5-2022

Refining, Testing, and Applying Thermal Species Distribution Models to Enhance Ecological Assessments

Donald J. Benkendorf
Utah State University

REFINING, TESTING, AND APPLYING THERMAL SPECIES DISTRIBUTION MODELS TO ENHANCE ECOLOGICAL ASSESSMENTS

by

Donald J. Benkendorf

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Ecology

Approved:

_____	_____
Charles Hawkins, Ph.D. Major Professor	D. Richard Cutler, Ph.D. Committee Member
_____	_____
Scott Miller, Ph.D. Committee Member	Edd Hammill, Ph.D. Committee Member
_____	_____
Brett Roper, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Interim Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2022

ABSTRACT

Refining, Testing, and Applying Thermal Species Distribution Models

to Enhance Ecological Assessments

by

Donald J. Benkendorf, Doctor of Philosophy

Utah State University, 2022

Major Professor: Dr. Charles P. Hawkins
Department: Watershed Sciences

Thermal regimes are changing rapidly, and temperature is strongly associated with the distributions of many species. Accurately modeling temperature – distribution associations provides a way to predict effects of changing temperature on species distributions. These models could also be used in the development of stressor-specific biotic indices, which are based on species-stressor tolerances, and could be used to diagnose if a stressor has altered aquatic life. Such stressor-specific indices could increase confidence in which stressors need to be managed. My broad research objectives were to improve understanding of how temperature affects aquatic invertebrate distributions, assess if new techniques can improve models describing temperature – distribution relationships, and develop a temperature-specific biotic index (TBI). In chapter two, I use chronic exposure laboratory experiments (>one week) to improve understanding of how temperature affects macroinvertebrate distributions. I found that lab-derived upper thermal limits based on survival were strongly associated with field-derived upper thermal limits (r^2 = 0.72), which supports the likelihood that temperature-based species distribution models (SDMs) have a mechanistic foundation. In chapter three, I compare how five methods for adjusting for data imbalance and four machine-learning algorithms affect SDM performance. I found that all methods for dealing with imbalanced presence-absence data improved SDM performance over the base models. In chapter four, I assess how sample size (100 to 10,000 observations) and

network depth (1 to 6 layers) affect the performance of deep-learning based SDMs. I found that deeper networks overfit the training data, but overfitting was reduced on the largest sample size. There was no benefit of additional layers on performance, and random forest models generally performed as well as all neural network models. In chapter five, I develop a TBI and assess if thermal alteration has potentially affected aquatic life in the Nation's streams and rivers. The TBI was generally sensitive and specific to spatial variation in mean summer stream temperature across sites. Applying the TBI to streams and rivers across the Nation implied that the invertebrate assemblages in approximately 2.6% of streams and rivers have been altered by thermal pollution.

(159 pages)

PUBLIC ABSTRACT

Refining, Testing, and Applying Thermal Species Distribution Models to Enhance Ecological Assessments

Donald J. Benkendorf

The temperature of streams and rivers is changing rapidly in response to a variety of human activities. This rapid change is concerning because the abundances and distributions of many aquatic species in streams and rivers are strongly associated with temperature. Linking observations of temperature effects on species distributions with observations of temperature effects on fitness is important for improving confidence that temperature (and not some other variable) is causing the distributions we observe. Furthermore, producing accurate models of temperature effects on species distributions may allow us to develop tools to diagnose whether or not thermal pollution has impaired aquatic life. Such a diagnostic tool could help us better target management efforts on the specific stressors impairing aquatic life. In chapter two, I describe several laboratory experiments designed to examine the link between the effects of temperature observed in the field with effects of temperature observed in the laboratory. I found that the effects of temperature on survival were correlated with the thermal limits inferred from species distributions, which supports the hypothesis that temperature influences distributions by affecting the survival of species. In chapters three and four, I assessed two techniques that could potentially improve our ability to model relationships between temperature and distributions. In chapter three, I show that methods for dealing with imbalanced data broadly improved our ability to model the relationship between predictor variables (temperature and other variables) and species distributions. In chapter four, I evaluated a recently developed technique (deep artificial neural networks) for modeling large complex datasets. I found that deep artificial neural networks did not improve predictions over that of standard artificial neural networks and random forest models. In chapter five, I developed and evaluated a diagnostic biotic index for diagnosing the likelihood

that temperature has affected macroinvertebrate species in streams and rivers. This index showed that 2.6% of streams across the continental United States had species with thermal tolerances higher than expected compared with thermally undisturbed conditions.

ACKNOWLEDGMENTS

The research in this dissertation was supported by a Utah State University Presidential Doctoral Research Fellowship, a National Science Foundation grant DEB-1456278, and Washington Department of Natural Resources Agreement No. 93-099051.

I thank my advisor Dr. Chuck Hawkins for his scientific training and for setting an example as a researcher that I will continue to try to emulate. I also thank my graduate committee members, Drs. Brett Roper, Edd Hammill, Richard Cutler, and Scott Miller, for their guidance and feedback. I was fortunate to work alongside, and learn from, many passionate and collaborative colleagues including Andrew Caudillo, Angela Merritt, Christian Perry, Daniel Nelson, Jamie Eddings, Jennifer Courtwright, John Olson, Katy Gardner, Matt Tagg, Ryan Hill, Samuel Schwartz, Trip Armstrong, and Umarfarooq Abdulwahab. I am grateful to everyone in the Department of Watershed Sciences at Utah State University that provided such a collaborative and welcoming atmosphere. I am also very grateful to Brian Bailey and Enid Kelley for all their help. Finally, I thank my family for their support and encouragement.

Donald J. Benkendorf

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The thermal regimes of freshwater ecosystems are changing quickly in response to anthropogenic activities (Poole and Berman 2001, Caisse 2006, Burgmer et al. 2007). These changes in temperature are alarming because temperature is strongly associated with the distribution of many aquatic species and ecosystem functioning (Petchey et al. 1999). Thus, as temperature regimes continue to change, aquatic ecosystems will likely change as well. Hypotheses have been posed regarding the causes by which temperature alters distributions (e.g., Sweeney and Vannote 1978), but thorough experimental validation of these hypothesized causes is still lacking. Empirically linking observations of temperature effects on distributions with temperature effects on fitness will improve confidence in predictions made from correlative thermal species distribution models (SDMs) (Dormann et al. 2012). Furthermore, understanding the causal mechanisms underlying SDMs provides the conceptual underpinnings for management tools. For example, management tools could include temperature-specific biotic indices that aim to diagnose alteration of aquatic life caused by temperature. Such diagnostic tools would complement the currently used indices that assess overall biological condition. These general condition indices do not diagnose the stressor or stressors causing impairment of biological condition. Improving our ability to interpret, and model, the effects of temperature on species distributions may lead to the development of tools that improve our confidence in identifying if temperature-caused changes to aquatic life have occurred and thus help target specific restoration activities. In chapter two, I assess if lab-derived upper thermal limits of seven species are consistent with the upper thermal limits derived from observational field surveys. In chapters three and four, I evaluate two different modeling techniques that may improve the performance of SDMs. In Chapter three, I compare the effectiveness of several methods for dealing with imbalanced data at improving machine learning-based SDM performance. Chapter four focuses

on the effects of number of hidden layers and sample size on neural network-based SDM performance. Finally, in chapter five, I describe the development of a temperature-specific biotic index that incorporates species-specific thermal tolerance values to assess if thermal alteration has likely altered macroinvertebrate species composition in streams and rivers. I then evaluate the effectiveness of the temperature-specific biotic index at diagnosing thermal alteration of aquatic life in streams and rivers across the United States.

Laboratory experiments are a common approach to derive and validate causal explanations for patterns observed in nature. Frequently, these experiments are short-term and expose species to acute temperatures that affect survival rapidly (often over the course of hours). However, thermal tolerances derived from field-data do not always parallel these short-term, laboratory-derived thermal tolerances (e.g., Sokolovska 2014). Unfortunately, comparisons of lab-derived thermal tolerances with field-derived thermal tolerances are rare. Longer-term experiments that expose species to different temperatures within the natural range of temperatures experienced in nature may be needed to generate ecologically meaningful measures of temperature tolerance. Longer-term experiments also allow for the assessment of non-lethal responses that can affect fitness. For example, Sweeney and Vannote (1978) hypothesized that temperatures beyond the optimal range for a species may cause reduced growth, size, and fecundity, which limits a species ability to persist long-term. In chapter 2, I describe laboratory experiments in which I reared seven macroinvertebrates over one to 11 weeks at several different temperatures, monitored two processes affecting fitness (growth and survival), and compared lab-derived thermal tolerances with field-derived thermal tolerances.

Thermal SDMs are increasingly being used to predict the effects of thermal alteration on species distributions. Therefore, our ability to accurately model temperature-distribution relationships is important. The rapid advance in machine learning has improved our ability to model species-environment relationships, which are often nonlinear and complex (Cutler et al. 2007). However, these machine-learning algorithms often struggle to make meaningful

predictions when the classes in the data are imbalanced (Japkowicz and Stephen 2002), as is often the case with species presence/absence data. There are generally far more absences than presences. In particular, the resulting models often severely over predict absences and under predict presences (high specificity, low sensitivity). The broader field of data science has developed methods to deal with this imbalance problem by balancing the tradeoff between model sensitivity and specificity. Imbalance-correction methods do not always balance the actual number of presences and absences used to train the model (e.g., down-sampling and up-sampling), but they do always attempt to balance the tradeoff between sensitivity and specificity (e.g., cutoff, weighting, down-sampling, and up-sampling). In chapter 3, I assess the effectiveness of four common imbalance-correction methods and four common machine-learning algorithms in balancing the tradeoff between sensitivity and specificity.

Advances in machine learning are happening at a fast rate. In particular, the field of deep learning (neural networks with more than one hidden layer) has led to improvements in model performance in a variety of disciplines, but it has been scarcely applied to species distribution modeling (Botella et al. 2018). However, the effects of network depth (i.e., number of hidden layers) on model performance is not well understood. Furthermore, the relative performance of deep learning is very dependent on the size and complexity of the dataset being modeled. In chapter 4, I assess the effects of sample size and network depth on the performance of stream macroinvertebrate SDMs.

Thermal regimes of freshwater ecosystems are changing quickly. We need tools that allow us to diagnose and track the effects of changing temperature on aquatic life. The ability to diagnose the stressor (e.g., temperature) that is causing impairment to biological condition could lead to more effective management, because managers could focus on mitigating the source of the impairment. A temperature-specific biotic index that incorporates species-specific thermal tolerance values may provide such a diagnostic tool. In chapter 5, I describe the development and evaluation of a temperature-specific biotic index (TBI). I apply the TBI to stream and river sites

distributed across the continental United States (CONUS) and infer the thermal alteration that has occurred at the CONUS-level based on the thermal response signatures of aquatic macroinvertebrate assemblages.

References

Botella, C., Joly, A., Bonnet, P., Monestiez, P., Munoz, F., 2018. A deep learning approach to species distribution modelling. *In* Multimedia Tools and Applications for Environmental & Biodiversity Informatics (pp. 169-199). Springer, Cham.

Burgmer, T., Hillebrand, H., Pfenninger, M., 2007. Effects of climate-driven temperature changes on the diversity of freshwater macroinvertebrates. Oecologia. 151, 93-103.

Caissie, D., 2006. The thermal regime of rivers: a review. Freshwater biology. 51, 1389-1406.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J., 2007. Random forests for classification in ecology. Ecology. 88, 2783-2792.

Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Romermann, C., Schroder, B., Singer, A., 2012. Correlation and process in species distribution models: bridging a dichotomy. Journal of Biogeography. 39, 2119-2131.

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. Intelligent data analysis. 6, 429-449.

Petchey, O. L., McPhearson, P. T., Casey, T. M., Morin, P. J., 1999. Environmental warming alters food-web structure and ecosystem function. Nature. 402, 69-72.

Poole, G. C., Berman, C. H., 2001. An ecological perspective on in-stream temperature: natural heat dynamics and mechanisms of human-caused thermal degradation. Environmental Management. 27, 787-802.

Sokolovska, I., 2014. Are experimentally derived estimates of thermal tolerance useful in interpreting species distribution models. Master of Science Thesis. Available from: https://digitalcommons.usu.edu/etd/3695.

Sweeney, B. W., Vannote, R. L., 1978. Size variation and the distribution of hemimetabolous aquatic insects: two thermal equilibrium hypotheses. Science. 200, 444-446.

CHAPTER 2

PREDICTING DISTRIBUTIONS OF FRESHWATER MACROINVERTEBRATES FROM LABORATORY-DERIVED ESTIMATES OF UPPER THERMAL LIMITS*

Abstract

Thermal regimes are strongly associated with the distributions of aquatic species, and these associations are often modeled to predict where species are likely to occur. These associations imply that temperature is an important driver of distributions, but they could also occur if temperature is correlated with one or more other factors that actually influence the fitness of species. Experimental validation is therefore needed to increase confidence in interpreting these associations as causal. Metrics like critical thermal maxima (CTMs) and median lethal concentrations (LC50s) based on short-term (i.e., hours to a few days), laboratory bioassays have been used to measure differences in the thermal tolerances of aquatic macroinvertebrates, but CTMs and LC50s can show little correspondence with tolerances derived from field survey data. Lack of correspondence between field- and laboratory-derived limits implies that either the short-term bioassays do not scale up well to natural settings or field-derived thermal limits are inaccurate. Despite the lack of robust experimental validation, field-derived thermal (and other environmental) limits are frequently used to inform environmental management and conservation planning, hence it is critical that we understand if they adequately characterize species preferences or tolerances. In this study, we tested the hypothesis that upper thermal limits (UTLs) derived from > one week long experiments would be predictive of field-derived limits. We used indoor, temperature-controlled water troughs to expose seven species of freshwater macroinvertebrates to chronic (one to several weeks) constant temperature treatments of 13 – 24 °C. We measured survival at weekly intervals and growth at bi-weekly intervals and estimated UTLs as the temperature at which zero survival or growth was predicted to occur. We then

* Coauthored by Charles P. Hawkins

compared these UTLs with UTLs derived from field survey data. Laboratory- and field-derived UTLs were strongly correlated (r^2 = 0.72), suggesting that field-derived UTLs do reflect a causal association between distributions and temperature in nature. Experiments > one week in duration conducted across temperatures observed in nature appear to be sufficient to adequately characterize differences among freshwater macroinvertebrates in their response to variation in temperature.

Introduction

Understanding the causes underlying environment – distribution associations is critical to effective environmental management and conservation planning. Environment – distribution associations are often modeled (typically referred to as species distribution models, SDMs) to predict the effects of environmental variation on species distributions (e.g., Dormann et al. 2012, Guisal et al. 2013). Such predictive models are frequently used in bioassessment (Moss et al. 1987, Wright 1995, Hawkins and Yuan 2016), causal assessment (Chessman and McEvoy 1997, Yuan 2006), and other areas of environmental management and conservation ecology (Kearney et al. 2010, Guisan et al. 2013). For example, in bioassessment, SDMs are used to predict the biota expected to occur at a site under natural (reference) conditions – a critical step in quantifying alterations or loss in biodiversity via indices of taxonomic completeness (e.g., Moss et al. 1987, Wright 1995, Hawkins et al. 2010). However, other correlated, and sometimes unmeasured, variables may actually be causing the distributional patterns (Kearney and Porter 2004, 2009, Crozier and Dwyer 2006, Dormann 2007, Braunisch et al. 2013, Allen-Ankins and Stoffels 2017). Despite the risk of spurious correlations, developers and users of these models often implicitly assume that the environmental predictors used in SDMs are causally related to species' distributions (Dormann et al. 2012). These assumptions must be tested and validated before we can interpret predictions with confidence.

Temperature is thought to be a key factor controlling the spatial distributions of freshwater invertebrate species and other ectotherms (Sweeney and Vannote 1978, Burgmer et al. 2007, Pearson and Dawson 2003, Deutsch et al. 2008). Both metabolic theory and empirical observations indicate that temperature can influence individual growth and survival and that these effects may scale up to affect abundances and distributions (Sweeney and Vannote 1978, Gillooly et al. 2001, Brown et al. 2004, Carlo et al. 2018). The thermal equilibrium hypothesis provides a valuable conceptual framework regarding these effects and states that temperatures beyond an optimal range lead to reduced growth, size, fecundity, and survival, which individually or in combination can limit a species' abundance and distribution (Sweeney and Vannote 1978, Vannote and Sweeney 1980). It follows that our confidence in inferring effects of temperature on species distributions should improve when fitness, or metrics of fitness, can be measured. For example, measurements of the effects of different concentrations of total dissolved solids on both growth and survival, two components of fitness, improved predictions of the distributions of several macroinvertebrate species relative to predictions based on just growth or survival alone (Olson and Hawkins 2017). However, we only found a few studies that have attempted to directly link measures of thermal tolerance obtained from controlled experiments to distributions observed in nature (Diamond et al. 2012a, Allen-Ankins et al. 2017, Ángeles-González et al. 2020, Rendoll-Cárcamo et al. 2020, Rezende et al. 2020).

Most experimentally-based estimates of thermal tolerance have been derived from bioassays that quantified acute responses to short-term (typically minutes to ≤ 96 hr) exposure to different temperatures, which often greatly exceed temperatures a species would encounter in nature. Critical thermal maximum (CTM) and median lethal concentrations (LC50) are commonly derived from such short laboratory experiments (e,g., Dallas and Ketley 2011). Few studies have compared predictions from these measures of thermal tolerance with limits derived from field survey data (Diamond et al. 2012b, Sokolovska 2014, Shah et al. 2017, Rendoll-Cárcamo et al. 2020). Furthermore, the strength of associations between field- and laboratory-

derived estimates of thermal tolerance vary substantially across the studies that have been conducted (Sokolovska 2014, Shah et al. 2017, Rendoll-Cárcamo et al. 2020). For example, Sokolovska (2014) found no association between CTMs and field-derived upper thermal limits (UTLs) for 32 stream macroinvertebrates collected from streams in Utah (USA). The inconsistencies among these studies imply that either the results obtained from short-term experiments do not always scale up to natural settings or that the thermal limits derived from survey data are inaccurate.

We suspect the use of acute response metrics, especially near-lethal CTMs, to predict how species will respond to spatial or temporal changes in temperature regimes may often be misleading because they may not accurately characterize how sublethal effects to temperature manifest over an organism's life cycle. For example, several studies have defined tolerance of a species to warming as the difference between the laboratory-derived CTM and a measure of the environmental temperature at which it is found (e.g., mean temperature over warmest yearly quarter) (Deutsch et al. 2008, Huey et al. 2009, Diamond et al. 2012b). They hypothesize that ectothermic species that occur at lower latitudes are more at risk from warming temperatures than those that occur at higher latitudes. This hypothesis is based on the idea that the margin of safety between a species' CTM and environmental temperature is lower than for species at higher latitudes, even though temperatures are predicted to rise faster at higher latitudes with projected climate change (Deutsch et al. 2008, Huey et al. 2009, Diamond et al. 2012b, Chown et al. 2015, Shah et al. 2017). However, if the acute physiological limits captured by CTMs are not strongly associated with how fitness responds to chronic thermal exposures, then using CTMs to define tolerance to warming of species may be misleading (Kim et al. 2017, Rezende et al. 2020).

Here we define a natural-temperature, chronic-exposure experiment (NTCEE) as including temperature treatments that a species could likely encounter in nature within its range and that runs long enough that fitness related metrics can be measured. We think that NTCEEs that subject individuals to prolonged (days to weeks) exposures are better suited than CTMs to

test the accuracy of field-derived temperature-distribution associations. NTCEEs will typically represent a compromise between full-lifecycle experiments, which are difficult and expensive to run, and short-term CTM- or LC50-like experiments that are easier and inexpensive to run. Full-lifecycle experiments are the gold standard for studying effects of environment factors on species in the laboratory and provide a wealth of detailed information on fitness responses (e.g., Sweeney et al. 2018). However, we hypothesized that experiments of one or more weeks should provide sufficient signal on fitness responses to improve our confidence in interpretating environment-distribution associations. Freshwater macroinvertebrates should be a good group to assess the potential advantages of NTCEEs for testing causal interpretations of the effects of temperature and other environmental factors on species distributions. For example, macroinvertebrates are ectotherms, have diverse environmental preferences and tolerances, and typically have lifecycles lasting several months to several years (Poff et al. 2006, Verberk et al. 2008).

In this study, we tested the hypothesis that UTLs derived from laboratory NTCEEs will be strongly associated with UTLs inferred from field distributions. We also hypothesized that measures of both growth and survival will be associated with field-derived UTLs and that a fitness index that incorporates measures of both growth and survival would produce the strongest associations with field-derived UTLs.

Methods

General approach

To test our hypotheses, we compared UTLs derived from 1-11 weeklong laboratory experiments with those derived from field survey data. The laboratory experiments included seven aquatic macroinvertebrate taxa (Table 2-1) that we collected from northern Utah (USA): *Pteronarcys californica* Newport, 1848 (Insecta, Ephemeroptera, Pteronarcyidae), *Drunella grandis* Eaton, 1884 (Insecta, Ephemeroptera, Ephemerellidae), *Hyalella azteca* Saussure, 1858 (Malacostraca, Amphipoda, Hyalellidae), *Gammarus lacustris* G. O. Sars, 1863 (Malacostraca,

Amphipoda, Gammaridae), *Drunella coloradensis* Dodds, 1923 (Insecta, Ephemeroptera, Ephemerellidae), *Cinygmula sp.* McDunnough, 1933 (Insecta, Ephemeroptera, Heptageniidae), and *Rhithrogena robusta* Dodds, 1923 (Insecta, Ephemeroptera, Heptageniidae). We reared each species at six or seven different constant temperatures and measured both survival and size (length) of each species at fixed intervals over the duration of the experiments. After each experiment, we used regression models to predict the temperatures at which survivorship or growth was zero. We used 90th percentiles of occurrences obtained from Richards et al. (2013) as the measure of field-derived UTLs for the seven species. We then tested our hypotheses by calculating the Pearson correlation between the laboratory- and field-derived UTLs for the seven species. We did not expect to observe a one-to-one correspondence between laboratory- and field-derived UTLs because exposure temperatures are typically calculated over different time periods in laboratory and field settings (e.g., daily or weekly means vs summer or annual means), but we did expect to see a strong correlation between the two estimates.

Table 2-1. The seven species used in the experiments. Collection date indicates the month and year each species was collected from the field, and experiment start date is the date the laboratory experiment began. FFG = functional feeding group. The instantaneous temperature recorded at each collection site is reported with the time of day (MST) the temperature was recorded. Collection site temperatures were all measured on 8/25/2020 to improve comparability and avoid confounding with day of collection. The instantaneous temperature for the spring from which *Hyalella azteca* was collected is not available because the spring was not accessible after animals were collected.

Species	Voltinism	FFG	Collection date	Experiment start date	Collection site temp	Collection site time
P. californica	semivoltine	SH	8/2018	8/29/2018	17.1 C	3:45 pm
D. grandis	univoltine	CG	1/2019	2/1/2019	17.2 C	3:30 pm
H. azteca	uni/multivoltine	CG	3/2019	3/23/2019	-	-
G. lacustris	uni/multivoltine	CG	5/2019	5/18/2019	9.4 C	2:50 pm
D. coloradensis	univoltine	CG/PR	6/2019	7/1/2019	16.2 C	4:40 pm
Cinygmula sp.	Univoltine	CG/SC	7/2020	7/28/2020	7.2 C	4:15 pm
R. robusta	univoltine	CG/SC	7/2020	7/30/2020	7.2 C	4:15 pm

Laboratory-derived UTLs

We conducted the NTCEEs in 3.5 by 0.5 m indoor, experimental flow-through troughs (Fig. 2-1) at Utah State University. We exposed each species to six or seven different temperature treatments (13, 14, 16, 18, 20, 22, 24 °C) that encompassed the range of field-derived UTLs reported by Richards et al. (2013) for many species of macroinvertebrates found in the western United States. This range of temperatures was appropriate for assessing trends in growth and survival that would allow us to estimate UTLs. Four of the species were only reared at six different temperatures (either omitting the 13 °C or the 24 °C treatment) because only six troughs were available at the time of those experiments. Temperature treatments were not independently replicated, but we did randomly assign temperature treatments to troughs to avoid potential confounding of temperature treatments with other unknown factors that might have varied systematically with the position of the troughs.

We used ThermoScientific™ recirculating heaters placed at the top of each trough to maintain near constant temperature treatments (Appendix Fig. A.3 and A.4). Each temperature treatment consisted of one flow-through trough with a spigot at the top that provided a constant inflow (~1.6 L/min) of well water. We were not able to chill water in the experimental troughs, so the lowest experimental temperature of 13 °C was constrained by the temperature (~12-13 °C) of the untreated well water that supplied the experimental troughs.

Within each trough, we placed five rearing chambers that were each made of 3.8 L plastic jars with two sides and the bottom removed and replaced with ~1mm Nitex™ mesh netting. The mesh netting permitted exchange of water in the rearing chamber with the water in the trough. We placed several air stones in each trough to ensure high O_2 concentrations (percent saturation $\geq$ 70). Oxygen in the troughs was similar to the oxygen in the mesh-sided rearing chambers (Appendix Fig. A.1 and A.2), and water temperatures in the troughs were also similar to water temperatures in the rearing chambers (Appendix Fig. A.3 and A.4). We used a 15-cm standpipe at the bottom end of each trough for outflow. The 15-cm standpipe ensured a constant water depth

in the troughs, regardless of small fluctuation in inflow from the spigot. For each experiment, we placed five individuals of a species in each of the five rearing chambers for all treatments (i.e., 25 individuals per treatment).

The experiment for each species began following an acclimation period, after which water temperatures in the troughs were increased by 2 °C approximately every 24 h until all treatment temperatures were reached. The initial water temperature in all troughs was ~12-13 °C. For several of the species, a few individuals died during temperature ramping and were promptly replaced so that day one of the experiment began with five individuals/rearing chamber for all treatments. However, for one of the most temperature-sensitive taxa (*R. robusta*), it was not possible to begin the warmest treatments with five individuals/rearing chamber because the death rate was too high during temperature ramping. We therefore began the *R. robusta* experiment at the very beginning of temperature ramping, when all chambers had five individuals. For all other species, day one of the experiment began on the day following the acclimation period. We stopped each experiment the week before all treatments had < 50% survival or the week before < three treatments contained survivors, whichever came first.

We fed all species ad libitum conditioned (colonized and softened by microbes) maple leaves with Tetramin™ fish flakes added as a supplement throughout the duration of each experiment. Tetramin™ fish flakes contain crude protein (~46%), crude fat (11%), phosphorous (1%), ascorbic acid (~446 mg/kg), and omega-3 fatty acids (~500 mg/kg).

We measured mortality at weekly intervals and individual size at bi-weekly intervals. We measured mortality by visually inspecting each rearing chamber for dead individuals. We measured the total length of each surviving individual with a dissecting microscope to the nearest 0.5 mm. The precision of size measurements was constrained to 0.5 mm because the length of live individuals varies as individuals constrict and expand. We chose not to measure head capsule widths because they took longer to measure than body lengths of live individuals, and we wanted to minimize the effects of stress on growth and survivorship. We returned individuals to their

rearing chamber immediately following measurement. Individuals within each rearing chamber were not uniquely identified, thus growth and survival results are reported as the average of chamber means within each trough. In the case of survival, we report results as mean survival in each treatment, averaged across all weeks of the experiment (i.e., mean survival from week zero was calculated at each week of the experiment and then averaged across weeks). We calculated mean weekly survival to avoid reporting survival = 0, which would have regularly been the case if we had reported survival only at the last week of the experiment. We used length-dry mass regression equations (Benke et al. 1999) to estimate mass and then calculated growth as instantaneous growth rate (day^{-1}) per treatment calculated at the last week of each experiment as $\log_e$(mean final mass / mean initial mass) / (days of growth). Mean final and initial masses refer to the means of chamber means within a treatment, resulting in treatment-level mean instantaneous growth rates. Calculating instantaneous growth in this way allowed us to control for small differences in mean initial mass among treatments, and instantaneous growth is a common measure of growth in the literature (Crane et al. 2020).

Our original plan was to estimate UTLs based on survival, growth, and the product of growth and survival (hereafter referred to as the fitness index), which we considered to be a more integrated surrogate of fitness than either growth or survival alone. When able, we estimated UTLs from each of these responses by regressing their mean values against temperature. Given that both survival and growth decreased as temperature increased, we used regression models to estimate the UTLs as the temperatures at which each measure was extrapolated to be zero. For *D. coloradensis* and *R. robusta*, more than one of the warmest temperature treatments had complete mortality during the first week of the experiment. For these two taxa, we excluded all but the first zero mortality datapoint from the regression analysis. We report UTLs based on survival data for all seven species, but we were only able to obtain robust growth estimates for *P. californica*, partially because of the coarseness of the 0.5-mm resolution length measurements and partly because of the low to modest growth that occurred over 4 weeks – the length of time most

experiments lasted. We therefore report UTLs based on growth and the fitness index only for *P. californica*. Having estimates of UTLs based on growth and the fitness index for just *P. californica* eliminated our ability to robustly evaluate the comparability of UTL estimates derived from the three different measures, but we include the results for *P. californica* as an initial assessment of their comparability.

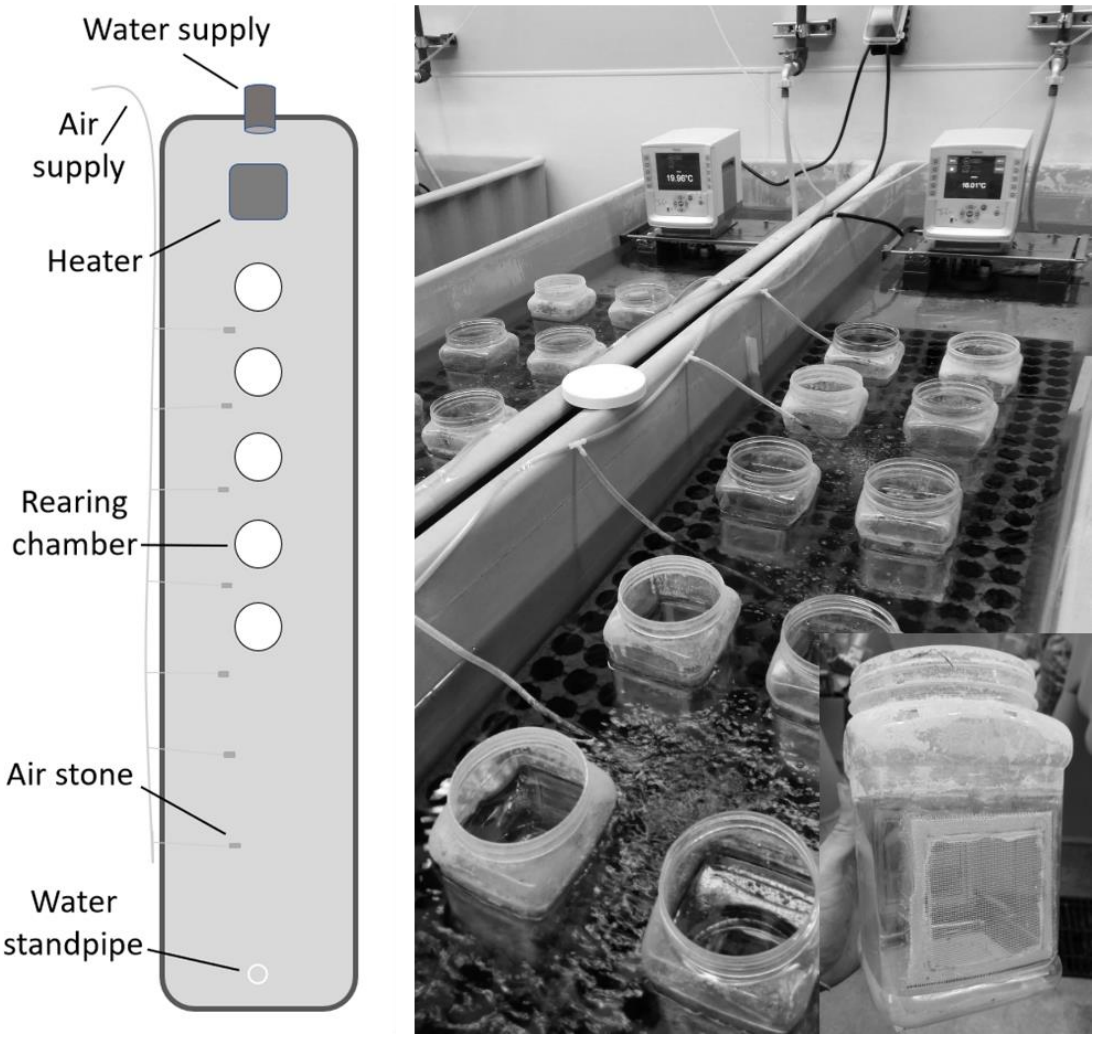


Figure 2-1. Aerial view schematic of the experimental flow-through troughs (left) and picture of the experimental troughs (right). The insert in the lower right-hand corner of the picture is a closeup of the mesh-sided and mesh-bottomed rearing chamber.

Field-derived UTLs

We used field-derived UTLs (Richards et al. 2013), estimated as the temperature below which 90 percent of presences for each species occurred (i.e., taxon-specific 90th percentile UTLs). The field data included macroinvertebrate occurrences and instantaneous temperature measurements collected between 1993 and 2010 for many locations across Idaho. We used field data from Idaho, even though our taxa were all collected from Utah because a similar database was not available for Utah, the Idaho dataset is large, and Utah and Idaho share many species in common. In addition, many of the Idaho taxa were identified to species including six of our seven experimental species. *Cinygmula sp.* was the only exception. The Idaho data used to estimate UTLs were collected between July 1 and September 30 of each year. Only macroinvertebrate taxa that occurred at ≥ 20 locations were used in their analysis.

Results

Laboratory-derived estimates of UTLs based on survival were highly variable among the seven species (range = 18 – 37 °C, Table 2-2). Mean weekly survival was negatively related to temperature for all seven species (r^2 = 0.56 – 0.92, Fig. 2-2). The duration of each experiment was also highly variable among the seven species and ranged from zero weeks (*R. robusta*) to 11 weeks (*H. azteca*). An experimental duration of zero weeks indicates the stopping criterion was met in the first week.

The UTL for *P. californica* inferred from three different performance metrics varied between 24 and 32 °C: 30 °C for survival, 32 °C for growth (Fig. 2-3), and 24 °C (Fig. 2-4) for the fitness index.

UTLs inferred from field data were also highly variable among the seven species. The 90th percentile UTLs ranged from 13 °C (*D. coloradensis*) to 23 °C (*H. azteca*) (Table 2-2).

UTLs inferred from the laboratory survivorship data and field data were strongly associated (r^2 = 0.72, Fig. 2-5).

Table 2-2. Field-derived and lab-derived upper thermal limits (UTLs) based on survivorship. Field-derived UTLs are 90th percentile temperatures from temperature-occurrence data collected across Idaho. Occurrences are the number of occurrences from which the field-UTL was derived. Lab-derived UTLs are from temperature-survival data measured in the laboratory.

Species	Field-UTL	Occurrences	Lab-UTL
Hyalella azteca	23.0	135	37
Pteronarcys californica	20.4	127	30
Drunella grandis	18.8	799	23
Gammarus lacustris	16.0	23	28
Cinygmula sp.	14.8	2807	21
Rhithrogena robusta	14.5	85	18
Drunella coloradensis	13.0	292	22

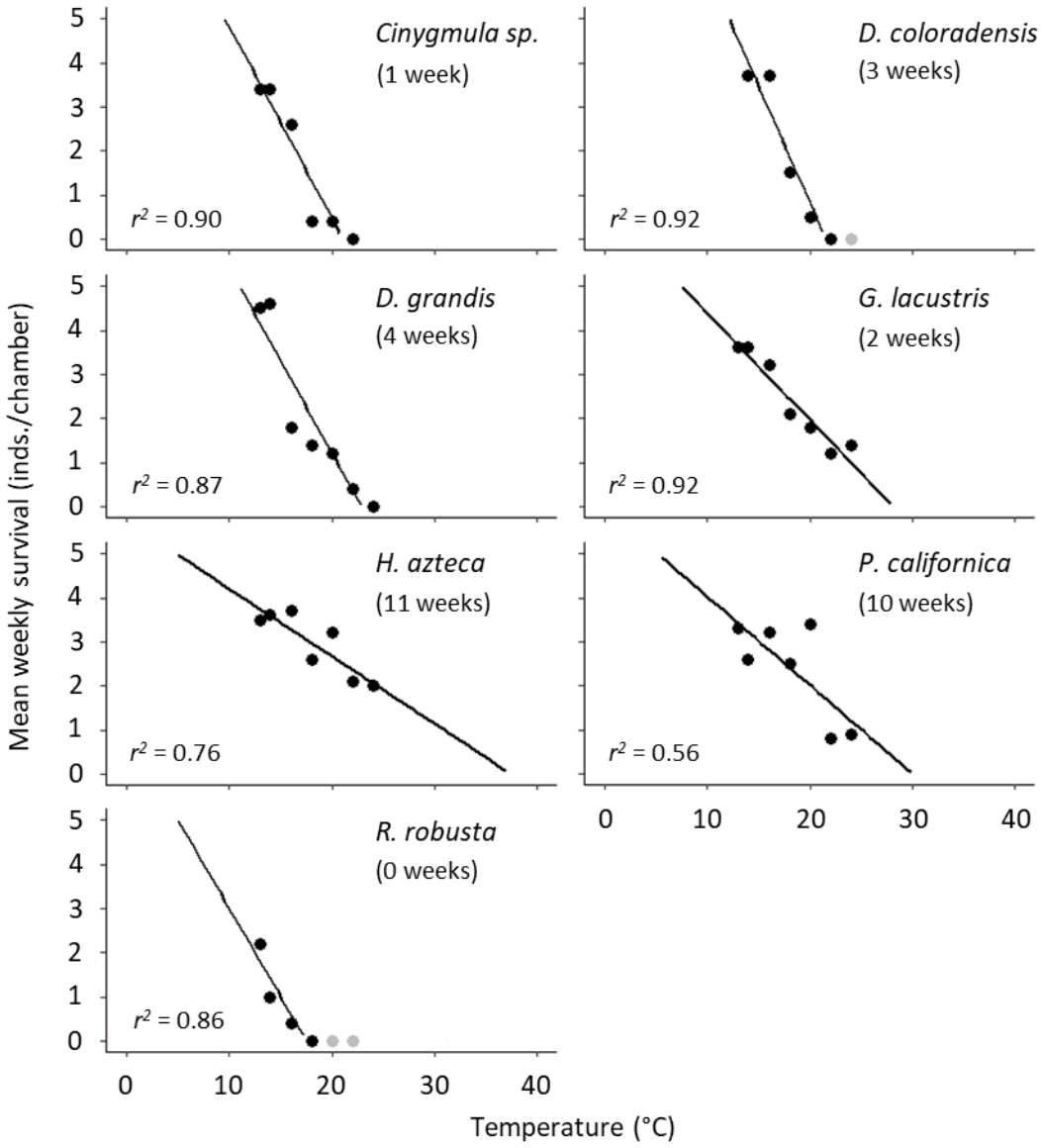


Figure 2-2. Mean weekly survival by treatment for the seven experimental species. The UTL for each species was inferred as the point where the best fit line intersected the x-axis. Average weekly survival was calculated over the duration (indicated in parentheses) of the experiment for each species. Grey points on the plots for *D. coloradensis* and *R. robusta* indicate treatments beyond the first instance where mean weekly survival was zero (i.e., complete mortality in the first week) and were not used to fit the best fit lines.

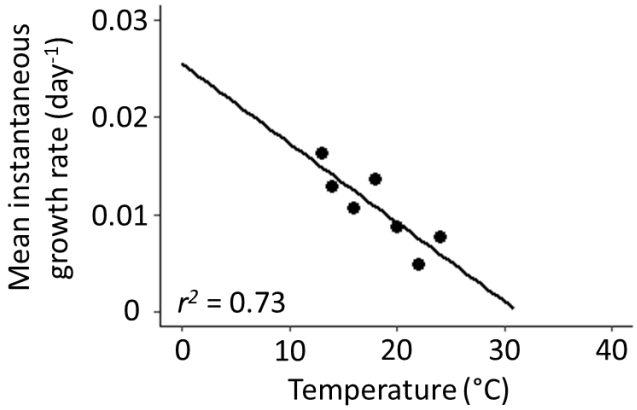


Figure 2-3. Variation across treatment temperatures in mean instantaneous growth rate (day^{-1}) of *P. californica* calculated at week 10 (last week of the experiment).

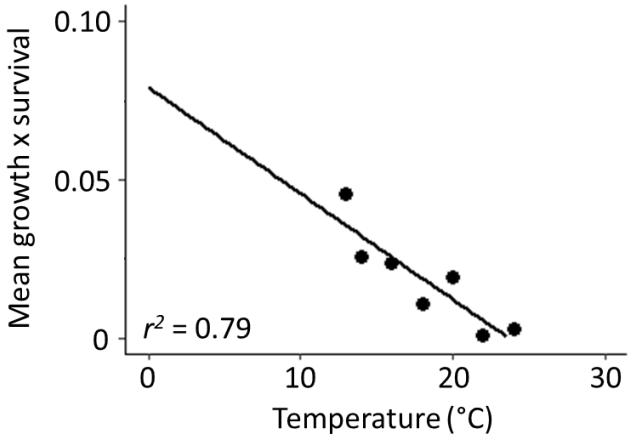


Figure 2-4. Variation across treatment temperatures in the fitness index (mean instantaneous growth rate (day^{-1}) × mean survival (average individuals/chamber) of *P. californica* calculated at week 10.

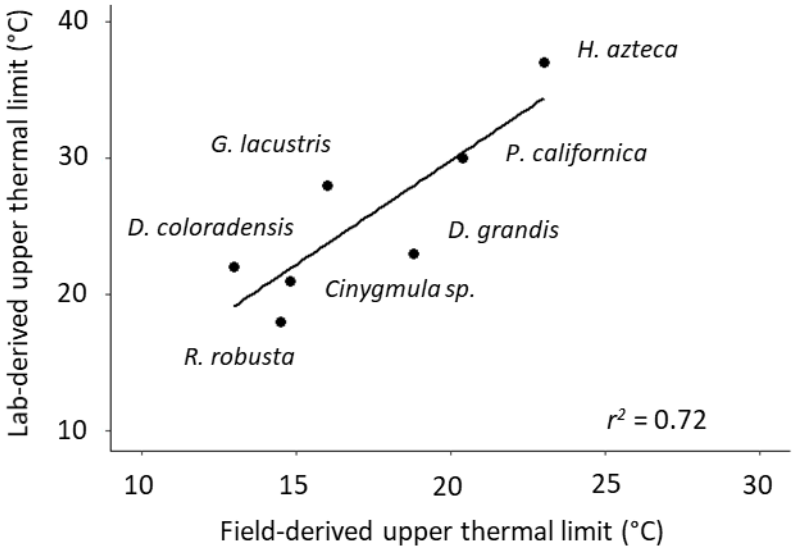


Figure 2-5. Association between laboratory-derived (based on survival) and field-derived (90th percentile) UTLs.

Discussion

Understanding the effects temperature has on the distributions of species is critical given the urgent need to manage and mitigate the effects that changing temperatures may have on aquatic life. Our NTCEE experiments increased our confidence that field-derived UTLs do describe meaningful differences among species in their thermal tolerances, hence they should be useful tools in informing conservation management and policy.

The vast majority of previous experiments that have examined temperature – distribution relationships were short-term, bioassay experiments that measure acute responses to either relatively rapid, and extreme, changes in temperature (CTMs) or temperature treatments that species would seldom, if ever, encounter in nature (many LC50 experiments). Unfortunately, the responses measured in these experiments do not always extrapolate well to longer-term processes of growth and survival (Kim et al. 2017). For example, the Oxygen- and Capacity-Limited Thermal Tolerance (OCLTT) hypothesis describes the incongruence between oxygen availability and demand with increasing temperature as the initial mechanism affecting the UTL of an

organism, and ultimately the distribution of its population (Pörtner 2010). The OCLTT hypothesis has been supported with acute exposure experiments with temperatures above those experienced in nature (Verberk and Calosi 2012), but it has not been supported with chronic exposure experiments within the range of temperatures experienced in nature (Kim et al. 2017, Sweeney et al. 2018, Funk et al. 2021). Thus, the OCLTT hypothesis may not provide a correct explanation of the mechanisms by which temperature affects distributions. Our study did not address the underlying physiological mechanisms by which temperature determines UTLs, but together with these other studies, it does support the need to focus on how temperatures routinely experienced in nature and over ecologically relevant timescales affect organism fitness and, ultimately, distributions. These longer-term experiments are more challenging and expensive to conduct than short-term bioassays, but they appear to more realistically characterize true differences among species in their thermal tolerances than short CTM and LC50 experiments. The thermal response experiments we conducted were longer than most experiments, but they were much shorter than the full lifecycle of any of the species we studied. For example, individuals from the population of *P. californica* we studied have three-year lifecycles, thus our 10-week experiment encompassed only a small fraction of the lifecycle. Nonetheless, the experiments were long enough to detect chronic effects of temperature exposure on survival of all species and effects on growth of *P. californica*.

Considering how temperature affects sublethal aspects of fitness should ultimately improve our ability to predict distributions from controlled experiments. The benefit of integrating responses of both growth and survival to different levels of total dissolved solids has already been shown to improve predictions of macroinvertebrate distributions (Olson and Hawkins 2017). Unfortunately, our assessment of the growth × survival fitness index was severely constrained because we could reliably estimate growth for only one species (*P. californica*). Directly weighing live individuals at the start and end of experiments will likely simultaneously minimize stress to experimental animals and allow more precise estimates of

growth. Direct measures of mass will also eliminate the need to use mass-length conversions to estimate growth which contain additional sources of error.

We used constant temperature treatments in this study but assessing how temperature fluctuations affect growth and survival should also be informative (Sweeney 1976). For example, Carlo et al. (2018) conducted experiments where they subjected lizard embryos to repeated sublethal warming. They found that more frequent sublethal warming (i.e., not causing acute stress) reduced embryo size and survival. When the survival results from repeated sublethal warming were incorporated into a species distribution model, they found the model predicted far lower survival than models that only accounted for effects of lethal temperatures. Thus, conducting temperature – fitness laboratory experiments with temperature regimes similar to those experienced in nature, and those forecasted to occur, could further improve predictions of species distributions.

Temperature – distribution associations are increasingly being used to model and predict the thermal niches of ectotherms. These predictions will play an important role in informing how to best conserve these species in an increasingly warm world. Having confidence that field-derived temperature – distribution associations have a mechanistic basis is essential to interpretation and proper application of these models. Chronic exposure experiments of modest length (1 – several weeks) may offer a good compromise between the inconsistent, and often misleading, characterizations of thermal tolerances derived from short-term experiments and the expense of full-lifecycle experiments for testing and validating field-derived temperature tolerances. The NTCEE-derived UTLs we observed based on survival alone improved our confidence that field-derived temperature – distribution associations have a casual basis. Incorporating additional sublethal metrics of fitness such as growth or fecundity into NTCEEs, and including assessments of the effects of fluctuating temperatures, should further improve predictions of how species distributions will change in response to thermal alterations.

References

Allen-Ankins, S., and R. J. Stoffels. 2017. Contrasting fundamental and realized niches: two fishes with similar thermal performance curves occupy different thermal habitats. Freshwater Science 36:635-652.

Ángeles-González, L. E., E. Martínez-Meyer, C. Yañez-Arenas, I. Velázquez-Abunader, A. Garcia-Rueda, F. Díaz, N. Tremblay, M. A. Flores-Rivero, P. Gebauer, and C. Rosas. 2020. Using realized thermal niche to validate thermal preferences from laboratory studies. How do they stand? Ecological Indicators 118:106741.

Angilletta Jr, M. J., T. D. Steury, and M. W. Sears. 2004. Temperature, growth rate, and body size in ectotherms: fitting pieces of a life-history puzzle. Integrative and Comparative Biology 44:498-509.

Benke, A. C., A. D. Huryn, L. A. Smock, and J. B. Wallace. 1999. Length-mass relationships for freshwater macroinvertebrates in North America with particular reference to the southeastern United States. Journal of the North American Benthological Society 18:308-343.

Bogert, C. M. 1949. Thermoregulation in reptiles, a factor in evolution. Evolution 3:195-211.

Braunisch, V., J. Coppes, R. Arlettaz, R. Suchant, H. Schmid, and K. Bollmann. 2013. Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change. Ecography 36:971-983.

Brown, J. H., J. F. Gillooly, A. P. Allen, V. M. Savage, and G. B. West. 2004. Toward a metabolic theory of ecology. Ecology 85:1771-1789.

Burgmer, T., H. Hillebrand, and M. Pfenninger. 2007. Effects of climate-driven temperature changes on the diversity of freshwater macroinvertebrates. Oecologia 151:93-103.

Carlo, M. A., E. A. Riddell, O. Levy, and M. W. Sears. 2018. Recurrent sublethal warming reduces embryonic survival, inhibits juvenile growth, and alters species distribution projections under climate change. Ecology Letters 21:104-116.

Chessman, B. C., and P. K. McEvoy. 1997. Towards diagnostic biotic indices for river macroinvertebrates. Hydrobiologia 364:169-182.

Chown, S. L., G. A. Duffy, and J. G. Sørensen. 2015. Upper thermal tolerance in aquatic insects. Current Opinion in Insect Science 11:78-83.

Crane, D. P., D. H. Ogle, and D. E. Shoup. 2020. Use and misuse of a common growth metric: guidance for appropriately calculating and reporting specific growth rate. Reviews in Aquaculture 12:1542-1547.

Crozier, L., and G. Dwyer. 2006. Combining population-dynamic and ecophysiological models to predict climate-induced insect range shifts. The American Naturalist 167:853-866.

Dallas, H. F., and Z. A. Ketley. 2011. Upper thermal limits of aquatic macroinvertebrates: comparing critical thermal maxima with 96-LT50 values. Journal of Thermal Biology 36:322-327.

Deutsch, C. A., J. J. Tewksbury, R. B. Huey, K. S. Sheldon, C. K. Ghalambor, D. C. Haak, and P. R. Martin. 2008. Impacts of climate warming on terrestrial ectotherms across latitude. Proceedings of the National Academy of Sciences 105:6668-6672.

Diamond, S. E., L. M. Nichols, N. McCoy, C. Hirsch, S. L. Pelini, N. J. Sanders, A. M. Ellison, N. J. Gotelli, and R. R. Dunn. 2012a. A physiological trait-based approach to predicting the responses of species to experimental climate warming. Ecology 93:2305-2312.

Diamond, S. E., D. M. Sorger, J. Hulcr, S. L. Pelini, I. D. Toro, C. Hirsch, E. Oberg, and R. R. Dunn. 2012b. Who likes it hot? A global analysis of the climatic, ecological, and evolutionary determinants of warming tolerance in ants. Global Change Biology 18:448-456.

Dormann, C. F. 2007. Promising the future? Global change projections of species distributions. Basic and Applied Ecology 8:387-397.

Dormann, C. F., S. J. Schymanski, J. Cabral, I. Chuine, C. Graham, F. Hartig, M. Kearney, X. Morin, C. Romermann, B. Schroder, A. Singer. 2012. Correlation and process in species distribution models: bridging a dichotomy. Journal of Biogeography 39:2119-2131.

Funk, D. H., B. W. Sweeney, and J. K. Jackson. 2021. Oxygen limitation fails to explain upper chronic thermal limits and the temperature size rule in mayflies. Journal of Experimental Biology 224:jeb233338.

Gillooly, J. F., J. H. Brown, G. B. West, V. M. Savage, and E. L. Charnov. 2001. Effects of size and temperature on metabolic rate. Science 293:2248-2251.

Guisan, A., R. Tingley, J. B. Baumgartner, I. Naujokaitis-Lewis, P. R. Sutcliffe, A. I. Tulloch, T. J. Regan, L. Brotons, E. McDonald-Madden, C. Mantyka-Pringle, T. G. Martin, J. R. Rhodes, R. Maggini, S. A. Setterfield, J. Elith, M. W. Schwartz, B. A. Wintle, O. Broennimann, M. Austin, S. Ferrier, M. R. Kearney, H. P. Possingham, and Y. M. Buckley. 2013. Predicting species distributions for conservation decisions. Ecology Letters 16:1424-1435.

Hawkins, C. P., Y. Cao, and B. Roper. 2010. Method of predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biology 55:1066-1085.

Hawkins, C. P., and L. L. Yuan. 2016. Multitaxon distribution models reveal severe alteration in the regional biodiversity of freshwater invertebrates. Freshwater Science 35:1365-1376.

Huey, R. B., C. A. Deutsch, J. J. Tewksbury, L. J. Vitt, P. E. Hertz, H. J. Álvarez Pérez, and T. Garland Jr. 2009. Why tropical forest lizards are vulnerable to climate warming. Proceedings of the Royal Society B: Biological Sciences 276:1939-1948.

Kearney, M., and W. P. Porter. 2004. Mapping the fundamental niche: physiology, climate, and the distribution of a nocturnal lizard. Ecology 85:3119-3131.

Kearney, M., and W. P. Porter. 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. Ecology Letters 12:334-350.

Kearney, M. R., B. A. Wintle, and W. P. Porter. 2010. Correlative and mechanistic models of species distribution provide congruent forecasts under climate change. Conservation Letters 3:203-213.

Kim, K. S., H. Chou, D. H. Funk, J. K. Jackson, B. W. Sweeney, and D. B. Buchwalter. 2017. Physiological responses to short-term thermal stress in mayfly (*Neocloeon triangulifer*) larvae in relation to upper thermal limits. Journal of Experimental Biology 220:2598-2605.

Moss, D., M. T. Furse, J. F. Wright, and P. D. Armitage. 1987. The prediction of the macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. Freshwater Biology 17:41-52.

Olson, J. R., and C. P. Hawkins. 2017. Effects of total dissolved solids on growth and mortality predict distributions of stream macroinvertebrates. Freshwater Biology 62:779-791.

Pacifici, M., W. B. Foden, P. Visconti, J. E. Watson, S. H. Butchart, K. M. Kovacs, B. R. Scheffers, D. G. Hole, T. G. Martin, H. R. Akçakaya, and R. T. Corlett. 2015. Assessing species vulnerability to climate change. Nature Climate Change 5:215-224.

Pearson, R. G., and T. P. Dawson. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology and Biogeography 12:361-371.

Poff, N. L., J. D. Olden, N. K. Vieira, D. S. Finn, M. P. Simmons, and B. C. Kondratieff. 2006. Functional trait niches of North American lotic insects: traits-based ecological applications in light of phylogenetic relationships. Journal of the North American Benthological Society 25:730-755.

Pörtner, H. O. 2010. Oxygen-and capacity-limitation of thermal tolerance: a matrix for integrating climate-related stressor effects in marine ecosystems. Journal of Experimental Biology 213:881-893.

Rendoll-Cárcamo, J., T. Contador, P. Convey, and J. Kennedy. 2020. Sub-Antarctic freshwater invertebrate thermal tolerances: an assessment of critical thermal limits and behavioral responses. Insects 11:102.

Rezende, E. L., F. Bozinovic, A. Szilágyi, and M. Santos. 2020. Predicting temperature mortality and selection in natural *Drosophila* populations. Science 369:1242-1245.

Richards, D. C., M. Bilger, and G. Lester. 2013. Development of Idaho Macroinvertebrate Temperature Occurrence Models: Final Report. Available from: https://www.ecoanalysts.com/development-of-idaho-macroinvertebrae-temperature-occurance-models

Shah, A. A., B. A. Gill, A. C. Encalada, A. S. Flecker, W. C. Funk, J. M. Guayasamin, B. C. Kondratieff, N. L. Poff, S. A. Thomas, K. R. Zamudio, and C. K. Ghalambor. 2017. Climate variability predicts thermal limits of aquatic insects across elevation and latitude. Functional Ecology 31:2118-2127.

Sofaer, H. R., C. S. Jarnevich, I. S. Pearse, R. L. Smyth, S. Auer, G. L. Cook, T. C. Edwards, G. F. Guala, T. G. Howard, J. T. Morisette, and H. Hamilton. 2019. Development and delivery of species distribution models to inform decision-making. BioScience 69:544-557.

Sokolovska, I. 2014. Are experimentally derived estimates of thermal tolerance useful in interpreting species distribution models. All Graduate Theses and Dissertations. 3695. Available from: https://digitalcommons.usu.edu/etd/3695

Sweeney, B. W. 1976. A diurnally fluctuating thermal system for studying the effect of temperature on aquatic organisms. Limnology and Oceanography 21:758-763.

Sweeney, B. W., D. H. Funk, A. A. Camp, D. B. Buchwalter, and J. K. Jackson. 2018. Why adult mayflies of *Cloeon dipterum* (Ephemeroptera: Baetidae) become smaller as temperature warms. Freshwater Science 37:64-81.

Sweeney, B. W., and R. L. Vannote. 1978. Size variation and the distribution of hemimetabolous aquatic insects: two thermal equilibrium hypotheses. Science 200:444-446.

Vannote, R. L., and B. W. Sweeney. 1980. Geographic analysis of thermal equilibria: a conceptual model for evaluating the effect of natural and modified thermal regimes on aquatic insect communities. The American Naturalist 115:667-695.

Verberk, W. C. E. P., and P. Calosi. 2012. Oxygen limits heat tolerance and drives heat hardening in the aquatic nymphs of the gill breathing damselfly *Calopteryx virgo* (Linnaeus, 1758). Journal of Thermal Biology 37:224-229.

Verberk, W. C. E. P., H. Siepel, and H. Esselink. 2008. Life-history strategies in freshwater macroinvertebrates. Freshwater Biology 53:1722-1738.

Wright, J. F. 1995. Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. Australian Journal of Ecology 20:181-197.

Yuan, L. L. 2006. Estimation and application of macroinvertebrate tolerance values. US EPA, ORD, National Center for Environmental Assessment, Washington, DC. Technical Report (EPA/600/P-04/116F).

CHAPTER 3

CORRECTING FOR THE EFFECTS OF CLASS IMBALANCE IMPROVES THE PERFORMANCE OF MACHINE-LEARNING BASED SPECIES DISTRIBUTION MODELS*

Abstract

Numerous methods have been developed to combat the unwanted effects of imbalanced training data on the performance of machine-learning based predictive models. These methods attempt to balance the tradeoff between sensitivity and specificity. However, the effects of specific imbalance-correction methods on the performance of different machine-learning algorithms are not well understood for ecological data. In this study, we used four machine-learning algorithms (random forest, artificial neural network, gradient boosting, support vector machine) and five imbalance-correction methods (base algorithm, cutoff, up-sampling, down-sampling, weighting) to produce species distribution models for 15 freshwater macroinvertebrate genera that varied from 2.5 – 29.0% in prevalence. All imbalance-correction methods substantially improved average model performance (true skill statistic) over the base machine-learning algorithms, except when up-sampling was applied to random forest models. Choice of machine-learning algorithm had little effect on model performance, although gradient boosting performed best when modeling taxa with the most imbalanced datasets. Our results suggest that the performance of species distribution models built with presence/absence data can generally be improved by correcting for imbalanced data.

1. Introduction

Species distribution models (SDMs) based on binary (presence/absence) data are frequently used to predict how probabilities of occurrence of species vary across environmental

* Coauthored by Samuel D. Schwartz, D. Richard Cutler, and Charles P. Hawkins

gradients (Elith and Leathwick, 2009). These SDMs are commonly built with machine-learning algorithms such as random forest, artificial neural network (ANN), gradient boosting, support vector machine (SVM), and maximum entropy models (Cutler et al., 2007; De'Ath, 2007; Olden et al., 2008; Hoang et al., 2010; Sor et al., 2017; Gobeyn et al., 2019). Machine-learning models are increasingly popular in ecology because of their ability to identify complex, nonlinear associations between response and predictors (Lek and Guégan, 1999; Breiman, 2001; Cutler et al., 2007; Hawkins et al., 2010).

Unfortunately, machine-learning algorithms often perform poorly when class occurrences are highly imbalanced, as is often the case for ecological datasets (Johnson et al., 2012). For example, species presences (positive class) usually are less frequent than species absences (negative class), which presents challenges when optimizing a machine-learning model. Optimizing a machine-learning model means finding the model parameters that best map input to expected output by minimizing the cost function of the algorithm and, thus, overall error rate of the model. Given a highly imbalanced binary dataset, a low overall error rate is easily achieved by consistently predicting the majority class, resulting in high specificity but low sensitivity (Akbani et al., 2004; Chen et al., 2004). However, ecologists are often more interested in correctly classifying where species are, thus we need methods that can better balance model sensitivity and specificity.

Data scientists have developed methods to improve classification performance of imbalanced data. These methods have been applied to datasets related to medical testing, financial management, and fraud and disaster detection (Chen et al., 2004; Haixiang et al., 2017). At least four methods have been proposed to deal with the class imbalance problem. However, as noted by Johnson et al. (2012), these methods have seldom been applied to species distribution modeling. Simple implementations of these methods are now generally supported in popular data science programming languages (e.g., R and Python), which should facilitate their use among ecologists.

A common approach to compensate for imbalanced data is to balance the data by up-sampling the minority class or down-sampling the majority class. This approach does not add any new information but does increase the weight of the minority class by balancing the class distribution (McCarthy et al., 2005). In the case of down-sampling, some information is lost because instances of the majority class are excluded from the analysis, whereas up-sampling can lead to overfitting and can be computationally expensive, sometimes to the point of being infeasible. (Chawla et al., 2004). Despite these shortcomings, up- and down-sampling have been shown to improve model performance in numerous studies (Chawla et al., 2004; McCarthy et al., 2005; Yap et al., 2014; Buda et al., 2018).

Another common approach is to directly apply class weights to the algorithm so the cost of misclassifying the minority class is elevated relative to the cost of misclassifying the majority class. Since machine-learning algorithms are designed to minimize some internal cost function, it is clear that increasing the relative cost of misclassifying minority class samples will cause the model to place higher weight on correctly classifying those samples, thus increasing model sensitivity but decreasing specificity. The weighting factor that determines the relative cost is considered a parameter to be tuned during model training, and an intuitive initial weighting factor of the minority class can be calculated from the ratio of samples in each of the classes (Chen et al., 2004). For example, if a dataset contains 100 presences and 400 absences, class weights would be assigned as [4, 1] for presences and absences, respectively. The model then incurs a 4× cost for the misclassification of a minority class sample relative to a majority class sample. This reweighting of classes has been shown to improve the tradeoff between model sensitivity and specificity (Chen et al., 2004; Hwang et al., 2011). Specifically, the tradeoff in model performance is that increases in sensitivity are often accompanied to some degree by decreases in specificity and overall model accuracy.

Finally, the predicted probability of occurrence threshold or cutoff value that implies presence can be optimized to balance the tradeoff between sensitivity and specificity (Greiner et

al., 2000; Freeman and Moisen, 2008; Freeman et al., 2012). For example, the default cutoff value is generally 0.5, where a value ≥ 0.5 implies presence and a value < 0.5 implies absence. For highly imbalanced species datasets, the optimal cutoff value is often close to the prevalence of the minority class, which is an easily calculated cutoff criterion (Liu et al., 2005; Freeman and Moisen, 2008). For example, a species with 10% prevalence would be predicted to occur at a site if the model predicted a 10% or higher probability of occurrence.

Imbalanced datasets are common in ecology, but few systematic studies have compared the performance of imbalance-correction methods, especially across different machine-learning algorithms. Freshwater macroinvertebrate distribution data are ideal for studying the effects of imbalance on model performance because they are readily available for a wide range of species that vary greatly in prevalence (e.g., rare species to common species). Additionally, the use of machine-learning algorithms to model the distributions of macroinvertebrates and other animals is increasingly common – e.g., see Dedecker et al. (2002, 2005), Goethals et al. (2003, 2007), Lin et al. (2016), Rocha et al. (2017), and Muñoz-Mas et al. (2019) for ANN; Kubosova et al. (2010) and Olaya-Marín et al. (2013) for random forest; Hoang et al. (2010) for SVM; and Maloney et al. (2012) for gradient boosting. In this study, we conducted a systematic comparison of several methods used to adjust for imbalanced data when modeling with random forest, ANN, gradient boosting, and SVM. We address two primary research questions: 1) What machine-learning algorithms and imbalance-correction methods perform best with imbalanced macroinvertebrate data? 2) Does the performance of machine-learning algorithms and imbalance-correction methods depend on prevalence?

2. Methods

2.1 Dataset

We used data from the National Rivers and Streams Assessment (NRSA) in this study (USEPA, 2016). This dataset contains presence/absence information on hundreds of

macroinvertebrate taxa found across the United States. We used data from 1,950 unique sites that were collected in 2008 and 2009 and between day 111 and 334 of each year, and for which associated data on 11 environmental predictors were available (Table 3-1). These 11 environmental variables are often associated with distributions of freshwater invertebrates (Moss et al., 1987; Vinson and Hawkins, 1998; Clarke et al., 2003; Berger et al., 2017). All predictors were normalized with a z-score transformation. Normalizing predictors can improve performance of certain machine-learning algorithms such as ANN, for which the training process is sensitive to differences in scaling among predictors (Olden and Jackson, 2002). To address our research questions, we selected 15 taxa that varied in prevalence from 2.4 to 29.4% (Table 3-2). These 15 taxa spanned 3 prevalence categories of 5 taxa each: < 10%, 10 – 20%, and 20 – 30%. We performed all analyses with R version 3.6.1 (R Core Team, 2019, Vienna, Austria).

Table 3-1 Variables included as predictors in the SDMs.

Predictor	Description
MSST	Predicted mean summer stream temperature (ºC)
Substrate	Log_{10} geometric mean substrate particle diameter (mm)
DOY	Day of the year sample was collected
Conductivity	Specific conductance (µS/cm)
ANC	Acid neutralizing capacity (µeq/L)
CA	Calcium (mg/L)
CL	Chloride (mg/L)
K	Potassium (mg/L)
MG	Magnesium (mg/L)
NA	Sodium (mg/L)
SO4	Sulfate (mg/L)

Table 3-2 The 15 macroinvertebrate genera for which SDMs were built from 1,950 sites.

Genus	Family	Order	Number of presences	Prevalence (%)
Malenka	Nemouridae	Plecoptera	49	2.5
Pteronarcys	Pteronarcyidae	Plecoptera	75	3.8
Zapada	Nemouridae	Plecoptera	92	4.7
Drunella	Ephemerellidae	Ephemeroptera	157	8.1
Callibaetis	Baetidae	Ephemeroptera	180	9.2
Rhyacophila	Rhyacophilidae	Trichoptera	223	11.4
Stenacron	Heptageniidae	Ephemeroptera	233	11.9

Sialis	Sialidae	Megaloptera	300	15.4
Gammarus	Gammaridae	Amphipoda	333	17.1
Argia	Coenagrionidae	Odonata	376	19.3
Hemerodromia	Empididae	Diptera	414	21.2
Optioservus	Elmidae	Coleoptera	471	24.2
Paratanytarsus	Chironomidae	Diptera	524	26.9
Hydroptila	Hydroptilidae	Trichoptera	540	27.7
Centroptilum/Procloeon	Baetidae	Ephemeroptera	565	29.0

2.2 Machine-learning algorithms and imbalance-correction methods

Each of the machine-learning algorithms used in this study has been shown to perform relatively well with at least one imbalanced dataset. Random forest classifies by constructing many individual classification trees (a forest) and uses this forest of trees to make a final class prediction based on the mode of class predictions from the individual trees (Breiman, 2001). Random forest has performed well on imbalanced data (Khalilia et al., 2011). We used version 4.6-14 of the *randomForest* package for random forest implementations (Liaw and Wiener, 2002). ANN is a nonlinear network structure that maps input to expected output (Lek and Guegan, 1999). In some studies, ANN has outperformed some other machine-learning algorithms on highly imbalanced data (e.g., Sor et al., 2017). We used version 7.3-15 of the *nnet* package for ANN implementations (Venables and Ripley, 2002). The gradient boosting algorithm produces an additive model by sequentially constructing an ensemble of weak learners that place higher weight on previously misclassified instances as the sequence progresses (Friedman 2001; De'Ath, 2007). Gradient boosting has also performed well on imbalanced classification tasks relative to other machine-learning algorithms (Moisen et al., 2006; Brown and Mues, 2012). We used version 2.1.8 of the *gbm* package for gradient boosting implementations (Greenwell et al., 2019). The SVM algorithm (Cortes and Vapnik, 1995) classifies by projecting the data space into a higher dimensional feature space and constructing a hyperplane in feature space that maximizes class separation. SVM has outperformed other classifiers on moderately imbalanced data for some datasets (Tang et al., 2009). We used version 1.7-6 of the *e1071* package for SVM implementations (Meyer et al., 2019).

The methods for dealing with imbalanced data considered here have numerous optimization criteria. For example, Freeman and Moisen (2008) compared 11 different cutoff optimization criteria applied to SDMs and noted that the optimal criterion is dependent on the intended research use of the SDM. Specifically, certain criteria were better choices if ecologists wanted to avoid over representing predicted presences (i.e., high false positives), whereas other criteria were better choices if balancing model sensitivity and specificity was important. We used core implementations of each method that are easily coded and interpreted (Table 3-3). To this end, the cutoff and weight criteria were based on the prevalence of each genus. Specifically, the threshold criterion chosen for each taxon was equal to the observed prevalence (Freeman and Moisen, 2008), and the weights were calculated from the ratio of samples in each class. We did not apply weighting to random forest because to our knowledge no reliable implementations were available for our selected package or any other R package. Initially, we implemented the *classwt* argument, however, we discovered this argument was broken at the time of implementation. Up-sampling was applied by replicating the minority class to match the number in the majority class, and down-sampling was applied by randomly selecting a subset of the majority class to match the number in the minority class. Up- and down-sampling implementations were done with version 6.0-86 of the *caret* package (Kuhn et al., 2019), except in the case of down-sampling with random forest, which had a built-in implementation that we used (Table 3-3).

Table 3-3 Methods and associated R packages, functions, and arguments for dealing with imbalanced data for each machine-learning algorithm compared in this study. Base refers to the machine-learning algorithm without any additional methods applied to deal with class imbalance. Also note all arguments listed in the table exist and are called within the base function of each machine-learning algorithm. Manual indicates that the implementation was coded manually. pkg = package.

Machine-learning algorithm	Imbalanced data method				
	Base	Up-sample	Down-sample	Cutoff	Weighting
Random forest	randomForest pkg and function	Caret pkg upSample function	Strata argument	cutoff argument	none available

ANN	nnet pkg and function	Caret pkg upSample function	Caret pkg downSample function	Manual	weights argument
Gradient boosting	gbm pkg and function	Caret pkg upSample function.	Caret pkg downSample function	Manual	weights argument
SVM	e1071 pkg svm function	Caret pkg upSample function	Caret pkg downSample function	Manual	class.weights argument

2.3 Model optimization and validation

To optimize each model, we applied a large hyperparameter grid search (see Appendix B for ranges and optimized hyperparameters for each model). Models were run on the University of Oregon supercomputer. A large grid search is computationally expensive, but it helps ensure that models are highly optimized in a standardized and reproducible manner, which was critical to the objectives of this study. In total, 285 models were optimized: (15 genera × 4 machine-learning algorithms × 5 imbalance-correction methods) minus the 15 weighted random forest models.

We validated models with stratified 5-fold cross validation. This validation procedure is commonly used for small imbalanced datasets (Johnson et al., 2012). Specifically, 5 randomly stratified 70/30 train/test split datasets were created for each species (data were shuffled after the creation of each train/test split). The stratification was done to preserve the ratio of presences to absences in the train and test datasets. Each model was run on these 5 train/test species datasets and the results averaged over the 5 runs to return the 5-fold cross validation metrics.

We used the true skill statistic (TSS) to identify optimal models and report results. The TSS is a performance metric that places equal weight on sensitivity and specificity, making it a good choice for evaluating responses of SDMs to imbalanced data (Allouche et al., 2006; Akosa, 2017). TSS is also well suited for comparing model performance across species that vary in prevalence (Freeman and Moisen, 2008). The formula for calculating the TSS is sensitivity + specificity – 1, and a model with no misclassified instances will produce a TSS of 1. In several places (Table 3-4 and Appendix C), we also report area under the receiver operating characteristic

curve (AUROC), percent classified correctly (PCC), and kappa because these metrics are common in the SDM literature (Johnson et al., 2012). However, caution is needed when interpreting PCC and kappa of models built with imbalanced datasets as their values are not necessarily independent of prevalence (McPherson et al., 2004; Vaughan and Ormerod, 2005; Allouche et al., 2006). For SVMs, we do not report the AUROC because preliminary analyses showed that classifications based on estimated probabilities (with decision threshold of 0.5) did not always match classifications based on decision values. Additionally, classifications based on decision values yielded higher TSSs for several of the models, so we decided to only use decision values with SVM. Thus, without consistent probability estimates, we could not calculate measures of AUROC comparable to those calculated for the other three machine-learning algorithms. Applying a modified cutoff to SVM was the only exception where we did use estimated probabilities, because this method relies on the use of class probabilities to make a final classification.

We evaluated and compared the effects of machine-learning algorithm and imbalance-correction methods on model performance in three ways. First, we assessed performance as the average model performance across all 15 genera (we also calculated standard errors as measures of consistency in model output). Second, we assessed how performance varied across the three prevalence groups as the average performance (with standard error) across the five genera within each group. Third, we used a linear model (*lm* function) to partition the variability in TSS values associated with machine-learning algorithm, imbalance-correction method, and prevalence. To further address our second research question regarding how performance of machine-learning algorithms and imbalance-correction methods varies with prevalence, potential interactions were also included in the linear model. Specifically, these interactions included imbalance-correction methods:prevalence and machine-learning algorithm:prevalence.

3. Results

3.1 Average model performance across all 15 genera

Base machine-learning algorithms

Overall model performance, averaged across the 15 taxa, was not strongly influenced by the particular base machine-learning algorithm (Table 3-4, see Appendix C for the performance metrics for each of the 285 individually optimized models) as evident by the slightly overlapping standard errors for model performance of all four machine-learning algorithms. However, gradient boosting appeared to perform slightly better than other base machine-learning algorithms across all 15 taxa (mean TSS = 0.34). ANN followed closely in average performance (mean TSS = 0.32). SVM and random forest were the two lowest performing base machine-learning algorithms on average (mean TSS = 0.26 and 0.23, respectively). The order of performance was similar based on kappa (Table 3-4). However, PCC ranked performance in the opposite order of the other performance metrics, and the threshold-independent metric, AUROC, was highest for base random forest (Table 3-4).

Imbalance-correction methods

Applying imbalance-correction methods to the machine-learning algorithms substantially increased performance over the base machine-learning algorithms (Table 3-4). The average performance improvement (TSS) of each machine-learning algorithm model built with imbalance-correction methods was ≥ 30%, ≥ 47%, ≥ 41%, ≥ 69% compared with models built with base random forest, base ANN, base gradient boosting, and base SVM, respectively (Table 3-4). Gradient boosting produced a mean TSS value of at least 0.48 with each of the four imbalance-correction methods. Random forest with down-sampling also produced a mean TSS of 0.48 and was followed closely by random forest with adjusted class prediction cutoff (mean TSS = 0.47). Mean TSS values for ANN and SVM based on all imbalance-correction methods were similar (≥ 0.47 and ≥ 0.44, respectively). Random forest with up-sampling was the only

noticeably underperforming model with a mean TSS value of 0.30, which was within the standard error of base random forest.

Table 3-4 Means and standard errors (SE) of TSS, AUROC, Kappa, and PCC across all 15 taxa for each imbalance-correction method × machine-learning algorithm. TSS = true skill statistic, AUROC = area under the receiver operating characteristic curve, PCC = percent classified correctly.

	Machine-learning algorithm							
Imbalance-	Random forest		ANN		Gradient boosting		SVM	
correction method	Mean	SE	Mean	SE	Mean	SE	Mean	SE
				TSS				
Base	0.23	0.05	0.32	0.05	0.34	0.05	0.26	0.05
Cutoff	0.47	0.06	0.48	0.06	0.48	0.06	0.44	0.06
Down-sample	0.48	0.06	0.47	0.06	0.49	0.06	0.48	0.06
Up-sample	0.30	0.05	0.48	0.06	0.49	0.06	0.48	0.06
Weighted	-	-	0.49	0.06	0.49	0.06	0.49	0.06
				AUROC				
Base	0.77	0.03	0.72	0.03	0.72	0.03	-	-
Cutoff	0.78	0.03	0.79	0.03	0.79	0.03	-	-
Down-sample	0.80	0.03	0.78	0.03	0.79	0.03	-	-
Up-sample	0.77	0.03	0.79	0.03	0.79	0.03	-	-
Weighted	-	-	0.79	0.03	0.79	0.03	-	-
				Kappa				
Base	0.27	0.05	0.30	0.04	0.31	0.05	0.28	0.05
Cutoff	0.27	0.04	0.29	0.04	0.32	0.04	0.27	0.04
Down-sample	0.33	0.04	0.27	0.04	0.29	0.04	0.27	0.03
Up-sample	0.32	0.05	0.29	0.04	0.31	0.04	0.28	0.04
Weighted	-	-	0.29	0.04	0.32	0.04	0.29	0.04
				PCC				
Base	85.7	2.3	82.3	2.5	80.9	2.4	83.9	2.6
Cutoff	70.3	2.9	72.3	2.8	75.1	3.1	72.7	2.9
Down-sample	78.3	2.3	70.4	2.9	72.3	2.9	71.7	2.7
Up-sample	84.4	2.5	72.9	2.8	75.2	2.6	71.7	2.8
Weighted	-	-	72.7	2.7	75.8	2.7	72.1	3.0

3.2 Average model performance across prevalence groups

Base machine-learning algorithms

Prevalence had little effect on which base machine-learning algorithms performed best, but the prevalence range did affect how variable performance was among machine-learning

algorithms (Fig. 3-1). Gradient boosting was the top average base-model performer for taxa in the 0 – 10% prevalence range (mean TSS = 0.47, Fig. 3-1A) and in the 20 – 30% prevalence range (mean TSS = 0.26, Fig. 3-1C). Taxa in the 10 – 20% prevalence range were best classified with base ANN (mean TSS = 0.31, Fig. 3-1B). Random forest and SVM were consistently the poorest performing base machine-learning algorithms across the three prevalence ranges, but model underperformance was very small for the higher prevalence range (Fig. 3-1C). Indeed, the range in mean model performance among the four different base machine-learning algorithms also decreased consistently as prevalence increased. Specifically, the ranges in base machine-learning algorithm performance were TSS of 0.24 – 0.47 for the 0 – 10% prevalence taxa (Fig. 3-1A), 0.24 – 0.31 for the 10 – 20% prevalence taxa (Fig. 3-1B), and 0.21 – 0.26 for the 20 – 30% prevalence taxa (Fig. 3-1C). Average model performance, calculated across three base machine-learning algorithms (RF, ANN, and GBM), also tended to decrease as prevalence increased (Fig. 3-2). The trend was most pronounced with performance measured as PCC.

Imbalance-correction methods

Prevalence had no observable effect on which imbalance-correction method best improved model performance but did affect the degree to which model performance improved by applying imbalance-correction methods (Fig. 3-1). On average, the highest performing models for taxa in the 0 – 10% prevalence range were ANN with cutoff and ANN with weighting (mean TSSs of 0.67; Fig. 3-1A). However, every combination of machine-learning algorithm and imbalance-correction method had a mean TSS of ≥ 0.63 and overlapping standard errors, except for random forest with up-sampling (mean TSS of 0.32). Gradient boosting with up-sampling had the highest performance for the 10 – 20% prevalence range with mean TSS of 0.47 (Fig. 3-1B), followed closely by every other combination with mean TSS ≥ 0.40, except for random forest with up-sampling (mean TSS of 0.32). Gradient boosting with down-sampling, cutoff, and weighting performed best (TSSs of 0.36) for taxa in the 20 – 30% prevalence range (Fig. 3-1C)

and every other combination of machine-learning by imbalance-correction method produced a mean TSS ≥ 0.26. The increase in mean performance between the base machine-learning algorithms and the models with imbalance-correction methods was generally higher for genera with lower prevalence (Fig. 3-1).

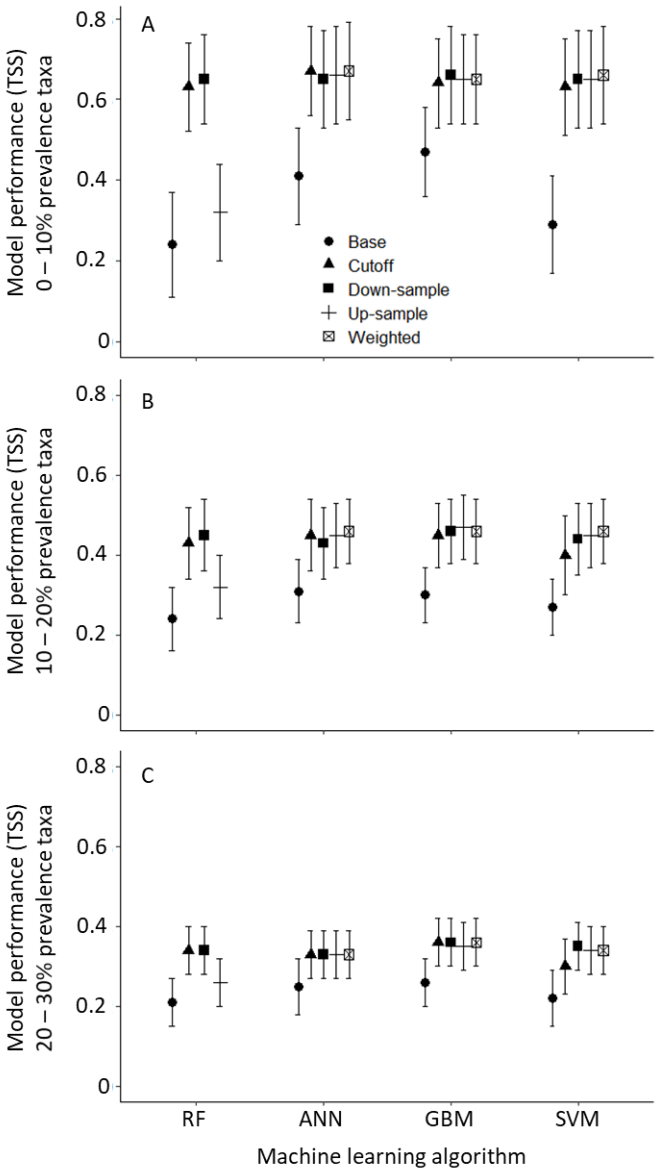


Fig. 3-1. Means ± standard errors of model performance (TSS) for each machine-learning algorithm and imbalance-correction methods and calculated across the five taxa in each prevalence range. A) taxa in the 0 – 10% prevalence range. B) Taxa in the 10 – 20% prevalence range. C) Taxa in the 20 – 30% prevalence range. Data are jittered for discernibility. RF = random forest, ANN = artificial neural network, GBM = gradient boosting machine, and SVM = support vector machine.

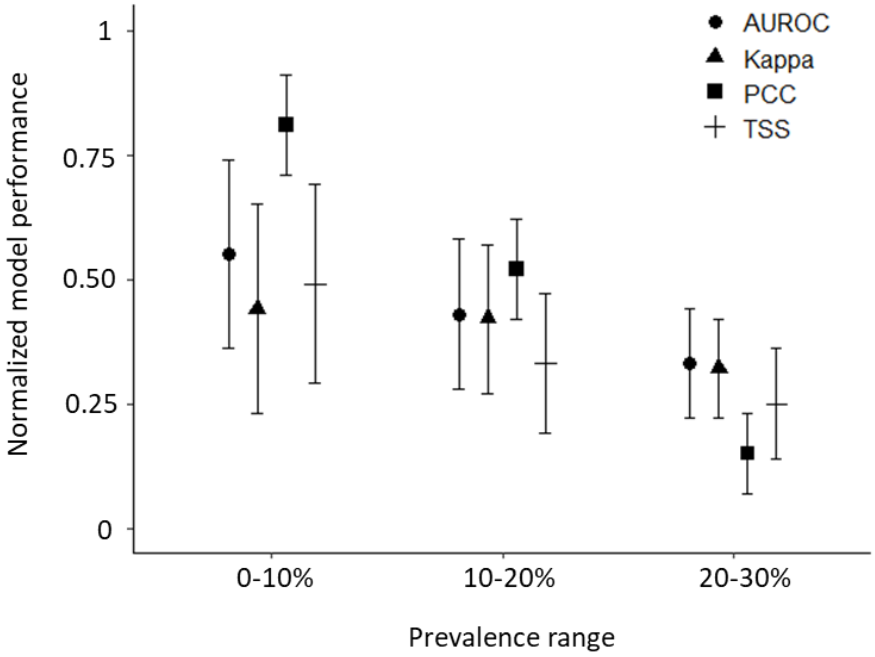


Fig. 3-2. Means and standard errors of the normalized performance metrics averaged across three base machine-learning algorithms (RF, ANN, GBM) and the five genera in each of the three prevalence ranges. Data are jittered for discernibility. AUROC = area under receiver operating characteristic curve, TSS = true skill statistic, PCC = percent correctly classified.

3.3 Parsing effects of machine-learning algorithm, imbalance-correction method, and prevalence on performance

The linear model showed that machine-learning algorithm, imbalance-correction method, and prevalence varied in their effect on model performance (Table 3-5). Prevalence affected variation in TSS the most (~28% of variation), then imbalance-correction methods (~8%), and lastly machine-learning algorithm (~2%). Pairwise interactions of imbalance-correction methods and machine-learning algorithm with prevalence had small effects on TSS. Notably, over 60% of the variation in TSS was not associated with these three factors.

Table 3-5 The amount of variation in TSS associated with imbalance-correction method, machine-learning algorithm, prevalence, and their pairwise interactions. Sum of squares are type III.

Source of variation	Sum of sq.	Df	Mean sq.	F-Ratio	% of var.
Imbalance-correction method	1.22	4	0.31	9.47	8.36
Machine-learning algorithm	0.24	3	0.08	2.51	1.64
Prevalence	4.02	1	4.02	124.57	27.55
Imbalance-correction method:Prevalence	0.33	4	0.08	2.59	2.26
Machine-learning algorithm:Prevalence	0.09	3	0.03	0.92	0.62
Error	8.69	269	0.03		

4. Discussion

4.1 Overall model comparison

We found that all four imbalance-correction methods broadly improved model performance, and therefore, should be considered by ecologists building species distribution models with machine-learning algorithms. Indeed, machine-learning based species distribution models are being built and applied more frequently than ever before (Sor et al., 2017, Gobeyn et al., 2019, da Silveira et al., 2021, Weinert et al. 2021). These models are used to address a variety of research questions about the effectiveness of conservation efforts and to inform decision-making (da Silveira et al., 2021, Weinert et al., 2021). For example, decisions related to the protection of rare or endangered species, often via identification and protection of their preferred habitats, are increasingly being informed and evaluated by species distribution models (da Silveira et al., 2021, Weinert et al., 2021). Whether or not conservation decisions informed by species distribution models built with imbalance-correction methods differ from those built without imbalance-correction methods remains to be seen. Areas of research, such as bioassessment, that directly compare predicted probabilities of occurrence of species across a diverse community, and across an inherently wide prevalence range (i.e., rare species to common species), may also be affected by the addition or omission of imbalance-correction methods in the model training process. For example, it is not yet clear what effect imbalance correction has on estimates of assessment endpoints, some of which compare observed taxa richness with expected

taxa richness, calculated as the sum of predicted probabilities of occurrence (REFS). Below, we discuss in more detail some of the advantages and disadvantages of building SDMs with the different imbalance-correction methods and machine-learning algorithms. We also discuss some of the observed effects of species prevalence on model performance.

Imbalance-correction methods advantages and disadvantages

There are advantages and disadvantages to several of the imbalance-correction methods we examined with respect to implementation and computational overhead. Up-sampling has a disadvantage of increased computational overhead during model optimization because of the larger resulting dataset, which can be an important consideration when dealing with large datasets and limitations in computing resources (Chawla et al., 2004). In contrast to up-sampling, down-sampling has an advantage of being extremely computationally efficient, which was evident in the comparatively short model optimization times we observed. Down-sampling also consistently performed well across machine-learning algorithms, making it a top candidate method for ecologists modeling imbalanced data. Up-sampling did not perform as consistently as other imbalance-correction methods and was a noticeable underperformer when applied to random forest. Results from previous studies are mixed regarding the effect of up-sampling on random forest performance. For example, Johnson et al. (2012) found that random forest with SMOTE (synthetic minority oversampling technique), a form of up-sampling, generally underperformed slightly relative to base random forest models when applied to imbalanced bird datasets. However, one of us (DRC) has previously observed that up-sampling led to higher model performance (TSS) on test data than down-sampling when modeling lichen distributions (*unpublished data*).

The tradeoffs that accompany up- and down-sampling and the contexts in which one method outperforms the others deserve further attention. Japkowicz and Stephen (2002) suggested that down-sampling is advantageous when the majority class of the dataset being

modeled has a lot of 'irrelevant' data (Japkowicz and Stephen, 2002). They used this reasoning to explain why a study by Domingos (1999) found that down-sampling was generally better than up-sampling at improving the performance of models built with various real-world datasets (Domingos, 1999). In contrast, Japkowicz and Stephen (2002) found that up-sampling performed better with simulated datasets where none of the data were irrelevant or noisy. In our study, down-sampling may have performed relatively well because the majority class (absences) had some noise in it. For example, aquatic macroinvertebrate samples never contain all of the species that occur at a site, so many false absences occur in these data sets, especially for species with low prevalence. Additional research is needed to better assess how false species absences generally influence the effect of up- versus down-sampling on SDM performance.

Machine-learning algorithm advantages and disadvantages

The base machine-learning algorithms used in our study also have advantages and disadvantages associated with them. For example, gradient boosting and ANN models have performed better than other approaches in several studies, including some with highly imbalanced datasets (Lawrence et al., 2004; Segurado and Araujo, 2004; Brown and Mues, 2012; Sor et al., 2017). In our study, choice of machine-learning algorithm did not greatly affect model performance, although gradient boosting and ANN did perform slightly better with more imbalanced datasets (Table 3-4). Together, these studies support the usefulness of gradient boosting and ANN for modeling imbalanced macroinvertebrate data. However, a disadvantage of ANNs, in particular, is that they can require extensive tuning of many hyperparameters to achieve good performance (Mendoza et al., 2016). Alternatively, random forest has the advantages of being less susceptible to overfitting issues and easier to optimize than other machine-learning algorithms (Breiman, 2001). Random forest often performs well with default hyperparameter values (e.g., Cutler et al., 2007). The ease with which random forest is implemented explains its popularity among ecologists.

Model improvements

We performed a large grid search to optimize the hyperparameters of each model, but we did not optimize the imbalance-correction methods. All of the imbalance-correction methods discussed in our study can be tuned to maximize some preselected measure of model performance. For the sake of comparability, we implemented a standard approach to determine the values for each imbalance-correction method based on the prevalence of each taxon modeled, which is generally considered a good starting point (Chen et al., 2004; Liu et al., 2005; Freeman and Moisen, 2008; Buda et al., 2018). For example, we updated the cutoff for each taxon to equal t the prevalence of that taxon in the given dataset. However, further improvements in model performance over the base machine-learning algorithms may have been achieved with additional fine-tuning of the imbalance-correction methods such as through a grid search approach.

Model performance may also have been improved if we used a more comprehensive suite of environmental predictors that were selected based on a priori knowledge of the primary habitat requirements of each unique taxon. For example, our models only had one coarse environmental predictor describing substrate size and no predictors describing aspects of flow, both of which can strongly influence which macroinvertebrate genera can inhabit a given stream (Statzner and Higler, 1986; Vinson and Hawkins, 1998; Poff and Zimmerman, 2010). Biotic predictors associated with predatory and competitive interactions between organisms were also absent from our models, but recent work has shown these biotic interactions can influence SDM performance (Van der Putten et al., 2010; Anderson, 2017; Wilkinson et al., 2019).

4.2 Effects of prevalence

Our linear model suggests that prevalence had a much larger effect on performance (~28%) than either machine-learning algorithm (~2%) or imbalance-correction method (~8%). However, examining model performance by prevalence range suggests that the choice of machine-learning algorithm may be slightly more important when modeling species of very low

prevalence compared with more common species. For example, on average, gradient boosting outperformed random forest by nearly 2x (TSS = 0.47 vs. 0.24, respectively) for the 0 – 10% prevalence taxa but by only about 1.25x (TSS = 0.26 vs. 0.21, respectively) for the 20 – 30% prevalence taxa (Fig. 3-1), though standard error bars were still overlapping. Sor et al. (2017) found that the choice of machine-learning algorithm became less important when modeling species with prevalence greater than 30%. Thus, more consideration in selecting a particular machine-learning algorithm may be warranted when developing SDMs for rare species.

Results from the linear model also highlight that a large amount of variation (>60%) in model performance was unexplained by prevalence, imbalance-correction methods, and machine-learning algorithm. Some of the variation in TSS that was unexplained in our study could have been caused by false absences that dampened the signals in our datasets. A high proportion of false absences in a dataset can be problematic when detection probability is low, as is often the case for rarer species, such as those used in our study (Tyre et al., 2003; Gu and Swihart, 2004; MacKenzie et al., 2005).

4.3 Performance metric comparison

We found that the specific performance metric used to assess model performance affected, to some degree, how model performance was interpreted. In general, TSS, AUROC, and kappa identified similar trends in model performance (Fig. 3-2). However, PCC showed a larger and more pronounced effect of prevalence on base machine-learning algorithm performance (Fig. 3-2) and also implied the reverse order in the performance of base machine-learning algorithms compared with the other performance metrics (Table 3-4). Shortcomings of using PCC to evaluate models built with highly imbalanced datasets are well documented (Manel et al., 2001; He and Garcia, 2009; Akosa, 2017). Less well documented is that even kappa may be dependent on prevalence and therefore not necessarily a robust choice to assess model performance with imbalanced datasets (McPherson et al., 2004; Allouche et al., 2006; Akosa, 2017). For example,

Allouche et al. (2006) showed that kappa had a unimodal relationship with species prevalence and recommended the use of the TSS, which maintained the benefits associated with kappa while presumably being independent of prevalence. In our study, kappa differed only slightly from AUROC and TSS with respect to the effects of prevalence on normalized model performance.

5. Conclusion

We found that all imbalance-correction methods improved model performance over the base machine-learning algorithms, and ecologists should therefore consider adjusting data for class imbalances when developing species distribution models. To our knowledge, our study is the first to systematically evaluate the degree to which imbalance-correction methods help balance the tradeoff between machine-learning model sensitivity and specificity with ecological presence/absence species data, thereby improving overall model performance at classifying both presences and absences. However, the research areas, such as bioassessment, where implementing imbalance-correction methods may lead to different outcomes compared to outcomes from models built without imbalance-correction methods requires further study.

We observed that up-sampling applied to random forest was the only imbalance-correction method that improved model performance noticeably less than the other imbalance-correction methods. Down-sampling, however, consistently improved model performance across machine-learning algorithms. The context (i.e., the dataset characteristics) in which down-sampling is more effective than up-sampling and vice versa is unclear and further research on the tradeoffs of excluding data (down-sampling) versus repeating data (up-sampling) during model training is also needed.

Finally, our linear model highlighted that prevalence explained more variation in model performance than machine-learning algorithm or imbalance-correction method. Additional research to assess the generality of this finding and the ecological reasons for it should further improve modeling efforts of rare and common species.

References

Akbani, R., Kwek, S., Japkowicz, N., 2004. Applying support vector machines to imbalanced datasets. European Conference on Machine Learning. Springer, Berlin Heidelberg, pp. 39-50.

Akosa, J., 2017. Predictive accuracy: a misleading performance measure for highly imbalanced data. In: Proceedings of the SAS Global Forum, pp. 2-5.

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology. 43, 1223-1232.

Anderson, R. P., 2017. When and how should biotic interactions be considered in models of species niches and distributions? Journal of Biogeography. 44, 8-17.

Berger, E., Haase, P., Kuemmerlen, M., Leps, M., Schaefer, R. B., Sundermann, A., 2017. Water quality variables and pollution sources shaping stream macroinvertebrate communities. Science of the Total Environment. 587, 1-10.

Breiman, L., 2001. Random forests. Machine Learning. 45, 5-32.

Brown, I., Mues, C., 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications. 39, 3446-3453.

Buda, M., Maki, A., Mazurowski, M. A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks. 106, 249-259.

Chawla, N. V., Japkowicz, N., Kotcz, A., 2004. Special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter. 6, 1-6.

Chen, C., Liaw, A., Breiman, L., 2004. Using random forest to learn imbalanced data. University of California Berkeley. Technical Report 666.

Clarke, R. T., Wright, J. F., Furse, M. T., 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecological Modelling. 160, 219-233.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning. 20, 273-297.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J., 2007. Random forests for classification in ecology. Ecology. 88, 2783-2792.

da Silveira, C. B. L., Strenzel, G. M. R., Maida, M., Gaspar, A. L. B., Ferreira, B. P., 2021. Coral Reef Mapping with Remote Sensing and Machine Learning: A Nurture and Nature Analysis in Marine Protected Areas. Remote Sensing. 13, 2907. https://doi.org/10.3390/rs13152907

De'Ath, G., 2007. Boosted trees for ecological modeling and prediction. Ecology. 88, 243-251.

Dedecker, A. P., Goethals, P. L., De Pauw, N., 2002. Comparison of artificial neural network (ANN) model development methods for prediction of macroinvertebrate communities in the Zwalm river basin in Flanders, Belgium. The Scientific World Journal. 2, 96-104.

Dedecker, A. P., Goethals, P. L., D'heygere, T., Gevrey, M., Lek, S., De Pauw, N., 2005. Application of artificial neural network models to analyse the relationships between *Gammarus pulex* L.(Crustacea, Amphipoda) and river characteristics. Environmental Monitoring and Assessment. 111, 223-241.

Domingos, P., 1999. Metacost: a general method for making classifiers cost-sensitive. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Pp. 155-164.

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, 677-697.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? The Journal of Machine Learning Research. 15, 3133-3181.

Freeman, E. A., Moisen, G. G., 2008. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecological Modelling. 217, 48-58.

Freeman, E. A., Moisen, G. G., Frescino, T. S., 2012. Evaluating effectiveness of down-sampling for stratified designs and unbalanced prevalence in random forest models of tree species distributions in Nevada. Ecological Modelling. 233, 1-10.

Friedman, J., 2001. Greedy function approximation: the gradient boosting machine. Annals of Statistics. 29, 1189-1232.

Gobeyn, S., Mouton, A. M., Cord, A. F., Kaim, A., Volk, M., Goethals, P. L., 2019. Evolutionary algorithms for species distribution modelling: a review in the context of machine learning. Ecological Modelling. 392, 179-195.

Goethals, P., Dedecker, A., Gabriëls, W., De Pauw, N., 2003. Development and application of predictive river ecosystem models based on classification trees and artificial neural networks. In: Recknagel F. (eds) Ecological Informatics. Springer, Berlin Heidelberg, pp. 91-107.

Goethals, P. L., Dedecker, A. P., Gabriels, W., Lek, S., De Pauw, N., 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquatic Ecology. 41, 491-508.

Greenwell, B., Boehmke, B., Cunningham, J., GBM Developers, 2019. Package gbm.

Greiner, M., Pfeiffer, D., Smith, R. D., 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Preventive Veterinary Medicine. 45, 23-41.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C. and Martin, T.G., 2013. Predicting species distributions for conservation decisions. Ecology Letters. 16, 1424-1435.

Gu, W., Swihart, R. K., 2004. Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. Biological Conservation. 116, 195-203.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: review of methods and applications. Expert Systems with Applications. 73, 220-239.

Hawkins, C. P., Cao, Y., Roper, B., 2010. Method of predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biology. 55, 1066-1085.

He, H., Garcia, E. A., 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 21, 1263-1284.

Hoang, T. H., Lock, K., Mouton, A., Goethals, P. L., 2010. Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. Ecological Informatics. 5, 140-146.

Hwang, J. P., Park, S., Kim, E., 2011. A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function. Expert Systems with Applications. 38, 8580-8585.

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. Intelligent Data Analysis. 6, 429-449.

Johnson, R. A., Chawla, N. V., Hellmann, J. J., 2012, October. Species distribution modeling and prediction: a class imbalance problem. Conference on Intelligent Data Understanding. IEEE, pp. 9-16.

Khalilia, M., Chakraborty, S., Popescu, M., 2011. Predicting disease risks from highly imbalanced data using random forest. BMC Medical Informatics and Decision Making. 11, 51.

Kubosova, K., Brabec, K., Jarkovsky, J., Syrovatka, V., 2010. Selection of indicative taxa for river habitats: a case study on benthic macroinvertebrates using indicator species analysis and the random forest methods. Hydrobiologia. 651, 101-114.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, Benesty, B., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T., 2019. Package caret.

Lawrence, R., Bunn, A., Powell, S., Zambon, M., 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote Sensing of Environment. 90, 331-336.

Lek, S., Guégan, J. F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecological Modelling. 120, 65-73.

Liaw, A., Wiener, M., 2002. Package randomForest.

Lin, Y., Chen, Q., Chen, K., Yang, Q., 2016. Modelling the presence and identifying the determinant factors of dominant macroinvertebrate taxa in a karst river. Environmental Monitoring and Assessment. 188, 318.

Liu, C., Berry, P. M., Dawson, T. P., Pearson, R. G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography. 28, 385-393.

MacKenzie, D. I., Nichols, J. D., Sutton, N., Kawanishi, K., Bailey, L. L., 2005. Improving inferences in population studies of rare species that are detected imperfectly. Ecology. 86, 1101-1113.

Maloney, K. O., Schmid, M., Weller, D. E., 2012. Applying additive modelling and gradient boosting to assess the effects of watershed and reach characteristics on riverine assemblages. Methods in Ecology and Evolution. 3, 116-128.

Manel, S., Williams, H. C., Ormerod, S. J., 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. Journal of Applied Ecology. 38, 921-931.

McCarthy, K., Zabar, B., Weiss, G., 2005. Does cost-sensitive learning beat sampling for classifying rare classes? In: Proceedings of the 1st International Workshop on Utility-based Data Mining. ACM, Chicago Illinois, pp. 69-77.

McPherson, J. M., Jetz, W., Rogers, D. J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? Journal of Applied Ecology. 41, 811-823.

Mendoza, H., Klein, A., Feurer, M., Springenberg, J. T., Hutter, F., 2016. Towards automatically-tuned neural networks. In: Workshop on Automatic Machine Learning. Pp. 58-65.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2019. Package e1071.

Moisen, G. G., Freeman, E. A., Blackard, J. A., Frescino, T. S., Zimmermann, N. E., Edwards Jr, T. C., 2006. Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. Ecological Modelling. 199, 176-187.

Moss, D., Furse, M. T., Wright, J. F., Armitage, P. D., 1987. The prediction of the macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. Freshwater Biology. 17, 41-52.

Muñoz-Mas, R., Sánchez-Hernández, J., McClain, M. E., Tamatamah, R., Mukama, S. C., Martínez-Capel, F., 2019. Investigating the influence of habitat structure and hydraulics on tropical macroinvertebrate communities. Ecohydrology & Hydrobiology. 19, 339-350.

Olaya-Marín, E. J., Martínez-Capel, F., Vezza, P., 2013. A comparison of artificial neural networks and random forests to predict native fish species richness in Mediterranean rivers. Knowledge and Management of Aquatic Ecosystems. 409, 07.

Olden, J. D., Lawler, J. J., Poff, N. L., 2008. Machine learning methods without tears: a primer for ecologists. The Quarterly Review of Biology. 83, 171-193.

Olden, J. D., Jackson, D. A., 2002. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling. 154, 135-150.

Poff, N. L., Zimmerman, J. K., 2010. Ecological responses to altered flow regimes: a literature review to inform the science and management of environmental flows. Freshwater Biology. 55, 194-205.

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rocha, J. C., Peres, C. K., Buzzo, J. L. L., de Souza, V., Krause, E. A., Bispo, P. C., Frei, F., Costa, L. S., Branco, C. C., 2017. Modeling the species richness and abundance of lotic macroalgae based on habitat characteristics by artificial neural networks: a potentially useful tool for stream biomonitoring programs. Journal of Applied Phycology. 29, 2145-2153.

Segurado, P., Araujo, M. B., 2004. An evaluation of methods for modelling species distributions. Journal of Biogeography. 31, 1555-1568.

Sor, R., Park, Y. S., Boets, P., Goethals, P. L., Lek, S., 2017. Effects of species prevalence on the performance of predictive models. Ecological Modelling. 354, 11-19.

Statzner, B., Higler, B., 1986. Stream hydraulics as a major determinant of benthic invertebrate zonation patterns. Freshwater Biology. 16, 127-139.

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., Possingham, H. P., 2003. Improving precision and reducing bias in biological surveys: estimating false-negative error rates. Ecological Applications. 13, 1790-1801.

USEPA (US Environmental Protection Agency), 2016. National aquatic resource surveys. National rivers and streams assessment 2008-2009 (data and metadata files). Available from U.S. EPA website: http://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys. Date accessed: 2019-3-01.

Van der Putten, W. H., Macel, M., Visser, M. E., 2010. Predicting species distribution and abundance responses to climate change: why it is essential to include biotic interactions across trophic levels. Philosophical Transactions of the Royal Society B: Biological Sciences. 365, 2025-2034.

Vaughan, I. P., Ormerod, S. J., 2005. The continuing challenges of testing species distribution models. Journal of Applied Ecology. 42, 720-730.

Venables, W. N., Ripley, B. D., 2002. Package nnet.

Vinson, M. R., Hawkins, C. P., 1998. Biodiversity of stream insects: variation at local, basin, and regional scales. Annual Review of Entomology. 43, 271-293.

Weinert, M., Mathis, M., Kröncke, I., Pohlmann, T., Reiss, H., 2021. Climate change effects on marine protected areas: Projected decline of benthic species in the North Sea. Marine Environmental Research. 163, 105230.

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., McCarthy, M. A., 2019. A comparison of joint species distribution models for presence–absence data. Methods in Ecology and Evolution. 10, 198-211.

Yap, B. W., Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., Abdullah, N. N., 2014. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In: Proceedings of the First International Conference on Advanced Data and Information Engineering. Springer, Singapore, pp. 13-22.

CHAPTER 4

EFFECTS OF SAMPLE SIZE AND NETWORK DEPTH ON A DEEP LEARNING APPROACH TO SPECIES DISTRIBUTION MODELING*

Abstract

Deep learning algorithms have improved predictive model performance in a variety of disciplines because of their ability to approximate complex functions. However, the amount of data and depth of the neural network needed to improve model performance is not well understood and may depend on many factors associated with the specific field of research. In ecology, ecologists rely on accurate species distribution models to inform conservation and management efforts. Here, we present the first study to systematically examine the effects of sample size and network depth on the performance of species distribution models built with artificial neural networks. We found that one or several deeper network architectures (>1 hidden layer) consistently led to slightly higher model performance than a shallow neural network on validation data when trained with a large sample size (10,000 sites). However, comparing deep network model performance with random forest model performance showed that random forest generally performed as well or slightly better. There was no clear or consistent benefit of using deep neural networks with smaller sample sizes (100 and 1,000 sites). Our results suggest that, given sufficiently big data, increasing the number of hidden layers in a neural network can potentially improve species distribution model performance. As datasets become larger and high performance computing resources become more available, a deep learning approach to species distribution modeling is likely to be used more frequently.

* Coauthored by Charles P. Hawkins

1. Introduction

Deep learning is a subgroup of machine-learning techniques that has received increased attention across a diverse set of fields because under some circumstances it has outperformed other machine-learning approaches when applied to very large datasets (Liu et al. 2017, Marcus 2018). However, the general size of the dataset and depth of the network needed to produce superior performance by deep neural networks (DNNs) is often uncertain and may be highly field and application specific (Karsoliya 2012, Cho et al. 2015, Knight et al. 2017). Deep learning is commonly applied in the fields of machine vision, speech recognition, finance, business, bioinformatics, and medicine, but it is just beginning to be explored in the field of ecology for purposes such as species detection and identification (Dyrmann et al. 2016, Villa et al. 2017, Buschbacher et al. 2020) and species distribution modeling (Chen et al. 2016, Botella et al. 2018, Christin et al. 2019). Elith and Leathwick (2009) provide a thorough review of the methods used in species distribution modeling. Their review mentions the use of artificial neural networks, but does not include examples of the use of deep learning methods, or DNNs, specifically. Since then, several studies have emphasized the potential of deep learning for species distribution modeling in ecology (Zhang and Li 2017, Botella et al. 2018, Christin et al. 2019), but few empirical studies exist evaluating and comparing the performance of these models. To our knowledge, Botella et al. (2018) first applied DNNs to species distribution modeling with a dataset of approximately 5,000 sites with associated species count data. One reason for the slow adoption of deep learning in ecology is that, traditionally, ecological datasets have been relatively small because of the expense associated with conducting large surveys (Stockwell and Peterson 2002). Even so, several studies have shown strong performance by deep learning approaches on relatively small classification datasets of only a few hundred observations per class (Guirado et al. 2018, Abrams et al. 2019). In comparison, in other fields, such as finance and business, data collection can be easily automated and large datasets quickly assembled from the internet (Begenau et al. 2018, Popovic et al. 2018). However, recent advances in automated data

collection (e.g., mass deployment of various environmental sensors that collect and transmit data remotely) and collaborative efforts have led to the creation of some large ecological datasets (Peters et al. 2014). When applied to these large datasets, deep learning can outperform other machine-learning approaches (e.g., Rammer and Seidl 2019). Additionally, as the technology surrounding automation progresses, it is likely that the size of ecological datasets will increase at a fast rate.

Although DNNs have seldom been applied to model species distributions, shallow neural networks (SNNs) have been used because of their flexibility in modeling nonlinear interactions (Lek et al. 1996, Lek and Guegan 1999, Park et al. 2003). SNNs imply few hidden layers, usually a single hidden layer, and have been applied in ecology since the 1990s (Lek et al. 1996, Lek and Guegan 1999, Mhaskar et al. 2017, Marcus 2018). The basic structure of a SNN consists of an input layer with the number of nodes corresponding to the number of predictors, a hidden layer consisting of a variable number of nodes, and an output or prediction layer. These layers and nodes are connected by weights, the values of which are learned during training. Optimizing or parameterizing a neural network requires learning the network weights that best map input to expected output.

DNNs are an extension of SNNs that add flexibility and efficiency to the model by incorporating >1 hidden layers that can handle higher complexity (Bianchini and Scarselli 2014, Mhaskar et al. 2017). DNNs became popular about a decade after SNNs when they were found to produce superior performance at speech and image recognition tasks (Krizhevsky et al. 2012, Marcus 2018). However, when training a DNN on a small dataset, the increased flexibility can lead to higher overfitting than a SNN because of an increased chance of modeling noise and random peculiarities, which may compromise how well they generalize to validation data (Marcus 2018). In contrast, when trained on very large datasets, DNNs appear to predict validation data more accurately than SNNs (Mhaskar et al. 2017, Marcus 2018). However, there is no general consensus about how many observations are needed to produce robust DNNs

because performance is dependent on many factors that can vary across datasets (Christin et al. 2019). Still, some studies have provided valuable insight regarding the sample size needed to reach certain performance thresholds with deep learning approaches. For example, Knight et al. (2017) found that a convolutional neural network (a type of DNN) outperformed humans and other software programs at identifying the Common Nighthawk (*Chordeiles minor*) from audio recordings when the convolutional neural network was trained on greater than 36 hours of audio data. Similarly, Cho et al. (2015) identified the optimal sample size needed by a DNN to achieve a specific performance when classifying medical images. Such empirical studies are critical to determining if deep learning is a viable option for a given area of research.

Freshwater macroinvertebrates are commonly used indicators of environmental quality because they are diverse, have highly variable environmental requirements, and differentially respond to environmental stressors (Goodnight 1973, Resh and Rosenberg 1993, Hawkins et al. 2010). Predicting how the distributions of different macroinvertebrates vary across complex, naturally occurring environmental conditions is a critical component of ecological assessments that assess ecological integrity by comparing observed taxa with those expected to occur under natural environmental conditions (Moss et al. 1987, Wright 1995, Hawkins 2006). SNNs and other machine-learning algorithms such as random forest have been used to model the relationships between macroinvertebrates (presence/absence, abundance, richness) and their environments (e.g., Park et al. 2003, Dedecker et al. 2005, Hoang et al. 2006, Olden et al. 2006, Goethals et al. 2007, Kubosova et al. 2010, Lin et al. 2016). However, to our knowledge, deep learning has not been applied to macroinvertebrate distribution modeling, perhaps in part because of limitations in the size of macroinvertebrate datasets.

The objective of this study was to use a large macroinvertebrate dataset to determine the effects of sample size and neural network depth on the performance of a deep learning approach to species distribution modeling in which occurrences (binary presences or absences) of individual species are predicted. A secondary objective was to compare neural network model

performance with random forest model performance. Random forest models are frequently used to predict species distributions (Cutler et al. 2007, Evans et al. 2011). We use the results to assess if there is a general sample size and network depth at which point DNNs generally outperform SNNs for macroinvertebrate distribution modeling. We hypothesized that one or several DNNs would outperform SNNs at large sample size, but would underperform SNNs on small sample size datasets because of overfitting issues.

2. Methods

2.1 Dataset

Macroinvertebrate presence/absence data were obtained from the National Aquatic Monitoring Center (https://www.usu.edu/buglab/). This repository contains thousands of records of macroinvertebrates collected at sites mostly across the western United States. The data we used were collected between 1980 and 2019 and between days 100 and 334 of each year. A total of 12,520 sites, each of which occurs in a unique catchment, are represented in the dataset. Macroinvertebrate and other species vary markedly in their prevalence, which creates imbalanced datasets. Machine-learning models such as neural networks often perform poorly when classifying imbalanced datasets, a phenomenon termed the imbalance problem (Chen et al. 2004, Johnson et al. 2012). To provide a representative assessment of model performance in the context of natural variation in prevalence, we chose 5 macroinvertebrate genera that varied almost 13 fold in prevalence (rare to common) across the 12,520 sites (Table 4-1). Each genus was modeled individually. In total, 90 models were optimized (5 genera × 6 network architectures × 3 dataset sizes).

Table 4-1. Taxa modeled in this study and their associated prevalence.

Genus	Family	Order	Number of presences	Prevalence
Caenis	Caenidae	Ephemeroptera	730	5.8%
Tricorythodes	Leptohyphidae	Ephemeroptera	2102	16.8%
Micrasema	Brachycentridae	Trichoptera	3771	30.1%
Rhyacophila	Rhyacophilidae	Trichoptera	4585	36.6%

Baetis	Baetidae	Ephemeroptera	9397	75.1%

StreamCat (Stream-Catchment) is a national dataset of 242 environmental variables that characterize geoclimatic conditions at 2.6 million stream segments and their associated catchments (Hill et al. 2016). We used 10 of these metrics in modeling that are often associated with macroinvertebrate distributions (Table 4-2) (Moss et al. 1987, Vinson and Hawkins 1998, Hawkins et al. 2010). These 10 predictors were then matched to the stream segment associated with each of the 12,520 sites where macroinvertebrate samples were collected. As recommended by Olden and Jackson (2002), all predictor data were standardized to give predictors equal weight with the formula.

$$z_n = \frac{x_n - X}{\sigma_x}$$

where z_n is the value of the nth observation after standardization, x_n is the original value of the nth observation, X is the mean, and σ_x is the standard deviation of the particular predictor variable.

We assembled training, testing, and validation datasets for each genus. Specifically, for each genus, a standard external stratified validation dataset consisting of 2,520 sites (~20% of sites) was set aside for final model validation. Next, we divided the remaining 10,000 sites into smaller datasets to test our hypothesis regarding effects of sample size and network depth on model performance. Specifically, we created three datasets for each genus that consisted of 100 (1X), 1,000 (10X), and 10,000 (100X) sites. These datasets were randomly split into 70/30 stratified training/testing sets. The stratification ensured that there was an equal proportion of presences and absences in the training and testing sets for each genus.

Table 4-2. Variables included as predictors in the species distribution models.

Predictor	Description
CatAreaSqKm	NHDPlus[1] catchment area (km^2)
HydrlCondCat	Mean catchment hydraulic conductivity of surface lithology (µm/s)

Mean_MSST	Predicted mean summer water temperature (C) for each segment averaged over years 2008, 2009, 2013, and 2014
Precip8110Cat	Catchment-scale PRISM[2] normal mean precipitation (mm) for 1981-2010
Tmax8110Cat	Catchment-scale PRISM normal maximum air temperature (C) for 1981-2010
Tmean8110Cat	Catchment-scale PRISM normal mean air temperature (C) for 1981-2010
Tmin8110Cat	Catchment-scale PRISM normal minimum air temperature (C) for 1981-2010
ElevCat	Mean elevation of the catchment (m)
BFICat	Catchment-scale base flow index describing the ratio of baseflow to total flow (%)
RunoffCat	Mean catchment runoff (mm)

[1] National Hydrography Dataset Plus
[2] Parameter elevation Regression on Independent Slopes Model

2.2 Model architectures

Model architectures consisted of neural networks with 1-6 hidden layers and several common architectural features shared among all models. Therefore, we modeled each genus 18 times (3 datasets (i.e., 100, 1,000, 10,000) × 6 neural networks (i.e., 1, 2, 3, 4, 5, 6 hidden layers). We selected specific model architectures for use in this study based on approaches commonly used in the artificial neural network literature as well as our observations during preliminary analyses. We used the Adam optimizer with a default learning rate of 0.001 in all models and applied the rectified linear unit (ReLU) activation function and batch normalization in all hidden layers of all models. In addition to batch normalization, dropout is a common technique for dealing with overfitting issues in neural networks (Srivastava et al. 2014). During preliminary analyses, we applied dropout at its default value of 0.50/hidden layer and did not see improvements in model performance, so we decided not to include dropout in our neural network models. The Adam optimizer is increasingly popular in deep learning and is considered relatively robust to choice of hyperparameters (Kingma and Ba 2014, Goodfellow et al. 2016). We chose to use the ReLU over the sigmoid activation function because it has been shown to improve parameter optimization and learning time (Nair and Hinton 2010, Krizhevsky et al. 2012, Botella et al. 2018). We used early stopping to determine the optimal number of epochs over which training occurred (specifically stopping occurred when no improvement in loss on the test dataset

occurred for 10 consecutive epochs). We used the binary cross entropy loss function and batch size was 50 for all models. Early stopping regularly improved model performance and markedly reduced optimization time compared with optimizing epochs via a grid search approach. We optimized the number of nodes in each hidden layer of each model as discussed below (section 2.5).

We also developed random forest models for each dataset and each genus to allow performance comparisons with a different classifier commonly used to model macroinvertebrate and other ecological datasets (e.g., Cutler et al. 2007, Kubosova et al. 2010, Olaya-Marín et al. 2013). The number of trees (500) and randomly selected variables to try at each node (3) were the same for all models.

2.3 Software

We implemented all neural networks with the Keras (https://keras.io/) open source neural network library in Python (Chollet et al. 2015). Keras offers numerous advantages for swiftly implementing and experimenting with neural networks, especially for scientists who are not very familiar with Python programming. Keras can run on top of the most popular deep learning frameworks such as Tensorflow, which has a well designed backend for handling the low-level mathematical operations (e.g., tensor products) needed during neural network training. However, Keras is implemented at a higher level than Tensorflow, making it far more accessible and expedient for creating and running neural networks. In Keras, it is possible to implement sophisticated neural networks in a very short amount of time, with limited programming experience.

Talos is a hyperparameter optimization library and workflow that works with Keras models (Autonomio 2019). We used Talos to reduce model implementation time. Talos enables fast implementation of grid, random, or probabilistic optimization strategies. We implemented

random forest models with the *randomForest* package in the R statistical software (Liaw and Wiener 2002, R Core Team 2019, Vienna, Austria).

2.4 Performance metrics

We used the true skill statistic (TSS) to evaluate model performance, which combines the information from sensitivity and specificity into a single value equal to sensitivity + specificity – 1 (Allouche et al. 2006). Sensitivity and specificity describe how well models correctly classify presences and absences, respectively. Specifically, sensitivity = true presences / (true presences + false absences) and specificity = true absences / (true absences + false presences). The TSS is a good metric for describing model performance given imbalanced binary class datasets because it places equal weight on the model's ability to predict both classes (Allouche et al. 2006, Akosa 2017). In the Appendix, we also include percent classified correctly (PCC) and area under the receiver operating characteristic curve (AUROC) as performance metrics because they are commonly applied in the species distribution modeling literature for model evaluation (Elith and Leathwick 2009, Akosa 2017).

2.5 Model optimization and validation

We used Talos to optimize the number of nodes in each hidden layer of each model with random grid search and probabilistic reduction. We tested a total of 20 different node configurations in each hidden layer. The number of nodes ranged from 10-380 in increments of approximately 20 nodes (i.e., 10, 29, 49, 68, 88, …, 380). In random grid search, the user specifies a fraction of hyperparameter combinations, which are randomly sampled from all possible combinations that make up the full hyperparameter space. This procedure is a necessary step given the computational infeasibility of optimizing numerous hyperparameters simultaneously with extremely high numbers of hyperparameter combinations (Bergstra and Benjio 2012). For example, the total number of combinations associated with optimizing the number of nodes in 6 hidden layers with 20 possible node configurations per layer is 20^6 =

64,000,000 combinations (Table 4-3). By randomly sampling 0.0001 of the total possible node configurations, the total combinations actually tested is 6,400. We chose 6,400 as the maximum number of combinations because the 6-layer neural network with dataset size of 10,000 took approximately 1 day to optimize on a desktop computer (Intel Core i7-3770 CPU). Probabilistic reduction further decreases optimization time by further narrowing the possible hyperparameter combinations. Specifically, probabilistic reduction applies a lookback window during optimization and determines the correlation between hyperparameter values (number of nodes/layer in this study) and a user selected measure of model performance (TSS in this study). If the correlation between a certain hyperparameter value and model performance is sufficiently negative, those hyperparameter values will be excluded from all future hyperparameter combinations tested. Optimized nodes per layer and epochs are presented for each model in Appendix E.

We used the training and testing sets during Talos model optimization to select the best hyperparameters for each model. Specifically, the best hyperparameters were those corresponding to the maximum training TSS + testing TSS. The reason we used both training TSS and testing TSS to select optimal hyperparameters was because the two TSS's were occasionally at odds. Specifically, on occasion, our grid search would produce hyperparameter combinations that led to low TSS when classifying the training data but a higher TSS when classifying the test data (i.e., in a sense overfitting the test data). These hyperparameters were likely not optimal if they did so poorly on the training data, even if they did relatively well on the test data. To avoid these rare cases, we used maximum training TSS + testing TSS to determine the optimal hyperparameters. These optimized hyperparameters were then used to train the final model on the full training dataset (training + testing sets). Final models were then validated on the external 2,520 site validation sets. We repeated the final training and validation procedure 5 times for each genus by sample size combination to calculate a TSS mean and standard error (SE) and a minimum and

maximum for both the training and validation datasets. We present results as the mean ± maximum and minimum for each genus and the mean ± SE of the means across the 5 genera.

Table 4-3. Optimization strategy showing the total possible combinations making up the hyperparameter space for each neural network architecture and the number of hyperparameter combinations actually tested in each neural network after applying the reduction fraction. Note that the 1-layer and 2-layer networks included a full grid search because the possible combinations were less than 6,400.

Network depth	Possible combinations	Reduction fraction	Combinations tested
1 layer	20	1	20
2 layer	400	1	400
3 layer	8,000	0.8	6,400
4 layer	160,000	0.04	6,400
5 layer	3,200,000	0.002	6,400
6 layer	64,000,000	0.0001	6,400

3. Results

3.1 Independent effects of sample size and network depth

Increasing sample size generally reduced model performance for the training data but increased it for the validation data (Fig. 4-1). Model performance assessed with PCC and AUROC showed a similar relationship to model performance assessed with TSS (Appendix D). Specifically, as sample size increased from 100 to 10,000, average neural network model performance based on the training data decreased from TSS = 0.54 to TSS = 0.45. However, average neural network model performance was highest for the 1,000-sample training data with TSS = 0.63. In contrast to model performance on the training data, average model performance based on the validation data improved linearly for neural networks as sample size increased from 100 to 10,000. The average validation TSS across all neural network models was 0.21, 0.29, and 0.38 when trained with the 100, 1,000, and 10,000-sample datasets, respectively. Random forest, averaged across genera, performed slightly better than neural network models with validation TSS of 0.24, 0.30, and 0.40 when trained with 100, 1,000, and 10,000 samples, respectively.

Increasing the number of hidden layers in the neural network models had a noticeable effect on model performance, averaged across sample sizes, with the training data, but the

number of hidden layers had no effect based on the validation data (Fig. 4-1). Specifically, based on the training data, the 1 hidden layer network, averaged across sample sizes, had a markedly lower mean TSS of 0.43 compared with the deeper networks which had mean TSS values ranging from 0.53 (6 hidden layer network) to 0.62 (4 hidden layer network). Effects of network depth averaged across sample sizes, however, had no effect on neural network model performance based on the validation data with mean TSS ranging from 0.29 to 0.30.

3.2 Interactions of sample size with network depth

The interaction between sample size and network depth further revealed trends in overfitting and indicated which models and sample sizes best generalized to the validation data. For example, average performance based on the training data increased systematically toward a performance plateau with increasing neural network depth when trained with 1,000 and 10,000 samples (Fig. 4-1). However, when contrasted with mean model performance based on the validation data, this trend disappeared. Moreover, the 1 hidden-layer network model was actually the top performing model when trained with 1,000 samples (mean validation TSS = 0.31 versus the lowest performing model with a 2 hidden layer network [TSS = 0.28]), but it was the lowest performing model when trained with 10,000 samples (mean validation TSS = 0.36 versus the highest performing neural network models with TSS = 0.39). No discernible trend in model performance was observed for the neural networks trained with 100 samples, though overfitting was generally a problem given the large differences between the model performances on the training and the validation data. On average, random forest generally performed comparably or slightly better than neural network models based on the validation data. For example, random forest trained with 10,000 samples performed slightly better on validation data (mean TSS = 0.40) than the 2-6 hidden layer network models (all with mean TSS = 0.39).

Increasing sample size generally also led to more stable model convergence among the deeper neural networks across the 5 runs for each genus (Fig. 4-2). For example, the 4, 5, or 6

hidden layer network model trained with 100 samples for *Tricorythodes*, *Baetis*, *Rhyacophila*, *Micrasema*, and *Caenis* had a minimum TSS very near or at zero, but a high maximum TSS (maximum range (max. – min.) TSS = 0.40 for the 4 hidden layer network for *Caenis*). However, the ranges in validation TSS occurring across the 5 runs for each genus for all neural networks trained with 10,000 samples was similar (maximum range (max. – min.) TSS = 0.13 for the 1 and 6 layer networks for *Tricorythodes*). Random forest models for each genus trained with any sample size consistently had low variation in validation TSS among model runs.

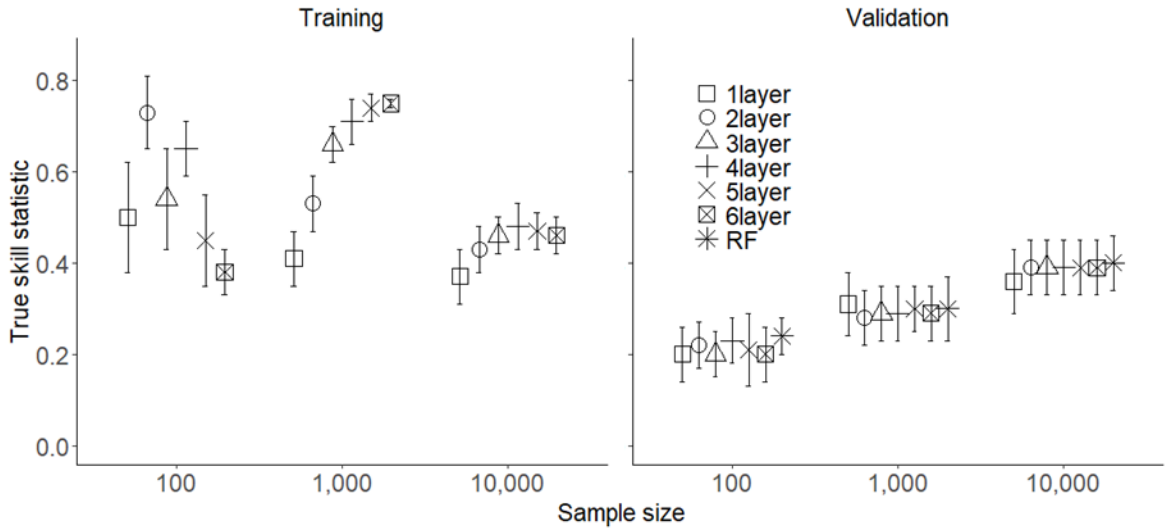


Fig. 4-1. Effects of dataset size and neural network depth on mean ± SE model performance (TSS) for the training dataset (left) and for the validation dataset (right) across the 5 macroinvertebrate genera modeled in this study. RF (random forest) was included for the validation dataset for comparison with a different classifier commonly used in species distribution modeling.

3.3 Effects of prevalence

The differences in model performance across genera appeared to be partly related to differences in prevalence. For example, the TSS for each sample size averaged across models generally decreased with increasing prevalence (Fig. 4-3A), and the TSS for each model averaged across sample sizes showed the same trend with increasing prevalence (Fig. 4-3B). *Rhyacophila* (prevalence = 36.6%) was an outlier in both cases, however, and performed better than the

models for genera with 16.8% and 30.1% prevalence. For each genus and prevalence, average model performance clearly increased with sample size (Fig. 4-3A) although no clear trend in average model performance with model architecture was observed (Fig. 4-3B).

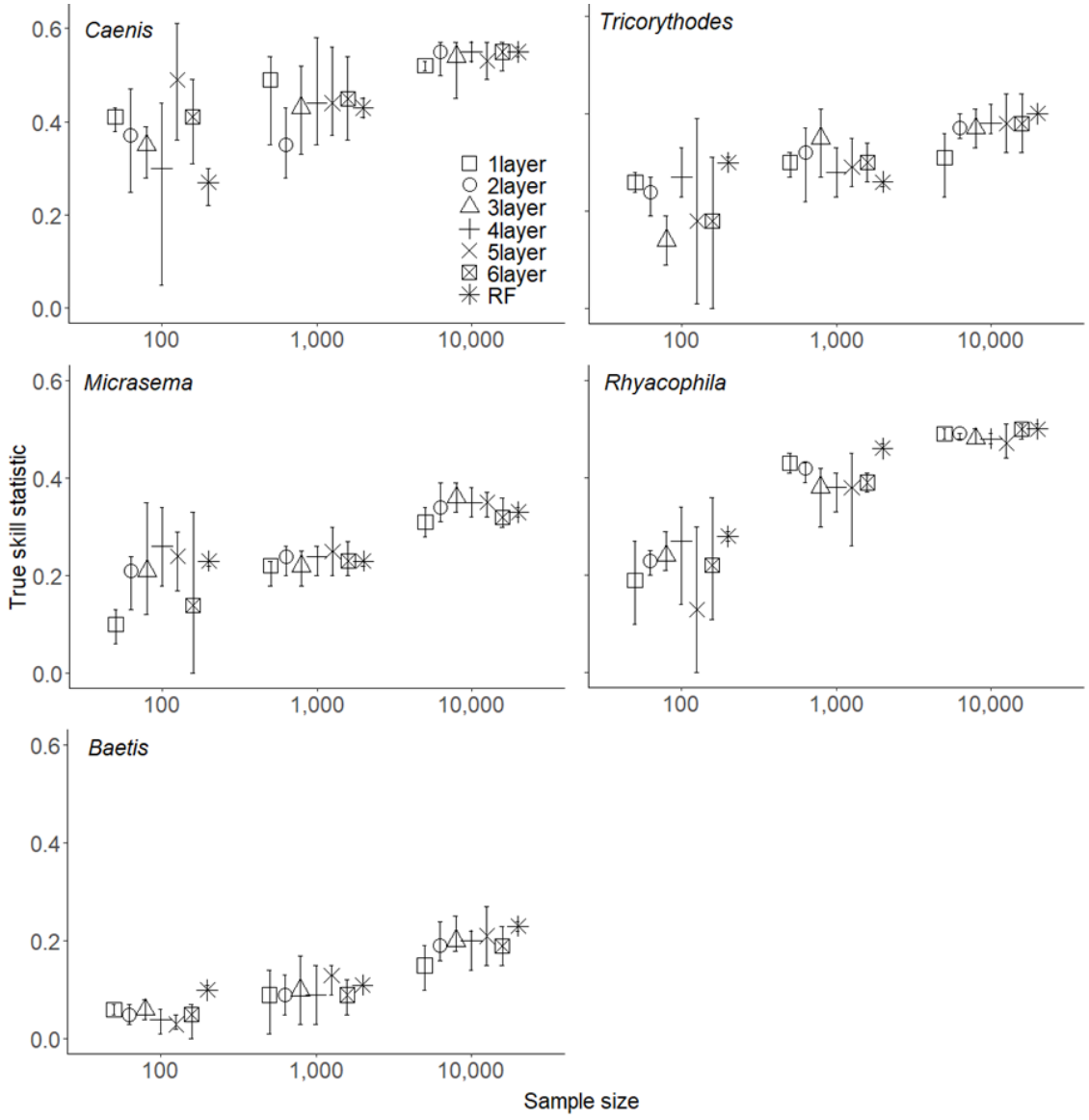


Fig. 4-2. Effects of dataset size and neural network depth on mean validation model performance (TSS) for the validation dataset for each of the 5 macroinvertebrate genera modeled in this study. The bars around each mean show the minimum and maximum TSS from the 5 model runs. RF = the random forest model results.

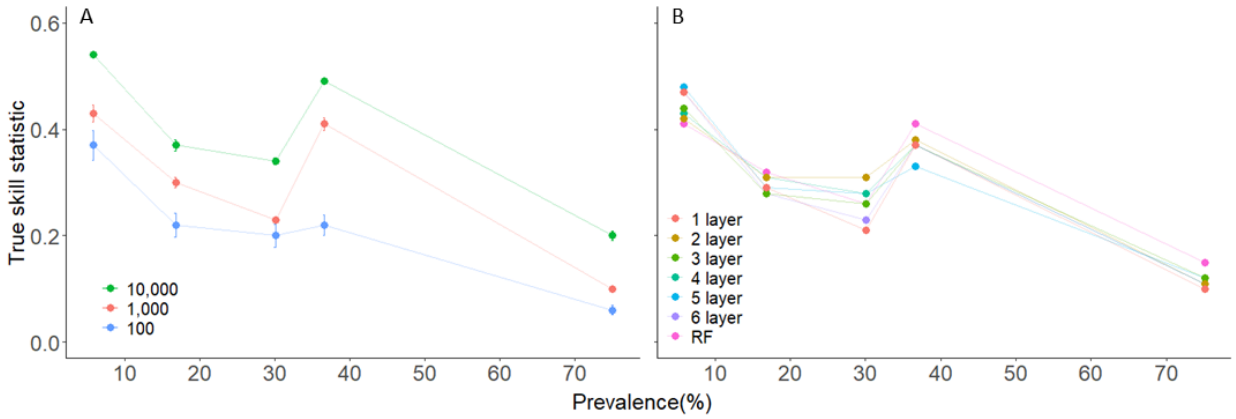


Fig. 4-3. Effects of genus prevalence on model performance. (A) Model performance averaged across models (neural networks and random forest) ± SE for each sample size at each prevalence of the 5 genera. (B) Model performance averaged across sample sizes for each model at each prevalence of the 5 genera. The range in SE is 0.02 – 0.10 across all data points on graph B.

4. Discussion

To our knowledge, our study is the first to systematically compare the effects of network depth (from 1 to 6 hidden layers) and sample size (from 100 to 10,000 sites) on a deep learning approach to species distribution modeling. The sample sizes we selected are representative of, or larger than, those used by ecologists for macroinvertebrate species distribution modeling. We found that as the sample size increased from 100 to 1,000 to 10,000, overfitting the training data generally lessened and a model's ability to generalize to validation data improved markedly. Many studies have noted that generalization improves with more training data (Stockwell and Peterson 2002, Wisz et al. 2008). No general effect of network depth averaged across the sample sizes was observed.

We hypothesized that one or several DNNs would outperform shallow neural networks at large sample size, but DNNs would underperform SNNs on small sample size datasets because of overfitting issues. We observed that deeper networks were generally more prone to overfitting the training data than shallower networks and the degree of overfitting was generally smaller on the largest datasets as we hypothesized. This trend was evident from the model performances on the 1,000 and 10,000 sample training and validation datasets (Fig. 4-1). For example, the 1 hidden

layer network was the worst performing model on the training data but the top performing model on the validation data when trained with 1,000 samples, and the deeper networks all performed slightly better than the shallow network on the validation data when trained with 10,000 samples. The enhanced capability of deep networks with many parameters to learn complicated relationships can lead to fitting noise present in the training data, which is more likely to occur with smaller datasets and reduces a model's ability to generalize (Srivastava et al. 2014). Our results suggest that DNNs for species distribution modeling may not be useful for small datasets in the 100s or low 1000s of samples. However, increasing the number of hidden layers (>1) can lead to slight improvements in species distribution model performance if enough data are available (~10,000 samples), but our study showed no advantage of going above 2 hidden layers (Fig. 4-1). Other studies, however, have shown that deeper network architectures can be advantageous. For example, a recent study by Botella et al. (2018) found that deep networks (6 hidden layers, 200 nodes/layer) outperformed shallow networks (1 hidden layer, 200 nodes) for species distribution models based on a dataset of about 5,000 sites. As datasets become larger and computing resources become more available, a deep learning approach to species distribution modeling may become more applicable in the future.

Random forest performed well in our study and implies this machine-learning technique may often be a preferred approach to species distribution modeling (Cutler et al. 2007). Random forest, on average slightly outperformed all neural network models when trained with 100 and 10,000 samples. Similarly, Shiferaw et al. (2019) found that a DNN clearly underperformed compared with random forest and several other machine-learning algorithms at mapping the distribution of an invasive plant given a dataset of 2,722 presence/absence records. However, the authors noted that further architectural and hyperparameter tuning may have been necessary for the DNN to perform well, but this tuning was not done. In contrast, Rammer and Seidl (2019) found that DNNs generally outperformed other machine-learning algorithms including random forest, gradient boosting machine, and generalized linear model at predicting bark beetle

outbreaks. However, in one experiment they did find that random forest was the top performer. Our analyses highlight an important aspect of DNNs - that they often require more consideration during the design and optimization process than other machine-learning algorithms to achieve comparable results. Random forest requires less consideration during the optimization process and generally performs well with default hyperparameter settings, so the time savings of not needing to tune models coupled with good performance explains its growing popularity among ecologists (Cutler et al. 2007). Additionally, in contrast to the large range in neural network model performance that often occurred over final model runs (5 runs total) for each genus, random forest model performance varied little over the 5 runs (Fig. 4-2). This higher variability affecting neural networks may be due to stochastic optimization methods such as Adam, which introduce variability among training runs (Kingma and Ba 2014). Additional variability among neural network training runs was introduced by the initialization (Xavier uniform initializer) of the weights which are selected randomly from a uniform distribution.

Prevalence appeared to have some effect on model performance (Fig. 4-3), and effects of prevalence on machine-learning model performance is well documented (Johnson et al. 2012, Sor et al. 2017, Buda et al. 2018). However, the performance of each genus model was also likely affected by the specific predictors we used, which were the same for all genera. Thus, some of the genera were likely modeled better than others because the specific set of predictors used probably better represented important niche requirements for some genera relative to others. For example, four of the predictors described aspects of temperature, which is known to affect the fitness of different genera differently (Sweeney and Vannote 1978, Besacier Monbertrand et al. 2019). Additionally, genera typically consist of one or more species that often vary in their niche requirements. For example, *Baetis* consists of several species that vary in their temperature preferences and tolerances (Richards et al. 2013). To more effectively assess the effect of prevalence on the performance of SDMs built with deep neural networks, we will need to model more taxa that vary in prevalence, use species-level data, and use a larger set of predictors that

more comprehensively characterize the environmental factors that can influence different species. Additionally, virtual species data and a systematic study design manipulating prevalence within datasets could be implemented to gain a more robust understanding of the effects of species prevalence on deep neural network model performance.

Compared with other machine-learning algorithms, neural networks are known to require, at times, extensive tuning (Mendoza et al. 2016, Diaz et al. 2017). For example, in this study, the node configurations in the deeper networks (>2 hidden layers) could have been optimized further by increasing the percentage of hyperparameter space sampled during model optimization. Additionally, methods designed to combat overfitting, such as dropout, could have been tuned at each layer (Srivastava et al. 2014). Applying further tuning could potentially lead to further improvements in model performance and should be explored by ecologists seeking to implement maximally performing neural networks, but doing so would also increase computation time. However, as high-performance computing resources continue to evolve and become more accessible, this limitation will be reduced in the future and further enable the optimization of deeper networks. All optimizations in this study were performed on a desktop computer, which supports the use of deeper network architectures by ecologists without access to high performance computing resources. Further, the ease of implementing model designs with Keras and hyperparameter optimization with Talos provides a straightforward approach for practitioners less familiar with more technical frameworks. Finally, the rapid advancement of automation and optimization approaches and associated software libraries that compare favorably to manual design and tuning by experts will continue to offer significant time savings advantages (Bergstra and Bengio 2012, Mendoza et al. 2016, Autonomio 2019). For example, Mendoza et al. (2016) used an automated neural network design and optimization approach to achieve higher model performance than those achieved by human experts.

A sample size of 10,000 sites with genus presence/absence information is large in the field of ecology for species distribution modeling (Peters et al. 2014), but it is small in many

fields where deep learning is currently applied. For example, datasets in computer vision often consist of millions of images (Najafabadi et al. 2015, Barbu et al. 2016). Our study was limited to a dataset of no larger than 10,000 sites. It would be useful to see if the slight trend we identified regarding improved performance by deeper networks continued to increase with even larger datasets in the 100,000s or millions of samples. For example, remote sensing and aerial survey datasets can provide records in the millions for certain scenarios such as outbreaks of terrestrial invasive plant species and DNNs have recently shown promise for modeling such scenarios (Rammer and Seidl 2019).

Finally, the number of environmental predictors we used in our study was small compared with the numbers routinely used in many other fields and applications (e.g., Reichstein et al. 2019). Deeper networks (>2 hidden layers) in our study may not have increasingly improved performance with added layers simply because the added ability to efficiently model complex relationships in the data was not needed. In the future, it would be useful to assess how data complexity and suites of specific predictor variables affect the performance of deep learning approaches for species distribution modeling.

References

Abrams, J. F., Vashishtha, A., Wong, S. T., Nguyen, A., Mohamed, A., Wieser, S., Kuijper, A., Wilting, A., Mukhopadhyay, A., 2019. Habitat-Net: Segmentation of habitat images using deep learning. Ecological Informatics. 51, 121-128.

Akosa, J., 2017. Predictive accuracy: A misleading performance measure for highly imbalanced data. In: Proceedings of the SAS Global Forum. pp. 2-5.

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology. 43, 1223-1232.

Autonomio Talos [Computer software]. 2019. Retrieved from http://github.com/autonomio/talos.

Barbu, A., She, Y., Ding, L., Gramajo, G., 2016. Feature selection with annealing for computer vision and big data learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 39, 272-286.

Begenau, J., Farboodi, M., Veldkamp, L. 2018. Big data in finance and the growth of large firms. Journal of Monetary Economics. 97, 71-87.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. Journal of Machine Learning Research. 13, 281-305.

Besacier Monbertrand, A. L., Timoner, P., Rahman, K., Burlando, P., Fatichi, S., Gonseth, Y., Moser, F., Castella, E., Lehmann, A., 2019. Assessing the vulnerability of aquatic macroinvertebrates to climate warming in a mountainous watershed: Supplementing presence-only data with species traits. Water. 11, 636.

Bianchini, M., Scarselli, F., 2014. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. IEEE Transactions on Neural Networks and Learning Systems. 25, 1553-1565.

Botella, C., Joly, A., Bonnet, P., Monestiez, P., Munoz, F., 2018. A deep learning approach to species distribution modelling. In: Multimedia Tools and Applications for Environmental & Biodiversity Informatics. Springer, Cham, pp. 169-199.

Buda, M., Maki, A., Mazurowski, M. A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks. 106, 249-259.

Buschbacher, K., Ahrens, D., Espeland, M., Steinhage, V., 2020. Image-based species identification of wild bees using convolutional neural networks. Ecological Informatics. 55, 101017.

Chen, C., Liaw, A., Breiman, L., 2004. Using random forest to learn imbalanced data. University of California Berkeley. Technical Report 666. Retrieved from: https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf.

Chen, D., Xue, Y., Chen, S., Fink, D., Gomes, C., 2016. Deep multi-species embedding. arXiv preprint arXiv:1609.09353.

Cho, J., Lee, K., Shin, E., Choy, G., Do, S., 2015. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?. arXiv preprint arXiv:1511.06348.

Chollet, F., et al., 2015. Keras. URL https://keras.io

Christin, S., Hervet, E., Lecomte, N., 2019. Applications for deep learning in ecology. Methods in Ecology and Evolution. 0, 1-13.

Dedecker, A. P., Goethals, P. L., D'heygere, T., Gevrey, M., Lek, S., De Pauw, N., 2005. Application of artificial neural network models to analyse the relationships between Gammarus pulex L.(Crustacea, Amphipoda) and river characteristics. Environmental Monitoring and Assessment. 111, 223-241.

Diaz, G. I., Fokoue-Nkoutche, A., Nannicini, G., & Samulowitz, H., 2017. An effective algorithm for hyperparameter optimization of neural networks. IBM Journal of Research and Development. 61, 9-1.

Dyrmann, M., Karstoft, H., Midtiby, H. S., 2016. Plant species classification using deep convolutional neural network. Biosystems Engineering. 151, 72-80.

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, 677-697.

Evans, J. S., Murphy, M. A., Holden, Z. A., Cushman, S. A., 2011. Modeling species distribution and change using random forest. In: Predictive Species and Habitat Modeling in Landscape Ecology. Springer, New York, pp. 139-159.

Goethals, P. L., Dedecker, A. P., Gabriels, W., Lek, S., De Pauw, N., 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquatic Ecology. 41, 491-508.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.

Goodnight, C. J., 1973. The use of aquatic macroinvertebrates as indicators of stream pollution. Transactions of the American Microscopical Society. 1, 1-13.

Guirado, E., Tabik, S., Rivas, M. L., Alcaraz-Segura, D., Herrera, F., 2018. Automatic whale counting in satellite images with deep learning. bioRxiv, 443671.

Hawkins, C. P., 2006. Quantifying biological integrity by taxonomic completeness: Its utility in regional and global assessments. Ecological Applications. 16, 1277-1294.

Hawkins, C. P., Cao, Y., Roper, B., 2010. Method of predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biology. 55, 1066-1085.

Hill, R. A., Weber, M. H, Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., 2016. The stream-catchment (StreamCat) dataset: A database of watershed metrics for the conterminous United States. Journal of the American Water Resources Association. 52, 120-128.

Hoang, H., Recknagel, F., Marshall, J., Choy, S., 2006. Elucidation of hypothetical relationships between habitat conditions and macroinvertebrate assemblages in freshwater streams by artificial neural networks. In: Ecological Informatics. Springer, Berlin, pp. 239-251.

Johnson, R. A., Chawla, N. V., Hellmann, J. J., 2012, October. Species distribution modeling and prediction: A class imbalance problem. In: 2012 Conference on Intelligent Data Understanding. IEEE, pp. 9-16.

Karsoliya, S., 2012. Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. International Journal of Engineering Trends and Technology. 3, 714-717.

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Knight, E., Hannah, K., Foley, G., Scott, C., Brigham, R., Bayne, E., 2017. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conservation and Ecology. 12, 14.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097-1105.

Kubosova, K., Brabec, K., Jarkovsky, J., Syrovatka, V., 2010. Selection of indicative taxa for river habitats: a case study on benthic macroinvertebrates using indicator species analysis and the random forest methods. Hydrobiologia. 651, 101-114.

Lek, S., Belaud, A., Baran, P., Dimopoulos, I., Delacoste, M., 1996. Role of some environmental variables in trout abundance models using neural networks. Aquatic Living Resources. 9, 23-29.

Lek, S., Guégan, J. F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecological Modelling. 120, 65-73.

Liaw, A., Wiener, M., 2002. Package randomForest.

Lin, Y., Chen, Q., Chen, K., Yang, Q., 2016. Modelling the presence and identifying the determinant factors of dominant macroinvertebrate taxa in a karst river. Environmental Monitoring and Assessment. 188, 318.

Liu, B., Wei, Y., Zhang, Y., Yang, Q., 2017. Deep neural networks for high dimension, low sample size data. In: International Joint Conference on Artificial Intelligence. pp. 2287-2293.

Marcus, G., 2018. Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631.

Mendoza, H., Klein, A., Feurer, M., Springenberg, J. T., Hutter, F., 2016. Towards automatically-tuned neural networks. In: Workshop on Automatic Machine Learning. pp. 58-65.

Mhaskar, H., Liao, Q., Poggio, T., 2017. When and why are deep networks better than shallow ones?. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. pp. 2343-2349.

Moss, D., Furse, M. T., Wright, J. F., Armitage, P. D., 1987. The prediction of the macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. Freshwater Biology. 17, 41-52.

Nair, V., Hinton, G. E., 2010. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10). pp. 807-814.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., Muharemagic, E., 2015. Deep learning applications and challenges in big data analytics. Journal of Big Data. 2, 1.

Olaya-Marín, E. J., Martínez-Capel, F., Vezza, P., 2013. A comparison of artificial neural networks and random forests to predict native fish species richness in Mediterranean rivers. Knowledge and Management of Aquatic Ecosystems. 409, 07.

Olden, J. D., Jackson, D. A., 2002. Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling. 154, 135-150.

Olden, J. D., Poff, N. L., Bledsoe, B. P., 2006. Incorporating ecological knowledge into ecoinformatics: An example of modeling hierarchically structured aquatic communities with neural networks. Ecological Informatics. 1, 33-42.

Park, Y. S., Céréghino, R., Compin, A., Lek, S., 2003. Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecological Modelling. 160, 265-280.

Peters, D. P., Havstad, K. M., Cushing, J., Tweedie, C., Fuentes, O., Villanueva-Rosales, N., 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere. 5, 1-15.

Popovič, A., Hackney, R., Tassabehji, R., Castelli, M., 2018. The impact of big data analytics on firms' high value business performance. Information Systems Frontiers. 20, 209-222.

Rammer, W., Seidl, R., 2019. Harnessing deep learning in ecology: An example predicting bark beetle outbreaks. Frontiers in Plant Science. 10, 1327.

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., 2019. Deep learning and process understanding for data-driven Earth system science. Nature. 566, 195-204.

Resh, V. H., Rosenberg, D. M., editors, 1993. Freshwater biomonitoring and benthic macroinvertebrates. New York, NY, USA:: Chapman & Hall.

Richards, D. C., Bilger, M., Lester, G., 2013. Development of Idaho macroinvertebrate temperature occurrence models. Final Report to Idaho Department of Environmental Quality. Boise, Idaho. Retrieved from: https://www.deq.idaho.gov/media/60177748/development-idaho-macroinvertebrate-temperature-occurrence-models.pdf

Shiferaw, H., Bewket, W., Eckert, S., 2019. Performances of machine learning algorithms for mapping fractional cover of an invasive plant species in a dryland ecosystem. Ecology and Evolution. 9, 2562-2574.

Sor, R., Park, Y. S., Boets, P., Goethals, P. L., Lek, S., 2017. Effects of species prevalence on the performance of predictive models. Ecological Modelling. 354, 11-19.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research. 15, 1929-1958.

Stockwell, D. R., Peterson, A. T., 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling. 148, 1-13.

Sweeney, B. W., Vannote, R. L., 1978. Size variation and the distribution of hemimetabolous aquatic insects: two thermal equilibrium hypotheses. Science. 200, 444-446.

Villa, A. G., Salazar, A., Vargas, F., 2017. Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. Ecological Informatics. 41, 24-32.

Vinson, M. R., Hawkins, C. P., 1998. Biodiversity of stream insects: variation at local, basin, and regional scales. Annual Review of Entomology. 43, 271-293.

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, 763-773.

Wright, J. F., 1995. Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. Australian Journal of Ecology. 20, 181-197.

Zhang, J., Li, S., 2017. A review of machine learning based species' distribution modelling. In: International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII). pp. 199-206.

CHAPTER 5

DIAGNOSING THE CAUSES OF ALTERED BIODIVERSITY IN FRESHWATER ECOSYSTEMS: DEVELOPMENT AND EVALUATION OF A TEMPERATURE-SPECIFIC BIOTIC INDEX*

Abstract

Human activities have profoundly altered aquatic environments on a global scale. These alterations have degraded aquatic ecosystems and threaten aquatic life and human wellbeing. However, it is often not clear which specific stressor or combination of stressors are primarily responsible for observed losses of aquatic life. We need tools that can accurately identify which stressors are responsible for observed losses of aquatic life and inform mitigation and restoration activities. Assemblage-level biotic indices derived from species-specific tolerances to different stressors (stressor-specific biotic indices) have the potential to provide these diagnoses. We derived and evaluated species-specific thermal tolerance values from USEPA National Rivers and Streams Assessment (NRSA) macroinvertebrate data, which we then incorporated into a temperature-specific biotic index (TBI). We found that thermal tolerance varied substantially among stream macroinvertebrate taxa and that temperature tolerances were temperature specific – i.e., they did not appear to be strongly confounded with tolerance to other stressors that we examined. We applied the TBI to 1,706 macroinvertebrate samples collected during the 2013/2014 USEPA NRSA probability survey and we extrapolated that 2.6% (47,000 km) of streams and rivers across the continental United States (CONUS) had assemblages that exhibited higher TBI values than expected under baseline (i.e., least altered) conditions. Our results indicate that a TBI can detect thermal alteration of aquatic life in the absence of direct temperature monitoring but that the CONUS-wide effects of temperature, relative to other stressors, may be less pervasive than we initially suspected.

* Coauthored by Charles P. Hawkins

1. Introduction

Human activities have altered the thermal environments of freshwater ecosystems at local to global scales (Poole and Berman, 2001; Caisse, 2006; Burgmer et al., 2007; Carpenter et al., 2011; Kaushal et al., 2010; Isaak et al., 2012), but we have a poor understanding of how important alterations in thermal regimes are relative to the effects of many other human-associated stressors on aquatic life (Durance and Ormerod, 2009; Moss, 2010; Craig et al., 2017). We have long known that point source discharges of heated (Sylvester, 1972; Lamberti and Resh, 1985; Lessard and Hayes, 2003) or cooled (Gore, 1977; Clarkson and Childs, 2000) waters can affect the viability of local populations of freshwater species, but we know much less about how the cumulative landscape-level alterations in freshwater thermal regimes have altered biodiversity at regional to global scales over the last ~200 years (Vinson and Hawkins, 1998; Heino et al., 2009). To improve ways of conserving aquatic biodiversity and restoring degraded freshwater ecosystems, we need quantitative, standardized methods that allow us to both compare how different groups of freshwater taxa have responded to thermal alteration and predict how they will respond to both projected climate change and different restoration or mitigation activities (e.g., Yuan, 2006).

Regulatory agencies often use indices of biological condition such as indices of taxonomic completeness, biotic indices based on general tolerance to organic pollution, and multimetric indices of biological integrity (Resh and Rosenberg, 1993) to assess the overall biological status of freshwater ecosystems, but these indices have limited, or no, power to identify the specific stressors that have harmed aquatic life. Stressor-specific biotic indices based on the environmental preferences and tolerances of individual species are promising tools that may aid in diagnosing the causes of change in local and regional patterns of biodiversity (Chessman and McEvoy, 1997; Feld et al., 2020). Such indices could complement general purpose indices of biological condition by helping managers target the specific factors in need of remediation. Research on stressor-specific indices is advancing on a number of fronts – e.g., sediment (Relyea

et al., 2012; Murphy et al., 2015; Hubler et al., 2016), nutrients (Smith et al., 2007), salinity (Horrigan et al., 2005), pH (Murphy et al., 2013), pesticides (Liess and Ohe, 2005; Bray et al., 2020), metals (Blanck, 2002), flow (Extence et al., 1999; O'Keeffe et al., 2002; Armanini et al., 2011; Monk et al., 2018), and temperature (Yuan, 2006; Huff et al., 2008; and Schuwirth et al., 2015). However, we know little regarding the sensitivity and specificity of most stressor-specific indices (but see Laini et al., 2018; Bray et al., 2020), two critical aspects affecting their utility. We define sensitivity as the degree to which an index responds to changes in the focal stressor and specificity as the degree to which an index responds only to the stressor of interest.

Species-specific thermal preferences or tolerances are usually measured in terms of either optima or minimum or maximum limits. In the biological assessment literature, these measures are often referred to as tolerance values (TVs) because they describe a species' tolerance or response to a stressor relative to other species (Chutter, 1972; Hilsenhoff, 1987; Lenat, 1993). For example, species with low thermal optima or upper limits should decline in abundance or go locally extinct in response to increases in temperature, whereas species with higher thermal optima or upper limits would likely increase in abundance (Burgmer et al., 2007; Domisch et al., 2011). Accurately quantifying thermal TVs is thus critical to both identifying which species are most at risk to thermal alterations (Li et al., 2013) and developing temperature-specific biotic indices (TBIs) based on aggregate, assemblage-wide responses for use in causal assessments (Yuan, 2006). For example, a TBI score for a site could be calculated as:

$$TBI = \frac{Observed\ mean\ assemblage\ TV}{Expected\ mean\ assemblage\ TV} = \frac{\frac{\sum O_i TV_i}{\sum O_i}}{\frac{\sum P_i TV_i}{\sum P_i}},$$

where O_i represents either presence (1) or absence (0) of taxon *i* at a site, TV_i represents the thermal TV for taxon *i*, and P_i represents the probability of occurrence of taxon *i* predicted to occur at a site under reference or natural environmental conditions. In freshwater biomonitoring, predicted probabilities of occurrence are often estimated with a RIVPACS-type predictive model (see Wright, 1995 for details on RIVPACS), a type of multitaxon distribution model (Hawkins

and Yuan, 2016). However, it is not yet clear if the method used to derive TVs affects the performance of stressor-specific indices. Additionally, we do not understand how quickly assemblage composition, and thus stressor-specific indices, respond to stress.

Several methods exist for deriving TVs from survey data based on simultaneous measures of environmental conditions and observations of either species abundance or occurrence. Two common methods include calculation of upper or lower limits based on cumulative distributions of abundances or occurrences across an environmental gradient (e.g., Lenat, 1993; Huff et al., 2005) and calculation of environmental optima expressed as either simple or weighted averages (WA) (Ter Braak and Looman, 1986; Yuan, 2006). The cumulative percentile method requires selection of an appropriate percentile as a standard criterion to identify conditions that presumably are either suboptimal to species (e.g., $\geq 75^{th}$ percentile as the upper limit) or identify limits beyond which species are unlikely to persist (e.g., $\geq 95^{th}$ percentile as the upper limit). Upper and lower limits conceptually match how ecologists think about how environmental conditions constrain distributions, but estimating limits is typically more prone to error than estimating optima because the number of samples with the occurrence of a target species is typically sparse near its limits (Yuan, 2006). In contrast, estimates of species optima may be less prone to error and still provide the same sort of environmental response signal needed to quantify biotic responses. However, estimates of optima can be sensitive to how completely the environmental gradient over which a species occurs is represented in a dataset (Ter Braak and Looman, 1986; Yuan, 2005). Incomplete gradients will produce biased estimates of both optima and limits. A third method of estimating TVs based on either limits or optima is to model the relationships between species' abundances or occurrences and the environmental factor of interest – e.g., species distribution models (SDMs) (Austin, 2002; Yuan, 2006; Li et al., 2013). This method may not be as susceptible to the potential errors associated with calculating cumulative percentiles and averages because good models can accurately describe the entire relationship between abundances or occurrences and the stressor of interest or clearly reveal incomplete range

data. Given the variety of methods available to calculate TVs, we need to understand if the methods used affect the potential diagnostic performance of stressor-specific indices.

In this study, we addressed five primary research questions. 1) How strongly is temperature associated with and predictive of taxa distributions compared with other environmental variables? 2) do different methods used to estimate thermal TVs scale TVs differently and affect the sensitivity of mean assemblage thermal tolerance values (hereafter MATTVs) to variation in stream temperature? 3) do MATTVs respond quickly to temporal variation in stream temperature or do responses lag for one or more years? 4) are thermal TVs specific enough that TBIs can isolate temperature-caused alteration of stream assemblages from the effects of other potential stressors? 5) Can a TBI be applied at the level of the continental United States (CONUS) to detect trends in thermal alteration of the Nation's streams and rivers? We used freshwater macroinvertebrates in our analyses because they have diverse ecological requirements and tolerances and occur in most freshwater ecosystems. These properties allow for both a potentially substantial scope of response to different environmental stressors (Chessman and McEvoy, 1997) and comparisons of responses across nearly all continents (Resh, 2008).

2. Material and methods

2.1. General approach

We performed several analyses to address our research questions. First, we built correlative SDMs to assess how important temperature might be as a potential driver of CONUS-level macroinvertebrate distributions relative to that of other major environmental predictors. Second, we used six different methods to derive thermal TVs from a nationally representative set of macroinvertebrate survey data. We compared these six sets of TVs by assigning TVs to the macroinvertebrate taxa observed at each site, calculating MATTVs for each site, and then measuring both the strengths of associations (r^2) and regression slopes between MATTVs and site temperatures. Third, we chose one set of the six TVs and applied them to macroinvertebrate

assemblages from sites that were sampled in two different years. We assessed the responsiveness of MATTVs to interannual variation in stream temperature by comparing the strength of the relationship between change in MATTV and change in site temperature. Fourth, we assessed the specificities of TBIs as the degree to which MATTVs for reference-condition sites varied only with predicted site temperature. Specifically, we fit a regression random forest model where MATTV was the response variable and four major environmental factors were the predictors: temperature, salinity, substrate size, and day of year samples were collected. We then calculated variable importance scores and inspected partial dependence plots that showed how MATTVs varied across different environmental gradients. Fifth, we chose one set of TVs and incorporated them into a TBI which we then applied to a set of macroinvertebrate samples collected at sites across the CONUS that were selected based on a probabilistic survey design. The probabilistic design allowed us to infer the percentage of total stream kilometers within the CONUS that were cooler than expected, similar to expected, or warmer than expected under reference condition.

2.2. How important is temperature to distributions relative to other potential drivers?

To address the first research question, we used the *randomForest* package (Liaw and Wiener, 2002) in R (R Core Team, 2019, Vienna, Austria) to build SDMs for 290 taxa. We used data from the National Rivers and Streams Assessment (NRSA; USEPA, 2016) collected in 2008 and 2009 (Fig. 5-1) to create SDMs. This dataset includes site-level presence/absence and abundance data for several hundred macroinvertebrate taxa as well as environmental attributes at hundreds of locations across the CONUS. Species-level identities are not available from the NRSA dataset, but we were able to use data on 290 unique taxa (251 taxa identified to genus, 37 taxa identified to family, 1 taxon identified to class (Arachnida), and 1 taxon identified to phylum (Platyhelminthes)), each of which occurred in ≥ 30 of the 2142 samples in the dataset. The NRSA dataset includes observations from 1954 unique sites. The additional samples are repeat (duplicate) samples taken on different dates at each of ~ 10% of the unique sites. Preliminary

analyses showed no differences in TV estimates based on the 2142 total samples and 1954 unique site samples, so we used the full dataset to build SDMs and estimate TVs as a means of maximizing the number of taxa we could use in analyses. We then used three environmental predictors plus the day of the year samples were collected (Table 5-1) to construct SDMs. Aspects of stream temperature, substrate, and water chemistry are often associated with the occurrence of different macroinvertebrate taxa (Moss et al., 1987; Hawkins et al., 1997; Clarke et al., 2003; Berger et al., 2017) as is day of year. Stream temperature measurements were not taken during the NRSA survey, but we were able to use predicted mean summer stream temperatures (MSST) derived from models of Hill et al. (2013), which have been mapped to all reaches of the National Hydrography Dataset Plus Version 2 (McKay et al., 2012) and are available in the StreamCat database (Hill et al., 2016). These predicted temperatures perform as well as measured temperatures in predicting taxa occurrences (Hill and Hawkins, 2014). Values of each predictor variable varied substantially across sites, which was ideal for building generalizable models and, ultimately, testing the specificity of thermal TBIs. We also checked if any of the predictors were correlated and could thereby potentially confound inferences.

We built separate models with presence/absence data and abundance data for each taxon. We used random forest because it is a flexible machine-learning algorithm capable of modeling nonlinear interactions between responses and predictors (Breiman, 2001; Cutler et al., 2007; Hawkins et al., 2010) and performs well compared with other modeling approaches (Cutler et al., 2007; Benkendorf et al. 2022). We used classification random forest for the presence/absence data and regression random forest for the abundance data. We then calculated variable importance metrics for each of the 290 models (separately for presence/absence and abundance models) and identified how many times each predictor was the most important predictor across the 290 SDMs. However, it is often difficult to model taxa for which presences and absences are highly imbalanced (typically far more absences than presences). In random forest models, down-sampling can improve the tradeoff between model sensitivity and specificity (Chawla et al., 2004;

Benkendorf et al. 2022) and potentially inferences regarding the importance of different predictors. We therefore conducted preliminary analyses to assess whether data imbalance affected model performance and variable importance inferences in this dataset. We compared variable importance results obtained from models built with abundance data, raw presence/absence data, and down-sampled presence/absence data in which the number of absences were down sampled to equal the number of presences for each taxon. Preliminary analyses showed that down-sampling affected model performance as measured by the True Skill Statistic, but did not affect either variable importance or partial dependence plots. We therefore present the variable importance metrics and TVs derived from random forest models (see below) built with just the raw presence/absence data.

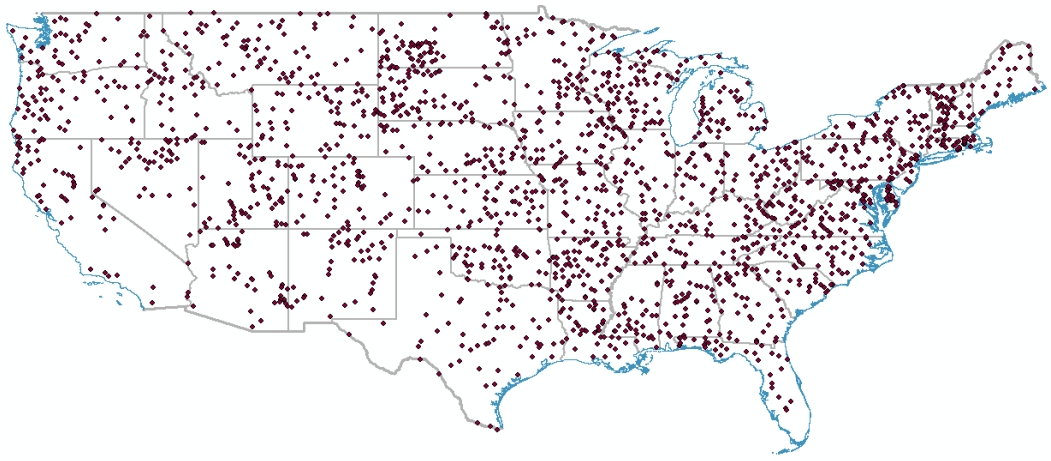


Fig. 5-1. Distribution of the 1954 NRSA 2008-2009 sites that were used to derive tolerance values.

Table 5-1 Variables included as predictors in the species distribution models.

Predictor	Description	Range	Mean	Median
MSST	Predicted mean summer stream temperature (°C)	8.9 – 28.2	20.5	21.2
Substrate	Log_{10} geometric mean substrate particle diameter (mm)	-2.1 – 3.8	0.2	0
Conductivity	Specific conductance (µS/cm)	9 – 62,2301	639	311
DOY	Day of the year sample was collected	111 – 334	213	211

2.3. Does the method of estimating TVs affect the relationship between MATTVs and site temperature?

To address the second research question, we first used six different methods to estimate thermal TVs from the NRSA 2008-2009 survey data: four methods estimated upper thermal limits and the other two estimated thermal optima. These methods included calculation of upper 95th percentiles (95th) derived from both presence/absence and abundance data, visual inspection of partial dependence plots (PDP) derived from both presence/absence and abundance data, and calculation of optima expressed as both simple and weighted (by abundance) averages of temperatures at reaches where a taxon was observed. We refer to these six TV estimates as 95th (p/a), 95th (abund), PDP (p/a), PDP (abund), A (p/a), WA (abund). The two 95th percentile TVs were calculated as the temperature below which 95 percent of taxon occurrences were observed in the presence/absence data and the temperature below which 95 percent of the individuals in a taxon were observed in the abundance data. From the individual species SDMs that we built, we generated partial dependence plots to infer upper limits as the temperature at which the partial dependence trend line showed minimal or near minimal probability of occurrence or abundance, with presence/absence and abundance data, respectively. For example, the inferred upper limit for *Baetis* was 25° C based on both presence/absence and abundance data (Fig. 5-2). Models for numerous taxa produced partial dependence plots in which trend lines continuously increased with increasing temperature. In these cases, we assigned an upper thermal limit of 28° C to the taxon because this was the highest predicted MSST in the NRSA dataset. In cases where partial dependence plots showed no clear trend, we did not assign a TV. For this reason, we report fewer than 290 TVs based on partial dependence plots. We estimated weighted averages as:

$$WA = \frac{\sum_{i=1}^{n} Y_{ij}\, x_i}{\sum_{i=1}^{n} Y_{ij}}$$

where *WA* = the weighted average, *n* indicates total sites and x_i is the MSST at site *i*. The variable Y_{ij} is equal to 1 when species *j* is present and 0 when species *j* is absent. When using abundance

data, Y_{ij} is the abundance of species j at site i. Note that with presence/absence data, the optimum is simply the average temperature across those sites at which the taxon occurred (i.e., no weighting is applied).

To evaluate the six methods of deriving thermal TVs, we applied them to an independent dataset – the 2013-2014 NRSA survey data (USEPA, 2020a). The NRSA 2013-2014 dataset has the same CONUS-level coverage as the 2008-2009 survey. We calculated MATTVs from samples collected at reference-quality sites (299 sites) that contained at least 10 macroinvertebrate taxa with associated TVs and for which temperature (MSST) data were available. We then assessed how strongly MATTVs were related to stream temperature. We calculated MATTV as:

$$MATTV = \frac{\sum O_i \, TV_i}{\sum O_i}$$

where O_i represents either presence (1) or absence (0) or the abundance of taxon i at the site, and TV_i represents the thermal TV for taxon i. We then regressed each of these MATTVs against MSST. We calculated MATTVs as the simple average of the thermal TVs across taxa observed at a site and as the average of the taxon thermal TVs weighted by their abundance at a site. We also assessed if MATTV – temperature relationships calculated from sites with ≥ 10 taxa differed from those calculated from sites with ≥ 20 taxa and ≥ 30 taxa. We included this analysis because the number of macroinvertebrate taxa for which we could assign TVs varied across the reference sites and errors in estimating MATTV might be sensitive to the number of taxa included in the calculations. We also regressed both the minimum observed thermal TV (MinTV) and the maximum observed thermal TV (MaxTV) observed at sites on MSST to assess if alternative ways of characterizing assemblage-level thermal tolerance differed from MATTV in their association with environmental temperature.

To assess if differences among MATTVs were sensitive to the specific temperature metric used to estimate MATTV, we regressed MATTVs against three different stream

temperature metrics available for an independent set of samples. For this analysis, we used a third dataset collected by the Pacfish/Infish Biological Opinion Monitoring Program (PIBO; Henderson et al., 2005). The PIBO data include estimates of macroinvertebrate occurrence and abundance as well as measurements of hourly summer stream temperatures (°C) from approximately 1300 streams in the western United States. We used a subset of the dataset that included 1000 unique sites that were sampled between 2001 and 2017 and days 150 and 285 of each year, for which hourly temperature measurements were available over the summer period 15 July to 31 August. Each sample included at least 10 different macroinvertebrate taxa for which we had TV estimates. We used the hourly temperature data to characterize the thermal environments of each site in three ways: mean summer temperature (MST), maximum temperature, and maximum mean weekly maximum temperature (MWMT). MST and maximum temperature are the mean and maximum temperature, respectively, across all hourly recordings. MWMT is the highest value of averaged weekly maximum temperatures (WMT) calculated over all possible continuous 7-day periods between 15 July and 31 August. MST is the temperature metric that most closely matched the predicted mean summer stream temperature (MSST) metric used in analysis of the NRSA data. We also used the PIBO dataset to assess if TVs derived from CONUS-level data were less sensitive when applied to a geographically-restricted subset of streams in the CONUS.

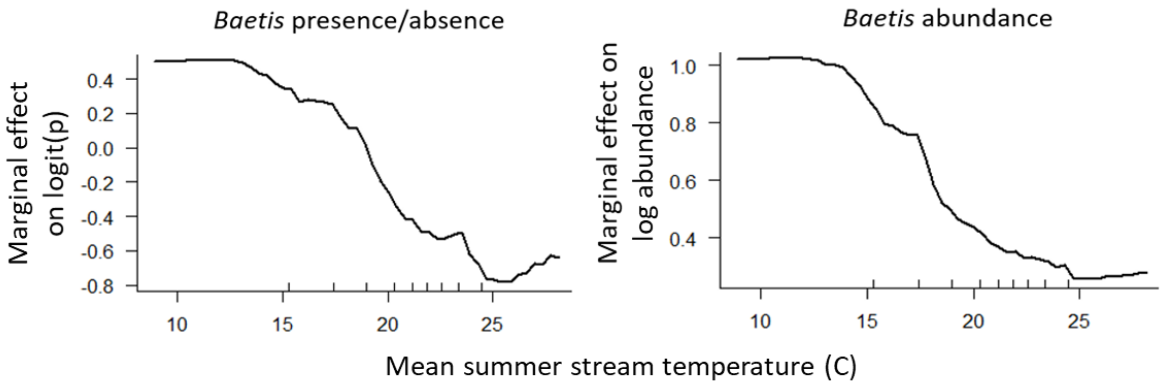


Fig. 5-2. Example of partial dependence plots used to infer thermal upper limits for each taxon based on presence/absence and abundance data. MSST = mean summer stream temperature (°C). The inner tick marks along the x-axis (the rug) show how each 10% of the observations are distributed across the temperature gradient.

2.4. Do MATTVs respond quickly to interannual variation in stream temperature?

To address the third research question, we identified PIBO sites that had two repeat samples (collected in nonconsecutive years). We assessed responsiveness by regressing the between-year differences in MATTVs against the between-year differences in site temperature (separately for MST, maximum temperature, and MWMT). We used the A (p/a) TVs to derive MATTVs because these TVs performed well and are simple to derive and interpret. Additionally, we calculated MATTVs weighted and not weighted by taxon abundance to assess if incorporating abundance data led to more responsive MATTVs.

2.5. Are thermal TVs specific enough that TBIs can isolate temperature-caused alteration of stream macroinvertebrate assemblages?

To address the fourth research question, we built a regression random forest model to assess how strongly MATTVs were associated with MSST and three other variables (Table 5-1) that could potentially confound interpretation of MATTVs. For this analysis, we calculated MATTVs and used environmental data from NRSA 2013-2014 reference sites. We assessed importance as the rank order of each variable in predicting variation in MATTVs (measured as

the increase in mean squared error that resulted from randomly permuting each variable and recomputing the model error). We also examined partial dependence plots for each predictor to visually assess the degree of potential confounding that existed between MSST and the other environmental variables.

2.6. A CONUS-level application of the TBI

To demonstrate the applicability of the TBI in detecting thermal alteration of aquatic life across large spatial extents, we applied the A (p/a) and PDP (abund) based TBIs to a second set of NRSA probability-based samples that were collected in 2013-2014. Following sampling, NRSA staff classified these sites as either being in most degraded, intermediate degraded, or reference condition based on land use and water chemistry analyses (USEPA 2020b). We calculated cumulative distribution functions of TBI scores for reference, most degraded, and all probability-based sites and used the 5^{th} and 95^{th} percentiles of reference site TBI scores as threshold values to assess how many most degraded and probability-based sites the TBI would diagnose as being thermally altered. Each probability-based site in the NRSA 2013-2014 dataset is accompanied by an estimated weight, which indicates the length of stream that it represents across the entire CONUS (USEPA, 2020b). For example, a site accompanied by a weight of 100, would represent 100 km of stream length. These weights allow for the extrapolation of observations from the probability-based samples to the entire population of streams and rivers within the CONUS. Thus, we could estimate the percentage of streams and the total length of streams within the CONUS that had TBI values implying they were cooler than, equivalent to, or warmer than expected. We used the *spsurvey* package in R to calculate lengths of streams and rivers in different categories (Kincaid et al., 2019).

3. Results

3.1. How important is temperature to distributions relative to other potential drivers?

Of the variables that we examined, temperature was the most important predictor of macroinvertebrate distributions. Temperature (MSST) was the most frequently ranked most important predictor (155 of 290 models) across SDMs for models built with presence/absence data. (Fig. 5-3). Additionally, MSST was only weakly correlated with the other predictors (range in r = -0.34 to 0.28, Fig. 5-4) implying little potential confounding. Substrate was the second most frequently ranked most important predictor (65 of 290 models), and conductivity was the third most frequently ranked most important predictor (55 of 290 models). The day of year (DOY) that samples were collected, was least often the most important predictor (15 of 290 models).

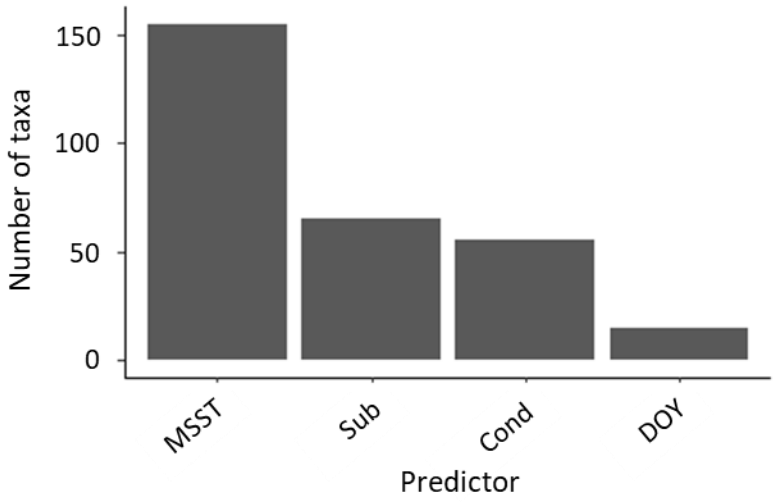


Fig. 5-3. Number of taxa for which each predictor was most important in predicting distribution based on variable importance metrics. Results are shown for SDMs built with presence/absence data. We modeled 290 taxa. MSST = mean predicted summer stream temperature, Cond = conductivity, Sub = substrate mean diameter, and DOY = day of year a sample was collected.

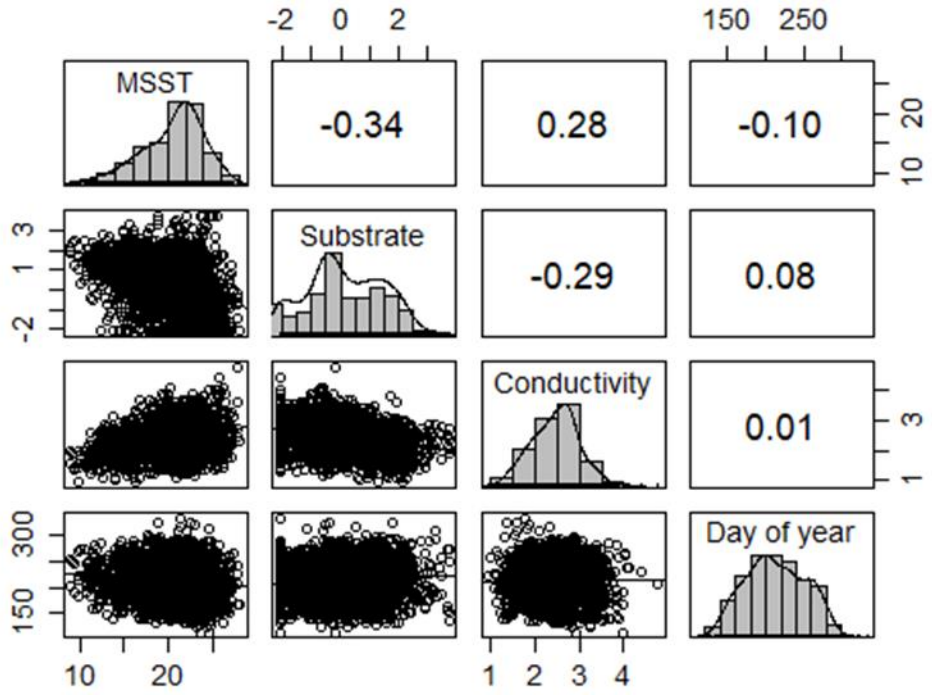


Fig. 5-4. Scatterplots comparing the relationships among the 4 predictors from the NRSA 0809 samples (n = 2142) that were used to build species distribution models. Conductivity was $\log_{10}$ transformed to better show the trend. Numbers in the boxes to the right of the frequency histograms are the Pearson's correlation coefficient for each predictor comparison.

3.2. Comparison of methods for deriving thermal TVs

All six TVs were strongly correlated (all pairwise $r \geq 0.84$, Fig. 5-5), but the range and distribution of values varied. For each set of TVs, the range was slightly larger for values derived from abundance data than presence/absence data. Each of the distributions of thermal TVs was left skewed indicating taxa were more frequently assigned TVs higher in the range. Distributions of TVs derived from the partial dependence plots were the most strongly skewed (because of the high number of taxa assigned a TV of 28°C) and differed the most from the other distributions. The number of taxa for which partial dependence plot TVs could be assigned was also lower than for the other methods because partial dependence plots did not always reveal observable trends. The TVs derived from optima (average and weighted average TVs) were shifted toward lower

temperatures relative to TVs based on upper limits (95th percentile and partial dependence plot TVs).

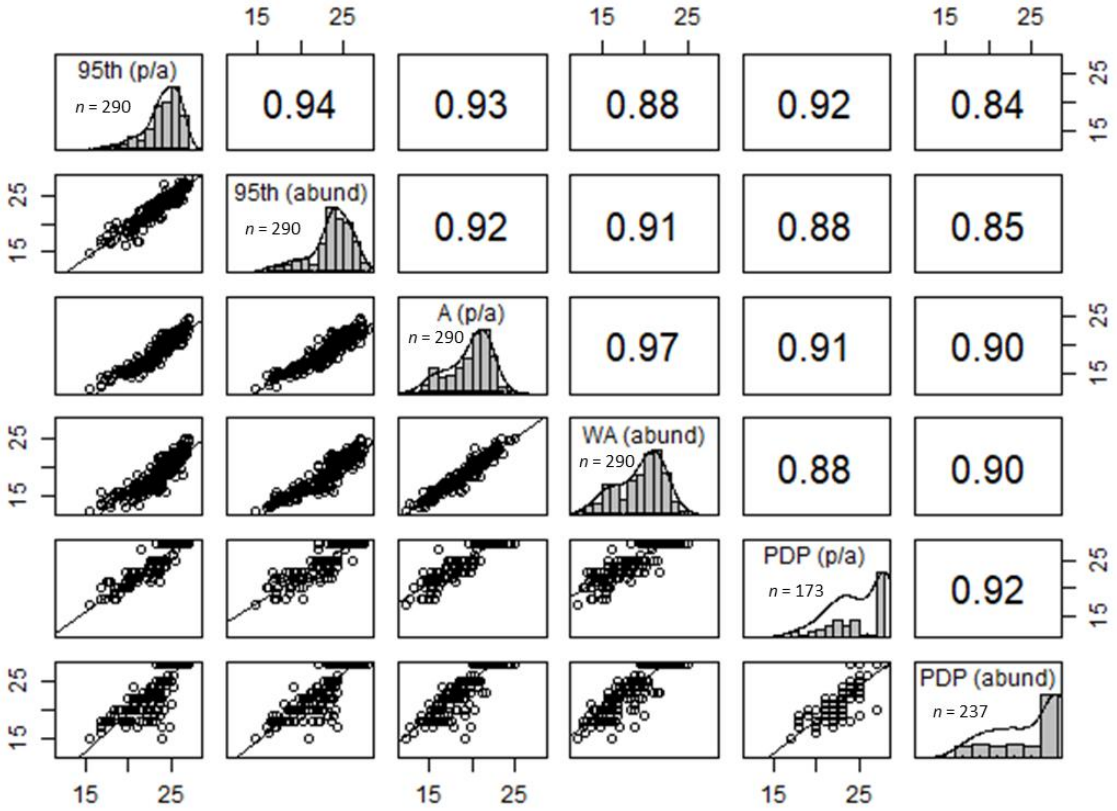


Fig. 5-5. Scatterplots comparing the relationships among the six tolerance values and frequency histograms showing the taxon assigned tolerance value distributions. The specific tolerance values are 95th = 95th percentile, A = average, WA = weighted average, PDP = partial dependence plot. The type of data used is specified in parentheses and is either p/a = presence/absence or abund = abundance. The number of taxa for which each tolerance value could be assigned = *n*. Numbers in the boxes to the right of the histograms are the Pearson's correlation coefficient for each tolerance value comparison.

3.3. Does the method of estimating TVs affect the relationship between MATTVs and site temperature?

The strength of the relationships between MATTVs (averaged across assemblages at reference quality sites with ≥ 10 taxa from the NRSA 2013-2014 dataset) and MSSTs varied only slightly among the six methods of estimating the thermal TVs. All six TVs resulted in relationships between MATTVs and MSSTs with r^2 = 0.72 to 0.75 (Fig. 5-6). The strength of the relationships between MATTVs and MSSTs did not increase when the number of taxa per site

with thermal TVs was increased to ≥ 20 taxa or ≥ 30 taxa (Appendix Fig. F.1 and F.2). The relationships between abundance-weighted MATTVs and MSSTs were consistently weaker than those for non-abundance weighted MATTVs (Appendix Fig. F.3). Compared with MATTVs, both MinTVs and MaxTVs (based on A (p/a) TVs) were less strongly related to MSSTs. However, the MinTVs were more strongly related to MSSTs than were MaxTVs (r^2 = 0.65 and 0.59, respectively, Fig. 5-7).

The slope of the best fit line through the regression of MATTVs on MSSTs varied substantially among the 6 methods of estimating TVs (range 0.32 – 0.59, Fig. 5-6). PDP (abund) TVs had the highest slope (0.59), and 95th (p/a) TVs had the lowest slope (0.32).

The strength of the relationship between MATTVs and stream temperature at PIBO sites revealed little effect of the particular temperature metric used, but the strength of the relationships between MATTVs and stream temperatures were markedly weaker than for the relationships based on CONUS-level data. Analyses of MATTVs (with A (p/a) TVs) calculated from the PIBO dataset showed that the strength of the relationship between MATTVs and site temperatures was similar for each of the three temperature metrics (range of r^2 = 0.40 to 0.42, Fig. 5-8). However, the relationship between each of the six MATTVs and MSTs for the PIBO sites (range of r^2 = 0.19 to 0.40 Appendix Fig. F.4) was lower than for the NRSA sites (range of r^2 = 0.72 to 0.75 Fig. 5-6).

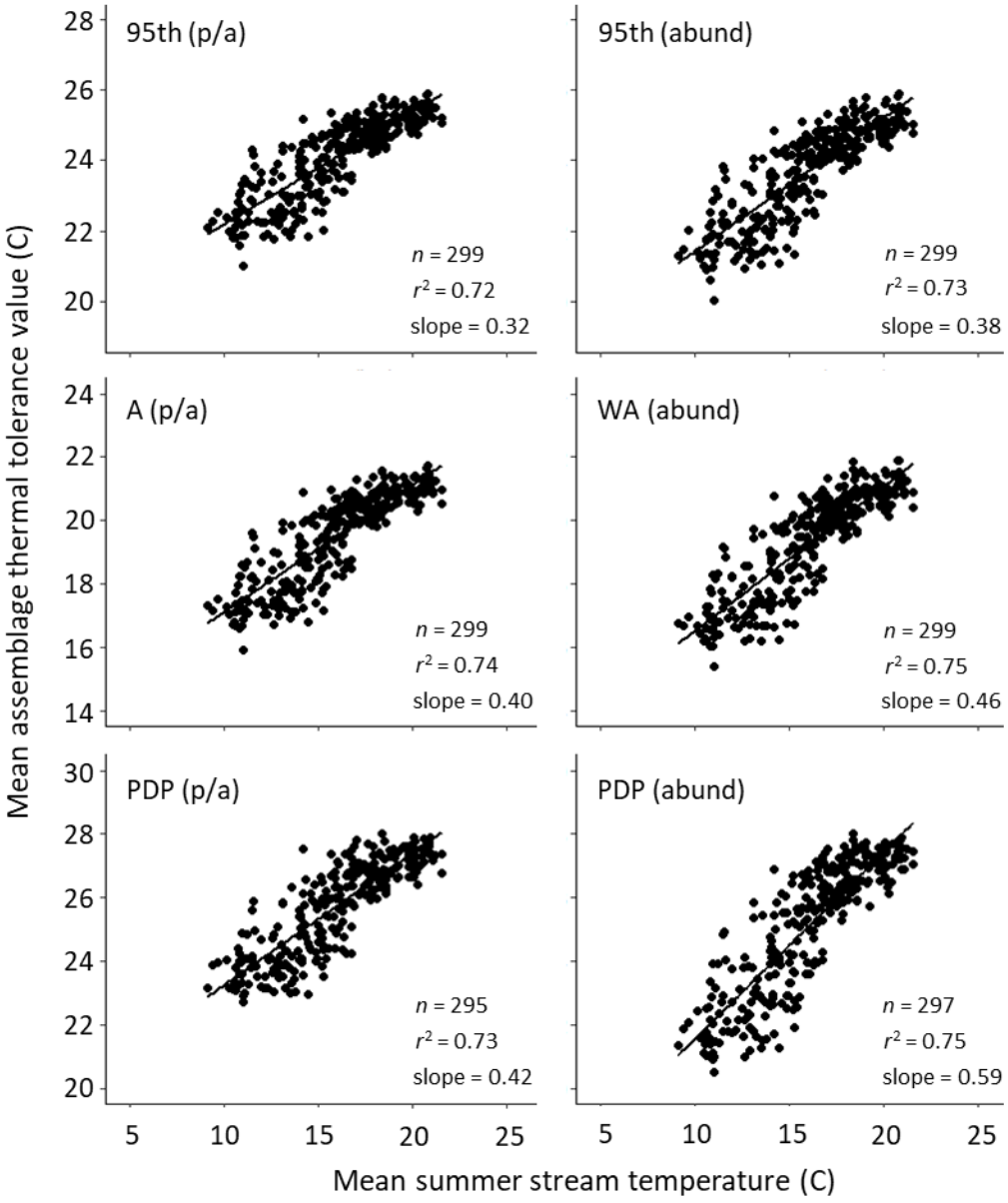


Fig. 5-6. Relationships between the six different MATTVs and predicted mean summer stream temperatures. All sites (*n*) were reference condition sites from the NRSA 2013-2014 dataset, and all sites had assemblages with at least 10 taxa with associated tolerance values. The specific tolerance values used were based on several methods of estimating TVs: 95^{th} = 95^{th} percentile, A = average, WA = weighted average, PDP = partial dependence plot. The type of data used is specified in parentheses and is either p/a = presence/absence or abund = abundance.

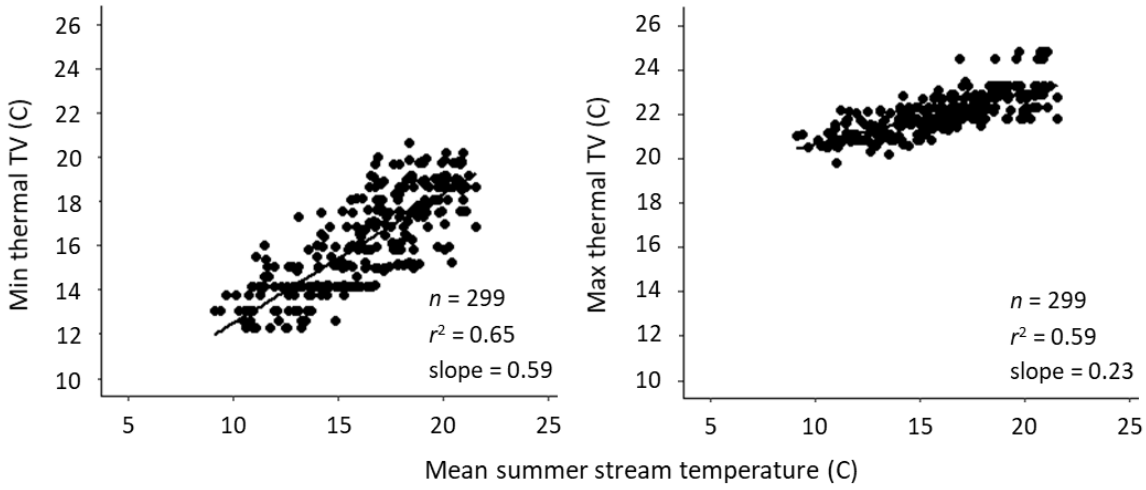


Fig. 5-7. Relationships between minimum (left) and maximum (right) thermal tolerance values and predicted mean summer stream temperatures where the assemblages were sampled. All sites (*n*) were reference condition sites from the NRSA 2013-2014 dataset, and TVs were derived with the average (presence/absence) tolerance values method. All sites had at least 10 taxa with assigned tolerance values.

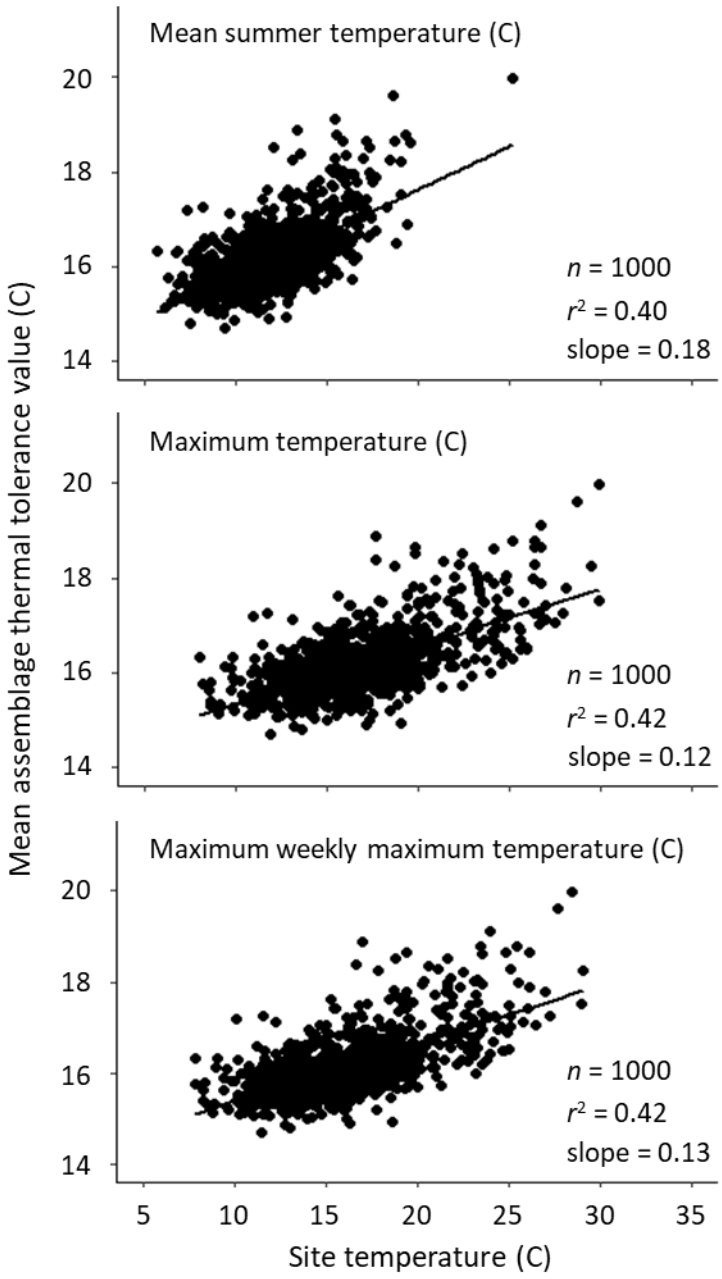


Fig. 5-8. Relationship between MATTVs and mean summer site temperature, maximum site temperature, and maximum weekly maximum site temperature for 1000 PIBO sites. TVs were estimated with the average (presence/absence) method. All sites had at least 10 taxa with assigned tolerance values.

3.4. Responsiveness of MATTVs to interannual variation in stream temperature

No association existed between change in site temperature and change in MATTV between years at PIBO sites (Fig. 5-9). Weighting MATTVs by site abundances did not increase the association (Appendix Fig. F.5).

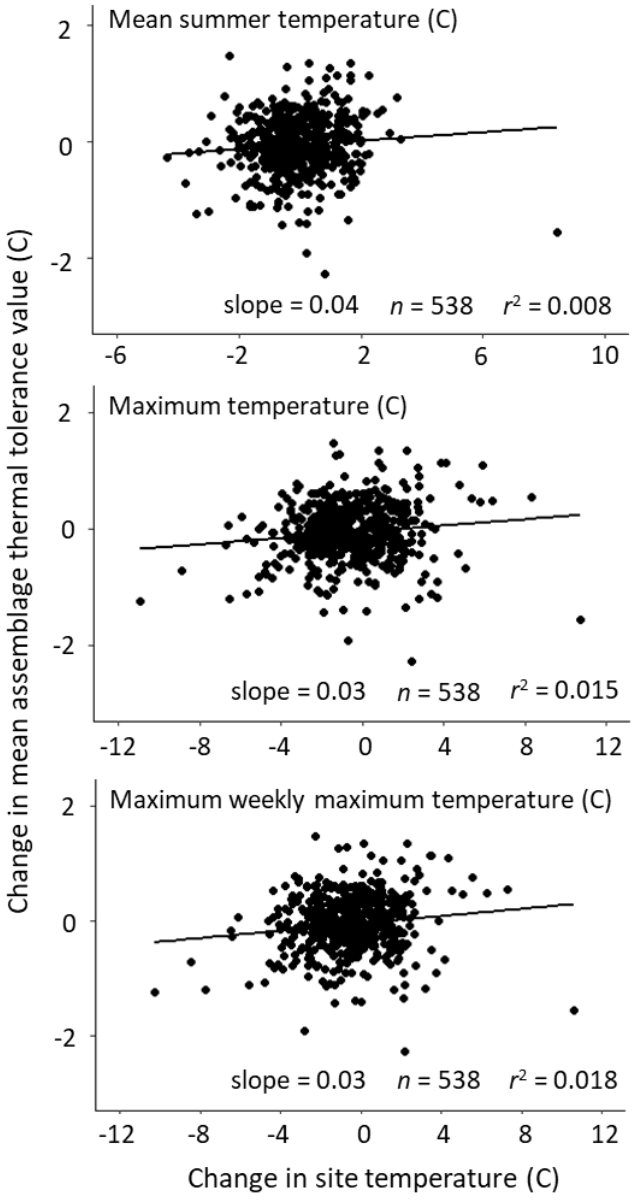


Fig. 5-9. Relationships between change in MATTVs and change in mean summer site temperature, maximum site temperature, and maximum weekly maximum temperature for 538 PIBO sites that were sampled in two different years. TVs were estimated with the average (presence/absence) method and all samples had assemblages with at least 10 taxa with assigned TVs.

3.5. Are thermal TVs specific enough that TBIs can isolate temperature-caused alteration of stream macroinvertebrate assemblages?

The TVs were generally specific to spatial variation in stream temperature. MSST was the most important predictor of MATTV, and its removal accounted for a greater than five-fold increase in model mean squared error compared to the next most important predictor (Fig. 5-10). Partial dependence plots for each predictor in the model also showed that MSST had the largest marginal effect on predicted MATTV (Fig. 5-11).

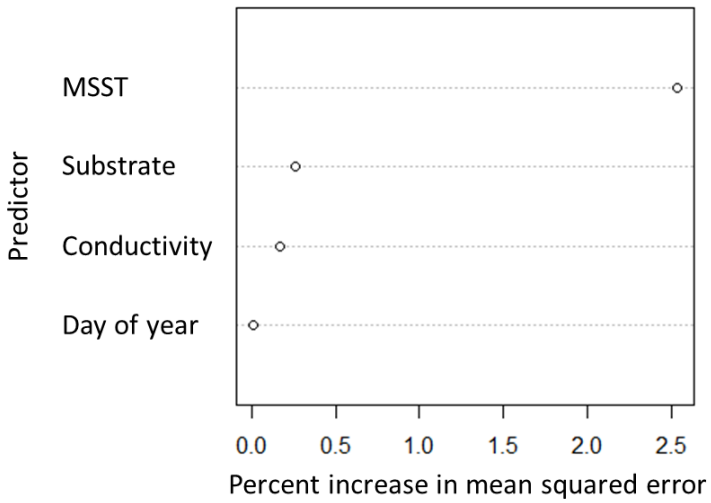


Fig. 5-10. The variable importance plot for the random forest model predicting MATTV from four predictors. The model was built with data from 299 NRSA 2013-2014 reference condition sites. MSST is predicted mean summer stream temperature.

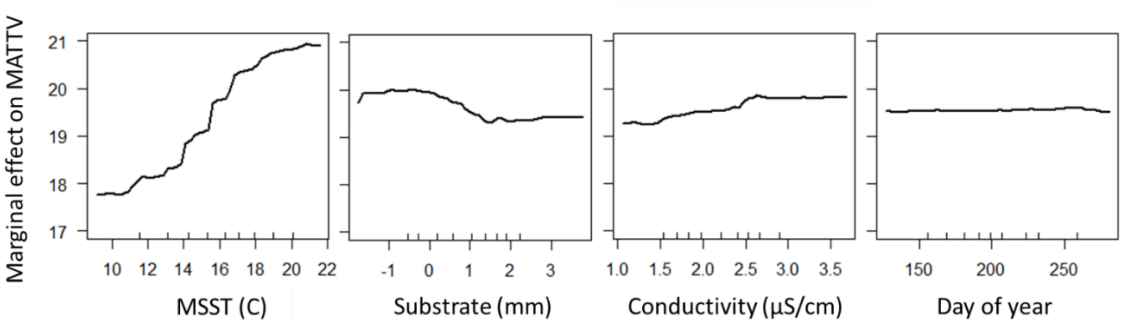


Fig. 5-11. Partial dependence plots showing the marginal effect of mean summer stream temperature (MSST), substrate, conductivity, and day of year on predicted MATTV. The partial dependence plots are based on a random forest model built with NRSA 2013-2014 data from 299 reference-condition sites. Substrate and conductivity were $\log_{10}$ transformed to enhance visualization of the responses.

3.6. A CONUS-level application of the TBI

The distribution of TBI values across the CONUS implied stream and rivers were generally warmer than expected. The cumulative distribution of TBI values for all streams was shifted toward higher values compared with reference-quality streams (Fig. 5-12), and values at degraded streams were shifted more markedly than for all streams (Fig. 5-12). Of total stream length, 7.6% (~138,500 km) was classified as warmer than expected, whereas only 5% was expected by chance alone (Table 5-2) (5% was expected by chance alone because our threshold was based on the 95th percentile of reference site TBI scores). Estimates of extents derived from a TBI based on TVs calculated from abundance-based SDMs yielded similar results (Appendix Table F.1).

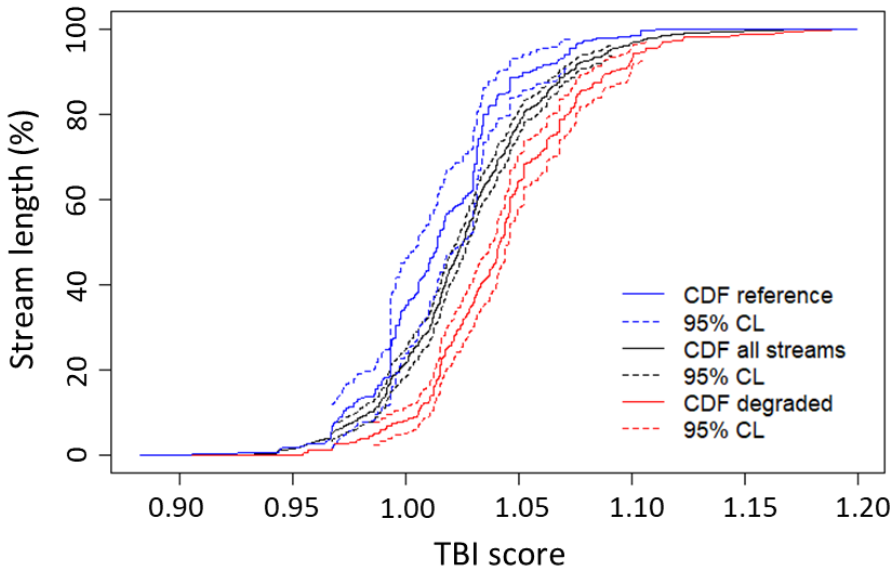


Fig. 5-12. Estimated cumulative distribution functions of TBI scores for reference streams (202,566 km), degraded streams (579,411 km), and all streams (1,814,925 km) across the CONUS.

Table 5-2 Extent estimates of TBI values across the CONUS. The TBI was calculated with the A (p/a) TVs. Cooler than expected means the TBI score was less than the 5th percentile of reference site TBI scores and warmer than expected means the TBI score was greater than the 95th percentile of reference site TBI scores. Expected means the TBI score falls within the 5th and 95th percentiles of reference site TBI scores and shows little evidence that the invertebrate assemblage has been thermally altered.

Stream class	Stream length (%)	SE	Stream length (km)	SE
Cooler	3.8	0.7	69,107	11,901
Expected	88.6	1.0	1,607,309	54,330
Warmer	7.6	0.8	138,509	15,036

4. Discussion

The thermal regimes of aquatic ecosystems are changing at unprecedented rates (Kaushal et al., 2010; Isaak et al., 2012; Hare et al., 2021). Therefore, assessing the importance of temperature in affecting aquatic assemblages is critical. Furthermore, effectively monitoring and mitigating these changes to aquatic life requires diagnostic tools that are sensitive to the effects of temperature on aquatic life. To be diagnostically useful, these tools also need to be capable of

isolating the effects caused by temperature from the effects caused by other stressors. Discerning the timeframe over which macroinvertebrate assemblages respond to changes is also critical to evaluating whether a diagnostic index is a potentially useful management tool and under what circumstances. For example, if there is a large lag between the shift in the thermal regime and the corresponding shift in the composition of the aquatic assemblage, a diagnostic index would not be very useful on short time scales. Finally, evaluating the spatial scale over which a diagnostic index can be applied is needed to determine if a single diagnostic index can be developed and applied over large areas or if a more targeted approach to smaller areas is more appropriate.

4.1. Importance of temperature

Our observation that temperature was most frequently the most important predictor of aquatic macroinvertebrate distributions is consistent with the results of several other studies (Sweeney and Vannote, 1978; Burgmer et al., 2007; Domisch et al., 2013; Bradie and Leung, 2017). For example, Sweeney and Vannote (1978) suggested that temperature effects on survival, growth, and fecundity play a large role in determining the distributional patterns of macroinvertebrate species. Numerous other studies have used species distribution models based on strong relationships between species distributions and temperature to forecast possible range shifts and extinctions caused by climate change (Kearney et al., 2010; Domisch et al., 2013; Li et al., 2013; Pyne and Poff, 2017). The importance of temperature in determining where species can persist coupled with the pervasive effects human activities have had on the thermal regimes of aquatic ecosystems support the need to develop diagnostic tools that can be used to isolate the effects of temperature on aquatic life.

4.2. Sensitivity of MATTVs to method of estimating TVs

We found that all six methods of estimating thermal tolerance yielded TVs that were similarly responsive and good candidates for incorporation into a TBI. Yuan (2006) also found that thermal TVs derived from weighted averages, cumulative percentile upper limits, generalized

linear models, and additive models produced similarly sensitive TVs (range in MATTV-temperature relationships: r^2 = 0.49 to 0.56). In a review of the effectiveness of using assigned traits for stressor assessment, Hamilton et al. (2020) reviewed five studies and found that the state of thermal preference traits (for aquatic life) matched predictions based on the state of the climate (e.g., warmer thermal preferences were associated with a warmer climatic state), which also suggests that TBIs should be potentially useful tools in aquatic ecosystem monitoring and management. The responsiveness and sensitivity of thermal traits (like MATTV) almost certainly depends on the taxonomic resolution with which thermal traits can be assigned. In our case, we derived TVs for genus and above levels of taxonomic resolution from a national dataset. However, intraspecific differences in thermal tolerance among local and regional populations are well documented (Feminella and Matthews, 1984; Huff et al., 2005; Stitt et al., 2014), especially among more isolated subpopulations (Eliason et al., 2011) indicating that TBIs would ideally be based on TVs derived from the highest taxonomic-resolution data possible. For example, Huff et al. (2005) found that the upper thermal limit for rainbow trout (*Oncorhynchus mykiss*), derived from field data, differed by as much as 5.5 °C among ecoregions in Oregon. Thermal tolerances among species in the same genus can be even more variable (Hildrew and Edington, 1979; Richards et al., 2013; Hamilton et al., 2020). As our ability to identify freshwater invertebrates to more resolved levels improves, the accuracy and precision of TVs and TBIs should also improve.

4.3. MATTV responsiveness to interannual changes in temperature

MATTVs were strongly associated with spatial variation in stream temperatures but did not appear to respond quickly to interannual variation in stream temperature at individual sites. This apparent lack of temporal responsiveness could have occurred because changes in mean annual temperature at the sites we studied were generally ± 2 °C and perhaps too small to elicit detectable responses in assemblage composition across the time intervals for which we had data (sites sampled twice between 2001 and 2017 in nonconsecutive years). In contrast, high-intensity

thermal disturbances that quickly surpass species tolerance limits have been shown to quickly alter assemblages (Voelz et al., 1994; Miller et al., 2007). For example, Voelz et al. (1994) observed that several species of thermally intolerant caddisflies were nearly extirpated below a reservoir immediately following an unexpected increase in water temperature that exceeded the normal maximum summer temperatures by greater than 4° C. Such marked between year-to-year differences in temperature did not generally occur in the PIBO dataset. In addition, the sites that we analyzed were sampled in two nonconsecutive years and the number of years between samples varied. It is possible that the year or years between survey years varied randomly in temperatures, which would obscure any strong directional response between the two years for which we had data. Monk et al. (2008) used a more complete dataset of 11 consecutive years of macroinvertebrate assemblages and flow to show that a flow biotic index (see Extence et al., 1999 for details) generally tracked inter-annual changes in flow, especially around drought years.

It is also possible that time lags associated with dispersal constraints may limit how quickly aquatic macroinvertebrate assemblages can track environmental changes in general (Parkyn and Smith, 2011; Heino, 2013; Sarremejane et al., 2017). For example, dispersal constraints can delay the reestablishment of aquatic communities following restoration efforts (Bond and Lake, 2003; Blakely et al., 2006; Lake et al., 2007; Parkyn and Smith, 2011, Tonkin et al. 2014). For context, Clements et al. (2021) reported that macroinvertebrate assemblages took 10 – 15 years to recover from severe metal pollution following the beginning of remediation efforts. The speed of recovery appeared to be affected by not only the time it took to reduce metal contamination but also the availability of nearby colonization sources. In our study, dispersal constraints coupled with modest and non-unidirectional between-year differences in temperature at most sites probably limited how quickly shifts in TBIs can occur. Analyses of longer-term data sets that include both larger between-year temperature differences, clear directional trends in temperature, or both are needed to better assess the responsiveness of TBIs to thermal alteration.

4.4. Thermal TV specificity

The thermal TVs we developed appear to be specific enough to temperature to isolate temperature-caused alteration of stream macroinvertebrate assemblages from the effects of other potential stressors. Of the factors that we examined, MSST was by far most closely associated with, and predictive of, MATTVs (Fig. 5-10). If a strong association existed between any of the other predictors (substrate, conductivity, or DOY), it would be difficult to infer with any certainty that changes in TBI values were caused by changes in temperature. Other stressor-specific indices appear to suffer from lack of specificity or specificity was not assessed at all. For example, Bray et al. (2020) concluded that a pesticide-specific biotic index also responded to other stressors associated with agriculture, limiting its effectiveness at isolating the effects of pesticides. Attempts to develop flow-specific biotic indices have also encountered some degree of confounding with physical habitat characteristics (see Lotic-invertebrate Index for Flow Evaluation (LIFE), Extence et al., 1999) and measures of water quality (see Armanini et al., 2011; Laini et al., 2018). It has even been suggested that physical habitat surveys should accompany the use of LIFE scores to assess the effects of flow, to avoid a confounded interpretation (Dunbar et al., 2010). Thoroughly evaluating the specificity of stressor-specific biotic indices is critical to their interpretation. In this study, we only examined three potentially confounding factors, and other, unmeasured stressors or naturally-occurring environmental conditions may have resulted in undetected confounding. Further examination of the specificity of stressor-specific biotic indices with different datasets and a broader range of potential stressors will strengthen our confidence in their interpretations (Blanck, 2005).

4.5. Detection of thermal alteration of aquatic life with large-scale survey data

When applied to streams and rivers across the CONUS, our TBI revealed a trend toward assemblages with higher thermal tolerance than expected under reference condition (i.e., mean difference between observed and expected MATTVs of 0.52 ºC and 2.6% of stream and river

length across the CONUS inferred as warmer than expected). This observed shift in assemblage-level thermal tolerance values across the CONUS is in line with observed trends showing increasing stream and river water temperatures (Kaushal et al., 2010; Isaak et al., 2012). For example, Kaushal et al. (2010) found that long-term (24 to 100 years) temperature data showed moderate to dramatic trends in warming (0.009 – 0.077 °C yr^{-1}) for 20 of 40 streams and rivers that they examined. Isaak et al. (2012) examined temperature data over three decades for seven unregulated stream and river sites in the western USA and found that summer water temperatures were increasing by approximately 0.2 °C per decade.

Numerous, and co-occurring, factors may be the cause of warming trends in certain streams and rivers. For example, broadscale landscape alteration from timber harvest, urban development, and agriculture is a major source of altered thermal regimes. Effects of these landscape level alterations are also well suited to being monitored by a temperature specific index. Other causes of altered thermal regimes, such as climate change, may be more difficult to monitor with a TBI that compares observed assemblage thermal tolerance with expected assemblage thermal tolerance based on reference conditions. However, a TBI can still be used to assess change in assemblage thermal tolerance, and avoid the potential effects of a shifting reference condition baseline, by anchoring the reference condition of the TBI at some standard time in the past. Thus, a temperature specific biotic index is a promising tool for diagnosing thermally-caused alteration of aquatic life.

References

Armanini, D.G., Horrigan, N., Monk, W.A., Peters, D.L., Baird, D.J., 2011. Development of a benthic macroinvertebrate flow sensitivity index for Canadian rivers. River Research and Applications. 27, 723-737.

Armitage, P.D., Moss, D., Wright, J.F., Furse, M.T., 1983. The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Water Research. 17, 333-347.

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling. 157, 101-118.

Berger, E., Haase, P., Kuemmerlen, M., Leps, M., Schaefer, R.B., Sundermann, A., 2017. Water quality variables and pollution sources shaping stream macroinvertebrate communities. Science of the Total Environment. 587, 1-10.

Blakely, T.J., Harding, J.S., McIntosh, A.R., Winterbourn, M.J., 2006. Barriers to the recovery of aquatic insect communities in urban streams. Freshwater Biology, 51, 1634-1645.

Blanck, H., 2002. A critical review of procedures and approaches used for assessing pollution-induced community tolerance (PICT) in biotic communities. Human and Ecological Risk Assessment. 8, 1003-1034.

Bond, N.R., Lake, P.S., 2003. Local habitat restoration in streams: constraints on the effectiveness of restoration for stream biota. Ecological Management & Restoration. 4, 193-198.

Bradie, J., Leung, B., 2017. A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. Journal of Biogeography. 44, 1344-1361.

Bray, J.P., O'Reilly-Nugent, A., King, G.K.K., Kaserzon, S., Nichols, S.J., Nally, R.M., Thompson, R.M., Kefford, B.J., 2020. Can SPEcies At Risk of pesticides (SPEAR) indices detect effects of target stressors amongst multiple interacting stressors? Science of The Total Environment. 142997.

Breiman, L., 2001. Random forests. Machine Learning. 45, 5-32.

Brown, G.W., Krygier, J.T., 1970. Effects of clear-cutting on stream temperature. Water resources research. 6, 1133-1139.

Burgmer, T., Hillebrand, H., Pfenninger, M., 2007. Effects of climate-driven temperature changes on the diversity of freshwater macroinvertebrates. Oecologia. 151, 93-103.

Burrows, M.T., Hawkins, S.J., Moore, J.J., Adams, L., Sugden, H., Firth, L., Mieszkowska, N., 2020. Global-scale species distributions predict temperature-related changes in species composition of rocky shore communities in Britain. Global Change Biology. 26, 2093-2105.

Caissie, D., 2006. The thermal regime of rivers: a review. Freshwater biology. 51, 1389-1406.

Carlo, M.A., Riddell, E.A., Levy, O., Sears, M.W., 2018. Recurrent sublethal warming reduces embryonic survival, inhibits juvenile growth, and alters species distribution projections under climate change. Ecology Letters. 21, 104-116.

Carpenter, S.R., Stanley, E.H., Vander Zanden, M.J., 2011. State of the world's freshwater ecosystems: physical, chemical, and biological changes. Annual review of Environment and Resources. 36, 75-99.

Chawla, N.V., Japkowicz, N., Kotcz, A., 2004. Special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter. 6, 1-6.

Chessman, B.C., McEvoy, P.K., 1997. Towards diagnostic biotic indices for river macroinvertebrates. Hydrobiologia. 364, 169-182.

Chutter, F.M., 1972. An empirical biotic index of the quality of water in South African streams and rivers. Water research. 6, 19-30.

Clarke, R.T., Wright, J.F., Furse, M.T., 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecological Modelling. 160, 219-233.

Clarkson, R.W., Childs, M.R., 2000. Temperature effects of hypolimnial-release dams on early life stages of Colorado River Basin big-river fishes. Copeia. 2000, 402-412.

Clements, W.H., Herbst, D.B., Hornberger, M.I., Mebane, C.A., Short, T.M., 2021. Long-term monitoring reveals convergent patterns of recovery from mining contamination across 4 western US watersheds. Freshwater Science. 40, 407-426.

Craig, L.S., Olden, J.D., Arthington, A.H., Entrekin, S., Hawkins, C.P., Kelly, J.J., Kennedy, T.A., Maitland, B.M., Rosi, E.J., Roy, A.H. Strayer, D.L., 2017. Meeting the challenge of interacting threats in freshwater ecosystems: a call to scientists and managers. Elementa: Science of the Anthropocene, 5.

Cutler, D.R., Edwards Jr, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology. 88, 2783-2792.

Domisch, S., Araújo, M.B., Bonada, N., Pauls, S.U., Jähnig, S.C., Haase, P., 2013. Modelling distribution in European stream macroinvertebrates under future climates. Global Change Biology. 19, 752-762.

Domisch, S., Jähnig, S.C., Haase, P., 2011. Climate-change winners and losers: stream macroinvertebrates of a submontane region in Central Europe. Freshwater Biology. 56, 2009-2020.

Dunbar, M.J., Pedersen, M.L., Cadman, D.A.N., Extence, C., Waddingham, J., Chadd, R., Larsen, S.E., 2010. River discharge and local-scale physical habitat influence macroinvertebrate LIFE scores. Freshwater Biology. 55, 226-242.

Durance, I., Ormerod, S.J., 2009. Trends in water quality and discharge confound long-term warming effects on river macroinvertebrates. Freshwater Biology. 54, 388-405.

Eliason, E.J., Clark, T.D., Hague, M.J., Hanson, L.M., Gallagher, Z.S., Jeffries, K.M., Gale, M.K., Patterson, D.A., Hinch, S.G., Farrell, A.P., 2011. Differences in thermal tolerance among sockeye salmon populations. Science. 332, 109-112.

Elith, J., and Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics. 40, 677-697.

Extence, C.A., Balbi, D.M., Chadd, R.P., 1999. River flow indexing using British benthic macroinvertebrates: a framework for setting hydroecological objectives. Regulated Rivers:

Research & Management: An International Journal Devoted to River Research and Management. 15, 545-574.

Feld, C.K., Saeedghalati, M., Hering, D., 2020. A framework to diagnose the causes of river ecosystem deterioration using biological symptoms. Journal of Applied Ecology. 57, 2271-2284.

Feminella, J.W., Matthews, W.J., 1984. Intraspecific differences in thermal tolerance of *Etheostoma spectabile* (Agassiz) in constant versus fluctuating environments. Journal of Fish biology. 25, 455-461.

Gore, J.A., 1977. Reservoir manipulations and benthic macroinvertebrates in a prairie river. Hydrobiologia. 55, 113-123.

Hamilton, A.T., Schäfer, R.B., Pyne, M.I., Chessman, B., Kakouei, K., Boersma, K.S., Verdonschot, P.F. M., Verdonschot, R.C.M., Mims, M., Khamis, K., Bierwagen, B., Stamp, J., 2020. Limitations of trait-based approaches for stressor assessment: the case of freshwater invertebrates and climate drivers. Global Change Biology. 26, 364-379.

Hare, D.K., Helton, A.M., Johnson, Z.C., Lane, J.W., Briggs, M.A., 2021. Continental-scale analysis of shallow and deep groundwater contributions to streams. Nature Communications. 12, 1-10.

Hawkins, C.P., Cao, Y., Roper, B., 2010. Method of predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biology. 55, 1066-1085.

Hawkins, C.P., Hogue, J.N., Decker, L.M., Feminella, J.W., 1997. Channel morphology, water temperature, and assemblage structure of stream insects. Journal of the North American Benthological Society. 16, 728-749.

Hawkins, C.P., Yuan, L.L., 2016. Multitaxon distribution models reveal severe alteration in the regional biodiversity of freshwater invertebrates. Freshwater Science. 35, 1365-1376.

Heino, J., 2013. Does dispersal ability affect the relative importance of environmental control and spatial structuring of littoral macroinvertebrate communities? Oecologia. 171, 971-980.

Heino, J., Virkkala, R., Toivonen, H., 2009. Climate change and freshwater biodiversity: detected patterns, future trends and adaptations in northern regions. Biological Reviews. 84, 39-54.

Henderson, R.C., Archer, E.K., Bouwes, B.A., Coles-Richie, M.C. Kershner, J.L., 2005. PACFISH/INFISH Biological Opinion (PIBO): Effectiveness Monitoring Program seven-year status report 1998 through 2004. Gen. Tech. Rep. RMRS-GTR-162. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 16 p.

Hildrew, A.G., Edington, J.M., 1979. Factors facilitating the coexistence of Hydropsychid caddis larvae (Trichoptera) in the same river system. The Journal of Animal Ecology. 48, 557-576.

Hill, R.A., Hawkins, C.P., 2014. Using modelled stream temperatures to predict macro-spatial patterns of stream invertebrate biodiversity. Freshwater Biology. 59, 2632-2644.

Hill, R.A., Hawkins, C.P., Carlisle, D.M., 2013. Predicting thermal reference conditions for USA streams and rivers. Freshwater Science. 32, 39-55.

Hill, R.A., Weber, M.H., Leibowitz, S.G., Olsen, A.R., Thornbrugh, D.J., 2016. The Stream-Catchment (StreamCat) dataset: a database of watershed metrics for the conterminous United States. Journal of the American Water Resources Association. 52, 120-128.

Hilsenhoff, W.L., 1987. An improved biotic index of organic stream pollution. The Great Lakes Entomologist. 20, 7.

Horrigan, N., Choy, S., Marshall, J., Recknagel, F., 2005. Response of stream macroinvertebrates to changes in salinity and the development of a salinity index. Marine and Freshwater Research. 56, 825-833.

Hubler, S., Huff, D.D., Edwards, P., Pan, Y., 2016. The biological sediment tolerance index: assessing fine sediments conditions in Oregon streams using macroinvertebrates. Ecological Indicators. 67, 132-145.

Huff, D.D., Hubler, S.L., Borisenko, A.N., 2005. Using field data to estimate the thermal niche of aquatic vertebrates. North American Journal of Fisheries Management. 25, 346-360.

Huff, D.D., Hubler, S.L., Pan, Y. and Drake, D.L., 2008. Detecting shifts in macroinvertebrate assemblage requirements: implicating causes of impairment in streams. Oregon Department of Environmental Quality Watershed Assessment. Technical Report: DEQ06-LAB-0068-TR.

Isaak, D.J., Wollrab, S., Horan, D., Chandler, G., 2012. Climate change effects on stream and river temperatures across the northwest US from 1980–2009 and implications for salmonid fishes. Climatic Change. 113, 499-524.

Kaushal, S.S., Likens, G.E., Jaworski, N.A., Pace, M.L., Sides, A.M., Seekell, D., Belt, K.T., Secor, D. H., Wingate, R.L., 2010. Rising stream and river temperatures in the United States. Frontiers in Ecology and the Environment. 8, 461-466.

Kearney, M.R., Wintle, B.A., Porter, W.P., 2010. Correlative and mechanistic models of species distribution provide congruent forecasts under climate change. Conservation Letters. 3, 203-213.

Kincaid, T.M., Olsen, A.R., Weber, M.H., 2019. spsurvey: Spatial Survey Design and Analysis. R package version 4.1.0.

Laini, A., Bolpagni, R., Cancellario, T., Guareschi, S., Racchetti, E., Viaroli, P., 2018. Testing the response of macroinvertebrate communities and biomonitoring indices under multiple stressors in a lowland regulated river. Ecological Indicators. 90, 47-53.

Lake, P.S., Bond, N., Reich, P., 2007. Linking ecological theory with stream restoration. Freshwater Biology. 52, 597-615.

Lamberti, G.A., Resh, V.H., 1985. Distribution of benthic algae and macroinvertebrates along a thermal stream gradient. Hydrobiologia. 128, 13-21.

Langford, T., 1990. Ecological Effects of Thermal Discharges. Springer Science & Business Media.

Lenat, D.R., 1993. A biotic index for the southeastern United States: derivation and list of tolerance values, with criteria for assigning water-quality ratings. Journal of the North American Benthological Society. 12, 279-290.

Lessard, J.L., Hayes, D.B., 2003. Effects of elevated water temperature on fish and macroinvertebrate communities below small dams. River Research and Applications. 19, 721-732.

Liaw, A., Wiener, M., 2002. Package randomForest.

Liess, M., Ohe, P.C.V.D., 2005. Analyzing effects of pesticides on invertebrate communities in streams. Environmental Toxicology and Chemistry: An International Journal. 24, 954-965.

Li, F., Chung, N., Bae, M.J., Kwon, Y.S., Kwon, T.S., Park, Y.S., 2013. Temperature change and macroinvertebrate biodiversity: assessments of organism vulnerability and potential distributions. Climatic Change. 119, 421-434.

McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., Reah, A., 2012. NHDPlus Version 2: User Guide. Accessed January 2021

Miller, S.W., Wooster, D., Li, J., 2007. Resistance and resilience of macroinvertebrates to irrigation water withdrawals. Freshwater Biology. 52, 2494-2510.

Monk, W.A., Wood, P.J., Hannah, D.M., Wilson, D. A., 2008. Macroinvertebrate community response to inter-annual and regional river flow regime dynamics. River Research and Applications. 24, 988-1001.

Moss, D., Furse, M.T., Wright, J.F., Armitage, P.D., 1987. The prediction of the macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. Freshwater Biology. 17, 41-52.

Moss, B., 2010. Climate change, nutrient pollution and the bargain of Dr Faustus. Freshwater Biology. 55, 175-187.

Murphy, J.F., Davy-Bowker, J., McFarland, B., Ormerod, S.J., 2013. A diagnostic biotic index for assessing acidity in sensitive streams in Britain. Ecological Indicators. 24, 562-572.

Murphy, J.F., Jones, J.I., Pretty, J.L., Duerdoth, C.P., Hawczak, A., Arnold, A., Blackburn, J.H., Naden, P.S., Old, G., Sear, D.A., Hornby, D., 2015. Development of a biotic index using stream macroinvertebrates to assess stress from deposited fine sediment. Freshwater Biology. 60, 2019-2036.

O'Keeffe, J., Hughes, D., Tharme, R., 2002. Linking ecological responses to altered flows, for use in environmental flow assessments: the flow stressor—response method. Internationale Vereinigung für theoretische und angewandte Limnologie: Verhandlungen. 28, 84-92.

Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling. 178, 389-397.

Olden, J.D., Naiman, R.J., 2010. Incorporating thermal regimes into environmental flows assessments: modifying dam operations to restore freshwater ecosystem integrity. Freshwater Biology. 55, 86-107.

Parkyn, S.M., Smith, B.J., 2011. Dispersal constraints for stream invertebrates: setting realistic timescales for biodiversity restoration. Environmental Management. 48, 602-614.

Pecl, G.T., Araújo, M.B., Bell, J.D., Blanchard, J., Bonebrake, T.C., Chen, I.C., Clark, T.D., Colwell, R.K., Danielsen, F., Evengard, B., Falconi, L., 2017. Biodiversity redistribution under climate change: impacts on ecosystems and human well-being. Science. 355, eaai9214.

Poole, G.C., Berman, C.H., 2001. An ecological perspective on in-stream temperature: natural heat dynamics and mechanisms of human-caused thermal degradation. Environmental Management. 27, 787-802.

Pyne, M.I., Poff, N.L., 2017. Vulnerability of stream community composition and function to projected thermal warming and hydrologic change across ecoregions in the western United States. Global Change Biology. 23, 77-93.

Raptis, C.E., van Vliet, M.T., Pfister, S., 2016. Global thermal pollution of rivers from thermoelectric power plants. Environmental Research Letters. 11, 104011.

R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Relyea, C.D., Minshall, G.W., Danehy, R.J., 2012. Development and validation of an aquatic fine sediment biotic index. Environmental Management. 49, 242-252.

Resh, V.H., 2008. Which group is best? attributes of different biological assemblages used in freshwater biomonitoring programs. Environmental Monitoring and Assessment. 138, 131-138.

Resh, V.H., Rosenberg, D.M., editors, 1993. Freshwater Biomonitoring and Benthic Macroinvertebrates. Chapman & Hall, New York, USA.

Richards, D.C., Bilger, M., Lester, G., 2013. Development of Idaho macroinvertebrate temperature occurrence models. Technical Report. Final Report to Idaho Department of Environmental Quality. Boise, Idaho.

Sarremejane, R., Mykrä, H., Bonada, N., Aroviita, J., Muotka, T., 2017. Habitat connectivity and dispersal ability drive the assembly mechanisms of macroinvertebrate communities in river networks. Freshwater Biology, 62, 1073-1082.

Schuwirth, N., Kattwinkel, M., Stamm, C., 2015. How stressor specific are trait-based ecological indices for ecosystem management? Science of The Total Environment. 505, 565-572.

Shah, D.N., Domisch, S., Pauls, S.U., Haase, P., Jähnig, S.C., 2014. Current and future latitudinal gradients in stream macroinvertebrate richness across North America. Freshwater Science. 33, 1136-1147.

Smith, A.J., Bode, R.W., Kleppel, G.S., 2007. A nutrient biotic index (NBI) for use with benthic macroinvertebrate communities. Ecological Indicators. 7, 371-386.

Stitt, B.C., Burness, G., Burgomaster, K.A., Currie, S., McDermid, J.L., Wilson, C.C., 2014. Intraspecific variation in thermal tolerance and acclimation capacity in brook trout (*Salvelinus fontinalis*): physiological implications for climate change. Physiological and Biochemical Zoology. 87, 15-29.

Sweeney, B.W., Vannote, R.L., 1978. Size variation and the distribution of hemimetabolous aquatic insects: two thermal equilibrium hypotheses. Science. 200, 444-446.

Sylvester, J.R., 1972. Possible effects of thermal effluents on fish: a review. Environmental Pollution. 3, 205-215.

Ter Braak, C.J., Looman, C.W., 1986. Weighted averaging, logistic regression and the Gaussian response model. Vegetatio. 65, 3-11.

Tonkin, J. D., Stoll, S., Sundermann, A., Haase, P. 2014. Dispersal distance and the pool of taxa, but not barriers, determine the colonisation of restored river reaches by benthic invertebrates. Freshwater Biology. 59, 1843-1855.

USEPA (US Environmental Protection Agency), 2016. National Aquatic Resource Surveys. National Rivers and Streams Assessment 2008-2009 (data and metadata files). Available from U.S. EPA website: http://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys. Date accessed: 2019-3-01.

USEPA (US Environmental Protection Agency), 2020a. National Aquatic Resource Surveys. National Rivers and Streams Assessment 2013-2014 (data and metadata files). Available from U.S. EPA website: http://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys. Date accessed: 2021-3-01.

USEPA (US Environmental Protection Agency), 2020b. National Rivers and Streams Assessment 2013-2014 Technical Support Document. EPA 843-R-19-001. Office of Water and Office of Research and Development. Washington, D.C. https://www.epa.gov/national-aquatic-resource-surveys/nrsa

Vannote, R.L., Sweeney, B.W., 1980. Geographic analysis of thermal equilibria: a conceptual model for evaluating the effect of natural and modified thermal regimes on aquatic insect communities. The American Naturalist. 115, 667-695.

Vinson, M.R., Hawkins, C.P., 1998. Biodiversity of stream insects: variation at local, basin, and regional scales. Annual Review of Entomology. 43, 271-293.

Voelz, N.J., Poff, N.L., Ward, J.V., 1994. Differential effects of a brief thermal disturbance on caddisflies (Trichoptera) in a regulated river. American Midland Naturalist. 132, 173-182.

Walther, G.R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T.J., Fromentin, J.M., Hoegh-Guldberg, O., Bairlein, F., 2002. Ecological responses to recent climate change. Nature. 416, 389-395.

Wright, J.F., 1995. Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. Australian Journal of Ecology. 20, 181-197.

Yuan, L.L., 2005. Sources of bias in weighted average inferences of environmental conditions. Journal of Paleolimnology. 34, 245-255.

Yuan, L.L., 2006. Estimation and application of macroinvertebrate tolerance values. US EPA, ORD, National Center for Environmental Assessment, Washington, DC. Technical Report (EPA/600/P-04/116F).

CHAPTER 6

CONCLUSION

My dissertation provides new insights into 1) experimentally validating the causal role that temperature plays in shaping distributions, 2) the application of several modeling techniques to potentially improve model performance, which have seldom been applied to species distribution modeling, and 3) the applicability of a stressor-specific biotic index for diagnosing thermal alteration of aquatic life. These insights should increase our understanding of the effects of temperature on aquatic ecosystems and improve our ability to model, predict, and diagnose effects of changing thermal regimes.

The chronic exposure laboratory experiments (>one week) I conducted in chapter two provide insight regarding a causal interpretation for the effects of temperature on species distributions. The association between upper thermal limits derived from longer-term survival and upper thermal limits derived from field data suggest that distributional constraints are in some part caused by limits to longer-term survival. Ideally, I would have obtained reliable growth data for all seven of my experimental macroinvertebrate species. Reliable growth data for all seven species would have allowed me to better assess the effectiveness of growth and a fitness index at predicting upper thermal limits to distributions compared with upper thermal limits based on survival alone. Still, the fact that longer-term measures of survival in the laboratory, at temperatures experienced in nature, provided a causal link with distributions is in some ways ideal because survival is an easy aspect of fitness to measure. Thus, chronic exposure laboratory experiments may be broadly applicable to assessing and validating the causes by which many potential environmental stressors affect distributions.

The two studies I described in chapters three and four provide insight regarding two approaches that may be useful for improving machine-learning models and predictions from temperature – distribution associations. Good models are essential, because they are commonly

applied to model and predict species distributions (Cutler et al. 2007), including with macroinvertebrate data for bioassessment purposes (e.g., Hawkins et al. 2010). In chapter 3, the systematic comparison of class imbalance-correction methods and machine-learning algorithms provided insight into the performance benefits of applying imbalance-correction methods when modeling imbalanced macroinvertebrate data. The results from chapter three showed that performance of machine-learning algorithm based SDMs, can be improved by applying imbalance-correction methods. In particular, when a balanced tradeoff between sensitivity and specificity are goals for the model, then imbalance-correction methods should be considered when building models with imbalanced datasets. Deep learning is another approach that has the potential to improve SDM performance relative to the methods that are currently used (Christin et al. 2019). The results in chapter four indicated that large datasets are required to train deep learning models, in order to avoid bad overfitting. As species-environment datasets in ecology continue to get larger due to automation and largescale collaborative efforts, deep learning approaches should continue to be considered and evaluated. The results in chapter four also showed that random forest performed as well or better than most deep learning models on the datasets examined. This good performance by random forest suggests that it is still a top choice for species distribution modeling.

The temperature-specific biotic index (TBI) I described in chapter five provides new insights into the design and applicability of a stressor-specific biotic index. The index had good specificity, indicated by the observation that mean assemblage thermal tolerances generally responded only to differences in temperature. However, we only assessed specificity relative to three other variables (conductivity, substrate, and day of year), thus additional variables should still be considered when assessing specificity of stressor-specific biotic indices. The index was also sensitive to spatial variation in temperature. This sensitivity was indicated by the observation that mean assemblage thermal tolerance was strongly related to predicted mean summer stream temperature, regardless of the method by which thermal tolerance values were derived. However,

the index was not sensitive to temporal variation in temperature, indicated by the lack of association between change in mean assemblage thermal tolerance and change in stream temperature. This finding was unexpected and leads to several pressing research questions: 1) what type (e.g., sustained warming or short-term thermal disturbance) and magnitude of thermal change is needed to elicit an assemblage level response and 2) over what timescales do assemblages respond (e.g., do responses lag and what factors determine lag time)? Additionally, how system dependent are the answers to these questions? For example, are the answers different in desert streams versus mountain streams or at low latitudes versus high latitudes? In chapter 5, I examined sites with generally only modest fluctuations in summer stream temperatures between years that were close together in time, albeit nonconsecutive. Longer-term datasets with temperature and assemblage data collected in consecutive years may be needed to better address questions about TBI responsiveness. Understanding the responsiveness of a TBI to changes in temperature is critical because it will determine in what way a TBI is applied for management purposes, or if it is appropriate for such purposes.

References

Christin, S., Hervet, É., Lecomte, N., 2019. Applications for deep learning in ecology. Methods in Ecology and Evolution. 10, 1632-1644.

Hawkins, C. P., Cao, Y., Roper, B., 2010. Method of predicting reference condition biota affects the performance and interpretation of ecological indices. Freshwater Biology. 55, 1066-1085.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J., 2007. Random forests for classification in ecology. Ecology. 88, 2783-2792.

APPENDICES

Appendix A. Summary of dissolved oxygen and temperature in the wet-lab troughs and rearing chambers over the duration of each laboratory experiment

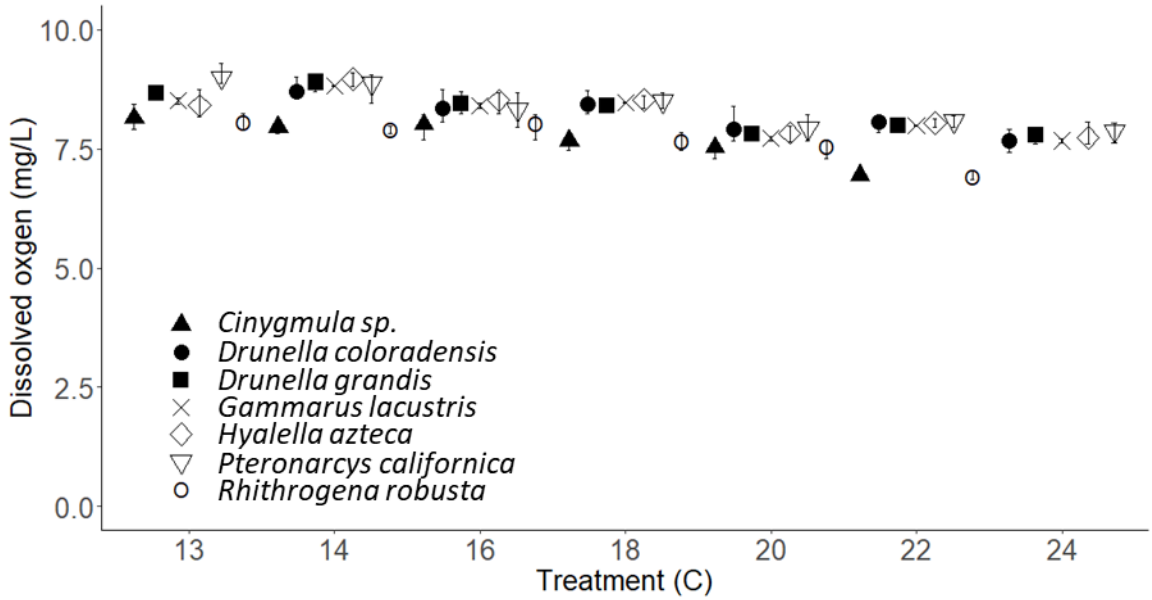


Figure A.1. Average dissolved oxygen (DO) (error bars are minimum and maximum readings) per trough calculated across all DO readings during the duration of each experiment. Two or three times per week, trough DO readings were taken at the top of each trough next to the heater. Data are jittered for discernibility among taxa at each treatment.

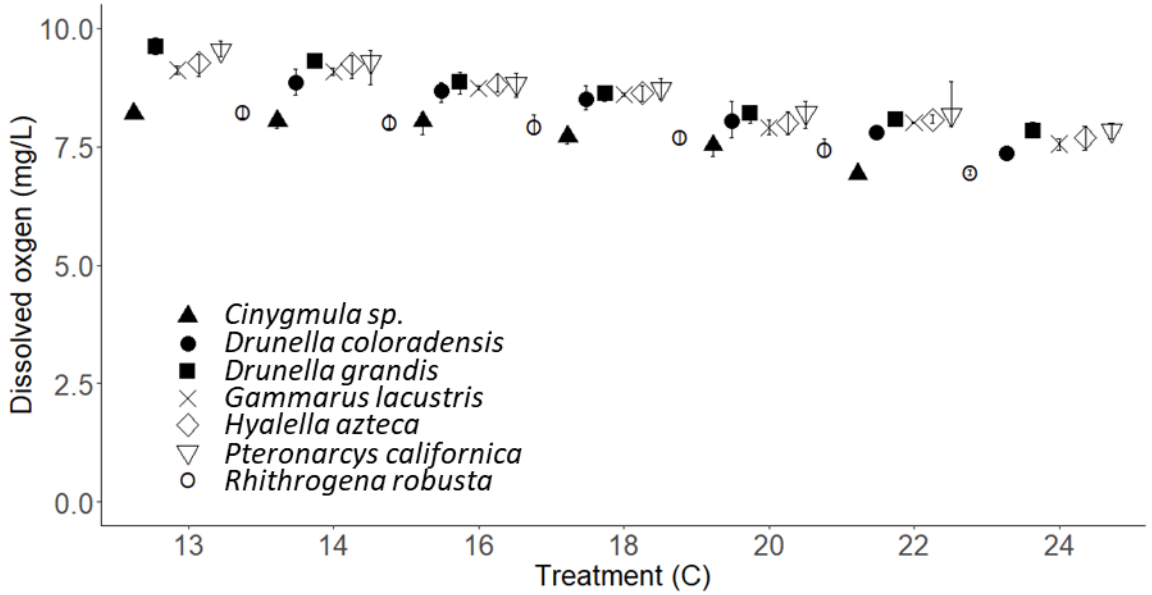


Figure A.2. Average dissolved oxygen (DO) (error bars are minimum and maximum readings) per rearing chamber calculated across all DO readings during the duration of each experiment. DO measurements were taken two or three times per week from one or two randomly selected rearing chambers per trough. Data are jittered for discernibility among taxa at each treatment.

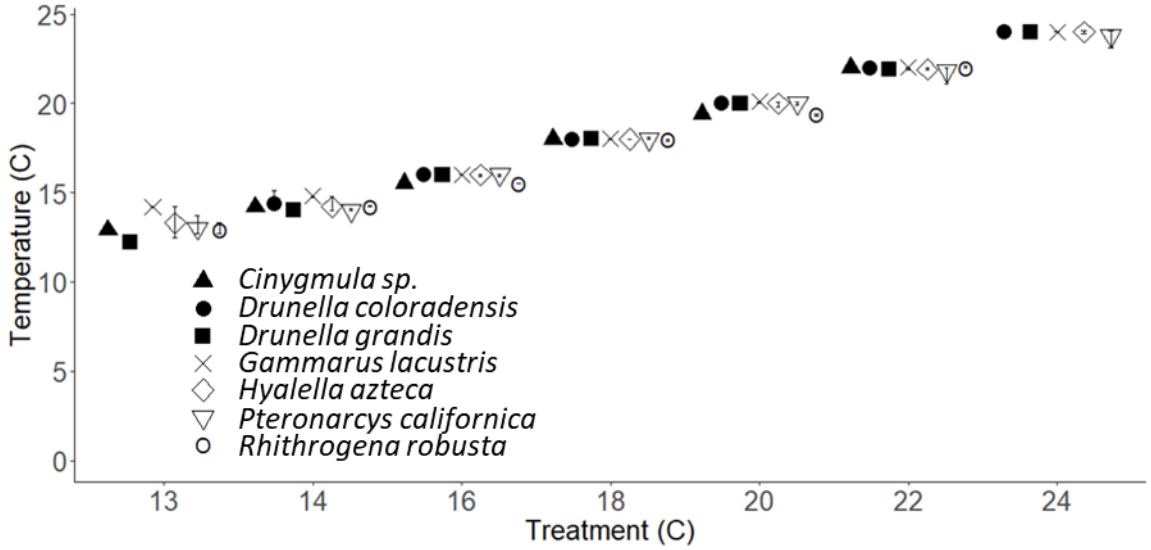


Figure A.3. Average temperature (°C) (error bars are minimum and maximum readings) per trough calculated across all temperature readings during the duration of each experiment. Two or three times per week, trough temperature readings were taken at the top of each trough next to the heater. Data are jittered for discernibility among taxa at each treatment.

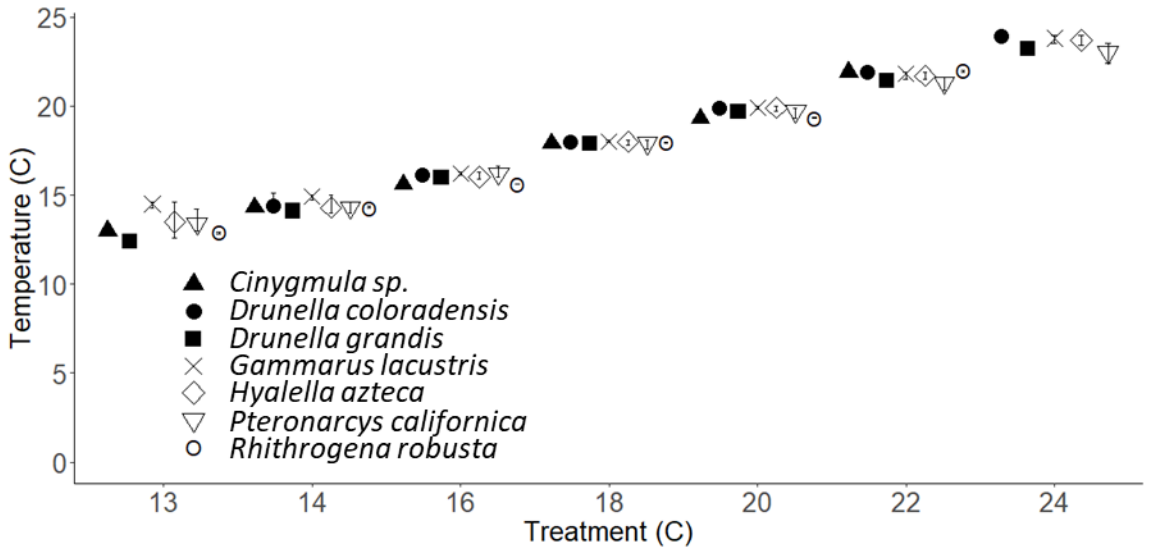


Figure A.4. Average temperature (°C) (error bars are minimum and maximum readings) per rearing chamber calculated across all temperature readings during the duration of each experiment. Temperature measurements were taken two or three times per week from one or two randomly selected rearing chambers per trough. Data are jittered for discernibility among taxa at each treatment.

Appendix B. Optimized hyperparameter values for each imbalance-correction methods by machine-learning algorithm model (species distribution model)

The values tried in each hyperparameter grid search are presented as (minimum value, maximum value, increment value). For example, (2, 10, 1) indicates the search began at 2 and incremented up by 1 until 10 was reached. The grid search for randomForest was mtry = (1, 11, 1) and ntree = (50, 800, 50). The grid search for hyperparameters in the nnet package were size = (1, 100, 1) and maxit = (10, 1000, 10). The grid search for hyperparameters in gbm were interaction.depth = (1, 10, 1), n.trees = (50, 800, 50), and shrinkage = 0.001 and (0.01, 1, 0.01). The grid search for svm was gamma = 0.001 and (0.03, 3, 0.03) and cost = (0.1, 20, 0.1).

Table B.1. Hyperparameters and optimized values for base models. The description of each hyperparameter can be found in the respective machine-learning algorithm R package.

	Base models								
	Random forest		ANN		Gradient boosting			SVM	
Taxa	mtry	ntree	size	maxit	interaction.depth	n.trees	shrinkage	gamma	cost
Malenka	11	50	24	360	3	250	0.85	1.33	16.7
Pteronarcys	8	50	49	120	2	450	1	2.15	19.9
Zapada	9	450	4	140	1	350	0.84	0.58	8.9
Drunella	11	50	22	280	4	100	0.21	0.15	12.7
Callibaetis	10	50	53	120	9	300	0.82	0.46	19.6
Rhyacophila	7	50	3	90	7	50	0.15	0.7	10.9
Stenacron	8	50	68	100	7	400	0.63	0.73	14.3
Sialis	9	250	93	90	5	150	0.99	1.55	19.5
Gammarus	9	200	60	50	10	300	0.35	0.61	18.7
Argia	10	50	60	50	5	250	0.6	0.21	14.5
Hemerodromia	10	150	44	50	7	100	0.52	0.94	17.5
Optioservus	3	250	94	10	10	200	0.03	0.03	10.4
Paratanytarsus	10	50	91	70	3	750	0.72	0.46	19.4
Hydroptila	10	200	57	30	2	250	0.33	0.52	12.7
Centroptilum/ Procloeon	7	50	79	60	8	650	0.15	0.73	6.2

Table B.2. Hyperparameters and optimized values for up-sampled models. The description of each hyperparameter can be found in the respective machine-learning algorithm R package.

	Up-sampled models								
	Random forest		ANN		Gradient boosting			SVM	
Taxa	mtry	ntree	size	maxit	interaction.depth	n.trees	shrinkage	gamma	cost
Malenka	11	500	87	10	1	450	0.01	0.001	14
Pteronarcys	6	150	67	10	4	800	0.001	0.03	0.7
Zapada	8	750	49	10	1	450	0.001	0.03	1.9
Drunella	9	650	94	10	4	700	0.001	0.03	0.1
Callibaetis	11	800	11	10	1	700	0.01	0.79	0.1
Rhyacophila	8	300	71	10	2	50	0.09	0.03	0.1
Stenacron	10	550	68	10	3	200	0.03	0.12	0.4
Sialis	10	650	71	30	5	100	0.01	0.06	13.8
Gammarus	10	600	10	30	10	500	0.001	0.09	12.7
Argia	7	750	29	10	2	400	0.01	0.03	0.5
Hemerodromia	10	50	74	20	1	150	0.17	0.03	1.7
Optioservus	3	350	66	10	1	700	0.05	0.03	5.6
Paratanytarsus	9	50	99	20	6	100	0.05	0.06	4.3
Hydroptila	9	450	80	20	7	50	0.05	0.15	1.5
Centroptilum/ Procloeon	10	750	10	50	10	50	0.06	0.46	0.3

Table B.3. Hyperparameters and optimized values for down-sampled models. The description of each hyperparameter can be found in the respective machine-learning algorithm R package.

	Down-sampled models								
	Random forest		ANN		Gradient boosting			SVM	
Taxa	mtry	ntree	size	maxit	interaction.depth	n.trees	shrinkage	gamma	cost
Malenka	7	250	33	10	2	250	0.06	0.03	2.5
Pteronarcys	5	500	1	40	2	150	0.01	1.76	0.9
Zapada	3	200	57	10	5	450	0.11	0.12	4.9
Drunella	7	400	20	10	3	150	0.04	0.06	1
Callibaetis	5	50	3	30	3	300	0.001	0.21	2.7
Rhyacophila	1	800	100	10	2	150	0.02	0.09	0.2
Stenacron	10	150	52	10	3	300	0.04	0.24	0.5
Sialis	6	550	24	20	2	50	0.25	1.46	0.6
Gammarus	7	200	42	20	8	150	0.02	0.24	20
Argia	8	150	42	10	2	100	0.02	0.03	0.7
Hemerodromia	8	350	75	20	9	50	0.02	0.03	3.3
Optioservus	2	100	9	10	3	100	0.6	0.03	10.4
Paratanytarsus	10	200	40	20	9	100	0.04	0.06	10.1
Hydroptila	7	150	99	10	3	50	0.1	0.12	0.8
Centroptilum/ Procloeon	6	200	64	20	10	200	0.02	0.43	0.2

Table B.4. Hyperparameters and optimized values for cutoff implemented models. The description of each hyperparameter can be found in the respective machine-learning algorithm R package.

	Cutoff models								
	Random forest		ANN		Gradient boosting			SVM	
Taxa	mtry	ntree	size	maxit	interaction.depth	n.trees	Shrinkage	gamma	cost
Malenka	3	150	9	10	1	350	0.19	0.001	17.6
Pteronarcys	3	250	22	10	5	50	0.04	0.001	19.6
Zapada	11	650	55	10	5	50	0.08	0.001	0.3
Drunella	4	50	12	20	9	100	0.04	0.001	14.7
Callibaetis	1	250	57	10	3	100	0.04	0.64	0.2
Rhyacophila	4	500	89	10	4	350	0.02	0.001	8.8
Stenacron	1	600	16	10	1	250	0.12	2.7	1.8
Sialis	3	350	92	20	4	300	0.02	2.82	3.2
Gammarus	3	200	73	30	10	700	0.001	0.12	6.2
Argia	2	700	53	10	1	200	0.05	1.46	0.2
Hemerodromia	4	150	92	20	1	50	0.08	0.001	17.8
Optioservus	9	550	93	20	9	50	0.09	0.06	0.8
Paratanytarsus	3	150	92	20	8	400	0.01	0.15	16.5
Hydroptila	1	150	11	10	7	150	0.001	0.55	2.2
Centroptilum/ Procloeon	6	100	11	40	7	800	0.01	0.49	0.1

Table B.5. Hyperparameters and optimized values for weighted models. The description of each hyperparameter can be found in the respective machine-learning algorithm R package. We did not apply weighting to random forest because no reliable implementations were available for our selected package, or for any package in R that we were aware of.

	Weighted models								
	Random forest		ANN		Gradient boosting			SVM	
Taxa	mtry	ntree	size	maxit	interaction.depth	n.trees	Shrinkage	gamma	cost
Malenka			35	10	1	300	0.001	0.03	2.8
Pteronarcys			89	10	7	150	0.001	0.12	0.3
Zapada			48	10	1	50	0.06	0.15	0.4
Drunella			42	10	2	450	0.01	0.3	2
Callibaetis			3	10	1	100	0.25	0.64	0.9
Rhyacophila			4	10	2	50	0.1	0.18	0.5
Stenacron			32	10	1	400	0.06	0.09	14.2
Sialis			39	20	4	100	0.03	0.15	9.4
Gammarus			13	40	9	50	0.02	0.43	20
Argia			2	20	2	150	0.02	0.03	0.7
Hemerodromia			74	20	1	300	0.03	0.06	2.6
Optioservus			29	10	3	50	0.19	0.21	0.1
Paratanytarsus			72	20	9	100	0.03	0.06	10.3

Hydroptila			45	10	10	50	0.001	0.15	0.1
Centroptilum/ Procloeon			40	10	7	200	0.03	0.97	0.9

Appendix C. Performance metrics for each imbalance-correction methods by machine-learning algorithm model (species distribution model)

Table C.1. Performance metrics for base random forest and ANN models for each species. Prev. = species prevalence.

		Base random forest				Base ANN			
Taxa	Prev.	TSS	k	AUROC	PCC	TSS	K	AUROC	PCC
Malenka	2.5	0.05	0.09	0.86	97	0.27	0.26	0.65	96
Pteronarcys	3.8	0.04	0.07	0.80	96	0.23	0.20	0.65	93
Zapada	4.7	0.50	0.57	0.97	97	0.69	0.59	0.94	96
Drunella	8.1	0.62	0.65	0.95	95	0.72	0.58	0.92	92
Callibaetis	9.2	0.01	0.02	0.61	90	0.13	0.12	0.56	84
Rhyacophila	11.4	0.54	0.60	0.93	93	0.60	0.61	0.93	92
Stenacron	11.9	0.15	0.22	0.73	89	0.24	0.23	0.64	83
Sialis	15.4	0.04	0.06	0.65	84	0.14	0.13	0.59	76
Gammarus	17.1	0.27	0.34	0.77	85	0.32	0.33	0.74	82
Argia	19.3	0.18	0.23	0.72	81	0.26	0.27	0.69	78
Hemerodromia	21.2	0.11	0.15	0.70	79	0.19	0.20	0.66	75
Optioservus	24.2	0.45	0.47	0.85	82	0.52	0.49	0.85	80
Paratanytarsus	26.9	0.14	0.16	0.68	73	0.17	0.17	0.63	68
Hydroptila	27.7	0.11	0.14	0.64	72	0.16	0.19	0.63	71
Centroptilum/ Procloeon	29	0.22	0.25	0.70	73	0.22	0.22	0.67	68

Table C.2. Performance metrics for base gradient boosting and SVM models for each species. Prev. = species prevalence.

		Base gradient boosting				Base SVM			
Taxa	Prev.	TSS	k	AUROC	PCC	TSS	K	AUROC	PCC
Malenka	2.5	0.54	0.19	0.72	88	0.09	0.08		96
Pteronarcys	3.8	0.31	0.14	0.66	85	0.15	0.18		95
Zapada	4.7	0.72	0.60	0.84	96	0.52	0.57		96
Drunella	8.1	0.65	0.67	0.96	95	0.64	0.68		95
Callibaetis	9.2	0.14	0.08	0.56	69	0.04	0.06		89
Rhyacophila	11.4	0.56	0.62	0.93	93	0.54	0.59		92
Stenacron	11.9	0.23	0.24	0.71	84	0.21	0.25		86
Sialis	15.4	0.14	0.12	0.57	72	0.13	0.15		80
Gammarus	17.1	0.33	0.37	0.74	84	0.28	0.31		82
Argia	19.3	0.23	0.25	0.68	78	0.17	0.22		81
Hemerodromia	21.2	0.20	0.22	0.65	76	0.17	0.18		75
Optioservus	24.2	0.47	0.48	0.86	82	0.49	0.48		81
Paratanytarsus	26.9	0.20	0.21	0.64	70	0.10	0.12		69
Hydroptila	27.7	0.16	0.18	0.63	70	0.13	0.15		71
Centroptilum/ Procloeon	29	0.25	0.27	0.69	72	0.20	0.21		69

Table C.3. Performance metrics for up-sampled random forest and ANN models for each species. Prev. = species prevalence.

		Up-sample random forest				Up-sample ANN			
Taxa	Prev.	TSS	k	AUROC	PCC	TSS	k	AUROC	PCC
Malenka	2.5	0.17	0.19	0.85	96	0.76	0.19	0.93	84
Pteronarcys	3.8	0.14	0.18	0.77	95	0.61	0.17	0.85	78
Zapada	4.7	0.57	0.57	0.93	96	0.87	0.47	0.98	91
Drunella	8.1	0.65	0.62	0.96	94	0.86	0.55	0.97	90
Callibaetis	9.2	0.05	0.07	0.62	90	0.22	0.08	0.66	59
Rhyacophila	11.4	0.62	0.65	0.94	93	0.76	0.57	0.95	88
Stenacron	11.9	0.26	0.31	0.76	88	0.43	0.22	0.76	68
Sialis	15.4	0.13	0.17	0.67	83	0.26	0.17	0.65	67
Gammarus	17.1	0.36	0.39	0.77	84	0.41	0.30	0.77	73
Argia	19.3	0.24	0.27	0.73	79	0.38	0.28	0.75	69
Hemerodromia	21.2	0.17	0.19	0.70	75	0.31	0.22	0.69	63
Optioservus	24.2	0.50	0.48	0.85	80	0.57	0.45	0.84	75
Paratanytarsus	26.9	0.19	0.20	0.67	71	0.26	0.22	0.67	63
Hydroptila	27.7	0.16	0.18	0.66	69	0.23	0.20	0.65	62
Centroptilum/ Procloeon	29	0.29	0.30	0.72	72	0.30	0.25	0.69	64

Table C.4. Performance metrics for up-sampled gradient boosting and SVM models for each species. Prev. = species prevalence. For SVMs, we do not report the AUROC because preliminary analyses showed that classifications based on estimated probabilities (with decision threshold of 0.5) did not always match classifications based on decision values. See methods for a more complete explanation.

		Up-sampled gradient boosting				Up-sampled SVM			
Taxa	Prev.	TSS	k	AUROC	PCC	TSS	k	AUROC	PCC
Malenka	2.5	0.77	0.19	0.93	84	0.75	0.17		82
Pteronarcys	3.8	0.58	0.20	0.84	83	0.61	0.15		74
Zapada	4.7	0.85	0.49	0.94	92	0.85	0.45		90
Drunella	8.1	0.83	0.54	0.96	90	0.86	0.50		87
Callibaetis	9.2	0.24	0.11	0.66	66	0.20	0.08		63
Rhyacophila	11.4	0.77	0.58	0.95	89	0.77	0.54		87
Stenacron	11.9	0.45	0.29	0.79	77	0.43	0.24		70
Sialis	15.4	0.27	0.16	0.68	64	0.27	0.15		61
Gammarus	17.1	0.45	0.34	0.76	75	0.43	0.31		73
Argia	19.3	0.40	0.29	0.75	69	0.37	0.27		68
Hemerodromia	21.2	0.34	0.25	0.72	66	0.35	0.22		59
Optioservus	24.2	0.58	0.49	0.85	78	0.56	0.45		75
Paratanytarsus	26.9	0.27	0.24	0.68	66	0.23	0.19		60
Hydroptila	27.7	0.26	0.22	0.67	63	0.23	0.19		60
Centroptilum/ Procloeon	29	0.33	0.29	0.72	68	0.31	0.26		64

Table C.5. Performance metrics for down-sampled random forest and ANN models for each species. Prev. = species prevalence.

		Down-sample random forest				Down-sample ANN			
Taxa	Prev.	TSS	k	AUROC	PCC	TSS	k	AUROC	PCC
Malenka	2.5	0.76	0.19	0.93	84	0.73	0.18	0.92	83
Pteronarcys	3.8	0.59	0.22	0.84	84	0.59	0.12	0.82	66
Zapada	4.7	0.86	0.53	0.97	93	0.86	0.40	0.97	88
Drunella	8.1	0.82	0.56	0.97	90	0.85	0.56	0.97	90
Callibaetis	9.2	0.24	0.14	0.65	75	0.23	0.09	0.65	61
Rhyacophila	11.4	0.78	0.63	0.94	91	0.76	0.53	0.95	86
Stenacron	11.9	0.41	0.30	0.79	80	0.40	0.21	0.76	68
Sialis	15.4	0.26	0.21	0.69	74	0.24	0.13	0.64	58
Gammarus	17.1	0.45	0.38	0.78	79	0.38	0.25	0.73	67
Argia	19.3	0.35	0.29	0.75	73	0.38	0.27	0.75	67
Hemerodromia	21.2	0.31	0.25	0.72	69	0.32	0.22	0.69	63
Optioservus	24.2	0.56	0.49	0.86	79	0.57	0.45	0.84	74
Paratanytarsus	26.9	0.24	0.22	0.68	67	0.24	0.20	0.66	61
Hydroptila	27.7	0.23	0.21	0.66	66	0.24	0.19	0.66	60
Centroptilum/ Procloeon	29	0.34	0.32	0.72	70	0.29	0.24	0.69	63

Table C.6. Performance metrics for down-sampled gradient boosting and SVM models for each species. Prev. = species prevalence. For SVMs, we do not report the AUROC because preliminary analyses showed that classifications based on estimated probabilities (with decision threshold of 0.5) did not always match classifications based on decision values. See methods for a more complete explanation.

		Down-sample gradient boosting				Down-sample SVM			
Taxa	Prev.	TSS	k	AUROC	PCC	TSS	k	AUROC	PCC
Malenka	2.5	0.78	0.20	0.92	84	0.73	0.15		79
Pteronarcys	3.8	0.57	0.15	0.83	75	0.59	0.20		83
Zapada	4.7	0.86	0.48	0.96	91	0.85	0.38		87
Drunella	8.1	0.84	0.55	0.97	90	0.85	0.50		87
Callibaetis	9.2	0.24	0.09	0.65	59	0.22	0.09		60
Rhyacophila	11.4	0.77	0.53	0.95	86	0.76	0.53		86
Stenacron	11.9	0.45	0.24	0.77	70	0.42	0.24		72
Sialis	15.4	0.27	0.15	0.66	62	0.24	0.17		70
Gammarus	17.1	0.42	0.30	0.77	71	0.38	0.26		69
Argia	19.3	0.39	0.28	0.75	68	0.37	0.25		64
Hemerodromia	21.2	0.34	0.24	0.71	63	0.34	0.22		59
Optioservus	24.2	0.58	0.48	0.85	77	0.58	0.45		74
Paratanytarsus	26.9	0.26	0.21	0.67	61	0.26	0.21		62
Hydroptila	27.7	0.28	0.23	0.67	61	0.24	0.18		57
Centroptilum/ Procloeon	29	0.33	0.28	0.71	65	0.32	0.28		66

Table C.7. Performance metrics for cutoff random forest and ANN models for each species. Prev. = species prevalence.

		Cutoff random forest				Cutoff ANN			
Taxa	Prev.	TSS	k	AUROC	PCC	TSS	k	AUROC	PCC
Malenka	2.5	0.72	0.15	0.91	79	0.75	0.19	0.91	84
Pteronarcys	3.8	0.53	0.12	0.83	70	0.61	0.15	0.86	74
Zapada	4.7	0.84	0.49	0.96	92	0.86	0.45	0.97	90
Drunella	8.1	0.81	0.49	0.96	87	0.86	0.56	0.97	90
Callibaetis	9.2	0.27	0.11	0.64	61	0.25	0.11	0.66	67
Rhyacophila	11.4	0.75	0.52	0.94	86	0.78	0.54	0.95	86
Stenacron	11.9	0.39	0.20	0.77	66	0.43	0.23	0.77	69
Sialis	15.4	0.23	0.12	0.66	58	0.27	0.16	0.67	63
Gammarus	17.1	0.42	0.28	0.78	69	0.41	0.29	0.75	72
Argia	19.3	0.36	0.24	0.74	64	0.38	0.27	0.75	67
Hemerodromia	21.2	0.32	0.22	0.71	62	0.31	0.22	0.70	63
Optioservus	24.2	0.56	0.45	0.84	75	0.57	0.47	0.85	76
Paratanytarsus	26.9	0.26	0.21	0.67	61	0.25	0.21	0.67	61
Hydroptila	27.7	0.21	0.17	0.64	59	0.24	0.19	0.65	59
Centroptilum/ Procloeon	29	0.33	0.28	0.71	64	0.30	0.25	0.68	64

Table C.8. Performance metrics for cutoff gradient boosting and SVM models for each species. Prev. = species prevalence. For SVMs, we do not report the AUROC because preliminary analyses showed that classifications based on estimated probabilities (with decision threshold of 0.5) did not always match classifications based on decision values. See methods for a more complete explanation.

		Cutoff gradient boosting				Cutoff SVM			
Taxa	Prev.	TSS	k	AUROC	PCC	TSS	k	AUROC	PCC
Malenka	2.5	0.72	0.24	0.91	88	0.72	0.13		75
Pteronarcys	3.8	0.58	0.21	0.83	84	0.51	0.09		61
Zapada	4.7	0.82	0.55	0.96	94	0.85	0.46		91
Drunella	8.1	0.82	0.62	0.97	92	0.85	0.53		89
Callibaetis	9.2	0.23	0.11	0.65	69	0.20	0.10		68
Rhyacophila	11.4	0.76	0.61	0.94	90	0.77	0.55		87
Stenacron	11.9	0.43	0.25	0.77	72	0.32	0.24		79
Sialis	15.4	0.26	0.16	0.67	66	0.20	0.14		69
Gammarus	17.1	0.43	0.32	0.76	75	0.40	0.30		73
Argia	19.3	0.38	0.27	0.74	68	0.33	0.26		71
Hemerodromia	21.2	0.33	0.23	0.71	63	0.24	0.14		49
Optioservus	24.2	0.58	0.50	0.86	78	0.57	0.49		78
Paratanytarsus	26.9	0.30	0.25	0.69	64	0.21	0.21		68
Hydroptila	27.7	0.26	0.20	0.67	58	0.18	0.17		64
Centroptilum/ Procloeon	29	0.33	0.29	0.71	67	0.29	0.29		71

Table C.9. Performance metrics for weighted random forest and ANN models for each species. Prev. = species prevalence. We did not apply weighting to random forest because no reliable implementations were available for our selected package, or for any package in R that we were aware of.

		Weighted random forest				Weighted ANN			
Taxa	Prev.	TSS	k	AUROC	PCC	TSS	k	AUROC	PCC
Malenka	2.5					0.77	0.21	0.93	85
Pteronarcys	3.8					0.61	0.17	0.86	78
Zapada	4.7					0.86	0.41	0.97	89
Drunella	8.1					0.86	0.55	0.97	90
Callibaetis	9.2					0.23	0.09	0.65	59
Rhyacophila	11.4					0.77	0.56	0.95	88
Stenacron	11.9					0.43	0.23	0.76	70
Sialis	15.4					0.28	0.17	0.67	65
Gammarus	17.1					0.41	0.31	0.75	74
Argia	19.3					0.39	0.28	0.75	69
Hemerodromia	21.2					0.32	0.22	0.70	63
Optioservus	24.2					0.56	0.43	0.85	73
Paratanytarsus	26.9					0.25	0.21	0.66	62
Hydroptila	27.7					0.25	0.22	0.65	63
Centroptilum/ Procloeon	29					0.29	0.25	0.69	64

Table C.10. Performance metrics for weighted gradient boosting and SVM models for each species. Prev. = species prevalence. For SVMs, we do not report the AUROC because preliminary analyses showed that classifications based on estimated probabilities (with decision threshold of 0.5) did not always match classifications based on decision values. See methods for a more complete explanation.

		Weighted gradient boosting				Weighted SVM			
Taxa	Prev.	TSS	k	AUROC	PCC	TSS	k	AUROC	PCC
Malenka	2.5	0.76	0.18	0.89	83	0.74	0.16		81
Pteronarcys	3.8	0.58	0.24	0.84	86	0.62	0.18		79
Zapada	4.7	0.86	0.51	0.97	92	0.86	0.42		89
Drunella	8.1	0.82	0.57	0.97	91	0.86	0.55		89
Callibaetis	9.2	0.26	0.13	0.66	71	0.21	0.07		55
Rhyacophila	11.4	0.77	0.59	0.95	89	0.77	0.57		88
Stenacron	11.9	0.45	0.28	0.79	75	0.44	0.24		70
Sialis	15.4	0.26	0.17	0.68	67	0.29	0.17		62
Gammarus	17.1	0.44	0.34	0.77	76	0.42	0.33		76
Argia	19.3	0.39	0.28	0.75	69	0.37	0.25		66
Hemerodromia	21.2	0.34	0.24	0.71	63	0.34	0.22		59
Optioservus	24.2	0.59	0.50	0.86	78	0.57	0.48		77
Paratanytarsus	26.9	0.27	0.24	0.69	66	0.24	0.19		60
Hydroptila	27.7	0.27	0.22	0.67	62	0.24	0.21		63
Centroptilum/ Procloeon	29	0.32	0.29	0.71	68	0.32	0.28		66

Appendix D. Artificial neural network-based species distribution model performance presented as two alternative performance metrics (area under the receiver operating characteristic curve and percent classified correctly)

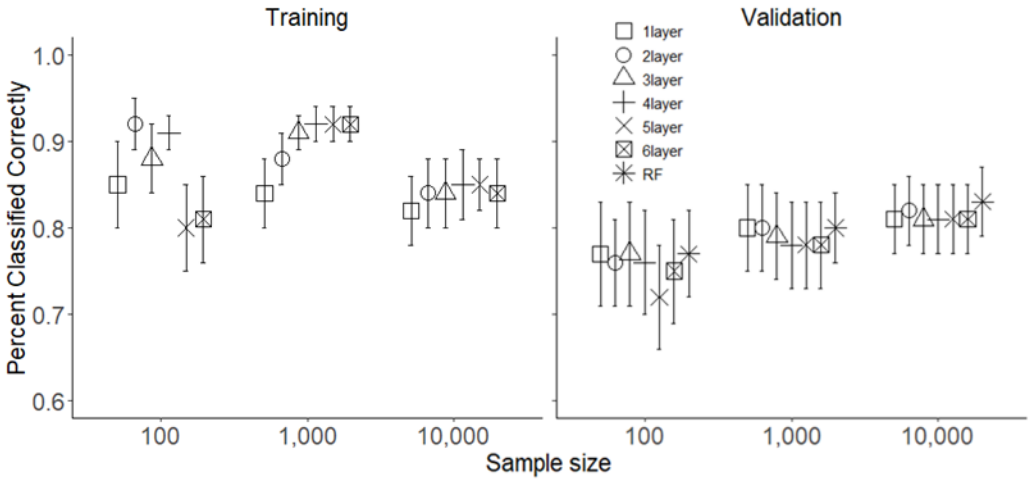


Fig. D.1. Effects of dataset size and neural network depth on mean ± SE model performance (Percent classified correctly) for the training dataset (left) and for the validation dataset (right) across the 5 macroinvertebrate genera modeled in this study. RF (random forest) was included for the validation dataset for comparison with a different classifier commonly used in species distribution modeling.

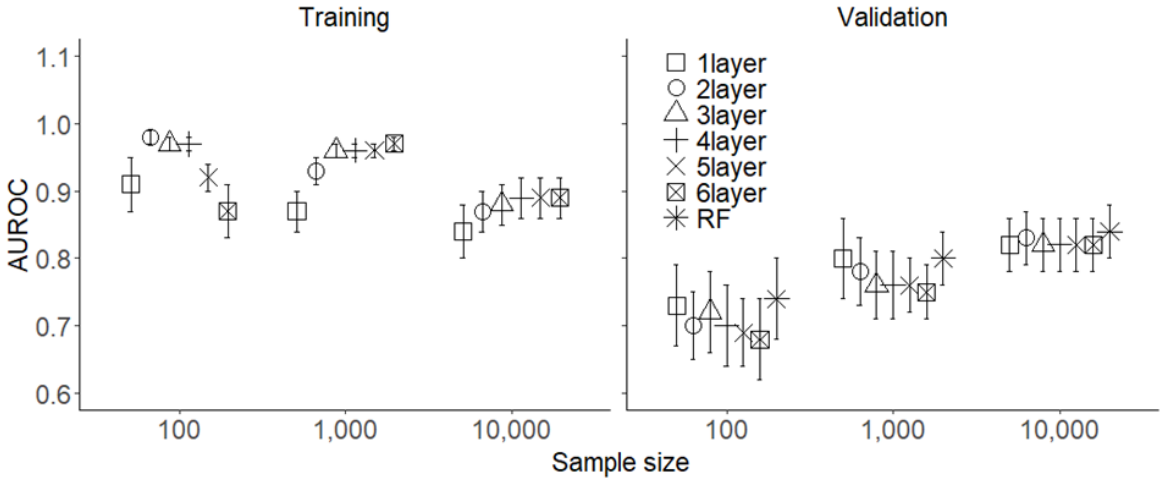


Fig. D.2. Effects of dataset size and neural network depth on mean ± SE model performance (area under the receiver operating characteristic curve) for the training dataset (left) and for the validation dataset (right) across the 5 macroinvertebrate genera modeled in this study. RF (random forest) was included for the validation dataset for comparison with a different classifier commonly used in species distribution modeling.

Appendix E. Optimized nodes/layer and number of epochs for each artificial neural network-based species distribution models

Table E.1. Optimized nodes/layer and number of epochs of all 18 models built for *Caenis*.

Caenis

1 hidden layer network

sample size	100	1000	10000
layer	1	1	1
nodes	302	283	88
epochs	131	47	47

2 hidden layer network

sample size	100		1000		10000	
layer	1	2	1	2	1	2
nodes	380	88	263	224	302	263
epochs	108		31		41	

3 hidden layer network

sample size	100			1000			10000		
layer	1	2	3	1	2	3	1	2	3
nodes	341	166	263	224	88	29	166	244	185
epochs	81			42			38		

4 hidden layer network

sample size	100				1000				10000			
layer	1	2	3	4	1	2	3	4	1	2	3	4
nodes	380	127	166	107	283	127	166	244	68	283	302	29
epochs	74				31				35			

5 hidden layer network

sample size	100					1000					10000				
layer	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
nodes	283	185	146	283	10	68	283	283	361	146	127	205	68	244	107
epochs	13					31					26				

6 hidden layer network

sample size	100						1000						10000					
layer	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
nodes	361	224	283	322	361	380	68	146	361	166	29	88	166	302	302	49	146	361
epochs	89						30						38					

Table E.2. Optimized nodes/layer and number of epochs of all 18 models built for *Tricorythodes*.

Tricorythodes

1 hidden layer network

sample size	100	1000	10000
layer	1	1	1
nodes	146	224	341
epochs	106	46	37

2 hidden layer network

sample size	100		1000		10000	
layer	1	2	1	2	1	2
nodes	88	322	127	380	244	224
epochs	186		33		63	

3 hidden layer network

sample size	100			1000			10000		
layer	1	2	3	1	2	3	1	2	3
nodes	322	127	146	185	322	88	185	244	302
epochs	54			29			34		

4 hidden layer network

sample size	100				1000				10000			
layer	1	2	3	4	1	2	3	4	1	2	3	4
nodes	88	341	185	68	68	361	29	49	166	224	146	88
epochs	77				32				38			

5 hidden layer network

sample size	100					1000					10000				
layer	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
nodes	10	107	322	146	107	224	49	244	224	224	380	68	380	283	68
epochs	69					29					39				

6 hidden layer network

sample size	100						1000						10000					
layer	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
nodes	166	68	244	185	49	29	302	263	361	302	166	29	49	185	107	107	185	88
epochs	61						29						31					

Table E.3. Optimized nodes/layer and number of epochs of all 18 models built for *Micrasema*.

Micrasema

1 hidden layer network

sample size	100	1000	10000
layer	1	1	1
nodes	283	185	380
epochs	18	36	65

2 hidden layer network

sample size	100		1000		10000	
layer	1	2	1	2	1	2
nodes	361	127	341	302	224	244
epochs	36		31		39	

3 hidden layer network

sample size	100			1000			10000		
layer	1	2	3	1	2	3	1	2	3
nodes	49	224	224	88	361	146	127	205	302
epochs	69			30			32		

4 hidden layer network

sample size	100				1000				10000			
layer	1	2	3	4	1	2	3	4	1	2	3	4
nodes	166	224	166	107	49	263	263	146	49	205	322	244
epochs	36				23				25			

5 hidden layer network

sample size	100					1000					10000				
layer	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
nodes	361	88	302	68	127	205	185	322	380	10	185	205	107	127	283
epochs	53					25					35				

6 hidden layer network

sample size	100						1000						10000					
layer	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
nodes	10	107	244	302	341	68	146	107	283	341	49	166	146	380	29	166	322	29
epochs	69						27						31					

Table E.4. Optimized nodes/layer and number of epochs of all 18 models built for *Baetis*.

Baetis

1 hidden layer network

sample size	100	1000	10000
layer	1	1	1
nodes	341	322	146
epochs	20	33	30

2 hidden layer network

sample size	100		1000		10000	
layer	1	2	1	2	1	2
nodes	49	185	88	361	107	380
epochs	63		22		42	

3 hidden layer network

sample size	100			1000			10000		
layer	1	2	3	1	2	3	1	2	3
nodes	49	244	263	107	244	302	127	283	380
epochs	50			21			39		

4 hidden layer network

sample size	100				1000				10000			
layer	1	2	3	4	1	2	3	4	1	2	3	4
nodes	224	127	10	107	88	341	283	244	244	341	302	205
epochs	72				19				41			

5 hidden layer network

sample size	100					1000					10000				
layer	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
nodes	29	244	68	127	49	107	224	322	10	29	88	205	263	263	68
epochs	60					29					37				

6 hidden layer network

sample size	100						1000						10000					
layer	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
nodes	380	361	283	322	88	10	49	146	341	185	205	380	49	263	185	244	107	341
epochs	67						23						42					

Table E.5. Optimized nodes/layer and number of epochs of all 18 models built for *Rhyacophila*.

Rhyacophila

1 hidden layer network

sample size	100	1000	10000
layer	1	1	1
nodes	244	361	380
epochs	20	41	33

2 hidden layer network

sample size	100		1000		10000	
layer	1	2	1	2	1	2
nodes	185	88	302	380	322	127
epochs	36		34		34	

3 hidden layer network

sample size	100			1000			10000		
layer	1	2	3	1	2	3	1	2	3
nodes	361	68	68	224	107	361	49	361	302
epochs	40			29			32		

4 hidden layer network

sample size	100				1000				10000			
layer	1	2	3	4	1	2	3	4	1	2	3	4
nodes	146	146	127	166	49	107	302	263	185	185	302	322
epochs	32				26				33			

5 hidden layer network

sample size	100					1000					10000				
layer	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
nodes	283	49	29	283	224	166	146	224	10	244	283	88	224	107	322
epochs	42					32					34				

6 hidden layer network

sample size	100						1000						10000					
layer	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
nodes	341	49	68	244	283	107	185	88	88	224	244	185	166	244	146	341	49	361
epochs	49						22						31					

Appendix F. Supplemental analyses regarding development and evaluation of a temperature biotic index

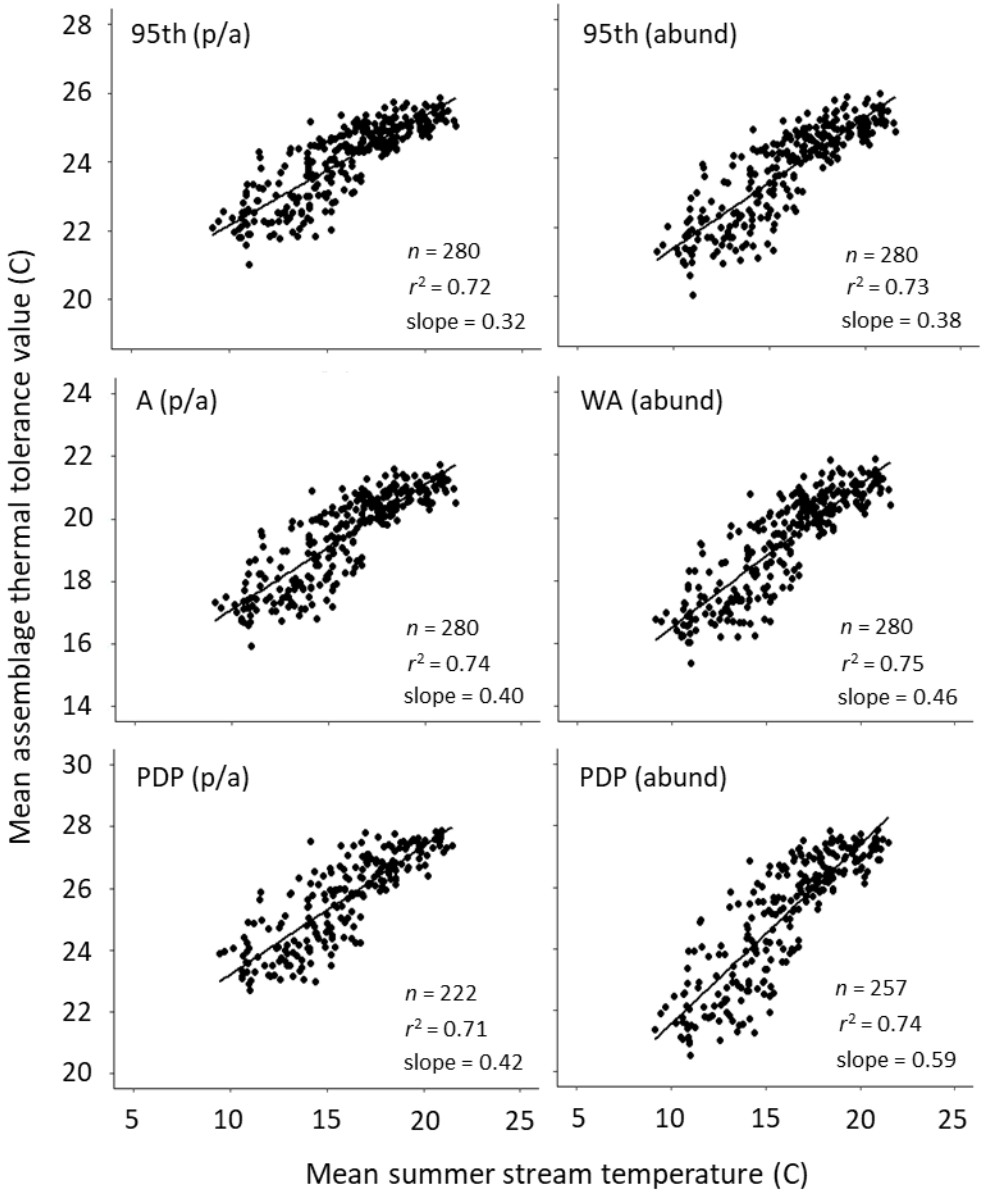


Figure F.1. Mean assemblage thermal tolerance values derived with the 6 different methods plotted against predicted mean summer stream temperature. The sites (*n*) were reference condition sites from the NRSA 2013-2014 dataset and all sites had assemblages with at least 20 taxa with associated tolerance values. The specific tolerance values are described as 95^{th} = 95^{th} percentile, A = average, WA = weighted average, PDP = partial dependence plot. The type of data used is specified in parentheses and is either p/a = presence/absence or abund = abundance.

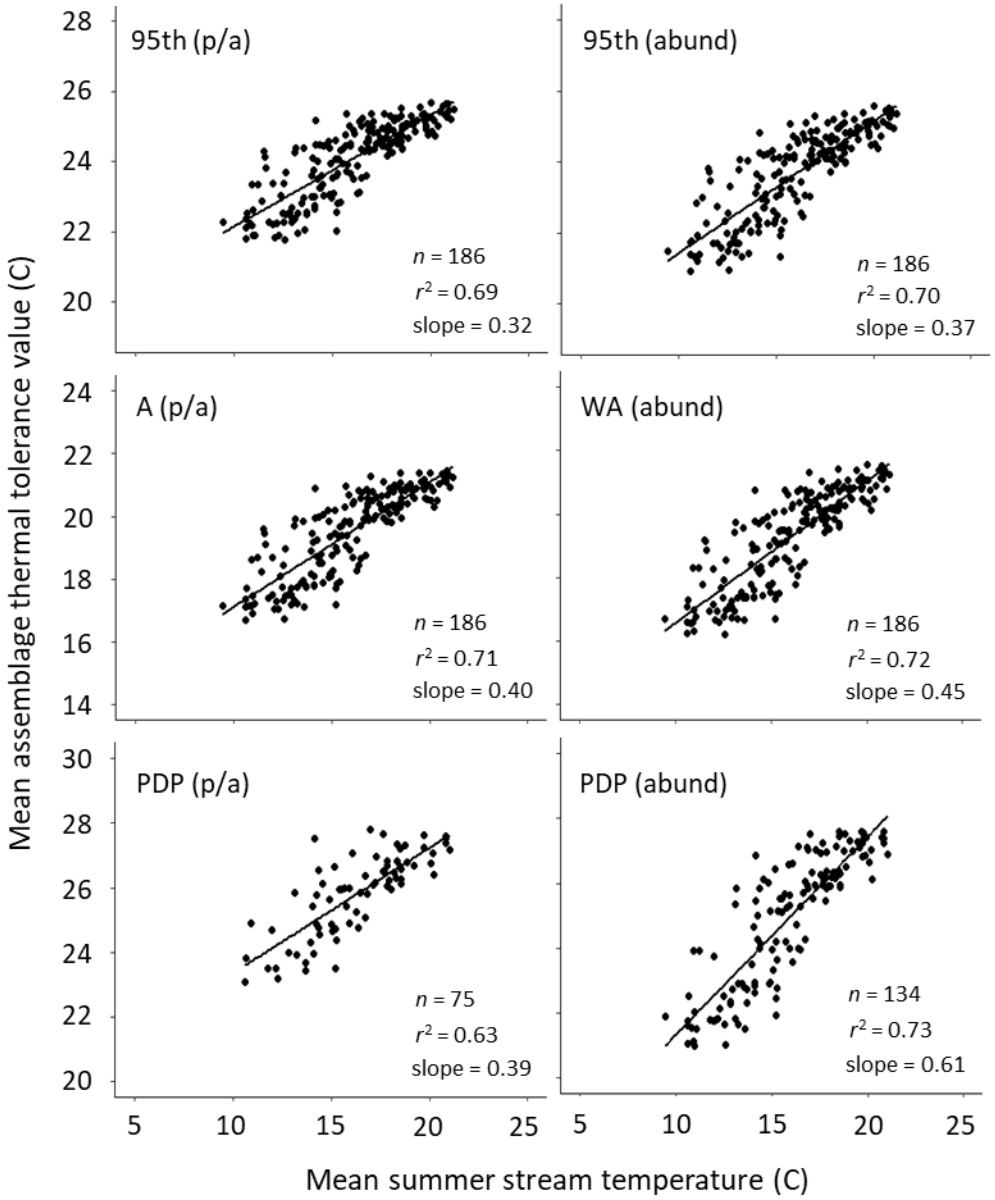


Figure F.2. Mean assemblage thermal tolerance values derived with the 6 different methods plotted against predicted mean summer stream temperature. The sites (*n*) were reference condition sites from the NRSA 2013-2014 dataset and all sites had assemblages with at least 30 taxa with associated tolerance values. The specific tolerance values are described as 95^{th} = 95^{th} percentile, A = average, WA = weighted average, PDP = partial dependence plot. The type of data used is specified in parentheses and is either p/a = presence/absence or abund = abundance.

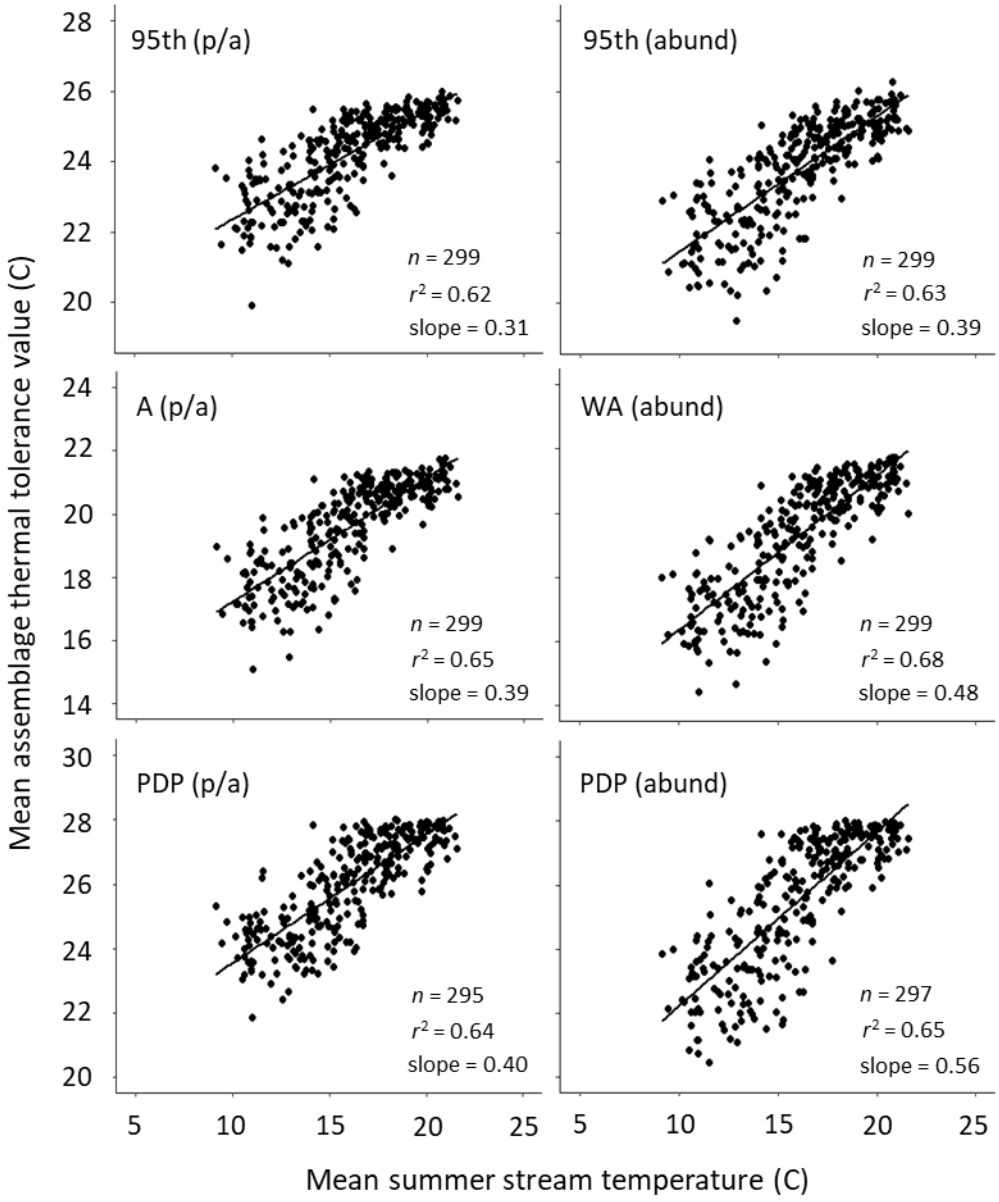


Figure F.3. Mean assemblage thermal tolerance values weighted by taxa abundances at each site and derived with the 6 different methods plotted against predicted mean summer stream temperature. The sites (*n*) were reference condition sites from the NRSA 2013-2014 dataset and all sites had assemblages with at least 10 taxa with associated tolerance values. The specific tolerance values are described as 95^{th} = 95^{th} percentile, A = average, WA = weighted average, PDP = partial dependence plot. The type of data used is specified in parentheses and is either p/a = presence/absence or abund = abundance.

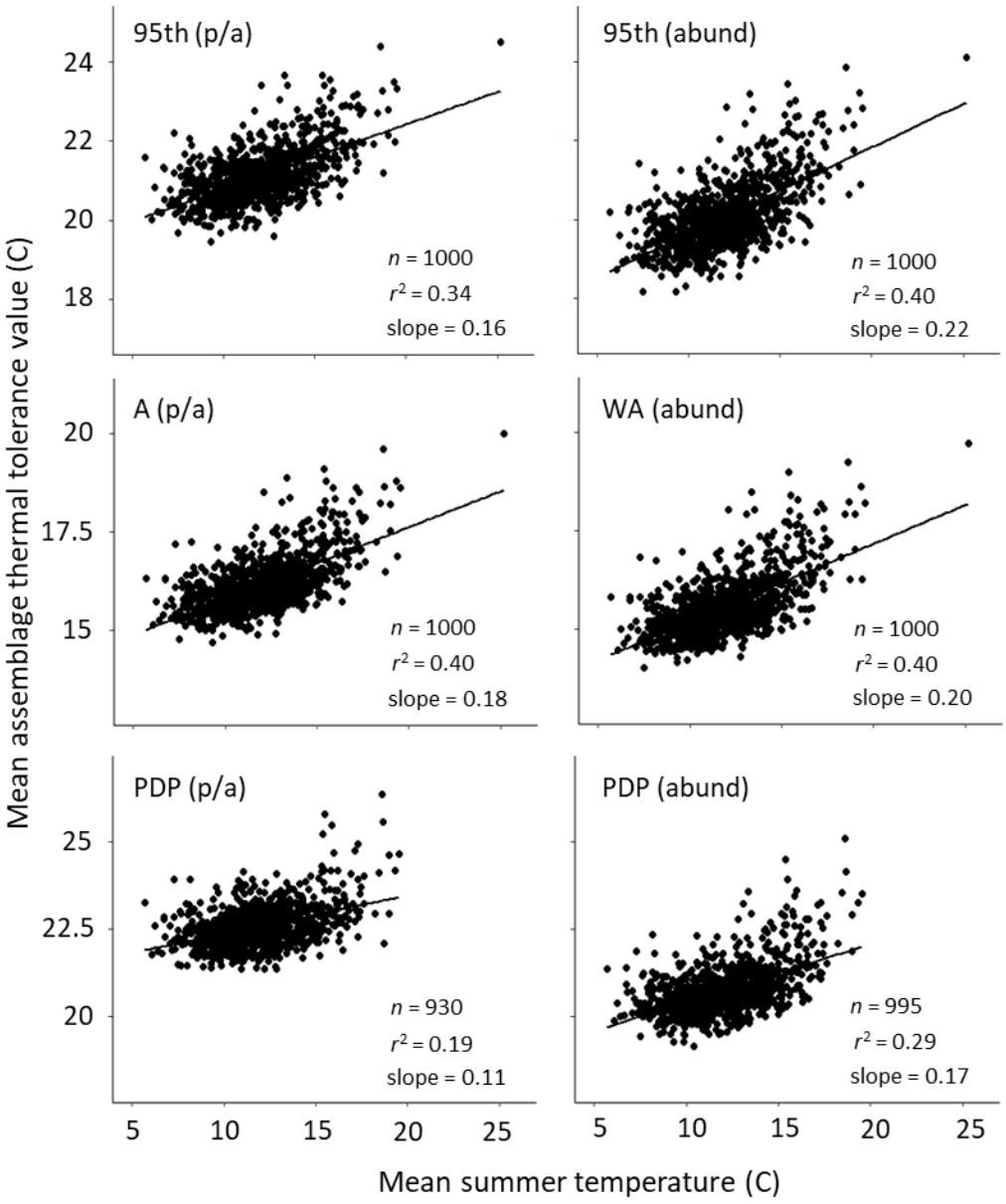


Fig F.4. Mean assemblage thermal tolerance values derived with the 6 different tolerance values plotted against mean summer site temperature. The sites (*n*) were from the PIBO dataset and all sites had assemblages with at least 10 taxa with associated tolerance values. The specific tolerance values are described as 95^{th} = 95^{th} percentile, A = average, WA = weighted average, PDP = partial dependence plot. The type of data used is specified in parentheses and is either p/a = presence/absence or abund = abundance.

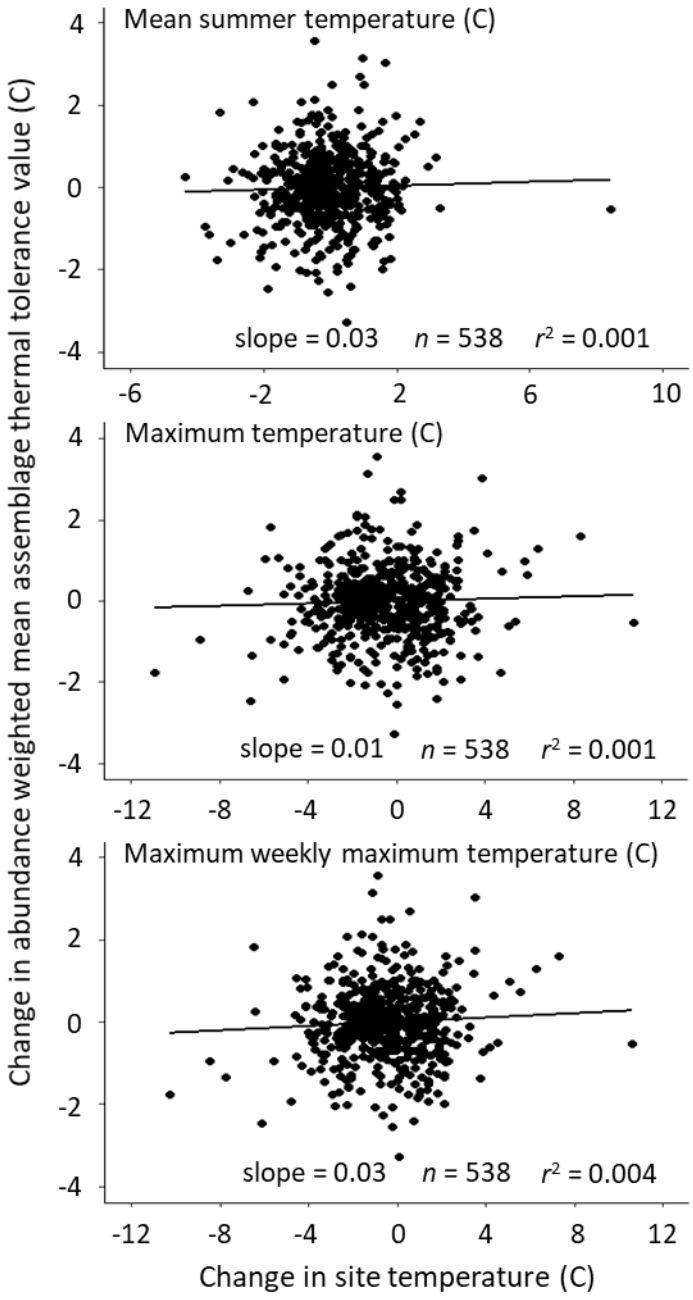


Figure F.5. Relationships between change in abundance weighted mean assemblage thermal tolerance values and change in mean summer site temperature, maximum site temperature, and maximum weekly maximum temperature for 538 PIBO sites that were sampled in two different years. The average (presence/absence) tolerance values were used to calculate mean assemblage thermal tolerance values and all samples had assemblages with at least 10 taxa with associated TVs.

Table F.1. Extent estimates of TBI values across the CONUS. The TBI was calculated with the PDP (abund) TVs. Cooler than expected means the TBI score was less than the 5th percentile of reference site TBI scores and warmer than expected means the TBI score was greater than the 95th percentile of reference site TBI scores. Expected means the TBI score falls within the 5th and 95th percentiles of reference site TBI scores shows little evidence that the invertebrate assemblage has been thermally altered.

Stream class	Stream length (%)	SE	Stream length (km)	SE
Cooler	3.4	0.7	60,899	12,274
Expected	90.6	1.0	1,619,640	53,638
Warmer	6.0	0.8	107,515	14,344

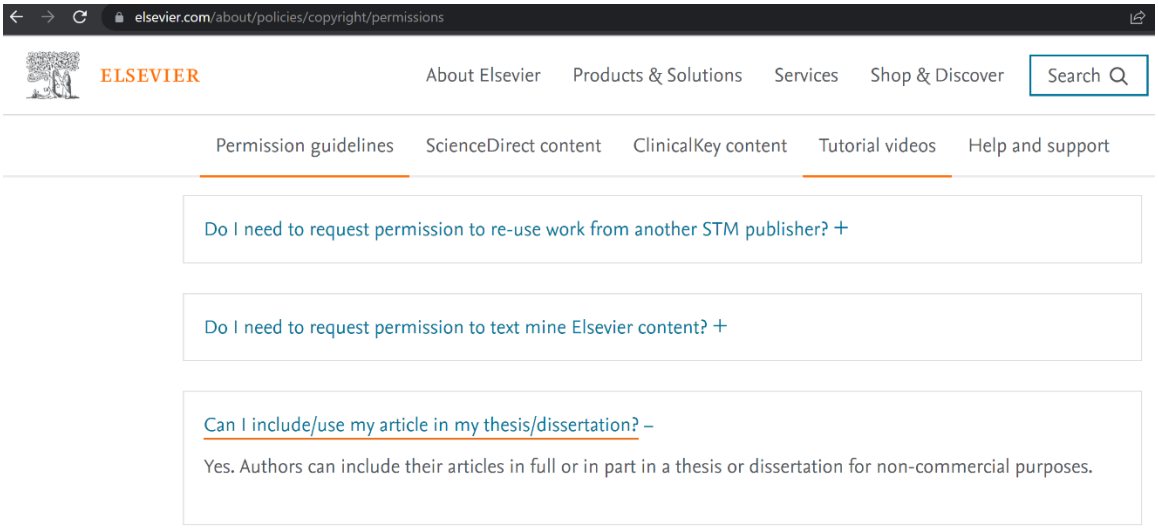
Appendix G. Permission-to-use confirmation

For chapter 3, permission was granted to reprint the chapter by the coauthor who was not also a signatory to the dissertation title page.

On Sun, Feb 27, 2022 at 5:30 PM Sam Schwartz <sam@cs.uoregon.edu> wrote:

> Absolutely!
>
> Sam
>
> Samuel David Schwartz
> Graduate Employee: Teacher and Researcher
> PhD Student | Cell: +1 801 739 3520
>
> UNIVERSITY OF OREGON | Computer and Information Science
>
> On 2/27/2022 1:22 PM, Donald Benkendorf wrote:
>
> > Hi Sam,
> >
> > Since you are a coauthor on the paper "CORRECTING FOR THE EFFECTS OF CLASS IMBALANCE IMPROVES THE PERFORMANCE OF MACHINE-LEARNING BASED SPECIES DISTRIBUTION MODELS", I need your permission in order to print the paper as a chapter in my dissertation. Do you give me permission to print our paper as a chapter in my dissertation? An email response from you will suffice and will be included in the appendix of my dissertation.
> >
> > Thank you!
> > Donald

For chapter 4, Elsevier, the publisher grants permission to authors to include their articles in dissertations.



Curriculum Vitae

Donald J. Benkendorf

Email: donald.benkendorf@usu.edu

Education

Utah State University

- Ph.D. in Ecology
  Defense Date: September 2021
- Dissertation: Refining, Testing, and Applying Thermal Species Distribution Models to Enhance Ecological Assessments
  -Advisor: Dr. Charles Hawkins

Murray State University

- M.S. in Watershed Science (Aquatic Ecology emphasis)
  Graduation Date: December 2017
- Thesis: Effects of Density and Size Structure on Top-Down Control by an Omnivore
  -Advisor: Dr. Howard Whiteman

Mansfield University of Pennsylvania

- B.S. in Fisheries Biology
  Graduation Date: May 2015
- Senior Research Project: The Effect of a Municipal Wastewater Treatment Plant Effluent on Macroinvertebrate Communities in the Conestoga River
  -Advisor: Dr. John Kirby

Professional Experience

EPA National Aquatic Resource Research Fellow (9/2021-current)

- Participated as an ORISE research participant with the US EPA's National Aquatic Resource Surveys team in the Office of Water

Graduate Research Fellow- Utah State University (8/2017-9/2021)

- Developed a temperature-specific biotic index and evaluated its utility for diagnosing thermal impairment of aquatic life.
- Evaluated the potential of several methods to improve performance of machine learning based species distribution models.

Graduate Field Technician- Eastside Type N Riparian Effectiveness Project (8/2017-9/2021)

- Oversaw collection of eDNA and traditional benthic samples from headwater streams in eastern Washington state and oversaw an undergraduate field technician. Samples were collected three times a year and were used to assess before and after effects of timber harvest.

Graduate Research Assistant- Murray State University (5/2015-8/2017)

- Conducted an experiment with a 35 mesocosm array to test effects of density and size structure on top-down control by Speckled Dace.

Graduate Field Technician- Kimball Creek Restoration Project (5/2015-8/2017)

- For two summers, led long-term sampling effort on a degraded third-order stream that was a candidate to undergo a large restoration.
- Managed an undergraduate research technician and trained new graduate students during the field season.

Native Fish Conservation Intern in Yellowstone National Park (5/2013-8/2013)

- Worked with Yellowstone National Park fisheries biologists and technicians and assisted in trout population assessments and conservation initiatives within Yellowstone National Park.

Publications (*in review or prep)

- ***Benkendorf DJ**, Hawkins CP. *In prep*. Validating field-derived estimates of the thermal niche with longer-term temperature experiments. Goal for submission to *Freshwater Science*.
- ***Benkendorf DJ**, Hawkins CP. *In prep*. Diagnosing the causes of altered biodiversity in freshwater ecosystems: development and evaluation of a temperature-specific biotic index. Goal for submission to *Ecological Indicators*.
- ***Benkendorf DJ**, Schwartz SD, Cutler DR, Hawkins CP. *In prep*. A systematic comparison of class imbalance methods and machine learning algorithms on species distribution model performance. Goal for submission to *Ecological Modelling*.
- **Benkendorf DJ**, Whiteman HH. *(2021*) Omnivore density affects community structure through multiple trophic cascades. *Oecologia* 195, 397-407. https://doi.org/10.1007/s00442-020-04836-0

- **Benkendorf DJ,** Hawkins CP. *(2020*) Effects of sample size and network depth on a deep learning approach to species distribution modeling. *Ecological Informatics* 60, 101137. https://doi.org/10.1016/j.ecoinf.2020.101137

Presentations

- **Society of Freshwater Science (5/2021)**
  -Online video presentation on "Diagnosing the Causes of Altered Biodiversity in Freshwater Ecosystems: Development, Evaluation, and Interpretation of a Temperature-Specific Biotic Index"
- **Ecological Society of America Conference (8/2020)**
  -Online Poster Presentation on "Effects of Sample Size and Network Depth on a Deep Learning Approach to Species Distribution Modeling"
- **Society of Freshwater Science (5/2019)**
  -Oral Presentation on "Growth and Survival Jointly Predict the Upper Thermal Limits of the Stonefly *Pteronarcys californica*"
- **WATS Graduate Research Symposium (4/2019)**
  -Oral Presentation on "Validating the Interpretation of Thermal Species Distribution Models to Test Macroecological Hypotheses and Enhance Ecological Assessments"
- **Ecological Society of America Conference (8/2017)**
  -Poster Presentation on "Effects of Density and Size Structure on Top-Down Control by an Omnivore"
- **Rocky Mountain Stream Restoration Conference (6/2017)**
  -Oral Presentation on "Intraspecific Variation and Ecosystem Function: Implications for more effective Post-Restoration Monitoring"
- **Watershed Studies Institute Research Symposium (4/2017)**
  -Oral Presentation on "Effects of Density and Size Structure on Omnivorous Trophic Cascades"
- **Midwest Ecology and Evolution Conference (3/2017)**
  -Oral Presentation on "Effects of Density and Size Structure on Omnivorous Trophic Cascades"
- **Sigma Xi Symposium (2/2017)**
  -Oral Presentation on "Effects of Density and Size Structure on Top-Down Control by an Omnivore"
- **BIO 330- Principles of Ecology (11/2016)**
  -Guest Lecture in Community Ecology on "Effects of Intraspecific Variation on Trophic Cascades"
- **Kentucky Academy of Sciences (11/2016)**

-Oral Presentation on "Effects of Density and Size Structure on Top-Down Control by an Omnivore" (2nd place finisher- best oral presentation)

- **Watershed Studies Institute Research Symposium (4/2016)**
  -Oral Presentation on "Effects of Density and Size Structure on Top-Down Control by an Omnivore"
- **Mansfield University Senior Seminar (05/2015)**
  -Oral Presentation on "The Effect of a Municipal Wastewater Treatment Plant Effluent on Macroinvertebrate Communities in the Conestoga River"
- **Commonwealth of Pennsylvania University Biologists Meeting (04/2015)**
  -Oral Platform Presentation on "The Effect of a Municipal Wastewater Treatment Plant Effluent on Macroinvertebrate Communities in the Conestoga River"
- **Mansfield University Showcase of Student Scholarship (04/2015)**
  -Oral Presentation on "The Effect of a Municipal Wastewater Treatment Plant Effluent on Macroinvertebrate Communities in the Conestoga River"

Awards

- Awarded EPA National Aquatic Resource Research Fellowship (current)
- Awarded Presidential Doctoral Research Fellowship (2017-2021)
- Sigma Xi Outstanding Student Research Award (2017)
- 2nd place for best Oral Platform Presentation at Kentucky Academy of Sciences on "Effects of Density and Size Structure on Top-Down Control by an Omnivore" (2016)
- North Hall Prize for Best Research Paper (2016)- "The Effect of a Municipal Wastewater Treatment Plant Effluent on Macroinvertebrate Communities in the Conestoga River"
- Mansfield University's Outstanding Senior in Biology (2015)
- Stanley Henry Nauman Memorial Award for academic achievement in Wildlife and Fisheries Science (2014)