6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

PREDICTING EXECUTION TIME OF CLIMATE-DRIVEN ECOLOGICAL FORECASTING MODELS

Scott Farley[1] and John W. Williams[1,2]

***Abstract*—Species distribution models are climate-driven ecological forecasting tools that are widely used to predict species range shifts and ecological responses to 21st century climate change. As modern and fossil biodiversity databases improve and statistical methods become more computationally intensive, choosing the correct computing configuration on which to run these models becomes more important. We present a predictive model for estimating species distribution model execution time based on algorithm inputs and computing hardware. The model shows considerable predictive skill and can inform future resource provisioning strategies. We also demonstrate a technique for predicting model accuracy that suggests that inclusion of training data from the fossil record can enhance the accuracy of distribution models.**

I. INTRODUCTION

21st century climate change is expected to significantly alter species distributions, both at global and regional scales. Species Distribution Models (SDMs) are statistical methods that estimate species-specific responses to climatic gradients, and are widely used to to predict species presence under future climate scenarios [1]. While the models are widespread in the literature, a thorough understanding of algorithm execution time and accuracy produced by different input datasets and on different computing hardware has not yet been established. Here we discuss models for predicting the accuracy and run time of three SDMs given these factors. Execution time and accuracy models can improve computing resource utilization and identify performance bottlenecks in popular SDM code repositories.

SDMs can be fit with both modern- and paleo-climate training data, the scale, size, and resolution of which has increased rapidly over the last several years. Emerging databases, such as the Neotoma Paleoecological Database (http://neotomadb.org) and the Global Biodiversity Information Facility (GBIF, http://gbif.org), provide biogeographical data for millions of species worldwide, both in the recent fossil record and for the modern era. Environmental covariates to species presences are obtained from widely-available climate model output, which can provide decadal or sub-decadal temporal resolution for the last 21,000 years. Downscaling techniques can improve the spatial resolution of gridded data to scales suitable for regional and sub-regional study.

While the size and resolution of climatic and biodiversity data used to train SDMs increase, the methods used to learn species' responses to climatic gradients are becoming more computationally expensive. Most competitive SDMs use statistical learning procedures to estimate the functional relationship between species presence and climatic patterns. Novel techniques, such as Bayesian learning have also demonstrated high accuracy in this setting [2]. Moreover, many researchers now model hundreds of species in a single study (e.g., [3]), or use joint modeling techniques to capture inter-species interactions [4], resulting in larger modeling workloads. More powerful computing hardware has the potential to reduce the execution time of SDMs, particularly those with high dimensionality, large training sets, and/or wide spatiotemporal extents. While work has been done to assess the characteristic complexity of machine learning models (e.g. big-O notation) [5], less has been done to characterize the differences in model execution time of SDM techniques due to different computing hardware configurations and algorithm inputs. Though internal variations in memory management make it is difficult to exactly define model runtime as a function of hardware [6], models of computer performance that consider input data and static hardware configuration may be capable of capturing high-level trends [7], [8].

Here we model algorithm speed using two static

Corresponding author: S Farley, sfarley2@wisc.edu [1]University of Wisconsin-Madison, Department of Geography, Madison, WI 53706 [2]University of Wisconsin-Madison, Center for Climatic Research, Madison, WI 53706

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

hardware components capable of improving performance: (1) main memory size (i.e., RAM) and (2) the number computing cores, and two algorithm inputs: (1) the size of the training data used to fit the model and (2) the spatial resolution of the output. While different learning techniques may have implementation- or algorithm-specific differences (i.e., tuning parameters, language differences) that may influence model execution time, we test several popular *R* implementations with experimental variables that extend across model classes.

We also examine the predictive accuracy gains made by fitting SDMs with different training data sizes. Recent studies have examined the best practices for using small numbers of training examples ($n < 300$), such as for rare species [9]. However, while very large training datasets ($n > 100,000$) are unlikely in the ecological domain, the fossil record can be used to fit SDMs with a larger set of training data than the modern era alone [10], perhaps by several times for some species. While there is a greater degree of uncertainty associated with fossil occurrences, their utilization may significantly enhance SDM skill when included in the fitting process.

II. METHOD

We systematically tested the accuracy and execution time of three popular species distribution modeling algorithms on four different training set sizes and four spatial resolutions on 44 computing configurations (4 x CPU, 11 x RAM). All experiments were done using the R programming language on virtual machines hosted on the Google Cloud Compute platform. Ten replicates of each combination of hardware and algorithm inputs were completed to improve understanding system-induced variance.

Fossil occurrences for the *Picea* (spruce) genus over the last 21,000 years were obtained from the Neotoma Paleoecological Database (http://neotomadb.org). Decadally-averaged climatic covariates for each fossil occurrence were extracted from 0.5 degree spatial resolution debiased and downscaled CCSM3 climate model output for North America [11], and used to fit the SDMs.

Three SDM algorithms that have shown competitive predictive skill in the literature were evaluated: (a) boosted regression trees (GBM-BRT) [12], (b) multivariate adaptive regression splines (MARS) [13], (c) generalized additive models (GAM) [14]. All models were fit using a randomized training data subset of a pre-specified size, and then projected onto a climatic grid for the year 2100. The output grid resolution was

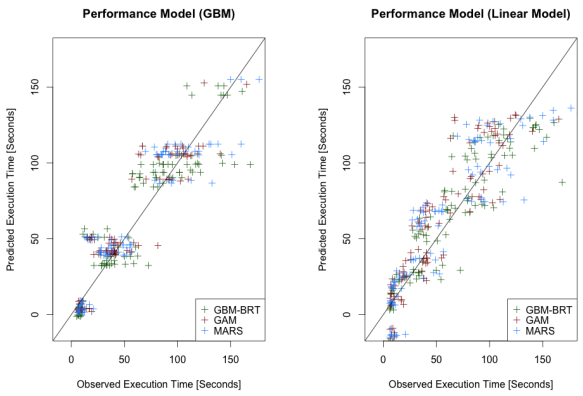

Fig. 1. Gradient Boosted Regression Tree (left) and Linear (Right) models of three different SDM modeling algorithms execution time under different algorithm and hardware parameters.

varied between 0.1 and 1 degree of latitude. SDM accuracy skill was evaluated using an independent testing set of 20% of the available data using the area under the receiver operator curve (AUC) statistic.

To predict model execution time, two predictive models were built for each SDM technique: (a) a multiple linear regression and (b) a gradient boosted regression tree model. Models were fit from the set of experiments ($n = 20,583$) using the number of training examples, CPU cores, memory, and spatial resolution as predictors of execution time. Model were evaluated using ANOVA and partial dependence plots, and skill was estimated using observed-to-predicted correlation metrics. Predictive models of SDM AUC score were also developed and fit using a boosted regression tree approach.

III. EVALUATION

Predictive models of SDM execution time demonstrate considerable skill when evaluated against a hold-out testing set of observed values. In general, the boosted regression tree model approach significantly outperformed the linear models. Regression trees are able better capture the potential non-linearities of the experimental dataset and can remove the negative predictions forecast by the linear model. However, both sets of models consistently showed $r^2 > 0.8$ correlation between observed and predicted values with a mean prediction error of less than 4 seconds.

Of all six execution time models (2 models x 3 SDMs), the regression tree prediction model of the MARS SDM performed the best, with a mean error of -0.457 ± 1.895 seconds and an r^2 value of 0.936. The regression tree models for GAM and GBM-BRT

SDMs both performed well with r^2 values of 0.892 and 0.880, respectively. The linear models all showed lower r^2 correlation values and had larger prediction variance and mean prediction errors than their decision tree counterparts. The best performing linear model was again for the MARS SDM, with an r^2 correlation of 0.876, with a significantly larger mean error of 2.17 ± 1.73 seconds. Figure 1 shows the observed and predicted values for each SDM for both of the prediction models.

Model interrogation using ANOVA (linear model) and partial dependency plots (GBM model) reveals that model execution time depends strongly on the number of training examples used to fit the SDM. In all cases, the number of training examples and spatial resolution of the output were shown to be highly significant ($p < 0.001$). Computer hardware variables were not shown to be significant predictors of execution time for these SDMs. In some cases, additional memory was shown to reduce model speed, perhaps due to increased overhead of memory management. Runtime logs indicate that model execution was bounded by CPU processing capability, rather than main memory capacity, suggesting that SDM workflows could be improved if the algorithms were written to run in parallel, rather than sequentially.

Models of SDM accuracy suggest that significant accuracy gains can be achieved by fitting the models with more than 2000 training examples. All three SDM algorithms showed a similar pattern of increasing accuracy as the number of input training examples increased, though the increase was not linear. Figure 2 demonstrates the accuracy of a GBM-BRT model with up to 9000 training training examples. The accuracy prediction model shows an observed-to-predicted r^2 of 0.900 and a mean prediction error of 0.001 ± 0.002 AUC. The model strongly suggests that use of the additional training data available in the fossil record can significantly enhance SDM accuracy.

Future work will be directed towards larger and more complex models of climate-species dynamics. Additional research should also investigate explicitly parallel machine learning techniques and their feasibility for SDM studies, as our results show that execution time is strongly limited by CPU-bound serial learning techniques.

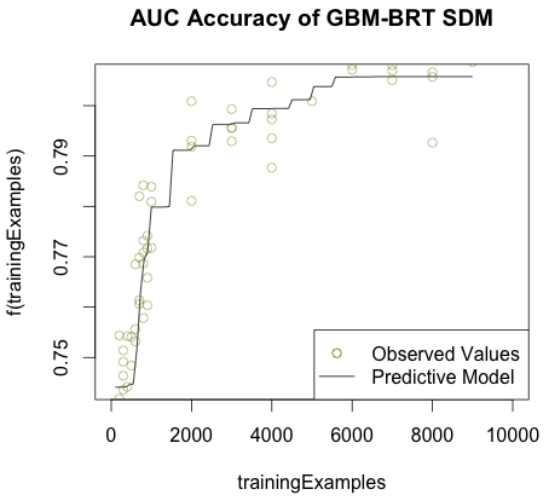

Fig. 2. Predictive model of GBM-BRT species distribution model achieved when using different input sizes.

ACKNOWLEDGMENTS

Funding for the authors was provided by University of Wisconsin-Madison Geography Department's Trewartha Research Award, the University of Wisconsin-Madison Vilas Research Trust, and the National Science Foundation (EAR-1550707).

REFERENCES

[1] J. Franklin, *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, 2010.

[2] N. Golding and B. V. Purse, "Fast and flexible bayesian species distribution modelling using gaussian processes," *Methods in Ecology and Evolution*, vol. 7, no. 5, pp. 598–608, 2016.

[3] J. Elith, C. Graham, R. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J Hijmans, F. Huettmann, J. R Leathwick, A. Lehmann, *et al.*, "Novel methods improve prediction of species distributions from occurrence data," *Ecography*, vol. 29, no. 2, pp. 129–151, 2006.

[4] J. S. Clark, A. E. Gelfand, C. W. Woodall, and K. Zhu, "More than the sum of the parts: forest climate response from joint species distribution models," *Ecological Applications*, vol. 24, no. 5, pp. 990–999, 2014.

[5] T. Hastie, J. Friedman, and R. Tibshirani, "Additive models, trees, and related methods," in *The Elements of Statistical Learning*, pp. 257–298, Springer, 2001.

[6] D. J. Lilja, *Measuring computer performance: a practitioner's guide*. Cambridge university press, 2005.

[7] Q. Wu and V. V. Datla, "On performance modeling and prediction in support of scientific workflow optimization," in *2011 IEEE World Congress on Services*, pp. 161–168, IEEE, 2011.

[8] B. C. Lee, D. M. Brooks, B. R. de Supinski, M. Schulz, K. Singh, and S. A. McKee, "Methods of inference and learning for performance modeling of parallel applications," in *Proceedings of the 12th ACM SIGPLAN symposium on Principles and practice of parallel programming*, pp. 249–258, ACM, 2007.

[9] M. S. Wisz, R. Hijmans, J. Li, A. T. Peterson, C. Graham, and A. Guisan, "Effects of sample size on the performance of species distribution models," *Diversity and Distributions*, vol. 14, no. 5, pp. 763–773, 2008.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

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

[11] D. Lorenz, D. Nieto-Lugilde, J. Blois, M. Fitzpatrick, and J. Williams, "Data from: Downscaled and debiased climate simulations for north america from 21,000 years ago to 2100ad," 2016.

[12] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, 2008.

[13] J. Leathwick, J. Elith, and T. Hastie, "Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions," *Ecological modelling*, vol. 199, no. 2, pp. 188–196, 2006.

[14] A. Guisan, T. C. Edwards, and T. Hastie, "Generalized linear and generalized additive models in studies of species distributions: setting the scene," *Ecological modelling*, vol. 157, no. 2, pp. 89–100, 2002.