

Predicting worldwide invasiveness for four major problematic decapods: an evaluation of using different calibration sets

César Capinha, Brian Leung and Pedro Anastácio

C. Capinha (capinha@uevora.pt) and P. Anastácio, IMAR, Centro de Mar e Ambiente do Depto de Paisagem, Ambiente e Ordenamento, Univ. de Évora, Rua Romão Ramalho, no. 59, PT-7000-671 Évora, Portugal. – B. Leung, Dept of Biology, McGill Univ., Montreal, QC H3A 1B1, Canada.

Recently, there has been much debate whether niche based models (NBM) can predict biological invasions into new areas. These studies have chiefly focused on the type of occurrence data to use for model calibration. Additionally, pseudo-absences are also known to cause uncertainty in NBM, but are rarely tested for predicting invasiveness. Here we test the implications of using different calibration sets for building worldwide invasiveness models for four major problematic decapods: *Cherax destructor*, *Eriocheir sinensis*, *Pacifastacus leniusculus* and *Procambarus clarkii*. Using Artificial Neural Networks models we compared predictions containing either native range occurrences (NRO), native and invasive occurrences (NIO) and invasive only (IRO) coupled with three types of pseudo-absences – based on sampling only 1) the native range (NRA), 2) native and invasive ranges (NIA), and 3) worldwide random (WRA). We further analysed the potential gains in accuracy obtained through averaging across multiple models. Our results showed that NRO and IRO provided the best predictions for native and invaded ranges, respectively. Still, NIO provided the best balance in predicting both ranges. Pseudo-absences had a large influence on the predictive performance of the models, and were more important for predictiveness than types of occurrences. Specifically, WRA performed the best and NRA and NIA performed poorly. We also found little benefit in combining predictions since best performing single-models showed consistently higher accuracies. We conclude that NBM can provide useful information in forecasting invasiveness but are largely dependent on the type of initial information used and more efforts should be placed on recognizing its implications. Our results also show extensive areas which are highly suitable for the studied species worldwide. In total these areas reach from three to nine times the species current ranges and large portions of them are contiguous with currently invasive populations.

Accurate information concerning the risk of a species becoming established outside its native range can provide a solid foundation for justifying preventive measures and has been a subject of increasing focus. As such, invasion biologists have sought predictive methods to forecast invasions (Côté and Reynolds 2002; for a recent review see Hayes and Barry 2008). Specifically, niche based models (NBM), which estimate the degree of environmental compatibility for the species in new areas, have become increasingly popular in the last few years. For example, Thuiller et al. (2005) found a close similarity between worldwide NBM invasiveness predictions and South African plant invaders distribution. Additionally, others have used NBM predictions coupled with propagule pressure estimates in order to provide final predictions of risk of establishment of a species (Leung and Mandrak 2007).

Recently, there has been much debate concerning the use of NBM for invasive species predictions (Mau-Crimmins et al. 2006, Loo et al. 2007, Broennimann and Guisan 2008, Pearman et al. 2008, Steiner et al. 2008, Beaumont et al. 2009). This has been chiefly focused on the type of

occurrence data to use for model calibration. Some studies using native-based models were able to provide accurate invasiveness predictions (Welk et al. 2002, Thuiller et al. 2005), but others found reduced predictability of the entire invaded ranges (Mau-Crimmins et al. 2006, Fitzpatrick et al. 2007, Broennimann and Guisan 2008, Beaumont et al. 2009). Most recent studies argue that the use of occurrence data from the invaded ranges improves predictions (Mau-Crimmins et al. 2006, Loo et al. 2007, Broennimann and Guisan 2008, Beaumont et al. 2009), since invaders may not conserve their niches across space (Broennimann et al. 2007, Pearman et al. 2008). Still, invasive occurrences would be constructed only after the invasion had already taken place and could have issues with data availability. Moreover, if the invasion process is not complete the equilibrium assumption underlying NBM may be violated and cause underestimation of the full invasiveness potential (Wilson et al. 2007). Further, native range data for invasive species is also often unavailable or difficult to collect (Mau-Crimmins et al. 2006). Thus, while coupling native and invasive occurrences may allow

the best overall characterization of species niches, there could be issues with data availability. As such, a direct comparison of the quantitative differences and marginal gains of using different occurrence data would be useful.

The inclusion of absence data is arguably also an issue of NBM for invaders. These data are usually hard to obtain from common sources of species distribution data such as museums or biodiversity databases (Chefaoui and Lobo 2008). To deal with these difficulties, the use of pseudo-absences has been common since it allows the use of “group-discrimination techniques” considered to provide more accurate predictions than presence only “profile techniques” (Hirzel et al. 2001, Brotons et al. 2004, Segurado and Araújo 2004). Still, the use of pseudo-absences is a known cause of uncertainty in NBM (Lobo 2008, Phillips 2008) and the best way to obtain them is still far from consensus (Chefaoui and Lobo 2008, VanDerWal et al. 2009). Moreover, extraction techniques have mostly been tested with species for which equilibrium with the environment is assumed (i.e. species within their native ranges) (for a review see Pearce and Boyce 2006). However, invasion biology is interested in estimating suitability in new areas and, thus, it is unclear how best to obtain pseudo-absences. Options include: 1) use sites located within the native range. The assumption is that there has likely been sufficient time for propagules to reach these sites. 2) Use sites within native and invasive ranges simultaneously since new ranges may provide additional information. However, this would not necessarily take into account the range of environments possible. 3) Use random points across the world. The consequences of these different forms of pseudo-absences for predictive ability of invasions have not been examined and are currently unknown.

For dealing with uncertainty in the predictions, researchers have been increasingly adopting the use of consensus methods – “ensembles” of single-model NBMs, with different architectures or different assumptions (i.e. averaging across the results of single models). Model ensembles have been applied for predicting distributions of threatened species (Marmion et al. 2009), impacts of climate change on species distributions (Araújo et al. 2005) or potential distributions of invasive species (Stohlgren et al. 2010). Ensembles have been mostly used to deal with the uncertainty caused by the use of distinct correlative models. Despite their potential to reduce uncertainties coming from the use of different calibration data, to our knowledge, ensembles have not been applied for dealing with this source of uncertainty in invasiveness predictions.

In this study, we build habitat suitability models and test the consequences of occurrence type and pseudo-absence type. Specifically we examine the use of occurrence data from their native range (NRO), invasive range (IRO) and both native and invasive occurrences (NIO) in the algorithm’s calibration sets. Simultaneously we also evaluate the effect of three different pseudo-absence methods based on sampling: 1) the native range (NRA), 2) native and invasive range (NIA), and 3) worldwide random (WRA). Finally, we also explore the use of consensus methods as a possibility for dealing with the uncertainty coming from the use of different calibration information.

We focus on four important invasive species: *Cherax destructor*, *Eriocheir sinensis*, *Procambarus clarkii* and

Pacifastacus leniusculus. These are wide-ranging invasive decapods, for which either populations or individuals are being systematically found in new areas. Their impacts in the invaded ecosystems are numerous (e.g. predation and competition with native species, habitat alteration and agricultural damage) and the major mechanisms of introduction have been identified. As such, it will be most effective for managers to target habitat suitability models to these species. Still, the uncertainty coming from the use of inadequate calibration information can undermine this objective.

Methods

Invaders and distribution data

Cherax destructor (yabby) is a crayfish indigenous to eastern Australia that currently invades several areas in Western Australia and Iberian Peninsula. As for other invaders, occurrence records for *C. destructor* are scarce and most have a low spatial accuracy (Souty-Grosset et al. 2006). For reducing the uncertainty of their spatial location of collected data we opted for using a cell resolution of 50 km. We used only one record per grid cell and gathered a total of 154 occurrence records for this species, 103 referring to its native range and 51 from invaded areas. This information was mostly collected from the Museum Victoria collections and several published works.

Eriocheir sinensis (Chinese mitten crab) is an invasive crab included in the 100 “World’s Worst” invaders by the World Conservation Union (Lowe et al. 2000). The native range of this catadromous crab encompasses eastern China, Japan and eastern Russia, being presently invasive in several coastal areas of North America and with particular expression in Europe (Gollasch 2006). For this species we obtained a total of 295 occurrence records, 101 from its native range and 194 from invaded areas. This information was collected using the global biodiversity information facility (GBIF) (<www.gbif.org/>) and a vast number of published works referring to this species.

Procambarus clarkii (red swamp crayfish) is a commercially harvested crayfish, native from northeast Mexico to south-central USA. This species currently has invasive populations across 5 continents (Africa, Asia, Europe, North America and South America) and was recently quoted as one of the 100 “Most invasive alien species in Europe” (DAISIE 2008). For this crayfish we collected a total of 598 occurrence records, 173 from its native range and 425 from invaded areas. Its distributional data were collected from the Smithsonian Inst. National Museum of Natural History, the Illinois Natural History Survey, the Atlas of crayfish in Europe (Souty-Grosset et al. 2006), GBIF and several published works.

Pacifastacus leniusculus (signal crayfish) is native from the north-western USA and south-western Canada and currently invades large portions of the European continent, south-western USA and some Japanese regions (Souty-Grosset et al. 2006). A total of 565 occurrence records were collected for this species, 125 from its native range and 440 from invaded areas. The data sources used were the same as for *P. clarkii*.

Environmental factors

To summarize the world environmental characteristics (Antarctica excluded) we considered 10 spatial coverages. All environmental predictors used were not collinear (Pearson's $|r| < 0.8$). Eight climatic variables concerning the period 1961–2000 were included: near surface annual mean temperature (amtemp); near surface mean maximum temperature of the warmest month (maxtwm); near surface mean minimum temperature of the coldest month (mintcm); near-surface mean diurnal temperature range (trange); mean number of frost days (frost); mean total annual precipitation (anpre); mean total precipitation of the wettest month (prewm) and mean total precipitation of the driest month (predr). These predictors were built using information from the CRU TS2.1 climate dataset (Mitchell and Jones 2005). Two physiographic variables were also included: altitude (alt) and in-stream distance to ocean (disto). Altitude was included since it can act as a surrogate of several environmental factors important for our species such as stream velocity and size – usually faster and smaller at higher elevations. In-stream distance to ocean was employed only for *E. sinensis* due to its catadromous nature. The digital elevation model was acquired on the United States Geological Survey (USGS) HYDRO1k geographical dataset (Verdin and Jenson 1996), from which in-stream distance to ocean was calculated using ILWIS 3.5 Open (<http://52north.org/>). All variables were resampled to a 50 km cell resolution using a bicubic method and projected to a Mollweide equal area world projection.

Pseudo-absences extraction

Pseudo-absences can be seen as a sample of the available conditions (Phillips et al. 2009) or as indicator of unsuitable conditions (Chefaoui and Lobo 2008). For pseudo-absences used in this study, we excluded all cells having occurrence of the species in order to potentially maximize the representativeness of unsuitable conditions. Following this principle, one approach for generating pseudo-absences is a simple spatially-random generation of records across the world, except from where presences are known (WRA).

Our second approach was entirely based on pseudo-absences in the native range distribution of the species (NRA). Given our relatively intense search for occurrence data and the fact that all four species have well documented native distributions, we assumed the areas without the species presence within native range boundaries were reliable representatives of unsuitable conditions. For NRA, we limited our sampling area to the inner boundary of the convex-hull defined by the occurrence records. Our third approach consisted of sampling both native and invasive ranges (NIA). By this we assume that the unoccupied areas within the invasive range can provide additional information regarding unsuitable conditions. Sampling was made within the inner boundary of the convex-hull defined by the occurrence records of each range. To avoid wide and unrealistic sampling areas distinct invasive populations of each species were delimited by independent convex-hulls.

Dataset assembly

Before calibration datasets were built we retained 20% of each species' occurrence to validate our predictions (i.e. they were not used to build the model). The remaining occurrence records were used to build nine different types of calibration datasets: 1) native range occurrences versus worldwide random pseudo-absences; 2) native range occurrences versus native range pseudo-absences; 3) native range occurrences versus native and invasive ranges pseudo-absences; 4) native and invasive ranges occurrences versus worldwide random pseudo-absences; 5) native and invasive ranges occurrences versus native range pseudo-absences; 6) native and invasive ranges occurrences versus native and invasive ranges pseudo-absences; 7) invasive range occurrences versus worldwide random pseudo-absences; 8) invasive range occurrences versus native range pseudo-absences and 9) invasive range occurrences versus native and invasive ranges pseudo-absences. For increasing the representation of the environment captured by the pseudo-absences we created 20 calibration datasets for each combination. Each of these had an independently drawn set of pseudo-absences. To avoid biasing predictions towards a more prevalent response each calibration dataset had a number of pseudo-absences equal to the number of occurrences (Supplementary material Table S1).

Model selection and predictions

NBM has been built using many distinct correlative models with several new approaches receiving great promise (Elith et al. 2006). For this study we have chosen to use Artificial Neural Networks (ANN) for predicting the probability of environmental suitability in each cell. ANN is a method used regularly in NBM and has also been recognized as one of the best performing techniques (Segurado and Araújo 2004). Moreover ANN are particularly appropriate when the relations between variables are not well known, which is often the case with ecological data (Lek and Guégan 1999). We used feedforward multilayer perceptron with back-propagation ANN models (MLP-ANN). MLP-ANN is one of the most common types of supervised ANN's used in ecology, being normally structured in one input layer representing the predictors (environmental variables), one or more hidden layers, each with a variable number of nodes, and one output layer representing the dependent variable (presence/absence).

Due to its large flexibility, ANN models are prone to overfitting, making the model less generalizable and decreasing their predictive power. To avoid overfitting in our models we used both a cross-validation procedure during the training episodes and tested different network configurations in order to optimize their degree of complexity and number of training cycles (Özesmia et al. 2006). While higher complexity increases the risk of overfitting, oversimplification can also result in poor fits. In the same way, excessive training of the network is prone to overfit the data while the inverse may result in failure to capture its regularities.

Before building final predictions we tested for the more appropriate network configuration for each of the dataset

types. To do this we compared the performances of single hidden layer MLP networks using three different levels of complexity. According to Burnham and Anderson (2002), the available data sample should be at least ten times larger than the number of parameters in a model. We adopted this principle for establishing the maximum complexity allowed in each of the tested models. Medium complexity networks were also considered, each containing half the hidden nodes of the previous models. Finally, for the least complex models we tested the performance of MLP-ANN containing no hidden nodes, which are equivalent to Generalized Linear Models.

Models were built using Weka 3.6 (Witten and Frank 2005). To comply with the binary response of the dependent variable, hidden nodes were automatically set to sigmoid functions. All training sessions included a weight decay function of the learning rate by dividing the starting value by the cycle number, forcing a low learning rate and by so reducing the risk of data overfit. A stopping rule was also included in order to avoid overtraining. Models were allowed to train for a total of 4000 cycles as long as the predictions did not exceed more than 500 consecutive cycles without performance improvement.

To analyse the predictive power of each of the three network configurations we used a 10 fold cross-validation procedure. That is, all models were calibrated using 90% of cases for model calibration while the remaining 10% were left-out for comparison with predicted values. This procedure was then repeated 10 times until the entire dataset had been compared against the predictions. These comparisons were evaluated using the mean values of the root mean squared error (RMSE) automatically supplied by Weka. The network configurations achieving lower mean RMSE for each dataset type were then selected and applied for predicting along the entire range of worldwide environmental conditions. Due to the use of 20 independent sets of pseudo-absences, the final prediction for each dataset type corresponded to the mean value obtained by these 20 calibration datasets.

Ensemble predictions

For dealing with the variability of single predictions, the combination of ensemble models has been adopted in studies with invasive species (Stohlgren et al. 2010). While the majority of efforts have been focused on the variability caused by the use of distinct modeling methods, this logic could apply to reduce the uncertainty coming from the use of different sets of calibration data. Here we explored the possibility of improving invasiveness predictions using ensemble models. We examined three types of ensembles – predictions based on a weighted average of all single models (WA(all)), averaged within each occurrence type (WA(NRO); WA(NIO); WA(IRO)), and averaged within each pseudo-absence type (WA(NRA), WA(NIA), WA(WRA)). For each, all ensembles were obtained through averaging single-models by their relative accuracy value (Marmion et al. 2009). In order to attain a fair comparison against the single-models predictions, relative accuracy value was based on the RMSE obtained from the 10-fold cross-validation process used in the network configuration

selection process as supplied by Weka (Supplementary material Table S2). The weighted averages of the single-models were performed as given by eq. 1

$$WA_i = \frac{\sum((1 - RMSE_{pj_i}) - pj_i)}{\sum j(1 - RMSE_{pj_i})} \quad (1)$$

where pj_i was the probability of environmental suitability for the i th decapod species in each of the j -selected single-models.

Models validation

After predictions were made for both single and combined-models their evaluation was performed. We used the 20% of each type of occurrence records (i.e. native and invasive) initially excluded from the calibration datasets. These were complemented with an equal number of worldwide random sample of areas without native or invaded occurrences. For increasing the representativeness of these areas in the evaluation datasets we made 10 datasets for each type of occurrence records. Each of these had an independently drawn sample of areas without native or invaded occurrences (see Supplementary material Table S3 for datasets description). Validation records were compared with the predicted values using the area-under-the-curve of the receiver-operating characteristic (ROC-AUC) (Hanley and McNeil 1982) and Cohen's Kappa (k) (Cohen 1960). Kappa was calculated across a range of thresholds along the 0 to 1 interval using a 0.05 amplitude increment and its maximum value selected (Elith et al. 2006). Both native and invasive ranges were evaluated. Final evaluation values were obtained by averaging the scores of the 10 replicate evaluation datasets. For assessing variability in predictive performance we also calculated the standard deviation of the obtained evaluation scores. For qualitatively describing the predictions values of k , we established the following classes: $k < 0.2$ poor; $0.2 < k < 0.4$ fair; $0.4 < k < 0.6$ moderate; $0.6 < k < 0.8$ good and $k > 0.80$ as very good (modified from Landis and Koch 1977). For ROC-AUC we considered ROC-AUC < 0.8 as poor accuracy; $0.8 < \text{ROC-AUC} < 0.9$ moderate; $0.9 < \text{ROC-AUC} < 0.95$ good and $\text{ROC-AUC} > 0.95$ as very good (adapted from Thuiller et al. 2005).

Species environmental space

We also examined environmental similarities between native and invaded ranges. If environmental conditions differ, this may indicate that a niche shift has occurred. Following Mau-Crimmins et al. (2006), we used a Principal Components Analysis (PCA) to simplify the species niche dimensionality and compared the position occupied by each occurrence record. By distinguishing native and invasive occurrences this procedure allowed us to verify the degree of environmental overlap between the two ranges occupied by each species. This was made through a score plot of the two Principal Components. Further we used the same method to compare the position of the species occurrences with the overall best performing pseudo-absences extraction method.

For improving visual interpretation we only included 300 randomly selected pseudo-absence records for plotting.

Results

Predictive performances

For single-models, native range occurrences and invasive range occurrences provided the best predictions for native and invaded ranges, respectively (Table 1 and 2). Still, combined native and invasive occurrences allowed a good balance between the two, with best models attaining good to very good accuracy values in both ranges (Table 1 and 2). Interestingly, native range occurrences allowed extrapolation to the invaded ranges of three of the four species examined with moderate to good accuracy values (*C. destructor* NRO best model: $k=0.63$ and ROC-AUC = 0.89; *E. sinensis* NRO best model: $k=0.9$ and ROC-AUC = 0.95; *P. leniusculus* NRO best model $k=0.89$ and ROC-AUC = 0.93, Table 2). In contrast, best predictions using native range occurrences for the invasiveness potential of *P. clarkii* were worse ($k=0.57$; ROC-AUC = 0.71). Models using invasive range occurrences were relatively modest in predicting the native ranges for all species (Table 1). Variability in accuracy was similar for the three types of occurrences in predicting native (mean SD of NRO ROC-AUC: 0.029; mean SD of NIO ROC-AUC: 0.033; mean SD of IRO ROC-AUC: 0.031) or invasive distributions (mean SD of NRO ROC-AUC: 0.033; mean SD of NIO ROC-AUC: 0.030; mean SD of IRO ROC-AUC: 0.029) (Supplementary material Table S4 and S5).

Type of pseudo-absence had a pronounced effect on predictive performance. Specifically, models using native range pseudo-absences and native and invasive ranges pseudo-absences had the lowest predictive performances (Table 1 and 2). Best performances were achieved unanimously with worldwide random pseudo-absences. Worldwide random pseudo-absences also provided the lowest mean variability of accuracy in predicting both native (mean SD of ROC-AUC: 0.019) and invasive distributions

(mean SD of ROC-AUC: 0.024) (Supplementary material Table S4 and S5). We found that the weighted average procedure provided poor to moderate predictive accuracy for both native or invaded ranges for WA(all) and WA(NIO) (Table 3). The WA(NRO) method provided moderate accuracy for predicting native distributions but was clearly unable to capture the species invasive ranges. Inversely, the WA(IRO) method provided good predictions concerning the species invasive distributions except for *C. destructor* ($k=0.64$ and ROC-AUC = 0.74), but was unable to predict native ranges. The WA(NRA) and WA(NIA) methods provided the lowest accuracies for predicting the species native ranges but provided moderate to good accuracy in predicting the invasive range of three species (*E. sinensis*, *P. leniusculus* and *P. clarkii*). Finally, the WA(WRA) model provided the best accuracies in both ranges. However, in general we found little benefit of combined models when compared to the best performing single models, specifically native and invasive occurrences coupled with worldwide random pseudo-absences.

Species environmental space

The PCA analysis allowed comparison of environmental space of presences versus pseudo-absences (Fig. 1), and between regions occupied in the native and exotic ranges (Fig. 2). Not surprisingly, the comparison between occupied ranges and best performing pseudo-absence extraction method (worldwide random) illustrated a much larger range of environmental conditions for pseudo-absences compared to occurrence records (Fig. 2). Interestingly, however, focusing on the occurrence data demonstrated that all species were occupying different environmental space in the naturalized areas compared to the exotic ranges (Fig. 2). For *C. destructor* the difference occurred along component 1, which was primarily a temperature composed gradient. The multivariate space occupied by *E. sinensis* in the invaded range was mainly differentiated along component 2, primarily associated with both near surface mean minimum temperature of the coldest month and mean total precipitation of the driest month. For *P. clarkii*, despite the existence

Table 1. Predicting distribution in native range: single-models validation results of kappa statistic (k) and area under the curve of receiver-operating characteristic (ROC-AUC) for native ranges using native range occurrences (NRO), native and invasive occurrences (NIO) and invaded range occurrences (IRO), coupled with within native range pseudo-absences random extraction (NRA), native and invasive ranges pseudo-absences random extraction (NIA) and the common spatially worldwide random pseudo-absences (WRA).

Species	Pseudo-absences	NRO		NIO		IRO	
		k	ROC-AUC	k	ROC-AUC	k	ROC-AUC
<i>C. destructor</i>	NRA	0.51	0.68	0.28	0.49	0.21	0.38
	NIA	0.51	0.67	0.52	0.63	0.17	0.29
	WRA	0.91	0.99	0.85	0.98	0.64	0.90
<i>E. sinensis</i>	NRA	0.81	0.89	0.52	0.63	0.32	0.15
	NIA	0.78	0.84	0.55	0.72	0.12	0.32
	WRA	0.82	0.90	0.81	0.86	0.74	0.86
<i>P. leniusculus</i>	NRA	0.68	0.75	0.50	0.60	0.49	0.35
	NIA	0.51	0.63	0.45	0.55	0.24	0.44
	WRA	0.95	0.99	0.88	0.96	0.82	0.91
<i>P. clarkii</i>	NRA	0.79	0.85	0.39	0.38	0.02	0.07
	NIA	0.79	0.86	0.52	0.74	0.05	0.21
	WRA	0.93	0.99	0.90	0.97	0.78	0.86

Table 2. Predicting distribution in introduced range: single-models validation results of kappa statistic (k) and area under the curve of receiver-operating characteristic (ROC-AUC) for invaded ranges using native range occurrences (NRO), native and invasive occurrences (NIO) and invaded range occurrences (IRO), coupled with within native range pseudo-absences random extraction (NRA), native and invasive ranges pseudo-absences random extraction (NIA) and the common spatially worldwide random pseudo-absences (WRA).

Species	Pseudo-absences	NRO		NIO		IRO	
		k	ROC-AUC	k	ROC-AUC	k	ROC-AUC
<i>C. destructor</i>	NRA	0.28	0.45	0.42	0.67	0.32	0.59
	NIA	0.23	0.39	0.28	0.42	0.64	0.72
	WRA	0.63	0.89	0.74	0.93	0.74	0.97
<i>E. sinensis</i>	NRA	0.57	0.67	0.80	0.90	0.68	0.81
	NIA	0.45	0.61	0.78	0.89	0.76	0.88
	WRA	0.90	0.95	0.90	0.96	0.91	0.97
<i>P. leniusculus</i>	NRA	0.58	0.70	0.66	0.78	0.83	0.85
	NIA	0.46	0.63	0.76	0.88	0.85	0.93
	WRA	0.89	0.93	0.90	0.96	0.91	0.96
<i>P. clarkii</i>	NRA	0.56	0.70	0.76	0.86	0.72	0.84
	NIA	0.33	0.56	0.80	0.91	0.81	0.95
	WRA	0.57	0.71	0.87	0.94	0.88	0.97

of some overlapping environmental conditions, native and invasive populations were differentiated along a gradient dominantly composed by the altitude and mean number of frost days variables. The PCA clouds for *P. leniusculus* indicated a less clear differentiation between the two ranges than for the previous species, but still had large non-overlapping portions along both components.

Discussion

In this study we aimed to test different yet plausible calibration data for predicting invasions for four major problematic decapods. Researchers have generally used either data from the species native or invaded ranges (Welk et al. 2002, Thuiller et al. 2005). More recent

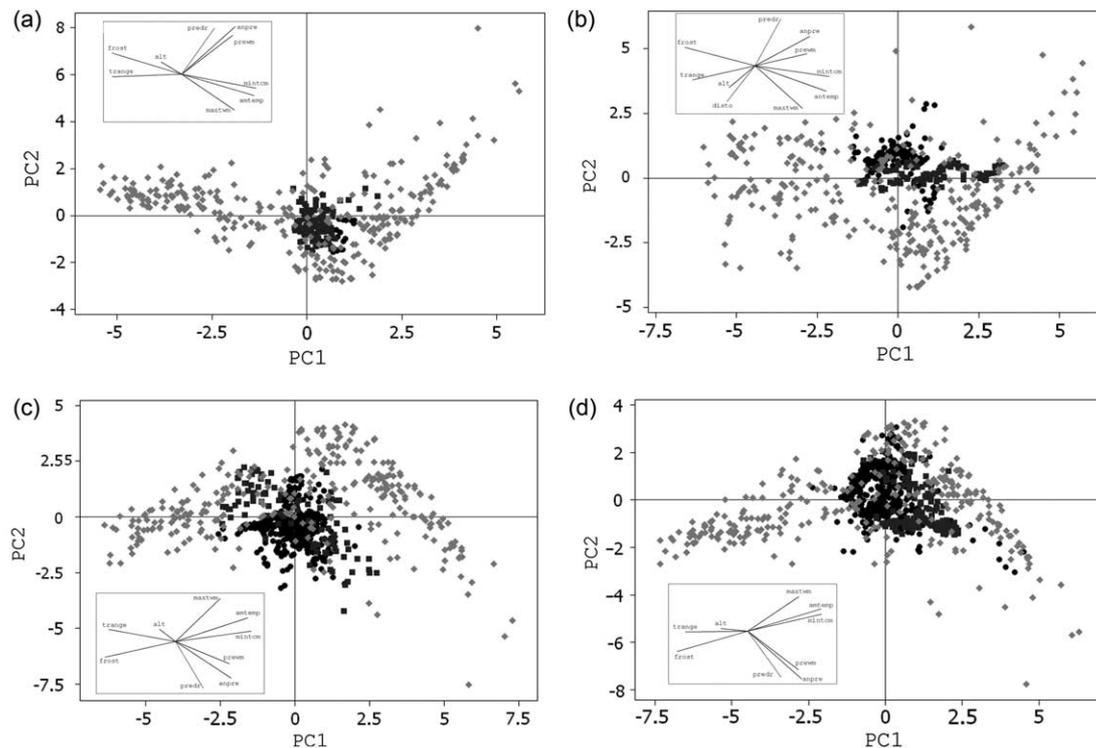


Figure 1. Position in the environmental space of native (grey squares) and invasive (black dots) populations and worldwide random pseudo-absences (light grey diamonds). A score plot of the two components was made from a PCA containing all environmental variables. PC1 and PC2 for (a) *Cherax destructor* with worldwide random pseudo-absences explained 76% of total variance, for (b) *Eriocheir sinensis* with worldwide random pseudo-absences explained 70% of total variance, for (c) *Pacifastacus leniusculus* with worldwide random pseudo-absences explained 75% of total variance, for (d) *Procambarus clarkii* with worldwide random pseudo-absences explained 73% of total variance. Components loadings are represented in the interior boxes.

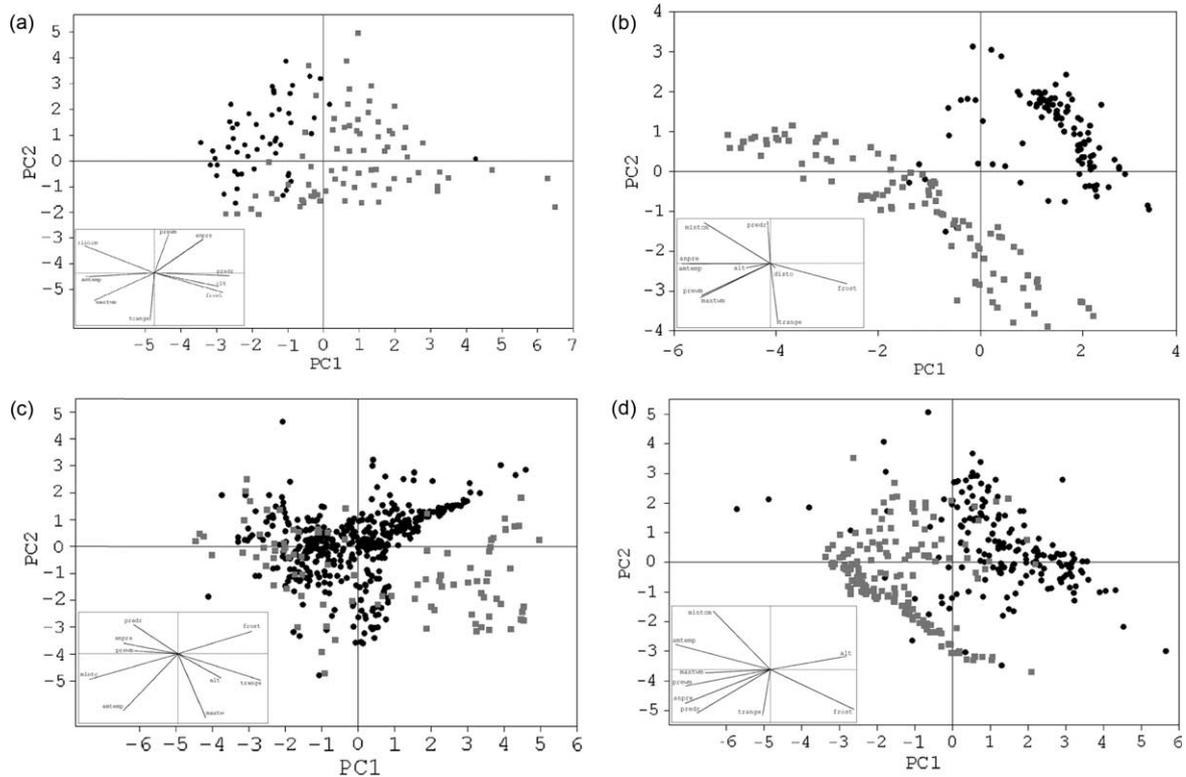


Figure 2. Position in the environmental space of native (grey squares) and invasive (black dots) populations. A score plot of the two components was made from a PCA containing all environmental variables. PC1 and PC2 for (a) *Cherax destructor* explained 69% of total variance, for (b) *Eriocheir sinensis* explained 66%, for (c) *Pacifastacus leniusculus* 77% and (d) *Procambarus clarkii* 67%. Components loadings are represented in the lower left corner.

studies have argued that due to possible changes in the species niches, coupling both native and invasive occurrence data in the calibration datasets might be the preferable option (Broennimann and Guisan 2008, Beaumont et al. 2009). In the same way pseudo-absences are also a known source of uncertainty in NBM (Lobo 2008, Phillips 2008), but to which little attention has been given in invasiveness predictions. Further, in invasion biology, NBM are typically applied to wide extents (continental or global scales), comprising a broad variety of environmental conditions that occur with different spatial frequencies and for which propagule pressure is usually unknown. Our results demonstrate that a large variability in predictive power can arise from the choice of both occurrence and pseudo-absence data.

Invasiveness predictability and the role of calibration data

Niche shifts occur when a species is occupying different environmental conditions in new areas or time periods than the ones found in initial populations. These shifts may be due to changes in the species realized niche (e.g. when a natural competitor is absent in new areas or the species moved to new environmental combinations in the invaded regions) or in its fundamental niche caused by changes in the species physiology (e.g. due to evolutionary change); either may undermine the ability of NBM to predict new

suitable areas for invaders (see Pearman et al. 2008, for a review). Our results are consistent with recent arguments that incorporating information from both native and introduced ranges yields the best estimate of the invasion potential (Broennimann and Guisan 2008, Beaumont et al. 2009), in that NBM built with both sets of occurrence data was able to simultaneously predict invasions in both ranges well. Thus, if such data are available, it should be used to make future predictions. However, while we should certainly be aware of the effects of possible niche shifts, our results also show that predictions based on native distributions could accurately forecast the invaded areas for three out of four species examined: *C. destructor*, *E. sinensis* and *P. leniusculus*, but not for *P. clarkii*, using the best performing pseudo-absence type (WRA, discussed below). These results suggest that, while caution is warranted, in the absence of information in the invaded range, NBM based on native distributions can still be useful in invasiveness forecasting, especially for forecasting the possible range of very new invasions or potential invasions that have not yet occurred. This is particularly relevant for invasive species modeling, where much of the interest in NBM has been its promise for forecasting invasions into new areas, before they actually occur. Further, despite environmental conditions differing between native and introduced ranges for all species (Fig. 2), NBM retained its predictive abilities. We argue that this occurred because NBM was able to identify which environmental variables were important for species

establishment, and down weighted those that were not important (Supplementary material Table S7).

While the majority of effort has been focused on the type of occurrence data (Mau-Crimmins et al. 2006, Loo et al. 2007, Broennimann and Guisan 2008, Beaumont et al. 2009), our results suggest that NBM models are even more sensitive to the type of pseudo-absences used. Native and native and invasive extraction methods attained the lowest performances for all species and types of occurrence data (Table 1 and 2). Although the ranges sampled by these pseudo-absences likely had the greatest chance of having received propagule pressure over time, this benefit appears to have been outweighed by having restricted environmental conditions which could not be extrapolated to new areas.

Worldwide random pseudo-absences had the stronger predictiveness achieving the highest evaluation scores for all predictions in both ranges (Table 1 and 2). This result seems supportive of this method. Still, the implications of sampling areas with unequal or unknown propagule pressure require further research. Consensus methods showed poor to high performances and could be a good alternative to deal with the wide array of possible outcomes. Still, their best performances in terms of predictiveness was no better than the best fitting single model – native and invasive occurrences with worldwide random pseudo-absences. While applying consensus methods for models obtained by distinct algorithms (e.g. regressions, classification models or machine learning) has been shown to improve predictive accuracy (Marmion et al. 2009, Stohlgren et al. 2010), we did not find obvious improvements for combining distinct calibration data. Still, the use of these ensembles may prove useful for cases where it is not possible to identify a single best performing model. Further, as we visually assessed in our predictions, single-models with similar accuracies may provide different spatial patterns of predictions. In such cases ensembles may also be of use by providing a consensual spatial pattern or model uncertainty.

Species invasiveness potential and conservation concerns

We used the best performing models in both native and introduced ranges to analyze their invasiveness potential. We found that each of the studied species still has large extents of suitable areas unoccupied (Fig. 3). This result is of high conservation concern, since the direct impact of these species in biological diversity is known to be high. Conservation problems caused by these decapods have been reported including disease transmission, competition, and active predation of native species and changes in the trophic webs of the invaded ecosystems (Gutierrez-Yurrita et al. 1999, Lynas et al. 2004, Correia and Anastácio 2008, Cruz et al. 2008, Dittel and Epifanio 2009). For example, Cruz et al. (2008) related the strong decline of both abundance and diversity of amphibian populations in a Portuguese protected wetland with the establishment of *P. clarkii*. These authors found that since the initial establishment of this invader, the number of amphibian species in the area was reduced from 13 to 6. It is thus worrying that the four species show suitable unoccupied areas that surpass the extents of the currently known invasive ranges (Fig. 4). Using the suitability threshold achieving higher kappa value to discretize predictions we found that higher suitability areas outside the species current ranges (both native and invaded) are nearly three times larger for *P. leniusculus*, four times for *P. clarkii*, seven times for *E. sinensis* and about nine times for *C. destructor*. Although also dependent on propagule pressure, these values indicate a large invasive potential for these species and detailed predictions should be conducted in areas of interest. Suitability for *C. destructor* and *P. clarkii* are noticeably similar (Fig. 3). This is a clear reflection of the similarity between the ecological preferences of these two species (Nyström 2002). Several biodiversity hotspots (Myers et al. 2000) fall under their environmental requirements. The Mediterranean basin and

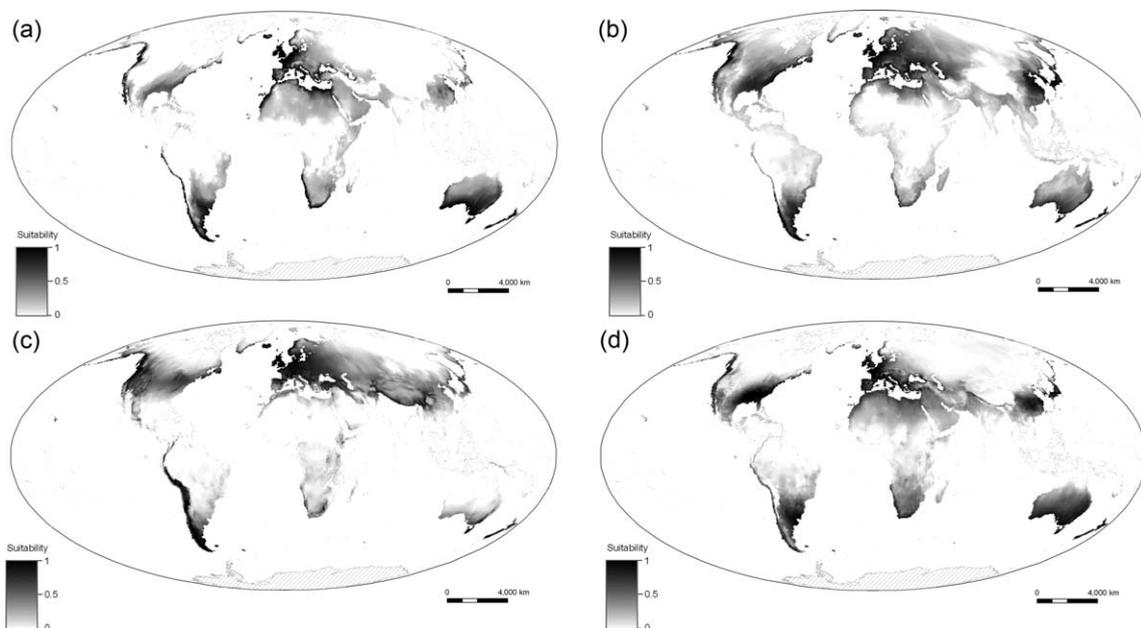


Figure 3. Best performing suitability models in both native and invaded areas for (a) *Cherax destructor*, (b) *Eriocheir sinensis*, (c) *Pacifastacus leniusculus* and (d) *Procambarus clarkii*.

Table 3. Validation results for consensus methods using kappa statistic (k) and area under the curve of receiver-operating characteristic (ROC-AUC) metrics. It compares for, each species, a weighted average of all single-models WA(all), a weighted average of all single-models using native range occurrences WA(NIO), a weighted average of all single-models using native and invasive occurrences WA(NIO), a weighted average of all single-models with invasive range occurrences WA(IRO), a weighted average of all single-models using native range pseudo-absences WA(NRA), a weighted average of all single-models using native and invasive ranges pseudo-absences random extraction WA(NIA) and a weighted average of all single-models using worldwide random pseudo-absences WA(WRA).

Species	Range	WA(all)		WA(NRO)		WA(NIO)		WA(IRO)		WA(NRA)		WA(NIA)		WA(WRA)	
		k	ROC-AUC												
<i>C. destructor</i>	Native	0.58	0.72	0.71	0.81	0.57	0.69	0.47	0.57	0	0.15	0.17	0.32	0.9	0.98
	Invaded	0.61	0.73	0.53	0.64	0.66	0.8	0.64	0.74	0.23	0.52	0.27	0.4	0.79	0.96
<i>E. sinensis</i>	Native	0.64	0.73	0.75	0.87	0.65	0.79	0.34	0.38	0.12	0.49	0.38	0.55	0.62	0.88
	Invaded	0.73	0.85	0.64	0.74	0.79	0.89	0.79	0.9	0.7	0.89	0.71	0.82	0.86	0.97
<i>P. leniusculus</i>	Native	0.61	0.73	0.75	0.85	0.59	0.73	0.27	0.56	0.11	0.55	0.53	0.61	0.74	0.94
	Invaded	0.73	0.84	0.58	0.7	0.74	0.88	0.74	0.92	0.77	0.95	0.63	0.76	0.87	0.97
<i>P. clarkii</i>	Native	0.62	0.77	0.8	0.91	0.61	0.7	0.27	0.32	0	0.34	0.12	0.24	0.84	0.98
	Invaded	0.69	0.84	0.55	0.67	0.73	0.88	0.76	0.92	0.67	0.85	0.67	0.79	0.79	0.94

southwest Australia are of special concern for *P. clarkii* and *C. destructor* respectively. The Mediterranean basin while being highly suitable for *P. clarkii* (Fig. 3) also encompasses the majority of *P. clarkii* invasive range in Europe (mostly found in the Iberian Peninsula). It is thus worrying that new nonadjacent invasions are also taking place here, such as in the Nile River (Cumberlidge 2009). Under this context the Mediterranean region may be particularly important because it contains a largely endemic biota and there is a great potential for invasion due to the high propagule pressure and environmental suitability. Likewise the currently largest invaded area for *C. destructor* occurs mostly within Western Australia – an area of high biological diversity – and its impact on the endemic crayfish species here is a concern (Lynas et al. 2004).

It is also worth noting that for these two species some high latitude areas such as several southern regions of Iceland and Greenland, Aleutian Islands and the southern tip of South America appear as suitable for this warm-water species. These areas, despite presenting colder temperatures than the ones verified in the majority of their distribution ranges, are under influence of the oceans moderating effect and their mean minimum temperatures reach fairly higher values than many other areas within the same latitudinal ranges. Although low temperatures have a known influence on some biological traits of *P. clarkii* (for a compilation see Anastácio et al. 1999) and *C. destructor* (Semple et al. 1995), these relatively unsuspecting areas are possibly on the edge of the thermal regimes required for these species to persist. Another possibility is that these values result from extrapolation errors. Whilst new methods for assessing that possibility exist (Elith et al. 2010) this evaluation would require a different modeling framework, beyond the scope of this work.

Also *P. leniusculus* show a large potential for future invasions. Despite already being the most widespread invasive crayfish in Europe, its invasion here seems to still be far from finished. The invasiveness projected for this continent shows a wide extent of suitable, yet uninvaded, areas mostly in eastern Europe, the Balkans and Turkey (Fig. 3). This potential expansion is of particular concern for native crayfish populations. Mostly due to competition and its role on transmission of the crayfish plague (*Aphanomyces astaci*), *P. leniusculus* is considered to have a major role in the extirpation of native crayfish populations (Souty-Grosset et al. 2006). Thus, crayfish conservationists should be aware that this species still has a large extent of environmentally suitable areas in contiguity to the existing invasive populations.

For *E. sinensis* five large suitable areas emerge as particularly vulnerable to new large-scale invasions: surroundings of Black and Caspian Seas, the Mediterranean basin (especially on the European side), and both eastern and western North American coasts (Fig. 3). In fact, besides presenting high suitability, small populations or isolated individuals have already been reported here (Dittel and Epifanio 2009). In contrast, it is not clear why *E. sinensis* has not been found in the southern Hemisphere. Our results also show that there are considerable extents of suitable environments in the southern areas of Australia, Africa and South America (Fig. 3). Ballast waters are the main vector of introduction of this species and some of

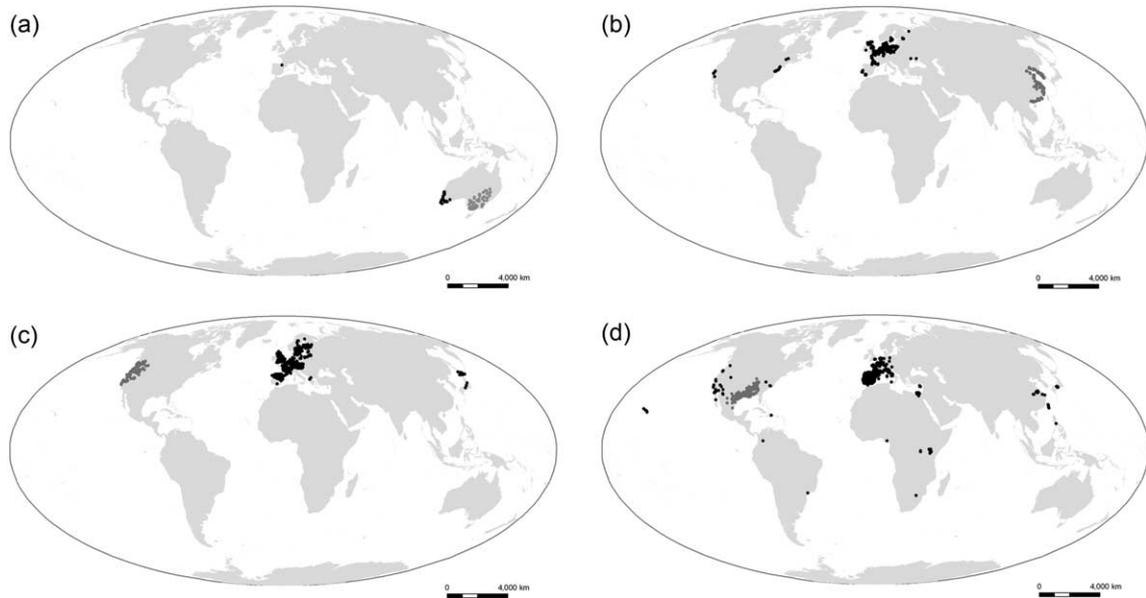


Figure 4. Native (grey circles) and invasive occurrences (black circles) of (a) *Cherax destructor*, (b) *Eriocheir sinensis*, (c) *Pacifastacus leniusculus* and (d) *Procambarus clarkii* used for model calibration.

these areas also have extensive international shipping connections. We therefore recommend that attention should be given to the possible existence of unknown populations or new introductions of *E. sinensis* in these areas since its establishment potential is high.

Caveats and future directions

While pseudo-absences are of current use in NBM, a large number of its implications remain poorly understood. We compared three distinct types of extraction methods and found substantial differences in accuracy. Still, more detailed comparisons should be addressed in the future. More specifically, different evaluation dataset types should also be tested. Our predictive evaluation relied on the use of worldwide pseudo-absences and, while to our mind this allows to test predictions for the entire study area, it may also provide overoptimistic scores for models calibrated with the same type of information. Further, increasing the number of invaders tested and more robust distribution datasets would also potentially increase the confidence of the obtained results. A comparison with other extraction approaches (and their paradigms) addressing issues such as sampling bias on the occurrence data (Phillips et al. 2009) or the use of presence-only NBM for driving extraction (Chefaoui and Lobo 2008) would also be highly valuable future directions, although they can require additional information that may not be available (e.g. some metric of sampling intensity or bias, Phillips et al. 2009).

Future work should also consider the effects that known issues of NBM such as spatial autocorrelation have in its predictive performance. While previous studies have found that simple spatial autocorrelation models can perform as well as NBM (Bahn and McGill 2007), this only indicates that we cannot distinguish between habitat suitability and spatial autocorrelation. However, because we were able to

extrapolate predictions from the native range to the exotic range, we were able to show predictive power in new spatially uncorrelated areas. Still, additional analyses on this topic are warranted. Other issues known to affect NBM such as unequal propagule pressure worldwide or potential lacks in our occurrence data (e.g. due to incomplete knowledge about the species invasive range) are of importance to our study and should also be explored in the future. One possibility to address the issues of unequal propagule pressure and spatial autocorrelation is to include a propagule pressure model (which provides a mechanistic model of spatial autocorrelation) together with our NBM (Leung and Mandrak 2007). Finally, obtained potential distributions for our invaders would also benefit from the availability of distribution data with higher spatial accuracy. Such data would allow the use of more detailed environmental data and thus increasing the detail of predictions which would also increase their value for managing purposes.

Despite the previous caveats, we present the first description of the worldwide invasiveness potential for four important invasive decapods. We also determined that the obtained projections are highly dependent on the type of data used for model calibration. While the type of occurrence data used could be important, accurate predictions were still obtained based solely on the occurrences in the native range for three of the four species examined. Moreover, our results suggest that pseudo-absences extraction methods were even more influential than the type of occurrence data used. Finally, despite its good results in other situations, the use of consensus models had limited benefit, with the best single modeling approach achieving consistently higher accuracies. Our results are supportive that worldwide predictions of invasiveness should be based on both native and invasive data, when available, and that worldwide random pseudo-absences seems the more favourable option if real absence information is lacking.

Acknowledgements – We thank the many researchers and institutions that provided species distribution data and three anonymous reviewers and the Subject Editor for their helpful comments. This research was performed within the scope of the DID (Dispersal of Invasive Decapoda) (PTDC/BIA-BEC/105182/2008) research project funded by the Portuguese Fundação para a Ciência e Tecnologia (FCT) and co-financed by FEDER, through the Eixo I of the Programa Operacional Factores de Competitividade (POFC) from QREN and a research grant provided to C. Capinha (SFRH/BD/41129/2007) (FCT).

References

- Anastácio, P. M. et al. 1999. CRISP (crayfish and rice integrated system of production): 2. Modelling crayfish (*Procambarus clarkii*) population dynamics. – *Ecol. Model.* 123: 5–16.
- Araújo, M. B. et al. 2005. Reducing uncertainty in projections of extinction risk from climate change. – *Global Ecol. Biogeogr.* 14: 538–529.
- Bahn, V. and McGill, B. J. 2007. Can niche-based distribution models outperform spatial interpolation? – *Global Ecol. Biogeogr.* 16: 733–742.
- Beaumont, L. J. et al. 2009. Different climatic envelopes among invasive populations may lead to underestimations of current and future biological invasions. – *Divers. Distrib.* 15: 409–420.
- Broennimann, O. and Guisan, A. 2008. Predicting current and future biological invasions: both native and invaded ranges matter. – *Biol. Lett.* 4: 585–589.
- Broennimann, O. et al. 2007. Evidence of climatic niche shift during biological invasion. – *Ecol. Lett.* 10: 701–709.
- Brotans, L. et al. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. – *Ecography* 27: 437–448.
- Burnham, K. P. and Anderson, D. R. 2002. Model selection and multimodel inference: a practical information-theoretic approach. – Springer.
- Chefaoui, R. M. and Lobo, J. M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. – *Ecol. Model.* 210: 478–486.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. – *Educ. Psychol. Meas.* 20: 37–46.
- Correia, A. and Anastácio, P. 2008. Shifts in aquatic macroinvertebrate biodiversity associated with the presence and size of an alien crayfish. – *Ecol. Res.* 23: 729–734.
- Côté, I. M. and Reynolds, J. D. 2002. Conservation biology: predictive ecology to the rescue? – *Science* 298: 1181–1182.
- Cruz, M. et al. 2008. Collapse of the amphibian community of the Paul do Boquilobo Natural Reserve (central Portugal) after the arrival of the exotic American crayfish *Procambarus clarkii*. – *Herpetol. J.* 18: 197–204.
- Cumberlidge, N. 2009. Freshwater crabs and shrimps (Crustacea: Decapoda) of the Nile Basin. – In: Dumont, H. J. (ed.), *The Nile*. Springer, pp. 547–561.
- DAISIE 2008. Handbook of alien species in Europe. – Springer.
- Dittel, A. I. and Epifanio, C. E. 2009. Invasion biology of the Chinese mitten crab *Eriocheir sinensis*: a brief review. – *J. Exp. Mar. Biol. Ecol.* 374: 79–92.
- Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – *Ecography* 29: 129–151.
- Elith, J. et al. 2010. The art of modelling range-shifting species. – *Methods Ecol. Evol.*, in press.
- Fitzpatrick, M. et al. 2007. The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? – *Global Ecol. Biogeogr.* 16: 24–33.
- Gollasch, S. 2006. NOBANIS – invasive alien species fact sheet – *Eriocheir sinensis*. – Online Database of the North European and Baltic Network on Invasive Alien Species – NOBANIS, <www.nobanis.org>.
- Gutiérrez-Yurrita, P. J. et al. 1999. The status of crayfish populations in Spain and Portugal. – In: Gherardi, F. and Holdich, D. M. (eds), *Crayfish in Europe as alien species: how to make the best of a bad situation?* A. A. Balkema, pp. 161–192.
- Hanley, J. A. and McNeil, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. – *Radiology* 143: 29–36.
- Hayes, K. and Barry, S. 2008. Are there any consistent predictors of invasion success? – *Biol. Invasions* 10: 483–506.
- Hirzel, A. H. et al. 2001. Assessing habitat-suitability models with a virtual species. – *Ecol. Model.* 145: 111–121.
- Landis, J. R. and Koch, G. G. 1977. The measurement of observer agreement for categorical data. – *Biometrics* 33: 159–174.
- Lek, S. and Guégan, J. F. 1999. Artificial neural networks as a tool in ecological modelling, an introduction. – *Ecol. Model.* 120: 65–73.
- Leung, B. and Mandrak, N. E. 2007. The risk of establishment of aquatic invasive species: joining invasibility and propagule pressure. – *Proc. R. Soc. B* 274: 2603–2609.
- Lobo, J. M. 2008. More complex distribution models or more representative data? – *Biodivers. Inform.* 5: 15–19.
- Loo, S. E. et al. 2007. Forecasting New Zealand mudsnail invasion range: model comparisons using native and invaded ranges. – *Ecol. Appl.* 17: 181–189.
- Lowe, S. et al. 2000. 100 of the world's worst invasive alien species. A selection from the global invasive species database. – ISSG, SSC and IUCN, <www.issg.org/booklet.pdf>.
- Lynas, J. et al. 2004. Is the yabby, *Cherax destructor* (Parastacidae) in Western Australia an ecological threat? – *Freshwater Crayfish* 14: 37–44.
- Marmion, M. et al. 2009. Evaluation of consensus methods in predictive species distribution modeling. – *Divers. Distrib.* 15: 59–69.
- Mau-Crimmins, T. M. et al. 2006. Can the invaded range of a species be predicted sufficiently using only native-range data? Lehmann lovegrass (*Eragrostis lehmanniana*) in the south-western United States. – *Ecol. Model.* 193: 736–746.
- Mitchell, T. D. and Jones, P. D. 2005. An improved method of constructing a database of monthly climate observations and associated high-resolution grids. – *Int. J. Climatol.* 25: 693–712.
- Myers, N. et al. 2000. Biodiversity hotspots for conservation priorities. – *Nature* 403: 853–858.
- Nyström, P. 2002. Ecology. – In: Holdich, D. M. (ed.), *Biology of freshwater crayfish*. Wiley-Blackwell, pp. 192–235.
- Özesmia, S. L. et al. 2006. Methodological issues in building, training, and testing artificial neural networks in ecological applications. – *Ecol. Model.* 195: 83–93.
- Pearce, J. L. and Boyce, M. S. 2006. Modelling distribution and abundance with presence-only data. – *J. Appl. Ecol.* 43: 405–412.
- Pearman, P. B. et al. 2008. Niche dynamics in space and time. – *Trends Ecol. Evol.* 23: 149–158.
- Phillips, S. J. 2008. Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al. (2007). – *Ecography* 31: 272–278.
- Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – *Ecol. Appl.* 19: 181–197.

- Segurado, P. and Araújo, M. B. 2004. An evaluation of methods for modelling species distributions. – *J. Biogeogr.* 31: 1555–1568.
- Semple, D. B. et al. 1995. *Cherax destructor*, *C. tenuimanus* and *C. quadricarinatus* (Decapoda: Parastacidae): a comparative review of biological traits relating to aquaculture potential. – *Freshwater Crayfish* 8: 495–503.
- Souty-Grosset, C. et al. (eds) 2006. Atlas of crayfish in Europe. – Museum national d'Histoire naturelle.
- Steiner, F. M. et al. 2008. Combined modelling of distribution and niche in invasion biology: a case study of two invasive *Tetramorium* ant species. – *Divers. Distrib.* 14: 538–545.
- Stohlgren, T. J. et al. 2010. Ensemble habitat mapping of invasive plant species. – *Risk Anal.* 30: 224–235.
- Thuiller, W. et al. 2005. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. – *Global Change Biol.* 11: 2234–2250.
- VanDerWal, J. et al. 2009. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? – *Ecol. Model.* 220: 589–594.
- Verdin, K. and Jenson, S. 1996. Development of continental scale DEMs and extraction of hydrographic features. – Proceedings of the Third Conference on GIS and Environmental Modeling, Univ. of California.
- Welk, E. et al. 2002. Present and potential distribution of invasive garlic mustard (*Alliaria petiolata*) in North America. – *Divers. Distrib.* 8: 219–233.
- Wilson, J. R. U. et al. 2007. Residence time and potential range: crucial considerations in modelling plant invasions. – *Divers. Distrib.* 13: 11–22.
- Witten, I. H. and Frank, E. 2005. Data mining: practical machine learning tools and techniques, 2nd ed. – Morgan Kaufmann.

Download the Supplementary material as file E6369 from www.oikos.ekol.lu.se/appendix.