Biologically-informed prediction of species-level responses to climate change

Introduction: Anthropogenic climate change is having a dramatic effect on earth's natural systems. To maintain functional ecosystems and preserve biodiversity in this changing world, the conservation community has come to rely on computational models of species' geographic distributions to inform decision making[1].

Two classes of models are employed in this context[2]: correlative models, which associate observed species' occurrences with observed climate in a statistical framework to evaluate habitat suitability[3], and mechanistic models, which use mathematical representations of species' physiology parameterized on laboratory-derived physiological metrics to evaluate habitat suitability with respect to environment. Correlative models have broad taxonomic scope, low data requirements, and are largely methodologically transparent, but they are limited by the challenges of extrapolation to *future* climate scenarios, many of which do not currently exist in today's climate regime[4]. Mechanistic models are biologically-informed and thus more robust to extrapolation in novel climatic regimes but are often proprietary, data-intensive, and limited in their taxonomic scope, reducing their appeal to conservation biologists. As in **Figure 1**, these models often produce dramatically different representations of a species' future distribution.

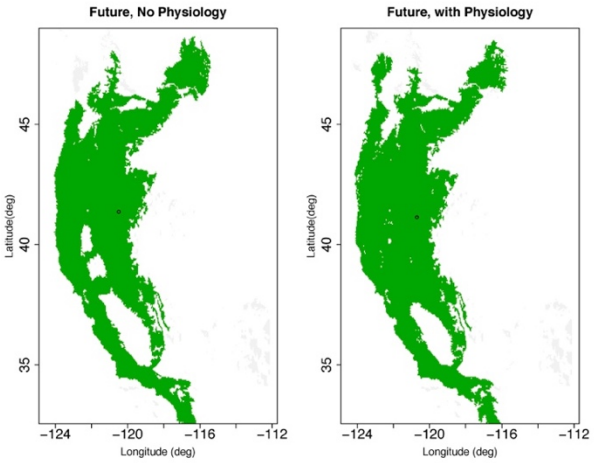

Figure 1: A lizard's *(Sceloporus occidentalis)* range is projected differently in novel climates with and without physiology. Columns, from left: correlative model projecting into 2050, correlative model with physiological priors projecting into 2050.

There is also an underexplored third modeling approach: by combining advanced statistical modeling techniques with the best available physiological research, we can develop models that leverage both decades of biophysiological research and the wealth of available species' occurrence data, together with high resolution models of climate, to create biologically-informed projections of species' distributions with high accuracy and realism.

I propose to develop novel computational approaches for the synthesis of physiological, trait, and environmental information to make more accurate, biologically informed predictions of species' responses to climate change. Leveraging this combination of factors will bring together the best features of correlative and mechanistic models, namely the broad taxonomic applicability and relatively low data requirements of correlative models, and the extrapolation robustness and biological realism of mechanistic models.

Aim 1: Develop an accurate, generalized species distribution model which incorporates laboratory-derived physiological information.

Hypothesis: A biologically-informed model will more accurately capture current species distributions when compared to standard correlative models and will more accurately project future distributions. We know that novel climatic regimes (that is, combinations of climatic conditions that don't currently exist) challenge current correlative species distribution models. I base my approach on the Gaussian Process (GP) modeling technique, which has been shown to have robust performance in capturing current species' distributions. Gaussian processes are built on a Bayesian modeling framework, which means that we can explicitly incorporate *prior knowledge* about the relationship between our covariates and the model output. I will use thermal physiological thresholds to define priors in a GP model. I will leverage the biophysical modeling expertise of my advisor, Dr. Lauren Buckley, whose collaborators have provided me with a wealth of laboratory-

derived physiological data which will be instrumental to this work. I will train and test these models on a dataset of 300 species across a variety of taxa for which both metrics of physiology and observed occurrences are available from the Global Biological Information Facility (GBIF). I intend to evaluate these biologically-informed models across a wide range of species at a global scale, and will therefore perform my computations on distributed cloud-based computation hardware and will utilize parallel database management systems to streamline analysis. Model performance evaluation is twofold: using standard methodology I will both **1)** test model accuracy in capturing current distributions and **2)** assess model performance in projecting into future climatic scenarios by measuring the shift in the species' range centroid across time, quantifying shifts in leading range edges, and using historical distribution and climate data to project backward in time and compare my prediction against known distributions. These metrics will be used to compare performance between standard correlative models, biologically-informed correlative models, and mechanistic models.

There is a wealth of functional trait data available in online databases for many genera in addition to physiology, which can similarly inform the development of more accurate, biologically-informed distribution models. Aim 2 explores the relationship between these other functional traits and their effect on range shifts to further enhance a biologically-informed modeling toolkit.

Aim 2: Develop statistical machine learning techniques to autonomously capture the complex relationships between functional traits and climate-induced distribution shifts with higher accuracy than linear modeling.

Hypothesis: The magnitude of species range shifts is more accurately predicted by nonlinear modeling techniques than by existing approaches. Previous work[6] has attempted to leverage functional trait data (e.g. dispersal ability, diet) to model shifts in species' geographic ranges. These efforts have used linear models to do this and found significant relationships but failed to explain enough variation to make the results useful in a predictive context. It is likely that the functional form of the relationship between functional trait and range shift magnitude is non-linear, and our initial experimentation suggests this. We will use functional trait and range shift data from a series of Swiss alpine plants, Yosemite small mammals, and seabirds from Dr. Buckley's collaborators to assess the performance of new nonlinear methods such as random forests, support vector machines, and others in predicting range shift magnitude from trait values.

Broader Impact: 1) Climate Mitigation: A significant effort in our lab is the **Tr**anslating **En**vironmental **Ch**ange Project (TrEnCh: http://trenchproject.github.io), wherein we build computational and visualization tools to translate environmental change into organismal responses. As a member and active participant in the TrEnCh project, together with my coursework in conservation decision making and collaborators at the UW Climate Impacts Group (CIG), I will work to **a)** combine physiology and traits into a single, coherent species distribution model which is flexible to data availability and **b)** create a web application and suite of programmatic tools in R and Python which make this model and the data necessary to run it available to anyone with a computer. I will then work with my connections to the conservation and climate adaptation community in Colorado (Aspen Center for Environmental Studies, Wilderness Workshop) and Washington (UW CIG) to hone and employ these tools. 2) Reproducible, Open Science: Just as clean lab protocols and detailed results are expected from laboratory experiments, so should clean, understandable code be in computational work. I will use this work as an opportunity to set an example for clean, open, and reproducible scientific software development.

References: [1]Evans TG, Diamond SE, Kelly MW (2015) *Conserv. Physiol.* **3:** cov056. [2]Dormann, CF *et al.* (2012) *J. Biogeogr.* **39:** 2119–2131. [3]Elith J and Leathwick JR (2009) *Annu. Rev. Ecol. Evol. Syst.* **40:** 677-697. [4]Williams JW and Jackson ST (2007) *Front. Ecol. Environ.* **5**: 475-482. [5]Golding N and Purse, BV (2016) *Methods Ecol. Evol.* **7**: 598-608 [6]Angert AL *et al.* (2011). *Ecol. Lett.* 14: 677-689

To me, one of the greatest things about being a scientist is the freedom: freedom to think, to ask, to be wrong, and to learn. I discovered this feeling on my first day of Introduction to Computer Science during my first semester at Tufts University. There I was: a biologist with aspirations in medicine, taking a Computer Science course on a good friend's suggestion and a whim. Within minutes of listening to Professor Ben Hescott introduce the study of computer science, I became curious. The opportunities to learn seemed endless, about topics I didn't even realize existed. Specifically, the idea that a computer could be programmed to discover patterns that occur in the natural world from huge quantities of information, and be interrogated for answers to questions about those patterns, thrilled me. The possibilities to apply computational methods to biological problems appeared enormous. I thought: *how could I use these computational techniques to be a better biologist?*

Days after that first lecture I went to Dr. Hescott's office hours to find out. He is a computational theorist who enjoys solving difficult molecular biology problems, and encouraged me to explore the field. I spent that summer working as a research assistant in a computational molecular biology laboratory at Brown University, under Dr. Ben Raphael and postdoctoral fellow Dr. Suzanne Sindi. It was my first exposure to writing code in a scientific context—**I helped to improve the algorithmic efficiency and usability of software for DNA structural variant analysis**. I loved how the study of biology could be aided by computational power, producing knowledge unattainable otherwise.

Having returned to Tufts, I was ready to embrace computational biology in whatever form I could get my hands on. As I decided to pursue a degree in Computer Science and Biology, **I began to study the data structures, machine learning algorithms, programming languages, and systems required to be a real-life computer scientist.** I began work on a project with Dr. Hescott to **predict the function of functionally-unannotated proteins from DNA and amino acid sequence data** using the known protein-protein interaction network and existing functional annotations. My work on that project evaluated the feasibility and performance of incorporating *genetic* protein-protein interactions (protein-protein interactions in which simultaneous mutation in the coding regions for both proteins exhibit a phenotype different than the mutations in isolate) into Dr. Hescott's existing prediction algorithm. I worked with several graduate students and faculty on this project, who were invaluable to the work and who served as great mentors. I worked on that project for **two funded summer research sessions, and I produced code for protein homology detection which is still in use today.** This was an invaluable experience which taught me what full-time research was like, how to work with others, and what the expectations of computational research are. I was simultaneously quite flexible with regard to the actual nature of the biology I was studying—my interest was wide and deep. I worked with my Biology advisor Dr. Erik Dopman to choose **an expansive curriculum of Biology courses** to take alongside my Computer Science curriculum, ranging from physiology, genetics, molecular biology, ecology, and conservation biology. As I began to study more biology, I began to appreciate that though my current work was on a very small (molecular) biological scale, I was also quite interested in larger-scale biology (ecosystems).

One of the origins of this interest is in my time spent outside. Throughout college I often escaped to the woods with the Tufts Mountain Club to hike, backpack, rock climb, and enjoy time out of doors. As I began to lead trips for the club, the time in New Hampshire's White Mountains brought me to realize my curiosity for ecosystems and their resiliency in the face of climate change. **How will organisms and the systems they comprise react to changing climate? Can we predict these responses computationally?** If so, **how can we build tools and**

systems to disseminate those predictions to the people who can use them? These questions motivated my search for researchers who studied these questions, and that search brought me to Dr. Lauren Buckley in UW's Biology department. When I wrote to Dr. Buckley, we talked about the myriad opportunities for merging my computational knowledge with her quantitative physiological background and our shared interest in publicly disseminating useful models. The opportunities to **hone my computational skills as a Big Data IGERT student** at the eScience Institute for Data-Intensive Discovery at UW drew me specifically to UW's program in Biology, and Dr. Buckley's funded initiatives to perform this work provided further incentive. I was resolved to study both the systems that inform the development of general models for an ecosystem's response to climatological stress and to develop tools and initiatives which disseminate these findings on a public stage.

Before I began graduate study I resolved to take some time to grow my passion for teaching and communicating science to others. **I became a Naturalist at the Aspen Center for Environmental Studies** in Aspen, Colorado, where for 14 months I studied the stories of my surroundings—from the ecological to the historic—and learned to bring them to bear in the pursuit of inspiring visitors to learn more about their place, both in Aspen and back at home. I led nature hikes, ski tours, bird-of-prey programs, snowshoe walks, and kids' camps with new families, senior center groups, certain tech executives, board members, politicians, honeymooners, and many others. I learned that **when employed correctly, and honed for each audience, rigorously-derived scientific knowledge can drive inspiration and action in anyone.**

Figure 1: Me while teaching a reptile education program in Colorado.

Only by informed action can we attempt to mitigate anthropogenic impacts on natural systems. I believe that the role of the scientist doesn't need to stop at the final sentence of our journal submissions; it is up to the researcher to advocate for their research, to bring life to it, and to produce actionable knowledge. I hope with my work to provide the **conservation and policymaking communities** with actionable, easy-to-understand models of climate change impacts on actual biological systems informed by the **best available biological information** and machine learning techniques. I believe that by leveraging my leadership experiences, teaching experiences, and computational skills I can bring this vision to fruition. In fact, I'm already hard at work: since I've started studying at UW, I've been awarded a **two-year Big Data IGERT PhD Traineeship** for this work, have **taken courses on connecting science and policy through UW's Climate Impacts Group,** and am taking part in the **Advanced Data Science Curriculum Option** here at UW.

To be clear: the challenge of **building resilience to climate change** into twenty-first century life and policy is almost unspeakably formidable. Yet I find working in a domain related to climate change and our response to it to be motivating and meaningful. I see the potential for computational and modeling techniques (paired with interfaces designed for their wide use) to dramatically lower the walls between science and informed action. I hope with this PhD degree to develop into a multifaceted interdisciplinary quantitative researcher with strong ties to the policy and conservation communities to advancing our ability to respond to climate change.

As a graduate student at the University of Washington, I will work to synthesize the best available statistical modeling techniques with available biophysical and physiological information to create a suite of species distribution models for the conservation community. I am

as passionate about the modeling approaches as I am about their widest possible dissemination and use. I will take my experience building high-traffic websites for college students, my experiences organizing groups together for a common goal, and my desire to advocate for informed climate action together to build a suite of publicly-accessible tools tailored for conservation. These tools, which I envision as a website and an R or Python package, will require extensive cyberinfrastructure to be easy-to-use, but I am willing to accept the challenge of building these systems. I also have the support of the team of data scientists and domain experts at the eScience Institute for Data Intensive Discovery at UW, my IGERT traineeship coursework and seminars, and of my co-advisor Magda Balazinska. Dr. Balazinska is a databases researcher in UW's computer science department, and is regarded as one of the best in her field; with her guidance, implementing systems which can leverage the global store of biological information quickly and with high accuracy will be quite possible.

Finally, I believe firmly in reproducible research, especially in computational science. In a recent conversation with Dr. David Beck, UW eScience Institute Director of Research and an IGERT advisor, he essentially told us students that for a researcher to create software, experiments, and protocols and keep them under lock and key is to sabotage the research from the beginning. Open science, where data, methods, and code are all available for anyone to view, is critical to both maintaining accountability and ensuring accuracy but also to the whole enterprise of science itself. Developing programming language literacy, using best practices, leveraging the most suitable tools and packages for a given task, and creating code which is as understandable as the text in a journal article are critical skills for the researcher who intends to create open, reproducible science. To this end I've embarked upon an effort to enhance the quality of scientific code in my sphere here at the University of Washington. Thus far I have trained as a Software Carpentry instructor, where I teach Python and R skills to researchers. I have also begun leading tutorials and workshops on software packages for geospatial data analysis, most recently at the eScience Institute's "GeoHackWeek," which is a weeklong workshop on best practices and new methods for geospatial data. In November of this year, I will attend the Moore-Sloan Foundation's Data Science Summit, where I will connect with other like-minded researchers to further progress on the quest for open, reproducible science.

This NSF GRFP opportunity will provide me with the resources to fulfill these goals above; the freedom granted by this generous support will not only allow me to grow into a contributing expert in the field of modeling species' responses to climate change, but will also allow for the creation of real and useful tools together with conservationists, and for the growth of open, reproducible science at UW and beyond.

Publications and Presentations

Lauren B. Buckley, Andrew J. Arakaki, Anthony F. Cannistra, Heather M. Kharouba, Joel G. Kingsolver; Insect Development, Thermal Plasticity and Fitness Implications in Changing, Seasonal Environments. Integr Comp Biol 2017 icx032. doi: 10.1093/icb/icx032

Anthony F. Cannistra; Intelligent Species Distribution Modeling via Traits and Physiology. eScience Institute Community Seminar, Seattle, WA. May 2017.

Anthony F. Cannistra, Lauren B. Buckley; Improving range shift predictions: Enhancing the power of traits. Ecological Society of America Meeting, Portland, OR. August 2017.

Randall J. Levesque, Anthony F. Cannistra; Tools for Geospatial Visualization in Python. GeoHackWeek 2017, University of Washington, Seattle, WA. September 2017.