

Modeling invasion risk for coastal marine species utilizing environmental and transport vector data

R. Eliot Crafton

Received: 8 May 2014 / Revised: 19 August 2014 / Accepted: 30 August 2014 / Published online: 12 September 2014
© The Author(s) 2014. This article is published with open access at Springerlink.com

Abstract Invasive species can cause ecological and economic damage and can be transported by several vectors, many of which are connected to socioeconomic activities. This research presents a model that combines introduction likelihood and environmental suitability to characterize global patterns of invasion risk in coastal marine areas by identifying where a species is both likely to arrive and able to survive. The model projects environmental suitability using MaxEnt and considers commercial port locations, as a proxy for commercial shipping, to map patterns of relative invasion risk on a near global scale. A case study of five coastal marine crab species is presented. These models identify several regions that are at risk of new invasion where modeled environmental suitability and introduction likelihood overlap. The distribution of large commercial ports is near global but not evenly distributed; northern hemisphere temperate locations have a higher density of ports and tend to

have more opportunities for invasion according to these models. This approach can be adapted to other marine and non-marine species and to current and future environmental and socioeconomic conditions, but it works best when occurrence data are representative of the complete range of conditions under which a species can survive.

Keywords Invasive species · Invasion risk · Species distribution modeling · MaxEnt · Dispersal vectors

Introduction

Invasive species are a major threat to biodiversity and can result in significant economic costs (e.g., Pimentel et al., 2005). These species can impact agriculture, aquaculture, industry, and recreation and have the potential to significantly alter terrestrial, freshwater, and marine ecosystems (Ruiz et al., 1997; Kolar & Lodge, 2000; Bax et al., 2003; Tylianakis et al., 2008). Socioeconomic pathways, notably merchandise trade, have been documented as primary routes of species transport (Westphal et al., 2008), and increased globalization and international trade are expanding opportunities for the spread of novel species (Hulme, 2009). In response, there have been many proposed policy and management strategies to curb species introductions (Bax et al., 2003) such as the International Maritime Organization's ballast water exchange

Guest editors: Sidinei M. Thomaz, Katya E. Kovalenko, John E. Havel & Lee B. Kats / Aquatic Invasive Species

Electronic supplementary material The online version of this article (doi:10.1007/s10750-014-2027-x) contains supplementary material, which is available to authorized users.

R. E. Crafton (✉)
Department of Evolution and Ecology, University of California Davis, 2320 Storer Hall, Davis, CA 95616, USA
e-mail: recrafton@ucdavis.edu

convention (see Gollasch et al., 2007 for an overview of this convention). While understanding the variety of species transport vectors is a priority for regulating introduction (Hulme et al., 2008; Williams et al., 2013), the success of efforts to manage existing vectors and prevent new invasions has been mixed (Simberloff et al., 2005).

Impacts of introduced species span a wide spectrum (Molnar et al., 2008), and the impacts caused by non-native species have been difficult to predict (Ricciardi et al., 2013). While non-native marine species include a wide range of taxa, one important taxon in coastal and estuarine systems is the crustacean class Decapoda, including crabs, shrimps, lobsters, and others. Within this class, *Carcinus maenas* (Linnaeus, 1758) and *Eriocheir sinensis* (H. Milne-Edwards, 1853) are two of the nine aquatic species listed in the “100 of the World’s Worst Invasive Alien Species” (Lowe et al., 2004). The transport of many marine species is often, though not exclusively, attributed to commercial shipping in ballast water and hull fouling (Carlton & Geller, 1993; Rodríguez & Suárez, 2001; Williams et al., 2013). Furthermore, it has been shown that select decapods can remain viable after transport by these vectors (Hamer et al., 1998).

The process by which a non-native species invades a new location can be a complex and multistage event (Carlton, 1996). This process includes the species in the source environment through to establishment and spread. Within the process, there are many stages that can be pinpointed as management opportunities to prevent an invasion from occurring (Kolar & Lodge, 2001). Research and management efforts have targeted species when they are exported, in transit, entering a new environment, after introduction, and once established. In many cases, it is difficult to eradicate a species once established, though there are examples of successful efforts (e.g., see Williams & Grosholz, 2008), and prevention and early detection can be the most viable management strategies (Lovell & Drake, 2009).

Predicting the factors that determine if a non-native species will successfully become established and invasive has been difficult. These factors include the characteristics of the non-native species, biotic interactions, suitability of the environment, and propagule pressure, among others (Stachowicz et al., 1999; Mack et al., 2000; Nyberg & Wallentinus, 2005; Kimbro et al., 2013). Given the complexity of the invasion

process, any foresight into where a species will likely invade can be used for more judicious and targeted management (Leung et al., 2004). For species that arrive in a new location, early detection can increase the likelihood for successful eradication prior to establishment and spread, reducing the costs associated with invasive species (Williams & Grosholz, 2008).

Species distribution modeling (SDM) has received considerable attention over the past several decades (Zimmermann et al., 2010). As SDM has evolved, statistical algorithms and software packages have been developed to improve predictive capacity, and increasingly complex simulations are being undertaken to take into account a greater number of predictive variables (Elith et al., 2006). However, the importance of considering ecological theory and underlying assumptions remains paramount, and a balance of complexity and accuracy is required (Austin, 2002; Wiens et al., 2009). While prevalence of these models is increasing, utilization has not been consistent between realms, with modeling in marine systems often lagging behind terrestrial counterparts (Robinson et al., 2011).

The goal of this project is to develop a model to assess invasion risk by combining environmental suitability and the availability of transport vectors to determine where species can both survive and are likely to arrive. While this model is not intended to replace multispecies and vector management, it can act as an additional tool to help understand regional and global patterns regarding where individual species may be able to invade. This project relies on freely available statistical and environmental modeling software and considers the efficacy of using open access data to generate the predictions.

Methods

This project utilized open access environmental and occurrence data, which has been supplemented with occurrence records from published literature, and freely available modeling and statistical software, MaxEnt and R, respectively. The methodology can be easily transferred to other marine and non-marine species for which occurrence and environmental data are available, and plausible future conditions, such as those due to changes in climate and socioeconomic

infrastructure, could be incorporated to project invasion risk in future scenarios.

Invasion risks for five crab species, all with documented non-native populations, were modeled. The five species modeled within are *Carcinus maenas* (Linnaeus, 1758), *Charybdis hellerii* (A. Milne-Edwards, 1867), *Charybdis japonica* (A. Milne-Edwards, 1861), *Hemigrapsus sanguineus* (De Haan, 1835), and *Rhithropanopeus harrisi* (Gould, 1841). These species represent a broad spectrum of invasion history, habitat preference, biology, and research focus (see Brockerhoff & McLay, 2011 for more species details).

Native ranges for these species include the northwest Pacific Ocean (*C. japonica* and *H. sanguineus*), Indo-Pacific (*C. hellerii*), and western (*R. harrisi*) and northeastern (*C. maenas*) Atlantic Ocean. Non-native populations are established around the globe, including in the Atlantic Ocean (*C. maenas* in the northwest, southwest, and southeast; *C. hellerii* in the Caribbean, northwest, and southwest; *H. sanguineus* in the northeast and northwest; and *R. harrisi* in the southwest and northeast) and Pacific Ocean (*C. maenas* in the northwest, southwest, and northeast; *C. japonica* in the southwest; and *R. harrisi* in the northeast) (Brockerhoff & McLay, 2011). Additionally, *C. hellerii* and *H. sanguineus* have been documented in the Mediterranean Sea. All five species have shipping related vectors (wet and dry ballast, hull fouling, and sea chests) as documented or potential transport mechanisms (McDermott, 1998; Carlton and Cohen, 2003; Tavares & Amouroux, 2003; Gust & Inglis, 2006; Roche & Torchin, 2007). Additional vectors, including seafood and aquaculture trade (e.g., *R. harrisi* transport from the eastern USA to the western USA) and natural dispersal (e.g., *C. maenas* on the USA West Coast and between Australia and Tasmania), are cited as responsible for secondary spread (Rodríguez & Suárez, 2001; Darling et al., 2008; Tepolt et al., 2009). While secondary spread is important for determining species distributions, this research is focused on predicting risk from initial introduction based on commercial shipping as a vector.

Species Environmental Suitability Modeling

Environmental suitability modeling was conducted using the maximum entropy method employed in the freely available software MaxEnt 3.3.3 k (available at

<http://www.cs.princeton.edu/~schapire/maxent/>; Phillips et al., 2006; Elith et al., 2011). MaxEnt relies on presence only species occurrence data to predict environmental suitability based on constraints derived from the relationship between training occurrence data and environmental variables (Phillips et al., 2006; Phillips & Dudík, 2008). Several types of input data were required to train the models including occurrence records for each species, environmental data, and sampling bias layers, which, although not required, were utilized to account for spatial bias in sampling (all described below). Three feature types (Linear, Quadratic, and Product) were selected out of the five available in MaxEnt to constrain the relationship between occurrence probability and environmental variables (see Elith et al., 2011). Selecting this subset excluded more complicated models that rely upon harder to conceptualize relationships between distribution and environmental variables (Syfert et al., 2013). Other important user defined options included setting Maximum Background Points to 50,000 to improve background representation, Cross-validation with ten Replicates, and increasing Maximum Iterations to 10,000 to ensure models had adequate opportunity to run to convergence (models required between 300 and 1,240 iterations to converge). Model averages are presented based on cross-validation with data partitioned into ten folds that were cycled through using nine for training and one for testing.

Results are presented in three formats. The logistic output provides a probability for environmental suitability between zero and one. The remaining two formats utilize thresholds provided in the MaxEnt result outputs, minimum and ten percentile (hereafter, 10%) based on training occurrence data, to differentiate between suitable and non-suitable environments. These are the thresholds at which all or all but 10% of the training presences are required to be included within projected suitable environments, respectively. The area under the receiver operating characteristic (ROC) curve (AUC) was used to assess the performance of the species distribution model. This metric provides a measure of performance across all possible thresholds, zero to one, and is presented for the test occurrence records. The ROC curve is a plot of sensitivity (also known as the true positive rate) on the y-axis and 1—specificity (or the false positive rate) on the x-axis. Details on this metric can be found in Phillips et al. (2006). In addition to AUC, average test

omission rates (proportion of test presence sites not modeled as suitable) are reported for both thresholds.

Occurrence data

This model was built using freely available, open access data for species occurrences, which were supplemented by data from a literature review. The open access data were acquired using the Global Biodiversity Information Facility (GBIF) data portal (GBIF, 2014). Data were retrieved by searching the scientific name of the species of interest and downloading the results using the default settings in csv format in March 2014. Citations for data sources are available by species in the Electronic Supplementary Material (ESM). Occurrence data outside of 70°N–70°S were not considered resulting in the exclusion of three records. For records that referred to non-marine locations, the nearest marine raster cell within 2 degrees was identified and the coordinates of this cell were used.

A literature search was utilized to supplement the open access data to ensure that the known distribution for each species was represented. The Thompson Reuters' Web of Science database (<http://apps.webofknowledge.com>) was used for the literature search in April 2014. The species name in quotation marks was entered for the search topic using all available years (1926–2014), except for *C. maenas* (see below). The resulting articles were sorted by title, with potentially useful articles marked for download. Occurrence record locations within articles were obtained as site descriptions, latitude and longitude coordinates, or maps. Google Earth version 6.2.2.6613 (Google Corporation, 2012) was utilized to determine latitude and longitude coordinates for site descriptions and map-based occurrence records. For map-based records, maps were overlaid onto the Google Earth projection and coordinates for the occurrence records were determined. For *C. maenas*, map-based non-native occurrence records were obtained from Compton et al. (2010, p. 247) with a supplementary search of the literature using Web of Science for 2010 to April 2014.

In order to address issues of spatial autocorrelation, two methods were employed. All duplicates were removed so that each species could have a maximum of one occurrence record per raster cell by converting all occurrence record coordinates to cell centered

coordinates using the R raster package and removing duplicates (Hijmans, 2013). The second method for addressing spatial autocorrelation was to create a bias grid for each species weighted using a Gaussian kernel,

$$\text{weight} = \exp(-d^2/2s^2),$$

where d is the least cost distance (see below) in kilometers between the center of a raster cell with an occurrence point and the center of a marine background raster cell, and s is the standard deviation (see Clements et al., 2012). The standard deviation was set at 20 km as this was between the distance from one cell center to the center of a cell two cells north, south, east, or west and the distance between a cell center and a cell two cells northwest, northeast, southwest, or southeast. Unique occurrence records for each of the five species were used to create the bias grids. To calculate the bias grids, least cost distance was calculated using the *gdistance* package available in R (van Etten, 2012). Least cost distance was used to ensure that distances followed a marine only path (as the fish swims). Least cost distance is particularly important for regions with bays and islands where the shortest absolute distance (including overland) differs significantly from the shortest distance a species could actually travel. For each species, distances for each marine background raster cell ($N = 428,768$) to each of the occurrence points of that species were calculated and then converted to Gaussian kernel based weights between zero and one. The individual weights for each cell were then summed to create a final weight for each marine background raster cell. Because MaxEnt requires bias grids to have a value greater than zero, 0.01 was added to all raster cells of the bias grid.

Environmental layers

Environmental layers were accessed from Bio-ORACLE as outlined in Tyberghein et al. (2012). The data package 70°N–70°S Real Values was utilized, which included 23 raster layers with 5 arcmin resolution. Correlations greater than or equal to 0.91 between layers cropped to coastal regions were considered to result in collinearity of the layers and only one of the collinear layers was considered based on a priori descriptiveness and relevance to the species. Of the remaining available layers, ten layers were selected based on a priori relevance for the considered species:

Mean Calcite, Chlorophyll Minimum, Chlorophyll Maximum, Mean Nitrate, Mean pH, Mean Phosphate, Mean Salinity, Mean Silicate, Sea Surface Temperature Maximum, and Sea Surface Temperature Minimum. More detail regarding these environmental layers can be found in Tyberghein et al. (2012) and at the Bio-ORACLE website (<http://www.oracle.ugent.be/>). Layers were masked so that environmental data were only available for the 80 km immediately offshore leaving the mid-ocean regions without associated data (NA).

Port data/introduction likelihood

Annual data on the world's largest ports were acquired for the time period 2008–2010 from the World Port Rankings available on the American Association of Port Authorities port statistics page (AAPA, 2013). These rankings listed the 125 largest ports by both total cargo volume (tons) and container traffic (Twenty Foot Equivalent Units—TEUs) for 2008 and 2009 and 150 ports by cargo volume and 128 ports by container traffic in 2010. Ports were compiled and a list of 208 unique ports was produced after inland ports were removed. Geolocations of these ports were acquired from the 2012 World Port Index (National Geospatial-Intelligence Agency, 2012). Least cost distances (as the fish swims) from non-native occurrence data points to the nearest port were calculated for each species using `gdistance` in R. The likelihood of observing a species at a specific distance from the port was calculated as the inverse cumulative probability. These values were calculated by starting with a probability of 1 at the port and cumulatively reducing the probability at a given distance by O/N , where O is the number of occurrences at a given distance from the closest port and N is the number of non-native occurrences for a given species. The inverse cumulative probability was aggregated across all species and used as a proxy for introduction likelihood.

Invasion risk

Invasion risk was calculated by overlaying the introduction likelihood on a minimum threshold binary environmental suitability map using the raster package in R. The minimum threshold was utilized in order to

minimize false negatives and to predict as close to a species' fundamental niche as possible.

Results

Environmental suitability

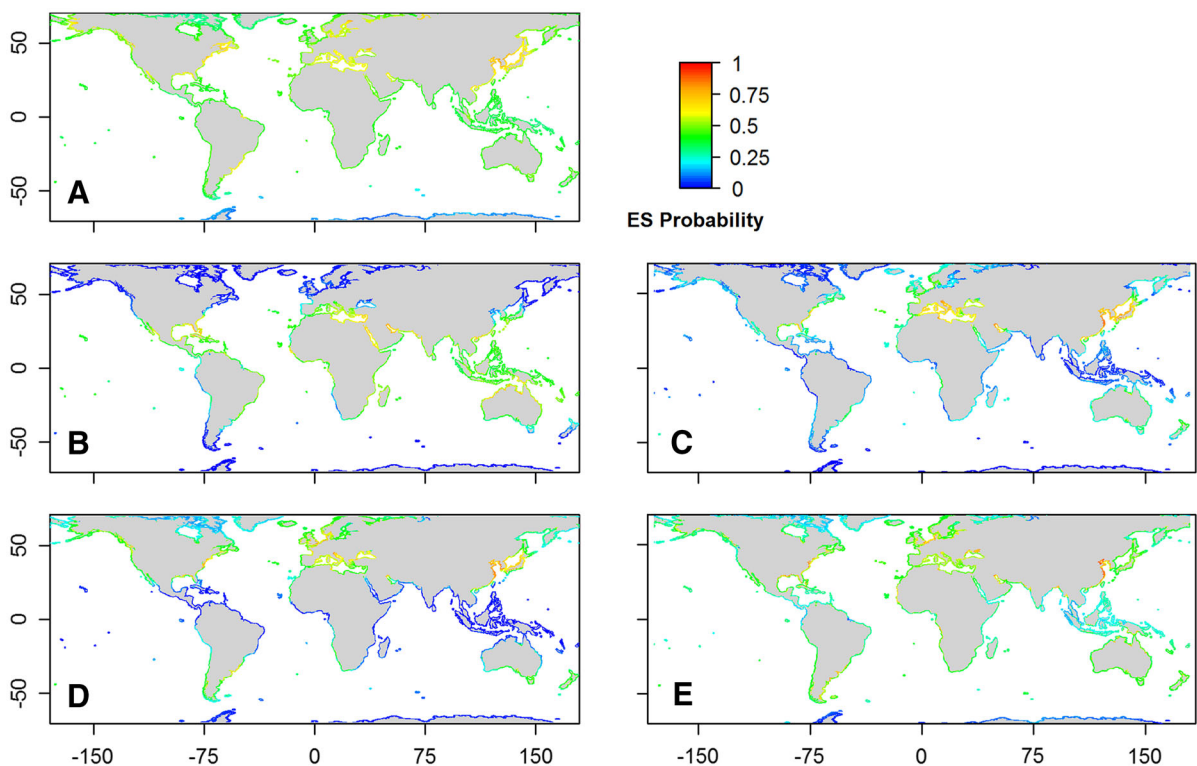
The number of available occurrence records varied widely between species ($N = 67$ –1,714) as did the modeled environmental suitability and the minimum and 10% training presence logistic thresholds (Table 1; Fig. 1). Test area under the receiver operating characteristic curve (AUC) (the metric used to assess model fit across all thresholds) averaged across runs at 0.676 or higher for each species with three species (*C. japonica*, *H. sanguineus*, and *R. harrisi*) returning a test AUC above 0.95. Minimum threshold ranged from roughly 0.26 to 0.45, and the 10% threshold ranged from 0.43 to 0.61. Test data omission rates were close to zero for the minimum threshold and between 10 and 15% for the 10% threshold. Within the model, sea surface temperature maximum and minimum (average importance of 30.1 and 44.8% across all species, respectively; Pearson correlation of 0.9 between cropped sea surface temperature layers) were the most important environmental layers for dictating suitability in the models based on permutation importance. The remaining variables were relatively unimportant for the models (importance between 0.69 and 6.0%).

Projected suitable environments for these species ranged from approximately 16 to 77% of the area modeled using the minimum thresholds and roughly 6–52% when the 10% thresholds were employed (Fig. 2). *C. japonica* had the narrowest range utilizing the minimum threshold (16.0%), and switching between thresholds reduced the suitable area to 10.4%. *C. maenas* and *R. harrisi* had projected suitable areas with the minimum threshold of 76.9 and 64.6%, respectively, but their 10% thresholds reduced this area to 51.7% and only 7.8%. *H. sanguineus* and *C. hellerii* had 19.3 and 43.15% suitable area with the minimum threshold. For the 10% threshold, suitable area was reduced to 24.9% for *C. hellerii*, and *H. sanguineus* had the smallest projected suitable area of only 6.0%.

Table 1 MaxEnt results for average test area under the receiver operating characteristic curve (AUC) with standard deviation (SD) in parentheses

Species	AUC (SD)	GBIF/Lit	MinT (Omm)	10%T (Omm)
<i>Carcinus maenas</i>	0.6755 (0.012)	1,533/181	0.3543 (0.00)	0.4254 (0.1057)
<i>Charybdis hellerii</i>	0.8847 (0.0379)	59/42	0.3342 (0.04)	0.4433 (0.1491)
<i>Charybdis japonica</i>	0.9594 (0.0139)	46/21	0.3966 (0.0143)	0.5387 (0.1333)
<i>Hemigrapsus sanguineus</i>	0.9677 (0.0062)	113/103	0.4503 (0.0091)	0.6108 (0.1113)
<i>Rhithropanopeus harrisi</i>	0.9568 (0.0107)	182/110	0.2638 (0.0035)	0.5748 (0.1203)

Each average is based on ten cross-validated runs. Number of occurrences by source (Global Biodiversity Information Facility, GBIF, or Literature, LIT), as well as, minimum (MinT) and 10 percentile (10%T) training presence logistic thresholds with their respective test omission rates in parentheses (Omm) are provided

**Fig. 1** Logistic output for environmental suitability (ES) based on MaxEnt models for each species. Scale is from low suitability (blue) to high suitability (red). **A** *Carcinus maenas*; **B** *Charybdis*

hellerii; **C** *Charybdis japonica*; **D** *Hemigrapsus sanguineus*; **E** *Rhithropanopeus harrisi*

Introduction likelihood

Introduction likelihood was calculated based on non-native occurrence locations by plotting the least cost distance from the nearest major world port to each occurrence record. *C. maenas* was found farthest from a port (~2,000 km), and *C. japonica* was the most

limited (131 km) (Fig. 3). For global introduction likelihood, aggregated data for all five species showed a nearly linear decline to approximately 500 km and then a long tail that ended at roughly 2,000 km. The regions with the greatest density of ports and highest continuous introduction likelihood are Europe, East Asia, India, the Gulf of Mexico, and southeastern

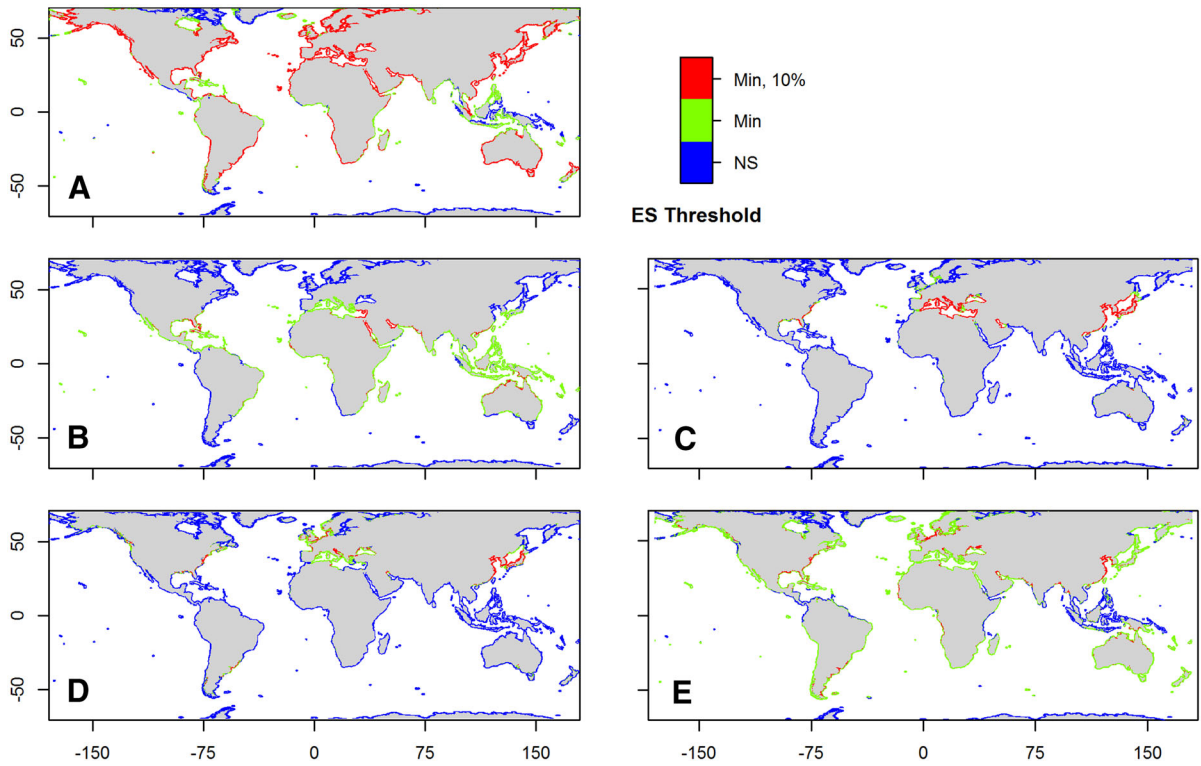


Fig. 2 Binary environmental suitability (ES) showing suitable habitat based on minimum only (green) and minimum and 10% training presence logistic thresholds by species.

A *Carcinus maenas*; **B** *Charybdis hellerii*; **C** *Charybdis japonica*; **D** *Hemigrapsus sanguineus*; **E** *Rhithropanopeus harrisi*

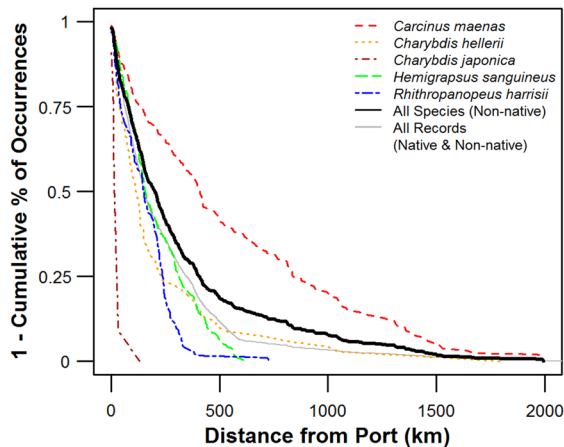


Fig. 3 Introduction likelihood presented as the inverse cumulative probability of non-native occurrence records by distance to nearest port for each species. The solid black line is based on non-native occurrence data for all five species; the solid gray line is for native and non-native occurrence records for all species. The line for ‘All Records (Native and Non-native)’ is restricted to records with distances to ports <2,000 km

United States (Fig. 4). Other regions have lower densities of ports and had non-continuous regions of high introduction likelihood. The only regions without any large ports are the higher northern and southern latitudes, Indo-Pacific, southern South America, eastern Africa, and southwestern Australia. As such, these regions have very low to zero modeled introduction likelihood.

Invasion risk

The modeled results for invasion risk varied widely between species. *C. maenas* had a broad invasion risk, which included several locations that already have documented non-native populations. These regions included the East and West Coasts of the United States and Canada, Australia, South Africa, and Japan (Fig. 5A). Other regions with high invasion risk included China, the Korean Peninsula, South America, northern New Zealand, and small regions of northern

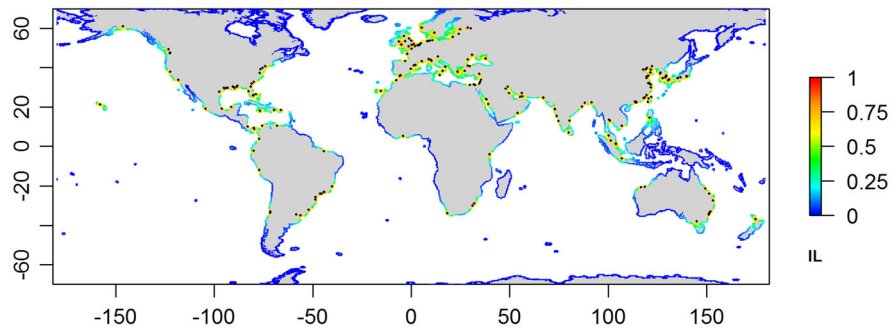


Fig. 4 Introduction likelihood (IL) applied to world's largest 208 ports using the inverse cumulative probability of non-native occurrence records by distance to nearest port aggregated for all

five species. Scale is from low introduction likelihood (blue) to high introduction likelihood (red). Black dots represent port locations

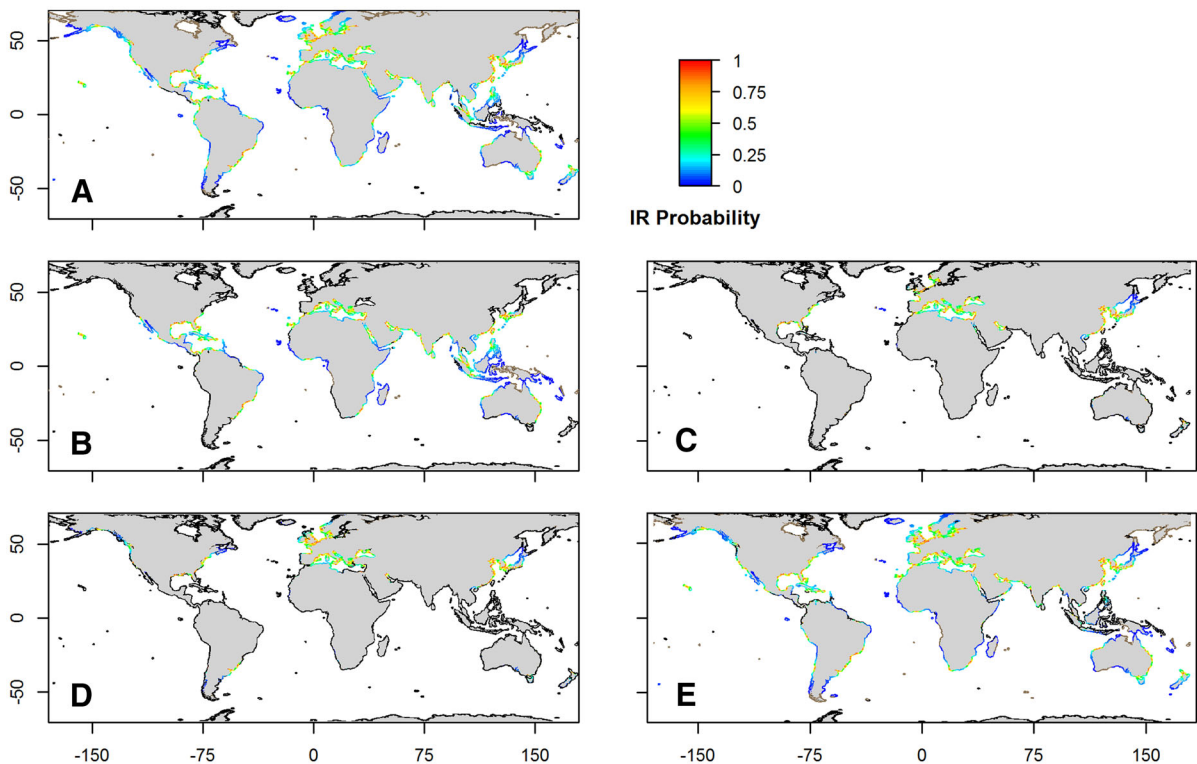


Fig. 5 Invasion risk (IR) for species calculated by overlaying introduction likelihood for the world's largest 208 ports on a binary suitability map based on the minimum training presence logistic threshold. Scale is from low invasion risk (blue) to high invasion risk (red). Black represents areas not predicted suitable

based on minimum threshold; brown represents areas suitable but with zero introduction likelihood. **A** *Carcinus maenas*; **B** *Charybdis hellerii*; **C** *Charybdis japonica*; **D** *Hemigrapsus sanguineus*; **E** *Rhithropanopeus harrisi*

and eastern Africa. While select higher northern latitudes were modeled as suitable, there are fewer ports in this region for introduction to occur. Much of the tropical and far southern latitudes were modeled as not suitable, regardless of introduction likelihood.

Charybdis hellerii had high invasion risk around the equatorial band reaching as far north as the Mediterranean Sea and as far south as Australia and eastern South America. (Fig. 5B). Hotspots of invasion risk included the regions currently reported to

have established invasions for this species, namely the southeastern United States, eastern South America, and the eastern Mediterranean Sea. As this species has a broad native distribution across the Indian and Pacific Oceans, there were fewer novel regions that had invasion risk. Novel regions for this species with relatively high risk of invasion were the western Mediterranean and western coasts of Central America.

Charybdis japonica had the narrowest predicted distribution with the Caribbean, Mediterranean, and Black Seas, Persian Gulf, eastern United States, and parts of Atlantic Europe all registering invasion risk in non-native regions (Fig. 5C). The one region that has an established invasive population, New Zealand, had a small region of high invasion risk modeled near the Ports of Auckland.

Hemigrapsus sanguineus had a predominately temperate distribution with invasion risk occurring in the Mediterranean and Black Seas, northern Europe, eastern South America, and both coasts of North America (Fig. 5D). New Zealand also had regions on both the North and South Islands of higher and lower invasion risk, respectively. Reported established non-native populations are along the East Coast of the United States and Atlantic Europe, which were included in the regions modeled as having high invasion risk.

Rhithropanopeus harrisi had a broad invasion risk similar to *C. maenas*, with many regions having modeled suitability and associated invasion risk (Fig. 5E). The northern Atlantic was nearly devoid of environmental suitability and introduction likelihood, though parts of coastal Northern Europe had low invasion risk. Established populations have been reported extensively from around Europe, including the Atlantic, northern, and Mediterranean coasts, inland seas, including the Baltic, Black, Aral, and Caspian Seas, and the west coast of the United States. These regions were modeled to have moderate to high invasion risk. Additionally, regions in South America, Africa, Asia, Australia, and New Zealand had regions that were modeled to have relatively high invasion risk that are yet to have reported establishments.

Discussion

Identifying regions that are susceptible to invasion is an important goal to understand where non-native

species may cause problems. The models presented within this study utilized a combination of modeled environmental suitability, based on maximum entropy, and likelihood of introduction, using ports as a proxy for commercial shipping, to inform our understanding of potential global patterns of invasion risk. For the five species presented here, this modeling approach has implications for invasion risk assessment and for the components of the models, environmental suitability and introduction likelihood. The models show considerable overlap between environmental suitability and proximity to ports, and there are several regions that are currently not invaded despite having high invasion risk.

Two of the species were projected to have primarily temperate suitable environmental distributions (*C. japonica* and *H. sanguineus*), *C. hellerii* was projected to have a tropical suitable distribution, and *C. maenas* and *R. harrisi* had very wide projected suitability with the exception of some and much of the Indo-Pacific region, respectively. These distributions are important as estimates of where the species will be able to survive upon arrival. Environmental suitability calculated here is dependent on the selection of environmental layers and the availability of occurrence data for the species of interest. Selecting environmental layers is limited to data availability at the scale of modeling, and having more detailed global coverage could likely improve the model. Environmental layers were selected based on their potential relevance either directly to the species being modeled or in the case of nutrient inputs to their prey species. While several of the environmental variables contributed a small amount of information to the ultimate models, maximum and minimum sea surface temperatures were the most important variables for dictating range limits for these species at this scale. While either summer or winter temperatures may limit species distribution, these two layers are highly correlated in coastal regions (Pearson correlation of ~ 0.9).

While occurrence data were abundant for one species, *C. maenas*, and representative for the native and invasive ranges for three additional species, *C. hellerii*, *H. sanguineus*, and *R. harrisi*, very few occurrence records were available for *C. japonica*, especially in the open access data. Test AUC values for four species suggest a high fit for these models based on the constraints provided by the occurrence data (*C. maenas* models had a moderate AUC value).

Values for test omission rates reflected the thresholds used for determining suitability, though omission rates at the 10% threshold were consistently higher than 10%. This discrepancy is likely in part because of the paucity of data for some of the species resulting in higher omission rates due to the significance of each test data point. Based on the higher test omission rates, models using the 10% threshold have the potential to wrongly characterize suitable regions as not suitable. To minimize false negatives, the more generous minimum threshold was utilized to calculate invasion risk in this study.

All five of these species were represented in the open access data utilized, at least in their native range. Collecting additional occurrence records from published literature was possible and necessary for underrepresented species and regions, but this collection increased both the time and resources needed to collect the data. While this added burden may be unnecessary for species that are well documented in the open access datasets such as *C. maenas*, *C. japonica* is a good example of a species that is underrepresented in the open access database, which lacked records for its non-native occurrences in New Zealand. When run without additional records, New Zealand was not modeled as suitable for *C. japonica* and not predicted as having invasion risk in that location (ESM Fig. 1C). Including records from the supplemental literature search resulted in a model that projected suitable environments near Auckland, New Zealand. Overall, there were slight changes to suitable environment projections using the minimum threshold (modeled increases between <1 and ~23% suitable area) when additional records were utilized (Fig. 2, ESM Fig. 1). In some cases, like *C. maenas* in Greenland, additional records resulted in the loss of environment that was classified as suitable when modeled with fewer records. In addition to the number of occurrence records, the inclusion of a bias grid was important to the projected environmental suitability. It was more important for species with regions of high occurrence record density, such as *C. maenas* and *R. harrisii* (Fig. 2, ESM Fig. 2A, F). However, the impact seemed to be lower for species with fewer occurrence records and narrower distributions (*C. japonica*, *C. hellerii*, and *H. sanguineus*).

Introduction likelihood for these five species illustrates two important findings. First, all non-native occurrence records considered in this study occurred

within 2,000 km of one of the 208 major world ports. Furthermore, 50% occurred within ~200 km from a port and 90% within ~840 km (Fig. 3, ESM Fig. 3). The assumption made in this research is that the clustering of non-native occurrence records near ports is due to the likelihood that these species are introduced via commercial shipping into ports (Drake & Lodge, 2004), at least in primary transport, and secondary spread will occur via intraregional transport mechanisms (Wasson et al., 2001). Given the extra steps involved and time lag to spread from the initial point of introduction (Crooks & Soulé, 1999; Byers et al., 2002), densities are expected to be highest close to ports in near term post invasion timescales with densities at distances scaling with time since invasion. In this study, results show that *C. maenas*, which has the longest invasion history, has been able to spread the farthest from ports, and *C. japonica*, the species with the shortest invasion history, has spread the least. While there are additional possible primary vectors for these species (e.g., historic oyster trade for *R. harrisii* between the East and West Coasts of the United States; see Roche & Torchin, 2007), commercial shipping is considered a likely vector for the spread of all five species. Additionally, even if these occurrences were not the result of commercial shipping, these records suggest that introductions are occurring near ports, which often coincide with significant human populations and other socioeconomic activities.

An alternate explanation for the observed pattern of occurrence record densities is that observation effort is not evenly distributed, and greater effort is undertaken near ports resulting in higher densities of occurrence records (e.g., Wasson et al., 2001). Recent research has shown that non-port embayments can be highly invaded, despite not having direct interregional shipping pressure, through intraregional boat traffic, aquaculture, or natural spread vectors (Wasson et al., 2001; Cohen et al., 2005). While greater observation effort may be supported by the similar trend in distance to nearest major port for the combination of native and non-native occurrence records (50% within ~150 km and 90% within 530 km; Fig. 3), this pattern is likely driven by the density of ports around Europe and the East Coast of the United States (Fig. 4). These regions are also the locations of the densest occurrence records, primarily for *C. maenas* and *R. harrisii*, which account for 71.7 and 12.2% of the records, respectively (ESM Fig. 3). In both cases,

the surveys for these species, as well as *H. sanguineus*, appear to be systematic and representative of the regions as opposed to limited to port locations. While there does not appear to be oversampling of port locals within these regions, higher observation effort in these regions in general may systemically bias the data. In this case, it is likely that density of ports and observation effort are both correlated with development status and location of institutions with higher education and natural history foci.

The second important point regarding introduction likelihood is that the abundance and distribution of major commercial cargo and container ports provides a nearly global, though spatially biased, transport mechanism for these species. Within the modeled region, only the Antarctic, Arctic, and Indo-Pacific have significant regions that fall outside of 2,000 km from a port. Smaller regions that are farther from one of these major ports also exist on the coasts of western Africa, southwestern Australia, southern South America, and many of the smaller island nations around the globe. There are also regions that have many ports that are close together representing hotspots of potential introduction. Introduction hotspots can be seen in China, Europe, and the Gulf Coast of the United States. For regions that do not have a major world port, it is possible that smaller regional ports could also act as primary introduction points or avenues for secondary spread. Regional models could be constructed to show relative introduction likelihood or secondary spread of these species by incorporating regional ports. Additionally, the construction of new ports has the potential to alter introduction likelihood and ultimately impact the invasion risk for these species. Inclusion of future port locations could provide an understanding of how introduction likelihood could change over time. It would also be possible to replace ports as the proxy and rerun the analysis for an alternate vector, such as aquaculture or ornamental trade, if a suitable proxy is identified.

The primary output of interest is the relative invasion risk maps presented for each species, which allow for regions of high relative threat for new invasions to be identified. For example, *C. maenas* has several regions that are at risk for new invasions based on this model, including regions near Brazil and in New Zealand. Additionally, while this species is already reported as successfully established in Japan (Carlton & Cohen, 2003), additional nearby regions

are modeled to have high invasion risk. Similar non-invaded regions exist for the other four species. However, knowing the limitations of this model is extremely important to the relative confidence placed in the specificity of these invasion risk models. *C. japonica* is a good example where the limited availability of occurrence data could have resulted in not having an accurate representation of the species plausible range and ultimately under predicting environmental suitability. Additionally, it is evident from this research that data collection effort is not evenly or randomly distributed between species or locations. While bias can be addressed using spatial corrections as conducted here, increases in data collection and reporting for underrepresented species and regions would benefit the ability to accurately conduct this type of modeling. Non-analogous conditions (conditions not represented in existing occurrence data), may not be identified as suitable if occurrence data is not available for the entire suitable range even if the species could survive in the conditions present. These non-analogous regions can represent either regions that are suitable but not represented in the existing range/occurrence records or regions with conditions that the species can adapt to. Species could be absent because of a lack of opportunity to access the region or range restriction due to another reason such as biotic limitation. These models are restricted in their ability to predict a species' fundamental niche as they are being trained on the realized niche conditions. To the extent possible, this limitation is alleviated by utilizing both native and non-native occurrence records to estimate the potential niche of the species (Jiménez-Valverde et al., 2011). Environmental data accuracy and availability is also a potential limitation of this type of modeling. Using variables that are at a global scale and averaged across years can be used to describe global patterns, but may not provide information regarding conditions present at small scales and extreme events that could restrict species' distributions. While the decision to not incorporate latitudes above and below 70° north and south (Tyberghein et al., 2012) did not meaningfully restrict occurrence data (only three records for *C. maenas* in Norway were outside the modeling window), it has implications for invasion risk as increases in access to shipping and future climate modification to the high arctic could alter susceptibility to invasion in this region (Ware et al., 2014).

Open access data and free software have the potential to increase access to this type of modeling, so long as existing data are available and representative of the complete range of the species of interest. Literature review should be used to assess the accuracy of occurrence records and to supplement open access data when needed. Additionally, as new collections of occurrence records are undertaken, researchers and research institutions should consider posting to open access data sharing and biodiversity repositories to improve modeling accuracy and reduce the need to seek additional occurrence records from literature sources that may not be readily available for all stakeholders. Repositories also need to continue to improve quality control efforts to ensure that occurrence data are reliable. Climate change and the modification of transport vectors have the potential to alter the environmental and socioeconomic landscapes, which could modify existing and potential future ranges. By incorporating scenarios for these potential changes, alterations to invasion risk in the future could be predicted.

In summary, environmental suitability and introduction likelihood are both integral components of the invasion cycle and to understand invasion risk. Modern modeling tools are useful in predicting where species will be able to survive, but they are limited by the availability of data to train the model. Of primary importance is the consideration that the model will only be able to predict novel environments based on the constraints imposed by current distribution rather than truly predicting the fundamental niche of a species; there is a high risk that environmental conditions that are suitable but not analogous to conditions represented by existing occurrence records will be modeled as false negatives. Using open access data has many benefits in terms of time and accessibility, but the results can under predict environmental suitability for species that are not well represented in these databases (e.g., *C. japonica*). Using literature based data in conjunction with open access data can help to ameliorate this limitation, but overall, these models work best for well-studied species. Non-native occurrence records for these five species mainly occur within close proximity to major world ports (90% within 840 km), suggesting that these ports act as a good proxy for introduction likelihood. These ports are spatially biased with the majority occurring in the northern hemisphere primarily in temperate regions.

By combining these two components, it is clear that there are large regions of overlap where these species could survive and are likely to be introduced. While the limitations to this type of modeling must be considered, these models show that it is possible to identify global patterns of invasion risk. Finally, this type of model can be used for other marine and non-marine species under current and future conditions.

Acknowledgments I would like to thank D. R. Strong, E. D. Grosholz, R. J. Hijmans, and previous reviewers for their comments on drafts of this manuscript. This material is based upon work supported by the National Science Foundation Graduate-12 Fellowship Program under DGE grant #0841297 to S.L. Williams and B. Ludaescher and the National Science Foundation Responding to Rapid Environmental Change (REACH) IGERT Program under DGE grant #0801430. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. Occurrence data citations and acknowledgements can be found in the Electronic Supplementary Material.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- AAPA, 2013. American Association of Port Authorities World Port Rankings (2008, 2009, 2010). Available at <http://www.aapa-ports.org/Industry/content.cfm?ItemNumber=900#Definitions%20Pertaining%20to%20Statistics>. Accessed 12/20/2013.
- Austin, M. P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* 157: 101–118.
- Bax, N., A. Williamson, M. Aguero, E. Gonzalez & W. Geeves, 2003. Marine invasive alien species: a threat to global biodiversity. *Marine Policy* 27: 313–323.
- Brockerhoff, A. & C. McLay, 2011. Human-mediated spread of alien crabs. In Galil, B. S., P. F. Clark & J. T. Carlton (eds) *In the Wrong Place—Alien Marine Crustaceans: Distribution, Biology and Impacts*. Invading Nature-Springer Series in Invasion Ecology, vol 6. Springer, Dordrecht: 27–106.
- Byers, J. E., S. Reichard, J. M. Randall, I. M. Parker, C. S. Smith, W. M. Lonsdale, I. A. E. Atkinson, T. R. Seastedt, M. Williamson, E. Chornesky & D. Hayes, 2002. Directing research to reduce the impacts of nonindigenous species. *Conservation Biology* 16: 630–640.
- Carlton, J. T., 1996. Pattern, process, and prediction in marine invasion ecology. *Biological Conservation* 78: 97–106.
- Carlton, J. T. & A. N. Cohen, 2003. Episodic global dispersal in shallow water marine organisms: the case history of the

- European shore crabs *Carcinus maenas* and *C. aestuarii*. *Journal of Biogeography* 30: 1809–1820.
- Carlton, J. T. & J. B. Geller, 1993. Ecological roulette: the global transport of nonindigenous marine organisms. *Science* 261: 78–82.
- Clements, G. R., D. M. Rayan, S. A. Aziz, K. Kawanishi, C. Traeholt, D. Magintan, M. F. A. Yazi & R. Tingley, 2012. Predicting the distribution of the Asian tapir in Peninsular Malaysia using maximum entropy modeling. *Integrative Zoology* 7: 400–406.
- Cohen, A. N., L. H. Harris, B. L. Bingham, J. T. Carlton, J. W. Chapman, C. C. Lambert, G. Lambert, J. C. Ljubenkov, S. N. Murray, L. C. Rao, K. Reardon & E. Schwindt, 2005. Rapid assessment survey for exotic organisms in Southern California bays and harbors, and abundance in port and non-port areas. *Biological Invasions* 7: 995–1002.
- Compton, T. J., J. R. Leathwick & G. J. Inglis, 2010. Thermogeography predicts the potential global range of the invasive European green crab (*Carcinus maenas*). *Diversity and Distributions* 16: 243–255.
- Crooks, J. & M. Soulé, 1999. Lag times in population explosions of invasive species: causes and implications. In Sandlund, O., P. Schei & Å. Viken (eds), *Invasive Species and Biodiversity Management*. Kluwer Academic Publishers, Dordrecht: 103–125.
- Darling, J. A., M. J. Bagley, J. O. E. Roman, C. K. Tepolt & J. B. Geller, 2008. Genetic patterns across multiple introductions of the globally invasive crab genus *Carcinus*. *Molecular Ecology* 17: 4992–5007.
- Drake, J. M. & D. M. Lodge, 2004. Global hot spots of biological invasions: evaluating options for ballast–water management. *Proceedings of the Royal Society of London Series B: Biological Sciences* 271: 575–580.
- Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC, A. T. Overton, S. J. Peterson, K. Phillips, R. Richardson, R. E. Scachetti-Pereira, J. Schapire, S. Soberón, M. S. Williams, Wisz & N. E. Zimmermann, 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129–151.
- Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee & C. J. Yates, 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17: 43–57.
- GBIF, 2014. Global Biodiversity Information Facility Data Portal. Available at <http://www.gbif.org/>. Accessed 03/16/2014.
- Gollasch, S., M. David, M. Voigt, E. Dragsund, C. Hewitt & Y. Fukuyo, 2007. Critical review of the IMO international convention on the management of ships' ballast water and sediments. *Harmful Algae* 6: 585–600.
- Google Corporation, 2012. Google Earth Release 6.2.2.6613. Available at <https://support.google.com/earth/answer/168344?hl=en&topic=2376075&ctx=topic>. Accessed 24 April 2104.
- Gust, N. & G. Inglis, 2006. Adaptive multi-scale sampling to determine an invasive crab's habitat usage and range in New Zealand. *Biological Invasions* 8: 339–353.
- Hamer, J. P., T. A. McCollin & I. A. N. Lucas, 1998. Viability of decapod larvae in ships' ballast water. *Marine Pollution Bulletin* 36: 646–647.
- Hijmans, R. J., 2013. Raster: Geographic Data Analysis and Modeling. R Package Version 2.1-25. Available at <http://cran.r-project.org/web/packages/raster/>. Accessed 08 May 2013.
- Hulme, P. E., 2009. Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology* 46: 10–18.
- Hulme, P. E., S. Bacher, M. Kenis, S. Klotz, I. Kühn, D. Minchin, W. Nentwig, S. Olenin, V. Panov, J. Pergl, P. Pyšek, A. Roques, D. Sol, W. Solarz & M. Vilà, 2008. Grasping at the routes of biological invasions: a framework for integrating pathways into policy. *Journal of Applied Ecology* 45: 403–414.
- Jiménez-Valverde, A., A. T. Peterson, J. Soberón, J. M. Overton, P. Aragón & J. M. Lobo, 2011. Use of niche models in invasive species risk assessments. *Biological Invasions* 13: 2785–2797.
- Kimbrow, D. L., B. S. Cheng & E. D. Grosholz, 2013. Biotic resistance in marine environments. *Ecology Letters* 16: 821–833.
- Kolar, C. S. & D. M. Lodge, 2000. Freshwater nonindigenous species: interactions with other global changes. In Mooney, H. A. & R. J. Hobbs (eds), *Invasive Species in a Changing World*. Island Press, Washington, DC: 3–30.
- Kolar, C. S. & D. M. Lodge, 2001. Progress in invasion biology: predicting invaders. *Trends in Ecology & Evolution* 16: 199–204.
- Leung, B., J. M. Drake & D. M. Lodge, 2004. Predicting invasions: propagule pressure and the gravity of allee effects. *Ecology* 85: 1651–1660.
- Lovell, S. J. & L. A. Drake, 2009. Tiny stowaways: analyzing the economic benefits of a U.S. Environmental Protection Agency permit regulating ballast water discharges. *Environmental Management* 43: 546–555.
- Lowe, S., M. Browne, S. Boudjelas & M. De Poorter, 2004. 100 of the World's Worst Invasive Alien Species. ISSG-IUCN, Auckland.
- Mack, R. N., D. Simberloff, W. Mark Lonsdale, H. Evans, M. Clout & F. A. Bazzaz, 2000. Biotic invasions: causes, epidemiology, global consequences, and control. *Ecological Applications* 10: 689–710.
- McDermott, J. J., 1998. The western Pacific brachyuran (*Hemigrapsus sanguineus*: Grapsidae), in its new habitat along the Atlantic coast of the United States: geographic distribution and ecology. *Ices Journal of Marine Science* 55: 289–298.
- Molnar, J. L., R. L. Gamboa, C. Revenga & M. D. Spalding, 2008. Assessing the global threat of invasive species to marine biodiversity. *Frontiers in Ecology and the Environment* 6: 485–492.
- National Geospatial-Intelligence Agency, 2012. Pub. 150 World Port Index Twenty-Second Edition. United States Government, Springfield, VA.
- Nyberg, C. & I. Wallentinus, 2005. Can species traits be used to predict marine macroalgal introductions? *Biological Invasions* 7: 265–279.

- Phillips, S. J. & M. Dudík, 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* 31: 161–175.
- Phillips, S. J., R. P. Anderson & R. E. Schapire, 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231–259.
- Pimentel, D., R. Zuniga & D. Morrison, 2005. Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics* 52: 273–288.
- Ricciardi, A., M. F. Hoopes, M. P. Marchetti & J. L. Lockwood, 2013. Progress toward understanding the ecological impacts of nonnative species. *Ecological Monographs* 83: 263–282.
- Robinson, L. M., J. Elith, A. J. Hobday, R. G. Pearson, B. E. Kendall, H. P. Possingham & A. J. Richardson, 2011. Pushing the limits in marine species distribution modeling: lessons from the land present challenges and opportunities. *Global Ecology and Biogeography* 20: 789–802.
- Roche, D. G. & M. E. Torchin, 2007. Established population of the North American Harris mud crab, *Rhithropanopeus harrisi* (Gould, 1841) (Crustacea: Brachyura: Xanthidae) in the Panama Canal. *Aquatic Invasions* 2: 155–161.
- Rodríguez, G. & H. Suárez, 2001. Anthropogenic dispersal of decapod crustaceans in aquatic environments. *Interciencia* 26: 282–288.
- Ruiz, G. M., J. T. Carlton, E. D. Grosholz & A. H. Hines, 1997. Global invasions of marine and estuarine habitats by non-indigenous species: mechanisms, extent, and consequences. *American Zoologist* 37: 621–632.
- Simberloff, D., I. M. Parker & P. N. Windle, 2005. Introduced species policy, management, and future research needs. *Frontiers in Ecology and the Environment* 3: 12–20.
- Stachowicz, J. J., R. B. Whitlatch & R. W. Osman, 1999. Species diversity and invasion resistance in a marine ecosystem. *Science* 286: 1577–1579.
- Syfert, M. M., M. J. Smith & D. A. Coomes, 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PLoS One* 8: e55158.
- Tavares, M. & J.-M. Amouroux, 2003. First record of the non-indigenous crab, *Charybdis hellerii* (A. Milne-Edwards, 1867) from French Guyana (Decapoda, Brachyura, Portunidae). *Crustaceana* 76: 625–630.
- Tepolt, C. K., J. A. Darling, M. J. Bagley, J. B. Geller, M. J. Blum & E. D. Grosholz, 2009. European green crabs (*Carcinus maenas*) in the northeastern Pacific: genetic evidence for high population connectivity and current-mediated expansion from a single introduced source population. *Diversity and Distributions* 15: 997–1009.
- Tyberghein, L., H. Verbruggen, K. Pauly, C. Troupin, F. Mineur & O. De Clerck, 2012. Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography* 21: 272–281.
- Tylianakis, J. M., R. K. Didham, J. Bascompte & D. A. Wardle, 2008. Global change and species interactions in terrestrial ecosystems. *Ecology Letters* 11: 1351–1363.
- van Etten, J., 2012. Gdistance: Distances and Routes on Geographical Grids. R Package Version 1.1–4. Available at <http://cran.r-project.org/web/packages/gdistance/index.html>. Accessed 19 Nov 2013.
- Ware, C., J. Berge, J. H. Sundet, J. B. Kirkpatrick, A. D. M. Coutts, A. Jelmert, S. M. Olsen, O. Floerl, M. S. Wisz & I. G. Alsos, 2014. Climate change, non-indigenous species and shipping: assessing the risk of species introduction to a high-Arctic archipelago. *Diversity and Distributions* 20: 10–19.
- Wasson, K., C. J. Zabin, L. Bedinger, M. Cristina Diaz & J. S. Pearse, 2001. Biological invasions of estuaries without international shipping: the importance of intraregional transport. *Biological Conservation* 102: 143–153.
- Westphal, M., M. Browne, K. MacKinnon & I. Noble, 2008. The link between international trade and the global distribution of invasive alien species. *Biological Invasions* 10: 391–398.
- Wiens, J. A., D. Stralberg, D. Jongsomjit, C. A. Howell & M. A. Snyder, 2009. Niches, models, and climate change: assessing the assumptions and uncertainties. *Proceedings of the National Academy of Sciences* 106: 19729–19736.
- Williams, S. & E. Grosholz, 2008. The invasive species challenge in estuarine and coastal environments: marrying management and science. *Estuaries and Coasts* 31: 3–20.
- Williams, S. L., I. C. Davidson, J. R. Pasari, G. V. Ashton, J. T. Carlton, R. E. Crafton, R. E. Fontana, E. D. Grosholz, A. W. Miller, G. M. Ruiz & C. J. Zabin, 2013. Managing multiple vectors for marine invasions in an increasingly connected world. *BioScience* 63: 952–956.
- Zimmermann, N. E., T. C. Edwards Jr, C. H. Graham, P. B. Pearman & J.-C. Svenning, 2010. New trends in species distribution modelling. *Ecography* 33: 985–989.