ABSTRACT

Title of Thesis:	A COMPARATIVE ANALYSIS OF RANDOM FOREST AND LOGISTIC REGRESSION FOR WEED RISK ASSESSMENT
	Chinchu Harris, Master of Science, 2018
Thesis Directed By:	Assistant Professor, Wendy Peer, Department of Environmental Science and Technology

Invasive species have largely negative impacts on the environment and the economy. The management and regulation of invasive plants are facilitated using screening tools, such as weed risk assessments (WRAs) to predict the invasive potential of non-native plants. The identification of these species and their subsequent regulation on importation helps to reduce the risk of future ecosystem and economic costs. Globally, there are many different types of highly useful WRAs already available. However, in this day of big data and powerful predictive analytics, there is an increasing demand for the development of new and more robust screening tools. In this thesis, I use the machine learning algorithm, Random forests, to develop a new WRA. I show that random forest model has greater predictive accuracies than an existing logistic regression model and that random forest is a better learner. In addition, variable importance analysis was performed to identify factors associated with invasive status classification of non-native plants. The study suggests that

random forests make powerful weed risk screening tools and should be utilized for assessing invasive risk potential along with other WRAs. An integrative approach for evaluating weed risk can greatly serve to facilitate the WRA process.

A COMPARATIVE ANALYSIS OF RANDOM FOREST AND LOGISTIC REGRESSION FOR WEED RISK ASSESSMENT

by

Chinchu Harris

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2018

Advisory Committee:
Assistant Professor Wendy Peer, Chair
Assistant Professor Lea Johnson
Associate Professor Paul Smith
Dr. John Payne

Foreword

This thesis is one of the fruits of my labor in the department of Plant Science and Landscape Architecture at the University of Maryland, College Park. Entering graduate school in PSLA of the college of Agriculture and Natural Sciences (AGNR) was a true learning experience. I started out my studies as an idealistic individual who wanted to grapple with many different aspects of the Invasion Ecology of plants. After a period of uncertainty I was able to pinpoint the direction my master's thesis would ultimately take.

I did not have much experience in programming prior to this experience, but I wanted to challenge myself and learn something new in my graduate student journey. Although learning the syntax of programming was challenging in the beginning, I soon got the hang of using such a tool and came to really enjoy my time coding and interacting with other R programmers whether by attending R Ladies meet ups or interacting with other R users on Stack Overflow and other programming centered resources. I became fascinated with the many different ways coding can be utilized to complete a wide variety of tasks. I was particularly excited by the way individuals in academia use Open Source tools like R to facilitate their research interests.

The basis for this research originally stemmed from my passion for finding ways to improve what is already available. After many conversations with Dr. Tony Koop at the United States Department of Agriculture (USDA) Animal and Plant Health Inspection Services (APHIS), I was given the opportunity to use the Plant

Protection and Quarantine (PPQ) Weed Risk Assessment (WRA) dataset to start my research project of a comparative analysis of random forests and logistic regression for weed risk assessment. I hope you enjoy reading my work.

Dedication

This thesis is dedicated to my parents, Bilhi and Kunjachan Harris.

Acknowledgements

Thank you to Wendy Peer for acting as my advisor for this research. Thank you Paul Smith for providing purposeful and insightful conversations on statistical modeling. Lea Johnson and John Payne for participating in the committee and their comments on the thesis. Angus Murphy for supporting this work from the beginning. Tony Koop for providing the dataset and spending many hours of correspondence. Benoit Parmetier for useful conversations. Kate Tully and Joe Sullivan for their continued encouragement.

Thanks for all the support of my former lab members Jun Zhang, Jemina Huang, Sarah Turner, Kasuni Wattarantenne, and my good friend Fernanda Mastrotti. I would like to also acknowledge fellow graduate students Angela Ferelli, Mary Theresa Callahan, Daniela Miller, Kayla Griffith, and Brianna Otte for their listening ears and advice throughout this graduate experience.

A very special thanks to my parents, siblings, and extended members for all their love and encouragement. Last but not least, I would like to thank my beloved husband, David. Thank you for your steadfast love and fostering my aspirations throughout graduate school.

Table of Contents

This Table of Contents is automatically generated by MS Word, linked to the Heading formats used within the Chapter text.

List of Tables

List of Figures

Chapter 1: Literature Review: The invasion process and the regulation of non-native invasive plants through predictive modeling

Non-native invasive plants can cause ecological and economic damage to new environmental ranges (Pimentel *et al.*, 2000; Richardson *et al.*, 2000; Hulme *et al.*, 2013). Biological invasions in agriculture account for a total of $120 billion in losses annually and out of that figure, alien weeds, which are plants that grow in unwanted geographical areas without being purposefully cultivated (Baker, 1991), cause $24 billion loss in the agricultural industry (Pimentel *et al.*, 2005). Strategies implemented to regulate invasive plant species include prevention through the implementation of screening tools like weed risk assessments (WRAs), and management practices such as, early detection, eradication, and control (Poorter *et al.*, 2005). Eradication and control of invasive plants can be costly, while prevention and early detection are the most economic strategies for minimizing invasive spread (Lockwood *et al.*, 2007).

Prevention in particular, through the use of screening tools such as WRAs, is the first line of defense against the spread of invasive plants and there is a need for the development of a faster scientific assessment of weed risk of plant taxa proposed for introduction. Conducting these assessments can better serve the interests of stakeholders if there is a concern for the plant taxon risk potential.

Inclusion of more types of data in the assessment will further help identify potential invasive plant taxa in the pending risk analyses process. In particular,

ecological and climate change models may inform the development of new, more dynamic and integrative weed risk models. As the climate is changing and ecosystems are not static, it becomes essential to have a dynamic WRA that incorporates predictions of climate change. However, developments of such integrative WRAs are limited to the availability of more and new types of data describing the relationship of non-native plants in terms of niche suitability, extreme weather event factors, plant-soil-microbe feedbacks and how it relates to species abundance, molecular biology studies that explore their genome attributes. Given the challenge in amassing ecological data specific to plant taxon proposed for introduction, a different approach to improve the predictive accuracies of WRAs could be to use improved machine-learning statistical methods, such as Random forests, on already existing data.

The invasion process

Invasion ecology has been a rapidly growing field in biology over the last six decades (Mooney *et al.*, 2005; Richardson & Pysek, 2008; Hoopes *et al.*, 2013). Understanding the invasion process is an important facet to the study of invasive species. The invasion process has been ably described elsewhere by Theoharides & Dukes (2007) and Mooney *et al.* (2005). The four stages of the invasion process will be described here. The invasion stages include the transport, establishment, spread, and impact of invasive species.

The first stage is the physical transport of the non-native species into a new geographical location. The rapid acceleration of large-scale anthropogenic

movements can be attributed to the increased spread of invasive plants (Hulme, 2015).

Establishment, the second stage of the invasion process, occurs when non-native species are able to survive and reproduce outside their native range. This may involve overcoming barriers such as abiotic and biotic factors of the new environment, as well as competition with native species for space and resources (Richardson *et al.*, 2000).

The third invasion stage is spread, where the non-native population spatially disperses into areas beyond the initial establishment. Dispersal through multiple vectors may help non-native plants to become widespread (PPQ, 2016). Dispersal mechanisms of successful plant invaders, which make them well-adapted to a large range of ecosystems, include unique characteristic traits, such as prolific and viable seed production, spreading roots, rhizomes and runner adventitious roots, or creeping stems (Richardson *et al.*, 2000).

Identification of plant traits that promote invasiveness can provide insight into the spread stage of the invasion process. For example, performance-related traits such as physiology, leaf-area allocation, shoot allocation, growth rate, and fitness showed a higher association with invasiveness for plants (see Van Kleunen *et al.*, 2010 for the complete meta-analysis of 117 studies comparing trait differences between invasive and non-invasive plant species).

The last stage of the invasion process is impact, where humans perceive the magnitude of impacts caused by invasive species. These impacts include negative effects on ecosystem structure and function (Neckles, 2015; Naeem *et al.*, 1994),

agriculture (Paini *et al.*, 2016), forestry (Anagnostakis, 1987; Maloy, 1997), fisheries (Griffiths *et al.*, 1991; Walton *et al.*, 2002; Rothlisberger *et al.*, 2012), and recreation (Eiswerth *et al.*, 2005; Pimentel *et al.*, 2000). See the example below for impacts of invasive plants in enactment of regulatory policies. Overall, each stage of the invasion process is complex and is influenced by myriad of factors.

Impacts of invasive cheatgrass and medusahead in action: Greater sage grouse

Populations of greater sage grouse, a keystone species found in the sagebrush ecosystem, have dwindled throughout the years (Crawford *et al.*, 2004). This can indirectly be contributed to increasing populations of cheatgrass and medusahead plants in the Great Basin. These invasive plants, cheatgrass and medusahead, outcompete native plants, such as sagebrush, and are easily combustible in the Sagebrush-Steppe habitat (Wambolt *et al.*, 2002; Taylor *et al.*, 2012). This becomes an issue for sage grouse, since they are heavily reliant on sagebrush for food, nesting, and cover from predators (Wambolt *et al.*, 2002; Schroeder *et al.*, 2006). While wildfires are a crucial component of the health of the sagebrush ecosystem, an increased frequency of fires jeopardizes the livelihood of sage grouse causing an overall decrease in their populations (Wambolt *et al.*, 2002). Decreases in sage grouse populations are culturally impactful for humans who game these birds (Guttery *et al.*, 2016).

In the past there have been failed attempts to list sage grouse under the Endangered Species Act (Hess, 2015). National Defense Authorization Act (NDAA), put together by the House Armed Services Committee, contains bill H.R. 4739, which prevents the greater sage grouse from being listed under the Endangered Species Act

of 1973 before September 30, 2026 (2016). This $600 billion annual bill for the U.S. defense policy argued that considering these birds as endangered would limit use of its rangeland for military training.

Since the sage grouse was not added to the endangered species list, the lands inhabited by sage grouse will continue to be used for oil drilling and windmills. These anthropogenic physical barriers, along with the expansive presence of invasive plants, compromise the integrity of the breeding grounds for sage grouse (Wambolt *et al.*, 2002). Female sage grouse tend to nest on larger sagebrush that is ≥ 2 miles away from where breeding occurs (Schroeder *et al.*, 2006), but nesting is jeopardized with the presence of urban equipment and smaller and fewer sagebrush.

Due to the ecological and cultural importance of these birds, restoration and conservation of sage grouse populations is considered a priority by some. Impacts of non-native invasive plants need to be further investigated to see if regulation that deters further degradation of the sagebrush-steppe ecosystem needs to be potentially enacted. Enforcing more regulatory action through vigorous weed risk assessments could help to assess the invasive potential of plants considered for introduction, thereby potentially preventing or decreasing the destruction of native species populations and ecosystem in the Sagebrush-Steppe habitat by new invasive species.

Predicting invasive potential of plants

According to the International Plant Protection Convention (IPPC) in 1997 of the Food and Agricultural Organization of the United Nations (FAO), plant protection

organizations in each country must conduct pest risk analysis in order to protect and preserve their native plant resources. The United States Department of Agriculture (USDA), Animal and Plant Health Inspection Service (APHIS), Plant Protection and Quarantine (PPQ) is one such national plant protection organization. One of the responsibilities of USDA-APHIS-PPQ is to safeguard native plants from noxious weeds.

PPQ WRA

The USDA-APHIS-PPQ WRA was developed as a preventative measure to limit the entry of potential non-native invasive plants into the United States. The PPQ WRA is based on the Australia WRA and has been a template for other screening tools such as the New Zealand WRA and the Hawaii-Pacific WRA. These models adapt questions from the Australia WRA to adjust for regional differences. For example, questions addressing Australia's "arid climate" were adapted to New Zealand's "equable climate" (Pheloung *et al.,* 1999). Refer to Koop *et al.* for a comparative analysis of the Australia WRA and PPQ WRA for estimates of accuracy, error, and predictive value (2012). One of the goals of the PPQ WRA is to evaluate the potential invasiveness of plant taxon as a candidate for Federal Noxious Weed listing. Once a plant taxon is listed as a Federal Noxious Weed, humans are prohibited from intentionally transporting it across state lines. (PPQ, 2016).

The principal risk elements of the PPQ WRA include questions evaluating the establishment/spread and impact potential of a plant taxon. Establishment and spread are discrete stages of the invasion process, but these stages are combined to form one

risk element in the APHIS PPQ WRA. These two risk elements (establishment/spread and impact potential) of the PPQ WRA evaluate the natural history of species or conspecific. Some of the questions on the PPQ WRA questionnaire refer to the known invasive potential of plants, such as invasive status outside native range (es1), weed status in natural systems (impn6), and weed status in production systems (immp6). A majority of the questions refer to ecological traits of plants that are known to contribute to invasiveness, such as shade tolerance (es4), nitrogen fixing capability (es9), and minimum generation time (es13). Refer to Appendix A for the full questions of the establishment/spread and impact potential sections. The establishment/spread potential section assesses the establishment and spread status of a plant outside of its native range, while impact potential evaluates the weed status of the taxon and its impact on trade. The PPQ WRAs are produced by the Plant Epidemiology and Risk Analysis Laboratory (PERAL) and scientifically reviewed by at least one trained PPQ WRA risk analyst (PPQ, 2016).

Uncertainty

Each answer and risk element in the PPQ WRA is explicitly evaluated for uncertainty in order to ensure that weed risk is assessed based on adequate scientific evidence. Uncertainty is contextualized by the quality and quantity of available literature support, which is circumscribed by missing or incomplete information, evidence that is inconsistent or conflicting, and old or wrong information.

Every PPQ WRA response contains a degree of associated uncertainty, which is categorized as negligible, low, moderate, high, or maximum. Maximum

uncertainty, the greatest degree of uncertainty, is assigned for answers lacking adequate literature support (PPQ, 2016). The PPQ WRA analyzes answer uncertainty, using a software program called @RISK, by running a Monte Carlo simulation 5,000 times. This simulation, based on the assigned uncertainty levels associated with each answer, replaces original answers with different answers, thus generating new responses for evaluating WRA risk scores in order to assess how uncertainty may affect the assessment outcome (PPQ, 2016).

The predictive model

The PPQ WRA model was developed with 204 plant species (Appendix B) with known invasive status in the U.S. The invasive status, i.e. whether the taxon is a non-invader (n=68), minor-invader (n=68), or major-invader (n=68), dictates the overall WRA risk score of the taxon. Non-invaders are plants that are not naturalized but have occupied the United States for 75 years or more. This minimum residency time ensures that the non-invader has had enough time to escape and establish, and 75 years was specifically chosen because the PPQ WRA used *Hortus* as a historical reference for plants cultivated in North America (Bailey & Bailey, 1930), but lag time for invasions can range from less than 50 years to 100 years or more (Kowarik, 1995). Plant naturalization status in the U.S. was determined primarily with the use of the USDA PLANTS database (http://plants.usda.gov/). Major-invaders are categorized with "I-rank" impact ranking of high or high-medium on NatureServe's categorization (NatureServe, 2009), or listed as "serious" or "principal" by Holm *et*

al. (1979), or listed as "troublesome" by Bridges (1992). Minor-invaders are plants that are naturalized in the United States but do not fit the major-invader requirements.

Collectively, the establishment/spread and impact potential risk elements of these 204 plant species were modeled with the Logit Generalized Linear Model (GLM) statistical method. This logistic regression model predicts the invasive status of a plant taxon. The risk score assigns the plant taxon to one of the three following risk categories: "Low Risk", "Evaluate Further", and "High Risk" (PPQ, 2016). This categorized risk score determined by the predictive model "quantifies a plant taxon's ability to escape, establish, spread, and cause harm" in the U.S. (PPQ, 2016). This prediction of invasive potential is based on data considered without significant reference to time, despite the fact that specific invasion processes may change over time due to factors such as climate variabilities. For example, a plant taxon predicted by the model to have the invasive status minor-invader may in fact be a minor-invader at this point in time, but may go onto become a major-invader or non-invader depending on future climate variabilities.

New considerations for assessing weed risk

Even though broad scale screening tools like the PPQ WRA are available, it is worthwhile to consider some additional types of data. For example, plant hardiness (Higgins & Richardson, 2014) and genomic plasticity (Des Marais, 2013) were found to be predictive factors of invasiveness. Moreover, incorporation of distribution and abundance data, which are not included in the PPQ WRA, of potential invasive plants both in native and invaded (if there any) ranges in WRAs may enhance model

applicability (Wilson *et al.*, 2007; Pearman *et al.*, 2008) at a smaller jurisdictional scale (e.g., regional, state, local). Further, incorporation of data focusing on niche, nutrient input and impact on species abundance, and extreme weather events could also increase robustness of regional, state, or local risk assessments.

In the absence of new types of data for small-scale WRAs, statistical methods like machine-learning algorithms, such as Random forests, could be implemented for broad-scale WRAs like the PPQ WRA. These newer statistical methods for assessing weed risk may improve upon the predictive accuracies of already existing screening tools like the PPQ WRA.

What follows is a sampling of additional data types that could be considered for the improvement of broad WRA models or ones that are more regionally based.

Niche data for predicting weed risk

Incorporation of niche data to broad scale weed risk assessments, such as the PPQ WRA, is problematic due to the large geographic scale of the United States (PPQ, 2016). Nonetheless, incorporation of niche datasets for small scale WRAs that address regional environmental differences could complement the broad scale WRA. For example, while there is a lack of data available for analyses of spatio-temporal niche dynamics during invasions (Broennimann *et al.*, 2014), such analyses could inform the development of smaller scale preventative measures that look into the velocity of the invasion process. One study found a much slower initial invasion of *Centaurea stoebe* in habitats dissimilar to the native niche (Broennimann *et al.*, 2014).

Another niche consideration that is not addressed by broad scale WRAs is the ability of alien populations to realize different climatic niches compared to their native populations. Most of these WRAs look into the degree in climate match of species between the native niche and the non-native potential niche (Pheloung *et al.,* 1999). This poses a challenge when assessing invasion risk because potential distributions may be either underestimated or overestimated. Suggesting one potential approach to overcome this challenge with respect to climate change, a study examined climatic suitability under current conditions and future scenarios, by creating models of distribution using the Maximum Entropy "MaxEnt" model (Beaumont *et al.,* 2014). They compared and assessed the realized climatic niche for subspecies of the Australian invasive plant *Chrysanthemoides monilifera* under current and future extreme weather event scenarios, and showed that alien populations can occupy a new climatic niche not present in their native habitat. Their study validated a ban that was in place for the importation of *C. monilifera* subspecies from South Africa, and supported the importance of taking niche shifts into account via modeling tools to guide policy decisions.

Despite improvements to WRAs that niche data can afford, a potential complication is that the ecological niche theory argues that for every species, only a fraction of its potential niches is ever realized (Shah & Shaanker, 2014). Therefore, a species introduced into a new environment may be simply expanding its original niche, in which case it should not be considered "invasive" (Shah & Shaanker, 2014), although the species may be alien in that niche. A further complication is that ecological niche theory does not consider the intentional or unintentional

transportation of species to new environments by humans. Application of the theory to WRA therefore represents a departure that requires closer consideration.

In cases where adaption of niche theory can be shown to be sufficiently robust, its application could better account for regional environmental differences and thereby improve the broad scale WRA.

Climatic change data for predicting weed risk

Incorporation of climate data to broad scale WRAs, such as the PPQ WRA, is not considered because the United States, due to its large size and land area distribution across different latitudes, is climatically diverse (PPQ, 2016). The PPQ WRA was designed to be climatically neutral in order to eliminate bias against smaller U.S. climatic regions (PPQ, 2016), but climate data could be implemented in small scale WRAs at regional, state, or local levels.

Extreme weather event data is one type of climate data that has a limited number of available empirical studies and should be further studied. An increase in frequency and severity of extreme weather events may facilitate future invasions by creating disturbances and altering resource availabilities (Jiménez *et al.,* 2011). Climate change could facilitate the expansion of invasive plants into new ranges where previously they were not able to survive and reproduce. For example, phenological events, which refer to the timing of plant growth and reproduction, will change in response to changes in climatic variables such as temperature and rainfall (Badeck *et al.,* 2004) and could potentially increase fecundity of invasive plants. Enhanced fecundity is an important trait associated with invasion success in a new

range (Pyšek & Richardson, 2007). As extreme weather events, like cyclones, heatwaves and frosts, droughts and floods, are becoming more common, incorporation of these types of data into small scale WRAs can show a different perspective to assessing weed risk, which has not yet been fully explored. Flooding events have been noted to benefit the invasion of *Tamarix aphylla* by dispersing seeds along the entirety of the Finke River (Griffin *et al.*, 1989). Extreme weather models should be combined with species distribution and small-scale weed risk models to help predict potential future species distributions of potentially invasive non-native plants with respect to occurrence of extreme weather events. This could be particularly useful in states like California where in recent years have experienced multi-year droughts followed by heavy precipitation with a high number of atmospheric river storms lasting several months and is predicted to have an increase in the dry season and in sudden precipitation events in the future (Swain *et al.*, 2018).

One study looked at the determinants of changes in biodiversity for the year 2100 based on atmospheric carbon dioxide, weather events, vegetation, and land use (Sala *et al.*, 2000). Factors such as changes in land use, climate, nitrogen deposition and acid rain, biotic exchanges (introduction of new species), and atmospheric CO_2 concentration were ranked based on importance to driving extreme weather events for a predictive model (Sala *et al.*, 2000). Different global models were used to measure the magnitude of change in climate and land use due to extreme weather events. Land use was found to have the most devastating impact due to effects on habitat availability and species extinctions. At higher latitudes the average temperature is predicted to increase, with less pronounced fluctuations in atmospheric CO_2 levels.

However, there were changes in competitive balance observed between species that differed in root depths, photosynthetic pathways, woodiness, and association with belowground organisms. The study also found that an increase in atmospheric CO_2 had the greatest effect on biodiversity in biomes with a mixture of C_3 and C_4 plant species (i.e. Grasslands and Savannas) and in biomes where limitations in plant growth are mostly due to the scarce availability of water (i.e. Mediterranean ecosystems and deserts). An increase in nitrogen deposition was predicted to have the largest impact on biodiversity in nitrogen-limited habitats (northern temperate forests close to cities) (Sala *et al*., 2000).

Another study used biogeography and current weather event data to configure the species distribution model "MaxEnt" to predict the invasive risk potential and future distributions of three non-native aquatic plants; *Alternanthera philoxeroides* **(**alligatorweed), *Limnophila sessiliflora* (limnophilia), and *Salvinia molesta* (giant salvini) in the United States under future climatic conditions (Koncki & Aronson, 2015). The study predicted a rise in temperature, and an increase in spatial distribution of all three species in the northeastern United States in years 2040 and 2080.

The above-mentioned studies show the importance of using climate change data to predict species distributions of non-native plants. Even though incorporation of these types of data describing extreme weather events to predict future occurrences of such events is difficult to integrate on a broad scale level, such as the PPQ WRA, due to the vast biogeographical diversity of the U.S., these types of data could be integrated to state or regional WRAs, such as the California WRA.

Plant-soil-microbe feedback and species abundance for predicting weed risk

The paradox of invasion is that some non-native species thrive in new environments and become invasive, even though native species are historically adapted to their local environmental conditions (Sax & Brown, 2000). This paradox can pose a challenge to predicting the invasive potential of non-native species. Looking into the differences in plant-soil-microbe interactions in the native and invaded ranges could provide one explanation to the paradox of invasion and potentially be an asset to small scale WRAs.

Altered interactions of soil microbial communities in new ranges are one way invasive plants increase their abundance and outcompete natives for resources. For example, a study investigated the influence of feedback with soil organisms in determining plant abundance of five invasive (i.e. *Alliaria petiolata, Cirsium arvense, Euphorbia esula* L., *Lythrum salicaria* L., and *Polygonum cuspidatum* Sieb. & Zucc) and five rare plant species (i.e. *Agalinis gattingeri*, *Aletris farinosa*, *Gentiana alba*, *Liatris spicata*, and *Polygala incarnata* L.). Measurement of the relative growth of plants in their own soil and soil from another species showed positive soil feedback responses for the five invasive species and negative feedback responses for the five rare plant species (Klironomos, 2002). These feedback responses are important considerations that affect the abundance of invasive plants. Incorporation of data obtained from plant-soil-microbe feedback studies to small scale WRAs is an innovative way to predict non-native plant abundance for new regions in a case-by-case basis, but not experimentally feasible to incorporate to a broad scale level, such as the PPQ WRA.

In a similar vein of investigating invasive plant abundance in new regions, species richness could also be explored to elucidate the role invasive plants have on plant-soil-microbe feedback. The abundance of invasive plants is known to have negative effects in ecosystem structure and function (Neckles, 2015; Naeem *et al.*, 1994). One study looked at the effect of this feedback on species richness in field plots with and without the invasive plant *Chromolaena odorata.* Significant differences in plant species richness in these two plots were noted, where plots without *C. odorata* showed greater plant species richness than the plots with *C. odorata*. Additionally, shoot height was measured for *Amaranthus spinosus* and *Bambusa arundinacea* in order to investigate phenotypic differences of native plants grown in non-sterilized, sterilized, and soil with activated carbon for soil collected from the rhizopheres of *C. odorata* and native neighboring plants. Decreased shoot height for *A. spinosus* and *B. arundinacea* were seen in plants grown in soil collected from the rhizopheres of *C. odorata* with carbon activated and non-sterilized soil samples, but no difference was observed for plants grown in sterilized soil (Mangla *et al.*, 2008).

The study further explored the plant-soil-microbe feedback by counting the number of fungi *Fusarium semitectum* spores for all the soil samples. The greatest number of *F. semitectum* spores were seen in non-sterilized soil collected from the rhizopheres of *C. odorata.* This finding suggests that *C. odorata* inhibits the growth of surrounding native plants through an indirect negative feedback by accumulating high concentrations of *F. semitectum,* thus enhancing the fungal infection potential for the native plants (Mangla *et al.*, 2008). While conducting a study similar to the

one just mentioned is impractical from a time and economic perspective, it does provide direction to the types of data and questions ecologists should explore when developing WRAs. For example, the PPQ WRA does not address the plant-soil-microbe feedback of invasive plants, but small scale WRAs could be developed to incorporate this type of data as a measure to assess invasive risk potential.

Investigation of nutrient runoffs from agricultural systems is another factor to consider for assessing invasive potential. Agricultural practices are known to increase resource availability, and therefore effect the abundance and ecological performance of native and invasive species, thus altering the community composition (Boudell & Stromberg, 2015; Chen, He, & Qiang, 2013; Gustafson & Wang 2002; Lambert, Dudley, & Robbins, 2014). For example, reductions in growth of native shrubs were shown in the presence of mycorrhizae with high nitrogen soils compared to invasive grass (Sigüenza *et al.*, 2006). Modeling these types of data describing nutrient runoffs in relation to prediction of weed risk is not feasible for broad scale applicability because the amount of nutrient runoff from agricultural systems may vary from one farm to the next and the effects of nutrient runoff may vary from one plant to the next; However, these types of data could be used as a first step to understanding how variables like nutrient runoff from agricultural systems affect species distribution of non-native plants on a small scale WRA.

The studies mentioned above highlight the importance of investigating the plant-soil-microbe feedback for invasive and non-invasive plants and could prove to be useful for assessing weed risk. These effects and factors need to be taken into

account when developing a small-scale integrative weed risk assessment model to predict the invasive potential of non-native plants.

Genomics, transcriptomic, and proteomics for predicting weed risk

It is important to understand and categorize the genetic traits that are characteristic of an invasive species to help understand which genes contribute to invasiveness. Environmental adaptation of organisms can be investigated through the use of next generation sequencing, proteomics, and transcriptomic analyses.

Expressed sequence tags, for transcriptomic analysis, have been used to characterize gene transcript expression differences in *Senecio madagascariensis* collected in native and introduced ranges. Differential gene expression was observed for defensive responses to biotic stimuli in the native range compared to the introduced range, most likely due to lack of natural enemies in the introduced ranges (Prentis *et al.*, 2010).

Next generation sequencing and quantitative proteomic analyses identified specific genes and proteins important for rhizome differentiation, development and function in *Phragmites australis*, which is highly invasive due to clonal reproduction via rhizomes (He *et al.*, 2012). Clonal reproduction, in particular, could lead to dense plant growth and may indicate invasion success (Liu *et al.*, 2006). In fact, this type of reproduction is important for predicting invasive potential and is included as a variable of interest in the PPQ WRA (PPQ, 2016). Elucidation of specific genes and proteins important for vegetative growth could help to predict the invasive potential of non-native plants. Specifically, a predictive model could be developed to assess the

potential that specific genes have for promoting vegetative growth for non-native plants. This would include conducting experimental work for each prospective plant undergoing evaluation.

Approaches more modest than –omics can also be effective in identifying invasive risk associated with genome level analysis. For example, quantitative trait locus (QTL) mapping is used to identify molecular markers that correlate genotypes to phenotypes of interest, and was used to study the genetics of adaptive introgression following hybridization. Molecular analysis revealed that the stabilized *Helianthus annuus texanus* was formed from *H. annuus* and *H. debilis* spp. *cucumerifolius.* This new hybrid is able to thrive in a new edaphic niche previously not inhabited by either parental species (Whitney *et al.,* 2015). Increased genetic fitness as a result of hybridization may support invasion potential through competitive advantage. An increase in competitiveness has been shown in hybrid offspring compared to non-hybrid parental lineages (Pareapa *et al.,* 2014). This may be the result of increased genetic variation through hybridization, which may provide genes necessary for rapid environmental adaptation (Ellstrand & Schierenbeck, 2000).

Evaluation of polyploidy could be another genome level analysis used for weed risk predictions. Polyploidy, a result of hybridization, is known to contribute to plant invasion success (Pandit & White, 2014; Hull-Sanders *et al.,* 2009) and are 20% more likely to be invasive than diploids (Pandit *et al.,* 2011) due to increased heterozygosity (Soltis & Solits, 2000) and hybrid vigor (Ni *et al.,* 2009). Increased chromosome numbers can be acquired either through autopolyploidy or allopolyploidy (te Beest *et al.,* 2012). Regardless of whether intra- or interspecies

hybridization occurs, studies have shown that polyploid cytotypes proliferate in invaded ranges while their diploid counterparts are restricted to native ranges (Thébault *et al.*, 2011; Broennimann *et al.*, 2014; Hahn *et al.*, 2012). For example, polyploidy in the spotted knapweed, *Centaurea* stoebe has been shown to contribute to invasion success in the U.S., where predominately the introduced tetraploid spotted knapweed is invasive, but not their diploid counterparts (Treier *et al.,* 2009). Even though both cytotypes are found in their native European habitat, only the polyploid cytotype is invasive in the U.S. and this invasion success may be attributed to population level adaptation in a novel environment compared to the diploids (Treier *et al.,* 2009). In addition, polyploidy can mask the negative effects of deleterious recessive mutations (te Beest *et al.,* 2012; Sattler *et al.,* 2016) by having more genetic material for growth and adaptation. Further, successful long-distance dispersal, which is a key aspect of invasion success, are seen in polyploid lineages rather than diploid (Linder & Barker, 2014). Genomic attributes such as hybridization, resulting in autopolyploidy or allopolyploidy, should be considered as a risk factor for predicting the invasive potential of plants. Cytological investigations of plants undergoing assessment for weed risk should be conducted to determine its genomic attributes, whether the plant is diploid or polyploid, to potentially help to predict its invasive potential. Ploidy level as a risk factor, which is not addressed in the PPQ WRA, could potentially be added to such broad scale risk-assessments.

A machine-learning method for weed risk assessment

Development of WRAs using ecological data describing niche suitability, climate change factors, and plant-soil-microbe relationships, and molecular biology of invasive plants are not readily available due to time and economic constraints. In the absence of these types of data, non-traditional non-parametric statistical methods could be used to improve the robustness of weed risk assessments leveraging already available data to improve predictive accuracies. In this respect, Random forests, a machine-learning algorithm, could be used as an alternative to traditional parametric statistical methods for assessing weed risk.

Random forests have already been used for classifying land cover with multisource remote sensing and geographic data (Gislason *et al.*, 2006), predicting conifer species occurrence (Evans & Cushman, 2009), and predicting soil properties for soil mapping in Africa (Hengl *et al.*, 2015). Random forests are ensembles of uncorrelated decision trees that can be used for classification, regression, and cluster analysis tasks (Breiman, 2001). Decorrelation of trees can be attributed to the two-step randomization process of Random forests. The first step in the randomization process is to use a bootstrap sample from the original data to grow a tree and then, for the second step, a subset of predictor variables is randomly selected for splitting a node in a tree (Truong *et al.*, 2004). "Growing" a "forest" with many decorrelated trees and using either averaging for regression tasks or majority voting for classification tasks aggregates the results, thus reducing model variance (Breiman, 2001). Contrastingly, for the PPQ WRA, variance is assessed outside the logistic

regression predictive model by running a Monte Carlo simulation (PPQ, 2016) which computes invasive risk potential 5,000 times.

For classification, the best split at each node of a tree is determined by choosing the split that minimizes the Gini index impurity (Maindonald, 2010). At each node, the Random forests classification algorithm computes the Gini index of impurity and best separates the bootstrap sample into two groups (Maindonald, 2010). The parameter *m*try indicates the number of predictor variables used to split the node (Liaw & Weiner, 2002). The Random forests algorithm calculates internal estimates of generalization error, classifier strength, and dependence to compute the number of predictor variables (features) needed to split the node.

These estimates, which are collectively referred to as out-of-bag (OOB) error reduce model bias (Breiman, 2001), which is the model error that is introduced when "approximating a real-life problem" (James *et al.,* 2013). Bias reduction in WRAs is particularly important because WRAs are developed to predict the invasive potential of a non-native plant that has not been introduced to the new environment, but WRAs are developed with species with known life-history data (Hulme, 2012) and already present in the new region (Koop *et al.,* 2012). In order to reduce model bias, the PPQ WRA for example, was developed with two subsets of data, one for training the model and one for testing (Koop *et al.,* 2012). Additionally, only two parameters (the number of trees in the forest and number of variables selected at each node) need to be tuned for a random forest. The high predictive accuracies of Random forests compared to logistic regression (Peters *et al.*, 2007), coupled with its applicability to

high-dimensional datasets (Li & Zhao, 2009) makes it an ideal statistical method for assessment of weed risk.

Chapter 2: A random forest approach for predicting invasive status of non-native plants for weed risk assessment

Introduction

In invasion ecology and related scientific fields, such as weed science and conservation biology, an important task is the prediction of outcomes (i.e. categorical response variables; dependent variables; class) with respect to available predictors (i.e. independent variables). Predictive modeling uses statistical methods to compute predictions for outcomes of interest. There are two broad classes of predictive modeling: parametric and non-parametric. This work aims to compare the predictive performance of the parametric Generalized Linear Model (GLM)-logistic regression and the nonparametric random forest models for Weed Risk Assessment (WRA). Logistic regression is used for multidimensional problems that linear regression cannot fit. The equation below expresses the logistic regression model for *p* quantitative independent variables for binary response variable *Y* (Agresti, 2014):

$$\text{logit}[\pi(x_1, \ldots, x_p)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p \quad (2.1)$$

- $\pi(x)$ denotes the probability Y=1 at value x.
- 1-$\pi(x)$ denotes the probability Y=0 at value x.
- $\alpha + \beta_1 x_1 + \cdots + \beta_p x_p$ denotes the formula for the regression function.

Taking the inverse of logit[$\pi(x)$], gives the probability $\pi(x)$ as a Sigmoid-shaped function of x:

$$\pi(x) = \frac{e^{(\alpha+\beta_1 x_1+\cdots+\beta_p x_p)}}{1 + e^{(\alpha+\beta_1 x_1+\cdots+\beta_p x_p)}} \quad (2.2)$$

Logistic regression models are fit using maximum likelihood estimates to predict probabilities for responses in non-linear solutions. The coefficients are expressed in the logistic regression formula for predictor variables. The formula for logistic regression provides a good model fit for datasets that have "large *n*, small *p*". Random forests, an algorithmic approach, allow many predictors and are not fixed to a form of the equation prediction (Breiman, 2001). The algorithm below expresses Random Forests Classification (Hastie *et al.*, 2009):

1. For b=1 to B (where B is the total number of trees in the forest): (2.3)
    a. Draw bootstrap sample of the training data.
    b. Grow a random forest tree to T_b using the bootstrapped data:
        i. Randomly select m number of variables from the total number of variables p.
        ii. Select the best m variable to split the data.
        iii. Make two daughter nodes from the node.
        iv. Recursively repeat steps i-iii until for tree's terminal nodes until minimum node size is achieved.
2. Transfer the information to the ensemble of trees $\{T_b\}_1^B$.
3. Make a classification prediction:
    a. Take majority vote of the trees in the random forest.

Majority vote refers to proportion of trees in the forest that predicts for a particular response.

Screening tools, such as WRAs, are utilized for assessment of potential invasive status of non-native plants in novel ranges. One of the challenges in development of WRAs includes limitations in available data. It is recommended that logistic regression models follow the "large n, small p" rule of thumb in order to reduce model noise from overfitting the data. For example, some suggest a minimum of 10-15 observations per covariate for logistic regression models (Babyak, 2004). This rule can prove to be a problem in WRAs due to limitations in data availability. In general, traditional statistical methods, such as logistic regression, used in risk analysis for predicting the invasive status of non-native plants afford no solution to the "small n, large p" problem. As an alternative, a non-parametric statistical method, Random forests has proved useful under such constraints and provide a more appropriate predictive modeling approach.

Unlike logistic regression, Random forests do not need to follow the "large n, small p" rule in order to fit the model (Matsuki *et al.*, 2016). Due to this fact, Random forests modeling has been readily utilized in various fields such as bioinformatics and computational biology (Boulesteix *et al.*, 2012), for predicting civil war onset in political science (Muchlinski *et al.*, 2015), and for ecohydrological modeling of vegetation distribution (Peters *et al.*, 2007) and for predicting presence of invasive plant species in lava beds in ecology (Cutler *et al.*, 2007).

Effective WRAs should provide reliable conclusions even in instances of limited data (Keese *et al.*, 2014), thus this flexibility of having a "small n, large p" makes Random forests a good candidate for the development of WRAs. Moreover, Random forests are well established for use in high-dimensional data (Li & Zhao, 2009), where (oftentimes) the number of predictors exceeds the number of available observations. The PPQ WRA is informed by a high-dimensional data set involving 41 predictors. Additionally, Random forests provide the flexibility to choose between constructions of multiclass classification or regression trees, whereas traditional logistic regression is limited to only binary classifications. Multiclass models are particularly important for assessing weed risk where there are more than two classes, such as the PPQ WRA (PPQ, 2016).

Random forests fits bootstrap samples of the complete data into numerous decorrelated decision trees yielding distinct grouped classes (James *et al.*, 2013). This reduces the overall variance of the Random forests model (Breiman, 2001). In contrast, logistic regression uses log odds to make binary predictive outcomes (Agresti, 2014) and requires linear relationships between log odds and the dependent variable. Further, studies show that Random forests have lower classification errors and higher predictability accuracies compared to logistic regression counterparts (Peters *et al.*, 2007).

The PPQ WRA, the U.S. plant protection organization, uses a logistic regression model to predict weed risk of non-native plants considered for introduction. The main objective of this study is to compare the predictive performance between statistical methods logistic regression and random forest for the

PPQ WRA dataset. Other objectives include: examine the effect different sampling methods have on predictive performance of all models, analyze variable importance for the random forest model, and investigate the effects values of predictor variables have on the change in predictive value of the random forest model.

Methods

WRA Data

The four models presented in this work were developed with a priori classes of the 204 plant species indicated in Koop *et al.* (2012). The a priori categories include non-invaders, minor-invaders, and major-invaders with N=64 for each. These data contain no missing values and is class-balanced since each a priori category has an equal number of observations. Definitions of non-invaders, major-invaders, and minor-invaders were retrieved from Koop et al. (2012). Non-invaders are plants that are not naturalized but have occupied the United States for 75 years or more. Major-invaders are categorized with "I-rank" impact ranking of high or high-medium on NatureServe's categorization (NatureServe, 2009), or listed as "serious" or "principal" by Holm *et al.* (1979), or listed as "troublesome" by Bridges (1992). Minor-invaders are plants that are naturalized in the United States but do not fit the major-invader requirements. The plant species used in the study are presented in Appendix B. Some of the questions on the PPQ WRA questionnaire refer to the known invasive potential of plants, such as invasive status outside native range (es1), weed status in natural systems (impn6), and weed status in production systems (imp6). A majority of the questions refer to ecological traits of plants that are known

to contribute to "invasiveness", such as shade tolerance (es4), nitrogen fixing capability (es9), and minimum generation time (es13). Refer to Appendix A for the full questions of the establishment/spread and impact potential sections. Scores of the establishment/spread and impact potential sections in the USDA-APHIS-PPQ WRA were summed to synthesize models with only two predictor variables in an effort to reduce model noise. In this study, we used the raw scores of each variable instead of the summed scores of each section, thus giving 41 predictors in order to calculate variable importance for the random forest model.

Model development and statistical analysis

All models, figures, and statistical analysis were developed in the R environment (version 3.3.2) and RStudio (version 1.0.153) on a Macintosh OS X El Capitan computer with the codes referenced in Appendix C.

Model comparisons

In binary classification, only two events can be evaluated; therefore, a first comparison of non-invaders and invaders and a second comparison between minor and major invaders were developed for logistic regression and random forest models. For model comparison A, between non-invaders and invaders, non-invaders were dummy coded as event 0, while minor and major invaders were dummy coded as event 1. For model comparison B, between major-invaders and minor-invaders, major-invaders were dummy coded as 0 and minor-invaders were dummy coded as 1. The logistic regression and random forest classifiers for model comparison A and B

were developed using *k*-fold cross-validation (CV) sampling methods, which is an estimate of model predictive accuracy (Cutler, 2010). In *k*-fold CV, the dataset is split into *k* folds where $k - 1$ folds are used for model training while the remaining fold is used for model testing (Hastie *et al.*, 2009). This process is repeated until each observation is used for model training and testing. Each fold contains unique observations; this ensures that training and testing datasets are different. The average estimates of performance across all *k* trails of the testing data are computed. Classifiers for model comparisons were built with *10*-fold CV with 10 repetitions, which gives 100 total resamples that are averaged for their estimates of performance, thus reducing variance. The typical choice for *k*-fold CV is 5 or 10 (Hashtie *et* al., 2009), but if the dataset is small then the *k* value needs to be larger. The predictive performances of each model were assessed for sampling folds: *k=2*, *5*, and *10*-fold CV. This validation step in model development is critical for evaluating the predictive performance of the models. Utilization of the training dataset for model evaluation can potentially cause model overfitting, which occurs when a model follows errors (noise) instead of the underlying relationships between the predictor variables (independent variable) and outcome (dependent variable) (Brownlee, 2016; Mainali *et al.*, 2015). A model that is overfit is less likely to compute accurate predictive estimates on new observations (cases not used in model development) (James *et al.*, 2013).

Random forest classifier

A separate random forest classifier was developed with all three classes of invaders, i.e. non-invaders, minor-invaders, and major-invaders. The original dataset was randomly and proportionally split into a model development set (70%) and an independent validation set (30%) (Appendix B). Out-of-bag (OOB) sampling, a feature unique to Random forests and not used in logistic regression, was used for model development with the 70% dataset, instead of *k*-fold CV. An OOB sample refers to observations that were not used to construct a random forest tree. Each tree in the random forest is constructed with a random subset of the original dataset and the OOB sample is used to assess the predictive ability of the tree. OOB error is used for downstream calculations for assessment of model variable importance and partial dependence. The forest was grown to 1000 trees (*n*tree=1000) and 6 variables were used to split the nodes of the trees (*m*try=6). For classification Random forests *m*try is the square root of the number of predictors. Theses value of the *n*tree and *m*try parameters yielded the lowest OOB error. The independent testing set (30%) was used for external validation of model predictive accuracy.

ROC plots

The predictive power of models is represented through the Area Under the Curve (AUC) values from Receiver Operating Characteristic (ROC) plots (Fawcett, 2006). The ROC-based metric can be used to calculate the predictive performance of binary classification models. This test uses evaluation datasets to calculate true positive rate (TPR; model sensitivity) and false positive rate (FPR; model specificity) over a range

of potential test thresholds of predictive models. The TPR indicates the proportion of instances that were correctly predicted to be in a particular class. The FPR indicates the proportion of instances that were incorrectly predicted to be of a certain class. In this case, the predicted class would be the species observed invasive status. TPR and FPR are used to calculate the AUC of a ROC plot. A good screening tool is characterized by an AUC that maximizes test sensitivity with minimal values of false positive predictions. For WRAs, a high proportion of potential invasive plants need to be rejected and non-invasive plants need to be accepted (Caley & Kuhnert, 2006). An AUC value of 1.0 is the greatest value for ROC plots. This value indicates that the predictive model is a perfect classifier. AUC values of 0.5 indicate that the predicted model outcome is purely random; therefore, the model did not learn to distinguish between the classes from the training dataset.

Variable importance plot—mean decrease accuracy

This method of calculating mean decrease accuracy is unique to Random forests and used to calculate the importance of variables (features) for prediction of outcomes. Variable importance was calculated for the Random forests classification task. In this study, variable importance was calculated for the 41-predictor variables (Appendix A) used in model development for predicting the invasive status of plant species for WRA. This methodology computes the importance of each variable X_j in classifying the data using mean decrease accuracy. The mean decrease in accuracy of X_j is calculated by permutation of the X_j values in the OOB samples, and then averaged across all the trees in the forest. In a random forest, the OOB samples are not used in

tree construction, but instead used in calculating prediction errors, which are known as OOB error values. These are the same calculations used to evaluate variable importance for each classification tree in a random forest (Genuer *et al.*, 2010). If model mean accuracy decreases when a variable is omitted, then that variable is deemed important for accurately classifying the data. The equation below expresses the calculation of mean decrease accuracy for variable importance (Han *et al.*, 2017):

$$VI_j = \frac{1}{ntree} \sum_{t=i}^{ntree} (EP_{tj} - E_{tj}) \tag{2.4}$$

- VI_j denotes variable importance of predictor variable *x*.
- *ntree* denotes the number of trees in the random forest.
- EP_{tj} denotes the OOB error on tree *t* after permuting values for predictor variable X_j.
- E_{tj} denotes the OOB error on tree *t* before permuting values for predictor variable X_j.

Partial dependence plots

Partial dependence plots graphically show the marginal effect variable values have on model predictions when all other variables in the model are held at their mean. Specifically, partial dependence plots visualize the delta log-odds for a particular class's sample probability of classification of "invasive status" (*y*-axis) as a function of a particular variable of interest (*x*-axis). The *y*-axis of a partial dependence plot shows how the predicted value changes in response to the value of the variable. Negative *y* values indicate that the specific variable value is less likely to be predictive for that particular class, whereas positive *y* values indicate the opposite.

The *x*-axis shows the range of values for the predictor variable. The equations expressed below mathematically define partial dependence plots (Liaw, 2015; Friedman, 2001):

$$\tilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, x_{iC}) \tag{2.5}$$

- *x* denotes the variable of interest for calculating partial dependence.
- x_{iC} denotes all the other variables in the dataset.
- The summand in (2.5) is the logit of the estimated probability of classification of "invasive status" as a function of a particular variable of interest.

When there are K classes, *f(x)* is defined as follows:

$$f(x) = \log[p_k(x)] - \frac{1}{K}\sum_{j=1}^{K} \log[p_j(x)] \tag{2.6}$$

- *K* denotes the number of classes in the model.
- *k* denotes the class of interest.
- p_j is the fraction of votes for class *j* in the classification model.

Results

Greater predictive accuracy in random forest classifiers

After developing the logistic regression and random forest classifiers for model comparison A (non-invader vs. invaders) and B (minor-invader vs. major-invader) with the PPQ WRA dataset, their ROC AUC values were compared in order to quantify classifier predictive accuracies. In model comparisons A and B, the random forest classifiers had higher AUC values compared to the logistic regression classifiers. For model comparison A, non-invaders versus invaders, the AUC values are 0.865 for logistic regression and 0.943 for random forest (Figure 1A). For model comparison B, minor-invaders versus major-invaders, the AUC values are 0.723 for logistic regression and 0.885 for random forest (Figure 1B). The significance of these differences show that the random forest classifiers have greater predictive accuracy than the logistic regression classifiers.

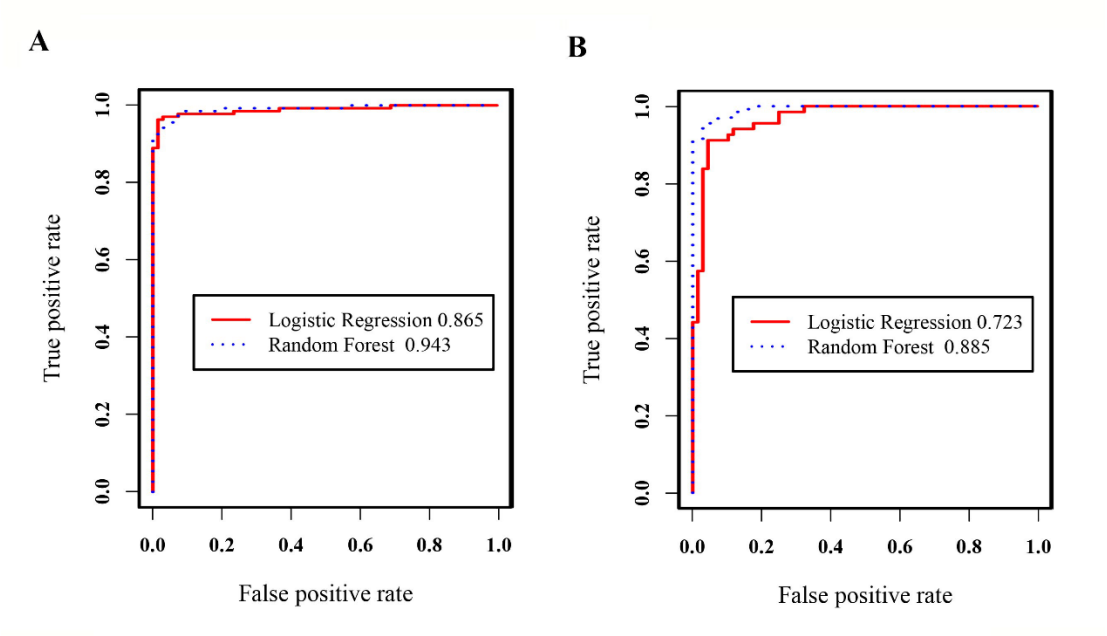

Figure 1. ROC curves for the logistic regression and random forest classifiers for model comparisons A (non-invader vs. invaders) and B (minor-invader vs. major-invader).

(*A)* For model comparison A, non-invaders versus invaders, the AUC value for random forest was greater than logistic regression in relation to the observed "invasive status". *(B)* For model comparison B, minor-invaders versus major-invaders, the AUC value for random forest was greater than logistic regression in relation to the observed "invasive status". Sampling *10*-fold CV with 10 repeats was used to generate ROC curves for all classifiers in both model comparisons. Random forest classifiers were grown to *n*tree = 1000 and *m*try=6.

Greater predictive accuracy in random forest classifiers in all *k*-fold CV

To assess model predictive performance for various sampling methods, additional models were developed for *2* and *5*-fold CV (see Methods: *I. Model* comparisons). The random forest classifiers (i.e. RF model A for non-invader vs. invaders and B for minor-invader vs. major-invader) showed greater ROC AUC values than logistic regression classifiers (LR model A for non-invader vs. invaders and B for minor-invader vs. major-invader and B) for *2*, *5*, and *10*-fold CV sampling (Figure 2). Random forest classifier A (RF model A for non-invader vs. invaders) has the highest AUC values and logistic regression classifier B (LR model B minor-invader vs. major-invader) has the lowest AUC value for all sampling methods. Using sampling method *10*-fold CV yielded the highest AUC for the logistic regression classifiers (LR model A (non-invader vs. invaders)= 0.865; LR model B (minor-invader vs. major-invader)=0.723) and *5*-fold CV yielded the highest AUC for the random forest classifiers (RF model A (non-invader vs. invaders)=0.943; RF model B (minor-invader vs. major-invader)=0.862), whereas sampling method *2*-fold CV yielded the lowest AUC for all classifiers (LR model A (non-invader vs. invaders)=0.732; LR model B (minor-invader vs. major-invader =0.638; RF model A (non-invader vs. invaders)=0.862; RF model B (minor-invader vs. major-invader =0.870). The random forest classifiers had less variability in AUC values across the three sampling methods than the logistic regression classifiers.

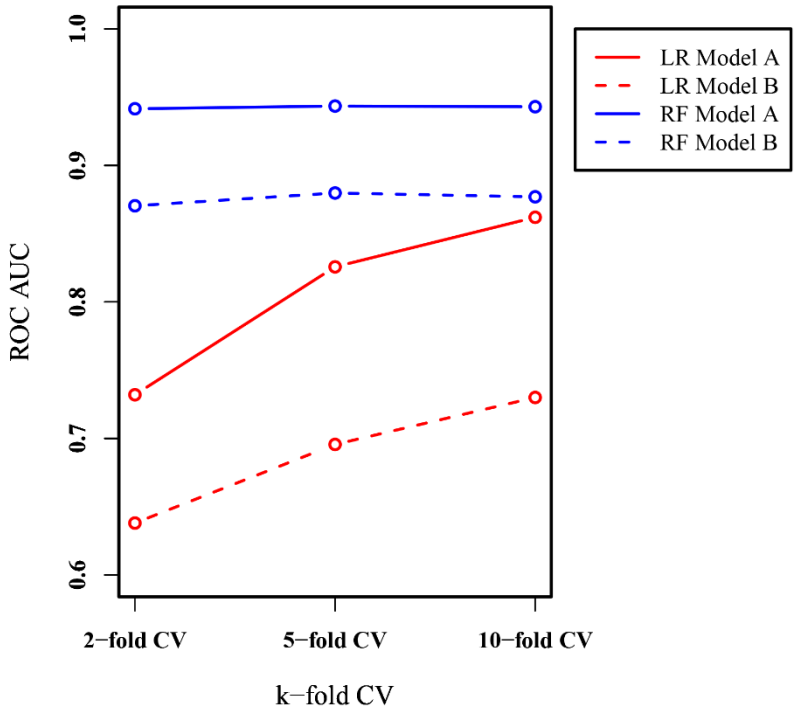

Figure 2. Classifier ROC AUC values for *2*, *5*, and *10*-fold CV for model comparison A (non-invader vs. invaders) and B (minor-invader vs. major-invader).

For model comparison A, non-invader vs. invaders, the AUC values for logistic regression are 0.732, 0.826, and 0.862 and 0.942, 0.944, and 0.943 for random forest using *k*=2, 5, and *10*-fold CV with 10 repeats, respectively. For model comparison B, minor-invader vs. major-invader, the AUC values for logistic regression are 0.638, 0.696, and 0.730 and 0.870, 0.879, and 0.877 for random forest using *k=2*, *5*, and *10*-fold CV with 10 repeats, respectively. The random forest classifiers have the highest ROC AUC values across all sampling methods for model comparison A and B. Solid lines in the figure indicate model comparison A (non-invader vs. invaders) and

dashed lines indicate model comparison B (minor-invader vs. major-invader). Random forest classifiers were grown to *n*tree = 1000.

Important variables for predictive accuracy in the random forest classifier

"Invasive status outside native range" (es1), "Weed status in natural systems" (impn6), and "Weed status in production systems" (impp6) had the highest mean decrease accuracy, 22.92, 20.74, and 20.14, respectively (Table 1), for the top ten important variables for predictive accuracy for random forest classifier (Figure 3). The top three important variables "Invasive status outside native range" (es1), "Weed status in natural systems" (impn6), and "Weed status in production systems" (impp6) occurred numerous times in the random forest (Table 1). The top three variables are representative of known invasive behavior of the taxon, while "Number of natural dispersal vectors" (es17) was the fourth most importance variable and the only variable out of the top four that related to biological characteristics of the plant taxon. "Climbing or smothering growth form" (es5) and "Minimum generation time" (es13) are the remaining top ten variables that were related to biological characteristics of the plant taxon. "Change in ecosystem processes and parameters that affect other species" (impn1) was the tenth most important variable, with the lowest mean decrease accuracy of 12.27, out of a total of 41 predictor variables (Figure 3). Impn1 occurred 518 times in the random forest (Appendix D). Variables "Is the species highly domesticated" (es2) and "Parasitic" (impg2), with mean decrease accuracy values of 0, ranked the least important variables for the classifier and occurred 25 and 38 times in the random forest, respectively (Appendix D). A gap of 3.02 in mean decrease accuracy values was present between "Number of natural dispersal vectors" (es17) and "Lowers commodity value" (impp2). In variable importance plots

assessing mean decrease accuracy, gaps between variables can be used as a break point for variable selection for further model development.

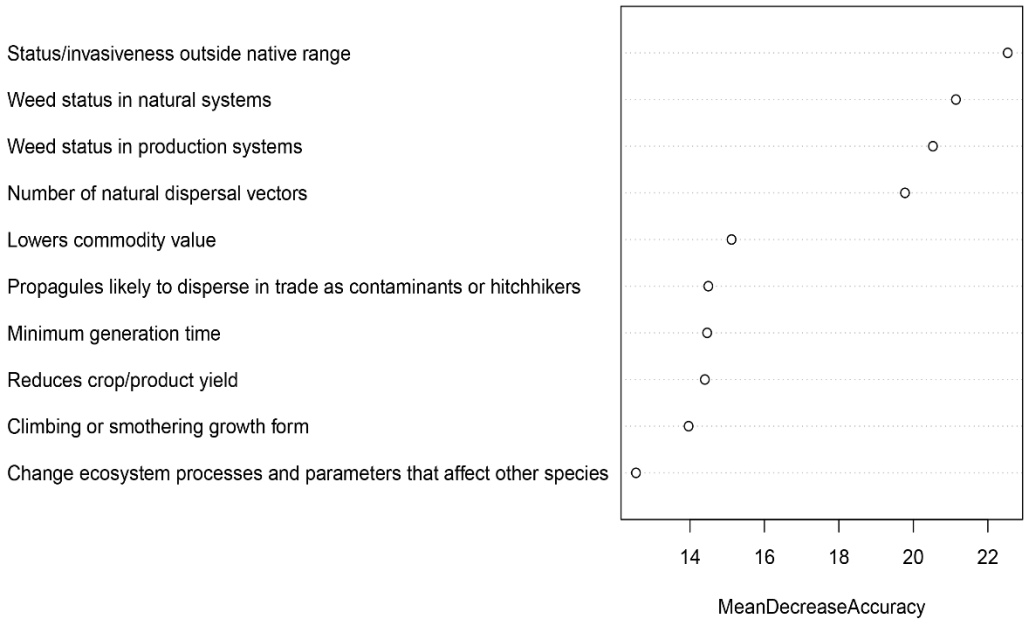

Figure 3. Variable importance by mean decrease accuracy for the random forest classifier.

A separate random forest model with all classes (non-invader, minor-invader, and major-invader) was constructed for variable importance calculations. Variables "Invasive status outside native range" (es1), "Weed status in natural systems" (impn6), and "Weed status in production systems" (impp6) have the highest mean decrease accuracy values for the random forest classifier. The mean decrease accuracy for es1, impn6, and impp6, are 22.92, 20.74, and 20.14, respectively. The most important variables for classifying the data are presented at the top-right of the variable importance plot, whereas variables of lesser importance are at the bottom-left. Refer to Appendix A. for more detailed descriptions of variables presented in this figure. The random forest classifier was grown to *n*tree = 1000.

Table 1. Top three important variables for each class in the random forest model

	Non-invader	Minor-invader	Major-invader	Mean Decrease Accuracy	# of times variable occurred in the random forest
Es1	20.90	-2.08	19.38	22.92	2267
Impn6	21.05	5.38	7.57	20.74	1452
Impp6	19.89	-5.68	16.79	20.14	1237

Random forest better at classification of non-invader and major-invader than minor-invader

The partial dependence plots show the change in predicted values of classification in response to predictor values (Friedman, 2001). Invasive status partial dependence was investigated for the top 3 important variables for the random forest classifier: "Invasive status outside native range" (es1), "Weed status in natural systems" (impn6), and "Weed status in production systems" (impp6) (Figure 4). The *y*-axis of a partial dependence plot shows the change in the predicted value in response to variable value (*x*-axis). A lower *y* value indicates that the variable value is less likely to predict for a particular class, whereas a greater *y* value indicates a higher likelihood (Machado *et al.,* 2015). In this study, the observed invasive status (i.e. non-, minor-, or major-invader) is the predicted classes. The random forest model predicted non-invader when es1 values were <2, major-invader when the value was >2, and minor-invader for all es1 values (-5 to 5) but more strongly predicted when values were <2 (Figure 4A). The random forest model predicted non-invader when impn6 values were < 0.2, major-invader when values were > 0.2, minor-invader for all impn6 values (0 to 0.6) (Figure 4B). The random forest model predicted non-invader when impp6 values were < 0.2, major-invader when values were > 0.2, minor-invader for all impp6 values (0 to 0.6) (Figure 4C). The partial dependence plots show that the random forest model predicted observed invasive status of major-invader when variable values were above a particular value, but non-invader and minor-invader were predicted for the same variable values. Random forest model could distinguish between non-invader and major-invader based on variable values, but cannot

distinguish minor-invader from the provided variable values. These results are consistent with the confusion matrix predictive accuracies for the random forest model. Confusion matrices are estimates of predictive accuracy that return detailed values for predicted and reference classes instead of just giving a general estimate of predictive accuracy (Fawcett, 2005). The confusion matrix showed that minor-invader predictions had low accuracy while high predictive accuracies were seen for non-invaders and major-invaders for the random forest classifier.

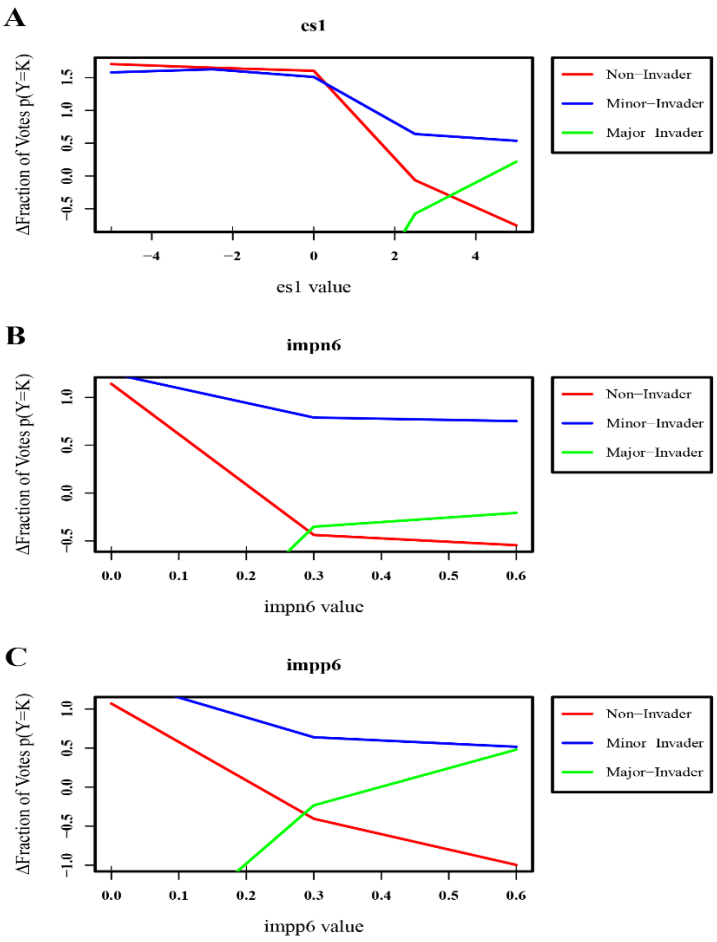

Figure 4. Partial dependence of top three important variables for the random forest classifier.

Variables with the highest mean decrease accuracy values were used to investigate the marginal effect raw values of variables have on model predictions for invasive status class: non-invader, minor-invader, and major-invader. *(A)* "Invasive status outside native range" (es1). *(B)* "Weed status in natural systems" (impn6). *(C)* "Weed status in production systems" (impp6). The *y*-axis shows the change in predictive value for invasive status (Δfraction of votes for class *K*) for an independent variable of interest, while all other variable predictions are held at their mean. The *x*-axis shows the corresponding variable values.

Table 2. Classification performance of Radom forest model for predictor variables. Confusion matrix for the random forest model validated, using plant species from the 30% original data subset, with all predictor variables (n=41). Random forest model is better at predicting for non-invader and major-invader than predicting for minor-invader.

		Reference		
		Non-invader	Minor-invader	Major-invader
Predicted	Non-invader	15	3	1
	Minor-invader	3	6	3
	Major-invader	0	7	18

Note: *Random forest model correctly predicted non-invader 15/18 times, minor-invader 3/16 times, and major-invader 18/22 times.

Discussion

WRA evaluations for model comparisons

In this study, logistic regression and random forest models were developed and then compared for predictive accuracy for assessing invasive status of non-native plants. The models are limited to the assessment of only non-native plants because the models were trained and validated with non-native plants that are known to have invasive "behavior" in the U.S. (PPQ, 2016). Both models have high predictive accuracies, with AUC scores exceeding 0.5, indicating non-random predictions, although, random forest classifiers resulted in higher predictive accuracies than the logistic regression classifiers. Applications of random forests in other studies also show improvements in predictive accuracies compared to logistic regression models (Cutler *et al.*, 2007; Muchlinski *et al.*, 2015; Peters *et al.*, 2007). Both logistic regression and random forest classifiers are better at differentiating the invasive status between non-invaders and invaders (model comparison A), than between minor-invaders and major-invaders (model comparison B). In previous studies, using the Australian WRA in Florida, a greater number of minor-invaders in the dataset needed to be further evaluated for weed risk compared to major-invaders (Gordon *et al.*, 2008a). In the PPQ WRA, minor-invaders have been shown to have a larger variability in invasive status compared to non-invaders and major-invaders and needed further evaluation (Koop *et al.*, 2012). This may relate to the gradients of invasion in the stages of the invasion process.

The dataset used to develop the USDA-APHIS-PPQ WRA logistic regression model was also used to develop the random forest model. However, PPQ-WRA was

developed using summed "risk scores" of the sections establishment/spread potential and impact potential, while the raw values of each variable in the establishment/spread and impact potential sections were used to develop the classifiers presented in the study. When comparing the predictive accuracy of the classifiers for model comparison A for *10*-fold CV, non-invader vs. invaders, the PPQ WRA (AUC=0.953) showed greater predictive accuracy. However, when comparing AUC scores with Koop *et al.*, the differences were minimal between random forest classifier A (AUC=0.943) developed in this study and the PPQ WRA (AUC=0.953), whereas the logistic regression classifier A (AUC=0.865), developed in this study, performed poorly compared to the PPQ WRA (Koop *et al.,* 2012). Overall, random forest classifiers showed greater predictive power than the logistic regression classifiers that were developed in this study and are an improvement to the PPQ WRA in the sense that all the variables of the PPQ WRA are considered in the analysis.

When comparing the predictive accuracy of all the classifiers developed for the model comparisons (between non-invader vs. invaders and minor-invader vs. major-invader) across *2, 5*, and *10*-fold CV methods, the results show that the random forest classifiers have greater predictive accuracies than the logistic regression classifiers across all CV methods. The *k*-fold CV sampling is indicative of how the dataset is split for training and testing and is an estimate for predictive accuracy of prediction of observed "invasive status", whether the taxon is a non-invader, minor-invader, or major-invader. The greatest predictive accuracy is shown using *5*-fold CV for the random forest classifiers, while the *10*-fold CV yields the best predictive

accuracy for the logistic regression classifiers. These results indicate that random forest classifiers need a lower ratio of training data compared to the logistic regression classifiers. That is, random forest classifiers are “better learners” than the logistic regression classifiers even in instances of limited number of observations in the dataset available for training the model (Figure 2). This is particularly important for assessing weed risk because of the limitation in data availability because alien species data tend to be scattered through disconnected data silos lacking interoperability (Quentin *et al*., 2017). Further, the performances of the random forest classifiers are more stable across the different sampling methods than the logistic regression classifiers (Figure 2)

Variable importance for the random forest classifier

The raw risk scores of the establishment/spread and impact potential sections of the data were used to develop the classifiers presented in this study; however, the PPQ WRA is developed with summed risk scores of the two sections which is useful in logistic regression for reduction of model overftting that occurs in ‘small n, large p’ dataset scenarios and masks the presence of multicollinearity within the dataset. Using raw scores in the random forest models is beneficial for assessing the model’s variable importance, whereas using summed risk scores results in only two independent variables available for assessing variable importance. By developing a separate multi-class (non-, minor-, and major-invader) random forest model, this study shows the importance of variables for predictor accuracy for the random forest model.

The best predictor of invasive status for the random forest classifier is es1, which quantifies the invasive status of the plant outside the native range. This result is consistent with the question analysis, using chi-square tests, conducted by Koop *et al.* prior to final model development, which found that the response to invasive status elsewhere has the greatest association with a priori invasive status for minor and major-invaders (2012). Additionally, other previous studies also indicate that the invasive status of plants elsewhere strongly predicts for invasiveness (Gordon *et al.*, 2008b; Dawson *et al.*, 2009; Herron *et al.*, 2007). Although, the invasive status of a non-native plant elsewhere is rated as the best predictor for assessing invasive potential in a new range, it is important to consider that non-native plants may respond differently to environmental factors at different spatio-temporal stages of the invasion process.

Other predictors such as, impn6 (weed status in natural systems) and impp6 (weed status in production systems) also have greater variable importance for the random forest classifier. This is consistent in industrial practice, where research develops new taxa for agricultural and horticultural purposes with characteristics that promote production which consequently also increase environmental weed risk (Driscoll *et al.*, 2014). Two predictors, es2 (is the species highly domesticated) and impg2 (parasitic) have no importance in predicting invasive status for the random forest classifier but this could be due to a low *n* of parasitic taxon in the dataset used to develop the model. Interestingly, a study found that native parasitic plants cause more damage to invasive hosts than the native, non-invasive hosts (Li *et al.*, 2012).

While a later study found that as the age of the invasive plant increases, there is a decrease in parasitic damage to the invasive plant (Li *et al.,* 2015).

Variables describing the biological traits of the taxon in regard to growth form (i.e. "whether it has climbing or smothering growth habit"; es5), "minimum generation time" (es13), "number of natural dispersal vectors" (es17), and "propagules likely to disperse in trade as contaminants or hitchhikers" (es16) are variables that are also important for predictive accuracy of the random forest model (Figure 3). This differs from a previous study where only a few questions regarding biological traits were found to be significant in chi-square tests for the PPQ WRA (Koop *et al.,* 2012). In general, variables describing biological traits of potential invasive plants are expected to be of significance in WRA because weed status elsewhere is most likely not known for plants undergoing evaluation.

Overall, assessing variable importance for the random forest classifier provides relevant information when looking for subsets of variables to use for WRA. Specifically, this provides insight into which variables are of the highest importance for invasive status prediction and which variables are of the least importance. Unimportant variables have low mean decrease accuracy, which means they have little to no effect on model accuracy. Moreover, variables of least importance may potentially be unreliable variables that could hinder model performance by introducing noise (Han *et al.*, 2017). An abbreviated model containing only variables with the greatest importance could potentially be developed in an effort to strengthen model predictive performance and to supplement the existing WRA process. When multiple predictive models yield similar predictions, this can further enhance the

confidence of the perceived outcome, which is valuable in WRAs to help inform policy-making.

Influence of predictor variables on the observed invasive status for the random forest classifier

Partial dependence functions can be used for interpretation of models produced with a "black box" prediction method such as Random forests (Friedman, 2001). In Ecology, partial dependence functions have been used to interpret the influence environmental variables has on presence of short-finned eels for a Boosted Regression Tree (BRT) model (Elith *et al.,* 2008). In this study, partial dependence plots were used to further investigate the relationship between prediction accuracy for the observed invasive classes (non-, minor-, and major-invader) for the top three important variables (es1, impn6, and impp6) of the random forest model. The results show that the marginal effect that variable values have on the change of class prediction can only be seen for observed invasive status non-invader and major-invader. That is, for the top three variables, the random forest classifier is best at classification of non-invaders and major-invaders than minor-invaders based on the variable value. Overall, none of the top three variables show strongly non-linear relationships between variable value and change in predicted value. If strong non-linear relationships were present then Generalized Linear Models (GLMs) like logistic regression would be unsuitable for prediction of invasive status of non-native plants for WRA. This result supports the conclusion that the logistic regression classifiers also show high predictive accuracy, even though greatest predictive accuracy is seen with random forest classifiers.

Closing remarks

Broadly, the conclusions presented here indicate that the prediction of invasive status using the PPQ WRA dataset is improved when using a Random forests model compared to predictions using the generalized linear model logistic regression. This was consistent across ROC AUC analysis under multiple *k*-fold conditions. Use of random forest methodology allowed additional analysis of variable importance with respect to predictive power and partial dependence of classification on variable values.

Partial dependence plots indicates the difficulty of predicting minor-invaders in WRAs apart from the non-invader and major-invader counterparts. This could be suggestive of room for improvement in the classification scheme for future analysis, it could be a consequence of the limitations of prediction of minor-invaders given the data and the model, or some combination thereof. On the one hand, while assignment of minor invader status is rooted in the literature, multiple of the pooled sources for defining the invasive status were ultimately subjective in their own classifications. This is indicative of the discretion involved in making such distinctions and of the care required when interpreting such analysis. Additional analysis of other variables ranked less important could also shed light on this issue. If the poor predictive accuracy for minor invaders is persistent across the vast majority of the most important variables, then this would be suggestive of the complexity involved in making such a prediction.

Of additional interest is the analysis of variable importance. One underlying assumption of a WRA approach that utilizes sets of traits associated with past

invaders is that said traits are reasonable grounds from which to predict future weed risk. This is a practical first order approach. Future efforts should strive to grapple with the complex temporal and ecological spatial relationships associated with potential invasive species relative to possible recipient communities across multiple geographic and time zones. Closer consideration of items such as the biological traits of the greatest importance determined here and integrating how these and other biological traits will influence future weed risk may be useful in efforts to improve future WRA implementations.

Chapter 3: An exploratory approach to invasive plant species distribution modeling for weed risk assessment

Introduction

Assessment of geographic potential of non-native plants for new ranges is integral to the WRA process. The USDA-APHIS-PPQ WRA *process* includes assessment of the geographic potential of non-native plants undergoing weed risk evaluation (PPQ, 2016). This portion of the WRA process is not part of the PPQ WRA model, but informs the overall report compiled for each plant taxon under WRA (PPQ, 2016). Geographic/climate suitability risk for regional establishment of species is assessed through climate variable matching, but does not assess species establishment from a climate change perspective. Future changes in climate are expected to promote weeds by increasing their negative impacts (Bradley *et al.,* 2010). Tools modeling plant species distributions (e.g. "MaxEnt", CLIMEX) are available to map current distributions and future distributions that incorporate climate change predictions. These models work under the assumption that climate variables are the primary risk factors for assessment of species potential ranges in the future.

This chapter of the thesis explores new avenues for modeling species distributions with respect to climate change predictions using open source software not currently utilized in potential species distribution of non-native plants. Distribution of *Alternanthera philoxeroides* (alligatorweed), *Acanthospermum australe* (paraguayan starbur), and *Abutilon megapotamicum* (trailing abutilon) in the

U.S. is visualized by mapping occurrence points to an overlay of climatic variables (i.e. mean monthly historical temperatures). These plants were part of the random forest validation dataset described in chapter two.

Methods

Climate data

Historical temperature data

The rWBclimate software package (Hart, 2014) was used to access historical temperature climate data from the World Bank Climate Change Knowledge Portal (CCKP). The Climatic Research Unit (CRU) of the University of East Anglia (UEA) originally produced the historical temperature data. Decade level temperatures for years 1990-2000 were retrieved at river basin spatial scales for the 78 major watersheds in the United States (excluding Hawaii and Alaska) and Mexico. A global map of river basins, with their respective basin IDs can be viewed at the CCKP website (http://sdwebx.worldbank.org/climateportal/index.cfm?page=basin_map_region&ThisMap=NA&ThisView=basin). KML files, a file format for the display of geographic data, were retrieved for mapping the historical temperatures of each river basin. KML files can also be retrieved using ISO country codes, but they result in lower map resolutions compared to river basins.

Climate change model

The rWBclimate software package (Hart, 2014) was used to access climate data from the World Bank Climate Change Knowledge Portal (CCKP) in order to compare modeled estimates of temperature with recorded historical temperatures. Climate data was downloaded from the General Circulation Model (GCM) Hadley Centre Coupled Model version 3 (HadCM3). The original historical climate data used by HadCM3 was provided by the CRU of UEA. HadCM3 is used by the North American Regional Climate Change Assessment Program (NARCCAP), Plant Species and Climate Profile Predictions, WorldClim, and the Intergovernmental Panel on Climate Change (IPCC) Third, Fourth, and Fifth Assessments (IPCC, 2001; IPCC, 2007; IPCC, 2014). The HadCM3, from the UK Met Office, is a coupled climate prediction model that has an atmospheric and oceanic component. There are 19 different categories to the atmospheric component, which includes a horizontal resolution of 2.5° latitude by 3.75° longitude, thus producing a global grid of 96 x 73 grid cells ("Met Office climate prediction model: HadCM3, 2018).

The HadCM3 model can forecast future climate predictions as well as backcast for 20-year intervals, starting from year 1920 to 2099 for two Green House Gas (GHG) emissions scenarios (a2 or b1). Both climate predictions are based on model estimates, not observed data. The a2 scenario is the future rate of GHG emissions remaining the same as the present, in a world that is regionally heterogeneous oriented in economic development with a continuous increase in

human population (Cubasch *et al.,* 2001). The optimistic b1 scenario is characterized by a decrease in rate of GHG emissions compared to the current rate with the introduction of clean technologies in a convergent world with improved equity and a continuous increase in human population (Cubasch *et al.,* 2001). Model data and detailed scenario descriptions are available at the International Panel on Climate Change Data Distribution Center (http://www.ipcc-data.org/). Climate model estimates of average monthly temperatures for GHG emission scenarios a2 and b1 were retrieved at country spatial scales for the United States for years 2080-2100 to ensure max change in the future years. These modeled estimates were compared to average monthly historical temperatures from years 1901-2009.

Species occurrence data

The spocc software package (Chamberlin, 2017) was used to retrieve species occurrence data for three plant species. They included, the aquatic/terrestrial plant *A. philoxeroides* (major-invader; family: Amaranthaceae), terrestrial plant *A. australe (*minor-invader; family: Asteraceae), and terrestrial plant *A. megapotamicum* (non-invader; family: Malvaceae). These plants were chosen because they are part of the data subset used to validate the random forest model (see methods on random forest classifier), they represent a wide variety of plant families, and each species represents a different observed invasive status in the PPQ WRA dataset. The spocc package was used to retrieve 500 occurrence records with coordinates for each species from the Global Biodiversity Information Facility (GBIF). Only 500 occurrence records were

retrieved for each plant to ensure feasible retrieval time. This returned 87 global occurrences with coordinates for *A. megapotamicum,* 500 records for *A. philoxeroides* out of a total of 2,729, and 500 records for *A. australe* out of a total of 972. GBIF is a network of over 200 million primary plant species occurrence data published by scientists all over the world, including 1,163 publishing institutions (GBIF; www.gbif.org). It is important to keep in mind that a lack of complete records for plant occurrences persists. However, for this work, the assumption was that the density of the data registered with the GBIF network is representative of species occurrence density gradientsEach record follows the Darwin Core Standard (DwC), which is a body of data standards used by GBIF (Wieczorek *et al.,* 2012). Along with coordinates from where the species was found, each occurrence record includes meta-data such as the event of the record (whether it was a human observation or a preserved specimen), location, geological context, occurrence, taxon, and identification (Wieczorek *et al.,* 2012).

Plant species distributions were visualized with respect to historical temperatures by mapping occurrence points to their recorded coordinates. The randomly selected occurrence points with coordinates were further modified to remove records with coordinates that were < 7 ° latitude (below the continental U.S. and Mexico). The remaining occurrence records were overlaid, with their corresponding coordinates, on a map of the continental United States and Mexico. Historical temperature (mean monthly temperature data from years 1990-2000) from the basin climate map was extracted for each remaining number of species occurrence point. Climatic variables, such as mean annual temperature are one of the most

important variables for predicting invasion success (Bellard *et al.*, 2016). Out of the original occurrence dataset *A. megapotamicum* (n=5), *A. australe* (n=11), and *A. philoxeroides* (n=158) occurrence points were overlaid onto the map of historical temperatures.

Results

HadCM3 climate model projects an increase in average monthly temperatures in the United States for years 2080-2100 than historical average temperatures

The observed historical temperature range for *A. megapotamicum* (trailing abutilon) is 26.33-29.66 ºC with an average of 26.99 ºC (Sd. 1.488) and 18.60 º latitude (Sd. 2.81). The trailing abutilon occurrence records are found in river basins located in southeast Mexico and the San Juan river basin and one occurrence in the U.S. Pacific Northwest. The observed historical temperature range for *A. australe* (paraguayan starbur) is 17.64-26.25 ºC with an average of 23.09 ºC (Sd. 2.84) and 31.31 º latitude (Sd. 3.22). The observed historical temperature range of *A. philoxeroides* (alligatorweed) is 16.19-26.25 ºC with an average of 23.55 ºC (Sd.1.49) and 31.57 º latitude (Sd. 1.88) (Figure 5). Majority of the recorded occurrences for alligatorweed were found in the Texas Gulf Coast river basin, while the rest were found in the South-Atlantic Gulf basin along with most of the recorded occurrences for paraguayan starbur. The climate model HadCM3 was used to forecast average monthly temperatures in future GHG scenarios a2 (pessimistic) and b1 (optimistic),

for years 2080-2100. The model predicted that temperatures would be higher than the reported averaged monthly historical temperatures for years 1901-2009 (Figure 6).

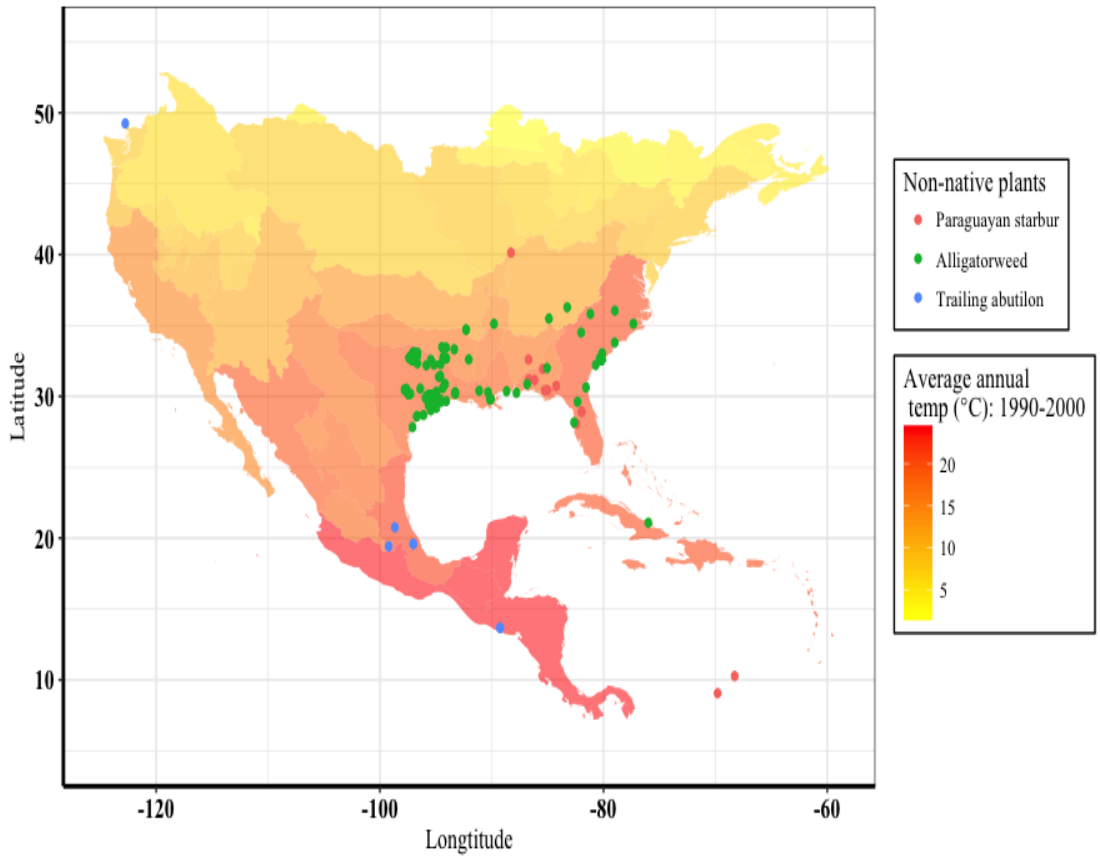

Figure 5. U.S. and Mexico mean historical temperature map for years 1990-2000 overlaid with species distribution occurrences of three non-native plant species, i.e. *A. australe* (paraguayan starbur), *A. megapotamicum* (trailing abutilon), *and A. philoxeroides* (alligatorweed). Historical temperature data were retrieved for each decade by averaging monthly temperatures from years 1990-2000. Species are shown in non-native ranges. Climate data from World Bank Climate API for basin level spatial scales. Species occurrence records were retrieved from GBIF.

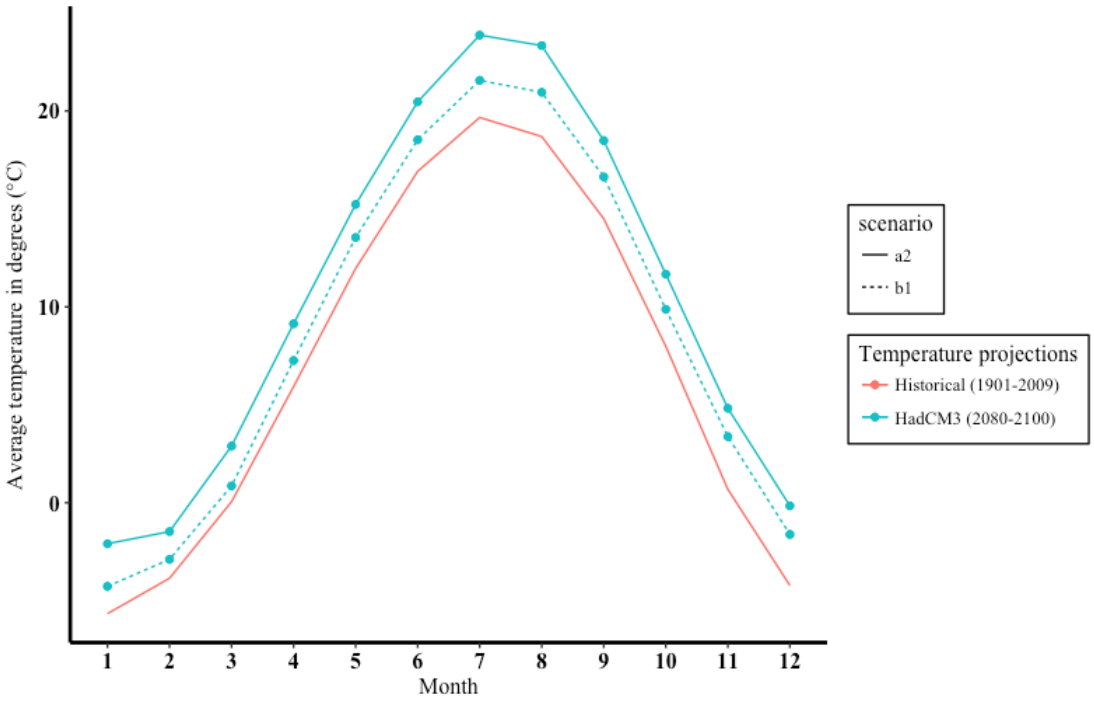

Figure 6. Averaged monthly temperature (°C) projections in the United States for years 1901-2009 and 2080-2100. Historical temperatures for years 1901-2009 are lower than HadCM3 modeled temperature projections for future years 2080-2100. Historical averaged monthly temperatures were derived for each decade. HadCM3 climate model temperature projections are presented for future GHG scenarios a2 (pessimistic) and b1 (optimistic). The Climatic Research Unit (CRU) of the University of East Anglia (UEA) provides the original dataset. Temperature data were retrieved for the U.S. ISO country code.

Discussion

Plant species distributions

Alligatorweed

Alligatorweed is a perennial aquatic and terrestrial invasive plant that is listed as a Federal Noxious Weed and a major-invader. The current distribution presented in this study and the distribution predicted by "MaxEnt" (Knocki & Aronson, 2015) are similar in their mapped distributions shows the majority of occurrence points located in the Southeast continental U.S. This is consistent with reports supporting that alligatorweed prefers subtropical to cool climates found in freshwater habitats in the Southeast (https://www.cabi.org/isc/datasheet/4403). The optimal growth temperature of alligatorweed in a greenhouse has experimentally been shown to be between 15-20°C (Julien *et al.*, 1995). However, another study showed a 90% survival rate of alligatorweed stem cuttings (4-5 cm in length; *N*=2000) harvested from 29°N-31°N in July 2008 and grown in a greenhouse where temperatures were between 15-40°C (15-25°C from October to November and 30-40°C from August to September) (Sun *et al.*, 2010), suggesting a wider habitat suitability. The HadCM3 climate model projection for mean monthly temperature, for years 2080-2100, is predicted to be highest in July, 23.87°C for scenario a2 and 21.56°C for scenario b1. The lowest predicted temperatures are -0.15°C and -1.61°C for a2 and b1, respectively. Historical temperatures and projected forecast of climate model temperature estimates presented in this study, and experimental temperatures (Julien *et al.*, 1995; Sun *et al.*, 2010) show the temperature in the continental U.S. will remain within physical tolerable

values for alligatorweed; therefore, future species range in the continental U.S. may continue to expand in freshwater habitats.

Paraguayan starbur

Paraguayan starbur is an annual and short-lived perennial that is a minor-invader. Its distribution presented in this study is similar to the distribution found with the "MaxEnt" species distribution model (Magarey *et al.*, 2017). Both distribution maps show occurrence points located in the Southeast continental U.S., where it thrives in warm, relatively dry habitats where the average warm temperature is > 10°C and the average cold temperature is > 0°C (https://www.cabi.org/isc/datasheet/118957). The mean temperature predictions from the HadCM3 climate model indicate that U.S. habitat temperatures will be increasingly tolerable for paraguyan starbur, indicating that species range could potentially expand in the future.

Trailing abutilon

Trailing abutilon is an annual ornamental plant and is considered a non-invader in the U.S. Its distribution presented in this study is mainly in Mexico and Canada, right above the Pacific Northwest. Trailing abutilon is found in USDA plant hardiness zones 8-10, where average annual extreme minimum temperature is estimated to be from -12.2°C to 4.4°C (http://planthardiness.ars.usda.gov/PHZMWeb/#) and was predicted to be a non-invader in the Florida WRA (Gordon *et al.*, 2008b). There are no species distribution models describing plant range for trailing abutilon available in scientific literature. This could be attributable to its status as a non-invader. From the available literature, one study showed that trailing abutilon is susceptible to Leaf Spot disease, caused by *Myrothecium roridum* Tode ex Fr., in the greenhouse and fields

(Ben *et al.*, 2016). Additionally, ornamental plant care manuals suggest yearly pruning to help prevent disease. Invasive plants tend to be characteristically hardy (Chai *et al.*, 2016) and relatively pest-free (Keane & Crawley, 2002; Lind & Parker, 2010). If perspective ranges become favorable in the future with an increase in mean temperature (as predicted by HadCM3 GCM), trailing abutilon may become invasive in the future.

Utility and future directions of simple SDMs

In this exploratory section of the study, species occurrence points were overlaid on historical temperatures (the mean monthly temperature recorded for years 1990-2000). This overlay of species occurrence points onto a climate map is a simple species distribution model (SDM) because species absences are not accounted for and only species occurrence records from a country are visualized (global distributions of occurrence records cannot be extrapolated for their associated climatic variables and then mapped to new particular region). Therefore, resultant species distributions are not quantitatively predicted for new regions.

The USDA-APHIS-PPQ WRA process uses a similar and more robust SDM, Proto3, to assess geographic/climatic suitability risk for regional establishment of species. The Proto3 model combines a Geographic Information System (GIS) overlay of three climate variables (i.e. plant hardiness zones (Magarey *et al.*, 2008), 10-inch global precipitation bands (Magarey *et al.*, 2008), and Köppen–Geiger climate classes (Peel *et al.*, 2007) to reveal a potential distribution of species (PPQ, 2016).

Although complex SDMs are known to have high predictive accuracies (Vorsino *et al.*, 2014), there are still benefits to using simple SDMs. For example, the simple SDM presented in this study used the spocc package, which can easily combine species occurrence data from multiple data sources, such as GBIF, Berkeley Ecoengine, iNaturalist, VertNet, Biodiversity Information Serving Our Nation (rbison), eBird, AntWeb, iDigBio, Ocean Biogeographic Information System (OBIS), and Atlas of Living Australia (ALA) (Chamberlin, 2017). However, some of these sources have overlaps in data resulting in duplicate occurrence points. This is an issue that could potentially be resolved in future versions of the spocc package (Chamberlin, 2017). Nevertheless, spocc package-derived tools could prove useful when modeling the presence of multiple species with niche overlaps. This species-level comparison can be a first step to get a better understanding of the underlying ecological relationships/trends in a region. Exploration of species-level comparisons are particularly important in WRAs because it is widely known that invasive species have detrimental effects on ecosystem structure and function (Neckles, 2015; Naeem *et al.*, 1994).

The most obvious advantage of this methodology for modeling species distributions is the availability of the software through open-source avenues like R (CRAN). Additionally, species occurrence data from spocc can be combined with other R software packages, as shown in this study, such as rWBclimate to generate climate maps with historical data or GCM data. Yet, climatic variable extrapolation from global species occurrences to predict for new geographic suitability is not yet available from these packages. Besides this lack in current functionality, these tasks

are more or less easily executable for individuals with at least an intermediate expertise in R.

One of the driving forces of the USDA-APHIS-PPQ is a demand for more timely and flexible solutions for assessing invasive potential risk for non-native species (PPQ, 2015). Ease of use for generating SDMs for the WRA process becomes particularly important for Risk Analysts conducting these assessments because they might not have experience developing complex SDMs and simpler SDMs take less time to conduct with respect to gathering data for the modeling process and to actually carry out the model run (Magarey *et al.*, 2017). It can be useful to incorporate more SDMs to facilitate a more integrative WRA process.

Appendices

Appendix A. WRA dataset variables used for development of all models. Questions ES1-IMPP6 are predictor variables and INVASIVE STATUS is the dependent variable. Adapted from Koop *et al.,* 2012. Corresponding numerical values are entered into model for lettered responses (i.e. A, B, C, D, E, and F). Responses with a "?" receive a numerical value of 0.

Variables	**Score Scale**
ESTABLISHMENT/SPREAD POTENTIAL	
ES1 (Status/invasiveness outside native range)	A=-5; B=-2; C=0; D=0; E=2; F=5; ?=0
ES2 (Is the species highly domesticated)	Yes=-3; No=0; ?=0
ES3 (Weedy congeners)	Yes=1; No=0; ?=0
ES4 (Shade tolerant at some stage of its life cycle)	Yes=1; No=0; ?=0
ES5 (Climbing or smothering growth form)	Yes=1; No=0; ?=0
ES6 (Forms dense thickets)	Yes=2; No=0; ?=0
ES7 (Aquatic)	Yes=1; No=0; ?=0
ES8 (Grass)	Yes=1; No=0; ?=0
ES9 (Nitrogen-fixing woody plant)	Yes=1; No=0; ?=0
ES10 (Does it produce viable seeds or spores)	Yes=1; No=-1; ?=0
ES11 (Self-compatible or apomictic)	Yes=1; No=-1; ?=0
ES13 (Minimum generation time)	A=2; B=1; C=0; D=-1
ES14 (Prolific reproduction)	Yes=1; No=-1; ?=0
ES15 (Propagules likely to be dispersed unintentionally by people)	Yes=1; No=-1; ?=0
ES16 (Propagules likely to disperse in trade as contaminants or hitchhikers)	Yes=1; No=-1; ?=0
ES17 (Number of natural dispersal vectors)	None=-4; One=-2; Two=0; Three= 2; Four or Five=4
ES18 (Evidence that a persistent (>1 year) propagule bank (seed bank) is formed)	Yes=1; No=-1; ?=0
ES19 (Tolerates/benefits from mutilation, cultivation or fire)	Yes=1; No=-1; ?=0
ES20 (Is resistant to herbicides or potential to acquire herbicide resistance)	Yes=1; No=0; ?=0

ES21 (Number of cold hardiness zones suitable for its survival)	Varies
ES22 (Number of climate types suitable for its survival)	Varies
ES23 (Number of precipitation bands suitable for its survival)	Varies

IMPACT POTENTIAL

IMPG1 (Allelopathic)	Yes= 0.1; No=0; ?=0
IMPG2 (Parasitic)	Yes=0.1; No=0; ?=0
IMPN1 (Change ecosystem processes and parameters that affect other species)	Yes=0.4; No=0; ?=0
IMPN2 (Changes community structure)	Yes=0.2; No=0; ?=0
IMPN3 (Changes community composition)	Yes=0.2; No=0; ?=0
IMPN4 (Is it likely to affect federal Threatened and Endangered species)	Yes=0.1; No=0; ?=0
IMPN5 (Is it likely to affect any globally outstanding ecoregions)	Yes=0.1; No=0; ?=0
IMPN6 (Weed status in natural systems)	A=0; B=0.2; C=0.6
IMPA1 (Impacts human property, processes, civilization, or safety)	Yes=0.1; No=0; ?=0
IMPA2 (Changes or limits recreational use of an area)	Yes=0.1; No=0; ?=0
IMPA3 (Outcompetes, replaces, or otherwise affects desirable plants and vegetation)	Yes=0.1; No=0; ?=0
IMPA4 (Weed status in anthropogenic systems)	A=0; B=0.1; C=0.4
IMPP1 (Reduces crop/product yield)	Yes=0.4; No=0; ?=0
IMPP2 (Lowers commodity value)	Yes=0.2; No=0; ?=0
IMMP3 (Is it likely to impact trade)	Yes=0.2; No=0; ?=0
IMMP4 (Reduces the quality or availability of irrigation, or strongly competes with plants for water)	Yes=0.1; No=0; ?=0
IMMP5 (Toxic to animals, including livestock/range animals and poultry)	Yes=0.1; No=0; ?=0

IMMP6 (Weed status in production systems)	A=0; B=0.2; C=0.6
INVASIVE STATUS	Non-invader; Major-invader; Minor-invader

Appendix B. Plant species used to develop all WRAs in this study. Adapted from Koop *et al.,* 2012. Note: *Plant species used for the 30% random forest validation subset.

Species	Family	Habit
	Major-invaders	
Abrus precatorius	Fabaceae	Vine
*Abutilon theophrasti**	Malvaceae	Herb
Alnus glutinosa	Betulaceae	Tree
*Ardisa elliptica**	Myrsinaceae	Shrub
Avena fatua	Poaceae	Graminoid
Cardaria draba	Brassicaceae	Herb
*Centaurea solstitialis**	Asteraceae	Herb
Cirsium arvense	Asteraceae	Herb
*Convolvulus arvensis**	Convolvulaceae	Vine
Cupaniopsis anacardioides	Anacardiaceae	Tree
Cyperus rotundua	Cyperaceae	Graminoid
*Datura stramonium**	Solanaceae	Herb
Eichhornia crassipes	Pontederiaceae	Herb

Eugenia uniflora	Myrtaceae	Shrub
Hydrilla verticillata	Hydrocharitaceae	Aquatic
Lactuca serriola	Asteraceae	Herb
Lonicera maackii	Caprifoliaceae	Shrub
Miconia calvescens	Melastomataceae	Tree
Mimosa pigra	Fabaceae	Shrub
Pittosporum undulatum	Pittosporaceae	Herb
Portulaca oleraceae	Portulaceae	Herb
Psidium guajava	Myrtaceae	Tree
Rumex crispus	Polgonaceae	Herb
Psidim guajava	Myrtaceae	Tree
Rumex crispus	Polgonaceae	Herb
Senecio vulgaris	Asteraceae	Herb
Setaria italica subsp. *Viridis*	Poaceae	Graminoid
Sisymbrium irio	Brassicaceae	Herb
*Solanum nigrum**	Solanaceae	Subshrub
Sorghum halepense	Poaceae	Graminoid
Tamarix ramosissima	Tamaricaceae	Shrub
Thlaspi arvense	Brassicaceae	Herb
Tradescantia fluminensis	Commelinaceae	Herb
Triadica sebifera	Euphorbiaceae	Tree
*Aegilops cylindrica**	Poaceae	Graminoid
*Albizia julibrissin**	Fabaceae	Tree

Allium vineasle	Lilliaceae	Herb
Alternanthera philoxeroides	Amaranthaceae	Aquatic
Barbarea vulgaris	Brassicaceae	Herb
Berberis thunbergii	Berberidaceae	Shrub
Bromus tectorum	Poaceae	Graminoid
Capsella bursa-pastoris	Brassicaceae	Herb
Carduus nutans	Asteraceae	Herb
Carpobrotus chilensis	Aizoaceae	Herb
Casuarina equisetifolia	Casuarinaceae	Tree
Cayratia japonica	Vitaceae	Vine
*Cytisus scoparius**	Fabaceae	Shrub
Daucus carota subsp.*	Apiaceae	Herb
Carota		
Elaeagnus umbellate	Elaeagnaceae	Shrub
Emex spinosa	Polygonaceae	Herb
Euphorbia esula	Euphorbiaceae	Herb
Galinsoga parviflora	Asteraceae	Herb
Hypericum perforatum	Hypericaceae	Herb
*Imperata cylindrical**	Poaceae	Graminoid
Lamium amplexicaule	Lamiaceae	Herb
Lygodium japonicum	Lygodiaceae	Vine
Lythrum salicaria	Lythraceae	Aquatic
Malva parviflora	Malvaceae	Tree

*Myriophyllum spicatum**	Haloragaceae	Aquatic
Nandina domestica	Berberidaceae	Shrub
*Neyraudia reynaudiana**	Poaceae	Graminoid
*Pennisetum ciliare**	Poaceae	Graminoid
Poa annua	Poaceae	Graminoid
Polygonum aviculare	Polygonaceae	Herb
Polygonum convolculus	Polygonaceae	Vine
*Rottboellia cochinchinesis**	Poaceae	Graminoid
*Schefflera actinophylla**	Araliaceae	Tree
Minor-invaders		
*Acer palmatum**	Aceraceae	Tree
*Artocarpus heterophyllus**	Moraceae	Tree
*Bellardia trixago**	Scrophulariaceae	Herb
Cichorium intybus	Asteraceae	Herb
Cissus rotundifolia	Vitaceae	Vine
Clematis terniflora	Ranunculaceae	Vine
Costus speciosus	Zingiberaceae	Herb
Cotoneaster coriaceus	Rosaceae	Shrub
Dioscorea oppositifolia	Dioscoreaceae	Vine
Epipactis helleborine	Orchidaceae	Herb
Euryops multifidus	Asteraceae	Subshrub
Geranium pusillum	Geraniaceae	Herb

Gloriosa superba	Colchicaceae	Vine
*Gomphrena globosa**	Amaranthaceae	Herb
Hiptage benghalensis	Malphigiaceae	Vine
*Hylotelephium telephium**	Crassulaceae	Herb
*Ilex paraguariensis**	Aquifoliaceae	Tree
Ligustrum obtusifolium	Oleaceae	Shrub
Linaraia vulgaris	Scrophulariaceae	Herb
Lysimachia punctata	Primulaceae	Herb
Melilotus indicus	Fabaceae	Herb
Orobanche minor	Orobanchaceae	Herb
Prunus armeniaca	Rosaceae	Tree
Pyracantha coccinea	Rosaceae	Shrub
Quisqualis indica	Combretaceae	Vine
Ranunculus acris	Ranunculaceae	Herb
Rhamnus utilis	Rhamnaceae	Shrub
Ribes rubrum	Grossulariaceae	Shrub
Rumex pulcher	Polgonaceae	Herb
Saponaria officinalis	Caryophyllaceae	Herb
*Senecio jacobaea**	Asteraceae	Herb
Spermacoce latifolia	Rubiaceae	Herb
*Stapelia gigantea**	Asclepiadaceae	Herb
Tillandsia gardneri	Bromeliaceae	Epiphyte
Abutilon hirtum	Malvaceae	Shrub

Acanthospermum australe	Asteraceae	Herb
Actinidia chinensis	Actinidiaceae	Vine
Agrostemma githago	Caryophyllaceae	Herb
Aira caryophyllea	Poaceae	Graminoid
Akebia quinata	Lardizabalaceae	Vine
*Antirrhimum majus**	Scrophulariaceae	Herb
Archontophoenix alexandrae	Arecaceae	Tree
Arctium minus	Asteraceae	Herb
Bassis hyssopifolia	Chenopodiaceae	Herb
*Betula pendula**	Betulaceae	Tree
Castilla elastica	Moraceae	Tree
Conium maculatum	Apiaceae	Herb
Costus dubius	Zingiberaceae	Herb
Dendrobium crumenatum	Orchidaceae	Epiphtye
Eucalyptus camaldulensis	Myrtaceae	Tree
*Euonymus alatus**	Celastraceae	Shrub
Glechoma hederacea	Lamiaceae	Herb
Guzmania lindenii	Bromeliaceae	Epiphyte
Helichrysum petiolare	Asteraceae	Subshrub
Hygrophila polysperma	Acanthaceae	Aquatic
*Ligustrum sinense**	Oleaceae	Shrub
Luma apiculata	Myrtaceae	Tree

Luziola subintegra	Poaceae	Aquatic
Pittosporum pentandrum	Pittosporaceae	Tree
Rosa multiflora	Rosaceae	Shrub
Setaria palmifolia	Poaceae	Graminoid
*Spartina densiflora**	Poaceae	Graminoid
*Theobroma cacao**	Sterculiaceae	Tree
*Trachelospemum jasminoides**	Apocynaceae	Vine
Ulmus procera	Ulmaceae	Tree
Verbena bonariensis	Verbenaceae	Subshrub
Wisteria sinensis/W. floribunda	Fabaceae	Vine
Xanthosoma atrovirens	Araceae	Herb
Non-invaders		
*Agave filifera**	Agavaceae	Shrub
Allium giganteum	Liliaceae	Herb
*Asarum europaeum**	Aristolochiaceae	Herb
Bombax ceiba	Bombacaceae	Tree
Brugmansia sanguinea	Solanaceae	Shrub
Buxus microphylla	Buxaceae	Shrub
Catalpa bungei	Bignoniaceae	Tree
*Cedrus libani**	Pinaceae	Tree
Centaurea dealbata	ASteraceae	Herb

*Cupressus sempervirens**	Cupressaceae	Herb
*Festuca amenthystina**	Poaceae	Graminoid
*Fortunella japonica**	Rutaceae	Shrub
Gazania rigens	Asteraceae	Herb
Kniphofia caulescens	Liliacae	Herb
Linaria alpina	Schrophulariaceae	Herb
*Listera ovate**	Orchidaceae	Herb
*Medicago arborea**	Fabaceae	Shrub
*Pistacia chinesis **	Anacardiaceae	Tree
Podophyllum hexandrum	Berberidaceae	Herb
*Polygonum amplexicaule**	Polygonaceae	Herb
Pouteria sapota	Sapotaceae	Tree
*Primula elatior**	Primulaceae	Herb
*Primula pulverulenta**	Primulaceae	Herb
Prunus japonica	Rosaceae	Shrub
Rhododendron simsii	Ericaceae	Shrub
Ribes orientale	Grossulariaceae	Shrub
Rondeletia odorata	Rubiaceae	Shrub
Saintpaulia ionantha	Gesneriaceae	Herb
Styphnolobium japonicum	Fabaceae	Tree
*Teucrium chamaedrys**	Lamiaceae	Subshrub
Tulipa gesneriana	Liliaceae	Herb

Yucca guatemalensis	Agavaceae	Tree
*Abutilon megapotamicum**	Malvaceae	Shrub
Acer buergerianum	Aceraceae	Tree
Acorus gramineus	Acoraceae	Aquatic
*Bergernia crassifolia**	Saxifragaceae	Herb
Blechnum brasiliense	Blechnaceae	Herb
*Brachycome iberidifolia**	Asteraceae	Herb
Ceiba speciosa	Bombacaceae	Tree
*Combretum coccineum**	Combretaceae	Shrub
Davallia canariensis	Polypodiaceae	Herb
*Dendrocalamus latifolrus**	Poaceae	Tree
Diospyros kaki	Ebenaceae	Tree
Erica carnea	Ericaceae	Subshrub
*Fatsia japonica**	Araliaceae	Shrub
Gardenia thunbergii	Rubiaceae	Shrub
Ginkgo biloba	Ginkgoaceae	Tree
Hydrangea anomala	Hydrangeaceae	Vine
*Lavandula latiflora**	Lamiaceae	Subshrub
Libertia grandiflora	Iridaceae	Herb
Lilium martagon	Liliaceae	Herb
Myrtus communis	Myrtaceae	Shrub

Penstemon companulatus	Scrophulariaceae	Herb
*Pinus wallichiana**	Pinaceae	Tree
Pittosporum bicolor	Pittosporaceae	Tree
Prunus maackii	Rosaceae	Tree
Quercus serrata	Fagaceae	Tree
Rodgersia sambucifolia	Saxifragaceae	Herb
Salix glabra	Salicaceae	Shrub
Stenocarpus sinuatus	Proteaceae	Tree
Stephanandra tanakae	Rosaceae	Shrub
Syzygium eucalyptoides	Myrataceae	Tree
*Torreya nucifera**	Taxaceae	Tree
Trollius europaeus	Ranunculaceae	Herb
Viburnum farreri	Adoxaceae	Shrub
*Wisteria brachybotrys**	Fabaceae	Vine

Appendix C. R codes for statistical modeling work

```
#########first model 2-fold CV###################################


library(randomForest) #random forests algorithm
library(lattice)
library(ggplot2)
library(caret) #for cross validation folds
library(ROCR) #for ROC plots and statistics
library(gplots)
library(pROC) #same as ROCR

data=read.csv(file="filename.csv") #data for prediction
#NON=non invader=0 #OTH=major invader + minor invader=1

###Using 41 variables specified in Appendix A###
full.data<-data[,c("invasion.status",
              "es1", "es2", "es3","es4", "es5","es6","es7",
              "es8","es9","es10","es11","es12","es13","es14",
              "es15","es16","es17","es18","es19","es20","es21",
              "es22","es23", "impg1", "impg2", "impn1", "impn2",
              "impn3", "impn4", "impn5", "impn6", "impa1", "impa2",
              "impa3", "impa4", "impp1", "impp2", "impp3", "impp4",
              "impp5", "impp6")]

###Conversion of the response variable “invasive status” into Factor with names for
Caret Library###
full.data$invasion.status<-factor(full.data$invasion.status,
                          levels=c(0,1),
                          labels=c("NON", "OTH"))


set.seed(100)  #for repeatability

#In the trainControl function, the resampling method is "repeatedcv" (repeated cross-
validation)
#number = 2 indicates that there are 2 folds in K-fold cross-validation
#repeats = 10 indicates that there are ten separate 2-fold cross-validations used as the
resampling scheme
#verboseIter is a logical for printing a training log
#returnData is a logical for saving the data into a slot called trainingData
#summaryFunction provides a ROC  AUC summary statistics

tc<-trainControl(method="repeatedcv", number=2,
            repeats=10, summaryFunction=twoClassSummary,
```

```
            verboseIter = T, returnData = T,
            classProb=T, savePredictions = T)

LR.first.model<-train(as.factor(invasion.status)~es1+es2+es3+
                      es4+es5+es6+es7+es8+es9+es10+es11+es12+
                      es13+es14+es15+es16+es17+es18+es19+es20+
                      es21+es22+es23+impg1+impg2+impn1+impn2+impn3+
                      impn4+impn5+impn6+impa1+impa2+impa3+impa4+impp1+
                      impp2+impp3+impp4+impp5+impp6,
                    metric="ROC", method="glm",
                    family="binomial", trControl=tc,
                    data=full.data)

#model coefficients for independent variables
LR.first.model$finalModel

# training log saved from the returnData argument = TRUE
LR.first.model$trainingData

# cross validation summary statistics
LR.first.model

### RF with 2-fold CV on data###
RF.first.model<-train(as.factor(invasion.status)~.,
                  metric="ROC", method="rf",
                  importance=T, proximity=F,
                  ntree=1000, trControl=tc,
                  data=full.data)

RF.first.model$finalModel
RF.first.model$trainingData

RF.first.model




#########second model 2-fold CV#################################


library(randomForest) #random forests algorithm
library(lattice)
```

```
library(ggplot2)
library(caret) #for cross validation folds
library(ROCR) # for ROC plots and statistics
library(gplots)
library(pROC) #same as ROCR

data=read.csv(file="filename.csv")
#MAJ=major invader=1 #MIN=minor invader=0

###Using 41 variables specified in Appendix A###
full.data<-data[,c("invasion.status",
                   "es1", "es2", "es3","es4", "es5","es6","es7",
                   "es8","es9","es10","es11","es12","es13","es14",
                   "es15","es16","es17","es18","es19","es20","es21",
                   "es22","es23", "impg1", "impg2", "impn1", "impn2",
                   "impn3", "impn4", "impn5", "impn6", "impa1", "impa2",
                   "impa3", "impa4", "impp1", "impp2", "impp3", "impp4",
                   "impp5", "impp6")]


###Conversion of the response variable “invasive status” into Factor with names for
Caret Library###
full.data$invasion.status<-factor(full.data$invasion.status,
                                  levels=c(0,1),
                                  labels=c("MIN", "MAJ"))

set.seed(100)  #for repeatability

#In the trainControl function, the resampling method is "repeatedcv" (repeated cross-
validation)
#number = 2 indicates that there are 2 folds in K-fold cross-validation
#repeats = 10 indicates that there are ten separate 2-fold cross-validations used as the
resampling scheme
#verboseIter is a logical for printing a training log
#returnData is a logical for saving the data into a slot called trainingData
#summaryFunction provides a ROC  AUC summary statistics

tc<-trainControl(method="repeatedcv", number=2,
                 repeats=10, summaryFunction=twoClassSummary,
                 verboseIter = T, returnData = T,
                 classProb=T, savePredictions = T)

LR.second.model<-train(as.factor(invasion.status)~es1+es2+es3+
                         es4+es5+es6+es7+es8+es9+es10+es11+es12+
                         es13+es14+es15+es16+es17+es18+es19+es20+
                         es21+es22+es23+impg1+impg2+impn1+impn2+impn3+
```

```
                  impn4+impn5+impn6+impa1+impa2+impa3+impa4+impp1+
                  impp2+impp3+impp4+impp5+impp6,
                metric="ROC", method="glm",
                family="binomial", trControl=tc,
                data=full.data)

#model coefficients for independent variables
LR.second.model$finalModel

# training log saved from the returnData argument = TRUE
LR.second.model$trainingData

# cross validation summary statistics
LR.second.model

### RF with 2-fold CV on data###
RF.second.model<-train(as.factor(invasion.status)~.,
                metric="ROC", method="rf",
                importance=T, proximity=F,
                ntree=1000, trControl=tc,
                data=full.data)

RF.second.model$finalModel
RF.second.model$trainingData

RF.second.model


#########first model 5-fold CV####################################



library(randomForest) #random forests algorithm
library(lattice)
library(ggplot2)
library(caret) #for cross validation folds
library(ROCR) # for ROC plots and statistics
library(gplots)
library(pROC) #same as ROCR

data=read.csv(file="filename.csv") #data for prediction
#NON=non invader=0 #OTH=major invader + minor invader=1

###Using 41 variables specified in Appendix A###
full.data<-data[,c("invasion.status",
              "es1", "es2", "es3","es4", "es5","es6","es7",
```

```
"es8","es9","es10","es11","es12","es13","es14",
"es15","es16","es17","es18","es19","es20","es21",
"es22","es23", "impg1", "impg2", "impn1", "impn2",
"impn3", "impn4", "impn5", "impn6", "impa1", "impa2",
"impa3", "impa4", "impp1", "impp2", "impp3", "impp4",
"impp5", "impp6")]

###Conversion of the response variable “invasive status” into Factor with names for Caret Library###
full.data$invasion.status<-factor(full.data$invasion.status,
                        levels=c(0,1),
                        labels=c("NON", "OTH"))


set.seed(100)  #for repeatability

#In the trainControl function, the resampling method is "repeatedcv" (repeated cross-validation)
#number = 5 indicates that there are 5 folds in K-fold cross-validation
#repeats = 10 indicates that there are ten separate 5-fold cross-validations used as the resampling scheme
#verboseIter is a logical for printing a training log
#returnData is a logical for saving the data into a slot called trainingData
#summaryFunction provides a ROC  AUC summary statistics

tc<-trainControl(method="repeatedcv", number=5,
            repeats=10, summaryFunction=twoClassSummary,
            verboseIter = T, returnData = T,
            classProb=T, savePredictions = T)

LR.first.model<-train(as.factor(invasion.status)~es1+es2+es3+
                    es4+es5+es6+es7+es8+es9+es10+es11+es12+
                    es13+es14+es15+es16+es17+es18+es19+es20+
                    es21+es22+es23+impg1+impg2+impn1+impn2+impn3+
                    impn4+impn5+impn6+impa1+impa2+impa3+impa4+impp1+
                    impp2+impp3+impp4+impp5+impp6,
                  metric="ROC", method="glm",
                  family="binomial", trControl=tc,
                  data=full.data)

#model coefficients for independent variables
LR.first.model$finalModel

# training log saved from the returnData argument = TRUE
LR.first.model$trainingData
```

```
# cross validation summary statistics
LR.first.model

### RF with 5-fold CV on data###
RF.first.model<-train(as.factor(invasion.status)~.,
                metric="ROC", method="rf",
                importance=T, proximity=F,
                ntree=1000, trControl=tc,
                data=full.data)

RF.first.model$finalModel
RF.first.model$trainingData

RF.first.model



#########second model 5-fold CV#################################



library(randomForest) #random forests algorithm
library(lattice)
library(ggplot2)
library(caret) #for cross validation folds
library(ROCR) # for ROC plots and statistics
library(gplots)
library(pROC) #same as ROCR

data=read.csv(file="filename.csv")
#MAJ=major invader=1 #MIN=minor invader=0

###Using 41 variables specified in Appendix A###
full.data<-data[,c("invasion.status",
                "es1", "es2", "es3","es4", "es5","es6","es7",
                "es8","es9","es10","es11","es12","es13","es14",
                "es15","es16","es17","es18","es19","es20","es21",
                "es22","es23", "impg1", "impg2", "impn1", "impn2",
                "impn3", "impn4", "impn5", "impn6", "impa1", "impa2",
                "impa3", "impa4", "impp1", "impp2", "impp3", "impp4",
                "impp5", "impp6")]


###Conversion of the response variable “invasive status” into Factor with names for
Caret Library###
full.data$invasion.status<-factor(full.data$invasion.status,
```

```
                    levels=c(0,1),
                    labels=c("MIN", "MAJ"))

set.seed(100)  #for repeatability

#In the trainControl function, the resampling method is "repeatedcv" (repeated cross-validation)
#number = 5 indicates that there are 5 folds in K-fold cross-validation
#repeats = 10 indicates that there are ten separate 5-fold cross-validations used as the resampling scheme
#verboseIter is a logical for printing a training log
#returnData is a logical for saving the data into a slot called trainingData
#summaryFunction provides a ROC  AUC summary statistics

tc<-trainControl(method="repeatedcv", number=5,
            repeats=10, summaryFunction=twoClassSummary,
            verboseIter = T, returnData = T,
            classProb=T, savePredictions = T)

LR.second.model<-train(as.factor(invasion.status)~es1+es2+es3+
                    es4+es5+es6+es7+es8+es9+es10+es11+es12+
                    es13+es14+es15+es16+es17+es18+es19+es20+
                    es21+es22+es23+impg1+impg2+impn1+impn2+impn3+
                    impn4+impn5+impn6+impa1+impa2+impa3+impa4+impp1+
                    impp2+impp3+impp4+impp5+impp6,
                  metric="ROC", method="glm",
                  family="binomial", trControl=tc,
                  data=full.data)

#model coefficients for independent variables
LR.second.model$finalModel

# training log saved from the returnData argument = TRUE
LR.second.model$trainingData

# cross validation summary statistics
LR.second.model

### RF with 5-fold CV on data###
RF.second.model<-train(as.factor(invasion.status)~.,
                metric="ROC", method="rf",
                importance=T, proximity=F,
                ntree=1000, trControl=tc,
                data=full.data)

RF.second.model$finalModel
```

```
RF.second.model$trainingData

RF.second.model


#########first model 10-fold CV###################################



library(randomForest) #random forests algorithm
library(lattice)
library(ggplot2)
library(caret) #for cross validation folds
library(ROCR) # for ROC plots and statistics
library(gplots)
library(pROC) #same as ROCR

data=read.csv(file="filename.csv") #data for prediction
#NON=non invader=0 #OTH=major invader + minor invader=1

###Using 41 variables specified in Appendix A###
full.data<-data[,c("invasion.status",
                   "es1", "es2", "es3","es4", "es5","es6","es7",
                   "es8","es9","es10","es11","es12","es13","es14",
                   "es15","es16","es17","es18","es19","es20","es21",
                   "es22","es23", "impg1", "impg2", "impn1", "impn2",
                   "impn3", "impn4", "impn5", "impn6", "impa1", "impa2",
                   "impa3", "impa4", "impp1", "impp2", "impp3", "impp4",
                   "impp5", "impp6")]

###Conversion of the response variable “invasive status” into Factor with names for
Caret Library###
full.data$invasion.status<-factor(full.data$invasion.status,
                                  levels=c(0,1),
                                  labels=c("NON", "OTH"))


set.seed(100)  #for repeatability

#In the trainControl function, the resampling method is "repeatedcv" (repeated cross-
validation)
#number = 10 indicates that there are 10 folds in K-fold cross-validation
#repeats = 10 indicates that there are ten separate 10-fold cross-validations used as
the resampling scheme
#verboseIter is a logical for printing a training log
#returnData is a logical for saving the data into a slot called trainingData
```

```
#summaryFunction provides a ROC  AUC summary statistics

tc<-trainControl(method="repeatedcv", number=10,
                 repeats=10, summaryFunction=twoClassSummary,
                 verboseIter = T, returnData = T,
                 classProb=T, savePredictions = T)

LR.first.model<-train(as.factor(invasion.status)~es1+es2+es3+
                        es4+es5+es6+es7+es8+es9+es10+es11+es12+
                        es13+es14+es15+es16+es17+es18+es19+es20+
                        es21+es22+es23+impg1+impg2+impn1+impn2+impn3+
                        impn4+impn5+impn6+impa1+impa2+impa3+impa4+impp1+
                        impp2+impp3+impp4+impp5+impp6,
                      metric="ROC", method="glm",
                      family="binomial", trControl=tc,
                      data=full.data)

#model coefficients for independent variables
LR.first.model$finalModel

# training log saved from the returnData argument = TRUE
LR.first.model$trainingData

# cross validation summary statistics
LR.first.model

### RF with 10-fold CV on data###
RF.first.model<-train(as.factor(invasion.status)~.,
                      metric="ROC", method="rf",
                      importance=T, proximity=F,
                      ntree=1000, trControl=tc,
                      data=full.data)

RF.first.model$finalModel
RF.first.model$trainingData

RF.first.model

###ROC Plots###
LR.first.model.pred<-predict(LR.first.model, full.data$invasion.status, type="prob")
RF.first.model.pred<-predict(RF.first.model, full.data$invasion.status, type="prob")
first.model.pred.LR<-prediction(LR.first.model.pred$OTH, data$invasion.status)
first.model.perf.LR<-performance(first.pred.LR, "tpr", "fpr")
first.model.pred.RF<-prediction(RF.first.model.pred$OTH, data$invasion.status)
first.model.perf.RF<-performance(first.pred.RF, "tpr", "fpr")
```

```
par(family="serif", fig=c(0.0,1,0.0,1),
   mar=par()$mar+c(2,2,2,13), xpd=TRUE,
   cex.lab=1.2, lwd=2, pty="m", font.axis=2)
plot(first.model.perf.LR,
    main="Logistic Regression and Random Forests",
    lty=1, col="red")
plot(first.model.perf.RF, xlab = "False Positive Rate",
    ylab = "True Positive Rate",
    add=T, lty=3, col="blue")


legend(1.2, 1.0, c("Logistic Regression 0.865",
              "Random Forest  0.943"),
      xpd=TRUE,
      lty = c(1,3), col = c("red","blue","green"),
      bty="o")


#########second model 10-fold CV#################################


library(randomForest) #random forests algorithm
library(lattice)
library(ggplot2)
library(caret) #for cross validation folds
library(ROCR) # for ROC plots and statistics
library(gplots)
library(pROC) #same as ROCR

data=read.csv(file="filename.csv")
#MAJ=major invader=1 #MIN=minor invader=0

###Using 41 variables specified in Appendix A###
full.data<-data[,c("invasion.status",
              "es1", "es2", "es3","es4", "es5","es6","es7",
              "es8","es9","es10","es11","es12","es13","es14",
              "es15","es16","es17","es18","es19","es20","es21",
              "es22","es23", "impg1", "impg2", "impn1", "impn2",
              "impn3", "impn4", "impn5", "impn6", "impa1", "impa2",
              "impa3", "impa4", "impp1", "impp2", "impp3", "impp4",
              "impp5", "impp6")]


###Conversion of the response variable “invasive status” into Factor with names for
Caret Library###
full.data$invasion.status<-factor(full.data$invasion.status,
```

```
                    levels=c(0,1),
                    labels=c("MIN", "MAJ"))

set.seed(100)  #for repeatability

#In the trainControl function, the resampling method is "repeatedcv" (repeated cross-
validation)
#number = 10 indicates that there are10 folds in K-fold cross-validation
#repeats = 10 indicates that there are ten separate 10-fold cross-validations used as
the resampling scheme
#verboseIter is a logical for printing a training log
#returnData is a logical for saving the data into a slot called trainingData
#summaryFunction provides a ROC  AUC summary statistics

tc<-trainControl(method="repeatedcv", number=10,
            repeats=10, summaryFunction=twoClassSummary,
            verboseIter = T, returnData = T,
            classProb=T, savePredictions = T)

LR.second.model<-train(as.factor(invasion.status)~es1+es2+es3+
                  es4+es5+es6+es7+es8+es9+es10+es11+es12+
                  es13+es14+es15+es16+es17+es18+es19+es20+
                  es21+es22+es23+impg1+impg2+impn1+impn2+impn3+
                  impn4+impn5+impn6+impa1+impa2+impa3+impa4+impp1+
                  impp2+impp3+impp4+impp5+impp6,
                metric="ROC", method="glm",
                family="binomial", trControl=tc,
                data=full.data)

#model coefficients for independent variables
LR.second.model$finalModel

# training log saved from the returnData argument = TRUE
LR.second.model$trainingData

# cross validation summary statistics
LR.second.model

### RF with 10-fold CV on data###
RF.second.model<-train(as.factor(invasion.status)~.,
              metric="ROC", method="rf",
              importance=T, proximity=F,
              ntree=1000, trControl=tc,
              data=full.data)

RF.second.model$finalModel
```

```
RF.second.model$trainingData

RF.second.model

###ROC Plots###
LR.seond.model.pred<-predict(LR.second.model, full.data$invasion.status,
type="prob")
RF.second.model.pred<-predict(RF.second.model, full.data$invasion.status,
type="prob")
second.model.pred.LR<-prediction(LR.second.model.pred$OTH,
data$invasion.status)
second.model.perf.LR<-performance(second.pred.LR, "tpr", "fpr")
second.model.pred.RF<-prediction(RF.second.model.pred$OTH,
data$invasion.status)
second.model.perf.RF<-performance(second.model.pred.RF, "tpr", "fpr")

par(family="serif", fig=c(0.0,1,0.0,1),
   mar=par()$mar+c(2,2,2,13), xpd=TRUE,
   cex.lab=1.2, lwd=2, pty="m", font.axis=2)

plot(second.model.perf.LR,
    main="Logistic Regression and Random Forests",
    lty=1, col="red")
plot(second.model.perf.RF,
    xlab = "False Positive Rate",
    ylab = "True Positive Rate", add=T, lty=3, col="blue")
legend(1.2, 1.0, c("Logistic Regression 0.723", "Random Forest  0.885"),
     xpd=TRUE,
     lty = c(1,3), col = c("red","blue"),
     bty="o")




########################Figure  2#####################################
##Data retrieved from previous calculations above
x <- as.factor(c("2-fold CV", "5-fold CV", "10-fold CV"))

sweet <- c(0.732, 0.8257, 0.862) #lr A
tart <- c(0.9415, 0.9435, 0.943) #rf A
sweet2 <- c(0.638, 0.6956, 0.730) #lr B
tart2 <- c(0.87045, 0.8798, 0.877) #rf B

###Plotting###
par(family="serif", #par(mar = c(5, 5, 4, 2)),
```

```
   xpd = T, mar = par()$mar + c(0,0,0,7),
   xpd=TRUE,
   cex.lab=1.2, lwd=2, pty="m", font.axis=2)

plot(seq_along(x), xlab="k-fold CV", ylab="ROC AUC", sweet, ylim=c(0.600, 1),
lty=1, col="red", type="b", xaxt="n")
##axis(1, at=seq_along(x), labels=c("2-fold CV", "5-fold CV", "10-fold CV"))
par(new=TRUE)
plot(seq_along(x),tart, ylim=c(0.600, 1), xlab="k-fold CV", ylab="ROC AUC",
type="b", add=TRUE, col="blue", xaxt="n")

par(new=TRUE)
plot(seq_along(x),sweet2, ylim=c(0.600, 1), xlab="k-fold CV", ylab="ROC AUC",
lty=2, type="b", add=TRUE, col="red", xaxt="n")

par(new=TRUE)
plot(seq_along(x),tart2, ylim=c(0.600, 1), xlab="k-fold CV", ylab="ROC AUC",
lty=2, type="b", add=TRUE, col="blue", xaxt="n")

axis(1, at=seq_along(x), labels=c("2-fold CV", "5-fold CV", "10-fold CV"))


legend(3.2, 1.0, c("LR Model A", "LR Model B", "RF Model A", "RF Model B"),
      xpd=TRUE, lty =c(1,2,1,2),col=c("red","red", "blue", "blue"),
      bty="o")



##############################Random forest model#########################
data <- read.csv("filename.csv", header = TRUE)

dataabbr<-data[,c("invasion.status",
                  "es1", "es2", "es3","es4", "es5","es6","es7",
                  "es8","es9","es10","es11","es12","es13","es14",
                  "es15","es16","es17","es18","es19","es20","es21",
                  "es22","es23", "impg1", "impg2", "impn1", "impn2",
                  "impn3", "impn4", "impn5", "impn6", "impa1", "impa2",
                  "impa3", "impa4", "impp1", "impp2", "impp3", "impp4",
                  "impp5", "impp6")]
str(dataabbr)
dataabbr$invasion.status <- as.factor(dataabbr$invasion.status)
table(dataabbr$invasion.status)
options(max.print=900000)

# Data Partition
set.seed(123)
```

```
ind <- sample(2, nrow(dataabbr), replace = TRUE, prob = c(0.7, 0.3))
train <- dataabbr[ind==1,]
test <- dataabbr[ind==2,]
options(max.print=900000)
train
test
write.csv(train, "traindatasetwra.csv")
write.csv(test, "testdatasetwra.csv")

library(randomForest) # Random Forest
set.seed(222)
rf <- randomForest(invasion.status~., data=train,
                   ntree = 1000,
                   mtry = 6,
                   importance = TRUE,
                   proximity = TRUE)
print(rf) #OOB estimate error rate: 22.3% for ntree=1000 mtry=6
        #   1  2  3 class.error
############1 41  8  1   0.1800000
############2 10 34  7   0.3333333
############3  0  7 40   0.1489362

attributes(rf)
rf$confusion
rf$err.rate

# Prediction & Confusion Matrix - train data
library(caret)
p1 <- predict(rf, train)
confusionMatrix(p1, train$invasion.status)
###Confusion Matrix and Statistics

########Reference
########Prediction  1  2  3
##################1 50  1  0
##################2  0 50  0
##################3  0  0 47



# # Prediction & Confusion Matrix - test data
p2 <- predict(rf, test)
confusionMatrix(p2, test$invasion.status)
#Confusion Matrix and Statistics:
```

```
#Reference
#Prediction  1  2  3
##########1 15  3  1
##########2  3  6  3
##########3  0  7 18

##########Overall Statistics

##########Accuracy : 0.6964
##########95% CI : (0.559, 0.8122)
##########No Information Rate : 0.3929
##########P-Value [Acc > NIR] : 4.143e-06
#Statistics by Class:
                #Class: 1 Class: 2 Class: 3
#Sensitivity          0.8333   0.3750   0.8182
#Specificity          0.8947   0.8500   0.7941
##Sensitivity(True positive rate) for Class 2 is really low (0.3750)



# Variable Importance
varImpPlot(rf,
        sort = T,
        n.var = 10, type=1, labels= c("Change ecosystem processes and parameters that affect other species", "Climbing or smothering growth form", "Reduces crop/product yield", "Minimum generation time", "Propagules likely to disperse in trade as contaminants or hitchhikers", "Lowers commodity value", "Number of natural dispersal vectors", "Weed status in production systems", "Weed status in natural systems", "Status/invasiveness outside native range"))



importance(rf)
###############Actual variable importance values for each class
################ 1         2         3 MeanDecreaseAccuracy MeanDecreaseGini
#es1   20.9077410 -2.0803047 19.3888251          22.9246610       9.19473316
#es2    0.0000000  0.0000000  0.0000000           0.0000000       0.01831460
#es3    1.8545260 -1.3865015  0.1108207           0.3154562       1.45901784
#es4   -3.1687545 -2.4245899 -1.8566209          -4.4761316       1.13815385
#es5   13.8174636  7.3215009  1.4018801          14.5241891       2.62737365
#es6    9.5839278 -3.2569912  8.3396054           9.7436463       1.47830128
#es7   -1.7284327 -1.0955790  1.1595379          -1.4648598       0.16808220
#es8    0.5583773 -0.5903296 -1.5353788          -1.0719336       0.31894988
#es9   -1.0972092  7.4534105  6.5055382           8.2661539       1.16109793
#es10  -2.4543880 -2.2503793  0.0000000          -3.1426419       0.42391230
#es11  -0.5110099  0.3727102  8.2991278           4.0427577       2.94826017
```

```
#es12  -4.0385710 -0.4334052  1.6203737         -3.4773321       0.70256893
#es13  15.9931308 -2.9257367  6.9632562         13.8320315       3.83445574
#es14   1.3045280 -1.6023560 16.3728518         11.4404989       4.34396400
#es15   8.9363975 -4.6687640  8.6962208          7.5826949       3.99227939
#es16  17.2796782 -3.6074208  7.9790571         13.6078543       3.84532243
#es17   9.7445245  4.8837234 17.5885127         18.0902551       6.11650676
#es18  -3.3808874  6.1531682 10.8406064          9.3330577       3.57144825
#es19   3.7496907  4.2083000 10.7431437         10.3355339       3.10193918
#es20   5.6865176  4.9192707  6.2409152          8.5782756       1.26109297
#es21   0.5448779 -0.6127891 -0.8657698         -0.4682052       1.06407360
#es22  12.3026495 -1.5326339  3.4367479          8.9389699       3.41825333
#es23   9.5216798  3.8165194  2.0717154          9.5964934       3.35591664
#impg1  0.9759378 -2.5087859 -0.2370607         -0.9259242       0.86774170
#impg2  0.0000000  0.0000000  0.0000000          0.0000000       0.06444133
#impn1  7.2431428 -1.1342385 12.1042891         12.2734430       1.64126082
#impn2  3.2571554 -2.2350409  7.9470701          6.1565847       1.25842775
#impn3 14.1738198 -6.1217933  5.6572710         11.4371387       1.84071123
#impn4  6.4249253 -2.6127886  5.4033326          6.0334833       1.31054863
#impn5  6.3930894 -3.1664783  9.4223193          7.8751462       1.79243584
#impn6 21.0503867  5.3872449  7.5793567         20.7441151       6.47437422
#impa1 -2.3400687  1.4857470 -1.1570218         -1.1117846       0.60499203
#impa2  6.3899186 -2.9985668  0.5829608          3.5679138       0.49378987
#impa3  5.5049030 -0.6669316 -0.9041189          3.3918606       0.61175650
#impa4  8.9244294 -2.1366781  6.8339701          8.5488364       3.18330280
#impp1 11.3405631  1.5549458  9.3559311         12.6102414       2.60334526
#impp2 11.7037795  0.1901616 13.6439690         15.0676896       3.35650684
#impp3  7.8231436 -1.3822586  3.4240562          6.8614587       1.41561606
#impp4  4.4052873  2.1898047  4.7343212          6.5796319       0.42345847
#impp5 -0.4323909 -2.9337907  2.1543595         -0.8728398       1.23916670
#impp6 19.8966966 -5.6865388 16.7956142         20.1448025       6.14611695

varUsed(rf)
####give the number of time each variable has occurred in the random forest
####es2 (the second variable) has only occurred a total of 22 times in the random
forest
#[1] 2267   25 1116  957 1063  670  139  243  497  314 1780  469 1747 1847 1635
#[16] 1148 2324 1809 1740  387  708 1772 1758  662   38  518  484  739  565  698
#[31] 1452  365  263  371 1191  585  603  497  199  907 1237

# Partial Dependence Plot
par(mfrow=c(3,3))
partialPlot(rf, train, es1, "1",xlab="Non-Invaders", main="es1",
ylab=expression(paste(Delta, "Fraction of Votes p(Y=K)")))
partialPlot(rf, train, es1, "2",xlab="Minor-Invaders", main="es1",
ylab=expression(paste(Delta, "Fraction of Votes p(Y=K)")))
```

```
partialPlot(rf, train, es1, "3",xlab="Major-Invaders", main="es1",
ylab=expression(paste(Delta, "Fraction of Votes p(Y=K)")))
partialPlot(rf, train, impn6, "1",xlab="Non-Invaders", main="impn6",
ylab=expression(paste(Delta, "Fraction of Votes p(Y=K)")))
partialPlot(rf, train, impn6, "2",xlab="Minor-Invaders", main="impn6",
ylab=expression(paste(Delta, "Fraction of Votes p(Y=K)")))
partialPlot(rf, train, impn6, "3",xlab="Major-Invaders", main="impn6",
ylab=expression(paste(Delta, "Fraction of Votes p(Y=K)")))
partialPlot(rf, train, impp6, "1",xlab="Non-Invaders", main="impp6",
ylab=expression(paste(Delta, "Fraction of Votes p(Y=K)")))
partialPlot(rf, train, impp6, "2",xlab="Minor-Invaders", main="impp6",
ylab=expression(paste(Delta, "Fraction of Votes p(Y=K)")))
partialPlot(rf, train, impp6, "3",xlab="Major-Invaders", main="impp6",
ylab=expression(paste(Delta, "Fraction of Votes p(Y=K)")))



########################################Figure 5#########################
####Adapted from Scott Chamberlin tutorial in R Open Sci
library("rWBclimate")
library("spocc")
library("plyr")
library("sp")
require(rWBclimate)
require("spocc")
dir.create("/path")
options(kmlpath="/path/kmlhist")
options(stringsAsFactors = FALSE)
usmex <- c(273:284, 328:365) #river basin IDs for Mexico and United States
str(usmex)
usmex.basin <- create_map_df(usmex)
str(usmex.basin)
## Download temperature data
temp.dat <- get_historical_temp(usmex, "decade")
temp.dat <- subset(temp.dat, temp.dat$year == 2000)
str(temp.dat)
write.csv(temp.dat, "temp.dathist.csv")
# Bind temperature data to map data frame
usmex.map.df <- climate_map(usmex.basin, temp.dat, return_map = F)
library(ggplot2)
splist <- c("Acanthospermum australe", "Abutilon megapotamicum", "Alternanthera
philoxeroides")
splist <-sort(splist)
splist
out <- occ(query=splist, from= "gbif", limit=500)
out <-fixnames(out, how="query")
```

```
out_df
write.csv(out_df, "out_df.csv")
out_df <- occ2df(out) #combine results from occ calls to a single data

library(taxize)
### grab common names
cname <- ldply(sci2comm(get_tsn(splist),
                        db = "itis", simplify = TRUE),
               function(x) { return(x[1]) })[, 2]

out_df <- out_df[order(out_df$name), ]
out_df <- out_df[!is.na(out_df$latitude), ]

str(out_df)
write.csv(out_df, "out_df_woNA.csv")
out_df$name
out_df <- out_df[out_df$latitude > 7, ]
str(out_df)
out_df$common <- rep(cname, table(out_df$name))
out_df$
  install.packages("extrafont")
install.packages("tidyverse")
library(tidyverse)
write.csv(out_df, "out_df_woNA_wolesslat7.csv")


usmex.map <- ggplot() +
  geom_polygon(data = usmex.map.df, aes(x = long, y = lat, group = group, fill =
data, alpha = 0.9)) +
  scale_fill_continuous("Average annual \n temp (°C): 1990-2000", low = "yellow",
high = "red") +
  guides(alpha = F) +
  theme_bw(12, base_family = "Times New Roman") +
  theme(axis.line = element_line(colour = "black",
                                 size = 1, linetype = "solid")) +
  theme(axis.text.x = element_text(face = "bold", colour = "black", size = 12),
        axis.text.y = element_text(face = "bold", colour = "black", size = 12)
  ) +
  xlab("Longtitude") +
  ylab("Latitude")

print(usmex.map)

usmex.map <- usmex.map +
  geom_point(data = out_df, aes(y = latitude, x = longitude, group = common, colour
= common)) +
```

```
  xlim(-125, -59) +
  ylim(5, 55) +
  scale_color_discrete(name = "Non-native plants",
                       labels = c("Paraguayan starbur", "Alligatorweed", "Trailing abutilon"))
+
  theme(legend.background = element_rect(colour= "black", fill="transparent",
size=.5, linetype="solid"))


print(usmex.map)

## Create a spatial polygon dataframe binding kml polygons to temperature
## data
temp_sdf <- kml_to_sp(usmex.basin, df = temp.dat)
### Now we can change the points to a spatial polygon:
library(sp)
library(maptools)
library(spocc)

occ_to_sp <- function(x, coord_string = "+proj=longlat +datum=WGS84",
just_coords = FALSE){
  points <- occ2df(x)
  # remove NA rows
  points <- points[complete.cases(points),]

  # check valid coords
  index <- 1:dim(points)[1]
  index <- index[(points$longitude < 180) & (points$longitude > -180) &
!is.na(points$longitude)]
  index <- index[(points$latitude[index] < 90) & (points$latitude[index] > -90) &
!is.na(points$latitude[index])]

  spobj <- sp::SpatialPoints(as.matrix(points[index,c('longitude','latitude')]),
proj4string = sp::CRS(coord_string))

  sp_df <- sp::SpatialPointsDataFrame(spobj, data =
data.frame(points[index,c('name',"prov")]))
  if (just_coords) spobj else sp_df
}



sp_points <- occ_to_sp(out)
str(sp_points)

tdat <- vector()
```

```
### Get averages
for (i in 1:length(splist)) {
  tmp_sp <- sp_points[which(sp_points$name == splist[i]), ]
  tmp_t <- over(tmp_sp, temp_sdf)$data
  tdat <- c(tdat, tmp_t)
}

### Assemble new dataframe
spDF <- data.frame(matrix(nrow = dim(sp_points)[1], ncol = 0))
spDF$species <- sp_points$name
spDF <- cbind(coordinates(sp_points), spDF)

### Alphebetically ordering points########
spDF <- spDF[order(spDF$species), ]

spDF$cname <- rep(cname, table(sp_points$name))
spDF$temp <- tdat
### Strip NA's
spDF <- spDF[!is.na(spDF$temp), ]

str(spDF)
write.csv(spDF, "spDF.csv")
## Create summary
summary_data <- ddply(spDF, .(cname), summarise, mlat = mean(latitude), mtemp =
mean(temp),
                  sdlat = sd(latitude), sdtemp = sd(temp))

str(summary_data)
write.csv(summary_data, "summary_data.csv")
ggplot(summary_data, aes(x = mlat, y = mtemp, label = cname)) +
  geom_text() +
  xlab("Mean Latitude") +
  ylab("Mean Temperature (C)") +
  theme_bw() +
  xlim(10, 50)

ggplot(spDF, aes(as.factor(cname), temp)) +
  geom_boxplot() +
  theme_bw(13) +
  ylab("Temperature") +
  xlab("Common Name") +
  theme(axis.text.x = element_text(angle = 45, hjust = 0.5, vjust = 0.5))
```

```
####################Figure 6#########################################
####Adapted from Scott Chamberlin tutorial in R Open Sci

library("rWBclimate")
library("spocc")
library("plyr")
library("sp")
require(rWBclimate)
require("spocc")
dir.create("/path")
options(kmlpath="/path/kmlmodel")
options(stringsAsFactors = FALSE)
usmex <- c(273:284, 328:365) #river basin IDs for Mexico and United States
usmex.basin <- create_map_df(usmex)
usa.dat <- get_model_temp("USA", "mavg", 2080, 2100)

###data.frame':          24 obs. of  7 variables:
#$ fromYear: num  2080 2080 2080 2080 2080 2080 2080 2080 2080 2080 ...
#$ toYear  : num  2099 2099 2099 2099 2099 ...
#$ gcm     : Factor w/ 15 levels "bccr_bcm2_0",..: 14 14 14 14 14 14 14 14 14 14 ...
#$ data    : num  -2.09 -1.47 2.9 9.14 15.23 ...
#$ scenario: chr  "a2" "a2" "a2" "a2" ...
#$ month   : int  1 2 3 4 5 6 7 8 9 10 ...
#$ locator : chr  "USA" "USA" "USA" "USA" ...
#usa.dat.bcc <- usa.dat[usa.dat$gcm == "bccr_bcm2_0", ]
usa.dat.had <- usa.dat[usa.dat$gcm == "ukmo_hadcm3", ]
write.csv(usa.dat.had, "usa.dat.had.csv")
###'data.frame':          24 obs. of  7 variables:
#$ fromYear: num  2080 2080 2080 2080 2080 2080 2080 2080 2080 2080 ...
#$ toYear  : num  2099 2099 2099 2099 2099 ...
#$ gcm     : Factor w/ 15 levels "bccr_bcm2_0",..: 14 14 14 14 14 14 14 14 14 14 ...
#$ data    : num  -2.09 -1.47 2.9 9.14 15.23 ...
#$ scenario: chr  "a2" "a2" "a2" "a2" ...
#$ month   : int  1 2 3 4 5 6 7 8 9 10 ...
#$ locator : chr  "USA" "USA" "USA" "USA" ...
str(usa.dat.had)
summary(usa.dat.had)

####historical temp####
hist.dat <- get_historical_temp("USA", "month") #monthly averages of temperatures
from 1901-2009
str(hist.dat)
#'data.frame':  12 obs. of  3 variables:
#$ month  : Factor w/ 12 levels "Jan","Feb","Mar",..: 1 2 3 4 5 6 7 8 9 10 ...
#$ data   : num  -5.662 -3.8577 0.0517 5.9262 11.9592 ...
#$ locator: chr  "USA" "USA" "USA" "USA" ...
```

```
write.csv(hist.dat, "hist.dat.csv")
str(hist.dat)
str(usa.dat.had)
hist.dat <- read.csv("hist.dat.csv", header = TRUE)

usa.dat.had$ID <- paste(usa.dat.had$scenario, usa.dat.had$gcm, sep = "-")

plot.df <- rbind(usa.dat.had, hist.dat)
str(hist.dat)
plot <- ggplot(usa.dat.had, aes(x = as.factor(month), y = data, group = ID, colour =
gcm,
                linetype = scenario)) + geom_point() + geom_path() +
  theme_classic(12, base_family = "Times New Roman") +
  theme(axis.line = element_line(colour = "black",
                           size = 1, linetype = "solid")) +
  theme(axis.text.x = element_text(face = "bold", colour = "black", size = 12),
       axis.text.y = element_text(face = "bold", colour = "black", size = 12)
  ) +
  ylab("Average temperature in degrees (°C)") +
  xlab("Month")


plot

plot <- plot +
  geom_line(data = hist.dat, aes(x = month, y = data, colour = "blue"),
          inherit.aes = FALSE) + geom_point() +
  scale_color_discrete(name = "Temperature projections",
                   labels = c("Historical (1901-2009)", "HadCM3 (2080-2100)")) +
  theme(legend.background = element_rect(colour= "black", fill="transparent",
size=.5, linetype="solid"))

print(plot)
```

Appendix D. Variable importance for each class for the random forest model. Variables presented here as the same as the variables presented in Appendix A.

	Non-invader	Minor-invader	Major-invader	Mean Decrease Accuracy	# of times variable occurred in the random forest
Es1	20.90	-2.08	19.38	22.92	2267
Es2	0	0	0	0	25
Es3	1.85	-1.38	0.11	0.315	1116
Es4	-3.16	-2.42	-1.85	-4.47	957
Es5	13.81	7.32	1.40	14.52	1063
Es6	9.58	-3.25	8.33	9.74	670
Es7	-1.72	-1.09	1.15	-1.46	139
Es8	0.55	-0.59	-1.53	-1.07	243
Es9	-1.09	7.45	6.50	8.26	497
Es10	-2.45	-2.25	0	-3.14	314
Es11	-0.51	0.37	8.29	4.04	1780
Es12	-4.03	-0.43	1.62	-3.47	469
Es13	15.99	-2.92	6.96	13.83	1747
Es14	1.30	-1.60	16.37	11.44	1847
Es15	8.93	-4.66	8.69	7.58	1635
Es16	17.27	-3.60	7.97	13.60	1148
Es17	9.74	4.88	17.58	18.09	2324

Es18	-3.38	6.15	10.84	9.33	1809
Es19	3.74	4.20	10.74	10.33	1740
Es20	5.68	4.91	6.24	8.57	387
Es21	0.54	-0.61	-0.86	-0.46	708
Es22	12.30	-1.53	3.43	8.93	1772
Es23	9.52	3.81	2.07	9.59	1758
Impg1	0.97	-2.50	-0.23	-0.92	662
Impg2	0	0	0	0	38
Impn1	7.24	-1.13	12.10	12.27	518
Impn2	3.25	-2.23	7.94	6.15	484
Impn3	14.17	-6.12	5.65	11.43	739
Impn4	6.42	-2.61	5.40	6.03	565
Impn5	6.39	-3.16	9.42	7.87	698
Impn6	21.05	5.38	7.57	20.74	1452
Impa1	-2.34	1.48	-1.15	-1.11	365
Impa2	6.38	-2.99	0.58	3.56	263
Impa3	5.50	-0.66	-0.90	3.39	371
Impa4	8.92	-2.13	6.83	8.54	1191
Impp1	11.34	1.55	9.35	12.61	585
Impp2	11.70	0.19	13.64	15.06	603
Impp3	7.82	-1.38	3.42	6.86	497

Impp4	4.40	2.18	4.73	6.57	199
Impp5	-0.43	-2.93	2.15	-0.87	907
Impp6	19.89	-5.68	16.79	20.14	1237

Bibliography

Agresti, A. (2014). Categorical Data Analysis. Hoboken: Wiley.

Anagnostakis, S. (1987). Chestnut blight: The classical problem of an introduced pathogen. *Mycologia, 79*(1), 23-23. doi:10.2307/3807741

Babyak, M. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-Type models. *Psychosomatic Medicine, 66*(3), 411-421.

Badeck, F., Bondeau, A., Böttcher, K., Doktor, D., Lucht, W., Schaber, J., & Sitch, S. (2004). Responses of spring phenology to climate change. *New Phytologist, 162*(2), 295-309. doi:10.1111/j.1469-8137.2004.01059.x

Baker, H.G. (1991). The continuing evolution of weeds. *Economic Botany, 45*(4), 445-449. doi:10.1007/BF02930705

Bailey, L., & Bailey, E. (1930). *Hortus : A concise dictionary of gardening, general horticulture and cultivated plants in north america*. New York: Macmillan.

Beaumont, L.J., Gallagher, R.V., Leishman, M.R., Hughes, L., & Downey, P.O. (2014). How can knowledge of the climate niche inform the weed risk assessment process? a case study of chrysanthemoides monilifera in australia. *Diversity and Distribution,s* 20 (6): 613–25. doi:10.1111/ddi.12190.

Bellard, C., Leroy, B., Thuiller, W., Rysman, J., Courchamp, F., & Collins, S. (2016). Major drivers of invasion risks throughout the world. *Ecosphere, 7*(3). doi:10.1002/ecs2.1241

Ben, H., Chai, A., Shi, Y., Xie, X., Li, B., Qu, H. & Gao, W. (2016). New host record

of myrothecium roridum causing leaf spot on abutilon megapotamicum from china. *Journal of Phytopathology, 164*(7-8), 563-566. doi:10.1111/jph.12439

Boudell, J.A., & Stromberg, J.C. (2015). Impact of nitrate enrichment on wetland and dryland seed germination and early seedling development. *Journal of Vegetation Science, 26*(3): 452–63. doi:10.1111/jvs.12258.

Boulesteix, A., Janitza, S., Kruppa, J., & König, I. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,2*(6), 493-507. doi:10.1002/widm.1072

Bradley, B.A., Blumenthal, D.M., Wilcove, D.S., & Ziska, L.H. (2010). Predicting plant invasions in an era of global change. *Trends In Ecology & Evolution, 25*(5), 310-8. doi:10.1016/j.tree.2009.12.003

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Bridges DC (ed) (1992) Crop losses due to weeds in the United States—1992. Weed Science Society of America, Champaign

Broennimann, O., Mráz, P., Petitpierre, B., Guisan, A., Müller-Schärer, H., & Pearson, R. (2014). Contrasting spatio-temporal climatic niche dynamics during the eastern and western invasions of spotted knapweed in North America. *Journal of Biogeography, 41*(6), 1126-1136. doi:10.1111/jbi.12274

Brownlee, J. (2016). Overfitting and underfitting with machine learning algorithms - machine learning mastery. Retrieved December 08, 2016, from

http://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Caley, P., & Kuhnert, P. (2006). Application and evaluation of classification trees for screening unwanted plants. *Austral Ecology, 31*(5), 647-655. doi:10.1111/j.1442-9993.2006.01617.x

Scott Chamberlain (2017). spocc: Interface to Species Occurrence Data Sources. R package version 0.7.0. https://CRAN.R-project.org/package=spocc

Chen, G.Q., He, Y., & Qiang, S. (2013). Increasing seriousness of plant invasions in croplands of Eastern China in relation to changing farming practices: A case study. *Plos One, 8*(9). doi:10.1371/journal.pone.0074136.

Crawford, J.A., Olson, R.A., West, N.E., Mosley, J.C., Schroeder, M.A., Whitson, T.D., Miller, R.F., Gregg, M.A., & Boyd, C.S. (2004). Ecology and management of sage-grouse and sage-grouse habitat. *Rangeland Ecology & Management, 57*(1), 2-19. doi:10.2111/1551-5028(2004)057[0002:EAMOSA]2.0.CO;2

Chai, S., Zhang, J., Nixon, A., Nielsen, S., & Li, B. (2016). Using risk assessment and habitat suitability models to prioritise invasive species for management in a changing climate. *Plos One, 11*(10), 0165292. doi:10.1371/journal.pone.0165292

U. Cubasch (Germany), X. Dai (China), Y. Ding (China), D.J. Griggs (UK), B. Hewitson (South Africa), J.T. Houghton (UK), I. Isaksen (Norway), T. Karl (USA), M. McFarland (USA), V.P. Meleshko (Russia), J.F.B. Mitchell (UK), M. Noguer (UK), B.S. Nyenzi (Tanzania), M. Oppenheimer (USA), J.E. Penner (USA), S. Pollonais (Trinidad and Tobago), T. Stocker (Switzerland),

K.E. Trenberth (USA), 2001: The technical summary of the Working Group I Report. In: *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change* [Houghton, J.T., Y. Ding, D.J. Griggs, M. Noguer, P.J. van der Linden, X. Dai, K. Maskell, and C.A. Johnson (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 881pp.

Cutler, D.R., Edwards, T.C. Jr, Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., & Lawler, J.J. (2007). Random forests for classification in ecology. *Ecology, 88*(11), 2783-92.

Dawson, W., Burslem, D., & Hulme, P. (2009). The suitability of weed risk assessment as a conservation tool to identify invasive plant threats in east african rainforests. *Biological Conservation, 142*(5), 1018-1024. doi:10.1016/j.biocon.2009.01.013

Des Marais, D.L., Hernandez, K.M., & Juenger, T.E. (2013). Genotype-by-environment interaction and plasticity: Exploring genomic responses of plants to the abiotic environment. *Annual Review of Ecology, Evolution, and Systematics* 44:5– 29.

Driscoll, D.A., Catford, J.A., Barney, J.N., Hulme, P.E., Inderjit, Martin, T.G., Pauchard, A., Pyšek, P., Richardson, D.M. Riley, S., & Visser, V. (2014). New pasture plants intensify invasive species risk. *Proceedings of the National Academy of Sciences of the United States of America,111*(46), 16622-7. doi:10.1073/pnas.1409347111

Eiswerth, M., Darden, T., Johnson, W., Agapoff, J., & Harris, T. (2005). Input–output modeling, outdoor recreation, and the economic impacts of weeds. *Weed Science, 53*(1), 130-137. doi:10.1614/WS-04-022R

Elith, J., Leathwick, J.R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology, 77*(4), 802-13. doi: 10.1111/j.1365-2656.2008.01390.x

Ellstrand, N. C., & Schierenbeck, K. A. (2000). Hybridization as a Stimulus for the Evolution of Invasiveness in Plants? *Proceedings of The National Academy Of Sciences Of The United States Of America*, *97*(13), 7043-7050.

Evans, J., & Cushman, S. (2009). Gradient modeling of conifer species using random forests. *Landscape Ecology, 24*(5), 673-683.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters, 27*(8), 861-874. doi:10.1016/j.patrec.2005.10.010

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist, 29*(5), 1189-1232. doi:10.1214/aos/1013203451

Genuer, R., Poggi, J., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters, 31*(14), 2225-2236. doi:10.1016/j.patrec.2010.03.014

Gislason, P., Benediktsson, J., & Sveinsson, J. (2006). Random forests for land cover classification. *Pattern Recognition Letters, 27*(4), 294-300. doi:10.1016/j.patrec.2005.08.011

Gordon, D.R., Onderdonk, D.A., Fox, A.M. & Stocker, R.K. (2008a). Consistent accuracy of the Australian weed risk assessment system across varied geographies. *Diversity and Distributions*, 14, 234–242.

Gordon, D., Onderdonk, D., Fox, A., Stocker, R., & Gantz, C. (2008b). Predicting invasive plants in Florida using the Australian weed risk assessment. *Invasive Plant Science and Management, 1*(2), 178-195. doi:10.1614/IPSM-07-037.1

Greater Sage Grouse Protection and Recovery Act of 2016, H.R. 4739, 114th Cong. (2015-2016). Retrieved from ProQuest Congressional Database.

Griffin, G.F., Smith, D.M.S., Morton, S.R., Allan, G.E., Masters, K.A., & Preece, N. (1989). Status and implications of the invasion of tamarix (Tamarix aphylla) on the Finke River, Northern Territory, Austraila. *Journal of Environmental Management, 29*(4), 297-315.

Griffiths, R., Schloesser, D., Leach, J., & Kovalak, W. (1991). Distribution and dispersal of the zebra mussel (Dreissena polymorpha) in the great lakes region. *Canadian Journal of Fisheries and Aquatic Sciences, 48*(8), 1381-1388. doi:10.1139/f91-165

Gustafson, S., & Wang, D. (2002). Effects of agricultural runoff on vegetation composition of a priority conservation wetland, Vermont, USA. *Journal of Environmental Quality, 31*(1): 350–57. doi:10.2134/jeq2002.0350.

Guttery, M.R., Messmer, T.A., Brunson, M.W., Robinson, J.D., & Dahlgren, D.K. (2016). Declining populations of greater sage-grouse: Hunter motivations when numbers are low. *Animal Conservation*, *19*(1), 26-34. doi:10.1111/acv.12213

Hahn, M.A., Buckley, Y.M., Müller-Schärer, H., & Gurevitch, J. (2012). Increased population growth rate in invasive polyploid *Centaurea stoebe* in a common garden. *Ecology Letters, 15*(9), 947-954. doi:10.1111/j.1461-0248.2012.01813.x

Han, H., Guo, X., Yu, H., & 7th IEEE International Conference on Software Engineering and Service Science ICSESS 2016 7th IEEE International Conference on Software Engineering and Service Science, ICSESS 2016 7 2016 08 26 - 2016 08 28. (2017). Variable selection using mean decrease accuracy and mean decrease gini based on random forest. *Proceedings of the Ieee International Conference on Software Engineering and Service Sciences, Icsess,219-224*, 219-224. doi:10.1109/ICSESS.2016.7883053

Hart E. (2014). _rWBclimate: A package for accessing World Bank climate data_. R package version 0.1.4.99, <URL: http://www.github.com/ropensci/rwbclimate>.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed). ed., Springer series in statistics). New York: Springer.

He, R., Kim, M.J., Nelson, W., Balbuena, T.S., Kim, R., Kramer, R., Crow, J.A., May, G.D., Thelen, J.J., Soderlund, C.A., & Gang, D.R. (2012). Next-generation sequencing-based transcriptomic and proteomic analysis of the common reed, phragmites australis (Poaceae), reveals genes involved in invasiveness and rhizome specificity. *American Journal of Botany 99*(2): 232–47. doi:10.3732/ajb.1100429.

Hengl, T., Heuvelink, G.B., Kempen, B., Leenaars, J., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., Tondoh, J.E. (2015). Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *Plos One, 10*(6), 0125814. doi:10.1371/journal.pone.0125814

Herron, P., Martine, C., Latimer, A., & Leicht-Young, S. (2007). Invasive plants and their ecological strategies: Prediction and explanation of woody plant invasion

in new england. *Diversity and Distributions, 13*(5), 633-644. doi:10.1111/j.1472-4642.2007.00381.x

Hess, L. (2015). A new look at the endangered species act and its effects on genetic diversity. *Journal Of Avian Medicine And Surgery, 29*(4), 354-359. doi:10.1647/1082-6742-29.4.354

Higgins, S.I., & Richardson, D.M. (2014). Invasive plants have broader physiological niches. *Proceedings Of The National Academy Of Sciences Of The United States Of America, 111*(29), 10610-4. doi:10.1073/pnas.1406075111

Barkley, T., Holm, L., Pancho, J., Herberger, J., & Plucknett, D. (1979). A geographical atlas of world weeds. *Brittonia, 32*(2), 127-127. doi:10.2307/2806777

Hoopes, M.F., Marchetti, M.P., & Lockwood, J.L. (2013). Invasion Ecology. Wiley.

Hull-Sanders, H.M., Johnson, R.H., Owen, H.A., & Meyer, G.A. (2009). Effects of polyploidy on secondary chemistry, physiology, and performance of native and invasive genotypes of Solidago gigantea (Asteraceae). *American Journal Of Botany, 96*(4), 762-770.

Hulme, P.E. (2012). Weed risk assessment: A way forward or a waste of time?" *Journal of Applied Ecology 49*(1): 10–19. doi:10.1111/j.1365-2664.2011.02069.x.

Hulme, P.E. (2015). Invasion pathways at a crossroad: policy and research challenges for managing alien species introductions. *Journal of Applied Ecology NOVOL*, n/a – n/a. doi:10.1111/1365-2664.12470.

Hulme, P. E., Pyšek, P., Jarošík, V., Pergl, J., Schaffner, U., & Vilà, M. (2013). Bias and error in understanding plant invasion impacts. *Trends In Ecology & Evolution, 28*(4), 212-218. doi:10.1016/j.tree.2012.10.010

International Plant Protection Convention. November 17, 1997. Food and Agriculture Organization.

IPCC, 2001: *Climate Change 2001: Synthesis Report. A Contribution of Working Groups I, II, and III to the Third Assessment Report of the Intergovernmental Panel on Climate Change* [Watson, R.T. and the Core Writing Team (eds.)]. Cambridge University Press, Cambridge, United Kingdom, and New York, NY, USA, 398 pp.

IPCC, 2007: *Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Core Writing Team, Pachauri, R.K and Reisinger, A. (eds.)]. IPCC, Geneva, Switzerland, 104 pp.

IPCC, 2014: *Climate Change 2014: Synthesis Report*. *Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland, 151 pp.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York: Springer.

Jiménez, M., Jaksic, F., Armesto, J., Gaxiola, A., Meserve, P., Kelt, D., & Gutiérrez, J. (2011). Extreme climatic events change the dynamics and invasibility of semi-arid annual plant communities. *Ecology Letters, 14*(12), 1227-1235. doi:10.1111/j.1461-0248.2011.01693.x

Julien, M., Skarratt, B., & Maywald, G. (1995). Potential geographical distribution of alligator weed and its biological control by agasicles hygrophila. *Journal of Aquatic Plant Management, 33*(7), 55-55.

Keane, R., & Crawley, M. (2002). Exotic plant invasions and the enemy release hypothesis. *Trends in Ecology & Evolution, 17*(4), 164-170. doi:10.1016/S0169-5347(02)02499-0

Keese, P.K., Robold, A.V., Myers, R.C., Weisman, S., & Smith, J. (2014). Applying a weed risk assessment approach to GM crops. *Transgenic Research*, *23*(6), 957-69. doi:10.1007/s11248-013-9745-0

Klironomos, J. (2002). Feedback with soil biota contributes to plant rarity and invasiveness in communities. *Nature, 417*(6884), 67-70. doi:10.1038/417067a

Koncki, N., & Aronson, M. (2015). Invasion risk in a warmer world: Modeling range expansion and habitat preferences of tree nonnative aquatic invasive plants. *Invasive Plant Science and Management, 8*(4), 436-449. doi:10.1614/IPSM-D-15-00020.1

Koop, A.L., Larry, F., Leslie, P.N., & Barney, P.C. (2012). Development and validation of a weed screening tool for the United States. *Biological Invasions, 14*(2): 273–94. doi:10.1007/s10530-011-0061-4.

Kowarik, I. (1995). Time lags in biological invasions with regard to the success and failure of alien species. In: Pyšek, P., Rejmánek, M., Wade, M. (eds) Plant invasions—general aspects and special problems. SPB Academic Publishing, Amsterdam, pp 15-38.

Lambert, A.M., Dudley, T.L., & Robbins, J. (2014). Nutrient enrichment and soil conditions drive productivity in the large-statured invasive grass arundo donax.

Aquatic Botany, 112. Elsevier B.V.: 16–22. doi:10.1016/j.aquabot.2013.07.004.

Li, J., Jin, Z., Song, W., & Herrera-Estrella, L. (2012). Do native parasitic plants cause more damage to exotic invasive hosts than native non-Invasive hosts? an implication for biocontrol. *Plos One, 7*(4), 34577. doi:10.1371/journal.pone.0034577

Li, J., Yang, B., Yan, Q., Zhang, J., Yan, M., & Li, M. (2015). Effects of a native parasitic plant on an exotic invader decrease with increasing host age. *Aob Plants, 7*. doi:10.1093/aobpla/plv031

Li, X., & Zhao, H. (2009). Weighted random subspace method for high dimensional data classification. *Statistics and Its Interface*, *2*(2), 153–159.

Liaw, A. (2015). Package "randomForest". https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Liaw A. & Wiener, M. (2002). Classification and regression by random-Forest. *R News, 2*(3). 18-22.

Lind, E.M., & Parker, J.D. (2010). Novel weapons testing: Are invasive plants more chemically defended than native plants? *Plos One, 5*(5). doi:10.1371/journal.pone.0010429

Linder HP, & Barker NP. (2014). Does polyploidy facilitate long-distance dispersal? *Annals of Botany, 113*(7), 1175-83. doi:10.1093/aob/mcu047

Liu, J., Dong, M., Miao, S., Li, Z., Song, M., & Wang, R. (2006). Invasive alien plants in china: Role of clonality and geographical origin. *Biological Invasions, 8*(7), 1461-1470.

Lockwood, J.L., Hoopes, M.F., & Marchetti, M.P. (2007). *Invasion ecology.* Malden, MA: Blackwell Pub.. http://catdir.loc.gov/catdir/toc/ecip0611/2006009954.html

Machado, G., Mendoza, M., & Corbellini, L. (2015). What variables are important in predicting bovine viral diarrhea virus? a random forest approach. *Veterinary Research, 46*, 85-85. doi:10.1186/s13567-015-0219-7

Magarey, R., Borchert, D., & Schlegel, J. (2008). Global plant hardiness zones for phytosanitary risk analysis. *Scientia Agricola, 65*(Spe), 54-59. doi:10.1590/S0103-90162008000700009

Magarey, R., Newton, L., Hong, S.C., Takeuchi, Y., Christie, D., Jarnevich, C., Kohl, L., Damus, M., Higgins, S.I., Millar, L., Castro, K., West, A., Hastings, J., Cook, G., Kartesz, J., & Koop, A. (2017). Comparison of four modeling tools for the prediction of potential distribution for non-indigenous weeds in the United States. *Biological Invasions*. doi:10.1007/s10530-017-1567-1

Mainali K.P., Warren D.L., Dhileepan K., Mcconnachie A., Strathie L., Hassan G., Karki, D., Shrestha, B.B., & Parmesan, C. (2015). Projecting future expansion of invasive species: Comparing and improving methodologies for species distribution modeling. *Global Change Biology*, *21*(12), 4464-4480. doi:10.1111/gcb.13038

Maindonald, J. (2010). Parametric vs nonparametric models for discrimination and classification. Retrieved from http://maths-people.anu.edu.au/~johnm/r-book/xtras/classif-notes.pdf

Maloy, O. (1997). White pine blister rust control in North America: A case history. *Annual Review of Phytopathology*, *35*(1), 87-109. doi:10.1146/annurev.phyto.35.1.87

Mangla, S., I., & Callaway, R. (2008). Exotic invasive plant accumulates native soil pathogens which inhibit native plants. *Journal of Ecology*, *96*(1), 58-67.

Matsuki, K., Kuperman, V., & Van Dyke, J.A. (2016). The Random forests statistical technique: An examination of its value for the study of reading. *Scientific Studies Of Reading*, *20*(1), 20-33. doi:10.1080/10888438.2015.1107073

Met Office climate prediction model: HadCM3. (2018). *Met Office*. Retrieved 2018, from https://www.metoffice.gov.uk/research/modelling-systems/unified-model/climate-models/hadcm3

Mooney, H.A., & International Council for Science. (2005). *Invasive alien species : A new synthesis* (SCOPE, 63; SCOPE report, 63). Washington, DC: Island Press. http://catdir.loc.gov/catdir/toc/ecip055/2004029420.html

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2015). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, (3), Mpv024. doi:10.1093/pan/mpv024

Naeem S., Thompson, L.J., Lawler, S.P., Lawton, J.H. & Woodfin, R.M. (1994). Declining biodiversity can alter the performance of ecosystems. *Nature*, 368:734-737.

Nakićenović, N., & Intergovernmental Panel on Climate Change. Working Group III. (2000). *Special report on emissions scenarios: A special report of working group III of the intergovernmental panel on climate change*. Cambridge: Cambridge University Press.

NatureServe. (2009). NatureServe Explorer. Online Database. http://www.natureserve.org/explorer/servlet/NatureServe?post_processes=PostReset&loadTemplate=nameSearchSpecies.wmt&Type=Reset.

Neckles, H. (2015). Loss of eelgrass in Casco Bay, Maine, linked to green crab disturbance. *Northeastern Naturalist, 22*(3), 478-500. doi:10.1656/045.022.0305

Ni, Z., Kim, E., Ha, M., Lackey, E., Liu, J., Zhang, Y., Sun, Q., & Chen, Z. (2009). Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature, 457*(7227), 327-331. doi:10.1038/nature07523

Paini, D.R., Sheppard, A.W., Cook, D.C., De Barro, P.J., Worner, S.P. & Thomas, M.B. (2016). Global threat to agriculture from invasive species. *Proceedings of the National Academy of Sciences of the United States of America, 113*(27), 7575-9. doi:10.1073/pnas.1602205113

Pandit, M.K., Pocock, M.J.O., & Kunin, W.E. (2011). Ploidy influences rarity and invasiveness in plants. *Journal of Ecology*, *99*(5).

Pandit, M.K., White, S.M., & Pocock, M.J.O. (2014). The contrasting effects of genome size, chromosome number and ploidy level on plant invasiveness: A global analysis. *New Phytologist, 203*(2), 697-703. doi:10.1111/nph.12799

Parepa, M., Fischer, M., Krebs, C., & Bossdorf, O. (2014). Hybridization increases invasive knotweed success. *Evolutionary Applications, 7*(3), 413-420. doi:10.1111/eva.12139

Pearman P.B., Guisan A., Broennimann O., & Randin C.F. (2008). Niche dynamics in space and time. *Trends in Ecology & Evolution, 23*(3), 149-58. doi:10.1016/j.tree.2007.11.005

Peel, M.C., Finlayson, B.L., & McMahon, T.A. (2007). Updated world map of the köppen-Geiger climate classification. *Hydrology and Earth System Sciences, 11*(5), 1633-1644.

Peters, J., Baets, B.D., Verhoest, N.E.C., Samson, R., Degroeve, S., Becker, P.D., & Huybrechts, W. (2007). Random forests as a tool for ecohydrological distribution modelling. *Ecological Modelling, 207*(2-4), 304-318. doi:10.1016/j.ecolmodel.2007.05.011

Pheloung, P., Williams, P., & Halloy, S. (1999). A weed risk assessment model for use as a biosecurity tool evaluating plant introductions. *Journal of Environmental Management, 57*(4), 239-251.

Pimentel, D., Zuniga, R., & Morrison, D. (2005). Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics, 52*(3), 273-288.

Pimentel, D., Lach, L., Zuniga, R., & Morrison, D. (2000). Environmental and economic costs of nonindigenous species in the United States. *BioScience, 50*(1), 53. doi:10.1641/0006-3568(2000)050[0053:EAECON]2.3.CO;2

de Poorter, M., Browne, M., Lowe, S., & Clout, M. (2005). The ISSG global invasive species database and other aspects of an early warning system. *SCOPE Report*, (63), 59-83.

Plant Protection and Quarantine (PPQ) (2016). *Guidelines for the USDA-APHIS-PPQ weed risk assessment process*. United States Department of Agriculture (USDA), Animal and Plant Health Inspection Service (APHIS), Plant Protection Quarantine (PPQ)

Plant Protection and Quarantine (PPQ) (2015). *Animal and plant health inspection service plant protection and quarantine: Strategic plan 2015-2019*. United States Department of Agriculture (USDA), Animal and Plant Health Inspection Service (APHIS), Plant Protection Quarantine (PPQ)

Prentis, P.J., Woolfit, M., Thomas-Hall, S.R. , Ortiz-Barrientos, D., Pavasovic, A., Lowe, A.J., & Schenk, P.M. (2010). Massively parallel sequencing and analysis of expressed sequence tags in a successful invasive plant. *Annals of Botany 106*(6), 1009–1017. doi:10.1093/aob/mcq201.

Pyšek, P., & Richardson, D.M. (2007). Traits associated with invasiveness: where do we stand? (pp. 97-125). Berlin, Heidelberg : Springer Berlin Heidelberg. doi:10.1007/978-3-540-36920-2_7

Quentin., T., P., A., A., I., . . . S. (2017). Seven recommendations to make your invasive alien species data more useful. *Frontiers in Applied Mathematics and Statistics, 3*. doi:10.3389/fams.2017.00013

Richardson, D.M., & Pyšek, P. (2008). Fifty years of invasion ecology - the legacy of Charles Elton. *Diversity and Distributions*, *14*(2), 161-168. doi:10.1111/j.1472-4642.2007.00464.x

Richardson, D.M., Pyšek, P., Rejmánek, M., Barbour, M.G., Panetta, F.D., & West, C.J. (2000). Naturalization and invasion of alien plants: concepts and definitions. *Diversity And Distributions, 6*(2), 93-107. doi:10.1046/j.1472-4642.2000.00083.x

Rothlisberger, J., Finnoff, D., Cooke, R., & Lodge, D. (2012). Ship-borne nonindigenous species diminish great lakes ecosystem services. *Ecosystems, 15*(3), 1-15. doi:10.1007/s10021-012-9522-6

Sala O.E., Chapin F.S. 3rd, Armesto J.J., Berlow E., Bloomfield J., Dirzo R., ... Wall D.H. (2000). Global biodiversity scenarios for the year 2100. *Science (New York, N.Y.), 287*(5459), 1770-4.

Sattler, M.C., Carvalho, C.R., and Clarindo, W.R. 2016. The polyploidy and its key role in plant breeding. *Planta, 243*(2). 281-296. doi: 10.1007/s00425-015-2450-x

Sax, D., & Brown, J. (2000). The paradox of invasion. *Global Ecology and Biogeography, 9*(5), 363-371. doi:10.1046/j.1365-2699.2000.00217.x

Schroeder, M.A., Connelly, J.W., Wambolt, C.L., Braun, C.E., Hagen, C.A., & Frisina, M.R. (2006). Society for range management issue paper: Ecology and management of sage-grouse and sage-grouse habitat—a reply. *Rangelands, 28*(3), 3-7. doi:10.2111/1551-501X(2006)28[3:SFRMIP]2.0.CO;2

Shah, M.A., & Shaanker, R.U. (2014). Invasive species: reality or myth? *Biodiversity and Conservation, 23*(6): 1425–26. doi:10.1007/s10531-014-0673-y.

Sigüenza, C., Corkidi, L., & Allen, E. (2006). Feedbacks of soil inoculum of mycorrhizal fungi altered by n deposition on the growth of a native shrub and an invasive annual grass. *Plant and Soil : An International Journal on Plant-Soil Relationships, 286*(1-2), 153-165. doi:10.1007/s11104-006-9034-2

Soltis, P.S., Soltis, D.E. (2000). The role of genetic and genomic attributes in the success of polyploids. *Proceedings of the National Academy of Sciences of the United States of America, 97*(13), 7051-57.

Sun, Y., Ding, J., & Frye, M. (2010). Effects of resource availability on tolerance of herbivory in the invasive Alternanthera philoxeroides and the native Alternanthera sessilis. *Weed Research, 50(6), 527-536. doi:10.1111/j.1365-3180.2010.00822.x*

Swain, D., Langenbrunner, B., Neelin, J., & Hall, A. (2018). Increasing precipitation volatility in twenty-first-century california. *Nature Climate Change*. doi:10.1038/s41558-018-0140-

Taylor M.H., Rollins K., Kobayashi M., & Tausch R.J. (2013). The economics of fuel management: wildfire, invasive plants, and the dynamics of sagebrush rangelands in the western United States. *Journal Of Environmental Management, 126*, 157-73. doi:10.1016/j.jenvman.2013.03.044

Taylor, S., Kumar, L., & Reid, N. (2012). Impacts of climate change and land-use on the potential distribution of an invasive weed: a case study of Lantana camara in Australia. *Weed Research, 52*(5), 391-401. doi:10.1111/j.1365-3180.2012.00930.x

te Beest M., Le Roux, J.J., Richardson, D.M., Brysting, A.K., Suda, J., Kubesová, M., & Pysek, P. (2012). The more the better? the role of polyploidy in facilitating plant invasions. *Annals of Botany, 109*(1), 19-45. doi:10.1093/aob/mcr277

Thébault, A., Gillet, F., Müller-Schärer, H., & Buttler, A. (2011). Polyploidy and invasion success: trait trade-offs in native and introduced cytotypes of two Asteraceae species. *Plant Ecology : An International Journal, 212*(2), 315-325. doi:10.1007/s11258-010-9824-8

Theoharides, K.A. & Dukes, J.S. (2007). Plant invasions across space and time: Factors affecting nonindigenious species success during the four stages of invasion. *The New Phytologist, 176*(2), 256-73.

Treier, U.A., Broennimann, O., Normand, S., Guisan, A., Schaffner, U., Steinger, T., Müller-Schärer, H. (2009). Shift in cytotypes frequency and niche space in the invasive plant Centaurea maculosa. *Ecology*, *90*(5), 1366-77.

Truong ,Y., Lin, X., & Beecher, C. (2004). Learning a complex metabolomic dataset using random forests and support vector machines, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA

Van Kleunen, M., Weber, E., & Fischer, M. (2010). A meta-analysis of trait differences between invasive and non-invasive plant species. *Ecology Letters, 13* (2): 235–45. doi:10.1111/j.1461-0248.2009.01418.x.

Vorsino, A.E., Fortini, L.B., Amidon, A.F., Miller, S.E., Jacobi, J.D., Price, J.P., Gon, S.O., Koob, G.A., & Adam, P. (2014). Modeling hawaiian ecosystem degradation due to invasive plants under current and future climates. *Plos One, 9*(5), 95427. doi:10.1371/journal.pone.0095427

Walton, W., MacKinnon, C., Rodriguez, L., Proctor, C., & Ruiz, G. (2002). Effect of an invasive crab upon a marine fishery: Green crab, carcinus maenas,

predation upon a venerid clam, katelysia scalarina, in tasmania (Australia). *Journal of Experimental Marine Biology and Ecology*, *272*(2), 171-189. doi:10.1016/S0022-0981(02)00127-2

Wambolt, C. L., & Policy Analysis Center for Western Public Lands. (2002). Conservation of greater sage-grouse on public lands in the Western U.S. : Implications of recovery and management policies (PACWPL Policy Paper, SG-02-02; PACWPL Policy Paper, SG-02-02). Caldwell, ID: Policy Analysis Center for Westerblic Lands. http://pacwpl.nmsu.edu/documents/Sage-grouse_policy.pdf

Whitney, K.D., Broman, K.W., Kane, N.C., Hovick, S.M., Randell, R.A., & Rieseberg, L.H. (2015). QTL mapping identifies candidate alleles involved in adaptive introgression and range expansion in a wild sunflower. *Molecular Ecology*, *24*(9), 2194–2211. http://doi.org/10.1111/mec.13044

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D. (2012). Darwin core: An evolving community-developed biodiversity data standard. *Plos One, 7*(1), 29715. doi:10.1371/journal.pone.0029715

Wilson, J.R., Richardson, D.M., Rouget, M., Proche, Amis, M.A., Henderson, L., & Thuiller, W. (2007). Residence time and potential range: Crucial considerations in modelling plant invasions. *South African Journal of Botany*, *73*(2), 322. doi:10.1016/j.sajb.2007.02.145