# Environmental Modelling in CRIA
## Issues and Ideas

**Arthur D. Chapman**
**Draft Version 1 - June 2003**
**Last modified – 29 January 2004**

**Aim:**
To examine modelling in CRIA, to provide comments and ideas, to identify common pitfalls, to comment on issues of data quality, to comment on different methods of modelling species' distributions, to examine some of the strengths and weaknesses of different methods and to make recommendations on possible changes to GARP and on future modelling projects that may be carried out within CRIA.

**Constraints:**
The paper looks at models for use in modelling the distribution of species, not for modelling either habitat or landscape, however habitat and landscape models that are also used for species' modelling are considered.

Generally, the models discussed here are models that model species' presence and not abundance. Species abundance models Population Viability Analysis (PVA) models are becoming increasingly more valuable and popular, however require an entirely different set of data, are not currently used at CRIA and are thus are not considered.

**Data:**
***Species data***
Most species' distribution models use either presence-only data (usually including records from herbaria or museums as well as observation data) or presence-absence data from systematic surveys. Most of these data are point-based, although some models also include area-based or grid data as input. All species' data are samples of geographic space and inevitably incorporate some degree of spatial bias (Williams *et al.* 2002). Sample sites are generally subsets of sites where the species actually occur and there are few, if any records of where a species may have been looked for, but not found (Margules and Austin 1994).

Plant and animal specimen data held in museums and herbaria provide a vast information resource, providing not only present day information on the locations of these entities, but also historic information going back several hundred years (Chapman and Busby 1994). This data does, however, have some major drawbacks when it comes to use for modelling distributions. Many of the records carry little geographic information other than a general description of the location where they were collected (Chapman and Milne 1998), the geocoding (latitude and longitude) where given is often not that accurate, especially with historic data (Chapman 1999) and is often added at a later date by those other than the collector (Chapman 1992), it can only supply information on the presence of the entity at a particular time and says nothing about absences in any other place or time (Peterson *et al.* 1998), and it is usually collected opportunistically rather than systematically (Chapman 1999, Williams *et al.* 2002) resulting in large biases – for example, collections that are highly correlated with road networks (Chapman 1999, Peterson and Stockwell 2002).

Austin (2002) states that presence-only data can only be used for statistical models that are conditional on the presence of a species. Models for predicting the potential occurrence of a species say nothing about where a species may be absent within the climatic envelope.

We cannot neglect this data, however, as it constitutes the largest database of biological information we are ever likely to have. It is estimated that there over three billion records (OECD 1999) worth around $135-150 billion held in this form around the world (Armstrong 1992). The cost of replacing these data with new surveys would be prohibitive. It is not unusual for a single survey to exceed $1 million to conduct (Burbidge 1991). They are an essential resource in any effort to conserve the environment, as they provide the only fully documented record of the occurrence of species in areas that may have undergone habitat change due to clearing for agriculture, urbanization, climate change, or been modified in some other way (Chapman 1999).

***Environmental data***

Just as important as the species data, are the environmental layers used to model the species against. The theory of most environmental models is that species have certain habitat preferences that have an environmental basis. Many models use climatological information such as temperature, rainfall, radiation, evaporation, soil moisture etc. as their basis on which to broadly define the habitat or ecological niche. Other models use vegetation characteristics including vegetation classes, detailed habitat information, correlated species, etc.

One of the important considerations in choosing appropriate environmental layers is that of scale. Too fine a scale will lead to errors due to mismatching with the biological data being modelled against it. Too coarse a scale may not adequately delineate the appropriate environmental niches. Too often modellers give little consideration to scale in their selection of environmental layers. This may, of course be due to availability as they may only be able to use those layers that are available at the time.

The preparation of the environmental layers is often one of the most time-consuming, and computer-intensive areas of modelling. Fortunately, it only has to be done occasionally and once the surfaces are prepared can be used for any number of models. Climate surfaces, for example, have been prepared for much of the world's land surface and are available for use by researchers at nominal charge. I do not intend delving into the creation and fitting of environmental surfaces in this paper as it is too large a topic to be covered here and is the subject of a separate paper (Chapman 2003a).

**Data Accuracy**

Errors in data are common and to be expected. Errors in spatial position (geocoding) and in taxonomic circumscription are two of the major causes of error in modelling. Assessment of the accuracy of input data is essential otherwise the results of any modelling will be meaningless. Correcting errors in data and weeding out the bad records is a time consuming and tedious process (Williams *et al.* 1994). A detailed discussion of data error is beyond the scope of this paper and is the subject of a separate paper to CRIA (Chapman 2004).

Although one can produce a perfectly good looking model from poor data; a model that may even be regarded as being “better than random”, it does not mean that the model has any meaning in reality when predicting the occurrence or non-occurrence of a particular species. The better and more accurate the input data, the better the resultant model is likely to be.

**Models:**
Species models can generally be placed into a number of categories of types of models based on the basic style of the algorithm used. Some broad examples of modelling types (in no particular order, and in no ways complete, and some may be overlapping) with some examples are given below. Some of these may not have been used with species data, however there is no reason that they could not be so applied:

1. Climatic Envelop Models (e.g. BIOCLIM, HABITAT)
2. Linear Models (e.g. Generalised Linear Models – GLM)
3. Additive Models (e.g. Generalised Additive Models – GAM)
4. Logistic Regression (e.g. Classification and Regression Trees – CART)
5. Spatial Statistics (e.g. World Map, Ripley’s K Function)
6. Genetic algorithms (e.g. GARP)
7. Principal Components Analysis (e.g. FloraMap)
8. Neural Networks (e.g. Artificial Neural Networks – ANN)
9. Habitat (e.g. GAP Analysis, Habitat Suitability Index – HIS, Habitat Viability Analysis – HVA, Biomapper, Habitat Evaluation Procedure - HEP)
10. Decision trees/ Multi Criterian Analysis
11. Sums of Squares (e.g. Locally Weighted Sums of Squares – LOWESS)
12. Domain (e.g. DOMAIN)
13. Automaton Cell
14. Probability Sensitivity Analyses
15. Subjective or Baysian Interpretation of Probability
16. Landscape Ecology Models
17. Metapopulation Models
18. Dynamic Models
19. Comparative Risk Assessment Techniques
20. Population Viability Analysis (e.g. PVA)
21. Latin Hypercube Sampling (LHS)
22. Fuzzy Set Analysis
23. Point Pattern Analysis
24. Gravity Models, (incl. Time/Distance Decay)
25. and many others

Each model has its strengths and weaknesses. Some work well with Presence-only data (BIOCLIM, HABITAT, GARP, FloraMap), others also require absence data (GLM, GAM).

This paper does not intend to delve into the different between these types of models, nor does it make recommendations on what models should be used at CRIA or elsewhere. Suffice to say, that considerable investment has already gone into GARP, and this should not be wasted. At the same time, other modelling techniques should not be ignored, and should be examined in the light of the purpose for which the model is to be used, and the type of data one has at hand, and the ease of use and interpretation of the particular model.

Users of modelling algorithms, whether at CRIA or elsewhere, should be clear about their goals before they start. Questions such as

- "Why am I modelling this species?"
- "What am I wanting this model to tell me?"
- "What input data do I have, and is this algorithm suitable for this type of data?"
- "Is this the right model for this particular task?"

"The right model in the right place" is an important consideration in modelling, and is important not to expect that any one model will provide all the answers. As stated by Weins (2002) – "Clearly different questions demand different approaches". After all, a model output is just one of many hypotheses on the possible distribution of a species. Wein (2002) goes on to say "In the realm of wildlife-habitat management, the most elegant and logically consistent model remains an esoteric toy unless it can actually predict something useful" (Wein, 2002).

**Issues**

Prior to modelling a species, one needs to consider the reason one is modelling. To what purpose is the model to be used. It would seem to me that a lot of the modelling done so far in CRIA has been for the purpose of testing GARP and for developing the modelling algorithms. There is a need for this to continue, as there are many improvements that can still be made. However, consideration also needs to be given to other uses for the models. It is these possible uses that should drive the future developments of the GARP algorithms.

CRIA has contributed to some major advances in species modelling in recent years. The importance of the development of Desktop GARP should not be underestimated. But where to from here?

**1. Algorithm Development**

There are a number of possible algorithm developments that may be worth pursuing.

a. ***Develop a probability surface output.***

At present, GARP outputs only give a +/- (predicted presence/predicted absence). It has usually been regarded that if the predicted presence/predicted absence line is "better than random" then that is all that is needed. This may be OK for a straight statistical exercise, but has little value in environmental decision-making, or for environmental management. Early versions of GARP produced a "pseudo"-

probability through a sliding scale of conflict with the rules. Consideration should be given to developing a probability layer, that the user can decide where they wish to draw the line. For example, for the preservation of an endangered species, one may like to know areas where there is a greater than 20% likelihood of occurrence. Alternatively, if one has limited resources to monitor the possible introduction and spread of an invasive species, one may like to know higher likelihood areas – the 80% probability level, for example. (See also comments by Anderson *et al.* 2003).

***b. Inclusion of true "absence" data***

At present, Desktop GARP does not allow for the inclusion of true absence data, and uses pseudo absences in its calculations. The inclusion of true absences would be a major enhancement.

***c. Incorporation of better environmental layers***

i. The layers being used in Desktop GARP until recently have been at a range of scales, with the key climate layers at 0.5 degree resolution. I have been able to obtain new climate layers at 10 minute resolution for South and Central America, and these are now being incorporated into GARP. In the long-term, I believe that the ideal scale for species modelling is around 3 arc minutes. Recently (January 2004) globally-consistent 30-minute (or 1km) climate surfaces have been produced (Hijmans *et al.* 2004). These are still in Beta format and are not likely to be released to the public until March or April 2004. At 30-minute resolution, they are perhaps a little to fine for use with GARP – firstly because the biological data may not be that accurate, and secondly because of the computing resources needed to use climate surfaces at that resolution. The developers of the 30-second data set are planning to reclassify the data to 2.5-minute and 5-minute resolution in the near future. It is the 2.5-minute resolution data that I recommend be used for modelling with GARP once they are publicly available.

ii The selection of environmental layers is something that needs considering. In the past, two months (January and July) were chosen to be representative of the climate. I have recommended (and I understand that this is beginning to be implemented) that consideration be given to using rainfall of the warmest quarter, rainfall of the coolest quarter, temperature of the warmest quarter, temperature of the dryest quarter, etc. Ideally, minimum and maximum temperatures should both be used , however, these layers are not presently available for South America at this scale. I would also recommend using the coefficient of variation of Rainfall. The reasons I suggest this, is that these layers have a lot more to do with where a plant (or animal) may occur than some of the raw layers (rainfall in January, etc.). A long dry period at the hottest period of the year will provide a lot more stress to a plant, than a long dry period in the cool part of the year. Similarly, rainfall during a warmer period will produce greater growth in many species than rainfall during the winter months, etc.

iii. Consideration needs to be given to the method of resampling the environmental layers. Until now, the Nearest Neighbour method was used to resample the grids. If, for example one is resampling a five by five grid to create one grid square, the Nearest Neighbour method gives the new larger grid the value of the single grid in the centre of the 5 X 5 array. Better methods for continuous data are either Bilinear Interpretation, which takes the weighted average of the 4 nearest cell centres, or Cubic Convolution, which fits a smooth curve through the sixteen nearest cell centres.

iv. I suggest that rather than sample all grids upwards to the smallest scale, that consideration be given to resampling to an intermediate scale (i.e. some upwards and some downwards). It makes little sense in resampling topographic layers (elevation, slope, aspect) to half a degree or even ten minutes. Slope and aspect have no meaning when scaled to a 60 km resolution. I thus recommend (and again, I understand this is being implemented) that the climate be resampled down to 5 minutes, and the topography up to the same scale (about 10 km). Even this is too coarse, but it is the best that can be done for now.

v. Consideration could be given to the incorporation of non-climate layers into GARP. Layers such as soil moisture, soil texture, National Difference Vegetation Index (ndvi) are some that come readily to mind. I would recommend, though, that where possible, only layers of a continuous nature rather than a categorical nature be used.

***d. Improved visualisation***
Visualisation of the GARP output could be significantly improved. An example is the FloraMap visualisations that include linking the input points to the mapped output, includes simple data cleaning tools, and includes a number of statistical visualisations. These all help the user in explaining the model output and in linking the outputs to the on-ground environment.

***e. Extension into the aquatic environment***
This may be a longer-term research project, but there is a lot of research going on in Brazil in the fresh-water aquatic environments. It makes no sense to use broad climatic layers in a model for modelling fish species. The water temperature, oxygen content, pH, water flows etc. may be far more important. The air temperature in a region, may have little affect on the water temperature of a stream. Much of the data to create such layers is being gathered by researchers in many areas, especially in São Paulo State, and this would be a great place to conduct such a trial.

***f. Extension into Marine areas***
Consideration could be given in the longer term to extending GARP modelling methodologies into the marine arena. There are very few, if any, good models available for modelling species in a marine environment. I believe that GARP would be ideal in this scenario. This is a major research project however, and would need funding.

***g. Incorporation of what-if scenarios***
This is again a long-term research project, but involves the inclusion of “what-if” scenarios into the algorithms. For example, what if we change the flow pattern of this river – what effect will that have on species downstream”, etc. This issue is become increasingly important in environmental decision making. If we dam this river, will it effect the species downstream, and how?

***h. Validation***

I have not examined the different methods of validation being run to test the GARP models, but it has been brought to my attention by Rafael Fonseca of CRIA, that there appears some problems, especially with the use of the chi-squared test. The result of the chi-squared test seems to vary with the size of the boundary around the model. In the example shown in figure 1 – a set of random points were selected and a GARP model run. Two boundaries were selected – one much closer (bottom) to the edges of the model than the other (top). When a chi-square test is run on these two models, the top one is accepted as being better than random ($p<0.05$), and the bottom one worse than random ($p>0.05$). This needs further examination.

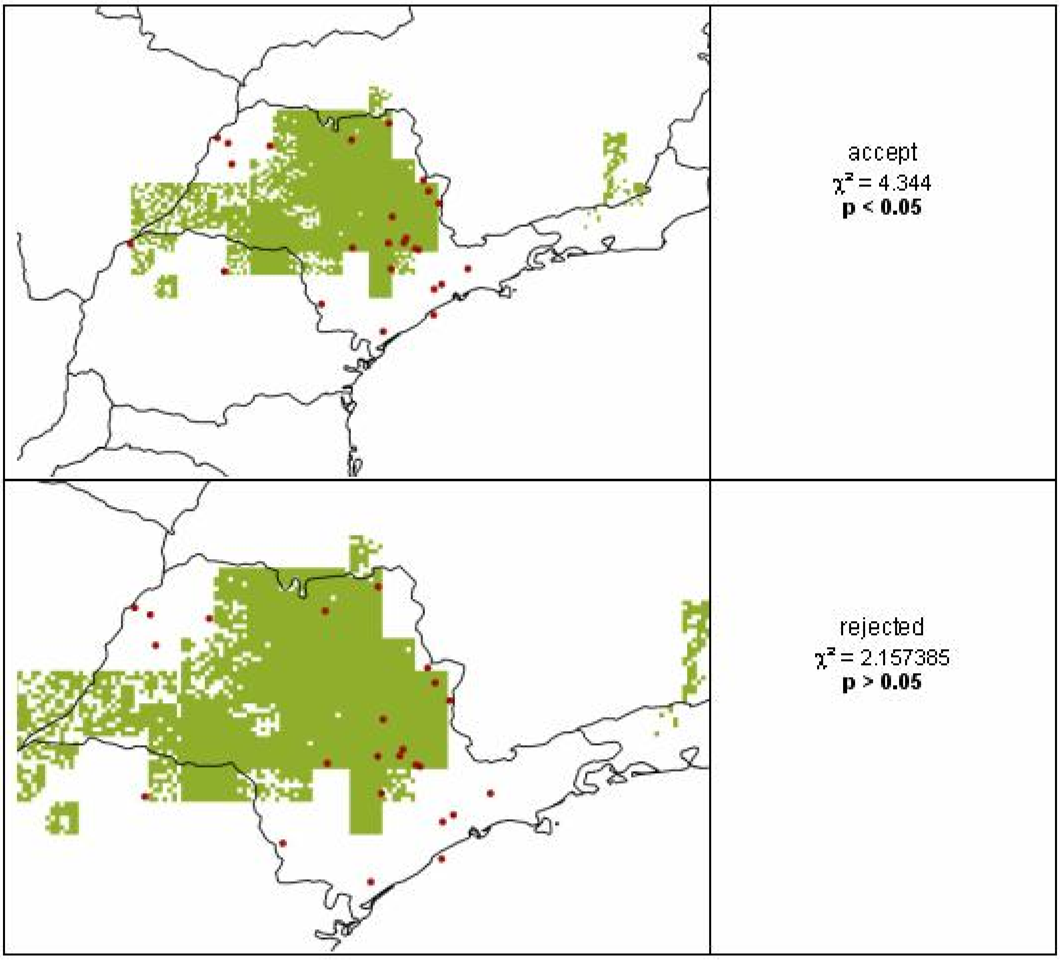


**Fig. 1.** Examples of a GARP model run using random points. The top model has a broad boundary and the chi- square test accepts the model as being better than random. The bottom model is zoomed in and has closer boundaries to the edge of the model. The chi-square test rejects the model as being worse than random.

**2. Possible Future Research Projects**

Beside the long-term research projects mentioned above, I see several other projects that will lead to both the improve of the GARP algorithms, as well as begin to provide environmental relevance to the GARP modelling being carried out. These include:

a. ***Affect of tree age***

Is modelling long-lived trees using today's climate a sensible thing to do? A tree that is several hundreds of years old may be surviving in today's climate, but that may not mean that that is the ideal climate. Is it reproducing, for example? One could test a number of hypotheses by testing against the climate data of the first half of the last century, compared to climate data from the latter part of the century.

***b. Examining C3/C4 boundaries***

It may be possible (including with the use of ndvi data) to look at modelling the boundary between C3 and C4 grass species. This boundary is highly dependent on climate and may well be an excellent indicator of climate change.

***c. Continental vs Regional Scale vs Local Scale***

There are a number of ways of examining the affect of the scale of the environmental layers on modelling at different environmental scales. Whereas climate layers at 0.5 degree may provide reasonable results at the Continental scale, do they show anything of value at the Regional Scale. Is the use of 5 km scale layers meaningful at the Continental Scale, etc.

***d. Testing the effect of abundance***

Testing of the effect of abundance, or duplicated specimens on the model needs to be carried out. Should one exclude duplicate records from the same grid cell or not? Examples of experiments could include running a model with just species presence and see the result. Then randomly increase the "abundance" at those points – firstly using a small (possibly linear) increase and later with a larger (possibly logarithmic) increase. See what effect there is on the modelled results.

***e. Studying the effects of species sub-setting***

A study could be conducted by running a model of a wide-spread species to just one region (i.e. modelling only a portion of the range and thus only accepting part of the environmental envelop as important). This could then be compared with the model for that region that also included the specimens from outside the region.

***f. Effects of grid size***

Run an series of models by resampling a small environmental grid and see the effects on the results.

***g. Inclusion of ndvi as a layer***

Normalised Difference Vegetation Index (ndvi) – a derived layer from Remote Sensing that has a strong link to vegetation, could be included as a layer in GARP as suggested above. This could be used (without necessarily knowing what the colours actually mean) to exclude cleared areas, agricultural land, etc. The inclusion of layers from different months would be able to increasing differentiate between different vegetation types.

***h. Test for inclusion of false-negatives***

GARP as used at present, includes false negatives where there are no presence records. This may work well in some cases (see papers by Stockwell and Peterson, etc.), but I have some doubts about applying this uniformly. It would be worth carrying out some tests using real absence data and compare the results with the pseudo-absence method.

***i. Training set selection methodology***
A series of tests need to be conducted on the method of selection of training sets. At present the tendency is to select by grid rather than random. I would like to see a comparison between these two methodologies.

***j. Validation methods***
Methods of validating models such as those produced by GARP need further investigation and testing.

***k. Temporal Modelling***
One area of species modelling that has yet to be carried out is the inclusion of a temporal scale to the model. For example, migratory species move over the space of a year in line with climate, and this may be something that could be modelled.

**Modelling myths and pitfalls**

When modelling species, it is important to think about what one is modelling, and about the data one is using – both the species data and the environmental data. Below I will mention a few common problems and pitfalls that occur in species modelling.

**1. Animals**
When modelling some animal species, all the available data is included in the model, but this may not be a sensible thing to do. Take large bats for example, Bats generally have a roosting site and a feeding site, and both these are of major importance to the animal. Quite a bit of time, however, is spent flying from one place to the other and back again. Often collectors will set up mist nets along this line to catch the bats and that is the recorded locality. When modelling, however, the environment at this intermediate point may have little to do with where the animal either roosts or where it feeds, and is completely irrelevant to the model. These records thus, may be best excluded.

**2. Jumping to conclusions**
All too often, a model is seen as being "the answer", rather than just a hypothesis. This can be particularly dangerous in a decision making process. It is important not to jump to conclusions about the model, and to look critically at the things the model is telling about the possible location of the species. The uncertainty within a prediction – any prediction – should not be ignored. Wherever possible, the uncertainty in a prediction should be documented and emphasised. One then has the task of selling that uncertainty to the decision makers – or "Selling Uncertainty to the Uncertain" as suggested by Mowrer (1999).

**3. Inaccurate input data**
A model is, at best, only as good as the input data – be it the specimen data, or the environmental layers themselves. Too often, a model is run after hours and hours of data cleaning for the specimen data, but with blind acceptance of the environmental layers that the species are being modelled against. The specimen data may only be

used the once, but the environmental layers are used over and over, and it is important that these be critically looked at for quality before being used.

**4. Better than Random**
It is often assumed that because a model is better than random it is a good model. I am sure, given my knowledge of many species, that in a reasonable number of cases, given a number of known points, I could draw a distribution on a map that would prove to be better than random. "Better than random" is **NOT** what environmental managers and policy makers require, and just because a model has some statistical significance does not mean the resultant model is of environmental significance or will be of value in environmental decision making. As environmentalists, we cannot be satisfied with models that are just "better than random".

**5. Pretty maps**
One can lie with maps just as easily as one can lie with statistics" (Weins 2002), and probably have them believed easier.

**6. Model uncertainty**
I have already mentioned the problem of model uncertainty. Clearly, there is a need to develop new approaches to spatiotemporal model validation. Part of this effort will require educating both producer and user communities about how to grapple intelligently with data and model uncertainty (Mowrer 1999, Henebry and Merchant 2002).

**7. Niche modelling**
We often here the term niche modelling for models at all sorts of scale, when often, I believe, we mean environmental or habitat modelling – especially at the smaller scales. At a scale of 10 minutes or half a degree (the scale at which much of the environment modelling in North and South America has been conducted) with the use of just climatic layers, all one is doing is modelling the broad scale climate requirements of the species. At that scale, one is not delineating the true niche of the species. It is only when one begins modelling at the finer (regional or local) scale and begins to use layers other than just climate, that I believe you can truly begin to be said to be niche-modelling.

**8. Modelling long-lived species**
In a number of cases, I have seen modellers modelling a long-lived species using the today's climate. In one case I know of, the Huon Pine in Tasmania (a species that lives to 2000 plus years) was modelled, and conclusions about its climatic preferences inferred. In this case, one is making the huge assumption that just because the species is alive today, that that is the ideal climate for it. But – this doesn't take into account the fact that the species may no longer be reproducing because the climate is unsuitable, but has managed to survive for any number of reasons. If one is doing this sort of modelling, then one needs to be aware of this and take account of the many assumptions being made.

**9. Suitable habitat**
Following on from the last point, as stated by Weins (2002). "If a species is present in an area, does that mean that the habitat is suitable? Suitable for what? If areas are

surveyed only once (as is often the case), records of presence may include transients that don't really "belong there".

**10. Boundaries and transitions zones (rare species)**
Many species (especially rare species) occur along transition zones between vegetation types. Often these transitions zones are not very well delineated in models, especially if the scale is too coarse to pick them up.

**11. Scale is not a discrete entity**
We often talk of scale as though it is a step-wise process, but it is not – it is continuous.

**12. Models don't model the same things!**
Elith and Burgman 2002, state that "Methods that create a rule or measure of environmental distance (such as GARP) often predict highest likelihood at the original record site. In contrast, methods that fit a prediction surface (such as GLMs and GAMs) do not necessarily predict high probabilities at presence locations, especially if identical environmental conditions are also associated with absence records. Each model has a different way of fitting the prediction to the original points and to the environmental layers being used. What method you use, may be determined by a number of factors, but be aware of what the model is doing before blindly accepting the results.

**Conclusion**
I believe GARP has come a long way since its early days as a command driven program back in Australia. The development of Desktop GARP has improved its useability and made it available to a much wider audience. CRIA's role in that should be recognised and applauded.

I also believe, however, that in many cases, the reason for using the model and for doing modelling, has been lost site of, and that the results have not been looked at critically enough. In many cases, environmental layers have been used without critical examination of why those layers are being used, and whether they are the best layers in the circumstances.

Both CRIA and GARP are at a critical step in their future development and use of modelling. I believe that modelling is an important adjunct to the work being carried out at CRIA, and is an essential tool (if used properly and critically) to future environmental management within Brazil.

The ideas and suggestions above, are just a few that I hope will help progress future modelling within CRIA.

**Updated Comments – January 2004.**

***Modelling Framework***. Many of the comments and suggestions made above have now been implemented at CRIA. Perhaps the largest development has been to look at the type of model being used and has led to the development of a Modelling Framework in CRIA that will allow the selection of model type and data layers so that choosing the "right model for the right place" as suggested above will soon be possible. This will be a major advancement in modelling, not only in CRIA, but internationally, and may lead to a new way at looking at environmental modelling.

***Resolution of Data Layers.*** When I first came to CRIA in March 2003, the scale at which most modelling was being carried out at a scale of about 0.5 degrees. In June 2003, I was able to obtain layers at 10-minute resolution – an improved resolution by a factor of 9 times. In January 2004, this has been increased to 30-seconds – a further improvement of 400 times, and a total improvement of 3600 times. I believe, however that this is too fine for use at CRIA – see comments earlier in this paper, and the further development of 2.5-minute surfaces in the next few months will prove just about ideal for modelling in South America. Layers at this scale will be an improvement in resolution in 12 months of 100 times, and over the modelling being done at the moment of 16 times. Modelling at this scale will have great environmental significance and make the models valuable for use in environmental decision making for conservation planning.

***Layer types.*** When I arrived at CRIA, the layers being used for modelling made very little sense from an environmental point of view (i.e. using rainfall and temperature for January and July). Following an early draft of this paper, these were altered to use environmentally meaningful layers such as rainfall of warmest and coolest quarters, temperature of wettest and driest quarters, etc.

***Resampling methods.*** The method of resampling grid layers fro use in GARP modelling has been altered from Nearest Neighbour to bileanear Interpretation and Cubic Convolution as recommended in an earlier draft.

***Research***. A number of the suggested research topics have been taken up by researchers at CRIA. These include looking at the effect of scale on modelling, the effect of sub-sampling, and the methods of model validation.

***Papers.*** A number of papers will arise out of my work at CRIA. These include a joint paper being written with Mauro Muñoz and Ingrid Koch on environmental models (Chapman *et al.* in press), a joint paper with Ricardo Scachetti-Pereira and Jorge Soberón on eliminating errors in modelling data (Scachetti-Pereira *et al.* in press), and well as individual papers arising out of some of the research recommendations in this paper.

***Conclusion.*** The speed at which many of my suggestions and recommendations have been accepted and acted upon is amazing. CRIA is at the forefront of species' model developments in the world, and especially with its planned developments on the Modelling Framework. The improvement is scale of environmental layers available, the move to more significant environmental layers and the move toward examing more critically model outputs are major steps forward. I look forward with interest to future developments.

**References:**

Anderson, R.P., Lew, D. and Peterson, A.T. (2003). Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modelling* **162:** 211-232.

Armstrong, J.A. (1992). The funding base for Australian biological collections. *Australian Biologist* **5(1):** 80-88.

Austin, M.P. (2002). Case Studies of the Use of Environmental Gradients in Vegetation and Fauna Modeling: Theory and Practice in Australia and New Zealand pp. 73-82 **in** Scott, M.J. et al. eds. *Predicting Species Occurrences. Issues of Accuracy and Scale*. Washington: Island Press.

Burbidge, A.A. (1991). Cost Constraints on Surveys for Nature Conservation **in** Margules, C.R. and Austin, M.P. (eds). *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*. Canberra: CSIRO

Chapman, A.D. (1992). Quality Control and Validation of Environmental Resource Data **in** *Data Quality and Standards: Proceedings of a Seminar Organised by the Commonwealth Land Information Forum, Canberra, 5 December 1991*. Canberra: Commonwealth land Information Forum.

Chapman, A.D. (1999). Quality Control and Validation of Point-Sourced Environmental Resource Data pp. 409-418 **in** Lowell, K. and Jaton, A. eds. *Spatial accuracy assessment: Land information uncertainty in natural resources*. Chelsea, MI: Ann Arbor Press.

Chapman, A.D. (2003a). *The Case for a 3 minute Climate Surface for South America*. Internal report to CRIA, Draft - May 2003.

Chapman, A.D. (2004). *Notes on Environmental Data Quality – A Discussion Paper*. Internal report to CRIA, Draft in Preparation – January 2004.

Chapman, A.D. and Busby, J.R. (1994). Linking plant species information to continental biodiversity inventory, climate and environmental monitoring 177-195 **in** Miller, R.I. (ed.). *Mapping the Diversity of Nature*. London: Chapman and Hall.

Chapman, A.D. and Milne, D.J. (1998). *The Impact of Global Warming on the Distribution of Selected Australian Plant and Animal Species in relation to Soils and Vegetation*. Canberra: Environment Australia

Chapman, A.D., Muñoz, M.E.S. and Koch, I. (in press). Environmental Information: Placing Biodiversity Phenomena in an Ecological and Environmental Context. *Biodiversity Informatics* **1:**

Elith, J. and Burgman, M. (2002). Predictions and Their Validation: Rare Plants in the Central Highlands, Victoria, Australia 303-313 **in** Scott, M.J. *et al.* eds. *Predicting Species Occurrences. Issues of Accuracy and Scale.* Washington: Island Press.

Henebry, G.M. and Merchant, J.W. (2002). Geospatial Data in Time": Predicting Species occurrences pp. 291-302 **in** Scott, M.J. et al. eds. *Predicting Species Occurrences. Issues of Accuracy and Scale*. Washington: Island Press...

Hijmans, R.J., Cameron, S., Para, J., Jones, P., Jarvis, A. and Richardson, K. (in prep.). *Worldclim Version 1.1.* Berkeley, CA: Museum of Vertebrate Zoology. [http://bnhm.berkeley.museum/gisdata/worldclim/worldclim.htm].

Margules, C.R. and Austin, M.P. (1994). Biological models for monitoring species decline: the construction and use of data bases. *Philos. Trans. R. Soc., London* **B344:** 69-74.

Mowrer, H.T. (1999). Accuracy (Re)assurance: Selling Uncertainty Assessment to the Uncertain 3-10 **in** Lowell, K. and Jaton, A. eds. *Spatial accuracy assessment: Land information uncertainty in natural resources*. Chelsea, MI: Ann Arbor Press.

OECD (1999). *Final Report of the Megascience Forum Working Group on Biological Informatics*. Paris: OECD

Peterson, A.T., Navarro-Siguenza, A.G. and Benitez-Diaz, H. (1998). The need for continued scientific collecting: A geographic analysis of Mexican bird specimens. *Ibis* **140:** 288-294. [not seen]

Peterson, A.T., Stockwell, D.R.B. and Kluza, D.A. (2002). Distributional Prediction Based on Ecological Niche Modelling of Primary Occurrence Data pp. 617-623 **in** Scott, M.J. *et al.* eds. *Predicting Species Occurrences. Issues of Accuracy and Scale.* Washington: Island Press.

Scachetti-Pereira, R., Soberón, J. and Chapman, A.D. (in press). Data Enhancement II. Detecting and Eliminating Errors from Biodiversity Datasets. *Biodiversity Informatics* **1:**

Wein, J. (2002). Predicting Species Occurrences: Progress, Problems, and Prospects pp. 739-749 **in** Scott, M.J. et al. eds. Predicting Species Occurrences. Issues of Accuracy and Scale. Washington: Island Press.

Williams, P.H., Margules, C.R. and Hilbert, D.W. (2002). Data requirements and data sources for biodiversity priority area selection. *J. Biosc.* **27(4):** 327-338.