1-1-2012

A simple post-hoc method to add spatial context to predictive species distribution models

Michael B. Ashcroft
University of Wollongong, ashcroft@uow.edu.au

Kristine O. French
University of Wollongong, kris@uow.edu.au

Laurie A. Chisholm
University of Wollongong, lauriec@uow.edu.au

A simple post-hoc method to add spatial context to predictive species distribution models

Abstract

Methods to incorporate spatial context into species distribution models (SDMs) are underutilised, with predictions usually based only on environmental space and ignoring geographic space. The goals of this study were to demonstrate a relatively simple post-hoc method to include spatial context in SDMs and to quantify the improvement over purely niche-based models. The method involved producing a standard niche-based model using established techniques, such as Maxent, and then calculating the neighbourhood average of the model output in geographic space. In effect, we tested whether the spatially averaged model output was better at predicting species distributions than the raw model output. We demonstrated the method using 32 tree species on the Illawarra Escarpment and found the area under the receiver operating characteristic curve (AUC) increased by a mean of 0.021 using this method. The improvements were largest for eucalypts, which have poor dispersal ability and clustered distributions. Improvements were smaller for moist rainforest species, which were restricted to small areas with sufficient shelter from hot, dry northwesterly winds. We conclude that it is relatively easy to add spatial context into species distribution models using this post-hoc method, and the resulting models are better for predicting species' distributions.

Keywords

distribution, species, predictive, context, spatial, models, add, simple, method, hoc, post

Disciplines

Life Sciences | Physical Sciences and Mathematics | Social and Behavioral Sciences

Publication Details

Ashcroft, M. B., French, K. O. & Chisholm, L. A. (2012). A simple post-hoc method to add spatial context to predictive species distribution models. Ecological Modelling, 228 17-26.

Article type: Original research paper

A simple post-hoc method to add spatial context to predictive species distribution models

Michael B Ashcroft [a, c, *]

Kristine O French [b]

Laurie A Chisholm [b]

[a] School of Earth and Environmental Sciences, University of Wollongong, Wollongong, NSW, Australia, 2522.

[b] Institute for Conservation Biology and Environmental Management, University of Wollongong, Wollongong, NSW, Australia, 2522.

[c] Australian Museum, 6 College Street, Sydney, NSW, Australia, 2010.

* Corresponding author. Tel.: +61 2 9320 6475; fax: +61 2 9320 6021; E-mail: Mick.Ashcroft@austmus.gov.au

Abstract Methods to incorporate spatial context into species distribution models (SDMs) are underutilised, with predictions usually based only on environmental space and ignoring geographic space. The goals of this study were to demonstrate a relatively simple post-hoc method to include spatial context in SDMs and to quantify the improvement over purely niche-based models. The method involved producing a standard niche-based model using established techniques, such as Maxent, and then calculating the neighbourhood average of the model output in geographic space. In effect, we tested whether the spatially averaged model output was better at predicting species distributions than the raw model output. We demonstrated the method using 32 tree species on the Illawarra Escarpment and found the area under the receiver operating characteristic curve (AUC) increased by a mean of 0.021 using this method. The improvements were largest for eucalypts, which have poor dispersal ability and clustered distributions. Improvements were smaller for moist rainforest species, which were restricted to small areas with sufficient shelter from hot, dry northwesterly winds. We conclude that it is relatively easy to add spatial context into species distribution models using this post-hoc method, and the resulting models are better for predicting species' distributions.

Keywords: Dispersal; Ecological niche models; Fragmentation; Landscape ecology; Neighbourhood averages; Spatial autocorrelation.

1. Introduction

The geographic distribution of a species is determined by factors such as its environmental niche, its dispersal ability and interspecific competition (Pulliam, 2000). However, species distribution models (SDMs, Guisan and Zimmermann, 2000; Rushton et al., 2004; Guisan and Thuiller, 2005) usually focus only on the environmental niche and neglect spatial processes such as dispersal (Guisan et al., 2006). That is, even though SDMs are frequently used to make spatial predictions, they are usually developed exclusively in environmental space and not geographic space. Although there have been numerous recommendations to consider spatial context in models (e.g. Guisan et al., 2006), the methods are still underutilised due to the difficulty of implementing them (Elith and Leathwick, 2009).

The proximity of presences in geographic space is often viewed as an issue of spatial autocorrelation, which can affect the perceived significance of predictors (Legendre, 1993) and bias model coefficients (Dormann, 2007). Indeed, the methods currently used to incorporate spatial context into SDMs have been assessed based on their effect on parameter estimates (Keitt et al., 2002) and are designed to remove spatial autocorrelation from the residuals or incorporate spatial autocorrelation into the statistical methods (Dormann et al., 2007). However, the ultimate usefulness of SDMs is their ability to predict species distributions, with spatially autocorrelated residuals or biased predictors of lesser importance unless they have a detrimental effect on predictive performance (Betts et al., 2009).

The goals of this study were to demonstrate a relatively simple method to include spatial context in SDMs and to quantify the improvement in predictive performance over purely environmental or niche based models. The method consists of two components. The first component is to develop a standard niche-based model using established techniques. Maxent, a popular machine-learning approach that is capable of producing complex models

without overfitting is used here (Phillips and Dudík, 2008), although the method could also be applied using methods such as Generalised Additive Models (GAMs), Generalised Linear Models (GLMs), and BioClim (Elith et al., 2006). The second component is to calculate the average model output for each location based on the surrounding geographic area. Although this is a crude and simple post-hoc method to add spatial context into models, the average amount of habitat in the surrounding area is a well established technique in landscape ecology (e.g. Betts et al., 2007), albeit often used with individual predictors rather than the actual output from an SDM (Ferrier et al., 2002; Wintle et al., 2005). Effectively, in this study, we tested whether the raw model output or the averaged model output in the local neighbourhood was better at predicting the distribution of species.

There are a number of reasons why neighbourhood averages of model output should improve predictions of distributions for, at least, some species. The raw model output ignores the amount and quality of habitat in the surrounding area. Therefore, a location with high habitat quality is predicted to be suitable even if it is isolated from other areas of high quality habitat. Averaging the model output over the surrounding neighbourhood lowers the predicted suitability in these circumstances (Fig. 1), which better reflects the lower colonisation rate and higher mortality rate that are predicted by island biogeography (MacArthur and Wilson, 1967) and fragmentation models (Fahrig and Merriam, 1994; Hill and Curran, 2003). Similarly, neighbourhood averages lower the predicted suitability of locations near edges (sharp transitions in habitat suitability) to capture edge-effects, and raise the predicted suitability of locations just outside the edges to simulate possible source-sink effects (Pulliam, 1988; Fig. 1). That is, neighbourhood averages suggest that species could be observed in sink locations that would be considered unsuitable by purely niche-based models. In addition, samples taken near sharp transitions in habitat suitability may also be erroneously recorded on the wrong side of boundaries (Mummery and Battaglia, 2002), and

neighbourhood averages create smoother transitions that cater for spatial errors and ecotones. In effect, raw model outputs only reflect environmental quality, whereas neighbourhood averages also estimate the effects of fragmentation and source-sink dynamics to provide a better estimate of probability of occurrence.

Our approach combines aspects of landscape ecology with the continuous output of species distribution models. Landscape ecology and island biogeography models were once based on the binary model of favourable or unfavourable habitat, but researchers have now realised that variations in the habitat quality of patches can have a profound influence on species' persistence as well as extinction and colonisation rates (Franken and Hik, 2004; Schooley and Branch, 2007; McAlpine et al., 2008). While these studies now consider differences in habitat quality between different patches, they are still limited to situations where the habitats are distinct, well-defined habitats such as snowbeds, wetlands or talus patches (Franken and Hik, 2004; Schooley and Branch, 2007; Dullinger et al., 2011), and they are based on the premise that these patches are relatively homogenous in habitat quality. Indeed, where relatively homogenous and distinct patches of habitat can be identified, it would be preferable to include patch size, distance from edge, and habitat quality as separate predictors so that the relative contributions of each can be isolated. However, species distribution models are generally used to predict continuous variations in habitat quality, and patches might not be readily identifiable when different species have complex and overlapping distributions in a more or less continuous forest. In these situations the sum (Betts et al., 2007) or average of habitat quality in the surrounding area provides a metric that combines both habitat quality and the amount of habitat without having to apply thresholds to classify and simplify the output of models to identify discrete patches.

It is also worth pointing out that neighbourhood averages have a different theoretical basis than methods currently used to cater for spatial autocorrelation. For example, methods

such as autologistic regression are designed to adjust predictions according to proximity in space, and hence neighbouring cells are weighted according to distance. The premise is that nearby locations are more similar than distant locations. However, our method is designed to reflect the total (or average) amount of habitat in the surrounding area, and hence we average all cells with equal weights. Our premise is that a greater area of favourable habitat leads to lower extinction rates, higher colonisation rates, higher mass effects and an overall higher probability of occurrence (MacArthur and Wilson, 1967). That is, our method is based on landscape ecology theory rather than the concept of spatial autocorrelation.

Neighbourhood averages have a number of other differences from other methods for incorporating spatial context into SDMs. Firstly, other methods of dealing with spatial context often utilise the presence or absence of the species in the surrounding area, or utilise the spatial autocorrelation in residuals. While these can potentially shed light on spatial patterns that are unrelated to the selected environmental factors, a drawback of these methods is that they are dependent on both presences and absences and cannot be used with presence only datasets (Dormann et al., 2007). Our method has the advantage that it can be used with presence-only datasets. Other methods also introduce circularity into the modelling process (the spatial context of the response variable or residuals is used as a predictor), and it is difficult to apply the models to other places or times when no survey data is available. Our method can be applied in this context, as it is only dependent on having environmental data available. Of course other methods could also use the modified Gibbs sampler (Augustin et al., 1996), which replaces the presence or absence in the surrounding area with the predicted probability of occurrence (as opposed to the standard Gibbs sampler which stochastically generates model predictions). In this respect, the neighbourhood averages we propose are somewhat similar to the modified Gibbs sampler, except the modified Gibbs sampler is applied iteratively. This makes the Gibbs sampler more robust, but also more complicated to

include in SDMs based on Maxent or other statistical methods. The simplicity of our approach allows more studies to consider spatial factors, especially those based on presence only datasets, or those that use modelling methods that do not have an in-built mechanism to consider spatial context (e.g. Maxent).

2. Methods

2.1 Study area

The study was conducted on the Illawarra Escarpment and Woronora plateau, 80 km south of Sydney, Australia (34.4 ^{o}S, 150.9 ^{o}E). The escarpment runs NE to SW, with Mt Keira and Mt Kembla rising over the city of Wollongong on the predominately cleared coastal plain in the south and east (Fig. 2). The uppermost geology in the study area, Hawkesbury sandstone, forms the summit of both mountains as well as the top of the escarpment. This geology supports vegetation communities that are vastly different from other substrates, and is dominated by eucalypt woodlands and upland swamps. The gullies on the Woronora plateau are predominately Bald Hill claystone and Bulgo sandstone, and support tall-open eucalypt forests, moist eucalypt forests, and rainforests (NPWS, 2002). The escarpment slopes and foothills also contain moist eucalypt forests and rainforests, but have a different species composition from the gullies on the plateau. The escarpment slopes consists of numerous layers of sandstones, claystones, and coal seams from the Narrabeen Group and Illawarra Coal Measures, with some species occurring more frequently on specific geological units.

2.2 Data

Presence-absence data were collected for 32 common canopy and subcanopy species (Table 1) between July 2005 and March 2006 at 600 sites (20m by 20m). The survey locations were randomly chosen subject to a number of constraints that were imposed to ensure that a representative and complete range of communities and environmental conditions were sampled. First, the proportion of each community in the study area was used to determine the approximate number of samples that should be taken from each community (NPWS, 2002). Once the number of samples for each community had been determined, a list of potential locations was placed in a random order. The highest ranked locations were selected provided the locations were spread geographically and environmentally, and we had permission to access the land. Sites were separated by approximately 300m on average, with adjacent sites from different communities or geologies where possible. Only 20 of the pairwise combinations between the 600 sites involved the same communities on the same geology separated by less than 200m. We were particularly careful to ensure that, where possible, each community was sampled on each geological unit on which it was found, and locations spanned all large and most small patches of the community. The constraints we imposed ensured the full range of conditions was sampled, which is more beneficial for modelling than ensuring the sample is random (Hirzel and Guisan, 2002).

We sought to avoid bias by not limiting our sites to those that were qualitatively homogenous or pristine, or by only surveying sites near or away from roads (to ensure easy access or avoid edge effects). While some sites were close together, this only occurred where the sites contained different vegetation communities, and usually on different geologies as well. The average walking distance between sites was approximately 300m, and many changes in vegetation were typically seen over this distance. Therefore, we believe that spatial auto-correlation in the survey was kept to a minimum. There are relatively few sites

on the coastal plain, and these are typically near creeks or in hilly areas, but this bias reflects land clearing preferences rather than a bias in our survey.

Eleven of the selected species were sclerophyll trees (hard leaved, evergreen), including 7 eucalypts (*Eucalyptus* spp.), red bloodwood (*Corymbia gummifera*), 2 acacias (*Acacia* spp.) and turpentine (*Syncarpia glomulifera*). Eight of the species were dry subtropical rainforest trees, which are also evergreen but have softer leaves. The remaining 13 species were moist rainforest species (Table 1), which were also predominately evergreen, mesic trees (*Toona ciliata* is deciduous, *Livistona australis* is a palm). The number of presences for each species ranged from 40 to 363 (mean 136, s.d. 91). Although five species had marginally less than the recommended 50 presences (Stockwell and Peterson, 2002; Coudun and Gégout, 2006) the fine spatial resolution in this study had the potential to produce better results with fewer presences than that recommended at coarser resolutions (Engler et al., 2004).

Models were produced using geology, summer maximum temperature, summer minimum temperature, and winter minimum temperature. Geology was a categorical layer (Moffit, 1999) that was used as a surrogate for soil properties that are known to influence the fine-scale distribution of vegetation (Beadle, 1954, 1966; Coudun et al., 2006). Boundaries of the geology polygons were quoted as having spatial errors of up to 150m. The three temperature layers were continuous predictors, derived at a fine scale using iButton temperature loggers at ground level (Ashcroft et al., 2008). These temperature surfaces were developed specifically for this study area, and were based on a number of climate-forcing factors including elevation, distance to streams, distance to coast, radiation, and exposure to winds. Surfaces were originally produced for three week periods between November 2004 and August 2006 (Ashcroft et al., 2008), but all surfaces from each season were later averaged to produce the seasonal surfaces. Summer minimum temperatures were well

correlated with elevation, while winter minimums were determined more by distance to coast, and summer maximum temperatures by exposure to hot, dry northwesterly winds. These three temperature surfaces were selected as they capture different spatial patterns, represent the extreme temperatures that are physiologically limiting for many species, and have been shown to explain species' distributions well (Ashcroft et al., 2008). Although we also developed temperature surfaces for other seasons, there is no significant improvement in model performance if extra temperature predictors are included in models (Ashcroft et al., 2011), and hence there was no benefit in including them.

2.3 Analysis

For each species, we randomly divided the 40 to 363 presences (Table 1) into ten pools of 10% for cross-validation purposes (i.e. pool sizes ranged from 4 to 37). Ten separate models were produced using Maxent version 3.2.19 (Phillips et al., 2006), where each model was produced using default parameters and the presences from 9 of the 10 pools (Maxent is a presence-only machine learning method which relies on background samples rather than observed absences). For each model we calculated the area under the receiver operating characteristic curve (AUC) for the raw logistic model output using all absences and the presences from the pool that was not used to produce the model. The AUC of each species was calculated as the average of the 10 respective models. We then calculated the neighbourhood average of the output from each of the 10 models using ArcGIS and the arbitrary radii of 50m, 100m, 200m, 400m, 600m, 800m and 1000m, and recalculated the AUC using the methods described above.

We calculated the improvement in predictive performance by subtracting the AUC of the raw logistic output from the AUC produced using the neighbourhood average of model

output. The optimal radius for each species was determined as the radius with the highest AUC, with a radius of 0 used when the raw logistic output performed best. We calculated the average improvement and optimal radius for each category of species (sclerophyll, dry rainforest, moist rainforest), where the improvement for each species was calculated using the optimal radius. We tested for global differences between the three categories of species using ANOVA, and if results were significant we examined pair-wise differences using multiple comparison analysis (TukeyHSD (Honestly Significant Difference) procedure in R). To produce maps of predicted distributions, we repeated the Maxent modelling process using all presences and calculated the neighbourhood average using the optimal radius for that species.

It is worth noting that any method of including spatial factors must make a decision on how to handle the boundary of the study area. If the study area is restricted strictly to the area that was surveyed, then sites near the boundary do not have the full spatial context (e.g. 1000m radius) like sites near the middle of the study area. In our study we extended the Maxent study area outside the surveyed area so that all sites would have the full spatial context. Model outputs will be less reliable in this extrapolated region, but we felt that it was better to include an estimate of suitability in these areas than reduce the spatial context for boundary locations. As the AUCs we calculated are based only on the surveyed sites, they are not directly affected by model extrapolations outside the surveyed region, although we still display the extrapolations in our figures for completeness.

3. Results

The AUCs for the 32 species using the raw logistic output ranged from 0.585 to 0.905 (mean 0.790, s.d. = 0.082). While there may be scepticism of the four models with an AUC of less than 0.7, it is worth pointing out that such a threshold, while commonly used, is not a

good basis to assess the relative merit of models (Lobo et al., 2008; Ashcroft et al., 2011), and these models did not impact the overall results (see below).

The average improvement in AUC of the neighbourhood averaged models (based on the optimal radius for each species) was 0.021 (s.d. = 0.017), and varied from 0 to 0.068 (Fig. 3). While improvements were correlated with optimal radii ($r^2 = 0.42$, $t = 4.62$, d.f. = 30, $P < 0.001$, Fig. 4a) and the performance of niche-based models was related to prevalence ($r^2 = 0.27$, $t = -3.29$, d.f. = 30, $P < 0.01$, Fig. 4b), neither the improvements nor optimal radii were related to prevalence or the performance of the niche-based models ($r^2 < 0.04$, $P > 0.31$, Fig. 4c-f). That is, there were no significant differences between the effects of neighbourhood averages on rare and common species, or between high and low performing models.

There were, however, significant differences between the three categories of species in terms of both optimal radius (ANOVA $F = 10.1$, d.f. = 2/29, $P = 0.0005$) and improvement in AUC ($F = 5.02$, d.f. = 2/29, $P = 0.013$). Multiple comparison tests using the TukeyHSD procedure indicated that moist rainforest species had significantly lower optimal radii than both sclerophyll ($P = 0.0047$) and dry rainforest species ($P = 0.0010$), but there was no difference between dry rainforest and sclerophyll species ($P = 0.67$; Fig. 5). Similarly, moist rainforest species had significantly smaller improvements (based on optimal radii) than sclerophyll species ($P = 0.013$) and there was no difference between sclerophyll and dry rainforest species ($P = 0.768$), but the difference between moist and dry rainforest species was non-significant in this case ($P = 0.115$; Fig. 5).

As predicted, the use of neighbourhood averages had pronounced effects on the predictions in locations where there were only limited amounts of suitable habitat. For example, the niche-based models for both *Eucalyptus sieberi* (Fig. 6a) and *Corymbia gummifera* (Fig. 6c) suggested that the highest quality habitat tended to be on Hawkesbury sandstone, and both raw logistic models predicted suitabilities of up to 0.8 for the small

Hawkesbury sandstone 'islands' of Mt Keira and Mt Kembla. However, the neighbourhood models for *C. gummifera* had a relatively large optimal radius (400m), and the suitability of these mountains was therefore reduced (<0.4; Fig. 6d). In contrast, the neighbourhood models for *E. sieberi* had a lower optimal radius (200m) and maintained a suitability of up to 0.75 on these mountains (Fig. 6b). This is consistent with the lack of observations of *C. gummifera* on these mountains, and indeed the species is less common here than it is on the Woronora plateau, but it is present in low density outside our sample sites. The fact that *C. Gummifera* was absent from areas where there were low amounts of favourable habitat in the surrounding area meant that the neighbourhood model improved the AUC for this species by 0.053, while *E. sieberi* was present even in locations where there was limited amount of habitat, and there was only an improvement of 0.019 in AUC when using neighbourhood averages.

Moist rainforest communities tend to occur in isolated patches on the Woronora Plateau where there is shelter from hot, dry northwesterly winds (NPWS, 2002), but individual species are not restricted to these patches and may be observed in the understorey of adjacent eucalypt forests. The models for these species were mostly influenced by summer maximum temperature (see also Ashcroft et al., 2008, 2011). Neighbourhood averages resulted in a loss of delineation between the rainforest patches, and hence model performance decreased if radii were too large (Fig. 3c). For example, the optimal radius for *Ceratopetalum apetalum* was 100m, and an improvement of only 0.013 was obtained (Fig. 7a–b).

In contrast, the models for dry rainforest species were influenced more by winter minimum temperature (see also Ashcroft et al., 2008, 2011), and the species were generally restricted to the lower slopes of the escarpment. The larger radii in the models for these species tended to produce one large band of suitable habitat (e.g. Fig. 7d). Isolated observations on the coastal plain and Woronora plateau were difficult to explain in the niche models for these species (e.g. Fig. 7c), however, neighbourhood averages could explain these

in terms of proximity to the band of habitat on the escarpment. Hence, larger improvements in model performance were observed for these species (e.g. 0.030 for *Croton verreauxii*; Fig 7c–d).

The large radii of sclerophyll species also meant that the models predicted they were restricted to rather large and distinct patches (E.g. Fig. 8b, d), however the locations of these patches varied widely. Two species were known to be restricted to the northern sections of the escarpment (*Eucalyptus pilularis* (Fig. 8a) and *Syncarpia glomulifera*), which remains difficult to explain with purely niche-based models because there are also areas with similar geology and elevation in southern areas of the escarpment. The temperature surfaces we produced suggested that the escarpment slopes north of Mt Keira were approximately 2^{o}C warmer in terms of winter minimum temperature than similar elevations south of Mt Kembla, yet the purely niche-based models still predicted that there were suitable locations in the south (logistic output > 0.8), while there were some presences in apparently unsuitable environments in the north (Fig. 8a). The neighbourhood average models captured the fact that there was, in general, a greater amount of higher quality habitat in the northern half of the study area. These models improved the AUC by 0.068 and provided a better contrast between the northern and southern portions of the escarpment (Fig. 8b). Improvements were not as large for all other sclerophyll species. For example, the distribution of *Eucalyptus cypellocarpa* could largely be explained by the niche based models (Fig. 8c), and an improvement of only 0.018 was observed when using neighbourhood averages (Fig. 8d).

4. Discussion

In this study we demonstrated that it is relatively simple to add spatial context into species distribution models. The neighbourhood averages of Maxent model output were

better able to explain the distribution of species, with the AUC increasing by up to 0.068 (mean 0.021). While this number may appear small, it is comparable to other studies. For example, Elith et al. (2006) found that the difference in performance between different statistical methods for producing SDMs was up to 0.082, with Maxent being one of the better performing methods. Therefore, the improvement that we recorded would be cumulative on the differences they noted. Phillips and Dudík (2008) noted improvements in AUC of up to 0.023 when optimising Maxent settings, so our results suggest that even simple methods to add spatial context into models may lead to bigger improvements than fine-tuning purely niche-based models.

4.1 What determines the optimal radius?

The 32 species we studied varied widely in optimal radius, with the full range of 0 to 1000m observed. Neighbourhood averages produced using small radii resulted in models that were little different from the purely-niche based models on which they were based. Neighbourhood averages produced using larger radii reduced the perceived suitability where there was a limited amount of favourable habitat in the surrounding area and increased the relative suitability of marginal habitat where there were locations with a lot of high quality habitat. The models with the largest improvements and largest radii restricted species to rather large ‘patches’, although strictly speaking there were no distinct patches present in our study, but rather continual variations in habitat quality. Therefore, the radius appeared to be determined by the ‘patch size’ that species occupied, although the number of presences in apparently unsuitable conditions in close proximity to favourable habitat or absences near the ‘edges of patches’ also played a role.

It is already well known that species vary in their ability to persist in areas where there is limited habitat (Hobbs and Yates, 2003). Common species are typically observed in both large and small patches, but rare species are affected more by fragmentation and are likely to be found only in large patches (Honnay et al., 1999; Davies et al., 2000; Hill and Curran, 2003; Debinski, 2006). There is reduced seed rain in small fragments (Hobbs and Yates, 2003), but it is worth noting that the ability for species to persist in small patches is determined by its mortality rate and colonisation capabilities, which is more than simply dispersal. Factors such as fecundity, dormancy, seedling establishment characteristics, species interactions, and habitat quality also influence colonisation ability (Fahrig and Merriam, 1994; Levin et al., 2003; Levine and Murrell, 2003; Franken and Hik, 2004; Guisan and Thuiller, 2005; Dullinger et al., 2011). We therefore predict that species that are better modelled using larger radii will have traits such as low fecundity, poor dispersal, or low probability of establishment, which is assessed below for the three categories of species we studied. Importantly, the optimal radius of neighbourhood averages is not expected to be proportional to dispersal distance, and we suggest that species with shorter dispersal distances would actually tend to have larger optimal radii (they would only survive in large patches where lower extinction rates counteracted their low colonisation ability).

The sclerophyll species in our study typically have limited dispersal mechanisms, and reproduce more abundantly following sporadic disturbances such as fire. They should therefore have relatively low colonisation ability, and consistent with our prediction above, they were better modelling using larger radii. Given that models for eucalypts have often been unsatisfactory (Austin et al., 1997), we suggest that neighbourhood averages should be pursued to improve predictive models for these species. Indeed, the largest improvements in AUC that we identified were for sclerophyll species (average improvement of 0.030).

In contrast, moist rainforest species should have good colonisation ability, as they typically produce many fleshy, bird dispersed seeds. These moist rainforest species were typically restricted to small areas where there was shelter from hot, dry northwesterly winds. We found that they were typically better modelled using low radii, and the use of neighbourhood average resulted in a smaller increase in AUC (mean 0.011). This is once again consistent with our prediction above.

Dry rainforest species also typically have fleshy, bird dispersed seeds, and so we would also predict them to occupy both large and small patches. Therefore, the larger optimal radii for these species are not consistent with our prediction. This may reflect that factors other than dispersal ability have a larger effect on colonisation ability for these species or that other spatial processes are operating. That is, the differences between the niche and neighbourhood models may be due to low fecundity, poor seedling establishment, species interactions or other spatial processes, because the species are generally expected to have good dispersal ability.

Although we have generalised species as either niche or dispersal limited in this section, these are opposite ends of a continuum (Gravel et al., 2006; Moore and Elmendorf, 2006), and neighbourhood averages can cater for this continuum by varying the radius.

4.2 The advantages and disadvantages of neighbourhood averages

Any method of catering for spatial context may simply be compensating for a missing, spatially structured environmental factor (Austin, 2002), and our method is no exception to this. Including factors such as soil nutrients or moisture may improve niche-based models and eliminate the benefits of using neighbourhood averages. Similarly, some niche-based models may be spurious correlations due to spatial structure in species'

distributions and environmental factors (Bahn and McGill, 2007), and the models produced here may not reflect causal processes. It is difficult to assess the degree to which this affected our results. Nevertheless, it is worth assessing the degree to which we learn anything about spatial processes from our results.

Environmental factors are often spatially autocorrelated, and hence it is inherently difficult to separate the effects of space and environment or determine which processes are causing spatial clustering (Wagner and Fortin, 2005; Currie, 2007). For example, methods that rely on clustering of survey data to incorporate spatial context into models are not determining the causal factors that caused those presences to be clustered in the first place. Importantly, our method does not simply inflate the predicted suitability in the areas where many presences have been observed. Instead, it makes a specific prediction that presences will be more common in areas where there is a lot of highly suitable habitat, and indeed that presences will be more common in marginal habitat near large patches of highly suitable habitat than they are near the core of small patches of highly suitable habitat. That is, our method is predicting the locations where clusters will occur based only on environmental conditions, and suggesting the reason the clusters occur there is due to the amount of suitable habitat. It could even be used to predict the locations of clusters in unsurveyed areas, as it relies solely on environmental data rather than survey data. However, the niche-based models we used to calculate habitat are not able to definitely separate environmental and spatial factors when they are correlated, and hence our method is still not able to definitively separate the two.

The main advantages of neighbourhood averages are that they can be used with presence-only data, and can be applied post-hoc to any statistical method. They are not as robust as the modified Gibbs sampler due to the lack of iteration, and indeed the purely-niche based models we produced originally may have over-estimated the width of the niches if they

included presences in sink locations (Austin, 2002). Locations that have relatively low output in the niche based model, but high in the neighbourhood average are expected to be the most likely to be sinks, however, we tested the effect of removing these from models for two species and found little difference in results (unpublished data). This supports the suggestion that considering spatial factors is more important than the actual method applied (Keitt et al., 2002).

There may also be other ways to improve upon the results we presented. For example, rather than considering the average model output within a given distance, cells could be weighted according to their distance from the centre (Ferrier et al., 2002). Although we deliberately avoided this because we wanted to our models to reflect the amount of habitat in the surrounding area, the probability of observing a species is almost certainly a function of both the habitat at a location as well as the amount of habitat in the surrounding area. The niche based models we produced only consider the former, while the neighbourhood averages only consider the latter. Including both factors separately in models may be required, which would require producing a second model that contains the neighbourhood average from the original model, as well as the environmental factors. This would bring our method even closer to autologistic regression with a modified Gibbs sampler, and iterations would need to be applied. Despite these potential improvements, the method we followed is a valuable addition to species distribution modelling, and could be used with any statistical method to improve the predictions of species distributions.

Acknowledgements

This research was conducted as part of a Ph.D. at the University of Wollongong with a University Postgraduate Award scholarship. Thanks to the many people who helped with the

fieldwork, granted access to their land, provided GIS data or reviewed earlier drafts of this manuscript, including BHP Billiton, the NSW National Parks and Wildlife Service, Bernard Ashcroft, Anders Boefeldt and Jane Elith.

References

Ashcroft, M.B., Chisholm, L.A., French, K.O., 2008. The effect of exposure on landscape scale soil surface temperatures and species distribution models. Landscape Ecol. 23, 211–225.

Ashcroft , M.B., French, K.O, Chisholm, L.A., 2011. An evaluation of environmental factors affecting species distributions. Ecol. Modell. 222, 524–531.

Augustin, N.H., Mugglestone, M.A., Buckland, S.T., 1996. An autologistic model for the spatial distribution of wildlife. J. Appl. Ecol. 33, 339–347.

Austin, M.P., Pausas, J.G., Noble, I.R., 1997. Modelling environmental and temporal niches of eucalypts. In: Williams, J.E., Woinarski, J.C.Z. (Eds.), Eucalypt Ecology. Cambridge University Press, Cambridge, UK, pp. 129–150.

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Modell. 157, 101–118.

Bahn, V, McGill, B.J., 2007. Can niche-based distribution models outperform spatial interpolation? Global Ecol. Biogeogr. 16, 733–742.

Beadle, N.C.W., 1954. Soil phosphate and the delimitation of plant communities in eastern Australia. Ecology 35, 370–375.

Beadle, N.C.W., 1966. Soil phosphate and its role in molding segments of the Australian flora and vegetation, with special reference to xeromorphy and sclerophylly. Ecology 47, 992–1007.

Betts, M.G., Forbes, G.J., Diamond, A.W., 2007. Thresholds in songbird occurrence in relation to landscape structure. Conserv. Biol. 21, 1046–1058.

Betts, M.G., Ganio, L.M., Huso, M.M.P., Som, N.A., Huettmann, F., Bowman, J., Wintle, B.A., 2009. Comment on "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review". Ecography 32, 374–378.

Coudun, C., Gégout, J.-C, 2006. The derivation of species response curves with Gaussian logistic regression is sensitive to sampling intensity and curve characteristics. Ecol. Modell. 199, 164–175.

Coudun, C., Gégout, J-C., Piedallu, C., Rameau, J-C., 2006. Soil nutritional factors improve models of plant species distribution: an illustration with *Acer campestre* (L.) in France. J. Biogeogr. 33, 1750–1763.

Currie, D., 2007. Disentangling the roles of environment and space in ecology. J. Biogeogr. 34, 2009–2011.

Davies, K.F., Margules, C.R., Lawrence, J.F., 2000. Which traits of species predict population declines in experimental forest fragments? Ecology 81, 1450–1461.

Debinski, D.M., 2006. Forest fragmentation and matrix effects: the matrix does matter. J. Biogeogr. 33, 1791–1792.

Dormann, C.F., 2007. Assessing the validity of autologistic regression. Ecol. Modell. 207, 234–242.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, D., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30, 609–628.

Dullinger, S., Mang, T., Dirnböck, T., Ertl, S., Gattringer, A., Grabherr, G., Leitner, M., Hülber, K., 2011. Patch configuration affects alpine plant distribution. Ecography 34, 576–587.

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.McC., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697.

Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J. Appl. Ecol. 41, 263–274.

Fahrig, L., Merriam, G., 1994. Conservation of fragmented populations. Conserv. Biol. 8, 50–59.

Ferrier, S., Watson, G., Pearce, J., Drielsma, M., 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. Biodivers. Conserv. 11, 2275–2307.

Franken, R.J., Hik, D.S., 2004. Influence of habitat quality, patch size and connectivity on colonization and extinction dynamics of collared pikas *Ochotona collaris*. J. Anim. Ecol. 73, 889–896.

Gravel, D., Canham, C.D., Beaudet, M., Messier, C., 2006. Reconciling niche and neutrality: the continuum hypothesis. Ecol. Lett. 9, 399–409.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Modell. 135, 147–186.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J.McC., Aspinall, R., Hastie, T., 2006. Making better biogeographical predictions of species' distributions. J. Appl. Ecol. 43, 386–392.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8, 993–1009.

Hill, J.L., Curran, P.J., 2003. Area, shape and isolation of tropical forest fragments: effects of tree species diversity and implications for conservation. J. Biogeogr. 30, 1391–1403.

Hirzel, A., Guisan, A., 2002. Which is the optimal strategy for habitat suitability modelling. Ecol. Modell. 157, 331–341.

Hobbs, R.J., Yates, C.J., 2003. Turner review No. 7. Impacts of ecosystem fragmentation on plant populations: generalising the idiosyncratic. Aust. J. Bot. 51, 471–488.

Honnay, O., Hermy, M., Coppin, P., 1999. Impact of habitat quality on forest plant species colonization. For. Ecol. Manage. 115, 157–170.

Keitt, T.H., Bjørnstad, O.N., Dixon, P.M., Citron-Pousty, S., 2002. Accounting for spatial pattern when modeling organism-environment interactions. Ecography 25, 616–625.

Legendre, P., 1993. Spatial autocorrelation: Trouble or new paradigm? Ecology 74, 1659–1673.

Levin, S.A., Muller-Landau, H.C., Nathan, R., Chave, J., 2003. The ecology and evolution of seed dispersal: A theoretical perspective. Annu. Rev. Ecol. Evol. Syst. 34, 575–604.

Levine, J.M., Murrell, D.J., 2003. The community-level consequences of seed dispersal patterns. Annu. Rev. Ecol. Evol. Syst. 34, 549–574.

Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecol. Biogeogr. 17, 145–151.

MacArthur, R.H., Wilson, E.O., 1967. The Theory of Island Biogeography. Princeton University Press.

McAlpine, C.A., Rhodes, J.R., Bowen, M.E., Lunney, D., Callaghan, J.G., Mitchell, D.L., Possingham, H.P., 2008. Can Multiscale models of species' distribution be generalized from region to region? A case study of the koala. J. Appl. Ecol. 45, 558–567.

Moffit, R.S., 1999. Southern Coalfield Regional Geology 1:100 000, 1st edition. Geological Survey of New South Wales, Sydney.

Moore, K.A., Elmendorf, S.C., 2006. Propagule vs. niche limitation: untangling the mechanisms behind plant species' distributions. Ecol. Lett. 9, 797–804.

Mummery, D., Battaglia, M., 2002. Data input quality and resolution effects on regional and local scale *Eucalyptus globulus* productivity predictions in north-east Tasmania. Ecol. Modell. 156, 13–25.

NPWS, 2002. Native Vegetation of the Illawarra Escarpment and Coastal Plain. NSW National Parks and Wildlife Service.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Modell. 190, 231–259.

Phillips, S.J., Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.

Pulliam, H.R., 1988. Sources, sinks and population regulation. Am. Nat. 132, 652–661.

Pulliam, H.R., 2000. On the relationship between niche and distribution. Ecol. Lett. 3, 349–361.

Rushton, S.P., Ormerod, S.J., Kerby, G., 2004. New paradigms for modelling species distributions? J. Appl. Ecol. 41, 193–200.

Schooley, R.L., Branch, L.C., 2007. Spatial heterogeneity in habitat quality and cross-scale interactions in metapopulations. Ecosystems 10, 846–853.

Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecol. Modell. 148, 1–13.

Wagner, H.H., Fortin, M-J., 2005. Spatial analysis of landscapes: concepts and statistics. Ecology 86, 1975–1987.

Wintle, B.A., Elith, J., Potts, J.M., 2005. Fauna habitat modelling and mapping: A review and case study in the lower Hunter Central Coast region of NSW. Austral Ecol. 30, 719–738.

Table 1 The 32 species that were modelled as part of this study. Species were classified as either sclerophyll[†], dry rainforest[‡] or moist rainforest[*] based on the communities in which they most frequently occur (NPWS 2002).

Scientific name	Common name	Abbreviation	Presences
Acacia binervata[†]	Two-veined hickory	TVH	207
A. mearnsii[†]	Green wattle	GW	51
Corymbia gummifera[†]	Red bloodwood	RB	65
Eucalyptus cypellocarpa[†]	Mountain grey gum	MGG	48
E. pilularis[†]	Blackbutt	BB	50
E. piperita[†]	Sydney peppermint	SPM	108
E. quadrangulata[†]	Coast white box	CWB	153
E. salignaXbotryoides[†]	Blue gum hybrid	BGH	219
E. sieberi[†]	Silvertop ash	SA	82
E. smithii[†]	Gully gum	GG	58
Syncarpia glomulifera[†]	Turpentine	TT	66
Cassine australis[‡]	Red olive plum	ROP	129
Clerodendrum tomentosum[‡]	Hairy clerodendrum	HC	143
Croton verreauxii[‡]	Native cascarilla	NC	112
Melicope micrococca[‡]	Hairy-leaved doughwood	HLD	40
Notelaea venosa[‡]	Veined mock-olive	VMO	363
Pittosporum undulatum[‡]	Sweet pittosporum	SP	280
Streblus brunonianus[‡]	Whalebone tree	WB	70
Synoum glandulosum[‡]	Scentless rosewood	SR	275
Acmena smithii[*]	Lilly pilly	LP	266
Ceratopetalum apetalum[*]	Coachwood	CW	135
Cryptocarya glaucescens[*]	Jackwood	JW	203
C. microneura[*]	Murrogun	MG	198
Cyathea leichhardtiana[*]	Prickly tree fern	PT	46
Dendrocnide excelsa[*]	Giant stinging tree	GST	46
Doryphora sassafras[*]	Sassafras	SF	177
Eupomatia laurina[*]	Bolwarra	BWR	132
Ficus coronata[*]	Creek sandpaper fig	CSF	86
Livistona australis[*]	Cabbage tree palm	CTP	320
Polyosma cunninghamii[*]	Featherwood	FW	57
Tasmannia insipida[*]	Brush pepperwood	BP	49
Toona ciliata[*]	Red cedar	RC	102

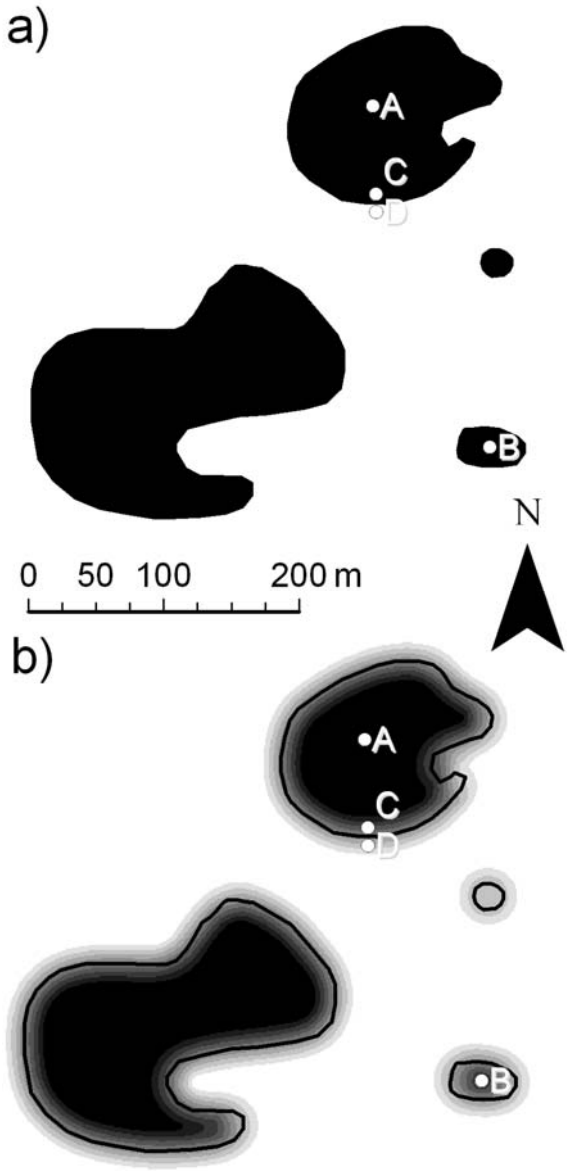

Fig. 1**:** A hypothetical example illustrating the influence of neighbourhood averages on model output. Panel a) illustrates a simplified model output, where locations have either suitable (black) or unsuitable (white) environmental conditions. Panel b) shows the 20m radius neighbourhood average of the output in panel a), with darker shades of grey representing a higher probability of occurrence. Points A and B illustrate how the neighbourhood average reduces the relative suitability of small habitat patches. Points C and D illustrate how they capture potential source-sink and edge effects, and cater for possible spatial errors in the sample locations and environmental layers.

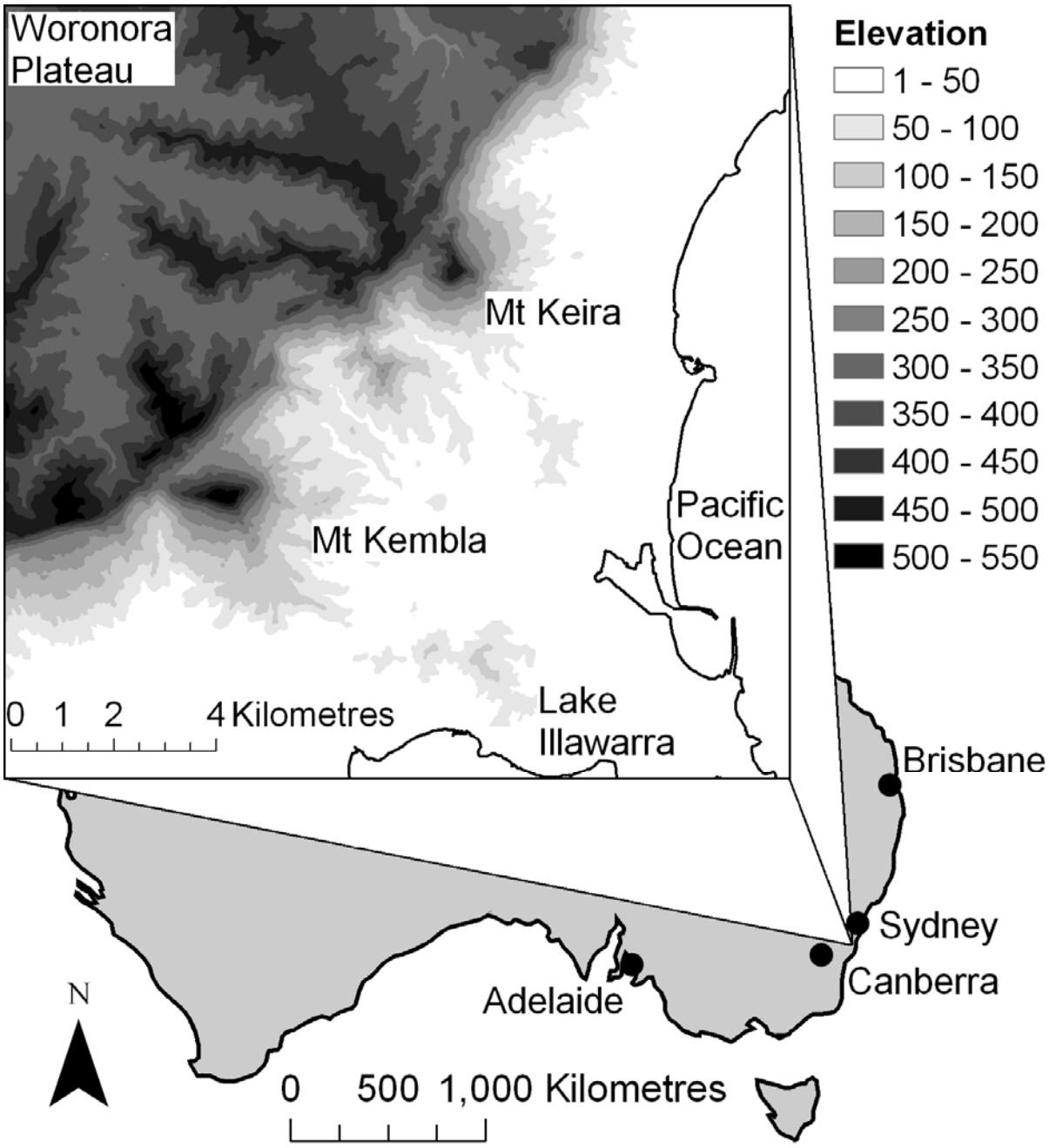

Fig. 2: The topography of the Illawarra Escarpment in the vicinity of Wollongong, Australia (34.4°S, 150.9°E). The inset is a Digital Elevation Model showing the rising elevation (m) from the coastal plain to the Woronora Plateau, with Mt Keira and Mt Kembla protruding eastward.

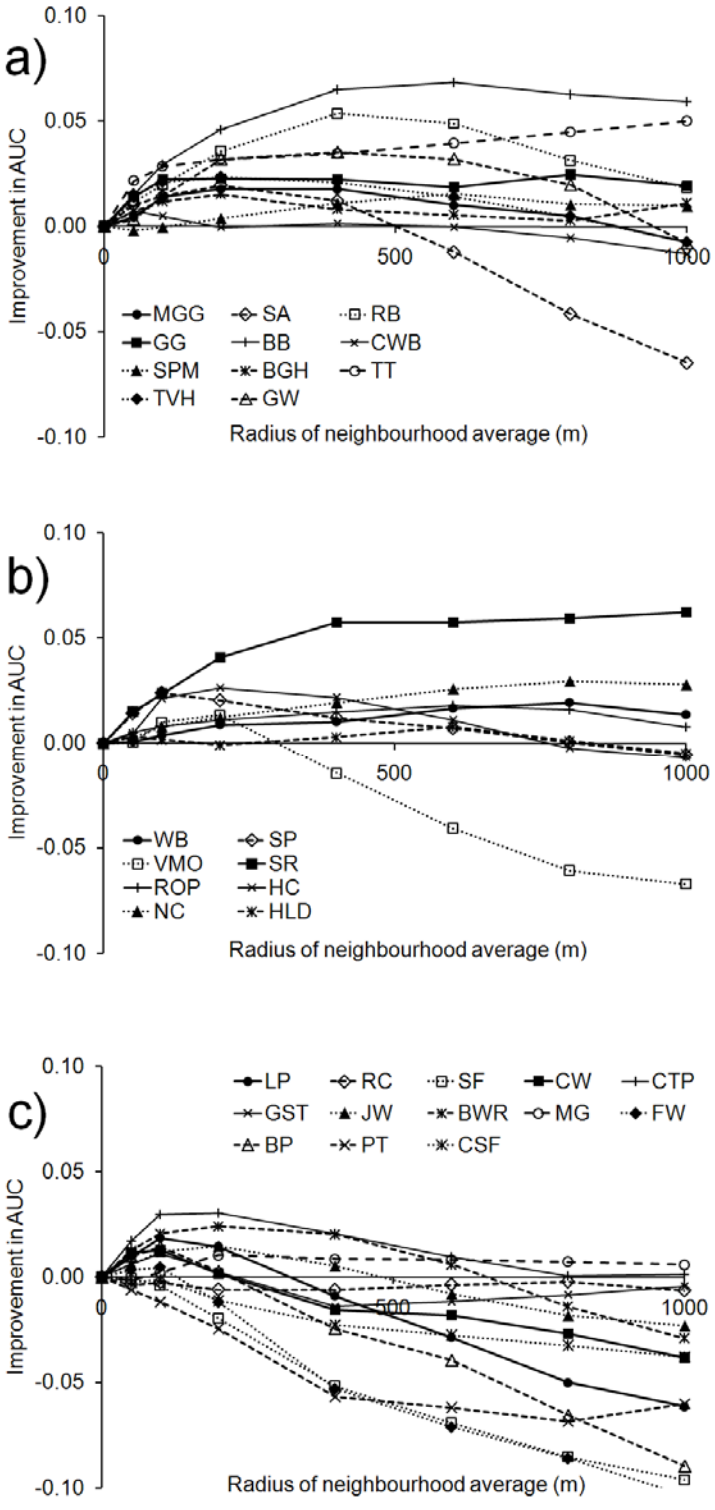

Fig. 3: The difference between the AUC of the neighbourhood averaged models and the raw logistic output of the Maxent models as assessed for sclerophyll (a), dry rainforest (b) and moist rainforest (c) species.

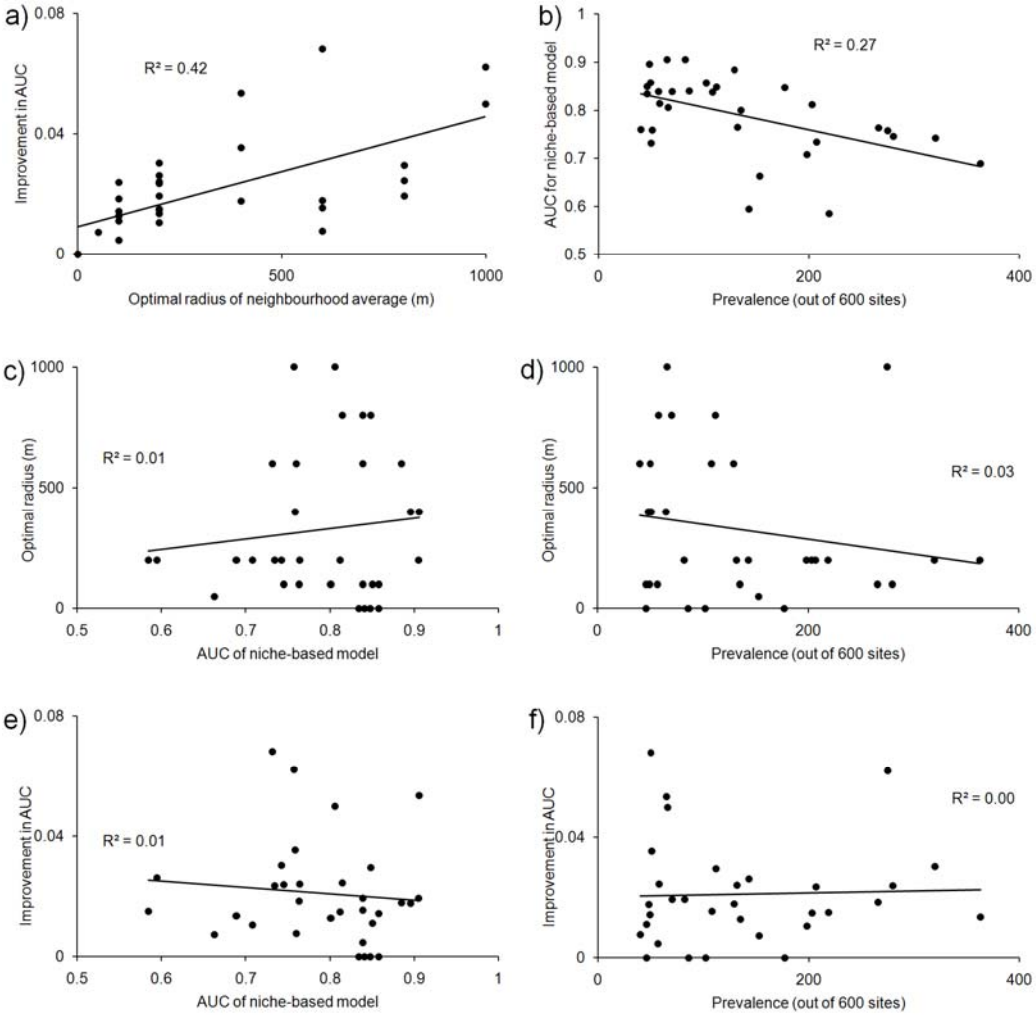

Fig. 4: Relationships are shown between the improvement in AUC when using neighbourhood averages instead of the raw logistic output of Maxent models, the samples prevalence of the respective species in a vegetation survey of 600 sites, the optimal radius of the neighbourhood average, and the AUC of original niche-based Maxent model (raw logistic output).

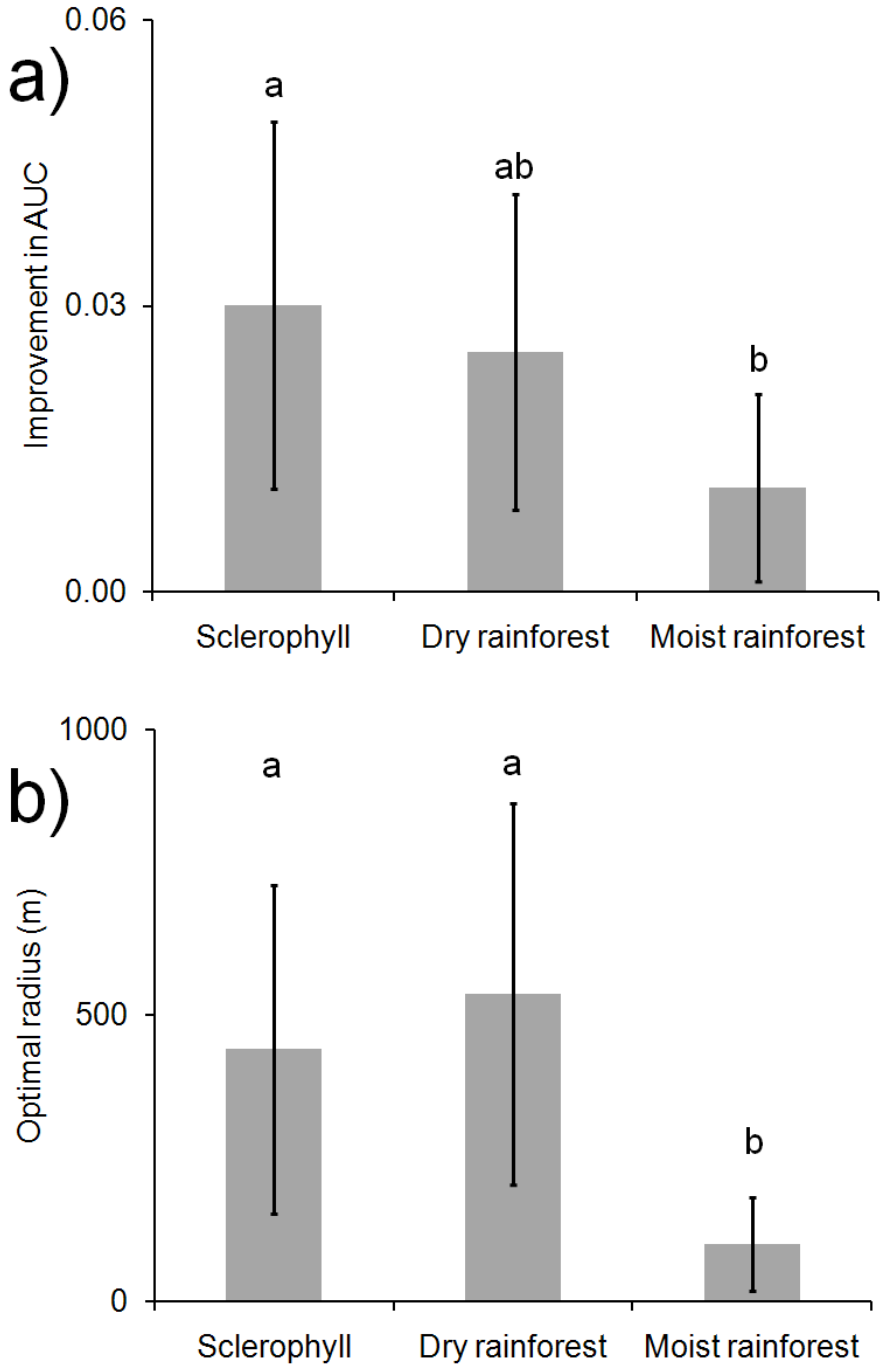

Fig. 5: The average improvement in AUC (a) and the average optimal radius (b) when the neighbourhood averages of Maxent logistic output were calculated for 32 species. Improvements are based on the optimal radii. Error bars indicate ± 1 standard deviation. Columns sharing the same letter are not significantly different ($P<0.05$) according to Student's t-test.

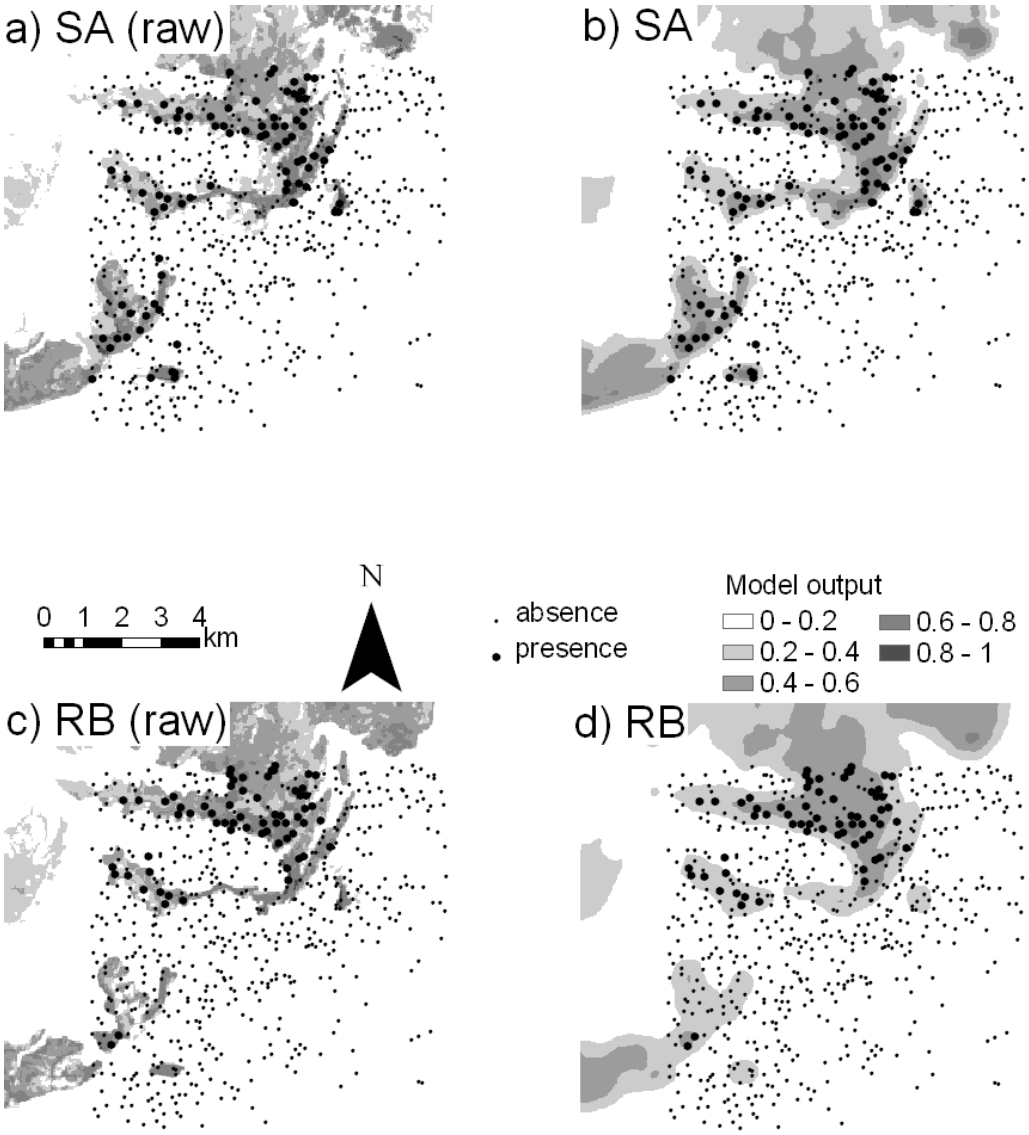

Fig. 6: The raw Maxent logistic model output (a, c) for *Eucalyptus sieberi* (SA) and *Corymbia gummifera* (RB) and the neighbourhood averages created using the optimal radii of 200m and 400m respectively (b, d).

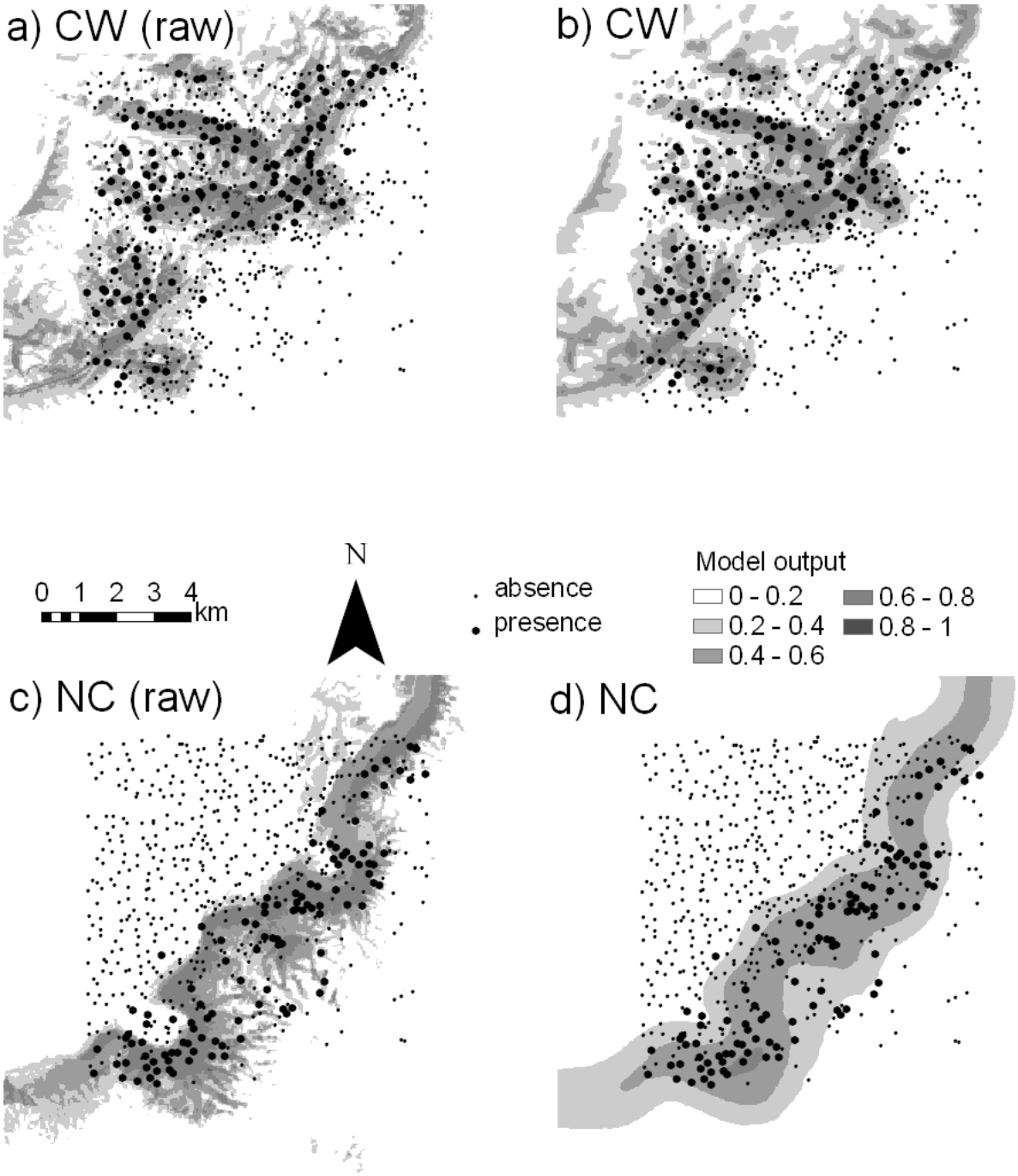

Fig. 7: The raw Maxent logistic model output (a, c) for *Ceratopetalum apetalum* (CW) and *Croton verreauxii* (NC) and the neighbourhood averages created using the optimal radii of 100m and 800m respectively (b, d).

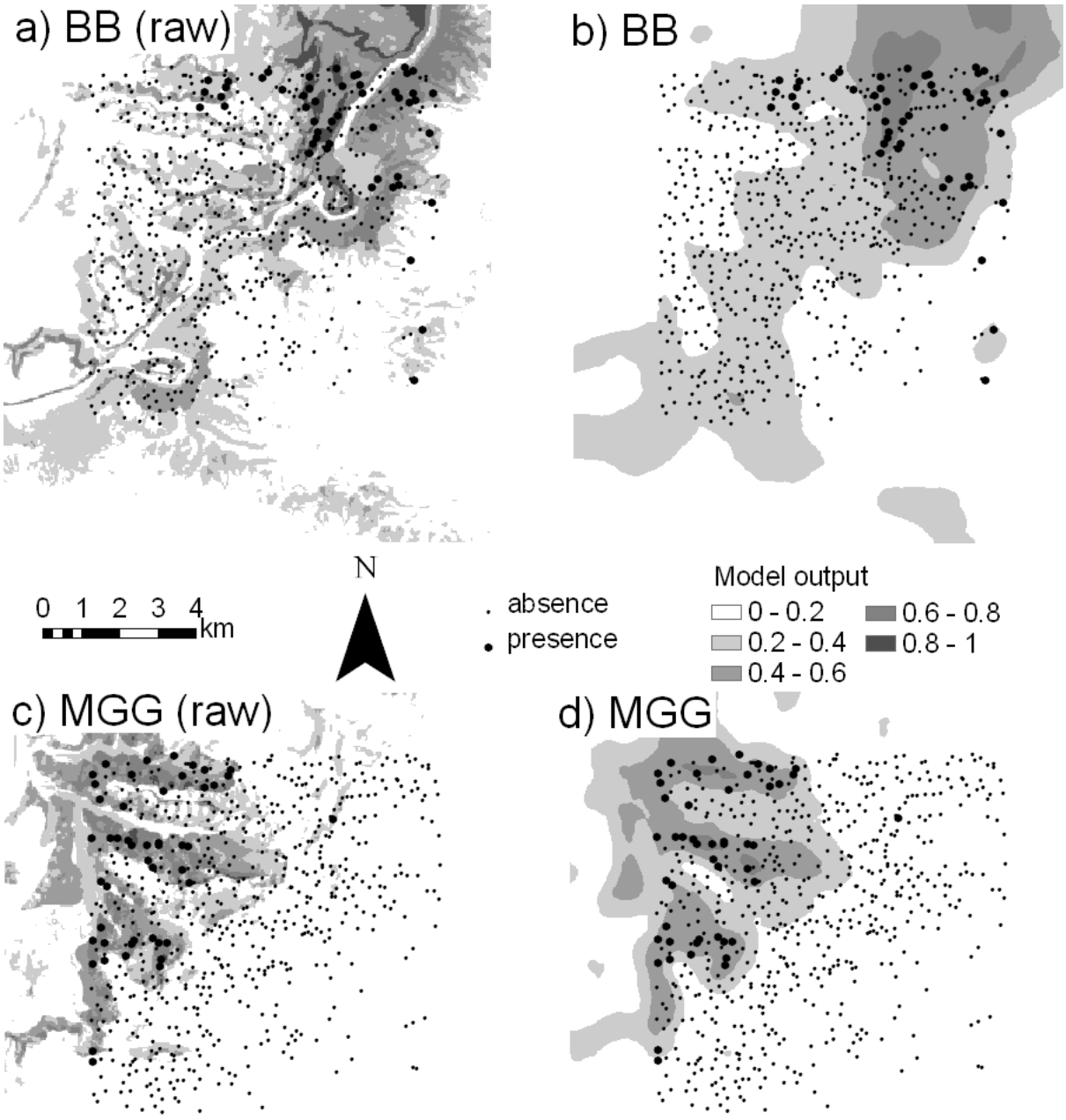

Fig. 8: The raw Maxent logistic model output (a, c) for *Eucalyptus pilularis* (BB) and *Eucalyptus cypellocarpa* (MGG) and the neighbourhood averages created using the optimal radii of 600m and 400m respectively (b, d).