I propose using the Amazon Web Services platform to assess the potential of using cloud computing in species distribution modeling (SDM) workflows. Species distribution models are a class of statistical models used by ecologists to understand past species distributions, predict future ecosystem dynamics, and test ecological hypotheses. These models use machine-learning algorithms to model a species’ response to modern environmental factors, such as temperature, precipitation, or soil type, and then use this model to predict the species distribution under a different assemblage of environmental covariates (often future conditions under climate change). The models are typically run on standard desktop computers in a laboratory setting, and can run for hours to days fitting a single species response. The cloud presents an interesting opportunity for this type of modeling by allowing investigators to use more powerful machines without the upfront cost of purchasing them. However, to date, few studies have been done to assess the potential of applying cloud computing technology to SDM models. I will use the AWS platform to understand which modeling situations are most suited for a migration to the cloud.

I will use the AWS EC2 platform to record the execution time of SDM models under different input conditions. Clearly, models that fit quickly using very little data are unlikely to be significantly impacted by using cloud-based workflows. However, models that required tens of thousands of training examples and must run at a very high spatial resolution seem to be prime candidates for running in the cloud. I will also qualitatively evaluate the ease of migrating the SDM workflow to the AWS infrastructure. Along the way, I will document my progress and develop open source scripts, so that, should the cloud prove useful for this type of modeling, other investigators will be able to easily migrate to the cloud platform.

Specifically, I propose testing the modeling workflow on the c4 (8xlarge) compute optimized instance type, with 36 virtual cores and 60 GB of memory. The computing power provided by this virtual machine should be able to quickly compute the species distribution models. I plan to use a single class of model, boosted regression trees (BRT), which have been recently shown to be highly accurate, but computationally expensive to fit. The BRT method fits many thousand weak learning decision trees in a stagewise manner to approximate the functional form of a species response to environmental conditions. My testing framework will systematically vary two experimental conditions: the spatial resolution of the input layers and the number of training examples used to fit the model.

Preliminary tests show that fitting the BRT models on more than 1000 training examples can take a considerable amount of time, however, the tradeoff between increased accuracy and more training examples has not been quantified. Preliminary tests on my laptop show an exponential relationship between the number of training examples and model execution time, though the results do not show any real change in accuracy. By running the model on the AWS infrastructure, I can reliably demonstrate the tradeoff by timing the model execution and then evaluating the model’s accuracy using a testing set of examples. State of the art models now are now implementing ‘niche pooling’, which uses the entire distribution of the species found in the fossil record to fit the model, requiring many tens of thousands of input training examples. I would like to systematically run the BRT models for between 10 and 50,000 training examples to more fully understand the time and accuracy effects of niche pooling.

Secondly, I will experiment with the affect of increasing the spatial resolution of the input layers. Increasing the spatial resolution of the predictors has been shown to provide better accuracy, however, it requires much larger file sizes and increases the time needed to predict the model onto different environmental covariate assemblages. Increasing the spatial resolution of the gridded input results in an exponential number of cells for the model to predict onto, posing a computational challenge for standard computers.

I plan to systematically test these two experimental variables on a case study using oak (genus quercus). Oak is an ecologically interesting taxon, and has been shown to be indicative of climatic changes in the fossil record. In this study, I will project the distribution of oak to the year 2100 under a realistic representative concentration pathway (RPC) specified by the UN Intergovernmental Panel on Climate Change (IPCC) in their last assessment report. I expect each model to take approximately 8 hours to fit, so I propose using 3600 hours of compute time on the c4 instance. This would let me test 450 different model configurations. The models can be designed to be fault-tolerant, and thus can utilize the spot instances available on AWS. I also request one t2 (micro) server to act a master node, hosting a database that records the results and automatically provisions and destroys additional computing nodes as necessary.

The results of this work will be distributed in three venues. First, the experiments described here form the core of my Master’s thesis, and will be documented and reported there. Secondly, I plan to disseminate my findings in a scientific paper published in a reputable journal. Because systematic testing of species distribution modeling workflows to the cloud has not been previously undertaken, the results of this study are likely to be interesting to others in the ecological community, and I have identified several journals that are likely to be interested in publishing a peer-reviewed paper on this topic. Finally, I will distribute all of the code used to setup, run, and evaluate the models on Github as an open-source project. By freely releasing annotated code and results I hope to reduce the difficulty and complexity of migrating a species distribution modeling workflow to the cloud.

If the cloud is shown to be an effective tool in the experiment described above, a permanent application will be developed to run these models as web services with an AWS backend. Models are typically run in the desktop R statistical programming environment. However, with a bit of development, they could be run through a RESTful API, a tool that I think would be very effective and extremely useful for the members of the ecological science community. Such a tool would make the models faster to compute and limit the amount of data management to be done by individual researchers. I plan to develop a prototype of this application as a section of my Master’s thesis. My research group has already expressed interest in using such a tool.

The dynamically downscaled CCAFS climate dataset, freely available on AWS, would benefit this project greatly. One of the challenges in running these models in effectively managing all of the climatic data used as model predictors. Using this preconfigured dataset would improve the ability to run the models quickly and easily.

I am requesting $4,000 in research credits to work on this project. Thank you for your thoughtful consideration. I look forward to your decision.