A GENERAL FRAMEWORK FOR PREDICTING THE OPTIMAL COMPUTING CONFIGURATIONS FOR CLIMATE-DRIVEN ECOLOGICAL FORECASTING MODELS

by

Scott Sherwin Farley

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Masters of Science

in Geography

at

The University of Wisconsin-Madison

April 2017

Approved \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Advisor Title \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dept. of Geography

Date \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ABSTRACT

A GENERAL FRAMEWORK FOR PREDICTING OPTIMAL COMPUTING CONFIGURATIONS FOR CLIMATE-DRIVEN ECOLOGICAL FORECASTING MODELS

by

Scott Sherwin Farley

The University of Wisconsin-Madison, 2017

Under the Supervision of Professor John Williams

Rapidly growing databases are rapidly transforming biodiversity modeling into a big data science, characterized by high volume, heterogeneous datasets with high uncertainty. As climate warming and habitat destruction accelerate, it is imperative that scientists leverage all available data to generate robust, high resolution, and accurate predictions of biotic reorganizations to protect vital ecosystem services and minimize biodiversity loss. In recent years, cloud computing’s flexibility and scalability has caused it to emerge as the standard for analyzing massive datasets in many fields. However, its role in biodiversity studies and climate-driven ecological forecasting specifically has so far been poorly explored. While the cloud’s novel operating model allows users to provision and release virtual server instances from a utility provider as they are needed, there is currently little guidance for ecological researchers about the most efficient configuration to use. The scheme introduces tradeoffs between model accuracy, computing cost, and model execution time, and the choice of configuration has scientific and financial ramifications for the researcher.

In this thesis, I present a general conceptual framework for approaching these tradeoffs and introduce a model for determining the optimal data-hardware configuration for a species distribution modeling (SDM) workflow. I develop and test three hypotheses relating model accuracy and cost to algorithm inputs and computer hardware and collected an empirical dataset of over 25,000 experimental trials of four leading SDMs (generalized additive models, GAM; boosted regression trees, GBM-BRT; multivariate adaptive regression splines, MARS; random forests, RF). Subsequently, I built predictive models of runtime and accuracy using Bayesian regression trees and used these to evaluate the drivers of SDM accuracy and execution time. These models demonstrated considerable skill, and can be used to improve allocation of time and money, as well as inform model developers on future priorities. Finally, I identified and appraised the data-hardware configuration that maximized accuracy that jointly minimized execution time and cost for each SDM.

In general, the SDMs examined were most accurate when fit with a large number of training examples and many covariates and SDM accuracy was largely unaffected by the underlying hardware configuration. The optimal hardware for GAM, GBM-BRT, and MARS were low memory with few CPUs. RF, an ensemble technique, can more easily leverage parallel infrastructure, causing its optimal hardware configuration to consist of between four and seven CPU cores. Many widely used implementations of biodiversity models are structurally unable to take advantage of the increased computing power offered by cloud computing. My findings suggest that as datasets continue to grow, new algorithms and software packages must be developed to explicitly take advantage of the modern high performance computing techniques. While this framework specifically examines the identification of optimal computing configurations in the context of ecological forecasting models, it is extensible to other forms of biodiversity and ecological modeling studies as well.

© Copyright by Scott Farley, 2017

All Rights Reserved

Table of Contents

Acknowledgements 1

Introduction 2

Background and Previous Work 3

Big Data in Ecology 3

Cloud Computing in the Sciences 6

Species Distribution Models 7

A Taxonomy of Species Distribution Models 9

Computational Challenges and Species Distribution Models 11

Assessing Algorithm Execution Time 11

Theoretical Problem Formulation 12

Hypotheses 14

Methods 14

Data Collection 14

SDM data preparation 14

Computing Infrastructure 15

Estimating and Modeling SDM Runtime, Cost, and Accuracy 15

Optimal Prediction 16

Limitations and Extensibility of this Framework 17

Results 18

Optimal data configuration 18

Model Performance 19

Controls on SDM Runtime 19

Optimization 20

Discussion 21

Overview 21

R-Based Biodiversity Models and High Performance VMs 21

Extensions of the Optimization Approach 23

Cloud Computing in Biodiversity Modeling: Recommendations and Prospects 24

References 26

Figures ii

Tables ii

Appendices ii

Appendix A: Literature Meta-Analysis ii

Table A1: Studies Evaluated in the Analysis ii

Appendix B: Data Collection Protocol xii

Figure B1: Conceptual Flowchart of Distributed System Used for Automated SDM xii

Appendix C: Bayesian Model Priors xiii

# **Acknowledgements**

# **Introduction**

Human-induced global environmental change, including climate warming, land clearance, and the spread of invasive species, threatens to severely alter biodiversity patterns worldwide in the coming century (Lowe et al., 2011; Root et al., 2005; Thuiller, 2007; Thuiller et al., 2008), potentially causing the extinction of over one-sixth of all species (Urban 2015). Species ranges, particularly those of vascular plants, are strongly controlled by climatic factors (Salisbury, 1926; Woodward, 1987), and global changes in climatic gradients are expected to have a substantial impact on future patterns of biodiversity (Lowe et al., 2011). Statistical methods that quantify a species’ biophysical response to environmental factors, known as species distribution models (SDMs), can be used to forecast future species distributions and biotic assemblages under different warming scenarios (Clark et al., 2014; Guisan & Thuiller, 2005; Guisan & Zimmerman, 2000; Guisan et al., 2013; Maguire et al., 2015; Thuiller et al., 2008). A rapidly growing volume of ecological data is available for modeling species-specific responses to the climate system, both at present and in the geologic past (Fig. 1). Environmental monitoring efforts, such as the Long Term Ecological Research program (LTER, Hobbie et al., 2003), the National Ecological Observatory Network (NEON, Schimel et al., 2009); databases of the fossil record, including the Neotoma Paleoecology Database (Neotoma; http://neotomadb.org) and the Paleobiology Database (PBDB, http://paleobiodb.org); and modern species occurrence databases, such as the Global Biodiversity Information Facility (GBIF, http://www.gbif.org), all organize, store, and distribute large amounts of information to researchers in pursuit of understanding and forecasting biological responses to perturbations in the earth system (Brewer et al., 2012; Michener & Jones, 2012; Howe et al., 2008; Uhen et al., 2013). While growing collections of modern and paleo biodiversity data has the potential to improve predictive ecological modeling studies, data volume and heterogeneity can make successful uptake and implementation challenging for ecologists (Hampton et al., 2013). Cloud computing may offer a technological solution to some of the problems posed by the increasing ‘bigness’ of ecological data (Hampton et al., 2013; Michener & Jones, 2012), by allowing users to easily gain access to configurable and convenient virtual resources (Mell & Grance, 2012). However, little guidance exists for the researcher approaching tradeoffs between model accuracy, performance and cost. The present study develops a method for identifying the optimal cloud computing configuration on which to run SDMs, describing a flexible and skillful framework that can be used to inform provisioning strategies as well as suggest future priorities for model developers.

With over 700 million modern and historical occurrence records in GBIF and 18 million paleo-occurrence records in Neotoma (Fig. 1), traditional statistical methods for analyzing, modeling, and forecasting ecological patterns often cannot be applied without compromising analysis scope. Many SDM methods, though popular in the literature and highly skillful, are not designed to leverage parallel processing or distributed computing, and cannot be scaled to huge datasets. Other scientific fields, including bioinformatics (Schatz et al., 2010), genomics (Stein, 2010), climate analytics (Schnase et al., 2015), as well as private industry (Mosco, 2014), have adopted these and other computing techniques to cope with large datasets. As the volume of ecological data increases and the need for accurate, high resolution projections of biotic distributions and extinction risk become more pressing, solving data volume challenges by reducing project scope (e.g., Bolker et al., 2009) can no longer be considered a valid option. Cloud computing provides a platform on which to undertake large-scale ecological analysis (Hampton et al., 2013; Michener & Jones, 2012), by providing “ubiquitous, convenient, and on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort” (Mell & Grance, 2012). The rapid commercialization of cloud computing and the widespread availability of public cloud products through providers like Amazon Web Services (AWS) and Google’s Cloud Compute Engine (GCE) have put a seemingly unlimited supply of computing resources at scientists’ disposal.

Moreover, the cloud’s novel business model of charging customers for the use of virtual machines (VMs) rather than the purchase of physical hardware provides a highly flexible platform for scientific computing. This operating model lets consumers scale their resources depending on computational demand (Hassan, 2011). Users are therefore not locked in to a single hardware configuration, and can choose to add or remove different hardware components as the problem changes (Armbrust et al., 2009). While the exact costs of migrating to a cloud environment are difficult to estimate (but see Sun & Li, 2013), the computational time gains achieved by running models on high-performance virtual instances can be significant (Yang et al., 2011a).

Given the cloud’s flexible computing options, biodiversity modelers are then faced with tradeoffs on at least three distinct axes: computing cost, model run time, and model accuracy. Moreover, in some situations, such as in web-based modeling systems or real-time modeling efforts, biodiversity modelers may seek to reduce variability in these three. At the present time, it is unclear when, if ever, the benefits in reduced computing time reaped from a high-performance computing solution outweigh the incurred financial costs, particularly in the context of climate-driven ecological forecasting models.

I hypothesize that, for any SDM modeling experiment, there exists an optimal data-hardware configuration that maximizes SDM accuracy while jointly minimizing the time and cost of modeling. In this thesis, I develop a framework for predicting this optimal configuration for four widely popular SDMs. I build a large empirical dataset (n=26,730) on the cost, runtime, and accuracy under different parameterizations and on different computing hardware. Subsequently, I fit a Bayesian learning model to predict the execution time and accuracy of modeling scenarios, understand the factors that contribute to the runtime and accuracy of these models, and identify the optimal hardware for the task. My findings suggest that if SDMs are to benefit from cloud computing, future effort should be directed towards developing models that take explicit advantage of parallelism and distributed computing frameworks. As the volume of biodiversity datasets continues to increase and global habitat change continues to threaten large numbers of species, the need for better engineering of climate-driven ecological forecasting model implementations becomes increasingly important.

# **Background and Previous Work**

## Big Data in Ecology

The contemporary influx of massive datasets, including genomic sequences, long term monitoring projects, phylogenetic histories, and biodiversity occurrence data, has required the development of robust quantitative methods in the biological sciences, and has promoted advances in techniques for data management, analysis, and accessibility (Howe et al., 2008). Worldwide data volume doubled nine times between 2006 and 2011, and successive doubling has continued into this decade (Chen et al., 2014). This rate of data volume increase is faster than the annual doubling of computing power predicted by Moore’s Law (Villars et al. 2011). Significant challenges include the inability to move large datasets across networks, increased metadata requirements for storage and data discovery, and the need to support novel data uses (Schnase et al., 2014).

Ecological occurrence data consist of spatiotemporally-explicit records of presence, absence, or abundance of individuals of a species or higher taxonomic grouping. These data form the backbone of many contemporary biodiversity analyses and environmental change forecasts. Increasingly, these data are being stored in large, dedicated, community-curated databases like Neotoma, GBIF, and PBDB (Uhen et al. 2013; Brewer et al. 2012). Since the early 1990s, the Internet and an increased willingness to share primary data between scientists precipitated rapid influxes of digital occurrence records (Soberón & Peterson, 2004). While there are known problems with the quality and consistency of data records in large occurrence databases (Soberón et al., 2002), they provide a low-friction way of consuming large amounts of data that would otherwise be prohibitively time consuming to obtain from the literature or via new fieldwork (Beck et al., 2014; Grimm et al., 2013). Entire new fields, e.g. ‘biodiversity informatics’ (Soberón & Peterson, 2004), ‘ecoinformatics’ (Michener & Jones, 2012), and ‘paleoecoinformatics’ (Brewer et al., 2012), have been delineated to address the growing challenges and opportunities presented by the management, exploration, analysis and interpretation of primary data regarding life now housed in biodiversity databases (Soberón & Peterson, 2004).

The term big data is typically used to describe very large datasets, whose volume is often accompanied by lack of structure and a need for real-time analysis. Big data, while posing significant management and analysis challenges, can provide new insights into difficult problems not tractable with smaller datasets (Chen et al., 2014). Though the precise definition of Big Data is loose, there are two prominent frameworks for identifying Big Data. One characterizes Big Data as “data sets so large and complex that they become awkward to work with using standard statistical software” (Snijders et al., 2012). This ambiguous delineation is often echoed in the advertising and marketing literature that accompanies products, such as Apache Hadoop, a popular distributed computing framework, which describes Big Data as “datasets which could not be captured, managed, and processed by general computers within an acceptable scope” (Chen et al., 2014).

Under this framework, the big data is specific to both the time of analysis and the entity attempting to analyze it. Manyika et al. (2015) suggest that the volume of data required to be ‘big’ can change over time, and may grow with time or as technology advances. Today, big data usually refers to datasets between terabytes and petabytes (240 to 250 bytes), but varies among problem domains (Chen et al., 2014), the size of datasets common to that domain, and the software tools that are commonly used there (Manyika et al., 2015). The rapid growth of databases and concurrent development of increasingly complex data models to store spatiotemporal biodiversity occurrence records and their metadata (e.g. Grimm et al., 2013) suggests that traditional methods of data handling were not sufficient for modern ecological analyses. Further developments, such as application programming interfaces (APIs) and language-specific bindings (e.g., R packages or python modules), facilitate accessing, filtering and working with large occurrence datasets (Goring et al., 2015; Hernández & Sgarbi, 2016; Chamberlain et al., 2016). Thus, new, custom-built tools to store, analyze, visualize, and use multiple large ecological occurrence datasets are required.

A second framework by which to assess Big Data is the ‘Four V’s Framework’. First introduced by IBM and used by large technological companies in the early 2000’s to characterize their data, it is now a popular and flexible framework under which to evaluate data bigness. Here a dataset is described by its volume, variety, veracity, and velocity, under which “volume refers to the size of the data; velocity indicates that big data are sensitive to time, variety means big data comprise various types of data with complicated relationships, and veracity indicates the trustworthiness of the data” (Yang & Huang, 2013).

Biodiversity data clearly meets at least three of these criteria. With respect to volume, since the late 1990s, the rapid growth and scale of biodiversity information has become challenging to manage (Fig. 1). Today, Neotoma holds over 14,000 datasets containing more than 18 million individual occurrence records and associated spatial, temporal, and taxonomic attributes, corresponding to an average growth rate of 1,825 occurrences per day for the past 27 years. All records in Neotoma were originally gathered during fieldwork (e.g. sediment cores, paleontological or archaeological digs, or other efforts) -- techniques that require large expenditures of time and effort (Glew et al., 2002). GBIF houses digital records of well over 700 million digital records of field observations, living and fossil specimens, and reports from the scientific literature. Since its conception in the early 2000s, GBIF’s holdings have grown nearly 300%, from about 180 million records in 2001 to over 700 million records in 2016. GBIF includes both contemporary and historical observations, though 98.9% of its holdings are from 1900 onwards[[1]](#footnote-1).

Biodiversity data also has a high variety, complicated by many interrelationships. As shown in Figure 2a, Neotoma’s holdings feature 23 dataset categories. Categories are separated for different taxonomic groups (e.g. plants, vertebrates, diatoms, ostracodes), fossil sizes (microfossils vs. macrofossils), geochronological datasets used to estimate time, and physical and geochemical measurements such as X-ray fluorescence (XRF) and loss-on-ignition (LOI). Each dataset type has somewhat different metadata requirements and is managed by different sets of Data Stewards and virtual Constituent Databases within Neotoma (Williams et al. in prep.) GBIF delineates nine record classes, including human observations, living and fossil specimens, literature review, and machine measurements (Figure 2b). Though the records coexist in GBIF, they are distinct, derived using different protocols by different communities of researchers. Moreover, the data is both spatial and temporal. All of Neotoma’s records and 87.6% of GBIF’s records are georeferenced[[2]](#footnote-2). Digital representations of spatial phenomena must grapple with unique challenges, including discrete representations of continuous physical processes, correlations between parameters in space and time, and differences in scale, that make storage and management difficult (Yang et al., 2011a). Finally, occurrence data represents the work of many dispersed individual researchers and research teams – the ‘long tail’ of ecological data (Hampton et al. 2014; Heidorn, 2008). While controlled vocabularies and defined data structures help to efficiently assemble large numbers of records, nearly every record was collected, analyzed, and published by a different scientist. While some scientists have contributed many datasets to occurrence databases, most have contributed only a handful. The median number of datasets contributed per investigator to Neotoma is two and the third quantile value is just seven datasets. While specific metadata is scarce, each researcher is likely to use somewhat different equipment, employ different lab procedures, follow different taxonomic guidelines, and utilize different documentation practices, contributing to variation among datasets.

Biodiversity data also faces issues of veracity, which is expressed as uncertainty in taxonomic identification or spatiotemporal position. Some of the sources of uncertainty in the data, including spatial or temporal positional uncertainty can be estimated (Wing et al., 2005) or modeled (Blaauw, 2010). Both Neotoma and GBIF show high levels of quantifiable uncertainty. Of a random sample of 10,000 records of the genus *Picea* from GBIF, over half did not report spatial coordinate uncertainty. Of the 4,519 records that did, the average spatial uncertainty was 305 meters, and the maximum was 1,970 meters. Such uncertainty may be problematic for modeling studies (Beck et al., 2014). Temporal uncertainties in Neotoma is also large. Out of a sample of 32,341 age controls (e.g., radiocarbon dates, varve counts) in the database, only 5,722 reported any age uncertainty at all. The remaining records indicate a mean uncertainty of ± 130 years of the actual date. Another key source of uncertainty in large aggregation databases is taxonomic uncertainty as incorrect identification of occurrences can have significant influence on the interpretations of synthetic studies that leverage these records. A recent study of benthic macroinvertebrates suggests that taxonomic identification error may be as high as 21% (Stribling et al., 2008). Other sources of uncertainty, including measurement errors and data loss incurred between field, lab, and database, may also be important.

The final piece of the Big Data framework is velocity, which characterizes the analytical sensitivity to time. High velocity data must be analyzed in real time as a stream to produce meaningful insights. For example, tweets from the Twitter microblogging service, can be analyzed for trends as they are posted. Users are drawn to participation in up-to-the minute discussion and significant effort has been put towards the development of sophisticated algorithms that can detect clusters and trends in social behavior in real time (Bifet et al., 2011; Kogan, 2014). Similar sensitivity to time is critical to state-of-the-art warning systems tornadoes, tsunamis, and earthquakes (e.g., Blewitt et al, 2009). Velocity is perhaps the weakest fit between big data criteria and biodiversity modeling. Unlike many private sector applications, there are few scientific factors driving biodiversity researchers to immediately analyze new occurrence records. Moreover, automated analyses of distributional data have been warned against, due to heterogeneous data quality (Soberón et al., 2002) and associated uncertainty. However, some kinds of environmental monitoring have a high temporal resolution (e.g. eddy flux towers) and could be in principle leveraged to provide, e.g., high-velocity, near-real-time estimates of biospheric carbon uptake and release.

In summary, ecological occurrence data requires advanced, sophisticated techniques to store and analyze, and demonstrates high volume, low veracity, and substantial variety, and sensitivity to time (velocity), and should therefore be considered big data. To fully and accurately derive value from new data being added to these databases, novel techniques for working with this data are required, while traditional statistical analyses may begin to suffer because they were not designed to handle big datasets.

## Cloud Computing in the Sciences

In recent years, large technology companies have promoted cloud computing as a way of overcoming the computational challenges associated with Big Data. The cloud leverages distributed networks of virtualized physical machines to create a computing utility, providing a pay-as-you-go business model and large economies of scale (Armbrust, 2009; Hassan, 2011) that delivers abstract resources and services, in addition to storage and compute resources (Foster et al., 2008). While some organizations and universities have developed ‘private clouds’ -- large collections of virtualized servers not made available to the general public, similar to computing grids -- many researchers have recognized the potential for incorporating public clouds -- provided as a service by a cloud provider -- into their workflows. With this technology, scientists with little or no computational infrastructure can access scalable and cost-effective computational resources (Hsu et al., 2013). Major federal agencies and scientific organizations in the United States, including the NSF and NASA, have actively promoted cloud computing. Spurred by the U.S. Office of Management and Budget’s “25 Point Plan to Reform Federal Information Technology Management” (Kundra, 2010), all federal agencies are now required to adopt a “Cloud First” policy when “contemplating IT purchases and evaluate secure, reliable, and cost-effective cloud computing alternatives when making new IT investments” (Office of the Inspector General, 2013). The federal plan also created programs to help agencies adopt cloud solutions, reducing the effort needed to screen cloud providers for data security policies and enable rapid procurement of cloud services (Kundra, 2010). NSF has launched new solicitations experimenting with and developing new cloud architectures for scientific computing (National Science Foundation, 2012; National Science Foundation, 2014). Public cloud providers, such as Amazon Web Services and Google Cloud Compute, support scientific enterprise by providing large, open-access datasets for public consumption, including Landsat images, real-time NEXRAD radar, and the 1000 Genomes project, and by soliciting grants for researchers incorporating cloud computing into their research.

Cloud technology, both public and private, has extended into many fields, including bioinformatics (Hsu et al., 2013, Issa et al., 2013; Stein et al., 2007) and climate analytics (Lu et al., 2011; Schnase et al., 2014, 2015). Cloud-based solutions for bioinformatics research relieve the large memory requirements often associated with genomics and drug-design data (Hsu et al., 2013), and have resulted in low latency, streaming methods for data analysis (Issa et al., 2013) and biology-specific operating systems for protein analysis (Kaján et al., 2013; Schatz et al. 2010). Contemporary climate analytics often requires working with massive datasets too large to be transferred across networks, resulting in the development of Climate Analytics as a Service, an effort to integrate data storage and high performance computing to perform data-proximal analytics (Schnase et al., 2014, 2015).

Cloud services have also been used in the biological and geosciences, and in ecological modeling problems specifically. Yang et al. (2011a) suggest that many kinds of geoscientific problems are strongly limited by computational ability and argue that the cloud provides a means of overcoming these challenges by leveraging distributed computational resources without increasing the carbon footprint or financial budget of research (Yang et al., 2011b). For example, cloud-optimized implementations of numerical models, such as real-time dust storm forecasting, have improved model performance significantly, while reducing cost by only using the intensive computing power required for forecasting during storm events (Yang et al., 2011a). Geoscientific, hydrological and environmental models can also be run in the cloud (Granell et al., 2013). Applications of cloud computing to biodiversity modeling and species distribution modeling are less common, though some SDM projects are starting to explore this area. For example, Candela et al. (2013) describe a novel platform that enables cloud-based SDM, arguing that a cloud-based approach can aid in data discovery and increase processing capabilities.

## Species Distribution Models

Species Distribution Models (SDMs) are a widely used class of statistical models that quantify the relationships between a species and its environmental determinants (Svenning et al., 2011). While these models may sometimes include mechanistic or process components, they most often refer to correlative models (Elith & Leathwick, 2009) that use supervised statistical learning algorithms to approximate the functional relationship between species occurrence and environmental covariates. Used extensively in both academic and management contexts, SDMs have been shown to provide reliable estimates of climate-driven range shifts when compared to independent datasets (Guisan & Zimmerman, 2000; Guisan et al., 2006). Widespread availability of powerful statistical software and large databases of environmental and occurrence data, have led to increased popularity of these techniques in recent years (Franklin, 2010; Svenning et al., 2011). Citations of SDM-focused studies outpaced the field average (National Science Board, 2016) by 3.8% per year between 1997 and 2015, according to an analysis of publications in Web of Science (Figure 3).

SDMs use a learning algorithm, along with occurrence records and environmental covariates, to approximate the functional form of the species niche that can be used to test ecological hypotheses or to predict to future scenarios. Hutchinson (1957) characterized a species’ fundamental niche as an *n*-dimensional hypervolume that defines the environmental spaces where the intrinsic population growth rate of the species is positive (Williams & Jackson, 2007). The realized niche describes the subset of environmental space that the species actually occupies at some point in time, and is smaller than the fundamental niche due to competing biotic interactions with other species. SDMs, through their reliance on observational data, approximate the species’ realized niche (Guisan & Zimmerman, 2000; Miller et al., 2007; Soberón & Peterson, 2005). Hence, a key uncertainty associated with SDMs is their possibly incomplete representation of the fundamental niche and derived predictions of species distributions under future or past climate changes. The inclusion of fossil data in the model fitting process can increase the likelihood that calibration captures the fundamental niche by exposing SDMs to states of the climate system not present on Earth today (Veloz et al., 2012; Nogués-Bravo, 2009).

Because of their relative ease of use, SDMs have been widely used by biodiversity conservationists to prioritize habitat protection and to predict species responses to past and future environmental change. SDMs are often used to confirm ecological hypotheses by comparing hindcast projections with the fossil record; for example, supporting hypotheses on the extinction of Eurasian megafauna (Nogués-Bravo et al., 2008; Svenning et al., 2011), identifying late-Pleistocene glacial refugia (Fløjgaard et al., 2009; Keppel et al., 2011; Waltari et al., 2007), and assessing the effect of post-glacial distributional limitations and biodiversity changes (Svenning et al., 2008). SDMs are sometimes combined with genetic, phylogeographic, and other methods to develop a complete assessment of a species biogeographical history (e.g., Fritz et al., 2013). In the context of contemporary environmental change, SDMs have been used to assess the effectiveness of modern reserve planning (Araújo et al., 2004), predict the distribution of both endangered (Thuiller et al., 2005) and invasive species (Ficetola et al., 2007; Václavík & Meentemeyer, 2009; Smith et al. 2013) and ecosystems (Hamann & Wang, 2006), and evaluate the effectiveness of conservation planning for the future (Loiselle et al., 2003).

SDMs rely on three important assumptions. First, as a fundamental justification for applying predictions across space and time, SDMs assume niche conservatism -- that the niche of the species remains constant across all spaces and times (Pearman et al., 2008). This assumption disregards niche evolution, based in part on evidence that species typically demonstrate niche conservatism on multi-million year time scales (Peterson et al. 1999). Second, SDMs rely on the assumption that species are at equilibrium with their environment (Nogués-Bravo, 2009), being present in all environmentally suitable areas while being absent from all unsuitable ones. Given dispersal limitations and biotic interactions between species, this may rarely be the case. For example, Svenning et al. (2008) showed that many European tree species are still limited by postglacial migrational lag. Finally, SDMs must account for extrapolation to novel climates for which there is no contemporary (or geological) data (Radeloff et al. 2015), which is a challenge given that many future climates are likely to lack current analogs (Williams and Jackson 2007). Inductive learning accuracy is declines when predicting cases not within the range of values used in training. Fitting SDMs with fossil data increases the likelihood that climatic assemblages will be included in the training data (Veloz et al., 2012; Nogués-Bravo, 2009), though the problem of projecting models onto novel future climates continues to be a major challenge to their application.

### A Taxonomy of Species Distribution Models

SDMs range in complexity from simple algorithms that characterize a ‘climate envelope’ for a species (Guisan & Zimmerman, 2000) to multivariate Bayesian techniques that use Markov Chain Monte Carlo simulations (MCMC) to develop probability distributions around projections and parameters. While all SDMs aim to model responses to climatic gradients, SDM algorithms can be broadly grouped into data-driven and model-driven categories (Franklin, 2010). The data-driven/model-driven dichotomy is introduced in Hastie et al. (2009) and can be thought of as differentiating between ‘statistical’ or ‘parametric’ (model-driven) and ‘machine learning’ or ‘nonparametric’ (data-driven) algorithms. I add, as a third category, the burgeoning set of methods that employ stochastic, probability-based Bayesian methods due to their structural differences, assumptions, and handling of uncertainty. Many scholars have attempted to assess variation between models (Araújo & Guisan, 2006; Elith et al., 2006) and parameterizations (Araújo & New, 2007; Thuiller et al., 2008; Veloz et al., 2012).

SDMs, and supervised learning techniques more broadly, must be calibrated against observational data prior to being used for predictions. During the calibration stage, SDMs use a set of training examples, , where both *x* (environmental covariates) and *y* (species presence) are known, to approximate the real relationship between the two, , with the learned approximation function, . The learned function is the relationship that minimizes a loss function based on the difference between the real and predicted value, . Each training example () is composed of a -dimensional vector of covariates, , (Haste et al, 2009). Model-driven algorithms make *a priori* decisions about the structure of , while data-driven algorithms can adapt to fit any given design matrix (Hastie et al., 2009; Franklin 2010).

Model-driven SDMs fit parametric statistical models to a dataset, making assumptions about how inputs and outputs are related, including e.g. linearity, error distribution, and independence. While these techniques can make poor predictions if the assumptions are not upheld, they were the first to see substantial use in SDM applications and continue to be widely used because of their strong statistical foundations and ability to realistically model ecological relationships (Austin, 2002). These models include boxcar algorithms, which build multidimensional bounding boxes around species presence in environmental space (Guisan & Zimmerman, 2000), as well as more complex methods, including generalized linear models (Guisan et al, 2002; Vincent & Haworth, 1983) and multiple linear and logistic regression (Franklin, 2010).

An increase in available computing power has spurred the development and application of data-driven learning algorithms, which take a non-parametric approach to approximating . While not reliant on stringent assumptions about the form of the relationship, any particular portion of parameter space depends on only a handful of input points, making the models highly sensitive to small changes in the input data (Hastie et al., 2009). In some cases, these models are shown to outperform their model-driven counterparts (Elith et al., 2006), and include genetic algorithms (Elith et al., 2006), classification and regression trees (Elith et al., 2008), artificial neural networks (Hastie et al., 2009), support vector machines (Drake et al., 2006), and maximum entropy techniques (Elith et al., 2010; Phillips & Dudík, 2008). MaxEnt, a maximum entropy algorithm for SDM and associated Java-based runtime environment is widely used and has been demonstrated to perform consistently, even on small sample sizes (Elith et al., 2010; Phillips & Dudík, 2008; Phillips et al., 2006). Recent literature suggests that MaxEnt is the most popular SDM method in use today, appearing in over 20% of all SDM studies published after 2008. Recent critiques of MaxEnt, however, suggest that its performance may be questionable particularly on small datasets that sample only a small portion of a species geographic range (Fitzpatrick et al., 2013). Data-driven models are often more computationally intensive than their model-driven counterparts because they usually take at least two passes over the input dataset to (1) process the data and (2) build the model (Hastie et al., 2009). Furthermore, data-driven SDMs are often combined with techniques like bagging -- building a collection of models based on random subsets of the input data -- and boosting -- combining many weakly predictive models into a single, highly predictive ensemble – which can further increase computational intensity.

Bayesian methods have also been used in the approximation of . Advantages of the Bayesian approach include the ability to include prior ecological knowledge in model formulation (Ellison, 2004) and the ability to estimate model uncertainty without the need for bootstrapping procedures (Dormann et al., 2012; Elith & Leathwick, 2009). With improved computational infrastructure and better MCMC sampling algorithms, Bayesian methods have become increasingly popular in recent years (Hegel et al., 2010). Golding & Purse (2016) introduce Bayesian SDMs that incorporate Gaussian random fields, which they claim demonstrate both high predictive accuracy and ecologically sound predictions. Clark et al. (2014) use the full joint probability distribution of all taxa in an ecosystem to model both the climatic range limitations of a species and its biotic interactions with other species. Though it can be challenging for ecologists trained in classical statistics to transition to a Bayesian approach (Ellison, 2004; Hegel et al., 2010), software packages are in development for implementing Bayesian SDMs in already-adopted languages such as R (e.g., Vieilledent et al., 2012). Because Bayesian algorithms rely generating and sampling from MCMC distributions (e.g. Gibbs Sampling, No-U-Turn Sampling), they are computationally expensive, though numerical approximations and analytical solutions can sometimes reduce computational burden (Golding & Purse, 2016).

A review of recent literature suggests that the majority of contemporary SDM users employ data-driven models. Of 100 randomly sampled recent publications in Web of Science that met the query “*(Species Distribution Model) OR (Ecological Niche Model) OR (Habitat Suitability Model)*”, the overwhelming majority used data-driven models. Of 203 modeling runs described, 131 were data-driven, 38 were model-driven, and 1 was Bayesian (Figure 4). An addition 33 experiments used unsupervised clustering analyses not suitable for prediction. Of all algorithms, MaxEnt was the most popular (64 runs). Algorithms in the model-driven category included generalized linear models (15), logistic regression (5) and multiple linear regression (2). Data-driven techniques included boosted regression trees (16), generalized additive models (11), genetic algorithms (11), random forests (8), artificial neural nets (6), and multivariate adaptive regression splines (4). Figure 4 shows the results of the literature meta-analysis and the classification into the taxonomy described here. The citation for each paper reviewed is presented in Appendix A, table A1.

### Computational Challenges and Species Distribution Models

Because of the strong popularity of data-driven SDMs, I focus my analyses on this class of algorithms. Many authors have alluded to the limitations imposed by computational complexity, though few have estimated or tested those limits explicitly. Elith et al. (2006) recorded the execution time of the runs they used in their often-cited review of novel SDM techniques, noting execution times of up to several weeks for some modeling algorithms. Popular data-driven models were all shown to be extremely computationally intensive, including boosted regression trees (80 h), generalized additive models (17h), generalized linear models (17h), and MaxEnt (2.75 h). The authors suggest that performance could be improved if model building was split over multiple processing cores. While processor speeds have increased since their 2006 analysis, models are still often unable to leverage multiple processors.

Methodological papers often advise against large modeling studies due to computational limitations. For example, a 2009 review by Bolker et al. (2009) suggests that, when fitting a generalized linear mixed model (GLMM), if a user encounters insufficient computer memory or time limitations, the user should reduce model complexity, perhaps using a subset of the original dataset. Many authors warn of the computational expense of running SDMs, for example, noting that “considerable computational capacity is necessary for the development of models even for a single species” (Peterson, 2003). Thuiller et al. (2008) cautions “limits to the broad application of this approach may be posed … by the computational challenges encountered in the statistical fitting of complex models.” Modern computing infrastructure may alleviate some of these problems, but often, the computational intensity of SDMs forces a reduction in model complexity or scope.

## Assessing Algorithm Execution Time

It is possible to theoretically estimate the upper, lower, and average run times of an algorithm using asymptotic complexity analysis. In this exercise, the first order term of an algorithm’s increase in runtime is determined as its input is increased to infinity (Knuth, 1976). The algorithm that is more efficient asymptotically will typically be the best choice for all but very small inputs (Cormen, 2009). While it is often not possible to produce a robust estimate of the lower-bound on runtime, given an infinite input, an estimate of the slowest or worst-case run time can usually be obtained by inspecting the structure of the algorithm and counting how many operations are required when the inputs is sufficiently large (i.e., Big-O; Cormen, 2009). Such theoretical complexity is often considered when considering scalability, though the actual runtime will vary with real-world inputs (Cormen, 2009; Goldsmith et al., 2007).

Empirical complexity studies have attempted to bridge the gap between asymptotic theory and real programs (Cannon et al., 2007). These studies use observations of algorithm runtime under different parameterizations and inputs to build models that predict the run time of future applications of the algorithm, seeking a method “with the generality of a Big-O bound by measuring and statistically modelling the performance … across many workloads” (Goldsmith et al., 2007). Brewer (1995) describes an initial attempt to develop a statistical model for the run and compile time of algorithms in a C library. While most contemporary empirical runtime models use data-driven pattern recognition, linear regression between input size and execution time has been shown to perform well in some cases (Fink, 1998). Empirical complexity models have become an important subfield of artificial intelligence and have important applications to algorithm selection (Hutter et al., 2014). Algorithms for solving very difficult (-Hard/-Complete) combinatorial problems can exhibit high runtime variance among different problem instances. Empirical models can be used to select the model that will most efficiently reach a solution (Hutter et al., 2014; Leyton-Brown et al., 2003; Hutter et al., 2013). Hutter et al. (2014) outline a comprehensive analysis of strategies and methods for empirical runtime models in the context of algorithm portfolio optimization. Parameterized algorithms can be treated the same way as nonparametric algorithms, by including model parameters as input features in the execution time model (Hutter et al., 2014). Nonlinear, tree based methods for empirical performance modeling, including random forests, were shown to be superior to other methods because of their ability to group similar inputs together and fit local responses, so that some large outliers do not interfere with the predictions of other groups (Hutter et al., 2014; Hutter et al., 2013).

Concurrently running programs, operating system tasks, and other processes may affect the execution time of a real computer program at any point in time. Changes in dynamic system state are stochastic and can cause unpredictable, non-linear and non-additive changes in program runtime (Jones & Kalibera, 2013; Lilja, 2009). Random variation in system state makes deterministic statistical modeling of hardware’s influence on execution time difficult. These variations result from the way in which memory access patterns differ in space and time when small changes are made to the operating system state, timing device, or algorithm and its inputs (Lilja, 2009), and few attempts have been made to model them explicitly. However, several recent studies that took dynamic system state into account as a predictor of algorithm runtime performed well when considering data center optimization (Sadjadi et al., 2008; Wu & Datla, 2011).

Jones & Kalibera (2013) suggest that models based on benchmarked runtime may provide an accurate estimate of an upper bound of execution time, though due to potentially large, nondeterminstic, system-induced variance in empirical results, it is important to perform the benchmarking experiment many times. Dongarra et al., (1987) warn that a failure to properly characterize the workload, running benchmarks that are too simplistic, or running benchmarks in inconsistent environments can lead to meaningless results.

# **Theoretical Problem Formulation**

In the present study, I use benchmarking and empirical performance modeling to develop predictive models useful in optimizing SDM workflows. The following framework presents an SDM workflow consisting of a series of steps that advance the SDM user towards her goal of obtaining scientific insight from a dataset. I assume the SDM modeler is a rational consumer in a supply- and demand-driven computing market and that the modeler has imperfect information regarding the covariate-species presence relationship. In this framework, the modeler will undertake several steps, including model computation, to minimize her costs, in both runtime and financial terms, and to maximize her utility, represented here as proximity to the knowledge of the true functional relationship, , between environment and species presence. This proximity between f and is measured as SDM accuracy (Simon, 1986). These steps are as follows:

1. Consider a pool of computing resources, , that is characterized by multiple possible hardware configurations consisting of memory, CPUs, and any other component that influences computing power.
2. Consumers of cloud computing services are part of a market driven by supply and demand, and consumers face costs that are priced by the hour by providers as a function of the computing power provided: For example, Google’s infrastructure-as-a-service (IaaS) cost surface closely tracks memory and CPUs (Figure 5). This differentiates cloud computing from traditional, non-cloud computing, which tends to have more fixed hardware configurations and costs, with less consumer flexibility. Nevertheless, even under fixed cost models, total purchase price can be converted to an hourly rate if several assumptions about computer use and lifespan are considered.
3. Costs associated with modeling application are multidimensional and are not limited to monetary costs. Additional costs may include the runtime of the model.
4. Every user of a modeling application has a particular set of goals for using it in the first place (Norman, 1984). Hence, we can conceptualize, for any given SDM, a finite set of use cases that fall within the bounds of existing or expected use (Carroll, 1999; Rosson, 2002). Let be a vector of characteristics that fully describe the user’s goals in these possible use cases. The components of include user traits, such as experience with the model and interface, motivation, skill, and desired accuracy, as well as the number and parameterizations of each modeling run required (
5. Assume, in addition to computing the model, the user must also undertake a number of other pre- and post-processing steps in a scientific workflow. The total time elapsed during this workflow can be expressed as

Where represents the portion of time that is spent by user gathering the resources needed to model, such as time needed to find and download occurrence points and covariates. is a function of user expertise, the computing resources available to the user (how fast can data be discovered and downloaded?) and the experiment (what is the data?). represents time required by the modeler to prepare the data for entry into an algorithm, including data cleaning, projection, and conversion. can vary widely among modelers, data source and quality, and user skill and motivation (Elith et al. 2006). is the time spent computing the model and predicting to future climate scenarios. represents time needed to transfer the output from the location of the computation to the user, which may be non-trivial if the model is run remotely and large-volume outputs are transferred over a network. Finally, represents the amount of time spent by the user evaluating model output and determining whether her goals were met during the modeling process. Like , this term will be highly variable between model users and applications.

1. Single experiments can be combined together to form workflows, so that a user’s time-to-goal for a workflow of modeling experiments can be expressed as a function of the experiments and the computing resources on which they are run.
2. Combining equations (2) and (5), the total time for a set of modeling experiments is the sum of total time of spent modeling, while the total monetary cost is the cost of provisioning computing resources for this time. The total workflow cost is then a function of the user and their set of required modeling experiments, the computing resources, and the cost surface that dictates the cost of these resources. Therefore, a multivariate cost function for all potential user activities is:
3. Each individual scenario, *u*, in *U*, the set of all possible scenarios, will have its own multidimensional cost curve that is subject to both the particular characteristics of the workflow and the cost surface imposed by the computing provider. If we select one and call it , we obtain a unique cost function for this workflow that depends only on the computing resources used to fit the model. The minimum along this curve in multidimensional space corresponds to the optimal hardware configuration for use in this modeling scenario.
4. Multiple experiments may meet the user’s goals, but have different costs. The optimal workflow for a user to pursue is that which jointly maximizes model accuracy while minimizing the costs of the modeling scenario. These costs include both the runtime of the model and monetary costs of provisioning the resources for the time required. If desired, a set of weights could be applied to preferentially weight one or more of these dimensions. Additionally, if a user faces constraints on time (e.g., latency requirements) or money (e.g., budget requirements), these can be incorporated to find the optimal configuration within the allowable space.

# **Hypotheses**

The remainder of this thesis addresses several hypotheses based on this framework. Specifically, I hypothesize that:

* 1. for any SDM, there exists an optimal configuration of data and hardware that maximizes SDM accuracy while jointly minimizing the time and cost of modeling;
  2. choice of hardware configuration will affect the runtime of the SDMs, but not the accuracy; and
  3. data volume will affect both the runtime and accuracy of the SDM.

In the study, I characterize data as the number of training examples and the number of environmental covariates used to fit the model. I characterize hardware as the number of CPUs and amount of memory, in gigabytes (GB), of the VM on which the SDM is run. I use four SDM algorithms that are widely used and have shown competitive accuracy results in the literature: multivariate adaptive regression splines (MARS, Leathwick et al., 2006), gradient boosted regression trees (GBM-BRT, Elith et al., 2008; Friedman, 2001; Natekin, 2013), generalized additive models (GAM, Guisan et al., 2002; Yee & Mitchell, 1991), and Random Forests (Breiman, 2006; Elith & Graham, 2009). Maxent is excluded because (1) it is written in Java, with only R bindings linking it to the R platform and (2) it is not open source, being instead distributed as a black-box algorithm. The experimental design is meant to mimic actual use cases, and is performed using popular implementations of the algorithms in R.

# **Methods**

## Data Collection

### SDM data preparation

Systematic, controlled observation of SDM run time and accuracy on a complete set of data and hardware configurations was completed using the R statistical environment (R Core Team, 2016). Each SDM was fit with the standard package for that model for use with SDM. Specifically, GBM-BRT models were fit using the dismo package version 1.1-1 (Hijmans et al., 2016), GAMs using the gam package, version 1.12 (Hastie, 2015), MARS using the earth package version 4.4.4 (Milborrow, 2016), and RF using the randomforest package version 4.6-12 (Liaw & Wiener, 2002).

Each SDM was fit using fossil pollen occurrence data obtained from the Neotoma Paleoecological Database in April 2016. Neotoma was selected as the provider of occurrence data due to its rich coverage in space and time in North America since the last glacial maximum. However, Neotoma is just one instance of occurrence data -- similar analyses could be undertaken with records from other databases, such as GBIF or PBDB. All records for the genera *Picea* (spruce), *Quercus* (oak), *Tsuga* (Hemlock), and *Betula* (birch) were downloaded in R using the neotoma package (Goring et al., 2015), and filtered to include only those records dated to within the last 22,000 years and located in North America. For each record, the latitude, longitude, age, and relative abundance of the taxon was retained and stored in comma-separated text format.

Climatic covariates were obtained from downscaled and debiased Community Climate System Version 3 (CCSM3) model simulations for North America (Lorenz et al., 2016). Post-processed model output was obtained in NetCDF format with a 0.5-degree spatial resolution and decadal temporal resolution for the last 22,000 years. Bioclimatic variables (BV, O’Donnell & Ignizio, 2012) were calculated for each timestep using the biovars function in the dismo R package (Hijmans et al. 2016). BV values were then extracted for the space-time location of each fossil occurrence. The dataset was then filtered to include only the six least correlated BV covariates, using the variance inflation factor (VIF, Obrien, 2007). The variables retained were BV2 (mean diurnal temperature range), BV7 (annual temperature range), BV8 (mean temperature of wettest quarter), BV15 (precipitation seasonality), BV17 (precipitation of warmest quarter), and BV18 (precipitation of driest quarter).

Downscaled future climate layers for the year 2100 CE were obtained for the HadCM3 climate model (Lorenz et al. 2016), for the CMIP5 RCP 8.5 scenario which assumes high population, moderate economic growth, and a sustained dependence on fossil fuels (Riahi et al., 2011). These layers were processed as above.

### Computing Infrastructure

Google Cloud Compute Engine (GCE) cloud-based virtual machines were used for all model runs. Google’s platform was chosen over other public cloud vendors because of its ability to create custom hardware configurations that adhere to user-defined specifications. Other vendors (e.g., Amazon Web Services) provide a larger number of predefined instance types, but do not support the creation of an instance with arbitrary hardware.

The experimental system is illustrated in full in Appendix B. In brief, a master node-compute node infrastructure was devised so that a single server monitored the progression through the experimental design and controlled the provisioning of computing nodes. The master node (e.g., cloud-based virtual machine) ran a python control script attached to a centralized MySQL relational database via a Node.js API. The database contained the parameters for all the experiments to be undertaken, including both hardware requirements and algorithm parameters. The control script drew rows at random from the database and executed the initialization of a computing node with the corresponding hardware configuration using the GCE API. The computing nodes all ran Debian Linux 8. All experimental data (e.g., occurrence records, environmental layers) required to compute the SDM were stored in a private GitHub repository and were transferred to the fresh VM using Git. Once booted and provided with data, the computing node automatically ran the SDM using the assigned algorithm. Upon completion, the node communicated the runtime and accuracy back to the central database on the master node, and was subsequently released.

## Estimating and Modeling SDM Runtime, Cost, and Accuracy

Once the hardware configuration (CPUs and memory) and the data parameters (number of covariates and number of training examples) were communicated to the computing node, the set of pre-processed occurrence points was partitioned into a testing set (20%) and a training set (80%) of the total number of training points. An SDM was fit to the training data, assessed for accuracy, and then projected to the modeled future climate. Accuracy was evaluated using the testing set and quantified using the Area Under the Curve (AUC) statistic. Runtime (in seconds) was estimated within R using the proc.time function. No database I/O was done inside the timing script, so network connection speed is not expected to influence the results.

In total, 26,730 experimental trials were made, with each trial consisting of a particular combination of CPU cores, server memory, and number of training examples, number of environmental covariates, and number of cells in the prediction layers. Configurations were chosen to maximize the parameter space covered in the analysis while maintaining at least three replicates per configuration. Where feasible (see “Limitations”), more replicates were made. Once data collection was completed, SDM runtime and accuracy were modeled using Bayesian additive regression trees (BART), fit with the bartMachine R package, version 1.2.3 (Kapelner & Bleich, 2016). The Bayesian model fits a probability distribution for the response at each leaf node, rather than the standard single maximum likelihood estimate. A boosted ensemble size of 50 trees was used, and models were fit using default priors on the parameters and hyperparameters as suggested by the authors (Appendix C). Runtime and accuracy were modeled separately for each SDM. Runtime was modeled on a log scale (log-seconds), which improves predictive skill for on high-variance datasets (Hutter et al., 2014). The observed runtime and accuracy data for each SDM was randomly split into a training set (80%) and testing set (20%) for evaluation. 1250 MCMC iterations were performed, each of which built an entire additive model ensemble. The first 250 iterations were discarded as burn-in, leaving 1000 posterior samples to analyze and evaluate.

The predictive skill of the runtime and accuracy models was evaluated using the mean squared error (MSE), the statistic between observed and predicted values from the mean of the posterior distribution, and the standard deviation of the prediction posterior. Model results were also visually assessed by plotting the predicted values against the observed data and qualitatively assessing deviations from the y=x line.

The influence of each predictor in the runtime and accuracy models was evaluated by leave-one-out cross-validation, in which models were separately built using four of the five predictors. Each predictor was left out of a model in turn, and the of the subset model was evaluated and compared to the of the full model. The reduction in was interpreted as the predictive strength of the left-out variable.

## Optimal Prediction

Prediction of the optimal data-hardware configuration for an SDM followed a four-step process (Figure 6). First, the accuracy model was used to identify the data configuration that maximized accuracy. Second, the performance model was used to predict the execution time of the accuracy-maximizing model run under various hardware configurations. Third, multidimensional hierarchical clustering was used to assemble groups of configurations similar to one another. Finally, the hardware cluster with the lowest runtime, cost, and uncertainty was selected as optimal. This approach assumes that the researcher chooses spatial resolution (number of cells in the prediction) *a priori*, and so this variable is not included in optimization.

For the first step, the accuracy model for a given SDM was used to predict the accuracy of 500 regularly spaced configurations. These configurations included training dataset sizes between 0 and 10,000 occurrences at an interval of 100 and covariates between one and five. Hardware configurations were chosen between 1 and 24 CPU cores and 1 and 24 GB of memory at an interval of 2 GB. Predictions were sorted, first by descending order of accuracy, then by ascending order of training dataset size, and finally by ascending number of covariates. Hence, given equal accuracy, the configuration that requires the smaller training dataset was preferentially chosen.

In the second step, the accuracy-maximizing data configuration, and thus expected accuracy, was held constant and used as an input for the performance model. The performance model was used to predict the runtime of a SDM experiment with the given data configuration for a set of 288 CPU and memory configurations. Candidate configurations were chosen from GCE allowable custom instance types, and covered the parameter space between 1 and 25 cores and 1 and 25 GB of memory. Each runtime prediction was evaluated as 1000 samples from the posterior distribution provided by bartMachine. The mean of the distribution was used to calculate runtime cost, using GCE rates (Cost = $0.03492/CPU + $0.00468/GB; Google, Inc, 2017), and the standard deviation was used as a measure of prediction uncertainty. The dataset was subsequently scaled and centered using the R function scale.

The runtime predictions were then clustered using complete linkage hierarchical clustering on runtime, run cost, and prediction uncertainty. Prior to clustering, the dataset was centered and scaled using the R scale function. The results were plotted as a dendrogram and demarcated into clusters using the silhouette rule for maximizing within-cluster homogeneity while maximizing out-of-cluster variance (Rousseeuw, 1987).

Finally, the clusters were plotted in time-cost-uncertainty space. The hypothetical ideal scenario would involve no time, no cost, and no uncertainty, which occurs at the origin of these three axes. The Euclidean distance between the centroid of each cluster and the origin was calculated and the cluster with the smallest distance to the origin was identified as the optimal set of hardware configurations for that SDM. The Euclidean distance metric normalizes each dimension by its mean and standard deviation, eliminating the problem of optimizing using data with different units and variances.

# **Limitations and Extensibility of this Framework**

This approach has several important limitations. Perhaps most importantly, while a real SDM workflow contains various pre- and post-processing steps (eq. 1), the analyses here focus only on computing time (T*Compute*). Overall workflow time depends on many factors that may be difficult or impossible to model, including data availability and user skill and motivation. Future work could be pointed towards modeling these factors in a form that could be incorporated into this predictive model framework. Furthermore, the hardware analysis includes only two components, CPU cores and memory. Other hardware components may improve model results, but are difficult to manipulate experimentally.

A second limitation of the approach described here is that these analyses are limited to virtual computing instances hosted on Google Cloud Compute Engine (GCE), rather than real-world physical machines, limiting the conclusions that can be drawn about optimization solutions based on physical hardware configurations. Nonetheless, this uniform experimental design adds validity to the benchmark estimates of computing time, by providing a consistent computing environment unaffected by other tasks or concurrent programs (Dongarra et al., 1987). Nonetheless, because real-world machines do have many concurrent and interacting processes running at any one time, this may bias the results. Moreover, by using GCE, I cannot experimentally vary the CPU clock rate as a hardware component. Similarly, using GCE VM instances limits this experiment to only the processors provided by Google, which may be changed in the future. At the current time, GCE provides only one processor type for the VMs I used, a 2.6 GHz Intel Xeon E5 processor.

This work is limited to the analysis of data-driven SDMs, because systematic literature review suggests that a majority of SDM users use these methods. Furthermore, I focus on the most popular R implementations of these SDMs, for similar reasons. While there are known limitations to the language design and speed of R (e.g., Morandat, 2012) the platform is the most widely used for SDM analysis.

A final limitation was computational cost. In order to gather enough data and replicates to develop predictive models of SDM runtime and accuracy, I limited the number of models running longer than several hours and the number of replicates. Similarly, experiments on virtual servers with expensive hardware configurations were limited to allow for more experimentation on less costly servers. More data collected in all portions of configuration space, particularly on virtual instances with high memory and many CPUs, may improve the robustness of the results presented here.

Most importantly, the specific estimates of runtime, accuracy, and optimal computing configuration are likely to quickly become outdated as hardware capabilities continue to improve and computing providers change pricing schemes. The optimization framework, however, is general and robust to both changing hardware capabilities and costs. To incorporate new hardware configurations, additional data can be collected using methods similar to those described here, the models refit, and the predictions re-run. If other parameters (e.g., processor speed) remain the same, new data can be directly appended to the existing dataset. If not, an entirely new dataset should be collected. To incorporate new pricing schemes, only the third and fourth steps of the optimization (clustering and distance calculations) would need to be re-run using the new costs.

Although these results are specific to the SDMs, parameterizations, and implementations described here, the framework could be applied to any predictive modeling workflow. This approach relies on algorithm inputs and hardware capabilities and has no intrinsic relationship to SDM. To extend the framework to additional algorithms or hardware or data components, a new set of empirical trials would be required to gather a dataset of runtime and accuracy. Care must be paid to completely and evenly covering the full parameter space. Even models with many parameters could be incorporated, by treating algorithm parameters as additional components (Hutter et al., 2014). Furthermore, the results presented here are likely to be robust between different cloud service providers. Because the data was collected on hardware dedicated to running the SDMs with no concurrently running programs, VMs with similar specifications are expected to behave similarly, regardless of provider. However, further work should be directed at confirming this hypothesis.

# **Results**

## Optimal data configuration

For most SDMs, the simulations with the highest predictive accuracies occur when dataset size and number of covariates are large (Fig. 10). Though in this study the total possible number of training examples and covariates was limited due by experimental design, and supplying additional data may further increase SDM accuracy, literature suggests that typical SDM workflows often use less than 10,000 training examples and five covariates. Interestingly, in contrast to other SDMs, MARS achieves its highest predicted accuracy with only 1000 training examples, and only the addition of more covariates can increase accuracy. However, for all SDMs, additional covariates continued to increase accuracy up to the five covariates included here. These findings are consistent with theory and prior empirical analyses (Hastie et al., 2009), and underscores the value of big data approaches in biodiversity modeling.

## Model Performance

Predictive models of runtime were skillful when compared to a holdout testing set (Table 1, Figure 7). While results varied across SDM classes, the models for each SDM explained more than 50% of the variance in the runtime data ( and more than 87% of the variance in the accuracy data (). The most explanative performance models were for GBM-BRT and MARS, with values of approximately 0.96, and an MSEs of approximately 0.05 log-seconds. The estimates from the runtime models were tightly constrained, with low mean standard deviation of the prediction posteriors, ranging between 0.01 and 0.035 log-seconds, suggesting rather low uncertainty in the predictions.

For the runtime models, GAM and RF had lower values than the other two models. Interestingly, however, the GAM runtime model had a relatively low MSE (0.01 log-s) compared to the other models, while RF had the highest MSE of all models (0.64 log-s). It is likely that several factors contributed to the lower explanation of variance by GAM and RF. The GAM trials tended to converge within several seconds (maximum 10.3s), regardless of data or hardware configuration, exposing these trials to a stronger influence by low-level system processes not explicit in the runtime model and resulting in a higher variance dataset with lower predictive power. In contrast, the other three SDMs took minutes to hours to terminate (maximum: GBM-BRT, 5285.0s). Secondly, the GAM (training set size = 2,636) and RF (2,861) models are fit with smaller datasets than the GBM-BRT (9,256) and MARS (6,632) models, which may partially explain their relatively lower predictive skill. While the parameter space appears to be relatively well covered for these models, additional data, including more replications and parameterizations, may enhance model skill. Nonetheless, all models explained a majority of the variance in SDM runtime. While stochastic variance in system processes is often suggested as an impediment to modeling algorithm execution time (Lilja, 2009; Kalibera & Jones, 2013), the results shown here indicate that a skillful model can be produced despite these processes.

The accuracy models were more skillful than the runtime models (Table 2, Figure 8), in terms of . While ensemble size, tree depth, and other hyper parameters, not included in this modeling exercise, are suggested to exert control over learning accuracy (Hastie et al., 2009), the models included here without those parameters still proved skillful. The RF accuracy models was the best performing of the four, with an of of 0.98 and a MSE of less than 3.5x10-5 AUC. The lowest performing accuracy model was the GBM-BRT, with an of 0.87 and an MSE of 2.45x10-4 AUC. All of the accuracy models indicate low uncertainty and well-constrained posterior estimates on their predictions, with posterior standard deviations ranging from 0.0006 to 0.005 AUC. GAM predictions have the highest uncertainty associated with them (0.005 AUC), again perhaps due to small training set size.

## Controls on SDM Runtime

As hypothesized, the factors that control SDM runtime vary among algorithms (Table 3). One of the most important contributions is the number of training examples with which the algorithm is fit. The influence of this term on GBM-BRT, GAM, and RF runtime is large (>0.36 reduction in ). As data-driven algorithms, these SDMs rely heavily on creating structure from the given input dataset, and their runtime should be tied asymptotically to the number of training examples (Hastie et al., 2009). However, the number of environmental covariates is not an important predictor of runtime for any SDM; only GBM-BRT is influenced by this predictor, and only slightly. This finding is surprising because theoretical complexity suggests that learning algorithms are often asymptotically influenced by both training examples and covariates (Hastie et al, 2009; Cormen 2009). However, under real workloads, it appears that the number of covariates does not strongly influence runtime. In contrast to other SDMs, GAM runtime was weakly influenced by number of training examples and strongly controlled by the number of cells on which to predict the fitted model. This SDM’s learning time, as described above, is quick; nearly all of the total time during each experiment was spent predicting the model onto the novel climate scenario grid.

Among the SDMs, RF shows the strongest sensitivity to number of CPU cores, with 5.55% of variance explained (Table 3). For GAM, GBM-BRT, and MARS, CPU capacity accounts for less than 1% of total variance. This difference is consistent with SDM model structure: RF models build many alternate trees and are better able to take advantage of parallel computing configurations. Figure 11 demonstrates the diminishing marginal returns of the RF algorithm when run on multiple processors and on dataset with different numbers of training examples. The difference between a model run sequentially and one run in parallel with just two cores is large, while the marginal benefit of adding the 16th or 24th core is comparatively small. An infinitely parallel algorithm (E=1), incurs no diminished returns as additional cores are added. RF efficiencies vary by the number of training examples, but range between 0.05 and 0.4 at 25 cores. Importantly, workloads with more training examples, are better suited to parallelization -- have higher efficiency -- than small modeling problems with few training examples. For example, SDM runs with 10,000 training examples experience a much slower decline in efficiency than runs with 1,000 training examples.

SDM model accuracy is closely tied to training dataset size (Table 3). Together, the number of training examples and covariates accounted for over 50% of the model skill in all models. Indeed, for RF and GBM-BRT, data volume accounted for nearly 80% of the model’s total predictive skill. As seen in the runtime model, GAM is different than the other three SDMs, and is only strongly controlled by the number of covariates in the training set. As hypothesized, hardware configuration has little influence on SDM accuracy.

## Optimization

In general, the optimization suggests that SDMs require few CPU cores (Fig. 12). GBM-BRT is best suited to only one CPU core, while the GAM optimum lies at 3 CPUs. As suggested by moderate dependence on CPU cores in the runtime model and the ability to build ensemble members on separate cores, the optimal configuration for RF is between four and seven CPU cores.

Memory requirements are generally low (Fig. 12). RF and GBM-BRT are both optimized at only one gigabyte of memory. Hence, the optimal configuration for RF should purchase time on several cores but relatively little memory. GAMs are best suited to between one and 20 GB. Because of the clustering, each member of the optimal cluster is statistically identical, suggesting very little dependence on memory for GAM; one GB of memory is as suitable for running these models as 25 GB. Looking back at the contributors to the runtime of this model confirms that virtually no structure in the data is explained by VM memory. Therefore, this SDM can be interpreted as requiring only one GB of memory.

The MARS optimization procedure yielded strange and as yet unexplained results. The routine suggested that MARS require anywhere between 1 and 25 cores, and 16 GB of memory. Upon further investigation, it was initially thought that an error in sampling design had incorrectly weighted some of the parameter space in this model, potentially causing an artificial bias towards some hardware configurations. To obviate this issue, the observations of MARS runtime used to train the model were subsampled into a reduced dataset that had an even sampling of all potential hardware configurations. This reduced the dataset size from 6,632 observations to 617 observations, potentially causing a reduction in the robustness of the model predictions, since within-configuration variance would not be modeled, because there were no replicate trials.

However, upon fitting the reduced model and rerunning the optimization routine, the same pattern of optimizing at 16 GB of memory and between 1 and 25 cores was still apparent. These results are peculiar for two reasons. First, the assessment of the model drivers suggests that the model is not dependent on memory to make predictions (Table 3). Indeed, information of the amount of memory of the configuration yields only 1.7% increase in model skill. Second, the optimization does not sequentially favor hardware configurations with higher and higher amounts of memory; rather, configurations with less than or more than 16 GB of memory are shown to be suboptimal, and only those with exactly 16 GB shown to be are optimal. Should memory be of true importance, it would be logical to expect that the configurations would become closer to optimal as additional memory is added to the configuration. At the present time, it is unclear why they are occurring, and only in the MARS model. For this reason, this finding will be left as unexplained and future work will collect additional data to uncover the reason for this strange behavior.

# **Discussion**

## Overview

The results presented here show how the general optimization framework can be applied to biodiversity modeling to identify the optimal data-hardware configuration on which to run these models. As biodiversity database growth continues and cloud computing gains popularity in the research sphere, more work to optimize computing platforms is needed. A key finding of this work is that many of the highly accurate models, widely used in the contemporary literature, are not well structured to harness the flexible computing power enabled by cloud computing. The optimization framework is general, and could be applied to other ecological models with similar structure.

The findings presented here confirm all three of the hypotheses developed at the outset of this thesis. First, for each model, there does appear to be an optimal region, at which financial and time costs are jointly minimized. Second, the hardware configuration appears to exert minimal influence over the accuracy of the model. This is a fundamental justification for the framework developed here. Finally, the data volume, in both number of training examples and number of environmental covariates, influences both the time required to fit the model and the model’s accuracy. A tension is therefore created between the running more accurate models and running faster models.

## R-Based Biodiversity Models and High Performance VMs

These findings suggest that many current R implementations are generally insensitive to the high performance hardware made available by cloud computing. The R implementations of GAM, GBM-BRT[[3]](#footnote-3), and MARS are all fit sequentially, one instruction after another, on a single processor. The algorithms underlying the corresponding R functions are not easy to parallelize, as the model building process involves loops over the entire dataset, a procedure not easily split into smaller tasks suitable for multiple processors (Hastie et al., 2009). A sequential model should theoretically have little dependence on CPU cores, since R may use only one core during model building. Empirical results support this claim, showing that less than one percent of the variance in these three SDMs’ runtimes is explained by the number of CPU cores on the VM.Both GBM-BRT and GAM have clearly defined optimal configurations at a low number of CPU cores. In both cases, there is no demonstrated advantage of a higher CPU configuration.

Furthermore, no SDM, except MARS, responds to increased memory allocations on the VM. Although not tested rigorously, datasets exceeding 100MB, or several hundred thousand training examples, caused fatal crashes in R. R is known for poor memory management, specifically, functions often create multiple copies of data objects in both built-in and external packages (Johnson, 2012). When data size becomes large, making in-memory copies is not possible without exceeding total allocation, resulting in program crashes. While packages exist to handle datasets too large to fit into an instance’s main memory (https://cran.r-project.org/web/views/HighPerformanceComputing.html accessed October 10, 2016), SDM functions in popular packages (e.g., dismo) require significant modification before they can incorporate these tools.

Conversely, the RF algorithm can be easily split across into small subtasks, and so can easily leverage additional cores in powerful hardware configurations. Specifically, individual tree building is done in parallel on multiple cores, after which the model ensemble is assembled and evaluated on a single processor. The maximum expected accuracy is approximately that of GBM-BRT, but can be achieved in a fraction of the time. Both SDMs achieve maximal accuracy with 10,000 training examples and 5 covariates. However, when parallelized, random forests can be fit in between 10% and 30% of the time to fit a GBM-BRT model with the same data. Given the approximately equal accuracy, the economically rational researcher would therefore be best served by employing an RF, run in parallel across many cores, rather than waiting for the GBM-BRT to converge.

The RF optimal hardware configuration demonstrates well the trade off between monetary cost and time. Because the algorithm can make effective use of additional CPU cores, configurations with additional cores are associated with decreased execution time, but are charged a higher rate. Conversely, VMs with fewer cores take additional time to fit, but have a lower rate. When taken together, the two balance out -- illustrating the tension between cost and time when considering hardware provisioning for parallel algorithms.

RF, like other algorithms designed for parallel computation, is subject to diminishing speed returns as it is spread across additional cores (Gustafson, 1988). All algorithms must run, at least in part, sequentially, for example, during setup and ensemble combination. Because of the portion of code executed on a single processor, it is impossible to obtain an infinite speedup across infinite processors (Amdahl’s Law, Amdahl, 1967). Inter-core communication increases as addition processors are added -- eventually causing the benefits of parallelization to be offset by the extra overhead. This phenomenon is typically measured in an algorithm’s parallel efficiency, the difference between parallel (*TN*) and serial (*T*1) runtimes of the algorithm, divided by the number of cores (*N*), given as:

Biodiversity database size trends suggests that much more data will be available to researchers in fitting SDMs over the next decade. Therefore, parallel SDMs, like random forests, are likely to see increased benefits of running in a cloud-based environment, because the hardware can be easily scaled to meet problem needs and optimize efficiency.

## Extensions of the Optimization Approach

The unconstrained maximization routine is useful when neither data nor runtime is externally limited. However, in many real-world situations, scenarios when neither of these factors is limited are unlikely. Indeed, many SDM analyses consider datasets with less than 100 occurrences (Wisz et al. 2008). Constrained optimization, either on runtime or data volume can be useful in these situations. For example, a constraint on the number of training examples or covariates available to the researcher seems probable, as the low-friction availability of biodiversity records in databases like Neotoma and GBIF if more data were available, it would be incorporated. In this analysis, the space of potential data configurations first would be truncated to include only those configurations possible under the constraint (e.g., configurations with less than the number in the dataset). The accuracy-maximizing point is then selected from the subspace, rather than the full space, to reflect a point feasible within the data limitations. The optimization then continues as above.

It can also be useful to place a hard maximum bound on the execution time or cost of the SDM. For example, consider a cloud-based SDM application that computes SDMs for arbitrary datasets remotely and returns the results to a client over the Internet. Users of interactive web applications are apt to loose interest and turn their attention to other tasks if the application takes more than several seconds to respond (Roth, 2013). If a maximum-accuracy experiment for GBM-BRT is requested, the client faces a response time of over 1 hour and would quickly stop interacting with the application. It would benefit the application developer, then, to limit the time it takes to complete the model run so that (a) their costs are minimized and (b) they retain the user’s interest and business. To accomplish this, a large set of potential configurations is generated and the runtime for each is predicted using the performance model. Those configurations that have predicted runtimes that fall within the constraint are fed into the accuracy model. These results are then sorted by accuracy to yield an estimate of the highest-accuracy configuration that would fall below the runtime threshold. By constraining the optimal with a hard maximum bound on the time it takes to compute the model, the management can provision the cheapest resources, return results of high accuracy, and continue to keep their audience engaged.

Another potential extension of this work is weighted optimization to preferentially influence one or two of the three components included in the determination of the optimal clusters. In a standard optimization, run time, run cost, and prediction uncertainty are all weighted equally. However, a given researcher may care different amounts about each of these axes. For example, she may care most about achieving low-cost, next about the level of certainty on the prediction, and finally about the time it takes to complete the modeling. Weighting each axis would allow the user customize the optimization to meet her own needs.

## Cloud Computing in Biodiversity Modeling: Recommendations and Prospects

The observations of collected here suggest fairly short runtimes and very low costs. Even the longest observations of runtime are only several hours, far short of the several weeks reported by Elith et al. (2006). However, as data volumes grow and modeling approaches to biodiversity problems become more common, even these short runtimes become formidable. Many studies now model hundreds or even thousands of species (e.g., Rezende et al., 2015). Some projects attempt to model entire biotas; for example, Candela et al. (2013) describe routinely modeling and mapping distributions of over 11,000 marine species. Most contemporary SDM studies use multiple model classes, predict to multiple time periods, and use multiple climate scenarios. When fitting multiple models for thousands of taxa, spending several hours for each model is not feasible. Therefore, while the individual results presented here may seem small, they add up quickly for all but the smallest modeling activities typical in contemporary SDM literature.

The relatively minor contributions of hardware to model execution time, suggest that the sequential models may not be good candidates for a transition to the pay-as-you-go, utility model of cloud computing. Because GAM, GBM-BRT, and MARS are all optimized with configurations of few cores and low memory, it may not benefit the model user to cloud-enable these SDMs. Rather, the user is likely better off using the lowest-cost hardware available. Provisioning additional cores or memory is unlikely to bring returns in execution time, reduce modeling cost, or increase prediction certainty. For these models, to achieve performance gains in a cloud-based system, a system would need to be developed that simultaneously provisioned multiple, isolated VM instances to run independently, each fitting a single SDM. To make this work automatically, efficiently, and without error would be difficult and require a significant amount of technical skill.

RF, however, may be a good candidate for incorporation into a cloud-based runtime environment because of its native support for multiple-CPU parallelism. Because RF efficiency scales with training set size, it is well suited to the scalability provided by the cloud. The ability to scale the number of CPUs based on problem-specific features would be helpful to model users interested in reducing model runtime. Public cloud providers make scalability easily, allowing users to put code and programs onto a virtual hard disk and scaling the underlying hardware configuration up or down.

Of course, even for RF, the difficulty of the transition to a cloud-based solution may outweigh any benefits, particularly for small workflows. It is not a trivial task to provision and configure virtual hardware and install and prepare modeling software effectively on a cloud instance. However, for large modeling workflows with thousands of species and many prediction scenarios, for researchers experienced with cloud-based VMs, or for server-based SDM geoprocessing (e.g, Souza Muñoz et al., 2009) as a service (e.g., Granell, 2013), cost and time optimization will be helpful.

The sensitivity of SDM execution time to training dataset size suggests that popular SDM algorithms like those examined here may be unable to cope with large training datasets possible as biodiversity database size increases. Austin (2007) argues that a solid foundation of ecological theory is the most essential factor in correctly predicting species ranges and testing hypotheses with SDMs. Indeed, he claims that the ecological underpinnings of the statistics may be even more important the statistical method itself. Elith & Leathwick (2009) follow, suggesting that additional improvements in species distribution modeling will come not from novel learning algorithms, but from the incorporation of more ecologically relevant information into the statistical modeling process, claiming, “further advances in SDM are more likely to come from better integration of theory, concepts, and practice than from improved methods per se.”

Due to rapidity of database growth, I contend that modelers should focus their effort not only on the incorporation of ecological realism, but on optimizing existing and novel models to take advantage of parallelism, high performance libraries, and cloud computing as well. While ecological datasets may not have been “big” in the past, they are now. New model development efforts should be undertaken to advance models that effectively leverage high performance computing infrastructure, multiple computing cores, effective memory management strategies, and scalability. While some attempts to refactor traditionally sequential models to run in parallel have been successful, these studies used entirely novel implementations of the traditional techniques (GBM-BRT, Tyree et al., 2011). Together the incorporation of additional ecological relevant information will help to strengthen the ecological inference made from these models and the computational optimization of modeling algorithms will facilitate inference from much more data.

# References

Amdahl, G. M. (1967). Validity of the single processor approach to achieving large scale computing capabilities (p. 483). Presented at the Proceedings of the Spring Joint Computer Conference. April 18-20, New York, New York, USA: ACM Press. http://doi.org/10.1145/1465482.1465560

Araújo, M., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, *22*(1), 42–47.

Araújo, M., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*(10), 1677–1688.

Araújo, M., Cabeza, M., Thuiller, W., Hannah, L., & Williams, P. H. (2004). Would climate change drive species out of reserves? An assessment of existing reserve-selection methods. *Global Change Biology*, *10*(9), 1618–1626.

Araújo, M., Whittaker, R. J., Ladle, R. J., & Erhard, M. (2005). Reducing uncertainty in projections of extinction risk from climate change. *Global Ecology and Biogeography*, *14*(6), 529–538.

Fox, Armando, et al. "Above the clouds: A Berkeley view of cloud computing." *Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS* 28.13 (2009): 2009.

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, *200*(1-2), 1–19.

Austin, M. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, *157*(2-3), 1–18.

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species’ geographic distributions. *Ecological Informatics*, *19*(C), 10–15.

Bifet, A., Holmes, G., Pfahringer, B., & Gavalda, R. (2011). Detecting Sentiment Change in Twitter Streaming Data. In *WAPA* (pp. 5-11).

Blaauw, M. (2010). Methods and code for “classical” age-modelling of radiocarbon sequences. *Quaternary Geochronology*, *5*(5), 512–518.

Blewitt, G., Hammond, W. C., Kreemer, C., Plag, H.-P., Stein, S., & Okal, E. (2009). GPS for real-time earthquake source determination and tsunami warning systems. *Journal of Geodesy*, *83*(3-4), 335–343. http://doi.org/10.1007/s00190-008-0262-5

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135.

Breiman, L. (2006). randomForest: Breiman and Cutler’s random forests for classification and regression. R package version 4.6-12.

Brewer, E. A. (1995). High-level optimization via automated statistical modeling. In *ACM SIGPLAN Notices* (Vol. 30, No. 8, pp. 80-91). ACM.

Brewer, S., Jackson, S. T., & Williams, J. W. (2012). Paleoecoinformatics: applying geohistorical data to ecological questions. *Trends in Ecology & Evolution*, *27*(2), 104–112.

Candela, L., Castelli, D., Coro, G., Pagano, P., & Sinibaldi, F. (2013). Species distribution modeling in the cloud. *Concurrency and Computation: Practice and Experience*, *28*(4), 1056–1079.

Cannon, A. R., & John, C. H. S. (2007). Measuring empirical computational complexity. *Organizational Research Methods*, *10*(2), 1–10.

Carroll, J. M. (2000). Five reasons for scenario-based design. *Interacting with Computers*, *13*(1), 43–60. http://doi.org/10.1016/s0953-5438(00)00023-0

Chamberlain, S., Ram, K., Barve V., & McGlinn, D.(2016). rgbif: Interface to the Global 'Biodiversity' Information Facility 'API'. R package version 0.9.4. https://CRAN.R-project.org/package=rgbif

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209.

Clark, J. S., Gelfand, A. E., Woodall, C. W., & Zhu, K. (2014). More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, *24*(5), 990–999.

Cormen, T. H. (2009). *Introduction to Algorithms*. MIT Press.

Davis, M. B. (1963). On the theory of pollen analysis. *American Journal of Science*, *261*(10), 897–912.

Dawson, A., Paciorek, C. J., McLachlan, J. S., Goring, S., Williams, J. W., & Jackson, S. T. (2016). Quantifying pollen-vegetation relationships to reconstruct ancient forests using 19th-century forest composition and pollen data. *Quaternary Science Reviews*, *137*(C), 156–175. http://doi.org/10.1016/j.quascirev.2016.01.012

Dongarra, J., Martin, J. L., & Worlton, J. (1987). Computer benchmarking: Paths and pitfalls. *IEEE Spectrum*, *24*(7), 38–43.

Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., et al. (2012). Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, *39*(12), 2119–2131.

Drake, J. M., Randin, C., & Guisan, A. (2006). Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, *43*(3), 424–432.

Elith, J., & Graham, C. H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, *32*(1), 66–77.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697.

Elith, J., H Graham, C., P Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J Hijmans, R., et al. (2006). Novel methods improve prediction of species distributions from occurrence data. *Ecography*, *29*(2), 129–151.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2010). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43–57.

Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology Letters*, *7*(6), 509–520.

Ficetola, G. F., Thuiller, W., & Miaud, C. (2007). Prediction and validation of the potential global distribution of a problematic alien invasive species - the American bullfrog. *Diversity and Distributions*, *13*(4), 476–485.

Fink, E. (1998). How to solve it automatically: Selection among problem solving methods. In *AIPS* (pp. 128-136).

Fitzpatrick, M. C., Gotelli, N. J., & Ellison, A. M. (2013). MaxEnt versus MaxLike: empirical comparisons with ant species distributions. *Ecosphere*, *4*(5), 55–15.

Fløjgaard, C., Normand, S., & Skov, F. (2009). Ice age distributions of European small mammals: insights from species distribution modelling. *Journal of Biogeography*, *36*(6), 1152–1163.

Foster, I., Zhao, Y., Raicu, I., & Lu, S. (2008). Cloud computing and grid computing 360-degree compared. *IEEE Grid Computing Environments and IEEE/ACM Supercomputing*, 1–10. http://doi.org/10.1109/gce.2008.4738445

Franklin, J. (2009). *Mapping Species Distributions*. Cambridge University Press. http://doi.org/10.1017/s0030605310001201

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5), 189-1232.

Fritz, S. A., Schnitzler, J., Eronen, J. T., Hof, C., Böhning-Gaese, K., & Graham, C. H. (2013). Diversity in time and space: wanted dead and alive. *Trends in Ecology & Evolution*, *28*(9), 509–516.

Glew, J. R., Smol, J. P., & Last, W. M. (2002). Sediment core collection and extrusion. In *Tracking environmental change using lake sediments* (pp. 73–105). Dordrecht: Springer Netherlands.

Golding, N., & Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, *7*(5), 598–608.

Goldsmith, S. F., Aiken, A. S., & Wilkerson, D. S. (2007). *Measuring Empirical Computational Complexity*. New York, New York, USA: ACM.

Google, Inc. (2017). *Google Compute Engine Pricing*. Retrieved from https://cloud.google.com/compute/pricing#custommachinetypepricing

Goring, S., Dawson, A., Simpson, G. L., Ram, K., Graham, R. W., Grimm, E. C., & Williams, J. W. (2015). neotoma: A Programmatic Interface to the Neotoma Paleoecological Database. *Open Quaternary*, *1*(1).

Granell, C., Díaz, L., Schade, S., Ostländer, N., & Huerta, J. (2013). Enhancing integrated environmental modelling by designing resource-oriented interfaces. *Environmental Modelling and Software*, *39*(C), 229–246.

Grimm, E. C., Bradshaw, R. H. W., Brewer, S., Flantua, S., Glesecke, T., Lézine, A.-M., Takahara, H., et al. (2013). Databases and Their Application. *Encyclopedia of Quaternary Science*, 831–838.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009.

Guisan, A., & Zimmerman, N. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2-3), 1–40.

Guisan, A., Edwards, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, *157*(2-3), 89–100.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J. M. C., Aspinall, R., & Hastie, T. (2006). Making better biogeographical predictions of species’ distributions. *Journal of Applied Ecology*, *43*(3), 386–392.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., et al. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*(12), 1424–1435.

Gustafson, J. L. (1988). Reevaluating Amdahl’s law. *Communications of the ACM*, *31*(5), 532–533.

Hamann, A., & Wang, T. (2006). Potential effects of climate change on ecosystem and tree species distribution in British Columbia. *Ecology*, *87*(11), 2773–2786.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C.S., & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156–162.

Hassan, Q. (2011). Demystifying cloud computing. *CrossTalk*, 16–21.

Hastie, T. (2015). *gam: Generalized Additive Models*. R package version 1.12.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York.

Hegel, T. M., Cushman, S. A., Evans, J., & Huettmann, F. (2010). Current state of the art for statistical modelling of species distributions. In *Spatial complexity, informatics, and wildlife conservation* (pp. 273–311). Tokyo: Springer Japan.

Heidorn, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, *57*(2), 280–299. http://doi.org/10.1353/lib.0.0036

Hijmans, R.J., Phillips, S., Leathwick, J. and Elith, J. (2016). dismo: Species Distribution Modeling. R package version 1.1-1. https://CRAN.R-project.org/package=dismo

Hobbie, J. E., Carpenter, S. R., Grimm, N. B., Gosz, J. R., & Seastedt, T. R. (2003). The US Long Term Ecological Research Program. *BioScience*, *53*(1), 21–32.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S. & Twigger, S. (2008). Big data: The future of biocuration. *Nature*, *455*(7209), 47-50.

Hsu, C., Lin, C.-Y., Ouyang, M., & Guo, Y. K. (2013). Biocloud: Cloud computing for biological, genomics, and drug design. *BioMed Research International*, *2013*, 1–3.

Huang, Q., Yang, C., Liu, K., Xia, J., Xu, C., Li, J., Gui, Z., et al. (2013). Evaluating open-source cloud computing solutions for geosciences. *Computers and Geosciences*, *59*(C), 41–52.

Huang, Q., Yang, C., Nebert, D., Liu, K., & Wu, H. (2010). Cloud computing for geosciences: deployment of GEOSS clearinghouse on Amazon's EC2. Presented at the Proceedings of the ACM. <http://doi.org/10.1145/1869692.1869699>

Hutchinson, G. E. (1957). Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, *22*(0), 1–7.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Identifying key algorithm parameters and instance features using forward selection. *Lecture Notes in Computer Science* (Vol. 7997, pp. 364–381). Berlin, Heidelberg: Springer Berlin Heidelberg. <http://doi.org/10.1007/978-3-642-44973-4_40>

Hutter, F., Xu, L., Hoos, H. H., & Leyton-Brown, K. (2014). Algorithm runtime prediction: Methods & evaluation. *Artificial Intelligence*, *206*, 79–111. http://doi.org/10.1016/j.artint.2013.10.003

Issa, S. A., Kienzler, R., El-Kalioby, M., Tonellato, P. J., Wall, D., Bruggmann, R., & Abouelhoda, M. (2013). Streaming support for data intensive cloud-based sequence analysis. *BioMed Research International*, *2013*(8), 1–16.

Johnson, K. (2012). Evaluating the design of the R language. *Designing Language Teaching Tasks*, (Chapter 8), 1–27.

Jones, R. & Kalibera, T. (2013). Rigorous benchmarking in reasonable time. *ACM SIGPLAN Notices*, *48*(11), 63–74. http://doi.org/10.1145/2555670.2464160

Kaján, L., Yachdav, G., Vicedo, E., Steinegger, M., Mirdita, M., Angermüller, C., Böhm, A., et al. (2013). Cloud prediction of protein structure and function with PredictProtein for Debian. *BioMed Research International*, *2013*(3), 1–6.

Kapelner A. & Bleich J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. Journal of Statistical Software, 70(4), 1-40. doi:10.18637/jss.v070.i04

Keppel, G., Van Niel, K. P., Wardell-Johnson, G. W., Yates, C. J., Byrne, M., Mucina, L., Schut, A. G. T., et al. (2011). Refugia: identifying and understanding safe havens for biodiversity under climate change. *Global Ecology and Biogeography*, *21*(4), 393–404.

Knuth, D. E. (1976). Big Omicron and big Omega and big Theta. *ACM Sigact News*, *8*(2), 18–24. http://doi.org/10.1145/1008328.1008329

Kogan, J. (2014). Feature selection over distributed data streams. In K. Yada (ed.), *Data Mining for Service*, *Studies in Big Data 3*, (pp. 11–26). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kundra, V. (2010). *25 Point Implementation Plan to Reform Federal Information Technology Management*. Executive Office Of The President/Office Of Management And Budget Office Of E-Government And Information Technology. Washington, DC, USA.

Leathwick, J. R., Elith, J., & Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling*, *199*(2), 188–196.

Leyton-Brown, K., Nudelman, E., & Andrew, G. (2003). A portfolio approach to algorithm selection. *IJCAI (1543)*.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Lilja, D. (2009). *Measuring Computer Performance: A Practitioner’s Guide*. Cambridge: Cambridge University Press.

Loiselle, B.A., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G., & Williams, P. H. (2003). Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology* *17*(6), 1591–1600.

Lorenz, D. J., Nieto-Lugilde, D., Blois, J. L., Fitzpatrick, M. C., & Williams, J. W. (2016). Downscaled and debiased climate simulations for North America from 21,000 years ago to 2100AD. *Scientific Data*, *3*, 160048–19.

Lowe, C. B., Kellis, M., Siepel, A., Raney, B. J., Clamp, M., Salama, S. R., Kingsley, D. M., et al. (2011). Rapid range shifts of species associated with high levels of climate warming. *Science*, *333*(6045), 1019–1024.

Lu, S., Li, R. M., Tjhi, W. C., Lee, K. K., Wang, L., Li, X., & Ma, D. (2011). A framework for cloud-based large-scale data analytics and visualization: A case study on multiscale climate data. In *2011 IEEE 3rd International Conference on Cloud Computing Technology and Science* (pp. 618–622). IEEE.

Maguire, K. C., Nieto-Lugilde, D., Fitzpatrick, M. C., Williams, J. W., & Blois, J. L. (2015). Modeling species and community responses to past, present, and future episodes of climatic and ecological change. *Annual Review of Ecology, Evolution, and Systematics*, *46*(1), 343–368.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., et al. (2011). Big Data. *McKinsey Global Institute,* 1-147.

Mell, P. M., & Grance, T. (2012). The NIST definition of cloud computing. *National Institute of Standards and Technology*.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, *27*(2), 85–93.

Milborrow, S. (2016). *earth: Multivariate Adaptive Regression Splines*. R package version 4.4.4. https://CRAN.R-project.org/package=earth

Miller, J., Franklin, J., & Aspinall, R. (2007). Incorporating spatial dependence in predictive vegetation models. *Ecological Modelling*, *202*(3-4), 225–242.

Morandat, F., Hill, B., Osvald, L., & Vitek, J. (2012). Evaluating the design of the R language. In *ECOOP 2012 – Object-Oriented Programming* (Vol. 7313, pp. 104–131). Berlin, Heidelberg: Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-31057-7\_6

Mosco, V. (2015). *To the Cloud: Big Data in a Turbulent World*. Boulder, Colorado, USA: Routledge.

Office of the Inspector General (2013). *NASA’s progress in adopting cloud-computing technologies.*(pp. 1–38).

Natekin, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*, 1–21.

National Science Board (2016). Science & engineering indicators 2016 (NSB-2016-1). National Science Foundation.

National Science Foundation. (2012). *NSF Report on Support for Cloud Computing* (No. 12040) (pp. 1–21).

National Science Foundation. (2014). Enabling a new future for cloud computing. *nsf.gov*. Accessed from https://nsf.gov/news/news\_summ.jsp?cntn\_id=132377.

Nogués-Bravo, D. (2009). Predicting the past distribution of species climatic niches. *Global Ecology and Biogeography*, *18*(5), 521–531.

Nogués-Bravo, D., Rodríguez, J., Hortal, J., & Batra, P. (2008). Climate change, humans, and the extinction of the woolly mammoth. *PLoS Biology*, *6*(4), e79.

Norman, D. A. (1984). Stages and levels in human-machine interaction. *International journal of man-machine studies*, *21*(4), 365–375.

Obrien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673–690.

O’Donnell, M. S., & Ignizio, D. A. (2012). Bioclimatic predictors for supporting ecological applications in the conterminous United States. *US Geological Survey Data Series*, *691*(10).

Pearman, P. B., Guisan, A., Broennimann, O., & Randin, C. F. (2008). Niche dynamics in space and time. *Trends in Ecology & Evolution*, *23*(3), 149–158.

Peterson, A. T. (2003). Predicting the geography of species’ invasions via ecological niche modeling. *The Quarterly Review of Biology*, *78*(4), 419–433.

Peterson, A. T., Soberon, J., & Sánchez-Cordero, V. (1999). Conservatism of Ecological Niches in Evolutionary Time. *Science*, *285*(5431), 1265–1267.

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3-4), 231–259.

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Radeloff, V. C., Williams, J. W., Bateman, B. L., Burke, K. D., Carter, S. K., Childress, E. S., et al., (2015). The rise of novelty in ecosystems. *Ecological Applications*, *25*(8), 2051-2068.

Rezende, V. L., de Oliveira-Filho, A. T., Eisenlohr, P. V., Kamino, L. H. Y., & Vibrans, A. C. (2015). Restricted geographic distribution of tree species calls for urgent conservation efforts in the Subtropical Atlantic Forest. *Biodiversity and Conservation*, *24*(5), 1057-1071.

Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., Kindermann, G., et al. (2011). RCP 8.5A scenario of comparatively high greenhouse gas emissions. *Climatic Change*, *109*(1-2), 33–57.

Ridgeway, G. (2015). gbm: Generalized Boosted Regression Models. R package version 2.1.1. https://CRAN.R-project.org/package=gbm

Root, T. L., MacMynowski, D. P., Mastrandrea, M. D., & Schneider, S. H. (2005). Human-modified temperatures induce species changes: joint attribution. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(21), 7465-7469.

Rosson, M. B. (2002). Scenario-based design. In J. Jacko & A. Sears (Eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technology and Emerging Applications* (pp. 1032–1050). CRC Press.

Roth, R. E. (2013). Interactive maps: What we know and what we need to know. *Journal of Spatial Information Science*, *2013*(6), 59-115.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. http://doi.org/10.1016/0377-0427(87)90125-7

Sadjadi, S. M., Shimizu, S., Figueroa, J., Rangaswami, R., Delgado, J., Duran, H., & Collazo-Mojica, X. J. (2008). A modeling approach for estimating execution time of long-running scientific applications. In *Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium on* *IEEE* (pp. 1-8).

Salisbury, E. J. (1926). The geographical distribution of plants in relation to climatic factors. *The Geographical Journal*, *67*(4), 312-335.

Schatz, M. C., Langmead, B., & Salzberg, S. L. (2010). Cloud computing and the DNA data race. *Nature Biotechnology*, *28*(7), 691–693.

Schimel, D., M. Keller, P. Duffy, L. Alves, S. Aulenbach, W. Gram, B. Johnson et al. "The NEON strategy: Enabling continental scale ecological forecasting." *NEON Inc., Boulder, CO* (2009).

Schnase, J. L., Duffy, D. Q., McInerney, M. A., Webster, W. P., & Lee, T. J. (2016). *Climate analytics as a service*. New York: Elsevier.

Schnase, J. L., Duffy, D. Q., Tamkin, G. S., Nadeau, D., Thompson, J. H., Grieg, C. M., McInerney, M. A., et al. (2014b). MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service. *Computers, Environment and Urban Systems*, 198-211.

Simon, H. A. (1986). Rationality in psychology and economics. *Journal of Business*, S209-S224.

Smith, S. E., Mendoza, M. G., Zúñiga, G., Halbrook, K., Hayes, J. L., & Byrne, D. N. (2013). Predicting the distribution of a novel bark beetle and its pine hosts under future climate conditions. *Agricultural and Forest Entomology*, *15*(2), 212–226.

Snijders, C., Matzat, U., & Reips, U. D. (2012). " Big Data": big gaps of knowledge in the field of internet science. *International Journal of Internet Science*, *7*(1), 1-5.

Soberón, J., & Peterson, T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *359*(1444), 689-698.

Soberón, J., & Peterson, A. T. (2005). Interpretation of Models of Fundamental Ecological Niches and Species Distributional Areas. *Biodiversity Informatics (2) 1-10*.

Soberón, J., Arriaga, L., & Lara, L. (2002). Issues of quality control in large, mixed-origin entomological databases. *Towards a Global Biological Information Infrastructure*, Saarenmaa H. and Nielsen, E.S. Eds. European Environment Agency, 15–22.

Souza Muñoz, M. E. de, De Giovanni, R., Siqueira, M. F. de, Sutton, T., Brewer, P., Pereira, R. S., Canhos, D. A. L., et al. (2009). openModeller: a generic approach to species potential distribution modelling. *GeoInformatica*, *15*(1), 111–135.

Stein, A. F., Isakov, V., Godowitch, J., & Draxler, R. R. (2007). A hybrid modeling approach to resolve pollutant concentrations in an urban area. *Atmospheric Environment*, *41*(40), 9410–9426.

Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome biology*, *11*(5).

Stribling, J. B., Pavlik, K. L., Holdsworth, S. M., & Leppo, E. W. (2008). Data quality, performance, and uncertainty in taxonomic identification for biological assessments. *Journal of the North American Benthological Society*, *27*(4), 906–919. http://doi.org/10.1899/07-175.1

Sun, K., & Li, Y. (2013). Effort estimation in cloud migration process. In *2013 IEEE 7th International Symposium on Service Oriented System Engineering (sose 2013)* (pp. 84–91). IEEE.

Svenning, J.-C., Fløjgaard, C., Marske, K. A., Nogués-Bravo, D., & Normand, S. (2011). Applications of species distribution modeling to paleobiology. *Quaternary Science Reviews*, *30*(21-22), 2930–2947.

Svenning, J. C., Normand, S., & Skov, F. (2008). Postglacial dispersal limitation of widespread forest plant species in nemoral Europe. *Ecography*, *31*(3), 316-326.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293.

Thomas, C. D. (2010). Climate, climate change and range boundaries. *Diversity and Distributions*, *16*(3), 488–495.

Thuiller, W. (2007). Biodiversity: Climate change and the ecologist. *Nature*, *448*(7153), 550–552.

Thuiller, W., Albert, C., Araújo, M. B., Berry, P. M., Cabeza, M., Guisan, A., Hickler, T., et al. (2008). Predicting global change impacts on plant species distributions: Future challenges. *Perspectives in Plant Ecology, Evolution and Systematics*, *9*(3), 137–152.

Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T., & Prentice, I. C. (2005). Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences of the united States of America*, *102*(23), 8245-8250.

Tyree, S., Weinberger, K. Q., Agrawal, K., & Paykin, J. (2011, March). Parallel boosted regression trees for web search ranking. In *Proceedings of the 20th international conference on the World Wide Web* (pp. 387-396). ACM.

Uhen, M. D., Barnosky, A. D., Bills, B., Blois, J., Carrano, M. T., Carrasco, M. A., et al. (2013). From card catalogs to computers: databases in vertebrate paleontology. *Journal of Vertebrate Paleontology*, *33*(1), 13–28. http://doi.org/10.1080/02724634.2012.716114

Urban, M. C. (2015). Accelerating extinction risk from climate change. *Science*, *348*(6234), 571–573. http://doi.org/10.1126/science.aaa4984

Varela, S., Hernández J.G., and Sgarbi L.F. (2016). paleobioDB: Download and Process Data from the Paleobiology Database. R package version 0.5.0.

Václavík, T., & Meentemeyer, R. K. (2009). Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling*, *220*(23), 3248–3258.

Veloz, S. D., Williams, J. W., Blois, J. L., He, F., Otto-Bliesner, B., & Liu, Z. (2012). No-analog climates and shifting realized niches during the late Quaternary: implications for 21st-century predictions by species distribution models. *Global Change Biology*, *18*(5), 1698–1713.

Vieilledent, G., Latimer, A. M., Gelfand, A. E., & Merow, C. (2012). *hSDM: hierarchical Bayesian species distribution models*. R package version 1.4. https://CRAN.R-project.org/package=hSDM

Villars, R. L., Olofson, C. W., & Eastwood, M. (2011). Big data: What it is and why you should care. *White Paper, IDC*, 14.

Vincent, P. J., & Haworth, J. M. (1983). Poisson regression models of species abundance. *Journal of Biogeography*, *10*(2), 153–160.

Waltari, E., Hijmans, R. J., Peterson, A. T., Nyári, Á. S., Perkins, S. L., & Guralnick, R. P. (2007). Locating pleistocene refugia: Comparing phylogeographic and ecological niche model predictions. *PLoS ONE*, *2*(7), e563.

Williams, J. W., & Jackson, S. T. (2007). Novel climates, no-analog communities, and ecological surprises. *Frontiers in Ecology and the Environment*, *5*(9), 475–482.

Wing, M. G., Eklund, A., & Kellogg, L. D. (2005). Consumer-grade global positioning system (GPS) accuracy and reliability. *Journal of forestry*, *103*(4), 169-173.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., NCEAS Predicting Species Distributions Working Group. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*(5), 763–773. http://doi.org/10.1111/j.1472-4642.2008.00482.x

Woodward, F. I. (1987). *Climate and Plant Distribution*. Cabmridge, UK: Cambridge University Press.

Wu, Q., & Datla, V. V. (2011). On performance modeling and prediction in support of scientific workflow optimization. In *2011 IEEE World Congress on Services* IEEE. 161-168.

Yang, C., & Huang, Q. (2013). *Spatial Cloud Computing: A Practical Approach.* Boca Raton, Florida, USA: CRC Press.

Yang, C., Goodchild, M., Huang, Q., Nebert, D., Raskin, R., Xu, Y., Bambacus, M., et al. (2011a). Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing? *International Journal of Digital Earth*, *4*(4), 305–329.

Yang, C., Wu, H., Huang, Q., Li, Z., & Li, J. (2011b). Using spatial principles to optimize distributed computing for enabling the physical science discoveries. *Proceedings of the National Academy of Sciences*, *108*(14), 5498–5503.

Yee, T. W., & Mitchell, N. D. (1991). Generalized additive models in plant ecology. *Journal of Vegetation Science*, *2*(5), 587–602.

# Figures

Figure 1: This figure tracks the growth in collection size of community biodiversity databases through time, for the Global Biodiversity Information Facility (A, left) and the Neotoma Paleoecology Database (B, right). A record is a single data point denoting presence, absence, or abundance of a taxonomic group at a spatiotemporal location.

Figure 2: This figure shows the diversity of record types in the Global Biodiversity Information Facility (A, left) and the diversity of datasets in the Neotoma Paleoecology Database (B, right). The vertical axis in panel A shows millions of occurrence records, while the axis in panel B shows numbers of datasets, due to the different data models of the two databases. Due to the larger scope of the GBIF project, the categorization of database records (horizontal axis) is more general in A than in B.

Figure 3: Time series of the number of papers returned in a Web of Knowledge search for query “Species Distribution Model\*” (blue), compared with average annual citation growth in science and engineering, as reported by the U.S. National Science Board (2016) (black).

Figure 4: This figure presents the classification of 100 recent SDM-focused studies into data-driven, model-driven, and Bayesian model types. Candidate studies were identified using a query to the Web of Knowledge on (“Species Distribution Model\*” OR “Ecological Niche Model\*” OR Habitat Suitability Model\*”), from which a randomly chosen subset of 100 studies was used for scoring. Scoring was based on Hastie et al. (2009), Franklin (2009), and Elith et al. (2006) and studies were scored for their use of parametric/statistical methods, non-parametric/machine-learning models, or Bayesian predictive models. Many studies evaluated multiple SDM algorithms, yielding more than 100 scored scores. Studies that used unsupervised modeling routines were scored into a separate category, because they cannot be used in prediction tasks.

Figure 5: Plot of the Google Cloud engine custom virtual machine (VM) pricing scheme, showing the price surface faced by consumers of computing utilities as a function of the VM hardware configuration. Number of cores refers to the number of CPU cores dedicated to the VM. Memory refers to the amount of virtual memory dedicated to the VM, measured in gigabytes (GB).

Figure 6: This flowchart illustrates the four-step process of optimizing the data-hardware configuration for an SDM. First, the configuration of training examples and environmental covariates that maximizes model accuracy is selected from a set of candidate configurations. Second, using the runtime model, the execution time of that model run is predicted on a set of candidate hardware configurations, limited to VM instance types allowed on the Google Cloud Engine. Third, complete linkage hierarchical clustering was performed on three axes: run time, run cost, and standard deviation of the prediction posterior. The theoretical optimal hardware configuration would involve no time, no cost, and no uncertainty; therefore, the cluster whose multidimensional center is closest to the origin (0,0,0), is selected as the optimal set of hardware configurations.

Figure 7: Plots of observed SDM runtime against the predicted SDM runtime for four species distribution models (GBM-BRT, MARS, GAM, RF) and the holdout testing set of observed runtime. R2 values indicate the percent variance explained by each runtime model, as calculated by the correlation between predicted and observed values. The black line shows the y=x line expected for a perfectly-predictive model. All plots are displayed on logarithmic axes.

Figure 8: Plots of expected SDM accuracy, measured by the AUC statistic, against the AUC predicted by the accuracy model. Figure design follows that of Figure 7.

Figure 9: Plot of the optimal hardware configuration, based on the accuracy-maximizing experiments, for each SDM type. Note that the findings from the MARS models are peculiar, and should be interpreted with caution.

Figure 10: This contour plot shows for each SDM algorithm the accuracy substitution rate as a function of training dataset volume and number of covariates. The substitution rate shows the relative tradeoff between number of training examples and number of covariates required to maintain a given accuracy. Contour interval is 0.01 AUC. AUC values of 1 indicate a perfect goodness of fit, while an AUC of 0.5 indicate a fit between model and data no better than random. For GBM-RT, MARS, and GAM, AUC is insensitive to training dataset volumes above 2500 samples, while AUC continues to improve with increasing training set size for RF.

Figure 11: This figure shows the parallel efficiency of the RF SDM on datasets with different numbers of training examples. Parallel efficiency is calculated by dividing the ratio of the runtime on multiple cores to the runtime on a single core by the number of cores used. Efficiency can be interpreted as the marginal return gained by provisioning additional cores on which to run the algorithm. Higher efficiencies are obtained for larger training datasets, but efficiency decreases as the numbers of cores increase. The perfect efficiency line is shown, although this is impossible to achieve under real circumstances (Amdahl, 1967).

Figure 12: This figure shows the mean distance of each hardware configuration from the origin over the GBM-BRT, GAM, and RF SDMs, in which distance is calculated as Euclidean distance from the origin of time, cost, and prediction uncertainty. Each configuration’s distance to the origin was averaged over all SDM types to generate a mean configuration position. Green indicates lower distances and hardware configurations that are closer to optimal for these SDMs, while orange indicates larger distances and hardware configurations that are further from optimal. MARS was omitted from this analysis, due to its potentially misleading sampling and spurious dependence on memory. The large green area between two and four cores is likely strongly influenced by the lack of dependence of GAMs on memory. Note the large portion of the space far from the origin (orange), which suggests that many configurations are suboptimal for running SDMs.

# Tables

*Table 1:* Performance model evaluation statistics. Training denotes the number of SDM runtime observation used to fit the runtime model. Testing represents the number of data points that were reserved for evaluation of the model (approximately 20% of the total dataset). MSE is the mean squared prediction error of the runtime model. is the coefficient of determination between observed and predicted values, and is interpreted as fraction of explained variance. Posterior SD is the mean standard deviation of the prediction posteriors, measured in log-seconds.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SDM | Training | Testing | MSE  (log-seconds) |  | Posterior SD |
| GBM-BRT | 9256 | 2314 | 0.0646 | 0.9615 | 0.0257 |
| GAM | 2636 | 659 | 0.0121 | 0.5213 | 0.01069 |
| MARS | 6632 | 1657 | 0.0561 | 0.9648 | 0.03397 |
| RF | 2861 | 715 | 0.64824 | 0.5851 | 0.10174 |

Table 2: Accuracy Model Evaluation Statistics. Fields are as in Table 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SDM | Training | Testing | MSE  (AUC) |  | Posterior SD |
| GBM-BRT | 9256 | 2314 | 0.000245 | 0.8748 | 0.0012 |
| GAM | 2636 | 659 | 0.000718 | 0.8993 | 0.004998 |
| MARS | 6632 | 1657 | 0.000168 | 0.9418 | 0.001421 |
| RF | 2861 | 715 | 0.000034 | 0.9818 | 0.00062 |

Table 3: Controls on SDM runtime and accuracy, calculated as the reduction in explanatory power when a predictive factor is removed, for the execution time model (top) and accuracy model (bottom). All values are expressed as percentage change in *r2* values for reduced model relative to full model. Note that the MARS values are from the reduced model that fixes a sampling issue, and should therefore be interpreted with caution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Runtime Model** | RF | MARS | GBM-BRT | GAM |
| Number of Covariates | -0.03% | 3.07% | -1.67% | 0.25% |
| CPU Cores | -5.55% | 3.00% | -0.04% | -0.63% |
| GB Memory | -0.65% | -1.70% | -0.12% | 0.08% |
| Number of Training Examples | -38.36% | -34.90% | -71.74% | -1.22% |
| Number of Predictive Cells | 0.01% | -1.81% | -5.69% | -33.67% |
|  |  |  |  |  |
| **Accuracy Model** | RF | MARS | GBM-BRT | GAM |
| Number of Covariates | -34.39% | -26.31% | -16.40% | -50.45% |
| CPU Cores | -0.09% | 0.04% | -0.07% | 0.01% |
| GB Memory | -0.02% | -0.01% | -0.30% | 0.00% |
| Number of Training Examples | -45.72% | -26.36% | -67.84% | -0.34% |
| Number of Predictive Cells | 0.00% | 0.14% | -0.03% | 0.01% |

*Table 4:* Accuracy-maximizing points for each SDM, calculated during optimization. Fixed accuracy is the estimated accuracy given the corresponding number of training examples and covariates for that model. Training examples is the number of training examples that optimize SDM accuracy, during the unconstrained optimization procedure. Covariates is the optimized number of environmental covariates with which to fit the model.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Fixed Accuracy | Training Examples | Covariates |
| GAM | 0.7131 | 9000 | 5 |
| GBM-BRT | 0.8087 | 10000 | 5 |
| MARS | 0.7722 | 1000 | 5 |
| RF | 0.8523 | 10000 | 5 |

# Appendices

## Appendix A: Literature Meta-Analysis

### Table A1: Studies Evaluated in the Analysis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Authors | Title | Journal | Issue | Number | Pages | DOI | Year |
| Diniz-Filho, Jose Alexandre F; Rodrigues, Hauanny; Telles, Mariana Pires De Campos; De Oliveira, Guilherme; Terribile, Levi Carina; Soares, Thannya Nascimento; Nabout, Joao Carlos | Correlation between genetic diversity and environmental suitability: taking uncertainty from ecological niche models into account | Molecular Ecology Resources | 15 | 5 | 1059-1066 | [10.5061/dryad.3cp3t](http://dx.doi.org/10.5061/dryad.3cp3t" \t "_blank) | 2015 |
| Khoury, Colin K.; Castaneda-Alvarez, Nora P.; Achicanoy, Harold A.; Sosa, Chrystian C.; Bernau, Vivian; Kassa, Mulualem T.; Norton, Sally L.; van der Maesen, L. Jos G.; Upadhyaya, Hari D.; Ramirez-Villegas, Julian; Jarvis, Andy; Struik, Paul C. | Crop wild relatives of pigeonpea [Cajanus cajan (L.) Millsp.]: Distributions, ex situ conservation status, and potential genetic resources for abiotic stress tolerance | BIOLOGICAL CONSERVATION | 184 |  | 259-270 | 10.1016/j.biocon.2015.01.032 | 2015 |
| Chust, Guillem; Castellani, Claudia; Licandro, Priscilla; Ibaibarriaga, Leire; Sagarminaga, Yolanda; Irigoien, Xabier | Are Calanus spp. shifting poleward in the North Atlantic? A habitat modelling approach | ICES JOURNAL OF MARINE SCIENCE | 71 | 2 | 241-253 | 10.1093/icesjms/fst147 | 2014 |
| Davis, Edward Byrd; McGuire, Jenny L.; Orcutt, John D. | Ecological niche models of mammalian glacial refugia show consistent bias | ECOGRAPHY | 37 | 11 | 1133-1138 | 10.1111/ecog.01294 | 2014 |
| Yang, Xuejun; Huang, Zhenying; Venable, David L.; Wang, Lei; Zhang, Keliang; Baskin, Jerry M.; Baskin, Carol C.; Cornelissen, Johannes H. C. | Linking performance trait stability with species distribution: the case of Artemisia and its close relatives in northern China | JOURNAL OF VEGETATION SCIENCE | 27 | 1 | 123-132 | 10.1111/jvs.12334 | 2016 |
| Ananjeva, Natalia B.; Golynsky, Evgeny E.; Lin, Si-Min; Orlov, Nikolai L.; Tseng, Hui-Yun | MODELING HABITAT SUITABILITY TO PREDICT THE POTENTIAL DISTRIBUTION OF THE KELUNG CAT SNAKE Boiga kraepelini STEINEGER, 1902 | RUSSIAN JOURNAL OF HERPETOLOGY | 22 | 3 | 197-205 |  | 2015 |
| Chlond, Dominik; Bugaj-Nawrocka, Agnieszka | Model of potential distribution of Platymeris rhadamanthus Gerstaecker, 1873 with redescription of species. | Zoological Studies | 53 |  | 1-14 |  | 2014 |
| Miller, Matthew J.; Lipshutz, Sara E.; Smith, Neal G.; Bermingham, Eldredge | Genetic and phenotypic characterization of a hybrid zone between polyandrous Northern and Wattled Jacanas in Western Panama | BMC EVOLUTIONARY BIOLOGY | 14 |  |  | 10.1186/s12862-014-0227-7 | 2014 |
| Fernandez-Mazuecos, Mario; Jimenez-Mejias, Pedro; Rotllan-Puig, Xavier; Vargas, Pablo | Narrow endemics to Mediterranean islands: Moderate genetic diversity but narrow climatic niche of the ancient, critically endangered Naufraga (Apiaceae) | PERSPECTIVES IN PLANT ECOLOGY EVOLUTION AND SYSTEMATICS | 16 | 4 | 190-202 | 10.1016/j.ppees.2014.05.003 | 2014 |
| Tainio, Anna; Heikkinen, Risto K.; Heliola, Janne; Hunt, Alistair; Watkiss, Paul; Fronzek, Stefan; Leikola, Niko; Lotjonen, Sanna; Mashkina, Olga; Carter, Timothy R. | Conservation of grassland butterflies in Finland under a changing climate | REGIONAL ENVIRONMENTAL CHANGE | 16 | 1 | 71-84 | 10.1007/s10113-014-0684-y | 2016 |
| Oke, Oluwatobi A.; Thompson, Ken A. | Distribution models for mountain plant species: The value of elevation | ECOLOGICAL MODELLING | 301 |  | 72-77 | 10.1016/j.ecolmodel.2015.01.019 | 2015 |
| Tsuyama, Ikutaro; Nakao, Katsuhiro; Higa, Motoki; Matsui, Tetsuya; Shichi, Koji; Tanaka, Nobuyuki | What controls the distribution of the Japanese endemic hemlock, Tsuga diversifolia? Footprint of climate in the glacial period on current habitat occupancy | JOURNAL OF FOREST RESEARCH | 19 | 1 | 154-165 | 10.1007/s10310-013-0399-9 | 2014 |
| Yen, Shih-Ching; Wang, Ying; Ou, Heng-You | Habitat of the Vulnerable Formosan sambar deer Rusa unicolor swinhoii in Taiwan | ORYX | 48 | 2 | 232-240 | 10.1017/S0030605312001378 | 2014 |
| Hertzog, Lionel R.; Besnard, Aurelien; Jay-Robert, Pierre | Field validation shows bias-corrected pseudo-absence selection is the best method for predictive species-distribution modelling | DIVERSITY AND DISTRIBUTIONS | 20 | 12 | 1403-1413 | 10.1111/ddi.12249 | 2014 |
| Patrao, Claudia; Assis, Jorge; Rufino, Marta; Silva, Goncalo; Jordaens, Kurt; Backeljau, Thierry; Castilho, Rita | Habitat suitability modelling of four terrestrial slug species in the Iberian Peninsula (Arionidae: Geomalacus species) | JOURNAL OF MOLLUSCAN STUDIES | 81 |  | 427-434 | 10.1093/mollus/eyv018 | 2015 |
| Del Toro, Israel; Silva, Rogerio R.; Ellison, Aaron M. | Predicted impacts of climatic change on ant functional diversity and distributions in eastern North American forests | DIVERSITY AND DISTRIBUTIONS | 21 | 7 | 781-791 | 10.1111/ddi.12331 | 2015 |
| Voda, Raluca; Dapporto, Leonardo; Dinca, Vlad; Vila, Roger | Why Do Cryptic Species Tend Not to Co-Occur? A Case Study on Two Cryptic Pairs of Butterflies | PLOS ONE | 10 | 2 |  | 10.1371/journal.pone.0117802 | 2015 |
| Laube, Irina; Graham, Catherine H.; Boehning-Gaese, Katrin | Niche availability in space and time: migration in Sylvia warblers | JOURNAL OF BIOGEOGRAPHY | 42 | 10 | 1896-1906 | 10.1111/jbi.12565 | 2015 |
| Askeyev, Oleg; Askeyev, Igor; Askeyev, Arthur; Monakhov, Sergey; Yanybaev, Nur | River fish assemblages in relation to environmental factors in the eastern extremity of Europe (Tatarstan Republic, Russia) | ENVIRONMENTAL BIOLOGY OF FISHES | 98 | 5 | 1277-1293 | 10.1007/s10641-014-0358-0 | 2015 |
| Meseguer, Andrea S; Lobo, Jorge M; Ree, Richard; Beerling, David J; Sanmartin, Isabel | Data from: Integrating fossils, phylogenies, and niche models into biogeography to reveal ancient evolutionary history: the case of Hypericum (Hypericaceae) | Dryad |  |  |  | [10.5061/dryad.5845b](http://dx.doi.org/10.5061/dryad.5845b" \t "_blank) | 2014 |
| Soto-Centeno, J. Angel; O'Brien, Margaret; Simmons, Nancy B. | The importance of late Quaternary climate change and karst on distributions of Caribbean mormoopid bats | AMERICAN MUSEUM NOVITATES | 3847 |  |  |  | 2015 |
| Blanchard, Ryan; O'Farrell, Patrick J.; Richardson, David M. | Anticipating potential biodiversity conflicts for future biofuel crops in South Africa: incorporating spatial filters with species distribution models | GLOBAL CHANGE BIOLOGY BIOENERGY | 7 | 2 | 273-287 | 10.1111/gcbb.12129 | 2015 |
| Paudel, Prakash Kumar; Hais, Martin; Kindlmann, Pavel | Habitat suitability models of mountain ungulates: identifying potential areas for conservation | ZOOLOGICAL STUDIES | 54 |  |  | 10.1186/s40555-015-0116-9 | 2015 |
| Hsu, Chorng-Bin; Hwang, Gwo-Wen; Lu, Jane-Fuh; Chen, Chang-Po; Tao, Hsiao-Hang; Hsieh, Hwey-Lian | Habitat Characteristics of the Wintering Common Teal in the Huajiang Wetland, Taiwan | WETLANDS | 34 | 6 | 1207-1218 | 10.1007/s13157-014-0581-7 | 2014 |
| Henrys, P. A.; Bee, E. J.; Watkins, J. W.; Smith, N. A.; Griffiths, R. I. | Mapping natural capital: optimising the use of national scale datasets | ECOGRAPHY | 38 | 6 | 632-638 | 10.1111/ecog.00402 | 2015 |
| Becker, Nina I.; Encarnacao, Jorge A. | Silvicolous on a Small Scale: Possibilities and Limitations of Habitat Suitability Models for Small, Elusive Mammals in Conservation Management and Landscape Planning | PLOS ONE | 10 | 3 |  | 10.1371/journal.pone.0120562 | 2015 |
| Drake, John M | Range bagging: a new method for ecological niche modelling from presence-only data. | Journal of the Royal Society, Interface / the Royal Society | 12 | 107 |  | 10.1098/rsif.2015.0086 | 2015 |
| Masin, Simone; Bonardi, Anna; Padoa-Schioppa, Emilio; Bottoni, Luciana; Ficetola, Gentile Francesco | Risk of invasion by frequently traded freshwater turtles | BIOLOGICAL INVASIONS | 16 | 1 | 217-231 | 10.1007/s10530-013-0515-y | 2014 |
| Ficetola, Gentile Francesco; Cagnetta, Massimo; Padoa-Schioppa, Emilio; Quas, Anita; Razzetti, Edoardo; Sindaco, Roberto; Bonardi, Anna | Sampling bias inverts ecogeographical relationships in island reptiles | GLOBAL ECOLOGY AND BIOGEOGRAPHY | 23 | 11 | 1303-1313 | 10.1111/geb.12201 | 2014 |
| Segurado, Pedro; Branco, Paulo; Avelar, Ana P.; Ferreira, Maria T. | Historical changes in the functional connectivity of rivers based on spatial network analysis and the past occurrences of diadromous species in Portugal | AQUATIC SCIENCES | 77 | 3 | 427-440 | 10.1007/s00027-014-0371-6 | 2015 |
| Zhao, C. S.; Yang, S. T.; Liu, C. M.; Dou, T. W.; Yang, Z. L.; Yang, Z. Y.; Liu, X. L.; Xiang, H.; Nie, S. Y.; Zhang, J. L.; Mitrovic, S. M.; Yu, Q.; Lim, R. P. | Linking hydrologic, physical and chemical habitat environments for the potential assessment of fish community rehabilitation in a developing city | JOURNAL OF HYDROLOGY | 523 |  | 384-397 | 10.1016/j.jhydrol.2015.01.067 | 2015 |
| Roger, Erin; Duursma, Daisy Englert; Downey, Paul O.; Gallagher, Rachael V.; Hughes, Lesley; Steel, Jackie; Johnson, Stephen B.; Leishman, Michelle R. | A tool to assess potential for alien plant establishment and expansion under climate change | JOURNAL OF ENVIRONMENTAL MANAGEMENT | 159 |  | 121-127 | 10.1016/j.jenvman.2015.05.039 | 2015 |
| Poulos, Davina E.; Gallen, Christopher; Davis, Tom; Booth, David J.; Harasti, David | Distribution and spatial modelling of a soft coral habitat in the Port Stephens-Great Lakes Marine Park: implications for management | MARINE AND FRESHWATER RESEARCH | 67 | 2 | 256-265 | 10.1071/MF14059 | 2016 |
| Lira-Noriega, Andres; Peterson, A. Townsend | Range-wide ecological niche comparisons of parasite, hosts and dispersers in a vector-borne plant parasite system | JOURNAL OF BIOGEOGRAPHY | 41 | 9 | 1664-1673 | 10.1111/jbi.12302 | 2014 |
| Olsson, Ola; Bolin, Arvid | A model for habitat selection and species distribution derived from central place foraging theory | OECOLOGIA | 175 | 2 | 537-548 | 10.1007/s00442-014-2931-9 | 2014 |
| Lee, Chang-Bae; Yun, Soon Jin; œÌÈTÖ | Effects of climate change on the geographic distribution of Quercus acuta Thunb. | Journal of Agriculture & Life Science | 49 | 6 | 47-56 | 10.14397/jals.2015.49.6.47 | 2015 |
| Skov, Henrik; Heinanen, Stefan; Thaxter, Chris B.; Williams, Adrian E.; Lohier, Sabine; Banks, Alex N. | Real-time species distribution models for conservation and management of natural resources in marine environments | MARINE ECOLOGY PROGRESS SERIES | 542 |  | 221-234 | 10.3354/meps11572 | 2016 |
| Afkhami, Michelle E.; McIntyre, Patrick J.; Strauss, Sharon Y. | Mutualist-mediated effects on species' range limits across large geographic scales | ECOLOGY LETTERS | 17 | 10 | 1265-1273 | 10.1111/ele.12332 | 2014 |
| Mimet, Anne; Maurel, Noelie; Pellissier, Vincent; Simon, Laurent; Julliard, Romain | Towards a unique landscape description for multi-species studies: A model comparison with common birds in a human-dominated French region | ECOLOGICAL INDICATORS | 36 |  | 19-32 | 10.1016/j.ecolind.2013.06.029 | 2014 |
| Owens, Hannah L. | Evolution of codfishes (Teleostei: Gadinae) in geographical and ecological space: evidence that physiological limits drove diversification of subarctic fishes | JOURNAL OF BIOGEOGRAPHY | 42 | 6 | 1091-1102 | 10.1111/jbi.12483 | 2015 |
| Rato, Catarina; Harris, David James; Perera, Ana; Carvalho, Silvia B.; Carretero, Miguel A.; Roedder, Dennis | A Combination of Divergence and Conservatism in the Niche Evolution of the Moorish Gecko, Tarentola mauritanica (Gekkota: Phyllodactylidae) | PLOS ONE | 10 | 5 | - | 10.1371/journal.pone.0127980 | 2015 |
| Munoz, Antonio-Roman; Jimenez-Valverde, Alberto; Luz Marquez, Ana; Moleon, Marcos; Real, Raimundo | Environmental favourability as a cost-efficient tool to estimate carrying capacity | DIVERSITY AND DISTRIBUTIONS | 21 | 12 | 1388-1400 | 10.1111/ddi.12352 | 2015 |
| Ortego, Joaquin; Gugger, Paul F.; Sork, Victoria L. | Climatically stable landscapes predict patterns of genetic structure and admixture in the Californian canyon live oak | JOURNAL OF BIOGEOGRAPHY | 42 | 2 | 328-338 | 10.1111/jbi.12419 | 2015 |
| Latif, Quresh S.; Saab, Victoria A.; Mellen-Mclean, Kim; Dudley, Jonathan G. | Evaluating Habitat Suitability Models for Nesting White-Headed Woodpeckers in Unburned Forest | JOURNAL OF WILDLIFE MANAGEMENT | 79 | 2 | 263-273 | 10.1002/jwmg.842 | 2015 |
| Branco, Paulo; Segurado, Pedro; Santos, Jose M.; Ferreira, Maria T. | Prioritizing barrier removal to improve functional connectivity of rivers | JOURNAL OF APPLIED ECOLOGY | 51 | 5 | 1197-1206 | 10.1111/1365-2664.12317 | 2014 |
| Callen, Steven T.; Miller, Allison J. | Signatures of niche conservatism and niche shift in the North American kudzu (Pueraria montana) invasion | DIVERSITY AND DISTRIBUTIONS | 21 | 8 | 853-863 | 10.1111/ddi.12341 | 2015 |
| Alberdi, Antton; Gilbert, M. Thomas P.; Razgour, Orly; Aizpurua, Ostaizka; Aihartza, Joxerra; Garin, Inazio | Contrasting population-level responses to Pleistocene climatic oscillations in analpine bat revealed by complete mitochondrial genomes and evolutionary history inference | JOURNAL OF BIOGEOGRAPHY | 42 | 9 | 1689-1700 | 10.1111/jbi.12535 | 2015 |
| Rezende, Vanessa Leite; de Oliveira-Filho, Ary T.; Eisenlohr, Pedro V.; Yoshino Kamino, Luciana Hiromi; Vibrans, Alexander Christian | Restricted geographic distribution of tree species calls for urgent conservation efforts in the Subtropical Atlantic Forest | BIODIVERSITY AND CONSERVATION | 24 | 5 | 1057-1071 | 10.1007/s10531-014-0721-7 | 2015 |
| Fresia, Pablo; Silver, Micha; Mastrangelo, Thiago; De Azeredo-Espin, Ana Maria L.; Lyra, Mariana L. | Applying spatial analysis of genetic and environmental data to predict connection corridors to the New World screwworm populations in South America | ACTA TROPICA | 138 |  | S34-S41 | 10.1016/j.actatropica.2014.04.003 | 2014 |
| Mod, Heidi K.; le Roux, Peter C.; Guisan, Antoine; Luoto, Miska | Biotic interactions boost spatial models of species richness | ECOGRAPHY | 38 | 9 | 913-921 | 10.1111/ecog.01129 | 2015 |
| le Roux, Peter C.; Luoto, Miska | Earth surface processes drive the richness, composition and occurrence of plant species in an arctic-alpine environment | JOURNAL OF VEGETATION SCIENCE | 25 | 1 | 45-54 | 10.1111/jvs.12059 | 2014 |
| Razgour, Orly; Salicini, Irene; Ibanez, Carlos; Randi, Ettore; Juste, Javier | Unravelling the evolutionary history and future prospects of endemic species restricted to former glacial refugia | MOLECULAR ECOLOGY | 24 | 20 | 5267-5283 | 10.1111/mec.13379 | 2015 |
| Belmaker, Jonathan; Zarnetske, Phoebe; Tuanmu, Mao-Ning; Zonneveld, Sara; Record, Sydne; Strecker, Angela; Beaudrot, Lydia | Empirical evidence for the scale dependence of biotic interactions | GLOBAL ECOLOGY AND BIOGEOGRAPHY | 24 | 7 | 750-761 | 10.1111/geb.12311 | 2015 |
| Wrege, Marcos Silveira; Coutinho, Enilton Fick; Pantano, Angelica Prela; Jorge, Rogerio Oliveira | POTENCIAL DISTRIBUTION OF OLIVE IN BRAZIL AND WORLDWIDE | REVISTA BRASILEIRA DE FRUTICULTURA | 37 | 3 | 656-666 | 10.1590/0100-2945-174/14 | 2015 |
| Dong, Xiaoli; Grimm, Nancy B.; Ogle, Kiona; Franklin, Janet | Temporal variability in hydrology modifies the influence of geomorphology on wetland distribution along a desert stream | JOURNAL OF ECOLOGY | 104 | 1 | 18-30 | 10.1111/1365-2745.12450 | 2016 |
| Liu, Canran; Newell, Graeme; White, Matt | On the selection of thresholds for predicting species occurrence with presence-only data | ECOLOGY AND EVOLUTION | 6 | 1 | 337-348 | 10.1002/ece3.1878 | 2016 |
| Recio, Mariano R.; Seddon, Philip J.; Moore, Antoni B. | Niche and movement models identify corridors of introduced feral cats infringing ecologically sensitive areas in New Zealand | BIOLOGICAL CONSERVATION | 192 |  | 48-56 | 10.1016/j.biocon.2015.09.004 | 2015 |
| Tocchio, Luana J.; Gurgel-Goncalves, Rodrigo; Escobar, Luis E.; Peterson, Andrew Townsend | Niche similarities among white-eared opossums (Mammalia, Didelphidae): Is ecological niche modelling relevant to setting species limits? | ZOOLOGICA SCRIPTA | 44 | 1 | 1-10 | 10.1111/zsc.12082 | 2015 |
| Guo, Chuanbo; Lek, Sovan; Ye, Shaowen; Li, Wei; Liu, Jiashou; Li, Zhongjie | Uncertainty in ensemble modelling of large-scale species distribution: Effects from species characteristics and model techniques | ECOLOGICAL MODELLING | 306 |  | 67-75 | 10.1016/j.ecolmodel.2014.08.002 | 2015 |
| Yin, Hengxia; Yan, Xia; Shi, Yong; Qian, Chaoju; Li, Zhonghu; Zhang, Wen; Wang, Lirong; Li, Yi; Li, Xiaoze; Chen, Guoxiong; Li, Xinrong; Nevo, Eviatar; Ma, Xiao-Fei | The role of East Asian monsoon system in shaping population divergence and dynamics of a constructive desert shrub Reaumuria soongarica | SCIENTIFIC REPORTS | 5 |  |  | 10.1038/srep15823 | 2015 |
| Huang, Jen-Pan | Modeling the effects of anthropogenic exploitation and climate change on an endemic stag beetle, Lucanus miwai, of Taiwan | Dryad |  |  |  | [10.5061/dryad.8mp13](http://dx.doi.org/10.5061/dryad.8mp13" \t "_blank) | 2014 |
| Planas, E.; Saupe, E. E.; Lima-Ribeiro, M. S.; Peterson, A. T.; Ribera, C. | Ecological niche and phylogeography elucidate complex biogeographic patterns in Loxosceles rufescens (Araneae, Sicariidae) in the Mediterranean Basin. | BMC Evolutionary Biology | 14 | 195 |  |  | 2014 |
| Naczk, Aleksandra M; Kolanowska, Marta | Glacial Refugia and Future Habitat Coverage of Selected Dactylorhiza Representatives (Orchidaceae). | PloS one | 10 | 11 | e0143478-e0143478 | 10.1371/journal.pone.0143478 | 2015 |
| Maher, Sean P.; Randin, Christophe F.; Guisan, Antoine; Drake, John M. | Pattern-recognition ecological niche models fit to presence-only and presence-absence data | METHODS IN ECOLOGY AND EVOLUTION | 5 | 8 | 761-770 | 10.1111/2041-210X.12222 | 2014 |
| Carolan, Kevin; Ebong, Solange Meyin A.; Garchitorena, Andres; Landier, Jordi; Sanhueza, Daniel; Texier, Gaetan; Marsollier, Laurent; Le Gall, Philipe; Guegan, Jean-Francois; Lo Seen, Danny | Ecological niche modelling of Hemipteran insects in Cameroon; the paradox of a vector-borne transmission for Mycobacterium ulcerans, the causative agent of Buruli ulcer | INTERNATIONAL JOURNAL OF HEALTH GEOGRAPHICS | 13 |  |  | 10.1186/1476-072X-13-44 | 2014 |
| Moraes R, Monica; Rios-Uzeda, Boris; Rene Moreno, Luis; Huanca-Huarachi, Gladis; Larrea-Alcazar, Daniel | Using potential distribution models for patterns of species richness, endemism, and phytogeography of palm species in Bolivia | TROPICAL CONSERVATION SCIENCE | 7 | 1 | 45-60 |  | 2014 |
| Lim, Haw Chuan; Zou, Fasheng; Sheldon, Frederick H. | Genetic differentiation in two widespread, open-forest bird species of Southeast Asia (Copsychus saularis and Megalaima haemacephala): Insights from ecological niche modeling | Current Zoology | 61 | 5 | 922-934 |  | 2015 |
| Caouette, J. P.; Steel, E. A.; Hennon, P. E.; Cunningham, P. G.; Pohl, C. A.; Schrader, B. A. | Influence of elevation and site productivity on conifer distributions across Alaskan temperate rainforests | CANADIAN JOURNAL OF FOREST RESEARCH | 46 | 2 | 249-261 |  | 2016 |
| Manuel Alvarez-Martinez, Jose; Suarez-Seoane, Susana; Stoorvogel, Jetse J.; de Luis Calabuig, Estanislao | Influence of land use and climate on recent forest expansion: a case study in the Eurosiberian-Mediterranean limit of north-west Spain | JOURNAL OF ECOLOGY | 102 | 4 | 905-919 | 10.1111/1365-2745.12257 | 2014 |
| Pabijan, Maciej; Brown, Jason L.; Chan, Lauren M.; Rakotondravony, Hery A.; Raselimanana, Achille P.; Yoder, Anne D.; Glaw, Frank; Vences, Miguel | Phylogeography of the arid-adapted Malagasy bullfrog, Laliostoma labrosum, influenced by past connectivity and habitat stability | MOLECULAR PHYLOGENETICS AND EVOLUTION | 92 |  | 11-24 | 10.1016/j.ympev.2015.05.018 | 2015 |
| Obolenskaya, Ekaterina V.; Lissovsky, Andrey A. | Regional zoogeographical zoning using species distribution modelling by the example of small mammals of South-Eastern Transbaikalia | Russian Journal of Theriology | 14 | 2 | 171-185 |  | 2015 |
| Zeng, Qing; Zhang, Yamian; Sun, Gongqi; Duo, Hairui; Wen, Li; Lei, Guangchun | Using Species Distribution Model to Estimate the Wintering Population Size of the Endangered Scaly-Sided Merganser in China | PLOS ONE | 10 | 2 |  | 10.1371/journal.pone.0117307 | 2015 |
| de Castro Pena, Joao Carlos; Yoshino Kamino, Luciana Hiromi; Rodrigues, Marcos; Mariano-Neto, Eduardo; de Siqueira, Marinez Ferreira | Assessing the conservation status of species with limited available data and disjunct distribution | BIOLOGICAL CONSERVATION | 170 |  | 130-136 | 10.1016/j.biocon.2013.12.015 | 2014 |
| Collevatti, Rosane G.; Terribile, Levi C.; Rabelo, Suelen G.; Lima-Ribeiro, Matheus S. | Relaxed random walk model coupled with ecological niche modeling unravel the dispersal dynamics of a Neotropical savanna tree species in the deeper Quaternary | FRONTIERS IN PLANT SCIENCE | 6 |  |  | 10.3389/fpls.2015.00653 | 2015 |
| Wang, Lifei; Jackson, Donald A. | Shaping up model transferability and generality of species distribution modeling for predicting invasions: implications from a study on Bythotrephes longimanus | BIOLOGICAL INVASIONS | 16 | 10 | 2079-2103 | 10.1007/s10530-014-0649-6 | 2014 |
| Paiva, Vitor H.; Geraldes, Pedro; Rodrigues, Isabel; Melo, Tommy; Melo, Jose; Ramos, Jaime A. | The Foraging Ecology of the Endangered Cape Verde Shearwater, a Sentinel Species for Marine Conservation off West Africa | PLOS ONE | 10 | 10 |  | 10.1371/journal.pone.0139390 | 2015 |
| Aguirre-Gutierrez, Jesus; Serna-Chavez, Hector M.; Villalobos-Arambula, Alma R.; Perez de la Rosa, Jorge A.; Raes, Niels | Similar but not equivalent: ecological niche comparison across closely-related Mexican white pines | DIVERSITY AND DISTRIBUTIONS | 21 | 3 | 245-257 | 10.1111/ddi.12268 | 2015 |
| Tererai, Farai; Wood, Alan R. | On the present and potential distribution of Ageratina adenophora (Asteraceae) in South Africa | SOUTH AFRICAN JOURNAL OF BOTANY | 95 |  | 152-158 | 10.1016/j.sajb.2014.09.001 | 2014 |
| Costa, Henrique C.; de Rezende, Daniella T.; Molina, Flavio B.; Nascimento, Luciana B.; Leite, Felipe S. F.; Fernandes, Ana Paula B. | New Distribution Records and Potentially Suitable Areas for the Threatened Snake-Necked Turtle Hydromedusa maximiliani (Testudines: Chelidae) | CHELONIAN CONSERVATION AND BIOLOGY | 14 | 1 | 88-94 |  | 2015 |
| Costa, J.; Dornak, L. L.; Almeida, C. E.; Peterson, A. T. | Distributional potential of the Triatoma brasiliensis species complex at present and under scenarios of future climate conditions. | Parasites and Vectors | 7 | 238 |  |  | 2014 |
| Saupe, Erin E.; Hendricks, Jonathan R.; Peterson, A. Townsend; Lieberman, Bruce S. | Climate change and marine molluscs of the western North Atlantic: future prospects and perils | JOURNAL OF BIOGEOGRAPHY | 41 | 7 | 1352-1366 | 10.1111/jbi.12289 | 2014 |
| Alves, Joana; da Silva, Antonio Alves; Soares, Amadeu M. V. M.; Fonseca, Carlos | Spatial and temporal habitat use and selection by red deer: The use of direct and indirect methods | MAMMALIAN BIOLOGY | 79 | 5 | 338-348 | 10.1016/j.mambio.2014.05.007 | 2014 |
| Qi, Xin-Shuai; Yuan, Na; Comes, Hans Peter; Sakaguchi, Shota; Qiu, Ying-Xiong | A strong "filter" effect of the East China Sea land bridge for East Asia's temperate plant species: inferences from molecular phylogeography and ecological niche modelling of Platycrater arguta (Hydrangeaceae) | TreeBASE |  |  |  |  | 2014 |
| Alexander, Neil; Medlock, Jolyon; Morley, David; Wint, Willy | A first attempt at modelling red deer (Cervus elaphus) distributions over Europe | Figshare | 1 |  |  | [10.6084/m9.figshare.1008334](http://dx.doi.org/10.6084/m9.figshare.1008334" \t "_blank) | 2014 |
| de Andrade, Andrey Jose; Gurgel-Goncalves, Rodrigo | New record and update on the geographical distribution of Pintomyia monticola (Costa Lima, 1932) (Diptera: Psychodidae) in South America. | Check List | 11 | 2 | 1566-1566 |  | 2015 |
| van Andel, T. R.; Croft, S.; van Loon, E. E.; Quiroz, D.; Towns, A. M.; Raes, N. | Prioritizing West African medicinal plants for conservation and sustainable extraction studies based on market surveys and species distribution models | BIOLOGICAL CONSERVATION | 181 |  | 173-181 | 10.1016/j.biocon.2014.11.015 | 2015 |
| Burger, J.; Edler, B.; Gerowitt, B.; Steinmann, H. H. | Predicting weed problems in maize cropping by species distribution modelling. | Julius-Kuhn-Archiv | 443 | 379 | 386- |  | 2014 |
| Sanchez-Montes, Sokani; Espinosa-Martinez, Deborah V.; Rios-Munoz, Cesar A.; Berzunza-Cruz, Miriam; Becker, Ingeborg | Leptospirosis in Mexico: Epidemiology and Potential Distribution of Human Cases | PLOS ONE | 10 | 7 |  | 10.1371/journal.pone.0133720 | 2015 |
| Williams, K. A.; Richards, C. S.; Villet, M. H. | Predicting the geographic distribution of Lucilia sericata and Lucilia cuprina (Diptera: Calliphoridae) in South Africa | AFRICAN INVERTEBRATES | 55 | 1 | 157-170 |  | 2014 |
| Wang, Hsiao-Hsuan; Koralewski, Tomasz E.; McGrew, Erin K.; Grant, William E.; Byram, Thomas D. | Species Distribution Model for Management of an Invasive Vine in Forestlands of Eastern Texas | FORESTS | 6 | 12 | 4374-4390 | 10.3390/f6124374 | 2015 |
| Pouteau, Robin; Meyer, Jean-Yves; Larrue, Sebastien | Using range filling rather than prevalence of invasive plant species for management prioritisation: the case of Spathodea campanulata in the Society Islands (South Pacific) | ECOLOGICAL INDICATORS | 54 |  | 87-95 | 10.1016/j.ecolind.2015.02.017 | 2015 |
| Carvalho, Bruno M.; Rangel, Elizabeth F.; Ready, Paul D.; Vale, Mariana M. | Ecological Niche Modelling Predicts Southward Expansion of Lutzomyia (Nyssomyia) flaviscutellata (Diptera: Psychodidae: Phlebotominae), Vector of Leishmania (Leishmania) amazonensis in South America, under Climate Change | PLOS ONE | 10 | 11 |  | 10.1371/journal.pone.0143282 | 2015 |
| Almalki, M.; Alrashidi, M.; O'Connell, M. J.; Shobrak, M.; Szekely, T. | MODELLING THE DISTRIBUTION OF WETLAND BIRDS ON THE RED SEA COAST IN THE KINGDOM OF SAUDI ARABIA | APPLIED ECOLOGY AND ENVIRONMENTAL RESEARCH | 13 | 1 | 67-84 |  | 2015 |
| Duff, Thomas J.; Bell, Tina L.; York, Alan | Recognising fuzzy vegetation pattern: the spatial prediction of floristically defined fuzzy communities using species distribution modelling methods | JOURNAL OF VEGETATION SCIENCE | 25 | 2 | 323-337 | 10.1111/jvs.12092 | 2014 |
| Fossog, Billy Tene; Ayala, Diego; Acevedo, Pelayo; Kengne, Pierre; Mebuy, Ignacio Ngomo Abeso; Makanga, Boris; Magnus, Julie; Awono-Ambene, Parfait; Njiokou, Flobert; Pombi, Marco; Antonio-Nkondjio, Christophe; Paupy, Christophe; Besansky, Nora J.; Costantini, Carlo | Habitat segregation and ecological character displacement in cryptic African malaria mosquitoes | EVOLUTIONARY APPLICATIONS | 8 | 4 | 326-345 | 10.1111/eva.12242 | 2015 |
| Razgour, Orly | Beyond species distribution modeling: A landscape genetics approach to investigating range shifts under future climate change | ECOLOGICAL INFORMATICS | 30 |  | 250-256 | 10.1016/j.ecoinf.2015.05.007 | 2015 |

## Appendix B: Data Collection Protocol

### Figure B1: Conceptual Flowchart of Distributed System Used for Automated SDM

## Appendix C: Bayesian Model Priors

Model Name: Bayesian Additive Regression Trees

Implementation: bartMachine version 1.2.3

Adam Kapelner, Justin Bleich (2016). bartMachine: Machine Learning with Bayesian Additive Regression Trees. Journal of Statistical Software, 70(4), 1-40. doi: 10.18637/jss.v070.i04

## Model Structure:

## Screen Shot 2016-12-09 at 8.01.37 PM.pngScreen Shot 2016-12-09 at 7.51.37 PM.png

Model Priors:

1. Node Depth Prior: P(Tt) ~ α(1+d)−β where α ∈ (0, 1) and β ∈ [0, ∞]
2. Leaf-Value Prior: P(Mt | Tt) = μl ~ *N*(μμ / m, σμ2 )

* is picked to be the range center,
* σμ2 is empirically chosen so that the range center plus or minus k =2 variances cover 95% of the provided response values in the training set

1. Error Variance Prior: σ2 ∼ InvGamma(ν/2, νλ/2)

* λ is determined from the data so that there is a q = 90% a priori chance (by default) that the BART model will improve upon the RMSE from an ordinary least squares regression.

1. Response likelihood: mean of response in leaf in given MCMC iteration with variance:
2. Default hyperparameters:
   * α: 0.95
   * β: 2
   * k: 2
   * ν: 3
   * q: 90%

1. As of March, 2017 [↑](#footnote-ref-1)
2. As of November, 2016. [↑](#footnote-ref-2)
3. The GBM-BRT code in the Ridgeway (2015) gbm package underlies the implementation of boosted regression trees in the dismo package, common in SDM applications (Hijmans et al., 2016). A review of the gbm code suggests that model validation can be done on multiple processing cores, but model fitting is done sequentially. [↑](#footnote-ref-3)