Prediction and error in multi-stage models for spread of aquatic non-indigenous species

Jim R. Muirhead[1,2*], Mark A. Lewis[1,2,3] and Hugh J. MacIsaac[4]

[1] Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada
[2] Centre for Mathematical Biology, University of Alberta, Edmonton, Alberta, Canada
[3] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada
[4] Great Lakes Institute for Environmental Research, University of Windsor, Windsor, Ontario, Canada

Correspondence: Jim R. Muirhead, Smithsonian Environmental Research Center, Edgewater, Maryland; Email: muirheadj@si.edu, Phone: (433) 482-2450

Keywords: *Cercopagis pengoi*; gravity model; invasion sequence; spread; stochasticity

Running title: Prediction and error in multi-stage models

Abstract

Aim

Predictions of spread of nonindigenous species allow for greater efficiency in managing invasions by targeting areas for preventative measures. The invasion sequence is a useful concept in predictions of spread as it allows us to test hypotheses about the transport and establishment of propagules in novel habitats. Our aims are two-fold: 1) to develop and validate multi-stage invasion models for the introduced fishhook waterflea, *Cercopagis pengoi,* and 2) assess how variability in the transport patterns of the propagules influences the accuracy and spatial extent for predictions of spread.

Location

New York State

Methods

We developed a two-stage model for the spread of *C. pengoi*. First, we developed a stochastic gravity model for dispersal based on surveys of recreational boat traffic in New York State as a proxy for propagule pressure. We then modeled the probability of establishment based on predicted levels of propagule pressure and measures of lakes' physicochemistry. In addition, we used Monte Carlo simulations based on the gravity model to propagate variability in boater traffic through the establishment model to assess how uncertainty in dispersal influenced predictions of spread.

Results

The amount recreationalists were willing to spend, lake area and population size of the city nearest to the destination lake were significant factors affecting boater traffic. In turn, boater traffic, lake area, specific conductance and turbidity were significant predictors of establishment. The inclusion of stochastic dispersal reduced the rate of false positives (i.e. incorrect prediction of an invasion) in detecting invasions at the upper 95% prediction interval for the probability of establishment.

Main conclusions

Combinations of measures of propagule pressure, habitat suitability, and stochastic dispersal allow for the most accurate predictions of spread. Further, multistage spread models may overestimate the extent of spread if stochasticity in early stages of the models is not considered.

Introduction

Invasions of ecosystems by nonindigenous species (NIS) are occurring at increasing rates globally (Gollasch, 2006; Ricciardi, 2007; Hulme, 2009). Proactive efforts to reduce invasions are the most cost-effective management option (e.g. Leung *et al*., 2002; Finnoff *et al.*, 2007), although managers may be unwilling to “risk” a preventative approach due to the high uncertainty inherent in preventative practices relative to post-establishment control (Simberloff, 2003; Finnoff *et al*., 2007). Thus, a key challenge for invasion biologists exists with respect to forecasting dispersal and establishment of NIS to inform the most appropriate management decision (see Lodge *et al.*, 2006).

An extensive literature exists for both theoretical and empirical approaches to predict invasions (e.g. Sakai *et al.*, 2001). Much of the earlier research on forecasting invasions focused on attributes intrinsic to the NIS (e.g. Thuiller *et al.*, 2006, Statzner *et al.*, 2007) or characteristics of the recipient community (e.g. Stachowicz *et al.*, 2002; Fridley *et al.*, 2007). More recently, models have used the conceptual model of the invasion sequence, which is characterized by a series of steps the NIS must overcome to become successfully established. These steps include the initial transport of propagules from native or recently invaded sources, survival during exposure to (or exploitation of) physical and chemical characteristics of the new habitat, and integration into the existing community (Richardson *et al.,* 2000; Kolar & Lodge 2001; Heger & Trepl, 2003; Melbourne *et al.,* 2007; Theoharides & Dukes, 2007). Colautti *et al.* (2006) proposed that hypothesis testing follow this sequence, as characteristics of the NIS or the recipient community that may facilitate or hinder invasion may be applicable only at certain stages of the sequence. Here, we follow this recommendation and develop a transport model for NIS propagules, and assess their establishment upon encountering the physical and chemical characteristics of the novel habitat.

Propagule pressure may broadly be defined as the rate of introduction comprised of the number of introduction events, and the number and quality of NIS individuals transported per event. Propagule pressure has been increasingly recognized as a key determinant of invasion success across a range of taxa as it may influence both the spatial extent of the invasion and level of impact (e.g. Lockwood *et al.,* 2005; Von Holle & Simberloff, 2005). Furthermore, propagule pressure may influence invasion success at different stages of the invasion sequence. For example, in addition to the initial transport of propagules into novel habitat, sufficient levels of inbound propagules may bolster the establishment of small populations that otherwise would have disappeared owing to demographic stochasticity, and may provide an ongoing source of genetic heterogeneity to the founding population (Simberloff, 2009). Since the introduction of propagules is often the stage of the invasion sequence at which management efforts to control the spread are most efficient, it is a major focus of policy recommendations (e.g., Lodge *et al.,* 2006, Reaser *et al.,* 2008).

Vector-based or transport predictive models are often successful at predicting NIS dispersal. In particular, gravity models that were initially developed to model immigration patterns (Zipf, 1946) and trade flows (Linneman, 1966) between spatially discrete sources and destinations, have been applied to model the flow of recreational boater traffic, a likely vector for the transport of aquatic NIS. For example, gravity models have been used to model dispersal of aquatic NIS based on single-trip recreationalist movement between invaded and non-invaded lakes (Schneider *et al.,* 1998; MacIsaac *et al.,* 2004) or multi-trip movement incorporating trips from the recreationalists' homes to invaded lakes (Bossenbroek *et al.,* 2001; Leung *et al.,* 2006).

Another approach to predict invasions has sought to link factors extrinsic to the NIS - such as environmental suitability or native species community composition - to invasion success. Environmental suitability in novel regions is most often forecast by matching correlates of species occurrence or abundance with environmental data from the native range with those in the new region. For example, Herborg *et al* (2007) used an environmental niche model to forecast suitable habitat in North America for the Chinese mitten crab *Eriocheir sinensis* based upon its current distributions in Asia and Europe. Alternatively, lower biotic resistance from the recipient community due to decreased competition from native species (e.g. Dzialowski *et al.,* 2007), or differences in traits between native and introduced species augmented by shifts in environmental conditions (Moles *et al.,* 2008) or environmental heterogeneity (Melbourne *et al.,* 2007) also may facilitate invasion.

The integration of multiple stages of the invasion sequence into single models or use of a hierarchical approach in which stage-specific hypotheses are sequentially tested, allows us to assess contributions of each stage to overall invasion success or to isolate the stage of the invasion sequence where the invasion first fails. Leung & Mandrak (2007), for example, developed a joint probability model for propagule pressure and environmental conditions to predict establishment of the zebra mussel *Dreissena polymorpha*. There, the inclusion of environmental characteristics in the joint model provided better predictive power than the model containing only propagule pressure. Similarly, Rouget & Richardson (2003) found that predictions of plant invasions were strongly related to environmental conditions after first accounting for the effect of propagule pressure. In their model, propagule pressure was a stronger predictor than environmental factors considered separately.

The fishhook waterflea *Cercopagis pengoi* (Crustacea: Cladocera) is a relatively recent Great Lakes invader, having first been reported in Lake Ontario in 1998 (MacIsaac *et al.,* 1999). It spread to Lakes Erie and Michigan by 2001, as well as to several inland lakes in New York State (Makarewicz *et al.,* 2001; Witt *et al.,* 2005). Introductions of *C. pengoi* have been linked to negative impacts in the native community including declines in abundance and diversity of small-sized zooplankton in the Baltic Sea (e.g. Ojaveer *et al.,* 2004, Kotta *et al.,* 2005) and Lake Ontario (e.g. Laxson *et al.,* 2003, Warner *et al.,* 2006). *Cercopagis* may compete with larval fish for zooplankton prey, thereby limiting fish production, but may also be of benefit as an alternate food source for planktivorous fish (Kotta et al., 2005).

Similar to dispersal of another invasive cladoceran, *Bythotrephes longimanus*, *Cercopagis* dispersal may be facilitated by overland transport associated with recreational movement of contaminated trailered boats or fishing gear. Like *Bythotrephes*, *Cercopagis* has alternating reproductive modes, with predominant parthenogenetic growth occasionally interrupted by sexual reproduction and production of resting stages. Despite its potential for high population growth and subsequent spread among inland lakes, there have been no studies to date to predict *Cercopagis*' spread in North America.

In this paper, we assess *Cercopagis* spread among lakes by developing a stochastic gravity model to explore propagule pressure associated with recreational boat traffic, and use these estimates of inbound propagule flow as input into a subsequent establishment model. Unlike previous gravity models of NIS dispersal (e.g. Bossenbroek *et al.,* 2001; Leung *et al.,* 2006), we treat recreational traffic as a stochastic process and model traffic between lakes as a random variable described by a statistical distribution. We use this model to test a series of hypotheses about economic, social and geographical factors governing boater traffic between lakes to derive the most parsimonious dispersal model. We then develop a baseline establishment model based on mean propagule flow and lake physicochemistry and propagate variability from the dispersal stage of the invasion sequence to evaluate the effect of stochastic boater movement on estimated probabilities of establishment. This approach allows us to evaluate how uncertainty in earlier stages of the invasion sequence affects our ability to predict lake invasion status and spatial extent of spread.

Methods

Data collection

We conducted creel surveys of recreational boaters at several boat launches in New York State including the Finger Lakes and lakes Erie and Ontario during the summer of 2004. We were able collect data on 534 outbound trips where the recreationalists trailered their boats from 11 lakes with established *Cercopagis* populations, including Lake Erie, Ontario, and nine inland lakes in the Finger Lakes region. Here, we treated Lakes Erie and Ontario similarly to other lakes and considered them as having a single access point where we collected survey data. Information from the surveys included the lake last visited by the boaters, lake(s) they were planning to visit, the length of time they were planning on spending at the lake (average (±1 SD), 1.94 ± 2.14 days), and the amount of money they spent to arrive at the lake ($141 ± 417). Additional data required to parameterize the dispersal model such as lake area and road distance between lakes were extracted from TIGER/Line geospatial data provided by the U.S. Census Bureau (http://www.census.gov/geo/www/tiger/).

Stochastic Gravity Model

To model recreational boater movement between lakes *i* (invaded) and *j* (invaded and noninvaded), we consider the number of pairwise trips as a random variable following a zero-inflated negative binomial (ZINB) distribution, which can be built up by a series of hierarchical submodels (Fig. 1, Stochastic Gravity model). The ZINB is a generalized form of a Poisson distribution and is useful for describing count data that is both overdispersed and contains an excess of zeros that are either structural in nature or arise due to sparse sampling effort. The ZINB distribution is described by three parameters, , where and *k* are the mean number of trips and dispersion parameter of the negative binomial (NB) distribution; and is a parameter describing zero-inflation as the probability that only $y_{ij} = 0$ can occur, and the probability that is occurring.

First, the mean number of trips between lakes *i* and *j* can be expressed in a submodel as:

, (1)

where w_j is a composite measure of destination lake attractiveness, and d_{ij} is the road distance between lakes *i* and *j*, and is a fitted parameter to account for distance decay. Here, trips between *i* and *j* are asymmetric since data was available on whether the recreationalist visited lake *i* before *j*. In addition, the model is limited to cases where .

Destination lake attractiveness, w_j, was in turn parameterized according to:

(2)

with a_j = lake area, c_j = average cost the recreationalist paid to arrive at the lake, p_j = population size of nearest city/town and $d_{p,j}$ is the distance of the nearest city/town to the destination lake. Lake attractiveness is constrained to a minimum 0 by exponentiating the vector of fitted parameters b, with the exception of the power functions associated with population size and distance of the nearest city/town.

Finally, zero-inflation of the ZINB distribution was estimated by:

, (3)

where is the distance between lakes *i* and *j* (in km), and is the surface area of the destination lake (in ha).

Since maximum likelihood estimates for parameters used in estimating , *k* and for the ZINB must be solved simultaneously; the EM (Expectation-Maximization) algorithm (Dempster *et al.,* 1977) is a useful approach for MLE in modeling mixture distributions when other methods fail to converge (see Appendix S1 in Supporting

Information). In general terms, the log-likelihood function for the ZINB is a sum of two components: one log-likelihood describing the probability of zero-inflation (hereafter referred as) and another log-likelihood conditional on this describing a negative binomial distribution weighted by zero-inflation ().

The statistical significance of variables used in measures of the mean number of trips , the dispersion parameter *k* and estimates of zero-inflation were assessed using likelihood ratio tests. Non-significant variables or parameter estimates were eliminated from the dispersal model with the most parsimonious model used in further analyses.

Predicting areas at risk for *Cercopagis* establishment

In the second stage of modelling *Cercopagis* establishment in lakes, significance of propagule pressure and water physicochemistry was tested in a logistic model using *Cercopagis* presence/absence data (Fig. 1, Establishment model). Here, expected propagule pressure to specified destinations is simply the summed inflow of the estimated number of trips into each destination *j* across the invaded source lakes adjusting for zero-inflation, .

Lake area and water quality data from the United States Geological Survey (http://waterdata.usgs.gov/nwis) including specific conductance, turbidity, pH, total phosphorus, total nitrogen and dissolved oxygen concentration were used as predictors of establishment success for 871 destinations. Water quality data for each lake were summarized by taking the average, minimum and maximum value from samples taken since 1998 to the present corresponding to the period beginning with the first record of

Cercopagis establishment. Missing water quality data for 24% of the destination lakes was estimated using kriging from a pool of 29886 sites across New York. Lake physicochemistry and characteristics of the nearest city or town to invaded and noninvaded lakes are summarized in Table 1.

We developed the establishment models in two stages, excluding and then including stochastic variability from the dispersal model. First, we estimated baseline probabilities of establishment by a logistic regression model relating *Cercopagis* presence/absence to propagule pressure and water quality data where expected propagule pressure is based on the best-fit gravity model. The logistic model was simplified as much as possible using the fewest significant predictors which were determined by stepwise evaluation of the model's AIC in both forward and reverse directions based on the expected number of inbound trips, and the complete suite of water physicochemistry. In order to cross-validate the logistic establishment models, we used a jackknife leave-one-out method where each destination lake was left out of the data set in turn and the models trained on the remaining lakes. Establishment probabilities were then estimated for the hold-out samples.

Second, we used a randomization approach to propagate variability from the dispersal model to the most parsimonious establishment model. Similar to the cross-validation of the baseline establishment model, data were repeatedly subset into training with a hold-out observation for testing. At each iteration, random numbers of pairwise trips were drawn from a ZINB distribution parameterized from the dispersal model, $Y_{ij} \sim ZINB(\mu_{ij},k,\omega_{ij})$, summed for each destination, and combined with water quality data (Fig. 1). A logistic regression model was fitted to the training subset, and confidence intervals for the fitted parameters were calculated from bias-corrected 2.5th and 97.5th percentiles of the resampled parameters (Efron & Tibshirani, 1986). In turn, prediction intervals for establishment probability for the testing data subset were based on the confidence limits with the propagated errors.

We evaluated the relationship between establishment probability and *Cercopagis* presence/absence data by means of the shape and area under the receiver operating characteristic curve (AUC) for baseline probability and prediction intervals with and without the propagated errors. This allows us to assess how the inclusion of stochastic variability affects estimates of hit rates (correctly predicting invasions when they occur) and false alarm rates (incorrectly predicting invasions as occurring when in fact, they have not) across a range of probability thresholds. In addition, we calculated the optimum threshold for establishment probability based on the receiver operating characteristic (ROC) curve and establishment probability from the baseline dispersal model without error propagation. The optimum threshold was calculated as the probability of establishment along the ROC curve that was closest to the curve if there was perfect model fit (i.e. 100% hit rate and 0% false alarms) (Liu *et al.,* 2005). Lakes were classified as invasible if their establishment probability was equal to or larger than this threshold. This, in turn, allowed us to compare the numbers of lakes predicted to be invaded based on this threshold and observed invasion status and quantify the effect of stochastic variability based on the frequency of correct invasion predictions and false alarms.

After determining the relationship between establishment probability and observed invasion status, we standardized the coefficients of the logistic regression to assess which variables had the most influence on establishment probability. We standardized the coefficients for each variable according to: , where *b* and *R* are the raw coefficients and square root of Pearson's correlation coefficient and and are standard deviations of the independent and logit-transformed predicted values for presence/absence, respectively (Menard, 2004). This method standardizes for variation in both the independent and dependent variables such that one standard deviation in the change of the independent variables can then be interpreted as producing *b** standard deviations in the dependent variable. Standardized coefficients were calculated for both the baseline model as well as model with stochastic variability. Confidence intervals for the standardized coefficients incorporating stochastic variability were calculated based on 95% bias-corrected bootstrap confidence limits (Efron & Tibshirani, 1986).

In the previous model, the timing of lake invasions was not specifically considered, as flow from currently invaded lakes (n = 11) was modeled to other invaded and noninvaded lakes. A subsequent dispersal and establishment model considered propagule flow from the first invaded lake (Lake Ontario) to Lake Erie and the inland lakes in New York. This scenario allows for a true validation of a model for *Cercopagis* spread since model predictions were compared against the invasion status of lakes that were invaded later on. In this scenario, we recalculated the fitted parameters for the stochastic gravity model using Lake Ontario as the sole source. Similar to the model for the 11 source lakes, we generated a random number of trips leaving Lake Ontario to each destination, and fitted a logistic establishment model with the same water chemistry variables selected from the previous scenario.

Results

Stochastic dispersal model

A summary of hypothesis tests and corresponding likelihood ratio tests are presented in Table 2. Lake area was a significant term in the logistical model to estimate zero-inflation (Eqn. 3) but not as a measure of destination attractiveness (Eqn. 2). Population size was also significant in the model for lake attractivity (Eqn. 2).

The estimated number of trips between pairs of lakes is given by with dispersion parameter for the ZINB distribution, $k = 0.92$. Nested within this model, lake attractiveness is modeled as , where c is the average cost to arrive at lake j and p_j is population size of the nearest city or town to the destination lake. This reduced model was not significantly different from the full model of lake attractiveness as determined by likelihood ratio testing (χ^2=0.669, d.f.=4, P = 0.955).

The final model for estimation of zero-inflation is given by:

for each pairwise trip with fitted parameters γ. Confidence limits for the fitted parameters , and are presented in Table 3.

Predicting *Cercopagis* establishment

After elimination of non-significant variables based on the stepwise procedure, the baseline logistic model for estimating the probability of *Cercopagis* establishment was determined by propagule pressure μ_j, lake area (a_j) maximum specific conductance (SC) and minimum and maximum turbidity (TB):

With the addition of the water chemistry data, there was a significant improvement in model fit over a logistic model using propagule pressure alone (χ^2 = 22.48, d.f. = 4, $P = 0.0002$). Finally, when coefficients for the logistic model were standardized, maximum and minimum turbidity had the most influence on estimated probability of establishment, followed by propagule pressure (Fig. 2).

There was high concordance with the baseline probability of establishment estimated for the testing data and observed *Cercopagis* presence/absence for each of the destination lakes. The area under the Receiver Operating Characteristic Curve (AUC = 0.984) was significantly different from 0.5 ($P < 0.001$) (Fig. 3(a)).

Effect of stochastic variability on predicting establishment

The impact from the inclusion of stochastic variability in the dispersal model on the ability of the establishment models to predict the invasion status of lakes depended on the sensitivity of the metric used to measure the change. At the upper 95% prediction limit excluding stochastic variability, there was a negligible decrease in AUC relative to the baseline model (0.991 vs. 0.994) (Fig. 3(a)). However, when we chose an optimum threshold for the baseline model ($P = 0.010$) above which lakes are predicted to be invaded, the chance of incorrectly predicting a lake as invaded when it is not (false alarm) increased from 3% to 100%, with only a 1% chance of correctly predicting the overall invasion status of the lakes (Table 4). Similarly, in a comparison of the upper 95% prediction intervals for the models excluding and including variability from the dispersal stage, there was no change in the overall AUC (Fig. 3a), but the chance of committing false alarms decreased from 100% to 78% (Table 4).

At the lower 95% prediction limits for the models without and with stochastic variability, estimated probabilities of establishment were poor predictors of observed invasion status. Probabilities of establishment estimated from the model without stochastic variation were able to significantly predict invasion status (AUC = 0.65, P = 0.03), but probabilities based on included stochasticity were unable to predict better than random (AUC = 0.60, P = 0.12). There was no change in either false alarm or hit rates at the lower 95% PI in a comparison of the models excluding and including stochasticity (Table 4).

Areas with high probabilities of *Cercopagis* establishment were concentrated primarily on a cluster of seven invaded lakes in the Finger Lakes region when the spatial extent of establishment probabilities is taken under consideration (Fig. 4a). When variability was propagated from the dispersal model, the spatial extent of predicted establishment based on the 95% prediction limits differed extensively. For the upper prediction limit, the spatial extent of establishment probability at the higher prediction limit was overly inclusive. At a probability threshold between 0.81 and 1.00, predicted areas of establishment extended from Lake Ontario to the Pennsylvania border (Fig. 4b). In contrast, predicted areas at high risk of invasion were focused on the cluster of lakes already invaded with the addition of several isolated hotspots in eastern New York State when stochasticity was propagated. (Fig. 4c).

Dispersal and establishment models with Lake Ontario as source for invasions

Parameter estimates for the dispersal model for Lake Ontario were not significantly different from the dispersal model estimated using all 11 source lakes as determined by overlapping confidence intervals (Table 3). In the sub-model estimating the mean number of pairwise trips leaving Lake Ontario, road distance between lakes was slightly less of a deterrent (i.e. less negative) than all 11 source lakes were considered (α = -0.485), but not significantly based on confidence intervals [-0.578, -0.393]. In terms of destination lake attractivity, w_j was negatively related to the average cost of travelling to a specified destination (β_4=-2.11) but the confidence intervals indicate that it was not significant from 0, unlike from the other scenario. The dispersion parameter for the ZINB, k = 1.39 [0.681, 2.858] was also not significantly larger than that for the 11 source lakes model, and parameters used in the estimation of zero inflation,

,

were also not different from parameter estimates for the other scenario.

Establishment model with Lake Ontario as sole source

In the scenario with Lake Ontario as the sole source, the probability of *Cercopagis* establishment was described by:

.

There was no significant difference between the fitted parameters for this establishment model in this scenario as compared to the 11 source lakes establishment model based on overlapping confidence limits (Table 3). Also, as in the scenario for the 11 source lakes, maximum and minimum levels of turbidity had the greatest influence on the probability of establishment, followed by propagule pressure (Fig. 2). When Lake Ontario was considered as the only source, performance of the establishment model was extremely high in being able to discriminate between true invasions and false alarms as evaluated from the Receiver Operating Characteristic Curve (AUC=0.984, P << 0.001).

Prediction limits for Cercopagis *establishment with Lake Ontario as single source*

Estimated probabilities of establishment varied greatly for invaded and non-invaded lakes when stochastic variability was propagated from the dispersal model. The average probability of establishment for invaded lakes was 0.69 but ranged from 4.5E-7 to 0.97 at the lower and upper 95% prediction limits when stochastic error was propagated. For non-invaded lakes, stochastic variability resulted in establishment probabilities that differed by two orders of magnitude between the lower and upper 95% prediction [averages of 5.96E-7 vs. 1.36E-1]. The average probability of establishment for non-invaded lakes was 4.53E-3 for the baseline model without error propagation. Similar to the scenario with the 11 currently invaded lakes as sources, there was negligible change in the AUC from probabilities estimated from the baseline model to probabilities at the upper 95% prediction limit (Fig. 2(b)). Likewise, hit rates and false alarm rates were more sensitive to the addition of stochasticity based on an optimum probability threshold of p = 0.013 derived from the baseline model. At the upper prediction interval, the probability of committing false alarms decreased from 1.00 to 0.38 when variability was propagated, resulting in an overall 73% chance of correctly predicting invasions overall (Table 4). At the lower prediction limit, the ability of the model to correctly predict invasions (hit rate) did not change in comparison to the model without stochastic dispersal, but remained at 0 (Table 4). Here, all estimated probabilities were low (< 0.2) and the model was unable to discriminate between true invasions and false alarms (AUC = 0.43, P = 0.78).

Spatial patterns of predicted establishment with Lake Ontario as the single source (Fig. 4d-f) are almost identical to the scenario with 11 lakes as sources (Fig. 4a-c). In the baseline models for both scenarios (Fig. 4a,d), a similar region of moderate probability of invasion (0.21 – 0.40) formed around the Finger Lakes. However, in the Lake Ontario scenario, lakes within this cluster were not at as a high a risk of invasion as in the 11 source lake scenario. One notable difference between the two scenarios is that small clusters of high and moderate risk areas in eastern New York State are contiguous in the 11 lake scenario at upper 95% PI (Fig. 4b,c) but are more isolated when only Lake Ontario is considered as single source (Fig. 4e,f) due to lower levels of propagule pressure. In other words, propagule flow from Lake Ontario is sufficient to create potential satellite colonies, but additional propagule flow from the Finger Lakes may allow for infilling of the satellite colonies.

Discussion

Multi-stage invasion models are a useful approach to implementing the invasion sequence conceptual model because they allow us to identify factors that limit or facilitate invasions at different stages of the sequence and provide flexibility in choosing suitable sub-models appropriate to the biology of the organism. In particular, estimating factors that influence dispersal of NIS is the significant first step in risk management because preventative measures are most readily applied at the introduction stage of the invasion sequence (Lodge et al. 2006). In this paper, we developed stochastic gravity and logistic establishment models to forecast spread of an aquatic NIS. Since we used a stochastic form of the gravity model, we were able to propagate variability from early stages of the invasion sequence to final predictions of spread.

The use of stochastic gravity models to model recreational traffic has some statistical advantages over deterministic approaches. First, stochastic gravity models require specifying a probability distribution about the probability of individual trips being made between sources and destinations, as opposed to deterministic models that model mean interaction traffic and assume a probability distribution about residual variability during the model fitting process. By specifying an appropriate trip distribution - such as Poisson or negative binomial if the number of trips is being modeled - estimated traffic would be less biased than if an incorrect distribution were assumed (Flowerdew and Aiken, 1982). Further, with these discrete distributions, cases where there are zero trips between a source and destination may be modeled explicitly instead of excluding them from the analysis. In other words, stochastic models based on discrete distributions are likely to perform better at smaller sample sizes. Finally, stochastic models that are fitted using maximum-likelihood estimation allow for statistical testing of hypotheses surrounding the number of trips through the use of likelihood ratio tests or information-theoretic approaches such as Akaike’s Information Criterion.

Invasions are ultimately stochastic processes, and even in strictly controlled experiments (e.g. Melbourne & Hastings, 2009), stochasticity will limit our ability to make predictions about invasion dynamics. However, an assessment of various sources of uncertainty in model predictions is useful. In this study, a comparison of model predictions among baseline models, and models including and excluding stochastic variability from earlier stages of the model, in addition to predicting spread under different scenarios, permits us to assess contributions of different sources of uncertainty in final predictions of establishment.

Uncertainty inherent in model parameterization may be considered as a key source of variability. Parameterization uncertainty was tested by comparing model predictions from the best-fit or baseline model, to the 95% prediction limits defined by the 95% confidence limits of the fitted parameters. Here, we consider this an evaluation of the input variables' statistical uncertainty, defined as uncertainty that may be described by statistical terms as deviations in the variables or parameters from the true value (Walker *et al.,* 2003). In our model, the high increase in false alarm rates (Table 4) and overly inclusive predicted area of establishment at the upper 95% prediction limit (Fig. 4) indicates that statistical uncertainty is a major source of uncertainty. That is, modest changes in the fitted parameters result in an overestimation of probabilities of establishment resulting in increased false alarms. Statistical uncertainty in models of predicted spread may be reduced through increased data collection in order to develop stronger model discriminators between invaded and non-invaded habitats.

Stochastic variability of model inputs can also be described as statistical uncertainty because it addresses whether the data set captures the true variability in the population. The decrease in false alarm rate at the upper 95% PI when stochasticity was propagated relative to the model without stochasticity indicates that accounting for sources of uncertainty is not as critical as addressing uncertainty inherent in the model. Nevertheless, we recommend the inclusion of stochastic variability in model inputs since estimated areas of spread were lower than the upper 95% PI of the model without propagated error. This reduced area of spread when stochastic input variables were included is consistent with population spread models of Clark *et al.* (2001), where propagated variability in the net reproduction rate, R_0, resulted in lower rates of forest spread.

A comparison of predictions in spread under the two scenarios with Lake Ontario only and 11 invaded lakes as sources addresses the issue of scenario uncertainty frequently used in policy analysis (Walker *et al.,* 2003). Scenario uncertainty involves uncertainty in how the system and processes driving the system may develop through time, and reflects alternative outcomes or conditions. In the context of this study, we are comparing two invasion trajectories about outbound propagule flow. The nearly identical patterns of spread under the two scenarios indicate that Lake Ontario is a major driving force in *Cercopagis* expansion since it is present in both scenarios, and thus may serve as a 'hub' for future expansion across a network of invaded lakes connected by boater traffic (see Muirhead & MacIsaac 2005).

Finally, one source of variability that our model does not contain is stochasticity present in local population growth over time. As time progresses, stochastic population growth may increase the probability of lake becoming a future source for invasions if the population is likely to expand faster than decreasing below a threshold where the population is likely to collapse or experience Allee effects (e.g., Drake & Lodge, 2006). That is, there is likely a greater chance of propagules being transported away from the lake depending on the source population size, although this relationship is difficult to quantify. In order to develop a complete population spread model based on the invasion sequence, population estimates of the NIS are required but are unfeasible for planktonic species.

The combination of propagule pressure and habitat suitability in determining establishment success for *Cercopagis* is similar for other zooplankton species. In particular, *Cercopagis* is more likely to found in larger, clear lakes (i.e. low turbidity) with high propagule inflow from human-mediated dispersal, similar to the confamilial *Bythotrephes longimanus.* European and North American lakes that support *Bythotrephes* have significantly greater surface area and are deeper and more transparent than those in which it is absent (MacIsaac *et al.,* 2000; Branstrator *et al.,* 2006; Weisz & Yan, 2010). In both the dispersal and establishment models for *Cercopagis*, the likelihood of travelling between two lakes (estimate of zero-inflation), as well as the probability of establishment, increased with increasing lake area. Lake area may be an indication of subjective attractiveness for recreationalists since it has been shown to be positively related to the probability of making a trip to that lake (Siderelis & Moore, 1998) as well as the average amount of boater traffic on a lake (Reed-Anderson *et al.,* 2000). In terms of *Cercopagis* establishment, lake area may influence habitat diversity, availability of refuge from fish predators and, consequently, population size of the NIS. Finally, specific conductance may not have a direct impact on *Cercopagis* establishment, but it is an indicator of lake position in a watershed. Lakes that are larger, deeper and at lower elevations tend to have higher specific conductance. In a zooplankton community analysis for Wisconsin lakes, Dodson *et al.* (2009) found that larger zooplankton species were found in these larger lakes.

The clustering of inland lakes invaded by *Cercopagis* is limited to the Finger Lakes region and raises a key question of why *Cercopagis* hasn't spread like *Bythotrephes*, despite ample time to do so. *Cercopagis* displays many of the life-history and physiological traits expected to confer an advantage in colonizing populations. For example, *Cercopagis* frequently has greater absolute abundance and fecundity, and thus we expect higher number of *Cercopagis* propagules based on relative abundance (Makarewicz *et al.*, 2001, Yan *et al.,* 2001).

One possibility of why *Cercopagis* has not spread extensively involves the timing of production of resting stages relative to seasonal patterns of recreational boat traffic. Like many cladocerans, *Cercopagis* produces gametogenic resting eggs that are likely the life stage to survive dispersal among lakes. Whereas recreational boating traffic occurs predominantly between May and early September, peak abundance of *Cercopagis* females with resting stages does not occur until mid-August in Lake Ontario (Makarewicz *et al.*, 2001) and late September in the Baltic Sea (Gorokhova *et al.,* 2000). Seasonal recreational boating activity is declining at this time, thereby reducing the likelihood of human-mediated species dispersal.

Allocation of sampling effort and optimum control to detect and contain spread

The spatial pattern of lakes at high risk of invasion by *Cercopagis* has implications on whether to allocate more effort to sampling and detecting new invasions versus controlling secondary spread. In terms of sampling effort to detect new invasions, risk maps allow the identification of invasion ‘hotspots’ where sampling can be prioritized to increase detection sensitivity for low population sizes. For example, in a survey for *Cercopagis* in Lake Ontario by Harvey *et al.* (2009), the probability of detecting the species was greater for targeted spatial arrangement of samples rather than lower-effort sampling over a great area when population densities were low. Based on our simulations, areas at high risk of *Cercopagis* invasion in New York State are primarily concentrated around a cluster of lakes that are currently invaded, although there exist several high-risk areas surrounding this cluster that currently are not invaded. If large volumes of recreational boating traffic depart from these high-risk areas, effort to detect new invasions should be focused on these potential satellite colonies rather than sampling for lakes proximal to those already invaded because satellite colonies may increase the overall rate of spread by exceeding that of the central core of the invasion (Moody & Mack, 1988; Blackwood *et al.,* 2010).

Alternatively, if management effort is targeted toward limiting secondary spread, the optimal solution is contingent on the spatial configuration of the invasion, rates of spread, and control strategies. Despite a relatively long time since its introduction into the Great Lakes, *Cercopagis* has not spread in New York State at the rate or geographic extent of *Bythotrephes* in Ontario (Muirhead & MacIsaac, 2005; Weisz & Yan 2010). Although eradication seems unfeasible for planktonic species such as *Cercopagis*, an optimal strategy of “slow the spread” based on creating barriers to dispersal is recommended since the area currently invaded is limited (Sharov, 2004). In instances where the invasible habitat is not homogenous, but is distributed in patches such as lakes, the optimal control strategy is to limit the amount of propagule dispersal among patches. Here, the spatial arrangement of patches has implications on where management efforts to control spread should be applied. For example, in an optimal model for zebra mussel spread, Potapov & Lewis (2008) show that disrupting the flow between clusters of lakes by washing trailered boats is preferred over disrupting propagule transfer within a cluster. Similarly, if we consider propagule dispersal between invaded sources and destinations as a source-sink model, Travis & Park (2004) illustrate how the optimal control strategy for overall population size does not always involve reductions in the source (i.e. invaded) habitat, but is strongly influenced by the level of dispersal and the strength of population decline in the sink habitat. In situations where density-dependent dispersal is low, as is often the case in invasions, they suggest a dynamic strategy where eradication effort switches between sources and sinks as opposed to splitting effort between the two. However, if a species exhibits high density-dependent dispersal, reducing the source population will be the most effective strategy. For planktonic NIS such as *Cercopagis*, management options may include education of recreationalists through awareness of invaded lakes in the Finger Lakes region; inspecting fishing lines or water-based vectors such as bait buckets, live wells or bilge water for animals; and washing boats and trailers after emerging from invaded lakes. Commercial solutions such as the development of fishing lines that prevent *Cercopagis* attachment may prove highly effective (Jacobs & MacIsaac, 2007).

Hierarchical approach to modeling invasions

The utility of combining stochastic dispersal and establishment models in a hierarchical approach provides a flexible framework to address a wide range of hypotheses in invasion biology. While we focus on aquatic NIS, the same approach could be used to assess the spread of terrestrial species. For example, stochasticity in wind currents affecting seed dispersal of plants or insect dispersal may be propagated into spatially- and temporally-explicit establishment models based on local environmental conditions. In a series of individual-based models for *Rhododentron* spread, Harris *et al,* (2009) show how differences between minimum and maximum reported windspeed can halve the time for seedlings to establish outside quarantine zones. Further, the authors were able to make recommendations on optimal control measures based on these simulations of stochastic dispersal of seedlings and age-dependent seed production.

The hierarchical approach also allows for testing multiple dispersal pathways. Pathways may be modeled separately and subsequently used as independent predictors in a combined establishment model allowing to test hypotheses of human-mediated versus natural dispersal. Variability in each of the pathways may be propagated through the establishment model to assess how uncertainty in each of the pathways can influence predicted rates of spread. Alternatively, comparisons of dispersal kernels derived for multiple pathways may be compared to observed patterns of spread to identify relative importance. For example, Wichmann *et al.* (2009) compared dispersal kernels derived from stochastic simulations of wind-mediated seed movement for black mustard (*Brassica nigra*) plants to human-mediated dispersal obtained from field experiments. Their results suggest that human-mediated dispersal is the primary pathway for observed long-distance dispersal. As a consequence, managed reductions in propagule supply transported over long distances may reduce the chance of establishing satellite colonies (Hulme, 2003).

In conclusion, models developed here address factors that influence human-mediated propagule pressure and establishment for the nonindigenous zooplankter *Cercopagis pengoi* following a conceptual model of the invasion sequence. Lake area, travel costs and population sizes were significantly related to the amount of boat traffic between lakes; in turn, boat traffic, lake area, specific conductance and turbidity were significant predictors of *Cercopagis* establishment. This study highlights how propagating stochasticity associated with dispersal throughout the invasion sequence reduces the rate of detecting false alarms for predicted spread.

Acknowledgements

We thank Drs. Alexei Potapov, Subhash Lele and Norman Yan for helpful discussions during this project. We are grateful for financial support from the Canadian Aquatic Invasive Species Network (CAISN), NSERC Discovery program (HJM and MAL), an Ontario Graduate Scholarship and NSERC postdoctoral fellowship to JRM, a Department of Fisheries and Oceans Invasive Species Research Chair to HJM, and Canada Research Chair to MAL.

References

Blackwood, J., Hastings, A. & Costello, C. (2010) Cost-effective management of invasive species using linear-quadratic control. *Ecological Economics*, **69**, 519-527.

Bossenbroek, J.M., Kraft, C.E. & Nekola, J.C. (2001) Prediction of long-distance dispersal using gravity models: zebra mussel invasion of inland lakes. *Ecological Applications*, **11**, 1778-1788.

Branstrator, D.K., Brown, M.E., Shannon, L.J., Thabes, M. & Heimgartner, K. (2006) Range expansion of *Bythotrephes longimanus* in North America: evaluating habitat characteristics in the spread of an exotic invader. *Biological Invasions*, **8**, 1367-1379.

Clark, J.S., Lewis, M. & Horváth, L. (2001) Invasion by extremes: Population spread with variation in dispersal and reproduction. *The American Naturalist*, **157**, 537-554.

Colautti, R.I., Grigorovich, I.A. & MacIsaac, H.J. (2006) Propagule pressure: A null model for biological invasions. *Biological Invasions*, **8**, 1023-1037.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, 1-38.

Dodson, S.I., Newman, A.L., Will-Wolf, S., Alexander, M.L., Woodford, M.P. & Van Egeren, S. (2009) The relationship between zooplankton community structure and lake characteristics in temperate lakes (Northern Wisconsin, USA). *Journal of Plankton Research*, **31**, 93-100.

Drake, J. & Lodge, D. (2006) Allee effects, propagule pressure and the probability of establishment: Risk analysis for biological invasions. *Biological Invasions*, **8**, 365-375.

Dzialowski, A.R., Lennon, J.T. & Smith, V.H. (2007) Food web structure provides biotic resistance against plankton invasion attempts. *Biological Invasions*, **9**, 257-267.

Efron, B. & Tibshirani, R. (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, **1**, 54-77.

Finnoff, D., Shogren, J.F., Leung, B. & Lodge, D.M. (2007) Take a risk: Preferring prevention over control of biological invaders. *Ecological Economics*, **62**, 216-222.

Flowerdew, R. & Aitkin, M. (1982) A method of fitting the gravity model based on the Poisson distribution. *Journal of Regional Science*, **22**, 191-202.

Fridley, J.D., Stachowicz, J.J., Naeem, S., Sax, D.F., Seabloom, E.W., Smith, M.D., Stohlgren, T.J., Tilman, D. & Von Holle, B. (2007) The invasion paradox: Reconciling pattern and process in species invasions. *Ecology*, **88**, 3-17.

Gollasch, S. (2006) Overview on introduced aquatic species in European navigational and adjacent waters. *Helgoland Marine Research*, **60**, 64-69.

Gorokhova, E., Aladin, N. & Dumont, H.J. (2000) Further expansion of the genus *Cercopagis* (Crustacea, Branchiopoda, Onychopoda) in the Baltic Sea, with notes on the taxa present and their ecology. *Hydrobiologia*, **429**, 207-218.

Harris, C.M., Park, K.J., Atkinson, R., Edwards, C. & Travis, J.M.J. (2009) Invasive species control: Incorporating demographic data and seed dispersal into a management model for *Rhododendron ponticum*. *Ecological Informatics*, **4**, 226-233.

Harvey, C.T., Qureshi, S.A. & MacIsaac, H.J. (2009) Detection of a colonizing, aquatic, non-indigenous species. *Diversity and Distributions*, **15**, 429-437.

Heger, T. & Trepl, L. (2003) Predicting biological invasions. *Biological Invasions*, **5**, 313-321.

Herborg, L.-M., Jerde, C.L., Lodge, D.M., Ruiz, G.M. & MacIsaac, H.J. (2007) Predicting invasion risk using measures of introduction effort and environmental niche models. *Ecological Applications*, **17**, 663-674.

Hulme, P.E. (2003) Biological invasions: winning the science battles but losing the conservation war? *Oryx*, **37**, 178-193.

Hulme, P.E. (2009) Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, **46**, 10-18.

Jacobs, M.J. & MacIsaac, H.J. (2007) Fouling of fishing line by the waterflea *Cercopagis pengoi*: a mechanism of human-mediated dispersal of zooplankton? *Hydrobiologia*, **583**, 119-126.

Jansakul, N. (2005) Fitting a zero-inflated negative binomial model via R. *Proceedings of the 20th International Workshop on Statistical Modelling*, 277-284.

Kolar, C.S. & Lodge, D.M. (2001) Progress in invasion biology: predicting invaders. *Trends in Ecology & Evolution*, **16**, 199-204.

Kotta, J., Kotta, I., Simm, M., Lankov, A., Lauringson, V., Pollumae, A. & Ojaveer, H. (2006) Ecological consequences of biological invasions: three invertebrate case studies in the north-eastern Baltic Sea. *Helgoland Marine Research*, **60**, 106-112.

Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 1-14.

Laxson, C.L., McPhedrin, K.N., Makarewicz, J.C., Telesh, I.V. & MacIsaac, H.J. (2003) Effects of the non-indigenous cladoceran *Cercopagis pengoi* on the lower food web of Lake Ontario. *Freshwater Biology*, **48**, 2094-2106.

Leung, B., Bossenbroek, J.M. & Lodge, D.M. (2006) Boats, pathways, and aquatic biological invasions, estimating dispersal potential with gravity models. *Biological Invasions*, **8**, 241-254.

Leung, B., Lodge, D.M., Finnoff, D., Shogren, J.F., Lewis, M.A. & Lamberti, G.A. (2002) An ounce of prevention or a pound of cure: bioeconomic risk analysis of invasive species. *Proceedings of the Royal Society of London, Series B*, **269**, 2407-2413.

Leung, B. & Mandrak, N.E. (2007) The risk of establishment of aquatic invasive species: joining invasibility and propagule pressure. *Proceedings of the Royal Society B*, **274**, 2603-2609.

Linneman, H.V. (1966) *An Econometric Study of International Trade Flows.* North-Holland Publishing Company, Amsterdam.

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

Lockwood, J.L., Cassey, P. & Blackburn, T. (2005) The role of propagule pressure in explaining species invasions. *Trends in Ecology & Evolution*, **20**, 223-228.

Lodge, D.M., Williams, S., MacIsaac, H.J., Hayes, K.R., Leung, B., Reichard, S., Mack, R.N., Moyle, P.B., Smith, M., Andow, D.A., Carlton, J.T. & McMichael, A. (2006) Biological invasions: Recommendations for US policy and management. *Ecological Applications*, **16**, 2035-2054.

MacIsaac, H.J., Borbely, J., Muirhead, J.R. & Graniero, P. (2004) Backcasting and forecasting biological invasions of inland lakes. *Ecological Applications*, **14**, 773-783.

MacIsaac, H.J., Grigorovich, I.A., Hoyle, J.A., Yan, N.D. & Panov, V.E. (1999) Invasion of Lake Ontario by the Ponto-Caspian predatory cladoceran *Cercopagis pengoi*. *Canadian Journal of Fisheries and Aquatic Sciences*, **56**, 1-5.

MacIsaac, H.J., Ketelaars, H.A.M., Grigorovich, I.A., Ramcharan, C.W. & Yan, N.D. (2000) Modeling *Bythotrephes longimanus* invasions in the Great Lakes basin based on its European distribution. *Archive für Hydrobiologie*, **149**, 1-21.

Makarewicz, J.C., Grigorovich, I.A., Mills, E., Damaske, E., Cristescu, M.E., Pearsall, W., LaVoie, M.J., Keats, R., Rudstam, L., Hebert, P., Halbritter, H., Kelly, T., Matkovich, C. & MacIssac, H.J. (2001) Distribution, fecundity, and genetics of *Cercopagis pengoi* (Ostroumov) (Crustacea, Cladocera) in Lake Ontario. *Journal of Great Lakes Research*, **27**, 19-32.

Melbourne, B.A., Cornell, H.V., Davies, K.F., Dugaw, C.J., Elmendorf, S., Freestone, A.L., Hall, R.J., Harrison, S., Hastings, A., Holland, M., Holyoak, M., Lambrinos, J., Moore, K. & Yokomizo, H. (2007) Invasion in a heterogeneous world: resistance, coexistence or hostile takeover? *Ecology Letters*, **10**, 77-94.

Melbourne, B.A. & Hastings, A. (2009) Highly variable spread rates in replicated biological invasions: Fundamental limits to predictability. *Science*, **109**, 1536-1539.

Menard, S. (2004) Six approaches to calculating standardized logistic regression coefficients. *The American Statistician*, **58**, 218-223.

Moles, A.T., Gruber, M.A.M. & Bonser, S.P. (2008) A new framework for predicting invasive plant species. *Journal of Ecology*, **96**, 13-17.

Moody, M.E. & Mack, R.N. (1988) The spread of plant invasions: the importance of nascent foci. *Journal of Applied Ecology*, **25**, 1009-1021.

Muirhead, J.R. & MacIsaac, H.J. (2005) Development of inland lakes as hubs in an invasion network. *Journal of Applied Ecology*, **42**, 80-90.

Ojaveer, H., Simm, M. & Lankov, A. (2004) Population dynamics and ecological impact of the non-indigenous *Cercopagis pengoi* in the Gulf of Riga (Baltic Sea). *Hydrobiologia*, **522**, 261-269.

Reaser, J.K., Meyerson, L.A. & Von Holle, B. (2008) Saving camels from straws: how propagule pressure-based prevention policies can reduce the risk of biological invasion. *Biological Invasions*, **10**, 1085-1098.

Reed-Andersen, T., Bennett, E.M., Jorgensen, B.S., Lauster, G., Lewis, D.B., Nowacek, D., Riera, J.L., Sanderson, B.L. & Stedman, R. (2000) Distribution of recreational boating across lakes: do landscape variables affect recreational use? *Freshwater Biology*, **43**, 439-448.

Ricciardi, A. (2007) Are modern biological invasions an unprecedented form of global change? *Conservation Biology*, **21**, 329-336.

Richardson, D.M., Pyšek, P., Rejmánek, M., Barbour, M., Panetta, F.D. & West, C.J. (2000) Naturalization and invasion of alien plants: concepts and definitions. *Diversity and Distributions*, **6**, 93-107.

Rouget, M. & Richardson, D.M. (2003) Inferring process from pattern in plant invasions: a semimechanistic model incorporating propagule pressure and environmental factors. *American Naturalist*, **162**, 713-724.

Sakai, A.K., Allendorf, F.W., Holt, J.S., Lodge, D.M., Molofsky, J., With, K.A., Baughman, S., Cabin, R.J., Cohen, J.E., Ellstrand, N.C., McCauley, D.E., O’Neil, P., Parker, I.M., Thompson, J.N. & Weller, S.G. (2001) The population biology of invasive species. *Annual Review of Ecology and Systematics*, **32**, 305-332.

Schneider, D.W., Ellis, C.D. & Cummings, K.S. (1998) A transportation model assessment of the risk to native mussel communities from zebra mussel spread. *Conservation Biology*, **12**, 788-800.

Sharov, A.A. (2004) Bioeconomics of managing the spread of exotic pest species with barrier zones. *Risk Analysis*, **24**, 879-892.

Siderelis, C. & Moore, R.L. (1998) Recreation demand and the influence of site preference variables. *Journal of Leisure Research*, **30**, 301-318.

Simberloff, D. (2003) How much information on population biology is needed to manage introduced species? *Conservation Biology*, **17**, 83-92.

Simberloff, D. (2009) The role of propagule pressure in biological invasions. *Annual Review of Ecology, Evolution and Systematics*, **40**, 81-102.

Stachowicz, J.J., Fried, H., Osman, R.W. & Whitlach, R.B. (2002) Biodiversity, invasion resistance, and marine ecosystem function: reconciling pattern and process. *Ecology*, **83**, 2575-2590.

Statzner, B., Bonada, N. & Dolédec, S. (2007) Biological attributes discriminating invasive from native European stream macroinvertebrates. *Biological Invasions*, **10**, 1387-3547.

Theoharides, K.A. & Dukes, J.S. (2007) Plant invasion across space and time: factors affecting nonindigenous species success during four stages of invasion. *New Phytologist*, **176**, 256-273.

Thuiller, W., Richardson, D.M., Rouget, M., Proches, S. & Wilson, J.R.U. (2006) Interactions between environment, species traits, and human uses describe patterns of plant invasions. *Ecology*, **87**, 1755-1769.

Von Holle, B. & Simberloff, D. (2005) Ecological resistance to biological invasion overwhelmed by propagule pressure. *Ecology*, **86**, 3212-3218.

Walker, W.E., Harremoës, P., Rotmans, J., Van der Sluijs, J.P., Van Asselt, M.B.A., Janssen, P. & Von Krauss, M.P.K. (2003) Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment*, **4**, 5-17.

Warner, D.M., Rudstam, L.G., Benoît, H.P., Mills, E.L. & Johannsson, O.E. (2006) Changes in seasonal nearshore zooplankton abundance patterns in Lake Ontario following establishment of the exotic predator *Cercopagis pengoi*. *Journal of Great Lakes Research*, **32**, 531-542.

Weisz, E.J. & Yan, N.D. (2010) Relative value of limnological, geographic, and human use variables as predictors of the presence of *Bythotrephes longimanus* in Canadian Shield lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, **67**, 462-472.

Witt, A.M., Dettmers, J.M. & Cáceres, C.E. (2005) *Cercopagis pengoi* in southwestern Lake Michigan in four years following invasion. *Journal of Great Lakes Research*, **31**, 245-252.

Wittmann, M.C., Alexander, M.J., Soons, M.B., Galsworthy, S., Dunne, L., Gould, R., Fairfax, C., Niggemann, M., Hails, R.S. & Bullock, J.M. (2009) Human-mediated dispersal of seeds over long distances. *Proceedings of the Royal Society of London, Series B*, **276**, 523-532.

Zipf, G.K. (1946) The P1P2/D hypothesis: on the intercity movement of persons. *American Sociological Review*, **11**, 677-686.

Biosketches:

J. R. Muirhead is currently a Post-doctoral Fellow at the Smithsonian Environmental Research Center in Edgewater, Maryland. His research focuses on developing statistical models for invasive species dispersal and establishment. M.A. Lewis is a professor who specializes in mathematical ecology and invasion biology. H. J. MacIsaac is a professor who studies invasion ecology, particularly factors affecting invasions of lakes and coastal marine ecosystems.

Author contributions: J.R.M. and H.J.M. conceived the original project, M.A.L. greatly expanded its scope, and J.R.M. analyzed data and led the writing.

Table 1. Means (+/-1 standard deviation) for measures of lake physicochemistry and factors influencing lake attractivity to recreational traffic (N=871 lakes). Abbreviations for water chemistry: SC – specific conductance, DO – dissolved oxygen, P – phosphorus, TN – total nitrogen, Turb – Turbidity.

Parameter	Noninvaded Lakes	Invaded Lakes
Lake area (ha)	3.47E+02 (3.96E+03)	3.76E+05 (8.76E+05)
Nearest city area (km^2)	30.71 (90.26)	7.51 (11.90)
Population Size of nearest city	4.21E+04 (5.45E+05)	6.65E+03 (1.47E+04)
Distance to City (km)	10.18 (8.64)	14.48 (28.85)
min SC (mS cm^{-2})	1.28E+02 (2.67E+02)	3.31E+02 (2.55E+02)
mean SC (mS cm^{-2})	2.30E+02 (1.01E+03)	6.95E+02 (1.02E+03)
max SC (mS cm^{-2})	4.09E+02 (2.15E+03)	2.03E+03 (5.26E+03)
min DO (mg L^{-1})	6.68 (2.10)	7.67 (2.22)

max DO (mg L^{-1})	12.84 (3.15)	14.07 (1.13)
min P (mg L^{-1})	0.02 (0.01)	0.04 (0.06)
mean P (mg L^{-1})	0.04 (0.04)	0.09 (0.07)
max P (mg L^{-1})	0.22 (0.30)	0.42 (0.49)
min pH	6.35 (1.00)	7.32 (0.83)
mean pH	7.01 (0.72)	7.94 (0.52)
max pH	7.75 (0.71)	8.39 (0.49)
min TN (mg L^{-1})	0.62 (0.44)	1.63 (1.12)
mean TN (mg L^{-1})	1.15 (0.63)	2.47 (1.16)
max TN (mg L^{-1})	2.64 (1.48)	4.34 (2.06)
min Turb (NTU)	1.45 (1.67)	4.37 (3.70)

mean Turb (NTU)	5.02 (9.57)	17.81 (8.31)
max Turb (NTU)	29.17 (63.75)	105.16 (111.04)

Table 2. Summary of hypothesis tests for variables used as measures of lake attractivity, covariates in the logistic regression and fitted parameters. and refer to the log-likelihood estimates for the zero-inflated and weighted negative binomial components of the ZINB.

Model	Parameter/ Hypothesis tested	LL_{NB}	LL_{ω}	LL_{Total}	$LL_{Total} -$ max(LL_{Total})	d.f	χ^2	$\Pr\left(\chi^2\right)$
1	Full model	-397.807	-310.909	-708.716				
2	Testing for significance of intercept $\exp(\beta 0) = 0$	-397.819	-310.927	-708.746	-0.031	1	0.062	8.04E-01
3	Testing for significance of lake area (a=0) (removed from LL_{NB} only)	-398.153	-310.776	-708.929	-0.213	2	0.426	8.08E-01
4*	Testing for significance of lake area (a=0) (removed from both LL_{NB} and LL_{ω})	-404.451	-662.348	-1066.889	-358.174	3	716.348	5.99E-155
5*	Testing for cost of travel to attractiveness. $\left(c_j = 0\right)$	-399.796	-311.094	-710.890	-2.174	1	4.348	3.71E-02

6*	Testing for linear relationship between population size and attractiveness	-402.638	-310.752	-713.389	-4.674	1	9.348	2.23E-03
7*	Testing for population size to attractiveness	-404.564	-310.650	-715.214	-6.499	2	12.998	1.51E-03
8	Testing for linear relationship of population distance to lake	-399.254	-310.586	-709.840	-1.125	1	2.250	1.34E-01
9	Testing for population distance to attractiveness	-397.796	-310.930	-708.726	-0.011	2	0.022	9.89E-01
10*	Testing for inverse linear relationship of source to destination distance	-440.131	-296.855	-736.986	-28.271	1	56.542	5.50E-14

* Significant at

Table 3. Summary table of parameter values with 95% confidence limits for the invasion scenarios with 11 currently invaded lakes as sources and Lake Ontario as sole source. Confidence limits for parameters in the establishment model excluding and including stochastic variability from the dispersal model are enclosed by square and curly brackets, respectively.

Model	Sub models	Parameter	11 source lakes	Lake Ontario
			Parameter value and confidence limits	Parameter value and confidence limits
Stochastic gravity model		☐	0.92 [0.664, 1.167]	1.39 [0.681, 2.858]
	Mean number of trips, ☐	☐	-0.58 [-0.845, -0.308]	-0.49 [-0.578, -0.393]
	Lake attractivity, ☐	☐	0.26 [0.086, 0.431]	-2.11 [-4.409, 0.422]
	Zero-inflation, ☐	☐	9.60 [8.605, 10.699]	8.36 [6.029, 11.260]
		☐	0.01 [0.010, 0.014]	8.83E-3 [3.860E-3, 1.424E-2]
		☐	-1.37 [-1.533, -1.230]	-1.25 [-1.681, -0.906]
Establishment model		☐	-11.89 [-18.268, -7.861]	-13.28 [-20.057, -8.237]

	{-13.337, -9.443}	{-13.603, -9.544}
□	0.20 [0.093, 0.428] {0.104, 0.678}	0.81 [0.484, 3.284] {0.097, 0.755}
□	0.84 [0.253, 1.532] {0.170, 1.008}	1.17 [0.301, 1.786] {-0.117, 1.021}
□	1.82E-4 [2.03E-5, 3.16E-4]	1.90E-4 [3.458E-5, 3.337E-4]
	{1.221E-4, 2.087E-4}	{1.312E-4, 2.124E-4}
□	0.82 [0.382, 1.416] {0.615, 1.299}	0.67 [0.455, 1.540] {0.617, 1.860}
□	-2.36E-2 [-5.82E-2, -6.16E-3]	-1.58E-2 [-6.78E-2, -8.93E-3]
	{-0.087, -0.014}	{-0.130, -0.014}

Table 4. Predicted (baseline and 95% prediction intervals) and observed invasion status for 871 destination lakes under scenarios with 11 lakes or only Lake Ontario as a source. Optimal probability thresholds above which lakes are classified as invasible are p = 0.010 for the 11 lakes scenario and p = 0.013 for the Lake Ontario scenario calculated from the baseline establishment models. Invasion status is abbreviated as 1=Present and 0=Absent.

Prediction scenario	Pred. 0, Obs. 0	Pred. 0, Obs. 1	Pred. 1, Obs. 0	Pred. 1 Obs. 1	Hit rate	False alarm rate	Overall percent correct
11 source lakes	834	0	26	11	1.00	0.03	0.97
11 source lakes, lower 95% PI	860	11	0	0	0.00	0.00	0.99
11 source lakes, upper 95% PI	0	0	860	11	1.00	1.00	0.01
11 source lakes, stochastic variability, lower 95% PI	860	11	0	0	0.00	0.00	0.99
11 source lakes, stochastic variability, upper 95% PI	182	0	678	11	1.00	0.78	0.22
Lake Ontario	834	1	26	9	0.9	0.03	0.97

Lake Ontario, lower 95% PI	860	10	0	0	0.00	0.00	0.99
Lake Ontario, upper 95% PI	1	0	859	10	1.00	1.00	0.01
Lake Ontario, stochastic variability, lower 95% PI	859	10	1	0	0.00	0.001	0.99
Lake Ontario, stochastic variability, upper 95% PI1	532	0	328	10	1.00	0.38	0.62

Figure Captions

Figure 1. Flowchart of model inputs for the stochastic dispersal model and Monte Carlo error propagation to the establishment models. Circles indicate model inputs, rectangles indicate intermediate or sub-models, and parallelograms represent model outputs or parameters estimated during the model fitting process. Gray boxes indicate the main dispersal and logistic establishment models.

Figure 2. Standardized logistic regression coefficients with bias-corrected 95% CL for establishment models with 11 lakes and Lake Ontario as sources. Plot markers indicate logistic regression coefficients for the baseline model and confidence limits incorporate stochastic variability from the dispersal model. Distance from horizontal dotted line indicates relative change in standard deviations in the probability of establishment with 1 standard deviation change in predictor variable.

Figure 3. Empirical Receiver Operating Characteristic curves for means and upper 95% PI for establishment probability excluding and including stochastic variation from the dispersal model. Establishment probability was estimated based on model validation sub-samples. Panels (a) and (b) represent the scenarios of 11 lakes and Lake Ontario as sources, respectively.

Figure 4. Estimated probabilities of *Cercopagis* establishment based on logistic models of propagule pressure and water chemistry for scenarios of 11 source lakes and Lake Ontario as single source only. Panels (a-c) show the probability of establishment with upper 95% prediction intervals excluding and including stochastic dispersal for the scenario of propagule flow leaving 11 sources, and panels (d-f)

show estimated probability of establishment with prediction intervals with Lake Ontario as the single source.

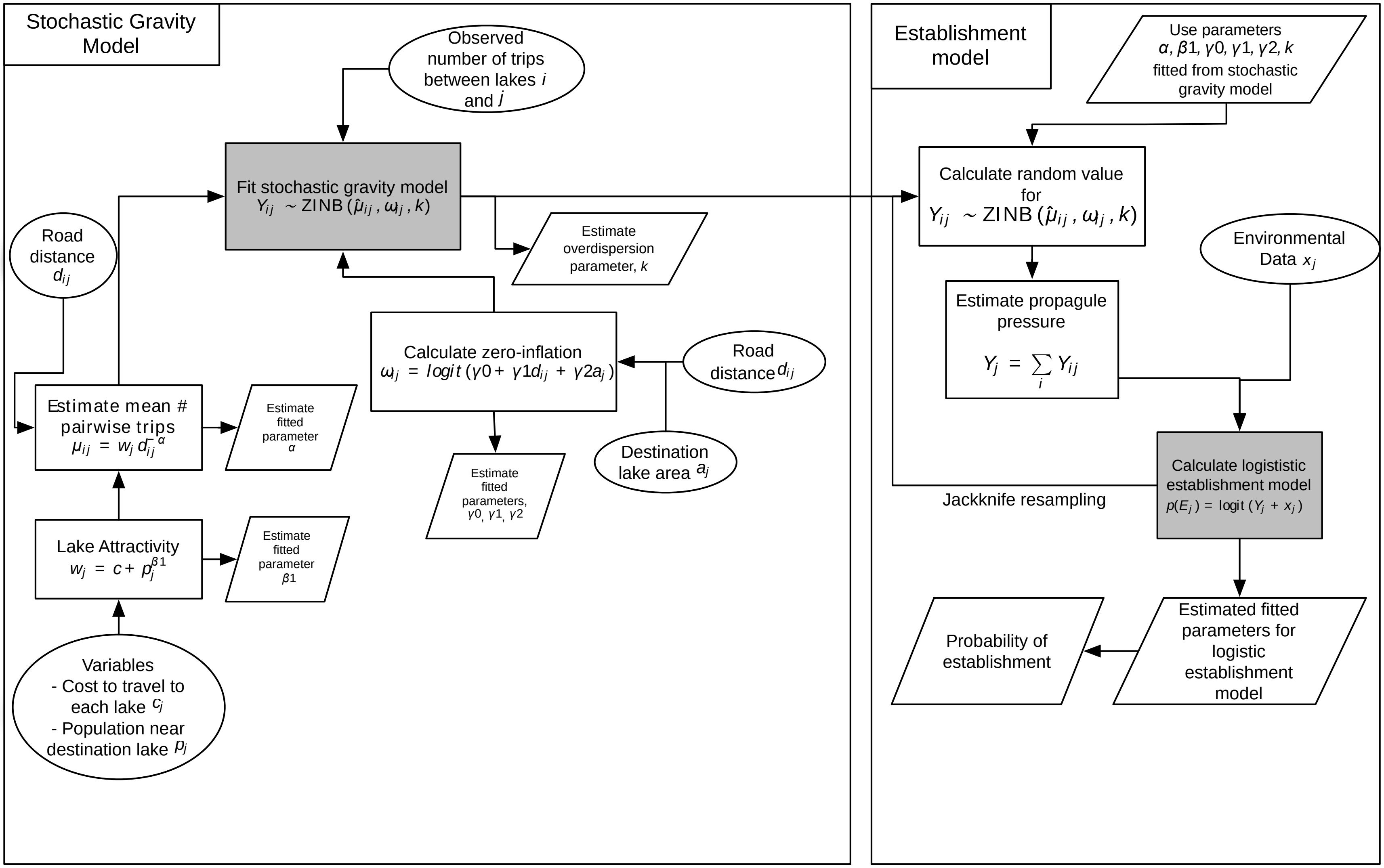
Stochastic Gravity Model
Observed number of trips between lakes i and j
Fit stochastic gravity model $Y_{ij} \sim \text{ZINB}(\hat{\mu}_{ij}, \omega_j, k)$
Road distance d_{ij}
Estimate overdispersion parameter, k
Calculate zero-inflation $\omega_j = logit(\gamma 0 + \gamma 1 d_{ij} + \gamma 2 a_j)$
Road distance d_{ij}
Destination lake area a_j
Estimate fitted parameters, $\gamma 0, \gamma 1, \gamma 2$
Estimate mean # pairwise trips $\mu_{ij} = w_j d_{ij}^{-\alpha}$
Estimate fitted parameter α
Lake Attractivity $w_j = c + p_j^{\beta 1}$
Estimate fitted parameter $\beta 1$
Variables
- Cost to travel to each lake c_j
- Population near destination lake p_j
Establishment model
Use parameters $\alpha, \beta 1, \gamma 0, \gamma 1, \gamma 2, k$ fitted from stochastic gravity model
Calculate random value for $Y_{ij} \sim \text{ZINB}(\hat{\mu}_{ij}, \omega_j, k)$
Estimate propagule pressure $Y_j = \sum_i Y_{ij}$
Environmental Data x_j
Calculate logistic establishment model $p(E_j) = \text{logit}(Y_j + x_j)$
Jackknife resampling
Estimated fitted parameters for logistic establishment model
Probability of establishment

Figure 1.

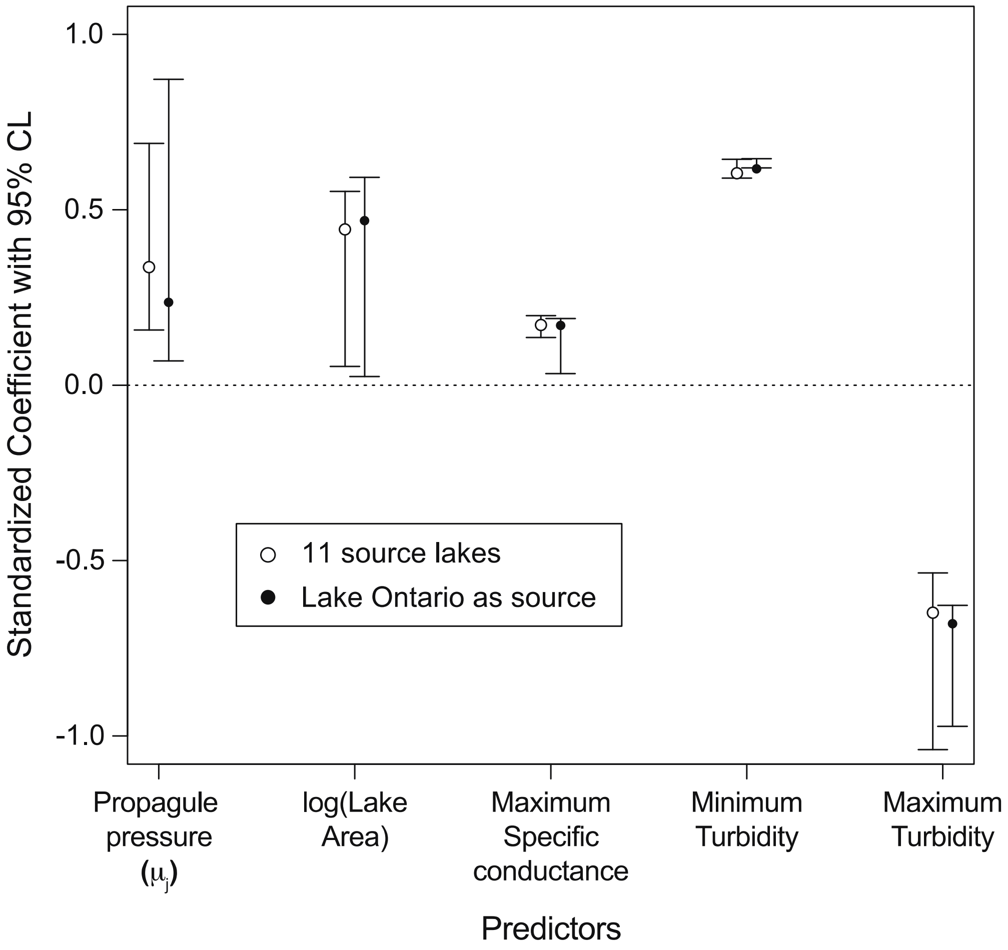
1.0
0.5
0.0
-0.5
-1.0
Standardized Coefficient with 95% CL
○ 11 source lakes
● Lake Ontario as source
Propagule pressure (μ_j)
log(Lake Area)
Maximum Specific conductance
Minimum Turbidity
Maximum Turbidity
Predictors

Figure 2.

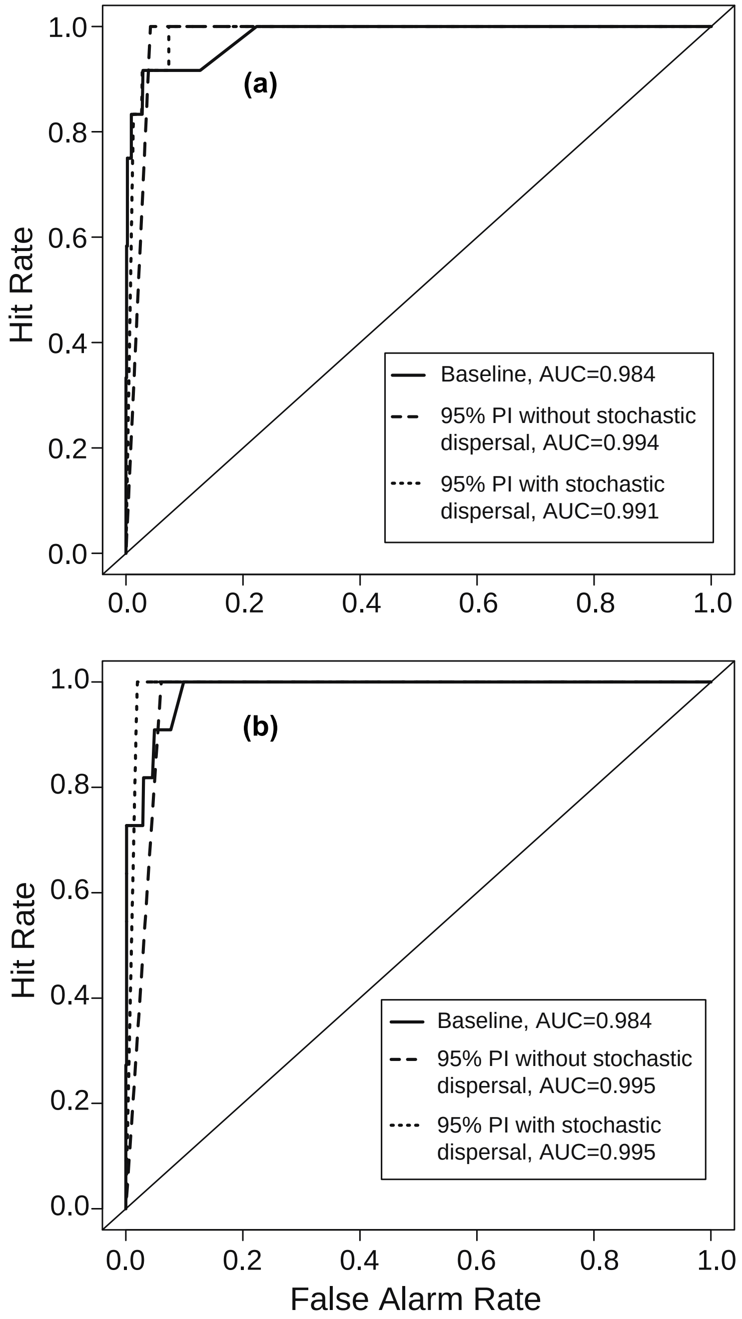
(a)
Hit Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Baseline, AUC=0.984
95% PI without stochastic dispersal, AUC=0.994
95% PI with stochastic dispersal, AUC=0.991
(b)
Hit Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Alarm Rate
Baseline, AUC=0.984
95% PI without stochastic dispersal, AUC=0.995
95% PI with stochastic dispersal, AUC=0.995

Figure 3.

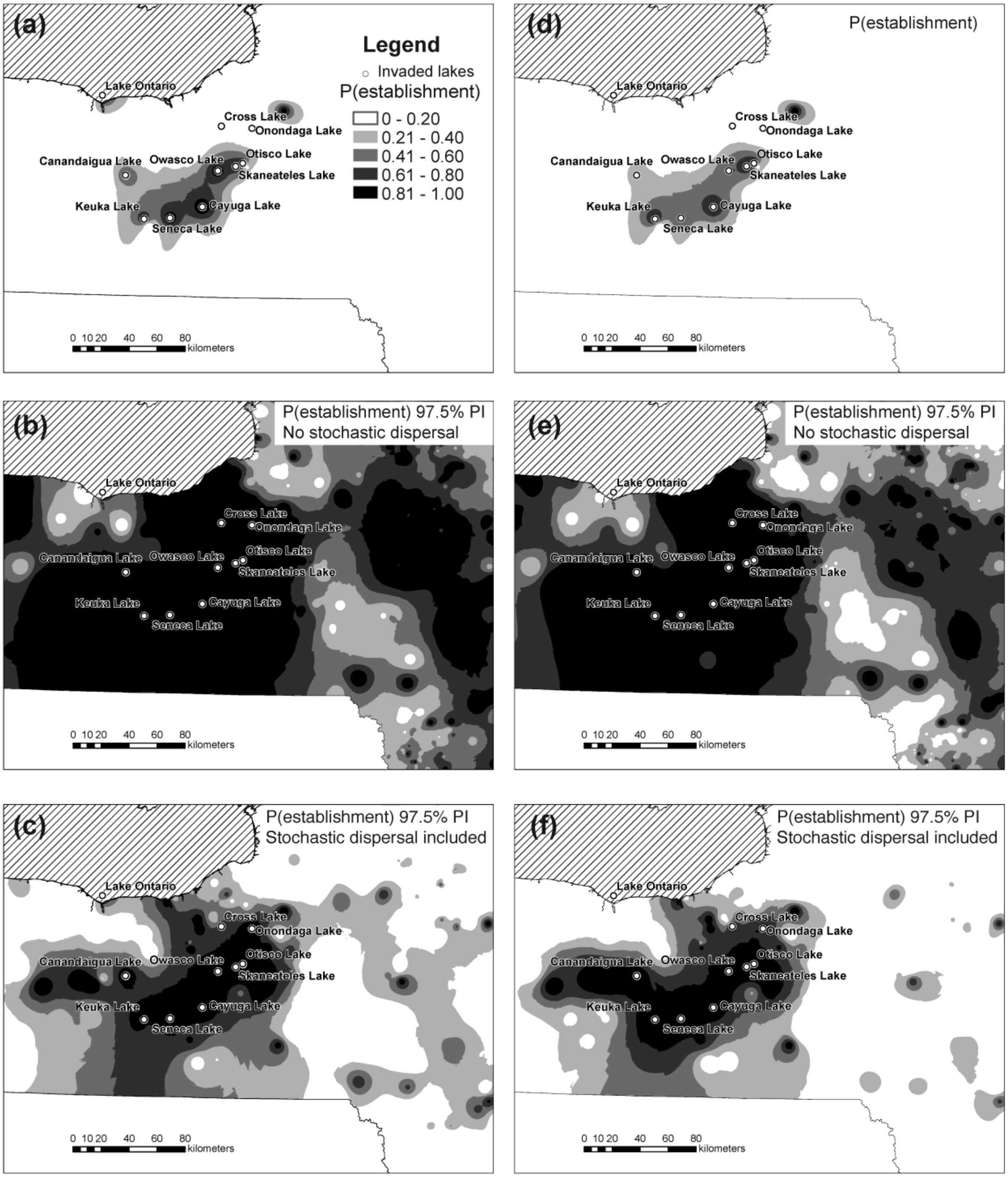
(a)
Legend
Invaded lakes
P(establishment)
0 - 0.20
0.21 - 0.40
0.41 - 0.60
0.61 - 0.80
0.81 - 1.00
(b)
P(establishment) 97.5% PI
No stochastic dispersal
(c)
P(establishment) 97.5% PI
Stochastic dispersal included
(d)
P(establishment)
(e)
P(establishment) 97.5% PI
No stochastic dispersal
(f)
P(establishment) 97.5% PI
Stochastic dispersal included
Lake Ontario
Cross Lake
Onondaga Lake
Otisco Lake
Canandaigua Lake
Owasco Lake
Skaneateles Lake
Keuka Lake
Cayuga Lake
Seneca Lake
0 10 20 40 60 80
kilometers

Figure 4.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Probability mass function and Expectation-Maximum Algorithm for fitting a Zero-inflated Negative Binomial Distribution

The probability mass function for the number of pairwise trips, assuming a ZINB is characterized by:

(A1)

where , *k* and represent the mean, dispersion and zero-inflation components of the distribution. Since difficulty frequently arises in finding maximum likelihood estimates for the three parameters using conventional optimization methods, the E-M (Expectation-Maximum) approach is recommended (Jansakul, 2005). In this approach, the E-steps and M-steps of the algorithm alternate until convergence is reached.

To begin, let Z_{ij} be a binary indicator of whether the trip had occurred with probability between lake pairs *i* and *j* with as the zero, or perfect state; and 0 otherwise (i.e.). Thus, conditional on this indicator variable, Z_{ij}, the number of trips, Y_{ij} follows either or (Jansakul, 2005). The log-likelihood function for parameter estimates ω_{ij}, μ_{ij}, *k* may then be separated into a weighted binomial likelihood for and a weighted negative binomial likelihood:

, (A2)

where

$$LL_{\omega} = \sum_{ij=1}^{n}\left[z_{ij}log\left(\frac{\omega_{ij}}{1-\omega_{ij}}\right)+log\left(1-\omega_{ij}\right)\right]$$

and

.

E-step for Z_{ij}

The expected value for may be estimated following Bayes' Theorem by its posterior mean given the observed number of trips y_{ij} and current estimates of , , and *k*. Here, parameter and the vector of parameters are used in estimating the mean number of pairwise trips and lake attractivity respectively (Eqns. 1 and 2), and the vector of parameters used in estimating the level of zero-inflation (Eqn. 3).

(Modified from Jansakul, 2005). (A3)

During the model fitting process, the initial estimates for μ_{ij}, *k* were set at $\mu_{ij} = 0.5$, and $k = 1$. Useful initial estimates for ω_{ij} and γ were based on the unweighted logistic regression,

. (A4)

In this model, covariates included road distance from lake *i to j*, d_{ij}, (in km) and log-transformed lake area of the destination lake, a_j (in ha).

M-step for μ_{ij}, *k*

Parameters α, β_0 through β_6, and the dispersion parameter for the negative binomial distribution, *k*, required for the estimated number of trips m_{ij} were fitted using MLE with loss function LL_{NB} (Eqn. A2) and weighted by from the previous E-step estimation for z_{ij}.

M-step for γ

Parameter values for g are updated by minimizing the loss function for ω_{ij} as a function of g and current estimates of z_{ij} from the previous E-step. This step is identical to a weighted logistic regression with weights, $h_{ij} = z_{ij}$ when the number of trips, $y_{ij} = 0$, and $h_{ij} = (1 - z_{ij})$ when $y_{ij} > 0$ (Lambert, 1992).