are widespread throughout Africa and Eurasia; other tick vectors have been suspected, but evidence for their competence is limited. (Papa *et al.*, 2017) In Africa, *Hyalomma truncatum* in particular is common throughout rangeland and is one of the strongest candidates for a primary vector. (Logan *et al.*, 1989; Wilson *et al.*, 1991)

A global map of Crimean-Congo haemorrhagic fever has been previously been produced with boosted regression trees; a significant amount of the Black Sea region was suitable, while areas outside had highly localized predictions of suitability, presumably because of data sparsity in Africa especially. (Messina *et al.*,

2015b) However, some major areas of presence appeared under-predicted, such as the western Congo Basin. To demonstrate the use of BARTs, we re-mapped transmission risk in Africa using the same original CCHF occurrence dataset (Messina *et al.*, 2015a). Just as studies of dengue have included suitability for *Aedes aegypti* as a covariate, our model included a suitability layer for *Hyalomma truncatum*, which we created using the canonical dataset on African tick distributions. (Cumming, 1998); all code for these models and their visualization is available as a detailed 30-page vignette with the package.

Both the *Hyalomma* model and the CCHF model were also run with 11 environmental covariates: eight average and monthly variables from WorldClim (BIO 1, 2, 5, 6, and 12–15; Hijmans *et al.* 2005), layers for mean and amplitude of NDVI previously used to map anthrax (Carlson *et al.*, 2019), and a layer of percentage cropland. (Ramankutty *et al.*, 2010) For both models we ran a BART with default `dbarts` settings using the full variable set. We then ran automated variable set reduction, made spatial predictions using the recommended variables, and recorded the accuracy (*Hyalomma*: AUC = 0.911; CCHF: AUC = 0.898). Our final model of Crimean-Congo risk included six variables: *H. truncatum* suitability, mean and amplitude of NDVI, mean annual precipitation (BIO12), and precipitation of wettest month (BIO13). The variable importance diagnostic for the CCHF model is shown in Figure 2.

Our model predicts that the distribution of CCHF may be more geographically expansive than previous studies have indicated (Figure 3). Areas of the highest risk are still heavily concentrated in Sahel rangeland and east African highlands, but also far more extensive in southern Africa and along the Atlantic coast than previously believed. Although *H. truncatum* had a high importance in the final model, and scaled positively with CCHF risk, our model still predicted some areas outside of its range. In particular, our final CCHF map captured an area in Gabon and the western Congo basin where occurrences have been recorded but *H. truncatum* is likely absent, and another vector may be involved. This may raise some interesting questions for future research.

5 Next steps

The `embarcadero` package is designed as a beginning framework for using BART as a species distribution modeling method. Currently, `embarcadero` version 1.0.1

is available open source on Github, at github.com/cjcarlson/embarcadero and soon it will be available on CRAN. We hope to continue to expand the package based on user input and collaborative coding. The top priority for development is faster (multithread) prediction, which remains computationally limiting for global mapping projects.

Acknowledgements

Thanks to Shweta Bansal and Jason Blackburn for research support, to Ethan Beaman for telling me about BARTs in the first place, to Zack Susswein for helpful comments on design of Bayesian models, and to Tad Dallas for helpful comments. This work was supported with a Georgetown Environment Initiative (GEI) postdoctoral fellowship.

242

Figures

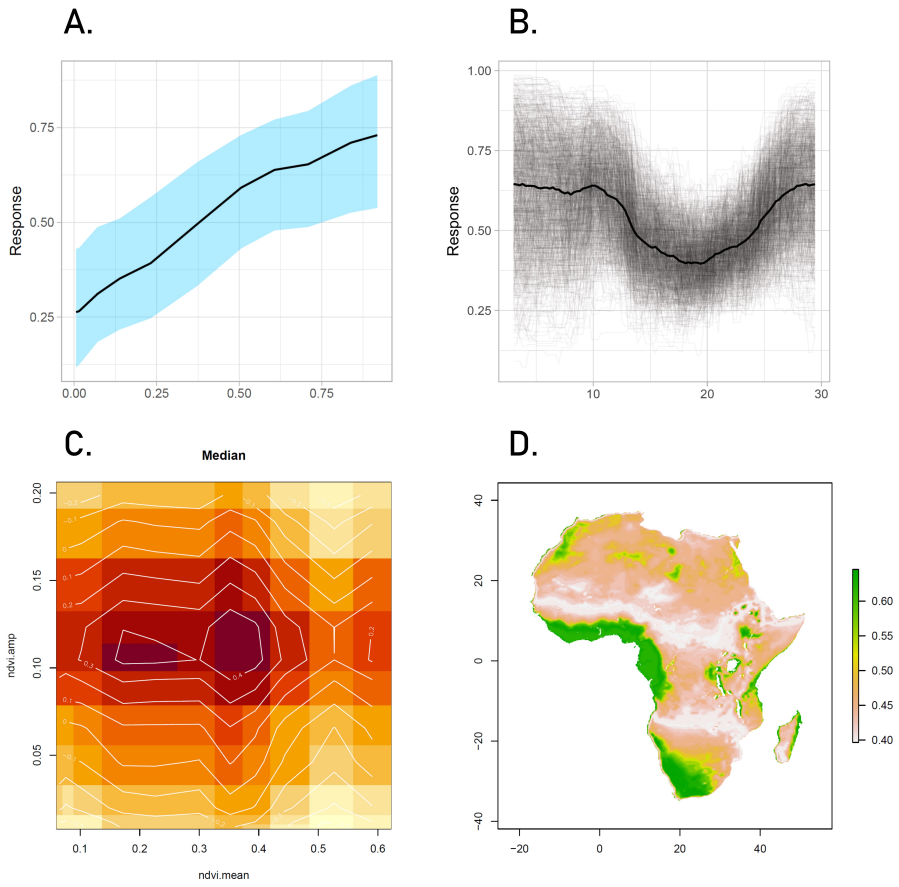

Figure 1: Partial dependence plots from the CCHF model (see Section 4), as examples of four different model visualization styles. (A) Partial dependence plot for *Hyalomma truncatum* suitability, blue bars show 95% posterior CI; (B) partial plot for BIO2 (mean diurnal range), where individual traces show every posterior draw; (c) a two-predictor partial plot for mean and amplitude of NDVI; and (d) a "spartial" plot for BIO12 (annual precipitation).

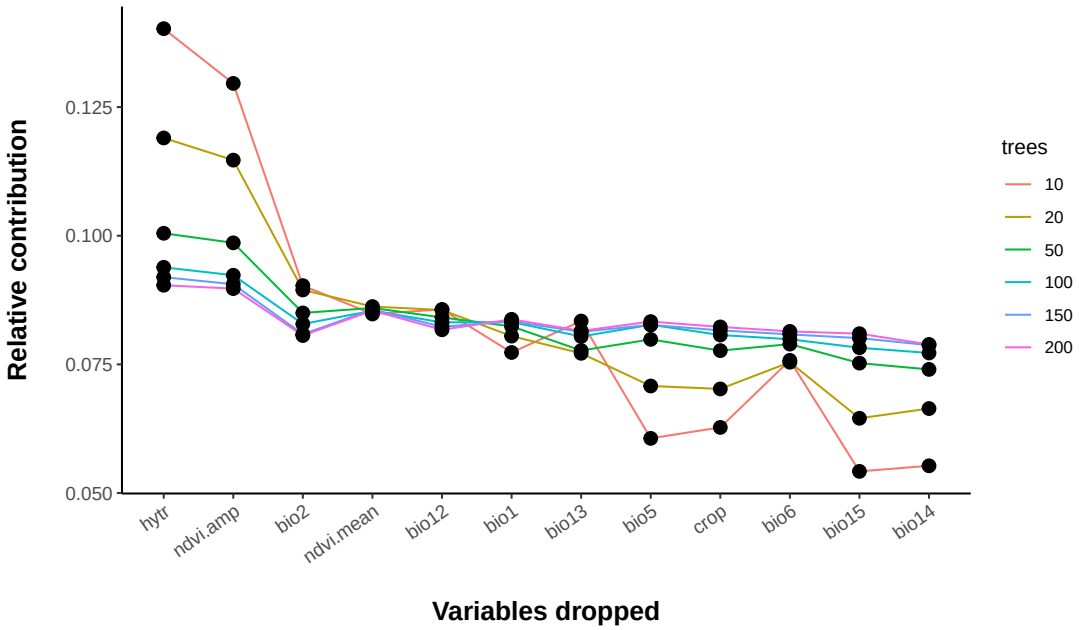

Figure 2: The variable selection diagnostic plot for the CCHF model, in the style of Chipman *et al.* (2010). Variables that are included more often in decision splits (the relative contribution) as the number of trees becomes smaller are more likely to be influential, real predictors. Variables that have increasing contributions as the number of trees increases, on the other hand, should be dropped.

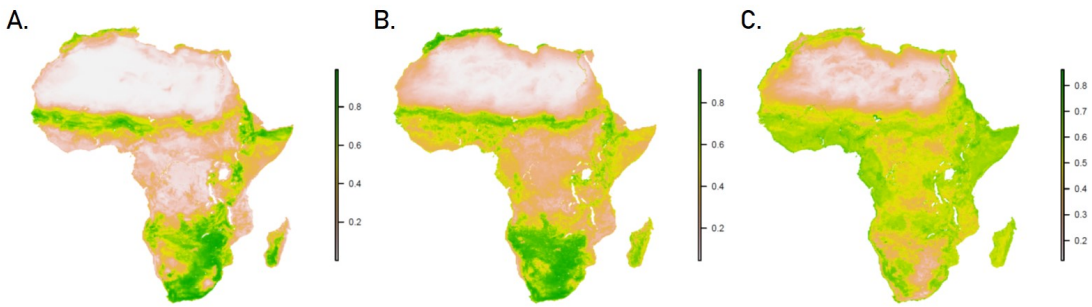

Figure 3: (A) Final posterior mean map of suitability for *Hyalomma truncatum*. (B) Final posterior mean map of suitability for Crimean-Congo haemorrhagic fever virus. (C) The 95% posterior credible interval width for the CCHF model.

References

Babayan, S.A., Orton, R.J. & Streicker, D.G. (2018) Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science*, **362**, 577–580.

Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., Drake, J.M., Brownstein, J.S., Hoen, A.G., Sankoh, O. *et al.* (2013) The global distribution and burden of dengue. *Nature*, **496**, 504.

Carlson, C.J., Kracalik, I.T., Ross, N., Alexander, K.A., Hugh-Jones, M.E., Fegan, M., Elkin, B.T., Epp, T., Shury, T.K., Zhang, W. *et al.* (2019) The global distribution of *Bacillus anthracis* and associated anthrax risk to humans, livestock and wildlife. *Nature Microbiology*, p. 1.

Chipman, H., McCulloch, R. & Dorie, V. (2014) dbarts: Discrete Bayesian Additive Regression Trees Sampler. R package version 0.8-5.

Chipman, H. & McCulloch, R. (2016) BayesTree: Bayesian Additive Regression Trees. R package version 0.3-1.3.

Chipman, H.A., George, E.I., McCulloch, R.E. *et al.* (2010) BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, **4**, 266–298.

Cumming, G. (1998) Host preference in African ticks (Acari: Ixodida): a quantitative data set. *Bulletin of Entomological Research*, **88**, 379–406.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.

Farley, S.S. (2017) *A General Framework for Predicting the Optimal Computing Configurations for Climate-driven Ecological Forecasting Models*. Ph.D. thesis.

Fourcade, Y., Besnard, A.G. & Secondi, J. (2018) Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, **27**, 245–256.

Golding, N. & Purse, B.V. (2016) Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, **7**, 598–608.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R. & Wintle, B.A. (2015) Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, **24**, 276–292.

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: A Journal of the Royal Meteorological Society*, **25**, 1965–1978.

Kapelner, A. & Bleich, J. (2013) bartMachine: Machine learning with Bayesian additive regression trees. *arXiv preprint arXiv:13122171*.

Logan, T.M., Linthicum, K.J., Bailey, C.L., Watts, D.M. & Moulton, J.R. (1989) Experimental transmission of Crimean-Congo hemorrhagic fever virus by *Hyalomma truncatum* Koch. *The American Journal of Tropical Medicine and Hygiene*, **40**, 207–212.

McCulloch, R., Sparapani, R., Gramacy, R., Spanbauer, C. & Pratola, M. (2018) BART: Bayesian additive regression trees. R package version 1.0.

Messina, J.P., Kraemer, M.U., Brady, O.J., Pigott, D.M., Shearer, F.M., Weiss, D.J., Golding, N., Ruktanonchai, C.W., Gething, P.W., Cohn, E. *et al.* (2016) Mapping global environmental suitability for Zika virus. *eLife*, **5**, e15272.

Messina, J.P., Pigott, D.M., Duda, K.A., Brownstein, J.S., Myers, M.F., George, D.B. & Hay, S.I. (2015a) A global compendium of human Crimean-Congo haemorrhagic fever virus occurrence. *Scientific Data*, **2**, 150016.

Messina, J.P., Pigott, D.M., Golding, N., Duda, K.A., Brownstein, J.S., Weiss, D.J., Gibson, H., Robinson, T.P., Gilbert, M., William Wint, G. *et al.* (2015b) The global distribution of Crimean-Congo hemorrhagic fever. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, **109**, 503–513.

Norberg, A., Abrego, N., Blanchet, F.G., Adler, F.R., Anderson, B.J., Anttila, J., Araújo, M.B., Dallas, T., Dunson, D., Elith, J. *et al.* (2019) A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, p. e01370.

Papa, A., Tsergouli, K., Tsioka, K. & Mirazimi, A. (2017) Crimean-congo hemorrhagic fever: tick-host-virus interactions. *Frontiers in Cellular and Infection Microbiology*, **7**, 213.

Pigott, D.M., Golding, N., Mylne, A., Huang, Z., Henry, A.J., Weiss, D.J., Brady, O.J., Kraemer, M.U., Smith, D.L., Moyes, C.L. *et al.* (2014) Mapping the zoonotic niche of Ebola virus disease in Africa. *eLife*, **3**, e04395.

Ramankutty, N., Evan, A., Monfreda, C. & Foley, J. (2010) Global agricultural lands: Croplands, 2000. *Data distributed by the Socioeconomic Data and Applications Center (SEDAC) [online]: http://sedac ciesin columbiaedu/data/set/aglands-croplands-2000 (Accessed on 1 January 2019)*.

Redding, D.W., Lucas, T.C., Blackburn, T.M. & Jones, K.E. (2017) Evaluating Bayesian spatial methods for modelling species distributions with clumped and restricted occurrence data. *PloS One*, **12**, e0187602.

Smith, A.B. & Santos, M.J. (2019) Testing the ability of species distribution models to infer variable importance. *bioRxiv*, p. 715904.

Sparapani, R., Dabbouseh, N., Gutterman, D., Zhang, J., Chen, H., Bluemke, D., Lima, J., Burke, G. & Soliman, E. (2018) Novel electrocardiographic criteria for the diagnosis of left ventricular hypertrophy derived with Bayesian additive regression trees: the multi-ethnic study of atherosclerosis. *Circulation*, **138**, A10908–A10908.

Tan, Y.V., Flannagan, C.A. & Elliott, M.R. (2018) Predicting human-driving behavior to help driverless vehicles drive: random intercept Bayesian additive regression trees. *Statistics and Its Interface*, **11**, 557–572.

VanDerWal, J., Murphy, H.T., Kutt, A.S., Perkins, G.C., Bateman, B.L., Perry, J.J. & Reside, A.E. (2013) Focus on poleward shifts in species' distribution underestimates the fingerprint of climate change. *Nature Climate Change*, **3**, 239.

Warren, D.L., Matzke, N.J. & Iglesias, T.L. (2019) Evaluating species distribution models with discrimination accuracy is uninformative for many applications. *BioRxiv*, p. 684399.

Warren, R., VanDerWal, J., Price, J., Welbergen, J.A., Atkinson, I., Ramirez-Villegas, J., Osborn, T.J., Jarvis, A., Shoo, L.P., Williams, S.E. *et al.* (2013) Quantifying the benefit of early climate change mitigation in avoiding biodiversity loss. *Nature Climate Change*, **3**, 678.

Wilson, M., Gonzalez, J.P., Cornet, J.P. & Camicas, J.L. (1991) Transmission of Crimean-Congo haemorrhagic fever virus from experimentally infected sheep to *Hyalomma truncatum* ticks. *Research in Virology*, **142**, 395–404.

Yen, J.D., Thomson, J.R., Vesk, P.A. & Mac Nally, R. (2011) To what are woodland birds responding? Inference on relative importance of in-site habitat variables using several ensemble habitat modelling techniques. *Ecography*, **34**, 946–954.