Hermoso, V., Kennard, M.J. & Linke, S. (2014). Assessing the risks and opportunities of presence-only data for conservation planning. *Journal of Biogeography*. doi: 10.1111/jbi.12393

Original Article

LRH V. Hermoso *et al.*

RRH Assessing presence-only data for conservation planning

Assessing the risks and opportunities of presence-only data for conservation planning

Virgilio Hermoso*, Mark J. Kennard and Simon Linke

Australian Rivers Institute and Tropical Rivers and Coastal Knowledge, National Environmental Research Program Northern Australia Hub, Griffith University, Nathan, Queensland, 4111, Australia

*Correspondence: Virgilio Hermoso, Australian Rivers Institute and Tropical Rivers and Coastal Knowledge, National Environmental Research Program Northern Australia Hub, Griffith University, Nathan, Queensland, 4111, Australia.

E-mail: virgilio.hermoso@gmail.com

ABSTRACT

Aim Presence-only data represent a significant source of information for quantifying biodiversity distributions and provide opportunities for use in conservation planning. The large databases of presence-only records that are available and the lower cost of acquisition could help overcome the traditional problem of lack of data for conservation. However, there are risks associated with the use of presence-only data inherent with the lack of true absences that might cause omission errors (species are erroneously thought to be absent) and loss of efficiency (more areas are thought to be necessary than needed). These errors could constrain the economic viability of conservation plans and thus the success of conservation practice. We therefore evaluated the opportunities and risks of using presence-only data for conservation planning.

Location Northern Australia.

Methods The effects of using two different types (presence-only and presence–absence) and different quantities of data were simulated by building predictive models on different subsets of data with increasing numbers of presence–absence or presence-only records or a combination of both for 80 freshwater fish species. We then compared the performance of conservation planning outcomes with the best information attainable (a true model built on the complete set of presence–absence data). We measured omission and commission errors in conservation planning outcomes, and the efficiency of and return on the investment in data acquisition.

Results Including presence-only data helped reduce commission and omission errors in conservation planning outcomes, but only when used in combination with at least some presence–absence data. The use of just a large quantity of presence-only data resulted in significant reductions in the efficiency of conservation planning outcomes, as more areas than actually needed were required to achieve conservation targets. This reduction in efficiency was mainly related to inflated omission errors.

Main conclusions We recommend using presence-only data cautiously if this is the only source of data available; whenever possible, presence-only data should be complemented with presence–absence data.

Keywords

Australia, commission error, conservation biogeography, efficiency, freshwater biodiversity, omission errors, predictive model, sensitivity, systematic planning, trade-offs.

INTRODUCTION

Conservation planning and decision-making benefit from robust and accurate information regarding the distribution of conservation features, normally species (Balmford & Gaston, 1999; Hirzel & Guisan, 2002; Hermoso *et al.*, 2013), and other socio-economic factors, such as conservation cost (e.g. land acquisition or stewardship). In an ideal scenario, the spatial distribution of all the conservation features would be known from recent ground surveys. However, stakeholders are constantly challenged to deal with sparse data subject to uncertainties and spatial/taxonomic bias (Possingham *et al*., 2000) because of the limited budgets and time available (Halpern *et al*., 2006; Grantham *et al*., 2008). The errors inherent in imperfect datasets can lead to poor decision-making, the misuse of limited resources and ultimately the failure of conservation practice (Possingham *et al*., 2007). For example, errors in estimates of species distributions can reduce the effectiveness (e.g. through commission errors, arising when a species is erroneously thought to be present) and efficiency (e.g. through omission errors, arising when a species is erroneously thought to be absent, forcing the selection of additional and unnecessary areas; Loiselle *et al*., 2003) of conservation recommendations. Consequently, the uncertainties in the information regarding the spatial distribution of biodiversity and the errors derived when the information is used for conservation planning must be assessed, communicated and taken into consideration (Rondinini *et al*., 2006).

The two main types of biological data normally available for conservation assessments are presence–absence and presence-only records. While both of them contain observed presences, only in presence–absence data have the absences been scrutinized (Tsoar *et al*., 2007). The acquisition cost of these types of data can differ substantially and result in unavoidable trade-offs between the investment and quality of data required. Collecting presence–absence records is time-consuming and requires the application of standardized survey methods that result in higher costs (Burbidge, 1991) but are less prone to omission errors. On the other hand, a potential advantage in using presence-only data for conservation is the existence of extensive records available in museums and from volunteer field surveys (Tulloch *et al*., 2012) that can help reduce the data-acquisition cost significantly (Danielsen *et al*., 2009). However, omission errors in presence-only datasets might reduce the efficiency of conservation planning outcomes. Assessing these trade-offs could highlight the value of the important quantity of presence-only data already available and guide future investment in data collection (whether in presence-only or presence–absence data) and lead to better informed decisions for conservation.

Sensitivity analyses can be used to evaluate the effect of using different types and quantities of data by testing the response of conservation planning outcomes to simulated or controlled manipulations of available datasets (e.g. by deleting some of the available records; Gaston & Rodrigues, 2003; Langford *et al*., 2009; Hermoso *et al*., 2013). Using this approach, previous studies have reported a decline in efficiency when deleting some of the available occurrence records (e.g. Freitag & Van Jaarsveld, 1998; Grand *et al*., 2007; De Ornellas *et al*., 2011) and recommended data acquisition as a general rule to reduce errors in planning outcomes (but see Gaston & Rodrigues, 2003; Hermoso *et al*., 2013). However, the sensitivity analyses in these studies were carried out by comparing the performance of conservation assessments (e.g. efficiency) only on subsets of the original point locality data available. This does not relate directly to the type of data usually used for conservation assessments, namely estimates of the complete spatial distribution of biodiversity (e.g. Loiselle *et al*., 2003; Vaughan & Ormerod, 2003; Wilson *et al*., 2005). Moreover, following the principle of complementarity (Kirkpatrick *et al*., 1983), a sensitivity analysis carried out on reduced sets of point locality data could lead to an overestimation of inefficiency. As fewer localities are made available for conservation planning, it becomes increasingly difficult to find sites that complement each other, and more of them will be necessary to represent all species. In order to provide a more robust evaluation of the value of presence-only records for conservation assessments, a more standardized sensitivity analysis is required where only the quality of the information associated with presence-only records is assessed.

To overcome this problem, sensitivity analyses could be carried out on complete distributions of conservation features. Predictive models have been highlighted as the best approach to providing data on biodiversity distributions for conservation planning (Loiselle *et al.*, 2003; Wilson *et al.*, 2005; Rondinini *et al.*, 2006). Estimates of habitat suitability derived from predictive models help overcome the inflated commission or omission errors associated with other methods, such as point locality data or estimates of geographical ranges (Rondinini *et al.*, 2006; Bombi *et al.*, 2011). Moreover, advances in predictive modelling techniques to enable the use of presence-only records (e.g. Phillips *et al.*, 2006; Elith & Leathwick, 2007) have facilitated the use of this type of data for conservation planning (e.g. De Ornellas *et al.*, 2011; Rodrigues, 2011).

We used a data-rich area in northern Australia as a case study to demonstrate the trade-offs associated with the use of different quantities of presence–absence and presence-only data in conservation planning. We compiled an extensive dataset containing presence–absence and presence-only records and used it to simulate the effect of using different types and quantities of data. We first modelled the spatial distribution of 80 freshwater fish species using the complete presence–absence dataset, and used it as the true distribution of each species representing the best available information (Grantham *et al.*, 2009; Hermoso *et al.*, 2013). We then used these predictions to assess the performance of alternative distribution maps obtained from predictive models built on different subsets of the database (subsampled from the complete dataset). Three different strategies were tested: (1) the use of increasing quantities of presence–absence data; (2) the use of increasing quantities of presence-only data; and (3) the use of increasing quantities of presence-only data to complement a sparse presence–absence dataset. The distributions maps produced from these different datasets were assumed to represent the information that a planner would have available if the area had not been extensively surveyed. We evaluated the performance of each distribution map for conservation planning with measures of commission and omission errors in conservation planning outcomes (*sensu* Rondinini *et al*., 2006), relative efficiency (the ratio of the number of planning units needed to achieve the conservation targets based on imperfect data/the actual number of planning units needed) and return on investment (the ratio of species representation/sampling cost). We conclude with a set of clear and transferable recommendations concerning the trade-offs involved in using data of varying quality and quantity for conservation planning.

MATERIALS AND METHODS

Study area and biological data

The study area spanned northern Australia's coastal river basins, from the Fitzroy River in the Kimberley region eastwards to the Jardine River in Cape York Peninsula. It encompassed a total area of about 1.19 million km^2, about 15% of the Australian continental area. The region supports approximately 60% of Australia's freshwater fish diversity (Pusey *et al*., 2011). We compiled all the available sampling records for 107 freshwater fish species across the study area (Hermoso *et al*., 2013). This dataset contained records for more than 2328 sampling records, including presence–absence data (30% of the records) and presence-only data (70% of the records). These records covered a wide range of environmental conditions in northern Australia and were the best available data for the area (see Appendix S1 in Supporting Information). For subsequent modelling purposes, we translated these sampling records into a network of predictive units. We delineated 11,508 subcatchments (on average 102.7 km^2) using ARC HYDRO (Maidment, 2002) for ARCGIS 9.3 (ESRI, 2002) from a 9-s digital elevation model (Geoscience Australia, 2011). We discarded from the dataset all species with less than five occurrences in the presence–absence dataset because of difficulties in modelling the distribution of these extremely rare species and the potential bias they would introduce to the analyses. Our final dataset comprised 80 freshwater fish species with an average frequency of occurrence in the presence–absence dataset of 106 subcatchments (range 5–581). The proportion of presence-only records varied across species, representing from none to up to 86% of the sampling records available (see Appendix S2).

Predictive modelling of species distributions

Nine ecologically relevant landscape-scale environmental variables were selected for use in the predictive models from a larger set of 38 candidate variables, and were derived from the National Environmental Stream Attributes database for rivers (for details see Geoscience Australia, 2011). We used principal components analysis (PCA) to select a set of nine non-redundant (Pearson’s $r < 0.25$ in all cases) environmental attributes that explained a high proportion of the environmental variability in the data (see Appendix S3).

We used multivariate adaptive regression splines (MARS; Leathwick *et al.*, 2005) to model the spatial occurrence of the 80 fish species with more than five occurrences in the presence–absence dataset. The model was built on the entire presence–absence dataset (hereafter referred to as the true model). Model performance was evaluated using two complementary approaches: deviance explained and the area under the receiver operating characteristic (ROC) curve (AUC) (Fielding & Bell, 1997). The AUC was assessed with a *k*-fold cross-validation procedure (Hastie *et al.*, 2001). With this process, the dataset is randomly divided into 10 exclusive subsets and model performance is calculated by successively removing each subset, refitting the model with the remaining data and predicting the omitted data. Deviance complements AUC because it expresses the magnitude of the deviations of the fitted values from the observations.

The model was then used to predict the probability of occurrence of each species in all of the unsurveyed areas (all within the range of the environmental conditions covered in the dataset used to build the models). Probabilities of occurrence were transformed into presence–absence data for subsequent analyses using the optimal threshold obtained from the ‘cost’ method in the ‘PresenceAbsence’ package in R (Freeman, 2007) as recommended by Benito *et al.* (2013). Given that these predictions were made with the best and most accurate dataset available, we treated them as our true species distributions for subsequent analyses (for a similar approach see Grantham *et al*., 2009; Langford *et al*., 2009; Hermoso *et al*., 2013).

Simulation of presence-only and presence–absence data acquisition

In order to test the effect of using different types (presence–absence and presence-only) and quantities of data for conservation planning, we built predictive models on different subsets of the data available, including presence–absence, presence-only and mixed datasets (presence–absence and presence-only; Fig. 1). Both presence–absence and presence-only models included two subsets of data, with a sparse (250 records) and large (500 records) quantity of data, respectively. The records for the sparse models were selected randomly from the whole dataset. Records for the large models were selected randomly from the remaining pool of records and added to the sparse dataset, so they reflected the addition of 250 extra sampling records to the existing sparse datasets. These models were used to evaluate the benefit of acquiring presence–absence versus presence-only data for conservation planning.

We also built four additional mixed models (mixed 1–4) with increasing quantities of presence-only data (150, 300, 600 and 1200 presence-only records for mixed 1, 2, 3 and 4, respectively) but a constant quantity of presence–absence data (250 records). Presence-only records for mixed models were selected randomly and added progressively to the initial sparse presence–absence models (Fig. 1). These mixed models were used to evaluate the suitability of investing in presence-only data acquisition to complement already existing sparse presence–absence datasets. Each of the eight datasets was used as if no additional data were available, so we followed the same procedure as detailed above for evaluating model performance (no additional data from the remaining sets were used). We used the same set of environmental predictors across all models (the same set of variables used for constructing the true model; see Appendix S3) and applied the same minimum threshold of five occurrences for a species to be included in the predictive model built for each dataset. Given that some of the models were built on just presence-only data, we also carried out an external validation to confirm the quality of these models by using an independent set of 219 presence–absence records (the remaining data not used for any of the models apart from the true model).

Identification of conservation priority areas using different data-availability scenarios

We used the predictions from each of the eight models as surrogates of biodiversity patterns to identify priority areas for conservation using the spatial prioritization software MARXAN (Ball *et al*., 2009). Our aim was to find an optimal set of planning units to represent 25% of each species' true distribution at the minimum cost. MARXAN uses simulated annealing, an optimization algorithm, to minimize an objective function that includes the cost of subcatchments (our units of planning) in the solution and penalties for not achieving the conservation target for all the species. Our goal was not to design a reserve network suitable for implementation of on-the-ground conservation, but to examine the effects of data type and quantity on the characteristics of hypothetical priority area networks. Given our special interest in evaluating the effect of using different types and quantities of data, we used a constant cost for each planning unit, so our objective translated into finding the minimum set of planning units to achieve the conservation targets (Hermoso & Kennard, 2012). We ran an independent conservation planning scenario for each predictive model output and retained 100 solutions obtained after 2.5 million iterations each for further analyses.

Commission and omission errors in conservation planning outcomes

We measured the commission error rate in conservation planning outcomes (when a species is erroneously thought to be present within a conservation priority area) and omission error rate (when a species is erroneously thought to be absent from a conservation priority area) by comparing the expected and observed representation achieved for each species (note that the use of commission and omission error rates here refers to errors in conservation planning outcomes rather than errors in predictions, as is usually the case with predictive modelling approaches). We did this for each of the eight scenarios × 100 solutions obtained from MARXAN. The expected representation was measured as the number of occurrences within solutions according to the surrogate data used in the conservation prioritization process (each of the eight different predictive models). This was treated as the expected representation because it indicated the potential representation that would be achieved if the predictions used had no associated errors. The observed representation was measured as the number of occurrences within solutions according to the true spatial distribution (i.e. based on the true model using the entire presence–absence dataset; Fig. 1). We then measured the rate of commission and omission errors in conservation planning outcomes as the proportion of expected representation that was not actually achieved [error = (expected representation – observed representation)/expected representation]. Whenever the expected and observed representations were similar, the error obtained was close to 0, indicating a low commission or omission error. However, when the expected representation was significantly higher or lower than the observed representation, the error value departed from 0 and was negative or positive, indicating commission and omission errors, respectively (Hermoso *et al*., 2013). Commission and omission errors were calculated independently for each species.

Relative efficiency and return on data-acquisition investment

We measured the relative efficiency of solutions under alternative data-availability scenarios by comparing the number of planning units required for each model to achieve targets against the number of planning units required when using the best available data (the true model). We ran an independent optimization exercise using the true spatial distribution and then compared the average number of planning units required to achieve the same target across 100 solutions against the eight previous scenarios [relative efficiency = (PU_{true} – $PU_{sceanrio}$)/PU_{true}], where PU_{true} and $PU_{scenario}$ are the average number of planning units across 100 solutions under the true distribution and each of the different data-availability scenarios, respectively. Negative values of this index of relative efficiency indicated inefficient solutions, as a higher number of planning units than actually needed was selected. Positive values indicated high efficiency, as the conservation targets could be achieved with less planning units than using the true distribution.

We estimated the investment necessary to assemble each dataset assuming a constant survey cost per site that only varied between presence–absence and presence-only data. Other approaches, for example accounting for distance between sampling sites, have been used to make more accurate estimates of biological surveys (e.g. Wessels *et al*., 1998). However, given that travelling distances were not considered here when selecting new sampling records (they were added randomly), we assumed a constant cost of $1500 and $500 per site for presence–absence and presence-only data, respectively. This reflected the relative survey effort required to characterize presence–absence versus presence-only data adequately. The return on a given investment for each subset of data was assessed as the average ratio across species between the observed representation and the cost of survey.

RESULTS

Predictive model performance

On average, the predictive models showed a high performance, with AUC > 0.75 and deviance explained > 0.3, even for the sparse data models (Fig. 2a,b). Data acquisition (independent of the type of data used) resulted in improved model accuracy. These results were confirmed with the external validation, as there was < 5% difference in AUCs obtained from the independent presence–absence dataset for all models.

Commission and omission errors in conservation planning solutions

The use of exclusively presence-only data for conservation planning resulted in high omission errors that could be up to three times the omission error rate obtained when using only presence–absence data (Fig. 3a). These errors could be reduced by combining the presence-only data with presence–absence data, although a large quantity of presence-only records was required to achieve similar values of omission error rates to the solutions obtained from presence–absence data (Fig. 3a). Commission errors were up to 50% higher when using sparse presence–absence and presence-only data independently of the quantity of data available (Fig. 3b).

Effectiveness and relative efficiency of conservation planning solutions

None of the alternative species distribution scenarios resulted in 100% effective conservation planning solutions (e.g. the observed representation of all species was enough to achieve the target). Note that MARXAN always achieved the target with the data provided, but given the errors in predictions some of the expected representation was not true and this translated into loss of effectiveness. Effectiveness ranged from 69% for the sparse presence-only scenario to more than 85% of species for the large presence-only, large presence–absence and mixed 1 scenarios (85%, 88% and 88%, respectively; Fig. 4a).

None of the scenarios tested produced solutions more efficient than the ones derived from the true distribution model, reflecting the implications of errors in the spatial distribution of species. The use of presence-only data resulted in a drastic reduction of efficiency, as up to 35% more planning units were needed to achieve the targets for the large presence-only scenario. The large presence–absence dataset showed the best efficiency of all the data-availability scenarios tested (Fig. 4b).

Cost and return on investment of data acquisition for conservation planning

The return on investment of data acquisition for conservation planning declined drastically with data addition, independently of the type of data added (Fig. 5b). Given the low cost of acquiring presence-only datasets (Fig. 5a), investing in sparse presence-only data resulted in returns three times higher than using sparse presence–absence data. The investment in presence–absence data produced higher returns than presence-only data only when large quantities of presence-only data were combined with the sparse presence–absence data (mixed 4; Fig. 5b). Investing in a large presence–absence dataset showed similar costs and returns as investing in the mixed 3 scenario (sparse presence–absence data and a large quantity of presence-only data). However, the mixed 3 scenario resulted in priority areas that were 6% less effective and 5% less efficient than the large presence-absence scenario.

DISCUSSION

The results demonstrate clear trade-offs in using data of variable quality and quantity for predictive modelling and conservation planning. The use of presence-only data allows the identification of priority areas for conservation at a relatively lower acquisition cost given the large quantity of information already available in public databases, etc. However, we have shown that combining presence-only data with at least some presence–absence data is needed to help reduce commission and omission errors in conservation planning outcomes. Moreover, there was a significant reduction in efficiency when using exclusively presence-only data, as more areas were required to achieve conservation targets. For this reason we recommend using presence-only data cautiously if this is the only source of data available, and complementing it with presence–absence data whenever possible.

Predictive models built on presence–absence data have performed better than models built on just presence-only data (e.g. Hirzel *et al*., 2001; Brotons *et al*., 2004). However, recent advances in predictive modelling techniques have made it possible to use presence-only records to derive reasonably accurate estimates of species distributions (Phillips *et al*., 2006). This is the case with the MARS models used in this study, which have been shown to outperform other traditional methods (Elith *et al*., 2006) and to be suitable for modelling species distributions when only presence data are available (Elith & Leathwick, 2007). Moreover, multispecies models like MARS help to overcome the traditional problem faced by ecologists and modellers attempting to model rare species (species with very few records; see Boitani *et al*., 2011) and circumvent the usual issue with selecting pseudo-absences or background data where species are assumed to be absent (Elith & Leathwick, 2007). We have shown that predictive models built on just presence-only data or in combination with presence–absence data can be significantly better than random models and are not significantly worse than models built on presence–absence data (see also MacLeod *et al.*, 2008). The performance of presence-only models in this study was similar to that reported for presence–absence MARS models in previous studies (e.g. Leathwick *et al.*, 2005; Hermoso *et al.*, 2011). Data acquisition proved to be beneficial in terms of increased modelling performance, measured by the AUC and the deviance explained independent of the type of data used. Given that no true absence data were available for some models, predictive models validated on presence-only data were tested for their ability to predict presences correctly, which could indicate potential commission errors (a high AUC could indicate low commission errors) but not absences (Phillips *et al.*, 2006; Elith & Leathwick, 2007). There was no significant difference between the internal cross-validation performed on the data available for each model and the external validation carried out on the independent presence–absence dataset. Thus the original estimates of modelling performance could be taken as indicators of suitability for predictions for conservation planning. However, the apparent benefit of data acquisition for modelling performance was not coupled with improvements in conservation planning outcomes. We agree with Lobo *et al.* (2008) and Jiménez-Valverde (2012) about the shortcomings of traditional measures of modelling performance such as AUC and deviance, and also argue that they do not provide appropriate indicators of model suitability for conservation applications.

Commission and omission errors can compromise conservation efforts from different perspectives: a lack of effectiveness (inflated commission errors) and a lack of economic viability, respectively. Loiselle *et al.* (2003) and Rodrigues (2011) suggested that minimizing commission errors is more important than omission errors to ensure the adequacy of conservation plans. Commission errors are particularly hazardous in conservation given that species might be erroneously thought to be protected. We expect that the use of presence-only data could help reduce commission errors in conservation planning outcomes as more information on confirmed presences could be incorporated in the planning process. However, the lack of certainty on absences might incur a high risk of omission errors, given that the species could be erroneously thought to be absent from an area. Omission errors lead to loss in efficiency, given that more areas than actually needed are selected to achieve the conservation targets (Rondinini *et al*., 2006), undermining the potential viability of the conservation plan. Our results align with this logic, as the use of just presence-only data resulted in high rates of omission errors and helped reduce commission errors, although only moderately. Given the inflated omission errors produced from predictive models based on large presence-only datasets, we recommend cautious use and interpretation of conservation plans derived from just presence-only data. We have demonstrated that these errors can be significantly reduced by combining the presence-only data with at least some presence–absence data, or by using presence–absence data alone if only sparse data are available. This highlights the value of even relatively small datasets for conservation planning (Gaston & Rodrigues, 2003; Grantham *et al*., 2009; Hermoso *et al*., 2013) and shows that a small investment in collecting presence–absence data, which could be used to complement existing presence-only datasets, might be enough to enhance the reliability of conservation planning outcomes.

Although none of the predictive scenarios produced conservation plans as efficient as the one based on the true model, the loss in efficiency was especially marked when using a large quantity of presence-only data, and efficiency was only close to the true model when using a large quantity of presence–absence data. On the other hand, the use of a large quantity of presence-only data resulted in highly effective solutions similar to that obtained with the use of large quantity of presence–absence data. This result highlights a trade-off between effectiveness and efficiency similar to the trade-offs reported in previous studies (Rodrigues *et al*., 2000; Wilson *et al*., 2005). In general, an improvement in effectiveness (less risk-prone solutions) is usually achieved at the expense of efficiency (e.g. larger areas are needed to make sure all the species are adequately protected). Interestingly, in our case this was only true with the use of presence-only data, as solutions obtained from presence–absence data were both effective and efficient. This again emphasizes the main limitation related to the use of presence-only data for conservation: the use of accurate information on presence is a safe option in terms of species representation (effectiveness), while the lack of information on absence affects efficiency, consequently undermining the conservation plan's viability.

There are additional trade-offs between effectiveness and efficiency once the cost of acquisition of different types of data is considered. Data acquisition was generally a good option in terms of increasing effectiveness (more species achieved the true target as more data on their spatial distribution became available) but did not always increase efficiency because it depended on the type of data acquired. For example, the same effectiveness was achieved when using the large datasets for both types of data (presence–absence and presence-only). However, the investment required to achieve the same effectiveness would be 2.7 times higher if using presence–absence data instead of presence-only data. On the other hand, the cheaper investment in presence-only data resulted in conservation plans that were 3.4 times less efficient than using presence–absence data. So, investing in acquisition of presence-only data would lead to enhanced effectiveness of conservation recommendations only if enough resources were available to facilitate the implementation of less efficient plans. This is unlikely to happen given the usually limited resources available for the implementation of conservation action on the ground (Knight *et al*., 2007). We recommend avoiding this economically risky strategy because it could undermine the viability of the conservation plan. Instead of ‘shopping’ for cheap data that could have negative consequences on the economic viability of the conservation action, we recommend enhancing efficiency by acquiring presence–absence data. However, we know that the quantity of presence–absence data that can be collected is constrained by limited budgets as well, while large datasets on presence-only data are already available or can be acquired at a lower cost. Therefore, we believe that an optimal strategy for enhancing both effectiveness and efficiency of conservation plans within budgetary constraints would be to combine cheap presence-only data with some more expensive but reliable presence–absence data.

Finally, there was a diminishing return on investment in data acquisition (similar to Grantham *et al*., 2008; Hermoso *et al*., 2014). The average species representation (effectiveness) for each dollar spent in data collection declined with increasing quantities of data collected, independent of the type of data (presence only versus presence–absence). For this reason, we recommend planning the investment in data acquisition carefully (e.g. prioritizing the collection of data necessary to reduce a spatial or taxonomic bias). Given the value of reduced datasets for conservation, the resources that could be spent in data acquisition might be better directed towards other actions, such as implementation of on-the-ground management or acquisition of data on other factors involved in decision-making, such as monitoring programmes for population trends, occurrence and intensity of threats or conservation costs. This would enhance the success of conservation practice, especially in areas where biodiversity is under pressure from increasing threats (Grantham *et al*., 2009).

ACKNOWLEDGEMENTS

We thank J. Stein for providing environmental data, and B. Pusey and D. Burrows for assistance with compilation of the fish data. We acknowledge the funding support provided by the Australian Research Council (Discovery Grant DP120103353 to S.L. and M.K., DECRA DE130100565 to S.L.), the Australian Government Department of Sustainability, Environment, Water, Population and Communities, the Tropical Rivers and Coastal Knowledge (TRaCK) Research Hub, the National Environmental Research Program Northern Australia Hub, and the Australian Rivers Institute, Griffith University.

REFERENCES

Ball, I.R., Possingham, H.P. & Watts, M. (2009) Marxan and relatives: software for spatial conservation prioritisation. *Spatial conservation prioritisation: quantitative methods and computational tools* (ed. by A. Moilanen, K.A.Wilson and H.P. Possingham), pp. 185–195. Oxford University Press, Oxford.

Balmford, A. & Gaston, K.J. (1999) Why biodiversity surveys are good value. *Nature*, **398**, 204–205.

Benito, B.M., Cayuela, L. & Albuquerque, F.S. (2013) The impact of modelling choices in the predictive performance of richness maps derived from species-distribution models: guidelines to build better diversity models. *Methods in Ecology and Evolution*, **4**, 327–335.

Boitani, L., Maiorano, L., Baisero, D., Falcucci, A., Visconti, P. & Rondinini, R. (2011) What spatial data do we need to develop global mammal conservation strategies? *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**, 2623–2632.

Bombi, P., Luiselli, L. & D'Amen, M. (2011) When the method for mapping species matters: defining priority areas for conservation of African freshwater turtles. *Diversity and Distributions*, **17**, 581–592.

Brotons, L., Thuiller, W., Araujo, M.B. & Hirzel, A.H. (2004) Presence–absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437–448.

Burbidge, A.A. (1991) Cost constraints on surveys for nature conservation. *Nature conservation: cost effective biological and data survey*s (ed. by C.R. Margules and M.P. Austin), pp. 3–6. CSIRO, Melbourne.

Danielsen, F., Burgess, N.D., Balmford, A. *et al.* (2009) Local participation in natural resource monitoring: a characterization of approaches. *Conservation Biology*, **23**, 31–42.

De Ornellas, P., Milner-Gulland, E.J. & Nicholson, E. (2011) The impact of data realities on conservation planning. *Biological Conservation*, **144**, 1980–1988.

Elith, J. & Leathwick, J. (2007) Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions*, **13**, 265–275.

Elith, J., Graham, C.H., Anderson, R.P. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

ESRI (2002) *ArcGIS*. Environmental Systems Research Institute, Redlands, CA.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Freeman, E. (2007) *Presence/absence: an R package for presence–absence model evaluation.* USDA Forest Service, Rocky Mountain Research Station, Ogden, UT.

Freitag, S. & Van Jaarsveld, A.S. (1998) Sensitivity of selection procedures for priority conservation areas to survey extent, survey intensity and taxonomic knowledge. *Proceedings of the Royal Society B: Biological Sciences*, **265**, 1475–1482.

Gaston, K.J. & Rodrigues, A.S.L. (2003) Reserve selection in regions with poor biological data. *Conservation Biology*, **17**, 188–195.

Geoscience Australia (2011) *Environmental attributes database*. Available at: http://www.ga.gov.au/ (last visited 29 November 2012).

Grand, J., Cummings, M.P., Rebelo, T.G., Ricketts, T.H. & Neel, M.C. (2007) Biased data reduce efficiency and effectiveness of conservation reserve networks. *Ecology Letters*, **10**, 364–374.

Grantham, H.S., Moilanen, A., Wilson, K.A., Pressey, R.L., Rebelo, T.G. & Possingham, H.P. (2008) Diminishing return on investment for biodiversity data in conservation planning. *Conservation Letters*, **1**, 190–198.

Grantham, H.S., Wilson, K.A., Moilanen, A., Rebelo, T. & Possingham, H.P. (2009) Delaying conservation actions for improved knowledge: how long should we wait? *Ecology Letters*, **12**, 293–301.

Halpern, B.S., Regan, H.M., Possingham, H.P. & McCarthy, M.A. (2006) Accounting for uncertainty in marine reserve design. *Ecology Letters*, **9**, 2–11.

Hastie, T., Tibshirani, R.J. & Friedman, J.H. (2001) *The elements of statistical learning: data mining, inference and prediction.* Springer-Verlag, New York.

Hermoso, V. & Kennard, M.J. (2012) Uncertainty in coarse conservation assessments hinders the efficient achievement of conservation goals. *Biological Conservation*, **147**, 52–59.

Hermoso, V., Linke, S., Prenda, J. & Possingham, H.P. (2011) Addressing longitudinal connectivity in the systematic conservation planning of fresh waters. *Freshwater Biology,* **56**, 57–70.

Hermoso, V., Kennard, M.J. & Linke, S. (2013) Data acquisition for conservation assessments: is the effort worth it? *PLoS ONE*, **8**, e59662.

Hermoso, V., Kennard, M.J. & Linke, S. (2014) Evaluating the costs and benefits of systematic data acquisition for conservation assessments. *Ecography*, DOI: 10.1111/ecog.00792.

Hirzel, A. & Guisan, A. (2002) Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling*, **157**, 331–341.

Hirzel, A.H., Helfer, V. & Metral, F. (2001) Assessing habitat suitability models with a virtual species. *Ecological Modelling*, **145**, 111–121.

Jiménez-Valverde, A. (2012) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography,* **21**, 498–507.

Kirkpatrick, J.B. (1983) An iterative method for establishing priorities for the selection of nature reserves: an example from Tasmania. *Biological Conservation*, **25**, 127–134.

Knight, A.T., Smith, R.J., Cowling, R.M., Desmet, P.G., Faith, D.P., Ferrier, S., Gelderblom, C.M., Grantham, H., Lombard, A.T., Maze, K., Nel, J.L., Parrish, J.D., Pence, G.Q.K., Possingham, H.P., Reyers, B., Rouget, M., Roux, D. & Wilson, K.A. (2007) Improving the key biodiversity areas approach for effective conservation planning. *BioScience*, **57**, 256–261.

Langford, W.T., Gordon, A. & Bastin, L. (2009) When do conservation planning methods deliver? Quantifying the consequences of uncertainty. *Ecological Informatics*, **4**, 123–135.

Leathwick, J.R., Rowe, D., Richardson, J., Elith, J. & Hastie, T. (2005) Using multivariate adaptive regression splines to predict the distribution of New Zealand´s freshwater diadromous fish. *Freshwater Biology*, **50**, 2034–2052.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Loiselle, B.A., Howell, C.A., Gaham, C.H., Goerck, J.M., Brooks, T., Smith, K.G. & Williams P. (2003) Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology*, **17**, 1591–1600.

MacLeod, C.D., Mandleberg, L., Schweder, C., Bannon, S.M. & Pierce, G.J. (2008) A comparison of approaches for modelling the occurrence of marine animals. *Hydrobiologia*, **612**, 21–32.

Maidment, D.R. (2002) *Arc Hydro: GIS for water resources*. ESRI Press, Redlands, CA.

De Ornellas, P., Milner-Gulland, E.J. & Nicholson, E. (2011) The impact of data realities on conservation planning. *Biological Conservation*, **144**, 1980–1988.

Phillips, S.J., Anderson, R.B. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Possingham, H.P., Ball, I.R. & Andelman, S. (2000) Mathematical methods for identifying representative reserve networks. *Quantitative Methods for Conservation Biology* (ed. by S. Ferson and M. Burgman), pp. 291–305. Springer-Verlag, New York.

Possingham, H.P., Grantham, H.& Rondinini, C. (2007) How can you conserve species that haven't been found? *Journal of Biogeography*, **34**, 758–759.

Pusey, B., Kennard, M., Burrows, D., Perna, C., Kyne, P., Cook, B. & Hughes, J. (2011) Freshwater fish. *Aquatic biodiversity in northern Australia: patterns, threats and future* (ed. by B.J. Pusey), pp 71–92. Charles Darwin University Press, Darwin.

Rodrigues, A.S.L., Gaston, K.J. & Gregory, R.D. (2000) Using presence–absence data to establish reserve selection procedures that are robust to temporal species turnover. *Proceedings of the Royal Society B: Biological Sciences*, **267**, 897–902.

Rodrigues, A.S.L. (2011) Improving coarse species distribution data for conservation planning in biodiversity-rich, data-poor, regions: no easy shortcuts. *Animal Conservation*, **14**, 108-110.

Rondinini, C., Wilson, K., Boitani, L., Grantham, H. & Possingham, H.P. (2006) Tradeoffs of different types of species occurrence data for use on systematic conservation planning. *Ecology Letters*, **9**, 1136–1145.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D. & Kadmon, R. (2007) A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, **13**, 397–405.

Tulloch, A.I.T., Mustin, K., Possingham, H.P., Szabo, J.K. & Wilson, K.A. (2012) To boldly go where no volunteer has gone before: predicting volunteer activity to prioritize surveys at the landscape scale. *Diversity and Distributions*, **19**, 465–480.

Vaughan, P. & Ormerod, S.J. (2003) Modelling the distribution of organisms for conservation: optimising the collection of field data for model development. *Conservation Biology*, **17**, 1601–1611.

Wessels, K.J., Van Jaarsveld, A.S., Grimbeek, J.D. & Van der Linde, M.J. (1998) An evaluation of the gradsect biological survey method. *Biodiversity and Conservation*, **7**, 1093–1121.

Wilson, K.A., Westphal, M.I., Possingham, H.P. & Elith, J. (2005) Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation*, **122**, 99–112.

SUPPORTING INFORMATION

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

Appendix S1 The spatial extent of the northern Australia study area and distribution of sampling records.

Appendix S2 A list of the 80 freshwater fish species included in the study.

Appendix S3 A summary of the environmental predictors used.

BIOSKETCHES

Virgilio Hermoso's research focuses on the application of systematic conservation planning to freshwater ecosystems, especially on how to address the peculiarities of the ecosystem processes of these systems to enhance the adequacy of conservation recommendations. He is also interested in studying the threats to the conservation of freshwater biodiversity, especially on the interactive effects of habitat degradation and introduced species, in order to inform conservation decisions better.

Mark Kennard's main research interests include the ecology and biogeography of freshwater fish, freshwater biodiversity and conservation planning, hydro-ecology and environmental flow management, and freshwater biomonitoring and bioassessment.

Simon Linke is a senior research fellow at the Australian Rivers Institute, Griffith University, Brisbane. After an early career in freshwater bioassessment and species distribution modelling, he is now primarily working in freshwater conservation planning and on generalized approaches to optimal allocation of resources to protect aquatic biodiversity. The latter includes planning for ecological restoration and environmental water allocations.

Author contributions: V.H., M.J.K. and S.L. conceived the ideas; M.J.K. and V.H. collected the data; V.H. analysed the data; V.H. led the writing helped by M.J.K. and S.L.

Editor: Richard Pearson

FIGURE CAPTIONS

Figure 1 A flow diagram of the analyses carried out to assess the risks and opportunities of presence-only data for conservation planning. MARS, multivariate adaptive regression splines, for 80 freshwater fish species in northern Australia.

Figure 2 Indicators of modelling performance for 80 freshwater fish species in northern Australia, (a) area under the receiver operating characteristic curve (AUC) and (b) proportion of deviance explained, for different data-availability scenarios for an assessment of the risks and opportunities of presence-only data for conservation planning.

Figure 3 Measures of (a) omission and (b) commission error rates (average ± SE across all species modelled) for conservation planning solutions using different data-availability scenarios for 80 freshwater fish species in northern Australia.

Figure 4 The (a) effectiveness (the proportion of species that actually achieve the target, from a set of 80 freshwater fish species) and (b) relative efficiency (the ratio between the number of planning units required to achieve targets under different data-availability constraints and the number of planning units required when using the best available data) of conservation planning solutions (average ± SE across all species modelled) for different data-availability scenarios in northern Australia. PU, planning units.

Figure 5 Measures of (a) data acquisition (number of sites × average cost/site) and (b) the return on the investment (species representation/cost of data acquisition) in data acquisition for conservation planning using different data-availability scenarios (average ± SE across all species modelled) in northern Australia. AU$, Australian dollars.

Figure 1

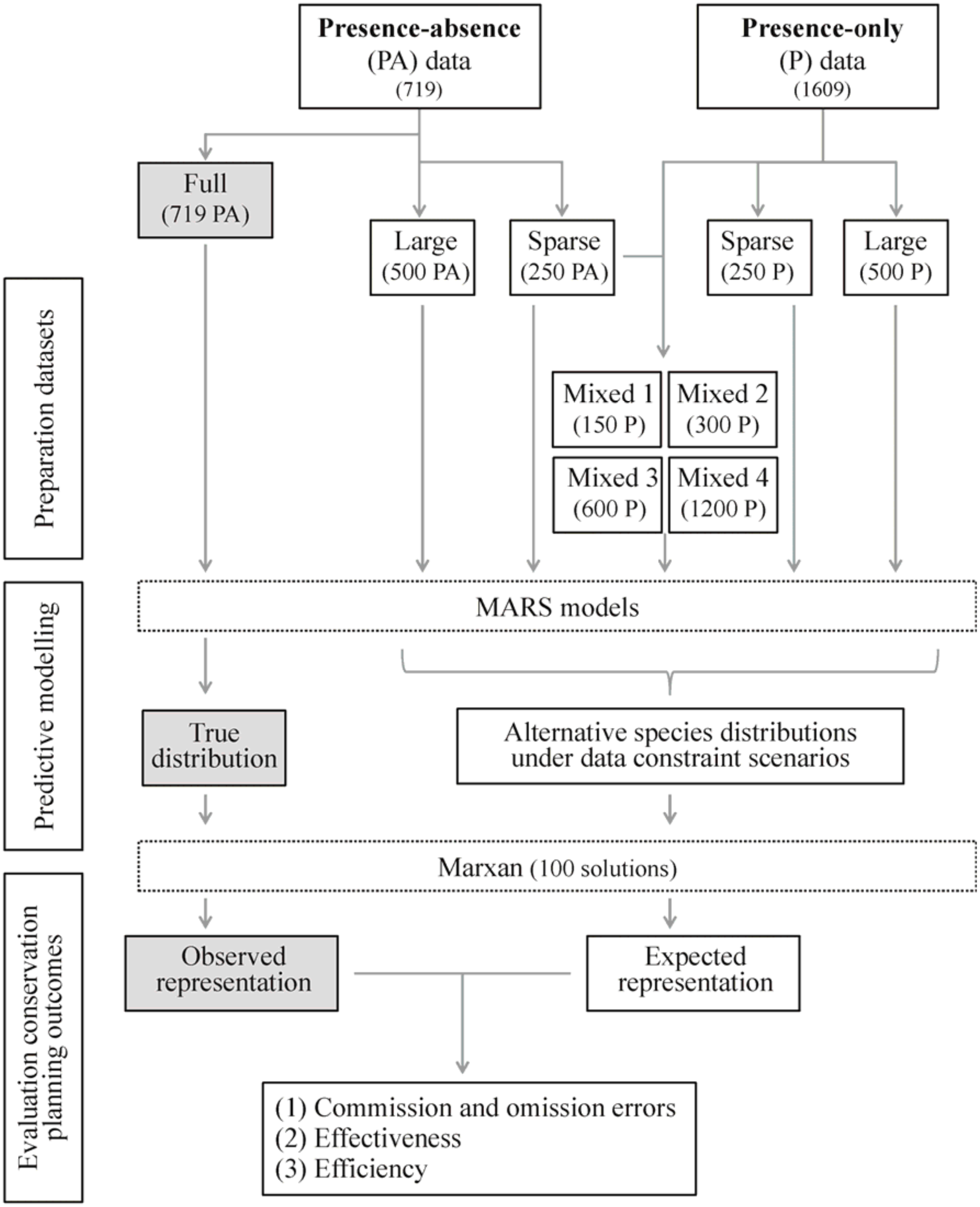
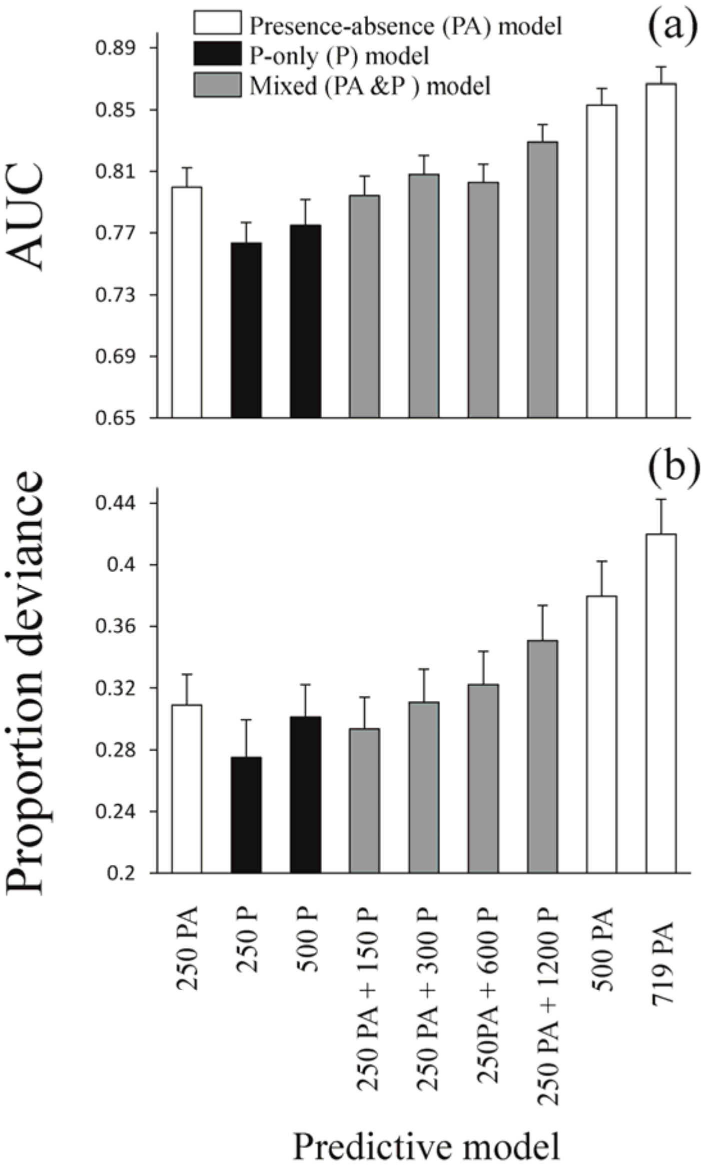
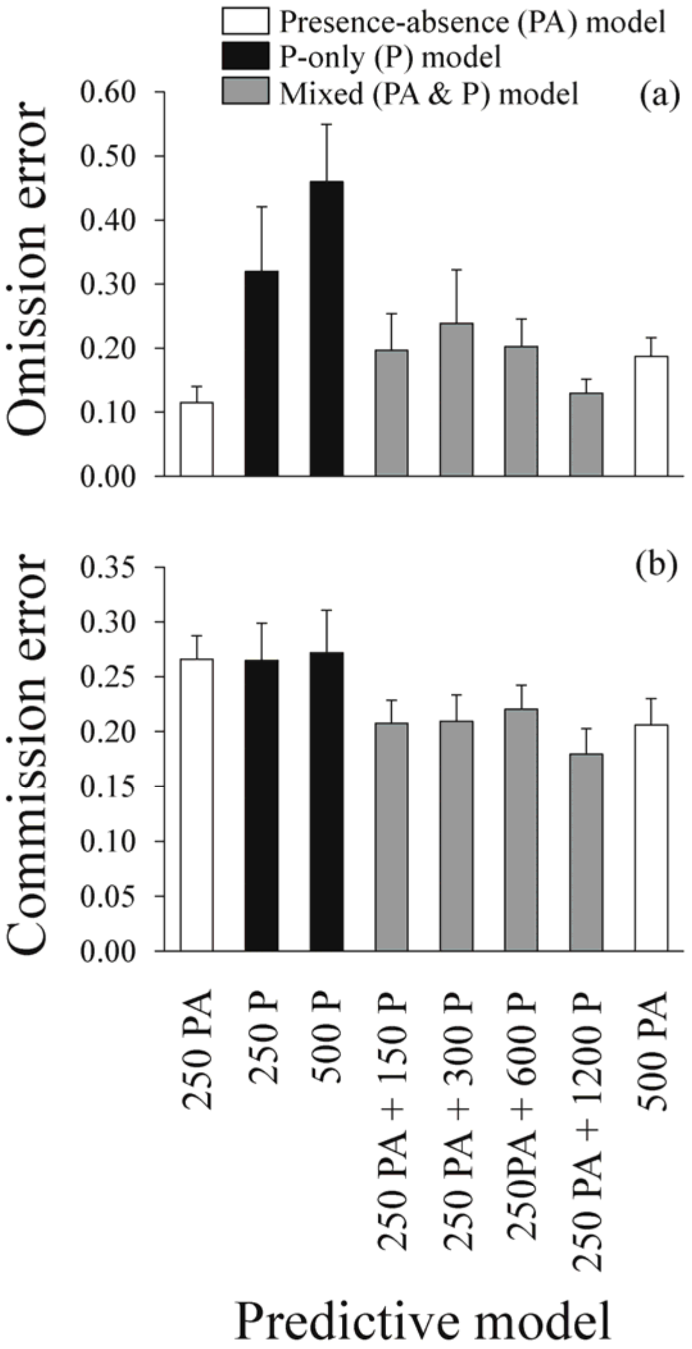

Presence-absence
(PA) data
(719)
Presence-only
(P) data
(1609)
Full
(719 PA)
Large
(500 PA)
Sparse
(250 PA)
Sparse
(250 P)
Large
(500 P)
Mixed 1
(150 P)
Mixed 2
(300 P)
Mixed 3
(600 P)
Mixed 4
(1200 P)
Preparation datasets
MARS models
Predictive modelling
True
distribution
Alternative species distributions
under data constraint scenarios
Marxan (100 solutions)
Evaluation conservation
planning outcomes
Observed
representation
Expected
representation
(1) Commission and omission errors
(2) Effectiveness
(3) Efficiency

Figure 2



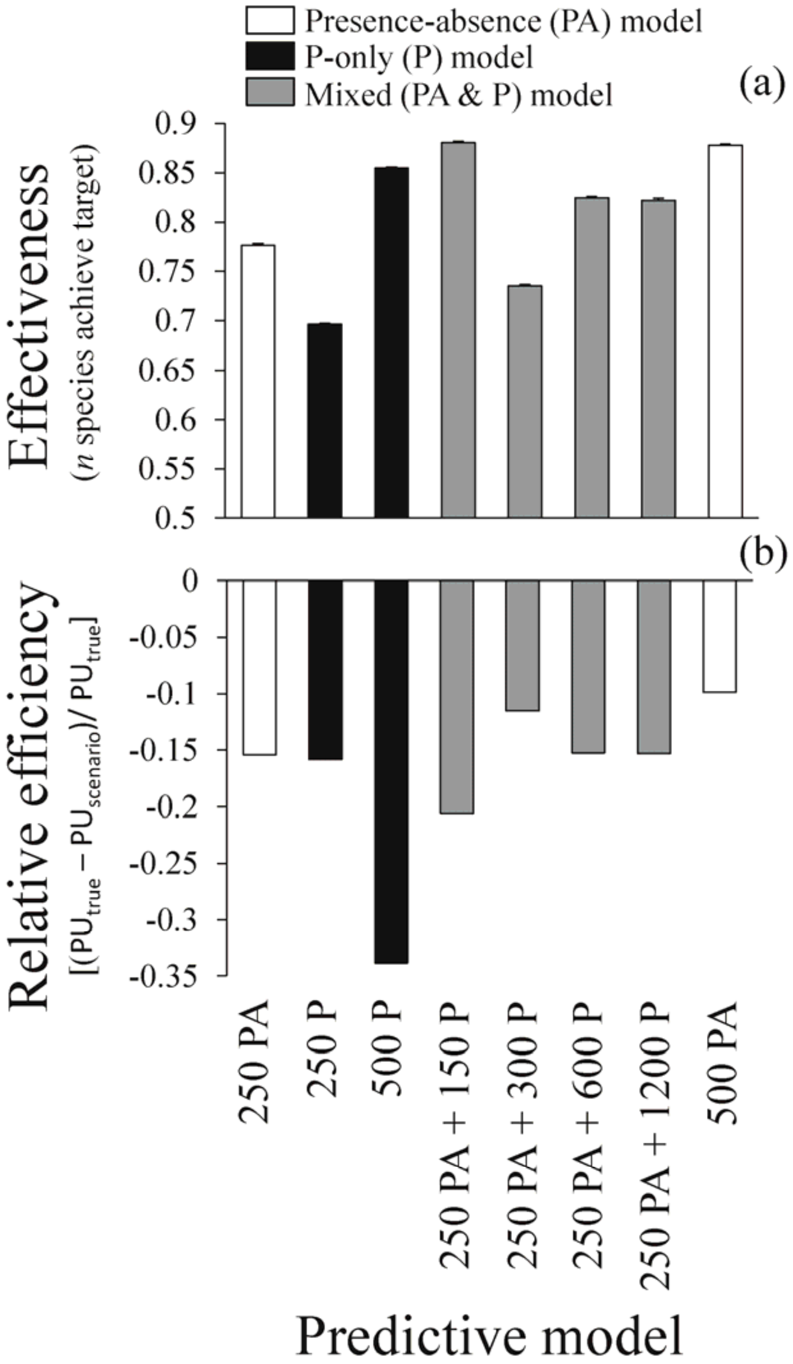
□ Presence-absence (PA) model
■ P-only (P) model
▨ Mixed (PA &P) model
(a)
AUC
0.89
0.85
0.81
0.77
0.73
0.69
0.65
(b)
Proportion deviance
0.44
0.4
0.36
0.32
0.28
0.24
0.2
250 PA
250 P
500 P
250 PA + 150 P
250 PA + 300 P
250PA + 600 P
250 PA + 1200 P
500 PA
719 PA
Predictive model

Figure 3



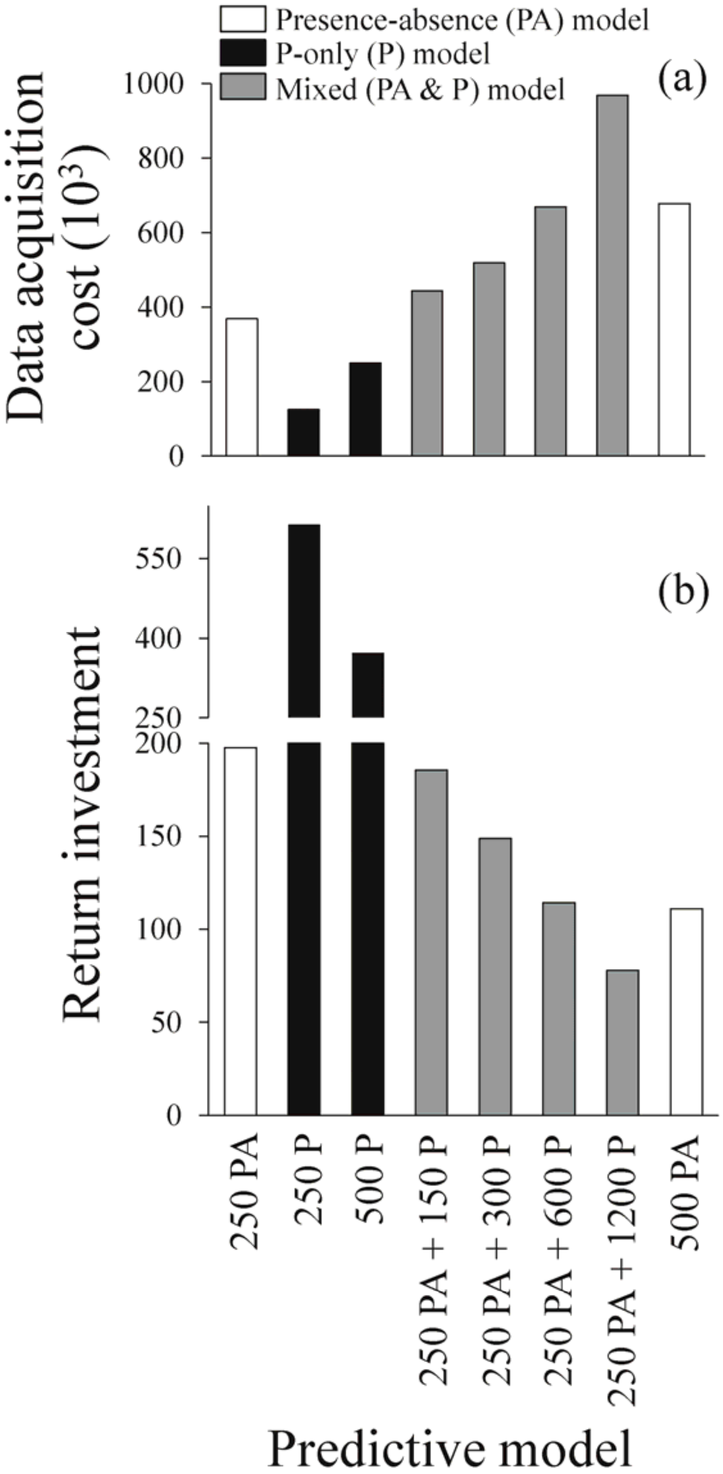
Presence-absence (PA) model
P-only (P) model
Mixed (PA & P) model
(a)
Omission error
0.60
0.50
0.40
0.30
0.20
0.10
0.00
(b)
Commission error
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
250 PA
250 P
500 P
250 PA + 150 P
250 PA + 300 P
250PA + 600 P
250 PA + 1200 P
500 PA
Predictive model

Figure 4



Presence-absence (PA) model
P-only (P) model
Mixed (PA & P) model
(a)
Effectiveness
(*n* species achieve target)
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
(b)
Relative efficiency
[(PU_true – PU_scenario)/ PU_true]
0
-0.05
-0.1
-0.15
-0.2
-0.25
-0.3
-0.35
250 PA
250 P
500 P
250 PA + 150 P
250 PA + 300 P
250 PA + 600 P
250 PA + 1200 P
500 PA
Predictive model

Figure 5


Presence-absence (PA) model
P-only (P) model
Mixed (PA & P) model
(a)
Data acquisition cost (10^3)
1000
800
600
400
200
0
(b)
Return investment
550
400
250
200
150
100
50
0
250 PA
250 P
500 P
250 PA + 150 P
250 PA + 300 P
250 PA + 600 P
250 PA + 1200 P
500 PA
Predictive model