Published in: ***Ecological Modelling*****, 2011; 222(15):2606-2614**

A novel method for mapping reefs and subtidal rocky habitats using artificial neural networks

Michael J. Watts[1]*, Yuxiao Li[1,2], Bayden Russell[2], Camille Mellin[1,3], Sean D. Connell[2], and Damien A. Fordham[1]

[1]The Environment Institute and School of Earth & Environmental Sciences, University of Adelaide, South Australia 5005, Australia

[2]Southern Seas Ecology Laboratories, School of Earth & Environmental Sciences, University of Adelaide, South Australia 5005, Australia

[3]Australian Institute of Marine Science, PMB No.3, Townsville MC, Townsville, Queensland 4810, Australia

*corresponding author, e: mjwatts@ieee.org

Abstract

Reefs and subtidal rocky habitats are sites of high biodiversity and productivity which harbour commercially important fish and invertebrate species. Although the conservation management of reef associated species has been informed using species distribution models (SDM) and community based approaches, to date their use has been constrained to specific regions where the locality and spatial extent of reefs is well known. Much of the world's subtidal habitats remain either undiscovered or unmapped, including coasts of intense human use. Consequently, to facilitate a stronger understanding of species-environmental relationships there is an urgent need for a cost and time effective standard method to map reefs at fine spatial resolutions across broad geographical extents. We used bathymetric data (~ 250 m resolution) to calculate the local slope and curvature of the seabed. We then constructed artificial neural networks (ANNs) to forecast the probability of reef occurrence within grid cells as a function of bathymetric and slope variables. Testing over an independent data set not used in training showed that ANNs were able to accurately predict the location of reefs for 86% of all grid cells (Kappa = 0.63) without over fitting. The ANN with greatest support, combining bathymetric values of the target grid cell with the slope of adjacent grid cells, was used to map inshore reef locations around the Southern Australian coastline (~ 250 m resolution). Broadly, our results show that reefs are identifiable from coarse-scale bathymetry data of the seabed. We expect that our research technique will strengthen systematic conservation planning tools in many regions of the world, by enabling identification of rocky substrate and mapping in localities that remain poorly surveyed due to logistics or monetary constraints.

Introduction

Some of the most diverse marine ecosystems are founded on subtidal rock or corals that fringe the world's coasts or occur as isolated reefs. Such subtidal habitats are generally known as 'subtidal rocky habitats', 'rocky reefs' or simply 'reefs' (hereafter referred to as reefs). Identifying the presence and extent of reefs is fundamental if we are to quantify their contribution to biological and socio-economic productivity through fisheries production, biological diversity and economic value in marine ecosystems (Connell and Gillanders, 2007).

Species distribution models (SDM; see review by Guisan and Thuiller (2002)) have been used to predict the distribution of some marine and reef biota (Robinson et al., 2010), informing conservation management (Mellin et al., 2010a). These models have for example investigated the environmental and spatial predictors of the diversity and abundance of coral reef fish (Mellin et al., 2010a); and have been used to map habitat suitability and range extents of marine invertebrates living in reef environments, for example, Galparsoro (2009).

Unfortunately, the use of SDMs for mapping the habitat suitability and range extent of reef species, and marine species in general, is often limited by a dearth of spatial data, including the location of suitable habitat. While fine-scale, remotely sensed maps of reef distributions are readily available for iconic and well-studied regions (e.g. Great Barrier Reef, Australia; U.S. Virgin Islands), in most parts of the world the location and extent of reefs is unknown, and therefore, the fundamental basis for considering their biological and economic importance is missing. Whilst the acoustic classification of habitats provides information to scales of 0.1 m (Cochrane and Lafferty, 2002), the cost and time involved in acquiring such information with side-scan sonar across large tracts of coast (e.g. > 2000 kilometres of coast in Southern Australia) hampers even the most basic exploration for such habitats. We developed a method based on artificial neural networks (ANNs) that used coarse-scale bathymetric data to predict the location and extent of reefs off the southern coast of Australia.

Artificial neural networks (ANNs) are a useful technique for modelling problems that involve complex but unknown processes (Crick, 1989; Haykin, 1994; Reed and Marks, 1999; Tarassenko, 1998). They have many advantages over statistically based techniques (Kasabov, 1996) because they can learn from existing data and therefore do not require an *a priori* model. If over fitting is avoided ANN can also generalise well, in other words, they can accurately classify data they have not been trained on. Additionally, ANNs can learn from noisy data (Kasabov, 1996) and model systems that involve multiple dependent variables and complex non-linear relationships between variables and outcomes. In ecological studies where ANNs have been compared to traditional statistical models, the ANNs have consistently out-performed the statistical models with respect to prediction accuracy (Brosse et al., 1999a; Ibarra et al., 2003; Jeong et al., 2006; Laë et al., 1999; Lek et al., 1996; Manel et al., 1999; Mastrorillo et al., 1997; Soltic et al., 2004; Wagner et al., 2000; Wagner et al., 2006).

While ANNs have been previously applied to classifying reefs from video images (Marcos et al., 2005) and classifying sediments on the seafloor (Zhou and Chen, 2005), we hereby provide the first evidence that ANNs can be used to identify the presence of reefs from coarse-scale bathymetric data. As a first attempt at addressing this problem with ANN, the goal of this paper was not to exhaustively test the myriad ANN architectures and training algorithms available (Reed and Marks, 1999) to get the absolute best model possible. Rather, the goal was to determine whether ANN are applicable to the generic problem of detecting reefs.

We were able to create a model of sufficient accuracy to provide the basis for modelling the spatial abundance patterns of two commercially significant Abalone species (*Haliotis rubra* and *H. laevigata*) inhabiting inshore rocky reefs of Southern Australia (Mellin et al., 2010b).

Method

Data

We present a map of the study area, from which we sourced the data for this study, in Figure 1. Two sets of data were combined for this study: (1) bathymetric measurements at a 250 m resolution, with a depth precision of six metres (Geoscience Australia, 2009); (2) point sample data that specified the location of observed reefs around South Australia at a bathymetric depth of less than 30 m. These sample points were acquired by visual surveys. There were 121 sample points that corresponded to reefs and 56 sample points that corresponded to non-reef seafloor. Also included were 297 randomly selected points that were known to be on land: the purpose of these samples were so that the ANNs could learn to distinguish between reefs that reached or breached the waterline and on-land features that were present in the on-land coastal buffer region. A second set of sample points (n=317, presence=126, absence=191), from a different survey, was the validation data set, or independent test set which used to test the generalisation accuracy of the ANNs. 147 points were randomly removed from the on-land data set and added to the validation data set, as this data set did not include any on-land survey points. Two of the sample points in the training set were unusable, as they appeared in the same cells as other sample points and therefore were redundant. There were thus a total of 325 vectors in the training set and 464 vectors in the validation set. Additional out-of-area validation data, that is, data from a region other than that used to train the ANNs, was sourced for the southern coastline of Victoria. This data set had 222 presences but no absences were available. A schematic of the way in which these data sets were combined is presented in Figure 2.

Data Preparation

We used ArcGIS v9.2 to calculate the slope (°) and curvature (unit-less) of the seabed from the bathymetric data. We excluded areas known to be on-land, although a two-cell (i.e. 500 m) landwards buffer was included. This was because the spatial resolution of the bathymetric data was such that strictly following the shoreline as a cut-off would have excluded known inshore reefs. Seabed areas that were deeper than 30 m, were also excluded, as there were no reef samples taken from deeper than 30 m.

A sliding window method was used to extract the slope data of grid cells surrounding each cell of the bathymetric grid, thus defining the input vectors for the artificial neural network (ANN). The sliding window conversion involved moving a sliding window of the specified size over the source matrix, as shown by the hypothetical example illustrated in Figure 3. Here, a three-by-three window starts centred on cell ‘g’ (that is, the ANN will predict the presence of a reef in cell ‘g’). The first vector produced is therefore composed of the contents of cells a, b, c, f, g, h, k, l and m. The second vector is produced by sliding the window one cell to the right so that the window is centred on cell 'h' (that is, the ANN will predict the presence of a reef in cell 'h'). This vector is therefore composed of the contents of cells b, c, d, g, h, i, l, m and n. As this method only predicts for the centre cell of the window, the cells on the periphery of the matrix (a, b, c, d, e, f, j, k, o, p, t, u, v, w, x and y) do not have corresponding predictions of reef presence.

A window was not included in the final data set if it contained any missing data, that is, if part or the entire window were on land. The purpose of this process was to provide the context of the target cell (the middle of the sliding window) to the ANN. That is, rather than classifying from the value of the target cell, the classification decision was made from the cell and its context. The central assumption in this work is that reefs exist in similar contexts in the seabed, that is, the area of seabed surrounding a reef has similar characteristics to the area surrounding other reefs. We utilised a window size of 5 x 5 matrix elements to ensure that sufficient geomorphological context was presented to the ANN. There were 8 004 860 vectors extracted for this window.

We assigned output values to the vectors according to whether the target cell contained a point known to be either a reef or not a reef. Vectors that did not have a corresponding reef sample point (that is, where it was not known from survey data whether or not there was a reef in the corresponding target cell) were not included in the ANN training sets, but were retained for later use to generate the final map as described below in the subsection "ANN Model Application".

Bathymetric and curvature data were linearly rescaled according to the following formula:

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x_n is the rescaled value of x, x_{max} is the maximum value of variable x, and x_{min} is the minimum value of x. The rescaling used the maximum and minimum possible values for these variables: for example, a depth of zero is the absolute minimum possible for bathymetric data, while the maximum was the 30 m cut-off depth used in the seabed data processing. The slope data was not linearly rescaled, because values were either zero or greater than 89. This data was rescaled by the simple process of subtracting 89 from any value that was not zero.

ANN Algorithms

The ANNs we used were three-neuron layer multi-layer perceptrons (MLP). These networks consisted of an input neuron layer, a single hidden neuron layer, and an output neuron layer. Each neuron layer was fully connected, that is, every neuron in a layer was connected to every neuron in the preceding neuron layer. The training algorithm used was unmodified backpropagation of errors with momentum (Rumelhart et al., 1986). This algorithm has been widely used in applications in ecology (Brosse et al., 2007; Brosse et al., 1999b; Bryant and Shreeve, 2002; Cocu et al., 2005; Dedecker et al., 2004; Dimopoulos et al., 1999; Fedor et al., 2008; Francl, 2004; Gutiérrez-Estrada and Bilton, 2010; Joy and Death, 2002, 2004; Paul and Munkvold, 2005) and elsewhere (Bourland and Wellekens, 1987; Chandonia and

Karplus, 1995; Diederichs et al., 1998; Franzini, 1988; Haskey and Datta, 1998; Lippmann, 1989; Qian and Sejnowski, 1988; Rost, 1996) and is in many ways the *de facto* standard for training MLP. An advantage of MLP is that their outputs can be interpreted as probabilities (Kasabov, 1996).

ANN Training and Evaluation

We used ten-fold cross-validation to select the topology and training parameters of the MLP. That is, the training data set was randomly divided into ten equally sized subsets. One testing subset was held out and a MLP with randomly initialised connection weights trained over the remaining nine (the training fold). The trained MLP was recalled and its accuracy assessed over the training fold (giving the training accuracy) and the held-out testing set (the testing accuracy). The process was repeated ten times, with a different subset held out as the testing set each time. This gave an estimate of the generalisation accuracy of the MLP over the entire data set, and is not only widely recommended in the ANN literature (Flexer, 1996; Prechelt, 1996; Zhang, 2007) but has also been previously applied to ecological applications of ANN (Joy and Death, 2002, 2004).

The combinations of input variables we investigated were: slope of all cells in the window and bathymetric value of the target cell; curvature of all cells in the window and bathymetric value of the target cell; bathymetric values of the target cell with curvature and slope values of all cells in the window; bathymetric values of all cells in the window with curvature and slope of all cells in the window.

We investigated a number of different MLP topologies (number of hidden-layer neurons) and training parameters (learning rate, momentum, and training epochs) which were selected heuristically using expert knowledge. The mean generalisation accuracy (that is, the accuracy over the cross-validation testing subsets that the MLP were not trained on) of the MLP was used to select the combination of input variables and MLP parameters that gave the best accuracy.

Accuracy was measured firstly using Cohen's Kappa statistic (Cohen, 1960). A kappa of less than 0.2 is considered poor accuracy, 0.2 to 0.4 fair, 0.4 to 0.6 moderate, 0.6 to 0.8 good and over 0.8 very good, with 1 being perfect accuracy. Kappa was used because it is a simple and well-known statistic (Manel et al., 2001) that is not biased by different proportions of presences or absences, and gives results that are qualitatively similar to more complex measures such as area under the receiver operating characteristic curve (Elith et al., 2006; Graham et al., 2008). The formula for kappa is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where κ is the Kappa statistic, $\Pr(a)$ is the observed agreement between the predicted and actual data, and $\Pr(e)$ is the probability of chance agreement between the predicted and actual data.

The second accuracy measure used was percentages of data sets correctly classified. The first percentage measured was the percentage of examples out of each data set that were correctly classified, which while easily interpreted can also be biased by uneven proportions of classes in the data set, that is, an uneven number of presence and absence examples. To address this, the true positive and true negative percentage accuracies were also measured. The true positive percentage is the percentage of positive examples that are correctly classified as positive, while the true negative percentage is the percentage of negative examples that are correctly classified as negative. While the final output of the MLP was the probability of each cell containing a reef, a threshold value of 0.5 was applied when calculating accuracies. That is, an output below 0.5 was considered to be negative, while an output above 0.5 was considered to be positive. This threshold is reasonable given the interpretation of MLP outputs as probabilities. Also, the MLP output values for this problem (results not shown) consistently followed a U-shaped distribution, with most outputs close to either zero or unity.

The training parameters that yielded the best accuracies are presented in Table 1. In this work, the 'best accuracies' means that the networks had the best balance between learning the training data and being able to generalise to unseen data.

ANN Model Application

The parameters that gave the best cross-validated results were used to train MLP over the entire training set. As the random initialisation of connection weights in MLP can cause variations in the final accuracy of the trained networks, 100 MLP were trained. Accuracies were assessed over both the training data set (training accuracy) and the independent validation data set (giving the validation accuracy), that was sourced from a different survey as the training set. The final predictions of reef presence over the entire study region were generated using the MLP that had the highest accuracy over the validation data set (that is, that generalised the best). This gave the probability that each cell contained a reef. As the validation data set was not used to select the topology or training parameters of the MLP, it remained independent. The process of selecting the optimal training parameters via ten-fold cross-validation, then further training over the complete training set and selecting the optimal MLP by validation error, is presented schematically in Figure 4.

ANN Contribution Analysis

Contribution analysis is a way of determining the relative importance of each input variable of the MLP with respect to the output. The method of contribution analysis used here was that of Olden and Jackson, (2002) which has been previously shown to be less biased in its assessment than other methods (Olden et al., 2004). This method yields unit-less values that show the relative positive or negative contribution of an input, where a positively contributing variable increases the activation of the output as the input variable increases and a negatively contributing variable decreases the activation as the variable increases. In the context of this application, a high value of a positively contributing variable is interpreted to be associated with the presence of a reef, while a high value of a negatively contributing variable is associated with the absence of a reef.

Results

The parameters that yielded the best mean cross-validated accuracies are presented in Table 1. These parameters yielded MLP that had the highest mean accuracies over the testing data sets, that is, they produced MLP that generalised to new data the best. The accuracies of the corresponding MLP are presented in Table 2. The input variables that produced the highest test accuracies were the bathymetric value of the target cell, a 5 x 5 window of seafloor curvature and a 5 x 5 window of seafloor slope. For each combination of input variables, the cross-validated training accuracies were significantly higher (two-tailed *t*-test, p=0.001) than the cross-validated testing accuracies. This implies that all networks over-trained to some extent; although the over-training was not severe in any case as the mean test kappas were all moderate to good. The true-positive and true negative percentage accuracies were similar, however, for both training and testing across all input data sets, which indicates that the cross validation data set was not badly unbalanced.

The parameters that produced the best cross-validated results were used to train MLP over the entire training set for each combination of variables. There were no significant differences between the cross-validated training accuracies and the accuracies over the complete training sets (two-tailed *t*-test, p=0.001). The trained networks were assessed over the South Australian validation data set. The performance for all networks over the validation data set were poor, and in all cases the true positive percentage was low, indicating that the networks found it difficult to identify reefs in the South Australian validation data set. The exception to this was those trained with the bathymetric value of the target cell and a 5 x 5 window of slope, which was able to detect more than half of the reefs present. The best of these networks gave the validation accuracies presented in Table 4, which is a good kappa score and true positive detection rate of over 68%, with an overall accuracy of 85.99%. As this network had the highest validation accuracy, that is, it classified unseen data the most accurately, it was used to create the final prediction map. The map generated by this MLP is displayed in Figure 5. Assessing this network over the

Victorian validation set gave a prediction accuracy of 52.25%, that is, the network correctly classified 52.25% of the reef cells.

The results of the MLP input contribution analysis of this network are presented in Figure 6. This shows that a high value of slope in the cells neighbouring the target cell contribute strongly to a prediction of a reef being present, and that a high bathymetric value contributes strongly to a prediction of a reef being absent.

Discussion

The need for a cost-effective standard method to map reefs across broad scales is likely to become an issue of increasing urgency as the world's coasts continue to bear the burden of the ecological costs of increasing human activity. South Australia presented one such locality in which this study sought to provide leadership in raising the challenges and solutions to what to date has been an intractable problem. In doing so we show that ANN provide a cost effective method for broadly mapping the probability of reef occurrence as a function of bathymetric and slope variables.

There were significant differences between the cross-validated training and testing accuracies for all combinations of input variables investigated; however the mean testing kappa scores were all moderate to good, which shows that the ANN did not badly over-train. The similarity between the true negative and true positive accuracies for both training and testing indicate that the cross-validation data set was not badly unbalanced in terms of positive and negative examples. There were no significant differences between the cross-validated training accuracies and the training accuracies over the complete data set. This was expected, as the purpose of cross-validated training was to approximate the optimal training parameters. The poor performance over the South Australian validation data was the result of under-prediction of reef presences by the networks. Although the networks trained on bathymetric value, slope and curvature had the highest accuracies over the cross-validated testing sets, they exhibited poor performance at detecting reefs in the validation set. Conversely, the networks trained on bathymetric value and slope, while scoring the lowest accuracies over the cross-validated testing accuracies, achieved the highest accuracies over the validation data set.

It is likely that over-training was a major contributing factor to the under-prediction of reefs from the networks trained on a combination of bathymetric value, slope and curvature. It is well-known that a larger number of input features makes over-training of ANN more likely (Kasabov, 1996). This reinforces the importance of using an independent validation data set to verify the performance of any classifier, but especially so for data-driven models such as ANN.

Model performance over the spatially disparate Victorian validation set was lower (52.25% accuracy), indicating that a level of caution should be shown when using the MLP to extrapolate outside the region for which it was not trained. This is also likely to be a contributing reason that the performance over the South Australian validation set was slightly lower, as the points for this set also came from a slightly different area to the cross-validation training set. Validation using data sets that fall outside the area from which the model was trained are a stringent test of model performance, often resulting in a reduction in model performance (Barry and Elith, 2006).

However, it is possible that the lower performance index for Victoria could be the result of a difference in the geological context of reefs in South Australian compared to Victorian waters. To overcome problems with extrapolating outside the model region, future work should concentrate on the construction of region-specific classifiers wherever possible to account for this.

The results of the input contribution analysis show that the most important variables for the detection of reefs were the slope of the cells next to the target cell. A high slope value next to the target cell indicates the presence of a reef, while a high slope value in the target cell itself indicates the absence of a reef. This is reasonable, as a reef is likely to have a greater slope on its side than on its top. A high value of the bathymetric variable for the target cell indicated the absence of a reef, while a low value indicated reef presence. This also is reasonable, as a reef is an outcropping from the sea floor: reefs can therefore be expected to have a low bathymetric value, that is, they will not be as deep as cells without reefs. Of course, bathymetric depth alone is not enough to identify reefs, because of the range of depths at which reefs occur.

Although ANN were able to predict the location of reefs from bathymetric data (at least at depths less than 30 m) and measures of the slope of the seafloor, as data-driven methods, ANN are strongly affected by the quality of the data. There are two issues with the method used to prepare the data. The first is that the windowing technique excludes cells around the edges of the matrix, with the number of cells excluded being equal to half the window size. It also excludes cells that are less than half the window size from any area of no data, such as those that are close to areas deeper than 30 m, including those near to the continental shelf. Thus, a window size of five will miss any reef within 500 m of the continental shelf. The second issue is that the resolution of the bathymetric data was 250 m. Therefore, this will miss any reefs smaller than 250 m. The coarse resolution of the data also caused problems with the placement of reef sample points. Some grid squares had two reef samples located within them. While for some grid squares this was the same reef, for others they were different reefs. Other grid squares had reefs in the same grid square as the coastline. Such data issues will cause problems for any classification algorithm (Kasabov, 1996).

Our future work will focus on improving the accuracy of the predictions. One way of doing this would be to use ensembles of ANNs (Sharkey, 1996), which is where the predictions of several ANNs are combined to make one final prediction. In this approach, the individual ANN would be trained over particular geographic areas, and would thus be highly specialised. This has been shown to yield superior accuracies in other applications (Sharkey and Sharkey, 1997). While this would imply that the individual ANN were over-fitted to their target region, such diversity among members has been shown to be beneficial to ensembles (Brown, 2004; Minku et al., 2010). We will also investigate identifying specific types of reefs (steep, flat, etc.) based on structure and form. Whereas the work reported in this paper focussed on detecting reefs in general, the morphologies of different reef types may be different enough that specialising ANN on reef types may lead to better predictions overall. We will also investigate other ANN training algorithms, such as Levenberg-Marquardt (Masters, 1995), resilient backpropagation (Riedmiller and Braun, 1993), and evolutionary programming (Fogel et al., 1997). Finally, there are several methods of variable selection (Abarbanel, 1993; Fernando et al., 2009; Gutiérrez-Estrada and Bilton, 2010; Sharma, 2000) that can be applied to the data set before constructing the ANN, which may yield improved ANN performance by reducing the number of variables to be modelled. The prevalence of reefs and unbalanced data sets is such that problems are likely to arise in modelling (Mouton et al., 2010). Boostrapped training may help mitigate the effects of low prevalence and unbalanced training sets. Setting output thresholds using Bayesian statistics (Tarassenko, 1998) is also a possibility.

In conclusion, we have demonstrated a novel, but simple tool that may be used to uncover the location and extent of subtidal reefs and rocky habitats. While side-scan sonar is often used to establish fine scale information of habitat types, much of the world's reefs are yet to be identified at spatial extents that are sufficiently useful for ecologists and natural resource managers. This lack of fundamental information represents a critical gap in knowledge for basic predictions about the ubiquity of subtidal ecosystems and their contribution to the world's coastal ecology and economy. The knowledge gap persists because large parts of the world's coasts are inaccessible to traditional methods of mapping; i.e. SCUBA diving in seas of low visibility or high exposure to physical injury by the elements and wildlife and acoustic mapping is expensive and time consuming. Consequently, the methods developed here removes one of the largest obstacles to allowing marine biologists and resource managers to uncover the location and extent of some of the most diverse and productive marine ecosystems of the globe.

Acknowledgements

We thank J. Tanner for kindly supplying data for the non-reef habitat used to parameterise the model. The research was funded by Australian Research Council grants awarded to D.A. Fordham and S.D. Connell.

References

Abarbanel, H.D.I., 1993. The analysis of observed chaotic data in physical systems. Reviews of Modern Physics 65, 1331-1392.

Barry, S.C., Elith, J., 2006. Error and uncertainty in habitat models. Journal of Applied Ecology 43, 413-423.

Bourland, H., Wellekens, C.J., 1987. Multiplayer Perceptrons and automatic speech recognition, IEEE First Annual Conference on Neural Networks, pp. 407-416.

Brosse, S., Grossman, G.D., Lek, S., 2007. Fish assemblage patterns in the littoral zone of a European reservoir. Freshwater Biology 52, 448-458.

Brosse, S., Guegan, J.-F., Tourenq, J.-N., Lek, S., 1999a. The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. Ecological Modelling 120, 299-311.

Brosse, S., Guegan, J.F., Tourenq, J.N., Lek, S., 1999b. The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. Ecological Modelling 120, 299-311.

Brown, G., 2004. Diversity in Neural Network Ensembles, School of Computer Science. University of Birmingham, Birmingham.

Bryant, S.R., Shreeve, T.G., 2002. The use of artificial neural networks in ecological analysis: estimating microhabitat temperature. Ecological Entomology 27, 424-432.

Chandonia, J.-M., Karplus, M., 1995. Neural Networks for Secondary Structure and Structural Class Predictions. Protein Science 4, 275-285.

Cochrane, G.R., Lafferty, K.D., 2002. Use of acoustic classification of sidescan sonar data for mapping benthic habitat in the Northern Channel Islands, California. Continental Shelf Research 22, 683-690.

Cocu, N., Harrington, R., Rounsell, M.D.A., Worner, S.P., Hulle, M., 2005. Geographical location, climate and land use influences on the phenology and numbers of the aphid, Myzus persica, in Europe. Journal of Biogeography 32, 615-632.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37-46.

Connell, S.D., Gillanders, B.M., 2007. Marine Ecology. Oxford University Press, Melbourne 630 pp.

Crick, F., 1989. The recent excitement about neural networks. Nature 337, 129-132.

Dedecker, A.P., Goethals, P.L.M., Gariels, W., De Pauw, N., 2004. Optimization of Aritifical Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwlam river basin (Flanders, Belgium). Ecological Modelling 174, 161-173.

Diederichs, K., Freigang, J., Umhau, S., Zeth, K., Breed, J., 1998. Prediction by a neural network of outer membrane ß-strand protein topology. Protein Science 7, 2413-2420.

Dimopoulos, I., Chronopoulos, J., Chronopoulou, S., Lek, S., 1999. Neural network models to study the relationships between lead concentration in grasses and permanent urban decriptors in Athens city (Greece). Ecological Modelling 120, 157-165.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129-151.

Fedor, P., Malenovsky, I., Vanhara, J., Sierka, W., Havel, J., 2008. Thrips (Thysanoptera) identification using artificial neural networks. Bulletin of Entomological Research 98, 437-447.

Fernando, T.M.K.G., Maier, H.R., Dandy, G.C., 2009. Selection of input variables for data driven models: An average shifted histogram partial mututal information estimator approach. Journal of Hydrology 367, 165-176.

Flexer, A., 1996. Statistical evaluation of neural network experiments: minimum requirements and current practice, 13th European Meeting on Cybernetics and Systems Research. Austrian Society for Cybernetic Studies, pp. 1005-1008.

Fogel, D.B., Wasson, E.C., Boughton, E.M., Porto, V.W., 1997. A step toward computer-assisted mammography using evolutionary programming and neural networks. Cancer Letters 119, 93-97.

Francl, L.J., 2004. Squeezing the turnip with artificial neural nets. Phytopathology 94, 1007-1012.

Franzini, M.A., 1988. Learning to recognize spoken words: A study in connectionist speech recognition, in: Touretzky, D., Hinton, G., Sejnowski, T. (Eds.), Proceedings of the 1988 Connectionist Models Summer. Morgan Kaufmann, pp. 407-416.

Galparsoro, 2009. Predicting suitable habitat for the European lobster (*Homarus gammarus*), on the Basque continental shelf (Bay of Biscay), using Ecological-Niche Factor Analysis Ecological Modelling 220, 556-567.

Geoscience Australia, D.o.R., Energy and Tourism, Canberra, Australia, 2009. Australian Bathymetriy and Topography Grid, GeoCat # 67703 ed. Australian Government, Geoscience Australia.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Peterson, A.T., Loiselle, B.A., Group, T.N.P.S.D.W., 2008. The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology 45, 239-247.

Guisan, A., Thuiller, W., 2002. Predicting species distribution: offering more than simple habitat models. Ecological Letters 8, 993-1009.

Gutiérrez-Estrada, J.C., Bilton, D.T., 2010. A heuristic approach to predicting water beetle diversity in temporary and fluctuating water. Ecological Modelling 221, 1451-1462.

Haskey, S.J., Datta, S., 1998. A Comparative Study of OCON and MLP Architectures for Phoneme Recognition, Proceedings of ICSLP 98.

Haykin, S., 1994. Neural networks: a comprehensive foundation. MacMillan Publishing Company.

Ibarra, A.A., Gevrey, M., Park, Y.-S., Lim, P., Lek, S., 2003. Modelling the factors that influence fish guilds composition using a back-propagation network: Assessment of metrics for indices of biotic integrity. Ecological Modelling 160, 281-290.

Jeong, K.-S., Kim, D.-K., Joo, G.-J., 2006. River phytoplankton prediction model by artificial neural network: Model performance and selection of input variables to predict time-series phytoplankton proliferations in a regulated river system. Ecological Informatics 1, 235-245.

Joy, M.K., Death, R.G., 2002. Predictive modelling of freshwater fish as a biomonitoring tool in New Zealand. Freshwater Biology 47, 2261-2275.

Joy, M.K., Death, R.G., 2004. Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. Freshwater Biology 49, 1036-1052.

Kasabov, N.K., 1996. Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering. MIT Press.

Laë, R., Lek, S., Moreau, J., 1999. Predicting fish yield of African lakes using neural networks. Ecological Modelling 120, 325-335.
Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecological Modelling 90, 39-52.
Lippmann, R.P., 1989. Review of Neural Networks for Speech Recognition. Neural Computation 1, 1-38.
Manel, S., Dias, J.-M., Ormerod, S.J., 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distribution: a case study with a Himalayan river bird. Ecological Modelling 120, 337-347.
Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38, 921-931.
Marcos, M.S.A.C., Soriano, M.N., Saloma, C.A., 2005. Classification of coral reef images from underwater video using neural networks. Optics Express 13, 8766-9771.
Masters, T. (ed.) 1995. Advanced Algorithms for Neural Networks, A C++ Sourcebook. Wiley, New York.
Mastrorillo, S., S., L., Dauba, F., Belaud, A., 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. Freshwater Biology 38, 237-246.
Mellin, C., Bradshaw, C.J.A., Meekan, M.G., Caley, M.J., 2010a. Environmental and spatial predictors of species richness and abundance in coral reef fishes. Global Ecology and Biogeography 19, 212-222.
Mellin, C., Russell, B.D., Connell, S.D., B.W., B., Fordham, D.A., 2010b. Geographic range determinants of two commercially important marine molluscs. submitted to Diversity and Distributions.
Minku, L.L., White, A.P., Yao, X., 2010. The Impact of Diversity on Online Ensemble Learning in the Presence of Concept Drift. IEEE Transactions on Knowledge and Data Engineering 22, 730-782.
Mouton, A.M., De Baets, B., Goethals, P.L.M., 2010. Ecological relevance of performance criteria for species distribution models. Ecological Modelling 221, 1995-2002.
Olden, J.D., Jackson, D.A., 2002. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecological Modelling 154, 135-150.
Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling 178, 389-397.
Paul, P.A., Munkvold, G.P., 2005. Regression and artificial neural network modeling for the prediction of gray leaf spot of maize. Phytopathology 95, 388-396.
Prechelt, L., 1996. A quantitative study of experimental evaluations of neural network learning algorithms: Current research practice. Neural Networks 9, 457-462.
Qian, N., Sejnowski, T.J., 1988. Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. Journal of Molecular Biology 202, 865-884.
Reed, R.D., Marks, R.J., 1999. Neural smithing. MIT Press.
Riedmiller, M., Braun, H., 1993. A direct adaptive method for faster backpropagation learning: The RPROP algorithm, IEEE International Conference on Neural Networks. IEEE, San Francisco.
Robinson, L.M., Elith, J., Hobday, A.J., Pearson, R.G., Kendall, B.E., Possingham, H.P., Richardson, A.J., 2010. Pushing the limits in marine species distribution

modelling: lessons from the land present challenges and opportunities. Global Ecology and Biogeography.
Rost, B., 1996. PHD: Predicting One-Dimensional Protein Structure by Profile-Based Neural Networks. Methods in Enzymology 266, 525-539.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533-536.
Sharkey, A.J.C., 1996. On Combining Artificial Neural Nets. Connection Sciences 8, 299-313.
Sharkey, A.J.C., Sharkey, N.E., 1997. Combining diverse neural nets. The Knowledge Engineering Review 12, 231-247.
Sharma, A., 2000. Seasonal to interannual rainfall probabilistic forecast for improved water supply management: Part 1 - A strategy for system predictor identification. Journal of Hydrology 239, 232-239.
Soltic, S., Pang, S., Peacock, L., Worner, S., 2004. Evolving computation offers potential for estimation of pest establishment. International Journal of Computers, Systems and Signals 5, 36-43.
Tarassenko, L., 1998. A guide to neural computing applications. Wiley, London.
Wagner, R., Dapper, T., Schmidt, H.-H., 2000. The influence of environmental variables on the abundance of aquatic insects: a comparison of ordination and artificial neural networks. Hydrobiologica 422/423, 143-152.
Wagner, R., Obach, M., Werner, H., Schmidt, H.-H., 2006. Artificial neural nets and abundance prediction of aquatic insects in small streams. Ecological Informatics 1, 423-430.
Zhang, G.P., 2007. Avoiding Pitfalls in Neural Network Research. IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews 37, 3-16.
Zhou, X., Chen, Y., 2005. Seafloor Classification of Multibeam Sonar Data Using Neural Network Approach. Marine Geodesy 28, 201-220.

Inputs	Hidden Neurons	Epochs	Eta	Alpha
Bathy-1 Curva	16	14000	0.25	0.15
Bathy-1 Curva Slope	15	10000	0.3	0.1
Bathy-1 Slope	17	15000	0.15	0.15
Bathy Curva Slope	15	10000	0.3	0.1

Table 1 – training parameters for best performing MLP. 'Bathy' is a window of bathymetric values; 'Bathy-1' is the bathymetric value of the target cell; 'Curva' is the curvature of the seabed; 'Slope' is the slope of the sea bed. 'Eta' is the backpropagation learning rate parameters. 'Alpha' is the backpropagation momentum parameter.

	Bathy-1 Curva	Bathy-1 Curva Slope	Bathy-1 Slope	Bathy Curva Slope
Train K	**0.88±0.01**	**0.93±0.02**	**0.77±0.04**	**0.90±0.07**
Train overall %	94.70±0.52	96.65±0.89	89.10±1.83	95.49±2.97
Train true pos. %	85.95±1.37	95.49±1.94	92.94±2.70	91.46±8.32
Train true neg. %	99.89±0.23	97.34±0.92	86.84±2.84	97.82±0.83
Test K	**0.70±0.12**	**0.77±0.06**	**0.51±0.13**	**0.71±0.09**
Test overall %	86.77±6.16	89.54±2.91	76.91±5.48	87.37±3.46
Test true pos. %	66.67±11.60	85.45±9.57	72.90±11.48	73.73±11.56
Test true neg. %	99.00±2.11	93.47±6.18	79.08±7.70	95.84±3.27
Complete K	**0.87±0.03**	**0.91±0.04**	**0.72±0.09**	**0.92±0.03**
Complete overall %	94.03±1.46	95.76±1.69	82.3±14.03	96.32±1.58
Complete true pos. %	85.48±3.81	93.60±3.42	89.65±5.72	94.66±3.73
Complete true neg. %	99.10±2.97	97.04±2.19	86.9±3.55	97.31±1.74
Validate K	**0.0±0.0**	**0.17±0.12**	**0.46±0.10**	**0.0±0.02**
Validate overall %	72.84±0.0	73.8±3.55	79.91±3.18	72.08±1.27
Validate true pos. %	0.0±0.0	19.59±9.70	53.25±10.50	2.17±1.17
Validate true neg. %	100.0±0.0	94.04±3.88	89.85±2.89	98.15±1.86

Table 2 – accuracies of MLP trained on 5 x 5 windows. Column labels describe the input variables of the networks: 'Bathy' is a window of bathymetric values; 'Bathy-1' is the bathymetric value of the target cell; 'Curva' is the curvature of the seabed. Rows labelled 'Train' are the accuracies over the training data sets. Rows labelled 'Test' are accuracies over the test sets, that is, the data sets that the MLP have not been trained on. Rows labelled 'Complete' are accuracies over the complete, combined training and testing set, that is, the training accuracies of the MLP over which the validation accuracies were assessed. Rows labelled 'Validate' are the accuracies over the independent validation data set. 'K' denotes Cohen's kappa and '%'

Kappa	0.63
Overall %	85.99
True Positive %	68.25
True Negative %	92.60

Table 3 – accuracies over independent validation set of MLP used to generate final prediction maps. Row and column heading are as for Tables 2 and 3.

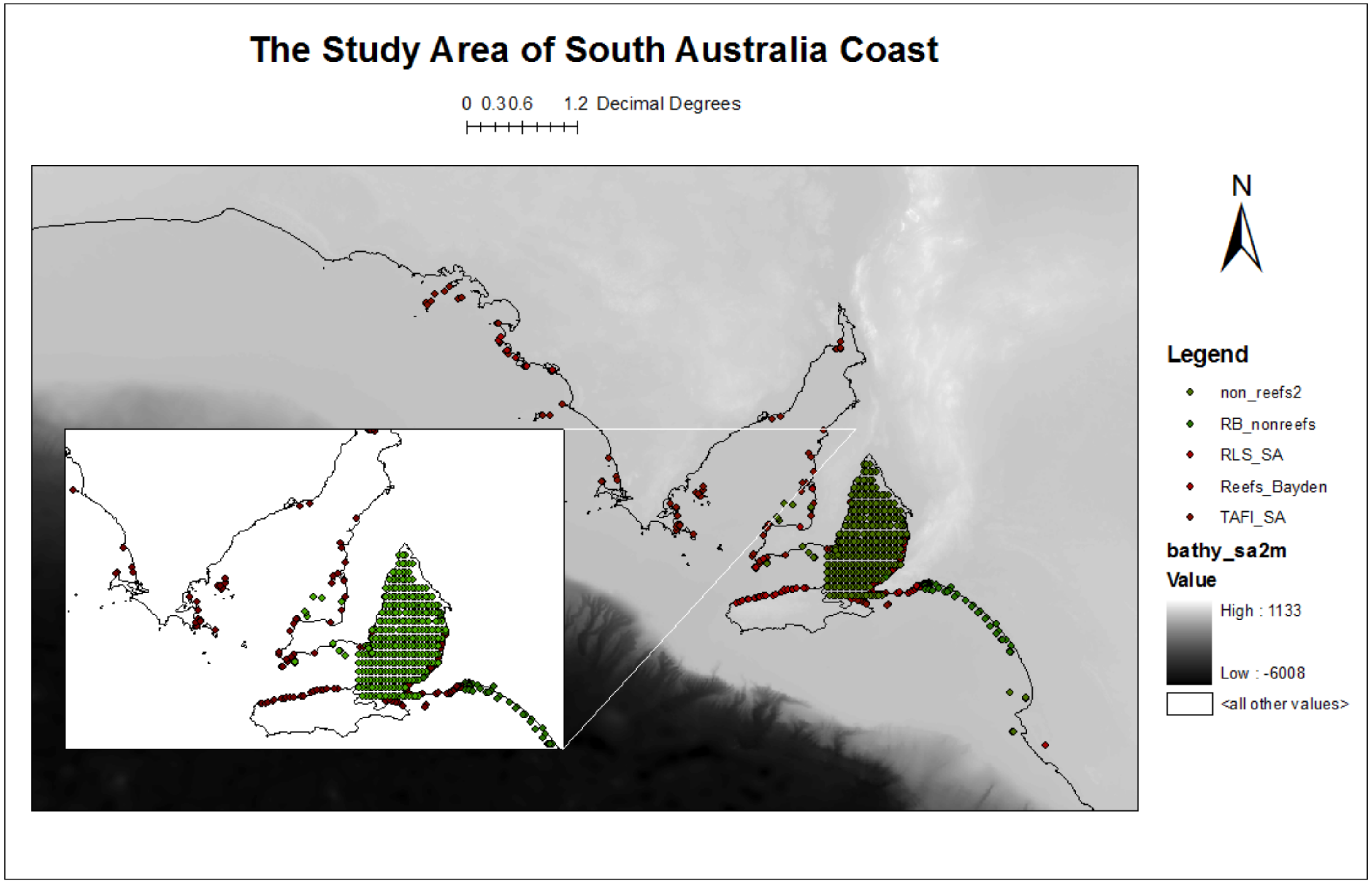

Figure 1 –map of the study area, the South Australian coastline and Kangaroo Island. The insert is a zoomed-in view of the central study area within the Spencer Gulf and the Gulf St Vincent. The study area goes from latitude -38 to -31 degrees, and longitude 129 to 141 degrees.

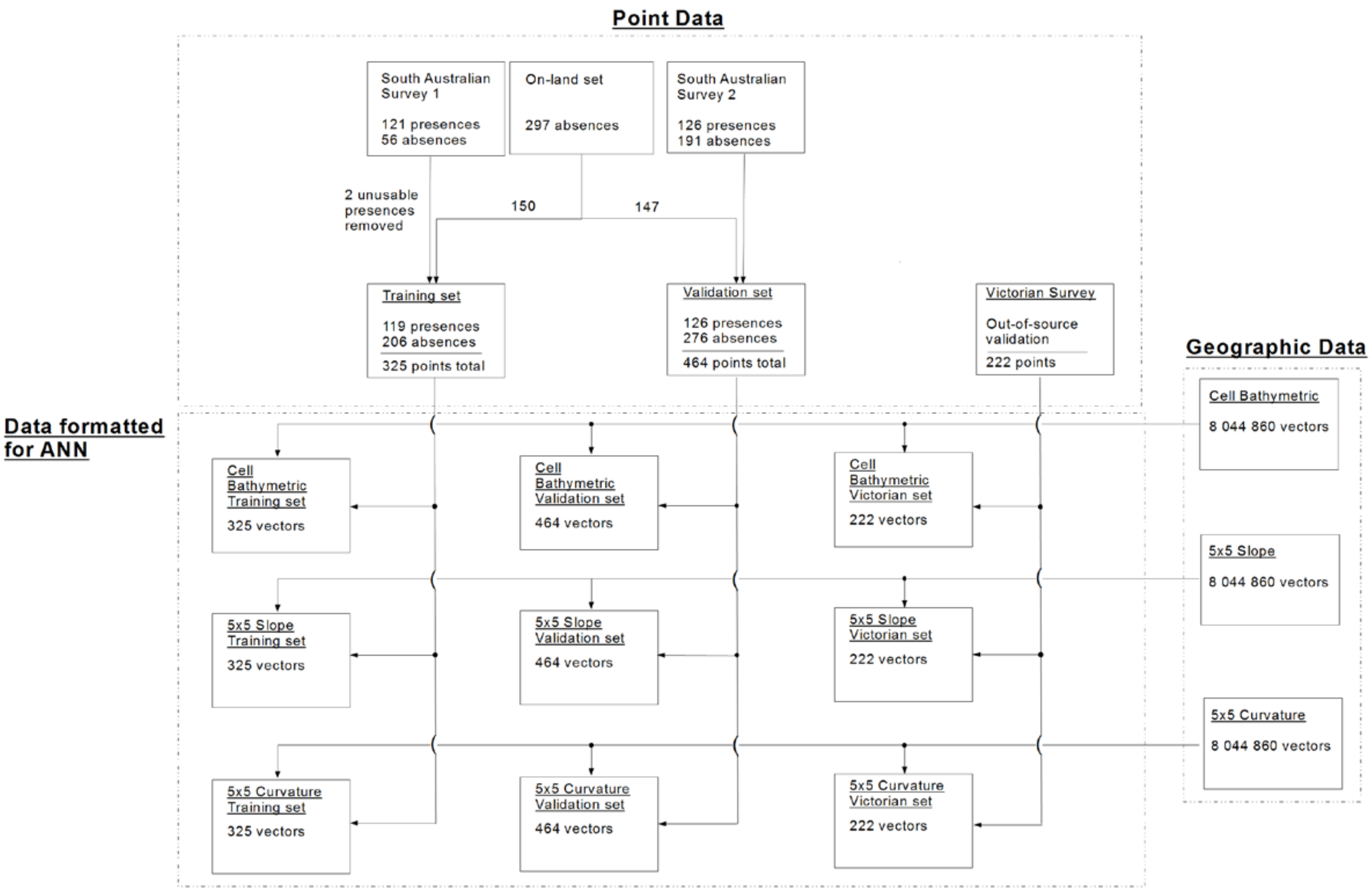

Figure 2 – combining source data sets to create ANN training data sets. The arrows show the flow of data from one set to another.

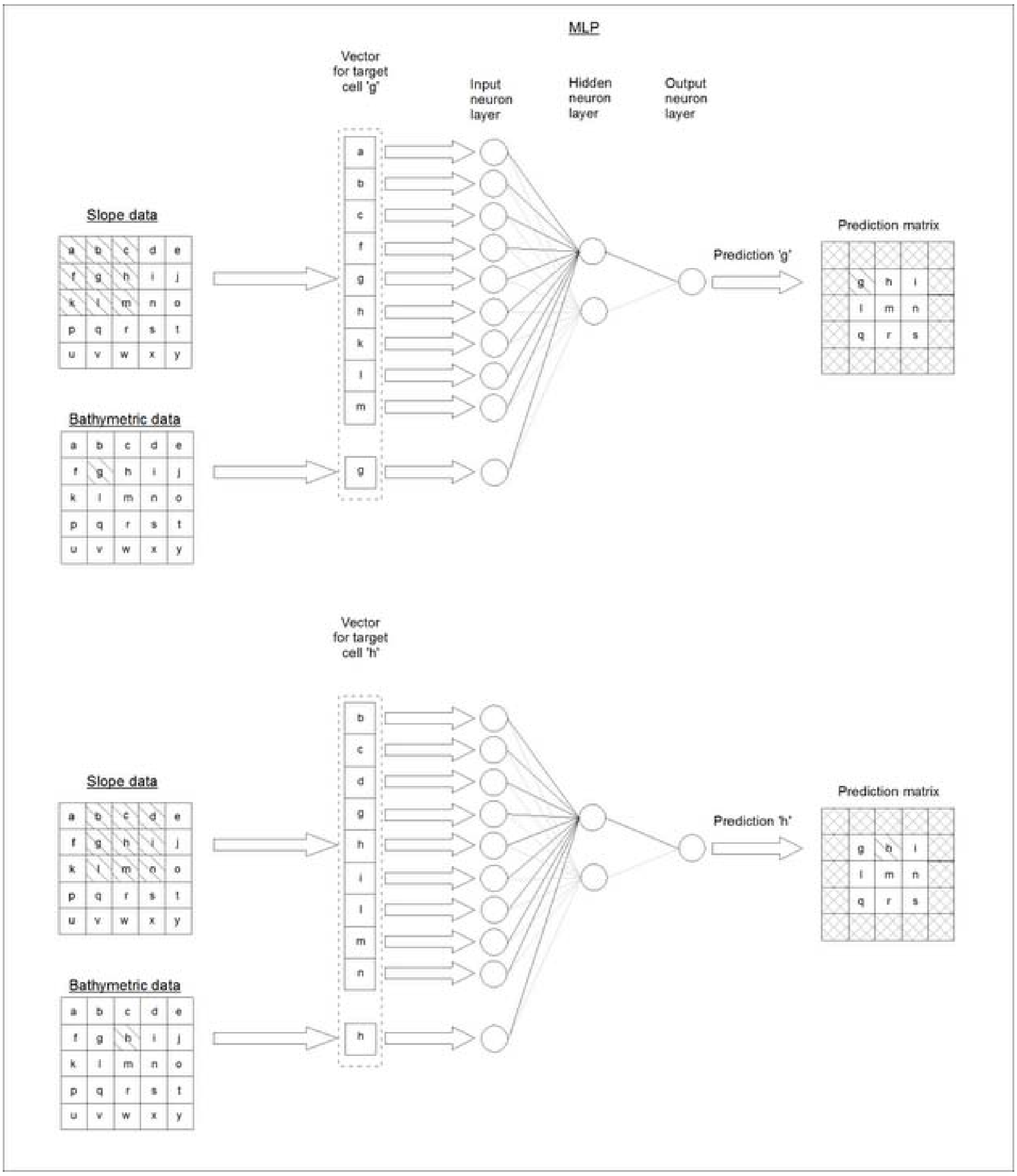

Figure 3 - encoding and prediction process for a 3-by-3 window of slope data and a single cell for bathymetric data. Two example vectors are being produced here. For the first, the target cell is cell ‘g’. For the second, the target cell is cell ‘h’. The cross-hatching in the prediction matrix shows the cells that are excluded from the predictions by the windowing process. Note that not all input neurons of the multi-layer perceptron (MLP) are shown

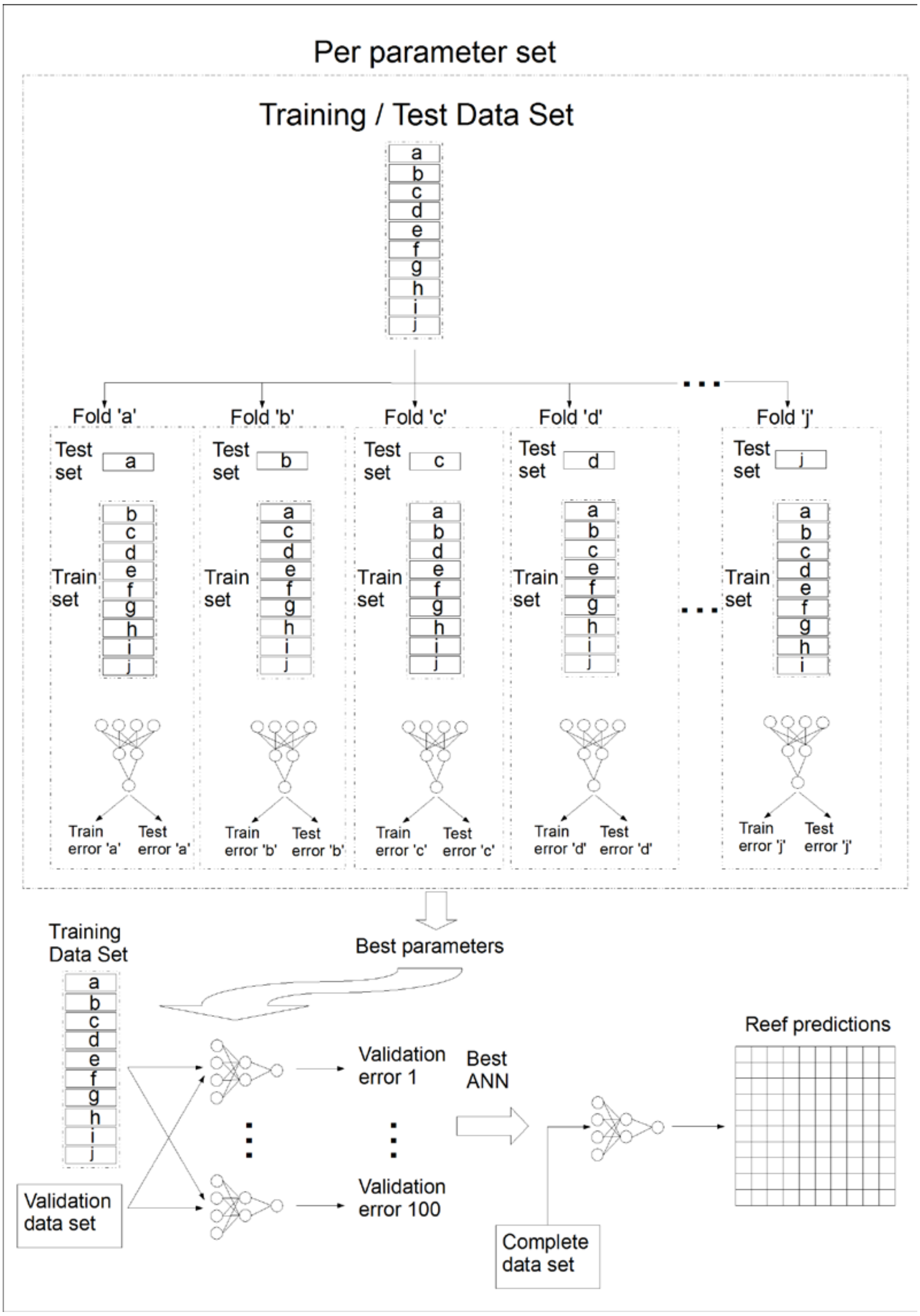

Figure 4 – schematic of cross-validation selection of training parameters, training over complete training set, selection of most accurate network and production of reef predictions.

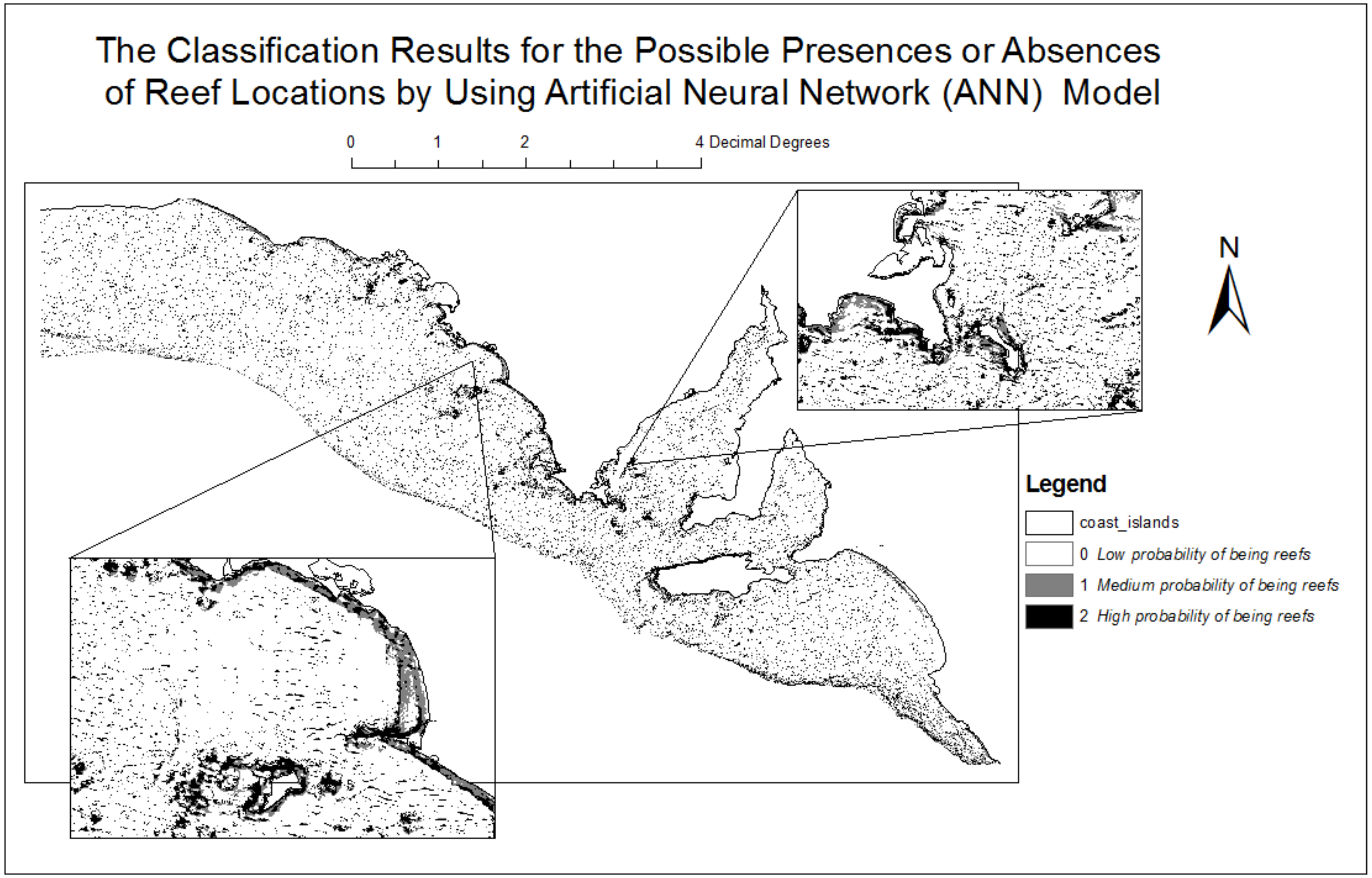

Figure 5 - Map of South Australian reefs generated by MLP trained on 5x5 input windows. Brighter colours are equal to higher probabilities of reef presences. Land masses are white and delimited by lines. The inserts are zoomed-in views of two areas on the South Australian coast and the Eyre peninsula. Areas deeper than 30 m have been masked out, as have areas on land.

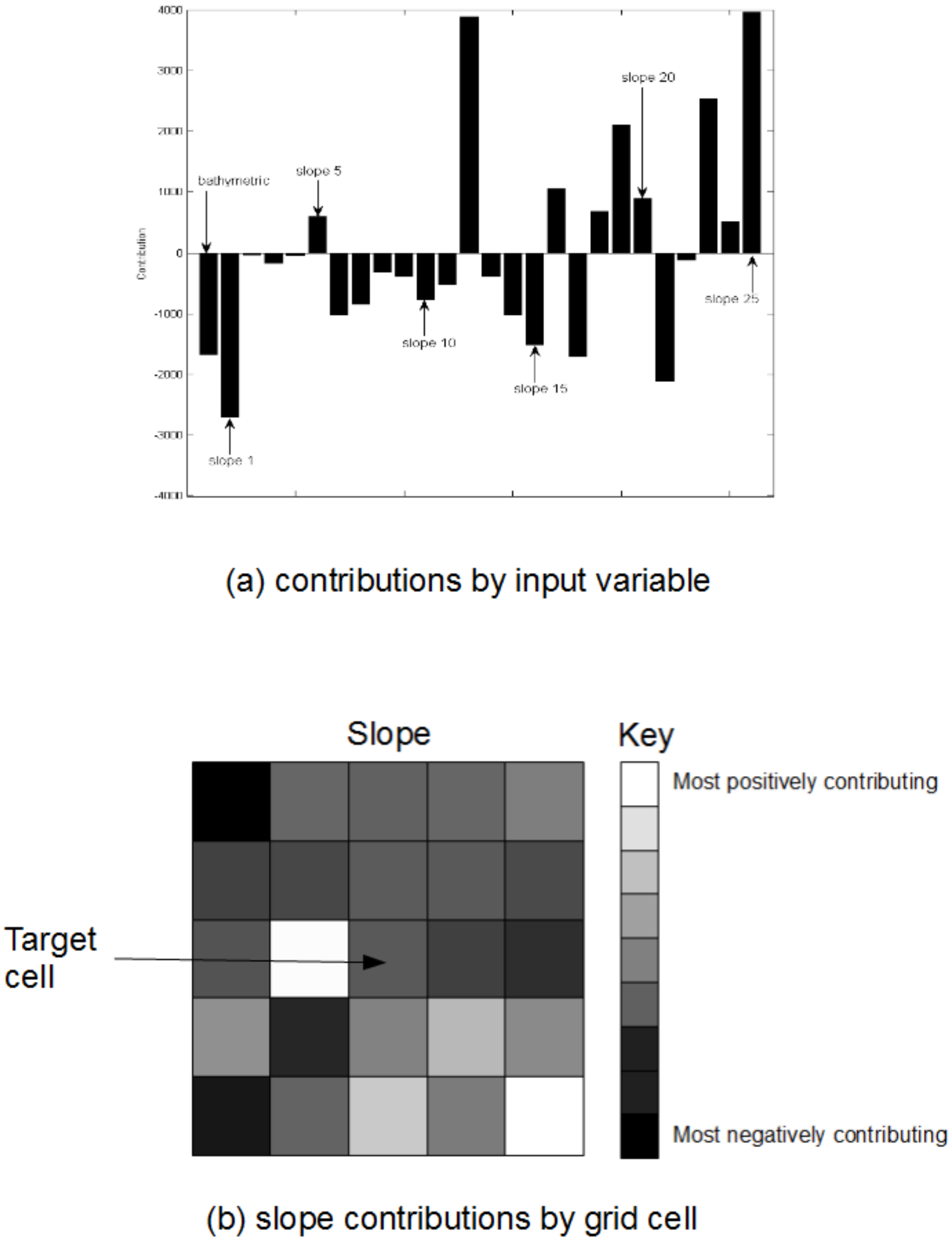

Figure 6 - Results of MLP input contribution for bathymetric value of target cell and a 5x5 window of seabed slope. In (a) the values of each variable are charted, with the variable corresponding to the bathymetric and slope variables labelled. In (b) the contributions of the slope variables are gridded according to their position in the sliding window and shaded according to their contribution, with the most positive contributions being white and the most negative contributions black.