2022

FACILITATING AQUATIC INVASIVE SPECIES MANAGEMENT USING SATELLITE REMOTE SENSING AND MACHINE LEARNING FRAMEWORKS

Sean C. Carter
University of Montana, Missoula

FACILITATING AQUATIC INVASIVE SPECIES MANAGEMENT USING SATELLITE REMOTE SENSING AND MACHINE LEARNING FRAMEWORKS

By
SEAN C. CARTER

B.A. Organismal Biology and Ecology, Colorado College, Colorado Springs, CO, 2016

Thesis

presented for partial completion of the requirements

for the degree of

Master of Science

in Systems Ecology

The University of Montana

Missoula, MT

May 2022

Approved by:

Scott Whittenberg, Dean of The Graduate School
Graduate School

John S. Kimball (chair)
W.A. Franke College of Forestry & Conservation, University of Montana

Marco P. Maneta
Department of Geosciences, University of Montana

Paul M. Lukacs
Department of Ecosystem and Conservation Sciences, University of Montana

ABSTRACT

Carter, Sean C., February 2022 Systems Ecology

Facilitating Aquatic Invasive Species Management using Satellite Remote Sensing and Machine Learning Algorithms

Chairperson: John Kimball

The urgent decision-making needs of invasive species managers can be better met by the integration of biodiversity big data with large-domain models and environmental data products in the form of new workflows and tools that facilitate data utilization across platforms. Timely risk assessments allow for the spatial prioritization of monitoring that could streamline invasive species management paradigms and invasive species' ability to prevent irreversible damage, such that decision makers can focus surveillance and intervention efforts where they are likely to be most effective under budgetary and resource constraints. I present a workflow that generates rapid spatial risk assessments on aquatic invasive species by combining occurrence data, spatially explicit environmental data, and an ensemble approach to species distribution modeling using five machine learning algorithms. For proof of concept and validation, I tested this workflow using extensive spatial and temporal occurrence data from Rainbow Trout (RBT; *Oncorhynchus* mykiss) invasion in the upper Flathead River system in northwestern Montana, USA. Due to this workflow's high performance against cross-validated datasets (87% accuracy) and congruence with known drivers of RBT invasion, I developed a tool that generates agile risk assessments based on the above workflow and suggest that it can be generalized to broader spatial and taxonomic scales in order to provide data-driven management information for early detection of potential invaders. I then use this tool as technical input for a management framework that provides guidance for users to incorporate and synthesize the component features of the workflow and toolkit to derive actionable insight in an efficient manner.

ACKNOWLEDGEMENTS

I would like to thank my personal community for the company and inspiration they provided as I worked on this thesis. I am honored to have shared my joys and suffering with so many supportive, authentic, and vulnerable people over the past few years. This journey would have not been the same without you, and for that I am tremendously grateful. In addition, I would like to thank my advisor, John, for providing me so many varied and engaging opportunities to develop my research skills at the Numerical Terradynamic Simulation Group, for providing a vision of academia that is aligned with my values, and for giving me the space to explore and find my own voice within science as a whole. Thanks to my committee members, John, Marco, and Paul, for providing tireless and insightful feedback on my work. Thanks to my project collaborators, Dr. Charles van Rees and Dr. Brian Hand, for providing patient and thoughtful input throughout the whole project, and interns Randy Flores, Myles Stokowski and Kevin Christensen, for sharing the workload and pushing me to be better. Lastly, I would like to thank my many academic mentors, in particular Dr. Jon Graham and Dr. Maury Valette, whose selfless instruction and committed coursework empowered me to think critically and understand the meaning of true learning. Thanks to all of your guidance, I'm a better person, better scientist, and better student. I'm excited to use my newfound skills for an engaging and gratifying career in Earth system science.

Table of Contents

Chapter 1

1 INTRODUCTION

1.1 BACKGROUND

Non-native, Invasive Species (IS) have drastic impacts on biodiversity and ecosystem services (Bellard et al., 2016, Walsh et al., 2016). The pace of biological invasions show no evidence of slowing down (Seebens et al., 2017). This situation results in an urgent need to both understand and mitigate IS establishment. Although the underlying mechanism is contextually dependent, the conditions necessary for successful establishment of a given non-native species are a combination of high propagule pressure, adequate resource availability, and favorable ecological circumstances (Enders et al., 2020; Pyšek and Richardson, 2010). Despite the lack of comprehensive understanding of the mechanism for biological invasions, averting damage from IS is highly time sensitive. Thus, preventative measures must prioritize mitigation over furthering mechanistic understandings of IS establishment.

Efficient and proactive strategies to avoid effects of IS establishment involve some combination of eradication and anticipatory prevention (Zanden et al., 2010). For example, there is evidence that early and aggressive measures can prevent major economic damages, but only when search efforts are targeted in areas of high risk (Kaiser and Burnett, 2010). In addition, preventative measures have historically been determined to be the most cost-effective approach to IS management (Leung et al., 2002), although this understanding has shifted in recent years

(Zanden et al., 2010). Regardless of the measures taken to mitigate severe ecological and economic repercussions, spatial prioritization of management efforts are greatly furthered with proactive approaches and predictive modeling (Ricciardi et al., 2017).

Various management frameworks have emerged to provide integrated and coordinated actions to mitigate the potentially drastic effects of IS establishment. These frameworks range from sustainability-oriented and objective-based policies that emphasize the equilibrium among various target "pillars" (Larson et al., 2011) to adaptive management paradigms that couple management actions with feedback from key performance indices (Foxcroft and Mcgeoch 2011). Although the efficacy of different management frameworks has not been evaluated, the choice of an optimal strategy requires a proactive assessment of potential costs of action and inaction that incorporates both heuristic preconceptions and technical inputs (Hyytiäinen et al., 2013, Hastings et al., 2005). Although each of these frameworks differs in their implementation, most emphasize proactive and rapid approaches.

Perhaps the most systematic and extensive IS management paradigm in the United States is Early Detection and Rapid Response (EDRR), a guiding doctrine designed to integrate and synthesize the conceptual and practical merits of various management frameworks such as those described above (Reaser et al., 2020). In calling for widespread and coordinated monitoring across agencies, it consists of a series of iterative, step-wise management actions that integrate informational and technical inputs with complex directives in a context-specific manner. Each action has different technological requirements and management goals. For example, the goal of the response measures action is to coordinate an informed and adept management reaction to the

detection of an IS that requires a sophisticated understanding of the potential risk and optimal strategy for minimizing the impact of the invader. Furthermore, the goal of target analysis is to prioritize species surveillance efforts in order to maximize the effectiveness of surveys through the use of ecological modeling and forecasting tools. Each stage of EDRR is coupled with informational and technical inputs from various sources. Providing and improving effective, yet efficient expediency in such input information remains an open challenge.

For instance, an integral difficulty of target analysis is the proactive modeling of potential establishment areas in a reliable and rapid manner. Computational approaches must represent, among other considerations, the habitat requirements of potential invaders and project these to geographic space while maintaining the flexibility to incorporate new training data as it is released (Morisette et al., 2020). Spatially prioritizing management areas in this way provides a favorable strategy for conducting monitoring efforts by identifying high-risk areas for the success of initial colonizers (Russell et al., 2017). Still, there remains a challenge in creating responsive and accurate representations of areas of that maximize the effectiveness and efficiency of invasive species detection (Berec et al., 2015; Wang et al., 2014). Major improvements could be made in the component spatial occurrence information used in such analytical techniques (Darling et al., 2007), the expansion of decision support tools that enable spatial prioritization of sampling efforts, and the development of rapid workflows that can integrate species occurrence information with geospatial data products representing key environmental controls on species distribution (Russel et al., 2017, van Rees et al., 2021).

The improvement of proactive modeling efforts necessitates a tradeoff between highly credible (and thus time-intensive) efforts and automated methods (Young et al., 2020). Striking the balance between time intensiveness and credibility remains a challenge and hinges on assembling either higher quality, narrow-use input information or readily available databases and data products. In particular, the availability of environmental data forces users to consider the limitation of various products against the value obtained by representing key aspects of organismal niche requirements. For example, the timing and duration of peak flow events is known to drive Rainbow Trout distributions in the northern Flathead River system (Muhfeld et al., 2017; Muhfeld et al., 2014), but there are few data products that are both temporally explicit and (spatially) high resolution, so modeling efforts might use static, proximal cues such as a surface water extent index (e.g., Pekel et al., 2016), consolidated to ecologically relevant polygons, and low-resolution spatially interpolated precipitation data (e.g., gridMET - Abatzoglou 2013; NLDAS -Mitchell 2004) to represent this essential driver of Rainbow Trout occupancy. Indeed, selection of appropriate environmental data is an integral part of improving proactive modeling for target analysis and requires knowledge on the constraints of various data types as well as information on the species' distributional requirements.

Data products appropriate for EDRR differ in terms of their spatiotemporal domains and resolutions. Possible data sources include in situ sensor networks, spatially interpolated climatic products, and remotely sensed imagery. Each source of environmental information has potential advantages and drawbacks. For example, because stream sensors are placed in specific site locations, they offer continuous, high spatiotemporal resolution information, but only in those specific locations. Similarly, due to being derived from networks of weather stations,

meteorological data products provide high temporal resolution (i.e. daily or hourly), but compromise their spatial resolution because their spatial continuity is a result of interpolation. In addition, this interpolation results in high uncertainty in areas with sparse geographic coverage. Lastly, remote sensing data, due to being directly observed with regular pass-over intervals, offers a balance between these two extremes. Although affected by observational constrains including solar illumination and atmospheric contaminants, remote sensing imagery has various advantages. For example, because these products are derived from direct observations of the Earth's surface, the burden of geographic uncertainty is mitigated. In addition, these data offer spatially contiguous observations, and open access products are available at relatively low cost to the user.

Remote sensing data include static and temporally dynamic products that are either highly processed or relatively unrefined. Static products such as digital elevation models and high level structural products (e.g. percent tree cover – Hansen et al., 2003; surface water occurrence – Pekel et al., 2016) provide ecological information that is distinctly interpretable, but suffers from not being temporally explicit. Temporally dynamic products can be low level (e.g. raw LANDSAT imagery) or high level (e.g. MODIS Land Surface Temperature- Wan et al., 2015), and their usability depends on the specific EDRR application.

The constraints of different data types must be balanced with the potential insight that can be derived from each product. For example, if one is interested in prediction alone, it would be logical to use lower-level data products that may offer higher resolution than high-level products that are a result of secondary modeling efforts. On the other hand, if one is interested in deriving

ecological insight, higher-level products may be more suited to this purpose due their more direct link to biophysical conditions experienced by organisms.

Capturing, representing, modeling, and projecting the environmental processes that drive species distributions at spatiotemporal scales relevant to IS managers is a major challenge. For the most part, integrating open access remote sensing data with biodiversity big data provides a favorable opportunity for addressing the timely needs of EDRR that compromise precision for automation (Randin et al., 2020; Reaser et al., 2020). However, it remains unclear the degree to which the rapid demands of EDRR can be met by linking species occurrence information with remote sensing and other geospatial imagery. Even though remote sensing data are expected to shape the "next generation" of species distribution models (He et al., 2015) and hold innumerable supposed advantages, they are unable to capture all relevant dimensions of the organismal niche space, particularly for aquatic species. Thus, there is a clear need for workflows that leverage the advantages of various types of environmental data layers in order to facilitate IS management within the EDRR framework.

1.2 RESEARCH QUESTIONS

To test whether it is possible to improve modeling efforts for the technical input required by EDRR, this thesis introduces an empirical machine learning framework to facilitate monitoring and forecasting of the risk of colonization, secondary spread, or establishment for aquatic IS using a habitat suitability framework. In doing so, it proposes a workflow that integrates species occurrence information with various readily available data products, evaluates the efficacy of

this workflow, and examines how that workflow might fit into the EDRR management paradigm. To guide this research, I ask the following research questions:

1. How can species occurrence information be integrated with remotely sensed and other geospatial imagery to inform invasive species management decisions?

2. How can workflows linking existing databases with modeling technologies facilitate more efficient, effective spatial prioritization of IS monitoring and intervention while augmenting existing management frameworks?

To answer question 1, I develop, implement, and validate a data pipeline that links point-level species occurrence information with readily available environmental data that has balanced the advantages of different types of geospatial environmental data with capturing the relevant habitat requirements of the species of interest. These rasterized data products were chosen to leverage the advantages inherent to their method of observation (i.e., directly observed -vs- spatially interpolated; static -vs- dynamic) and represent vital aspects of freshwater aquatic ecosystems. The pipeline generates rapid spatial risk assessments on aquatic IS using an ensemble approach to species distribution modeling with five machine learning algorithms. In order to determine the degree to which my approach can inform management decisions, this workflow is tested against a well-studied invasion system to determine whether it can result in insights and predictions that are well-aligned with expert knowledge.

To answer question 2, the workflow developed in question 1 is used to develop a graphical user interface tool that generates agile risk assessments in a user-friendly manner. In

addition, it is published in a software package that allows users to better investigate the underlying drivers of occurrence and model performance against various covariates. These tools are then used to consider a hypothetical management exercise that feeds into the early detection and rapid response target analysis action.

1.3 THESIS OUTLINE

This thesis is broken into four chapters. In Chapter 2, a workflow for linking species occurrence information is developed, implemented, and tested against a well-known invasion system in the northern Montana region of the United States. This chapter provides the basis to answer Research Question 1 and has been published in the journal *Frontiers in Big Data* (Carter et al. 2021). Chapter 3 of this thesis addresses Research Question 2 and develops a geospatial toolbox that implements the Chapter 2 workflow in a user-friendly manner. In addition, it proposes a general management framework that uses these tools as technical input for target analysis, a vital action in the EDRR paradigm.[1] The exercise provided in Chapter 3 will help managers use the workflow to facilitate the spatial prioritization of management actions within a given area. In addition, it will enable the use of my workflow to confront and support existing knowledge of species' niche requirements in order to direct the monitoring of shifting environmental conditions as temporally explicit data products are continually released. Lastly, this thesis concludes with chapter four that directly answers both research questions, summarizes major findings, limitations, and broader impacts from this work and recommends next steps and future research.

[1] I use the term *technical input* to describe the practical information used to evaluate and inform potential management actions. This information can come from modeling output or biological knowledge.

Chapter 2

2 TESTING A GENERALIZABLE MACHINE LEARNING WORKFLOW FOR AQUATIC INVASIVE SPECIES ON RAINBOW TROUT (*ONCORHYNCHUS MYKISS*) IN NORTHWEST MONTANA

ABSTRACT

Biological invasions are accelerating worldwide, causing major ecological and economic impacts in aquatic ecosystems. The urgent decision-making needs of invasive species managers can be better met by the integration of biodiversity big data with large-domain models and data-driven products. Remotely sensed data products can be combined with existing invasive species occurrence data via machine learning models to provide the proactive spatial risk analysis necessary for implementing coordinated and agile management paradigms across large scales. We present a workflow that generates rapid spatial risk assessments on aquatic invasive species using occurrence data, spatially explicit environmental data, and an ensemble approach to species distribution modeling using five machine learning algorithms. For proof of concept and validation, we tested this workflow using extensive spatial and temporal hybridization and occurrence data from a well-studied, ongoing, and climate-driven species invasion in the upper Flathead River system in northwestern Montana, USA. Rainbow Trout (RBT; Oncorhynchus mykiss), an introduced species in the Flathead River Basin, compete and readily hybridize with native West-slope Cutthroat Trout (WCT; O. clarkia lewisii), and the spread of RBT individuals and their alleles has been tracked for decades. We used remotely sensed and other geospatial data as key environmental predictors for projecting resultant habitat suitability to geographic space. The ensemble modeling technique yielded high accuracy predictions relative to 30-fold cross-

validated datasets (87% 30-fold cross-validated accuracy score). Both top predictors and model performance relative to these predictors matched current understanding of the drivers of RBT invasion and habitat suitability, indicating that temperature is a major factor influencing the spread of invasive RBT and hybridization with native WCT. The congruence between more time-consuming modeling approaches and our rapid machine-learning approach suggest that this workflow could be applied more broadly to provide data-driven management information for early detection of potential invaders.

2.1 INTRODUCTION

Non-native, Invasive Species (IS) are causing severe biological and economic disruption worldwide (Sepulveda et al., 2012; Shackleton et. al 2019). IS are the second most prevalent driver of species extinctions (Bellard et al. 2015), with estimated financial damages amounting to over a hundred billion dollars annually in certain individual countries (Pimentel 2002; Bradshaw et al. 2016). Continued anthropogenic landscape change and climate change may favor invaders by shifting competitive relationships with native species (Hellmann et al. 2008). Aquatic IS represent a particular threat to freshwater ecosystems due to their high potential for establishment and spread, and severe ecosystem impacts (Havel et al., 2015). The current and predominant paradigm for IS management is Early Detection and Rapid Response (EDRR), but the intensive resources and surveillance involved in this framework's implementation may be prohibitive without new and innovative uses of technology (Martinez et al., 2020). EDRR depends on frequent, widespread, and ongoing monitoring to enable timely response, but such monitoring is extremely labor intensive and likely beyond the capabilities of many management actors. Timely risk assessments allow for the spatial prioritization of monitoring that could

streamline EDRR and its ability to prevent irreversible damage (Reaser et al., 2020a; Martinez et al., 2020), such that decision makers can focus surveillance and intervention efforts where they are likely to be most effective under budgetary and resource constraints. Such prioritizations are often based on heuristic preconceptions rather than data-driven approaches, and as such are neither repeatable nor transparent for system stakeholders. By contrast, scientifically-informed, formal target analysis may lack adequate temporal agility and accurate risk assessments. Many conventional modeling approaches to knowledge creation operate on long time scales (months to years) which may not be helpful to managers. Indeed, current modeling methodologies fail to provide managers with sufficient decision-making information in near real-time (Bayliss et al., 2013).

Given the finite supply of resources and quick timelines for IS management, there is a need for improved expediency and accuracy in identifying areas of highest vulnerability to IS establishment.

Species Distribution Models (SDMs) have been widely applied as spatial decision support tools for IS managers (Srivastava et al., 2019) and can be broadly categorized into mechanistic and correlative model classes (Elith et al., 2015). Process-based, or mechanistic, models require considerable developmental and computational effort (Kearney and Porter 2009) and can thus be out of sync with the needs for timely analyses for EDRR (Merow et al., 2011). These models rely on exhaustive, experimentally derived functional characteristics (Shabani et al., 2016) or hierarchal frameworks that are built to elucidate or test hypotheses about ecological

relationships rather than simply predict patterns in species occurrence (see Muhlfeld et al., 2014 and 2017; Berthon 2015; Farley et al., 2018).

On the other hand, correlative SDMs require less mechanistic understanding and instead rely on apparent relationships between species and environmental characteristics. Such models are comparatively quick to train and develop, but are often built using low-resolution spatially interpolated climatic data, such as WorldClim (Elith et al. 2010; Fourcade et al. 2014; Hijmans et al. 2005). Since the WorldClim data (Fick and Hijmans 2017) are not temporally explicit, and static covariates, by definition, cannot adequately provide a temporally continuous evaluation of risk, the value of these data for EDRR is hampered. Although a major drawback of these correlative models is that long-term extrapolation is more difficult, this disadvantage is outweighed by the acute need for rapid risk assessments to inform IS monitoring and biosurveillance. Indeed, facilitating IS management within the EDRR framework would be significantly improved by new workflows that can identify readily available drivers of invasion and establish relative invasion risk within the operational time scales of managers.

Many of the challenges outlined above can be met by data-driven and iterative workflows made possible by machine learning (ML) and the big data revolution (Runting et al, 2020). For instance, one challenge is the need for scalable and fast modeling workflows to guide managers and decision-makers (Reaser et al., 2020a). ML algorithms are an increasingly viable method for many modeling problems involving big data, particularly when the primary objective is to achieve high levels of predictive accuracy rather than develop a mechanistic understanding of the study system. ML algorithms, particularly non-parametric iterative algorithms (e.g. random

forests), are free from many strict assumptions such as independent observations and the need to avoid collinearity (Thessen 2016; Olden et al., 2008). In addition, ML models are well suited to the iterative modeling framework due to their automated approach, fast development process (Tarca et al., 2007), and highly scalable nature (Farley et al., 2018). This enables them to take advantage of other big data attributes, including its widespread proliferation, global coverage, and rapid updating (Whitehead et al., 2020). As new data become available, ML frameworks can be updated to reflect new understanding.

However, ML models are not a panacea: because they are immensely complex and, with the exception of intricate Bayesian ML models, do not incorporate the underlying uncertainty of the data (Cressie et al., 2009), making inferences about underlying processes less straightforward and dependent on the type of model being used (Farley et al., 2018; Parr et al., 2020). Nevertheless, the rapid, iterative, and predictive characteristics of ML approaches are an excellent match for the analytical needs of EDRR implementation, which prioritize speed and adaptiveness over mechanistic understanding.

Another challenge of EDRR is the availability and distribution of environmental data typically used to assess relative habitat suitability (Randin et al., 2020). Conventional spatially interpolated climate data often require enormous developmental effort (Daly et al., 2005; Hijmans et al., 2005), which, when temporally explicit, can hinder their utility in developing models that meet the adaptive (e.g. annually repeating) demands of EDRR. Moreover, because they are based on interpolations from global weather stations, such products yield high model uncertainty in areas with sparse geographic coverage (Bedia et al., 2013).

In contrast, Remote Sensing (RS) products available from global polar-orbiting environmental satellites have regular revisit intervals ranging from 1-16 days and are derived from spatially explicit observations, so the burden of geographic uncertainty is mitigated. Indeed, because of the complimentary nature and spatial and temporal continuity of many operational satellite records, RS observational data are expected to shape the next generation of SDMs (He et al., 2015), and are the preferred or perhaps the only option for regional, continental, and global scale prediction of IS spread (Vaz et al., 2019). These products are sensitive to many environmental properties, such as surface temperature, that constrain and explain species' occurrences (Randin et al., 2020). These and other satellite-based measurements have rarely been applied to SDMs relative to spatially interpolated climate data products (Dittrich et al., 2019), and their use for assessing species' distributions has been increasing in recent years (Lausch et al., 2016; Randin et al., 2020).

Although the spatial and temporal continuity of RS data improves the transferability and precision of capturing ecological niche requirements in many terrestrial environments (Randin et al., 2020), stream environments represent a particular challenge in integrating technological advances with IS management. Because the 2-dimensional footprint of RS products is often larger than the footprint of streams, such products can only provide proxies for physiologically relevant conditions within the aquatic environment. Thus, models trained to link species occurrences with environmental remotely sensed information may fail to capture the actual processes experienced by aquatic organisms, and care must be taken to avoid spurious conclusions. Coherent workflows that link remote sensing data and machine learning

functionalities are especially needed for freshwater systems to mobilize myriad spatial products in data-driven aquatic IS risk analysis.

Here, we demonstrate one such workflow linking these technologies to produce rapid and adaptable species distribution modeling for spatial risk assessments of aquatic IS. To provide proof of concept, we implemented this workflow on a well-documented case study of a climate-assisted species invasion. This worked case study allowed us to assess not only the predictive accuracy of this approach but also whether it gives meaningful insights into the environmental drivers of habitat suitability for a focal IS. Our study objectives were to: 1) Identify the most effective remotely sensed proxies for characterizing habitat suitability (a proxy for invasion risk) for our focal IS (RBT; *Oncorhynchus mykiss*); 2) Construct habitat suitability maps for spatial risk assessments using a combination of RS data products and ML methods; and 3) Test the feasibility of ML models for iterative reassessment of IS risk screening efforts within the EDRR framework.

2.2 METHODS

2.2.1 STUDY SYSTEM

The study area encompassed the tributaries of upper Flathead River system extending over portions of northwestern Montana USA, and southern British Columbia and Alberta CA (Figure 1). These mountain streams flow through forested landscapes and host several native fish species including Westslope Cutthroat Trout (WCT; *Oncorhynchus clarki lewisi*). Stream temperature and the timing and duration of peak streamflow events are key ecological drivers in these

streams (Hauer et al., 2007), while the timing and intensity of snowmelt is a key driver influencing spring runoff in this system (Pederson et al. 2011; Wu et al. 2012).

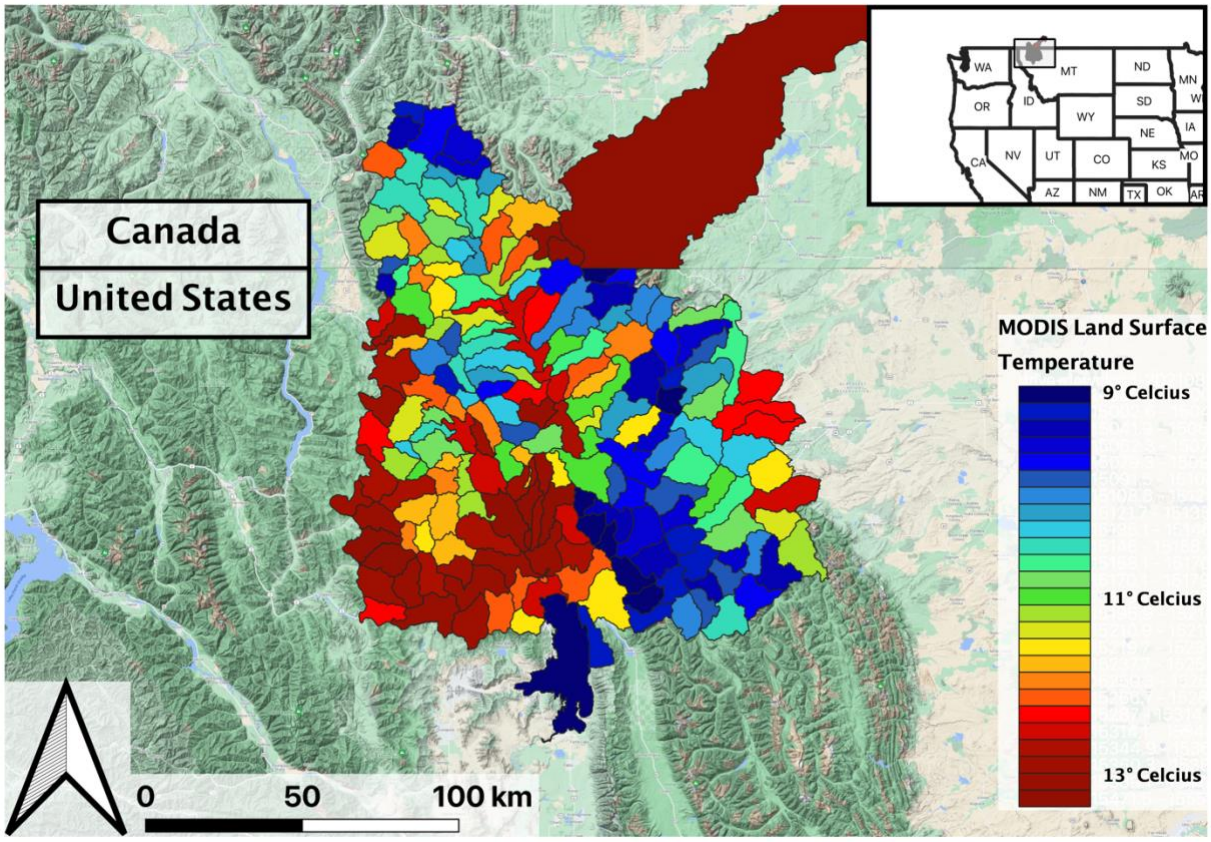

Figure 1. Overview of study area, including a sample data product (LST) aggregated by hydrologic units (HUCs).

Rainbow trout (*O. mykiss*) were artificially propagated and introduced into watersheds across the Continental US for recreational purposes between 1870 and 1971 (Pister 2001; Bennet et al. 2010). Since their introduction into the Flathead River in 1880 (Hitt et al., 2003), RBT have been hybridizing with native WCT (Allendorf et al. 2004; Hitt et al. 2003; Boyer et al. 2008; Muhlfeld et al. 2017). The impacts of RBT on WCT populations, particularly due to the spread of RBT individuals and their alleles, has been tracked for decades (Kovach et al., 2016). The spread of alleles appears to be driven more by legacy introductions, and thus propagule pressure,

than environmental conditions (Muhlfeld et al., 2017; Boyer et al., 2008). Relative to WCT, RBT prefer warmer temperatures, lower spring flows, earlier spring runoff, and tolerate greater environmental disturbance (Fausch et al. 2001; Muhlfeld et al., 2009a,b; Bear et al. 2007). During spawning, WCT generally migrate greater distances and spawn during peak flows, whereas RBT spawn earlier (i.e., during periods of lower flows) and lower in the river system (Muhlfeld et al., 2009b). High flows can affect both RBT and WCT, although reduced spring flows and warmer water temperatures have been associated with increased spread of RBT hybridization in the Flathead and across the northern Rockies (Muhlfeld et al., 2014; Muhlfeld et al. 2017), which are strongly influenced by spring precipitation, winter snowpack, and the timing of spring snowmelt (Pederson et al., 2011).

2.2.2 DATA ACQUISITION – GENETIC AND GENOMIC DATA

Trout have been periodically captured, sampled, and genotyped to assess the degree of RBT genetic admixture (the proportion of RBT alleles at the population level) in the study system since 2000. We used the associated long-term genetic monitoring data between years 2002 and 2019 as an index of RBT invasion. United States Geological Survey and Montana Fish Wildlife and Parks personnel selectively sampled streams where there was concern that WCT were hybridizing with non-native RBT, collecting fin clips from electrofished individuals and genotyping these individuals using various markers (microsatellites, SNPs, RAD-Capture sequencing). The genetic data were used to calculate RBT admixture in sampled populations.

2.2.3 DATA ACQUISITION – PRESENCE ABSENCE DATA

We generated a presence-absence dataset by classifying all occurrence records of less than 10% admixture to be “absent”. Although 10% still represents the presence of RBT alleles, conditions at these locations are less favorable for the establishment of this invasive taxon. Considering the difficulty of acquiring actual absence data (Jimenez-Valverde et al. 2008) and that many SDM's rely on ‘pseudo absences’ — background points used to characterize the range of environmental conditions in a given study area (Lobo et al. 2010) — we assume that these genotypic absences contain insightful information regarding the distribution of RBT, particularly in comparison to pseudo absences. We supplemented these absences with a RBT dataset acquired from the Non-indigenous Aquatic Species (NAS; USGS 2020) database and clipped these records to the bounding box of the RBT genetics dataset. We included only data records acquired after year 2002 to match the availability of RS data. We also corrected for the influence of spatial autocorrelation by systematically subsampling data records so that no two points fell within 500 meters of each other in a given year (Fourcade et al., 2014). The resultant occurrence dataset included 323 RBT presence locations and 167 absence point locations distributed across the study region over a 14 year record; the occurrence data were then joined to Hydrologic Unit Catchment polygons (HUC; Seaber et al., 1987). HUC polygons represent the landscape catchment area that drains to a portion of the stream network, whose hierarchical structure allows for a multi-scale delineation of drainage systems.

2.2.4 DATA ACQUISITION – ENVIRONMENTAL DATA LAYERS

To test whether proximal remote sensing cues contain sufficient environmental information to capture RBT niche requirements, we selected a number of readily available

satellite RS data products based on a priori assumptions of ecologically relevant drivers of hybridization and distribution (see below; Table 1). To avoid scale mismatch issues among predictors, we modeled environmental variables aggregated over HUC-12 polygons at the sub-watershed scale. Aggregating each covariate to HUC polygons mitigates the potential footprint mismatch between the RS observations and stream network within a catchment and is a common technique used in building freshwater SDMs in order to handle issues of scale relating to predictor variables (Friedrichs-Manthey et al., 2020). In addition, this method alleviates the inconsistent sampling inherent in the data and implicitly accommodates the mobile nature of RBT. Here, we give a brief description of the data products selected for model training and their connection to RBT niche requirements. The data products were preprocessed before being spatially aggregated to HUC-12 polygons as follows.

Environmental Covariate	Source	Description	Hypothesized Ecological Connection	Units	Resolution
Land Surface Temperature	MODIS AQUA LST MYD11A2 (V6; Wan et al., 2015)	Temperature on the surface of the Earth measured using thermal infrared passive sensors	Stream Temperature; Maximum annual temperature record	Kelvin	1 km
Precipitation[a]	National Land Data Assimilation System (NLDAS; Mitchell 2004)	Rain and snow accumulation, interpolated from weather stations and integrated with actively sensed radar products	Magnitude of peak flow events	kg/m^2	0.125 arc degrees 10 km
Flashiness[a]	USGS Dynamic Surface Water Extent Product (Jones 2018)	Annual per-pixel variation of a dynamic surface water extent algorithm; Derived from Landsat satellite imagery	Flood disturbances; Seasonal flow variation	Unitless	30 m
Surface Water Occurrence	JRC Global Surface Water Mapping Layers (Pekel et al. 2016)	Persistence of water on the surface; Derived from Landsat satellite imagery	Stream flow rates (at HUC - level aggregation); Habitat connectivity	Unitless	30 m
Topographic Diversity	Theobald et al. (2015)	Variation in temperature and moisture conditions available to species	Habitat structure and diversity	Unitless	90 m
Gross Primary Productivity	Robinson et al. (2018)	Amount of carbon captured by plants in an ecosystem; Derived from Landsat satellite imagery	Carbon available to the system	kg C/m^2/ 16-days	30 m
Normalized Difference Vegetation Index	MODIS AQUA MYD13A2 (V6) Vegetation Indices	Density of "greenness" on landscape	Photosynthetic Activity	Unitless	250 m
Enhanced Vegetation Index	MODIS AQUA MYD13A2 (V6) Vegetation Indices	Modified vegetation index that reduces atmospheric contamination and maintains sensitivity over dense vegetation	Photosynthetic Activity relative to Canopy Structure	Unitless	250 m
Percent Tree Cover	MODIS TERRA MOD44B (Hansen et al., 2003)	Percent of woody vegetation	Stream structure and habitat diversity	Percent cover	250 m
Heat Insolation Load	Theobald et al. (2015)	Incident radiation derived from latitude, slope, and aspect	Daily temperature variation; Stream Temperature	Unitless	90 m

Table 1. Library of ecologically relevant data products. [a]Preprocessed further from published products (see methods)

Land Surface "skin" Temperature (LST) observations were obtained from thermal-infrared measurements from the Moderate Resolution Imaging Spectroradiometer (MODIS) mounted on the NASA EOS Aqua satellite (Z. Wan et al., 2015; Li et al., 2013). The MODIS LST product is mapped to a 1-km resolution spatial grid similar to the sensor footprint. LST retrievals are acquired on a daily basis and composited over coarser eight-day intervals to reduce cloud and atmosphere contamination effects. The MODIS Aqua LST retrievals are acquired at 1330 local time from the sun-synchronous polar orbiting satellite and reflect mid-day conditions close to the maximum diurnal temperature range. Because trout species are limited by high temperature (Wenger 2011), we constructed a maximum composite image by capturing the maximum LST recorded in each grid cell for each year in our study period.

The National Land Data Assimilation System (NLDAS) uses a land surface model to integrate ground and space based observing systems, providing spatially explicit and temporally continuous estimates for various environmental variables including precipitation, potential evaporation, and specific humidity (Mitchell 2004) at 0.125 arc° and hourly resolutions. We aggregated the NLDAS precipitation product with a per-pixel sum composite at three-month seasonal intervals (i.e. Spring Precipitation, Summer Precipitation, etc).

The Dynamic Surface Water Extent (DSWE) product provides high temporal (8-day) repeat, moderate spatial resolution (30m) data on surface water inundation across broad spatial scales (Jones 2019). It uses an experimentally derived spectral mixture model and 5 rule-based decision criteria to classify Landsat surface reflectance pixels as "not water", "open water", or "partial surface water" in a spatially and temporally explicit manner. For each week in our study

period (i.e. 2002 - 2018), we gathered DSWE observations and generated a weekly per-pixel estimate of surface water inundation in our study area. We produced a surface water variation metric by finding the per-pixel temporal standard deviation within each year. The temporal standard deviation (as opposed to the IQR or variance) of the water variation was chosen as a proximal cue for stream flashiness due to its sensitivity to outliers, since RBT spawning is known to be sensitive to variations in stream flow rates.

In contrast to the DSWE product, the Landsat global surface water extent product identifies the presence of water over time using a mix of expert systems, visual analytics, and evidential reasoning (Pekel et al., 2016). Using this algorithm, Pekel et al. (2016) developed several thematic mapping layers including the Surface Water Occurrence metric, which quantifies the overall location and persistence of surface water cover at 30m spatial resolution from 1984 to present. The surface water persistence metrics are derived from the Landsat satellite series record, which provides consistent 30m spatial resolution and potential 16-day repeat coverage over the globe. However, actual spatial and temporal coverage of surface water dynamics is degraded by cloud and atmosphere contamination, seasonal reductions in solar illumination at higher latitudes, and overlying vegetation cover. Slow moving main-stem rivers generally have larger surface areas than lower order streams, so when spatially aggregated to HUC-level polygons, this product encapsulates information about flow rates and overall aquatic habitat connectivity.

Gross Primary Productivity (GPP) quantifies the plant photosynthetic uptake of atmospheric CO_2 and represents the amount of carbon and energy flow into the ecosystem. In

this study, a 30m resolution daily GPP record for the continental USA was used to characterize energy (and nutrients) available to ultimately support aquatic food webs. The GPP record is calculated using a modified form of the MOD17 light use efficiency algorithm driven by satellite observed fraction of photosynthetic active radiation (FPAR) derived from Landsat 30m spectral reflectances, gridded (4-km resolution) daily surface meteorology observations (i.e. gridMET; Abatzoglou 2013), and the national land cover database (Robinson et al. 2018). GPP has been used to predict freshwater fish species richness across the globe (Pelayo-Villamil et al., 2015), and previous research supports the link between primary production and fish productivity (Downing et al., 1990). Thus, this proximal product may contain information pertaining to the invertebrate community or vegetation structure. We calculated the accumulated annual GPP during each year of interest as a temporal sum composite, hypothesizing that the Landsat based GPP record captures bioenergetic constraints at scales relevant to RBT.

The MODIS Enhanced Vegetation Index (EVI; Didan 2015) is a modified version of the Normalized Difference Vegetation Index (NDVI), has improved sensitivity to green vegetation cover in high biomass regions, and minimizes atmospheric contamination effects. The MODIS (MOD13Q1) EVI product is derived globally at 250m, 16-day spatiotemporal resolutions. Because plants both absorb radiation in the visible spectrum and emit radiation in the near-infrared spectrum, the EVI is sensitive to the photosynthetic activity of terrestrial systems. Massicotte et al., (2015) used EVI as a proxy for aquatic vegetation biomass to predict larval fish abundance. Here, we used EVI as a proxy for the potential productivity of stream and riparian systems, where higher productivity systems would be more susceptible to invasion (i.e. hot

spots). Thus, we calculated a temporal EVI mean composite for each year to capture average conditions relevant to RBT.

The NASA MODIS Vegetation Continuous Fields (VCF) product provides a spatially continuous land cover estimate of general vegetation traits such as percent tree cover, percent non-tree cover, and percent barren land at 250m resolution and annual temporal fidelity (Hansen et al., 2003). The MODIS (MOD44B) VCF product is derived using a decision tree classification trained on MODIS surface reflectance and LST; we used the VCF percent tree cover metric to define the vegetative structure of the system within each HUC. The vegetation structure of various riparian areas has been linked to macro-invertebrate species richness (Death and Collier 2009; Sweeney 1993). We chose the VCF product to represent the overall disturbance and shadiness of a given HUC. Although GPP, EVI, and Percent Tree Cover quantify similar aspects of bioenergetic constraints, macro-invertebrate potential, and habitat structure, we expected to see differences in predictive power due to their differing resolutions, underlying algorithms, and retrieval accuracy.

In addition, topographic indices such as Topographic Diversity and Heat Insolation Load (Theobald et al., 2015) provide information about the topographic structure, microclimate variability, and resultant thermal dynamics of a given HUC. Topographic diversity is also congruent with the measurement of the heterogeneity of various landforms including valley bottom constraints, hills, and ridges as derived from a multi-scale neighborhood analysis. This metric indicates the structural diversity, and therefore the likelihood of connectivity of stream networks within watersheds. Heat Insolation Load reflects variations in latitude and incident

solar radiation to quantify the heat-loading capacity of different regions. Together with LST, heat insolation load provides a proximal cue to the overall stream temperature of a given HUC.

Covariates were obtained through data preprocessing performed within Google Earth Engine (GEE; Gorelick et al., 2017). We subjected each lower-level remote sensing variable (e.g. LST, GPP, EVI, Percent Tree Cover) to stringent quality filtering based on pre-published quality bands included in each product (see supplemental materials S1 for details). We kept the quality control filters inherent in the higher-level development products (e.g. Surface Water Occurrence, Heat-insolation Load). We intersected the RBT survey locations to their encompassing HUC12 catchments and calculated a weighted average of genetic admixture relative to the number of individuals in a dataset. For the RBT occurrence dataset, we simply aggregated occurrence points to the HUC level. We classified any HUC containing at least one presence location to be suitable. We then averaged each environmental covariate across all HUCs in our study area. This resulted in a tabular dataset with each column corresponding to the spatial average of an environmental covariate, or — depending on what our dependent variable was— a HUC-level weighted admixture percentage or HUC-level occurrence boolean. By taking HUC-level aggregates, we controlled for the effects of steep topography that concentrate environmental gradients at small spatial scales and the potential footprint mismatch between environmental data pixels and stream conditions. Although the same HUC may have been sampled in multiple years, we treated each HUC - year pair as an independent observation.

Data were exported from GEE, and due to the reliance of variable importance techniques on predictors being independent of one another, all covariates with a Pearson's correlation

coefficient > 0.7 were dropped (Dormann et al. 2013). In addition, because covariates may contain similar explanatory information but may not be represented by a linear relationship, we tested for multicollinearity (Mansfield and Helms 1982) by fitting Random Forest models with each covariate as an independent variable, and we dropped each variable that was shown to have a feature dependence score > 0.7 in predicting another variable. This process was repeated until no two columns had a partial dependency exceeding 0.6. This process resulted in 12 covariates: land surface temperature, surface water occurrence, heat insolation load, percent tree cover, flashiness, winter precipitation, fall precipitation, topographic diversity, summer precipitation, spring precipitation, gross primary productivity, and enhanced vegetation index. An overview of model inputs, outputs, and overall workflow can be found in Figure 2.

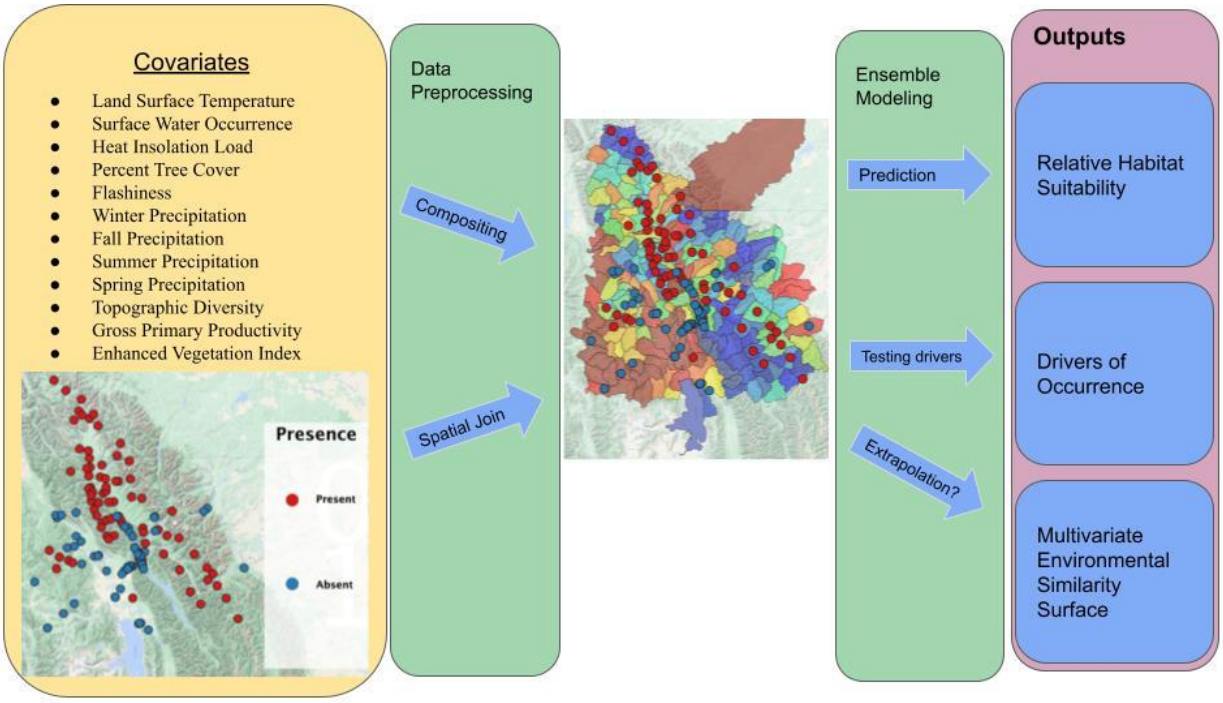

Figure 2. Overall workflow, model inputs, and model outputs. Yellow box (left) indicates model inputs. Green boxes indicate steps as referenced in the methods. Purple box (right) indicates each model output. RBT presence and absence observation locations are denoted by respective red and blue points on the associated study area maps.

2.2.5 ADMIXTURE MODEL TRAINING

Using the above covariates, we trained an ensemble of Linear Regression (GLM), Gradient Boosted Regressor (GBM), Classification Tree Regressor (CTA), Artificial Neural Network Regressor (ANN), XGBoost Regressor (XGB), and Random Forest Regressor (RF) models using sklearn version 0.23.1 in Python 3.7.7, with 20% of data randomly withheld for testing. We used the ensemble method because it has been shown to be an improvement over single models by reducing model-based uncertainty (Marmion et al., 2009; Elith et al., 2010). For a brief description of each component model, see supplemental materials (S2). Because the distribution of RBT hybridization was severely skewed toward higher rates (i.e. right skewed), we visually confirmed that testing data had similar distributions to training data. To consolidate model estimates, we implemented an ensemble method consisting of each of the above models, weighting the overall prediction by the mean absolute error (Willmott and Matsuura 2005) and omitting the artificial neural network due to severe inaccuracy.

2.2.6 PRESENCE ABSENCE MODEL TRAINING

The same covariates were used for both the hybridization and occurrence models. We implemented an ensemble method consisting of the classification analogues for the above regression models, again using Scikit-learn version 0.23.2 (Pedregosa et al., 2011). We took a weighted average of each component model prediction by the area under the receiver operative characteristic curve statistic (i.e. AUC score; Bradley 1997), omitting the GLM and ANN due to the unrealistic predictions (see below; Elith et al., 2010). For example, if the random forest model were to have a higher accuracy score than the decision tree model, the overall ensemble model prediction would be more influenced by the random forest than the decision tree. We

evaluated the predictive accuracy of the resultant ensemble model by computing a 30-fold cross validation accuracy score, where the training data was partitioned into 30 random segments of equal size, 29 of which were used to train the model, while the remaining segment was used to calculate the accuracy score. We calculated this accuracy score by computing the fraction of correct predictions of each segment, averaging the scores over all 30 folds for an overall metric of ensemble model accuracy. We then generated choropleth range maps (i.e. thematic maps showing summary statistics over a set number of polygons) by applying the ensemble of models to predict suitable habitat for mean covariates across two vector datasets representing the "first decade" (years 2002-2010) and the "second decade" (2010-2018) of the study period, each spatially aggregated to HUC level. Although each ensemble model predicted different presence amounts for the testing dataset, both the GLM and ANN did not show any variation of predicted suitability among first decade and second decade HUCs, so were removed from further analysis. To examine the degree of extrapolation, we calculated the Multivariate Environmental Similarity Surface (Elith et al., 2010) for each vector dataset. To examine the model prediction certainty, we calculated the standard deviation of prediction probabilities for each remaining estimator.

2.2.7 DISCERNING TOP PREDICTORS

To identify top predictors of RBT distributions, we implemented an ensemble of different feature importance techniques with each of the aforementioned ML models trained to predict occurrence and their analogues trained to predict hybridization. Each model was subject to Recursive Feature Elimination (Chen et al. 2018), Permutation Importance (Altmann et al. 2010), and Backwards Elimination (Draper and Smith 1981). These feature importance methods are similar, but with some important distinctions. Recursive Feature Elimination iteratively drops

features which have the smallest impact on model prediction until a pre-defined number of features is leftover. Permutation Importance iteratively shuffles the values of a given predictor, predicts using all covariates including the artificially permuted feature, and measures the subsequent drop in classification accuracy. The predictor whose permutation yields the largest drop in classification accuracy is identified as the most important predictor. Backwards Selection drops a single predictor entirely, retraining a different model for each iteration and again measuring the drop in predictive performance. The top three predictors were selected for each remaining model and importance technique, and we tallied the number of times a given predictor was found in the top three. We also interrogated partial dependency plots for known mechanisms driving occurrence and hybridization.

2.3 RESULTS

The tree-based methods (i.e. Random Forest, Decision Tree, Gradient Boosted Trees, XGBoost) yielded higher predictive accuracy than the linear and deep learning models for the RBT application (Table 2). Although the occurrence ANN and logistic regression models predicted a mix of RBT presence and absence for an unseen test dataset, both models predicted homogenous vectors of presence or absence for the first and second decades. For instance, the logistic regression predicted that all HUCs in both decades were suitable, and conversely, the ANN predicted that all HUCs in both decades were unsuitable. Similarly, both the hybridization ANN and linear regression models predicted unrealistic hybridization levels of 100% for every HUC, whereas all the tree-based regressors predicted RBT hybridization levels between 0 and 100%.

Occurrence	Model	Area Under the Curve Score
	Random Forest	**0.89**
	Logistic Regression	0.69
	Artificial Neural Network *	0.62
	Gradient Boosted Trees	0.84
	XGBoost	0.83
	Classification Tree	0.81
Hybridization	**Model**	**Mean Absolute Error**
	Random Forest	**0.05**
	Linear Regression	0.07
	Artificial Neural Network *	121.79
	Gradient Boosted Trees	**0.05**
	XGBoost	0.06
	Classification Tree	**0.05**

Table 2. Predictive capability of each ensemble model. Bold indicates highest accuracy models. Asterisk indicates models that were removed due to unrealistic predictions.

In evaluating the hybridization predictor (i.e. the ensemble of regression models), Land Surface Temperature, Heat Insolation Load, and Gross Primary Productivity were the most predictive features explaining RBT hybridization trends. The ensemble model also produced a favorable Mean Absolute Error of 5.5%. 90% of the residuals were less than 15% hybridization, although some predicted hybridization values had errors greater than 15%. Although observed hybridization percentages ranged from 0 to 100 %, admixture predictions only ranged from 0 to 60%. Choropleth maps trained on the hybridization dataset did not corresponded with known hybridization levels within the study area and showed unrealistic spatial patterning (i.e. checkerboarding rather than being spatially correlated) (Figure 3).

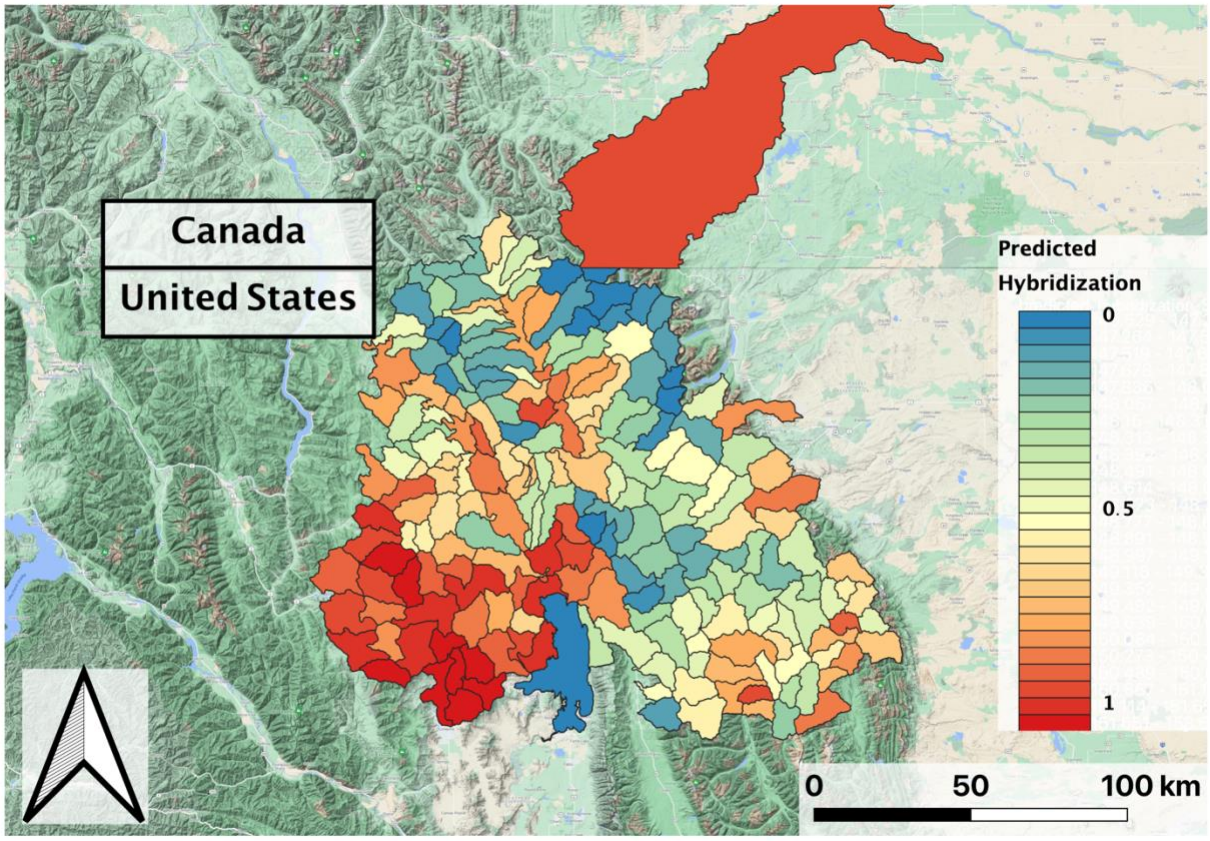

Figure 3. Predicted RBT hybridization for the second decade (2010-2018) composite, with dimensionless hybridization levels ranging from low (0) to high (1); black lines delineate individual HUCs within the larger study basin.

In evaluating the ensemble RBT occurrence model, we identified Land Surface Temperature, Surface Water Occurrence, and Heat Insolation Load as key predictive indices explaining RBT presence and absence (Figure 4). The model results also showed a favorable 30-fold cross validation accuracy score of 0.87. Surprisingly, Gross Primary Productivity did not show up as a top predictor of RBT occurrence, even though it was identified as a key predictor of RBT hybridization. Choropleth maps showed spatial patterns that agreed with known RBT occurrence records within the study area and reveal a strong tendency to predict high RBT relative suitability in main-stem rivers (Figure 5). In particular, the ensemble model predicted high relative suitability in the North Fork of the Flathead River basin and in the upper Flathead

River system for both the first and second decade. For a comparison of the component classifier predictions, see the supplemental materials (S3). The predicted RBT occurrences showed relatively small changes between the first and second decades. Although most predicted suitability differences were negligible, the ensemble model predicted a large degree of decreasing RBT suitability in the Salish and Lewis mountains, with increased suitability in the northern Mission mountains and East Glacier Park regions (Figure 6). The multivariate environmental similarity surface map shows that most HUCs fall within reasonable extrapolation distance from training locations (Figure 7).

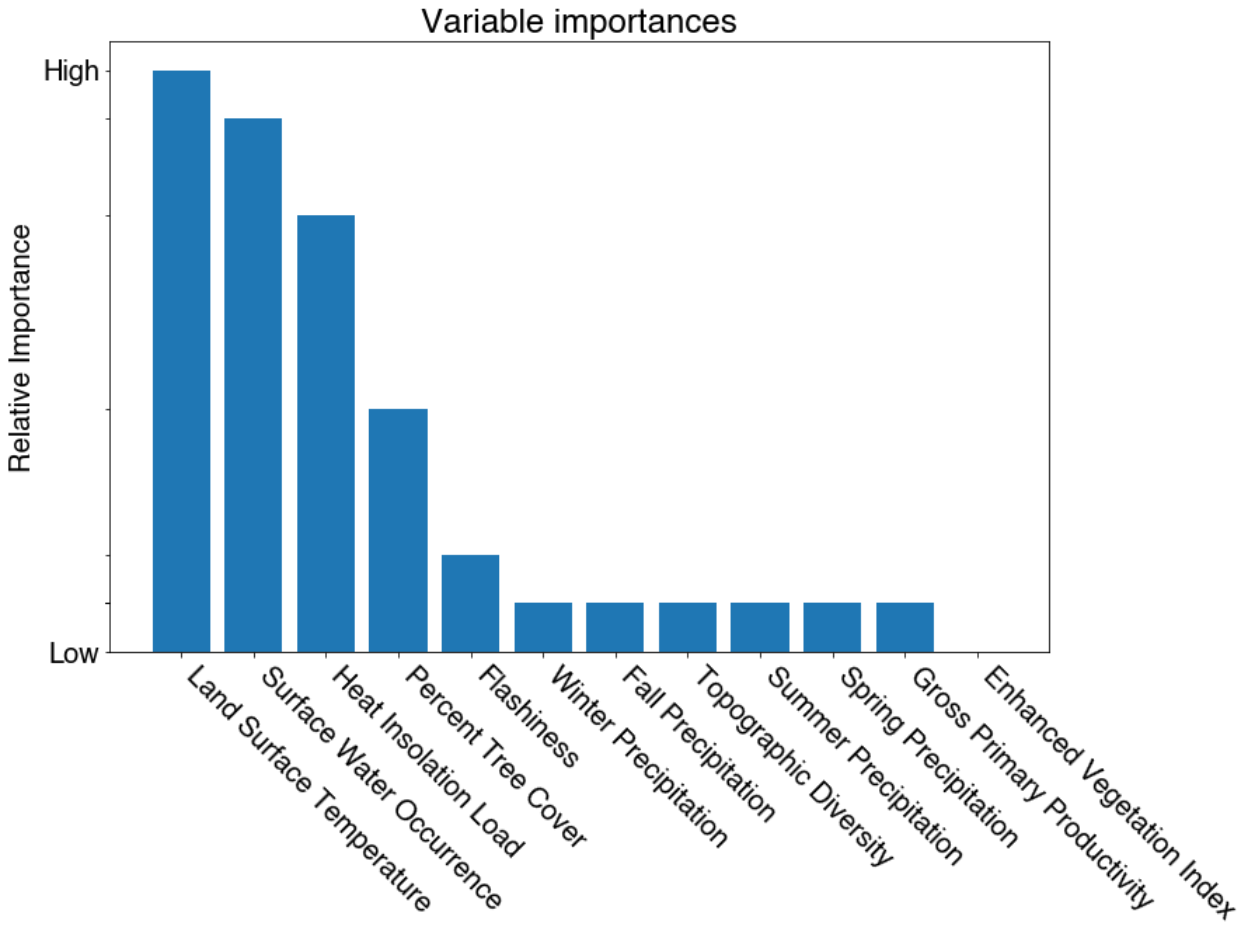

Figure 4. Top predictors of RBT occurrence as identified by the occurrence model.

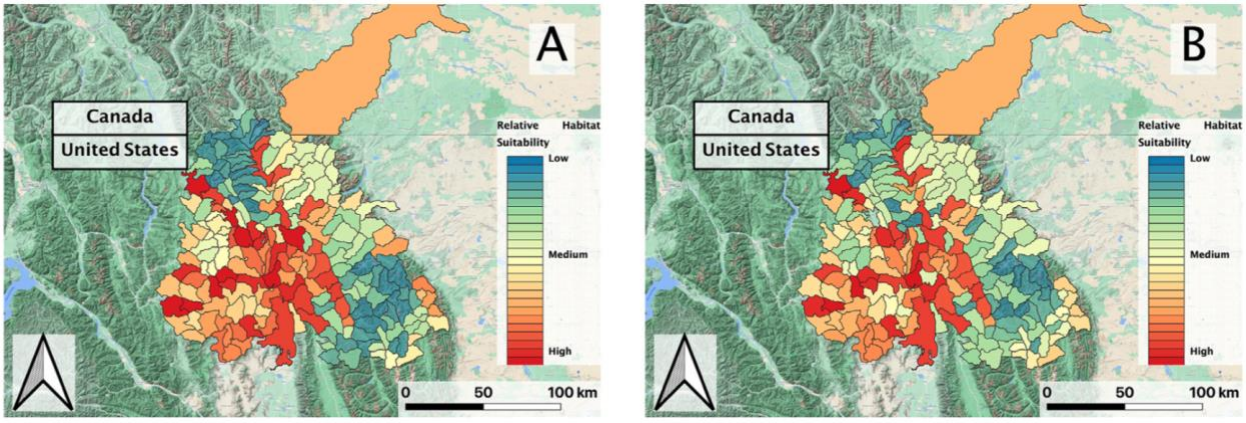

Figure 5. **(A)** Predicted RBT relative suitability of first decade (2002-2010,) and **(B)** second decade (2010-2018) vector composites within the Flathead basin study region; black lines delineate individual HUCs within the larger basin.

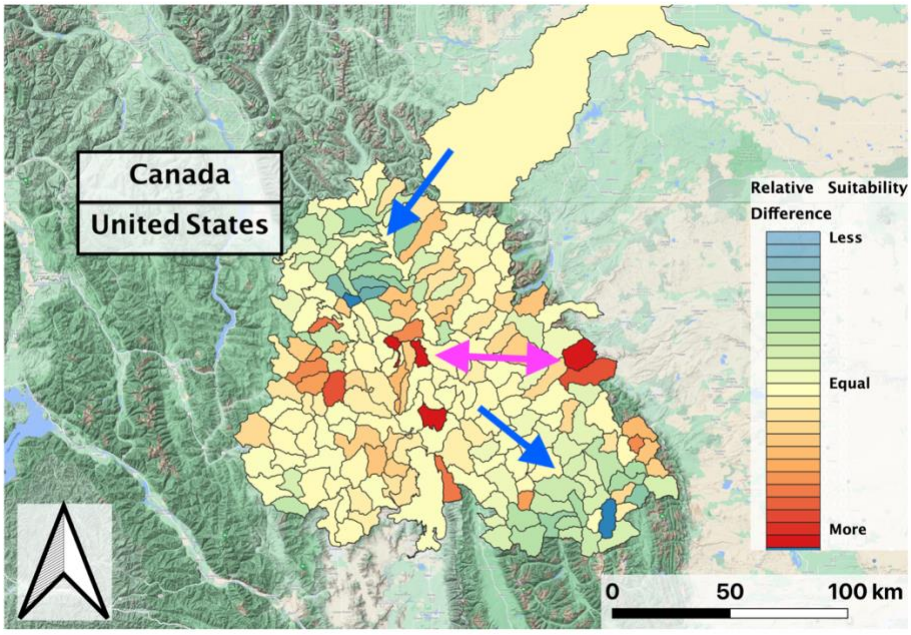

Figure 6, Normalized predicted relative RBT suitability change between the second and first decades of the study period (2002-2018) within the Flathead basin. The Salish Mountains and Lewis Range sub-regions decreased in suitability (blue-green shades; blue arrow), while suitability marginally increased in other regions and increased more drastically in portions of the northern Mission mountains and East Glacier Park regions (red shades; red arrow).

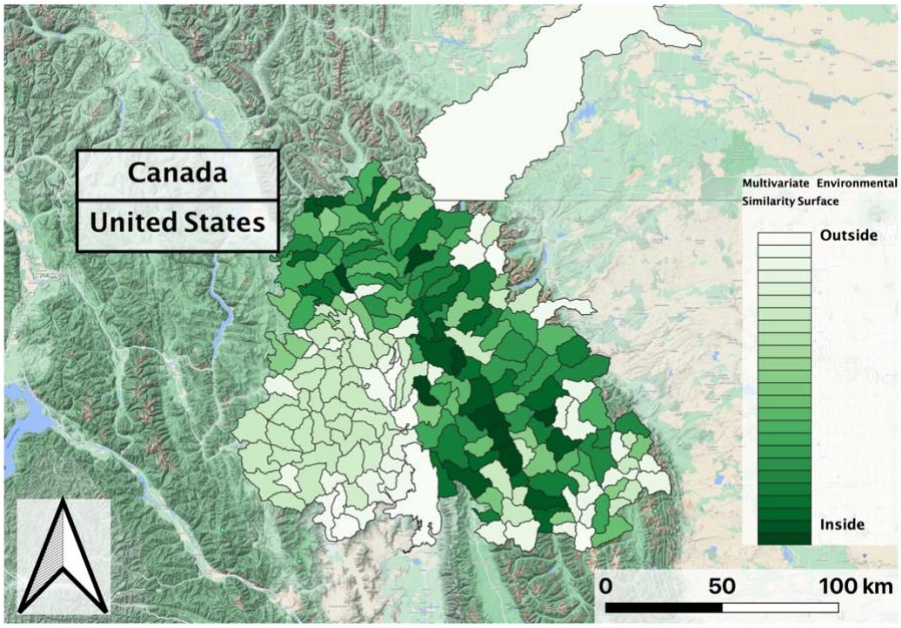

Figure 7. Multivariate Environmental Similarity Surface in the Flathead basin for the second decade (2010-2018) vector composite, which was consistent with the first decade (2002-2010) composite. Greener shades in the similarity surface indicate that most HUCs fall within a reasonable extrapolation distance from RBT training locations.

Partial Dependency Plots (PDP) for the RBT occurrence and hybridization models revealed differing model performances relative to the top predictors, although the PDPs for the RBT occurrence model are more reliable because this model revealed more realistic spatial patterns of habitat suitability (Figure 3 vs 5). For example, the occurrence PDP for flashiness predicted the highest suitability relative to (unitless) flashiness values of 3, whereas the hybridization PDP for flashiness predicted the highest hybridization levels at 7 (Figure 8). The PDPs for both Land Surface Temperature and Surface Water Occurrence showed similar performance between models, and both models showed increasing suitability at temperatures below 34°C. Although both ensemble models identified Heat Insolation Load as a top predictor, the shape of this PDP differed substantially for both models (Figure 9).

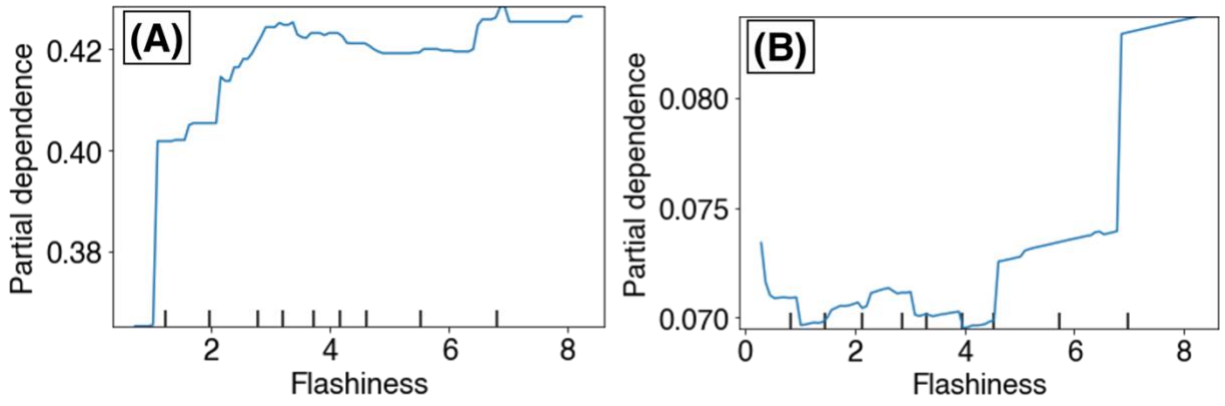

Figure 8. Partial dependency plots for surface water flashiness in both the RBT occurrence ensemble **(A)** and the hybridization ensemble **(B)** models.

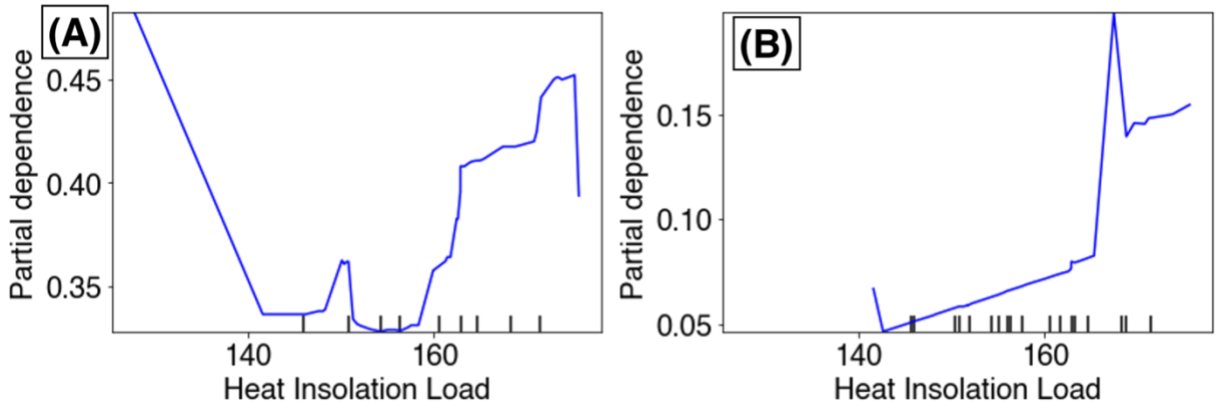

Figure 9. Partial dependency plots for Heat Insolation Load in both the RBT occurrence ensemble **(A)** and the hybridization ensemble **(B)** models.

2.4 DISCUSSION

We present a streamlined workflow that can be used for identifying top predictors of species occurrence and evaluating areas of high risk for invasion and establishment of IS in freshwater ecosystems. This case study allowed us to identify strengths, pitfalls, and opportunities for refinement of this workflow. We attained high cross-validation accuracy and identified key environmental predictors. Model performance relative to the top predictors reinforced known assumptions about RBT distributional requirements in the case of the occurrence model.

We place the utility of this methodology squarely in the realm of prediction-first objectives, to be used in tandem with other management tools. Our methodology provides pivotal advancement towards integrating research insights between managers, stakeholders, and decision-makers, a crucial step towards proactive IS management (Reaser et al., 2020b). The effectiveness and efficiency of this data-driven approach not only permit managers to objectively prioritize "high-risk pathways" (Pyšek et al., 2020), but also enable frequent sharing of maps created from rapidly mobilized occurrence data (Groom et al., 2019). These advantages allow for weighing the costs and benefits of potential management actions at intervals and time scales relevant to managers. As species occurrence data and temporally dynamic environmental information are received, they can be readily mobilized into actionable products using methodologies similar to the current study.

The lack of spatial continuity of RBT hybridization predictions suggests that our workflow was unable to accurately model this process in part due to a non-random field sampling effort. Understandably, sampling protocols prioritized streams where there was concern that RBT were hybridizing with native WCT, resulting in an overrepresentation of recent hybrids which may have skewed the distribution of hybridization training data or at least underrepresented hybridization values in the 40-70% range. It remains unclear whether the unreliable model performance was due to the weaknesses of the training information or the difficulty in representing this process from remotely sensed data products. Indeed, modeling hybridization may not be possible without incorporating a clear dispersal mechanism in the model. In fact, RBT hybridization appears to be driven more by propagule pressure than environmental

conditions (Muhlfeld et al., 2017). Thus, results of the hybridization model must be interpreted cautiously — unless stated explicitly, the remainder of this discussion addresses the RBT occurrence model.

Correlative approaches to evaluating relative habitat suitability are well suited to the EDRR framework, although the tree-based models (both hybridization and occurrence) performed relatively well without additional tuning steps and could be better suited to EDRR. Reaser et al. (2020a) define EDRR as a "guiding principle for minimizing the effects of IS in an expedited, yet effective and cost-efficient manner". Here, we demonstrate that readily-available data products and empirical machine learning models can facilitate these foundational principles and specifically address the *target analysis* portion of the EDRR paradigm. Due to their flexibility and swiftness without the need of tuning procedures, tree-based ML models are especially suited to this stage, which is characterized by intensive surveys and proactive biosurveillance to detect the presence of IS with limited resources (Ricciardi et al., 2017). This spatial prioritization tool is critical during the early stages of invasion (Carlson et al., 2019), and managers using our workflow could prioritize high suitability areas to maximize the effectiveness and cost-efficiency of field efforts. For example, our occurrence model predicts high RBT suitability in the North Fork of the Flathead River and therefore suggests that monitoring efforts could be focused in that region. In addition, identifying top environmental drivers of RBT occurrence allows for more robust assessments of shifting conditions as observational data products are updated and released.

The fact that LST was still identified as a top predictor in both the hybridization and occurrence models suggests that temperature is an important driver of RBT distributions in this region. In addition, our connectivity metric (Surface Water Occurrence) was identified as another top predictor in the case of the more robust RBT occurrence model. However, the steep topography and dense riparian vegetation of stream ecosystems create a challenge for interpretation. For example, the global surface water extent algorithm does not include water bodies of less than 30 x 30m, is known to underestimate water occurrence under emergent vegetation, and resolves the effects of terrain shadows via slopes derived from a 30m DEM (Pekel et al., 2016). Indeed, the diverse vegetation communities and structural heterogeneity of aquatic systems biases the detection capability of this product towards open areas and larger stream orders. Similarly, although the LST product has been linked to stream temperature at the basin or reach level, the connection is less clear in smaller streams, particularly in those with mixed inputs (McNyset et al., 2015). Aggregating at a HUC scale mitigates some adverse effects but does not preclude all issues of scale mismatch. Still, given the above caveats, a cautious interpretation of model performance against such predictors is insightful.

Specifically, the sign and magnitude of PDPs (i.e. Partial Dependency Plots) relative to proximal predictors of known niche requirements of RBT can be interrogated for realism. For example, the occurrence model predicts increasing relative suitability with increasing LST. Previous research has revealed that LST and stream temperature follow a linear relationship at roughly a 3:1 slope in the Columbia River Basin (McNyset et al., 2015). After adjusting for this relationship, the occurrence model predicts increasing suitability at our highest observed stream temperature of 13°C, and Wenger et al. (2011) found that RBT have optimal temperatures at

16°C (Figure 10). However, not all PDPs showed realistic model performance. For example, the PDP for GPP showed an unrealistic dip at 250 kg C / m^2 / 16-days (Figure 11).

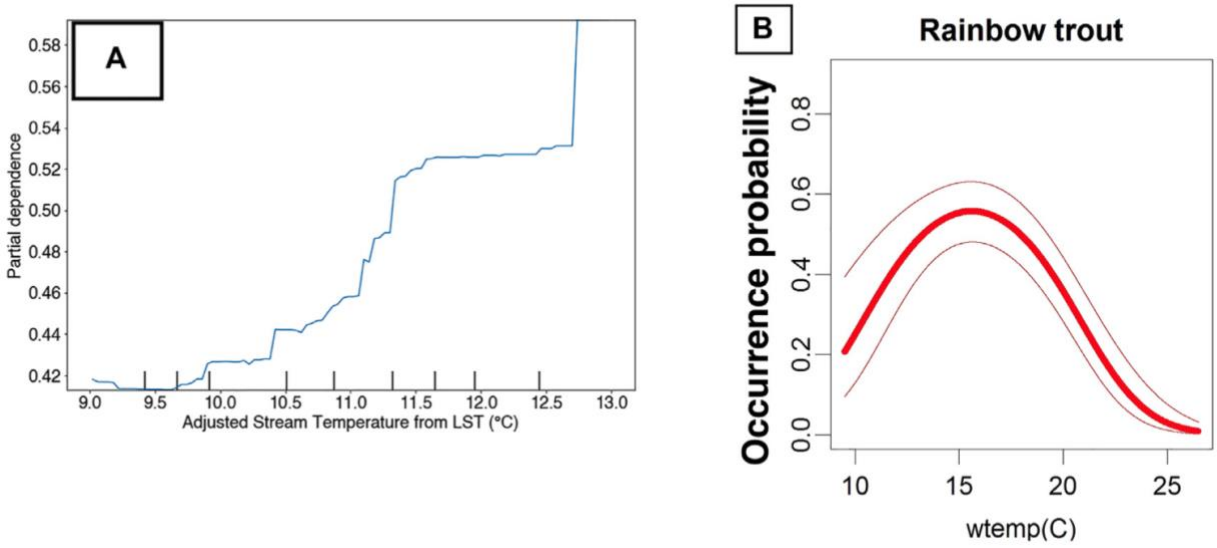

Figure 10. Partial dependency plot showing RBT occurrence model performance against stream-temperature adjusted Land Surface temperature in the Flathead basin **(A)** versus predicted water temperature (wtemp) niche requirements of RBT **(B)** from Wenger et al. (2011).

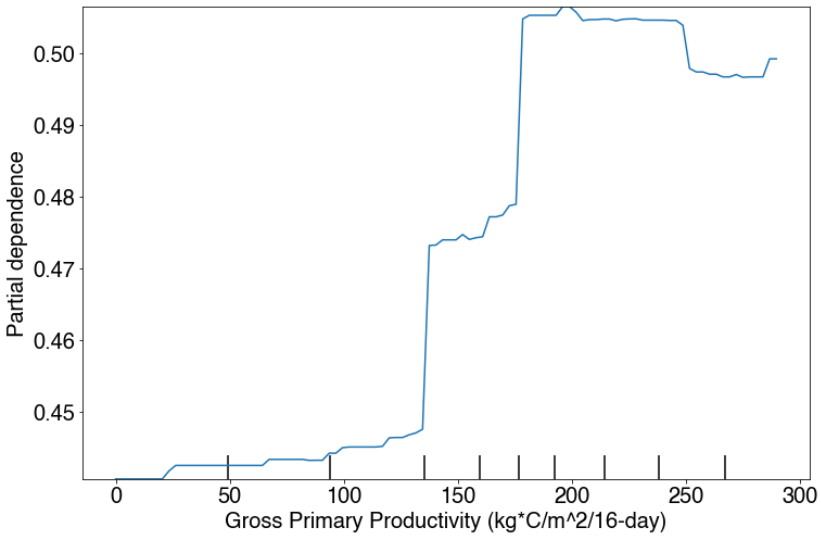

Figure 11. Partial dependency plot showing RBT occurrence model performance against Gross Primary Productivity in the Flathead basin study region.

Interrogating relatively low-importance model predictors can also be valuable. There were a few such products whose lack of explanatory power can be attributed to temporal lag effects,

scale mismatch, or model uncertainty. For example, EVI has been used as a proxy for submerged aquatic vegetation in open water systems (Massicotte et al., 2015), although the connection to species richness in streams is less clear (Vieira et al., 2015). Thus, EVI may not translate to ecologically relevant conditions for RBT within the spatial and temporal scale of our study. Similarly, a terrestrial GPP metric was the most important variable in predicting global-scale species richness of freshwater fish (Pelayo-Villamil et al., 2015) and is correlated with fish production in lakes (Downing et al., 1990). However, our analysis did not reveal GPP as an important predictor for RBT.

Given that GPP represents terrestrial carbon available to primary producers (Robinson et al., 2018) and provides the basis for energy flows supporting aquatic food webs (Welti et al., 2017), it may not drive the higher-level trophic response of stream vertebrates until after a lagging period. In addition, the NLDAS seasonal precipitation metrics did not show up as top predictors, even though RBT are known to be sensitive to peak flow events (Fausch et al., 2001). One possible explanation is the geographic bias present in such spatially interpolated climatic data. Indeed, an examination of the weather stations used in the NLDAS product reveals that geographic coverage of the regional weather station network may be too sparse to fully represent the climate distribution imposed from relatively complex terrain and orographic effects in the Pacific Northwest (Mo et al., 2012). Thus, we recommend the use of landscape scale RS products because of their spatial contiguity. Lastly, although the seasonal additive aggregate model inputs (i.e. Spring Total Precipitation, Summer Total Precipitation) may have captured the magnitude of peak flow events, these aggregates did not inform the timing and duration of flow.

More work is needed to integrate the temporal variability of dynamic data products into our workflow.

Our workflow compromises interpretability for speed, accuracy, and efficiency. Top predictors are correlative at best, and without explicitly modeling the dispersal potential of these organisms, our model predicts relative habitat suitability alone. In addition, using temporally composited covariates results in a loss of information relating to the timing and duration of environmental conditions. However, such improvements would compromise the speed and agility strengths of this workflow. As the rate of new biological invasions shows no sign of slowing (Seebens et al., 2017), early detection and rapid response is becoming more vital to prevent irreversible ecological damage and massive economic costs to societies. New technological integrations are needed to facilitate aquatic IS detection and promote proactive management. We present and test one such generalizable workflow for integrating occurrence information with readily available data products to generate spatiotemporally explicit habitat suitability (i.e. risk) maps. While this application case study was for RBT, the underlying models and workflow can be readily extended to other aquatic and terrestrial species.

Given further testing and validation, this workflow could be expanded in its geographic and taxonomic breadth by exploiting web-hosted databases of species occurrence data (e.g. GBIF, www.gbif.org; USGS NAS, http://nas.er.usgs.gov). Future considerations include accounting for sampling bias, integrating presence-only rather than presence-absence datasets, and working toward fully automating the data acquisition and preprocessing steps. The advancement of data sharing capabilities in ecological sciences, born out of the field's recent

rebirth as a big-data science, has enabled robust methodologies and automated pipelines that can produce actionable insight based on continuous occurrence and environmental data streams. Leveraging workflows such as this provide a major step in the way of integrating these data with management action at broad spatial and ecological scales.

2.5 ACKNOWLEDGEMENTS

This research was funded by NASA (80NSSC19K0185, NNH18ZDA001N) and a USGS Northwest Climate Adaptation Science Center award (G17AC000218). USGS researcher and co-author Clint Muhlfeld and Montana Fish Wildlife and Parks researcher Ryan Kovach provided the RBT admixture data used in the models. We thank two anonymous reviewers and the handling editor, Dr. Bin Peng, for insightful comments that helped improve and clarify this manuscript.

Chapter 3

3 TOOLKIT AND MAGEMENT FRAMEWORK

3.1 BACKGROUND

Despite the enthusiasm for, and pervasive use of, SDMs in a conservation context (e.g. Elith et al., 2010), there are difficulties associated with their integration in structured IS management frameworks such as EDRR. In the previous chapter, I suggest that my workflow can be used to inform the target analysis action of EDRR but incorporating model output into management decision-making remains challenging. Target analysis is vital to the EDRR framework, and I define it as a strategic approach for detecting IS using predictive technologies (Morisette et al., 2021). The requirements of spatial target analysis cannot be met without technical input that proactively illustrates potential establishment areas (Morisette et al., 2021), yet high requirements for technical skill, computing resources, and modeling tools makes this difficult (Guisan et al., 2013). Indeed, given the urgent and timely demands of IS target analysis, the development of SDMs for this purpose requires considerable technical knowledge that is a major implementation barrier to end users and managers (Addison et al., 2013; Guisan et al., 2013).

The "knowing-doing" gap, created by the mismatch between technical expertise and decision-making power, can be mitigated by a combination of simple models and heuristic approaches that prioritize expediency (Pyšek et al., 2020; Sutherland and Freckleton 2012). I

contributed to a recent review that encouraged the coordinated use and uptake of predictive technologies for proactive IS management and research (van Rees et al., 2022). This framework identified the transfer of research insights from researchers to IS managers as a crucial final step in the analytical pipeline for proactive IS management. In particular, it recommended the use and development of user-friendly products to decrease the methodological expertise necessary to guide conservation actions and close the implementation gap between analytical approaches and on-the-ground results. Although these products are essential to the EDRR paradigm guiding such actions (Russel et al., 2017; Berec et al., 2015; Wang et al., 2014), the effective implementation of such tools is sparse.

To address this gap, I lead the development of the Chapter 2 workflow in two online tool prototypes. I then provide management guidance using tool outputs that considers the technical details of the workflow, model outputs, and relevant demands of the EDRR paradigm. This information provides a basis from which to answer Research Question 2, and a more targeted response is given in section 4.1.2.

3.2 DEVELOPMENT OF TOOLKIT

The development of this toolkit blended the essential specifications of EDRR with technical constraints. By allowing users to directly interface with the data, the toolkit addresses a major need to make analytical results be relevant and communicable (van Rees et al., 2022). The information necessary for coordinated management action can broadly be divided into two classes: rapid and cursory risk assessments on the one hand and thorough examinations of environmental drivers and modeled responses to environmental change on the other. To reflect these differing demands, I split the toolbox into a front-facing Graphical User Interface (GUI)

and a lower-level, open-source Command Line Interface (CLI). The GUI was built using Google Earth Engine's User Interface API that allows users to construct dynamic user interfaces with the same functionality that can be found in the JavaScript code editor. Thus, the GUI uses JavaScript as its programming language. On the other hand, I constructed the CLI using Python 3.7.7 with various open-source libraries including Sklearn (Pedregosa et al., 2011), Pandas (McKinney et al., 2010), and Google Earth Engine's Python API (Gorelick et al., 2017). While the GUI can be accessed with any web browser, the CLI must be downloaded and installed using a package manager such as Anaconda (Anaconda 2020).

During the development of the GUI, I needed to compromise some features of the Chapter 2 workflow due to technical constraints. This decision was made to address the rapid and timely demands of IS management. For the most part, the workflow stays the same, but because the Google Earth Engine platform lacks the implementation of classifiers other than random forest, I decided to remove the ensemble protocol from this tool. This decision was made with the knowledge that the random forest model performed relatively well in the Chapter 2 case study and in a multitude of other modeling efforts (Norberg et al., 2019). As such, predictions are slightly less robust, but all outputs are similar, and it produces maps in a much more streamlined manner without the need of client-side resources. Indeed, it is still able to predict the relative habitat suitability of a given area (given by the probability estimate provided by the model) and estimate the uncertainty of these predictions using the variation of the vector of predictions given by the set of all estimators within the model.

By design, the GUI and CLI complement one another's respective advantages and shortcomings, allowing for flexibility in the implementation experience. For example, by compromising speed for control, the GUI does not allow for an examination of the top environmental drivers of occurrence, nor does it include the ability to automate multi-species analyses. In contrast, the technical interface of the CLI allows for more detailed analyses that can be used for deriving biological insight of invaders to inform future management. Each tool in the toolbox reflects different management needs and fits into different parts of the EDRR framework.

3.2.1 DESCRIPTION OF TOOL FEATURES

This toolbox provides the technical input necessary for target analysis of aquatic invasive species. It requires that the spatially explicit species observations be coded in a presence-absence binary format and that such observations include a year of observation. With this input information, the toolbox can be used for forecasting of aquatic species distributions, assessing the uncertainty associated with such forecasts, and examining species-environment relationships.

The toolbox currently houses two different tools. Each tool in the toolbox consists of a stack of various software packages, with some subtle differences. Both tools leverage the cloud computing power of Google Earth Engine in order to assimilate and process environmental information. From there, the GUI continues to use Google Earth Engine to learn, predict, and visualize the relative habitat suitability of a given area. On the other hand, the CLI pulls the training data onto the client's computer and uses a stack of open-source Python libraries that allow the user more control over the modeling process. Thus, the major benefit of the GUI is an

easy-to-use visualization tool that can be used to document and share results from a rapid and generalizable habitat suitability model, and the major benefit of the CLI is the detailed and complex modeling of habitat suitability that can be used to examine species-environment relationships and drivers of species occurrence.

The GUI tool can be found at this Google Earth Engine link (https://mstokowski.users.earthengine.app/view/aismodel). This tool generates two major outputs: a habitat suitability assessment and an estimate of prediction uncertainty over the domain of interest (Figure 12). These two outputs represent major components of the technical input for target analysis, and in the following section I provide a brief overview of how they can be incorporated into a decision-making process.

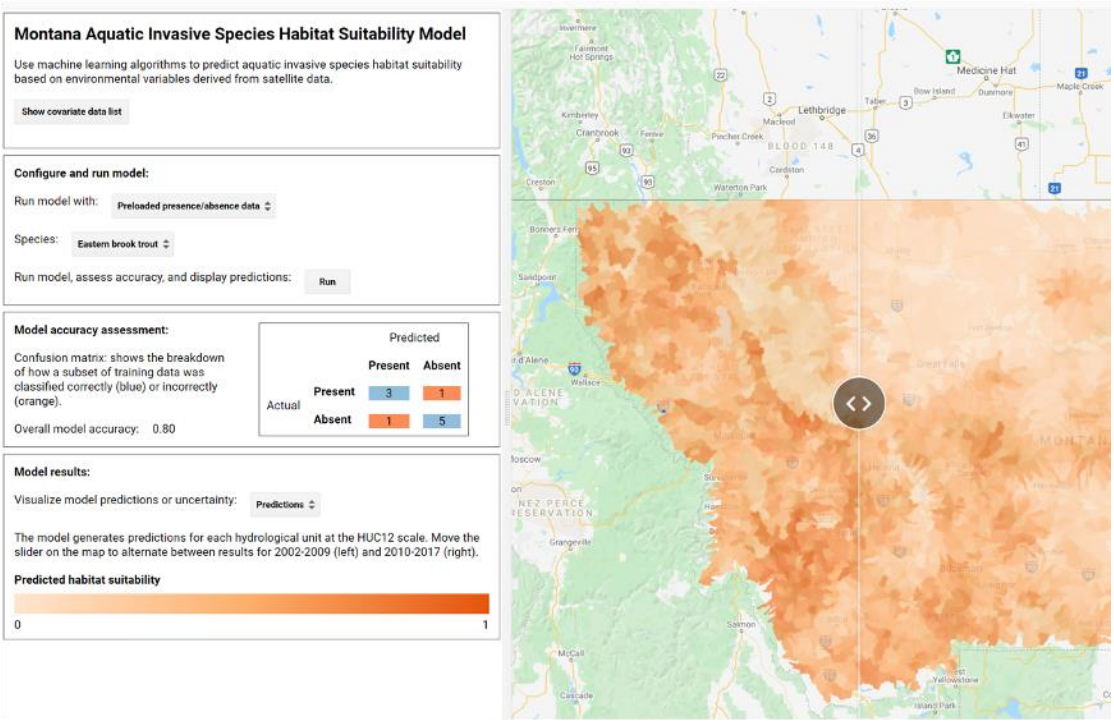

Figure 12. Illustration of GUI that can be found at this link: https://mstokowski.users.earthengine.app/view/aismodel

The CLI can be downloaded and installed at this github link: (https://github.com/COYE-Coder/AIS; Figure 13). Because it is a complete implementation of the Chapter 2 workflow, outputs consist of relative rankings of environmental covariates and species responses to shifting environmental conditions. In the github repository, users will find an introduction, documentation, and guidance for running the tool. It is built using Python 3.7.7 and uses a combination of various open-source software packages such as Pandas and Sklearn. Because of the lower-level nature of this tool, it has advantages over the GUI, although it does require a larger degree of technical skill. For example, it allows for users to modify various hyperparameters such as the number of tallies required for a predictor to be given high importance in the assessment of top drivers of occurrence (see section 2.2.7). In addition, it can be used to automate batch processes for multiple species or different areas of concern. Because of its capacity to examine species-environment relationships, the CLI is intended to be used as the technical input for the risk screening portion of EDRR (Reaser et al., 2020).

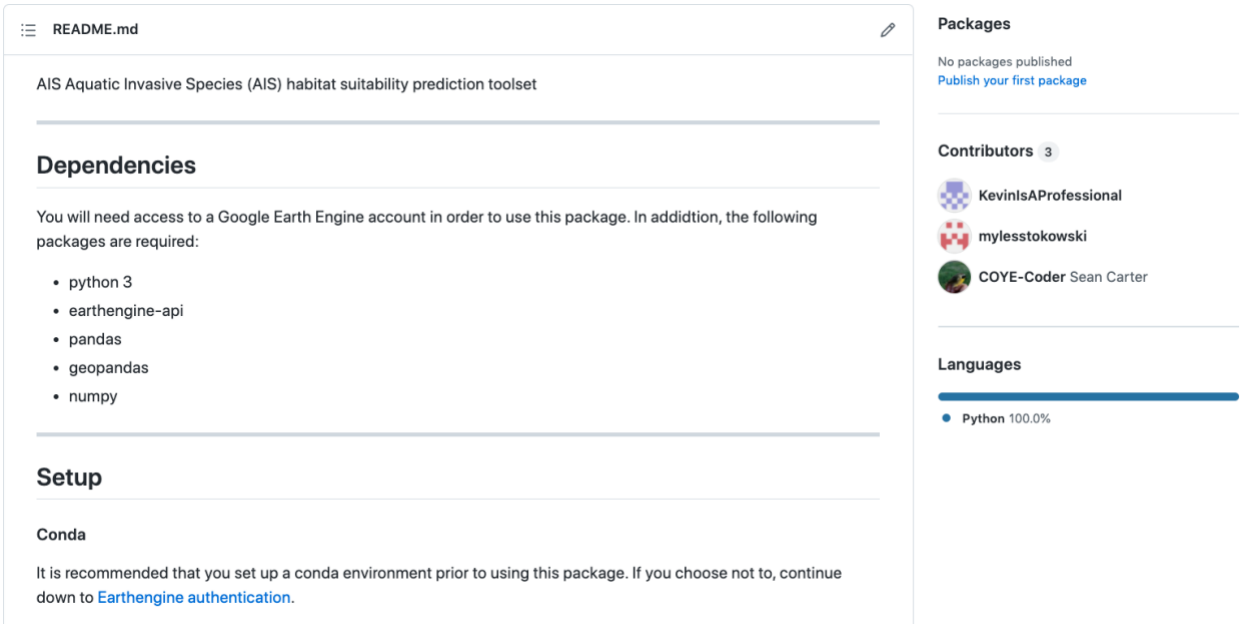

Figure 13. Illustration of the CLI that can be found at this link: https://github.com/COYE-Coder/AIS

Although these tools are readily available and relatively easy to use, incorporating tool output into management decision-making remains challenging. In the following section, I provide general guidance for IS target analysis that incorporates the technical input provided by the GUI described above. This framework provides guidance for users to incorporate and synthesize the component features of the workflow and toolkit to derive actionable insight in an efficient manner.

3.3 GUIDANCE FOR INVASIVE SPECIES TARGET ANALYSIS USING THE TOOLKIT

In order to maximize survey effectiveness and cost efficiency, target analysis requires 1) identifying priority areas for monitoring or intervention by examining mechanistic dispersal constraints and habitat requirements, 2) considering the uncertainty associated with such predictions, 3) developing robust and efficient sampling efforts, and 4) conducting thorough survey efforts whose component field observations may be used to validate and recursively improve initial modeling inputs over iterative cycles (Morisette et al., 2020). Each of the above steps can incorporate either technical or heuristic inputs and has many potential considerations (Figure 14). This section describes general guidance from which to view these considerations, providing an avenue through which the online GUI can be used to address them. In doing so, I address both the implementation barrier of SDMs into structured decision-making frameworks and Research Question 2.

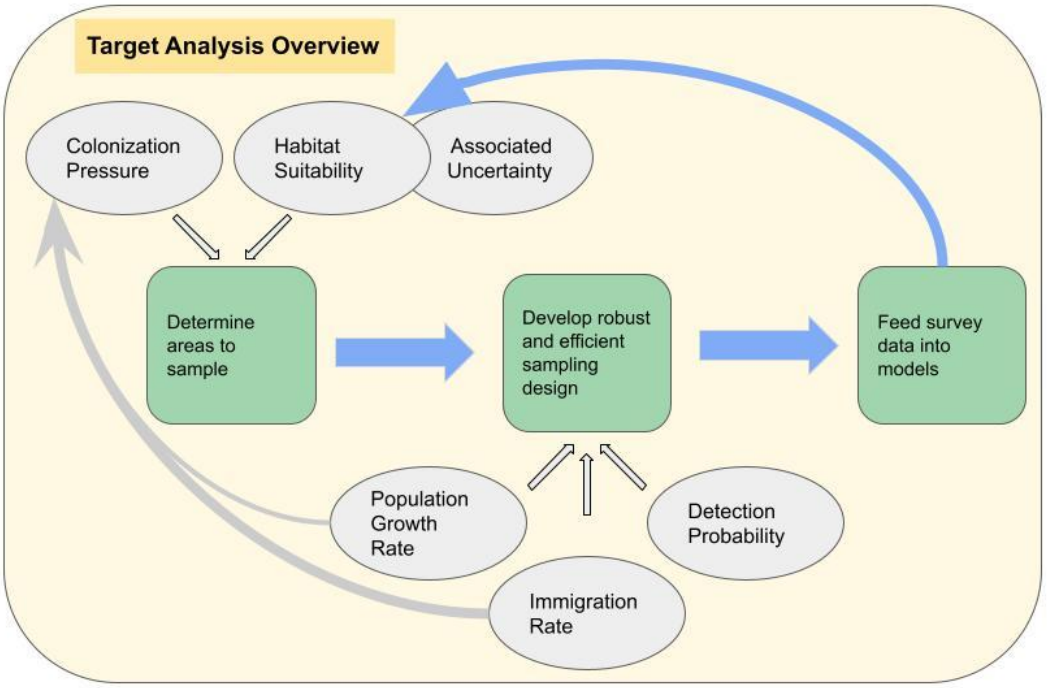

Figure 14. A general workflow for target analysis. Rectangles indicate major steps and ovals indicate technical inputs

Identifying priority locations for monitoring must integrate the colonization pressure (Lockwood et al., 2009), the niche requirements of a given species, and the uncertainty associated with such input information. This vital step makes the most valuable contributions to management when undertaken before appropriate sampling efforts are planned and initialized. Including colonization pressure into decision making must balance expertise, speed, and information availability. Colonization (propagule) pressure can be assessed in a number of different ways (Lockwood et al., 2005) ranging from complex agent-based models to qualitative composites of propagule size and frequency that incorporate heuristic rules approximating the dispersal behavior of a given species and system. In this case, propagule size refers to the number of individuals within a potential dispersal source, and propagule frequency refers to the relative rate of introduction events from a given source population. Agent-based modeling

requires, among other considerations, quantifying the dispersal tendencies of each potential source population of the IS of concern, quantifying the difficulty of dispersal, and simulating potential establishment scenarios given some adaptive feedback process (Macal and North 2005). Because these considerations are typically unavailable to managers either due to the lack of requisite expertise of input data, the complexity of such models is prohibitively time consuming.

Alternatively, it is possible to qualitatively assess the colonization pressure based on first principles and heuristic approximations of animal movement from natural history knowledge. For example, the hypothetical example shown in Figure 15 shows four different scenarios of similar colonization pressure. By considering first approximations of propagule size and frequency, managers can rank different areas from high to low colonization pressure based on their heuristic knowledge of the study system. This can be coupled with natural history knowledge of the population growth and immigration rates. Thus, regarding colonization pressure, the rapid implementation demands and short time scales of EDRR are better addressed using qualitative composites that are generated from first principles than complex agent-based modeling efforts.

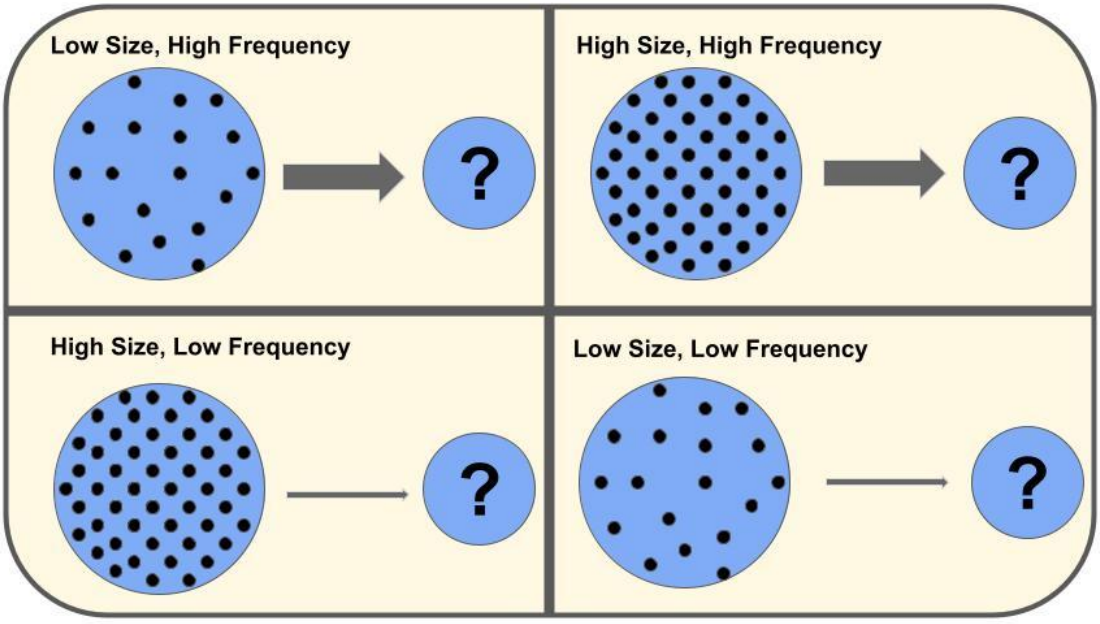

Figure 15. Four different scenarios that yield various levels of colonization pressure to an area that is being considered for sampling. The top two boxes display similar colonization pressure, which is different than the bottom two boxes. By incorporating the two fundamental drivers of colonization pressure, propagule size and propagule frequency, managers can rank the relative urgency of conducting sampling efforts in different areas.

On the other hand, predicting the habitat suitability for an IS and the uncertainty associated with such modeling predictions can be accomplished in a rapid and effective manner using correlative modeling techniques. User-friendly tools such as the ones detailed in section 3.2 can provide technical input that, with consideration of colonization pressure described above, delineate areas that will be most beneficial to sample. By balancing the urgency described by the colonization pressure, the relative habitat suitability, and the uncertainty associated with predicting habitat suitability, managers can trade-off the value of sampling under-represented areas and prioritizing areas of high projected risk with the ultimate goal of either improving model generalizability or detecting organisms before they become too far established to be managed effectively. For example, using this tool, managers might consider that areas which are

likely highly suitable should garner more attention, granted that they are in areas of high colonization pressure and moderate to high model confidence (Figure 16).

Evaluating the appropriate compromise in the above situation will differ with the context of each individual management case and the needs of stakeholders and decision-makers (Figure 16). For instance, sampling areas that are predicted to be highly suitable that are well-within possible dispersal areas could either redirect the model, if it was wrong, or allow for a deeper confidence in predictions to other areas. Similarly, although counter-intuitive, conducting a sampling effort in areas that are projected to be low risk may be fruitful if these areas are also associated with high uncertainty by providing validation for future modeling projections. Thus, gathering input information in such areas would improve model generalizability by increasing the size of the environmental envelope captured by the model. Considering this option would likely be a result of having determined that there were no areas with crucial danger levels. The first step to a diligent target analysis effort requires the interactive consideration of three different technical inputs, two of which are provided by the tools described in section 3.2, and the third can either be developed qualitatively or quantitatively.

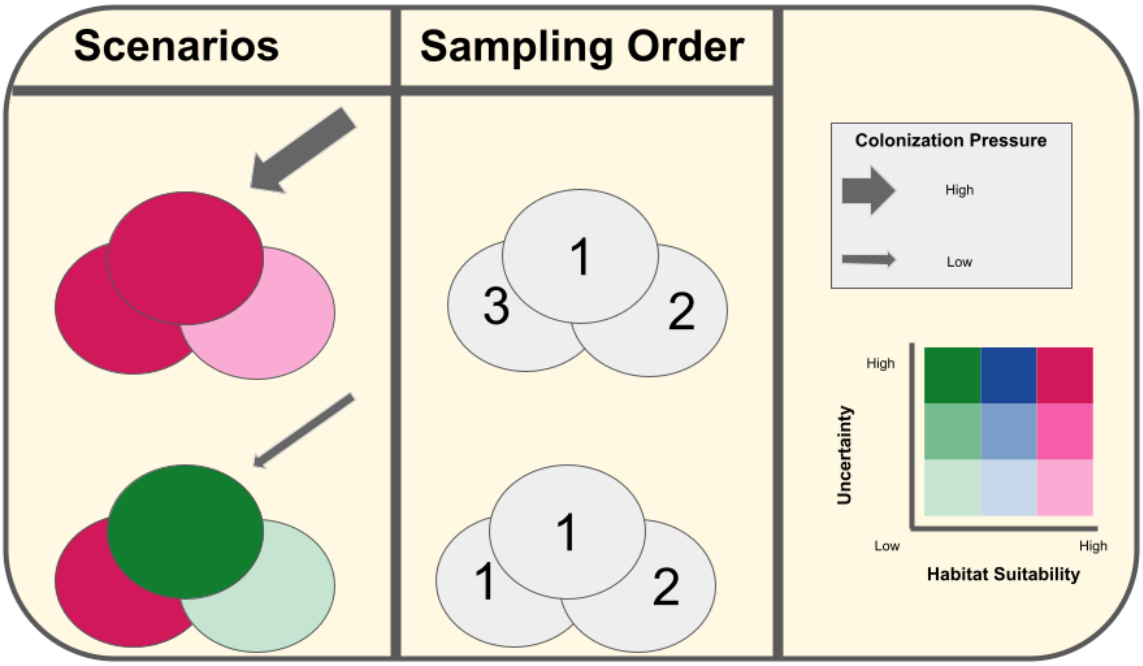

Figure 16. Synthesizing three different considerations for determining the sampling order for various areas. On the left are two different scenarios based on differing model output and estimated colonization pressure. Model output is shown in color scheme. In the first row, a high colonization pressure causes the top region to be high priority, even though the model is highly uncertain. The right region is less important due to less colonization pressure, even though model uncertainty is lower. In the second row, small colonization pressure allows the managers to decide about how they might wish to improve future modeling efforts. Due to the lack of urgency, either the top or left regions can be sampled in order to improve model generalizability because the model is highly uncertain in either region. The right region has lower uncertainty, and thus does not take priority over higher uncertainty regions.

Once an area has been identified that is of primary concern or yielding high insight, target analysis then requires the development of a robust sampling effort. This action must consider the relative population growth rate, the probability of detection, and the rate of immigration (Mehta et al., 2007). Simulation modeling suggests that the efficiency of sampling effort is driven more by the density of observation points than their spatial arrangement, but only when survey sensitivity (i.e., detection probability) is high. In addition, high population growth rates lead to the most efficient performance of grid-based survey efforts, and low population growth rates lead to the most efficient performance of random survey efforts (Berec et al., 2015). Lastly, high

colonization pressure increases the efficacy of grid-based survey efforts through the consequent increase of detection probability. Although the consideration of these biophysical parameters must be preeminent, managers can be confident that maximizing the detection probability at small scales is unlikely to diminish the probability of detection at broad scales. Finally, because sampling efforts must be constrained by operational costs, exhaustive design of sampling efforts might weigh the risk of potential damages with bioeconomic pre-screenings (e.g. Kller et al., 2007). However, this work must be made in consideration of the urgent requirements of EDRR. Efficient sampling efforts are contingent upon biophysical parameters and operational costs, and as with any aspect of this workflow, compromises must be made between speed and precision.

A crucial final step in the target analysis action is the reincorporation of ground-based data from survey efforts into the next model iteration. Labeled data remains the most pressing deficiency in machine learning (Sarker 2021). This problem is accentuated within the earth and biological sciences due to the difficulty of acquiring robust and accurate ground-based data. Indeed, after the acquisition of such training information, it can be fed into the next model iteration, where the automated approach and fast development process will facilitate the cyclical nature of target analysis. The potential for quick and rapid incorporation of this new information is a major advantage of the ML workflow described in Chapter 2 and the general guidance given in Chapter 3.

The challenges of IS Target Analysis can be mitigated with technical support and guidance for management decisions. For example, identifying priority areas requires considering the colonization pressure and the habitat suitability with its associated uncertainties. My tool

provides the technical input necessary for two of those considerations: the predicted habitat suitability and its associated uncertainty. From there, Target Analysis requires developing robust and efficient sampling efforts and conducting thorough survey efforts whose component field observations may be used to validate and recursively improve initial modeling inputs over iterative cycles. The guidance in this chapter provides a framework through which these requirements can be viewed. Target Analysis is a vital action of the EDRR paradigm. By creating tools and guidance to cater to the urgent needs of the EDRR paradigm, this chapter serves as a major step towards preventing irreversible damage from IS establishment.

Chapter 4

4 SUMMARY AND CONCLUSIONS

In Chapters 2 and 3 of this thesis, I developed methods and provided avenues to address the guiding research questions asked in section 1.2. These questions steered the overall direction of the thesis and were developed to address science objectives identified by the presiding grant. Indeed, by addressing these questions, I provide the basis for achieving major project goals including 1) providing tools for mapping current distributions of focal IS and 2) promoting use of remote sensing data and data integration with IS among managers and researchers through new workflows and management frameworks.

RQ 1: How can species occurrence information be integrated with remotely sensed and other geospatial imagery to inform invasive species management decisions?

Species occurrence information can be integrated with environmental data layers to represent and project areas of high risk to persistence and reproduction of IS of concern within the habitat suitability framework. This modeling framework can inform management decisions by providing the technical input necessary for the rapid demands of existing management paradigms such as Early Detection and Rapid Response. New workflows, such as the one provided in Chapter 2, confront these rapid demands, presenting a technique to delineate high risk areas that has been validated against a well-known system. Species occurrence information can be integrated with remotely sensed and other geospatial imagery in the following manner: 1)

selecting appropriate environmental data products, 2) compositing and consolidating them to ecologically relevant metrics and modeling units, 3) and projecting habitat suitability using iterable and generalizable algorithms. Integrating these disparate data sources in a ML habitat suitability workflow provides actionable management insight, as has been demonstrated in Chapter 2.

RQ 2: How can workflows linking existing databases with modeling technologies facilitate more efficient, effective spatial prioritization of IS monitoring and intervention while augmenting existing management frameworks?

New workflows can augment existing management frameworks to facilitate effective and efficient target analysis by providing the technical input necessary for determining areas of primary concern. The guidance provided in Chapter 3 suggests that the consideration of the relative habitat suitability, the uncertainty of that prediction, and the propagule pressure of a given area are integral portions of the target analysis action within the EDRR framework. The tools that I developed serve as a foundation to design and implement more methodical target analysis used for IS management and intervention. The guidance provided in Chapter 3 facilitates efficient, effective spatial prioritization of IS by synthesizing technical input and management concerns while also augmenting the EDRR paradigm.

This thesis addresses the need for spatial prioritization of proactive measures for IS management via providing both a technical framework and management guidance for evaluating areas of high concern. The rapid demands of IS management, coupled with the large computational costs to developing technical input, create a large implementation hurdle, which

this thesis partially addresses. Correlative machine learning approaches to evaluating relative habitat suitability of focal IS are well-suited to these rapid demands of IS managers due to their flexibility and swiftness. Workflows, such as the one developed in Chapter 2, can be used as technical input for management decision-making actions, such as Target Analysis, as described in Chapter 3. The case study presented in Chapter 2 provides evidence that a rapid workflow can be used to derive both biological and management insights on the drivers of IS occurrence and areas of high priority for field efforts. This workflow can facilitate efficient and proactive strategies to avoid irreversible effects of IS establishment and spread. The guidelines for one such management strategy are provided in Chapter 3. Target analysis, as with any structured decision-making workflow, requires the compromise between different pressing needs. Although contributing to only a part of a comprehensive decision process, the guidance provided in Chapter 3 can assist in overcoming implementation barriers by structuring the problem, providing suggested considerations at each stage within the management action, and supplying examples that can be applied to different scenarios. The information provided in this chapter fills a gap in the lack of structured and sequential guidance for target analysis in invasive species management.

This thesis addresses a major challenge in IS management but is not free from limitations. For example, a major driver of IS spread is propagule pressure, and there remains a huge challenge to integrate dispersal information into rapid and effective modeling efforts. In addition, the constraints inherent to the environmental information are significant. For example, the spatiotemporal resolution of some data products constrains the resolution of prediction maps. Furthermore, compositing temporally explicit products results in the loss of information relating to the timing and duration of dramatic events. Lastly, there remains a pressing need to generalize

the Chapter 2 workflow to presence-only datasets and multiple taxonomic groups simultaneously. Steps that can be considered to address these limitations include quantifying the uncertainty associated with the model through error propagation (Andriew et al., 2003), incorporating the seasonal variation of some data products (Schneider 2012), and accounting for spatially biased sampling effort of the species occurrence data (Stolar and Nielsen 2015).

In addition to improvements that can be made on the above limitations, future research can focus on specific technical improvements. For example, with the qualitative framework proposed in Chapter 3 in place, future actions could feed these technical inputs into a multiple criteria evaluation by quantifying the urgency associated with each technical input (e.g., the relative habitat suitability, the uncertainty, the colonization pressure), rasterizing these inputs, and providing weights associated with each raster layer. This quantification framework could be further improved by feeding all management and technical considerations into a cost of outcome analysis (e.g., Blaalid et al., 2021). In addition, the GUI and CLI tools can be improved in their speed and usability. Still, this thesis addresses a major need to enhance decision-making by developing a support system to facilitate monitoring and forecasting of the spread of aquatic IS. In doing so, it allows for users to mitigate and prevent drastic damages associated with IS establishment.

REFERENCES

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. Int. J. Climatol., 33: 121–131.

Allendorf, F.W., Leary, R.F., Hitt, N.P., Knudsen, K.L., Lundquist, L.L., Spruell, P. (2004). Intercrosses and the U.S. Endangered Species Act: Should Hybridized Populations be Included as Westslope Cutthroat Trout?. Conservation Biology.. doi:10.1111/j.1523-1739.2004.00305.x

Altmann, A., Toloşi, L., Sander, O., Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics.. doi:10.1093/bioinformatics/btq134

Anaconda Software Distribution. (2020). *Anaconda Documentation*. Anaconda Inc. Retrieved from https://docs.anaconda.com/

Andrieu, C., De Freitas, N., Doucet, A., Jordan, M.I., 2003. None. Machine Learning 50, 5–43.. doi:10.1023/a:1020281327116

Bayliss, H. R, Stewart, G. B, Wilcox, A., & Randall, N. P. (2013). A perceived gap between invasive species research and stakeholder priorities. *NeoBiota*, *19*, 67-82. doi: 10.3897/neobiota.19.4897

Bear, E.A., Mcmahon, T.E., Zale, A.V. (2007). Comparative Thermal Requirements of Westslope Cutthroat Trout and Rainbow Trout: Implications for Species Interactions and Development of Thermal Protection Standards. Transactions of the American Fisheries Society.. doi:10.1577/t06-072.1

Bedia, J., Herrera, S., & Gutiérrez, J. M. (2013). Dangers of using global bioclimatic datasets for ecological niche modeling. Limitations for future climate projections. *Global and Planetary Change*, *107*, 1–12. https://doi.org/10.1016/j.gloplacha.2013.04.005

Bellard, C., Cassey, P., Blackburn, T.M. (2016). Alien species as a driver of recent extinctions. Biology Letters.. doi:10.1098/rsbl.2015.0623

Bennett, S.N., Olson, J.R., Kershner, J.L., Corbett, P. (2010). Propagule pressure and stream characteristics influence introgression: cutthroat and rainbow trout in British Columbia. Ecological Applications.. doi:10.1890/08-0441.1

Berthon, K. (2015). How do native species respond to invaders? Mechanistic and trait-based perspectives. Biological Invasions.. doi:10.1007/s10530-015-0874-7

Bhattacharya, M. (2013). Machine Learning for Bioclimatic Modelling. International Journal of Advanced Computer Science and Applications.. doi:10.14569/ijacsa.2013.040201

Blaalid, R., Magnussen, K., Westberg, N.B., Navrud, S., 2021. A benefit-cost analysis framework for prioritization of control programs for well-established invasive alien species. NeoBiota 68, 31–52.. doi:10.3897/neobiota.68.62122

Boyer, M., Muhlfeld, C., & Allendorf, F. (2008). Rainbow trout (Oncorhynchus mykiss) invasion and the spread of hybridization with native westslope cutthroat trout (Oncorhynchus clarkii lewisi). *Canadian Journal of Fisheries and Aquatic Sciences, 65(4), 658-669.*

Bradley, A. P.. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/s0031-3203(96)00142-2

Bradshaw, C.J.A., Leroy, B., Bellard, C., Roiz, D., Albert, C., Fournier, A., Barbet-Massin, M., Salles, J.-M., Simard, F., Courchamp, F. (2016). Massive yet grossly underestimated global costs of invasive insects. Nature Communications.. doi:10.1038/ncomms12986

Breiman, L. (2001). Random Forests. Machine Learning.. doi:10.1023/a:1010933404324

Brett, J.R. (1952). Temperature Tolerance in Young Pacific Salmon, Genus Oncorhynchus. Journal of the Fisheries Research Board of Canada.. doi:10.1139/f52-016

Carter, S., van Rees, C. B., Hand, B. K., Muhlfeld, C. C., Luikart, G., & Kimball, J. S. (2021). Testing a Generalizable Machine Learning Workflow for Aquatic Invasive Species on Rainbow Trout (*Oncorhynchus mykiss*) in Northwest Montana. *Frontiers in big data*, *4*, 734990. https://doi.org/10.3389/fdata.2021.734990

Carlson, A. K., Taylor, W. W., Kinnison, M. T., Sullivan, S. M. P., Weber, M. J., Melstrom, R. T., Venturelli, P. A., Wuellner, M. R., Newman, R. M., Hartman, K. J., Zydlewski, G. B., Devries, D. R., Gray, S. M., Infante, D. M., Pegg, M. A., & Harrell, R. M.. (2019). Threats to Freshwater Fisheries in the United States: Perspectives and Investments of State Fisheries Administrators and Agricultural Experiment Station Directors. *Fisheries*, *44*(6), 276–287. https://doi.org/10.1002/fsh.10238

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chen, Q., Meng, Z., Liu, X., Jin, Q., Su, R. (2018). Decision Variants for the Automatic Determination of Optimal Feature Subset in RF-RFE. Genes.. doi:10.3390/genes9060301

Cramer, J. S. (2003). The Origins of Logistic Regression. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.360300

Cressie, N., Calder, C.A., Clark, J.S., Hoef, J.M.V., Wikle, C.K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications.. doi:10.1890/07-0744.1

Death, R. G., & Collier, K. J. (2009). Measuring stream macroinvertebrate responses to gradients of vegetation cover: when is enough enough?. *Freshwater Biology*, *55*(7), 1447–1464. https://doi.org/10.1111/j.1365-2427.2009.02233.x

Didan, K. (2015). *MYD13A2 MODIS/Aqua Vegetation Indices 16-Day L3 Global 1km SIN Grid V006* [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 2021-04-15 from https://doi.org/10.5067/MODIS/MYD13A2.006

Dittrich, A., Roilo, S., Sonnenschein, R., Cerrato, C., Ewald, M., Viterbi, R., Cord, A.F. (2019). Modelling Distributions of Rove Beetles in Mountainous Areas Using Remote Sensing Data. Remote Sensing.. doi:10.3390/rs12010080

Dormann C.F., Elith J., Bacher S. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36, 27–46.

Downing, J. A., Plante, C., & Lalonde, S. (1990). Fish Production Correlated with Primary Productivity, not the Morphoedaphic Index. *Canadian Journal of Fisheries and Aquatic Sciences*, *47*(10), 1929–1936. https://doi.org/10.1139/f90-217

Draper, N. and Smith, H. (1981) Applied Regression Analysis, 2d Edition, New York: John Wiley & Sons, Inc

Ebersole, J.L., Liss, W.J., Frissell, C.A. (2001). Relationship between stream temperature, thermal refugia and rainbow trout Oncorhynchus mykiss abundance in arid-land streams in the northwestern United States. Ecology Of Freshwater Fish.. doi:10.1034/j.1600-0633.2001.100101.x

Elith, J., Kearney, M., Phillips, S. (2010). The art of modelling range-shifting species. Methods in Ecology and Evolution.. doi:10.1111/j.2041-210x.2010.00036.x

Elith, J. (2015). Predicting distributions of invasive species.

Farley, S.S., Dawson, A., Goring, S.J., Williams, J.W. (2018). Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions. BioScience.. doi:10.1093/biosci/biy068

Fausch, K. D., Taniguchi, Y., Nakano, S., Grossman, G. D., & Townsend, C. R. (2001). Flood disturbance regimes influence rainbow trout invasion success among five holarctic regions. Ecological Applications, 11, 1438–1455.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Fourcade, Y., Engler, J.O., Rödder, D., Secondi, J. (2014). Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. PLoS ONE.. doi:10.1371/journal.pone.0097122

Friedrichs-Manthey, M., Langhans, S. D., Hein, T., Borgwardt, F., Kling, H., Jähnig, S. C., & Domisch, S. (2020). From topography to hydrology—The modifiable area unit problem impacts freshwater species distribution models. *Ecology and Evolution*, *10*(6), 2956–2968. https://doi.org/10.1002/ece3.6110

Gallien, L., Münkemüller, T., Albert, C.H., Boulangeat, I., Thuiller, W. (2010). Predicting potential distributions of invasive species: where to go from here?. Diversity and Distributions.. doi:10.1111/j.1472-4642.2010.00652.x

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment.*

Groom, Q., Strubbe, D., Adriaens, T., Davis, A. J. S., Desmet, P., Oldoni, D., Reyserhove, L., Roy, H. E., & Vanderhoeven, S. (2019). Empowering Citizens to Inform Decision-Making as a Way Forward to Support Invasive Alien Species Policy. *Citizen Science: Theory and Practice*, *4*(1). https://doi.org/10.5334/cstp.238

Hansen, M. C., R. S. DeFries, J. R. G. Townshend, M. Carroll, C. Dimiceli, and R. A. Sohlberg. (2003). Global Percent Tree Cover at a Spatial Resolution of 500 Meters: First Results of the MODIS Vegetation Continuous Fields Algorithm. *Earth Interact.*, 7, 1–15, https://doi.org/10.1175/1087-3562(2003)007<0001:GPTCAA>2.0.CO;2.

Hauer, F. R., Stanford, J. A., & Lorang, M. S. (2007). Pattern and Process in Northern Rocky Mountain Headwaters: Ecological Linkages in the Headwaters of the Crown of the Continent1. *JAWRA Journal of the American Water Resources Association*, *43*(1), 104–117. https://doi.org/10.1111/j.1752-1688.2007.00009.x

Havel, J. E., Kovalenko, K. E., Thomaz, S. M., Amalfitano, S., & Kats, L. B. (2015). Aquatic invasive species: challenges for the future. *Hydrobiologia*, *750*(1), 147–170. https://doi.org/10.1007/s10750-014-2166-0

He, K. S., Bradley, B. A., Cord, A. F., Rocchini, D., Tuanmu, M., Schmidtlein, S., Turner, W., Wegmann, M., & Pettorelli, N. (2015). Will remote sensing shape the next generation of species distribution models?. *Remote Sensing in Ecology and Conservation*, *1*(1), 4–18. https://doi.org/10.1002/rse2.7

Hellmann, J.J., Byers, J.E., Bierwagen, B.G., Dukes, J.S. (2008). Five Potential Consequences of Climate Change for Invasive Species. Conservation Biology.. doi:10.1111/j.1523-1739.2008.00951.x

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology.. doi:10.1002/joc.1276

Hitt, N.P., Frissell, C.A., Muhlfeld, C.C., Allendorf, F.W. (2003). Spread of hybridization between native westslope cutthroat trout, Oncorhynchus clarki lewisi , and nonnative rainbow trout, Oncorhynchus mykiss. Canadian Journal of Fisheries and Aquatic Sciences.. doi:10.1139/f03-125

Jones, J. (2019). Improved Automated Detection of Subpixel-Scale Inundation—Revised Dynamic Surface Water Extent (DSWE) Partial Surface Water Tests. Remote Sensing.. doi:10.3390/rs11040374

Jiménez-Valverde, A., Lobo, J.M., Hortal, J. (2008). Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions.. doi:10.1111/j.1472-4642.2008.00496.x

Kearney, M., Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. Ecology Letters.. doi:10.1111/j.1461-0248.2008.01277.x

Kennedy, C.M., J.R. Oakleaf, D.M. Theobald, S. Baurch-Murdo, and J. Kiesecker. (2019). Managing the middle: A shift in conservation priorities based on the global human modification gradient. Global Change Biology 00:1-16. https://doi.org/10.1111/gcb.14549\u2028

Kovach, R. P., Hand, B. K., Hohenlohe, P. A., Cosart, T. F., Boyer, M. C., Neville, H. H., Muhlfeld, C. C., Amish, S. J., Carim, K., Narum, S. R., Lowe, W. H., Allendorf, F. W., & Luikart, G. (2016). Vive la résistance: genome-wide selection against introduced alleles in invasive hybrid zones. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1843), 20161380. https://doi.org/10.1098/rspb.2016.1380

Leitão, P.J., Santos, M.J. (2019). Improving Models of Species Ecological Niches: A Remote Sensing Overview. Frontiers in Ecology and Evolution.. doi:10.3389/fevo.2019.00009

Li, Z.-L., Tang, B.-H., Wu, H., Ren, H., Yan, G., Wan, Z., Trigo, I.F., Sobrino, J.A. (2013). Satellite-derived land surface temperature: Current status and perspectives. Remote Sensing of Environment.. doi:10.1016/j.rse.2012.12.008

Lobo, J.M., Jiménez-Valverde, A., Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. Ecography.. doi:10.1111/j.1600-0587.2009.06039.x

Martinez, B., Reaser, J.K., Dehgan, A., Zamft, B., Baisch, D., Mccormick, C., Giordano, A.J., Aicher, R., Selbe, S. (2020). Technology innovation: advancing capacities for the early detection of and rapid response to invasive species. Biological Invasions.. doi:10.1007/s10530-019-02146-y

Mansfield, E. R., & Helms, B. P. (1982). Detecting Multicollinearity. The American Statistician, *36*(3), 158. https://doi.org/10.2307/2683167

McCulloch, Warren S. and Pitts, Walter. (1943). A logical calculus of the ideas immanent in nervous activity The bulletin of mathematical biophysics 5.4 115-133.

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Mcnyset, K., Volk, C., & Jordan, C.. (2015). Developing an Effective Model for Predicting Spatially and Temporally Continuous Stream Temperatures from Remotely Sensed Land Surface Temperatures. *Water*, *7*(12), 6827–6846. https://doi.org/10.3390/w7126660

Merow, C., Lafleur, N., Silander Jr., J.A., Wilson, A.M., Rubega, M. (2011). Developing Dynamic Mechanistic Species Distribution Models: Predicting Bird-Mediated Spread of Invasive Plants across Northeastern North America. The American Naturalist.. doi:10.1086/660295

Mishina, Y., Murata, R., Yamauchi, Y., Yamashita, T., & Fujiyoshi, H. (2015). Boosted Random Forest. *IEICE Transactions on Information and Systems*, *E98.D*(9), 1630–1636. https://doi.org/10.1587/transinf.2014opp0004

Mitchell, K.E. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. Journal of Geophysical Research Atmospheres.. doi:10.1029/2003jd003823

Mo, K. C., Chen, L.-C., Shukla, S., Bohn, T. J., & Lettenmaier, D. P. (2012). Uncertainties in North American Land Data Assimilation Systems over the Contiguous United States. *Journal of Hydrometeorology*, *13*(3), 996–1009. https://doi.org/10.1175/jhm-d-11-0132.1

Muhlfeld CC., Kovach RP., Al-Chokhachy R., et al. (2017). Legacy introductions and climatic variation explain spatiotemporal patterns of invasive hybridization in a native trout. *Glob Chang Biol*.2017;23(11):4663-4674. doi:10.1111/gcb.13681

Muhlfeld, C.C., Kovach, R.P., Jones, L.A., Al-Chokhachy, R., Boyer, M.C., Leary, R.F., Lowe, W.H., Luikart, G., Allendorf, F.W. (2014). Invasive hybridization in a threatened species is accelerated by climate change. Nature Climate Change.. doi:10.1038/nclimate2252

Muhlfeld, C.C., Mcmahon, T.E., Boyer, M.C., Gresswell, R.E. (2009a). Local Habitat, Watershed, and Biotic Factors Influencing the Spread of Hybridization between Native Westslope Cutthroat Trout and Introduced Rainbow Trout. Transactions of the American Fisheries Society.. doi:10.1577/t08-235.1

Muhlfeld, C.C., Mcmahon, T.E., Belcer, D., Kershner, J.L. (2009b). Spatial and temporal spawning dynamics of native westslope cutthroat trout, Oncorhynchus clarkii lewisi, introduced rainbow trout, Oncorhynchus mykiss, and their hybrids. Canadian Journal of Fisheries and Aquatic Sciences.. doi:10.1139/f09-073

Muhlfeld, C.C., Kovach, R.P., Jones, L.A., Al-Chokhachy, R., Boyer, M.C., Leary, R.F., Lowe, W.H., Luikart, G., Allendorf, F.W. (2014). Invasive hybridization in a threatened species is accelerated by climate change. Nature Climate Change.. doi:10.1038/nclimate2252

NASA/GSFC, Greenbelt, MD, USA, NASA Goddard Earth Sciences Data and Information Services Center (GES DISC)

National Elevation Dataset (2002); Web site; U.S Geological Survey

Olden, Julian D., Lawler, Joshua J., Poff, LeRoy N. (2008). Machine Learning Methods Without Tears: A Primer for Ecologists. The Quarterly Review of Biology.. doi:10.1086/587826

Parr, T., Wilson, J., Hamrick, J. (2020). Nonparametric Feature Impact and Importance. arXiv:2006.04750

Pederson, G.T., Graumlich, L.J., Fagre, D.B., Kipfer, T., Muhlfeld, C.C. (2010). A century of climate and ecosystem change in Western Montana: what do temperature trends portend?. Climatic Change.. doi:10.1007/s10584-009-9642-y

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python*Journal of Machine Learning Research, 12*, 2825–2830.

Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature.. doi:10.1038/nature20584

Pelayo-Villamil, P., Guisande, C., Vari, R. P., Manjarrés-Hernández, A., García-Roselló, E., González-Dacosta, J., Heine, J., González Vilas, L., Patti, B., Quinci, E. M., Jiménez, L. F., Granado-Lorencio, C., Tedesco, P. A., & Lobo, J. M.. (2015). Global diversity patterns of freshwater fishes - potential victims of their own success. *Diversity and Distributions*, *21*(3), 345–356. https://doi.org/10.1111/ddi.12271

Philippe Massicotte, Andrea Bertolo, Philippe Brodeur, Christiane Hudon, Marc Mingelbier, & Pierre Magnan (2015). Influence of the aquatic vegetation landscape on larval fish abundance*Journal of Great Lakes Research, 41*(3), 873-880.

Pimentel, D., ed. (2002). Biological Invasions. Economic and Environmental Plants, Animals, and Microbe Species. Boca Raton, FL: CRC. 369 p

Pister, E.P. (2001). Wilderness Fish Stocking: History and Perspective. Ecosystems.. doi:10.1007/s10021-001-0010-7

Pyšek, P., Hulme, P. E., Simberloff, D., Bacher, S., Blackburn, T. M., Carlton, J. T., Dawson, W., Essl, F., Foxcroft, L. C., Genovesi, P., Jeschke, J. M., Kühn, I., Liebhold, A. M., Mandrak, N. E., Meyerson, L. A., Pauchard, A., Pergl, J., Roy, H. E., Seebens, H., … Richardson, D. M. (2020). Scientists' warning on invasive alien species. *Biological Reviews, 95*(6), 1511–1534. https://doi.org/10.1111/brv.12627

Randin, C. F., Ashcroft, M. B., Bolliger, J., Cavender-Bares, J., Coops, N. C., Dullinger, S., Dirnböck, T., Eckert, S., Ellis, E., Fernández, N., Giuliani, G., Guisan, A., Jetz, W., Joost, S., Karger, D., Lembrechts, J., Lenoir, J., Luoto, M., Morin, X., … Payne, D. (2020). Monitoring biodiversity in the Anthropocene using remote sensing in species distribution models. *Remote Sensing of Environment, 239*, 111626. https://doi.org/10.1016/j.rse.2019.111626

Reaser, J.K., Burgiel, S.W., Kirkey, J., Brantley, K.A., Veatch, S.D., Burgos-Rodríguez, J. (2020a). The early detection of and rapid response (EDRR) to invasive species: a conceptual framework and federal capacities assessment. Biological Invasions.. doi:10.1007/s10530-019-02156-w

Reaser, J. K., Simpson, A., Guala, G. F., Morisette, J. T., & Fuller, P. (2020b). Envisioning a national invasive species information framework. *Biological Invasions, 22*(1), 21–36. https://doi.org/10.1007/s10530-019-02141-3

Ricciardi, A., Blackburn, T. M., Carlton, J. T., Dick, J. T. A., Hulme, P. E., Iacarella, J. C., Jeschke, J. M., Liebhold, A. M., Lockwood, J. L., Macisaac, H. J., Pyšek, P., Richardson, D. M., Ruiz, G. M., Simberloff, D., Sutherland, W. J., Wardle, D. A., & Aldridge, D. C. (2017). Invasion Science: A Horizon Scan of Emerging Challenges and Opportunities. *Trends in Ecology & Evolution, 32*(6), 464–474. https://doi.org/10.1016/j.tree.2017.03.007

Robinson, N.P., B.W. Allred, W.K. Smith, M.O. Jones, A. Moreno, T.A. Erickson, D.E. Naugle, and S.W. Running. (2018). Terrestrial primary production for the conterminous United States derived from Landsat 30 m and MODIS 250 m. Remote Sensing in Ecology and Conservation. doi:\xa0https://doi.org/10.1002/rse2.74\u2028

Runting, R.K., S. Phinn, Z. Xie, O. Venter and J.E.M. Watson. (2020). Opportunities for big data in conservation and sustainability. Nature Communications 11(1).

Schneider, A., 2012. Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach. Remote Sensing of Environment 124, 689–704.. doi:10.1016/j.rse.2012.06.006

Seaber, P.R., F.P. Kapinos, and G.L. Knapp. (1987). Hydrologic units maps. Water-Supply Paper 2294, U.S. Geological Survey, Reston, VA.

Seebens, H., Blackburn, T. M., Dyer, E. E., Genovesi, P., Hulme, P. E., Jeschke, J. M., Pagad, S., Pyšek, P., Winter, M., Arianoutsou, M., Bacher, S., Blasius, B., Brundu, G., Capinha, C., Celesti-Grapow, L., Dawson, W., Dullinger, S., Fuentes, N., Jäger, H., … Essl, F. (2017). No saturation in the accumulation of alien species worldwide. *Nature Communications*, *8*(1), 14435. https://doi.org/10.1038/ncomms14435

Sepulveda, A., A. Ray, R. Al-Chokhachy, C. Muhlfeld, R. Gresswell, J. Gross, and J. Kershner. (2012). Aquatic invasive species: lessons from cancer research. *American Scientist* 100: 234–242.

Shackleton, R.T., Shackleton, C.M., Kull, C.A. (2019). The role of invasive alien species in shaping local livelihoods and human well-being: A review. Journal of Environmental Management.. doi:10.1016/j.jenvman.2018.05.007

Sobrino, J. A., Jiménez-Muñoz, J. C., & Paolini, L.. (2004). Land surface temperature retrieval from LANDSAT TM 5. *Remote Sensing of Environment*, *90*(4), 434–440. https://doi.org/10.1016/j.rse.2004.02.003

Srivastava, Vivek & Lafond, Valentine & Griess, Verena. (2019). Species distribution models (SDM): applications, benefits and challenges in invasive species management. CAB Reviews Perspectives in Agriculture Veterinary Science Nutrition and Natural Resources. 14. 1-13. 10.1079/PAVSNNR201914020.

Stolar, J., Nielsen, S.E., 2015. Accounting for spatially biased sampling effort in presence-only species distribution modelling. Diversity and Distributions 21, 595–608.. doi:10.1111/ddi.12279

Sweeney, B. (1993). Effects of Streamside Vegetation on Macroinvertebrate Communities of White Clay Creek in Eastern North America. *Proceedings of the Academy of Natural Sciences of Philadelphia, 144*, 291-340. Retrieved March 23, 2021, from http://www.jstor.org/stable/4065013

Tarca, A. L., Carey, V. J., Chen, X.-W., Romero, R., & Drăghici, S. (2007). Machine Learning and Its Applications to Biology. *PLOS Computational Biology*, *3*(6), e116. https://doi.org/10.1371/journal.pcbi.0030116

Tomlinson, C.J., Chapman, L., Thornes, J.E., Baker, C. (2011). Remote sensing land surface temperature for meteorology and climatology: a review. Meteorological Applications 18, 296–306.. doi:10.1002/met.287

Theobald, D.M., Harrison-Atlas, D., Monahan, W.B., Albano, C.M. (2015). Ecologically-Relevant Maps of Landforms and Physiographic Diversity for Climate Adaptation Planning. PLoS ONE.. doi:10.1371/journal.pone.0143619

Thessen, A. (2016). Adoption of Machine Learning Techniques in Ecology and Earth Science. One Ecosystem.. doi:10.3897/oneeco.1.e8621

Thuiller, W. (2003). BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology.. doi:10.1046/j.1365-2486.2003.00666.x

Thuiller, W., Lafourcade, B., Engler, R., Araújo, M.B. (2009). BIOMOD - a platform for ensemble forecasting of species distributions. Ecography.. doi:10.1111/j.1600-0587.2008.05742.x

U.S. Geological Survey. (2020). Nonindigenous Aquatic Species Database, Gainesville, FL. http://nas.er.usgs.gov, 29 July 2020.

van Rees, C.B., Hand, B.K., Bargeron, C., Carter, S.C. et al. (2022). *Accepted*. A Framework to Integrate Innovations in Invasion Biology Science for Proactive Management. *Biological Reviews.*

Vaz, A. S., Alcaraz-Segura, D., Vicente, J. R., & Honrado, J. P. (2019). The Many Roles of Remote Sensing in Invasion Science. *Frontiers in Ecology and Evolution*, *7*. https://doi.org/10.3389/fevo.2019.00370

Vieira, T. B., Dias-Silva, K. & Pacifico, E. S. (2015). Effects of riparian vegetation integrity on fish and Heteroptera communities. — Applied Ecology and Environmental Research 13: 53–65.

Wan Z., Hook, S., Hulley, G. (2015). MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD11A2.006

Waring, R.H., Coops, N.C., Fan, W., Nightingale, J.M. (2006). MODIS enhanced vegetation index predicts tree species richness across forested ecoregions in the contiguous U.S.A.. Remote Sensing of Environment.. doi:10.1016/j.rse.2006.05.007

Westbrooks, Randy G. (2004). New Approaches for Early Detection and Rapid Response to Invasive Plants in the United States. *Weed Technology*, vol. 18, pp. 1468–1471. *JSTOR*, www.jstor.org/stable/3989673. Accessed 11 Aug. 2020.

Welti, N., Striebel, M., Ulseth, A. J., Cross, W. F., Devilbiss, S., Glibert, P. M., Guo, L., Hirst, A. G., Hood, J., Kominoski, J. S., Macneill, K. L., Mehring, A. S., Welter, J. R., & Hillebrand, H.. (2017). Bridging Food Webs, Ecosystem Metabolism, and Biogeochemistry Using Ecological Stoichiometry Theory. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.01298

Wenger, S. J., Isaak, D. J., Luce, C. H., Neville, H. M., Fausch, K. D., Dunham, J. B., Dauwalter, D. C., Young, M. K., Elsner, M. M., Rieman, B. E., Hamlet, A. F., & Williams, J. E. (2011). Flow regime, temperature, and biotic interactions drive differential declines of trout species under climate change. *Proceedings of the National Academy of Sciences*, *108*(34), 14175–14180. https://doi.org/10.1073/pnas.1103097108

Whitehead, D.A., F.G. Magaña, J.T. Ketchum, E.M. Hoyos, R.G. Armas, F. Pancaldi, and D. Olivier. 2020. The use of machine learning to detect foraging behavior in whale sharks: a new tool in conservation. Journal of Fish Biology 98(3). https://doi.org/10.1111/jfb.14589

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*, 79–82. https://doi.org/10.3354/cr030079

Wu, H., J.S. Kimball, M.M., Elsner, N. Mantua, R.F. Adlere, and J. Stanford. (2012). Projected climate change impacts on the hydrology and temperature of Pacific Northwest rivers. *Water Resources Research*, 48, W11530, doi:10.1029/2012WR012082.

SUPPLEMENTAL MATERIALS

S1) Detailed quality control filtering actions

Data Product	Quality Control (QC) Flag Description	Quality Filtering Action
Land Surface Temperature	Bits 0 and 1: - 0: Pixel produced, good quality, not necessary to examine more detailed QA - 1: Pixel produced, unreliable or unquantifiable quality, recommend examination of more detailed QA - 2: Pixel not produced due to cloud effects - 3: Pixel not produced primarily due to reasons other than cloud (such as ocean pixel, poor input data) *(Etc)*	Only used pixels where bits 0 and 1 equal 0
Gross Primary Productivity	Value 10: Clear not smoothed Value 11: Clear smoothed Value 20: Snow or water not smoothed Value 21: Snow or water smoothed Value 30: Climatology not smoothed Value 31: Climatology smoothed Value 40: Gap filled not smoothed Value 41: Gap filled smooth	Only used pixels where QC band equaled 10 or 11
Enhanced Vegetation Index	Bits 0 and 1: - 0: Pixel produced with good quality - 1: Pixel produced, but check other QA - 2: Pixel produced, but most probably cloudy - 3: Pixel not produced due to other reasons than clouds *(Etc)*	Only used pixels where bits 0 and 1 equal 0
Percent Tree Cover	Bit 0: State of input layers DOY 065-097 - 0: Clear - 1: Bad Bit 1: State of input layers DOY 113-145 - 0: Clear - 1: Bad Bit 2: State of input layers DOY 161-193 - 0: Clear - 1: Bad Bit 3: State of input layers DOY 209-241 - 0: Clear - 1: Bad Bit 4: State of input layers DOY 257-289 - 0: Clear - 1: Bad Bit 5: State of input layers DOY 305-337 - 0: Clear - 1: Bad Bit 6: State of input layers DOY 353-017 - 0: Clear - 1: Bad Bit 7: State of input layers DOY 033-045 - 0: Clear - 1: Bad	Only used “Clear” pixels for all bits

Table S1. Detailed description of pre-published quality filtering heuristic rules and our stringent quality masking procedures.

S2) Description of Machine Learning model implementation

Logistic Regression functions within the maximum likelihood framework by performing gradient descent on the error surface characterized by the difference between observed and predicted suitability (Cramer 2003). A classification tree is built by splitting the input data into successive "leaves" to minimize the gini impurity between consecutive layers of the canopy (Quinlan 1986). Random forests are trained using a series of such classification trees on independently bootstrapped samples (Breiman 2001). Boosted Regression Trees are similar to random forests, but instead recursively build each classification tree using the remaining errors left by previous, intentionally "shallow" trees rather than independently bootstrapped samples of the same data used to train stronger learners (Mishina et al., 2015). XGBoost is a more generalizable form of Boosted Regression Trees that incorporates regularization and a more accurate gradient descent algorithm (Chen and Guestrin 2016). Neural networks are trained using a gradient descent algorithm that minimizes the error between observations and predictions (McCulloch and Pitts 1943).

S3) Figure S3

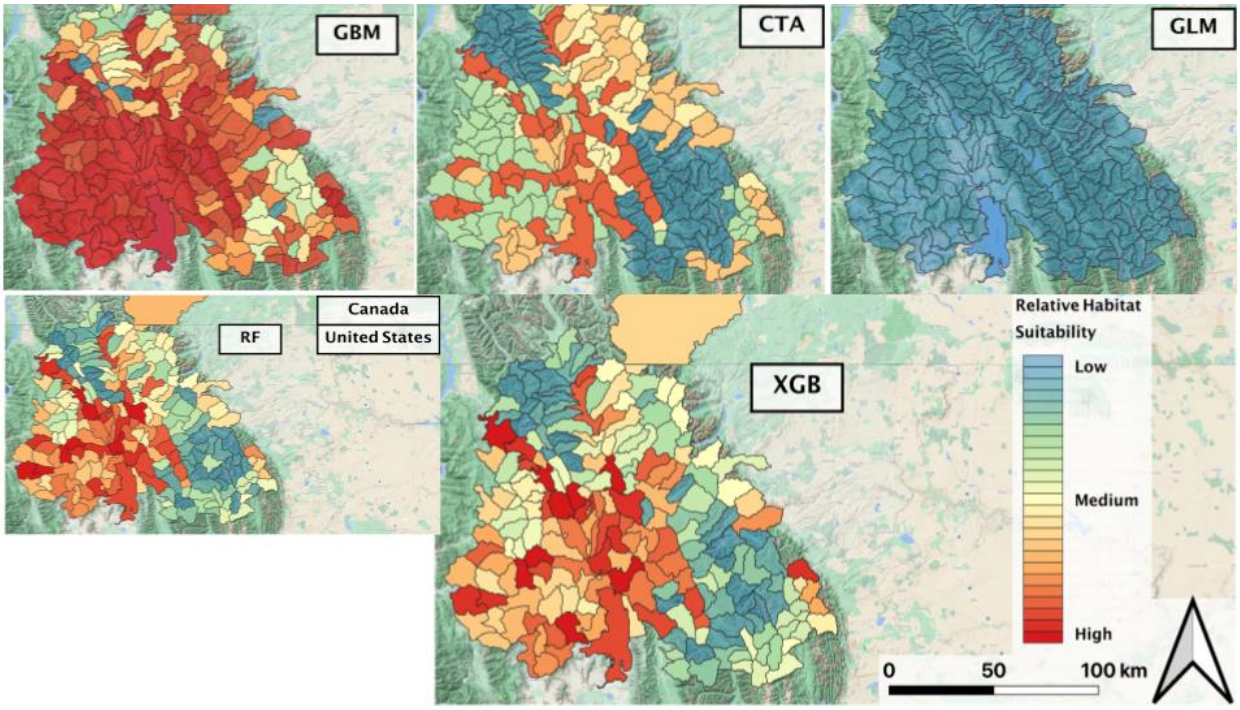

Figure S3. Comparison of estimated relative habitat suitability over the Flathead study region derived from the different ML models (i.e. GBM, CTA, GLM, RF, XGB). All of the models predict higher habitat suitability in the southwest portion of the study area, an area which was shown to be outside of the training envelope (Figure 7). The ANN was not included in this figure; predictions are the exact opposite of the GLM (i.e. high suitability for all regions of the study area).