





# A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels

ANNA NORBERG <sup>1,34</sup> NEREA ABREGO,<sup>2,3</sup> F. GUILLAUME BLANCHET,<sup>4</sup> FREDERICK R. ADLER,<sup>5,6</sup> BARBARA J. ANDERSON,<sup>7</sup> JANI ANTILA,<sup>1</sup> MIGUEL B. ARAÚJO,<sup>8,9,10</sup> TAD DALLAS,<sup>1</sup> DAVID DUNSON,<sup>11</sup> JANE ELITH,<sup>12</sup> SCOTT D. FOSTER,<sup>13</sup> RICHARD FOX,<sup>14</sup> JANET FRANKLIN,<sup>15</sup> WILLIAM GODSOE,<sup>16</sup> ANTOINE GUISAN,<sup>17,18</sup> BOB O'HARA,<sup>19</sup> NICOLE A. HILL,<sup>20</sup> ROBERT D. HOLT,<sup>21</sup> FRANCIS K. C. HUI,<sup>22</sup> MAGNE HUSBY,<sup>23,24</sup> JOHN ATLE KÅLÅS,<sup>25</sup> ALEKSI LEHIKONEN,<sup>26</sup> MISKA LUOTO,<sup>27</sup> HEIDI K. MOD,<sup>18</sup> GRAEME NEWELL,<sup>28</sup> IAN RENNER,<sup>29</sup> TOMAS ROSLIN <sup>3,30</sup> JANNE SOININEN <sup>27</sup> WILFRIED THUILLER,<sup>31</sup> JARNO VANHATALO,<sup>1</sup> DAVID WARTON,<sup>32</sup> MATT WHITE,<sup>28</sup> NIKLAUS E. ZIMMERMANN,<sup>33</sup> DOMINIQUE GRAVEL,<sup>4</sup> AND OTSO OVASKAINEN <sup>1,2</sup>

<sup>1</sup>Organismal and Evolutionary Biology Research Programme, University of Helsinki, P.O. Box 65, Helsinki FI-00014 Finland

<sup>2</sup>Department of Biology, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim N-7491 Norway

<sup>3</sup>Department of Agricultural Sciences, University of Helsinki, P.O. Box 27, Helsinki FI-00014 Finland

<sup>4</sup>Département de Biologie, Université de Sherbrooke, 2500 boulevard de l'Université, Sherbrooke, Québec J1K 2R1 Canada

<sup>5</sup>Department of Mathematics, University of Utah, 155 South 1400 East, Salt Lake City, Utah 84112 USA

<sup>6</sup>School of Biological Sciences, University of Utah, 257 South 1400 East, Salt Lake City, Utah 84112 USA

<sup>7</sup>Manaaki Whenua Landcare Research, Private Bag 1930, Dunedin 1954 New Zealand

<sup>8</sup>Departamento de Biogeografía y Cambio Global, Museo Nacional de Ciencias Naturales, Consejo Superior de Investigaciones Científicas (CSIC), Calle José Gutiérrez Abascal 2, Madrid 28006 Spain

<sup>9</sup>Rui Nabeiro Biodiversity Chair, Universidade de Évora, Largo dos Colegiais, Evora 7000 Portugal

<sup>10</sup>Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Copenhagen 2100 Denmark

<sup>11</sup>Department of Statistical Science, Duke University, P.O. Box 90251, Durham, North Carolina 27708 USA

<sup>12</sup>School of BioSciences, University of Melbourne, Parkville, Victoria 3010 Australia

<sup>13</sup>Commonwealth Scientific and Industrial Research Organisation (CSIRO), Hobart, Tasmania, Australia

<sup>14</sup>Butterfly Conservation, Manor Yard, East Lulworth, Wareham BH20 5QP United Kingdom

<sup>15</sup>Department of Botany and Plant Sciences, University of California, Riverside, California 92521 USA

<sup>16</sup>Bio-Protection Research Centre, Lincoln University, P.O. Box 85084, Lincoln 7647 New Zealand

<sup>17</sup>Department of Ecology and Evolution (DEE), University of Lausanne, Biophore, Lausanne CH-1015 Switzerland

<sup>18</sup>Institute of Earth Surface Dynamics (IDYST), University of Lausanne, Geopolis, Lausanne CH-1015 Switzerland

<sup>19</sup>Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim N-7491 Norway

<sup>20</sup>Institute for Marine and Antarctic Studies, University of Tasmania, Private Bag 49, Hobart, Tasmania 7001 Australia

<sup>21</sup>Department of Biology, The University of Florida, Gainesville, Florida 32611 USA

<sup>22</sup>Mathematical Sciences Institute, The Australian National University, Acton, Australian Capital Territory 2601 Australia

<sup>23</sup>Nord University, Røstad, Levanger 7600 Norway

<sup>24</sup>BirdLife Norway, Sandgata 30B, Trondheim 7012 Norway

<sup>25</sup>Norwegian Institute for Nature Research, P.O. Box 5685, Torgarden, Trondheim NO-7485 Norway

<sup>26</sup>The Helsinki Lab of Ornithology, Finnish Museum of Natural History, University of Helsinki, P.O. Box 17, Helsinki FI-00014 Finland

<sup>27</sup>Department of Geosciences and Geography, University of Helsinki, P.O. Box 64, Helsinki 00014 Finland

<sup>28</sup>Biodiversity Division, Department of Environment, Land, Water & Planning, Arthur Rylah Institute for Environmental Research, 123 Brown Street, Heidelberg, Victoria 3084 Australia

<sup>29</sup>School of Mathematical and Physical Sciences, The University of Newcastle, University Drive, Callaghan, New South Wales 2308 Australia

<sup>30</sup>Department of Ecology, Swedish University of Agricultural Sciences, Box 7044, Uppsala 750 07 Sweden

<sup>31</sup>CNRS, LECA, Laboratoire d'Écologie Alpine, University Grenoble Alpes, Grenoble F-38000 France

<sup>32</sup>School of Mathematics and Statistics, Evolution & Ecology Research Centre, University of New South Wales, Sydney, New South Wales 2052 Australia

<sup>33</sup>Dynamic Macroecology, Swiss Federal Research Institute WSL, Zuercherstrasse 111, Birmensdorf CH-8903 Switzerland

*Citation:* Norberg A., N. Abrego, F. G. Blanchet, F. R. Adler, B. J. Anderson, J. Anttila, M. B. Araújo, T. Dallas, D. Dunson, J. Elith, S. D. Foster, R. Fox, J. Franklin, W. Godsoe, A. Guisan, B. O'Hara, N. A. Hill, R. D. Holt, F. K. C. Hui, M. Husby, J. A. Kålås, A. Lehtikoinen, M. Luoto, H. K. Mod, G. Newell, I. Renner, T. Roslin, J. Soininen, W. Thuiller, J. Vanhatalo, D. Warton, M. White, N. E. Zimmermann, D. Gravel, and O. Ovaskainen. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs* 89(3):e01370. 10.1002/ecm.1370

**Abstract.** A large array of species distribution model (SDM) approaches has been developed for explaining and predicting the occurrences of individual species or species assemblages. Given the wealth of existing models, it is unclear which models perform best for interpolation or extrapolation of existing data sets, particularly when one is concerned with species assemblages. We compared the predictive performance of 33 variants of 15 widely applied and recently emerged SDMs in the context of multispecies data, including both joint SDMs that model multiple species together, and stacked SDMs that model each species individually combining the predictions afterward. We offer a comprehensive evaluation of these SDM approaches by examining their performance in predicting withheld empirical validation data of different sizes representing five different taxonomic groups, and for prediction tasks related to both interpolation and extrapolation. We measure predictive performance by 12 measures of accuracy, discrimination power, calibration, and precision of predictions, for the biological levels of species occurrence, species richness, and community composition. Our results show large variation among the models in their predictive performance, especially for communities comprising many species that are rare. The results do not reveal any major trade-offs among measures of model performance; the same models performed generally well in terms of accuracy, discrimination, and calibration, and for the biological levels of individual species, species richness, and community composition. In contrast, the models that gave the most precise predictions were not well calibrated, suggesting that poorly performing models can make overconfident predictions. However, none of the models performed well for all prediction tasks. As a general strategy, we therefore propose that researchers fit a small set of models showing complementary performance, and then apply a cross-validation procedure involving separate data to establish which of these models performs best for the goal of the study.

**Key words:** *community assembly; community modeling; environmental filtering; joint species distribution model; model performance; prediction; predictive power; species interactions; stacked species distribution model.*

## INTRODUCTION

One of the key challenges in ecology is to predict how species and communities respond to spatiotemporal variation in abiotic and biotic conditions. The last two decades have seen a proliferation of species distribution models (SDMs) addressing the challenge of predicting the occurrences of individual species (Guisan and Zimmermann 2000, Guisan and Thuiller 2005, Elith et al. 2006, Leathwick et al. 2006, Zimmermann et al. 2010). Methodological advances in multiple-species distribution modeling have lagged behind, but are recently experiencing a rapid expansion (Leathwick et al. 2006, Dunstan et al. 2011, Guisan and Rahbek 2011, Warton et al. 2015, Wilkinson et al. 2019). Many previous studies (Table 1) have compared the predictive performance of SDMs for single-species analyses (Moisen and Frescino 2002, Thuiller et al. 2003, Elith et al. 2006, Leathwick et al. 2006, Elith and Graham 2009, Guisan and Rahbek 2011). Some studies have compared single-species and multi-species distribution models (Araújo and Luoto 2007, Heikkinen et al. 2007, Baselga and Araújo 2009, 2010, Elith and Leathwick 2009, Chapman and Purse 2011, Bonthoux et al. 2013, Madon et al. 2013, Maguire et al. 2016, Harris et al. 2018), while a few have examined the performance of alternative multiple species modeling approaches (Baselga and Araújo 2010, Madon et al. 2013, Wilkinson et al. 2019). Yet, a comprehensive comparison among SDM methods and many of the newly emerged joint SDM (JSDM) methods is still lacking. Furthermore, previous comparisons have largely focused on asking how well SDMs predict species-level

occurrences, but communities of interacting species are more than the sum of their constituent species. Hence, it is critical to also learn how well SDMs perform at a community level, i.e., in predicting how community composition covaries with environmental conditions. Variation in community composition can arise, for instance, because of chains of indirect interactions in multispecies networks and it is not clear how such processes might complicate multispecies distributional modeling efforts.

Communities of species result from numerous deterministic and stochastic assembly (and disassembly) processes, including the response of each species to its environment (environmental filtering, including episodic disturbances), to each other (biotic filtering), and to stochastic processes (e.g., dispersal, temporal variability, and ecological drift; Vellend 2010, Weiher et al. 2011, Götzenberger et al. 2012). Each statistical modeling method is based on different assumptions that can be viewed as hypotheses about how ecological communities are structured (D'Amen et al. 2017). Therefore, the capability of a modeling method to make predictions can be expected to depend on how well the underlying assumptions align with those assembly processes that shape the community. However, as most SDMs are phenomenological and based on finding statistical dependence between environmental and distributional data (so-called correlative models), they do not directly model the assembly processes themselves, but instead the patterns emerging from those processes (Baselga and Araújo 2009, Elith and Leathwick 2009). Thus, the link between the assumptions of SDMs and the assembly

TABLE 1. A review on recent species distribution model comparison studies.

Study	Data	Type	Model name abbreviations
Fielding and Haworth (1995)	R	SDM	DFA, GLM
Lek et al. (1996)	R	SDM	MR, NN
Mastrorillo et al. (1997)	R	SDM	ANN, DFA
Bio et al. (1998)	R	SDM	GAM, GLM
Franklin (1998)	R	SDM	CT, GAM, GLM
Manel et al. (1999)	R	SDM	GLM, NN, LDA
Vayssières et al. (2000)	R	SDM	CART, GLM
Moisen and Frescino (2002)	R, S	SDM, SSDM	ANN, CART, GAM, LM, MARS
Olden and Jackson (2002)	R, S	SDM	ANN, CFT, GLM, LDA
Loiselle et al. (2003)	R	SDM	BIOCLIM, DOMAIN, GLM, GARP
Thuiller et al. (2003)	R	SDM	CART, GAM, GLM
Segurado and Araújo (2004)	R	SDM	CT, ENFA, GAM, GLM, GOWER, NN, SI
Thuiller (2004)	R	SDM	ANN, CT, GAM, GLM
Elith et al. (2006)	R	SDM, SSDM	BIOCLIM, BRT, BRUTO, DOMAIN, GAM, GARP, GDM, GLM, LIVES, MARS, MAXENT
Austin et al. (2006)	S	SDM	GLM, GAM
Leathwick et al. (2006)	R	SDM, SSDM	GAM, MARS
Maggini et al. (2006)	R	SDM	GAM
Pearson et al. (2006)	R	SDM	ANN, CER, CGM, CT, GA, GAM, GARP, GLM
Randin et al. (2006)	R	SDM	GAM, GLM
Guisan et al. (2007a)	R	SDM, SSDM	BIOCLIM, DOMAIN, GLM, GAM, BRUTO, MARS, BRT, GARP, GDM, MAXENT
Guisan et al. (2007b)	R	SDM, SSDM	BIOCLIM, DOMAIN, GLM, GAM, BRUTO, MARS, BRT, GARP, BRT, MAXENT
Heikkinen et al. (2007)	R	SDM	GAM
Meynard and Quinn (2007)	S	SDM	GLM, GAM, GAM, CART, GARP
Peterson et al. (2007)	R	SDM	GARP, MAXENT
Wisz et al. (2008)	R	SDM, SSDM	BIOCLIM, DOMAIN, GLM, GAM, BRUTO, MARS, BRT, GARP, MAXENT, LIVES
Elith and Graham (2009)	S	SDM	GLM, BRT, RF, MAXENT, GARP
Santika and Hutchinson (2009)	S	SDM	BIOCLIM, GLM, GAM, CART
Syphard and Franklin (2009)	R	SDM	GAM, GLM, CT, RF
Baselga and Araújo (2010)	R	SDM, SSDM	GLM, CQO
Hoffman et al. (2010)	S	SDM	GLM, GAM, MAXENT, DCM
Santika (2011)	S	SDM	GLM, GAM, CART
Wenger and Olden (2012)	R	SDM	ANN, GLM, RF
Bahn and McGill (2013)	R	SDM	BRT, GAM, GARP, MARS, MAXENT, RF
Hui et al. (2013)	R	SDM, JSMD	GLM, SAM
Owens et al. (2013)	S, R	SDM	GAM, GARP, MAXENT
Madon et al. (2013)	R	SDM	GLM
Miller (2014)	S	SDM	-
D'Amen et al. (2015)	R	SDM, SSDM	GLM, GAM, BRT, RF (SESAM <sup>2</sup> )
Maguire et al. (2016)	R	ORD, SDM, SSDM	CAO, CQO, MANN, MARS, MRT, GLM, GAM, ANN, MARS, CART
D'Amen et al. (2017)	R	SDM, JSMD	GLM, GAM, BRT, BORAL
Sor et al. (2017)	R	SDM	ANN, GLM, RF, SVM
Harris et al. (2018)	R	SDM, JSMD	BRT, RF, MISTNET
Nieto-Lugilde et al. (2018)	-	ORD, SDM, SSDM, JSMD	CLO, CQO, CAO, GDM, GF, HBM, MANN, MARS, MRT, GLM, GAM, RF, ANN, CART
Zhang et al. (2018)	S, R	JSMD	HMSC, BORAL, GJAM, MISTNET, BC
Wilkinson et al. (2019)	R	JSMD	BC, GJAM, BORAL, HMSC

*Notes:* Data indicates whether the comparisons were based on models fitted to simulated (S) and/or real empirical data (R); Type refers to whether the compared model types were single-species distribution models (SDM), stacked species distribution models (SSDM), joint species distribution model (JSMD) or ordination-based models (ORD). The last column provides the names of the modeling frameworks compared.

processes is typically indirect and challenging to discern. In a somewhat simplified view, environmental filtering will result in an association between local environmental conditions and species occurrences, whereas biotic filtering will result in species co-occurrence that cannot be attributed solely to correlated responses to the environment (Cazelles et al. 2016). Stochastic processes, as well as historical contingencies (e.g., evolutionary processes, founder effects, alternative stable states or past environmental conditions), can be expected to produce distributions with unexplained residual spatial autocorrelation, thus being best captured by spatial predictors (and ideally, historical information). All of these factors need to be woven into statistical analyses of ecological patterns.

The aim of this study is to compare the predictive performances of a large number of SDM methods applied to a common suite of community data sets and to ask how their predictive performance relates to their structural properties. To do so, we first classify SDM methods based on their structural properties (later referred to as “Features A–G”; Table 2), and discuss how these can be translated into hypotheses about how communities are structured. In short, these methods differ in regard to whether they are parametric or semi-parametric (Feature A); whether or not they account for interactions among environmental covariates when estimating species responses to the environment (Feature B); whether or not they assess shared responses by species to the environment (Feature C); whether or not they explicitly include species co-occurrences not related directly to environmental variables (Feature D); whether or not they explicitly account for spatial structure (Feature E); whether or not the statistical inference framework applies shrinkage when estimating the response of each species to its environment (Feature F); and whether the statistical framework accounts for parameter uncertainty when generating the predictions (Feature G). The next paragraphs explain these structural properties in more detail.

Species distribution models vary in how they represent the relationship between local environmental conditions and species occurrences (Guisan and Thuiller 2005, Peterson et al. 2011). They range from purely data-driven SDMs allowing for very flexible predictor functions (e.g., random forest and generalized additive models) to more rigid ones (e.g., generalized linear models; Guisan et al. 2002, James et al. 2013, Merow et al. 2014; Table 2, Feature A). Even if there are expectations about the unimodal relationship that species distributions should have with main environmental predictors (Austin et al. 2009), there is evidence that the relationship is likely skewed and there is complete lack of information regarding the actual relationships when several variables interact to shape the distribution of a species (Normand et al. 2009, Araújo et al. 2013). However, more flexibility carries the cost of increasing the number of degrees of freedom, which, in turn, increases the risk of statistical overfitting and thus modeling noise rather than

signal (Araújo et al. 2005, Randin et al. 2006, Wenger and Olden 2012, Merow et al. 2014, García-Callejas and Araújo 2016). The same consideration holds when asking whether to include interactions among environmental predictors (Table 2, Feature B): while both ecological theory and empirical studies suggest that how ecological processes depend on one covariate may depend on the value of other covariates (Harpole et al. 2011), including interactions among covariates increases model complexity and, therefore, the risk of statistical overfitting (Guisan et al. 2006, Merow et al. 2014).

With inventory data on multiple species, one can additionally make assumptions about how the relationship between environmental covariates and species occurrences is structured among species (Table 2, Feature C). The widely used stacked species distribution models are first fit separately for each species, after which their predictions are combined. They thus assume that species respond individualistically to variation in environmental conditions (Williams and Jackson 2007, Guisan and Rahbek 2011). By comparison, the more recently developed joint species distribution models (JSDMs) represent the response of entire species assemblages to environmental variation, assuming, for example, that species with similar traits have similar responses (Warton et al. 2015, Ovaskainen et al. 2017). In complex communities, it is difficult to predict a priori the joint structure of species responses to environmental variation and thus one might assume that treating each species individually is more in line with our limited current understanding of community assembly. However, treating each species individually may come with a higher risk of overfitting, while borrowing information from other species may increase predictive performance if the species respond similarly enough to abiotic variation (Ovaskainen and Soininen 2011, Hui et al. 2013, Madon et al. 2013, Maguire et al. 2016). Intermediately common species may show more statistically reliable relationships with environmental variables than rare species with wide and scattered distributions (Segurado and Araújo 2004), so treating assemblages as a whole can in effect increase the statistical power of detecting true environment–species relationships for rarer species within communities (Ovaskainen and Soininen 2011, Hui et al. 2013).

Species distribution models also vary in their assumptions whether and how biotic interactions influence species occurrences (Kissling et al. 2012, Wisz et al. 2013). Biotic interactions can be expected to result in nonrandom co-occurrence patterns, with the caveat that nonrandom co-occurrence patterns can also result from species responses to unmeasured environmental variation (Araújo et al. 2011, Pollock et al. 2014, Ovaskainen et al. 2017). Most SDMs assume that species distributions are statistically independent of each other after controlling for the effects of environmental covariates (Table 2, Feature D). Yet, it is possible to account for



TABLE 2. Summary of statistical modeling frameworks considered in this paper.

Model	Model name	Variant	Statistical inference framework	Type	Feature							Reference
					A	B	C	D	E	F	G	
BC	Bayesian community ecology analysis	BC.1	Bayes	JSDM	0	0	0	0	0	1	1	Golding and Harris (2015)
BC	BC with species associations	BC.2	Bayes	JSDM	0	0	0	1	0	1	1	Golding and Harris (2015)
BORAL	Bayesian ordination and regression analysis	BORAL.1	Bayes	JSDM	0	0	0	1	0	1	1	Hui (2017)
BRT	boosted regression trees	BRT.1	ML	SSDM	1	1	0	0	0	0	0	Hijmans et al. (2017), Ridgeway (2017)
GAM	generalized additive models	GAM.1	ML	SSDM	1	0	0	0	0	0	0	Wood (2011)
GAM	GAM with spatial structure	GAM.2	ML	SSDM	1	0	0	0	1	0	0	Wood (2011)
GJAM	generalized joint attribute modeling	GJAM.1	Bayes	JSDM	0	0	0	1	0	1	1	Clark et al. (2017)
GLM	generalized linear models	GLM.1	ML	SSDM	0	0	0	0	0	0	0	R Core Team (2018)
GLM	GLM fitted with PQL	GLM.2	ML	SSDM	0	0	0	0	0	0	0	Venables and Ripley (2002)
GLM	GLM with PQL and spatial random effect	GLM.3	ML	SSDM	0	0	0	0	1	0	0	Venables and Ripley (2002)
GLM	Bayesian (single-species HMSC)	GLM.4	Bayes	SSDM	0	0	0	0	0	1	1	Ovaskainen et al. (2017)
GLM	Bayesian and spatial (single-species HMSC)	GLM.5	Bayes	SSDM	0	0	0	0	1	1	1	Ovaskainen et al. (2017)
GLM	GLM fitted with MVABUND	GLM.6	ML	SSDM	0	0	0	0	0	0	0	Wang et al. (2012)
GLM	Bayesian (BORAL with no latent variable)	GLM.7	Bayes	SSDM	0	0	0	0	0	1	1	Hui (2017)
GLM	same as GLM.1, but predictions incorporate parameter uncertainty	GLM.8	ML	SSDM	0	0	0	0	0	0	1	Foster and Dunstan (2010), R Core Team (2018)
GLM	GLM fitted with MVABUND with LASSO	GLM.9	ML	SSDM	0	0	0	0	0	1	0	Wang et al. (2012)
GLM	GLM fitted with GLMNET with LASSO	GLM.10	ML	SSDM	0	0	0	0	0	1	0	Friedman et al. (2010)
GLM	same as GLM.10, but predictions incorporate parameter uncertainty	GLM.11	ML	SSDM	0	0	0	0	0	1	1	Friedman et al. (2010)
GLM	same as GLM.1, but the model includes interactions between covariates	GLM.12	ML	SSDM	0	1	0	0	0	0	0	R Core Team (2018)
GLM	same as GLM.4, but the model includes interactions between covariates	GLM.13	Bayes	SSDM	0	1	0	0	0	1	0	Ovaskainen et al. (2017)
GNN	gradient nearest neighbor	GNN.1	ML	SSDM	1	1	1	0	0	0	0	Crookston and Finley (2008)

TABLE 2. (Continued)

Model	Model name	Variant	Statistical inference framework	Type	Feature							Reference
					A	B	C	D	E	F	G	
HMSC	hierarchical modeling of species communities	HMSC.1	Bayes	JSDM	0	0	1	0	0	1	1	Ovaskainen et al. (2017)
HMSC	HMSC with species associations	HMSC.2	Bayes	JSDM	0	0	1	1	0	1	1	Ovaskainen et al. (2017)
HMSC	HMSC with species associations implemented as spatial random effects	HMSC.3	Bayes	JSDM	0	0	1	1	1	1	1	Ovaskainen et al. (2017)
HMSC	same as HMSC.1, but the model includes interactions between covariates	HMSC.4	Bayes	JSDM	0	1	1	0	0	1	1	Ovaskainen et al. (2017)
MARS	multivariate adaptive regression spline (MARS-COMM)	MARS.1	ML	SSDM	1	0	0	0	0	0	0	Milborrow (2017)
MARS	MARS with interactions in covariate selection (MARS-INT)	MARS.2	ML	SSDM	1	1	1	0	0	0	0	Milborrow (2017)
MISTN	multivariate stochastic neural networks	MISTN.1	ML	JSDM	1	1	0	1	0	0	0	Harris (2015)
MRTS	multivariate regression tree	MRTS.1	ML	SSDM	1	1	1	0	0	0	0	De'Ath et al. (2014)
RF	random forest	RF.1	ML	SSDM	1	1	0	0	0	0	0	Liaw and Wiener (2002)
SAM	species archetype model	SAM.1	Bayes	JSDM	0	0	1	0	0	1	1	Hui et al. (2013) (the exact implementation provided by the developer)
SVM	support vector machines	SVM.1	ML	SSDM	1	1	0	0	0	0	0	Meyer et al. (2017)
XGB	gradient extreme boosting	XGB.1	ML	SSDM	1	1	0	0	0	0	0	Chen et al. (2018)

*Notes:* The column “Statistical inference framework” describes whether the model was fitted to data in the Bayesian or in the maximum likelihood (ML) framework. The column “Type” classifies each model either as stacked species distribution model (SSDM) or joint species distribution model (JSDM). Feature A refers to the assumed form of species response to their environment, classified as semi-parametric (1) or parametric (0). Feature B describes whether the statistical inference framework accounts (1) or does not account (0) for interactions among environmental covariates when estimating the responses of species to them. Feature C classifies the models according to whether models share (1) or do not share (0) information among the species when modeling their responses to environmental covariates. Feature D describes whether the modeling method accounts (1) or does not account (0) for species co-occurrences not explained by their environmental niches. Feature E describes whether the model accounts (1) or does not account (0) explicitly for spatial variation. Feature F describes whether the statistical inference framework involves (1) or does not involve (0) shrinkage when estimating the responses of species to environmental covariates. Feature G describes whether the statistical inference framework accounts for (1) or does not account for (0) parameter uncertainty when generating the predictions. For more detailed descriptions of the models, information on their practical implementations, as well as more references for the methods and their use in practice, see Appendix S1.

interspecific associations even in the context of single-species SDMs by using the occurrences of some species as predictors (Leathwick and Austin 2001, Meentemeyer et al. 2001, Stephens and MacCall 2004, Araújo and Luoto 2007, Pellissier et al. 2010, Meier et al. 2011, Kissling et al. 2012, Mod et al. 2015, Mäkinen and Vanhatalo 2018). This seems particularly appropriate when some species play disproportionately large roles in the lives of others (e.g., keystone or foundation species, and host plants for host-specific herbivores). Alternatively, JSDMs model the occurrences of all species in a

community simultaneously and include a covariance structure to capture species-to-species associations, without necessarily assuming rigid species-by-species relationships (Clark et al. 2014, Pollock et al. 2014, Thorson et al. 2015, Ovaskainen et al. 2017). A model that accounts for species-to-species associations can be expected to be superior in predicting community-level features (e.g., community composition or species richness) for those communities in which biotic interactions are in fact a strong driver of local coexistence (Wisz et al. 2013).

The impact of stochastic processes such as dispersal and ecological drift on species distributions has received relatively little attention in the SDM literature, partly because it is challenging to derive straightforward hypotheses about these processes from non-manipulative observational data (Araújo and Guisan 2006, Thuiller et al. 2013) and partly because stochastic process models are inherently challenging and still under development in ecology (Pásztor et al. 2016). The most appropriate way to account for such processes in the context of SDMs is to incorporate model structures and parameters describing directly the demographic processes underlying the community (Morin et al. 2008, Boulangeat et al. 2012, Dormann et al. 2012, Thuiller et al. 2013, Talluto et al. 2016, Zurell et al. 2016). These might for instance incorporate greater impacts of stochasticity on rare species within communities (Umaña et al. 2017). An alternative way to account for, e.g., dispersal or missing covariates is to include spatial predictors or covariance structures that control for the variation in the data that cannot be attributed to the variation in observed abiotic or biotic environmental conditions (Augustin et al. 1996, Dormann 2007, Dormann et al. 2007, Miller 2012; Table 2, Feature E). The inclusion of spatial structure can be expected to provide increased predictive performance for interpolation (predictions made for similar environmental conditions and same region as data used for model fitting), by borrowing information about species occurrences from nearby sites, which are likely linked by dispersal (Latimer et al. 2006). A model failing to account for spatial autocorrelation can in some cases (but not necessarily) lead to biased or spurious relationships between environmental variation and species occurrence, decreasing predictive power both for interpolation as well as extrapolation (predictions made for dissimilar environmental conditions or different region as data used for model fitting; Diniz-Filho et al. 2003, Diggle and Ribeiro 2007, Fieberg et al. 2010, Thibaud et al. 2014).

In addition to model structure and the selection of predictors, the statistical inference framework within which the model is fit to data can have a major impact on predictive performance. In comparison to the maximum likelihood (ML) framework, parameterization with Bayesian inference is not only influenced by the data but also by prior information (Ellison 2004). Bayesian inference (or, more generally, shrinkage estimators, including penalized maximum likelihood; Table 2, Feature F), allows the researcher to utilize prior information and assumptions regarding how species respond to the abiotic environment or to each other, thus influencing parameter estimates, especially when data are scarce. Whether guiding the model parameterization with the help of prior information improves predictive performance, or instead deteriorates it, clearly depends on the accuracy of the prior information. Another important choice is how parameter uncertainty is accounted for in model predictions (Beale and Lennon 2012), if at all

(Table 2, Feature G). While ML applications typically generate predictions utilizing solely point estimates and only generate confidence intervals (if at all) through resampling, applications utilizing the Bayesian inference framework often propagate parameter uncertainty by resampling the parameters from the posterior distribution for each replicate prediction (Clark 2005).

Here, we evaluate the predictive performance of different modeling methods, all varying accounting for the features presented above. To achieve this goal, we used five spatially explicit data sets on species occurrence for different types of communities (birds, butterflies, herbaceous plants, trees, and vegetation data; Table 3) from different geographical regions. Specifically, we asked how well 33 variants of 15 modeling frameworks perform in predicting species occurrences under spatial and environmental conditions that were either similar to (interpolation) or different from (partial or full extrapolation) those in the training data. Earlier studies comparing SDMs have evaluated predictive power mainly on a per species basis (Fielding and Haworth 1995, Allouche et al. 2006, Elith et al. 2006). Here, we compare the models' predictive ability using performance measures defined both at the species and community levels. Moreover, while most earlier comparisons have assessed predictive performance in terms of discrimination (e.g., using the area under the curve [AUC] statistic), we evaluate predictive performance in terms of accuracy, discrimination, calibration, and precision (Fig. 1, Table 4). This suite of metrics provides distinctive assessments of model performance.

Based on the reasoning above, our overarching hypothesis is that variation in predictive performance can be linked to structural variation among statistical models, as classified by Features A–G (Table 2). In particular, we hypothesize that semi-parametric models that allow for flexible responses of species to environmental covariates (Feature A; Table 2), models that account for interactions among environmental predictors (Feature B; Table 2), models that do not assume joint responses among the species (Feature C; Table 2), models that use spatial predictors (Feature E; Table 2), and models that do not apply shrinkage (Feature F; Table 2), are superior in predicting occurrence probabilities for common species with a large number of occurrences. In contrast, we hypothesize that for rare species with limited data the superior models will include some of the following: parametric responses, no interactions among environmental predictors, joint responses among the species, shrinkage, or no spatial predictors. The reasons for these hypotheses are several-fold: (1) semi-parametric models and models with interaction terms require more data than parametric models and models without interaction terms to be successfully fitted; (2) borrowing information from other species is expected to be especially beneficial for rare species for which fitting species-specific models is difficult (Madon et al. 2013); (3) spatial autocorrelation is pervasive in natural ecosystems (Dormann et al.

TABLE 3. Descriptions of the data sets used to test the performance of the statistical modeling approaches.

No.	Data set	Type	Species	Species prevalence range as median (min–max)	Reference
1	breeding bird surveys in Finland, Sweden, and Norway	atlas data	141	0.16 (0.0066–0.97)	Lindström et al. (2015)
2	butterflies in Great Britain	atlas data	50	0.43 (0.018–0.94)	Asher et al. (2001)
3	plants from Victorian Biodiversity Atlas	community	162	0.018 (0.0033–0.23)	<a href="https://www.environment.vic.gov.au/biodiversity/victorian-biodiversity-atlas">https://www.environment.vic.gov.au/biodiversity/victorian-biodiversity-atlas</a>
4	trees in the United States	community	63	0.04 (0.0067–0.36)	<a href="http://fia.fs.fed.us/">http://fia.fs.fed.us/</a>
5	vegetation in northern Norway	community	242	0.045 (0.0017–0.69)	Niittynen and Luoto (2017)

Notes: The columns show for each data set (1) the types of organisms included in the set, (2) whether the set is true community data or based on atlas data, (3) the number of species in the data set, (4) the prevalence of species in the data set (including data for both for training and validation), (5) a reference to the data.

2007), as dispersal couples local communities into broader, regional metacommunities, but the proper estimation of spatial residual structure requires considerable data; and (4) bringing prior information is expected to make important differences especially for modeling rare species. We further hypothesize that models that account for species-to-species associations (Feature D; Table 2) will exhibit better predictive performance especially in terms of community-level features that depend on co-occurrences, i.e., variability in species richness and community composition. Finally, we hypothesize that models that account for parameter uncertainty in their

predictions (Feature G) are not necessarily more accurate nor have higher discrimination power, but that they are better calibrated than models that do not account for parameter uncertainty.

MATERIALS AND METHODS

We evaluate the predictive performance of 33 variants of 15 SDMs (Table 2) using five data sets on species-rich communities (Table 3). The general workflow of our study is summarized in Fig. 1.

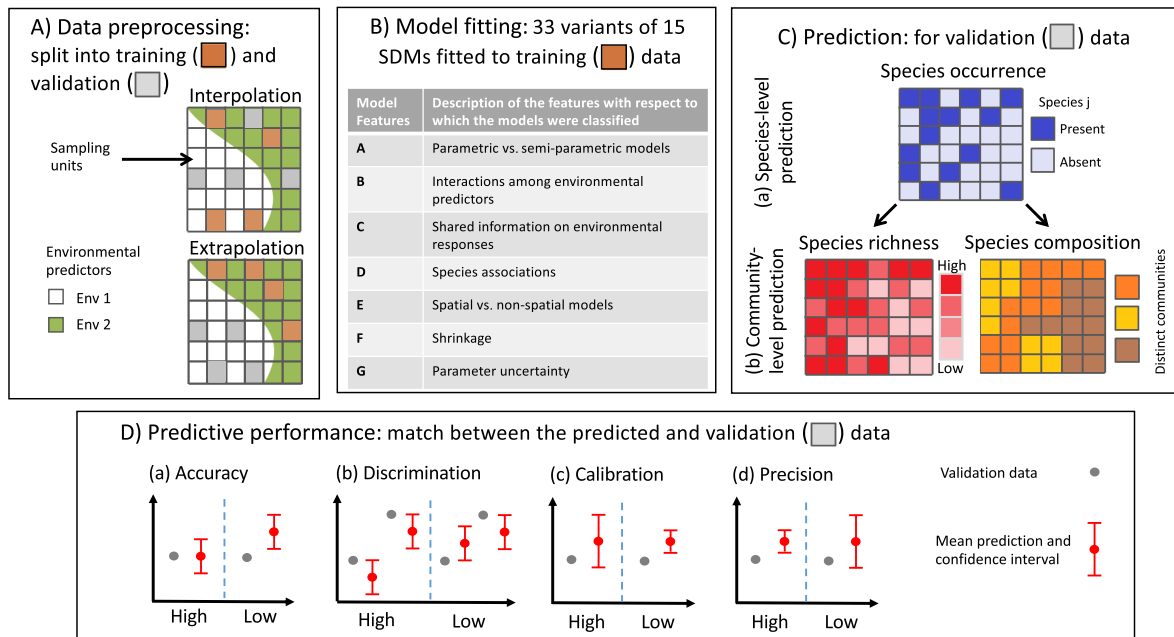


FIG. 1. Workflow of the study. We split data sets into (A) training and validation data, (B) fitted models to training data, and (C) compared model predictions to validation data in terms of species occurrences, species richness, and community composition. (D) We evaluated the predictive power in terms of accuracy, discrimination, calibration and precision. Panel B describes the features A–G with respect to which the models have been classified, as detailed in Table 2. In panel D, an accurate prediction is close to the true value (a), predictions with high discrimination can separate, e.g., sites where species occurs from those where it does not (b), well-calibrated predictions have valid confidence intervals (c), and precise predictions present little uncertainty (d).



TABLE 4. The performance measures used to assess how well the different statistical frameworks are able to predict held out validation data.

Ecological level (rows) and aspect of performance (columns) to be measured	(a) Accuracy	(b) Discrimination	(c) Calibration	(d) Precision
1. Species-specific occurrence	absolute difference between expected (probability) and observed (0/1) occurrence, averaged over species and sites	AUC, averaged over species	absolute difference between predicted and observed numbers of occurrences in 10 probability bins (each including same number of data points, based on quantiles), averaged over species	$\sqrt{p(1-p)}$ , where $p$ is the probability of species occurrence, averaged over species and sampling units
2. Species richness	root mean squared error (RMSE) between mean prediction and observed richness	Spearman rank correlation among sites/regions, based on predictive mean	$ p - 0.5 $ , where $p$ is the proportion of predictions that fall within 50% prediction interval	average of predictive standard deviations
3. Community composition measured by Sorensen, Simpson and nestedness indices	root mean squared error (RMSE) between predictive mean and observed composition	Spearman rank correlation among pairs of sites	$ p - 0.5 $ , where $p$ is the proportion of predictions that fall within 50% prediction interval	average of predictive standard deviations

#### *Analysed data sets*

All of our data are presence–absence data in the sense that they consist of 0s and 1s for all species and sampling units (rather than only coordinates of known occurrences of species), but with some of the data sets a proportion of the zeros are likely to result from lack of observation or observation error rather than true absences (Guillera-Arroita 2017). The herbaceous plant, tree, and vegetation data sets were all collected at a spatial scale at which the organisms can be expected to interact within each community, and thus can be considered as data on local ecological communities. In contrast, the data on butterfly and bird distributions represent atlas data on species assemblages sampled at broader spatial scales, which likely comprise many local communities. The tree and vegetation data were acquired with exhaustive sampling of study plots, and thus can be considered true presence–absence data, whereas absences in the other data sets may to a degree represent inadequate sampling, and so conservatively should be viewed as “presence-only” data. All data sets are spatially explicit, in that the sampling units involve information on their geographical coordinates. However, the data for the different functional groups come from different geographical regions, so the analyses presented here do not delve into some community ecology processes that can bear importantly on distributions (e.g., butterfly dependencies on plant host species, or impacts of vertebrate herbivores on herbaceous species assemblages).

As some of the statistical methods are computationally intensive (Appendix S3), their application to the original full data was not possible. To enable comparison among all methods, we subsampled each data set to 1,200 sampling units and included only those species that were present in at least 10 sampling units and that were present at least once in all three training data sets (see below). The main features of subsampled data are described in the following paragraphs and in Table 3.

*Bird data.*—The data originate from national common bird monitoring programs in Finland, Sweden, and Norway (Lindström et al. 2015). Between 2013 and 2014, a total of 141 bird species were surveyed using line transects (Finland and Sweden) and point counts (Norway). The largest distance between the sampling units was 1,853 km. The covariates (which are detailed in Appendix S2 for all five data sets) include 21 variables related to land cover, climate, and variation in sampling effort. There is substantial overlap in the species composition within these countries, and so it is reasonable to consider the data set as a cohesive Fennoscandian faunal survey.

*Butterfly data.*—The data originate from the Butterflies for the New Millennium recording scheme in Great Britain (Asher et al. 2001). The data on 50 butterfly species were recorded in 1995–1999 on a 10 × 10 km grid, and the largest distance between sampling units was 640 km. The environmental covariates include 34 variables related to land cover, topography, and climate.

*Herbaceous plant data.*—The data originate from the Victorian Biodiversity Atlas, which is a state database that collaborates with the Atlas of Living Australia (*available online*).<sup>35,36</sup> The presence–absence data on 161 herbaceous species were collected in years 1984–2014 on sampling plots of size 900 m<sup>2</sup>, and the largest distance between the sampling units was 895 km. The environmental covariates include 19 variables related to soil, topography, and climate.

*Tree data.*—The data originate from the US Forest Service’s Forest Inventory and Analysis (*available online*).<sup>37</sup> The data on 89 tree species were recorded in 2012 on sampling plots of 672 m<sup>2</sup> across the eastern United States, and the largest distance between the sampling units was 3,500 km. The environmental covariates include 38 variables related to soil, topography, and climate.

*Vegetation data.*—The vegetation data originate from a community ecological study conducted in northern Norway (Niittynen and Luoto 2017). The data on 245 species of plants, bryophytes, and lichens were surveyed in 2014–2016 on sampling plots, each of which consisted of four 1-m<sup>2</sup> squares. The largest distance between the sampling units is 18 km. The environmental covariates include six variables related to soil, topography, and climate.

#### *Selection of covariates and subsampling the data sets into training and validation data*

While covariate selection is an important part of any statistical modeling exercise, we utilized the same set of pre-selected covariates in all statistical models to ensure the comparability of the results by minimizing the number of model-specific subjective choices. To reduce the number of potential predictors and thus the risk of overfitting, we reduced the raw predictors using principal components of the environmental covariates at the sampling locations. We then included the first five principal components (PC) as predictors, except if a smaller number was sufficient to explain at least 80% of the variation. The numbers of principal components included (and their proportions of explained variance) were, respectively, five (56%) for the bird data, five (47%) for the butterfly data, five (78%) for the herbaceous plant data, three (83%) for the tree data, and four (88%) for the vegetation data.

We split each data set into two parts to form training data and validation data. We did this in three ways to mimic the tasks of interpolation, partial extrapolation, and full extrapolation. Interpolated validation data

represent environmental and spatial conditions that are similar to those in the training data, whereas the conditions in the partially and, especially, the fully extrapolated validation data differ systematically from those in the training data, making the task of prediction more challenging. The predictive ability of a model to interpolate tests the ability to capture species occurrence within known environments, while extrapolation tests that model’s ability to predict to environmental conditions outside of the training data (Randin et al. 2006). The interpolated validation data were constructed by randomly selecting half of the sampling units and leaving the remaining half for training. The fully extrapolated validation data include those sampling units for which the PC1 value was higher than the median value. To construct partially extrapolated validation data, we grouped the sampling units randomly into pairs and selected from each pair the one with the lower PC1 value for training data, and the other one for validation data. This resulted in the training data having, on average, lower PC1 values. While we split the data into training and validation data based on the distributions of the environmental covariates, at the same time these splits resulted in related patterns of spatial partitioning: in the case of interpolation, the training and validation data are spatially randomly distributed with respect to each other, whereas in the case of full extrapolation, they are spatially well separated from each other (Appendix S2). Thus, in the interpolated cases the validation and training data cannot be considered fully independent, whereas for the extrapolated data the assumption of independence holds better (Roberts et al. 2017).

The data used for fitting the statistical models (i.e., the training data) are the  $n \times m$  matrix  $\mathbf{Y}$  of species occurrences, the  $n \times k$  matrix  $\mathbf{X}$  of environmental covariates, and the spatial coordinates of the sampling units. Here,  $n$  is the number of sampling units,  $m$  the number of species, and  $k$  the number of environmental predictors. The validation data consist of the corresponding matrices  $\mathbf{Y}^v$  and  $\mathbf{X}^v$  and their spatial coordinates. To examine the effect of the size of the data set on our outcomes, we included either  $n = 600$  or  $n = 150$  sampling units in the training data. To do so, we either used the full training data, or randomly sampled 150 units from it. The validation data always consisted of  $n = 600$  sampling units. The reason for not following alternative possible protocols (e.g., a leave-one-out cross-validation strategy) was that some of the models were computationally too intensive to be fitted repeatedly.

#### *Modeling methods considered*

We selected 15 SDM methods that are suitable for modeling presence–absence data (hence excluding, e.g., Maxent; Guillerá-Arroita et al. 2014) based on reviewing recent literature and selecting both routinely used and recently emerged methods (Table 2). We included several variants of some of the SDMs in order to provide

<sup>35</sup> <https://www.environment.vic.gov.au/biodiversity/victorian-biodiversity-atlas>

<sup>36</sup> <http://www.ala.org.au>

<sup>37</sup> <http://fia.fs.fed.us/>

resolution on how different types of underlying assumptions (Features A–G, Table 2) influence predictive capability. In particular, we included 13 variants of the widely applied GLM (out of which 11 were non-spatial and two spatial; six were without and seven with shrinkage) in order to examine the sensitivity of the results to the statistical inference framework and how it is implemented. For all 33 SDM variants included, we utilized the same environmental predictors, but the spatial coordinates of the sampling units were included only for spatially explicit models. We classified 23 of the 33 SDM variants as stacked species distribution models (SSDM; Dubuis et al. 2011, Guisan and Rahbek 2011), since they essentially model species individually and then stack the model predictions together to build up a compound prediction at the community level (Ferrier and Guisan 2006; Table 2). The remaining 10 model variants were classified as joint species distribution models (JSDM), as they construct a single model that connects the species together, with some of the model parameters being at the community level (Warton et al. 2015).

When fitting models that make strict assumptions about the functional forms of the response to the environment (Feature A classified as 0, Table 2), we included the linear and squared effects of the PCs as predictors in accordance with niche theory, which predicts that species will usually have their maximum occurrence at some interior position within their multidimensional niche space, say nearer the centroid than on the edge (Austin 2002). When fitting models that do not make such assumptions (Feature A classified as 1, Table 2), we did not include squared predictors, since those models test and account for non-linear relationships by default. To examine the influence of interactions among the environmental predictors, we included three comparisons (GLM.12 vs. GLM.1; GLM.13 vs. GLM.4; HMSC.4 vs. HMSC.1) out of which one included and the other one excluded such interaction. In cases where model fitting failed technically (e.g., due to quasi-complete separation), we fitted an intercept-only model, except for the case of spatial models that failed technically, for which we first attempted to fit the corresponding non-spatial model. Further technical details on how the statistical models were fitted to the data are presented in Appendix S1.

As many of the communities included a high proportion of rare species, and predicting their occurrences can be challenging, we further considered either all species, or included only species with a prevalence of at least 10%, henceforth called common species. Thus, for the SSDMs we fitted the species-specific models once, and stacked them either for all species or for the common species only. For JSDM, model fitting is influenced by the selection of the species, and thus we fitted the JSDM models separately for all and for the common species.

To summarize, we fitted 33 statistical model variants to five data sets. Each of these data sets was split in three different ways into training and validation sets, and, in

each case, two different sizes of data set were assessed, and two types of species communities (all or common) were included. Thus, the total number of cases that we considered was 1980.

### *Evaluating predictive performance*

We compared the predictive performance of the different statistical frameworks both at the species and at the community levels. To do so, we fitted the models based on the training data  $\mathbf{X}$  and  $\mathbf{Y}$ , then used the fitted model and the environmental conditions  $\mathbf{X}^v$  to predict species occurrences in the validation data, and finally compared the predicted occurrences to the true occurrences  $\mathbf{Y}^v$ . Community-level tests require joint predictions for all species, which we did by using the models to predict 100 random realizations of species occurrence matrices, i.e., matrices of zeros and ones. The mean of the predicted occurrences equals occurrence probability (up to sampling error), but the predicted occurrences involve also information on dependencies among species (and sometimes among spatial units) beyond occurrence probabilities (see also Appendix S1). Typical applications of Bayesian models account for parameter uncertainty when making predictions, whereas predictions derived from ML models are often based on point estimates. To follow these conventions, in models fitted with Bayesian inference, the 100 random realizations corresponded to Monte Carlo estimates from the posterior predictive distribution, whereas for models fitted with maximum likelihood (ML) inference, we used the point estimates for each prediction and applied 100 realizations of Bernoulli randomization based on the predicted occurrence probabilities. As an exception, to examine specifically the influence of parameter uncertainty, we included two SDM variants (GLM.8 and GLM.11) that were fitted in the ML framework, but for which we accounted for parameter uncertainty in the predictions by a parametric bootstrap routine (used in, e.g., Foster and Dunstan 2010). We did so by drawing the parameters for each of the 100 predictions from the estimated asymptotic distribution and transforming to the response scale, using the inverse link function.

The samples of  $\mathbf{Y}^v$  provide a Monte Carlo approximation for the joint predictive distribution of all species. We note that many previous applications of SDMs have evaluated them based on either the predicted species-specific marginal occurrence probabilities, or occurrences derived by thresholding the occurrence probabilities (Liu et al. 2005, Jiménez-Valverde and Lobo 2007, Lawson et al. 2014). The reason why we did not solely use the marginal (species-specific) occurrence probabilities is that these probabilities neglect correlations among species occurrences, thus predicting inevitably that two species with marginal occurrence probabilities 0.5 are found from the same sampling unit with probability 0.25. In contrast, our predictions accommodate possible co-occurrence as estimated by joint species distribution

models, thus allowing for the prediction where both of the above-mentioned species are present in half of the sampling units and both are absent in the remaining half of the sampling units. By predicting the joint distribution of  $\mathbf{Y}^v$  we can evaluate both marginal species- and sampling unit-specific predictions and the joint species distribution predictions.

To further examine the performance of ensemble modeling (Thuiller 2004, Marmion et al. 2009), we averaged predictions produced by the individual model variants. As one approach to ensemble modeling, we averaged the predictions of all 33 model variants. To do so, we generated 99 random realizations of species occurrence matrices by randomly selecting three such matrices generated for each model variant, and we then added one prediction of randomly selected model variant to obtain 100 matrices as for the other models. As an alternative approach to ensemble modeling, we averaged the predictions of the best performing model variants of the five best performing models (see below on how these were selected). In this case we generated 100 random realizations of species occurrence matrices by randomly selecting 20 such matrices generated for each selected model variant.

#### *Measures of predictive performance*

In order to compare predictive performance in a comprehensive and coherent manner, we evaluated the ability of the models to predict withheld validation data at three levels: (1) species occurrence, (2) species richness, and (3) community composition. For each of these levels, we measured predictive performance in terms of accuracy, discrimination power, calibration, and precision (Fig. 1, Table 4). In statistical terminology, accuracy is the opposite of bias, and measures the degree of proximity between the predicted and the true value (here the observed value in the validation data). Discrimination power does not examine the absolute match between predicted and true values, but how well (some) predictive value can discern different types of true values (e.g., presence-absence). Calibration refers to statistical consistency between distributional predictions and the true values; that is, in calibrated predictions the relative frequency of test values with predictive probability  $p$  should be  $p$  (Gneiting and Raftery 2007). Precision (also referred to as sharpness) measures the width of the predictive distribution and thus its information content.

*Performance measures related to species-specific occurrence probabilities.*—For the measures of predictive performance at the species level, we averaged the 0/1 predictions over the 100 replicate matrices, thus obtaining species- and site-specific predicted occurrence probabilities. As a measure of accuracy, we used the absolute difference between the observed occurrence (0 or 1) and the predicted probability of occurrence, averaged over species and sampling units. As a measure of

discrimination power, we used AUC values of species-specific predictions, which we then averaged over species. We note that while AUC has often been considered to be a measure of accuracy, it is not so in the statistical meaning of the word “accuracy”: AUC does not compare the predictive point estimate to a corresponding test value. Instead, it measures how well the occurrence probabilities discriminate sampling units to either occupied or empty. As a measure of calibration, we used the mean error between predicted and observed numbers of occurrences in 10 probability bins (each including the same number of sampling units based on quantiles), averaged over species. As a measure of precision, we used the standard deviation of the predicted species occurrence, i.e., the square root of the product of the probability of species presence and the probability of species absence. We averaged precision over species and sampling units.

*Performance measures related to species richness.*—To evaluate predictive performance at the level of species richness, we summed species occurrences separately for each of the 100 replicate matrices, thus producing 100 replicate vectors of predicted species richness for each sampling unit. The measures of accuracy and discrimination power are based on the mean prediction, i.e., the average over the 100 replicate predictions. As a measure of accuracy, we used the square root transformed mean squared error between mean prediction and observed species richness. As a measure of discrimination power, we used the Spearman rank correlation between mean prediction and observed species richness, the correlation being computed among the sampling units. The quantification of calibration was assessed with the relative frequency,  $p$ , of test values within the corresponding predictive 50% central interval and we report  $|p - 0.5|$  so that smaller values indicate higher performance. To assess precision, we calculated the standard deviation of the prediction intervals, and averaged these standard deviations over the sampling units.

*Performance measures related to community composition.*—Using all pairs of sampling units to evaluate predictive performance at the level of community composition would have led to excessive computations. Thus, we selected a random sample of 300 pairs of sampling units. For each of these pairs, we calculated three measures of pairwise community similarity: the Sørensen-based dissimilarity  $\beta_{\text{SOR}}$ , the Simpson-based dissimilarity  $\beta_{\text{SIM}}$ , and the nestedness-resultant dissimilarity  $\beta_{\text{NES}}$  (Baselga 2010). We computed each of these separately for the 100 replicate predictions. We then evaluated the accuracy, discrimination power, calibration, and precision exactly as we did with species richness, but replacing species richness with one of the dissimilarity indices, and by comparing the predicted and observed values over pairs of sampling units rather than over individual sampling units.



*Computing details.*—All analyses were carried out in the R statistical environment (R Core Team 2018) or Matlab (MathWorks 2015). The R and Matlab packages used for model fitting are described in Appendix S1. As the Bayesian models are computationally intensive, we ran the MCMC chains for 50,000 iterations (see Appendix S1). To examine the level of MCMC convergence, we fitted all Bayesian models twice, and computed the correlation among the predicted species occurrence probabilities between the two chains. We note that while MCMC convergence should ideally be examined based on all model parameters, the convergence should be checked at least for the key model parameters to be used in subsequent inference (Gelman et al. 2013). Hence, we chose to base our analyses on predicted occurrence probabilities as that is the primary parameter controlling the performance of models' predictive performance. We note that convergence is an issue also in optimization related to ML estimation. However, we did not check whether the optimization algorithms had found true (global) maxima, but assumed that if optimization stopped before the maximum number of iterations it had reached or was very near the maxima. For calculating the performance measures, we used several packages available in R, details of which can be found from the pipeline used for producing the results (see *Data Availability*; Norberg 2019).

#### *Synthesizing the results*

As described above, we generated 60 predictions (five data sets, three prediction types, two data sizes, two community sizes) for each of 35 model variants (the original 33 and the two ensemble models) and assessed the quality of these predictions by 20 performance measures, resulting in a total of 42,000 performance measure values. To simplify the interpretation of the results, we reversed the signs of the performance measures as needed, so that higher values of the performance measures always corresponded with higher accuracy, greater discrimination power, more accurate calibration, and higher precision. We further standardized each performance measure to have zero mean and unit variance among the SDM variants, separately for each data set and for each prediction task. As some of the models failed completely in some of their predictions, this produced outliers that would have dominated the variation over performance measures, hampering the comparison among the non-failed models. To avoid this effect, we delimited the values of performance measures to a maximum (and minimum) of plus (and minus) two standard deviations. To obtain a single summary of predictive performance at the level of community composition, we averaged the normalized performance measures obtained for  $\beta_{\text{SOR}}$ ,  $\beta_{\text{SIM}}$ , and  $\beta_{\text{NES}}$ , and thus our results involve 12 instead of 20 performance measures. The raw results for all the performance measures are provided in Appendix S3.

To compare the 35 model variants, we first averaged each of the 12 performance measures over the 60 predictions. To obtain an overall measure of performance, we further averaged the nine measures of accuracy, discrimination, and calibration, but excluded the three measures for precision. The reason for this is that while the quality of the predictions unambiguously increases with increasing accuracy, increasing discrimination power, and increasing calibration, the interpretation of precision depends on the accuracy of the predictions (Gneiting and Raftery 2007). If the predictions are accurate, their quality increases with precision. However, if the predictions are not accurate, with increasing precision, the true value will increasingly fall outside the prediction interval, meaning that a high value of precision actually *decreases* the calibration of predictive distributions (as illustrated in the precision panel of Fig. 1D). We selected the best performing variants of the five best performing models based on this overall ranking as a basis of ensemble modeling.

To examine how much ranking among the model variants depends on the type of the data and the prediction task, we also produced rankings separately for different subsets of the data. Specifically, we examined (1) interpolation, partial extrapolation, and full extrapolation; (2) each of the five data sets; (3) small vs. large data sets; and (4) each of the 12 performance measures. Further, to evaluate which model variants and their combinations perform generally well in many kinds of prediction tasks, we examined the performances of the model variants over all of the performance evaluations for the data sets with all species. We classified a model variant as “well performing” in a given performance evaluation if its performance measure exceeded  $\min + 0.9 \times (\max - \min)$ , where min and max were the performance measures of the worst and the best model variant. We computed for each model variant the proportion of the performance evaluations in which it was ranked as well performing. To identify a set of model variants of complementary value, we first selected the model variant that was scored as well performing the highest number of times. We then restricted the analysis to those performance evaluations in which the selected model variant did not perform well, and selected a second model variant that performed well in the highest number of times. We continued iteratively to produce an ordering of model variants out of which at least one model performed well in as many performance evaluations as possible.

To explore the factors influencing predictive performance in more detail we used a multivariate GLM framework (as implemented via HMSC; Ovaskainen et al. 2017) to analyze the results, where we consider the performance measures as response variables, and the properties of the data and the model variants as explanatory variables. We performed this analysis in two ways. In the first analysis, we included the size of the data set and the type of prediction as fixed explanatory variables, and the model variant and the identity of the data set as

random effects. With this analysis, we aimed to examine the variation and covariation (i.e., correlations between the 12 different performance measures) in predictive performance among model variants. In the second analysis, we included the Features A–G (Table 2) used to classify the model variants as additional fixed explanatory variables. We further included the SDM model (i.e., the 15 models that the 33 variants represent, Table 2) as an additional random effect. With this second analysis, we aimed to assess how much of the variation in predictive performance among model variants could be attributed to the modeling framework and in particular to its characteristics, which we included as explanatory variables. To test our hypotheses related to the influence of rare species, we also conducted these analyses basing the performance measures either on all species or only on the common species.

## RESULTS

Based on the overall performance, the five best-performing model variants (including only one from each modeling framework) were HMSC.3, GLM.5, MISTN.1, MARS.1, and GNN.1 (Fig. 2A). The ensemble model ENS.BEST5 consisting of the above mentioned five variants performed worse than HMSC.3 but better than the other four model variants of which it was composed. The ensemble model ENS.ALL performed worse than seven, and better than 26 of the 33 model variants of which it was comprised. The variants of the same models ranked close to each other, with the major exception of GLM, for which some variants performed well but others poorly. When restricting the evaluation of predictive power to the common species (Fig. 2B), the relative performance of some of the models (e.g., BORAL, some of the GLM variants, the BC models and GJAM) increased substantially.

A variance partitioning among the performance measures showed that the properties of the data, the prediction tasks, and the model variant that was applied all strongly influenced predictive performance, whereas the size of the data sets had only a minor effect (Fig. 2C). Averaged over the 12 measures of predictive performance and considering all species, 33% of the explained variance was attributed to the model variant, 38% to the properties of the data, and 29% to whether the prediction task was interpolation, or partial, or full extrapolation (Fig. 2C). When considering only common species, 30% of the explained variance was attributed to the model variant, 49% to the properties of the data, and 21% to whether the prediction task was interpolation, or partial, or full extrapolation. So, predictive performance is influenced by both the model employed and by the predictive goal, as well as by qualities of the available data. The choice of the model variant is especially important for communities with a high proportion of rare species. The measures of accuracy, discrimination and calibration were positively correlated with each

other among the model variants (Fig. 2D). This result suggests that some model variants performed generally well with respect to many performance measures, while others performed generally poorly, justifying the comparison based on overall performance (Fig. 2A, B). In contrast, the measures of precision were positively associated with each other, but negatively associated with some measures of accuracy, discrimination and calibration (Fig. 2D), meaning that those model variants that produced the least uncertain predictions performed otherwise the poorest.

Out of the sources of variation among the model variants, Features A–G explained together 58% (54% if considering common species only) of the variation, the random effect of model (i.e., the 15 models as listed in Table 2) 18% (15%), whereas the remaining 24% (31%) remained as idiosyncratic variation among the model variants. Thus, even if we classified the models with seven different features that we expected to play a major role, one-half of the variation remained unexplained by these. When considering all species, the most important features were Feature F, i.e., whether the model involved shrinkage (35% of all variation attributed to all Features A–G), Feature A, i.e., whether the model was parametric or semi-parametric (23%), and Feature D, i.e., whether the model accounted for species associations (17%), the remaining features explaining only minor parts of the variation. When instead considering common species, Feature F (35%) remained as important, Feature A (17%) was somewhat less important, whereas Feature D became more important (20%).

Regardless of the data set, degree of extrapolation, or data set size, the ranking of the model variants was generally, but not entirely, consistent. Concerning the influence of the data set, perhaps the clearest contrast emerged between the butterfly data, collected at large spatial scale and including a relatively small number of species, and the vegetation data, collected at a small spatial scale and including a large number of rare species. For the butterfly data, the best model was the stacked species distribution model GLM.5, whereas for the vegetation data, the best models were joint species distribution models (Fig. 3A, B). As expected a priori, extrapolation was much more difficult than interpolation (Fig. 3C, D), but in general the same models performed well for both interpolation and for extrapolation. The rankings among the models for other subsets of results, as well as separately for each performance measure, are shown in Appendix S3.

The models that performed well in a large proportion of performance evaluations (Fig. 4A) were generally the same models that achieved the highest average performance scores (Fig. 2A, B), suggesting the robustness of the results. In particular, HMSC.3, which achieved the highest average performance (Fig. 2A, B) and was also most frequently (in 44% of the performance evaluations) classified as a well performing model (Fig. 4A). However, many of the other well performing models

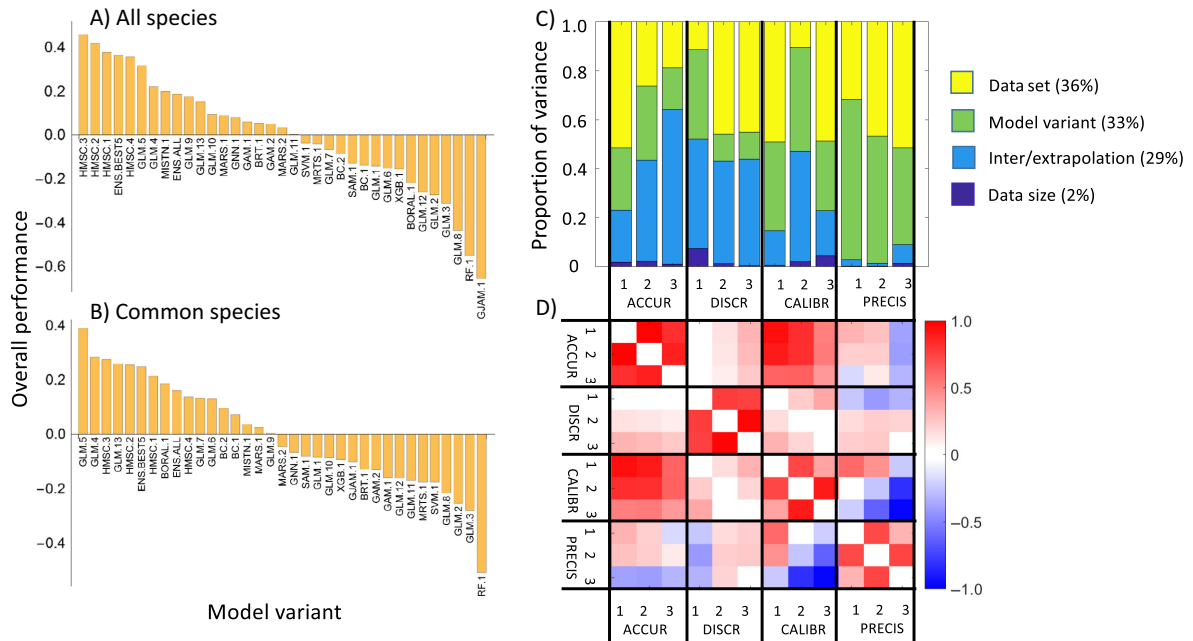


Fig. 2. Variation in predictive performance among model variants. Panels A (based on all species) and B (restricted to species with prevalence at least 0.1) rank the model variants based on their overall predictive performance, i.e., the average among measures of accuracy, discrimination and calibration. Models are described in Table 2. Panel C partitions variation in predictive performance among the properties of the data (data set and data size), type of prediction (interpolation/partial or full extrapolation), and the model variant. Panel D shows correlations among the different measures of predictive performance (accuracy [ACCUR], discrimination [DISCR], calibration [CALIBR], and precision [PRECIS]). Red color refers to positive correlation and blue color to negative correlation, and cases with lower than 75% posterior support for positive or negative association are shown by white. Panels C and D are based on analyses on all species at the levels of species (1), species richness (2) and community composition (3).

performed well in the same cases as did HMSC.3. The model that provided the highest amount of complementarity in its performance was GLM.5 (Fig. 4B), which was also the second best in the variant-specific comparisons (Fig. 4A). The second most complementary model was SAM.1, which was only the 15th best model in the model-variant specific comparisons (Fig. 4A). At least one of the four models HMSC.3, GLM.5, SAM.1 and GLM.12, performed well in 76% of all the evaluation tasks (Fig. 4B).

DISCUSSION

Statistical models cannot mimic the complexity of the real world, but in order to help understand this complexity, a useful model should predict reality as accurately as possible (Burnham et al. 2011, Hand 2014, Houlahan et al. 2017). In the absence of detailed process knowledge, which is the norm for ecological systems (Urban et al. 2016), statistical models such as those we have explored here are essential tools in many applied ecological arenas. Given the wide range of models now available, it is important to provide a degree of guidance to practitioners attempting to apply these models, including an articulation of the limits in model performance.

The differences we have found in the predictive performance among models arise from a large number of factors, including differences in their structural assumptions, their statistical inference frameworks, qualities of the available data sets, and software implementations. The SDM variants that we compared showed consistent variation in their performance, with some performing generally well and others poorly across most data sets, prediction tasks, and measures of predictive performance. This tentatively points to some models as being the initial “go-to” models in analyses of distributional data. Despite this consistency, however, our results do not yield any straightforward explanation of why some models performed better than others, as much of the variation among model variants remained unexplained. In particular, our results failed to give strong support for the hypothesis that the structural model assumptions (Features A–G, Table 2) would explain differences in predictive performance (see *Introduction*). An intriguing question that remains is identifying which model features explain the consistent variation that we observed in predictive performance. As the models simultaneously differ from each other in many aspects, it is difficult, in general, to conclusively pinpoint the causal and inferential reasons for differences in their

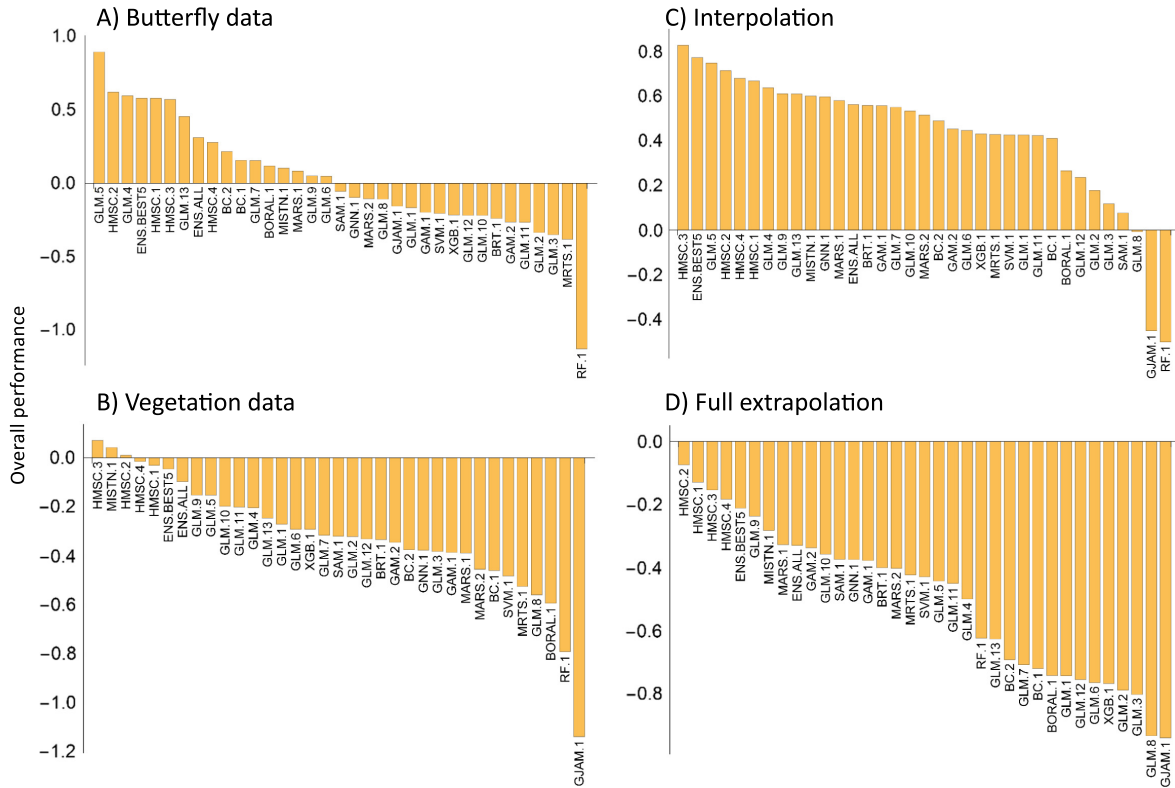


FIG. 3. Variation in predictive performance among data sets and prediction tasks. The panels show the overall performance of the models based on predicting all species (as in Fig. 2A) but evaluated for (A) butterfly data or (B) vegetation data separately, or only for (C) interpolation or (D) full extrapolation tasks. Models are described in Table 2.

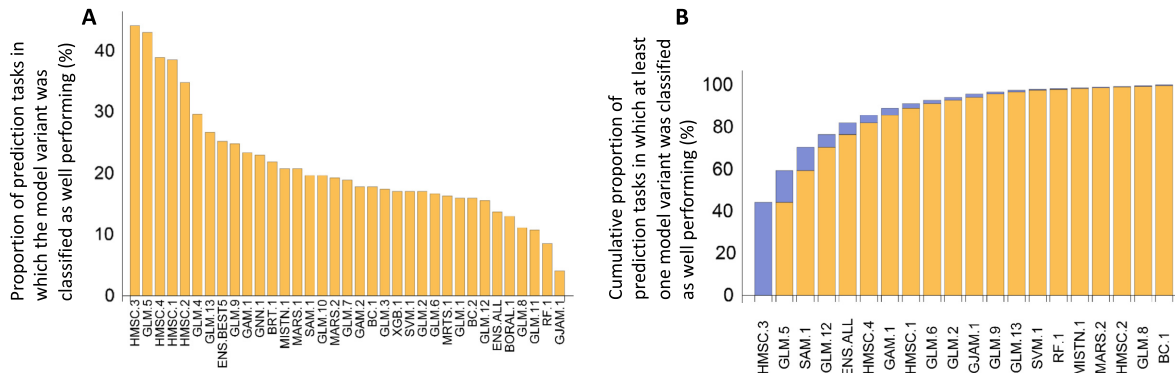


FIG. 4. The proportion of prediction tasks for the case of all species, among which model variants and their combinations performed well. Panel A shows how the proportion of the prediction tasks for which each model variant was classified among the well performing models (see *Synthesizing the results* for how this was defined). Panel B shows the cumulative proportion of prediction tasks among which at least one of the included model variants performed well. In panel B, the model variants were added one by one from left to right, the orange bar shows the proportion achieved by model variants included before the focal one, and the blue bar shows the additional proportion achieved by the focal variant. The model variants were added in the order of the proportion of prediction tasks for which the candidate model was well performing but for which none of the already included models was well performing. Thus, unlike panel A, panel B accounts for complementarity among the prediction tasks. Models are described in Table 2.

performance. However, our study includes specific sets of model variants differing only in single features, and thus it provides suitable cases for comparison. We next

discuss the results on the influence of each model feature, based on such controlled comparisons when possible to do so.



#### *Feature A: parametric vs. semi-parametric models*

In our results, the majority of the best performing models were based on the parametric GLM framework. One reason for the success of parametric models might have been that we considered presence–absence data on species-rich communities that involve a large proportion of rare species. In other situations, such as those involving a large amount of data for a few common species, more flexible semi-parametric models are likely to be more informative (Merow et al. 2014). Further, as discussed above, the model variants differ simultaneously in many aspects, and it is difficult to make controlled comparisons where the only difference would be whether the model is parametric or not. In one such comparison, GLM.1 (parametric) and GAM.1 (semi-parametric) performed roughly equally well, both being in the intermediate category of models.

#### *Feature B: interactions among environmental covariates*

To pinpoint the influence of interactions between environmental covariates, we included three controlled comparisons: between HMSC.4 and HMSC.1, between GLM.12 and GLM.1, and between GLM.13 and GLM.4. The sole difference in each of these comparisons is that the first model variant includes interactions among the environmental predictors while the second variant does not. In all of these comparisons, the models without interactions performed better, suggesting that models including interactions were generally too complex to be estimated with the data considered here. However, for some specific prediction tasks GLM.12 performed well, and we found it to be among the model variants that provided most complementary information after HMSC.3 (Fig. 4B).

#### *Feature C: shared information on environmental responses*

To pinpoint the influence of sharing information among the species, we included the controlled comparison between GLM.4 and HMSC.1. The sole difference between these two models is that while GLM.4 estimates the influence of covariates independently for each species, HMSC.1 shares information among the species. Our results showed that HMSC.1 performed better when all species were considered (Fig. 2A), but with only common species included, GLM.4 performed better (Fig. 2B). This is in line with other recent literature on species distribution modeling showing that assuming shared responses to the environment can improve predictive performance especially for rare species through “borrowing information from other species” (Guisan et al. 1999, Ovaskainen and Soininen 2011, Hui et al. 2013, Madon et al. 2013, Ovaskainen et al. 2016, Tikhonov et al. 2017). Since most ecological communities consist of a few common and many rare species, and given that rare species are often the focus of study in

community-level analyses, particularly those with a conservation bent (Aizen et al. 2012, Moullot et al. 2013), we expect the assumption of joint responses to be generally beneficial in community ecology studies. The concept of “shared responses to environmental covariates” can be incorporated in many different ways. For example, HMSC assumes that the species-specific regression parameters are sampled from a multivariate normal distribution, whereas SAM classifies them into distinct groups. As HMSC and SAM also differ in many other aspects, it is difficult to resolve whether the difference in model performance relates to how shared responses are modeled or instead to how the models are implemented.

#### *Feature D: species co-occurrences*

To pinpoint the influence of accounting for species co-occurrences, we included the three controlled comparisons between HMSC.2 and HMSC.1, between BC.2 and BC.1, and between BORAL.1 and GLM.7. In each of these comparisons, the principal difference is that the first of the variant pairs accounts for residual species-to-species associations, while the second does not. A general comparison (Fig. 2A, B) among these models supports the hypothesis that models that account for statistical non-independence among species have better predictive performance, except that GLM.7 performed better than BORAL.1 for the case that included all species. However, compared to sharing information among the species on their responses to covariates (Feature C), accounting for residual co-occurrences (Feature D) provided only a minor improvement (HMSC.2 performed only a little better than did HMSC.1, which, in turn, performed better than GLM.4 when all species were included).

It is important to note that our evaluation of model performance entailed generating predictions for new sampling units, in which the occurrences of all species were unknown. However, if one knows the occurrences of some of the species at the validation sites, it is possible to improve predictions for other species by including potentially interacting species as predictors (Araújo and Luoto 2007, Heikkinen et al. 2007, Wisz et al. 2013, Mod et al. 2015, but see Godsoe et al. 2016), or by using joint species distribution models to predict occurrences of a target species conditional on the occurrences of all other species (Ovaskainen et al. 2017). This suggests that in other kinds of prediction tasks, the utility of including species-to-species associations can be greater. That models that account for associations produce better predictions could be either due to species having real ecological interactions with each other, or to unrecognized environmental covariates not included in the model (Pollock et al. 2014, Ovaskainen et al. 2017).

#### *Feature E: spatial vs. non-spatial models*

To pinpoint the influence of including spatial predictors, we included the controlled comparison between

HMSC.3 and HMSC.2, between GLM.5 and GLM.4, between GLM.3 and GLM.2, and between GAM.2 and GAM.1. The sole difference in each of these comparisons is that the first model variant includes an explicit spatial structure while the second variant does not. In our overall evaluation (Fig. 2A), the spatial models performed better in two comparisons (HMSC.3 vs. HMSC.2 and GLM.5 vs. GLM.4), whereas the non-spatial model performed better in the other two comparisons (GLM.3 vs. GLM.2 and GAM.2 vs. GAM.1). Results were similar for the case of common species, except that GAM.2 outperformed GAM.1 (Fig. 2B). Thus, Bayesian methods tended to improve when spatial effects were added, whereas ML methods did not, suggesting that the inclusion of prior information (even if weak) was important for the proper estimation of spatial structure, especially when also the rare species are included.

The result that spatial structure increased performance for some models is in line with previous studies on single-species SDMs highlighting the importance of accounting for spatial autocorrelation (Dormann et al. 2007, Record et al. 2013, Crase et al. 2014). As discussed in previous studies, this is because dispersal processes, historical contingencies, and missing covariates (Foster et al. 2012) generate spatial variation in species communities (Bokma et al. 2001, Fernando et al. 2007, Kessler 2009). Although the degree to which dispersal and historical processes influence species occurrences might vary depending on the community type or spatial coverage of the study (Record et al. 2013), including a spatially structured random effect is recommended so as not to violate the assumption of independence among sampling units (and consequently overestimating confidence in ecological inferences or in model predictions). However, the utility of spatial information depends also on the prediction task: for the case of full extrapolation, the non-spatial HMSC.2 actually performed somewhat better than the spatial HMSC.3 (Fig. 3C, D), as can be expected from the grounds that the use of spatial information is especially useful for making predictions for sampling units near the training data.

#### *Feature F: shrinkage*

To identify the influence of shrinkage, we may compare GLM.1 to GLM.4 and GLM.6 to GLM.9. In these comparisons, the latter model variant includes shrinkage, whereas the former one does not. However, we note that GLM.1 and GLM.4 differ also in how parameter uncertainty is accounted for in the prediction, the influence of which is discussed below. In our results, GLM.4 (fitted in the Bayesian framework) was among the best performing model variants, whereas GLM.1 (fitted in the ML framework) showed average performance. As adding parameter uncertainty to ML models decreased their performance (see *Feature G: parameter uncertainty*), we attribute the superior

performance of GLM.4 specifically to the influence of the Bayesian prior and thus to shrinkage. In GLM.4, the prior shrinks the regression parameters toward zero, thus restricting the effect sizes of the environmental covariates.

Consistent with the comparison between GLM.4 and GLM.1, we found that GLM.9 (with shrinkage through penalized likelihood) performed better than GLM.6 (which does not involve shrinkage) for the case of all species (Fig. 2A), but the opposite was found for common species (Fig. 2B). As data on rare species inherently have limited potential to estimate parameters, the inclusion of shrinkage can indeed be expected to make a major difference. GLM.10 (with shrinkage through penalized likelihood) performed worse in our overall evaluation (Fig. 2A, B) than GLM.9, suggesting that the way in which shrinkage is implemented can make a difference.

#### *Feature G: parameter uncertainty*

To pinpoint the influence of accounting for parameter uncertainty in making predictions, we included the two controlled comparisons of GLM.8 vs. GLM.1 and GLM.11 vs. GLM.10. In these comparisons the model variants are otherwise identical, except that, when making predictions, GLM.8 and GLM.11 account for parameter uncertainty using the standard asymptotic distribution approximation (Foster and Dunstan 2010), while GLM.1 and GLM.10 use only ML estimate. In both cases, we found the model variant that was based on point estimates to perform generally better (Fig. 2A, B) than the one that accounted for uncertainty using the asymptotic distribution approximation. However, we note that GLM.4, which is the Bayesian version of GLM.1, generally performed well (Fig. 2A), especially when considering only common species (Fig. 2B). This is again likely to be related to the fact that our data comprised of species-rich communities large species communities containing many rare species. In this case, the asymptotic approximations might not work well for finite samples, and the disparity seems to be an over-estimation of the uncertainty, making the predictions uninformative. On the other hand, the uncertainty estimate in GLM.4 is based on a Bayesian joint posterior distribution and moreover, but unlike GLM.8, it also includes shrinkage.

#### *Other factors affecting model performance*

While the comparisons discussed above in the context of the Features A–G yielded results that were largely consistent with our hypotheses, the overall comparison among the models showed a large amount of idiosyncratic, unexplained variation in model performance. One source of such variation is that while we attempted to optimize the performance of each individual model, doing so was more challenging for some models than for

others. All models used in this study were implemented in freely available software, but these packages varied in their level of documentation and the amount and transparency of the user-defined tuning parameters. One reason for the popularity of modeling frameworks such as GAM, GLM, and MARS might simply be the relative availability of their user-friendly and well-documented software, and that they are computationally efficient. One important further difference among the models, which we have not explored in this study, is that additional data types could be incorporated in some of the modeling frameworks, which could have improved their predictive performance. For example, including species traits can both bring more ecological insight (McGill et al. 2006) as well as improve predictive performance (Brown et al. 2014). Only some of the models have the capacity to incorporate traits directly, and thus we did not include traits in these analyses so as to keep the results more comparable among the models. Another direction is to tie SDMs more directly to models of community dynamics with strongly interacting species. In some cases (e.g., specialist herbivores tracking their required host plants, or generalist predators constraining the distribution of vulnerable prey), there can be large-scale distributional imprints of locally strong interactions (Gilman et al. 2010, Godsoe et al. 2017).

Previous studies have shown that one of the main sources of variation in SDM performance is the structure of the data (Fielding and Haworth 1995), especially the prevalence of species (Leathwick et al. 2006, Meynard and Quinn 2007, Syphard and Franklin 2009, Santika 2011, Madon et al. 2013) and the strength and shape of the environmental gradient (Thuiller et al. 2003, Austin et al. 2006, Santika and Hutchinson 2009, Hoffman et al. 2010, Santika 2011). Consistent with this, our results demonstrate that the specific data set studied has a major impact on predictive performance, as well as the type of prediction task. In particular, our results pinpoint the difficulty of extrapolative predictions, which has direct implications for model transferability across systems, space and time (Wenger and Olden 2012, Owens et al. 2013). Furthermore, a detailed inspection of the results (Appendix S3) shows that the rank order of the models differs considerably with respect to the measure used for evaluating their performance. This is for instance illustrated by the fact that even the generally best performing model variant (HMSC.3) belonged to the well-performing models in only 44% of the evaluation tasks, and applying just this model means it would perform substantially less well in 56% of the cases than some other models. Thus, it is important for the researcher to evaluate which aspect of model performance is especially critical given the aim of the modeling. For example, if the goal is to predict the probability that a focal species is present in a site, or the expected species richness in a site, or the expected level of beta-diversity between a pair of sites, then measures of accuracy are likely to be the most relevant criterion. If

the goal instead is to prioritize sites in terms of their species occurrences, species richness, or community composition, then measures of discrimination are likely to be the most relevant. If the goal is to make statements about prediction uncertainty, e.g., whether the predicted species occurrence probabilities are reliable, or whether the uncertainty estimates involved in predictions of species richness or community composition are valid, then measures of calibration are likely to be important. In theory, measures of precision would be relevant if one wishes to minimize uncertainty, but, as we have shown, the models that involve the least uncertainty in their predictions tend to behave badly with respect to the other measures of performance.

Overall, our analyses show that there is considerable variation in performance among models, and that it may be difficult to predict a priori which kinds of model features do, or do not, improve model performance. Which model works best will not only depend on how the assumptions of the model relate to the assembly processes shaping a particular community, but also on other characteristics such as the amount, quality, and spatial structure of the data. Two data sets, even with apparently similar characteristics, might be best modeled by different methods (James et al. 2013). A general strategy that we recommend is to apply at least a few alternative models, and use cross-validation or other model selection approaches to assess critically how well the models predict the aspects of the data that are relevant, given the aims of the study. Based on our results on model complementarity (Fig. 4B), including, e.g., model variants HMSC.3, GLM.5, SAM.1, and GLM.12 among the set of the candidate models is likely to lead to a good result, in the sense that at least one of these models will perform almost as well as any of the 35 model variants considered here. The results of the cross-validation exercise will then tell which of these models is to be trusted most. The recommendation of using these specific models as the set of candidate models is of course conditional on the data and the prediction tasks being similar to those considered here: presence-absence data on large ecological communities with many rare species. We hope that our results provide a helpful starting point for researchers applying species distribution modeling in community ecology, both in terms of gauging the potential pitfalls and advantages in the models available to choose from among, and in defining the characteristics of the predictions that they may wish to validate.

#### ACKNOWLEDGMENTS

This work was funded by the Research Foundation of the University of Helsinki (A. Norberg), the Academy of Finland (CoE grant 284601 and grant 309581 to O. Ovaskainen, grant 308651 to N. Abrego, grant 1275606 to A. Lehtikoinen), the Research Council of Norway (CoE grant 223257), the Jane and Aatos Erkkö Foundation, and the Ministry of Science, Innovation and Universities (grant CGL2015-68438-P to M. B. Araújo). Contributions were also made by members of the Biotic Interactions Working Group (<http://www.nimbios>).

org/workinggroups/WG\_biotic\_interactions) at the National Institute for Mathematical and Biological Synthesis, sponsored by the National Science Foundation, the U.S. Dept. of Homeland Security, and the U.S. Department of Agriculture through NSF Awards #EF-0832858 and #DBI-1300426, with additional support from The University of Tennessee, Knoxville. We would like to thank the thousands of researchers and volunteers who have contributed to the data sets utilized in this study. In particular, we thank Åke Lindström for making Swedish bird monitoring data available to us. The bird surveys in Sweden were supported by grants from the Swedish Environmental Protection Agency and carried out in collaboration with all 21 County Administrative Boards of Sweden, and the bird surveys in Norway were supported by the Norwegian Environment Agency.

## LITERATURE CITED

- Aizen, M. A., M. Sabatino, and J. M. Tylianakis. 2012. Specialization and rarity predict nonrandom loss of interactions from mutualist networks. *Science* 355:1486–1489.
- Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43:1223–1232.
- Araújo, M. B., F. Ferri-Yáñez, F. Bozinovic, P. A. Marquet, F. Valladares, and S. L. Chown. 2013. Heat freezes niche evolution. *Ecology Letters* 16:1206–1219.
- Araújo, M. B., and A. Guisan. 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography* 33:1677–1688.
- Araújo, M. B., and M. Luoto. 2007. The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography* 16:743–753.
- Araújo, M. B., R. G. Pearson, W. Thuiller, and M. Erhard. 2005. Validation of species – climate impact models under climate change. *Global Change Biology* 11:1504–1513.
- Araújo, M. B., R. Rozenfeld, C. Rahbek, and P. A. Marquet. 2011. Using species coexistence networks to assess the impacts of climate change. *Ecography* 34:897–908.
- Asher, J., M. Warren, R. Fox, P. Harding, G. Jeffcoate, and S. Jeffcoate. 2001. *The millennium atlas of butterflies in Britain and Ireland*. Oxford University Press, Oxford, UK.
- Augustin, N. H., M. A. Muggleston, and S. T. Buckland. 1996. An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology* 33:339–347.
- Austin, M. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* 157:101–118.
- Austin, M., L. Belbin, J. A. Meyers, M. D. Doherty, and M. Luoto. 2006. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. *Ecological Modelling* 199:197–216.
- Austin, M. P., T. M. Smith, K. P. Van Niel, and A. B. Wellington. 2009. Physiological responses and statistical models of the environmental niche: a comparative study of two co-occurring *Eucalyptus* species. *Journal of Ecology* 97:496–507.
- Bahn, V., and B. J. McGill. 2013. Testing the predictive performance of distribution models. *Oikos* 122:321–331.
- Baselga, A. 2010. Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography* 19:134–143.
- Baselga, A., and M. B. Araújo. 2009. Individualistic vs. community modelling of species distributions under climate change. *Ecography* 32:55–65.
- Baselga, A., and M. B. Araújo. 2010. Do community-level models describe community variation effectively? *Journal of Biogeography* 37:1842–1850.
- Beale, C. M., and J. J. Lennon. 2012. Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B* 367:247–258.
- Bio, A. M. F., R. Alkemade, and A. Barendregt. 1998. Determining alternative models for vegetation response analysis: a non-parametric approach. *Journal of Vegetation Science* 9:5–16.
- Bokma, F., J. Bokma, and M. Mönkkönen. 2001. Random processes and geographic species richness patterns: Why so few species in the north? *Ecography* 24:43–49.
- Bonthoux, S., A. Baselga, and G. Balent. 2013. Assessing community-level and single-species models predictions of species distributions and assemblage composition after 25 years of land cover change. *PLoS ONE* 8:e54179.
- Boulangeat, I., D. Gravel, and W. Thuiller. 2012. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology Letters* 15:584–593.
- Brown, A. M., D. I. Warton, N. R. Andrew, M. Binns, G. Casalis, and H. Gibb. 2014. The fourth-corner solution—using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution* 5:344–352.
- Burnham, K. P., D. R. Anderson, and K. P. Huyvaert. 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology* 65:23–35.
- Cazelles, K., M. B. Araújo, N. Mouquet, and D. Gravel. 2016. A theory for species co-occurrence in interaction networks. *Theoretical Ecology* 9:39–48.
- Chapman, D. S., and B. V. Purse. 2011. Community versus single-species distribution models for British plants. *Journal of Biogeography* 38:1524–1535.
- Chen, T., et al. 2018. xgboost: extreme gradient boosting. R package version 0.71.2. <https://CRAN.R-project.org/package=xgboost>
- Clark, J. S. 2005. Why environmental scientists are becoming Bayesians. *Ecology Letters* 8:2–14.
- Clark, J. S., A. E. Gelfand, C. W. Woodall, and K. Zhu. 2014. More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications* 24:990–999.
- Clark, J. S., D. Nemergut, B. Seyednasrollah, P. Turner, and S. Zhang. 2017. Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. *Ecological Monographs* 87:34–56.
- Crane, B., A. Liedloff, P. A. Vesik, Y. Fukuda, and B. A. Wintle. 2014. Incorporating spatial autocorrelation into species distribution models alters forecasts of climate-mediated range shifts. *Global Change Biology* 20:2566–2579.
- Crookston, N. L., and A. O. Finley. 2008. VaImpute: an R package for kNN imputation. *Journal of Statistical Software* 23:1–16.
- D’Amen, M., J. N. Pradervand, and A. Guisan. 2015. Predicting richness and composition in mountain insect communities at high resolution: a new test of the SESAM framework. *Global Ecology and Biogeography* 24:1443–1453.
- D’Amen, M., C. Rahbek, N. E. Zimmermann, and A. Guisan. 2017. Spatial predictions at the community level: from current approaches to future frameworks. *Biological Reviews* 92:169–187.
- De’Ath, G., T. M. Therneau, B. Atkinson, B. Ripley, and J. Oksanen. 2014. Mvpart: multivariate partitioning. R package version 1.6-2. <https://CRAN.R-project.org/package=mvpart>



- Diggle, P., and P. Ribeiro. 2007. Model-based geostatistics. Springer, Berlin, Germany.
- Diniz-Filho, J. A. F., L. M. Bini, and B. A. Hawkins. 2003. Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography* 12:53–64.
- Dormann, C. F. 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography* 16:129–138.
- Dormann, C. F., et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30:609–628.
- Dormann, C. F., et al. 2012. Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography* 39:2119–2131.
- Dubuis, A., J. Pottier, V. Rion, L. Pellissier, J.-P. Theurillat, and A. Guisan. 2011. Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches. *Diversity and Distributions* 17:1122–1131.
- Dunstan, P. K., S. D. Foster, and R. Darnell. 2011. Model based grouping of species across environmental gradients. *Ecological Modelling* 222:955–963.
- Elith, J., and C. H. Graham. 2009. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. *Ecography* 32:66–77.
- Elith, J., and J. R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* 40:677–697.
- Elith, J., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29:129–151.
- Ellison, A. M. 2004. Bayesian inference in ecology. *Ecology Letters* 7:509–520.
- Fernando, T., J. Alexandre, F. D. Filho, K. Robert, and T. F. L. V. B. Rangel. 2007. Species richness and evolutionary niche dynamics: a spatial pattern-oriented simulation experiment. *American Naturalist* 170:602–616.
- Ferrier, S., and A. Guisan. 2006. Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology* 43:393–404.
- Fieberg, J., J. Matthiopoulos, M. Hebblewhite, M. S. Boyce, and J. L. Frair. 2010. Correlation and studies of habitat selection: Problem, red herring or opportunity? *Philosophical Transactions of the Royal Society B* 365:2233–2244.
- Fielding, A. H., and P. F. Haworth. 1995. Testing the generality of bird-habitat models. *Conservation Biology* 9:1466–1481.
- Foster, S. D., and P. K. Dunstan. 2010. The analysis of biodiversity using rank abundance distributions. *Biometrics* 66:186–195.
- Foster, S. D., H. Shimadzu, and R. Darnell. 2012. Uncertainty in spatially predicted covariates: Is it ignorable? *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61:637–652.
- Franklin, J. 1998. Predicting the distributions of shrub species in southern California from climate and terrain-derived variables. *Journal of Vegetation Science* 9:733–748.
- Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33:1–22.
- García-Callejas, D., and M. B. Araújo. 2016. The effects of model and data complexity on predictions from species distributions models. *Ecological Modelling* 326:4–12.
- Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. Bayesian data analysis. Third edition. Chapman and Hall/CRC, New York, New York, USA.
- Gilman, S. E., M. C. Urban, J. Tewksbury, G. W. Gilchrist, and R. D. Holt. 2010. A framework for community interactions under climate change. *Trends in Ecology and Evolution* 25:325–331.
- Gneiting, T., and A. E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102:359–378.
- Godsoe, W., J. Franklin, and F. G. Blanchet. 2016. Effects of biotic interactions on modeled species' distribution can be masked by environmental gradients. *Ecology and Evolution* 7:654–664.
- Godsoe, W., J. Jankowski, R. D. Holt, and D. Gravel. 2017. Integrating biogeography with contemporary niche theory. *Trends in Ecology and Evolution* 32:488–499.
- Golding, N., and D. J. Harris. 2015. BayesComm: Bayesian community ecology analysis. R package version 0.1-2. <https://cran.r-project.org/package=BayesComm>
- Götzenberger, L., et al. 2012. Ecological assembly rules in plant communities—approaches, patterns and prospects. *Biological Reviews* 87:111–127.
- Guillera-Arroita, G. 2017. Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography* 40:281–295.
- Guillera-Arroita, G., J. J. Lahoz-Monfort, and J. Elith. 2014. Maxent is not a presence-absence method: a comment on Thibaud *et al.* *Methods in Ecology and Evolution* 5:1192–1197.
- Guisan, A., T. C. Edwards, and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* 157:89–100.
- Guisan, A., A. Lehmann, S. Ferrier, M. Austin, J. M. C. Overton, R. Aspinall, and T. Hastie. 2006. Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology* 43:386–392.
- Guisan, A., and C. Rahbek. 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.
- Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. *Ecology Letters* 8:993–1009.
- Guisan, A., S. B. Weiss, and A. D. Weiss. 1999. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology* 143:107–122.
- Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* 135:147–186.
- Guisan, A., N. E. Zimmermann, J. Elith, C. H. Graham, S. J. Phillips, and A. T. Peterson. 2007a. What matters for predicting the occurrences of trees: Techniques, data, or species' characteristics? *Ecological Monographs* 77:615–630.
- Guisan, A., et al. 2007b. Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions* 13:332–340.
- Hand, D. J. 2014. Wonderful examples, but let's not close our eyes. *Statistical Science* 29:98–100.
- Harpole, W. S., et al. 2011. Nutrient co-limitation of primary producer communities. *Ecology Letters* 14:852–862.
- Harris, D. J. 2015. Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution* 6:465–473.
- Harris, D. J., S. Taylor, and E. P. White. 2018. Forecasting biodiversity in breeding birds using best practices. *PeerJ* 6:e4278.
- Heikkinen, R. K., M. Luoto, R. Virkkala, R. G. Pearson, and J.-H. Körber. 2007. Biotic interactions improve prediction of

- boreal bird distributions at macro-scales. *Global Ecology and Biogeography* 16:754–763.
- Hijmans, R. J., S. J. Phillips, J. R. Leathwick, and J. Elith. 2017. Dismo: Species distribution modeling. R package version 1.1-4. <https://CRAN.R-project.org/package=dismo>
- Hoffman, J. D., N. Aguilar-Amuchastegui, and A. J. Tyre. 2010. Use of simulated data from a process-based habitat model to evaluate methods for predicting species occurrence. *Ecography* 33:656–666.
- Houlahan, J. E., S. T. Mckinney, T. M. Anderson, and B. J. McGill. 2017. The priority of prediction in ecological understanding. *Oikos* 126:1–7.
- Hui, F. K. C. 2017. Boral: Bayesian ordination and regression analysis. R package version 1.4. <https://CRAN.R-project.org/package=boral>
- Hui, F. K. C., D. I. Warton, S. D. Foster, and P. K. Dunstan. 2013. To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. *Ecology* 94:1913–1919.
- James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. An introduction to statistical learning with applications in R. Springer International Publishing, New York, New York, USA.
- Jiménez-Valverde, A., and J. M. Lobo. 2007. Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecologica* 31:361–369.
- Kessler, M. 2009. The impact of population processes on patterns of species richness: lessons from elevational gradients. *Basic and Applied Ecology* 10:295–299.
- Kissling, W. D., et al. 2012. Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography* 39:2163–2178.
- Latimer, A. M., S. Wu, A. E. Gelfand, and J. A. Silander, Jr. 2006. Building statistical models to analyze species distributions. *Ecological Applications* 16:33–50.
- Lawson, C. R., J. A. Hodgson, R. J. Wilson, and S. A. Richards. 2014. Prevalence, thresholds and the performance of presence-absence models. *Methods in Ecology and Evolution* 5:54–64.
- Leathwick, J. R., and M. P. Austin. 2001. Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology* 82:2560–2573.
- Leathwick, J. R., J. Elith, and T. Hastie. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling* 199:188–196.
- Lek, S., M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga, and S. Aulagnier. 1996. Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling* 90:39–52.
- Liaw, A., and M. Wiener. 2002. Classification and regression by random forest. *R News* 2:18–22.
- Lindström, Å., M. Green, M. Husby, J. A. Kålås, and A. Lehtikoinen. 2015. Large-scale monitoring of waders on their boreal and arctic breeding grounds in northern Europe. *Ardea* 103:3–15.
- Liu, C., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28:385–393.
- Loiselle, B. A., C. A. Howell, C. H. Graham, J. M. Goerck, T. Brooks, K. G. Smith, and P. H. Williams. 2003. Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology* 17:1591–1600.
- Madon, B., D. I. Warton, and M. B. Araújo. 2013. Community-level vs species-specific approaches to model selection. *Ecography* 36:1291–1298.
- Maggini, R., A. Lehmann, N. E. Zimmermann, and A. Guisan. 2006. Improving generalized regression analysis for the spatial prediction of forest communities. *Journal of Biogeography* 33:1729–1749.
- Maguire, K. C., D. Nieto-Lugilde, J. L. Blois, M. C. Fitzpatrick, J. W. Williams, S. Ferrier, and D. J. Lorenz. 2016. Controlled comparison of species- and community-level models across novel climates and communities. *Proceedings of the Royal Society B* 283:1–10.
- Mäkinen, J., and J. Vanhatalo. 2018. Hierarchical Bayesian model reveals the distributional shifts of Arctic marine mammals. *Diversity and Distributions* 24:1381–1394.
- Manel, S., J. Dias, S. Buckton, and S. Ormerod. 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river. *Journal of Applied Ecology* 36:734–747.
- Marmion, M., M. Parviainen, M. Luoto, R. K. Heikkinen, and W. Thuiller. 2009. Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions* 15:59–69.
- Mastorillo, S., S. Lek, F. Dauba, and A. Belaud. 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. *Freshwater Biology* 38:237–246.
- McGill, B. J., B. J. Enquist, E. Weiher, and M. Westoby. 2006. Rebuilding community ecology from functional traits. *Trends in Ecology and Evolution* 21:178–185.
- Meentemeyer, R. K., A. Moody, and J. Franklin. 2001. Landscape-scale patterns of shrub-species abundance in California chaparral. *Plant Ecology* 156:19–41.
- Meier, E. S., T. C. Edwards, F. Kienast, M. Dobbertin, and N. E. Zimmermann. 2011. Co-occurrence patterns of trees along macro-climatic gradients and their potential influence on the present and future distribution of *Fagus sylvatica* L. *Journal of Biogeography* 38:371–382.
- Merow, C., M. J. Smith, T. C. Edwards, A. Guisan, S. M. McMahon, S. Normand, W. Thuiller, R. O. Wüest, N. E. Zimmermann, and J. Elith. 2014. What do we gain from simplicity versus complexity in species distribution models? *Ecography* 37:1267–1281.
- Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. 2017. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. <https://cran.r-project.org/package=e1071>
- Meynard, C. N., and J. F. Quinn. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography* 34:1455–1469.
- Milborrow, S. 2017. Earth: Multivariate adaptive regression splines. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. R package version 4.5.1. <https://cran.r-project.org/package=earth>
- Miller, J. A. 2012. Species distribution models: spatial autocorrelation and non-stationarity. *Progress in Physical Geography* 36:681–692.
- Miller, J. A. 2014. Virtual species distribution models: using simulated data to evaluate aspects of model performance. *Progress in Physical Geography* 38:117–128.
- Mod, H. K., P. C. le Roux, A. Guisan, and M. Luoto. 2015. Biotic interactions boost spatial models of species richness. *Ecography* 38:913–921.
- Moisen, G. G., and T. S. Frescino. 2002. Comparing five modelling techniques for prediction forest characteristics. *Ecological Modelling* 157:209–225.
- Morin, X., D. Viner, and I. Chuine. 2008. Tree species range shifts at a continental scale: new predictive insights from a process-based mod. *Journal of Ecology* 96:784–794.

- Moullot, D., et al. 2013. Rare species support vulnerable functions in high-diversity ecosystems. *PLoS Biology* 11:1–11.
- Nieto-Lugilde, D., K. C. Maguire, J. L. Blois, J. W. Williams, and M. C. Fitzpatrick. 2018. Multiresponse algorithms for community-level modelling: review of theory, applications, and comparison to species distribution models. *Methods in Ecology and Evolution* 9:834–848.
- Niittynen, P., and M. Luoto. 2017. The importance of snow in species distribution models of arctic vegetation. *Ecography* 40:1–13.
- Norberg, A. 2019. *aminorberg/SDM-comparison*: Norberg et al. (2019) (Version publication). <https://doi.org/10.5281/zenodo.2637812>
- Normand, S., U. A. Treier, C. Randin, P. Vittoz, A. Guisan, and J. C. Svenning. 2009. Importance of abiotic stress as a range-limit determinant for European plants: insights from species responses to climatic gradients. *Global Ecology and Biogeography* 18:437–449.
- Olden, J. D., and D. A. Jackson. 2002. A comparison of statistical approaches for modelling fish species distributions. *Freshwater Biology* 47:1976–1995.
- Ovaskainen, O., N. Abrego, P. Halme, and D. B. Dunson. 2016. Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution* 7:549–555.
- Ovaskainen, O., A. Gleb Tikhonov, F. G. Norberg, L. Blanchet, D. B. Duan, T. R. Dunson, and N. Abrego. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters* 2:561–576.
- Ovaskainen, O., and J. Soininen. 2011. Making more out of sparse data: hierarchical modeling of species communities. *Ecology* 92:289–295.
- Owens, H. L., et al. 2013. Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecological Modelling* 263:10–18.
- Pásztor, L., Z. Botta-Dukát, G. Magyar, T. Czárán, and G. Meszéna. 2016. *Theory based ecology: a Darwinian approach*. Oxford University Press, Oxford, UK.
- Pearson, R. G., W. Thuiller, M. B. Araújo, E. Martínez-Meyer, L. Brotons, C. McClean, L. Miles, P. Segurado, T. P. Dawson, and D. C. Lees. 2006. Model based uncertainty in species range prediction. *Journal of Biogeography* 33:1704–1711.
- Pellissier, L., K. A. Bråthen, J. Pottier, C. F. Randin, P. Vittoz, A. Dubuis, N. G. Yoccoz, T. Alm, N. E. Zimmermann, and A. Guisan. 2010. Species distribution models reveal apparent competitive and facilitative effects of a dominant species on the distribution of tundra plants. *Ecography* 33:1004–1014.
- Peterson, A. T., M. Papes, and M. Eaton. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. *Ecography* 30:550–560.
- Peterson, A. T., J. Soberón, R. G. Pearson, R. P. Anderson, E. Martínez-Meyer, M. Nakamura, and M. B. Araújo. 2011. *Ecological niches and geographical distributions: a modeling perspective*. Monographs in population biology. Princeton University Press, Princeton, New Jersey, USA.
- Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O’Hara, K. M. Parris, P. A. Vesik, and M. A. McCarthy. 2014. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). *Methods in Ecology and Evolution* 5:397–406.
- R Core Team. 2018. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Randin, C. F., T. Dirnböck, S. Dullinger, N. E. Zimmermann, M. Zappa, and A. Guisan. 2006. Are niche-based species distribution models transferable in space? *Journal of Biogeography* 33:1689–1703.
- Record, S., M. C. Fitzpatrick, A. O. Finley, S. Veloz, and A. M. Ellison. 2013. Should species distribution models account for spatial autocorrelation? A test of model projections across eight millennia of climate change. *Global Ecology and Biogeography* 22:760–771.
- Ridgeway, G. 2017. *Gbm*: Generalized boosted regression models. R package version 2.1.3. <https://cran.r-project.org/package=gbm>
- Roberts, D. R., et al. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40:913–929.
- Santika, T. 2011. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. *Global Ecology and Biogeography* 20:181–192.
- Santika, T., and M. F. Hutchinson. 2009. The effect of species response form on species distribution model prediction and inference. *Ecological Modelling* 220:2365–2379.
- Segurado, P., and M. B. Araújo. 2004. An evaluation of methods for modelling species distributions. *Journal of Biogeography* 31:1555–1568.
- Sor, R., Y. S. Park, P. Boets, P. L. Goethals, and S. Lek. 2017. Effects of species prevalence on the performance of predictive models. *Ecological Modelling* 354:11–19.
- Stephens, A., and A. A. MacCall. 2004. Multispecies approach to subsetting logbook data for purposes of estimating CPUE. *Fisheries Research* 70:299–310.
- Syphard, A. D., and J. Franklin. 2009. Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography* 32:907–918.
- Talluto, M. V., et al. 2016. Cross-scale integration of knowledge for predicting species ranges: a metamodeling framework. *Global Ecology and Biogeography* 25:238–249.
- The MathWorks Inc. 2015. *MATLAB and statistics toolbox release 2015a*. MathWorks, Natick, Massachusetts, USA.
- Thuiller, E., B. Petitpierre, O. Broennimann, A. C. Davison, and A. Guisan. 2014. Measuring the relative effect of factors affecting species distribution model predictions. *Methods in Ecology and Evolution* 5:947–955.
- Thorson, J. T., M. D. Scheuerell, A. O. Shelton, K. E. See, H. J. Skaug, and K. Kristensen. 2015. Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution* 6:627–637.
- Thuiller, W. 2004. Patterns and uncertainties of species’ range shifts under climate change. *Global Change Biology* 10:2020–2027.
- Thuiller, W., M. B. Araújo, and S. Lavorel. 2003. Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science* 14:669–680.
- Thuiller, W., T. Münkemüller, S. Lavergne, and D. Moullot. 2013. A road map for integrating eco-evolutionary processes into biodiversity models. *Ecology Letters* 16:94–105.
- Tikhonov, G., N. Abrego, D. B. Dunson, and O. Ovaskainen. 2017. Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods in Ecology and Evolution* 8:443–452.
- Umaña, M. N., C. Zhang, M. Cao, L. Lin, and N. G. Swenson. 2017. A core-transient framework for trait-based community ecology: an example from a tropical tree seedling community. *Ecology Letters* 20:619–628.
- Urban, M. C., et al. 2016. Improving the forecast for biodiversity under climate change. *Science* 353:aad8466.

- Vayssières, M. P., R. E. Plant, and B. H. Allen-Diaz. 2000. Classification trees: an alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science* 11:679–694.
- Vellend, M. 2010. Conceptual synthesis in community ecology. *Quarterly Review of Biology* 85:183–206.
- Venables, W. N., and B. D. Ripley. 2002. *Modern applied statistics with S*. Fourth edition. Springer International Publishing, New York, New York, USA.
- Wang, Y., U. Naumann, S. T. Wright, and D. I. Warton. 2012. Mvabund—an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* 3:471–474.
- Warton, D. I., F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. C. Hui. 2015. So many variables: joint modeling in community ecology. *Trends in Ecology and Evolution* 30:766–779.
- Weiher, E., D. Freund, T. Buntton, A. Stefanski, T. Lee, and S. Bentivenga. 2011. Advances, challenges and a developing synthesis of ecological community assembly theory. *Philosophical Transactions of the Royal Society B* 366:2403–2413.
- Wenger, S. J., and J. D. Olden. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution* 3:260–267.
- Wilkinson, D. P., N. Golding, G. Guillera-Arroita, R. Tingley, and M. A. McCarthy. 2019. A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution* 10:198–211.
- Williams, J. W., and S. T. Jackson. 2007. Novel climates, no-analog communities, and ecological surprises. *Frontiers in Ecology and the Environment* 5:475–482.
- Wisz, M. S., et al. 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions* 14:763–773.
- Wisz, M. S., 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.
- Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society B (Statistical Methodology)* 73:3–36.
- Zhang, C., Y. Chen, B. Xu, Y. Xue, and Y. Ren. 2018. Comparing the prediction of joint species distribution models with respect to characteristics of sampling data. *Ecography* 41:1876–1887.
- Zimmermann, N. E., T. C. Edwards Jr., C. H. Graham, P. B. Pearman, and J. C. Svenning. 2010. New trends in species distribution modelling. *Ecography* 33:985–989.
- Zurell, D., et al. 2016. Benchmarking novel approaches for modelling species range dynamics. *Global Change Biology* 22:2651–2664.

## SUPPORTING INFORMATION

Additional supporting information may be found online at: <http://onlinelibrary.wiley.com/doi/10.1002/ecm.1370/full>

## DATA AVAILABILITY

Data are available on the Zenodo repository: <https://doi.org/10.5281/zenodo.2637812>

**Supporting Information.** Anna Norberg, Nerea Abrego, F. Guillaume Blanchet, Frederick R. Adler, Barbara J. Anderson, Jani Anttila, Miguel B. Araújo, Tad Dallas, David Dunson, Jane Elith, Scott D. Foster, Richard Fox, Janet Franklin, William Godsoe, Antoine Guisan, Bob O’Hara, Nicole A. Hill, Robert D. Holt, Francis K.C. Hui, Magne Husby, John Atle Kålås, Alekski Lehtikoinen, Miska Luoto, Heidi K. Mod, Graeme Newell, Ian Renner, Tomas Roslin, Janne Soininen, Wilfried Thuiller, Jarno Vanhatalo, David Warton, Matt White, Niklaus E. Zimmermann, Dominique Gravel, and Otso Ovaskainen. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*.

## **Appendix S1. Supporting information about statistical models and how they were fitted.**

In the following listing of the modelling frameworks used in this study, we describe the name, main references for the implementation used, and briefly the overall functioning of the framework and model settings applied, with an emphasis on how they deviate from the default, as well as references for some (other) practical ecological examples of the modelling framework. The approach for this supplement was adopted from the elaborate supplementary descriptions of Elith et al. (2006).

We describe here all the model variants and the settings used to fit them. **The descriptions are based on the main references of the implementation used, indicated right after the model title, unless otherwise stated.** We note that most of the applied models have also other implementations, which are more or less similar to the ones we have used, and for which there might be several methodological as well as practical references not given here. In addition, here we only technically describe the settings related to the model functions and their usage, since the data used is described elsewhere (Table 3 in the main text) and is treated similarly throughout the models. Hence, only if there are e.g. some inbuilt variable selection procedures within the functions, such procedure is applied. All the models were applied in the R environment (R Core Team 2018), except for the method ‘HMSC’, for which Matlab (MATLAB 2015a) was used.

After describing the models, we clarify why the models were evaluated based on predicted distributions of realized species occurrences, rather than predicted species occurrence probabilities.

The fitting of the models and producing predictions were implemented by the first author with contributions from the developers of the methods and coauthors, and the pipeline was inspected as a whole by coauthor T. Dallas.

*BayesComm* – *Bayesian Community Ecology Analysis* (Golding and Harris 2015)

**Description:** ‘BayesComm’ is a joint species distribution model based on multivariate binomial Bayesian probit regressions. The method is based on a model described by Edwards & Allenby (2003). Four types of models can be implemented: 1) a null model, with intercept only; 2) an environmental model with intercept and covariates; 3) a community model, with intercept and community matrix; or 4) a full model with intercept, covariates as well as community matrix. The community matrix refers to the matrix describing the residual correlation patterns between species.

**Implementation and modification from its default settings:** We used the ‘BC’ function of the package ‘BayesComm’. We fitted two versions of the model: the ‘environment’ model, i.e. including the intercept and environmental covariates (model variant BC.1); and the ‘full’ model, i.e. including the intercept, environmental covariates as well as the community matrix in the model (model variant BC.2). We used default prior distributions of the function. The models were fitted with 30 000 burn in and 50 000 as the total number of MCMC iterations.

**Practical ecological examples:** Golding et al. (2015)



*Boral* — Bayesian Ordination and Regression Analysis (Hui 2016, 2017)

**Description:** ‘boral’ is a joint species distribution modelling method using Bayesian ordination and regression models. With ‘boral’, three types of models can be implemented: 1) with covariates but no latent variables, i.e. stacked single species GLMs; 2) no covariates, i.e. a joint latent variable model; 3) with both covariates and latent variables, i.e. a joint community model with shared responses as well as latent variables. In addition, row effects (e.g. for differences in sampling intensity) can be accounted for.

**Implementation and modification from its default settings:** We used the ‘boral’ function of the package ‘boral’. We fitted a joint model with covariates and two latent variables, as recommended in the documentation as the minimum (model variant BORAL.1). We used default prior distributions of the function. The models were fitted with 40 000 burn in and as 50 000 as the total number of MCMC iterations. We used probit-link function.

**Practical ecological examples:** Hansen et al. (2016), Harabiš & Dolný (2018)

*BRT* — Boosted Regression Trees (Friedman 2001, 2002, Elith et al. 2008)

**Description:** ‘BRT’ is a single species distribution modelling method based on additive regression models in which individual terms are simple regression trees, fitted in a forward, stage-wise fashion. ‘BRT’ uses two algorithms: regression trees, in which the prediction space is partitioned, and boosting, which combines the collection of the regression trees by fitting a tree after another to the residuals of the previous tree.

**Implementation and modification from its default settings:** We used the ‘gbm’ and ‘dismo’ packages (model variant BRT.1). We use the ‘gbm.step’ –function, which performs step-wise cross-validation for identifying the optimal number of trees. For each species of each data set we fit versions of boosted regression trees with learning rates 0.1, 0.01, 0.001 and 0.0001, and with tree complexities 2, 3, 4 and 5, so that we fit models with all possible combinations of these learning rates and tree complexities and select the model with the lowest mean residual deviance. For all the model, we used the default initial number of trees, i.e. 50.

**Practical ecological examples:** Elith et al. (2008), Lovelock et al. (2015), Feld et al. (2016)

*GAM* — Generalised Additive Model (Wood 2004, 2011)

**Description:** ‘GAM’ is a single species distribution model, in which each explanatory variable has an additive effect resulting in a sum of functions of each variable. ‘GAM’ with penalized regression splines is a data-driven, non-parametric extension of a generalised linear model, and can thus also fit nonlinear relationships. There are alternative smoothers available, and the degree of the smoothness of the model terms is estimated as a part of the fitting process.

**Implementation and modification from its default settings:** We used the ‘mgcv’ package and functions ‘gam’ and ‘gamm’. We use cubic regression splines with shrinkage (‘cs’) as smoothing basis. Otherwise we used the default settings of the functions. We fitted variants with (GAM.2) and without (GAM.1) spatial structure. For the model variant with spatial structure, we defined the correlation structure (‘corStruct’) as Gaussian spatial correlation with default options. With all models, we used probit-link function.

**Practical ecological examples:** Soininen et al. (2013), Harrison et al. (2015)

*GJAM* — Generalised Joint Attribute Modelling (Clark et al. 2017, Taylor-Rodríguez et al. 2017)

**Description:** ‘GJAM’ is a Bayesian joint species distribution modelling framework enabling the usage of joint attribute data, e.g. species abundances, counts, as well as presence-absence, in the same individual JSDM. The

integration of discrete and continuous data on the observed scales makes use of censoring. Censoring is used with the effort for an observation to combine continuous and discrete variables with appropriate weight, and discrete observations can be viewed as censored versions of an underlying continuous space. The method applies the Dirichlet process in order to reduce the dimension of the joint covariance structure among species, enabling computational flexibility.

**Implementation and modification from its default settings:** We used the function ‘gjam’ of the package ‘gjam’ (model variant GJAM.1). We used default prior distributions of the function. The models were fitted with 30 000 burn in and as 50 000 as the total number of MCMC iterations. As the data is presence-absence, we used ‘PA’ as the only data type. ‘gjam’ applies Tobit-link function.

**Practical ecological examples:** Clark et al. (2014), Bachelot et al. (2018)

*GLM – Univariate Generalised Linear Models* (Venables and Ripley 2002, Wang et al. 2012, Hui 2016, 2017, Ovaskainen et al. 2017, R Core Team 2018)

**Description:** We fitted several variants of univariate generalised linear models, which all are versions of single species distribution models. ‘GLMs’ are often default models used in various application, including species distribution modelling. We fitted model variants with both maximum likelihood and Bayesian inference.

**Implementation and modification from default settings:** With ‘glm’ function of the package ‘stats’, we used the default fitting method (‘glm.fit’) defined in the function, which uses iteratively reweighted least squares (IRLS). For producing predictions with account for parameter uncertainty (model variant GLM.8), we used the same fitted model variant GLM.1 but we simulated the observations by first sampling a set of parameters from the asymptotic distribution of the parameters, then transforming them to the response scale using the inverse link function, and as a final step, simulating an observation for each of the predictions as is done with the predictions without parameter uncertainty (model variant GLM.1). This method for accounting parameter uncertainty adopted here mirrors the approach used in prediction applications (e.g. Foster and Dunstan 2010) and is a simple alternative to the delta-method (Ver Hoef 2012).

With ‘glmmPQL’ of the package ‘MASS’ we used the default fitting method, which relies on the ‘lme’ function of the package ‘nlme’ (Pinheiro et al. 2017). For the model variant with no spatial structure (GLM.2), we applied a random effect using the sampling unit for grouping, resulting in independent sampling units. For the model with spatial structure ‘corSpatial’, we used exponential spatial correlation structure (GLM.3). As implied by the name, ‘glmmPQL’ relies on penalized quasi-likelihood. For all the aforementioned, we used probit-link function. With the package ‘mvabund’ we fit two models: with the ‘manyglm’ function, using the default settings (model variant GLM.6), as well as with the function ‘traitglm’ which includes a LASSO penalty via ‘glm1path’ (GLM.9). For both, logit-link function was used and the best model chosen (automatically, within the function) based on minimizing the BIC.

With ‘glmnet’ function of the package ‘Glmnet’ we fitted two variants of generalised linear models fitted with maximum likelihood, applying LASSO penalty. The regularisation path for the penalty was computed at a grid of 200 values for the regularization parameter lambda. For the first variant (GLM.10), the best model was selected based on minimizing the BIC, and the predictions produced with the default prediction function. The second variant (GLM.11) was implemented by fitting a model with the same lambda value as the first variant, but the model was fitted (and predictions made) 100 times for bootstrapped data, so we obtained a variant incorporating parameter uncertainty.

For the HMSC models, we used the package ‘HMSC’ for Matlab, and implemented variants with only environmental variables (model variant GLM.4), with spatial structure (spatially structured latent variable, model variant GLM.5), using default priors defined in the function. The individual HMSC models for each species were fitted with 6 000 MCMC iterations, of which the first half was cut. We fitted another Bayesian GLM with the package ‘boral’, using the function ‘boral’ (GLM.7). We fitted the models with 50 000 MCMC iterations and a 40 000 burn in. With all the Bayesian GLM variants we used probit-link function and default priors.

**Practical ecological examples:** Burger et al. (2012); Rousset & Ferdy (2014); Doherty et al. (2015); Nguyen & Landfald (2015)

*GNN – Gradient Nearest Neighbour (imputation)* (Crookston and Finley 2008)

**Description:** ‘GNN’ is a nearest neighbour technique, which is a class of multivariate, non-parametric approaches to continuous or categorical prediction. Predictions are calculated as linear combinations of observations for population units in a sample that are similar or nearest in a space of auxiliary variables to population units requiring predictions. Possible missing values within a data set are filled with known values from that same data set with imputation. Gradient refers to the strategy used to compute the distance, which is calculated by using a projected ordination of environmental covariates, which are found using canonical correspondence analysis (CCA), or in case it fails, redundancy analysis (RDA).

**Implementation and modification from its default settings:** We used the package ‘yaImpute’ and function ‘yai’ (model variant GNN.1). For each data set we fitted models with varying k nearest neighbours (1, 5, 10, 30, 100) and chose the best model based on the Area Under Curve (AUC), with otherwise default settings of the function.

**Practical ecological examples:** Chirici et al. (2015)

*HMSC – Hierarchical Modelling of Species Communities* (Ovaskainen et al. 2017)

**Description:** ‘HMSC’ is a joint species distribution modelling framework, based on Bayesian generalised linear models, that relates the variation in species occurrence and co-occurrence to the influences of environmental variables and species-to-species associations. The model can be used jointly so that the species-specific parameters influence each other, or also including community-level parameters to be estimated.

**Implementation and modification from its default settings:** We used the ‘HMSC’ package for Matlab. We implemented variants with only environmental variables (HMSC.1), with a sampling unit level latent variable (HMSC.2), and with spatial structure (spatially structured latent variable; HMSC.3). We used default prior distributions of the function. The models were fitted with a total of 50 000 MCMC iterations, of which the first 40 000 iterations were cut. For all versions, we used the probit-link and the default priors defined in the function.

**Practical ecological examples:** Abrego et al. (2016, 2017), Ovaskainen et al. (2016)

*MARS – Multivariate Adaptive Regression Splines* (Friedman 1991, Leathwick et al. 2005, Milborrow 2017)

**Description:** ‘MARS’ is an adaptive non-linear regression that uses piecewise linear basis functions to correlate the response variable and predictors, resulting in a linear continuous regression surface. Binomial data are analysed by fitting an initial model, after which a generalised linear model that relates species occurrences to these, assuming a binomial error distribution. ‘MARS’ is a stacked single species distribution modelling method, but with a joint SDM flavour in the sense that the environmental covariates are selected simultaneously for all the species: the model identifies the set of explanatory variables that explains the overall variation in species occurrence patterns best, but individual regression models are used to calculate species-specific parameters.

**Implementation and modification from its default settings:** We used the package ‘earth’, and function ‘earth’. For each data set we fitted a model variant that includes interactions between environmental covariates (degree=2, MARS.2), as well as a variant without interactions (an additive model, degree=1, MARS.1). For both models, we used probit-link function.

**Practical ecological examples:** Moisen and Frescino (2002), Yen et al. (2004), Leathwick et al. (2005)

*MISTNET — Stochastic Feedforward Neural Networks with Random Variables* (Harris 2015)

**Description:** ‘mistnet’ is a Bayesian joint species distribution modelling method. A neural net applies a series of nonlinear transformations to covariates, and after a suitable transformation, logistic regressions are applied in the transformed space to make predictions of probabilities. Training the neural network requires adjusting the transformation parameters simultaneously for optimising the overall likelihood, or posterior density. The estimated probabilities are allowed to depend also on unobserved environmental factors. Coefficients are initialised randomly, after which they are iteratively adjusted towards values with higher likelihoods via gradient-based hill climbing. To avoid overfitting, prior distributions are set to favour smaller magnitude parameter values over larger ones.

**Implementation and modification from its default settings:** We used the package ‘mistnet2’ and function ‘mistnet’ (model variant MISTN.1). We defined three neural network layers: input, hidden and output. The number of nodes in the input layer equalled the number of explanatory variables plus the number of latent variables plus one. We used one latent variable. The number of output nodes equals the number of species. The number of nodes in the hidden layer was set to its recommended minimum, i.e. the same as the number of nodes of the input layer. The activation function was set to a sigmoid form, and the prior distributions were normal with mean zero in standard deviation of 1. As error distribution for the response variables, we used binomial distribution with probit-link function. Otherwise, we used the default options of the function.

**Practical ecological examples:** Lacasella et al. (2016), Marta et al. (2016), Naujokaitis-Lewis and Fortin (2016)

*MRT — Multivariate Regression Trees* (De’Ath 2002)

**Description:** ‘MRT’ is a stacked single species distribution method, and a multivariate extension to univariate regression trees. It is both a form of multivariate regression and constrained clustering. ‘MRT’ makes no assumptions of the form of the species responses to their environment but fits a single tree for the whole multivariate data set, and the tree splits are based on minimizing the sums of impurity measures (which can be defined as the sum of squares of the multivariate mean) of the node over the multivariate response. Tree pruning and deciding tree size are done with cross-validation. ‘MRT’ has a slight joint SDM flavour in the sense that it clusters the species according to their responses, identifies species that characterise these groups, as well as species that strongly influence the splits of the common tree, but in the end produces single species distributions.

**Implementation and modification from its default settings:** We used the ‘mvpart’ package and the ‘mvpart’ function. We fitted models using 10-fold cross-validation for evaluation, with method ‘min’, which picks the overall best tree, and otherwise default settings (model variant MRTS.1).

**Practical ecological examples:** Cappo et al. (2007)

*RF — Random Forest* (Breiman 2001)

**Description:** ‘RF’ is a single species distribution modelling method based on an ensemble learning technique, in which many classifiers are generated and the results aggregated. The diversity in the classification trees of a random forest is increased by resampling the data with replacement, as well as randomly selecting a subsample of the covariates for each tree building process. Hence, random forest models include a random component also, as with each run, the subsamples are selected again.

**Implementation and modification from its default settings:** We use the ‘randomForest’ function of the ‘randomForest’ package (model variant RF.1). We tried versions of the model by varying the number of trees (500, 1000, 1500 or 2000); the number of variables randomly sampled as candidates at each split (from two to rounded half of the total amount of variables, resulting in e.g. with the total number variables being 5, to 2 or 3 randomly selected variables at a time); and the minimum size of terminal nodes (1, 2, 5 or 10). The best model was selected based on mean squared error (built in the tuning function ‘tune’).

**Practical ecological examples:** Yatsuneneko et al. (2009), Heikkinen et al. (2012)

*SAM – Species Archetype Model* (Hui et al. 2013, Dunstan et al. 2013)

**Description:** ‘SAM’ is a joint species distribution method based on a multivariate generalised linear mixture model using maximum likelihood, which assumes that all species can be classified into a small number of archetypal responses to the explanatory variables, reducing the dimension of the regression and thus the number of parameters that has to be estimated. By clustering based on the response to the explanatory variables, strength is borrowed across species, allowing responses for more rare species to be estimated with greater precision by grouping them with prevalent species having statistically similar responses.

**Implementation and modification from its default settings:** We used the ‘psams.coord’ function and the source code for the method was provided by the developer F. K. Hui (see refs. for the method). We fitted models with the number of archetypes ranging from 3 to 13, and otherwise default options (model variant SAM.1). The best model was selected based on their BIC values.

**Practical ecological examples:** Leaper et al. (2014), Foster et al. (2015)

*SVM – Support Vector Machine* (Drake et al. 2006)

**Description:** ‘SVM’ is a single species distribution modelling method based on simple maximal margin classifiers but incorporating nonlinear boundaries. The support vectors are projected into a multidimensional feature space by applying a kernel trick, which replaces the predictors with a kernel function, and fitting an optimal hyperplane using an optimisation function, which maximises the margin, i.e. is the distance between the closest training samples, or support vectors, and the hyperplane itself.

**Implementation and modification from its default settings:** We used the package ‘e1071’ and function ‘svm’ (model variant SVM.1). We used the ‘eps-regression’ and as model kernel we used the Gaussian radial basis function (‘radial’), which relies on tuning only one parameter, gamma, for which we tried values from 0.001 to 100, with one order of magnitude intervals (the default is 1 divided by the number of explanatory variables). We used values 1, 5, 10, 50 or 100 as the cost parameter. The cost parameter allows the user to set the balance between bias and variance: increasing the cost parameter adds variance to the model and decreases bias. We used 10-fold cross-validation for assessing the quality of the model and selected the best one based on mean squared error (built in the tuning function ‘tune’).

**Practical ecological examples:** Pouteau et al. (2012), Bari et al. (2014)

*XGB – Extreme Gradient Boosting* (Chen et al. 2018)

**Description:** ‘XGB’ is a single species distribution modelling method based on gradient boosting, where individual terms are simple regression trees, fitted in a forward, stage-wise fashion. In addition, ‘XGB’ uses boosting, as well as gradient descent to optimise the cost function.



**Implementation and modification from its default settings:** We used the package ‘xgboost’ and function ‘xgboost’ (model variant XGB.1). We used 1000 as maximum number of boosting iterations and the default (6) maximum depth of a tree, and ‘bilinear’ as booster.

**Practical ecological examples:** Perez Saez (2018)

### **Generation of predicted distributions: why distribution of occurrences rather than occurrence probabilities?**

As explained in the main text, we evaluated model performance based on 100 random realizations of predicted species occurrence matrices, i.e. matrices of zeros and ones, rather than occurrence probabilities. Here we explain the reasoning behind this choice. Briefly, the 0/1 predictions involve the correlation structure among species occurrences, for those models variants that are able to capture it: those that model co-occurrences beyond those explained by environmental predictors (Feature D). Let us illustrate by assuming first that there are 100 species, all of which are predicted to occur on a given site independently of each other with probability 0.5. Then expected species richness is 50, and its 95% CI is ca. [40...60], based on Binomial(100,0.5). Let us then assume that the 100 species are still predicted to occur on that site with marginal probability 0.5, but the occurrences are not independent of each other. Then expected species richness is still 50, but the 95% CI may be e.g. [10...90] (if associations are predominantly positive), or [45...55] (if associations are predominantly negative), instead of [40...60] that assumes independent occurrences. Such information (relevant for some but not all of the performance measures) cannot be propagated if the starting point is just the vector of occurrence probabilities rather than a distribution of realized occurrences. In contrast, a sample of predicted occurrences (0/1) contains information about the entire joint distribution. The mean of the predicted occurrences equals occurrence probability (up to sampling error) and thus the predictions include the probabilistic information marginally for each species, but the predicted occurrences include also information on dependencies among species (and sometimes among spatial units) beyond occurrence probabilities. To further illustrate, consider many sites with identical environmental predictors, and assume that two species are predicted to be present in all of these with probability 0.5. Assume further that the two species show a high positive association that is captured by the fitted statistical model. Then the occurrence probabilities for (species 1, species 2) are (0.5,0.5) for all sites, suggesting that the cases of (0,0), (1,0), (0,1) and (1,1) are equally common. However, the predicted binary realizations will mostly be (0,0) or (1,1) due to the positive association between the species.

### **References**

- Abrego, N., D. B. Dunson, P. Halme, I. Salcedo, and O. Ovaskainen. 2017. Wood-inhabiting fungi with tight associations with other species have declined as a response to forest management. *Oikos* 126:269–275.
- Abrego, N., A. Norberg, and O. Ovaskainen. 2018. Measuring and predicting the influence of traits on the assembly processes of wood-inhabiting fungi. *Journal of Ecology* 105:1070–1081.
- Bachelot, B., M. Uriarte, R. Muscarella, J. Forero-Montaña, J. Thompson, K. McGuire, J. Zimmerman, N. G. Swenson, and J. S. Clark. 2018. Associations among arbuscular mycorrhizal fungi and seedlings are predicted to change with tree successional status. *Ecology* 99:607–620.
- Bari, A., A. Amri, K. Street, M. Mackay, E. De Pauw, R. Sanders, K. Nazari, B. Humeid, J. Konopka, and F. Alo. 2014. Predicting resistance to stripe (yellow) rust (*Puccinia striiformis*) in wheat genetic resources using focused identification of germplasm strategy. *The Journal of Agricultural Science* 152: 906–916.
- Breiman, L. E. O. 2001. Random Forests. *Machine Learning* 45:5–32.
- Burger, C., E. Belskii, T. Eeva, T. Laaksonen, M. Mägi, R. Mänd, A. Qvarnström, T. Slagsvold, T. Veen, M. E. Visser, K. L. Wiebe, C. Wiley, J. Wright, and C. Both. 2012. Climate change, breeding date and nestling diet: how temperature differentially affects seasonal changes in pied flycatcher diet depending

on habitat variation. *Journal of Animal Ecology* 81:926–936.

- Cappo, M., G. De'ath, and P. Speare. 2007. Inter-reef vertebrate communities of the Great Barrier Reef Marine Park determined by baited remote underwater video stations. *Marine Ecology Progress Series* 350:209–221.
- Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li. 2018. *xgboost: Extreme Gradient Boosting*. R package version 0.71.2. <https://CRAN.R-project.org/package=xgboost>.
- Chirici, G., M. Mura, D. McInerney, N. Py, E. O. Tomppo, L. T. Waser, D. Travaglini, and R. E. McRoberts. 2016. A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. *Remote Sensing of Environment* 176:282–294.
- Clark, J. S., A. E. Gelfand, C. W. Woodall, and K. Zhu. 2014. More than the sum of the parts: Forest Climate response from joint species distributions. *Ecological Applications* 24:990–999.
- Clark, J. S., D. Nemergut, B. Seyednasrollah, P. Turner, and S. Zhang. 2017. Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs* 87:34–56.
- Crookston, N. L., and a O. Finley. 2008. *yaImpute: An R package for  $k$ NN imputation*. *Journal of Statistical Software* 23:1–16.
- De'Ath, G. 2002. Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships. *Ecology* 83:1105–1117.
- Doherty, T. S., R. A. Davis, E. J. B. van Etten, D. Algar, N. Collier, C. R. Dickman, G. Edwards, P. Masters, R. Palmer, and S. Robinson. 2015. A continental-scale analysis of feral cat diet in Australia. *Journal of Biogeography* 42:964–975.
- Drake, J. M., C. F. Randin, and A. Guisan. 2006. Modelling ecological niches with support vector machines. *Journal of Applied Ecology* 43:424–432.
- Dunstan, P. K., S. D. Foster, F. K. C. Hui, and D. I. Warton. 2013. Finite Mixture of Regression Modeling for High-Dimensional Count and Biomass Data in Ecology. *Journal of Agricultural, Biological, and Environmental Statistics* 18:357–375.
- Edwards, Y. D., and G. M. Allenby. 2003. Multivariate analysis of multiple response data. *Journal of Marketing Research* 40:321–334.
- Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. S. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberón, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29:129–151.
- Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77:802–813.
- Feld, C., S. Birk, D. Eme, M. Gerisch, D. Hering, M. Kernan, K. Maileht, U. Mischke, I. Ott, F. Pletterbauer, S. Poikane, J. Salgado, C. Sayer, J. van Wichelen, and F. Malard. 2016. Disentangling the effects of land use and geo-climatic factors on diversity in European freshwater ecosystems. *Ecological Indicators* 60:71–83.
- Foster, S. D., and P. K. Dunstan. 2010. The analysis of biodiversity using rank abundance distributions. *Biometrics* 66:186–195.

- Foster, S. D., P. K. Dunstan, F. Althaus, and A. Williams. 2015. The cumulative effect of trawl fishing on a multispecies fish assemblage in south-eastern Australia. *Journal of Applied Ecology* 52:129–139.
- Friedman, J. H. 1991. Adaptive Regression Splines. *The Annals of Statistics* 19:1–67.
- Friedman, J. H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29:1189–1232.
- Friedman, J. H. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38:367–378.
- Golding, N., and D. J. Harris. 2015. BayesComm: Bayesian Community Ecology Analysis. R package version 0.1-2.
- Golding, N., M. A. Nunn, and B. V Purse. 2015. Identifying biotic interactions which drive the spatial distribution of a mosquito community. *Parasites & Vectors* 8:1–10.
- Hansen, R. R., O. L. P. Hansen, J. J. Bowden, U. A. Treier, S. Normand, and T. Høye. 2016. Meter scale variation in shrub dominance and soil moisture structure Arctic arthropod communities. *PeerJ* 4:e2224.
- Harabiš, F., and A. Dolný. 2018. Military training areas as refuges for threatened dragonfly species: Effect of spatial isolation and military activity. *Biological Conservation* 217:28–35.
- Harris, D. J. 2015. Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution* 6:465–473.
- Harrison, P. J., Y. Yuan, S. T. Buckland, C. S. Oedekoven, D. A. Elston, M. J. Brewer, A. Johnston, and J. W. Pearce-Higgins. 2015. Quantifying turnover in biodiversity of British breeding birds. *Journal of Applied Ecology* 53:469–478.
- Heikkinen, R. K., M. Marmion, and M. Luoto. 2012. Does the interpolation accuracy of species distribution models come at the expense of transferability? *Ecography* 35:276–288.
- Ver Hoef, J. M. 2012. Who invented the delta method? *American Statistician* 66:124–127.
- Hui, F. K. C. 2016. BORAL - Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R. *Methods in Ecology and Evolution* 7:744–750.
- Hui, F. K. C. 2017. boral: Bayesian Ordination and Regression AnaLysis. R package version 1.4.
- Hui, F. K. C., D. I. Warton, S. D. Foster, and P. K. Dunstan. 2013. To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. *Ecology* 94:1913–1919.
- Lacasella, F., S. Marta, A. Singh, K. Stack Whitney, K. Hamilton, P. Townsend, C. J. Kucharik, T. D. Meehan, and C. Gratton. 2016. From pest data to abundance-based risk maps combining eco-physiological knowledge, weather and habitat variability. *Ecological Applications* 27:575–588.
- Leaper, R., P. K. Dunstan, S. D. Foster, N. S. Barrett, and G. J. Edgar. 2014. Do communities exist? Complex patterns of overlapping marine species distributions. *Ecology* 95:2016–2025.
- Leathwick, J. R., D. Rowe, J. Richardson, J. Elith, and T. Hastie. 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology* 50:2034–2052.
- Lovelock, C. E., D. R. Cahoon, D. A. Friess, G. R. Guntenspergen, K. W. Krauss, R. Reef, K. Rogers, M. L. Saunders, F. Sidik, A. Swales, N. Saintilan, L. X. Thuyen, and T. Triet. 2015. The vulnerability of Indo-Pacific mangrove forests to sea-level rise. *Nature* 526:559–U217.
- Marta, S., F. Lacasella, P. Gratton, D. Cesaroni, and V. Sbordoni. 2016. Deciphering range dynamics: effects

of niche stability areas and post-glacial colonization on alpine species distribution. *Journal of Biogeography* 43:2186–2198.

MATLAB and Statistics Toolbox Release 2015a. (n.d.). The MathWorks Inc., Natick, Massachusetts, United States.

Milborrow, S. 2017. earth: Multivariate Adaptive Regression Splines. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. R package version 4.5.1.

Naujokaitis-Lewis, I., and M. J. Fortin. 2016. Spatio-temporal variation of biotic factors underpins contemporary range dynamics of congeners. *Global Change Biology* 22:1201–1213.

Nguyen, T. T., and B. Landfald. 2015. Polar front associated variation in prokaryotic community structure in Arctic shelf seafloor. *Frontiers in Microbiology* 6:1–12.

Ovaskainen, O., D. B. Roy, R. Fox, and B. J. Anderson. 2016. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution* 7:428–436.

Ovaskainen, O., G. Tikhonov, A. Norberg, F. G. Blanchet, L. Duan, D. B. Dunson, T. Roslin, N. Abrego, G. F. Blanchet, L. Duan, D. B. Dunson, T. Roslin, and N. Abrego. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters* 2:561–576.

Perez Saez, F. J. 2018. A field-based modelling framework of the ecohydrology of schistosomiasis. EPFL, Lausanne, Switzerland.

Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, and R Core Team. 2017. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131.

Pouteau, R., J.-Y. Meyer, R. Taputuarai, and B. Stoll. 2012. Support vector machines to map rare and endangered native plants in Pacific islands forests. *Ecological Informatics* 9:37–46.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rousset, F., and J.-B. Ferdy. 2014. Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography* 37:781–790.

Soininen, J., J. J. Korhonen, and M. Luoto. 2013. Stochastic species distributions are driven by organism size. *Ecology* 94:660–670.

Taylor-Rodríguez, D., K. Kaufeld, E. M. Schliep, J. S. Clark, and A. E. Gelfand. 2017. Joint species distribution modeling: Dimension reduction using Dirichlet processes. *Bayesian Analysis* 12:939–967.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. 4th edition. Springer International Publishing, New York, USA.

Wang, Y., U. Naumann, S. T. Wright, and D. I. Warton. 2012. mvabund - an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* 3:471–474.

Wood, S. N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* 99:673–686.

Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 73:3–36.



Yatsunenko, T., F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. a. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. I. Gordon. 2009. Human gut microbiome viewed across age and geography. *Nature* 457:222–227.

**Supporting Information.** Anna Norberg, Nerea Abrego, F. Guillaume Blanchet, Frederick R. Adler, Barbara J. Anderson, Jani Anttila, Miguel B. Araújo, Tad Dallas, David Dunson, Jane Elith, Scott D. Foster, Richard Fox, Janet Franklin, William Godsoe, Antoine Guisan, Bob O’Hara, Nicole A. Hill, Robert D. Holt, Francis K.C. Hui, Magne Husby, John Atle Kålås, Aleksi Lehikoinen, Miska Luoto, Heidi K. Mod, Graeme Newell, Ian Renner, Tomas Roslin, Janne Soininen, Wilfried Thuiller, Jarno Vanhatalo, David Warton, Matt White, Niklaus E. Zimmermann, Dominique Gravel, and Otso Ovaskainen. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*.

## **Appendix S2. Supporting information about the included datasets and their splitting to training and validation data**

### **Environmental covariates of the five data sets**

Below we list the original covariates for each of the five data sets. As described in the main text, the original covariates were transformed into principal components that were then used as predictors in the statistical models.

#### ***Bird data. Breeding Bird Surveys in Finland, Sweden and Norway***

The environmental covariates include 21 variables related to land cover, weather, and variation in sampling effort.

*Land cover* (grouped based on Corine land cover; <https://land.copernicus.eu/pan-european/corine-land-cover>)

1. Proportion of grid cell consisting of urban habitats (URBAN)
2. Proportion of grid cell consisting of arable land (ARABLE)
3. Proportion of grid cell consisting of pastures (PASTURE)
4. Proportion of grid cell consisting of agriculture + natural vegetation (ARABLE\_EDGE)
5. Proportion of grid cell consisting of broad-leaved forest (BROADLEAF)
6. Proportion of grid cell consisting of coniferous forest (CONIFER)
7. Proportion of grid cell consisting of mixed forest (MIXED)
8. Proportion of grid cell consisting of natural grasslands and sparsely vegetated areas (MOUNTAIN)
9. Proportion of grid cell consisting of scrub land (SHRUB)
10. Proportion of grid cell consisting of beaches, dunes and sands (BEACH)
11. Proportion of grid cell consisting of inland marshes (WETLAND)
12. Proportion of grid cell consisting of peat bogs (PEATLAND)
13. Proportion of grid cell consisting of water courses and bodies (INLANDWATER)
14. Proportion of grid cell consisting of sea (MARINEWATER)

*Weather variables (based on national meteorological institutes)*

15. Mean temperature in June and July (JUNJUL)
16. Mean temperature in December, January and February (DJF)
17. Mean temperature in April (APRIL)
18. Mean temperature in May (MAY)

*Sampling effort (sampling method and effort)*

19. Indicator variable for whether the sampling was conducted as point or line count (LINE)
20. Length of transect for line count (LINE Effort)
21. Number of counting points for point effort (POINT Effort)

#### ***Butterfly data. Butterflies in the Great Britain***

The environmental covariates include 34 variables related to land cover, topography and climate.

### *Topography based on the 10km x 10km British National Grid*

1. Mean elevation of the grid cell (ELE)

### *Climate – 30 year means based on the CRU ts2.1 data*

2. Growing degree days above five degrees (GDD5)
3. Mean temperature of the coldest month (MTCO)
4. Mean temperature of the warmest month (MTWA)
5. Ratio of the actual to potential evapotranspiration (APET)
6. Total annual precipitation (TPRE)

### *Land cover – based on the vector LCM 2000 data each variable represents the proportion of the 10km x 10km British National Grid (BNG) cell classified as that Land cover class.*

7. The proportion NOT classified as any terrestrial land cover class (SEA)
8. Inland water (>0.5ha), standing open water, canals, rivers and streams (INLANDWATER)
9. Littoral rock, including algae covered rock (ROCKL)
10. Littoral sand including with and without algae (SAND)
11. Saltmarsh including grazed saltmarsh (SALTMARSH)
12. Supra-littoral rock, including shingle vegetated and not vegetated ROCKSL
13. Supra-littoral dunes including vegetated and not vegetated (DUNE)
14. Deep peat bog (>0.5m), including cover of herb, grass, shrub and mixtures of these (BOG)
15. Dwarf shrub heath including dense ericaceous and gorse (HEATH)
16. Open heath (OPENHEATH)
17. Montane, vegetated ground above 600m (MONTANE)
18. Broadleaved and mixed woodland, including deciduous, mixed, open birch and scrub (BLWOOD)
19. Coniferous woodland, including conifers, felled and new plantations (CONWOOD)
20. Arable fields in grain, e.g., wheat, barley, maize, oats (GRAIN)
21. Arable fields in horticulture, non-grain, crops (CROPS)
22. Arable fields not in rotation (SETASIDE)
23. Improved grassland (GRASSIMP)
24. Setaside grassland (GRASSET)
25. Neutral grassland, rough unmanaged and unimproved grassland (GRASSNEUTRAL)
26. Calcareous grassland, managed and rough grassland, pH > 5.5 (GRASSCAL)
27. Acid grassland, including rough, with *Juncus*, *Nardus*, *Festuca* and *Molina*; pH<4.5 (GRASSACID)
28. Bracken (BRACKEN)
29. Fen, marsh or swamp, including fen willow (SWAMP)
30. A mixture of built and vegetated surfaces (SUBURBAN)
31. Continuous urban, including industrial and commercial and residential urban (URBAN)
32. Inland bare ground including semi natural and despoiled areas, e.g. quarries (ROCKIN)

Calcareous base rock derived from the British Geological Survey (BGS) vector data and transformed to a 10km x 10km British National Grid.

33. Proportion of grid cell where upper most layer of bedrock is classified as chalk (CHALK)
34. Proportion of grid cell where upper most layer of bedrock is classified as limestone (LIMESTONE)

### ***Plant data. Plants from Victorian Biodiversity Atlas***

The environmental covariates include 19 variables related to soil, topography and climate.

### *Topography*

1. Diurnal ‘heating’ associated with north-westerly aspects. Implemented on 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation (see Boehner and Conrad 2008). In high latitudes there is differentiation between shady south facing slopes and sunny north facing slopes. In addition there is a global asymmetry between western and eastern slopes. While levels of radiation maybe

identical, direct solar radiation on eastern slopes is typically associated with cooler ambient air and soil surface temperatures. As applied to a DEM the alpha(max) (or slope aspect[max]) variable is set at 315 degrees (North West) - giving a result that shows the 'north-westerlyness' of the aspect where Anisotropic heating =  $\cos(\text{slope aspect}(\text{max}) - \text{slope aspect}) * \arctan(\text{slope angle})$ . Spatial resolution 75m (ANISOTROPICHEATINGRUGGEDNESS75M)

2. Mean pan Evaporation January (mm) developed with ANUCLIM (Houlder et al 2000) applied to 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model. Spatial resolution 75m. (EVAPORATIONJAN75M)
3. Mean pan Evaporation July (mm) developed with ANUCLIM (Houlder et al 2000) applied to 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model. Spatial resolution 75m. (EVAPORATIONJUL75M)
4. Relative potential incoming solar radiation (RPISR) measured in kWh/metre squared. RPISR is equal to the modelled PISR on the Winter solstice + PISR modelled on the Summer solstice + PISR modelled on the equinox (for methods see Boehner and Antonic 2009) Wetness Index a compound terrain attribute (sensu Bevan and Kirby 1979) implemented using the Shuttle Radar Topography Mission (SRTM) Digital Elevation Model and TOPOCROP Version 2.1 (Schmidt 2002). Spatial resolution 75m (INSOLATIONTOTAL75M)
5. The amount of sky visible from a cell as a percentage of an unobstructed hemisphere (see Boehner and Antonic 2009). Transformation performed on the 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model. Spatial resolution 75m (VISIBLESKY75M)
6. Topographic Wetness Index a compound terrain attribute (sensu Bevan and Kirby 1979) implemented using the Shuttle Radar Topography Mission (SRTM) Digital Elevation Model and TOPOCROP Version 2.1 (Schmidt 2002). Spatial resolution 75m (WETNESSNOFILL75M)
7. Topographic Wetness Index a compound terrain attribute (sensu Bevan and Kirby 1979) implemented using the Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) and TOPOCROP Version 2.1 (Schmidt 2002). The DEM was pre-processed using an algorithm developed by Wang and Liu (2006) to remove surface depressions and preserve the downward slope gradient. Spatial resolution 75m (WETNESSWITHFILL75M)
8. Log transformed vertical distance above mapped saline lakes and the ocean (Conrad 2002). Implemented on 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model. Spatial resolution 75m (LOG\_SALINE)
9. Log transformed vertical distance major rivers (Conrad 2002). Implemented on 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model. Spatial resolution 75m (LOG\_MAJOR)
10. Log transformed vertical distance above the mapped stream channel network (Conrad 2002). Implemented on 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model Spatial resolution 75m (LOG\_MINOR)

### *Climate*

11. Mean maximum temperature January (degrees Celsius) developed with ANUCLIM (Houlder et al 2000) applied to 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model. Spatial resolution 75m (MAXTEMPJAN75M)
12. Mean minimum temperature July (degrees Celsius) developed with ANUCLIM (Houlder et al 2000) applied to 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model. Spatial resolution 75m (MINTEMPJUL75M)
13. Mean number of Rain days in January developed with ANUCLIM (Houlder et al 2000) applied to 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model. Spatial resolution 75m (RAINDAYSJAN75M)
14. Mean number of Rain days in July developed with ANUCLIM (Houlder et al 2000) applied to 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model. Spatial resolution 75m (RAINDAYSJUL75M)
15. Mean Rainfall January (mm) developed with ANUCLIM (Houlder et al 2000) applied to 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model. Spatial resolution 75m (RAINFALLJAN75M)



16. Mean number of Rain days in July developed with ANUCLIM (Houlder et al 2000) applied to 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model. Spatial resolution 75m (RAINFALLJUL75M)

### *Soil*

17. The Normalised Difference Ratio of the radioelement counts for Thorium (%) and Potassium (ppm). Derived from 50-m gridded rasters, Department of Economic Development, Jobs, Transport and Resources, Victoria, for airborne gamma radiometric spectrometry surveys (see also Read et al 2018). Spatial resolution 100m (RADIOMETICS\_ND\_KTH\_STAND\_2009)
18. The Ratio of the radioelement count of Thorium (%) and the inverse of Potassium (ppm) . Derived from 50-m gridded rasters, Department of Economic Development, Jobs, Transport and Resources, Victoria, for airborne gamma radiometric spectrometry surveys (see also Read et al 2018). Spatial resolution 100m (RADIOMETICS\_RATIO\_KINVTH\_STAND\_2009)
19. The Ratio of the radioelement counts of Thorium (%) and Potassium (ppm). Derived from 50-m gridded rasters, Department of Economic Development, Jobs, Transport and Resources, Victoria, for airborne gamma radiometric spectrometry surveys (see also Read et al 2018). Spatial resolution 100m (RADIOMETICS\_RATIO\_KTH\_STAND\_2009)

### ***Tree data. Trees in the USA***

The environmental covariates include 38 variables related to soil, topography and climate. The past-climate was provided by the Natural Resources Canada database covering the entire North-America with a resolution of 10 km<sup>2</sup> (McKenney et al. 2011). Spatial coordinates (using the 0-1 mile jigger and swapping 20% of the sites among a county, applied in the FIA method to mask the exact locations of the plots) were used to locate plots and interpolate climate variables. Climate variables consist of a set of variables directly related to temperature (e.g. annual maximal temperature, annual mean temperature, temperature annual range) and precipitation (e.g. total annual precipitation, precipitation per period). Growing degree days are interpreted from temperature data. To limit the inter-annual variability, climate was aggregated using the average on the last 15 years before the plot measurement. Slope and pH measurements are provided from the FIA data.

The environmental covariates include 38 variables related to climate, topography and soil.

### *Climate*

1. Annual maximum temperature (ANNUAL\_MAXIMUM\_TEMP)
2. Annual mean temperature (ANNUAL\_MEAN\_TEMP)
3. Annual minimum temperature (ANNUAL\_MINIMUM\_TEMP)
4. Growing degree days (above 5°C) for 3 weeks prior to growing season (GDD\_ABOVE\_BASE\_TEMP\_PERIOD)
5. Growing degree days (above 5°C) for first 6 weeks of growing season (GDD\_ABOVE\_BASE\_TEMP\_PERIOD2)
6. Growing degree days (above 5°C) for growing season (GDD\_ABOVE\_BASE\_TEMP\_PERIOD3)
7. Growing degree days (above 5°C) for growing season, excluding the first 6 weeks of it (GDD\_ABOVE\_BASE\_TEMP\_PERIOD4)
8. The average of monthly temperature ranges divided by the temperature annual range (see McKenney et al. 2011) (ISOTHERMALITY)
9. Julian date for the end of growing season (JD\_NUMBER\_END\_GROW\_SEASON)
10. Number of dates in the growing season (JD\_NUMBER\_GROW\_SEASON)
11. Julian date for the start of growing season (JD\_NUMBER\_START\_GROW\_SEASON)
12. Maximal temperature of the warmest month (MAX\_TEMP\_WARMEST\_PERIOD)
13. Mean diurnal temperature range (MEAN\_DIURNAL\_RANGE)
14. Mean temperature of the three coldest months (MEAN\_TEMP\_COLDEST\_QUARTER)
15. Mean temperature of the three driest months (MEAN\_TEMP\_DRIEST\_QUARTER)
16. Mean temperature during the growing season (MEAN\_TEMP\_PERIOD)
17. Mean temperature of the three warmest months (MEAN\_TEMP\_WARMEST\_QUARTER)

18. Mean temperature of the wettest months (MEAN\_TEMP\_WETTEST\_QUARTER)
19. Minimum temperature of the three coldest months (MIN\_TEMP\_COLDEST\_PERIOD)
20. Mean precipitation of the three coldest quarter (PP\_COLDEST\_QUARTER)
21. Mean precipitation during the driest month (PP\_DRIEST\_PERIOD)
22. Mean precipitation during the three driest months (PP\_DRIEST\_QUARTER)
23. Standard deviation of the monthly precipitation estimates expressed as a percent of their mean (see McKenney et al., 2011) (PP\_SEASONALITY)
24. Mean precipitation of the three warmest months (PP\_WARMEST\_QUARTER)
25. Mean precipitation during the wettest month (PP\_WETTEST\_PERIOD)
26. Mean precipitation during the three wettest months (PP\_WETTEST\_QUARTER)
27. Annual temperature range (TEMP\_ANNUAL\_RANGE)
28. Temperature range during growing season (TEMP\_RANGE\_PERIOD\_3)
29. Standard deviation of monthly-mean temperature estimates expressed as a percent of their mean (see McKenney et al., 2011) (TEMP\_SEASONALITY)
30. Total annual precipitation (TOT\_ANNUAL\_PP)
31. Total precipitation during 3 weeks prior to growing season (TOT\_PP\_PERIOD1)
32. Total precipitation for first 6 weeks of growing season (TOT\_PP\_PERIOD2)
33. Total precipitation in growing season (TOT\_PP\_PERIOD)
34. Total precipitation in for growing season, excluding the first 6 weeks of it (TOT\_PP\_PERIOD4)

#### *Topography*

35. Slope (SLP)

#### *Soil*

36. Soil depth (SOIL)
37. pH measured at 2cm depth (PH\_2CM)
38. pH measured at 10 cm depth (PH\_10CM)

#### ***Vegetation data. Vegetation in northern Norway***

The environmental covariates include six variables related to soil, topography and climate (see Niittynen and Luoto, in press).

*Soil variables are based on based on field surveys and visual interpretation of aerial and satellite images, and bedrock geology database maintained by geological survey of Norway (<http://geo.ngu.no/kart/berggrunn/>)*

1. A six-class classification of surface geology (peat, fluvial sediments, moraine, bolder field, and bare rock (SOILQUALITY)
2. An index determined as the downhill Euclidean distance to the shale belt (the only base-rich rock type in the area) scaled from 1 to 100. Regions located on the shale belt were valued at 100 and contrarily, areas uphill from shale belt or outside the area where the shale belt drains were set as zero (SURFDEP).

*Topography variables are derived from a 10 m resolution digital elevation model (DEM) [data portal of Norwegian Mapping Authority (<http://data.kartverket.no/download/>)]*

3. The elevational difference between the focal cell and mean elevation of the pixels within a 100 m radius using a 10-m resolution digital elevation model (TPI100)
4. The potential annual direct solar radiation was calculated (MJ/cm<sup>2</sup>/yr<sup>-1</sup>) by using the equations of McCune and Keon (2002) assuming clear sky conditions (RADIATIONMCCUNE)

*Climatic variables are modelled with generalized additive models (GAM) for the study area using climate information from 942 climate stations in Scandinavia (Norway, Finland and Sweden; Aalto et al. 2017).*

5. Mean temperature of the coldest month (January) (TAVG\_1)

6. Variable sums the daily mean temperature of the days when the mean temperature exceeds the selected temperature threshold, in this case 3°C. (GDD3)

## References

- Aalto, J., Riihimäki, H., Meineri, E., Hylander, K., and Luoto, M. (2017). Revealing topoclimatic heterogeneity using meteorological station data. *International Journal of Climatology*, 37, 544-556.
- Beven, K.J., Kirkby, M.J. (1979): A physically-based variable contributing area model of basin hydrology' *Hydrology Science Bulletin* 24(1), p.43-69.
- Boehner J. and Antonic, O. (2009). Land surface parameters specific to topo-climatology, in Hengl T. and Reuter H.I., eds., *Geomorphometry: concepts, software, applications*, Elsevier.
- Boehner, J and Conrad, O. (2008) *ChannelNetwork\_Altitude.cpp*, <http://sourceforge.net/saga-gis:24.05.2008>.
- Conrad, O. (2002) *ChannelNetwork\_Altitude.cpp*, <http://sourceforge.net/saga-gis:24.05.2008>.
- Houlder, D., Hutchinson, M., Nix, H. and MacMahon, J, 2001. ANUCLIM5.1 Users Guide. The Australian National University, enter for Resource and Environmental Studies, Canberra.
- McCune, B. and Keon, D. 2002. Equations for potential annual direct incident radiation and heat load. – *J. Veg. Sci.* 13: 603–606.
- McKenney, D. W., Hutchinson, M. F., Papadopol, P., Lawrence, K., Pedlar, J., Campbell, K., Milewska, E., Hopkinson, R. F., Price, D., and Owen, T. 2011. Customized spatial climate models for North America. *Bull. Am. Meteorol. Soc.*, 92(12): 1611–1622.
- Niittynen, P., and Luoto, M. 2018. The importance of snow in species distribution models of arctic vegetation. *Ecography*, in press.
- Read, C. F., Duncan, D. H., Ho, C. Y. C., White, M., & Vesk, P. A. (2018). Useful surrogates of soil texture for plant ecologists from airborne gamma-ray detection. *Ecology and Evolution*, 8(4), 1974–1983. <http://doi.org/10.1002/ece3.3417>
- Riley, S.J., De Gloria, S.D., Elliot, R. (1999): A Terrain Ruggedness that Quantifies Topographic Heterogeneity. *Intermountain Journal of Science*, Vol.5, No.1-4, pp.23-27.
- Schmidt, F. and Persson, A. 2003. Comparison of DEM data capture and topographic wetness indices. *Precision Agriculture*, 4, pp. 179 – 192.
- Wang, L. & H. Liu (2006): An efficient method for identifying and filling surface depressions in digital elevation models for hydrologic analysis and modelling. *International Journal of Geographical Information Science*, Vol. 20, 2: pp193-213.

## Distributions of environmental and spatial predictors in training and validation data

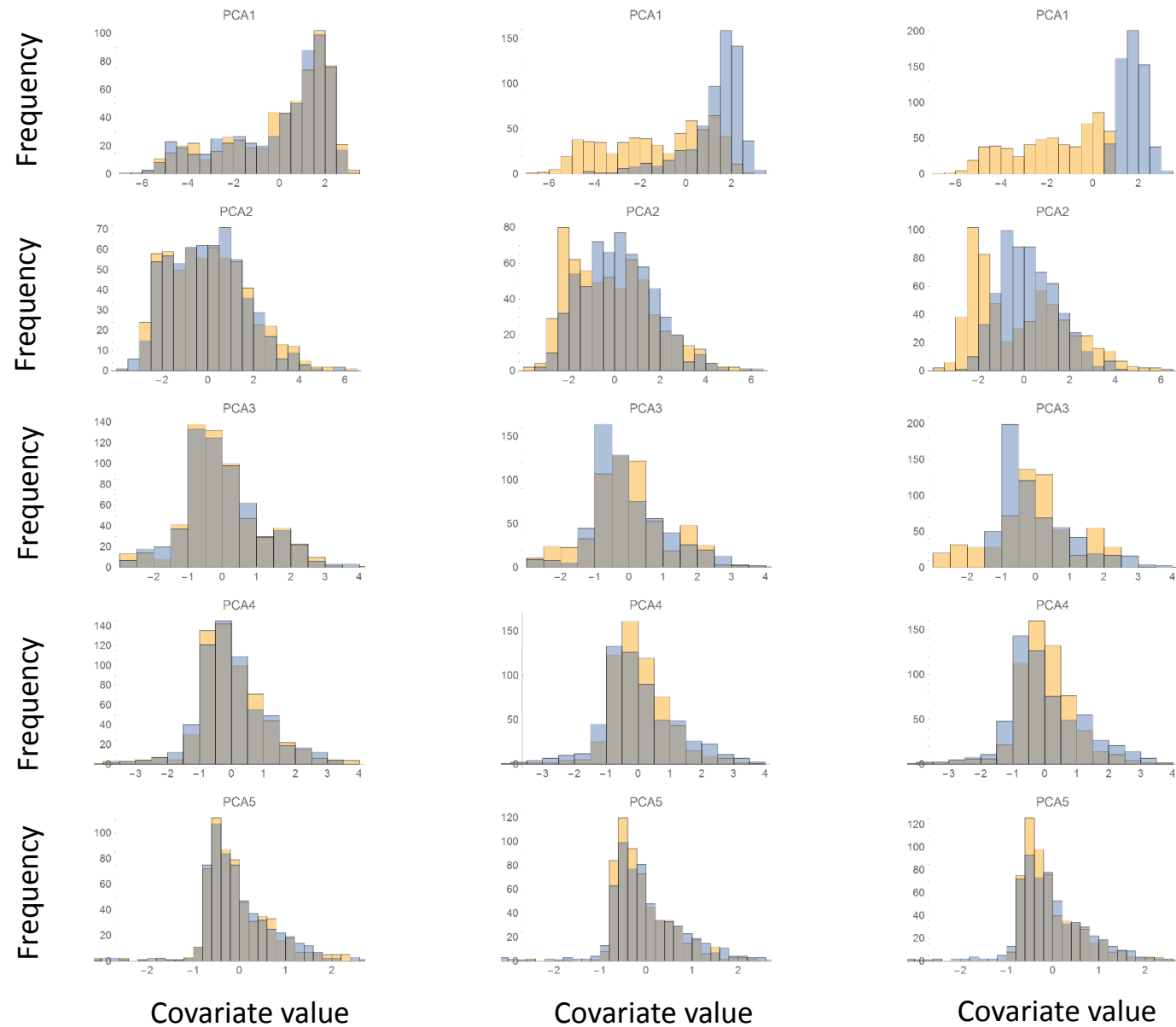
The figures below show the distributions of the environmental predictors (the principal components) and spatial predictors (the coordinates of the sampling units) for the three ways the data were split to training and validation: interpolation, partial extrapolation, and full extrapolation (see main text on how these were done). For cases related to extrapolation, the values of PC1 are lower for training data (yellow parts of the histograms) than for the validation data (blue parts of the histograms), whereas the distributions of the other principal components do not differ between training and validation. For cases related to extrapolation, the spatial distributions are also different for training sampling units (black dots) and validation sampling units (red dots).

# BIRD DATA

## INTERPOLATION

## PARTIAL EXTRAPOLATION

## FULL EXTRAPOLATION









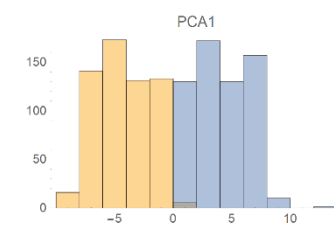
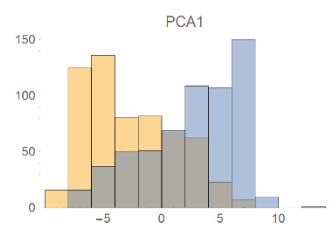
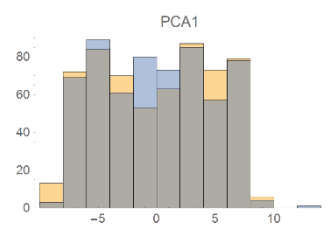
# TREE DATA

## INTERPOLATION

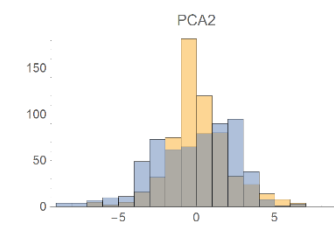
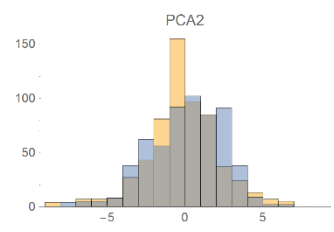
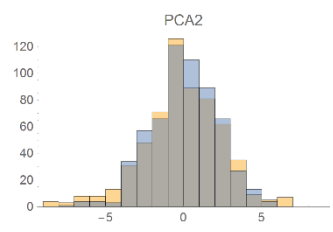
## PARTIAL EXTRAPOLATION

## FULL EXTRAPOLATION

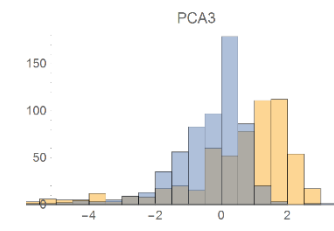
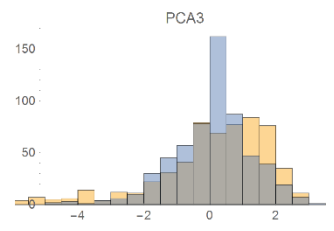
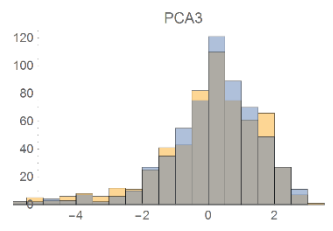
Frequency



Frequency



Frequency



Covariate value

Covariate value

Covariate value

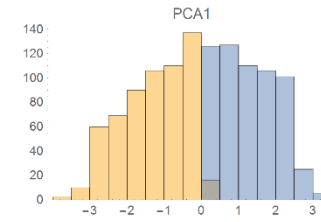
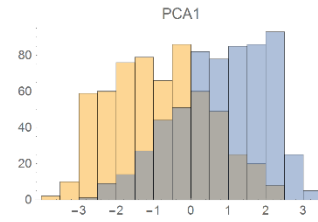
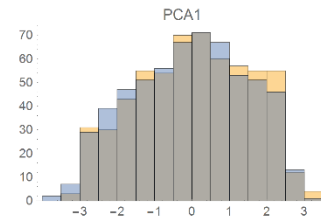
# VEGETATION DATA

INTERPOLATION

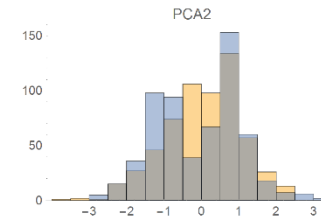
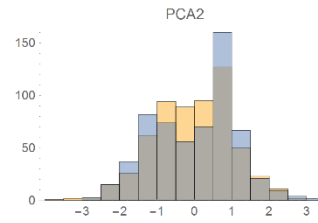
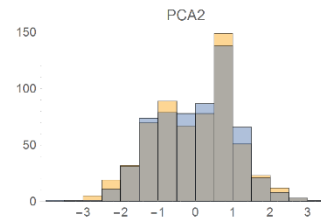
PARTIAL EXTRAPOLATION

FULL EXTRAPOLATION

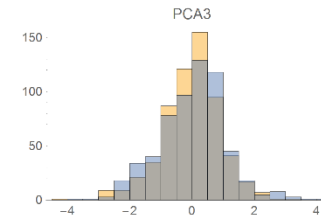
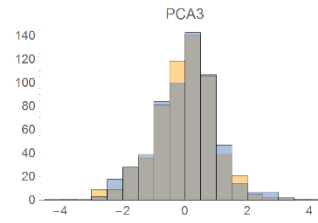
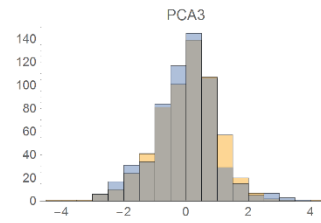
Frequency



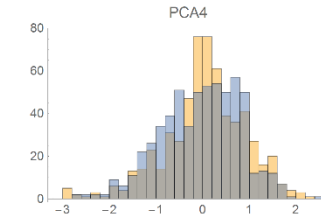
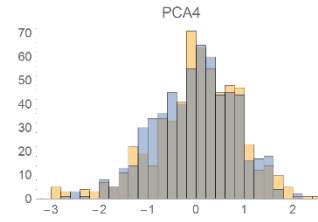
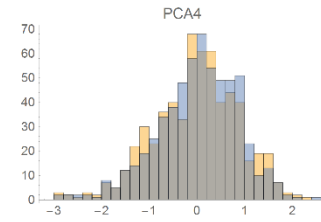
Frequency



Frequency



Frequency



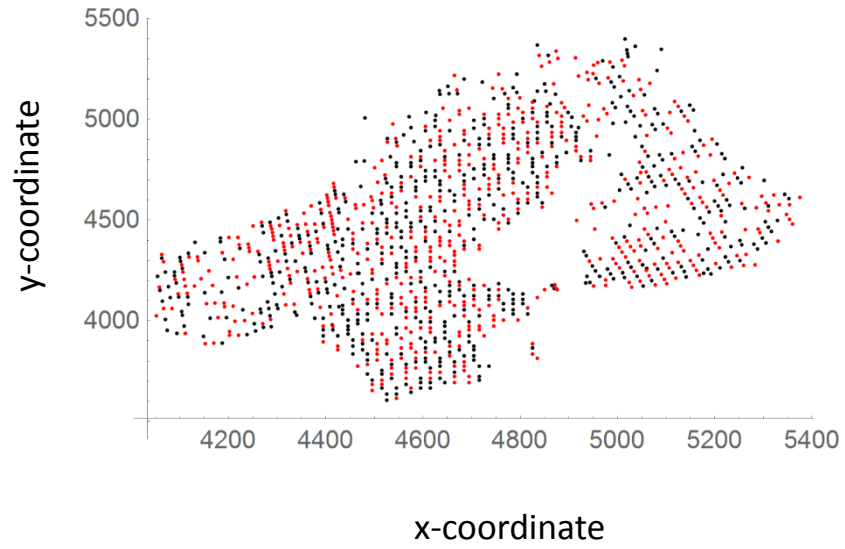
Covariate value

Covariate value

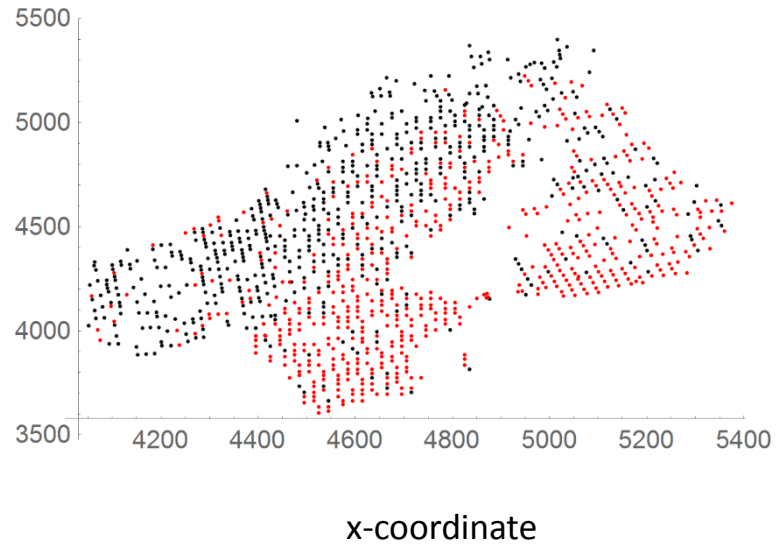
Covariate value

# BIRD DATA

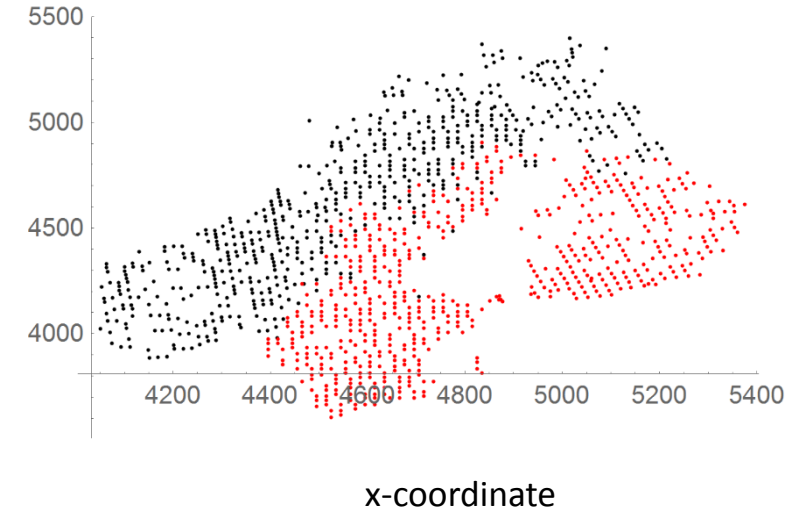
## INTERPOLATION



## PARTIAL EXTRAPOLATION

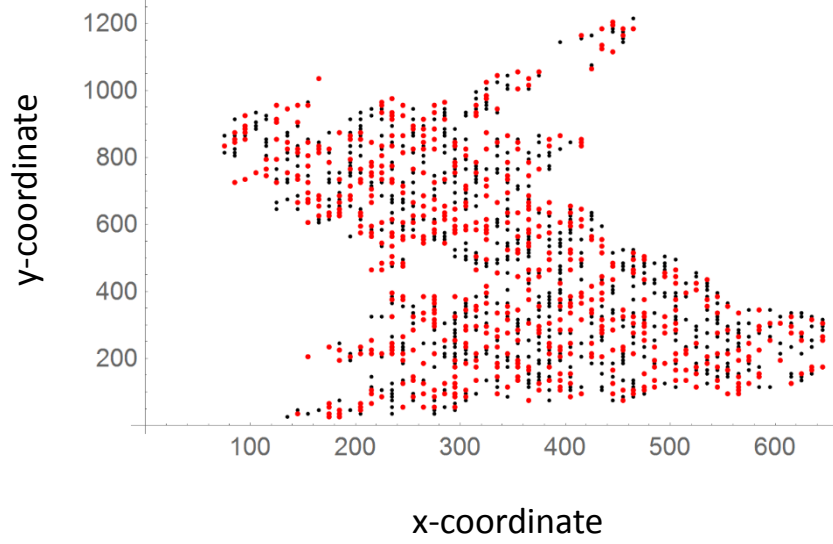


## FULL EXTRAPOLATION

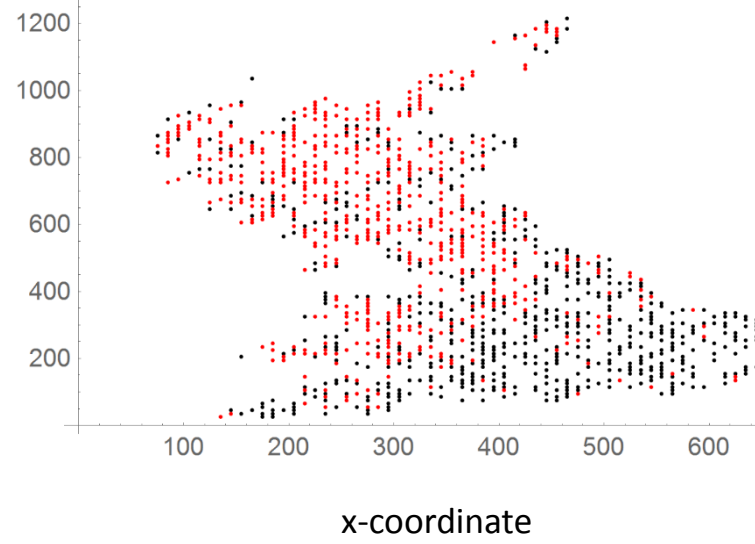


# BUTTERLY DATA

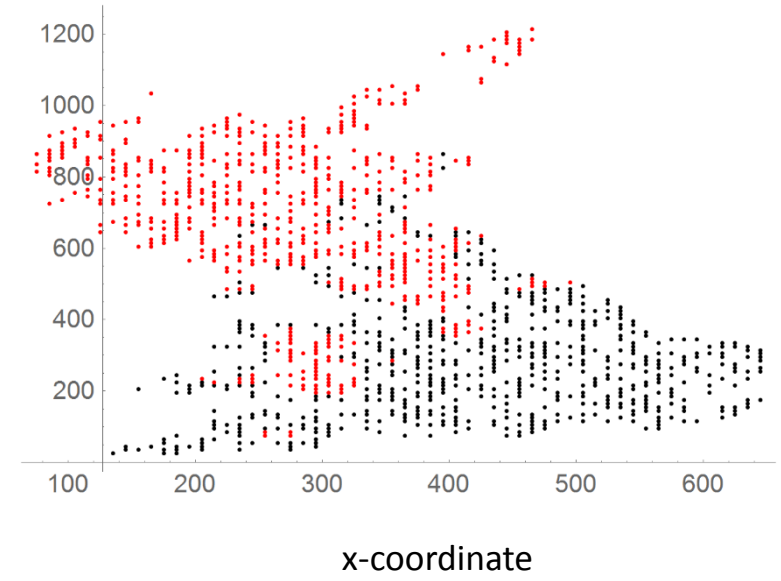
## INTERPOLATION



## PARTIAL EXTRAPOLATION

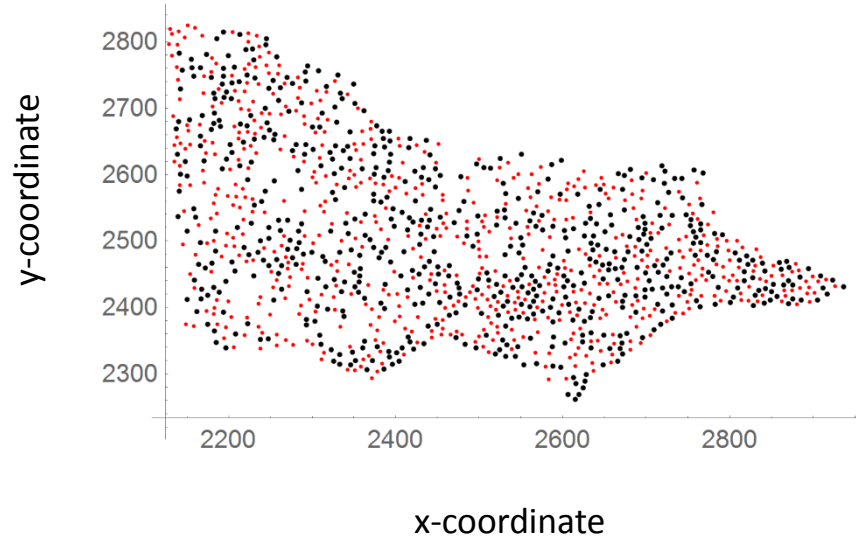


## FULL EXTRAPOLATION

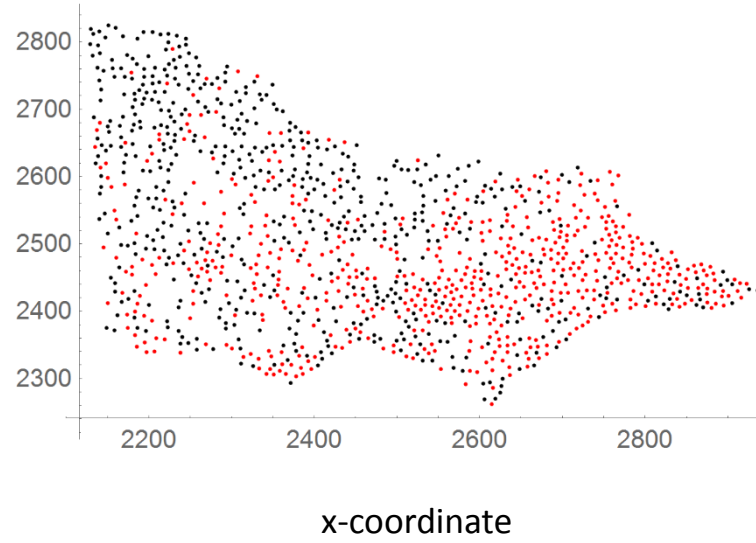


# PLANT DATA

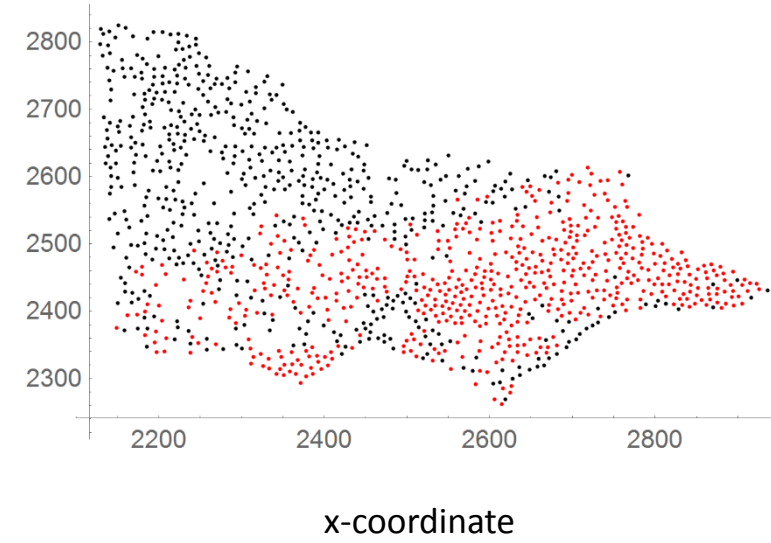
## INTERPOLATION



## PARTIAL EXTRAPOLATION

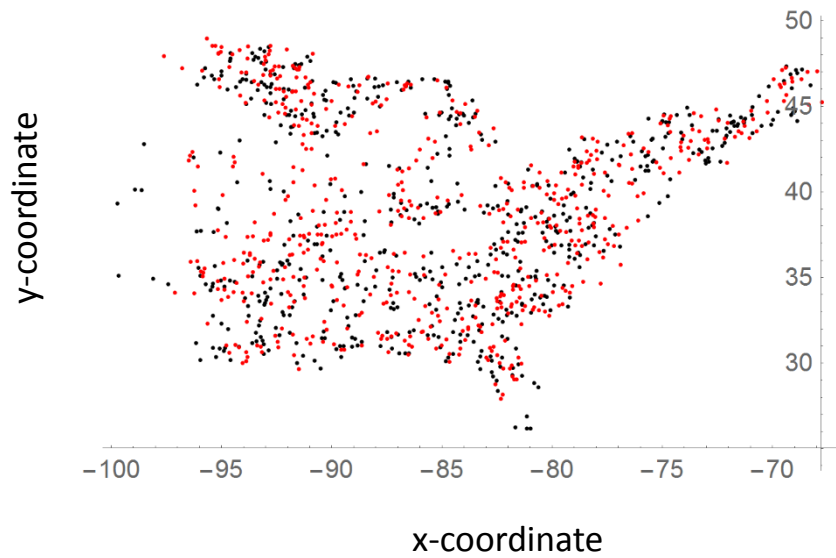


## FULL EXTRAPOLATION

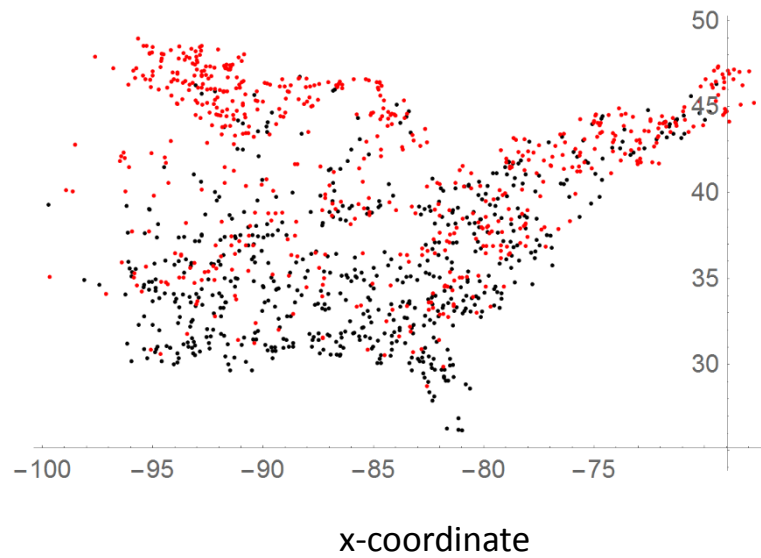


# TREE DATA

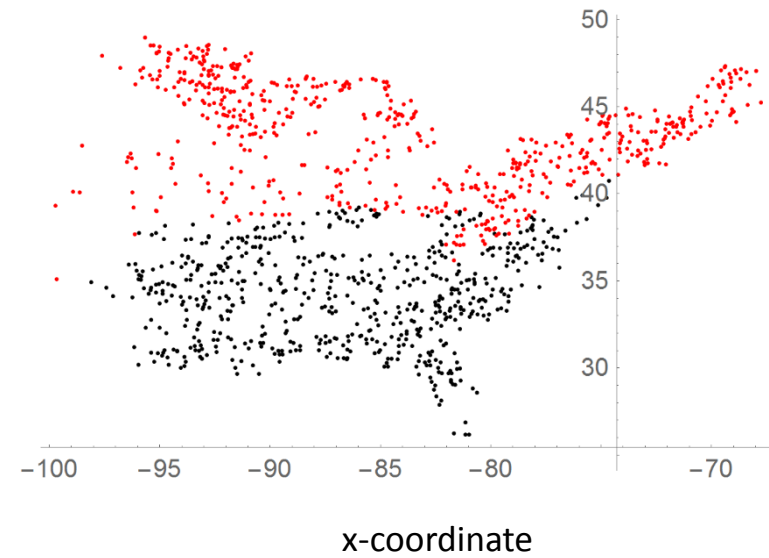
## INTERPOLATION



## PARTIAL EXTRAPOLATION



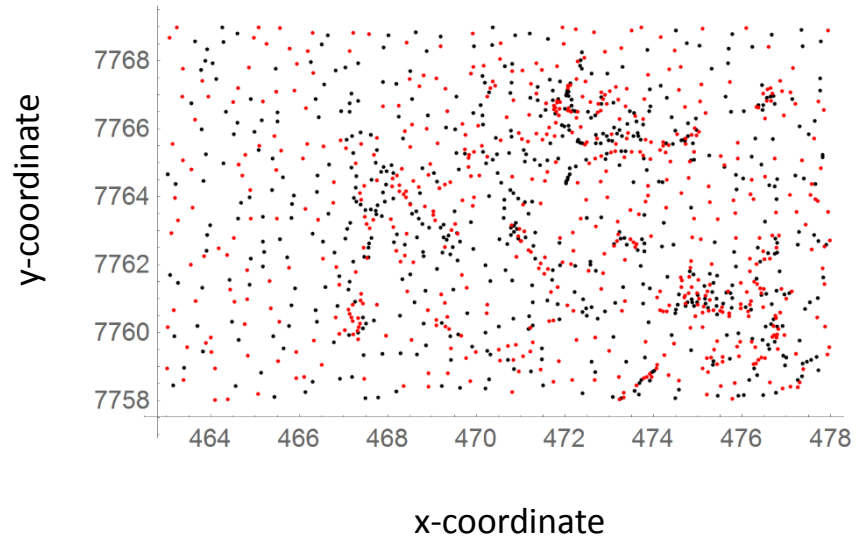
## FULL EXTRAPOLATION



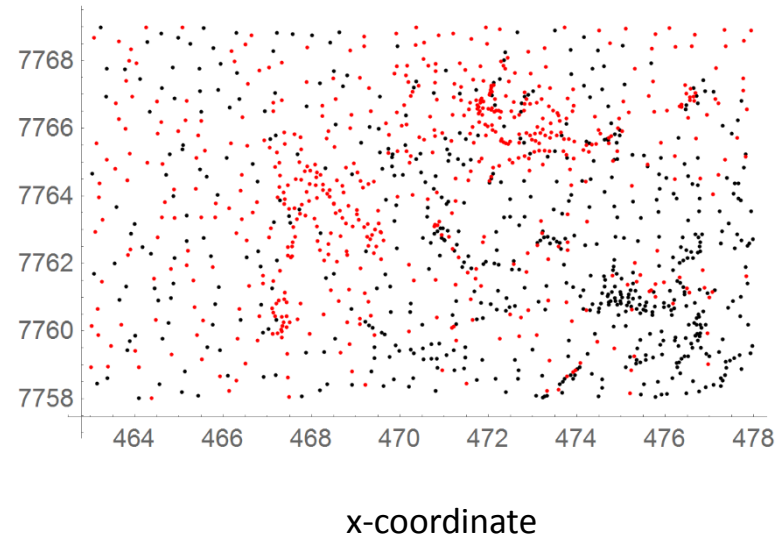


# VEGETATION DATA

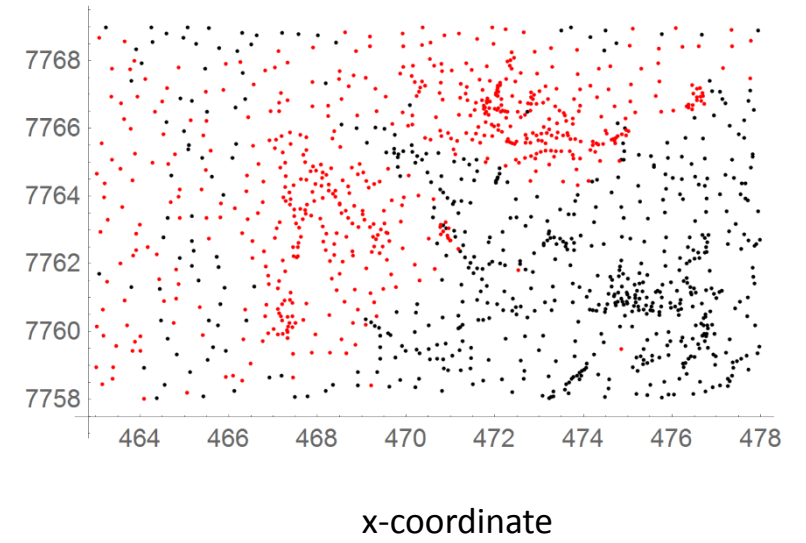
## INTERPOLATION



## PARTIAL EXTRAPOLATION



## FULL EXTRAPOLATION



**Supporting Information.** Anna Norberg, Nerea Abrego, F. Guillaume Blanchet, Frederick R. Adler, Barbara J. Anderson, Jani Anttila, Miguel B. Araújo, Tad Dallas, David Dunson, Jane Elith, Scott D. Foster, Richard Fox, Janet Franklin, William Godsoe, Antoine Guisan, Bob O’Hara, Nicole A. Hill, Robert D. Holt, Francis K.C. Hui, Magne Husby, John Atle Kålås, Aleksi Lehikoinen, Miska Luoto, Heidi K. Mod, Graeme Newell, Ian Renner, Tomas Roslin, Janne Soininen, Wilfried Thuiller, Jarno Vanhatalo, David Warton, Matt White, Niklaus E. Zimmermann, Dominique Gravel and Otso Ovaskainen. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*.

### **Appendix S3. Supporting results.**

#### **Supporting results for the meta-analysis**

In Figures S1- S8 we provide additional results based on the meta-analysis not included in the main text. Figures S1-S2 show the differences in the overall ranking of the model variants between individual data sets for all species, S3-S4 show the differences in the ranking between prediction types (interpolation and extrapolation) and data sizes ( $n = 150$  and  $n = 600$ ), and S5-S8 show the differences in ranking between the different statistical aspects.

#### **Raw results for the performance measures**

In Figures S9-S10 we provide the raw results for all the performance measures. The measures are pooled for all data sets and prediction types, and thus the figures illustrate the range of values gained for the measures (vertical axis). The models are organised according to their overall ranking (horizontal axis). For all the measures of discrimination, a larger the measure value indicates better performance. For all the rest of the measures, a smaller value indicates better performance, except that for the sharpness this is conditional on how the model performs regarding other measures. See Table 3 and the main text for explanations of the performance measures.

In Figure S11 we show the variation among data sets and prediction types regarding the commonly used measure for predictive performance AUC (area under curve) in more detail. Note that the same performance measure is shown for the pooled results in figure S9 as discrimination of species-specific occurrence probabilities. Figure S11 illustrates the variation among the data sets in terms of model ranking (horizontal axis) and among interpolative and extrapolative predictions (vertical axis). Black points indicate interpolation, grey partial extrapolation and white full extrapolation.

#### **Convergence of MCMC chains for Bayesian models**

To examine the level of MCMC convergence, we fitted all Bayesian models twice, and computed the correlation among the predicted species occurrence probabilities between the two MCMC chains.

While in the application of a single modelling framework MCMC convergence should be examined based on the primary model parameters, we chose to base our analyses on predicted occurrence probabilities as that provided a common currency among all Bayesian models that differ in their primary parameters. As we used each model variant to five datasets, three prediction tasks, and two data sizes, we obtained for each model variant 30 correlations among the predicted occurrence probabilities between the two MCMC chains. Figure S12 indicates that MCMC convergence was good in the model variants BC.1, BC.2, BORAL.1 and GLM.7, intermediate in the model variants GLM.4, GLM.5, GLM.13, HMSC.1, HMSC.2, HMSC.3 and HMSC.4, and poor in the model variant GJAM.1.

Running a longer MCMC chain for the models with intermediate and especially with poor convergence could have resulted in better predictive performance. However, achieving satisfactory convergence could require excessive model fitting times, making it difficult to achieve convergence in practice. While lack of MCMC convergence makes parameter inference based on parameter estimates unreliable (e.g. because credible intervals may be biased), it does not necessarily compromise much predictive performance. For example, in our comparisons HMSC.3 was generally found to be the model with highest predictive performance in spite of the fact that it did not fully converge. Moreover, moderate lack of MCMC convergence is comparable to not finding the true maximum likelihood estimate with optimization, which is a typical problem in models with many parameters.

### Failed models

For some single species models the model fitting failed, due to e.g. not reaching convergence. In table S1 (below), the proportion (%) of failed models are shown for all the data sets and all the model variants for which some individual model fits failed.

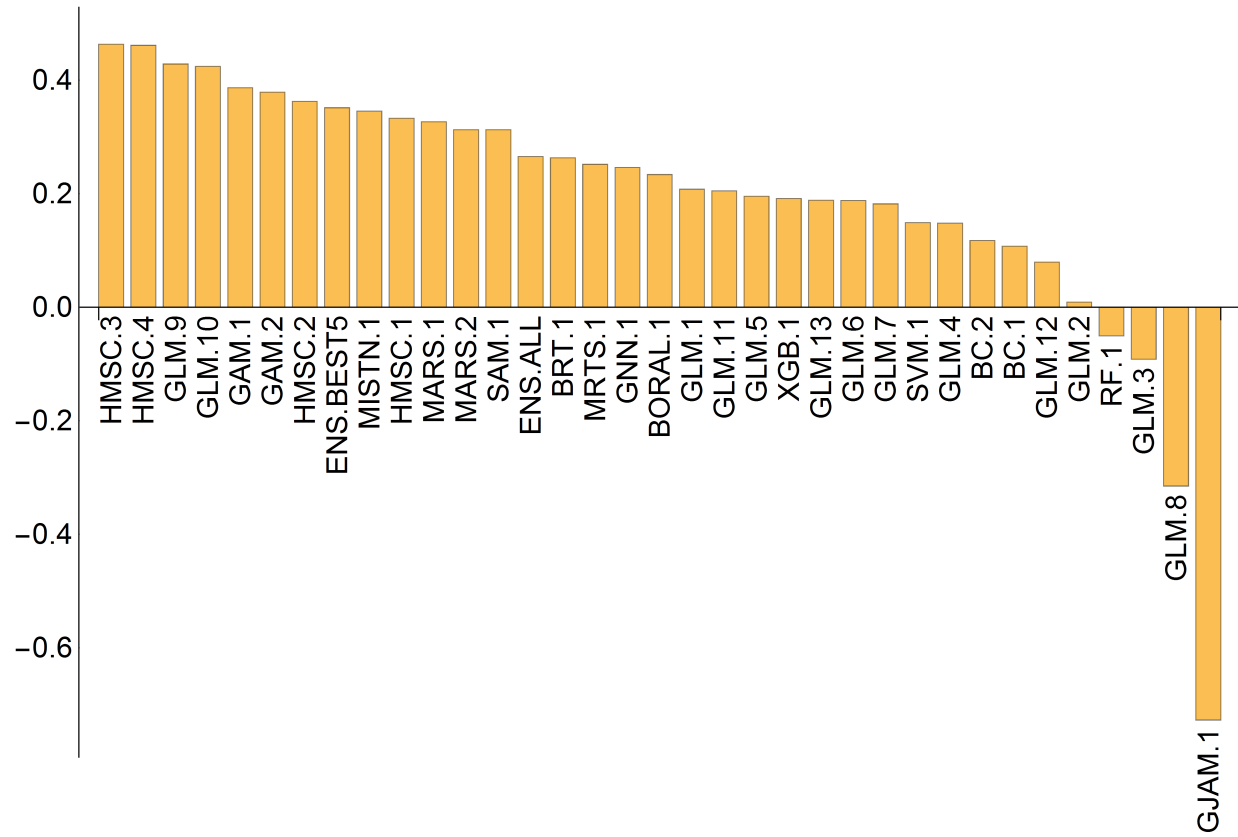
Model	Taxon	Interpolation	Partial extrapolation	Full extrapolation	Data_size
SVM.1	birds	4	9	7	150
	butterfly	2	0	16	150
	plant	20	16	16	150
	trees	13	6	6	150
	vegetation	13	8	6	150
	birds	0	0	1	600
	butterfly	0	0	4	600
	plant	1	0	1	600
	trees	0	0	0	600
	vegetation	0	0	0	600
GLM.10	birds	4	8	7	150
	butterfly	2	0	16	150
	plant	20	15	13	150

	trees	10	6	5	150
	vegetation	12	7	6	150
	birds	0	0	1	600
	butterfly	0	0	4	600
	plant	1	0	1	600
	trees	0	0	0	600
	vegetation	0	0	0	600
GLM.2	birds	4	4	6	150
	butterfly	0	0	10	150
	plant	20	15	13	150
	trees	10	5	10	150
	vegetation	8	8	4	150
	birds	0	0	1	600
	butterfly	0	0	0	600
	plant	0	0	0	600
	trees	2	2	0	600
	vegetation	0	0	0	600
GLM.3	birds	57	35	40	150
	butterfly	64	70	66	150
	plant	59	68	77	150
	trees	43	44	43	150
	vegetation	45	44	43	150
	birds	48	55	56	600
	butterfly	56	50	40	600
	plant	54	55	53	600
	trees	51	30	30	600
	vegetation	45	40	30	600
GAM.2	birds	45	35	38	150
	butterfly	60	56	76	150
	plant	72	65	69	150
	trees	25	19	11	150
	vegetation	38	28	31	150
	birds	28	27	28	600
	butterfly	60	42	38	600
	plant	56	62	58	600
	trees	13	10	8	600
	vegetation	33	18	19	600

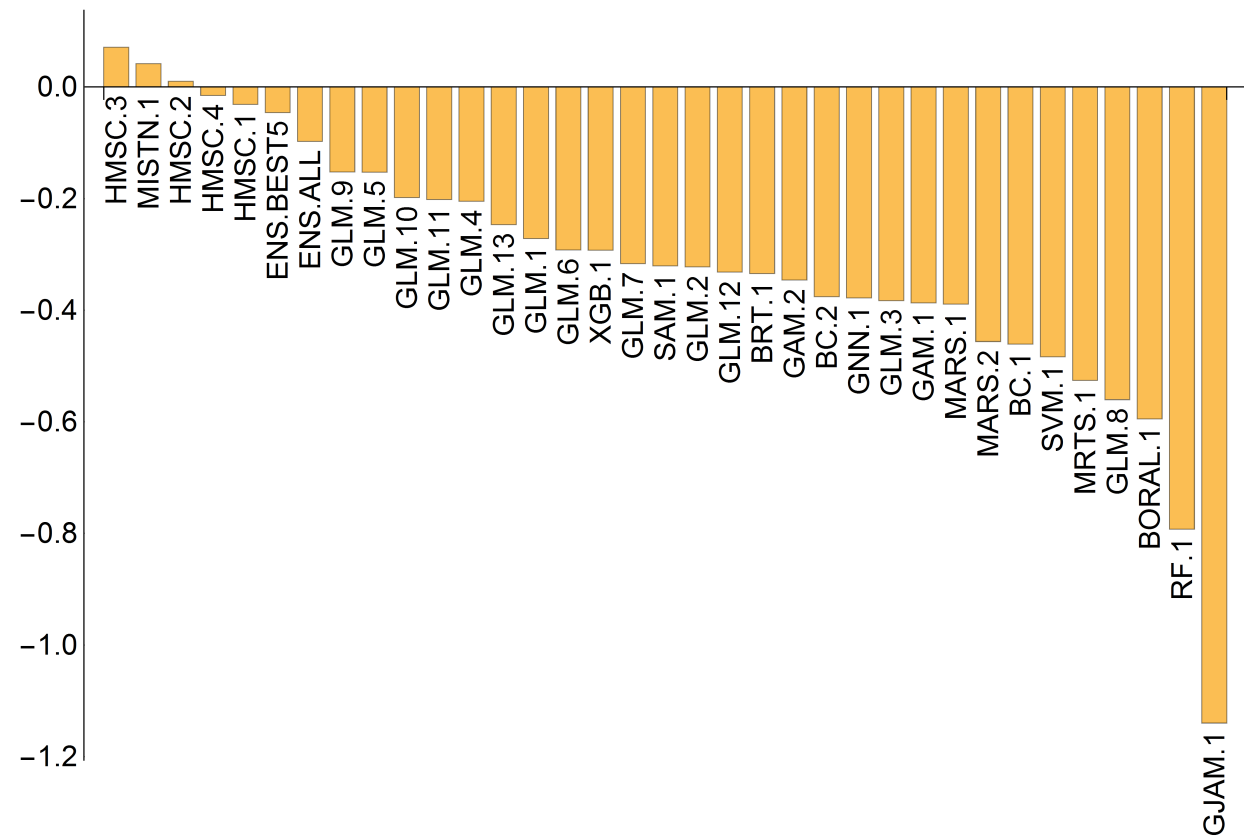
# S1 OVERALL RANKING

community data

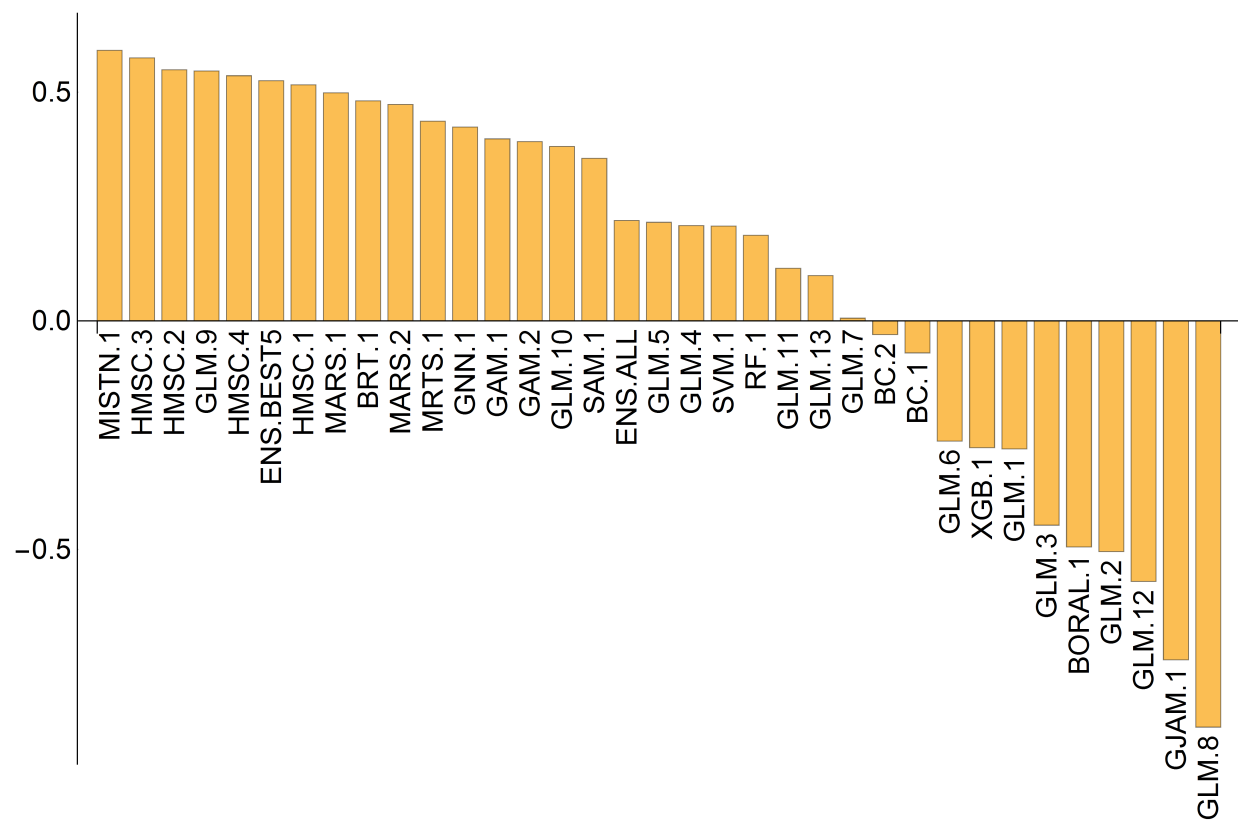
## Trees



## Vegetation



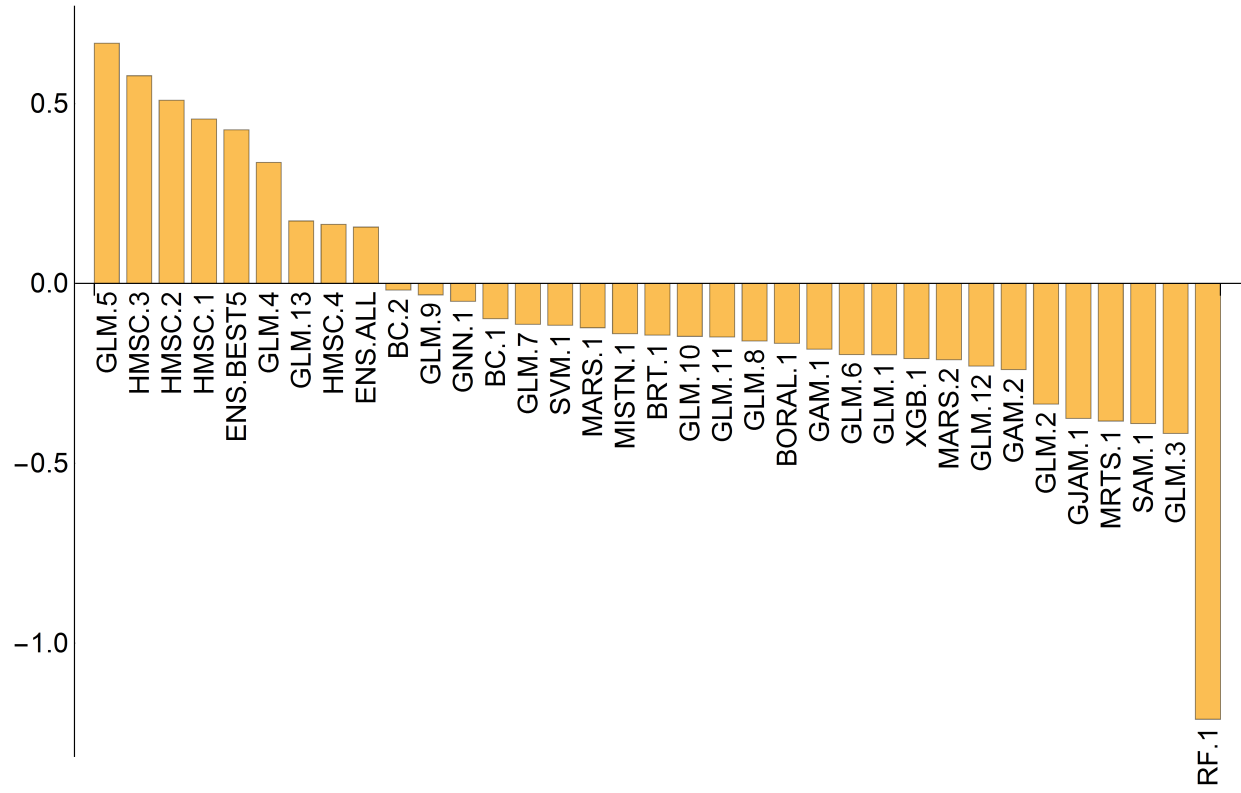
## Plants



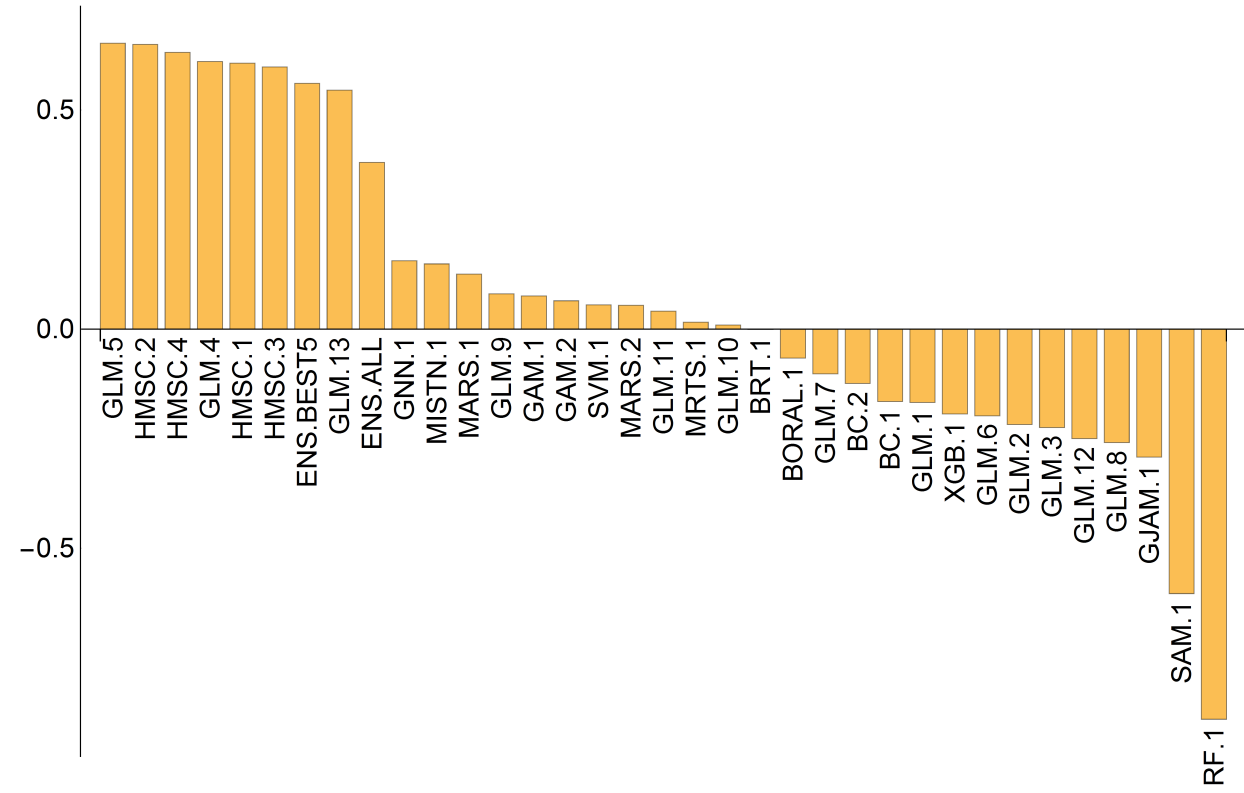
# S2 OVERALL RANKING

## assemblage data

### Butterflies



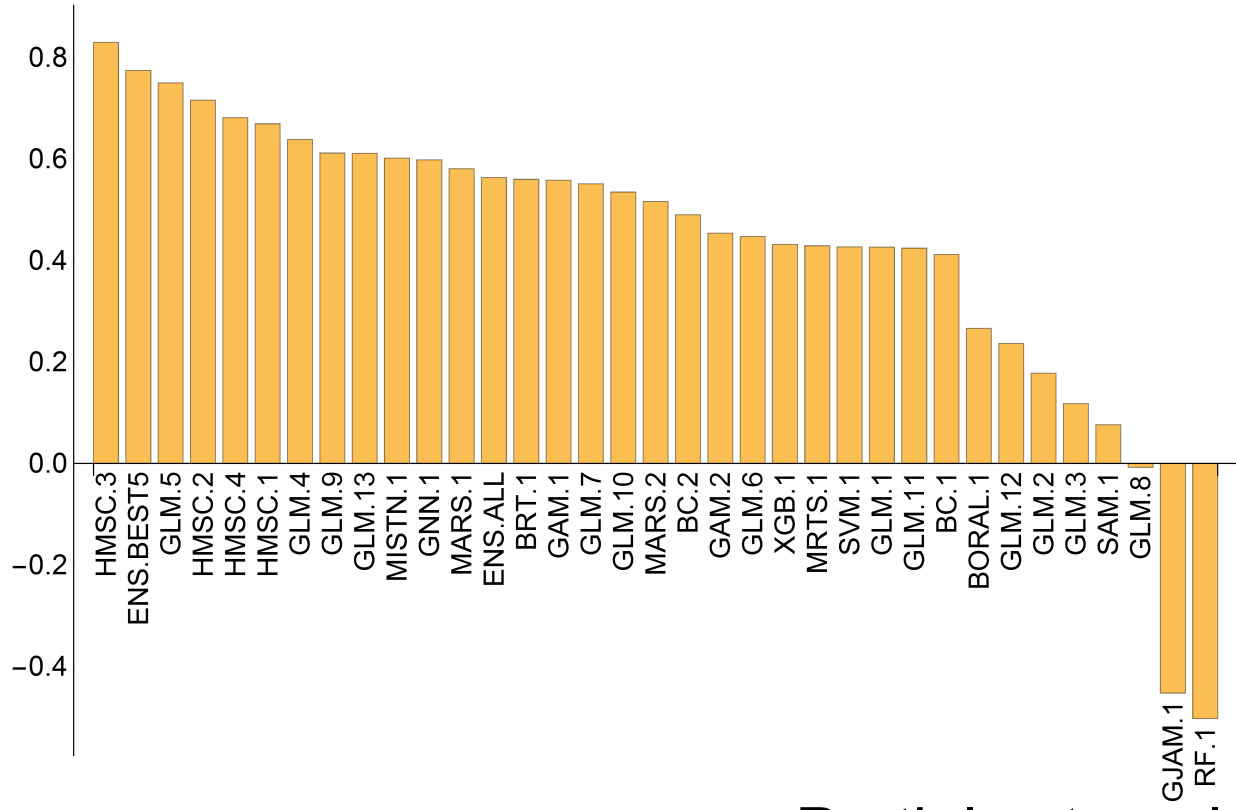
### Birds



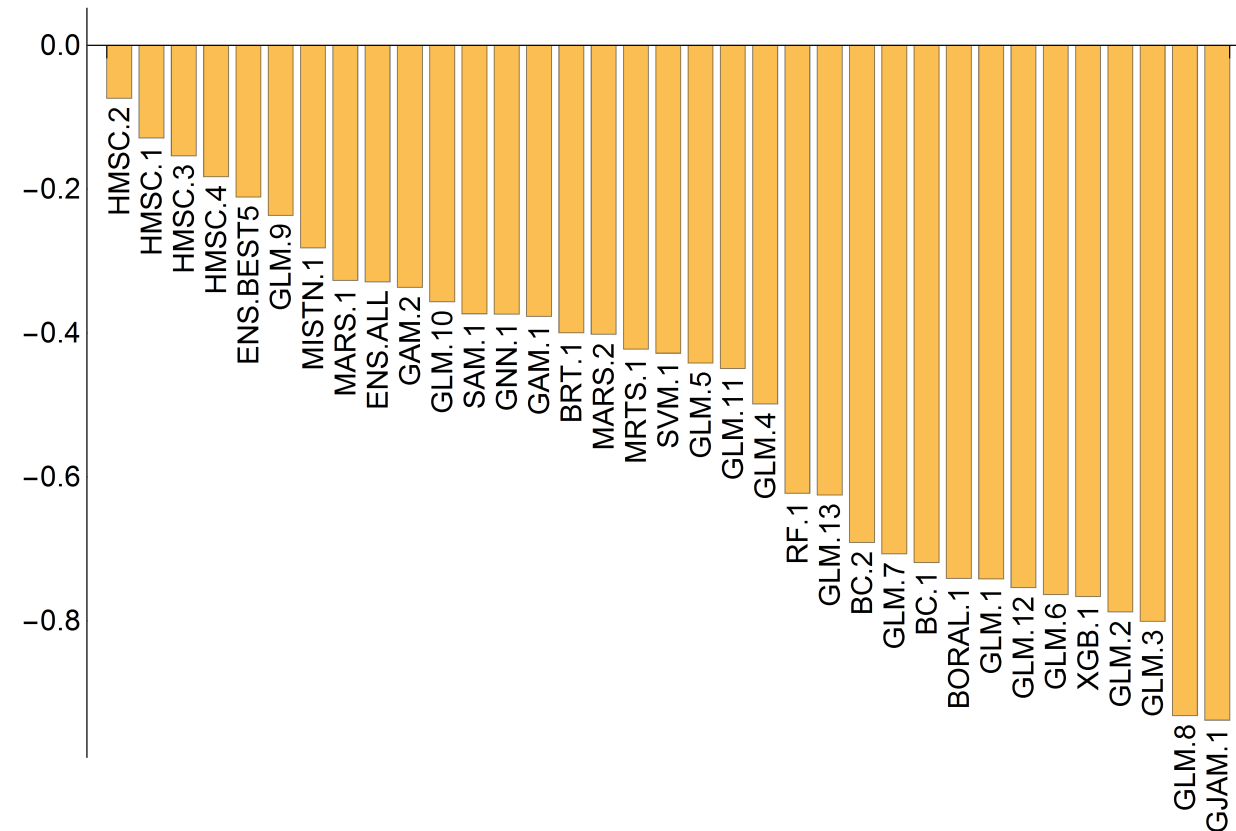


# S3 PREDICTION TYPE

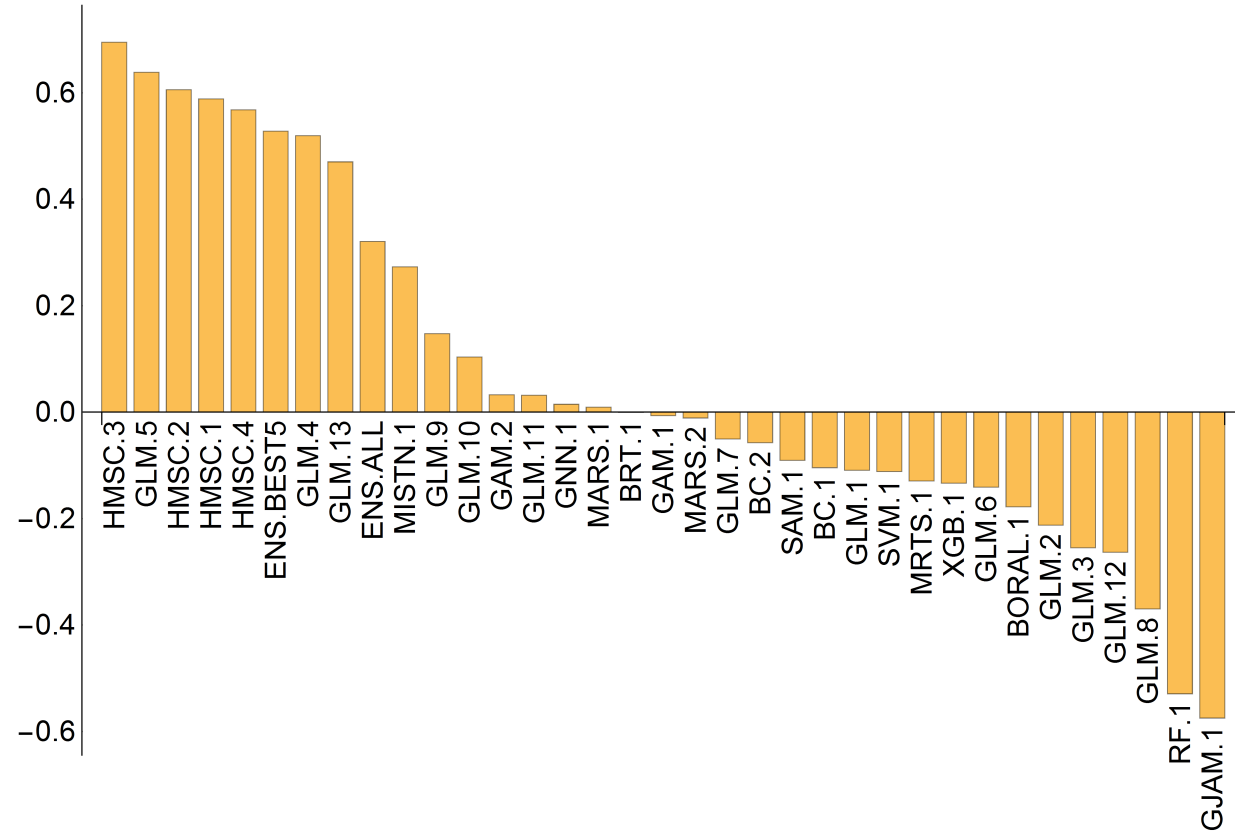
## Interpolation



## Full extrapolation

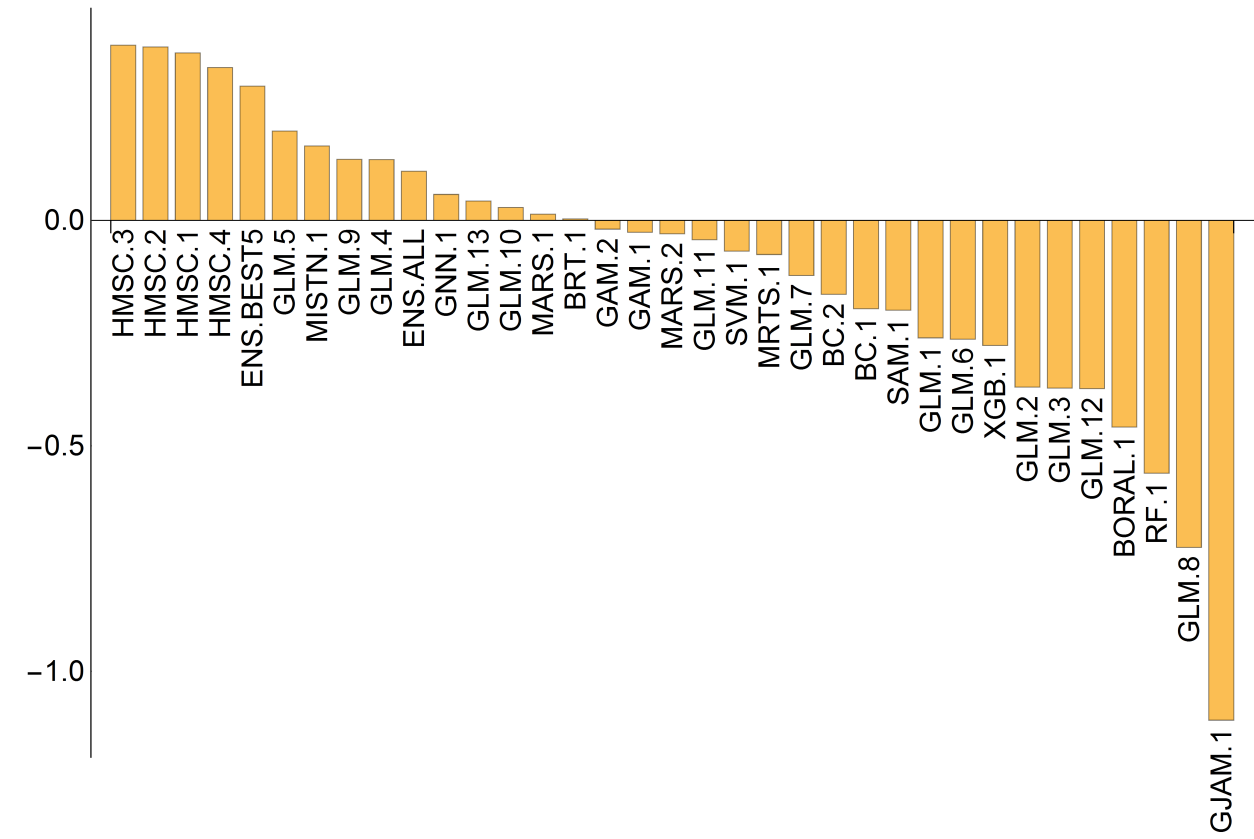


## Partial extrapolation

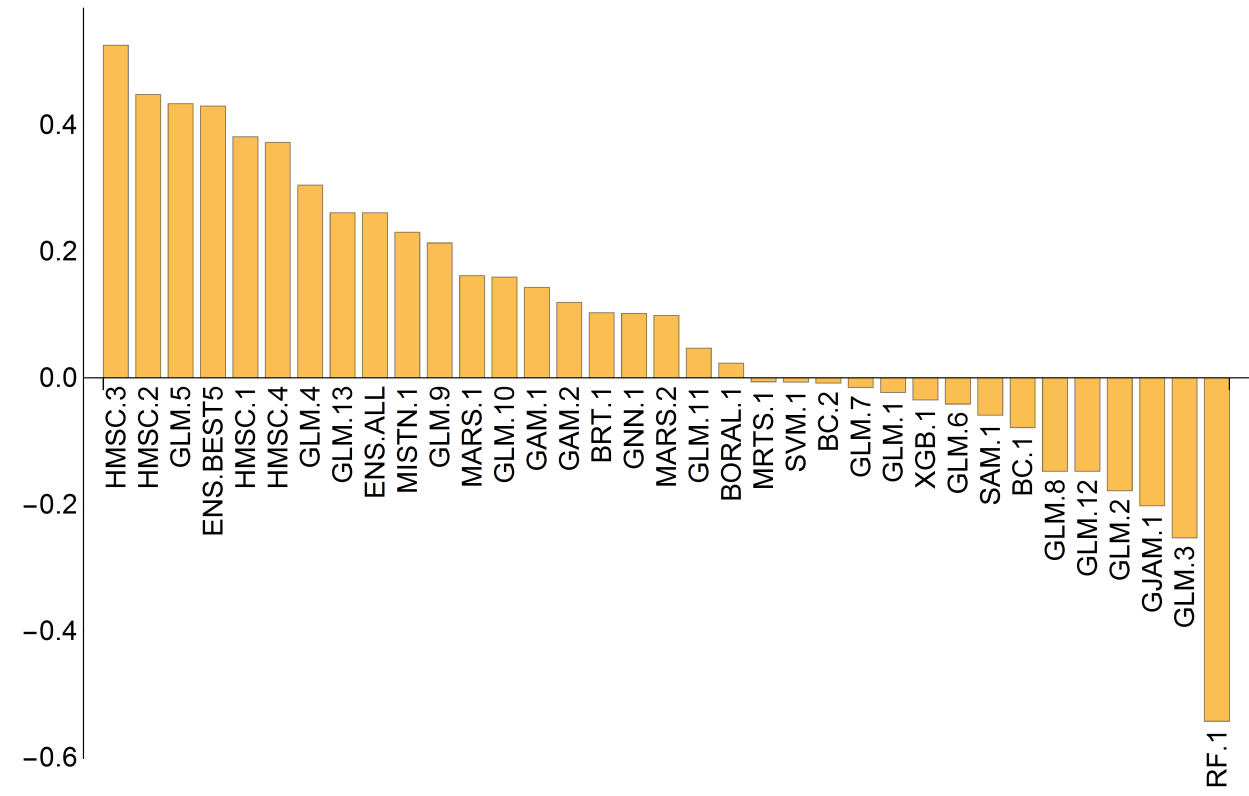


# S4 DATA SIZE

## Small ( $n = 150$ )

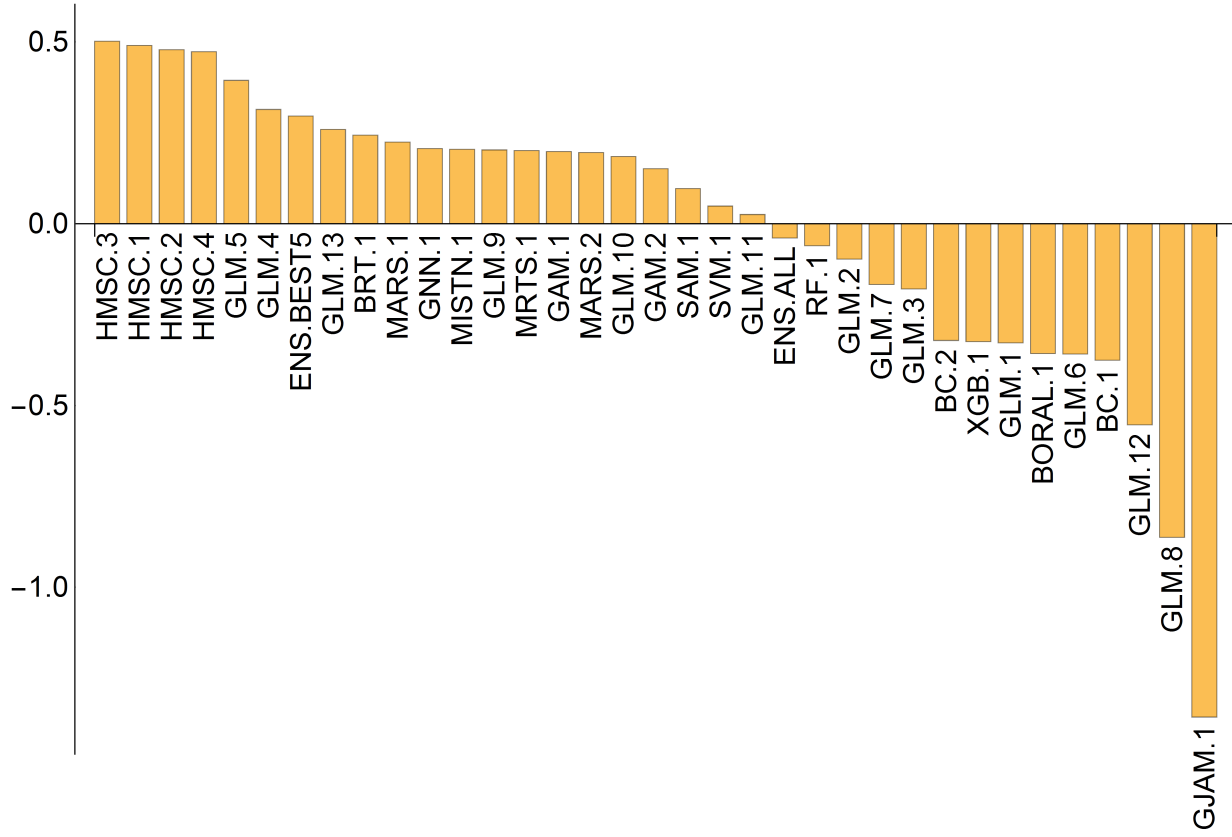


## Large ( $n = 600$ )

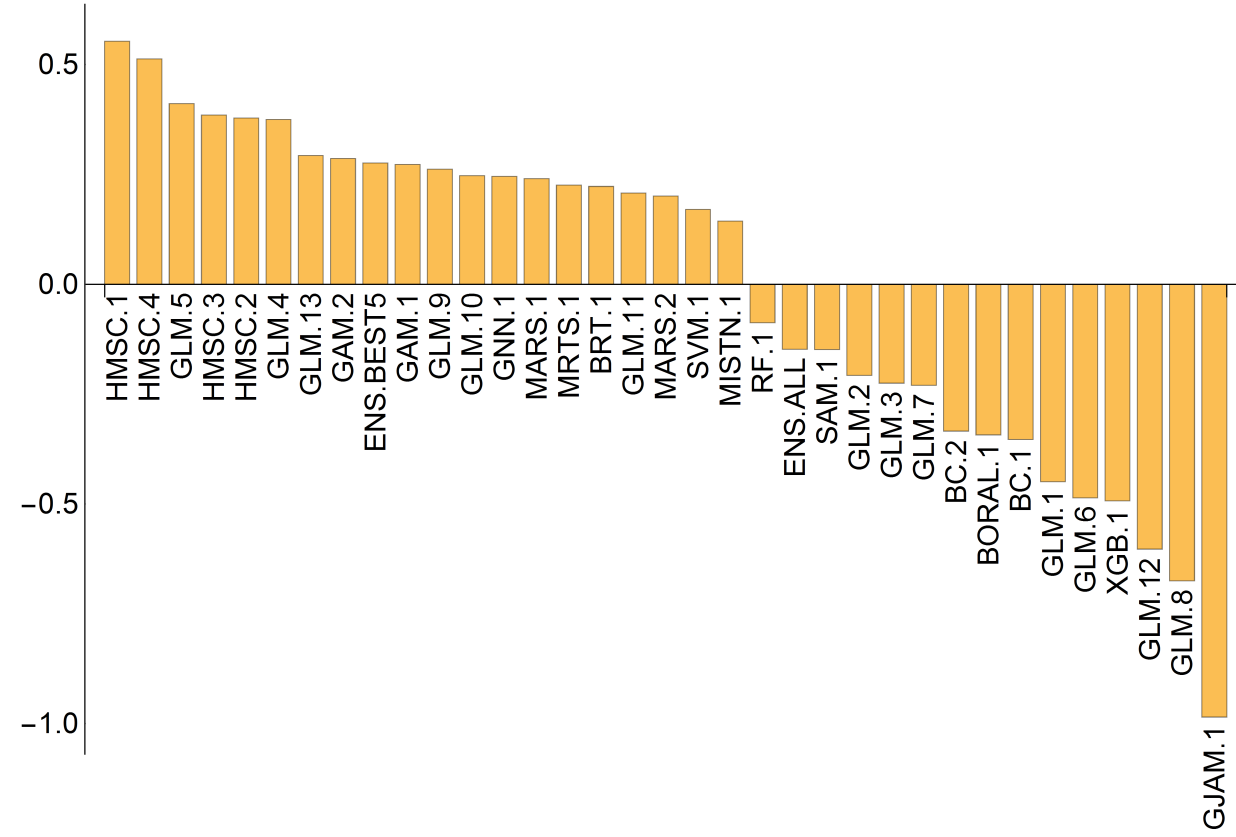


# S5 ACCURACY

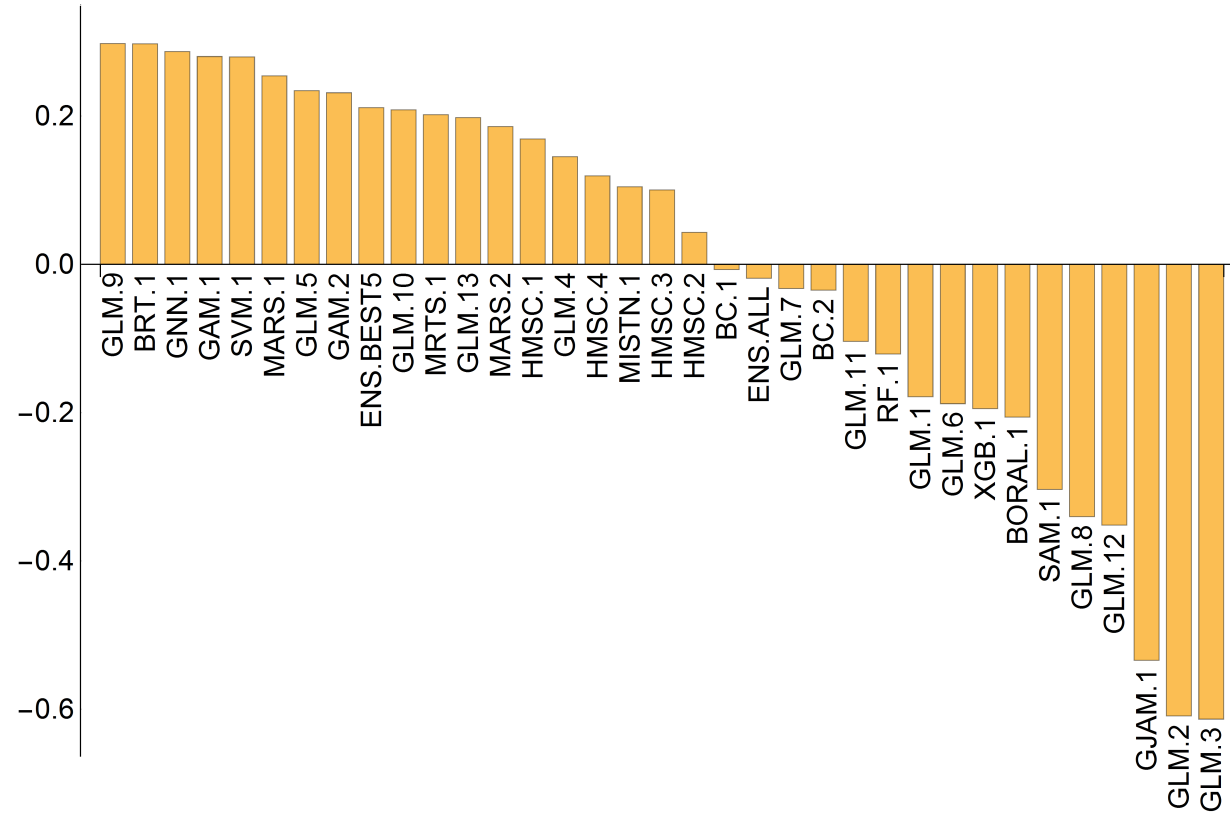
## 1 Species level



## 2 Species richness

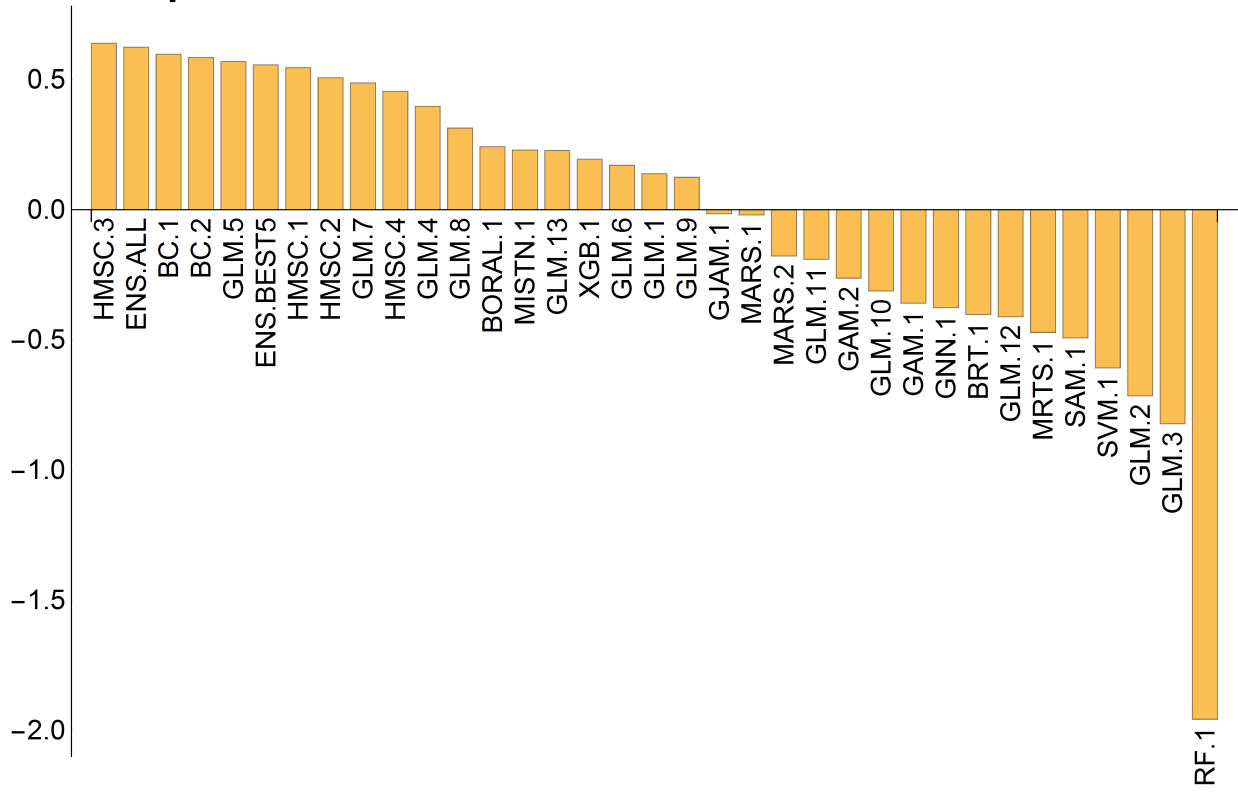


## 3 Community composition

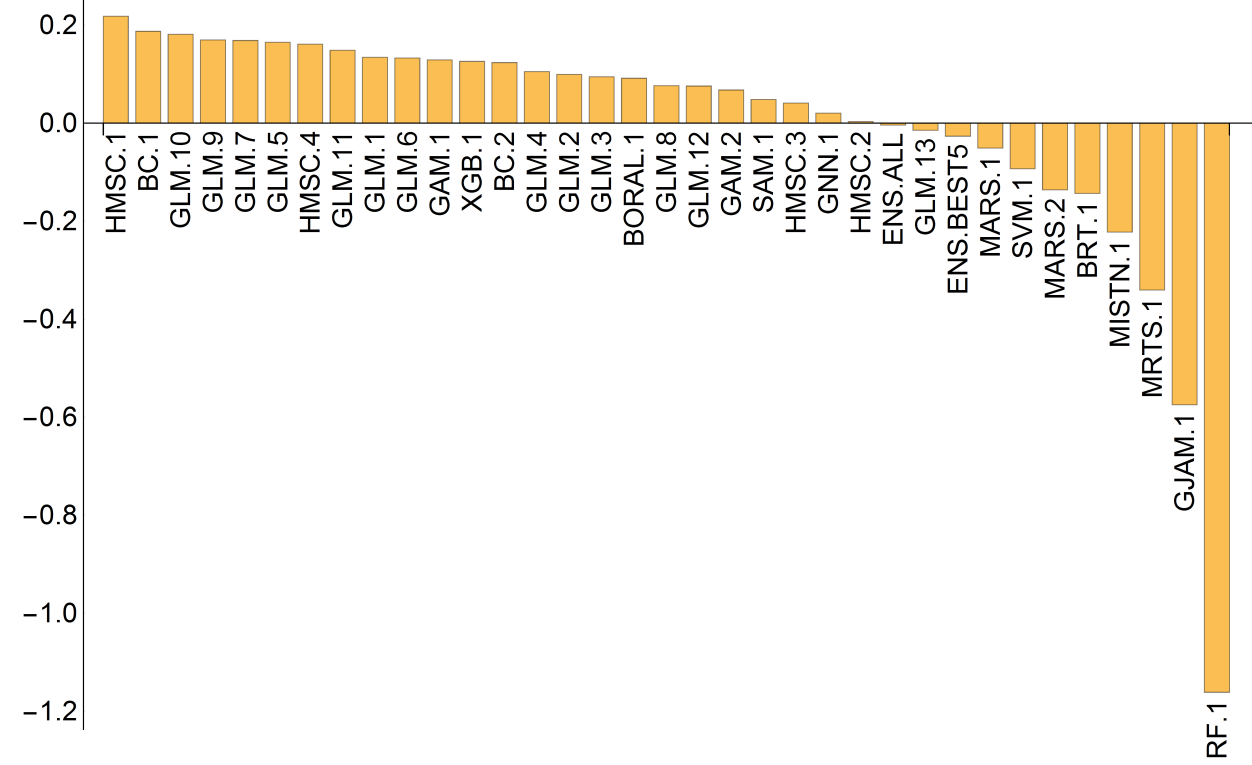


# S6 DISCRIMINATION

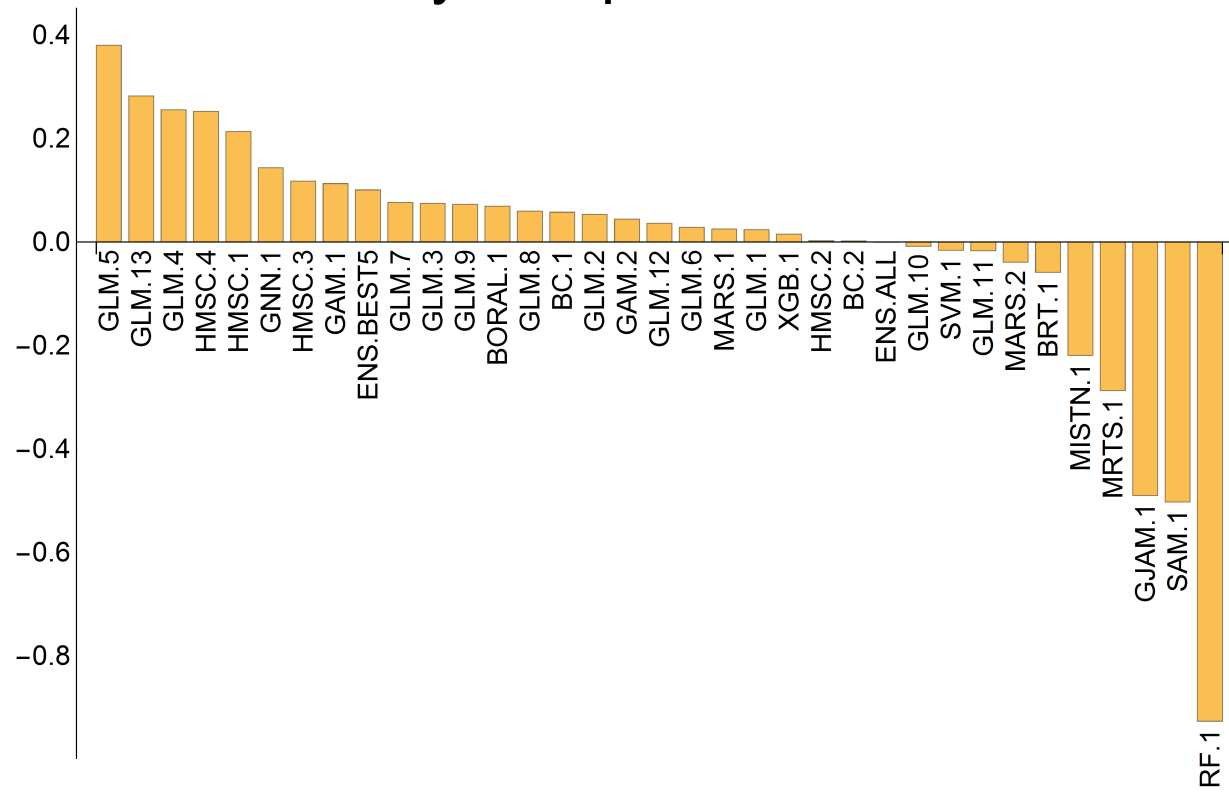
## 1 Species level



## 2 Species richness

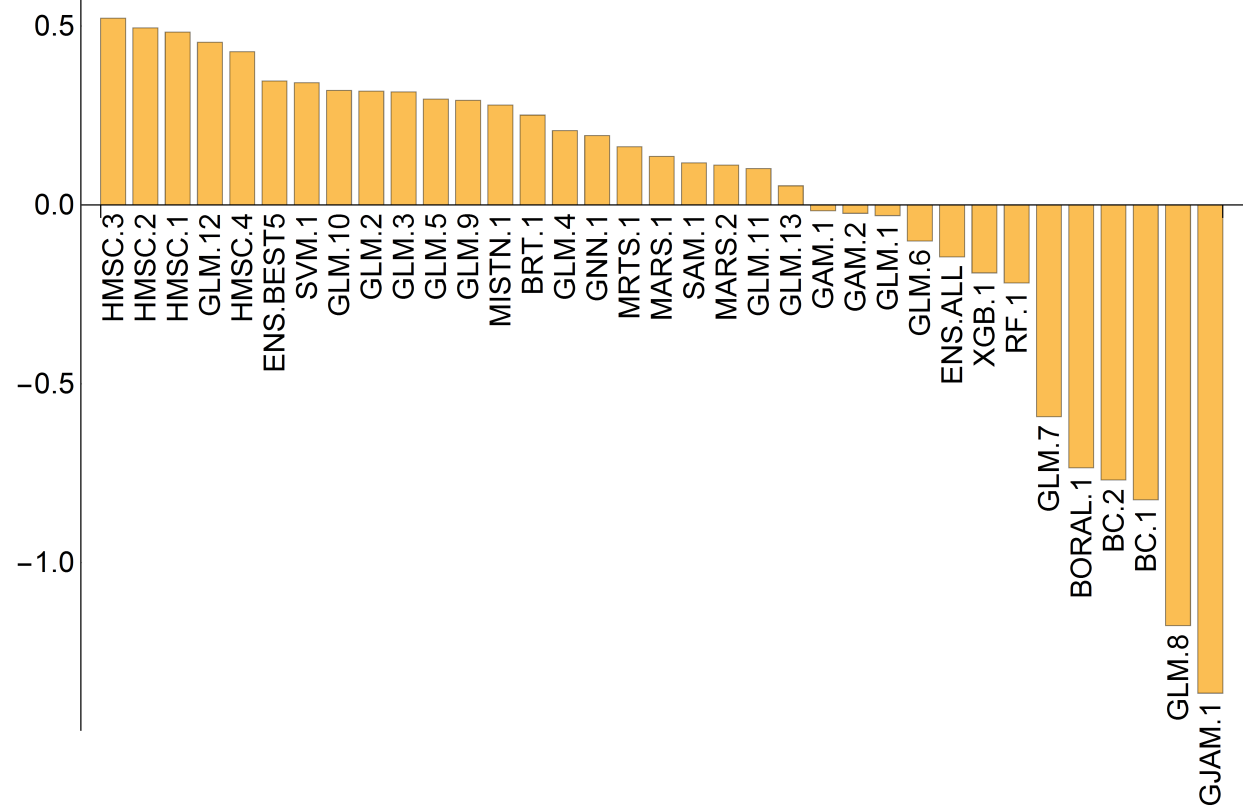


## 3 Community composition

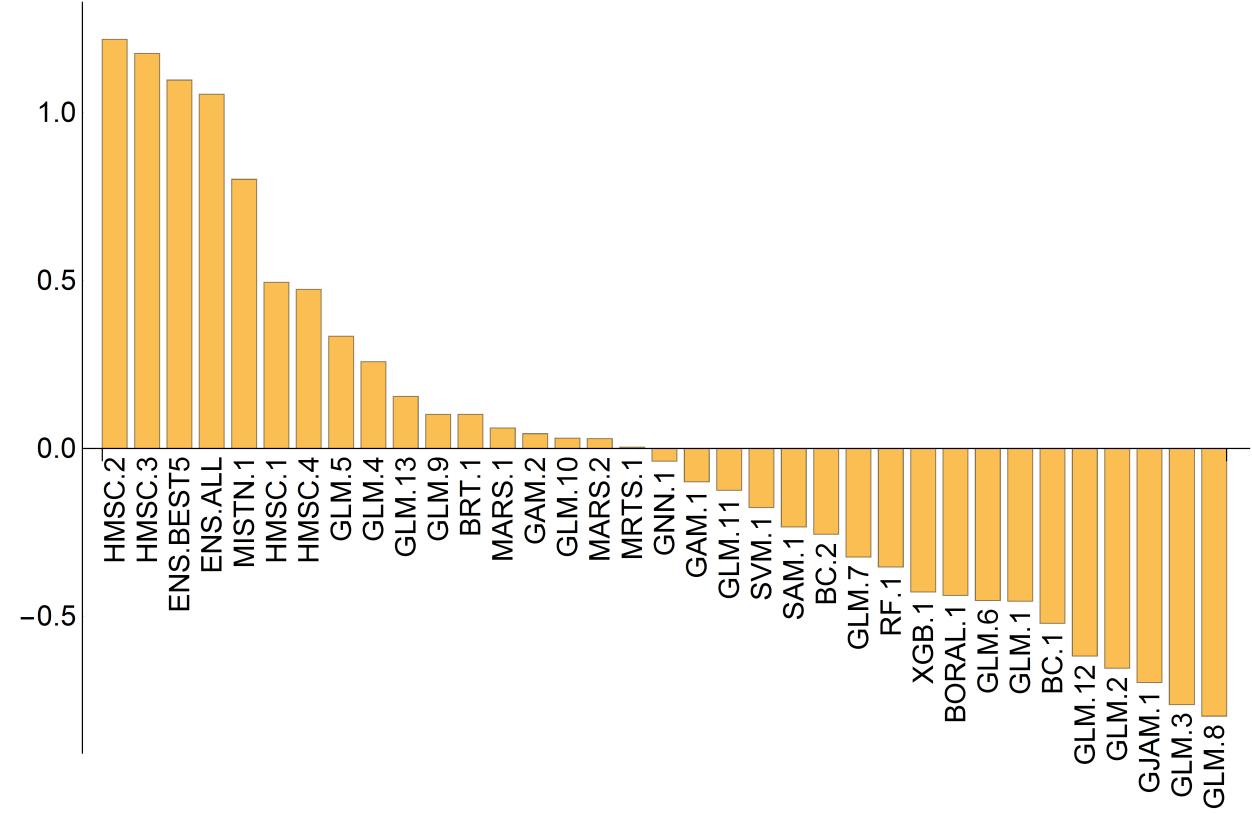


# S7 CALIBRATION

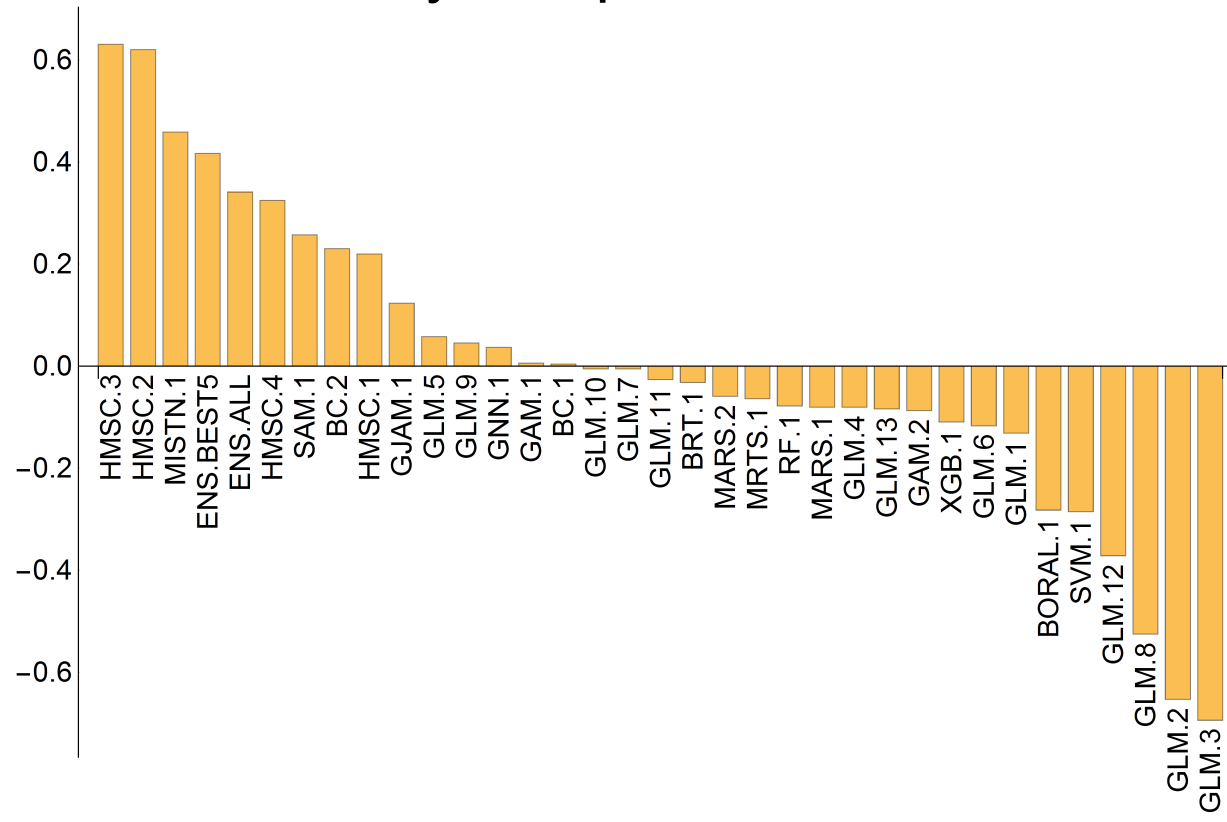
## 1 Species level



## 2 Species richness

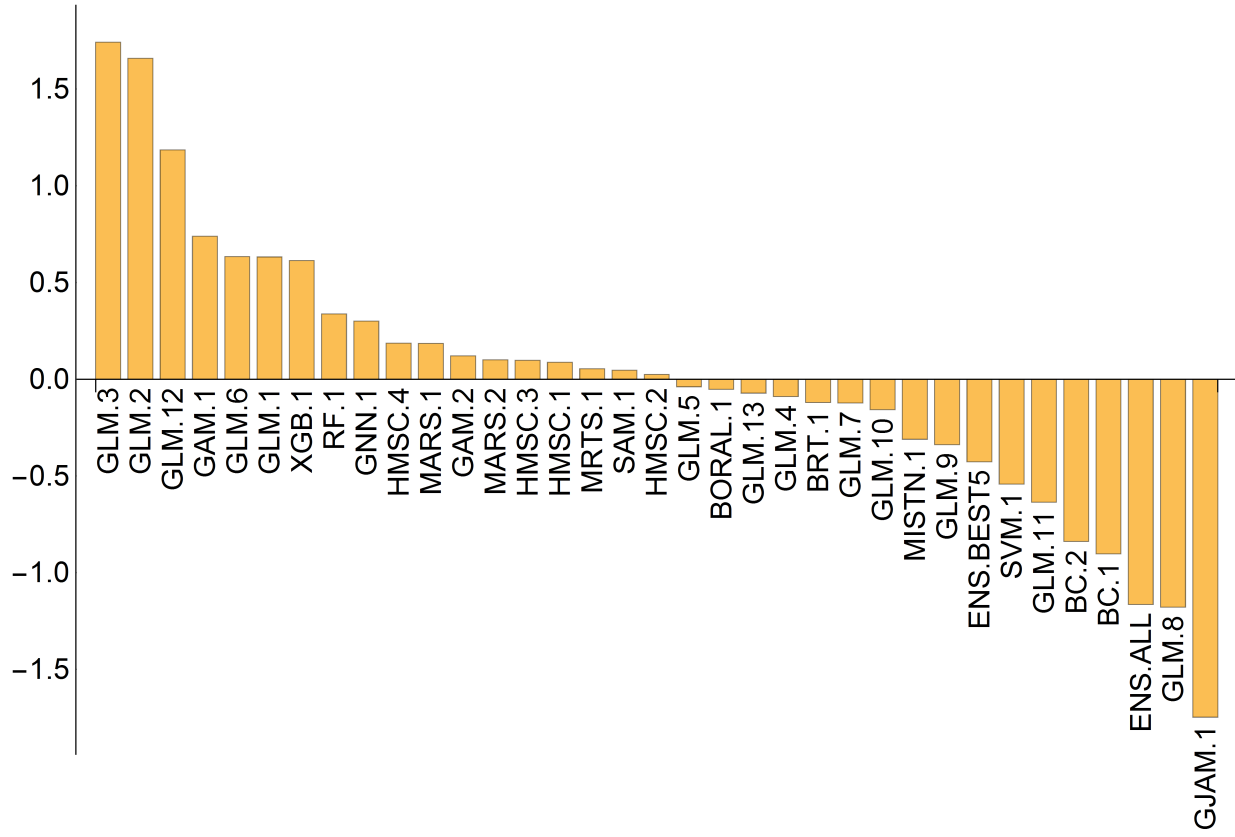


## 3 Community composition

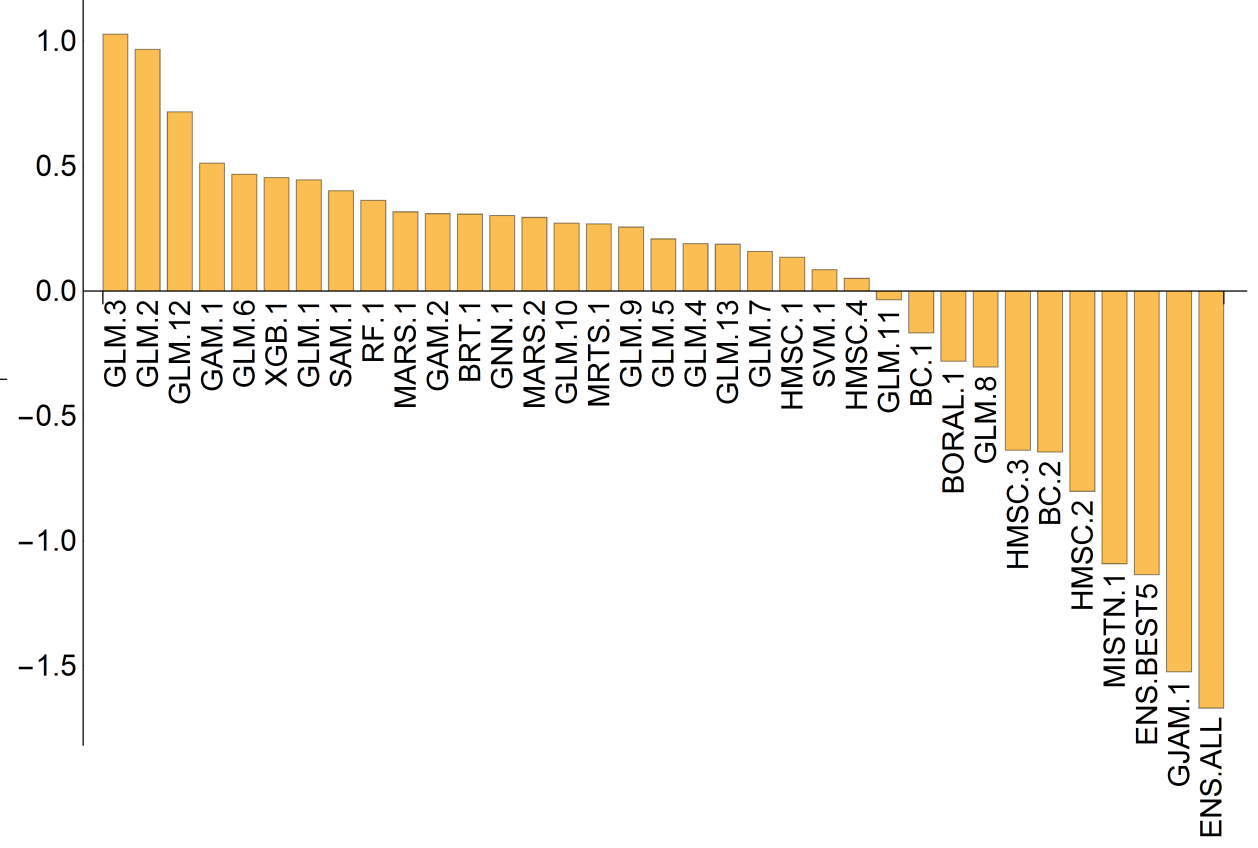


# S8 PRECISION

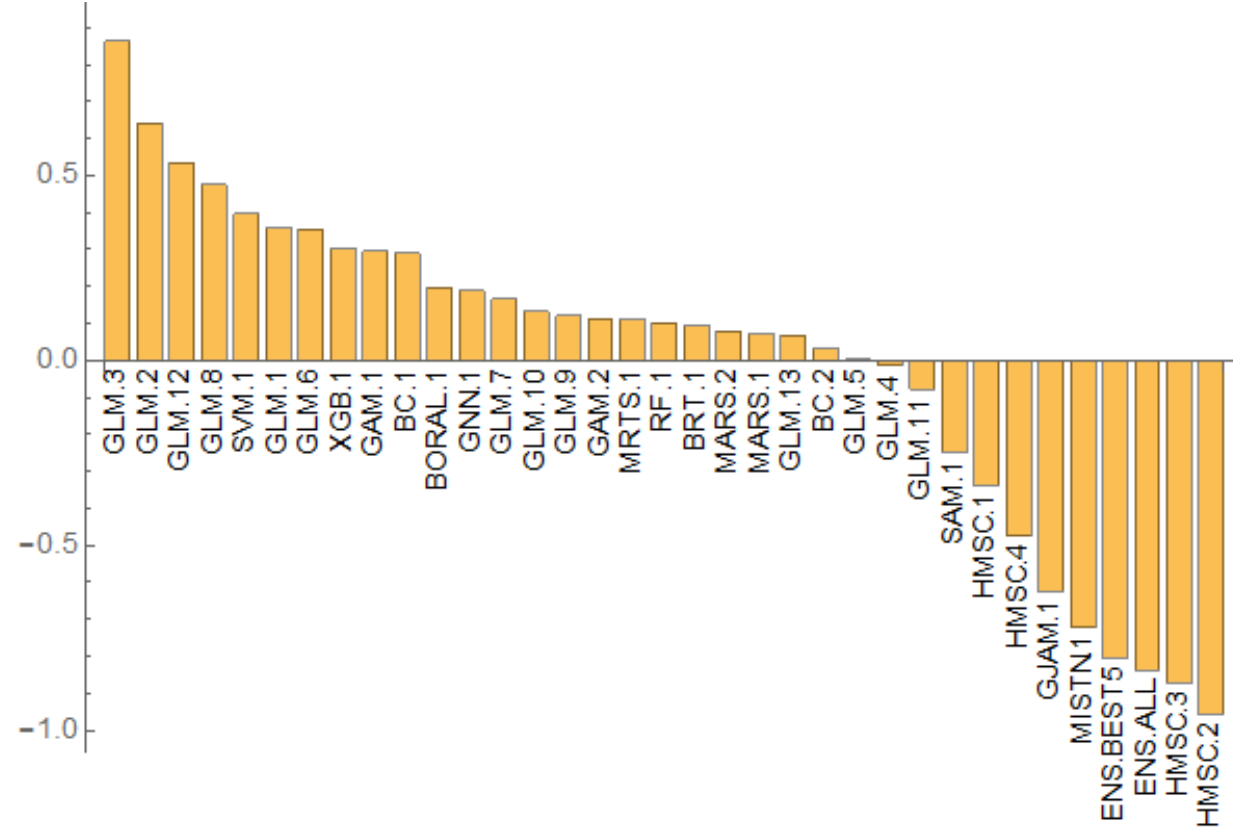
## 1 Species level



## 2 Species richness



## 3 Community composition

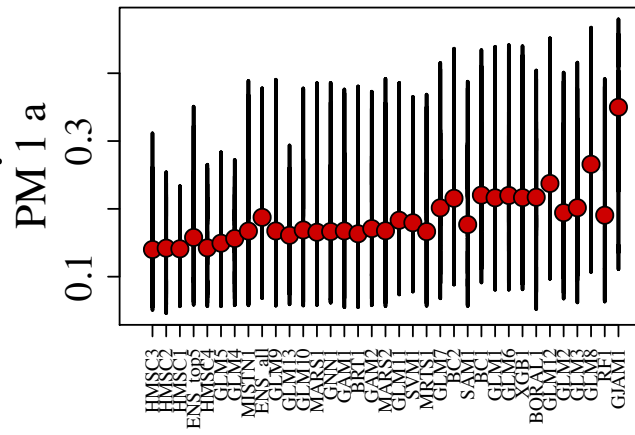




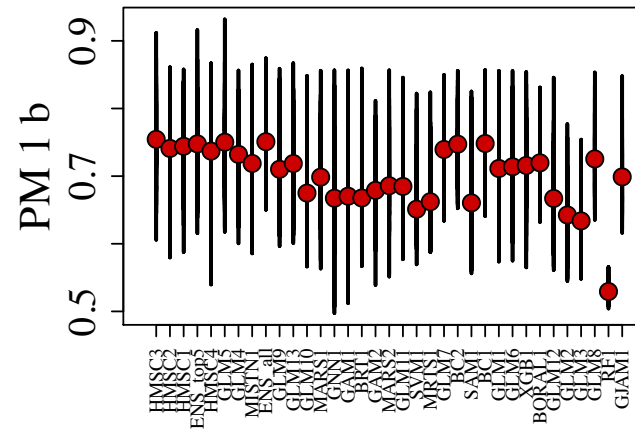
S9 Raw results for all model variants and all performance measures:  
 species-specific occurrence probability (1) and species richness (2)

Species-specific  
occurrence probability

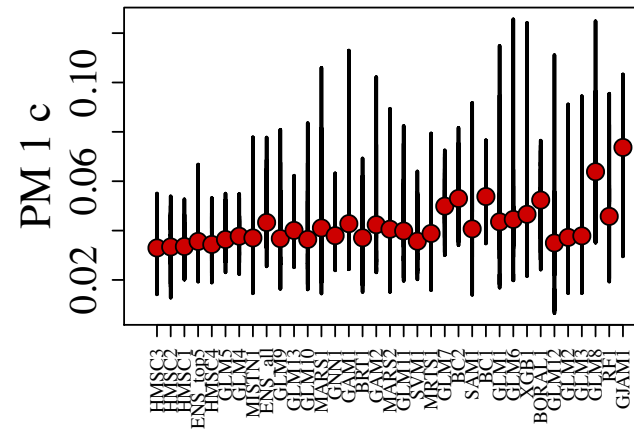
Accuracy



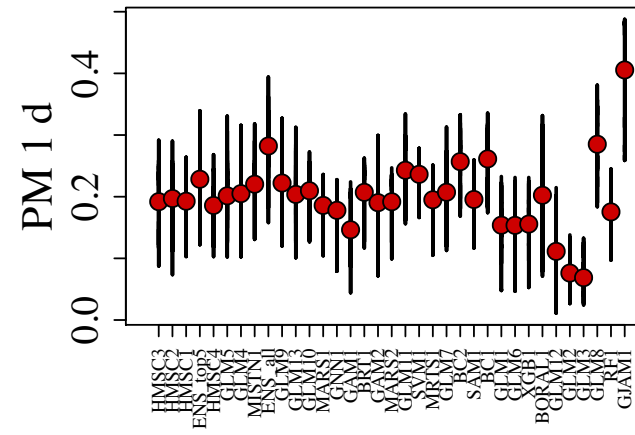
Discrimination



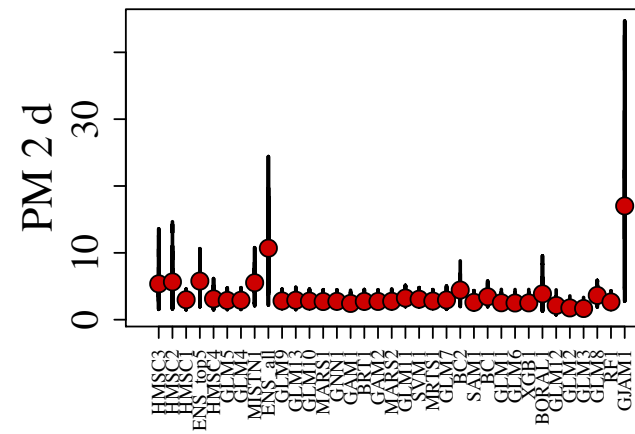
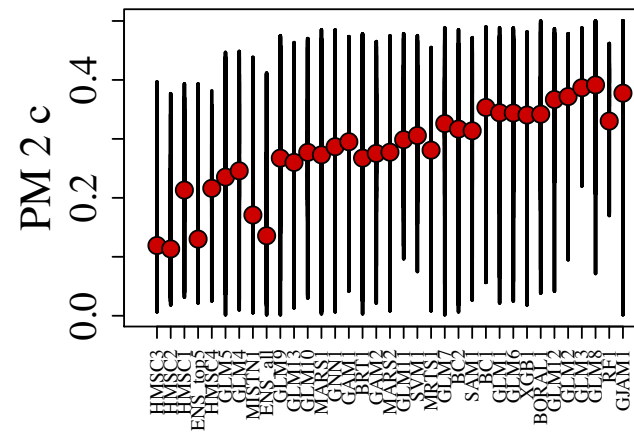
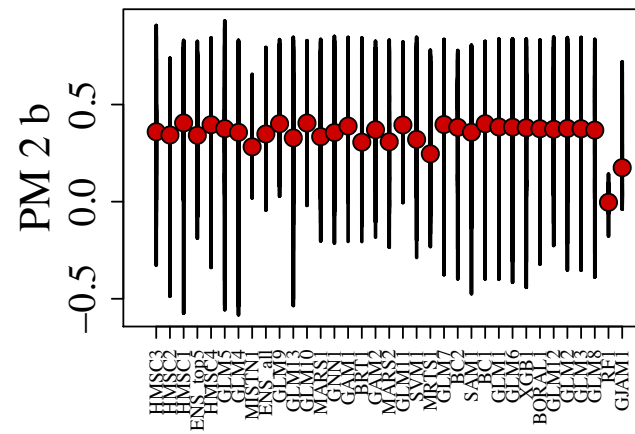
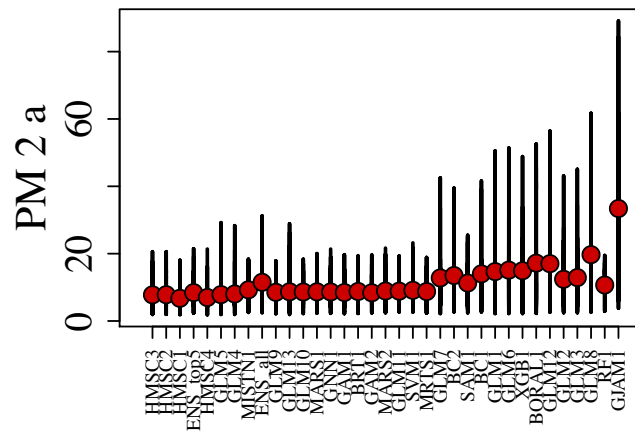
Calibration



Sharpness

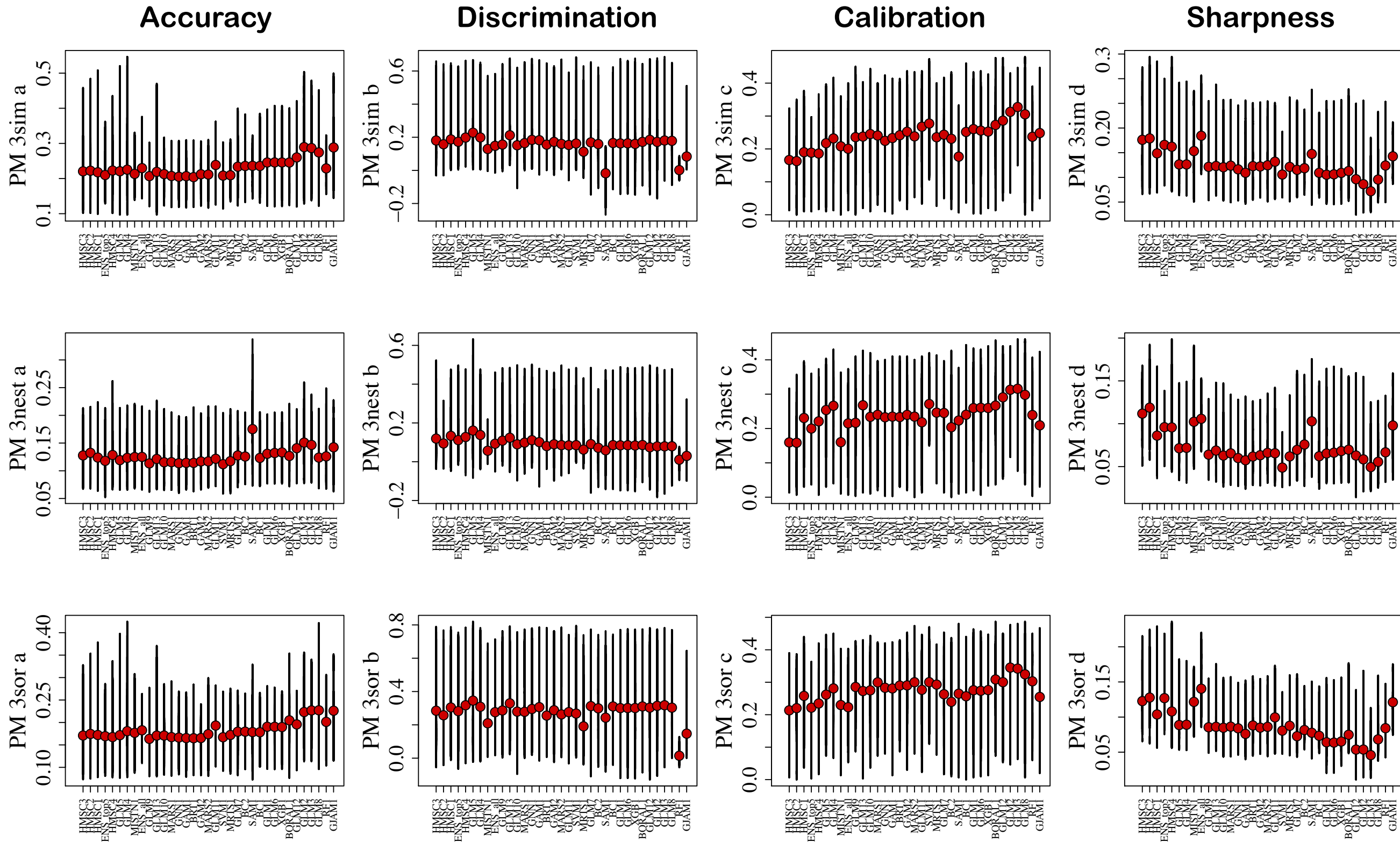


Species  
richness



S10 Raw results for all model variants and all performance measures:  
 Community composition measured by Simpson (sim), nestedness (nest) and  
 Sørensen (sor) indices

Community composition

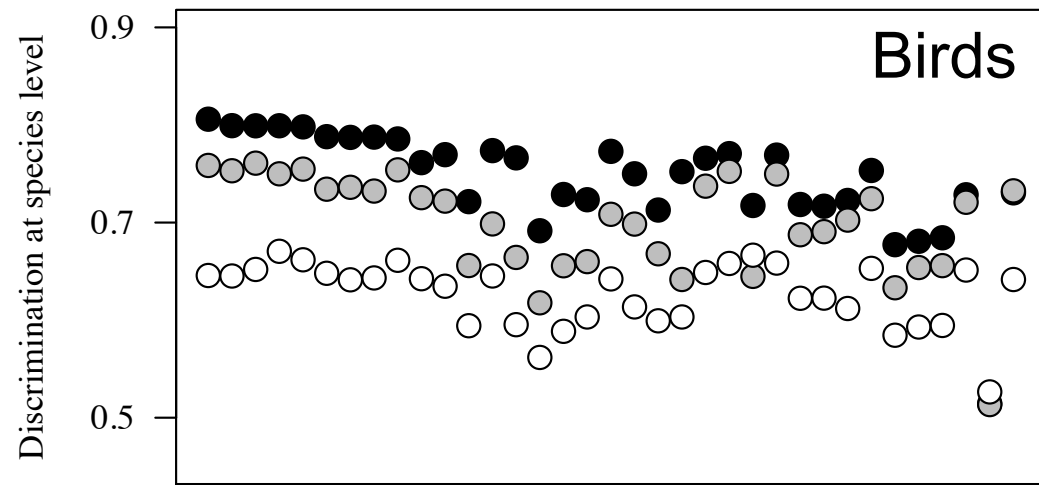


S11 Area under the receiving operator curve (AUC) based on all species, for the smaller data size ( $n = 150$ )

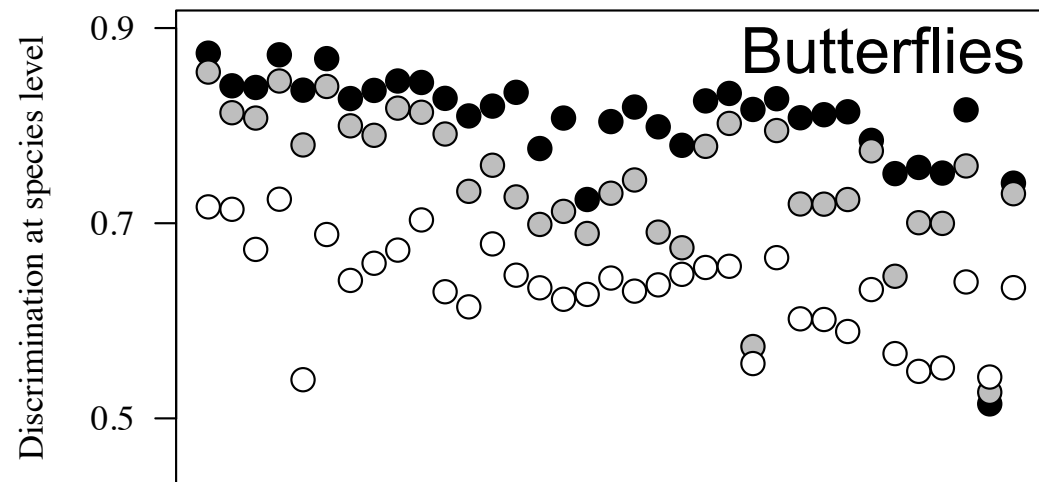
black: interpolation

grey: partial extrapolation

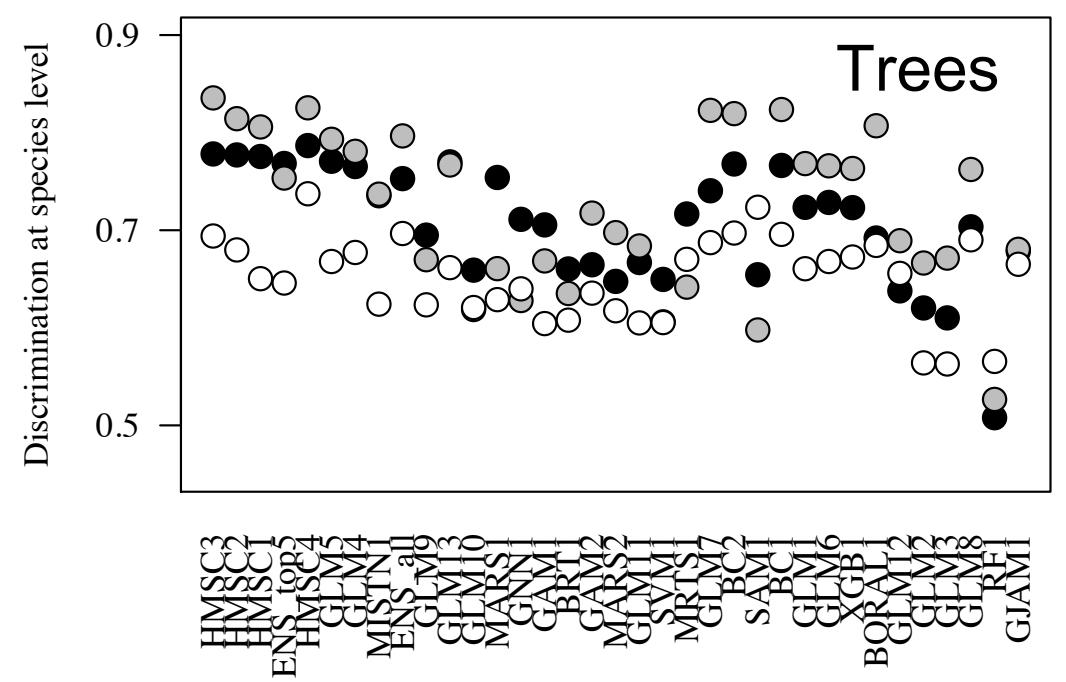
white: full extrapolation



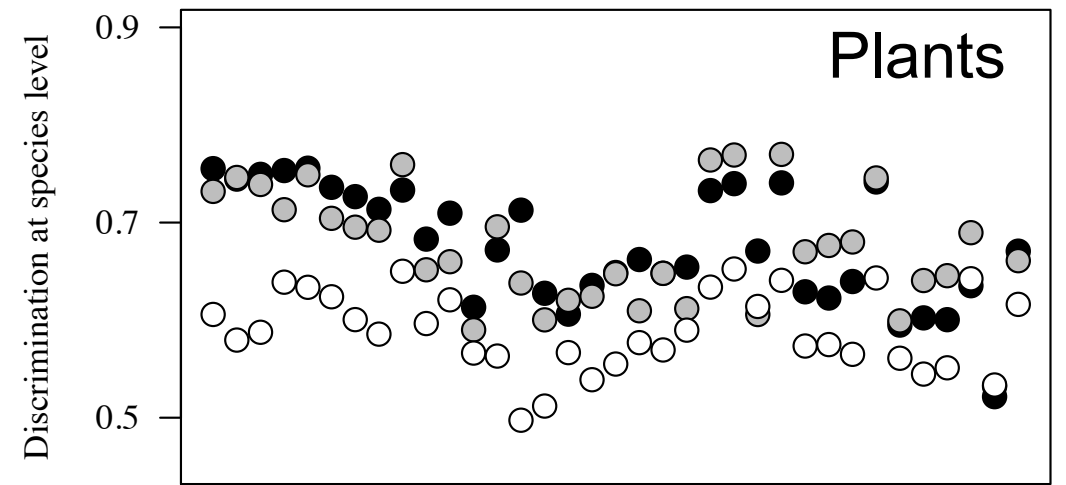
HM3 SC3  
HM SC2  
HM SC1  
ENS\_top5  
HM SC4  
GLM5  
GLM4  
MIS\_top1  
ENS\_all  
GLM9  
GLM3  
GLM10  
MARS  
GAM1  
BART  
GARS2  
GLM11  
SVM1  
MRTS  
GLM7  
SVM2  
GLM6  
GLM12  
BOR\_all  
GLM2  
GLM38  
RF1  
GJAMI



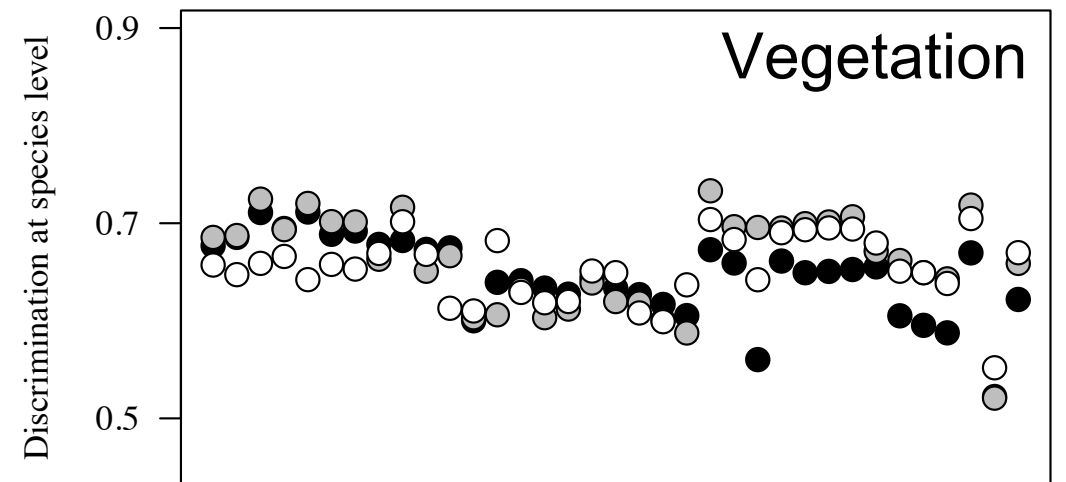
HM3 SC3  
HM SC2  
HM SC1  
ENS\_top5  
HM SC4  
GLM5  
GLM4  
MIS\_top1  
ENS\_all  
GLM9  
GLM3  
GLM10  
MARS  
GAM1  
BART  
GARS2  
GLM11  
SVM1  
MRTS  
GLM7  
SVM2  
GLM6  
GLM12  
BOR\_all  
GLM2  
GLM38  
RF1  
GJAMI



HM3 SC3  
HM SC2  
HM SC1  
ENS\_top5  
HM SC4  
GLM5  
GLM4  
MIS\_top1  
ENS\_all  
GLM9  
GLM3  
GLM10  
MARS  
GAM1  
BART  
GARS2  
GLM11  
SVM1  
MRTS  
GLM7  
SVM2  
GLM6  
GLM12  
BOR\_all  
GLM2  
GLM38  
RF1  
GJAMI



HM3 SC3  
HM SC2  
HM SC1  
ENS\_top5  
HM SC4  
GLM5  
GLM4  
MIS\_top1  
ENS\_all  
GLM9  
GLM3  
GLM10  
MARS  
GAM1  
BART  
GARS2  
GLM11  
SVM1  
MRTS  
GLM7  
SVM2  
GLM6  
GLM12  
BOR\_all  
GLM2  
GLM38  
RF1  
GJAMI



HM3 SC3  
HM SC2  
HM SC1  
ENS\_top5  
HM SC4  
GLM5  
GLM4  
MIS\_top1  
ENS\_all  
GLM9  
GLM3  
GLM10  
MARS  
GAM1  
BART  
GARS2  
GLM11  
SVM1  
MRTS  
GLM7  
SVM2  
GLM6  
GLM12  
BOR\_all  
GLM2  
GLM38  
RF1  
GJAMI

# S12 CONVERGENCE OF MCMC CHAINS FOR BAYESIAN MODELS

