Department of Spatial Sciences

Modelling and Communicating the Effects of Spatial Data Uncertainty on Spatially Based Decision-Making

Marco Antonio Marinelli

This thesis is presented for the degree of
Doctor of Philosophy (Spatial Sciences)
of
Curtin University

May 2011

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signature: .....

Date: .....

Abstract

Important economic and environmental decisions are routinely based on spatial/temporal models. This thesis studies the uncertainty in the predictions of three such models caused by uncertainty propagation. This is considered important as it quantifies the sensitivity of a model's prediction to uncertainty in other components of the model, such as the model's inputs. Furthermore, many software packages that implement these models do not permit users to easily visualize either the uncertainty in the data inputs, the effects of the model on the magnitude of that uncertainty, or the sensitivity of the uncertainty to individual data layers. In this thesis, emphasis has been placed on demonstrating the methods used to quantify and then, to a lesser extent, visualize the sensitivity of the models. Also, the key questions required to be resolved with regards to the source of the uncertainty and the structure of the model is investigated. For all models investigated, the propagation paths that most influence the uncertainty in the prediction were determined. How the influence of these paths can be minimised, or removed, is also discussed.

Two different methods commonly used to analyse uncertainty propagation were investigated. The first is the analytical Taylor series method, which can be applied to models with continuous functions. The second is the Monte Carlo simulation method which can be used on most types of models. Also, the later can be used to investigate how the uncertainty propagation changes when the distribution of model uncertainty is non Gaussian. This is not possible with the Taylor method.

The models tested were two continuous Precision Agriculture models and one ecological niche statistical model. The Precision Agriculture models studied were the nitrogen (N) availability component of the SPLAT model and the Mitscherlich precision agricultural model. The third, called BIOCLIM, is a probabilistic model that can be used to investigate and predict species distributions for both native and agricultural species.

It was generally expected that, for a specific model, the results from the

Taylor method and the Monte Carlo will agree. However, it was found that the structure of the model in fact influences this agreement, especially in the Mitscherlich Model which has more complex non linear functions. Several non-normal input uncertainty distributions were investigated to see if they could improve the agreement between these methods. The uncertainty and skew of the Monte Carlo results relative to the prediction of the model was also useful in highlighting how the distribution of model inputs and the models structure itself, may bias the results.

The version of BIOCLIM used in this study uses three basic spatial climatic input layers (monthly maximum and minimum temperature and precipitation layers) and a dataset describing the current spatial distribution of the species of interest. The thesis investigated how uncertainty in the input data propagates through to the estimated spatial distribution for Field Peas (Pisum sativum) in the agriculturally significant region of south west Western Australia. The results clearly show the effect of uncertainty in the input layers on the predicted specie's distribution map. In places the uncertainty significantly influences the final validity of the result and the spatial distribution of the validity also varies significantly.

Acknowledgements

I would like to thank my supervisors Dr. R. Corner and Prof. G. Wright for their advice and direction during the course of this work.

I would also like to acknowledge Amos Maggi, Simon Cook, Peter Fearns, Jim Davies, Prof. M.J. Lynch, Margot and Mike Harness, Barbarra Hart and Dr Susan Ho.

Finally, a special thanks must also go to my parents, and to Chris and Melissa for their support and patience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With increasing frequency, decisions are made with inputs that are the result of modeled real world situations. In a historical context, modeling can be seen as a relatively new tool that has evolved with advances in science, mathematics and statistics. The degree to which a model accurately reflects a true system is dependent on a range of factors including the quality of the data from which the model was developed and the understanding of the "laws" that control the system being modeled. For example, modeling the motion of a ball down a smooth slope is easy as the physical laws of motion are well understood. In a complex system this is rarely the case, especially in an earth/biological system where there can be many unknowns in the interactions that occur within the system. For these, the model developed is usually empirical and hence based upon observations rather than physical and biophysical laws. Ideally, in the physical sciences, physically based models are preferred. However, if a system is complex and data limited, statistical or combined physical models are used. Extensive decades of research have found that these models, if well tested, can simulate a true system even though the system's components are not well understood.

The models used to study the Earth's systems can be classified as Geospatial Information System (GIS) models, which is the application of the Geographical Information Science. Traditionally, this was not the case as most models evolved from studies in the older established sciences. For example, climate

models evolved primarily from the work of physical scientists, statisticians and mathematicians. But, as they are a representation of the earth's climate, they are also temporal geospatial models. The same could be said for most fields where a spatial component is included in an analysis, such as models used in the the study of agriculture and biophysical systems.

All models have a level of uncertainty in their prediction and minimising this uncertainty is an important aim of a model developer. The initial way to test this is to see how well a model simulates a true situation with no uncertainty in the model's input data. If the prediction's accuracy is high then it can be assumed that the knowledge of the system, and the algorithms in the model's construction, are mostly correct. However, as there is always uncertainty in the input data, another key question arises: *Is the model's accuracy sensitive to the uncertainties in the inputs?* This is important as it raises two important questions.

1. How does the prediction change in the domain of these uncertainties.

2. How does the *uncertainty of the prediction* change in this domain.

In the first, the prediction will change depending on where, within the range determined by the uncertainty, the input(s) values are taken. The prediction may be correct or incorrect in this range, so models are tested against a known situation. For example, if a meteorologic model accurately predicts the path of all cyclones which occurred in the last fifty years, then it will be more trusted in predicting the path of a tropical cyclone during the next summer period. In short, what is important is the degree to which the uncertainty determined range in the known meteorological inputs changes the prediction of the cyclone's path. This must be minimal, or the influence of the uncertainty well quantified, for the model to be trusted.

The second question regards the sensitivity of the prediction's uncertainty to a change in the model inputs. This is important as the prediction's uncertainty can be more sensitive to the input uncertainty than the prediction. Therefore,

a small change in the model inputs may result in a large change in the uncertainty of the prediction. Also, the sensitivity of the uncertainty of the prediction may change significantly across the models domain. In recent decades, in the Geospatial Sciences, there has been an increasing interest in this, with research aimed at more accurately quantifying this relationship. This additional information can contribute to the assessment of the model's validity, improvement in its development and the degree to which an end user can *trust* the model. In the literature, this area of research is often referred to as *error* or *uncertainty propagation* analysis.

The degree to which an end user's trust of a model is influenced by the uncertainty propagation analysis does depend in part on the background of the user, how much they understand about what a model tells them and whether they consider the uncertainty significant in their use of the model. Also, as the prediction may be used along with other data in making a decision, it becomes a part of a broader assessment of a *risk*. In non-scientific fields such as insurance, assessing a problem from this viewpoint is normal. The same can be said in many other fields, including agriculture. This does not mean that uncertainty propagation analysis may not be of value in the traditional risk analysis methodology. In fact, it can be argued that the opposite is the case as it quantifies the trustworthiness of the prediction provided by the model, which in turn helps in the assessment of risk. Uncertainty and risk are further discussed in Chapter 2.

1.0.1 The Models Investigated

Modeling is extensively used in a wide range of environmental study areas. These include meteorological and oceanographic studies (e.g. "Adjoint Models" for sensitivity analysis (Errico 1997) and their use in 3-D and 4-D data assimilation models (Pu, Kalnay, Derber & Sela 1997)), species distribution and abundance models using climatic variables, spatial prediction and surface modeling (as discussed by Atkinson (2005)), spatio-temporal analysis using multiple scale ecoregionalisation (Handcock & Csillag 2004) and agricultural models.

This thesis investigates uncertainty propagation in two *Precision Agriculture* models and the *Ecological Niche* BIOCLIM model described in Nix (1986). For the Precision Agriculture models, two different analysis methods are applied; the Taylor Series and Monte Carlo simulation methods - which are often used in the investigation of uncertainty propagation. The algorithms in the Precision Agriculture models have been published, so the details of how the model is structured and the underlying theory can be seen. These models are continuous and are not probabilistic so the Taylor Series method can be used as well as the Monte Carlo method (which is not constrained by these factors). The Ecological Niche model is more complex, has continuous and noncontinuous functions and also has a probabilistic component. Therefore the only method that can be applied in its analysis is the Monte Carlo method. In complex models, even when the functions are non probabilistic and continuous, the Monte Carlo Method is the method of choice because it is generally easier to apply.

1.1 Project Aims

Accordingly, the aims of this research are to:

1. Investigate the methods used to quantify a model's sensitivity to propagated uncertainty. If the results are not clearly visible in the plots and images of the results, investigate methods to resolve this.

2. Assess the sensitivity of two types of GIS models. To increase the scientific and applied practical value of this thesis, the chosen models are very different in their structure and applied purposes.

3. When both the Taylor and Monte Carlo methods can be used, compare results. It is assumed that the analysis methods should produce the same results. Therefore, if there is a difference, determine why it occurred and if it can be explained by the structure of the model.

4. Where possible, investigate how the model's structure influences the sensitivity of the uncertainty of the model.

5. Where possible, investigate the components of the models that are contributing to the uncertainty in input − prediction relationship. This includes the domain of the input uncertainty and its distribution.

6. Combine these to produce a result that improves the confidence that a user will have in their understanding of the models accuracy.

1.2 Thesis Outline

The thesis has three parts. The first introduces the reader to the theory of uncertainty and how this knowledge is applied in this thesis. The second describes the theory, methods and results of the analysis of the Precision Agriculture models and the third describes the same for the Ecological Niche Model investigated:

Uncertainty Theory

Chapter 2 discusses uncertainty and risk. Uncertainty propagation and the methods used in its analysis are detailed as are questions relevant to the estimated uncertainties in the input data layers in a spatial model. The models studied in this thesis are further discussed.

Precision Agriculture Models

Chapter 3 discusses the theory of Precision Agriculture and the influence of uncertainty on their predictions. The models studied in this thesis are described in this chapter. Section 3.5 discusses the uncertainty in the model inputs and details how their propagation was analysed using the Taylor and Monte Carlo Methods.

Chapter 4 presents the results and conclusions of the analysis.

Ecological Niche Model

Chapter 5 discusses the basic theory of ecological niche models. It discusses the key factors which determine their accuracy. BIOCLIM, the Ecological Niche

model studied in this thesis, is detailed.

Chapter 6 discusses the input climate grids of the BIOCLIM model. The interpolation methods used in their generation is detailed as are the estimated uncertainty in these interpolated grids. This is one source of uncertainty included in this thesis.

Chapter 7 describes the application and testing of the Monte Carlo method in the analysis of the BIOCLIM model. Also, as the climate grids represent a monthly and seasonal climatology, this section describes how the uncertainty in these, caused by inter-decadel variation in rainfall and temperature, was quantified and a second uncertainty grid interpolated (for this thesis).

Chapter 8 presents the uncertainty propagation analysis of the BIOCLIM model in several scenarios. Briefly, one group of predictions was made for a current climate. The difference was only in the uncertainties in the input climate grids. The second group investigated if the uncertainties in the input climate grids had significant influence on the conclusion that could be made from future BIOCLIM predictions. The sensitivity of the prediction to the uncertainty in the input grids is not linear, so conclusions were not easy to make from simple plots. Therefore, this Section also describes the method used to quantify this relationship and conclusions made are discussed.

Chapter 9 analyses why the observations, in the present prediction Model discussed in Chapter 8, are occurring. This analysis focuses on the structure of the model.

Summary of Thesis

Chapter 10 provides a summary of the thesis' conclusions and contains suggestions for future research.

Chapter 2

Uncertainty Theory in Modeling

While there are uncertainties in all the components of a model, the uncertainty in the model's result is of primary interest to most end users. For the model's developer, minimising as well as effectively communicating this uncertainty requires knowledge and understanding of:

1. The system that the model is simulating.
2. The known sources of uncertainty in the model, especially in the model inputs.
3. If possible, the algorithms in the model and their known limitations.
4. The methods that can be applied in testing the sensitivity of a model to these limitations and all known uncertainties.

In a complex model the influence of these components can be intertwined and unclear, which then makes the methodology that can be applied in their analysis difficult to choose. Therefore, the initial step is to clearly understand and then categorise the components of the uncertainty.

2.1 Defining Uncertainty

As discussed by Rowe (1994), uncertainties can be classified into four groupings:

Temporal

Uncertainty in future states and past states. The sources of this are variable, ranging from the inherent randomness of nature, inconsistent human behaviour, non-linear dynamic (chaotic) systems behaviour and the uncertainty associated with measurement such as sparse rare events and rare events embedded in noise.

Metrical

Uncertainty due to measurement. We make observations about the empirical world using nominal, ordinal, cardinal or ratio scale (Rowe 1994). Accuracy addresses how correctly we have measured and interpreted measurement about scale values. Statistical models are used to describe the results, but one must keep in mind that the accuracy of the model is heavily influenced by the underlying process by which the data are generated.

Structural

Uncertainty due to complexity. This falls into two classes: The first, Retrodiction, which includes variables such as incomplete historical data (measurement error) and changing system parameters (systematic error). The second, interpretation of data, includes factors such as a lack of external data references (which limits the validity of the data), imposition of political correctness and conflicting reports.

Translational

Uncertainty in explaining uncertain results. This is caused by the poor communication of results to the "stakeholders." As the final users have differing degrees of knowledge (regarding the issue studied) and differing interests and biases, the method by which the results are displayed can be critical in their accurate use.

In all situations, all four of these classes occur. But, the dominance of each class will vary depending on the situation. Furthermore, even though it is considered that although each of these classes is not necessarily independent, the nature of each of the classes is quite different and hence each class can be addressed separately and then their interaction examined (Rowe 1994).

The user of a model's results should have an understanding of the uncertainty in the models results. But, the way that this uncertainty is communicated depends on the user of the information. If the end user has little understanding or experience with the model, then an easily understandable communication of the uncertainty is very important, as this information is critical to understanding the "risk" associated with using model outputs. Quantifying or visualising this "risk" is especially important when the output is considered key information by commercial and government decision makers. On the other hand, if the end user has an understanding of the model and the limitations of its inputs, then the uncertainty propagation analysis results are more important. Scientific researchers and model developers/programmers generally fall into this grouping. In either case, the calculation and communication of the uncertainty is of fundamental importance.

2.2 Uncertainty and Risk

The analysis of risk is important in many decision making areas. This section will discuss its relevance to the models used in commercial and research fields, which are studied in this thesis: Precision Agriculture - Chapter 3 and Ecological Niche Modeling - Chapter 5. The former's primary aim is to improve the efficiency of a farm's output. The later's aim is to understand ecological systems and use this information to predict where a species (both commercial and non commercial) is most likely to thrive. It may also be used in the prediction of how crop viability is likely to change due to a change in the climate of an area.

In both of these, there is a commonality in the approach that is taken to access the risk and then determine if the risk is acceptable. This requires the propagation of the data uncertainty into decision uncertainty, which is followed by a formal risk scenario identification and analysis. An example of this decision making analysis is discussed by Agumya and Hunter (2002) and is summarised Figure 2.1. That work aimed to understand uncertainty so as to (a) avoid data

which is not suitable for intended purposes i.e. data whose consequences are unacceptable, (b) to reduce undesirable consequences to an acceptable level and (c) devise ways of living with undesirable data when the adverse consequences of using this data do not alter the ultimate decision choice. In their work the method used to achieve this is called "the risk management approach" which concentrates on minimising the parts of the three part risk function:

1. Reduction of the likelihood of a risk scenario occurring (directly related to uncertainty in data, the algorithms in the model and the decision model used to produce the product).

2. Reduction of the consequences of a risk scenario (to the end user, the magnitude of the risk and its cost or value (Dobran 1995)).

3. Reduction in the degree of data utilisation i.e. continue to use the dataset but put less reliance on it which, in turn, will minimise the influence of the uncertainty in this data.

It then concludes that "by providing data users with a broad range of responses for dealing with this impact, risk management offers them greater flexibility in containing and responding to the consequences of any adverse scenarios that might occur" (Agumya & Hunter 2002). In this conclusion, "impact" refers to the impact of geographical data uncertainty, as this was their field of interest. This is valuable for the research aims of this thesis, as well as being applicable to other fields.

2.2.1 Risk in Agricultural and Ecosystem Analysis

There is extensive research on how uncertainty and risk assessment has been applied in determining the impact of climate change on environmental systems and resources. This is of special interest as, since the early 1990's, it has been recognised that climate change would negatively impact developing

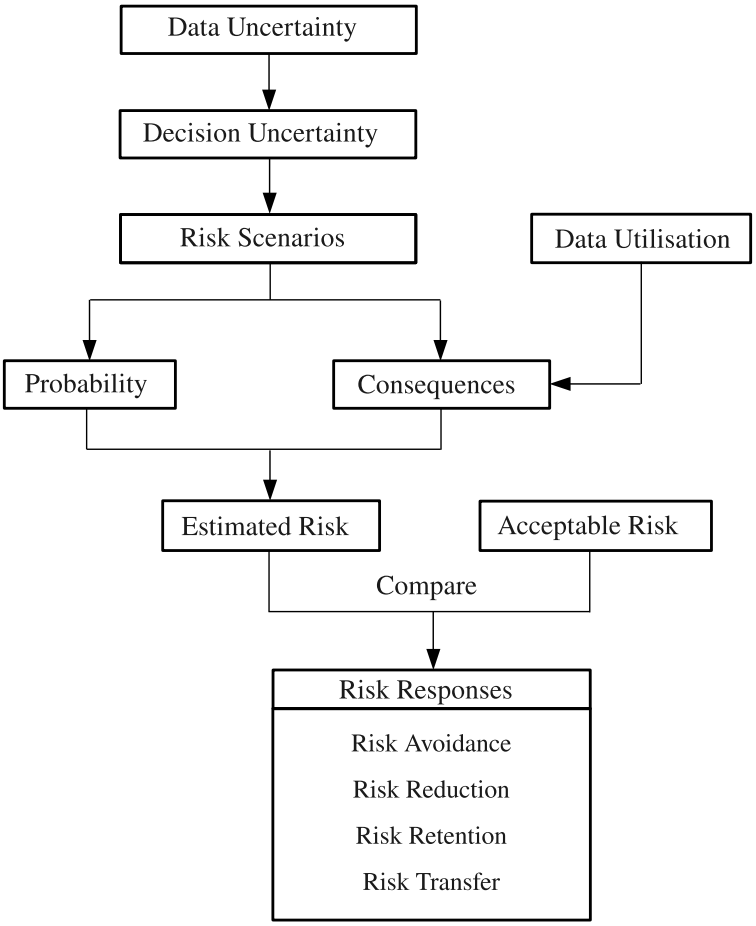


Figure 2.1: Risk-management of the impact of data uncertainty (Agumya & Hunter 2002).

countries to a larger extent than developed countries. Examples of relevant research include ecosystem studies (Hulme, Mitchell, Ingram, Lowe, Johns, New & Viner 1999), food security (Hulme et al. 1999, Parry, Rosenzweig, Iglesias, Fischer & Livermore 1999, Parry, Rosenzweig, Iglesias, Livermore & Fischer 2004, Rosenzweig & Parry 1994), malaria (vector born pathogens) and coastal flooding (Hulme et al. 1999). Other relevant areas investigated include carbon sinks (White, Cannell & Friend 1999), carbon dioxide fertilization of crops (Darwin & Kennedy 2000) and water security (Arnell 1999).

Studies such as these use future climate projections, which are calculated using global climate models, such as the United Kingdom Hadley Centre's third generation coupled atmosphere-ocean global climate model (HadCM3) (Johns, Gregory, Ingram, Johnson, Jones, Lowe, Mitchell, Roberts, Sexton, Stevenson, Tett & Woodage 2003). The uncertainties in these model's future predictions are quantified and the limitation in the accuracy of the future climate scenarios is important and extensively covered in the literature. For example, as discussed in Hulme (1999) "limitations should be recognised when interpreting the results of the impacts studies that make use of climate scenarios described here. A full risk assessment of climate change impact would attempt to sample the various sources of uncertainty mentioned ..."

2.3 Spatial Models

In most literature, a spatial model is described as a model of a terrestrial system. This name suggests that these models are representative of an area or location in space at a certain point in time. But, spatial models have also been developed to study changes over time. Therefore, and not unexpectedly, the complexity of these models varies significantly. However, the main component(s) of a model's can be categorised into two groups:

Interpolation Models.

The purpose of an interpolation model is to generate an accurate layer in

an area of interest by analysing known data from that area. The interpolated layer could be an end product of an analysis, or a step within a more complex model (such as - but not only - a process model). Interpolation models vary in their complexity depending on the system being studied. For example, the Kriging family of linearised regression techniques is a complex method that has multiple versions for specific purposes. The commonly known ones are Simple Kriging (which is the mathematically least complicated), Ordinary Kriging, Universal Kriging, Block Kriging and Cokriging (Davis 2002). They require a prior knowledge of the area being studied in the form a semivariogram (Burrough & McDonnell 1998, Davis 2002), which is a statistical estimation of the relationship between the known data points.

Process Models.

A process model is a model that aims to simulate a complete system, such as an ecological or coupled atmospheric-oceanographic climate system. In theory, a model would represent the complete system, or at least all the principal components of that system. In practice, it is usually limited by the availability of data and the understanding of the system being modeled. Also, the more complex the modeled system, the more likely it is to be a combination of differing model types such as statistical and physical models. Therefore, a process model could also contain interpolation steps.

2.4 Uncertainty Propagation in Spatial Models

In all models, the structure of their algorithms will contribute to uncertainty propagation. The often unique spatial algorithms, in what can loosely be defined as a "spatial model," introduce unique uncertainty influencing factors. For example - and very important - is that input point data is usually gridded with each grid representing a large area. Also, each grid location is often referred to as a point confusing its true spatial character. This gridded data can be further interpolated to produce a higher spatial resolution model input layer, which may

be adequate if it is a true representative of the courser input grid. If this is not the case, then the interpolated surface will contain a greater level of uncertainty. Also, the aggregation of input data by a process model can have a negative influence on a process model's output. In some scenarios this can be minimised by aggregating the process models results rather than its inputs. Therefore, this question can be simplified as to whether the model input or its outputs should be interpolated. A step wise representation of the two paths in this process are shown in Figure 2.2.

This interpolate first - calculate later or calculate first - interpolate later question is important when process modeling a spatial system and it must be addressed specific to the process model being studied. For example, if the point locations of the input layers are different then an interpolation step must occur first. Examples of studies where this decision step was considered important are Stein, Staritsky, Bouma, van Eijnsbergen and Bregt (1991), Bosma, Marinussen and van der Zee (1994) and Addiscott and Tuck (1996).

Further examples of work aimed at improving the accuracy of spatial models include the papers by:

1. Kerry and Oliver (2007*a*, 2007*b*), which investigate the influence of asymmetric data on a variogram - of key importance for kriging interpolation.

2. Wratt et al. (2006), which discusses the preparation of GIS maps used in agricultural modeling (using inputs such as climate and soil properties). Of particular relevance (to this thesis), is its discussion on the range of variables used and the typical uncertainties of these climate mapping techniques.

3. Refsgaard et al. (2007), which discusses the terminology and typology of uncertainty and presents a framework for the modeling process - its interaction with the broader water management process and the role of uncertainty at different stages.

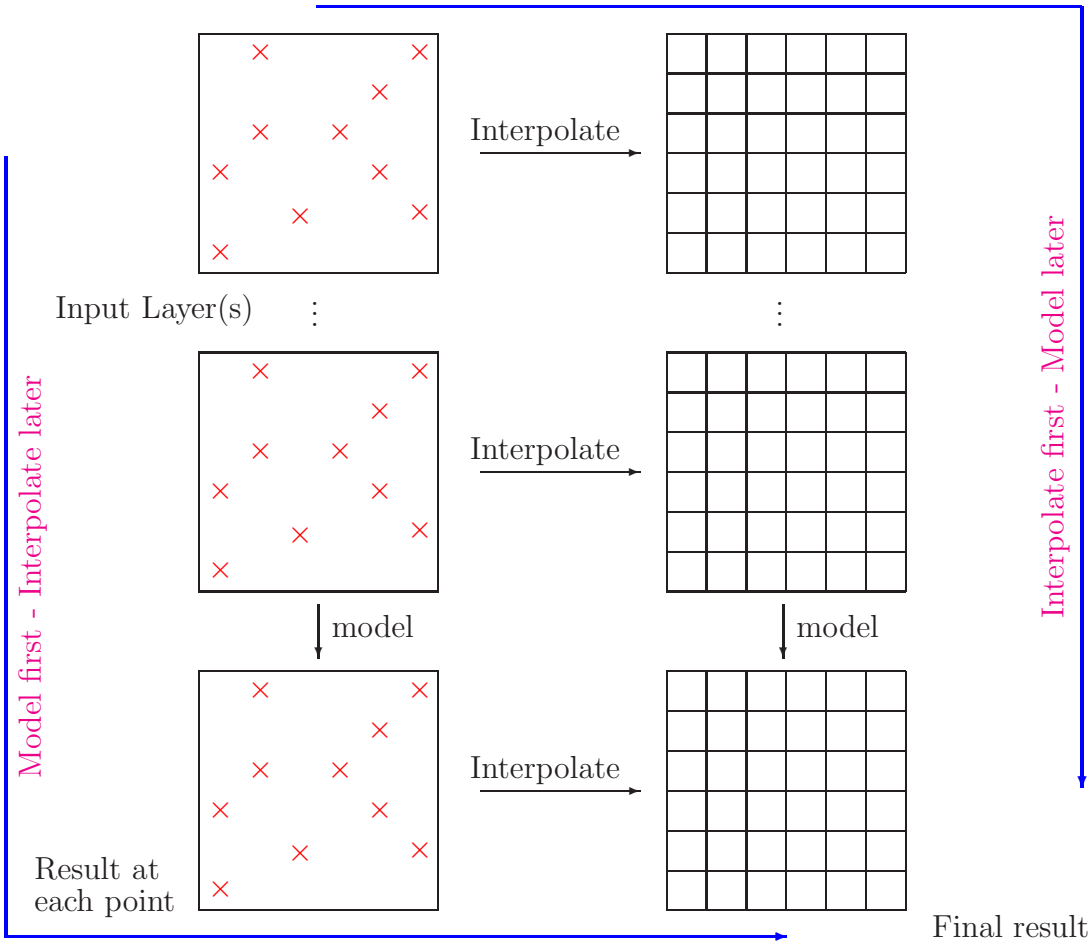


Figure 2.2: A grahical representation of the paths when either model inputs or outputs are interpolated.

2.5 Analysis of Uncertainty in Spatial Models

The quality of results from quantitative mathematical and statistical models that involve the combination of data sets depends on the following factors (Heuvelink & Burrough 2002, Burrough & McDonnell 1998):

1. the quality of data,
2. the quality of the model used in processing the data,
3. the way data and model interact.

Therefore, to get reliable results it is important to know how uncertainties in both the model parameters and in the input data propagate through the model. Hence analysis of uncertainty propagation requires estimates of the sources of uncertainty and an uncertainty propagation theory and tools usable in the models analysis. The following section (2.5.1) describes the fundamentals of uncertainty when studying spatial systems. Section 2.7 describes the basic mathematical and statistical methods used in this thesis.

2.5.1 Description of Uncertainty

Although we are aware that a particular attribute of interest in a spatial system (such as the seasonal rainfall totals in the south west of Western Australian) has one fixed deterministic value, the uncertainty about that fixed value allows us to treat it as the outcome of some random mechanism. To understand this mechanism we must proceed to define the constants, variables and statistical/mathematical rules that are applied in the analysis of the spatial system.

Uncertainty at a Single Location

In general, the definition of uncertainty at one point in space, is simply the arithmetic difference

$$v(x) = a(x) - b(x), \tag{2.1}$$

where the true value of $a(x)$ is unknown and $b(x)$ is the known estimate of $a(x)$. This *model of uncertainty* refers to a stochastic or statistical representation of attribute uncertainty, whose range of values is based upon the known range of what the value for $a(x)$ should be and hence the uncertainty $v(x)$ is based upon this known (or assumed) range. This assumed knowledge of the uncertainty $v(x)$ then allows it to represented as a *random variable* $V(x)$, even though this variable is determined from the deterministic variable $a(x)$. This assumption can be made as our uncertainty about $a(x)$ allows us to treat is as the outcome of some random mechanism $A(x)$. Therefore, the simple uncertainty model becomes:

$$A(x) = b(x) + V(x) \tag{2.2}$$

where $A(x)$ and $V(x)$ are random variables and $b(x)$ is a deterministic variable.

For uncertainty at a single location, the mean and variance of $V(x)$ are denoted by $E[V(x)] = \xi(x)$. and $var(V(x)) = \sigma^2(x)$. As discussed by Heuvelink (1998), "The mean $\xi(x)$ is often referred to as the systematic error or bias, because it says how much $b(x)$ systematically differs from $A(x)$. The standard deviation $\sigma(x)$ of $V(x)$ characterises the non-systematic, random component of the error $V(x)$." As implied by Equation 2.2, the attribute $A(x)$ and error $V(x)$ have the same distribution, except for a change in the mean. This reflects the assumption in standard uncertainty analysis that uncertainty always follows the normal (Gaussian) distribution, as defined in the central limit theorem (Ott & Longnecker 2001, Davis 2002).

For this random statistical model to be valid, we must specify with a solid level of certainty, the rules of the random mechanism (such as the size and distribution of the uncertainty). The importance of these is logical enough. However, the assumption that the uncertainty is always normally distributed can be incorrect, especially in environmental systems.

Multi Layer Spatial Model

The expansion of single point model into multidimensional space is defined as $A(\cdot) \equiv A(x) \mid \epsilon D$ on the domain of interest D in n-dimensional space $\mathbb{R}$, where we refer to the value of $A(\cdot)$ at a specific location $x \epsilon D$ as $A(x)$. The single point uncertainty model then becomes:

$$A(x) = b(x) + V(x) \quad for \quad all x \epsilon D \tag{2.3}$$

Let x and x' be elements of D, where these elements are either in the same data layer or in another layer. The correlation of two points in a single layer, $\rho(x, x')$ of $V(x)$ and $V(x')$, defined as:

$$\rho(x, x') = \frac{R(x, x')}{\sigma(x)\sigma(x')} \tag{2.4}$$

where $R(x, x')$ is the covariance of $V(x)$ and $V(x')$. The expansion of this to a multivariate (multiple layer) stochastic model has multiple attributes $A_i(x)$ and uncertainties $V_i(x)$, $i = 1, ..., m$. As discussed in Heuvelink (1998), "for each of the attributes an uncertainty model $A_i(x) = b_i(x) + V_i(x)$ is defined, where the" uncertainty "$V_i(x)$ follows some distribution with mean $\xi_i(x)$ and variance $\sigma_i^2(x)$. Let $\rho_{i,j}(x, x')$ be the correlation of $V_i(x)$ and $V_j(x')$, defined as:

$$\rho_{i,j}(x, x') = \frac{R_{i,j}(x, x')}{\sigma_i(x)\sigma_j(x)'} \tag{2.5}$$

where $R_{i,j}(x, x')$ is the covariance of $V_i(i)$ and $V_j(x')$. The cross-covariance function $R_{i,j}(\cdot, \cdot)$ thus defines the covariance of different attribute uncertainties, possibly at different locations." This relationship is graphically illustrated in Figure 2.3.

In many situations it is not possible or necessary to determine all the dimen-

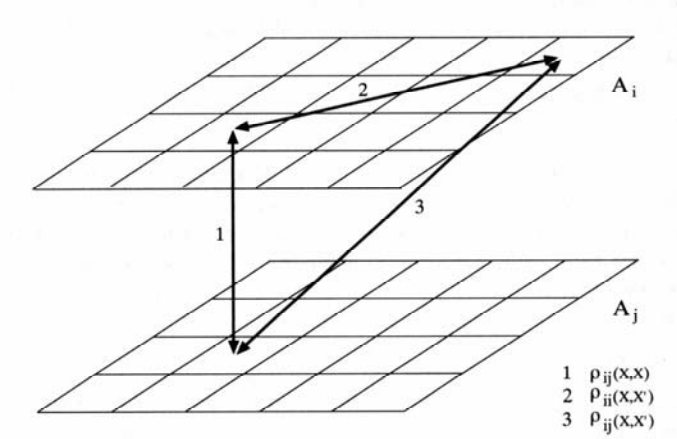


Figure 2.3: Graphical representation of the differences between the correlations $\rho_{i,j}(x, x)$, $\rho_{i,j}(x, x')$ and $\rho_{i,j}(x, x')$. (Heuvelink 1998).

sions of this relationship. For example, in cases where the relationship between known attributes at the same location is considered of greatest importance. So, as discussed by Heuvelink (Heuvelink 1998) "the correlation between attributes at the same location " will be of particular interest. However, where there is a significant relationship between variables at different geographical locations, these locations should be included in the uncertainty model as well.

2.6 Skewed Spatial Uncertainty Patterns

As discussed by Heuvelink (1998), Aguilar, Aguilar and Agüera (2007), Lucas (2010) and Wang, Chen, Wu, Feng and Pu (2010), the uncertainty in an environmental data set may not be normally distributed. For example, the potential sources of Digital Elevation Map (DEM) uncertainties are ascribable to the uncertainties in sample data error, the sum of the interpolation error and error due to sampling the continuous terrain surface with a finite grid interval (Aguilar, Aguilar & Aguera 2007). Most models used to compute the residuals in a DEM assume a Gaussian Distribution of the residuals but this is "sometimes incompatible with the fact that vertical errors in DEM often follow a non-normal distribution" (Aguilar et al. 2007).

How interpolation methods and sampling locations can influence uncertainty distributions is shown in Marinelli (2009). In that study, the input data used in the interpolations was sourced from a DEM of western Australia: Latitude $116^o16' - 117^o14'$ E., Longitude $27^o10' - 27^o8'$ S. This DEM has a resolution of 692 by 735 pixels with each pixel covering an area of 86.73 metres squared. The maximum and minimum values in the map are 537 and $323m$ above sea level. This image was derived from the Skylab 3 S-190B Earth Terrain Camera.

In order to investigate the effect of different interpolation models data was sampled from regularly and randomly spaced grid cells which, in both cases, equaled to 0.1% of the original DEM surface. The interpolation methods used were inverse distance weighting (IDW), spline and ordinary kriging. These layers were subtracted from the original DEM to generate *difference layers* showing how much the interpolated surfaces differed from the original DEM and the spatial distribution of this difference (Figure 2.4). Their associated histograms are illustrated in Figure 2.5. The actual uncertainty (per cell) in the input DEM layer was unknown and hence could not included in this analysis.

Error Statistics	Randomly spaced samples		Equally spaced samples	
	Skew	Kurtosis	Skew	Kurtosis
IDW	2.66	14.22	2.73	14.26
spline	1.51	9.50	1.14	12.36
kriging	3.36	21.34	2.28	15.68

Table 2.1: Skew and Kurtosis of Interpolated Surfaces.

For the spline and kriging techniques, the uncertainty layers with the lowest skew (and hence higher normality) occurred when the sampled points were equally spaced (Table 2.1). The exception was the result for the IDW, which was less skewed (but not by a large extent when compared with the other changes observed). It is also noted that the greatest agreement between these three methods occurred when the sampled data were evenly spaced. A general statement that can be made from these results is that most of the results appear relatively normally distributed. But, there are some points in the generated data layers

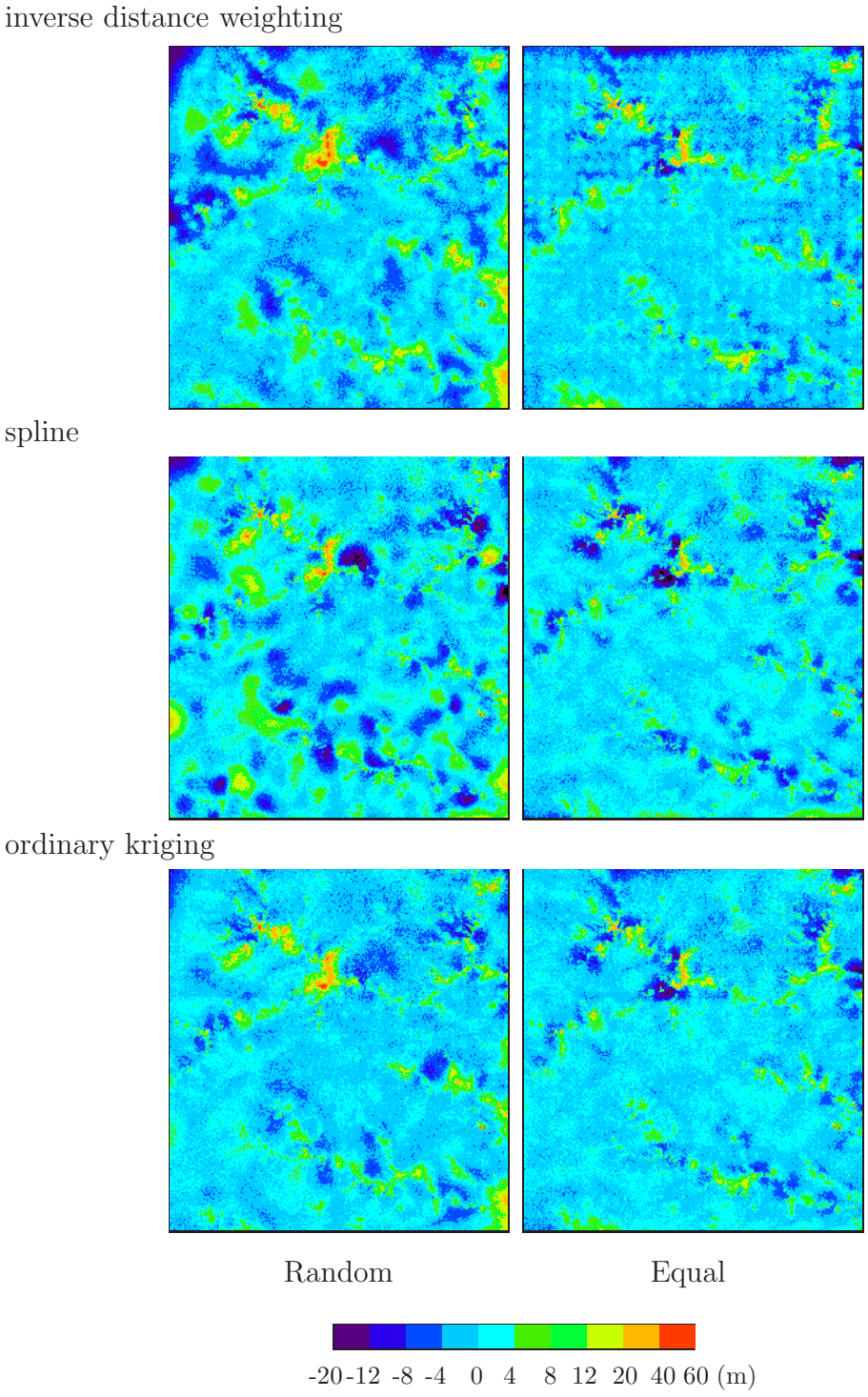


Figure 2.4: Uncertainty in Interpolated Digital Elevation Map from randomly and equally spaced samples.

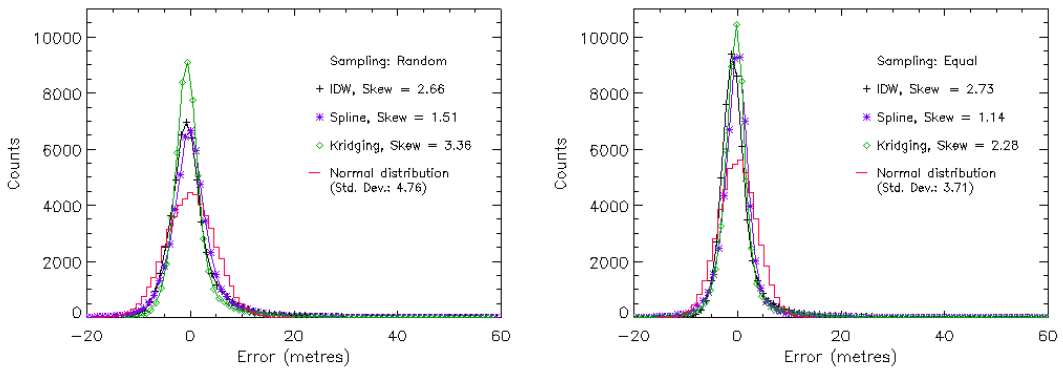


Figure 2.5: Interpolated DEM error distributions from random and equal sampling distances.

where the difference from the original DEM is considerably higher in the positive range. This non normal distribution is reflected in the skew of the uncertainty results.

Several questions relating to the accuracy of an interpolated data layer arise from these results:

1. What interpolation method gives the most accurate interpolated surface (and hence the least uncertainty).

2. Is the skew a true representation of the distribution of the uncertainty and therefore,

3. Can the skew be used in a data simulation to generate a valid random data set from which uncertainty propagation can be investigated.

These may be answered if the interpolated data layer can be compared to sufficient field measurements which, in environmental and agricultural studies, is often not the case. Finally, if the distribution of the uncertainty is known, it could determine which method can be applied in the uncertainty sensitivity analysis.

2.7 Analysis of Uncertainty Sensitivity

The way in which the uncertainty propagates through a spatial process model depends on a number of factors in the structure of the model. In models where the algorithms are known and continuous (in a data range that is valid), analysing uncertainty propagation can be done mathematically or statistically using a number of techniques. In models that are statistical or have discontinuities in the algorithms, the methods are limited to statistical analysis techniques. This is also the case in a model that has "black box" components where the algorithms are unknown.

In this thesis, precision agriculture models and an ecological niche model are investigated. In the first case, the algorithm's mathematical structure is known, differentiable and continuous (see Chapter 3). In the second a component of the model is a "black box" and so some of its algorithms are unknown. Also, many of the known algorithms are statistical and have step functions (see Chapter 5). In this thesis, the analysis methods used are the Taylor Series and Monte Carlo Simulations.

2.7.1 Taylor Power Series.

The Taylor Series is a convergence power series (Stewart 2003, Anton 1984) used extensively to closely approximate the sensitivity of an algorithms output relative to its input. More specifically, by analysing the partial derivatives of the algorithms in a process model, the uncertainties in the input(s) which cause the greatest change in the output can be determined. For example, in the sciences it is used extensively in areas such as ocean current and atmospheric modeling (Errico 1997). In these and other fields of research, the term "sensitivity analysis" is commonly used when discussing the application of this method, as its users aim to quantify which of the model inputs cause the greatest change in the output. As discussed in Heuvelink (1998, 1989), Burrough and McDonnell (1998) and Marinelli (2009), Taylor series is used in the sensitivity analysis of spatial

science/environmental/agriculture models.

2.7.2 Theory

The Taylor series estimates the value of a function f *about* a, where a is a range of normally distributed (Gaussian) values on axis x (such as the uncertainty defined in Section 2.5.1). If the function can be differentiated an infinite number of times at a, then the n-th Taylor function for f *about* $x = a$ is defined as

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \tag{2.6}$$

As with all convergent power series, "this means that $f(x)$ is the limit of the sequence of partial sums" (Stewart 2003), which for the Taylor series are

$$\begin{aligned} T_n(x) &= \sum_{i=0}^{n} \frac{f^{(i)}(a)}{i!}(x-a)^i \\ = f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 &+ \\ \frac{f'''(a)}{3!}(x-a)^2 + ... + \frac{f^{(n)}(a)}{n!}(x-a)^n. & \end{aligned} \tag{2.7}$$

Where $T_n(x)$ is a polynomial of degree n called the nth-degree Taylor polynomial of f at a. For example, for

$$f(x) = e^x, \tag{2.8}$$

the Taylor polynomials at 0 with n = 1, 2, 3 are

$$T_1(x) = 1 + x \tag{2.9}$$

$$T_2(x) = 1 + x + \frac{x^2}{2!} \tag{2.10}$$

$$T_3(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} \tag{2.11}$$

as drawn in Figure 2.6.

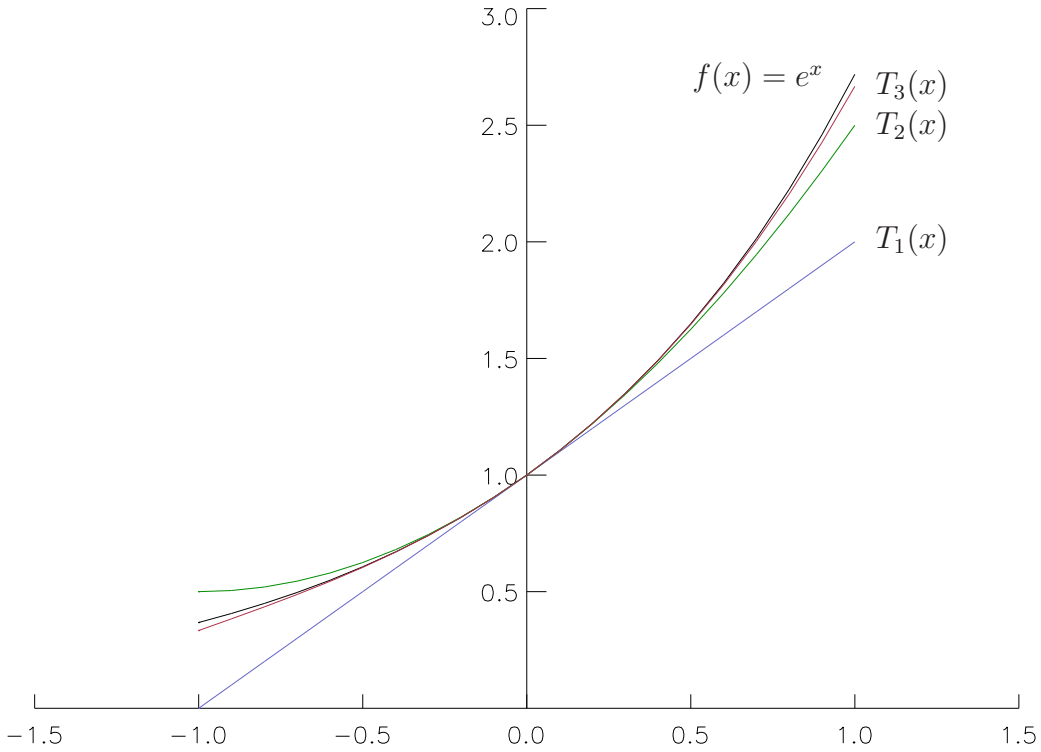


Figure 2.6: The graph of the $f(x) = e^x$ and its first Taylor polynomials, $T_{1..3}(x)$, when there is no uncertainty in the x values.

As is clear, as n increases, $T_n(x)$ appears to approach e^x. As discussed in Stewart (2003), "this suggests that this function is equal to the sum of its Taylor series", which is why it can be used to estimate the value of a function at a value or across a range of values, such as is defined in an uncertainty value [1].

A function's range of values is dependent on the sensitivity of the function to uncertainty about a. However, this is dependent on the functions structure (Arras 1998). A simple example where the function f has only only one input is illustrated in Figure 2.7, where the uncertainty X is normally distributed, mean

[1]When there is no uncertainty in x, the Taylor Polynomial is the same as the Maclaurin Polynomial.

μ_x and standard deviation σ_x.

It can be seen that the shaded interval on x, which is the 34% probability interval $[U_x - \sigma_x, U_x + \sigma_x]$, propagates through function (or model) $f(.)$ and maps onto the y axis as a non Gaussian (asymmetric) distribution. The first order Taylor Series expansion approximates $f(X)$ about the point $X = \mu_x$,

$$Y \approx f(\mu_x) + \frac{\partial f}{\partial X}|_{X=\mu_x} \quad (X - \mu_x), \tag{2.12}$$

which is the linear relationship illustrated in Figure 2.7. From this can be determined its parameters μ_x and σ_x (Anton 1984, Stewart 2003, Arras 1998),

$$\mu_y = f(\mu_x), \tag{2.13}$$

$$\mu_\sigma = \frac{\partial f}{\partial X}|_{X=\mu_x} \quad \sigma_x. \tag{2.14}$$

The output distributions represented by μ_x and σ_x are a representation of some unknown truth which is non linear, non normal and assymetric.

As illustrated in Figure 2.8, the sensitivity of output distributions is dependent on the shape of the function, with the approximation of uncertainty being substantially larger when a function is non linear, steep and concave upwards. Therefore, it is reasonable to assume that either in a simple function $f(x)$ whose range varies across the domain of x, or in more complex function f which has two variables x_1 and x_2, two unique functions and most likely, two domains. As illustrated in Figure 2.8(a), in the first of these functions, the output distributions μ_x and σ_x will vary depending on the value of x. In the second case, there will be two output distributions which will have a combined effect on the output distribution of f (Figure 2.8(b)).

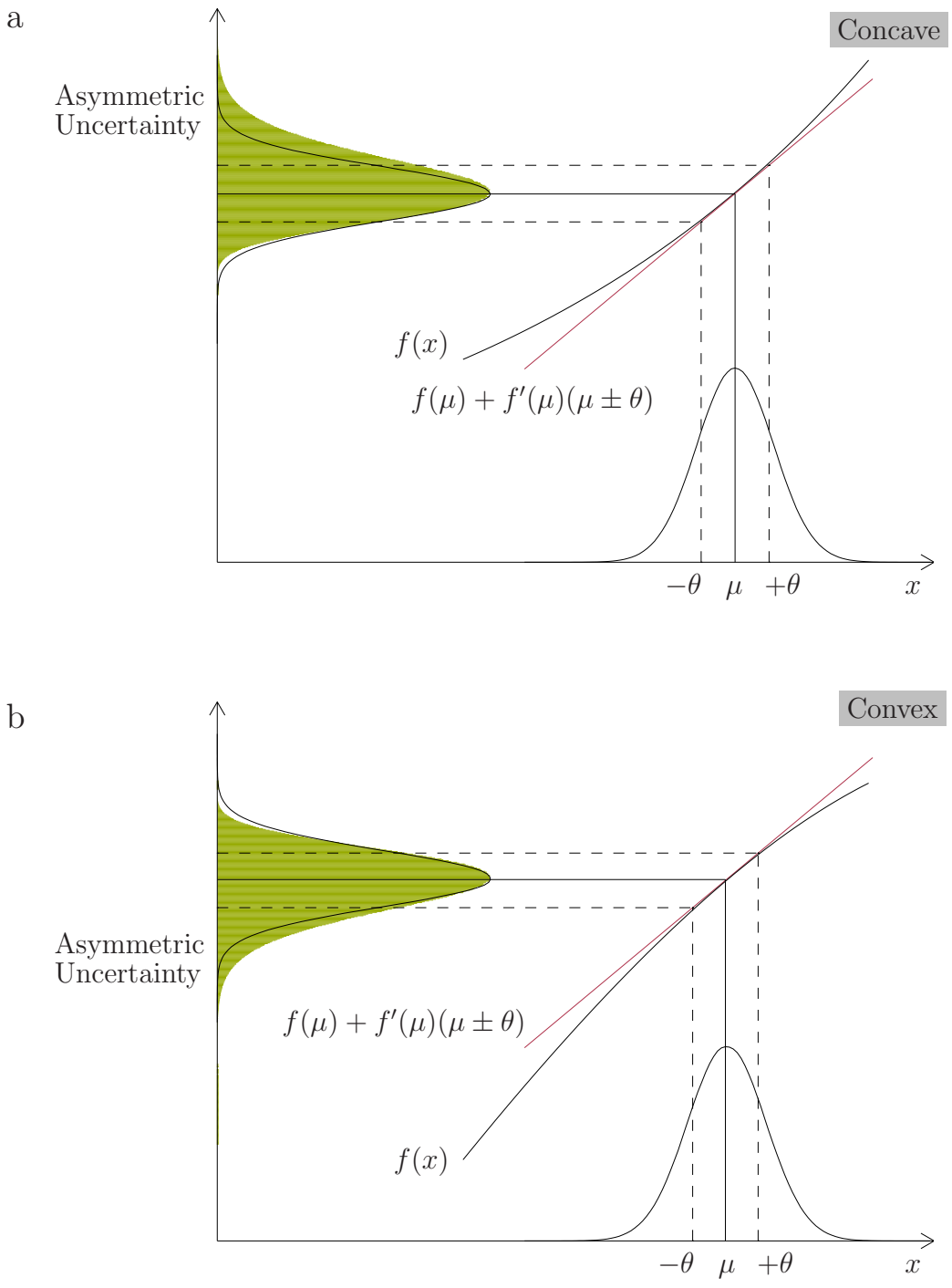


Figure 2.7: A graph of uncertainty propagation as determined by the first order Taylor Method for two simple non linear functions. These concave and convex non linear functions have one symmetrically distributed input and one asymmetric (non Gaussian) output - as graphically illustrated in the green coloured histogram.

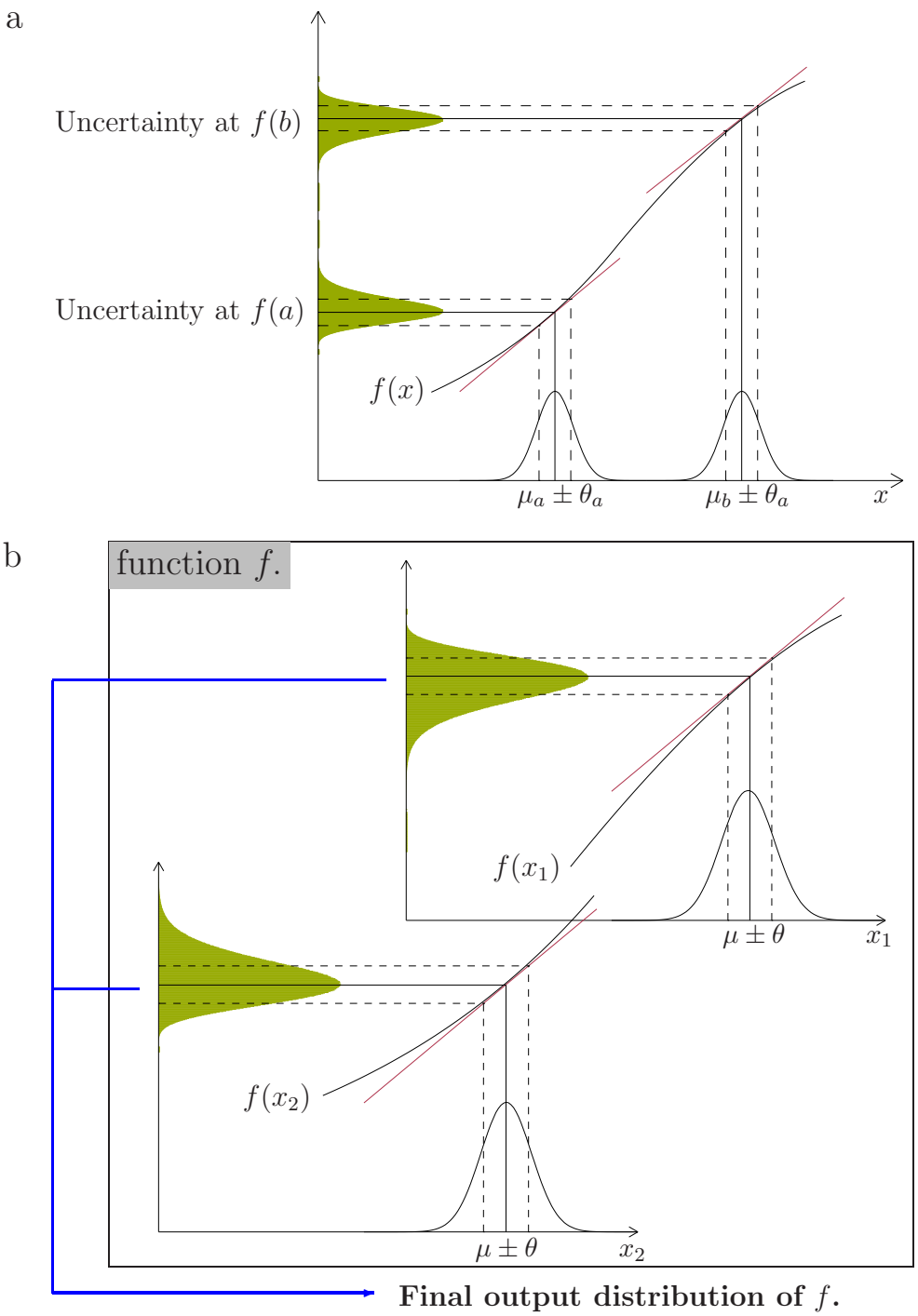


Figure 2.8: Uncertainty propagation analysis using the first order Taylor method; a simple and more complex function: (a) $f(x)$ is a simple function with a convex upward and concave downward component. The possible distribution of $f(x)$, for two mean values and uncertainty on x, is shown. (b) Concave and convex components of the more complex function f; which has two independent variables x_1 and x_2. As is clear, the output distributions of these two components is different. The final f output distribution μ_{x_1,x_2} and σ_{x_1,x_2} will result from these.

2.7.3 Application in the Analysis of a GIS Model

Theoretically, in GIS models, the output point map $U(\cdot)$ can be calculated from the point maps $A_i(\cdot)$ by means of the GIS point operation model $g(\cdot)$:

$$U = g(A_{(1)}(\cdot), ..., A_{(m)}(\cdot)) \tag{2.15}$$

The input maps of this GIS operation are the random fields $A_i(\cdot)$ as defined by some domain D and satisfy the equation $A_i(\cdot) = b_i(\cdot) + V_i(\cdot)$. The output map $U(\cdot)$ is effectively a randomly distributed field of values, represented by its mean $\zeta(\cdot)$ and variance $\tau^2(\cdot)$ (Heuvelink 1998, Burrough & McDonnell 1998, Bailey & Gatrell 1995). In a real situation where the number of locations in a polygon map or grid cell are finite, this equation becomes:

$$U = g(A_{(1)}(x), ..., A_{(m)}(x)) \tag{2.16}$$

Uncertainty propagation with point operations are in fact a non spatial problem so "x plays a dummy role in the analysis" (Heuvelink 1998). Therefore, for the rest of this section the index x is omitted.

When studying uncertainty propagation, the main interest is in the influence of the difference $A_i - b_i$, on the models output. Using the Taylor Series to determine this first requires defining the Taylor series of $g(\cdot)$ around $\overline{b}$,

$$U = g(\overline{b}) + \sum_{i=1}^{n} \{(A_i - b_i) g_i'(\overline{b})\} + remainder \tag{2.17}$$

where $g_i'(\cdot)$ is the first derivative of $g(\cdot)$ with respect to its i-th argument, where the number of arguments is equivalent to the number of propagation paths in the model. The remainder of the equation contains the higher order Taylor terms of $g(\cdot)$. The first order terms of the variance of U is

$$\tau^2 = E[(U - E[U])^2] \approx E[(g(\overline{b}) + \sum_{i=1}^{n}\{(A_i - b_i)g_i'(\overline{b})\} - g(\overline{b}))^2] \tag{2.18}$$

Note that to obtain only the uncertainty τ^2, $g(\overline{b})$ is subtracted from $(A_i - b_i)g_i'(\overline{b})$. This simplifies the equation to

$$\tau^2 = E[(U - E[U])^2] \approx E[\sum_{i=1}^{n}\{(A_i - b_i)g_i'(\overline{b})\}]. \tag{2.19}$$

When the model contains multiple variables n, which may or may not be independent of each other, the equation expands to

$$\begin{aligned}\tau^2 = E[(\sum_{i=1}^{n}\{(A_i - b_i)g_i'(\overline{b})\}) \cdot (\sum_{j=1}^{n}\{(A_j - b_j)g_j'(\overline{b})\})] \\ = \sum_{i=1}\sum_{j=1}\rho_{ij}\sigma_i\sigma_j g_i'(\overline{b})g_j'(\overline{b})\end{aligned} \tag{2.20}$$

where the variance of U is the sum of several terms containing the correlations (ρ) and standard deviations (σ) of A_i and the first derivatives of $g(\cdot)$ at $\overline{b}$. Therefore, as discussed in 2.7.2, τ^2 depends on the variance of, and correlations between, the inputs of the function $g(\cdot)$ as well as the sensitivity of the model's output to its inputs. To reduce the approximation in τ^2, the Taylor series could be extended to include higher order derivatives. However, the first and second order is considered adequate for most GIS models.

A Taylor Series Example

A simple example analysing the sensitivity of a two fields (inputs) is presented in Burrough and McDonnell (1998):

$$G = Y \times P \tag{2.21}$$

where G is the gross returns from a wheat farm, Y is the yield and P is the crop price. The partial derivatives of the gross returns function are:

$$\frac{\partial G}{\partial Y} = 1 \times P \tag{2.22}$$

$$\frac{\partial G}{\partial P} = Y \times 1. \tag{2.23}$$

For each field in the farm, Y is 6 ± 2 tonnes per hectare and P is 100 ± 10 currency units per tonne of wheat. So the gross return is equal to 600 currency units with a standard deviation of:

$$\begin{aligned} \sigma &= \sqrt{\{P^2 \cdot \sigma_Y^2 + Y^2 \cdot \sigma_P^2\}} \\ &= \sqrt{\{10000 \cdot 4 + 36 \cdot 100\}} \\ &= 208.81 \end{aligned} \tag{2.24}$$

currency units.

2.7.4 Monte Carlo Simulation Method.

As described in the general modeling and Geospatial Information Science literature (e.g. Metropolis and Ulam (1949), Burrough and McDonnell (1998) and Heuvelink (Heuvelink 1998)), the name Monte Carlo is given to a widely-used class of approaches which model a physical or mathematical system. It is an approach which tends to follow a particular pattern:

1. Define a domain of possible inputs.

2. Generate inputs randomly from the domain using a certain specified probability distribution.

3. Perform a deterministic computation using the inputs.

4. Aggregate the results of the individual computations into the final result.

Therefore, the idea of this methodology when applied to a GIS model, is to compute the results for the model repeatedly from inputs that are randomly sampled from their known distributions. In relation to the aims of this thesis, this "domain of possible inputs" and the final result are of primary importance. The first is equivalent to the values of f about a, as discussed in the Taylor Series method description (section 2.7.2). Each realisation g_i, $i = 1, ..., m$ will result in a unique value for each g_i which in turn results in a unique $u = g(a_1, ..., a_m)$. If there are N simulations, there will be N outputs of U from which the statistics of the Monte Carlo simulation will be calculated, as graphically illustrated in Figure 2.9.

The Monte Carlo method allows the domain of the input uncertainty to contain any valid probability distribution about a . In contrast, the Taylor method calculates only the sensitivity analysis result of an uncertainty with a Gaussian distribution. Additional reasons favouring the Monte Carlo method are that it allows the investigation of complex models containing many types of algorithms as well as those which have a "black box" component. The number of simulations N required for the final result to accurately represent the studied system will vary depending on the structure and complexity of the model. The test for this is when N is sufficiently high to produce a consistent result. That is, when the result from $> N$ simulations is sufficiently close to the result from $N + 1$ simulations.

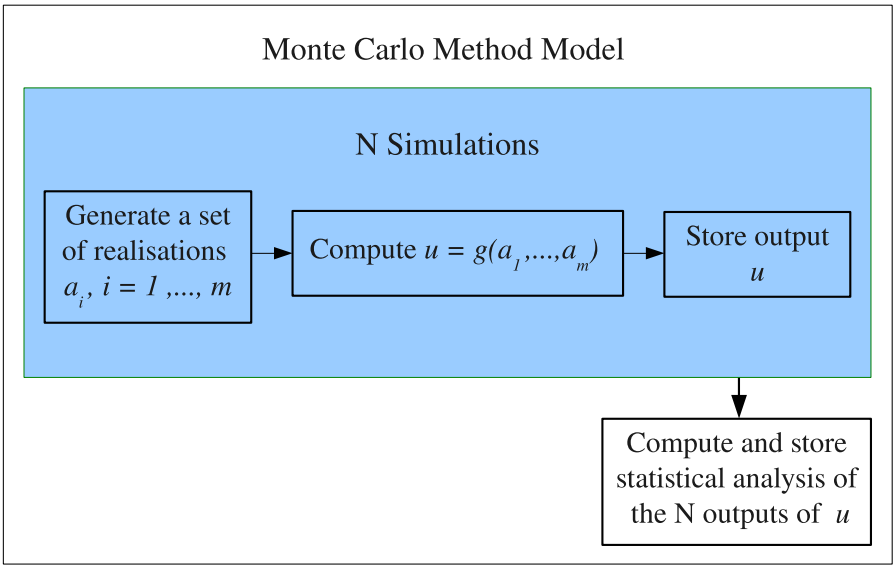


Figure 2.9: Illustration of the Monte Carlo Simulation Method

2.8 Summary

Spatial process models are used in a broad range of applications. They vary in type, complexity and often contain a combination of both interpolation and process models. To study their sensitivity to uncertainty propagation, the uncertainty in their input fields must first be represented by a stochastic or statistical uncertainty model. Then, the method used to quantify the sensitivity of the spatial process model to input uncertainty is limited by a combination of factors. If its algorithms are complex in structure, not continuous, contain step functions or are statistical, then the Monte Carlo Method is preferentially used. Furthermore, if the model contains components who's algorithms are not known, then the Monte Carlo method is the only option. The other method discussed in this section, the Taylor Method, is only used if the algorithms are differentiable and continuous within the domain of interest. Also, with the Taylor series, the uncertainty in the models input fields must be normally distributed.

Chapter 3

Precision Agriculture Models

3.1 Introduction

Precision agriculture is an agricultural concept that uses suitable technologies to improve agricultural practices and achieve targeted aims. The data input sets must embrace a range of crop, soil, landscape and other environmental attributes over significant areas (McBratney, Whelan, Walvoort & Minasny 1999). The technologies used to extract and then further analyse/interpret these data include traditional in-situ measurement and analysis techniques such as measuring soil attributes (soil nutrient levels, salinity and acidity etc.) and the use of modern technologies such as global positioning systems (GPS), remote sensing (Corner 1997), information management tools (GIS) and modeling of the agricultural region of interest. These more recent technologies are especially useful when studying large areas as they allow estimation of a variable's spatial distribution from point data measurements. Furthermore, the combination of these technologies has moved land assessment agencies from a regime of traditional mapping towards quantitative methods of land resource assessment as discussed in Corner, Hickey and Cook (2002).

The final aims of this scientific field vary depending on the unique characteristics of an agricultural region. For example, where the environmental conditions are good, such as good soil quality and consistent predictable rain pattern, the

primary aim may be to improve the production output while maintaining or lowering input costs. However, in an area with poor soils and less consistent rainfall patterns, the aim may be to minimise costs, maintaining previous yields and to minimise soil degradation. Furthermore, the degree and the reasons for the use of precision agriculture can be influenced by other priorities. For example, in the American Midwest agricultural region precision agriculture's primary attraction to farmers is to minimise costs by allowing map controlled variable rate fertiliser application across a field (The map is calculated on the known nutrient composition of the field). In Europe, the aim of minimising costs is also important, but an additional aim has been to minimise the environmental impact of modern farming practices. Originally, the purpose of this was to minimise nutrient levels in river systems and coastal run-off areas with the aim of avoiding or repairing "dead zones" (as they are commonly referred to). This was the main reason why, since the 1980's, European Union and member state fertilizer-use regulations aimed at reducing nutrient surpluses have become increasingly stringent (Vitousek, Naylor, Crews, David, Drinkwater, Holland, Johnes, Katzenberger, Martinelli, Matson, Nziguheba, Ojima, Palm, Robertson, Sanchez, Townsend & Zhang 2009). However, since that time, the aims have broadened to include, for example, reducing the contribution of modern agricultural practices to greenhouse gases.

The methodologies used in precision agriculture can also be used to minimise other agriculture-related environmental damage, while aiming to maximise yield. However, this is often referred to as "sustainable agriculture," even though there are clear overlaps between the fields. Sustainable agriculture will not be further discussed.

3.2 Application in Western Australia

In Western Australia, precision agriculture has been applied mostly in cereal grain crops grown in the "Wheat Belt" agricultural region (see Figure 3.1).

Wheat is the primary crop grown in this region, with pulses being the most prominent secondary crop (Leff, Ramankutty & Foley 2004). Nutrient run-off is not a major influence on the environment in this region and so precision agriculture methods have not been applied for environmental management purposes. Instead, it has been used primarily with the aim of increasing farming efficiency in a region that is low in production (per hectare) by world standards but still significant due to its spatial size.

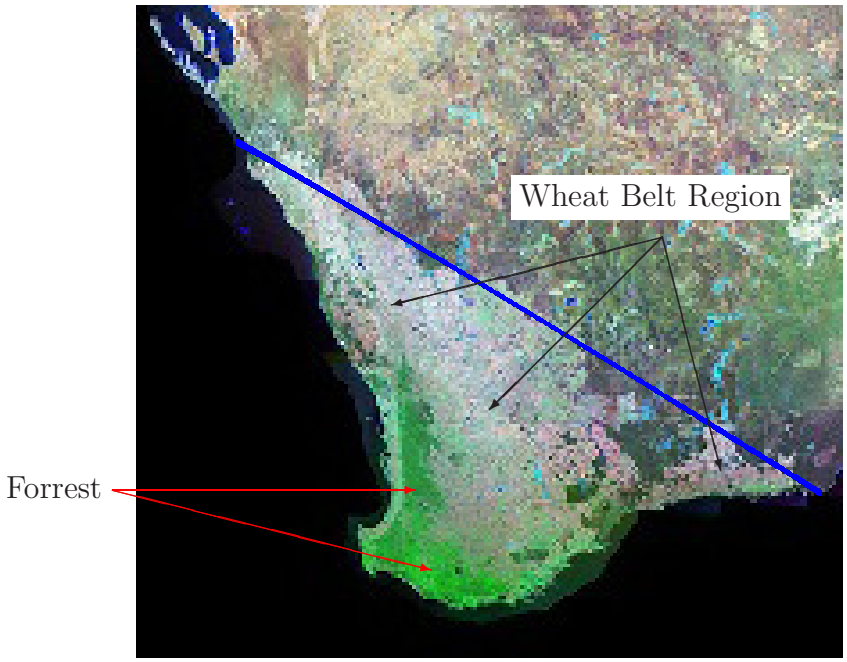


Figure 3.1: The "Wheat Belt" in Western Australia. This cleared agricultural area is clearly visible in this satellite image (Geoscience Australia (2011)), with the blue line showing its approximate north eastern border. The darker green area in the state's south west corner is mostly uncleared forest.

The two major controlling factors in the low productivity levels of this region are nutrient poor soils and the lack of a continuous water source such as a major river system. Therefore, the agriculture in this region is "rain fed," being dependent on the natural precipitation brought by the passage of the cold winter fronts which travel west to east from the Indian Ocean. Sufficient consistency in this precipitation during the mid year months is critical to a good yield in this area. There is summer rainfall which occurs due to the south-east monsoon

during the southern hemisphere summer months. But, it is sporadic and not suitable for these crops. Also, the lack of a permanent water source makes this region unsuitable for crop growth during this period. Therefore, the main limiting factor(s) which can be directly controlled are the nutrient level of the soils (by the addition of artificial fertiliser) and the control of pests and weeds using pesticides and herbicides.

In nutrient poor soils, the critical nutrients are nitrogen (N) and phosphorous (P). But, other nutrients may be lacking in a particular soil type (such as potassium (K)) and so may also be included in the Model. For example, the influence of P content in the yield of clover was investigated by Brennan and Bolland (2003). This work clearly showed the P to yield relationship and how it can vary depending on the soil type, as illustrated in Figure 3.2, which shows the P to relative yield response. This information is used in models such as the Mitscherlich Fertiliser requirement model (Wong, Corner & Cook 2001, Brennan & Bolland 2003), a model that uses both in-situ measurements and inputs calculated from information measured by remote sensing. More specifically, this model requires an estimated desired yield as a proportion of the maximum achievable yield in the calculation of the required fertiliser (see Chapter 4 for more details on the algorithm of this model). The maximum achievable yield is taken as input from the Normalised Difference Vegetation Index (NDVI), which is an estimation of green biomass.

$$NDVI = \frac{(NIR - VIS)}{(NIR + VIS)} \tag{3.1}$$

This method uses remote sensing to measure the adsorption in the visible (VIS) wavelengths ($400 - 700\ nm$) by chlorophyll and associated pigments and the increased reflectance in the near infra-red (NIR) ($750 - 850\ nm$), which is due to multiple scattering by leaf tissue (Smith, Wallace, Hick, Gilmore, Belford, Portmann, Regan & Turner 1994).

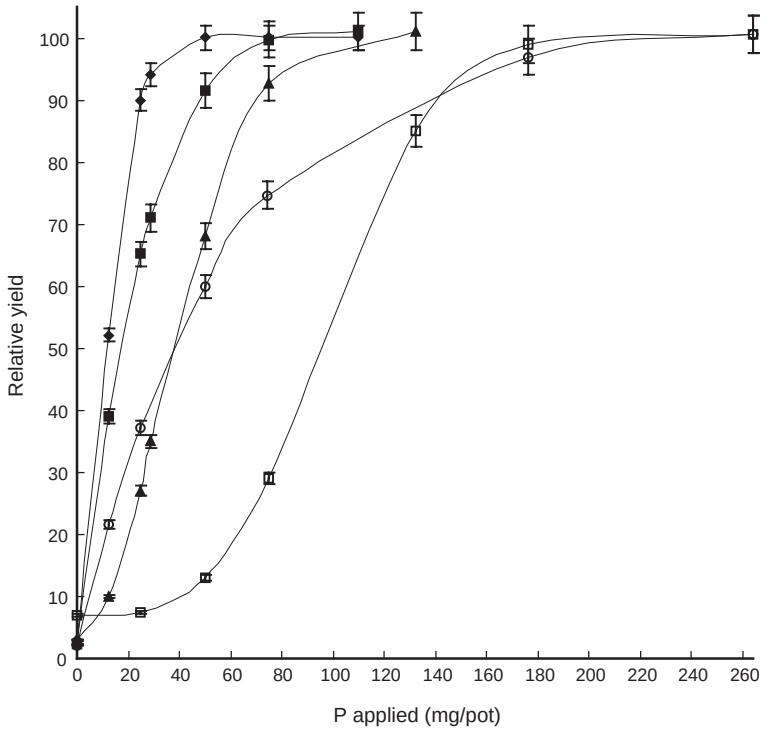


Figure 3.2: Relationship between percentage of maximum (relative) yield of dried clover shoots and the amount of P applied for 5 soils. Diagram from Brennan and Bolland (2003).

The application of Precision Agriculture to determining the optimum fertilizer requirement for an agricultural field is a clear example of how it can be used to help resolve a traditional question in agriculture, "How much fertilizer should I add?" More broadly, it can be used to increase the yield and/or efficiency of production by investigating a range of other factors. For example, in the Western Australian wheat belt region, as well as other regions of Australia that do not have irrigation, there has been research aimed at improving our understanding of how moisture content in soils influences yield response, and how this moisture content can be maintained between the seasons. Also of importance is the rate at which moisture is lost from the row crop canopies. For example, earlier work in this field by Ritche (1972) developed a model to predict daily evaporation rates from a row crop with incomplete cover. In later work by French and Schultz (1984), work was done to quantify the relationship between wheat yield and water use in Mediterranean like climates. Understanding this relationship is important in the development of crop growth models, predicting yield and the general use of water in similar environments. This is especially the case in areas where there is a very clear relationship between yield and rainfall (Stephens & Lyons 1998).

3.3 Influence of Uncertainties in Precision Agriculture Models

The influence of uncertainties in precision agricultural models is often discussed in the literature. For example:

1. The precision of the input data varies in quality and spatial density. In some cases, this uncertainty is unknown and, as discussed in McBratney et al. (1999), "While there is no doubting the quality/density of data obtainable from modern crop yield sensors, the quality and density of data gathered on other variables is less than optimal. The ultimate solution is

the development of real-time sensors to measure these attributes" and

2. The mathematical structure of a model may make it more sensitive to input uncertainties, as discussed in Purnomo, Corner and Adams (Purnomo, Corner & Adams 2003).

These uncertainties fall broadly into the temporal, metrical and structural classes of uncertainty (discussed in Chapter 2.1). Quantifying this uncertainty in a "less than optimal" agricultural situation is complex as they are not easily identified or measurable. Furthermore, the input data mapping methodology adopted by a farmer may mask the scale of the input uncertainty. For example, in modern farming, fertilizer equipment variable rate technology allows a mapped application of fertilizer that can be made to constantly change depending on the true quality of the soil. This is only possible if there is high spatial density of data for the field of interest and that the interpolation method used (and its limitations) is adequately understood - the combination of which results in an accurate input map of the field of interest. Unfortunately, it is often the practice to partition a field into aggregated zones, resulting in a stepped application of fertilizer. This aggregation of the input data has the effect of at best smoothing or, at worst, disregarding the uncertainty in the inputs, which in turn introduces another level of uncertainty in the calculated fertilizer requirement of a field. The practical effect of this is that the nutrient input is less likely to be optimal.

The practice of aggregation into zones also falls into the "translational error" class as it may reflect a lack of understanding of a models or technologies limitations. If these limitations were communicated more clearly it may encourage the end user to minimize the uncertainties in a model's inputs as well as increase the user's trust, understanding and acceptance of a model's results. Also, communicating the level of uncertainty will, depending on the environment, determine which model is best to use. For example, in data-poor situations knowledge driven models may be preferred (by a farmer) but in data rich situations, data driven models may be more appropriate (Adams, Cook & Corner 2000).

To summarise, four broad classes of uncertainty must be considered when determining how uncertainty may influence the accuracy of a precision agriculture model and how this can be minimised. In a complex precision agriculture model that is part of a GIS software package, this could require a multiple stage (input - algorithm - visualize result) approach to resolve. However, in a complex model that provides an easily useable result (such as the specified nutrient level input to a GPS guided agricultural tool), the main question to resolve and quantify is the model's sensitivity to uncertainties in the inputs.

3.4 The Precision Agriculture Models Investigated

The precision agriculture models studied in this thesis are the nitrogen (N) availability component of the SPLAT model (Adams, Cook & Bowden 2000) and the Mitscherlich precision agricultural model (often referred to as the "Mitscherlich Equation"). The N-availability (in soil) model is linear whereas the Mitscherlich is not, a key factor expected to effect the shape and size of the propagated uncertainty. This, their continuity (through their valid input range) and their relative simplicity qualifies them as ideal for sensitivity to uncertainty analysis using both Taylor and Monte Carlo uncertainty propagation analysis techniques. The details of the models are:

N-availability.

$$N(available) = (RON \times RONDep(T-1) \times RONEff) + 10000 \times (OC \times (1 - GravProp) \times SONEff) + (15 \times FerTeff)$$

where the input data layers are the residual organic nitrogen (RON),organic carbon in the soil (OC) and the gravel proportion in the soil ($GravProp$). The other four parameters are the $RONDep$ depletion coefficient and three efficiency

coefficients $RONEff$, $SonEff$ and $FertEff$. These coefficients are constants that were determined by productivity trials. The known uncertainties of the input layers and the coefficients were obtained from discussion with experts in soil testing. The other variable is time (T) in years, since the last lupin crop. The N-available is in Kg/Ha.

The Mitscherlich model.

An inverted form of this model (Edwards 1997) has been proposed (Wong et al. 2001) as a method of determining the spatially variable potassium fertiliser requirements for wheat. This relationship, which describes the response of wheat plants to potassium, is shown in Equation 3.2,

$$Y = A - B^{-CR} \tag{3.2}$$

where Y is the yield in Tonnes per Hectare; A is the maximum achievable yield with no other limitations; B is the response to potassium; C is a curvature parameter; and R is the rate of applied fertiliser.

It has been shown (Edwards 1997) that the response, B, to potassium fertiliser for a range of paddocks in the Australian wheat belt may be determined by Equation 3.3,

$$B = A(0.95 + 2.6 \times e^{-0.095 \times K_0}) \tag{3.3}$$

where K_0 is the soil potassium level. Substituting Equation 3.3 into Equation 3.2 and inverting provides a means of calculating the potassium requirements for any location with any given soil potassium value. This is shown in Equation 3.4,

$$R = \frac{-1}{C} \times \ln(\frac{Yt - A}{-A(0.95 + 2.6e^{(-0.095K_0)})}) \tag{3.4}$$

where R is the fertiliser requirement (Kg/Ha) to achieve a target yield of Y_t Tonnes per Hectare.

3.5 Propagation Analysis: Data and Methods

This section describes the input data of the Precision Agriculture models and the methodology used to apply the Taylor and Monte Carlo Methods of uncertainty propagation analysis.

3.6 Input Data Layers

The input data layers of the N-availability equation were constructed from data collected at a 20 hectare paddock in the northern wheat belt. The data for the Mitsherlich model is from an 80 hectare paddock in the central wheat belt, where potassium fertilization is often required. Achievable yield was calculated by aggregating NDVI representations of biomass, derived from Landsat five images over a period of three years and estimating water limited achievable yield using the method of French and Schultz (1984). This method of deriving achievable yield is described in greater detail in Wong et al. (2001). Soil potassium was determined at 74 regularly spaced sample points using the Colwell K test (Rayment & Higginson 1992). These values were interpolated into a potassium surface using Inverse Distance Weighting. All data were assembled as raster layers with a spatial resolution of 25m. For the work described here a Target Yield of two Tones per Hectare was set. This is within the Achievable Yield value for 97% of the paddock.

3.7 Application of Uncertainty Propagation Analysis

As discussed in Section 2.7, the analysis of a Precision Agriculture model's sensitivity to input uncertainty was done with both the Taylor Series methods and the Monte Carlo Simulation methods. The results will be assessed to (a) determine how the input uncertainties propagate through the models and (b) how the

results of this analysis differs between the analysis methods.

3.7.1 The Taylor Series Method

The Taylor series method relies on using either the first, or both first and second, differentials of the function under investigation (Stewart 2003). In the case of when the uncertainty is normally distributed and the algorithm is continuous, it is effectively considered a "gold standard" and widely used. Its main limitation is that it can only be used in the analysis of the parts of an algorithm which are continuous. Since the functions in the two investigated Precision Agriculture models are continuous and differentiable, that is not a constraint here. The partial derivatives of these functions, and the matrix data used in their analysis, are in Appendix A. This was implemented in a procedure written in the Interactive Data Language (IDL) (Research Systems, 2006).

The N Availability Model

The variables in the N availability model are the residual organic nitrogen (RON), the organic carbon fraction (OC), the gravel proportion ($GravProp$). The $RONeff$, $SONeff$ and $FertEff$ efficiency coefficients are constants, by do have a quantified uncertainty. Equation 3.2 was partially differentiated with respect to each of these inputs, to the first order. These variables were converted to spatially variable data layers by combining with an absolute uncertainty layer. These uncertainty layers were generated as follows:

1. For the RON, OC and Grav Prop a relative uncertainty of $\pm 10\%$ has been chosen for each data point. Therefore, the uncertainty was first calculated by multiplying the data by 0.1. This value was assumed to represent the full width of the normally distributed uncertainty distribution. In order to provide the same approximate uncertainty magnitude as is being used in the Monte Carlo simulations (described below) the uncertainty was represented by 3.33% being equivalent to 1 standard deviation.

2. The $RONeff$, $SONeff$ and $FertEff$ efficiency coefficients are not spa-

tially variable and nor are their estimated uncertainties of ±0.4, 0.025 and 0.025 respectively. These were divided by their respective coefficients to obtain an uncertainty ratio representing the full width of an uncertainty distribution. From this, the values representing three standard deviations from the mean were calculated as the difference between the fertiliser coefficients and their extreme values (the coefficients plus their uncertainties).

3. The uncertainty in the RONdep depletion coefficient is difficult to quantify and so was not included in this study.

4. The output generated using the Taylor Series method is an uncertainty error surface for N-availability.

The Mitscherlich Model

The input variables in the Mitsherlich model are the achievable yield (A_y), the soil potassium level (K_0) and the curvature term (C). Equation 3.2 was partially differentiated with respect to each of these inputs to both the first and second order. The Uncertainties in these layers were generated as follows:

1. For the A and the K_0 data layers a relative error of ±10% was assumed for each data point.

2. Curvature term C is not spatially variable but is known to contain uncertainty. In this case, the value is derived from a series of regional experiments on potassium uptake by wheat crops and is quoted in the literature as having a value of between 0.011 and 0.015 for Australian Standard Wheat (Edwards 1997). The work described here used the mean of those two values as the "true value" for C. Using the same logic as above, the absolute error was regarded as being one third of the difference between the mean and the extreme values quoted.

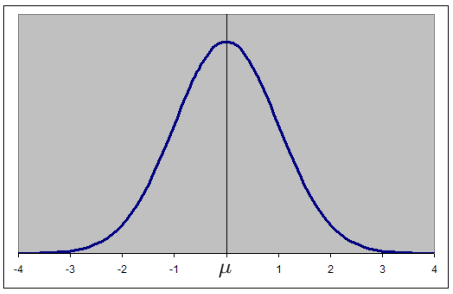
3. The output generated using the Taylor Series method is an absolute uncertainty surface for R. The uncertainty surface produced incorporated any

correlation which exists between the data layers. Correlation was only able to be determined between the A and K_0 input surfaces, with a ρ value of 0.53. There may be other cross correlated variables. But, these could not be included in the analysis because they could not be quantified.

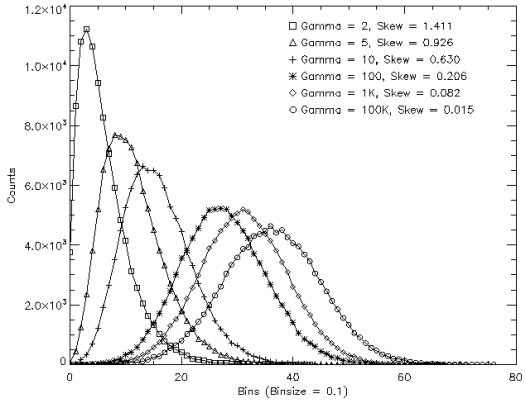
3.7.2 Monte Carlo Method

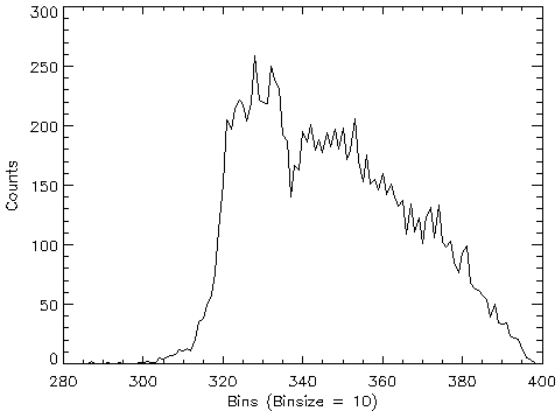
The Monte Carlo method of uncertainty propagation analysis assumes that the distribution of the uncertainty at each grid cell in all input data layers is known. The distribution is frequently assumed to be Gaussian with no positive or negative bias. For each of the data layers an uncertainty surface is simulated by drawing, at random, from an uncertainty pool defined by this distribution. Those uncertainty surfaces are added to the input data layers and the model is run using the resulting combined data layers as input. The process is repeated many times with a new realisation of an uncertainty surface being generated for each input data layer. The results of each run are accumulated and both a running mean and a surface representing deviation from that mean are calculated. Since the uncertainty surfaces are zero centered, the stable running mean may be taken as the true model output surface, and the deviation surface as an estimate of the uncertainty in that surface. Another important point is that the Monte Carlo method can be used in the analysis of disjoint functions, whereas the Taylor method can not. Again, the reader is referred to Heuvelink (1998) for a full description.

It reality, uncertainties in input data layers are not always evenly distributed. Therefore, for the Mitsherlich model, uncertainty simulations were drawn from distributions that were skewed to differing degrees. The skewed distribution was generated using the "RANDOMN" command in IDL with the "Gamma" option set to differing levels. This produces a family of curves with a variety of skews; a selection of these and an unbiased normal distribution are compared in Figure 3.3.



(a)





(b) (c)

Figure 3.3: (a) Normal (Gaussian), (b) Gamma and (c) typical soil test distributions. In (a), the positions of the mean (μ) and $\pm$ $1st$, $2nd$, $3rd$ and $4th$ standard deviations are marked on the horizontal axis.

3.8 Implementation

A procedure was written in IDL to perform the process described above. Simulated random data sets were generated for the appropriate inputs with the incorporation of appropriate uncertainty realisations. For each run, 100000 simulated data values were generated for each valid grid cell in each of the input data layers and coefficients. From these, the mean, absolute error and relative error for R was calculated for each grid cell. The number of simulations chosen was the minimum required to obtain stable statistical results at each each grid cell.

For the Mitsherlich model, in some cells either the achievable yield (A_y) is less than the target yield (Y_t) or soil K values are adequate for the achievable

yield and hence a calculation of the fertiliser requirement (R) returns a negative value. Where this happened the result was classified as invalid and the cell value set to null. For the N-availability, the same procedure was implemented if the values in the input layers or the simulated results where less than zero.

The level of agreement between the calculated values of N-availability and fertiliser recommendation (R) and their associated error surfaces calculated by the error propagation methods was determined by performing pair-wise linear regressions between the various outputs. Two surfaces that agree completely should have a slope of 1 and a correlation coefficient of 1.

3.9 Summary

Precision Agriculture is an agricultural concept that uses suitable technologies to improve agricultural practices and achieve targeted aims and has been applied in the wheat belt region of Western Australian.

The influence of uncertainties on the accuracy of Precision Agricultural models is discussed in the literature. This thesis aims to expand this knowledge by investigating the sensitivity to uncertainty of the algorithms in the SPLAT and Mitscherlich Equations.

Chapter 4

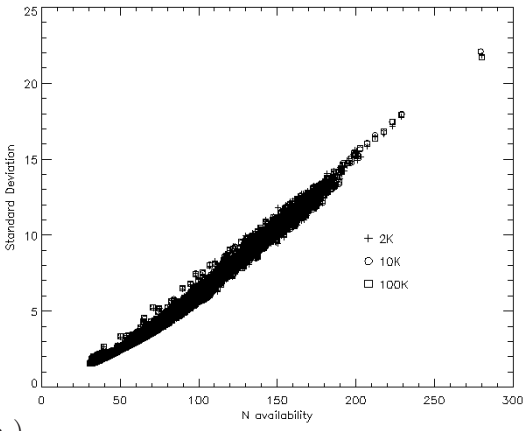
Precision Agriculture Results

The sensivity analysis of the simple and linear N-available model and the more complex Mitsherlich Model are discussed in the following sections. For both, when the uncertainty distribution was Gaussian, both analysis methods were applied and their results compared. When the distribution was skewed only the Monte Carlo method could be applied.

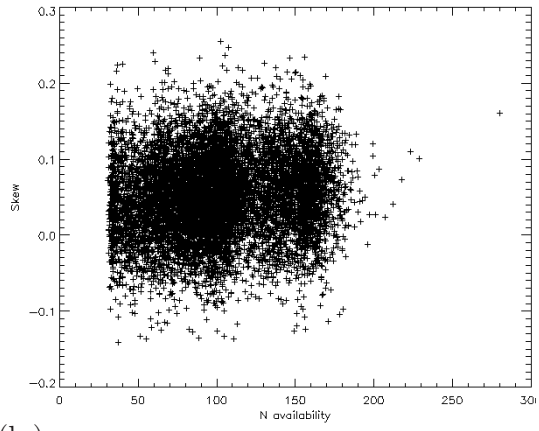
4.1 N-availability Linear Algorithm

For the Monte Carlo synthesised results, there is a high agreement between the N-available results calculated from the default (no uncertainty added) input layers and the synthesised input layers (correlation: 0.999, slope 0.999). In the synthesised results, there is also a high agreement in the calculated uncertainty even though the number of simulations investigated varied significantly (2000 to 100000, see Figure 4.1(a)). Also, in all cases the mean prediction to uncertainty relationship is almost linear with a small upward trend in the uncertainty vs mean N-availability relationship. Therefore, it can be concluded that the function does not significantly change the uncertainty of the N-available result.

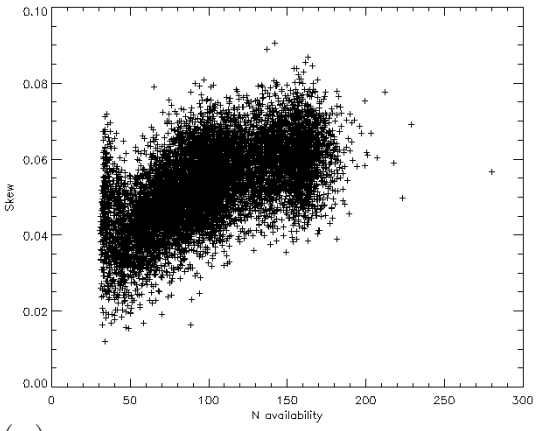
This is further reflected in the skew of the synthesised results (per point, see Figure 4.1(b) and (c)). At first impression, it would appear that there is a significant difference in these skew results. However, closer inspection shows



(a.)



(b.)



(c.)

Figure 4.1: Mean N-availability versus Standard Deviation (a) and Skew (b) and (c). The skew plots show the difference between when 2000 and 100000 simulations occur.

that for the low and high values of N, the centre of the skew is approximately the same ($\sim$0.4 and 0.6 respectively). The major difference is the range of the skew results, which is lower for the greater number of simulations suggesting that a higher number of simulations is required for a more accurate and easily interpreted results e.g. as seen in Figure 4.1(b), the increase in skew with higher N-availability is more easily seen.

Also of note is the fact that the skew is not centered on zero. As the skew of the synthesised input layers are centred on zero this suggests that the model itself is influencing not only the propagated uncertainty results but also the shape of the synthesised results. This influence is most likely to be greater in the more complex, non linear Mitsherlich model (as may also be the case for non normal inputs) and both are investigated in the following sections.

There is a good linear fit between the Taylor and Monte Carlo simulated uncertainty results, with a slope of 1.0 and correlation of 0.999. The relative uncertainty is also small, with a minimum and maximum of 0.048 (4.8%) and 0.078 (7.8%) respectively. This agrees with the change in % uncertainty observed in the Monte Carlo analysis results, allowing a greater confidence in the conclusion that the N-availability component of the $SPLAT$ model does not propagate uncertainty to any large or varying degree.

4.2 Mitsherlich Non Linear Model

Figure 4.2 shows the uncertainty of the Monte Carlo synthesised results versus the Taylor Methods results, for both a Gaussian and Gamma distribution (+ and − distribution, Gamma = 2 and 100000; a high and low skewed distribution). The maximum number of simulations (per grid point) is 100000 for all the following results.

Clearly, greatest agreement between the methods is when the uncertainty per grid cell calculated is low. The greatest agreement (with the Taylor method) is with the Gaussian distribution as would be expected. Closer inspection of

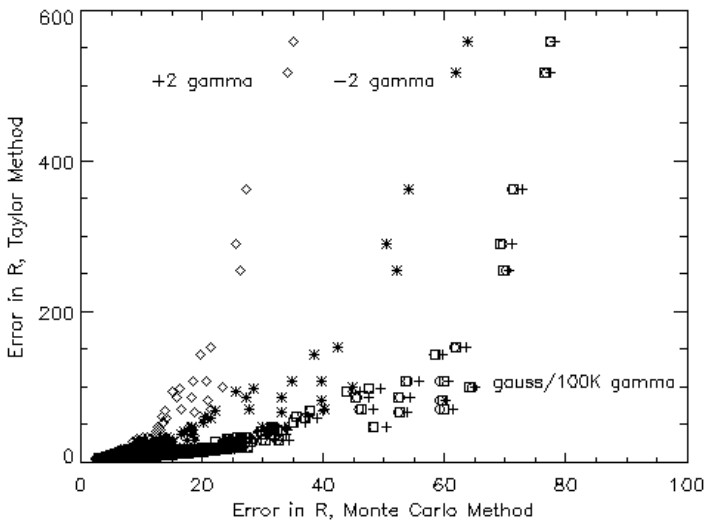


Figure 4.2: A comparison of the uncertainty of fertiliser requirement R (Kg/Ha), simulated (Monte Carlo) and Taylor methods.

the results show that the best agreement occurs at points where the fertiliser requirement R is less than or equal to 100 Kg/Ha (calculated uncertainty <30; regression analysis in this range gives a slope of 0.93). Also, in this uncertainty range the Gamma distribution of 100000 gives a similar result of 0.93. However, as is clearly seen, at higher values of R the Taylor Method error results increase significantly.

The heavily skewed distributions (Gamma = 2) clearly are in even less agreement with the Taylor Method result. Furthermore, in this case the positive and negative Gamma distributions are not in agreement. This is reflected (to a lesser degree) in Figure 4.3, which shows fertiliser recommendation values (R) plotted against Gaussian and Gamma distribution results. As one might expect, for the most part the best agreement occurs with the mean R calculated from the Gaussian distribution (slope of 0.935). However, above a value of 250 there is greater agreement with the negative (and to a lesser degree with the positive) Gamma distribution. The reason for this is due to the Gaussian distribution model producing simulations where R results are invalid and filtered out e.g. where A is

less than Y_t. This weighs the calculated mean in a negative direction. More importantly it highlights how biased results may occur depending on the structure of the model and the skew of the input variables. This is further discussed in the following sections.

Uncertainty relative to R

Figure 4.4 compares the calculated mean and standard deviations representing 1 σ of R per cell from the Gaussian and Gamma distribution synthesised inputs. It can be seen that there appears to be a similar pattern for all three distributions, with notable changes occurring in the R vs. uncertainty relationship at approximately 100−200 Kg/Ha and then at 250−400 Kg/Ha. The second of these changes is most likely due to the bias in the results due to the decrease in the number of valid data points. However, the first change suggests that at a point where one or more of the inputs contribute to a higher output R, a significant increase occurs in the uncertainty associated with that result. Also notable is that both positive and negative skew gamma inputs generally have lower uncertainty. This is most likely due to the concentration of the simulated inputs into a smaller range than occurs in a Gaussian distribution.

Figure 4.4(b) shows the results for a skewed distribution for which the gamma value is 100000. The R vs. uncertainty relationship is essentially the same as when gamma is set to 2, but notably smoother in the curve (as R increases). There is also very good agreement between the Gaussian and Gamma distribution results. This is expected as the gamma distribution of 100000 is equally biased (and hence the skew is very close to 0).

Skew relative to R

The skew in the R results calculated from the synthesised datasets show three features:

1. As in the uncertainty results, the skew values appear to remain approximately the same when R is equal to or less than 100, but then increases

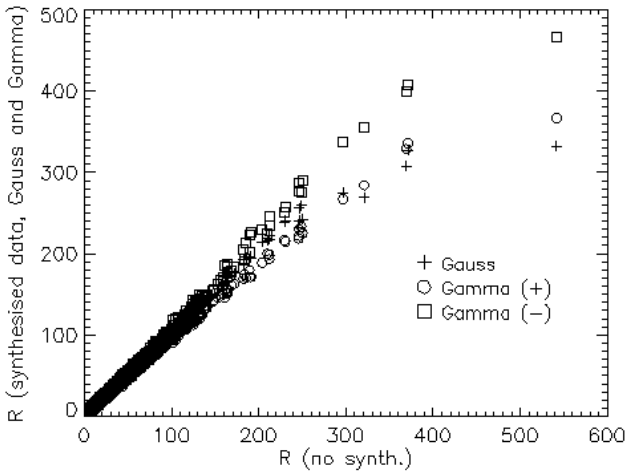
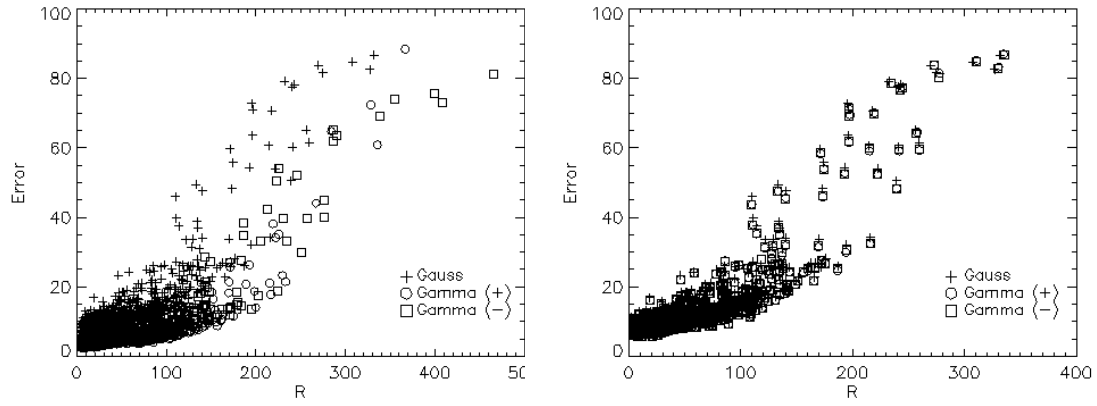


Figure 4.3: Comparison of simulated and directly calculated R values (Gamma = 2).



(a.) (b.)

Figure 4.4: Mean synthesised R versus uncertainty. (a) Gamma = 2 (b) Gamma = 100000. For comparison a Gauss distribution is included in both plots.

(see Figure 4.5). Three of the four Gamma distributions investigated eventually peak and then fall. However the negatively skewed Gamma=−2 distribution continues to increase. This mirrors the valid results pattern discussed earlier.

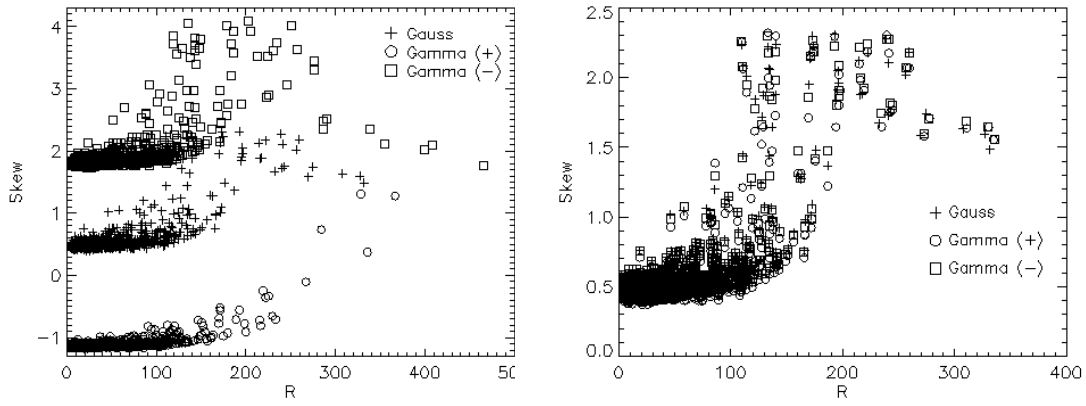
2. The heavily non normal distribution in the inputs is reflected in the position of the skew results relative to the Gaussian skew results, easily seen in Figure 4.5(a).

3. As shown in Figure 4.5(b), the skew of the Gaussian and equally weighted Gamma distribution is not centred on 0, even when the values of R are low (and hence considered valid) and the skew of the input layers is insignificant. Analysis of the Mitsherlich model shows that this is a due to the mathematical structure of the model and this is important as it may bias R and its associated uncertainty.

4.3 Conclusions

The values of N-available (nitrogen available) and R (fertiliser requirement) calculated from the given data layers are in close agreement with the mean values calculated from the Monte Carlo synthesised datasets under a Gaussian assumption. The exception occurs at grid cells where the higher mean R occurs, which is also where the lower number of valid simulation results occur.

As is more likely the case for linear models, uncertainty propagation in the linear N-available model is negligible. However, the model structure did influence the skew of the N-available results (calculated from the synthesised input layers). For the Mitsherlich model, uncertainty propagation increases as R increases and the rate of this increase can vary significantly and abruptly. In the case of this non-linear model, there appears to be several reasons for this which are dependent on how the input/model interaction can change as the input values change.



(a.) (b.)

Figure 4.5: Mean synthesised R versus skew. (a) Gamma = 2 (b) Gamma = 100000. For comparison a Gauss distribution is included in both plots.

The closest agreement in the absolute uncertainty trends is seen between the combined 1st and 2nd order Taylor series results and the Monte Carlo Gaussian distribution for calculated R values of less than or equal to 100. Above this value, there is considerably less agreement.

For the skewed Gamma distribution, the best in the calculated R agreement is seen when the synthesised dataset has little positive or negative bias (within a given valid range for R). However, the heavily negatively skewed distribution produces results that are less prone to the models bias at higher R values.

Both the uncertainty and skew statistical results (for the calculated R) can give an insight into how a model and/or its inputs may influence the validity of the final results.

Chapter 5

Niche Envelope Models

5.1 Distribution Modeling

Species distribution models (SDM) are models relating field observations [1] to environmental predictor variables (Guisan & Zimmermann 2000) to calculate the species distribution, which could then be used to predict the suitability of any other site for said species. Much effort has been put into the development of these models in the field of ecology as they elucidate spatial and temporal patterns of organism behavior and system dynamics. This type of analysis is also frequently referred to as "Bioclimatic Modeling." A list of commonly used models, their classification method and limitations is given in Table 5.1. A schematic description of two of these models is illustrated in Figure 5.1.

The "envelope of a species" is the set of environments within which it is believed a species can live (Walker & Cocks 1991). This 'delineation' of a species' distribution is frequently used to predict the full geographical range (of a species), particularly when an estimate of the probability of occurrence or the relative suitability at a given site is required (Guisan & Thuiller 2005). Most modeling approaches developed for predicting animal and plant distributions have their ori-

[1] Field observations are often described as, or part of, the entities of the system being studied. For example, Segers, Branquart, Caudron and Tack (Segers, Branquart, Caudron & Tack 2001) discusses the "entities of biodiversity" required for accurate and reliable "biodiversity indicators," where the "chosen indicators do not indicate biodiversity *per se*, but selected aspects or entities of biodiversity."

Model type	Example	Classification method	Features
Boxcar	BIOCLIM	multilevel rectilinear envelope	dimensions treated independently performs poorly on covariate data includes some dissimilar sites excludes some similar sites easily implemented simply described envelope
Convex hull	HABITAT	binary convex envelope	tightly constrained envelope excludes many similar sites difficult to implement computationally expensive
Distance Based	DOMAIN	continuous point-to-point symmetric	variable sensitivity performs well with limited site data gives similarity value to all sites easily implemented

Table 5.1: Commonly used model types. Modification of table from Carpenter et al. (1993)

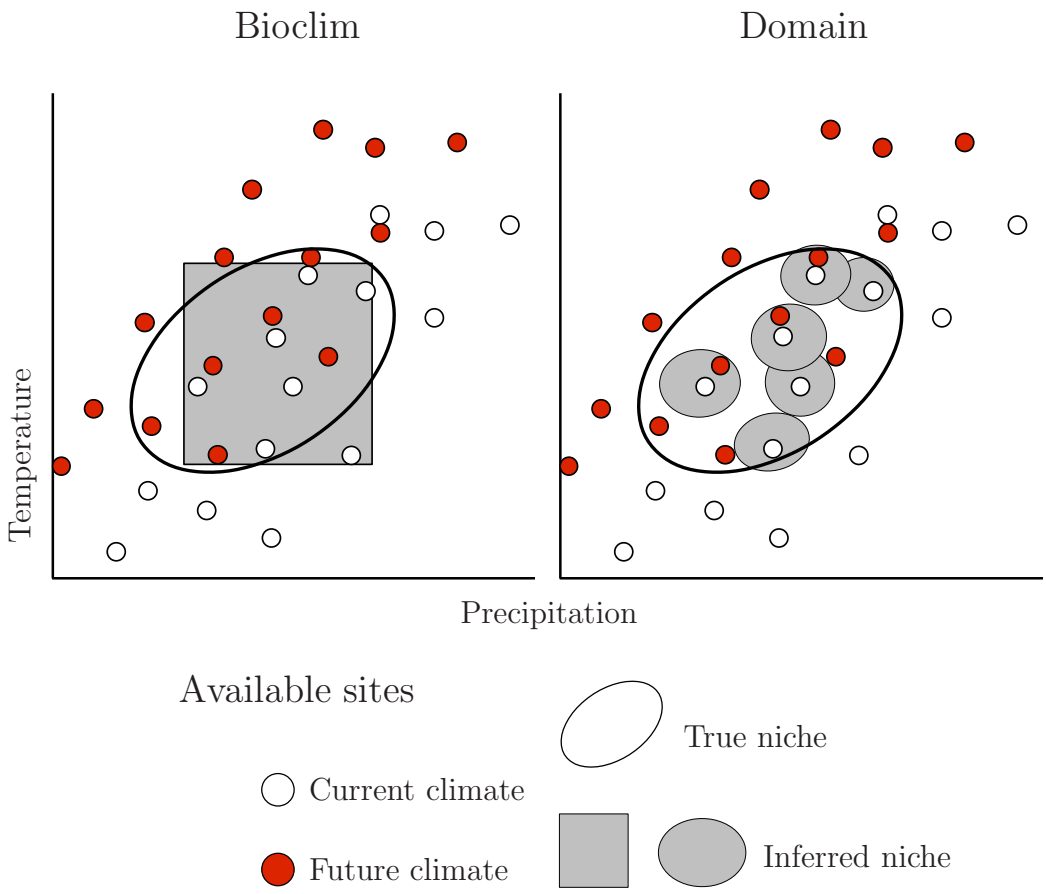


Figure 5.1: Schematic description of predicting the distribution of a species under different climates using two climate envelope models, Bioclim, and Domain. There are 15 sites, with different climates in the two time periods. The true requirements of the species are constant and indicated with an ellipsoid. The inferred requirements do not fully overlap with the true requirements because there are insufficient sites where the species has been observed and/or because parts of the true niche are currently not present on the landscape, and because the model methods are imperfect. Under future conditions, model performance is diminished because some sites are incorrectly classified as not having the species (false negatives) (Hijmans & Graham 2006).

gins in quantifying species−environment relationships (Guisan & Thuiller 2005). The development of these models occurred in three phases: Firstly, non-spatial statistical quantification of species-environment relationship based on empirical data, secondly expert based (non-statistical, non empirical) spatial modeling of a species distribution and thirdly, spatially explicit statistical and empirical modeling of species distribution.

As summarised in Elith and Leathwich (Elith & Leathwick 2010) and Graham et al. (Graham, Elith, Hijmans, Guisan, Peterson, Loiselle & Group 2008), SDM's are used in both applied and theoretical research to predict how species are distributed and to understand attributes of species' environmental requirements. For example, species distribution modeling has been used to manage species of conservation concern (Gabert, Papes & Peterson 2006), create richness maps for conservation planning (Loiselle, Howell, Graham, Goerck, Brooks, Smith & Williams 2003, Rissler, Hijmans, Graham, Moritz & Wake 2006) and predict the geographical spread of invasive species (Peterson 2003). Among the earliest found examples of modeling using correlations between species and climate are the prediction of the invasive spread of a cactus species in Australia by Johnson (1924) and the assessment of climatic determinants in the distribution of several European species by Hittinka (1963). This area has continued to evolve with the technique now applied in the analysis of ecology, biogeography, evolution, conservation biology and climate change research (Guisan & Thuiller 2005) (see Table 1 in Guisan and Thuiller (2005) for additional applications and details).The data required for these models vary depending on the model type, its complexity and structure. However, the inputs generally include species occurrence data and environmental spatial data layers which are then processed to create the species predictive model.

5.1.1 Species Distribution Modeling Theory

Bioclimatic models are generally static and probabilistic in nature since they statistically relate to the geographical distribution of species or communities

to their present environment. Traditionally, they have been correlative rather than mechanistic; an approach which characterises - in a statistical sense - the complex response of a species to a complex, changing (as the process/species relationship may vary) process (Barry & Elith 2006). The method relies on the *niche* concept (Guisan & Zimmermann 2000), where the concept is seen as either driven by the environmental requirements (the "requirement" niche) or by the impact the species can have on the environment (the "impact" niche) (Leibold 1995). As these concepts apply to different scales, only the requirement concepts and *environmental* niche (see Austin (1992)) are usually considered in SDM's. The classification of a model is determined by its intrinsic properties (Figure 5.2).

The bioclimate envelope modeling approach has its foundations in the ecological niche theory (Pearson & Dawson 2003). As described by Pearson and Dawson (2003), Hutchinson (1957) defined the fundamental ecological niche as comprising "those environmental conditions within which a species can survive and grow." Furthermore, Hutchinson proposed that the fundamental niche would completely define the ecological properties of a species as a "conceptual space whose axes include all of the environmental variables effecting that species" (Austin, Nicholls & Margules 1990, Leibold 1995).

The adequacy of models is dependent on ecological processes driving the true distribution and the processes used to observe and model it. Understanding these factors requires an understanding of the combined effects of a species' ecology and the measurement-prediction process; and we must also distinguish ecological patterns from statistical artifacts (McPherson, Jetz & Rogers 2004). As discussed in Section 5.1.3, the validity of bioclimatic modeling for certain applications therefore contributes to this uncertainty.

Ecological theory and observation suggest that the association between environmental variables and species presence and abundance should be non random (Barry & Elith 2006). However, the processes in this relationship are generally complex with both abiotic and biotic factors possibly influencing the distribution

and abundance of a species (Leathwick & Austin 2001). To compound this, and despite the wide use of predictive models, many applications give insufficient considerations to the error and uncertainty. These sources of "modeling error" result from the interactions of two broad error groupings (1) deficiencies in the data and (2) deficiencies in the model's ecological realism.

5.1.2 Empirical and Mechanistic Models

Species distribution models can be loosely classed as either empirical or mechanistic (Figure 5.2):

Empirical

In its purest form a bioclimate envelope can be defined as constituting the climatic component of the fundamental ecological niche, or the 'climatic niche.' Therefore, this bioclimatic model in its purest form considers only climatic variables in their processing and not other environmental factors. Bioclimatic models of this type are based on empirical relationships where the climate variable is correlated with the species distribution i.e., the best indicator of a species' climatic requirements is its current location. These correlation models then characterise bioclimatic envelopes based on the *realized* niche since the observed species distributions are, in reality, constrained by non climatic factors, including biotic interactions (Austin et al. 1990, Guisan & Zimmermann 2000, Guisan, Theurillat & Kienast 1998).

This methodology is used in the study of both present and future climate scenarios. For example, the study by Hutley et al. (1995) showed that the principal determinant of the distribution of eight European plant species, in the present climate envelope, was the macro climate. In the case of possible future climate envelopes, Bakkenes et al. (2002) modeled the climate envelope for 1400 plant species by multiple regression analysis and used this to obtain predictions about plant diversity and distributions in 2050; Peterson et al. (2001) modeled the effect of climate change on the ecological niches of a bird family; and Pearson et al. (Pearson, Dawson & P. M. Berry 2002) developed a model which

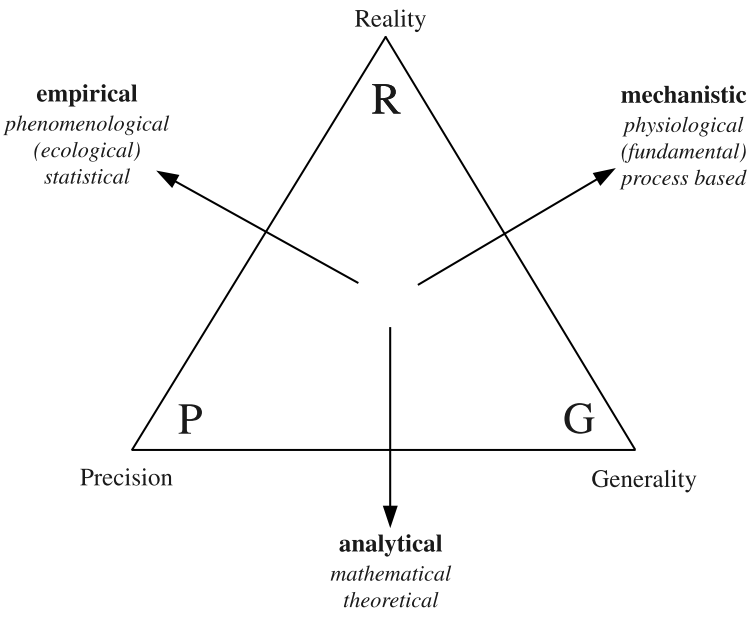


Figure 5.2: A classification of models based on their intrinsic properties (Guisan & Zimmermann 2000).

couples an artificial neural network with a climate-hydrological process model to ensure "encapsulation of the full extent of future climate scenarios within Great Britain without extrapolating outside of the models training dataset." In a later work (Pearson, Dawson & Lui 2003) a modeling framework was developed which integrated land-cover data into a correlative bioclimatic model in a scale-dependent hierarchical manner. In this, Artificial Neural Networks were used to "characterise species' climatic requirements at the European scale and land-cover requirements at the British scale."

Finally, these types of models are used extensively in predicting how climate change will affect the world's food crop productivity in the future. The results of these are then used in policy development. For example, Parry et al. (2004) considers the projected effects of "climate change on global food supply under different pathways of future socio-economic development" using complex models. Rosenzweig and Parry (1994) discusses how future scenarios, as calculated from such models, might impact world food supply and how this will vary across the

world.

Mechanistic

Other research has concentrated on determining the *fundamental* niche based on the *physiologically* based mechanistic relationship between climate parameters and species response. Examples of this include the model developed by Prentice et al. (1992) "to predict the global patterns in vegetation physiognomy from physiological considerations influencing the distributions of different functional types of plant;" the study by Haxeltine and Prentice (1996) which couples vegetation distribution directly to biogeochemistry and the study by Sykes et al. (1996); which developed a model to map the distribution of northern European major trees both in present and future climate scenarios. These bioclimatic models aim to identify "the *realized* niche by modeling the physiological limiting mechanisms in a species' climatic requirements" (Pearson & Dawson 2003).

5.1.3 Limitations of Bioclimatic Modeling

There are fundamental limitations to the predictive capacity (the Model's Realism) of a bioclimatic model, regardless of the methodology used to characterise the bioclimatic envelope (Pearson & Dawson 2003). Once a model is fitted, the next logical step is to evaluate its fit to the modeling data and/or its predictive ability. To determine the discrepancies between the model and data and/or its predictive ability, a number of measures are available. However, it is important to remember that the end use of the model dictates the most relevant measures and datasets for evaluation (Barry & Elith 2006). For example, if a bioclimatic model is used to predict a region's viability for a particular crop, several factors such as climate, soil type and nutrient levels are very important. However, given that the last two variables can be ameliorated but climate cannot, the climate variables − species' distribution relationship is critical in choosing the model and then in the assessment of the model's limitations.

Biotic Interactions

The difference between how a species functions on its own and when in the presence of other species (i.e. the inter-species interactions) can be significant but is absent from most examples of SDM research (Guisan & Thuiller 2005). This was seen as a significant flaw in bioclimate envelope modeling techniques by Davis et al. (1998). Earlier, the work by Leibold (1995) highlighted the importance of "environmental requirements and environmental impacts of species to highlight the dichotomy between the responses of organisms to the environment and their effects upon it." However, when these models are applied at macro-scales the impact of biotic interactions are minimised (Pearson & Dawson 2003). This is most likely due to the dominance of the climate influences on species distributions at larger scales, a possibility that is supported by the success of bioclimatic *fundamental* niche models in calculating present distributions across large areas. Examples of this include the hard-fern (*Blechnum spicant*) European scale comparison of observed and simulated distributions by Pearson et al. (2002) and the study by Beerling et al. (1995) which investigated the predictive capacity of climate response surfaces for the distribution of Japanese knotweed (*Fallopia japonica*). Close agreement between observed and simulated distributions was observed, suggesting that the European distribution of this species was climatically determined. Therefore, on a macro-scale bioclimatic models may be suitable for making broad predictions as to the likely impact of climate change on the distribution of a species (Pearson & Dawson 2003).

Evolutionary Change

It is usually expected that evolutionary change occurs over long time scales and that the tolerance range of a species does not change as it changes its geographical range. Because of this, in literature covering the biotic effects of past and potential future climate change, the genetic adaptation of a species is rarely considered (as range shifts were frequently seen as the expected response). In both cases this is not necessarily the case. With regard to the first case, the importance of the

potential of rapid evolutionary change has been shown by studies of several butterfly species and the emergence of dispersive phenotypes (Thomas, Bodsworth, Wilson, Simmons, Davies, Musche & Conradt 2001). These evolutionary changes have emerged during the recent change in climate and are now showing a greater ability to cross previous barriers. This is showing that rapid evolutionary change is not confined to the range margins of highly dispersive species. There is also growing evidence of species' ability to rapidly adapt to regions outside its known geographical boundaries. For example, the study by Woodward et al. (1990) of transplant population experiments has shown the rapid *in-situ* adaptation of the navelwort (*Umbilicus rupestris*) to a new geographical range. This winter-green species occurs naturally on the western maritime areas of southern Britain and, as a study of the impact of lower winter temperatures on this species, was transplanted to the cooler climate of Sussex (Crowburrough) at an elevation of 157m (a location outside the natural climate range of the species).

The implications of rapid evolutionary change for bioclimate envelope modeling are important since, for species that can rapidly adapt, the assumption of niche conservatism (where adaptation is assumed slower than the rate of climate change), will be wrong. However, it would be incorrect to assume that all species have this ability, as scientific research suggests that many plant's evolutionary rates are to slow (e.g. Etterson and Shaw (2001)). Bioclimate studies have also been used to show that adaptation to future climates will not occur for some species (e.g. Huntley et al. (1989, 1995)).

In conclusion, niche models cannot account for adaptive changes in bioclimate response. This limits the effectiveness of the models to species which are not expected to undergo rapid evolutionary changes over the time-scale being studied.

Species Dispersal

The unprecedented rates of climate changes anticipated to occur in the future, coupled with land use changes that impede gene flow, can be expected to disrupt the interplay of adaptation and migration of a species. This may be of less

significance when using bioclimate mapping for future crop predictions, but it is important when studying a specie's ability to adapt to environmental change and the ability of the species to disperse, especially in longer time-scales. For example, palaeoecological studies of the late-Quartery have shown that terrestrial plant and animal species adapt to long-term climate change by migrating to track the changing environment rather than evolving to adapt to it (Collingham, Hill & Huntley 1996).

The ability of a species to adapt to changing climate is determined by the individual dispersal characteristics of a species. Bioclimate envelope models do not account for species dispersal mechanisms, but instead aim to predict the potential range of organisms under changed climate. Though there is great potential to couple bioclimate envelope models and dispersal simulations (Carey 1996, Peterson, Sanchez-Cordero, Sobero'n, Bartley, Buddemeier & Navarro-Siguenza 2001), it is apparent that current predictions of potential distributions may differ greatly from actual future distributions due to migration limitations (Midgleya, Hannah, Millara, Thuiller & Booth 2003).

The ability of a species to disperse is also a function of the structure of the landscape (such as mountain ranges) and fragmentation of natural landscapes (through human induced land use changes that impede gene flow and can be expected to disrupt the interplay of adaptation and migration (Davis & Shaw 2001)). In these cases predictions derived from bioclimate models will be erroneous. Therefore, for these predictions to be accurate, it will be necessary to have a good understanding of a species' ability to migrate through dynamic heterogeneous landscapes within the constraint of changing bioclimate envelopes (Pearson & Dawson 2003). It is also important to remember that bioclimate studies across long time frames (such as the study by Davis and Shaw (2001), which studied the last 25 thousand years - the late Quaternary Period) will not provide the detail required to study subcontinental changes across a shorter time frame, such as 100 years.

These models may also identify the broader distribution trends but not the

finer details which may be due to human influenced disruption of species-climate equilibrium. Therefore, the degree to which a species can occupy its full ecological niche depends on the scale of the model, the dispersal ability of the organisms and the history and biology of the species (Pulliam 2000, Tyre, Possingham & Lindenmayer 2001).

Model Specification Error

As discussed by Barry and Elith (2006), "With a well-specified model and an adequately large random variable sample from the population, the statistical component of error is easily quantified and incorporated into predictions. But in most practical settings, the data sample is rarely random and the model not fully adequate, and these factors combine to induce error in the final predictions." Therefore, the principal questions relating to a model's accuracy are:

(1) Does our model approach the true model ?

Most studies conclude that, provided the 'true' model is nested within the model specification, the model estimated using maximum likelihood or other consistent techniques will converge to the true model as the sample size increases (Welsh 1996). If the 'true' relationship is not contained in the model, then over and underestimation will typically result in different parts of the covariate matrix. This will result in errors in inference and prediction and cannot be fixed by increasing the sample size.

(2) How do we estimate the response surface ?

A response surface maps how the probability of a species' presence varies within the environmental variables. For example, Figure 5.3 shows a hypothetical response surface where probability of occurrence depends on rainfall and temperature. Different techniques approach the problem of trying to estimate a complex surface with limited data in different ways, but all try to approximate the surface by simple components. For example, as discussed by Barry and Elith (2006), in climate envelope approaches:

1. The shape of the response surface is mostly ignored.

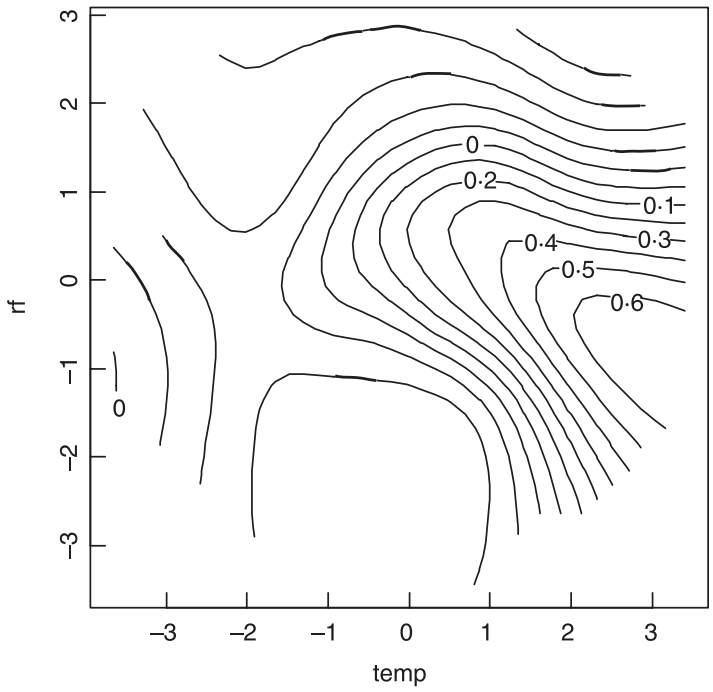


Figure 5.3: A response surface showing the response of the species to rainfall (rf) and temperature (temp). Diagram from Barry et al. (2006)

2. The use of percentiles implies a belief in core and non-core habitat and presupposes uni-modality of response pattern to a gradient.

3. Envelope approaches seek to delineate the non-zero components of the response along each gradient.

4. The region is assumed to be rectilinear and oriented with the environmental axes.

5. Distance-based approaches use observations close to the point that is being predicted to asses suitability. Thus they attempt to model the response surface as a local smoothing.

Generally, regardless of the model and its type, errors in model specifications are essentially ubiquitous and the majority of models are typically simpler than the real-world complexities they seek to describe.

5.1.4 Present and Future Predictions: Empirical versus Physiologically Based Models

It may be argued that the bioclimate envelope determined with a physiologically based model will better represent a species' absolute climate limits than that identified through the empirical (correlative) approach (which assumes equilibrium). Examples of work investigating this not-in-equilibrium situation can be found in Woodward (1990) and Peterson et al. (1999). This situation exist in many Australian ecosystems such as the Tasmanian temperate rain-forests, which have remnants of species from an earlier, wetter climate period, as well as the Eucalyptus species. Therefore, a large event (such as a severe bushfire) would shift the equilibrium in favor of the more fire tolerant species, resulting in the system being dominated by the Eucalyptus.

Despite the advantages of these more complex models, when using a physiologically based model type, the limitations of the model's ability to account for nonclimatic influences should be well understood as they will increase the model's inaccuracies (Pearson & Dawson 2003). More specifically, when applied to the prediction of future distribution: (1) Predicted future species distributions based on the physiologically determined fundamental niche are unlikely to be as accurate as those based on correlations between the observed distribution and the current realised niche and (2), there is increased evidence that the concept of undifferentiated species comprising individuals with broad tolerances is not correct as intra-species variation makes it impossible to define precise limits to a species' climatic tolerance. The potential importance of rapid evolutionary change means that some species' climatic tolerance may alter in the future, making the fundamental niche unstable over time.

All correlative species distribution models assume that the modeled species is in a pseudo-equilibrium with its environment (Barry & Elith 2006, Guisan & Thuiller 2005, Guisan & Theurillat 2000). This issue has been raised in the literature (see Araújo et al. (2005)) but this has not been extensively covered. There-

fore, an important criticism of this is that species distributions as observed today may not be in equilibrium with the current climate, nor indeed are they necessarily determined primarily by climate (Pearson & Dawson 2003, Woodward 1990). The limit on species distribution may be determined by a physiological response to the environment, but the strength of this response may be influenced by interactions with other species (Leathwick & Austin 2001). Other influencing factors may be physical barriers to dispersal and/or human management. As a result of this, fundamental niches in correlative bioclimate envelope methodologies may not represent absolute limits to species-ranges resulting in present and future distributions showing very different realised niches (Pearson & Dawson 2003).

Even though biological system are complex, the successful use of bioclimatic envelope models to simulate the distributions of vascular plants in Europe (Beerling, Huntley & Bailey 1995, Pearson et al. 2002, Huntley, Berry, Cramer & Mcdonald 1995) supports the hypothesis that continental-scale distributions are principally determined by climate (but this may not apply to all species). These models may also identify the broader distribution trends but not the finer details which may be due to human influenced disruption of species-climate equilibrium. Therefore, the degree to which a species can occupy it full ecological niche depends on the scale of the model, the dispersal ability of the organisms and the history and biology of the species (Pulliam 2000, Tyre et al. 2001).

While it is argued that physiologically based methods are superior, they also have limitations that, when applied at appropriate scales, make them no more accurate than correlative techniques. The constricting of fundamental response curves (in realised niche models) has been investigated (Austin et al. 1990), but the uni-laterality of biotic versus abiotic pressures has only rarely been discussed in the literature (Guisan et al. 1998). Correlative techniques also have the advantage of not requiring physiological data and hence can be applied to a larger number of species. This enables conclusions, regarding potential impacts of climate change, to be made on a wide range of species, including current and potential crop species (Jarvisa, Lanec & Hijmans 2008, Hijmans & Graham 2006).

5.1.5 Application of SDM and the Hierarchical Modeling Framework

Pearson and Dawson (2003) suggested that the use of bioclimate envelope models to identify a species' suitable climate space should form an important step in a broader modeling framework. The framework would be a hierarchy of factors operating at different scales. Therefore, the bioclimate envelope models would be used at a continental scale where climate is the dominant factor. Providing that the higher level conditions are satisfied, the factors further down the scale can be included such as topography and land type at local scales. Figure 5.4 illustrates how this might apply to a specific area, where there is sufficient information about the environmental variables. As one would expect, there would be differences in the 'scales domains' between different continental regions.

5.2 Uncertainties in Ecological Modeling: Data Introduced Error

This section summarises the principal ways in which uncertainty in input data can influence the prediction results of a model.

Missing Predictor Layers

No study aims to use a comprehensive suite of all known direct predictor data layers as, even with all these covariates known, prediction is not perfect because of other variations (Tyre et al. 2001) such as low genetic diversity in a crop species. The predictor variables of models are almost always incomplete in time and space, over the range of scales at which the processes operate. They may also be insufficient to explain the suite of species' interactions and historical and current disturbances that can affect a species' occurrence (Horne 1983, levin 1992). This lack of predictor variables missing from most models (i.e the missing covariates), generally reflects a lack of knowledge of which environmental factors constrain

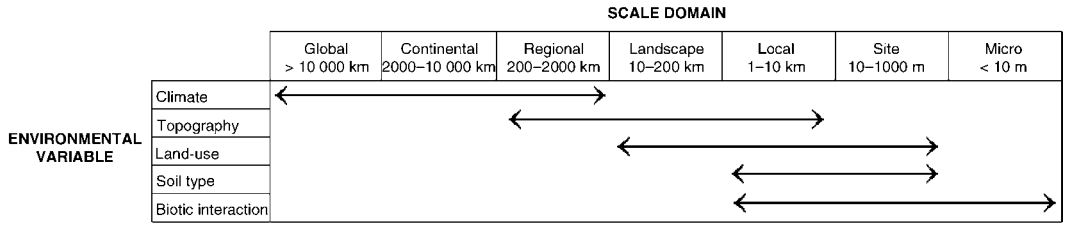


Figure 5.4: Schematic example of how different factors may affect the distribution of species across varying spatial scales. Characteristic scale domains are proposed within which certain variables can be identified as having a dominant control over species distributions. Approximate spatial extents have been assigned to categories of scale based in part on Willis & Whittaker (2002). It is assumed that large spatial extents are associated with coarse data resolutions, and small extents with fine data resolutions. Diagram from Pearson et al. (2003)

the distribution of a species throughout its range and a lack of spatial data sets describing attributes known to be important. As specified by Barry and Smith (2006), a "sufficient" set of covariates may be defined as that which allows a model to be specified that does not have significant spatial errors or global errors, with respect to a specific context and end use. Unfortunately, even for mechanisms that are well understood, directly relevant quantitative data that can be used for modeling are often unavailable. As discussed by Ferrier, Drielsma, Manion and Watson (2002), these entities of complex ecological systems that "have not yet been discovered, let alone had their distributions mapped at a spatial scale appropriate for regional conservation planning." To minimise this limitation, "a widely applied solution to this problem is to use those entities for which we do have distributional information as *surrogates* for spatial pattern in biodiversity as a whole. However, even for a surrogate species or groups of species, available information on fine-scaled spatial distribution is usually far from complete."

Small sample size in known predictor layers

As expected, in the layers chosen for the bioclimate mapping, the existence of adequate information on the spatial distribution of biodiversity is very important.

For example, models that have a spatial interpolation component need adequate local sampling in the species spatial range as this cannot be estimated accurately with small or unevenly distributed samples. Unfortunately, such information is usually grossly incomplete and problematic because absence data are generally unavailable, sample sizes are often small and geographic bias and spatial error in the data are generally known but not corrected or simply unknown (Hijmans & Spooner 2001, Graham, Ferrier, Huettman, Moritz & Peterson 2004, Wieczorek, Guo & Hijmans 2004, Rowe 2005). Also, small sample sizes will be a limitation if it is insufficient for a particular type of model. For example, in presence absence-data, sample size needs to be assessed in relation to the least frequent class rather than a count of the total number of sites (where that class is found). A species may be genuinely rare (i.e. not due to lack of measurements), but still be an inadequate sample for these modeling methods (Barry & Elith 2006).

In order to help resolve this limitation, the availability of collected data has become more easily accessible as a result of data generation, improvement and sharing initiatives between natural history museums and the broader scientific and conservation communities (Graham et al. 2004, Tsutsui 2004). Also, work is being done to document the effects of data sparsity on a model's accuracy. For example, Elith et al. (Elith, Graham, Anderson, Dudk, Ferrier, Guisan, Hijmans, Huettmann, Leathwick, Lehmann, Li, Lohmann, Loiselle, Manion, Moritz, Nakamura, Nakazawa, Overton, Peterson, Phillips, Richardson, Scachetti-Pereira, Schapire, Soberon, Williams, Wisz & Zimmermann 2006) showed that accurate models can be made with presence-only data and other studies have explored how the number of occurrences and a bias in data influences a model's accuracy (e.g. see Mcpherson et al. (2004) and Hernandez et al. (Hernandez, Graham, Master & Albert 2006)).

In summary, modeling options for a small number of presence records or limited predictor layers include (1), create a habitat suitability index model. This method is based on the judgments of experts who identify critical variables that can be used to identify suitable habitat through a conceptual model of how

the species responds to the environment. (2), Use a statistical model where the number of candidate variables is limited to those that can be supported numerically. However, restricting a model to few predictor variables averages the response over all the omitted variables, and may result in a misleading model. (3) Use models that can be used to model either the collective properties of the biota (Ferrier, Watson, Pearce & Michael 2002), or to make predictions for an individual species from these collective properties. This type of model, which is called a "community" model, uses information from a wider set of species to construct a context in which individual species are described (Barry & Elith 2006).

Biased samples

The ideal data for modeling are collected using a planned sampling regime, structured to sample the major environmental gradients likely to be important for the species and covering the spatial extent of the region of interest (Austin & Heylingers 1989, Cawsey, Austin & Baker 2002). Unfortunately this ideal is seldom (if ever) achieved as:

1. Species data may not have been collected for the purpose of the study and may comprise of an ad-hoc collection of existing data that is biased in geographical and environmental space.

2. The modeled relationships are dominated by the patterns at sampled sites, rather than the true pattern of the entire study area. This can lead to marked spatial variation in prediction uncertainty (spatial error).

3. In situ samples can be biased by inappropriate sampling techniques and less than adequate samples, which could introduce unknown biases.

The most obvious way to minimise bias related error is to supplement the current datasets with new data from targeted areas. The other is to use exploratory statistical techniques to diagnose the bias. These include making plots

in geographical space, analysing site density in environmental strata and using statistics to measure distances between sets of sites in multivariate environmental space.

Lack of absence records

Ecological theory and observation suggest that the association between environmental variables and species presence and absence should be non random (Barry & Elith 2006). However, the processes in this relationship are generally complex, with the reason for the absence of a species being just as important as the reason for its presence. This presence-absence relationship is often not directly related, as the abundance and distribution of a species can be due to a combination of both abiotic and biotic factors (Leathwick & Austin 2001). Therefore, a model fitted with presence-absence data is most likely to give a superior result to presence only models (Barry & Elith 2006). Despite this, bioclimate models that use presence-only data are still commonly used, in part, because of the limitations of available data sets. Data sets without absence records are common, particularly in natural history collections, and their development and application is ongoing.

Errors in variables

Predictor variables can have errors that can be random and/or biased. This can occur when input layers are interpolated from point data and will have errors consistent with those of the interpolation method and the quality of the original point data. As stated in Barry and Elith (2006), "the problem of such errors in predictor variables can become overwhelming, and a common reaction is to ignore them, an approach that can have some justification in a statistical sense." An example of when "ignoring" is valid is where the errors at the prediction sites are the same errors as those for model building. In this case the error can be ignored as "in cases where the prediction sites have the same errors as those used for model building, the model will already reflect the errors and the predictions will be consistent with the data." (Barry & Elith 2006).

However, even when such errors are ignored, large errors in the result can arise if there is error (in the input layers) of differing structure. For example, the relationship between "proximal" or "distal" variables may change significantly with location across the area being studied (Austin 2002, Barry & Elith 2006). Subtle errors related to this phenomenon are particularly likely when predictions are made for a geographical area X, from a model that was built using data describing a geographical area Y. Another possible source of error that can compound this problem is that strong spatial patterns in predictor variable errors will inevitably produce local prediction errors.

Interactions between data errors and model mis-specification

"Interactions between data and model errors can be illustrated through the impacts of missing covariates on a model robustness (Barry & Elith 2006)." A missing covariate is a variable that would provide additional predictive power if it was known and observed. As these are often poorly described, "the covariates in most models are typically coarse correlates with more proximate (closer) factors" controlling the species' prediction distribution (Barry & Elith 2006), as described in (Austin 2002).

Therefore, as the use of a correlated variable assumes that the correlation structure between the fitted predictor and it more proximate components is stable throughout the sampling domain, departures from this will result in spatially correlated errors. Also, significant spatial patterning in residuals (i.e. large local errors) are likely if missing covariates have a non random spatial distribution. This is a feature commonly observed in soil attributes illustrated in Figure 5.5. The impacts of this are summarised in Table 5.2. This example shows the value of the information in the soil properties layer. However, as is clear from the discussion in Chapter 3.3 (but relevant to scientific fields beyond Precision Agriculture) the quality of the soil property map is just as important as its presence.

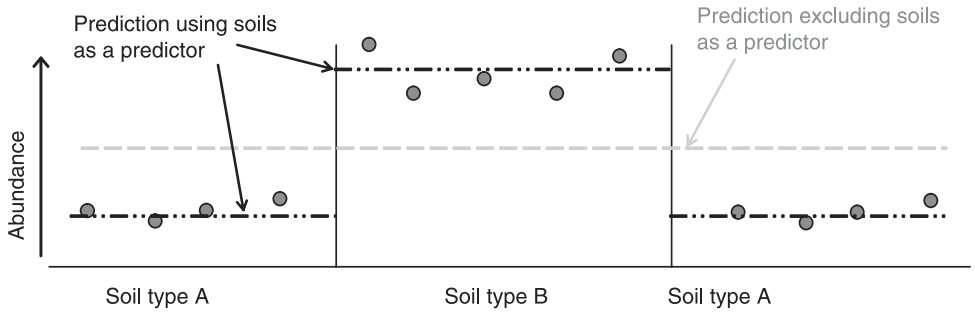


Figure 5.5: An illustration of the impact of a missing covariate on modeled predictions of species abundance. The X axis is in geographical space and circles represent the observations. When the covariate is missing, predictions are averaged across both soil types. Diagram from Barry and Elith. (2006)

Model	Model form error	Impact of missing covariates	Modeling recommendations
boxcar	High. Assumes independent rectilinear bounds and that all variables are known. Will cause over prediction with few variables and under prediction with many variables.	Increase area predicted introducing spurious predictions	
Distance-based	Medium. Estimates model non-parametrically but difficulties arise with data density and the definition of distance.	Algorithm will not choose the appropriate data points as being 'close' and biases will result.	Work needs to be performed to assess best distance measure. Cross validation?
Regression	High-low. Flexible techniques exists such as GAM and boosting. Simple model may suffer from considerable specification bias.	Spatial correlation in residuals.	Use flexible models unless clear theoretical reasons to ignore. Consider spatial patterns of errors to diagnose models problems. Truncate response range.

Table 5.2: Summary of types of model error and impact of missing covariates. Modification of table from Barry & Elith (2006).

5.3 Emperical Bioclimatic Models

As discussed in Section 5.1.1, Emperical Bioclimatic Models define the ecological niche primarily by the climate of the area being studied. These models use distribution records as surrogates for explicit organism performance parameters. Therefore, relatively modest data requirements allow this class of model to be applied to a wide variety of ecological estimates of potential distributions (Carpenter, Gillison & Winter 1993). However, the *fundamental niche* species distribution modeling techniques which fall into this category do vary significantly in their analysis methodology and data inputs. More specifically, the species data can be simple presence, presence-absence or abundance observations based on random or stratified field samplings or observations obtained opportunistically.

Modeling techniques that require absence or background data include the Generalised Linear Models (GLMs), Generalise Additive Models (GAMs) and Multivariate adaptive regression splines (MARS) (Hastie, Tibshirani & Friedman 2001, Guisan, Edwards & Hastie 2002) − these are "Regression models", a class of methods which, in a univariate setting, fit a curve through a set of points using some goodness-of-fit criterion (Barry & Elith 2006, Yee & Mitchell 1991). Other commonly used modeling techniques are Classification and Regression Tree analysis (CART) and Artificial Neural Networks (ANN) (Thuiller 2003).

Community models which fall under the empirical classification include generalised dissimilarity modeling (Ferrier, Drielsma, Manion & Watson 2002), neural networks (e.g. where Olden (2003) applied community predictive models that explicitly considered species membership, and thus each species' functional role, in the community) and multivariate adaptive regression splines (e.g. where Leathwich, Rowe, Richardson, Elith and Hastie (2005) used the multivariate adaptive regression splines (MARS) technique to describe non-linear relationships between species environment variables.

Modeling techniques that require only presence data include BIOCLIM (Nix

1986, Busby 1991): a Climate Envelope Model that uses species presence records to create a hyper-space which summarize how these records are distributed with respect to environmental variables, DOMAIN (Carpenter et al. 1993): a distance-based technique that estimates the environmental similarity using the Gower distance metric, between a site of interest and the nearest presence record in environmental space and LIVES (Carpenter et al. 1993): which uses a limiting factor method that postulates that the occurrence of a species is determined only by the environmental factor that most limits is distribution. This class of predictive models is used extensively for a range of practical and scientific purposes as it has the ability to cope with presence only data, a common limitation of biological datasets.

5.3.1 Climate Envelope Models

As described by Farber (2003), climate envelope models "involve two conceptual steps. The first step is the projection of the recording sites from the map into a multidimensional space defined by a set of climatic variables. The purpose of this step is to identify the climatic niche (also termed 'climatic envelope' or 'climatic profile') of the target species. The second step is the projection of the climatic niche from the multidimensional climatic space back into a two dimensional geographic space (i.e. a map)." The second step mentioned is termed "homoclime matching .... since a grid of the study area is scanned for locations with similar conditions to those of the species climatic profile" (Farber & Kadmon 2003). As described in Faber (2003), "the basic procedure applied for constructing these rectilinear models consisted of five steps..."

1. assembling the presence observations of the relevant species,

2. determining the climatic characteristics of each observation

3. removing outliers by choosing a percentile range

4. constructing a rectilinear climatic envelope based on the distribution of the remaining observations within the climatic space, and

5. projecting the climatic envelope back to the geographic space.

With the major limitations of climate envelope models summarised as:

1. The rectilinear nature of the climatic envelope, which bounds the climatic niche of the species within the multidimensional space by straight lines/surfaces. This enveloping approach may overestimate the distribution boundaries of the modeled species if climatic variables are correlated (Skidmore, Gauld & Walker 1996).

2. The fact that all climatic combinations within the boundaries of the climatic envelope are considered equally suitable for the modeled species (Shao & Haplin 1995).

3. Sensitivity to outliers, which originates from the boundaries of the climatic envelope being defined by the outermost observations. To reduce the impact of outlying observations on model predictions, users of climate envelope models often chop the outermost values of each climatic variable by using only a certain percentile range of the data (Busby 1991, Kershaw 1997) (see Figure 5.6). By reducing the performance of models by reducing the probability of making false predictions, this procedure may improve. However, it may also cause deterioration in predictive accuracy by increasing the rate of incorrect predictions of absences (Farber & Kadmon 2003).

BIOCLIM

The BIOCLIM "boxcar" ecological niche bioclimatic model is a popular and relatively statistically simple empirical climate envelope model, that is often used to predict the distribution of a species in a chosen area (Nix 1986, Beaumont, Hughes & Poulsen 2005). It uses the presence records and environmental data

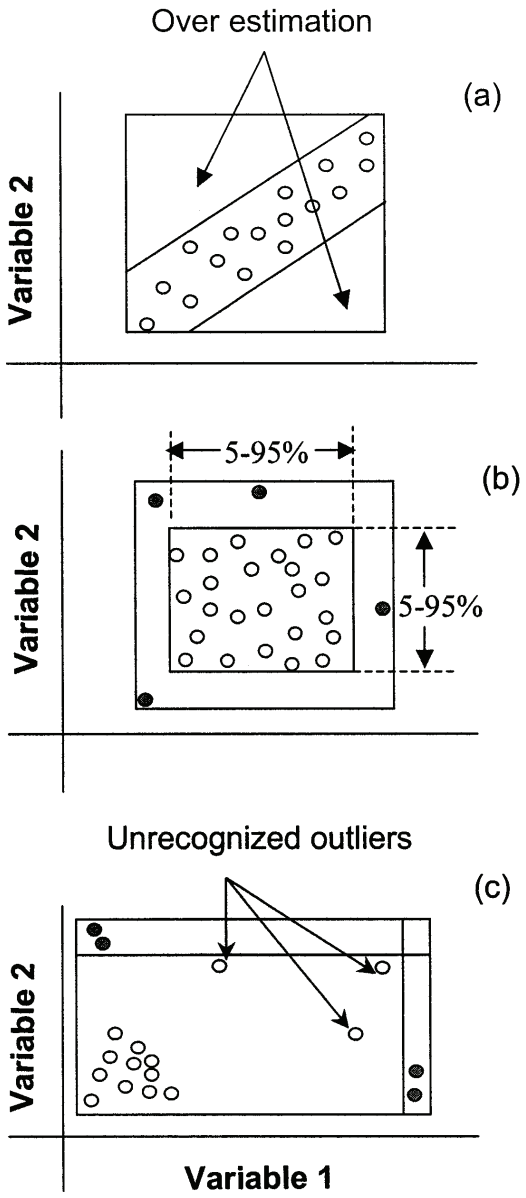


Figure 5.6: Schematic illustrations of some limitations of rectilinear models. Circles represent the distribution of observations in a climatic space defined by two hypothetical variables. Gray circles are observations recognized as outliers. (a) In cases of correlations or interactions between climatic variables, the rectilinear model tends to overestimate the domain of climatic combinations represented by the data. (b) The boundaries of the climatic envelope are determined by the outermost data, and are therefore, sensitive to outliers. Using a certain percentile range (e.g. 5-95%) can reduce the impact of such outlying observations on model predictions. (c) Removal of outliers using the percentiles-range method may prove inadequate if observations are outliers in a multidimensional sense without being outliers (marginal) in any single dimension (Farber & Kadmon 2003).

of temperature max, temperature min and precipitation to form a profile for a species that summarizes how the known presences are distributed with respect to the environmental variables. The envelope defined specifies the model in terms of upper and lower tolerances, and does not allow for regions of absence (i.e. "holes") within the envelope. Also, as the distribution is solely predicted on the basis of bioclimate (and providing a first approximation to limits to distribution) the influence of other environmental factors and interactions are excluded (Nix 1986). A BIOCLIM model has two steps:

(1). The first is the calculation of the bioclimatic variables for each raster cell in the area being studied. They are calculated from the temperature and precipitation records of each month to give, for example, the mean temperature of the grid points or the maximum rainfall in the coldest month of the year. These, often referred to as the BIO layers, are the bioclimatic environmental variables that form the multi-dimensional space (a hyper-rectangle or *environmental envelope*) that defines the environmental domain of the species.

The number of bioclimate layers used in the model, and their importance in a particular study, varies. For example, Beaumont et. al. (2005) discusses how BIOCLIM can summarise up to 35 bioclimatic variables, but that they may not all be needed (or relevant) for the study of a particular specie's distribution. This is considered important as it may lead to over-fitting of the model, which in turn may result in misrepresentations of a species' potential range. The implementation of the BIOCLIM model used in this study uses a maximum of 19 layers:

1 = Annual Mean Temperature

2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))

3 = Isothermality ((bioclimate 2/bioclimate 7) * 100)

4 = Temperature Seasonality (standard deviation * 100)

5 = Max Temperature of Warmest Month

6 = Min Temperature of Coldest Month

7 = Temperature Annual Range (bioclimate 5 - bioclimate 6)

8 = Mean Temperature of Wettest Quarter

9 = Mean Temperature of Driest Quarter

10 = Mean Temperature of Warmest Quarter

11 = Mean Temperature of Coldest Quarter

12 = Annual Precipitation

13 = Precipitation of Wettest Month

14 = Precipitation of Driest Month

15 = Precipitation Seasonality (Coefficient of Variation)

16 = Precipitation of Wettest Quarter

17 = Precipitation of Driest Quarter

18 = Precipitation of Warmest Quarter

19 = Precipitation of Coldest Quarter

These layers represent annual trends, seasonality and extreme or limiting environmental variables. A quarter of the year (a three month period) is defined as a "quarter." Most of the algorithms that calculate these rasters are not continuous functions and therefore not differentiable. Therefore, the sensitivity of the BIOCLIM model to uncertainty in the temperature and precipitation input layers can only be investigated using the Monte Carlo simulation method. The algorithms which calculate the Bioclimate Grids are described in Appendix B.

(2). The second step is the statistical component of the model that calculates the ecological niche. It calculates this habitat map by ranking each location according to its position in the species environmental profile. To calculate this profile, the model treats the bioclimate raster data values at the locations where the species is known to occur, as multiple one-tailed percentile distributions; that is, it creates a percentile distribution for each bioclimate layer (see Figure 5.7). As the percentile distributions are one tailed, the 5^{th} percentile is treated the same as the 95^{th} percentile in the calculation of the prediction.

From the information in these distributions, a prediction value can be assigned across the area being studied. For each raster cell in this area, the values of each bioclimate layer variable are assessed to determine their position in the

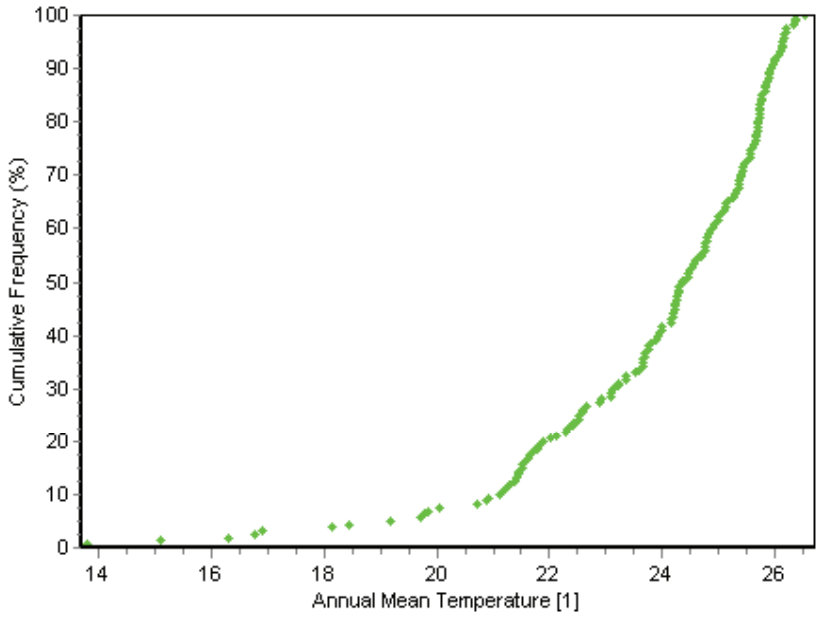


Figure 5.7: The frequency distribution for the Annual Mean Temperature at grid locations where a species is known to survive. Plot taken from (DIVA-GIS 2005).

percentile distribution. The lowest score across environmental values for a grid cell is mapped and can be "null" (outside the observed range of values) or range from 0 (low) to the theoretical maximum of 50 (very high). Therefore, the percentile assigned to the grid cell is the minimum percentile that is matched in all bioclimatic parameters. If one or more of the bioclimate layers at a grid location has no agreement within the chosen span, then the prediction assigned (to that grid location) will be null. Alternatively, if the minimum percentile is 5 for one bioclimate layer, but the others are higher, the grid location will be set to the 5^{th} percentile. Finally, the maps produced by this process are grid-based and classify each cell into one of several ranked classes of environmental suitability for the species (Barry & Elith 2006).

A graphical representation of the BIOCLIM model, climate prediction of the known *current* climate, is shown in Figure 5.8. Its components are (a) the current climate rasters, (b) the bioclimate layer generation algorithms (c), the known species location file (d) the component of the model which determines the ecological niches and (e) the species presence prediction grid map.

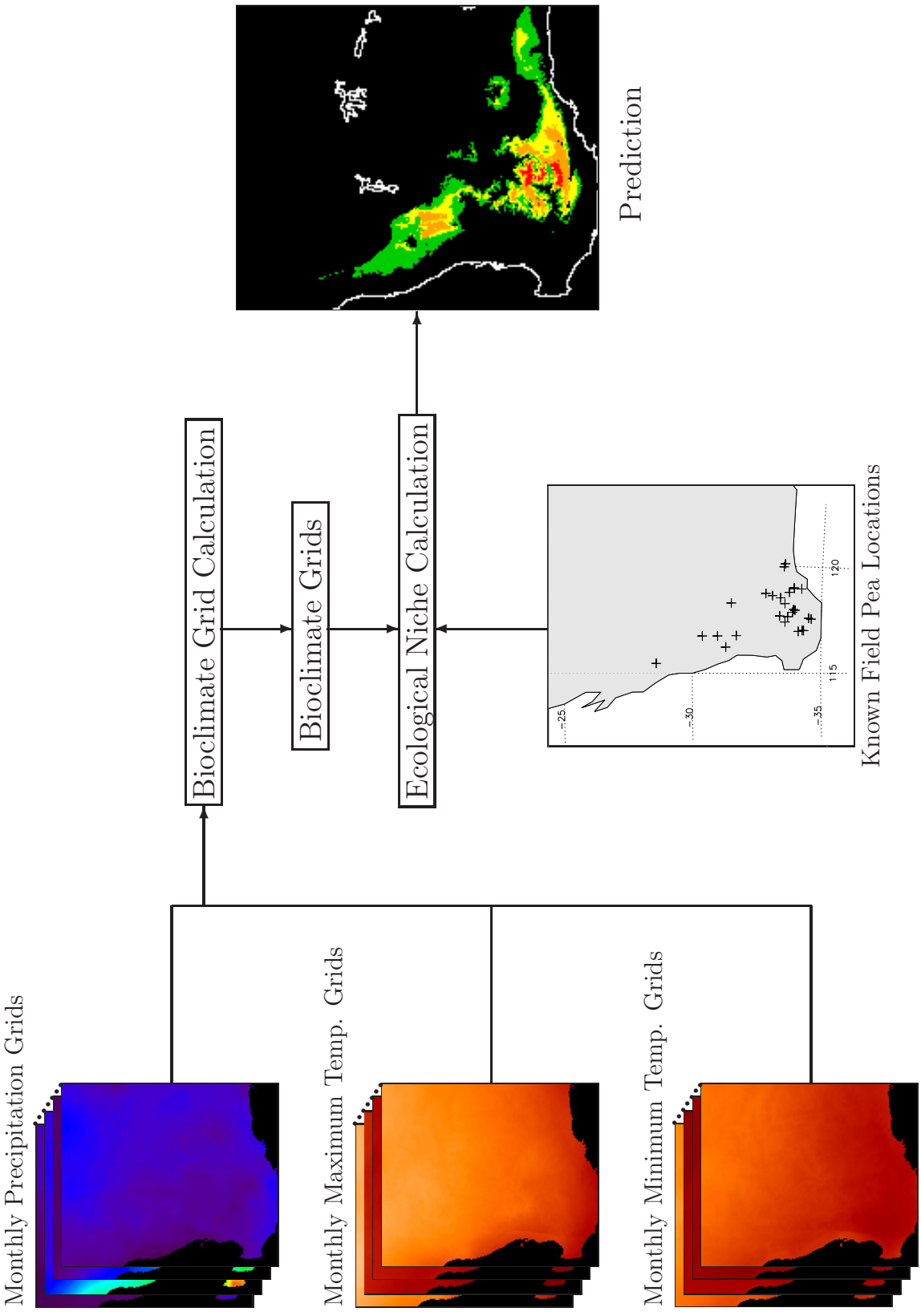


Figure 5.8: Graphical representation of the BIOCLIM present prediction model.

For a study of BIOCLIM's sensitivity to input uncertainty to be representative of the uncertainties in the input data rasters, several factors should be known. These include the source of the raw data (such as a temperature measurement stations) and the methods used in preparation of raster inputs (such as another model) from that raw data. Also, the uncertainty and limitations in either the raw data or the models used is usually documented in the literature. Knowing these, an estimation of how representative an input raster is of a real environment, can be calculated.

5.4 Summary

To summarise, the ecological niche models relate field observations to environmental predictor variables to calculate the species distribution model of that species. This modeling is also frequently referred to as "Bioclimatic Modeling." There are several types of models which fall into this classification, each with their own strengths that best suite the system to be studied.

There is extensive literature on the use, limitations and continuing development of these models. Also, the influence of uncertainty and error in the model's inputs, on the accuracy of its prediction is a growing area of research.

The BIOCLIM "boxcar" ecological niche bioclimatic model, is a popular and relatively statistically simple empirical climate envelope model, that is often used to predict the distribution of a species in a chosen area. Therefore, there is ongoing research into improving the accuracy of its predictions as well as understanding its limitations.

To accurately quantify this accuracy requires detailed historical knowledge of the input raw data and how this data has been processed to produce the input grids of the model. Of particular importance are the models and/or spatial statistical methods used to produce the climate input grids, their accuracy and limitations.

Chapter 6

Input Climate Layers

As discussed in Section 5.3.1, the BIOCLIM model inputs are (a) bioclimate grid layers and (b) a species presence point layer. The bioclimate grids are calculated from twelve temperature maximum, twelve temperature minimum and twelve precipitation grids (one for each month in each case, thirty six total). These grids were calculated from the data collected at meteorological stations that have passed the necessary quality control criteria (this climate grid generation step is the equivalent of process step 1 discussed in Section 2.4 and illustrated in Figure 2.2). This chapter discusses:

1. the accuracy of the climate point data,
2. the interpolation method(s) used to generate these monthly climate grids (from the station data) and
3. briefly discusses the uncertainties in future predictions.

The monthly climate grids used are the "Worldclim - Gobal Climate Data" grids (*Worldclim - Global Climate Data* 2009) and represent (a) *current conditions* (interpolations of observed data, representative of 1950-2000) and (b) two future conditions (downscaled from the Global Climate Model (GCM) output, IPPC 3rd Assessment Report (2001)). Present and future grids were downloaded from http://www.worldclim.org/current and

http://www.worldclim.org/futdown respectively. The data grids used have a spatial resolution of 2.5 and 5.0 arc minutes.

The current conditions WorldClim grids were interpolated by Hijmans, Cameron, Parra, Jones and Jarvis (2005), primarily from the Global Historical Climatology Network (GHCN) climate dataset. The monthly future climate projection grids were calculated from the future climates modeled by the CCCMA, CSIRO and HADCM3 models (Flato et al. (2000); Gordon et al. (2000)) and the emission scenarios reported in the Special Report on Emissions Scenarios (SRES) by the Intergovernmental Panel on Climate Change, IPCC (http://www.grida.no/climate/ipcc/emission/). From these, the CSIRO model A2a and B2a 2050 grids were used. Each future scenario described one possible demographic, politico-economic, social and technological future as expected for the year 2050. As discussed by Rödder (2009), "scenario B2a emphasizes more environmentally conscious, more regionalized solutions to economic, social and environmental sustainability. Compared to B2a, scenario A2a also emphasizes regionalized solutions to economic and social development, but it is less environmentally conscious." Therefore, of these two scenarios, the A2a grids represent the warmer climate conditions as expected with higher emissions.

The research methods applied to minimise the uncertainty in the data grids is discussed in the following Sections: Section (6.1) summarises the history of the GHCN (meteorological station) point data and the statistical quality control methods used to select which meteorological stations were of sufficient quality for use. Section 6.2 discusses the interpolation methods used by Hijmans et al. (Hijmans, Cameron, Parra, Jones & Jarvis 2005) to generate the WorldClim current grids from the point data and the uncertainty in the generated grids. Uncertainty in the future climate predictions is discussed in Section 6.3.

The other input layer shows where the species of interest are known to exist. The spatial uncertainty in this input is insignificant when compared to the climate raster grid cell size and has, for this reason, has been excluded in this study.

6.1 The Global Historical Climatology Network

The Global Historical Climatology Network (GHCN Monthly) "contains historical temperature, precipitation, and pressure data for thousands of land stations worldwide" (NOAA 2010*a*). Its aims are a continuation of many efforts to produce long−term monthly global climate databases. The earliest temperature record is greater than 100 years, with the average being ≈60 years old. Examples of where data has been sourced include the World Weather Record which started in 1923, the National Centre for Atmospheric Research's annually published World Monthly Surface Station Climatology dataset and the Jones dataset (Jones 1994, Jones 1995). The first release of the GHCN was in 1992 (Vose, Schmoyer, Steurer, Peterson, Heim, Karl & Eischeid 1992) and contained quality−controlled monthly climatic time series from 6039 land based temperature stations world wide. This information has been used extensively in basic form and derived products such as gridded temperature anomalies (Peterson & Vose 1997). It is a an important and popular dataset in climate change research (e.g. Brown et al. (1993); Groisman Karl and Night (1994)).

This database was later enhanced to the GHCN Version 2, which is used in this thesis. Improvements in the Version 2 dataset (a) the increased number of data points included, (b) the new suite of quality control checks and (c) the homogeneity testing and adjusting techniques (Peterson & Easterling 1994, Easterling & Peterson 1995, Easterling, Peterson & Karl 1996) applied to the data. As discussed in Peterson and Easterling (1994) and Easterling and Peterson (1995), for some stations "adjusted" data has been through several homogeneity control procedures (Hijmans, Cameron, Parra, Jones & Jarvis 2005). The geographical locations of all the stations used is shown in Figure 6.1

6.1.1 Dataset Extent and Quality

A principal aims of the GHCN project is the continuing improvement of the GHCN dataset quality and size (number of stations, temporaral length of record).

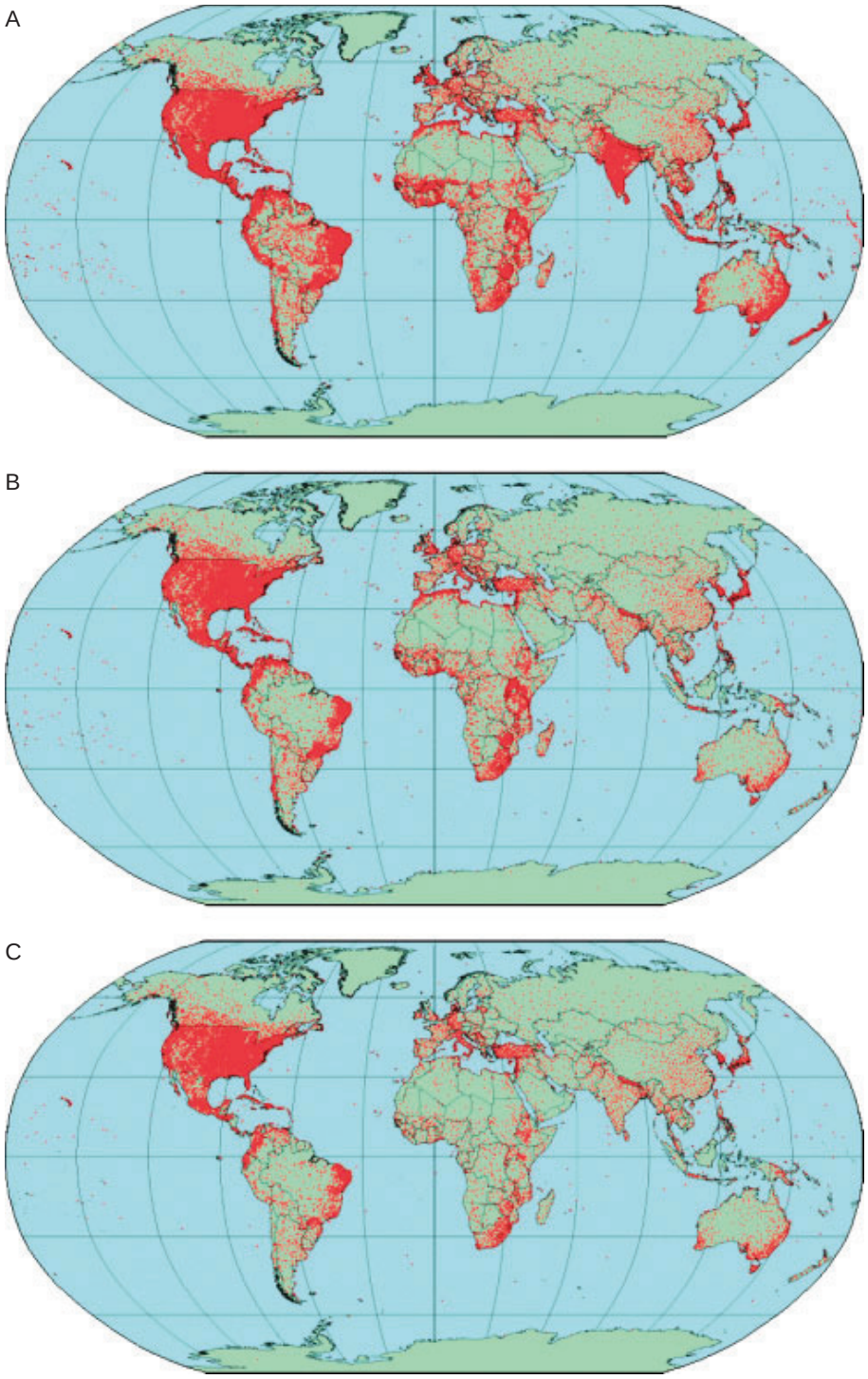


Figure 6.1: Locations of weather stations from which data was used in the interpolations. (A) precipitation (47 554 stations); (B) mean temperature (24 542 stations); (C) maximum or minimum temperature (14 930 stations). Diagram from Hijmans et al. (2005)

Resulting from this work, the main enhancements in the second version of the GHCN dataset include:

1. Data from additional stations to improve regional scale analysis.
2. The addition of maximum−minimum temperature data to provide important climate information not available in mean temperature alone.
3. Detailed assessments of data quality to increase the confidence in research results.
4. Rigorous and objective homogeneity adjustments (in both space and time) to decrease the effect of non climatic factors on the time series.
5. Detailed meta data such as topography and population.
6. An infrastructure for updating the archive at regular intervals.

The theory and methods applied in the preparation of the GHCN dataset is discussed in Peterson and Vose (1997). The methodology used to address the limitations in the data quality is discussed in Peterson and Griffiths (1997) and Peterson et al. (1998). The majority of these and other publications discuss improving the quality of the temperature record, so the majority of the following sections will discuss the methods used in testing the quality of the temperature record (Section 6.1.2). Section 6.1.3 discusses the methods applied to increase the number and quality of high quality precipitation stations in the GHCN dataset Durre, Matthew, Menne, Gleason, Houston and Vose (2010). Durre et al. (2010) discusses the most recent development of the statistical analysis methods and software development to automate this analysis.

6.1.2 Temperature Record Improvement

As cited in the GHCN Monthly Version 2 temperature records website (NOAA 2010*c*), the methodology used in the quality control of the GHCN temperature

data is discussed in Peterson and Vose (1997) and Peterson, Vose, Schmoyer and Razuvaev (1998). It has three steps:

Examination for clearly visible low quality data

Quality compromised by, for example, (a) homogeneity-adjusted data which could not be quality tested by comparison with the original data, (b) monthly data that were derived from incomplete synoptic reports (which caused unacceptable biases and errors) and (c) significant processing errors.

Time series analysis

The second stage examined individual station time series for anomalies such as stations that were temporally out of phase. Methods used include climatological evaluation (which compared the stations' data to the climatology of that location (Legates & Wilmott 1990, Peterson & Vose 1997)) and the cumulative sum test (to look for changes in the mean as an indicator of gross discontinuities). As described in Peterson (1998) and demonstrated by Rhoades and Salinger (1993), this method is used in determining the homogeneity of a station.

Individual data points

The final stage, using statistical methods, determined if the data points were outliers in both time and space (*temporal spatial*). As described in Lazante (1996), this method flags all data points that have greater than 2.5 standard deviations (σ). These were then compared to neighbouring stations to determine if the extreme value was an extreme climate event in the region. It was found that over 85% of the points were valid. Those that were invalid were removed from the main GHCN data file.

This method gives less weight to an observation the further it is from the centre of the distribution (Lanzante 1996), with values over 5 σ from the median being flagged with a 0 weight. It is a robust method because it uses almost all the data and the weighting factor makes it resistant to outliers. However, as with other methods, the more data points one has the more reliable the assessment of data. Therefore, the variance of each month was determined by also using the

data from the months before and after the month (of interest). As mentioned in Peterson et al. (1998), this three month approach produced excellent σ for the quality control of the temperature data. This statistical method, for quality control purposes, can have limits due to its sensitivity to outliers. In particular, a large outlier can inflate the σ such that smaller outliers are not identified (Peterson, Vose, Schmoyer & Razuvaev 1998). To reduce this problem, the data which was three to five times the σ was flagged and removed from a re-calculation of the σ (Boyer & Levitus 1994).

Given these method's sensitivity to outliers, defining an accurate outlier, or more specifically its threshold, is very important. The σ threshold is often set at 3 to 5, but a common practice is to also set a 3σ limit (e.g Guttman and Quayle (1990)), as it has been observed that the fraction of the data that are bad is highest at the most extreme outliers and decreases as data gets closer to the mean. Therefore, the worst data should be able to be found where the datum is bad. As described in Peterson et al. (1998), "it was decided that that positive outliers greater than 1.5σ could be considered bad only if all of the nearest five neighboring stations had negative anomalies for that month and *vice versa* for cold outliers." This work is based on the assumption that five regional neighbours represent the same regional climate as the studied station, and that the stations where this is not true should produce a background error rate that is independent of the magnitude of the outlier. The work concluded that any data that starts at 2.5σ's may be classified as suspect (see Lanzante (1996)), but "great care must be taken not to throw away good data that happens to be extreme" as these may "represent very important aspects of the climate." (Peterson et al. 1998)

Another method applied is the interquartile range (I.Q.R.), which measured the difference between the first quartile and fourth quartile. The results show that results, when using the I.Q.R. and σ to study large numbers of normally distributed data points, were comparable. However, the I.Q.R. is resistant to outliers, which is why it has been used in the quality control of climate data

(Eischeid, Baker, Karl & Diaz 1995), but it is not very robust when the time series are fairly short (Peterson et al. 1998).

6.1.3 Precipitation Record Improvement

As discussed in on the GHCN Monthly Version 2 precipitation records website (NOAA 2010*b*) the methods used to assemble the GHCN-Monthly precipitation dataset are "generally comparable to those used in developing the temperature dataset." The main objectives when applying these methods to the precipitation data is to (a) eliminate duplicates records using a range of methods and (b) quality control using the methods:

1. Comparing stations with a gridded climatology and plotting the stations for visual inspection.

2. The cummulative sum test (which looks for changes in the mean).

3. An analogous test that looks for changes in the variance or scale.

4. Evaluated for runs of three or more months of the same nonzero value.

5. Precipitation total was evaluated to determine if it was an outlier in space and/or time using a variety of nonparametric statistics.

6.2 Generation of the Current Conditions Grids

The WorldClim current climate data grids were interpolated from temperature maximum, minimum and precipitation meteorological inputs, covering the period 1950−2000. The thin-plate spline algorithm was the preferred technique (Hijmans, Cameron, Parra, Jones & Jarvis 2005) and the main data used was the GHCN dataset. The monthly GHCN dataset was preferred as it had undergone the most explicit quality control, but other data from other stations were included if they complied with certain criteria (see Hijmans et al. (2005) for

more details). Uncertainty layers where also calculated, as illustrated in Figure 6.2.

There are a number of different statistical interpolation methods which can be used to interpolate a surface. For example, three widely used methods include inverse distance weighting, kriging and splines (Davis 2002). Of these, thin plate smoothing splines was used in the generation of the WorldClim surfaces (Hijmans, Cameron, Parra, Jones & Jarvis 2005) as this procedure has been used in similar global studies (New et al. (1999, 2002)) and performed well in comparative tests of multiple interpolation techniques (Hartkamp, Beurs, Stein & White 1999). The procedure is further described in Hutchinson (1995). Splines share close connections with other geostatistical interpolation techniques, so there has been comparisons between these methods (e.g. Hutchinson and Gessler (1994) and Laslett (1994)) to evaluate the best method to use. As described in Hutchinson (1995), the main advantage that thin plate splines has (over competing geostatistical techniques) is that "splines do not require prior estimation of spatial auto−covariance structure," a structure which may be difficult to validate.

6.2.1 Uncertainties in the Worldclim Grids

As discussed by Hijmans (2005), the uncertainty and error in the datasets used to generate the Worldclim layers was minimised. Therefore, the uncertainty that is not due to natural variation across short time scales (months), in the GHCN data, is considered minimal. However, as discussed in Section 3.3 of Hijamns (2005), the uncertainty in the interpolated Worlclim current climate layers was mapped in Section 3.3 of Hijamns et al. (2005). In these course resolution maps, in all the grids in the western Australian study region, the uncertainty in the precipitation and temperature are the same, 2.5mm and 1.5°C. respectively. These uncertainties are included in this study.

The other significant potential source of uncertainty in the Worldclim Layers, is their representation of the natural variation in a regions climate. As this has

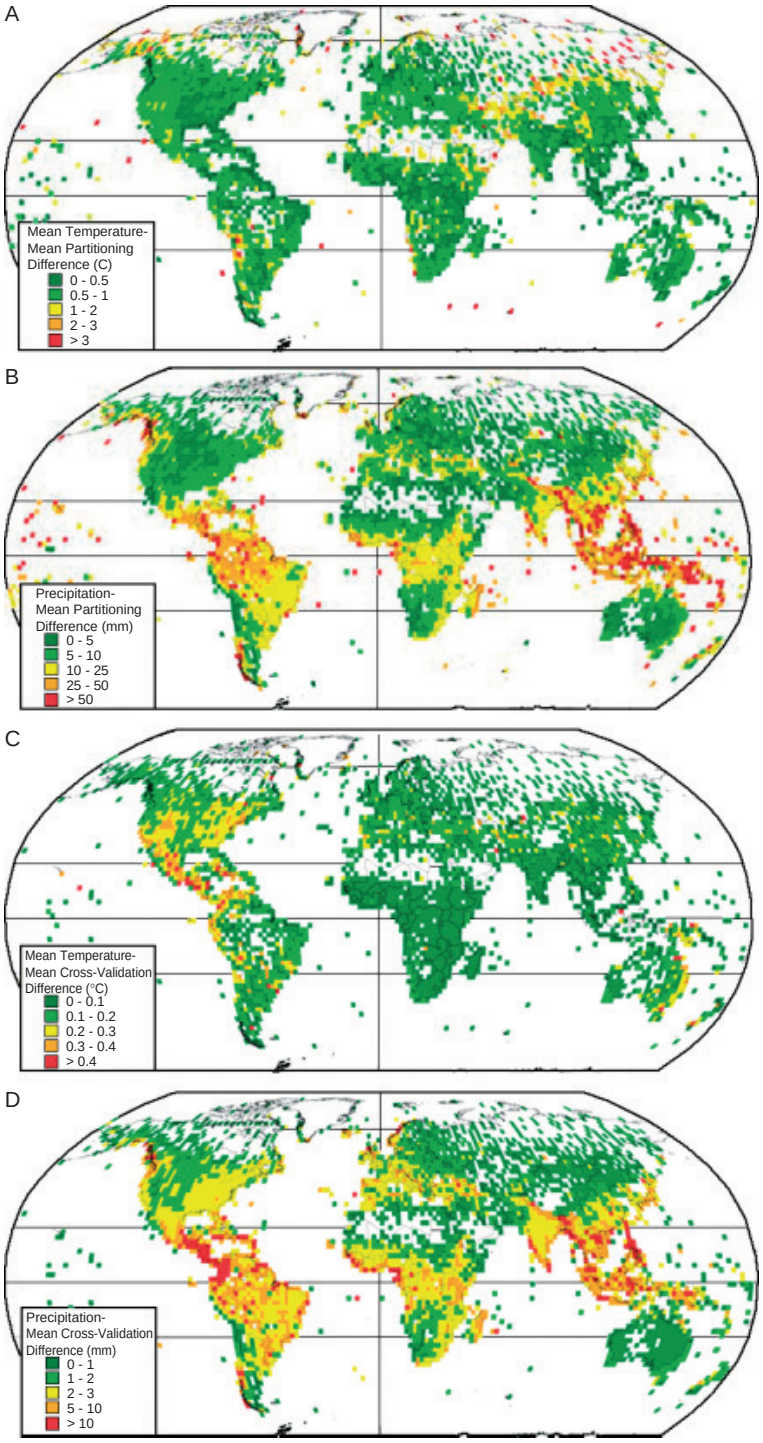


Figure 6.2: Uncertainty in the climate surfaces. Mean cross-validation deviations for precipitation (A) and mean temperature (C) and deviations when partitioning data in test and training sets for precipitation (B) and mean temperature (D). Values are averaged across 12 months and by 2-degree grid cell. Diagram from Hijmans et al. (2005)

not been quantified by other research in the areas studied in this thesis, it will done by examining how the decadel means (1950-59, 1960-69,..., 1990-99) vary from the 50 year mean, at each relevant station. The largest absolute differences calculated will then be used to interpolate uncertainty climate surfaces.

How the uncertainty-related data required for this analysis, of these two uncertainty sources, was determined and the method used to study their influence (on the BIOCLIM prediction) is discussed in Chapter 7. The results are discussed in Chapters 8 and 9.

6.3 Uncertainties in the Future Worldclim Layers

The uncertainty of the future climate grids is determined by the accuracy of the future climate model. As discused in the IPCC Third Assesment Report (2001), Section 9.2.2.4 "Projections of climate change are affected by a range of uncertainties and there is a need to discuss and to quantify uncertainty in so far as is possible. Uncertainty in projected climate change arises from three main sources; uncertainty in forcing scenarios, uncertainty in modeled responses to given forcing scenarios, and uncertainty due to missing or misrepresented physical processes in models." The complexity of these uncertainties makes generating a single grid that represents all uncertainties difficult. For example, as discussed "the ensemble standard deviation and the range are used as available indications of uncertainty in model results for a given forcing, although they are by no means a complete characterisation of the uncertainty."

While these uncertainties will influence the BIOCLIM prediction, the degree of this influence has not been investigated due to its complexity. Also, if a BIOCLIM user is interested in the difference between future predictions, the uncertainties in the future climate grids will not be relevant (as the difference between modeled scenarios is significant (Corner & Marinelli 2008)). What may be more important is what influence the uncertainties in the present climate

grids have on the future predictions. For example, will the difference between the two future climate (A2a and B2a, 2050) BIOCLIM predictions be significantly changed with the presence of uncertainties to the present climate grids? This is investigated using the Monte Carlo method (see Section 7.4 for a graphical representation of this model) and the results discussed in Section 8.3.2.

6.4 Summary

The Wordlcim data layers and their data sources, have been rigourously studied and uncertainty and error minimised using peer reviewed statistical methods. The uncertainty in the Worldclim layers has been quantified and its influence on the accuracy of the BIOCLIM model investigated. Also, the decadel variation in the Global Historical Climate Network datasets, at the relevant stations, will also be included as a source of uncertainty.

Chapter 7

Analysis Methods

The BIOCLIM ecological niche model studied is the version in the DIVA-GIS Geographical Information System, version 5.2 (*DIVA-GIS* 2005, Hijmans, Guarino, Jarvis, OBrien, Mathur, Bussink, Cruz, Barrantes & Rojas 2005). The ecological niche modeling option calculates and saves the bioclimate raster grids, inputs them into the BIOCLIM model and produces one prediction grid of the ecological niche. The Disk Operating System (dos) version of this software is AVID-GIS (version 4.3). Using the dos version requires that the bioclimate grids are initially generated and saved using DIVA-GIS or some other purpose written code. These are then read by AVID when its distrubution model option is executed with the following command:

DISTMODEL model in_stack out_stack points output

where:

model The ecological niche model to execute (BIOCLIM or DOMAIN).

in_stack A .grs file which contains the names of the present climate bioclimatic grids to include in the model.

out_stack A .grs file containing the names of other bioclimatic grid files. For present predictions, the in_stack and out_stack .grs files are the same. For future predictions, these are future bioclimate grids (which were generated using the A2a or B2a climate grids).

point A .csv file which contains the location (latitude/longitude) of *where the species/crop is known to grow/be present.*

output The BIOCLIM prediction output raster grid (image.grd) and its associated header file (image.gri).

In DIVA-GIS, the bioclimate generating algorithm was ported from the Arc Macro Language (AML) format bioclimate code, as documented at the DIVA-GIS website and authored by Robert Hijmans in 2004 (seen appendix B). For this thesis, it was ported into an IDL procedure which will be referred to as *IDL-bioclimate*. In both DIVA-GIS and IDL-bioclimate, the bioclimate products were saved in the DIVA-GIS default format, as this is the only AVID compatible format (for its bioclimate inputs).

DIVA-GIS can not generate predictions for multiple climatic conditions in an automated multiple prediction process, so it can not be used to study uncertainty propagation using the Monte Carlo method. However, the Dos version does allow this, as it can be spawned from a mother program which is written to carry out the Monte Carlo method. A combined *IDL-AVID* program (which also contained the IDL-bioclimate code) were written for this.

The initial IDL-AVID program written replicated what the DIVA-GIS BIOCLIM tool does, that is, make a prediction under one climate and species location scenario. As discussed in Section 7.2, this allowed an IDL-AVID prediction to be checked against its DIVA-GIS equivalent as they must be the same. Then, the IDL-AVID program was further developed to allow the Monte Carlo method to be applied. The quality control results of this and the methods used to determine valid climatic uncertainties as needed for the Monte Carlo simulation, is discussed in Section 7.3.

Finally, the bioclimate layers and the DIVA-GIS and AVID-GIS predictions are by default rounded to one decimal place.

7.1 Data Formats

Two data type conversions occurred for, or in, the IDL-AVID program:

1. The Worldclim input climate rasters are in the Band Interleaved by Line (.bil) format, which can be read by DIVA-GIS but not by IDL-AVID (due to a lack of a suitable IDL input filter). This was resolved by converting the worldclim data to the GeoTIFF format using the ENVI data conversion tools (*ENVI 4.4* 2007). Geotiff is a "TIFF based interchange format for georeferenced raster imagery," (http://trac.osgeo.org/geotiff/).

2. The geospatial analysis of IDL-AVID predictions was performed using the ENVI software package. As ENVI can not read the AVID-GIS format files, the IDL-AVID prediction's statistical analysis grids were also saved in the ENVI format by default.

In both cases their products where compared with the original or default format (geotiff with .bil; and ENVI with DIVA-GIS) to check for any anomalies.

7.2 Quality Testing IDL-AVID

Testing the IDL-AVID code was performed by comparing its (a) bioclimate and (b) BIOCLIM prediction gridded products with the DIVA-GIS equivalents. Figure 7.1 shows a graphical representation of the quality control steps. In both cases, the contents of their grids and associated header files must be identical.

7.2.1 The Bioclimate Grids

Bioclimate raster and header files were generated with both DIVA-GIS and the IDL-bioclimate code. The header files were compared by manual inspection and the data files compared using mapping and analysis tools in AVID-GIS as well as purpose written IDL statistical analysis programs. As .gri data files are in

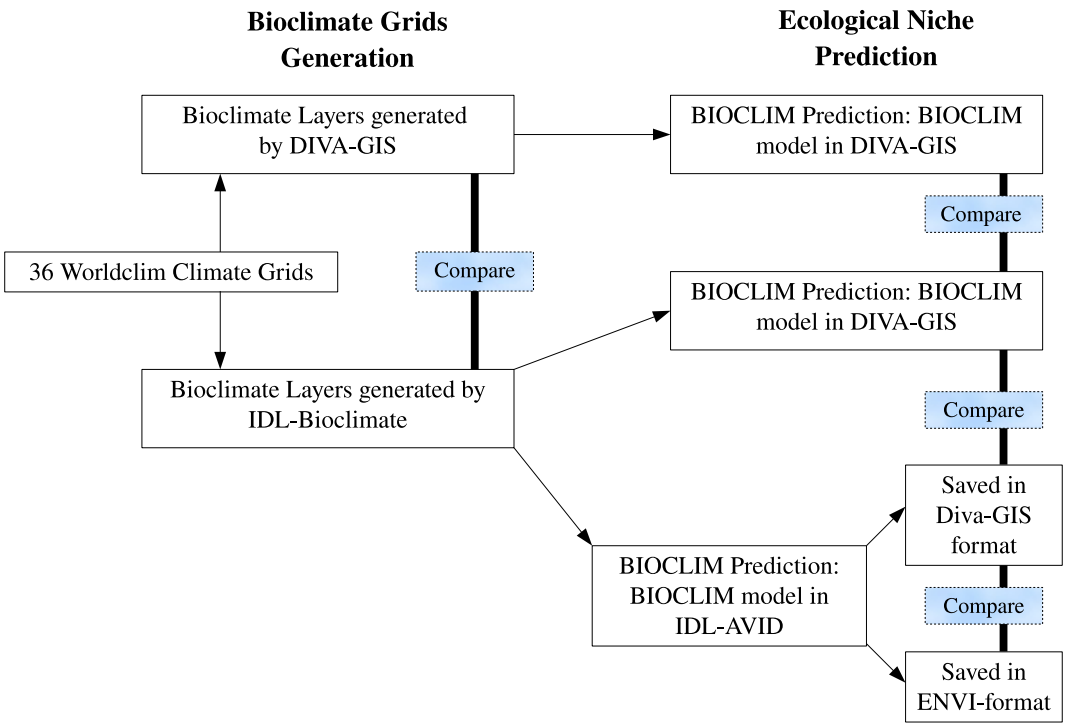


Figure 7.1: Steps in the comparison of IDL-Bioclimate and Avid-GIS Grids with DIVA-GIS Grids

binary format, to be read correctly (by AVID-GIS), the contents of their header files must be correct.

Appendix C contains an example of the bioclimate raster 1 (Annual Mean Temperature) header file. Note that in the AML code, as well as in the bioclimate rasters generated by DIVA-GIS, they are named as *BIO 1*, ..., *BIO 19*. The entries in the header files are separated into groups. Of these, the ones that can affect the prediction are the Georefference, Data and Application groups. The Data and Georefference groups are very important as their information will determine if the binary data is correctly read and then assigned to a correctly sized - and geographically postitioned - raster grid. Therefore, the entries in these three header groups, as calculated by the IDL-bioclimate script, were manually compared to their DIVA-GIS produced equivalents and any anomalies isolated and removed (by further IDL-bioclimate development). These tests were repeated at several different geographic areas to further test for anomalies. The other groups in the header files do not influence the BIOCLIM prediction as they are only

required by the DIVA-GIS visualisation tools (scale and colour settings).

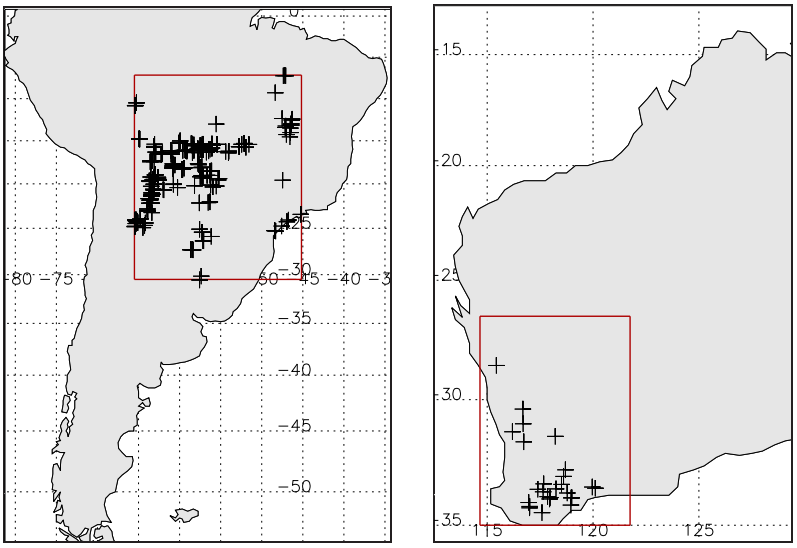
If the header file entries agreed, but the AVID-GIS and IDL-bioclimate generated data raster grids did not, this showed that there were other errors in IDL-bioclimate. Detailed code inspection showed that, in most cases, the anomalies in the data were the product of minor errors in the code's structure and/or rounding related problems. In particular, a floating point number that was incorrectly rounded to a single significant digit was a common problem found in the IDL-bioclimate code. The source of these and other errors were determined and removed. The final test compared the AVID-GIS BIOCLIM predictions made with DIVA-GIS and AVID-bioclimate generated bioclimate input grids, as the results must be the same.

7.2.2 BIOCLIM Prediction

The BIOCLIM ecological niche prediction, made with the Worldclim current climate layers, were generated using the DIVA-GIS and IDL-AVID models. The input climate layers were the same, except for the difference in format (.bil and geotiff). In both models, the output predictions were saved in the DIVA-GIS format and then compared using DIVA-GIS. These results were converted to and/or also saved in ENVI format, allowing statistical comparison using ENVI tools if required.

The first test of IDL-AVID replicated example 2 from the DIVA-GIS Exercises website, "Modeling the distribution of wild peanuts (*Arachis spp.*)" (2005). The result of this exercise is a prediction of where wild peanuts should grow in South America. As this is a documented example and the results are known, it was repeated for the comparison of AVID-GIS and IDL-AVID predictions.

This geographical area and the known location of wild peanuts in it, is shown in Figure 7.2(a). The default DIVA-GIS and IDL-AVIS predictions are mapped and in Figure 7.3(a and b) and compared in the plots shown. As described in (*DIVA-GIS* 2005), predictions equaling 0 are labeled ND in the images. For consistency, this labeling format is used in this thesis.



South America Western Australia

Figure 7.2: The areas studied in South America and Western Australia are within the red bordered region. The known locations of the Wild Peanuts and Field Peas in each respective study area are shown.

As expected the images are the same, the co-plots are linear and the correlation is 1.0. This was also the case when this test was repeated for the Western Australian region of interest (Figure 7.2(b) and 7.3(e-g)). The species of interest in the Western Australian region was the Field Pea agricultural crop. The locations of this crop used in this study were at the Western Australian Department of Agriculture and Food (http://www.agric.wa.gov.au/) field test sites as shown in Figure 7.2(b).

Futher testing included scenarios where the model had:

1. Future prediction out_stack entries.

2. Not all bioclimate layers were included.

In all cases tested, the DIVA-GIS and IDL-AVID predictions were the same. Therefore, it was concluded that the IDL-AVID program could be further developed for uncertainty-sensitivity analysis using the Monte Carlo method.

7.3 Monte Carlo Model

IDL-AVID was further developed to quantify BIOCLIM's sensitivity to uncertainty in the climate inputs, using the Monte Carlo method. This required that multiple simulations could be realised, with each having thirty six:

1. Random number grids (one for each climate grid); generated using the same methodology as described in the Precision Agriculture Methods Section (3.7.2).

2. Unique uncertainty grids; generated by multiplying constant uncertainty layers with the random number grids.

3. Unique climate grids; generated by adding the unique uncertainty grids to the present climate grids.

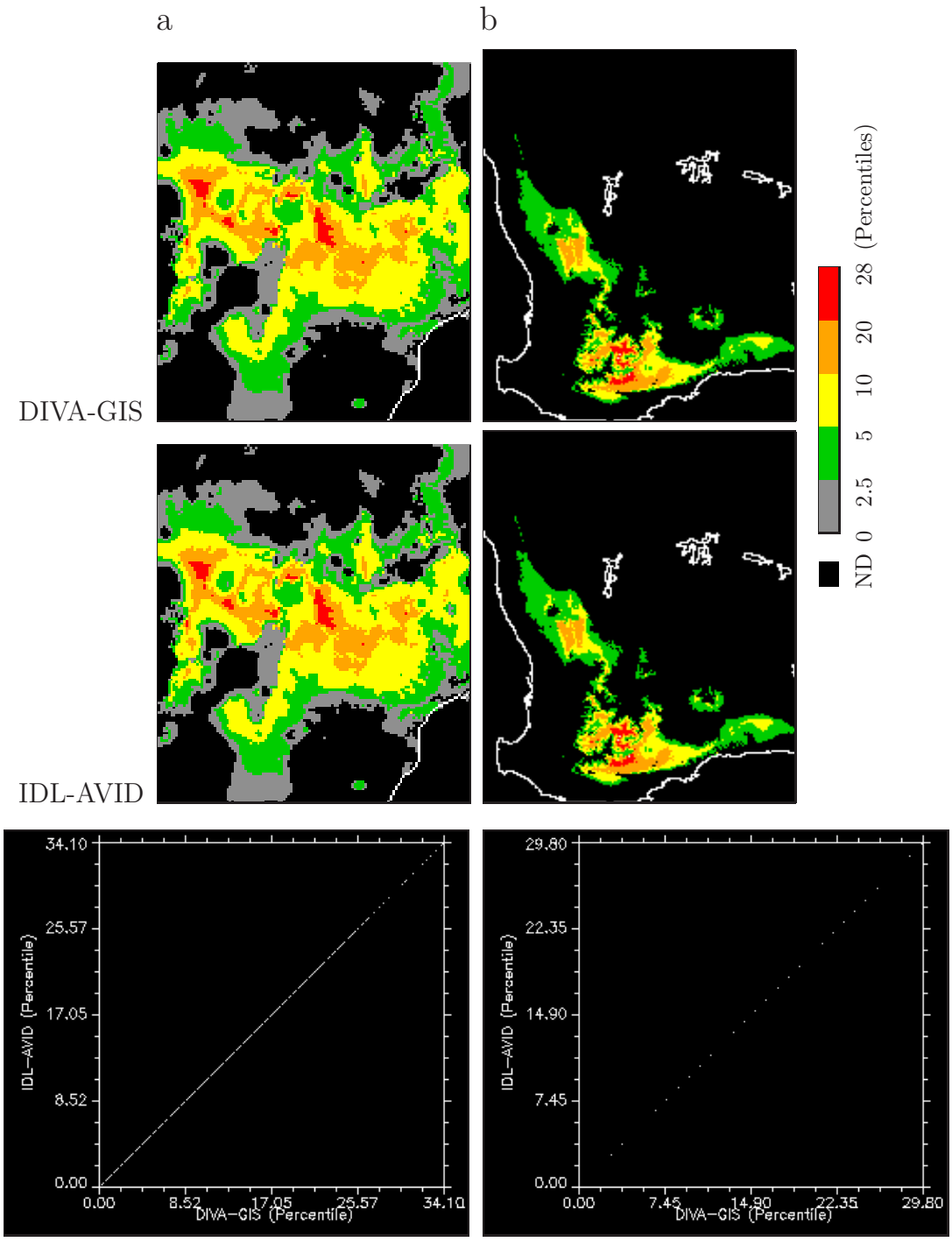


Figure 7.3: Prediction Grids and associated comparison plots, DIVA-GIS and IDL-AVID BIOCLIM Predictions: (a) South America and (b) Western Australia. As described in Section 7.2.2, Default or Mean predictions equaling 0 are labeled as ND.

Each simulation will produce a unique realisation of an ecological niche. For these predictions to sit within a range that is a propogated product of a realistic uncertainty, it is neccesary to accurately map the constant uncertainty grids. The methods used to determine the constant uncertainty grid was applied and tested in both the South American and Western Australian regions. However, as the area of interest in this thesis is in Western Australia, further discussion will be limited to the Western Australian results.

7.3.1 Constant Uncertainty Rasters

Uncertainty raster grids were generated for two sources of uncertainty in the Worldclim Raster layer. The first source of uncertainty is due to the limitations of the interpolation model used to generate the Worldclim layers. The second source is the accuracy in the Worldclim surfaces representation of the climate of the period 1950 to 2000. That is, the uncertainty due to the natural temporal variation in climate. As with all datasets, there are other uncertainties in the GHCN dataset. However, as discussed in Section 6.1.1, extensive work has been done in minimising the uncertainties in these datasets, so their influence on BIOCLIM predictions will be minimal and hence has not been included in this thesis.

(1). The uncertainty in the Worldclim Grids.

The uncertainty in the worldClim climate rasters has been investigated using data partitioning and cross validation (Hijmans, Cameron, Parra, Jones & Jarvis 2005) (see Figure 6.2). The data-partitioning method resulted in the highest deviations from the observed data due, in part, to it using only half the available data. The cross validation method is a *leaveoneoutmethod* which uses all of the available data and for this reason was used to estimate the error in the worldclim grids, despite the fact that it calculates lower uncertainties. A single value for each grid was used as the error for the geographic area of interest is the same. For Western Australia this was 1.5^{o}C (temperature minimum and maximum) and 2.5 mm (precipitation). Three raster grids containing these values were generated

and were called the "Worldclim" Uncertainty Grids.

(2). Uncertainty due to Temporal Variation.

One source of uncertainty in the WorldClim Grids is the natural variation of climate over time. This temporal uncertainty was estimated by analysis of the GHCN monthly mean point data (calculated from a maximum of 50 years of point measurements). In this case, "point" refers to measurement station. The GHCN location number and coordinates are in Table 7.1 and 7.2 and mapped in Figure 7.4. These stations were chosen as they occur across an area that exceeded in spatial extent, the area where the species are known to occur.

From these GHCN stations, all the points which had more than 42 years of data over all months (e.g 1950 to 1992, all months) were selected. This number of years was chosen so that the five decades from 1950 to 2000 were represented. Some stations did not have data post 1992, but these were considered adequate as they covered most of the 50 year period. For each station, the mean of all available data in the period 1950 to 1999 was calculated (maximum of 50 years, all months). The mean for each decade in this period was then calculated (e.g 1950-1959, 1960-69 etc, all months) and then subtracted from the overall 50 year mean to give five decadel difference values. The maximum of these values was then divided by the 50 year mean, converted to the absolute value and then assigned as the maximum decadel variation ratio (for that station, see Table 7.1 and 7.1). These maximum decadel values, for each station, were used to interpolate an uncertainty surface using the splining interpolation technique. This method was used as it was the method used in generating the Worldclim grids (Hijmans, Cameron, Parra, Jones & Jarvis 2005). Also, it is a preferred method when the number of data locations is low (such as the number of GHCN temperature stations in Western Australia) and is often used in climate related areas of research (Hutchinson 2004, Hutchinson & Gessler 1994, Hutchinson 1995). These surfaces were interpolated with the interpolation tools in ArcMap (*ARCMAP 9.2* 2004), using default settings. Therefore, three decadel uncertainty raster grids were generated, (1) temperature maximum, (2) temperature minimum and (3) pre-

cipitation; which will be referred to as the "GHCN" uncertainty grids (Figure 7.4.

As discussed in Section 2.6, the distribution of an interpolated surface is influenced by the number and spatial distribution of the interpolation input points. With surfaces interpolated from a low number of points, the edges of the uncertainty surfaces may be inaccurately skewed at their extremes (such as the north-west inland region). However, in these areas the uncertainty in the BIOCLIM prediction is 0, as the prediction is also always 0 due to there being no Field Pea sites in these climatic areas. Of greater significance to this study was the accuracy of the temperature uncertainty grid at, or between locations where Field Peas had been trialled but where there are no GHCN temperature stations, such as in the northern section of the Western Australian wheat belt. As can be seen in Figure 7.4(e), this area has negative values which are in part due to the high absolute decadel uncertainty at stations close to the coast or in the north east. These negative values are, as absolute values, not significantly different to the uncertainty which occurs at the same geographical location as the Field Pea trial sites. Therefore, it was still considered valid for this study.

7.4 Unique Uncertainty Layers

For a particular grid (max temp, min temp or precipitation) for a particular month, the unique uncertainty grid is either one of, or a summation of both, the corresponding GHCN and Worldclim unique uncertainty layers. As discussed in Section 7.3, these were generated by multiplying the constant uncertainty layers by a random number grid. As described in Section 3.7.2, the random numbers generated were either (a) normally (Gaussian) or (b) positively or negatively skewed. For the normal distribution, the random number values ranged from $+1$ to -1 and the IDL-AVID code written such that the resulting uncertainty surfaces represented $+/-3$ standard deviations. As in Section 3.7.2, the skewed distribution, random number setting is $+/-2$ gamma (see Figure 3.3 for this

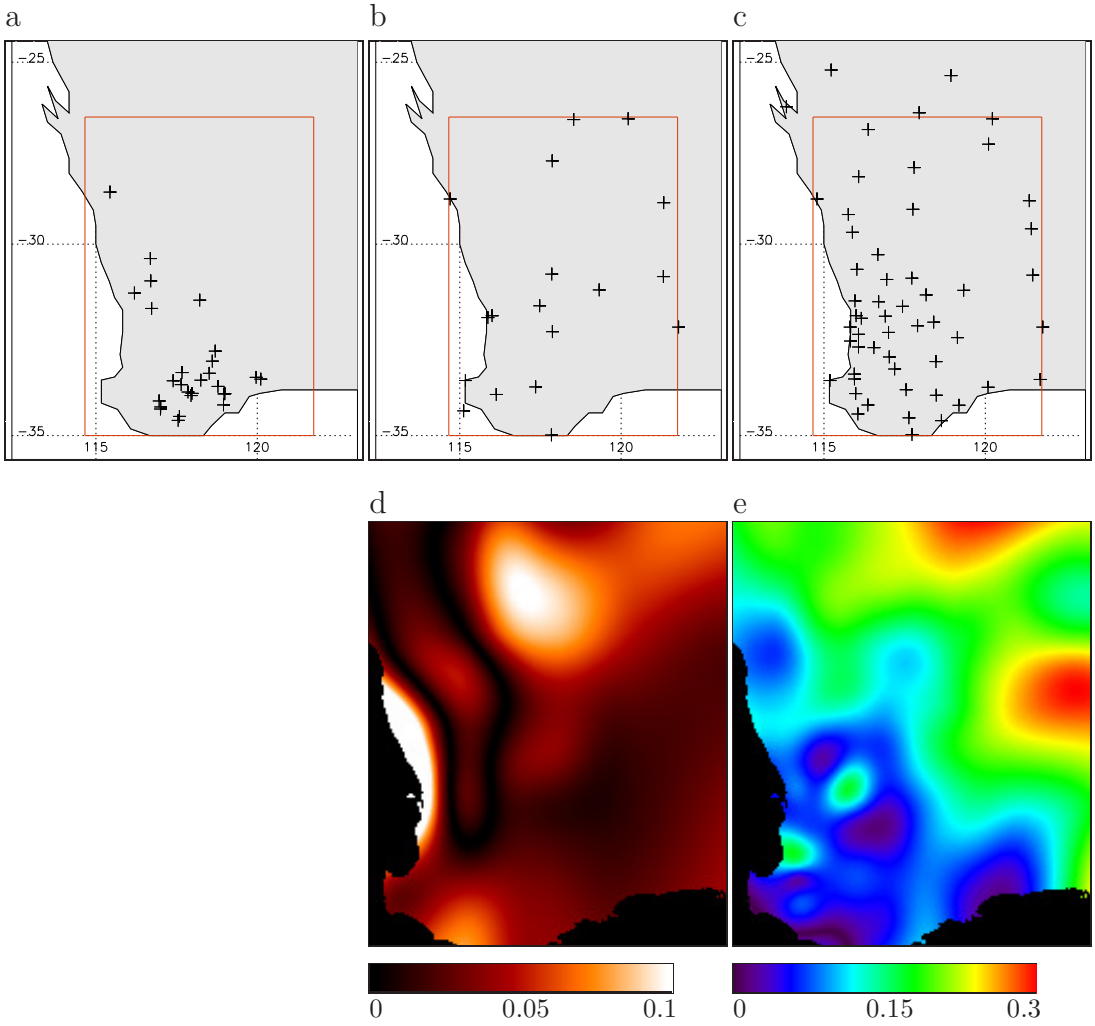


Figure 7.4: The sampling stations for (a) The Field Pea , (b) GHCN Temperature and (c) GHCN Precipitation. The red rectangle shows the modeled area of Western Australia. (d) Temperature and (e) Precipitation interpolated GHCN absolute ratio uncertainty surfaces.

GHCN Station No.	Absolute Deviation Ratio	Location (deg S/E)	
		S	E
50194403	0.0122	$28^o47'$	$114^o42'$
50194428	0.0461	$27^o44'$	$117^o52'$
50194430	0.0183	$26^o36'$	$118^o32'$
50194439	0.0299	$26^o35'$	$120^o13'$
50194448	0.0133	$28^o52'$	$121^o19'$
50194600	0.0234	$33^o35'$	$115^o10'$
50194601	0.0143	$34^o22'$	$115^o7'$
50194608	0.0529	$31^o57'$	$115^o52'$
50194610	0.0338	$31^o54'$	$116^o0'$
50194616	0.0195	$33^o57'$	$116^o8'$
50194626	0.0114	$31^o38'$	$117^o28'$
50194629	0.0177	$33^o45'$	$117^o21'$
50194632	0.0148	$30^o48'$	$117^o51'$
50194633	0.0083	$32^o19'$	$117^o52'$
50194634	0.0078	$31^o13'$	$119^o19'$
50194637	0.0111	$30^o52'$	$121^o19'$
50194639	0.0165	$32^o12'$	$121^o47'$
50194802	0.0193	$34^o59'$	$117^o50'$

Table 7.1: Absolute deviation in temperature as a ratio, WA region

distribution's histogram).

As the constant uncertainty grid is a *ratio* of uncertainty (or % uncertainty if multiplied by 100), the unique uncertainty grid in correct units (T or mm) was generated by multiplying the *ratio* of uncertainty by the related climate grid. It is this final unique uncertainty grid that is added to the Worldclim climate grids to generate the unique climate raster for each month. This is repeated for each climate surface, so there are 36 IDL-AVID input climate layers for each simulation. A single simulation of the IDL-DIVA executed BIOCLIM model is illustrated in Figure 7.5. A representation of the future-climate conditions BIOCLIM model is shown in Figure 7.6. As discussed in Section 6.3 uncertainty in the modeled future climate layers cannot be quantified simply. Therefore, for the future predictions the future A2a and B2a Climate Grids are constant. .

Each unique simulation was saved and, when the simulation was completed, the mean, standard deviation, skew and variance of every raster grid cell was calculated. Therefore, if 5000 simulations were created the statistics will be calculated from 5000 numbers for each grid cell. Finally the statistical result

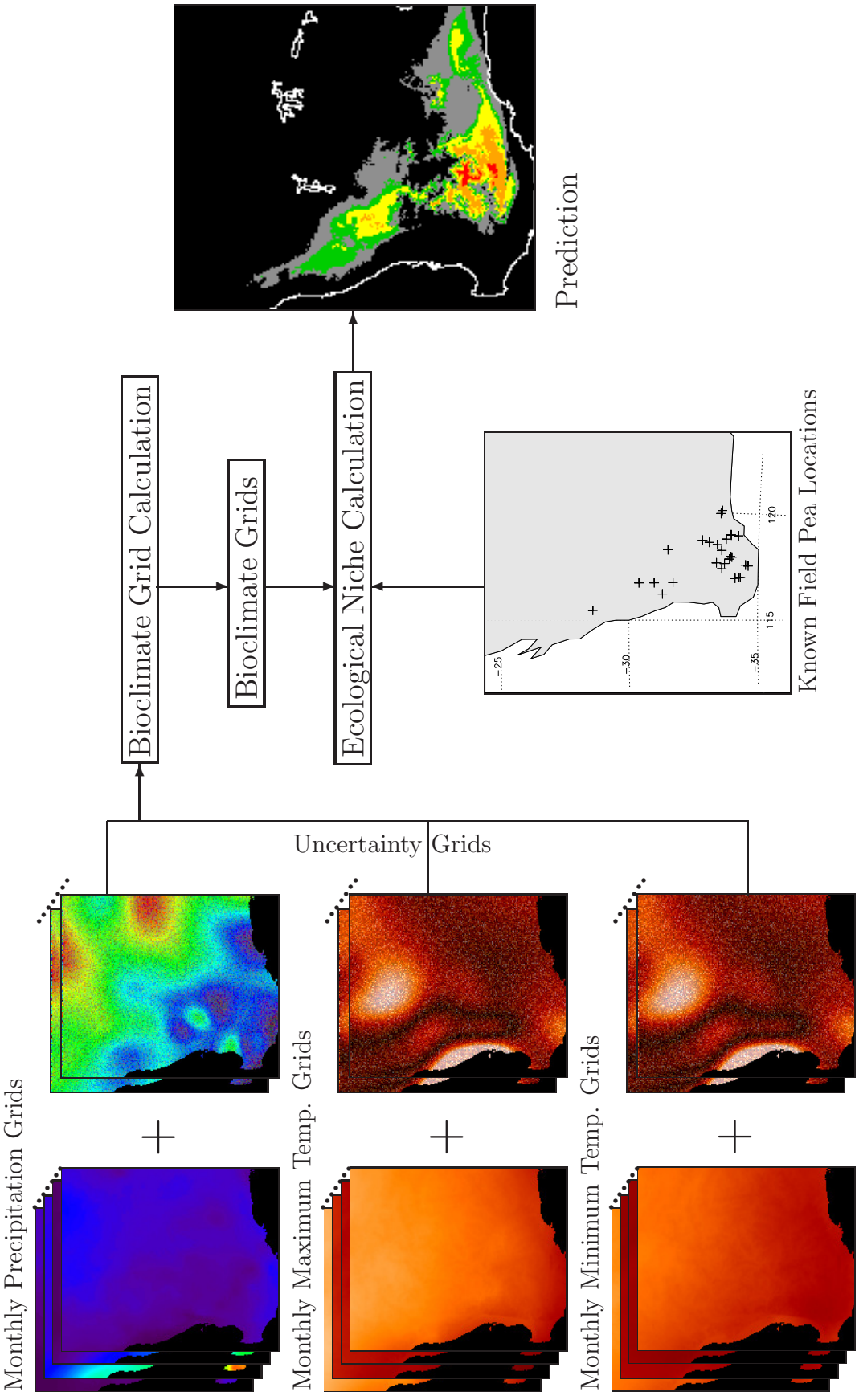


Figure 7.5: Graphical representation of the BIOCLIM present prediction model, with uncertainty added to the Climate grids. This sequence is repeated once for each simulation, with new uncertainty grids generated for each simulation.

GHCN Station No.	Absolute Deviation Ratio	Location (deg S/E) S	E	GHCN Station No.	Absolute Deviation Ratio	Location (deg S/E) S	E
50194402	0.1259	$26^o15'$	$113^o50'$	50194403	0.1017	$28^o47'$	$114^o45'$
50194404	0.1289	$29^o36'$	$117^o46'$	50194410	0.2664	$25^o13'$	$115^o13'$
50194411	0.1809	$28^o10'$	$116^o4'$	50194415	0.11	$29^o41'$	$115^o53'$
50194416	0.0875	$29^o12'$	$115^o45'$	50194422	0.1991	$26^o53'$	$116^o22'$
50194428	0.2294	$27^o55'$	$117^o48'$	50194430	0.2584	$26^o25'$	$117^o57'$
50194439	0.3005	$26^o35'$	$120^o13'$	50194440	0.25	$27^o17'$	$120^o6'$
50194446	0.3135	$29^o35'$	$121^o26'$	50194448	0.216	$28^o50'$	$121^o22'$
50194600	0.0641	$33^o35'$	$115^o11'$	50194604	0.115	$33^o26'$	$115^o56'$
50194605	0.0862	$32^o34'$	$115^o49'$	50194609	0.0716	$32^o12'$	$115^o49'$
50194610	0.0911	$31^o54'$	$116^o0'$	50194611	0.1143	$30^o40'$	$116^o1'$
50194612	0.0787	$31^o31'$	$115^o58'$	50194616	0.0603	$33^o56'$	$115^o59'$
50194617	0.062	$34^o13'$	$116^o21'$	50194619	0.1407	$30^o17'$	$116^o40'$
50194620	0.0923	$32^o43'$	$116^o4'$	50194621	0.0871	$31^o32'$	$116^o42'$
50194622	0.0683	$30^o57'$	$116^o57'$	50194623	0.1805	$31^o55'$	$116^o54'$
50194624	0.0683	$32^o44'$	$116^o33'$	50194625	0.0788	$32^o20'$	$117^o0'$
50194626	0.0994	$31^o39'$	$117^o26'$	50194627	0.041	$32^o59'$	$117^o2'$
50194629	0.0684	$33^o49'$	$117^o33'$	50194630	0.0501	$34^o32'$	$117^o38'$
50194632	0.0794	$30^o54'$	$117^o43'$	50194633	0.0464	$32^o10'$	$117^o54'$
50194634	0.2021	$31^o14'$	$119^o20'$	50194635	0.0905	$33^o6'$	$118^o28'$
50194636	0.0261	$33^o46'$	$120^o5'$	50194637	0.2701	$30^o50'$	$121^o28'$
50194639	0.193	$32^o12'$	$121^o47'$	50194802	0.0685	$34^o58'$	$117^o44'$
50195400	0.3733	$25^o22'$	$118^o56'$	50195608	0.0569	$33^o33'$	$115^o57'$
50195610	0.0788	$31^o58'$	$116^o10'$	50195611	0.1025	$34^o37'$	$118^o38'$
50195613	0.0543	$34^o27'$	$116^o3'$	50195614	0.0747	$32^o23'$	$116^o4'$
50195618	0.0669	$33^o18'$	$117^o11'$	50195624	0.1104	$31^o21'$	$118^o10'$
50195626	0.0952	$32^o4'$	$118^o24'$	50195627	0.1332	$32^o28'$	$119^o8'$
50195628	0.0947	$33^o58'$	$118^o29'$	50195636	0.0521	$34^o14'$	$119^o11'$
50195638	0.2327	$33^o34'$	$121^o42'$				

Table 7.2: Abslolute deviation in precipitation as a ratio, WA region

raster was saved in DIVA-GIS and ENVI format.

7.4.1 Influence of Number of Simulations

For the Monte Carlo Model to produce enough simulations to represent a real situation, the simulations must be repeated until the statistical results of the BIOCLIM predictions at each grid cell stabilises. Investigation concluded that 4500 simulations was sufficient to achieve this. For example, when compared to the 4500 simulation result, a prediction made with greater than 4500 simulations produced insignificantly small differences in the descriptive statistics of the predictions. As illustrated in Figure 7.7(a and b) this is seen in the high

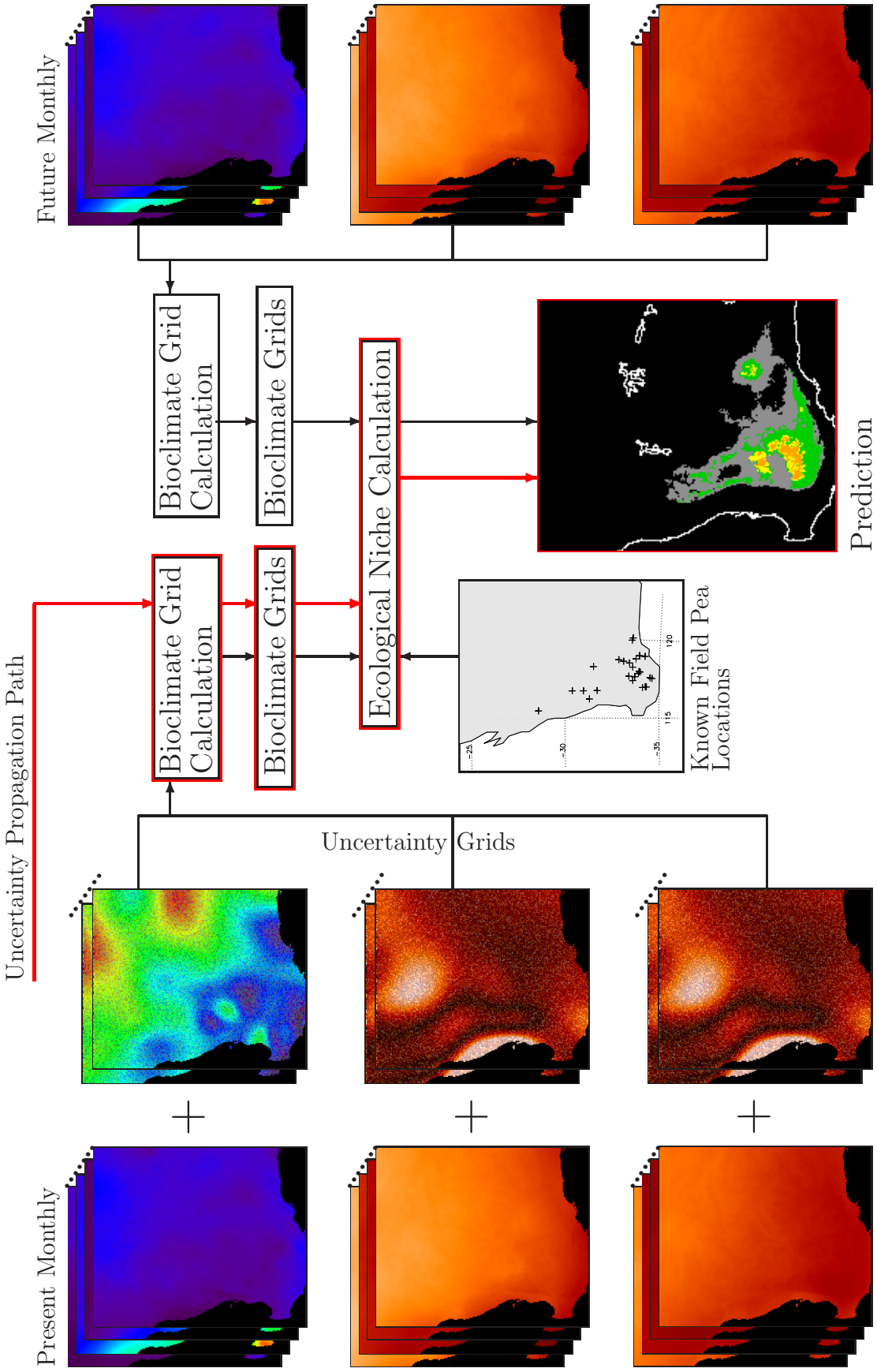


Figure 7.6: Graphical representation of the BIOCLIM future prediction model, with uncertainty added to the Present Climate grids. This sequence is repeated once for each simulation, with new uncertainty grids generated for each simulation. The uncertainty propagation path is coloured red - note that the steps in the generation of the future grids is not in it.

agreement of the mean and standard deviation of the prediction grids, 4500 and 7000 simulations (correlation of 0.99).

As shown in Figure 7.8 the predictions at almost all grid cells were skewed to some degree. This may explain why there was always some difference in the mean and standard deviation predictions. For most of the area (coloured red), the skew is negative but still relatively low. More notable is that the greater the skew value is from zero, there is a large disagreement, especially at values < -20 and > 20. Also, where this greater disagreement occurs is mostly along the border of where the BIOCLIM predictions greater than 0 occurs. Repeating this test with a higher number of simulations did not appear to change this, so it can not be attributed to the number of simulations. Also, this pattern is clearly present when the uncertainty grids are skewed. The same tests applied to the South American region produced the same diverging pattern, suggests that this was not a geographically influenced anomaly. Where and why this may be occurring is discussed in Chapters 8 and 9.

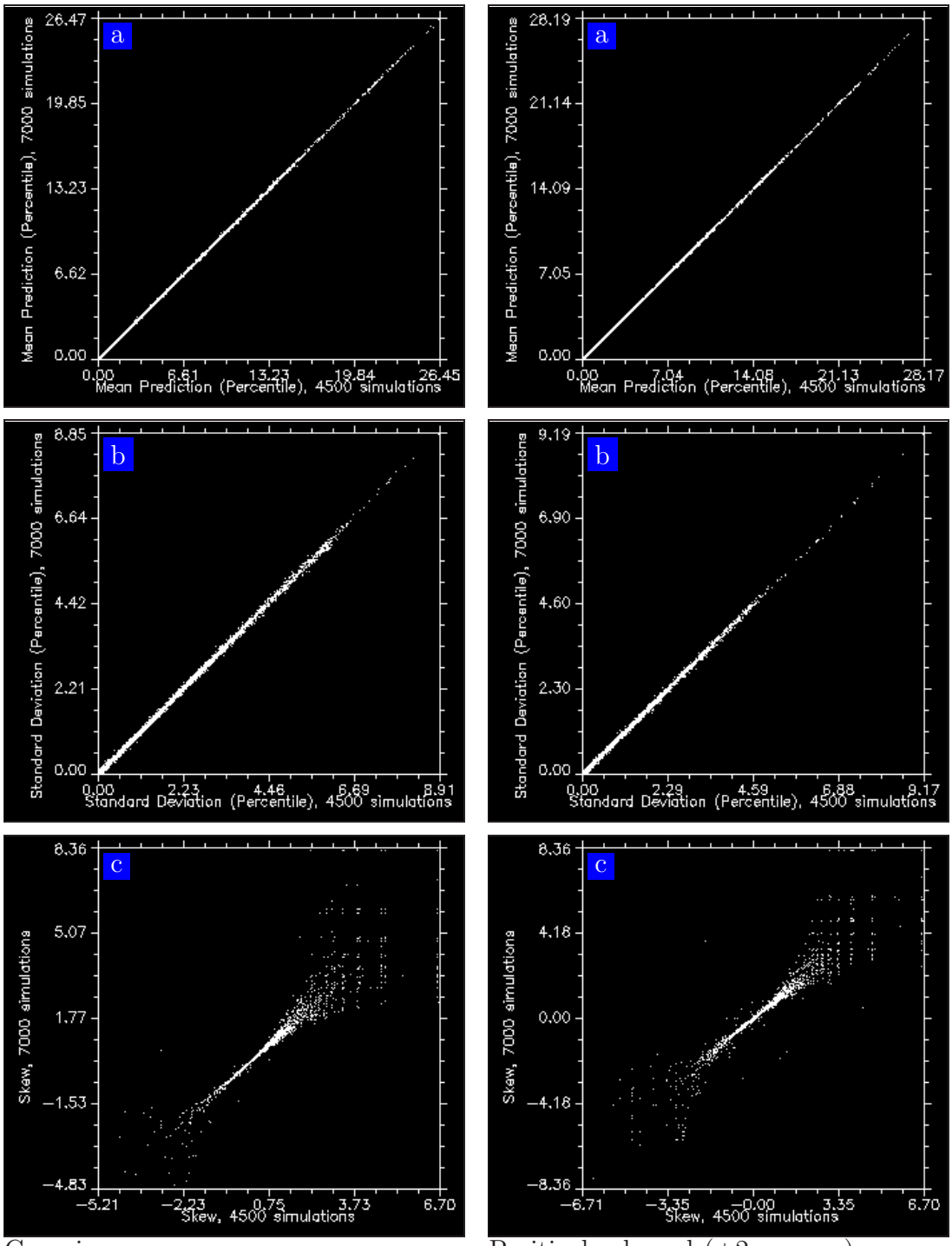
7.5 Prediction Examples

Figure 7.9 shows two examples of BIOCLIM model predictions, executed with the IDL-AVID program:

(a) The default prediction for the Western Australian region as discussed in Section 7.2.2.

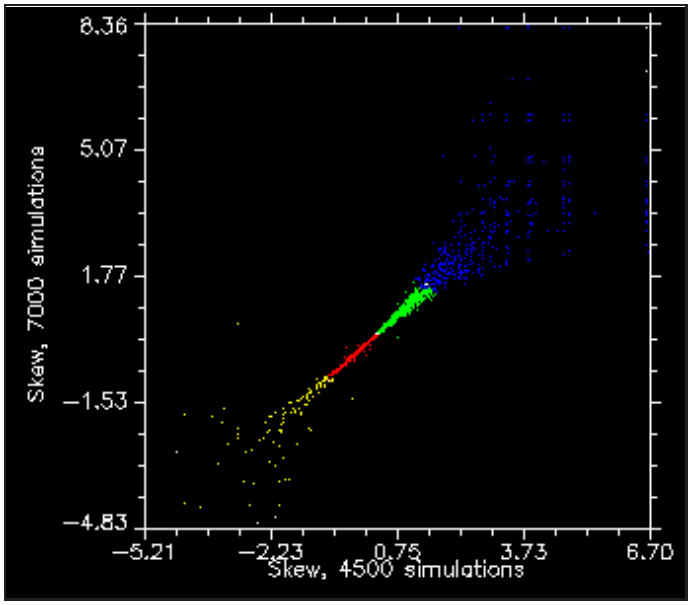
(b) The mean prediction of the Monte Carlo Similation (per raster grid); normally distributed uncertainty. (c) and (d) showing the standard deviation and skew of the predictions.

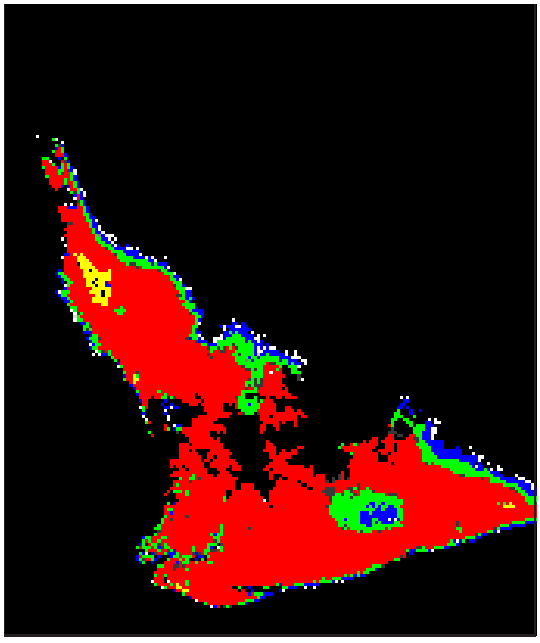
Figure 7.10(a) graphs the difference in the predictions of these two process models. As seen on the x axis, the default prediction has 23 distinct values from 0 to 28.8 percentiles. The y axis values shows how the prediction has changed as a result of the added uncertainty in the climate grids. Figure 7.10(b) shows the mean prediction versus standard deviation relationship at each grid cell. Clearly,



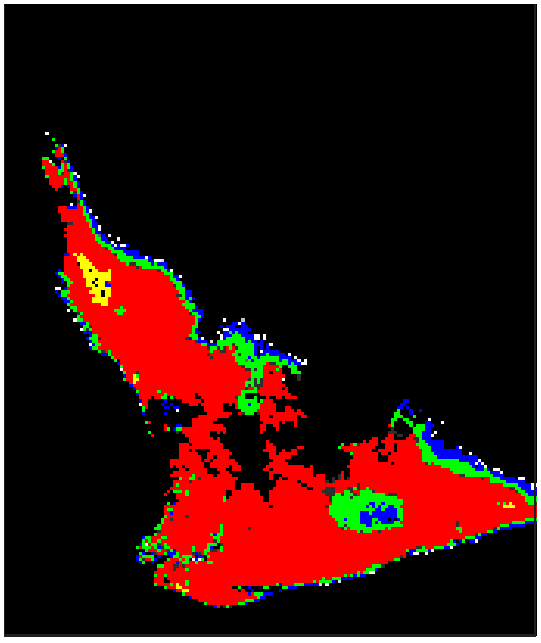
Gaussian Positively skewed (+2 gamma)

Figure 7.7: Comparison of BIOCLIM (a) mean, (b) standard deviation and (c) skew. (a) 4500 versus 7000 simulations. Normal and positively skewed uncertainty distribution model.





4500 simulations



7000 simulations

Figure 7.8: Comparison of the skew in two BIOCLIM models at each grid cell. 4500 and 7000 simulations, Gaussian Distribution. The colours in the maps illustrate the regions where skew values occur.

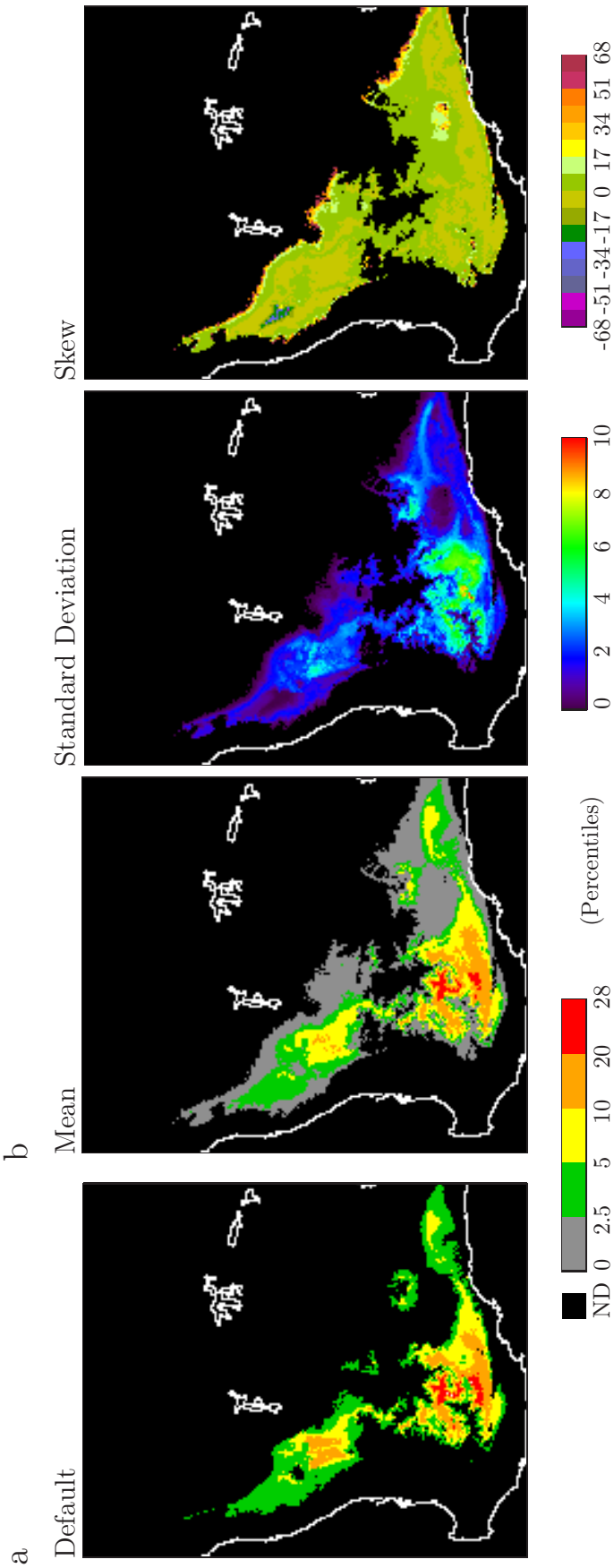


Figure 7.9: Examples of BIOCLIM Present Model Prediction and further analysis, (a) without and (b) with uncertainty in the climate grid inputs. Default, Mean and Standard Deviation are in Percentile units. As described in Section 7.2.2, Default or Mean predictions equaling 0 are labeled as ND.

no relationships can be determined from this graph alone. However, combining the interpretation of both these graphs can be used to put the predictions into specific *spatial regions boxes*, from which conclusions can be more easily drawn. This is discussed, and its use illustrated, in Section 8.1.1. The ENVI *regions of interest* tools were used for this.

7.6 Summary

To summarise, the Monte Carlo Simulation method is used to test BIOCLIM's sensitivity to uncertainty in the climate grids. With each simulation, new uncertainty grids are added to the Worldclim climate grids producing a unique prediction grid. For a stable statistical result, at least 4500 simulations per grid cell is required. Therefore, 4500 prediction layers were simulated and then statistically analysed.

Most statistical results, at each grid cell, are very similar above this number of simulations, with the exception of the skew at its higher values. This anomaly does not be appear to be due to any geographical or climatological reason.

Interpretation of the mean to standard deviation relationship of a prediction could not be easily determined. Grouping data into *spatial regions boxes* allowed further investigation of this relationship.

a b

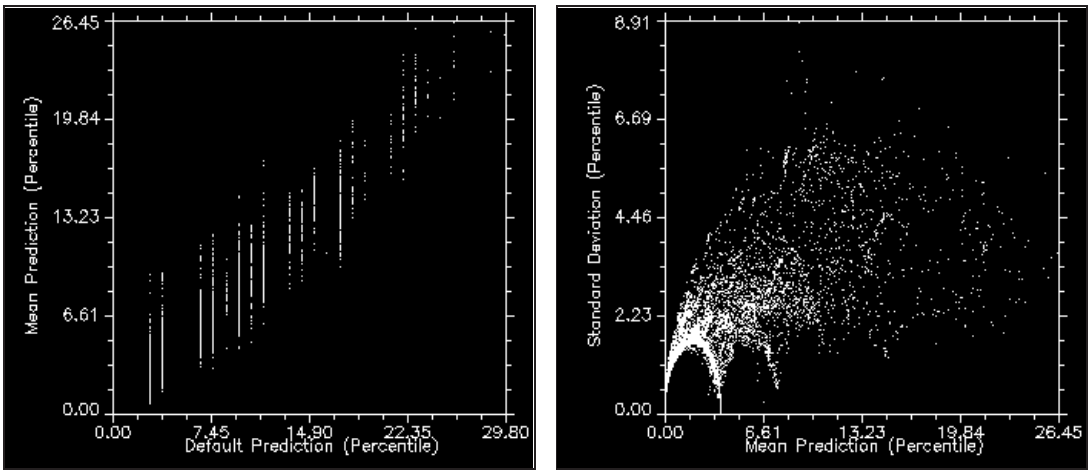


Figure 7.10: Plots of (a) default versus mean prediction and (b) mean versus standard deviation.. See Figure 7.9 for images of results.

Chapter 8

Sensitivity of Bioclim Predictions

The initial study of uncertainty propagation through the BIOCLIM statistical model (Section 7.5) clearly showed that (a) the model can change both the size and spatial distribution of the uncertainty and (b) that the distribution of the error or uncertainty (in each grid cell) in an input layer can also change both the prediction and its uncertainty. This chapter aims to increase the understanding of why this is occurring in the BIOCLIM model in the south west of Western Australia, which is a substantial and important agricultural area. The region is relatively flat so change in elevation will have minimal effect on the climate and hence a minimal effect on the prediction result. Instead, in Western Australia, the annual rainfall and temperature patterns will be the primary influence of the suitability of an area for a specific crop.

The south western region of Western Australia was, before the arrival of Europeans, largely covered by forests of differing densities and species. This region is now mostly used for agriculture, with a significant part of it used for wheat, barley and other suitable crops. This region is known as the "Wheat Belt", and is an important agricultural area. The rainfall in this region ranges from 200 to $1200mm$, with rainfed broadacre agriculture being practiced in the range of 350 to $700mm$. This rain occurs with the passage of cold fronts from west to east during winter, due to the passage of the high and low meteorological atmospheric pressure gradients which dominate during the mid year period. Therefore, in the

region of Western Australia studied, the greatest rainfall occurs in the south west, with the highest measurements being along the southern coast and in the western region's relatively close to the coast. This area of Western Australia is most suitable for the Field Pea (see Figure 8.1). It is possible to have high but patchy rainfall during the hot summer period, but this is not suitable for this crop's production.

There has been a significant drop in the rainfall in this region since 1970 which has already had a strong impact on the agriculture and native species of the region. Therefore, there is a significant interest in what impact the projected future climate scenarios will have on the region's suitability to traditional crops in the future.

The other most important variable is soil quality. However, this can be controlled and uncertainty effectively removed by, for example, controlling nutrient levels with well controlled use of fertilizers. Therefore, quantifying uncertainty in the soil quality is not as important as the uncertainty in climatic variables, especially when using tools such as BIOCLIM, which predict (using only climate data) which sections of a large area may be suitable for use in growing a particular crop. Also, because BIOCLIM is most accurate when modeling large areas, small regional studies were not included in this thesis.

This chapter discusses:

1. The propagation of uncertainty through the BIOCLIM model for both present and future predictions. The resulting uncertainty in the prediction will be referred to as "uncertainty" in this chapter and Chapter 9.

2. How the distribution and size of uncertainty changes (and hence influences the validity of the models results).

Chapter 9 discusses:

1. Why these changes may be occurring in the present prediction model which has a normal distribution of uncertainty and

2. How they might be minimised.

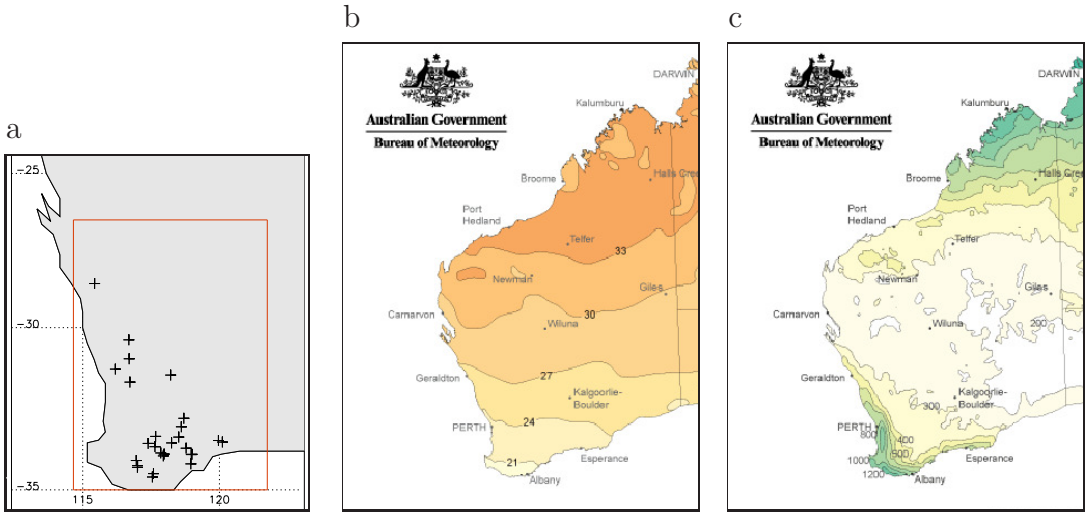


Figure 8.1: (a) Field Pea sampling stations, area of Australia studied (red bordered region), average (b) maximum temperature and (c) rainfall in Western Australia.

8.1 Present Predictions in the Western Australian Region: Gaussian Uncertainties

Figure 8.2 shows the default prediction for the Western Australian region. The reasons for the difference in predictions between the north-east and south-west/ southern regions of the Wheat Belt region is because the highest and most consistent mid year rainfall occurs along the western boundary (the western coastline) and in the south of this region. As with the study of the South American region, there are some differences between the default prediction and the uncertainty-included predictions (see Figure 8.2, Table 8.4 has a statistical comparison of the influence of the uncertainty layers on the mean and uncertainty for the bioclimatic present prediction). There is a high correlation between the mean result layers predicted when both the GHCN and Wordlclim uncertainty grids are added and when only the GHCN uncertainty layer is added. This suggests that the strongest uncertainty influence is from the GHCN uncertainty alone.

As is clearly seen in Figure 8.2, there is a relatively high agreement between

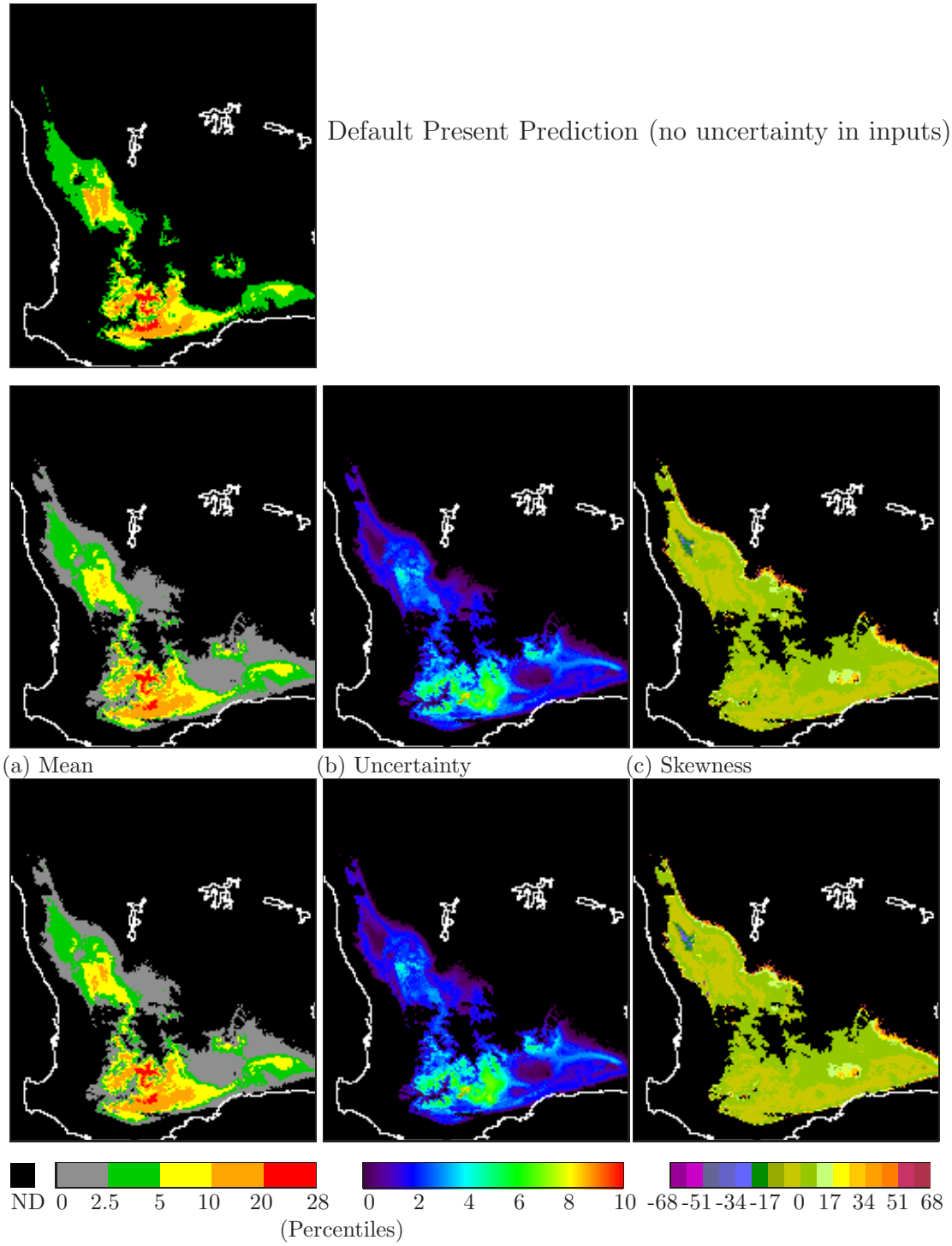


Figure 8.2: BIOCLIM present prediction results: (a) Mean, (b) Uncertainty and (c) Skewness. Error simulated, 4500 simulations. Middle row, GHCN and WorldClim uncertainty added. Bottom row, GHCN uncertainty added.

Normal Distribution (Worldclim and GHCN)	Normal Distribution (GHCN)	+skewed (GHCN)	-skewed (GHCN)
0.9756	0.9766	0.9869	0.9859

Table 8.1: Correlation between the Default predictions and predictions when uncertainty is present in the model.

the predictions made with and without uncertainty in the model input (with a correlation between predictions of 0.9756 (Table 8.1)), but there are some points of difference. The similarities are clearly where the areas of highest, medium and low prediction occurs, such as the high predictions made for the "northern" region and for the "south west" region. However, there are some significant differences such as the smaller spatial extent of the higher prediction in the northern region. The other significant difference - which is perhaps expected - is that the groups into which the predictions are boxed are less clear when uncertainty is added to the model inputs, as clearly shown when the default prediction is plotted against the uncertainty-included prediction (Figure 8.3(a) and (b)). In the default model, it is clear that the predictions have been boxed and that this clarity is lost when uncertainty is added to the precipitation and temperature model inputs.

A result of this fuzziness, which occurs when uncertainty is included in the models inputs, is that a percentile of less than 2.5 is calculated. A large fraction of this covers a large area where the default prediction was 0. In these areas the uncertainty is similar to the mean prediction at each grid cell, which suggests that the prediction in this area is of a low certainty. In the areas where the prediction is highest ($\approx$20-30), the uncertainty is also higher but significantly lower than the predicted value. However, it is clear that there is no simple linear or easily interpretable spatial relationship between the mean and uncertainty of the prediction across the Western Australian region. Also, the uncertainty visibly increases and decreases in an "arc" like relationship to the increasing mean (for some grid cells).

There is a close agreement (correlation of 0.9817; and clearly seen in Figure

8.2) in the skewness of the Prediction results (from which the mean Prediction was calculated), whether the Worldclim uncertainty layer is included or excluded. The skewness is mostly in the range of -17 to $+17$, with large continuous areas being clearly in the positive or negative. The spatial distribution of the Worldclim and GHCN uncertainty surfaces do not mirror this pattern and the synthesised uncertainty inputs were normally distributed.

8.1.1 Analysis of Uncertainty

As shown in Figure 8.3(a), the default prediction for the Western Australian region has 23 percentile values, as shown in Table 8.2). There appears to be a grouping of the results into six larger bins. Region of interest (ROI) 1 is where the default predicted value is 0 but, with uncertainty added, has results greater than this (see Figure 8.3(b)). Similarly, the predicted results for the other default locations also vary around the default value. For example, at the locations where the default prediction was 2.9 Percentile (ROI 2), the values range from 0.7 to 9.4, have a mean of 2.8 and a standard deviation of 1.1 Percentile. As seen in the analysis of the other ROI, there is a clear skew in the prediction such that the means are lower than the default values (see Table 8.3).

The uncertainty in the predictions, grouped into the ROI's, is shown in Figure 8.3(c). The highest variation in these results is clearly in the region of lowest prediction, but they are not evenly skewed. The greatest agreement in the uncer-

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
0	2.9	7.7	13.5	21.2	28.8
	3.8	8.7	14.4	22.1	
		6.7	15.4	23.1	
		7.7	16.3	24.0	
		8.7	17.3	25.0	
		9.6	18.3	26.0	
		10.6	19.2		
		11.5			

Table 8.2: Default Predictions in the Western Australian Region (Percentiles). The maximum possible prediction is 50 Percentiles. The predictions are rounded to one decimal place.

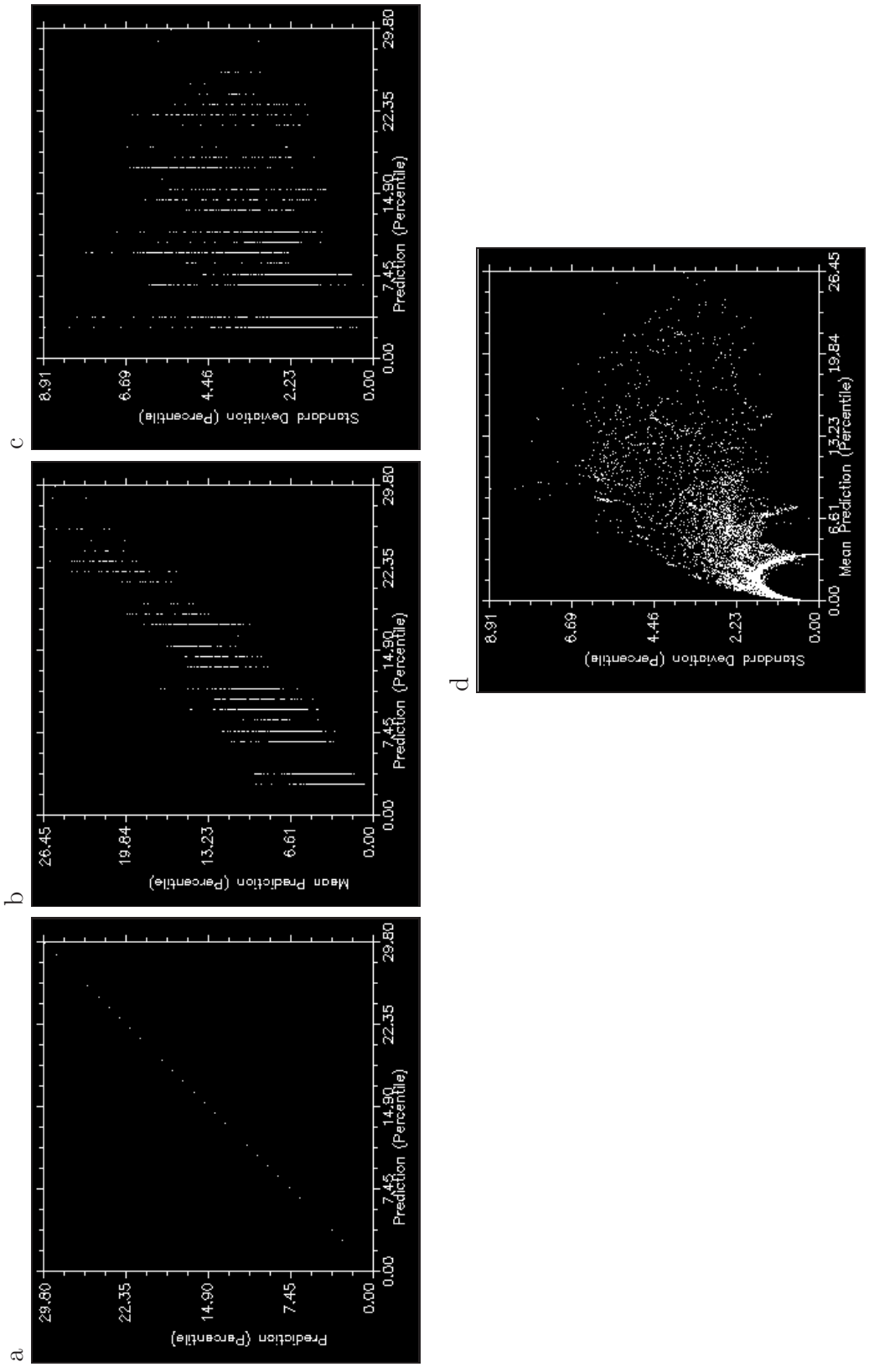


Figure 8.3: A comparison of Default Predicted Results with: (a) default only. (b) GHCN and Worldclim (c) Uncertainty (in uncertainty included results). (d) Mean versus Uncertainty - GHCN and Worldclim included

tainty occurs in the ROI where the highest values were predicted. There appear to be ROI where the grids cells have a lower uncertainty. This is more clearly seen in Figure 8.3(d), where there are points where the uncertainty reduces, as the mean increases, before increasing again. The reason for this "arc" like relationship cannot be easily determined from the plot, especially considering that it is information across a two dimensional area. However, one conclusion that can be drawn is that the lower uncertainties which occur as the prediction increases suggests that some prediction range(s) are less influenced by the uncertainty in the climate inputs.

To understand the relationship between the prediction mean and its uncertainty, the uncertainty at each grid cell in a particular ROI was investigated independently. As shown in Table 8.3, a ROI corresponds with the point where a particular default prediction occurred in the Western Australian region. As expected, the number of locations within each ROI varies considerably. The mean Prediction results which are in the ROI were then subdivided into four categories:

(1) minimum prediction, to default prediction minus one standard deviation (red).

(2) default prediction minus one standard deviation, to default prediction (green).

(3) default prediction, to default prediction plus one standard deviation (blue)

(4) default prediction plus one standard deviation, to maximum prediction (magenta).

The standard deviation of the predictions, within the ROI of interest, was chosen as a box within which most measurments should lie. However, the centre of the study range is set to the default prediction (which defined the ROI).

The mean versus uncertainty relationship for the only ROI in Group 1 (the "0" ROI), is shown in Figure 8.4. The predictions are always ≥ 0, so there are no red or green areas. For almost all of the locations in this ROI, the uncertainty increases with increasing prediction. Relative to the mean prediction, the uncertainty is clearly greatest where the mean is lowest. This relationship

Region of Interest	Default Prediction	Number of grid cells	Min	Max	Mean	Standard Deviation
1	0	3126	0.01	5.56	0.66	0.79
2	2.9	867	0.74	9.49	2.82	1.11
3	3.8	1280	1.58	9.61	3.71	1.06
4	6.7	433	3.27	12.23	6.36	1.31
5	7.7	381	3.24	12.23	7.01	1.40
6	8.7	31	4.52	10.49	8.08	1.08
7	9.7	246	4.53	14.77	8.89	1.76
8	10.6	100	4.92	12.76	9.84	1.76
9	11.5	171	6.13	17.14	10.06	1.63
10	13.5	77	8.51	14.95	12.56	1.57
11	14.4	56	9.04	15.16	12.32	1.52
12	15.4	77	11.16	16.64	14.41	1.14
13	16.3	1	–	–	–	–
14	17.3	124	10.00	18.50	13.81	1.90
15	18.3	34	13.28	19.91	16.40	1.97
16	19.2	8	–	–	–	–
17	21.2	14	–	–	–	–
18	22.1	48	15.88	24.28	20.30	1.71
19	23.1	29	19.17	26.04	22.00	1.50
20	24.0	11	–	–	–	–
21	25.0	2	–	–	–	–
22	26.0	8	–	–	–	–
23	28.8	2	–	–	–	–

Table 8.3: Statistics of Predictions at Default Prediction Grid Cells: Western Australian Region

then decrease as the prediction increases.

The general pattern observed for all ROI can be seen in Figure 8.5. The lowest uncertainty is present in the green and blue sub-regions, with the lowest of these values occurring as the prediction approaches the default prediction of the ROI, in this case 2.9 and 3.8. In the 2.9 ROI, the red sub-region is relatively small and has a lower uncertainty than many points in the green sub-region, even though it is further from the default prediction. However, the lowest uncertainty is clearly in the green and, to a greater extent, blue sub regions. In the points where the uncertainty-included prediction is higher than the default prediction, the uncertainty is clearly greater - especially the points in the majenta sub region.

The reason for this $\surd$ shaped mean to uncertainty relationship, in the Bioclim model, is relatively clear. As discussed in Section 5.1.1, the BIOCLIM climate envelope model assigns the lowest bioclimatic layer result, for a particular grid cell, as the final prediction (for that grid cell). For example, if the mean temperature bioclimate layer at a grid cell is in the 2.5 - 5.0 percentile span, but all other bioclimate layers are in the 5.0 - 10 percentile span, it will be assigned a value of "2.5 - 5.0" percentile. Therefore, adding uncertainty to the climate inputs that are used to calculate the climatic layers causes the percentile value assigned to change, especially when the climatic grid value is near the span boundaries. A higher uncertainty in the predictions occurred in the grid cells in these regions, because the likelihood of the predictions not falling into the same grouping is higher. If, on the other hand, the grid cell's climatic values are close to the median of the climatic envelope, the assigned percentile is not as likely to change to the same degree and this will be reflected in a lower uncertainty value for the prediction. An analysis of how this relationship influences the validity of the results in each group is discussed in the following section.

8.1.2 Spatial Patterns

In Group 1, ROI 0.0 (region where all default points in image are equal to 0 (referred to as *Null* in the text or ND in the images)), the most consistent

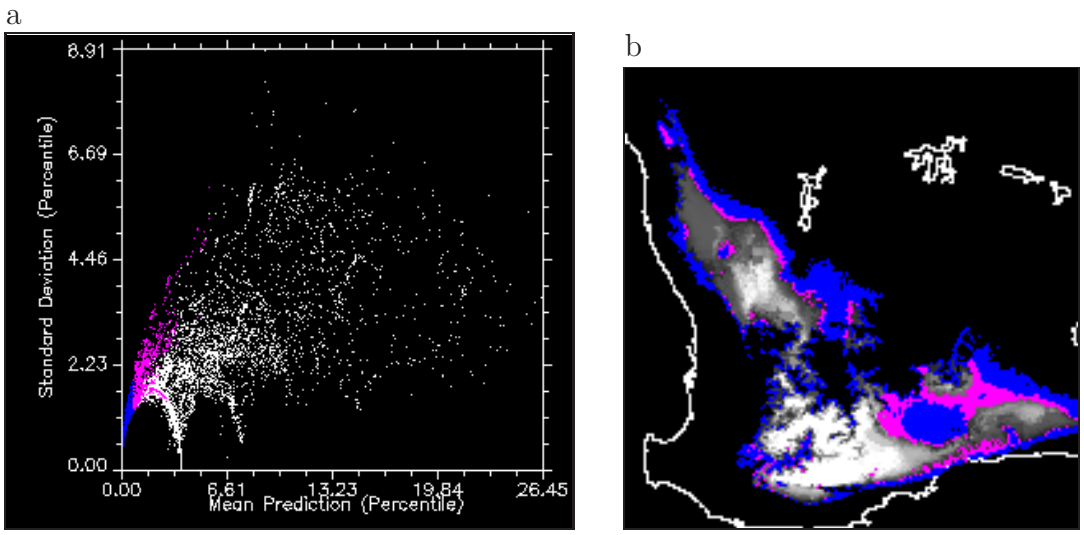


Figure 8.4: (a) Mean versus Uncertainty relationship in ROI 1. (b) ROI location. Where there are no entries for Min, Max, Mean, Standard Deviation, the predictions were not changed when uncertainty was added to the climate inputs. For ROI colour classifications see categories described on Page 129.

patterns (in uncertainty size) and largest number of points, is along the north east border region. For most of this area the default prediction is null. As expected, the prediction when uncertainty is added, is in the lowest percentile range (see Figure 8.2). This suggests that the range of predictions centres on the $0 - 2.5^{th}$ percentile range, but as the uncertainty is higher than the lowest predictions for most of this region, the quality of the prediction is very low. The highest uncertainty values occur at the boundaries of areas where the predictions are in $2.5 - 5.0^{th}$ percentile range, the starting percentile of the default prediction. This is consistent with the higher sensitivity to uncertainty observed at areas which border areas of different predictions (as discussed in Section 8.1.1). However, this is not a generalisation that can be applied to all of the default group borders. For example, in the south west corner, where the prediction quickly (spatially) increased from null to greater than 5^{th} Percentile, the grid points with higher uncertainty are not present. This suggests that:

1. The uncertainty in the climate in this area is low (see Figure 7.4). This is likely in the south western region of the area studied due to the higher and

more consistent rainfall of this region.

2. Bioclimate values at these grid cells fall in the centre of the range of the percentile group.

The first of these is the most likely, given the climate of that area. However, the combination of these would further reduce the uncertainty in the final prediction and should therefore be considered.

In the 3.8 ROI (Group 2), there is a repeat of the relationship observed in the 2.8 ROI sub-region pattern. However, the Uncertainty values are clearly much lower, with it reducing to and then rising from almost zero, as the prediction approaches and passes the default prediction value. There are points were there are similar predictions in both ROI, but this is not always the case with the prediction's associated uncertainty. For example, there is similar uncertainty in the red 3.8 sub-region and green 2.8 sub-region (just after the top of the first arch). However, as the prediction in the 2.8 ROI enters its blue region this is not the case, as the uncertainty is considerably higher (except for 3 points that have mean predictions of 3.0 Percentile). The initial conclusions that can be drawn from the patterns observed in Group 2 is (a), that some prediction "range" (in this case centred on the 3.8^{th} percentile) is less prone to uncertainty and (b) that it is larger at points where predictions are higher than the default prediction (for those points).

In the 3.8 ROI, the uncertainty can drop to close to zero, which clearly shows that for grid locations in this ROI, the climatic layer results are in closer agreement than for any of the grid locations in the 2.8 ROI. That is, they fall into a smaller range of percentile spans, centred on 3.8. This difference (in uncertainty patterns between the 2.8 and 3.8 ROI) are clear indicators of the sensitivity of the models result to uncertainty in a models inputs. While the pattern in the uncertainty may be similar i.e. dropping then increasing (as the prediction value increases), the significant difference in the uncertainties for each ROI clearly shows that the degree to which the climate envelope model's

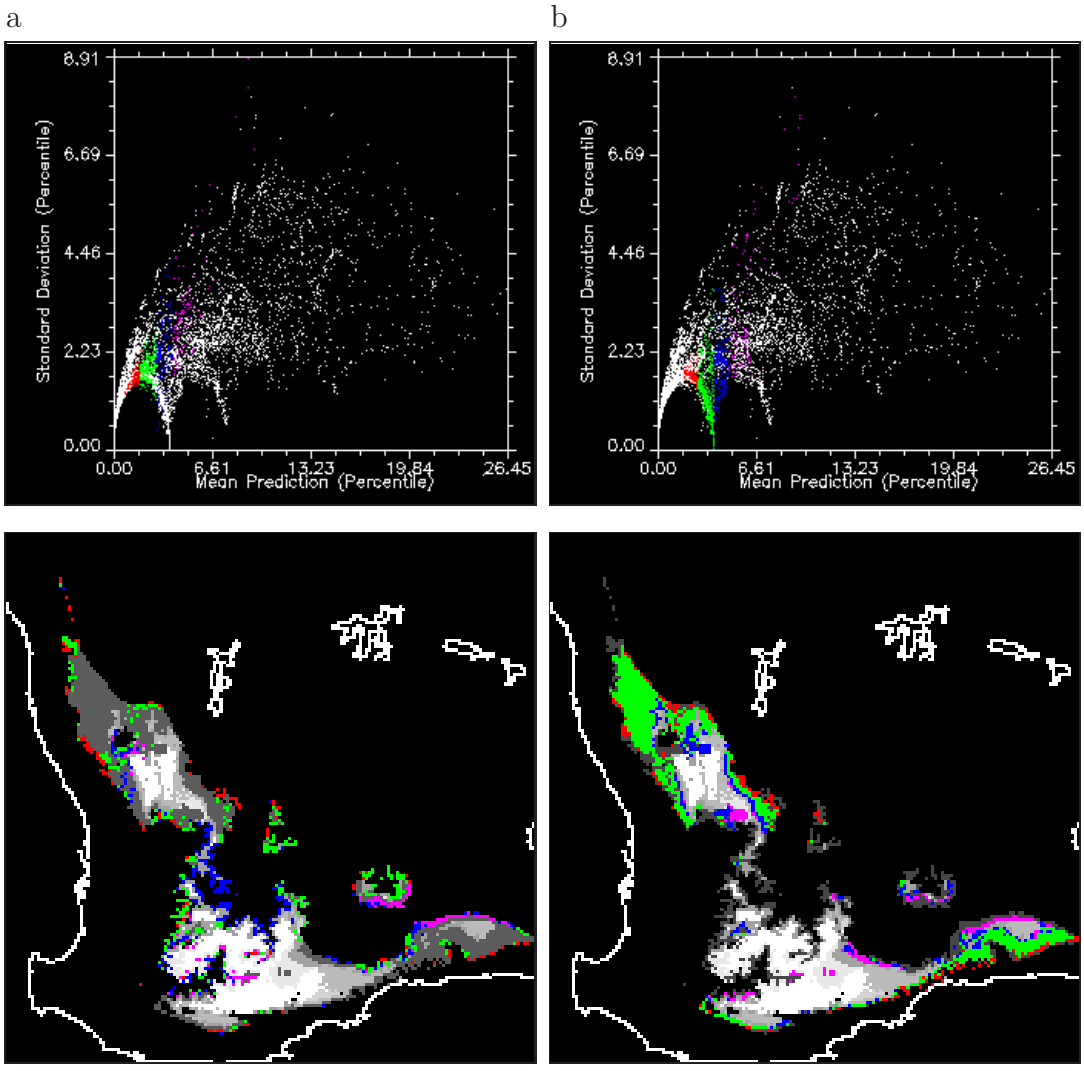


Figure 8.5: Mean versus Uncertainty Relationship in Group 2: (a) 2.9 (b) 3.8. For ROI colour classifications see categories described on Page 129.

sensitivity to uncertainty varies between the ROI's. In summary, for group 2 the results suggest that for the 2.9 ROI grid points, the climatic variables are close to the boundaries in the specie's climate profile for most of its grid points, even when the prediction is close to 2.9. For the 3.8 ROI, this occurs in less grid points as there are clearly points where the influence of uncertainty is significantly lower. It is at these points where variation in climate inputs has least influence on the validity of the Bioclim prediction results.

The mean to uncertainty relationship observed in the Group 2 ROI's are largely repeated in Group 3, as shown in Figure 8.6 and 8.7. This is most clear in the 6.6, 7.7 and 9.6 ROI (as there are more points) than in the 10.6 and 11.5 ROI. The other ROI in this group are excluded due to a low number of points (see Table 8.3 for details).

The reasons, related to the BIOCLIM envelope model's structure, for why this observed pattern is occurring are the same as for Group 2. However, the climatic conditions in the grid points of Group 2 clearly vary from those of Group 3 as the predictions are higher. The lowest uncertainty occurs in the 7.7 and 11.5 ROI, with a clear uncertainty to mean relationship as the prediction rises. These minimum uncertainties (which are still of a significantly high value as they occur at the top of the arcs) occur in the 6.6, 9.6 and 10.6 ROI. In these there are grid cells in the blue area, with a mean prediction with a notably lower uncertainty (6.3, 6.4 and 6.7 Percentiles), but the majority are notably higher than in the green sub-region. This relationship mirrors the relationship observed in the Group 2, (especially in the ROI which contain more grid cells), but the lowest uncertainty is not as low as occurs in Group 3. This could be due to the fact that most of the grid cells in Group 3 are not clearly grouped in one large area, except for the larger "green" area in the south of the state (which also is where the lowest uncertainty occurs).

The lowest uncertainty in Group 4 occurs in the 15.4 ROI (see Figure 8.8 and 8.9). Once again a similar pattern is observed as occurred in Groups 2 and 3, but the lower number of points in Group 4 makes this less easily seen. It appears

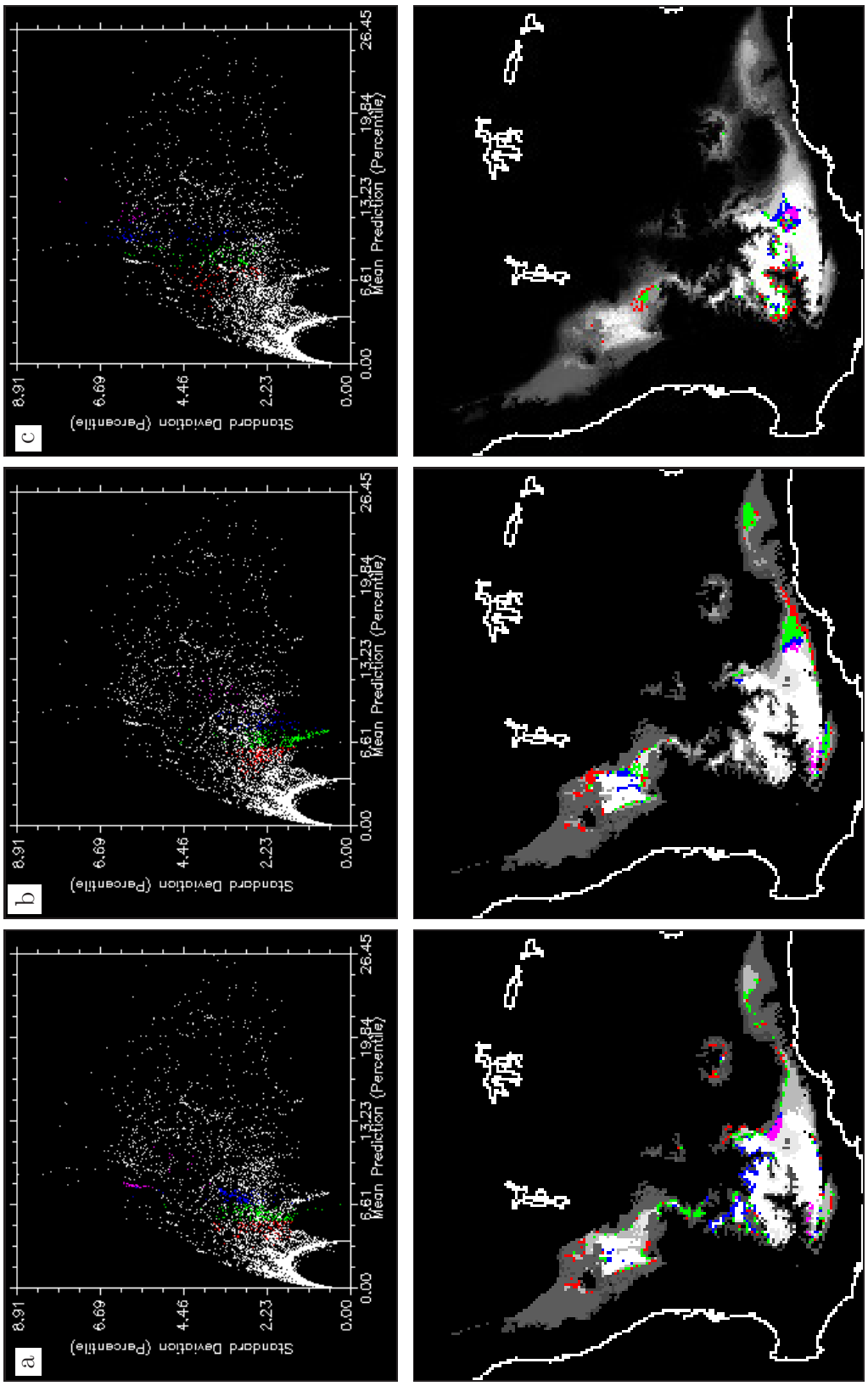


Figure 8.6: Mean versus Uncertainty Relationship in Group 3: (a) 6.6, (b) 7.7 and (c) 9.6. For ROI colour classifications see categories described on Page 129.

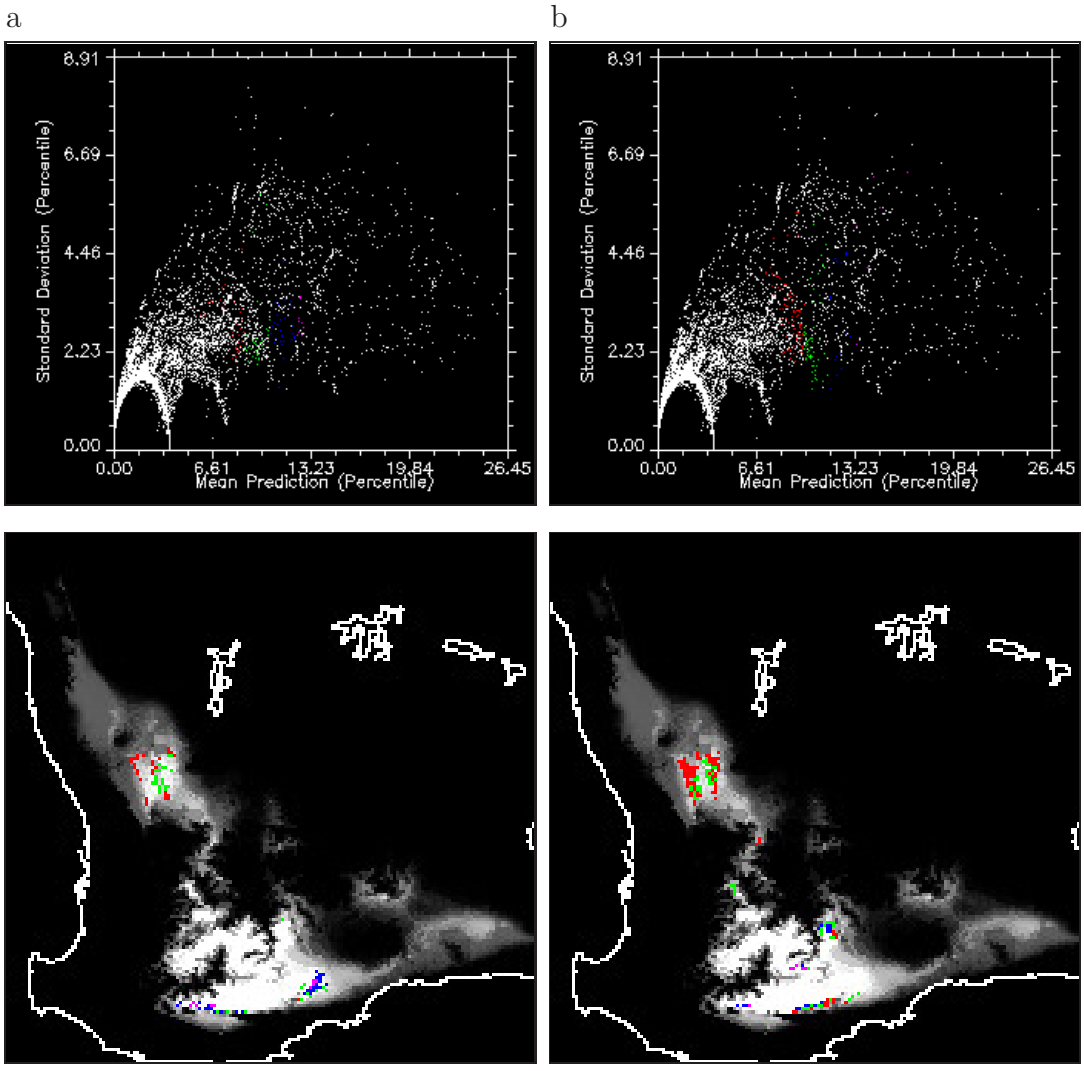


Figure 8.7: Mean versus Uncertainty Relationship in Group 3: (a) 10.6 and (b) 11.5. For ROI colour classifications see categories described on Page 129.

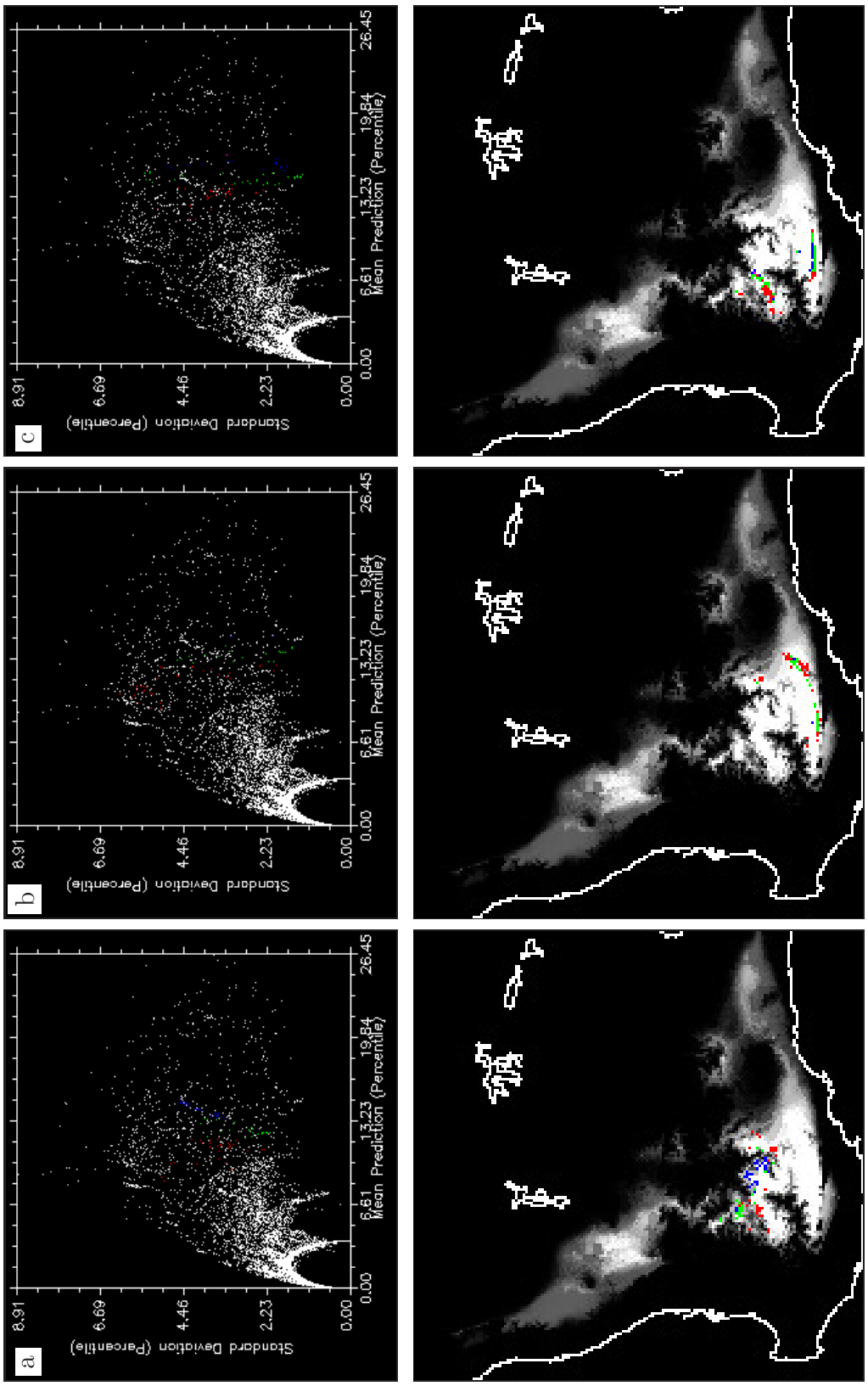


Figure 8.8: Mean versus Uncertainty Relationship in Group 4: (a) 13.5, (b) 14.4 and (c) 15.4. For ROI colour classifications see categories described on Page 129.

that the uncertainty values in the red sub-region are the highest, suggesting that the model's response to uncertainty changes as the prediction value increases. This mirrors the default to uncertainty relationship, as seen in Figure 8.3(c). The statistical analysis of grid cells in the 16.3 and 19.2 ROI was not carried out because these predictions only occurred in 1 and 8 cells respectively.

The number of grid cells in Group 5 (Figure 8.10) is lower than Group 4, so conclusions are even more difficult to make. However, it is still possible to see some agreement in the mean prediction to uncertainty relationship observed in Groups 2, 3 and 4. However, generally the uncertainty/mean ratio are lower than in the Group 4. This suggests that uncertainty in the climatic inputs has less influence on the validity of the percentile calculated for the grid cells in Group 5, especially with respect to the relative uncertainty (uncertainty/mean at a grid cell).

As discussed in Section 5.1.1, the BIOCLIM prediction result is determined by the known crop point sites, the climate at those points and the climatic similarity of these sites with the rest area of interest. As temperature varies significantly from the south to north of the region studied, it can be concluded that rainfall is the dominant climatic variable in these results. Therefore the highest prediction occurs in grid cells that are most similar in rainfall patterns, to where the Field Pea trial sites are. As discussed in Section 8.1 and shown in Table 8.1, the introduction of uncertainty into the model's climate inputs does change the prediction to some, mostly small extent. Therefore, it can be concluded that the model's prediction is most sensitive to geographical related changes in the climate and field sites and not the uncertainty in the climate grids. However, the results of this section clearly show that the uncertainty of the prediction can be significantly influenced by propagation of the climate grid uncertainty, with very rapid changes occurring quite abruptly across geographical space.

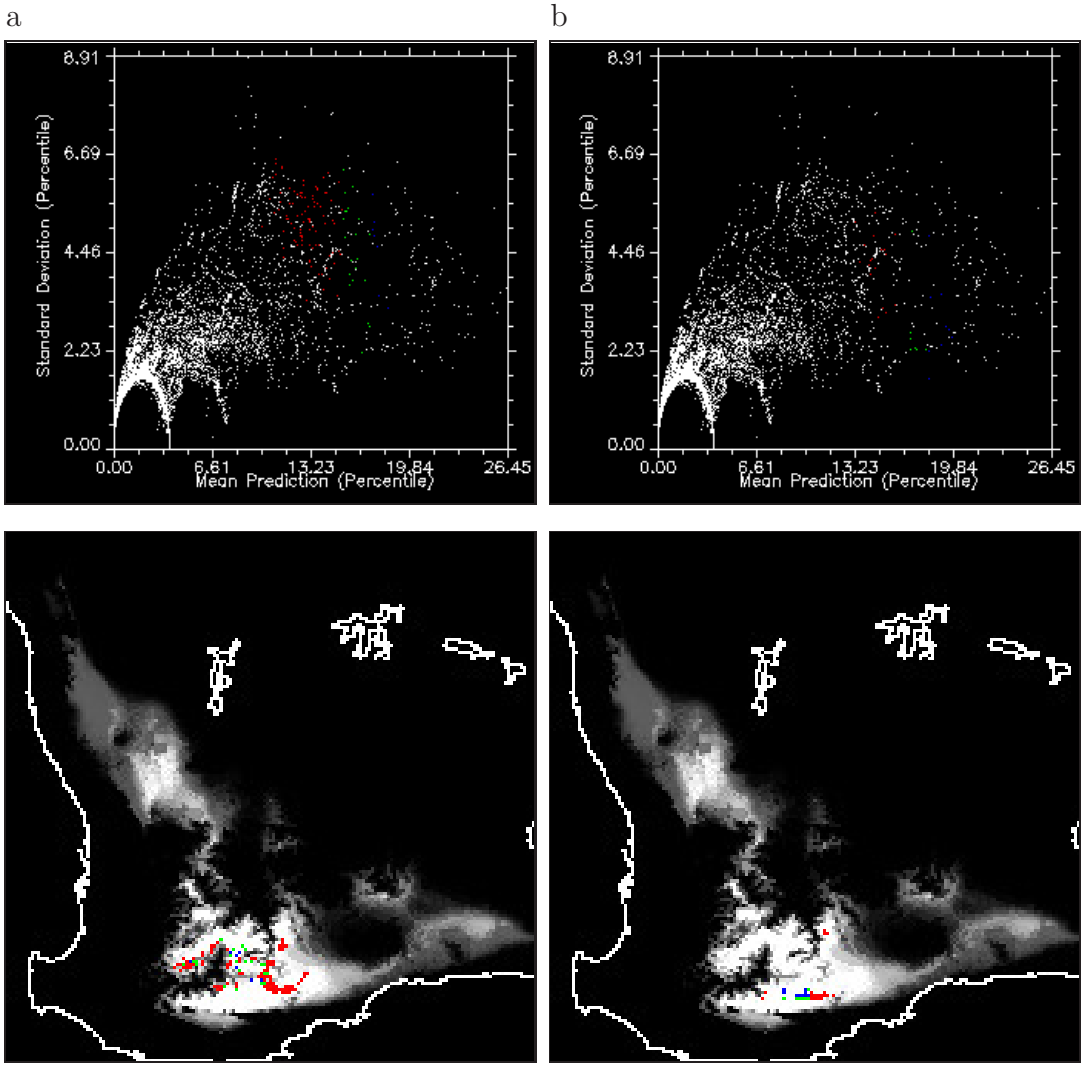


Figure 8.9: Mean versus Uncertainty Relationship in Group 4: (a) 17.3 and (b) 18.3. For ROI colour classifications see categories described on Page 129.

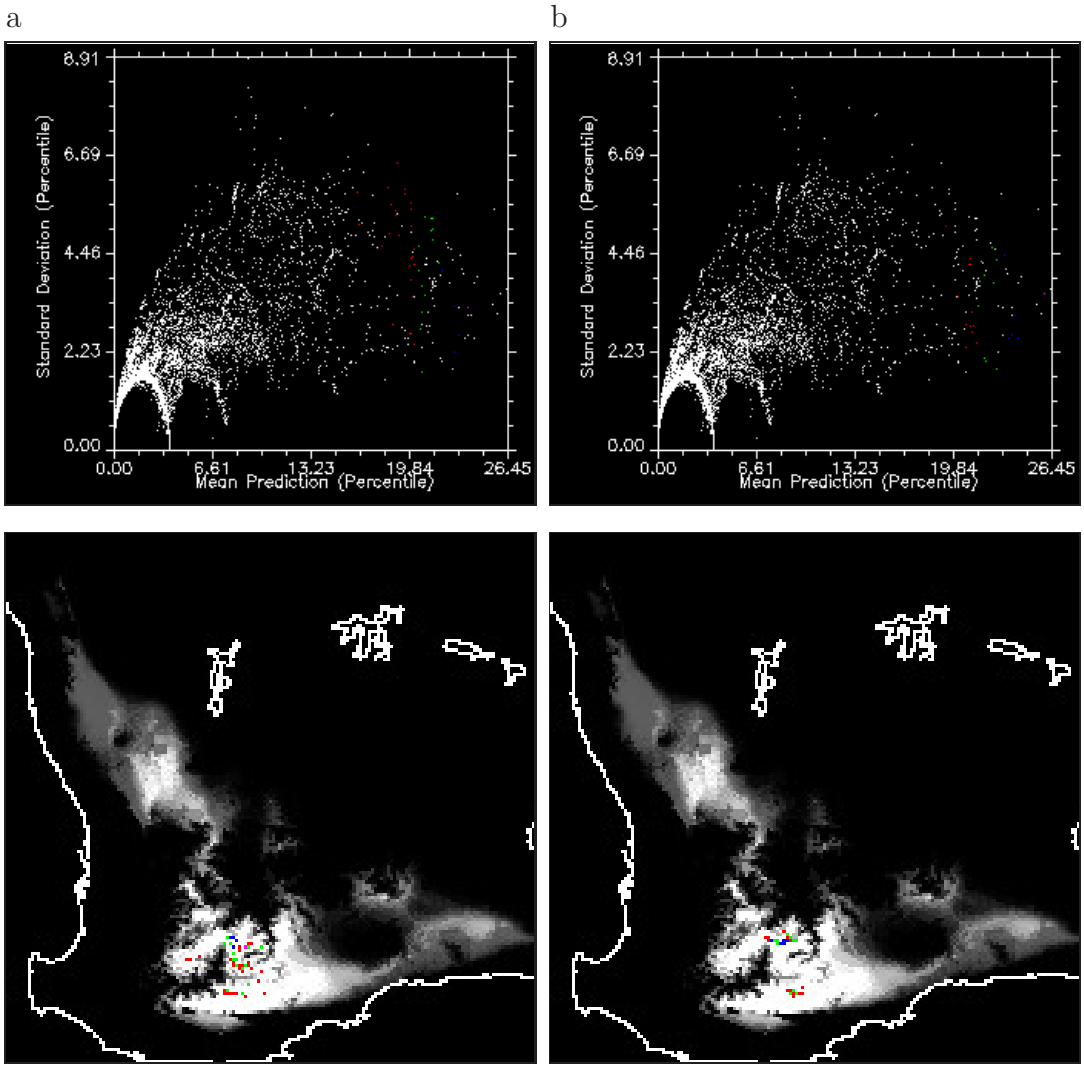


Figure 8.10: Mean versus Uncertainty Relationship in Group 5: (a) 22.1 and (b) 23.1. For ROI colour classifications see categories described on Page 129.

8.2 Skewed Uncertainty Distributions

As discussed in Chapter 4, a skewed uncertainty distribution can change a precision agriculture models output and the uncertainty of that output. This analysis was repeated on the BIOCLIM model and the results discussed in this section.

8.2.1 Quantified Similarities and Differences

The mean prediction and its associated uncertainty, when the added GHCN uncertainty layers are non-normal, is displayed in Figure 8.11 (the Worldclim uncertainty layer is not included in this analysis due to its small influence). As is clear, the mean prediction results are lower than the default prediction irrespective of the skew direction, but still in high agreement with each other: Positively and negatively skewed GHCN uncertainty - correlation of 0.9921 (see Table 8.4); Correlation of the predictions made with normally distribution to positively and negatively skewed GHCN uncertainty - 0.9934 and 0.9917 respectively. The high values of these correlations suggest that the predictions are very close, irrespective of the direction of the skew or whether the uncertainty is skewed or not.

Similarly, the calculated statistical relationship in the uncertainty results is very close: normally distributed to positively skewed GHCN uncertainty - correlation of 0.9555, normally distributed to negatively skewed GHCN uncertainty - correlation of 0.9441, positively and negatively skewed GHCN uncertainty - correlation of 0.9560. As expected, in all cases the correlation between the uncertainty results is less than between the mean results, but they are still high. These correlation results are quite easily visible in the uncertainty images in Figure 8.11, with differences in uncertainty being most easily visible where the prediction is low (< 2.5 Percentile) or in areas where the prediction is generally higher (> 5 Percentile).

As previously discussed, the degree of sensitivity of the model varies across the region studied, but the general pattern(s) observed are consistent whether the uncertainty inputs are skewed or normal. As before, the clearest examples

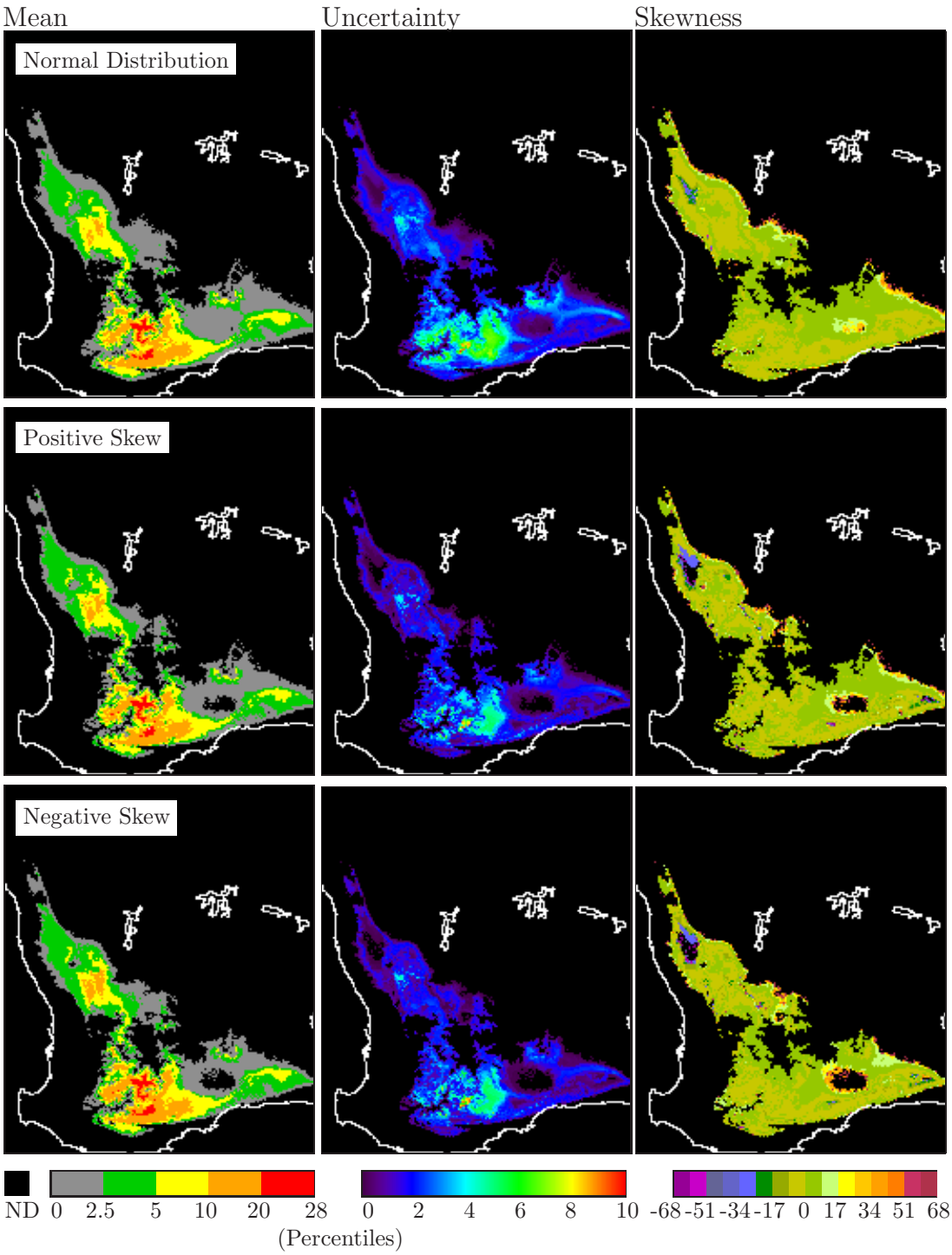


Figure 8.11: BIOCLIM present prediction results for normal vs skewed uncertainty (GHCN only): (a) Mean, (b) Uncertainty and (c) Skew. Middle row, GHCN and WorldClim uncertainty added. Top - Gaussian, Middle gamma +2, Bottom gamma −2.

	GHCN, World (normal)	GHCN (normal)	GHCN (skew +2)	GHCN (skew -2)
Mean				
GHCN, World (normal)	1.000000	0.999822	0.992934	0.991211
GHCN (normal)	0.999822	1.000000	0.993405	0.991742
GHCN (skew +2)	0.992934	0.993405	1.000000	0.992105
GHCN (skew -2)	0.991211	0.991742	0.992105	1.000000
Uncertainty				
GHCN, World (normal)	1.000000	0.998885	0.952304	0.941274
GHCN (normal)	0.998885	1.000000	0.955554	0.944184
GHCN (skew +2)	0.952304	0.955554	1.000000	0.955979
GHCN (skew -2)	0.941274	0.944184	0.955979	1.000000
Skew				
GHCN, World (normal)	1.000000	0.981726	0.714235	0.732270
GHCN (normal)	0.981726	1.000000	0.731408	0.727543
GHCN (skew +2)	0.714235	0.731408	1.000000	0.529652
GHCN (skew -2)	0.732270	0.727543	0.529652	1.000000

Table 8.4: Correlation of Present Prediction results.

of this are where both the mean and uncertainty are low, but the uncertainty relative to the mean is high (and therefore important); and where the uncertainty is high (but lower than the mean) and spatially quite variable (such as in the "south western" region. Also of note is the area centred at $\approx 118^o13'$ E. $33^o40'$ S. (see Section 8.1) has a lower uncertainty, but is still clear in its spatial extent. These results suggest:

1. That the uncertainty is generally lower when the GHCN uncertainty inputs are skewed.

2. Similar to the results in Section 8.1, there are some regions where the uncertainty is consistent, but across the whole region this generality can not be made.

3. There is a higher correlation between uncertainty results for the normal-uncertainty and positively-skewed uncertainty input BIOCLIM prediction. This is not easily visible in the uncertainty result images (Figure 8.11), but is in the plotted results, as the mean to uncertainty relationship is slightly more linear and more consistent (Figure 8.12(a) and (b)).

Knowing how the BIOCLIM model calculates its predictions, it can be generalised that heavily skewed non-normal uncertainty will reduce the width of the climate input layer uncertainty distribution of each grid cell. In turn, this favours minimising the uncertainty in the prediction at each grid cell, regardless of the direction of the skew (see Figure 8.12). This relationship reflects the conclusion drawn from the analysis of the R component of the Mitscherlich Equation, as discussed in Chapter 4.2, even though the models are quite different.

Analysis by Region of Interest

The mapped mean to uncertainty relationship, when the input uncertainty is normally distributed, was discussed in Chapter 8.1.1 and 8.1.2. This section applies the same ROI methodology to investigate how non-normally distributed uncertainty changes this relationship.

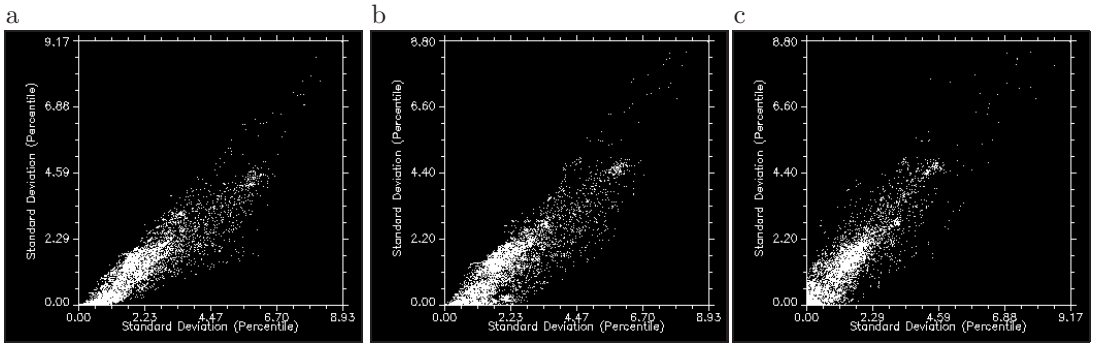


Figure 8.12: Uncertainty of Present Prediction in the Western Australian Region. Uncertainty Inputs: (a) Normal versus Positive Skew, (b) Normal versus Negative Skew and (c) Positive versus Negative Skew

As seen in Figures 8.13, 8.14 and 8.15, changing the skew in the inputs uncertainty does significantly change where the lowest uncertainty occurs. The mean - uncertainty relationship does mirror what occurs when the uncertainty in the input has a Gaussian distribution (generally lower when the mean is less than the ROI value and higher in the opposite case). However, when the input uncertainty is negatively skewed, the higher mean values can have a lower associated uncertainty. This is most clear in the 2.9 ROI (the red region), but is also visible in the 9.6 ROI in Figure 8.14 (Group 3, yellow region). The influence of non normality on the inputs is most clear in Group 2 and 3. To summarise,

1. The $\surd$ shaped prediction to uncertainty relationship is still clearly present. However, more of the ROI (and hence more grid cells) have uncertainties close to zero.

2. In Group 2, when the distribution is normal, the lowest uncertainty clearly occurs in the 3.8 ROI (see Figure 8.5). Also, the 2.9 ROI is not distinguishable (in the plot), as the uncertainty for the 2.9 ROI is high. However, when the input uncertainty is skewed the 2.9 ROI becomes easily seen as the uncertainty, for part of this ROI, is lower.

3. In group 3, the uncertainty is clearly lower, with the 6.6, 7.7 and 9.6 ROI having uncertainty values close to 0.

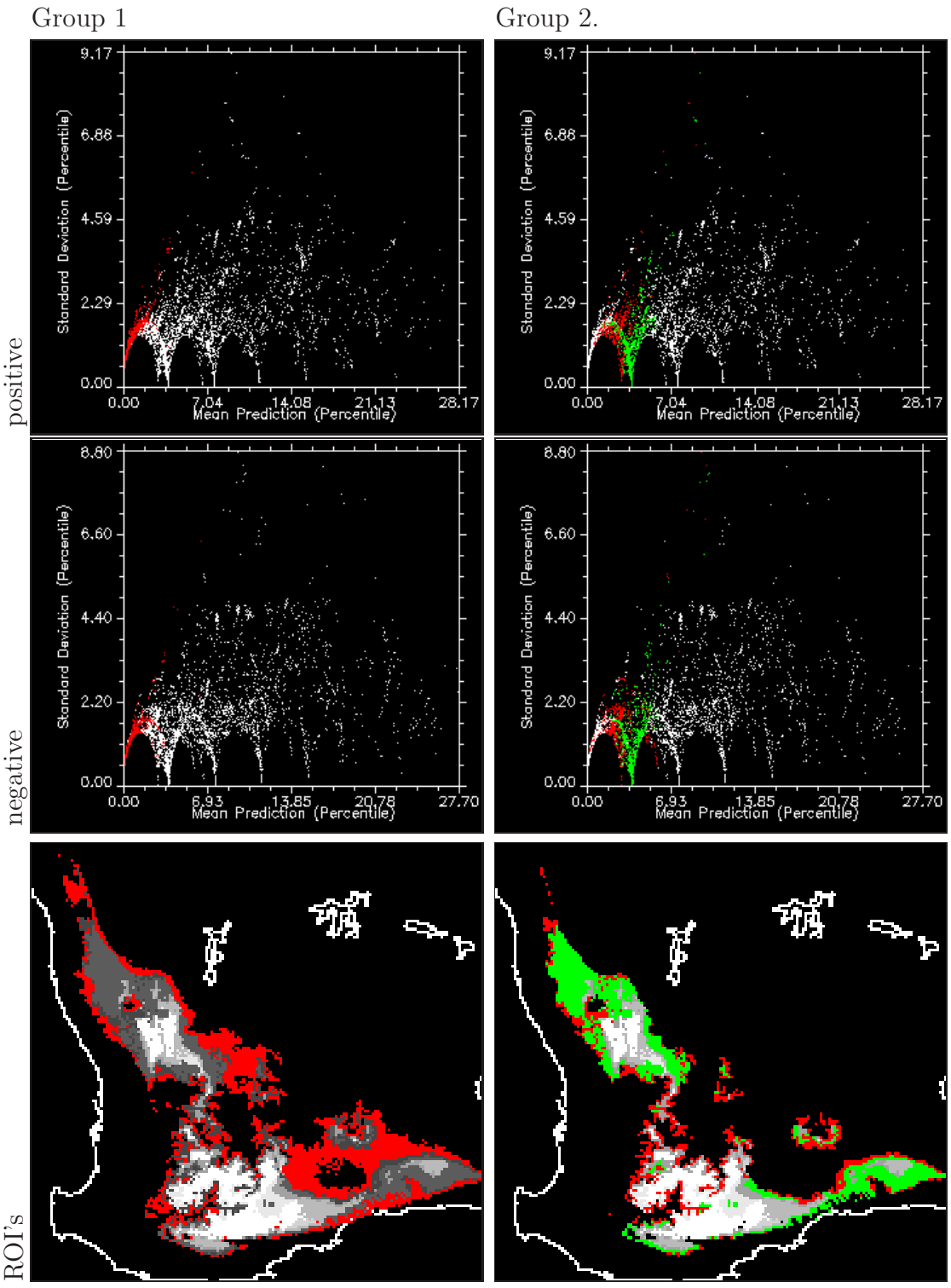


Figure 8.13: Mean versus Uncertainty in Group 1 and 2, when uncertainty input is positively and negatively skewed. For ROI colour classifications see Table 8.5

This clearly shows that a lower uncertainty in the model's output occurs when the uncertainty in the climate grids is skewed and that this occurs in several ROI. It also shows that the ROI with the lowest uncertainty can be significantly influenced by the shape of the uncertainty distribution in the climate inputs.

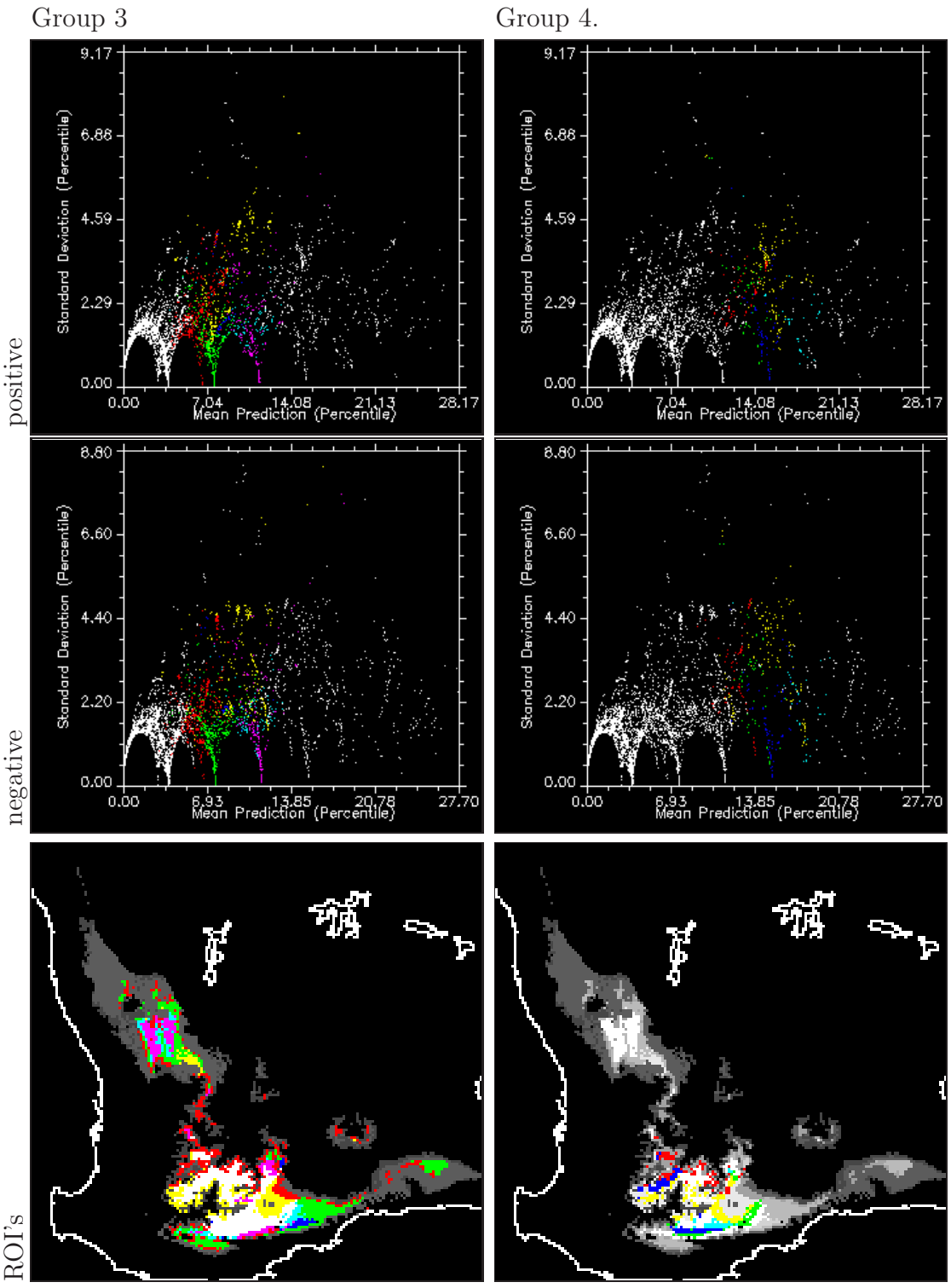


Figure 8.14: Mean versus Uncertainty in Group 3 and 4, when uncertainty input is positively and negatively skewed. For ROI colour classifications see Table 8.5

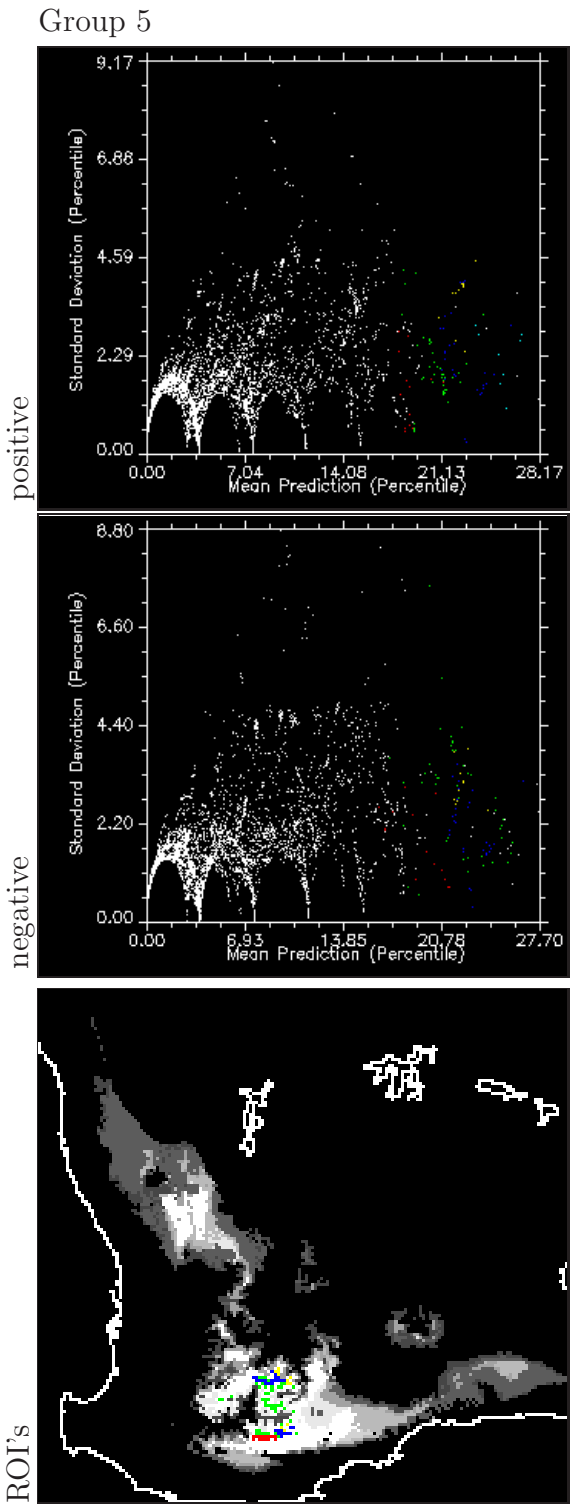


Figure 8.15: Mean versus Uncertainty in Group 5, when uncertainty input is positively and negatively skewed. For ROI colour classifications see Table 8.5

Group	ROI	Colour	No of points
1	0	red	2126 and 2079 (+/- skew)
2	2.9	red	867
	3.8	green	1280
3	6.7	red	433
	7.7	green	381
	8.7	blue	31
	9.6	yellow	246
	10.6	cyan	100
	11.5	majenta	171
4	13.5	red	77
	14.4	green	56
	15.4	blue	77
	16.3	–	1
	17.3	yellow	124
	18.3	cyan	34
	19.2	–	8
5	21.2	red	14
	22.1	green	48
	23.1	blue	29
	24.0	yellow	11
	25.0	–	2
	26.0	–	8
6	28.8	–	2

Table 8.5: Colour labels for the ROI in each group.

8.3 Influence of Uncertainty on Future Predictions

As discussed in Section 6.3, the complexity of the uncertainty in future climate predictions is not easily quantified and so has been excluded from the future prediction BIOCLIM model. Therefore, as discussed in Section 7.4, in the future prediction model uncertainty is present only in the present climate grids, the Bioclimate layers, the tailed percentile distributions and the prediction. This, is illustrated in Figure 7.6 and 8.16. In both figures the uncertainty propagation pathways through the model is shown and, as expected, it does not include the calculation of the future Bioclimate grids.

The areas predicted to be suitable for Field Pea in both present and the two CSIRO (A2a and B2a) future climates are illustrated in Figure 8.17. The climate grids are static i.e no uncertainty is present and will be referred to as default-A2a

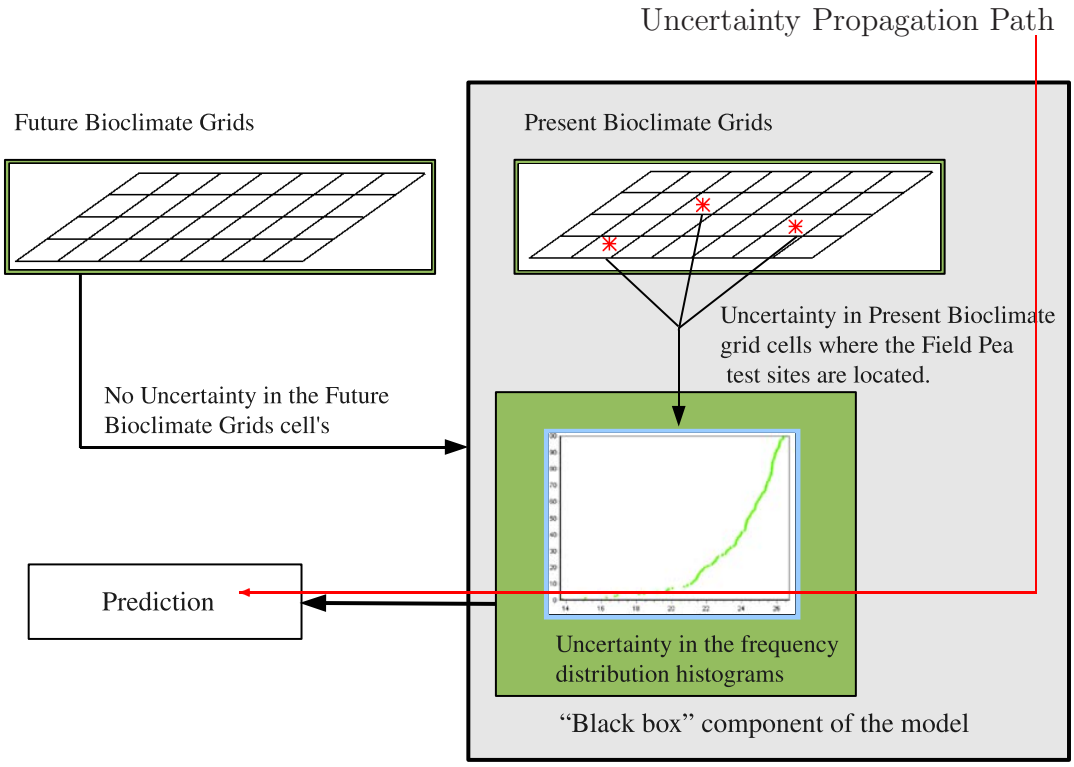


Figure 8.16: Sections of the future model where uncertainty is present and is not present. For the former, the components are in the grey boxed area and the prediction. The green boxed area is the "black box" component of the model. The uncertainty propagation path is shown.

and default-B2a. In both these future scenarios the area predicted as suitable is considerably less and the maximum percentile calculated is $\leq$ 20 Percentile. The closer agreement between the future predictions, at each grid cell, is seen in the statistical correlation between each (present−A2a: 0.4257, present−B2a: 0.3027, A2a−B2a: 0.7956). The spatial distribution of all grid cells with non null values, for both future scenarios, are also in closer agreement. However, in all three cases there are clear similarities, the most notable being the area where the highest prediction occurs i.e. in the "south western" region.

When compared to a present prediction, where the reduction in productive area has occurred is not unexpected (due to the predicted decrease in precipitation across the south of Western Australia). However, there is one clear anomaly

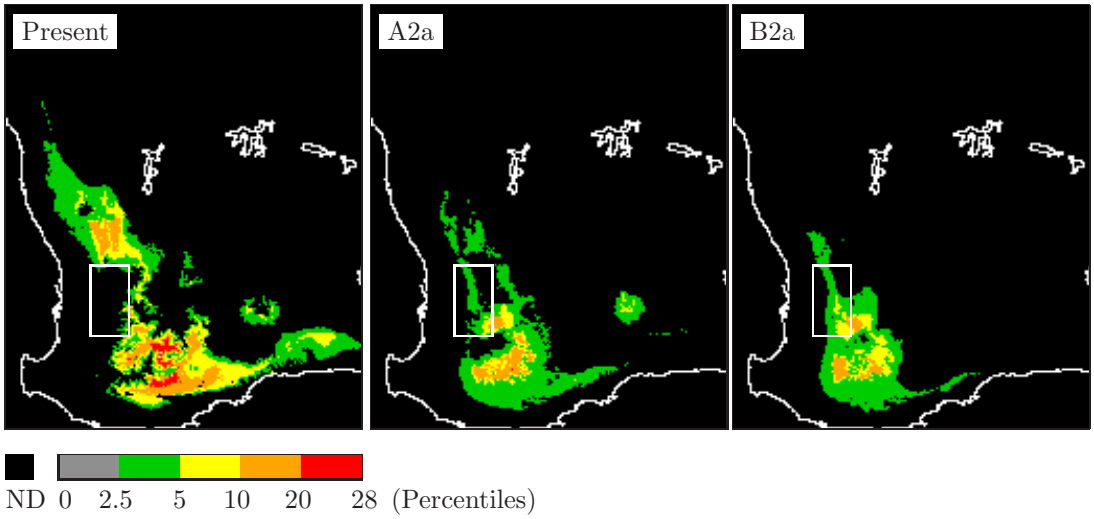


Figure 8.17: Default BIOCLIM predictions: Present, CSIRO A2a and B2a models. The white boxed area, at $\approx 116^o$ E. 32^o S., is approximately where productivity is greater than 0 in future predictions, but not in the present prediction.

in this pattern along the western boundary (centred at $\approx 116^o$ E. $32^o0'$ S., see Figure 8.17) that, in the future, is considered to be more suitable than at present. For a user of the BIOCLIM model to understand why this may be valid, would require a closer look at the future climate scenario in that area and an understanding of the conditions favourable to the Field Pea.

8.3.1 Future: CSIRO A2a and B2a Scenarios with Normally Distributed Uncertainty

As in the present scenario for the Western Australia region, the future predictions are clearly boxed at values ≥ 2.5 and < 20 Percentiles (except for one grid cell in A2a, see Figure 8.17). But, as discussed in Section 8.1, when uncertainty is added to the climate inputs, the prediction become less well defined (Figure 8.18 and 8.19 (Mean)). Also, there is a close agreement between outputs when the normal-distribution Worldclim uncertainty layer is present or excluded (correlation, A2a: 0.99997, B2a: 0.99998). Therefore, all following discussion and conclusions refer to the situation in only the GHCN uncertainty layer is added to the BIOCLIM

climate inputs.

When uncertainty is added, there is a considerable increase in the spatial extent of the area classified as $\leq$ 2.5 Percentile. This is most notable for the A2a future prediction, where the "grey" areas has significantly extended into areas that are graded as unsuitable (black) or $\geq$ 2.5 to $\leq$ 5 (green) when no uncertainty is added. The same does occur for the B2a future prediction, but to a much lesser extent. However, in the B2a future, the highest predictions cover a larger area at $\approx 117^o5'0''$ E. $32^o52'30''$ S and $\approx 117^o47'30''$ E. $33^o47'30''$ S (red rectangled area in Figure 8.19, B2a mean prediction map). This suggests that the propagation of present climate input uncertainty has a differing influence on the future predictions, depending on which of the future climate inputs is input (to the model). This difference is clear in the changes in both the prediction values and their spatial extent.

A summary of the uncertainty associated with the future predictions is:

1. The highest uncertainty does occur where the higher prediction occurs, most notably in the B2a model. But, there is also an almost equally high uncertainty where the predictions of $\approx$ 2 percentile occur.

2. The highest uncertainties of the simulated future prediction(s) are less than one third the uncertainty associated with the present prediction. For both scenarios (A2a and B2a), the lowest values ($\leq$ 0.5) are on the north east border region (where the prediction is also lowest), but also occurs in other low prediction border regions.

3. In the south there is a border region where the uncertainty is low, but the prediction here is consistently in the range of 2.5 to 5 Percentile. Heading south or south west, this prediction quickly drops to 0, with associated uncertainty as high as 2 Percentile (at various grid locations along this border).

4. In the south west of both future predicted regions (A2a and B2a), there is a "strip" and other grid locations in which the uncertainty is equal to 0, even

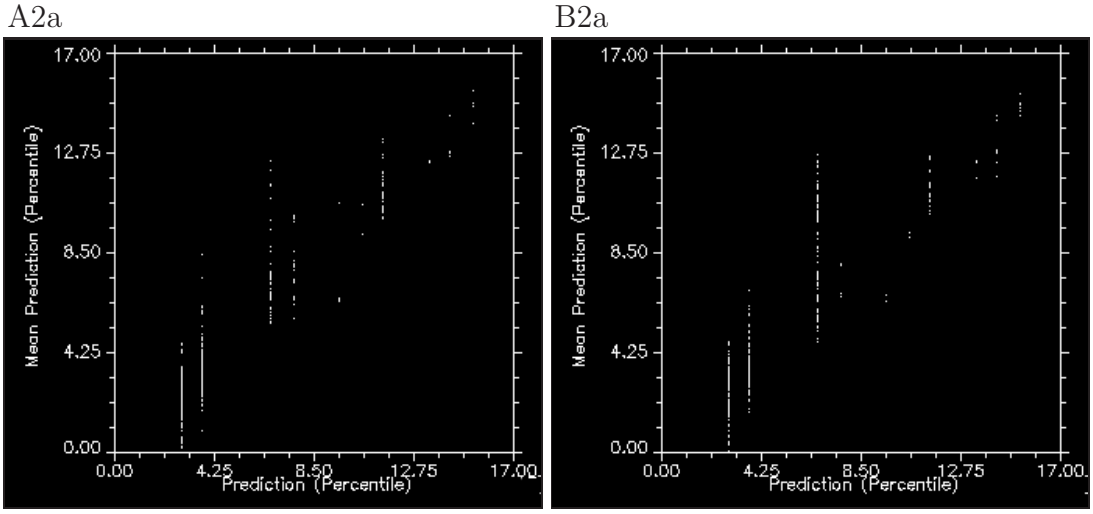


Figure 8.18: Comparison of CSIRO A2a and B2a Predictions, with and without added uncertainty.

though the prediction is as high as 5 Percentile. This clearly indicates that the input uncertainty (in the present climate surfaces) have no influence on the prediction in this sub region.

5. There are areas where the prediction is very low, but the uncertainty is high. This indicates that the model is sensitive to uncertainty in these regions. Where these sensitive sub regions occur is not neccesariy the same for the A2a or B2a future scenarios.

6. As in the present predictions results discussed in Chapter 8.1, the highest uncertainty mostly occurs where the highest predictions occur (see Figure 8.20), most notably in the B2a model. But, there is also an almost equally high uncertainty where the predictions of ≈ 2 percentile occur.

7. Finally, the $\surd$ shaped relationship between the prediction and uncertainty is clearly present.

For both of the future scenarios, the uncertainty to mean relationship appears to follow that observed in the present prediction (8.21). In the A2a and B2a scenario, this is most clear up to the Prediction of ≈ 4.5 and ≈ 3.5 respectively

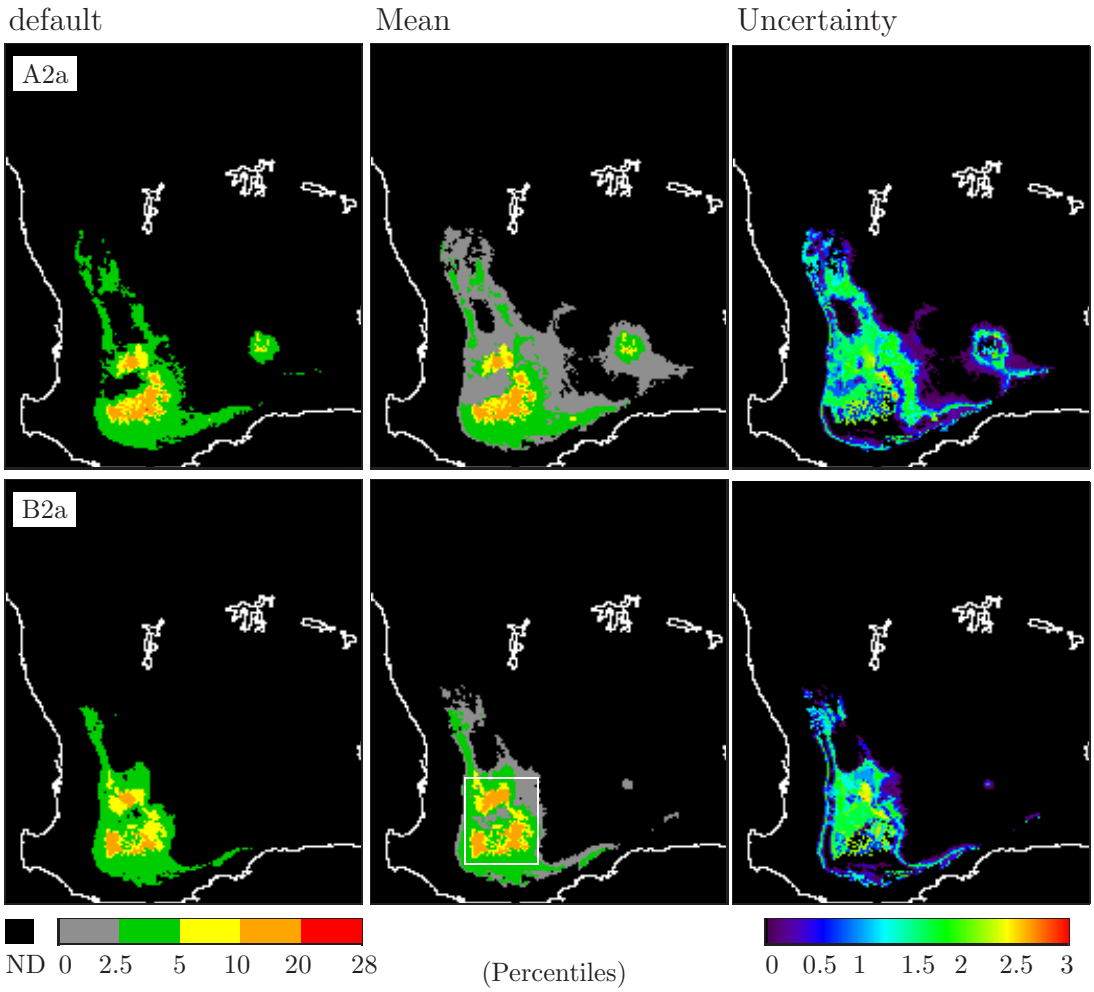


Figure 8.19: BIOCLIM Future Prediction CSIRO A2a and B2a. Default (no uncertainty in present climate grids). Mean and Uncertainty (with uncertainty). The white boxed area in the B2a mean map shows an area where the prediction has increased, as discussed in Section 8.3.1.

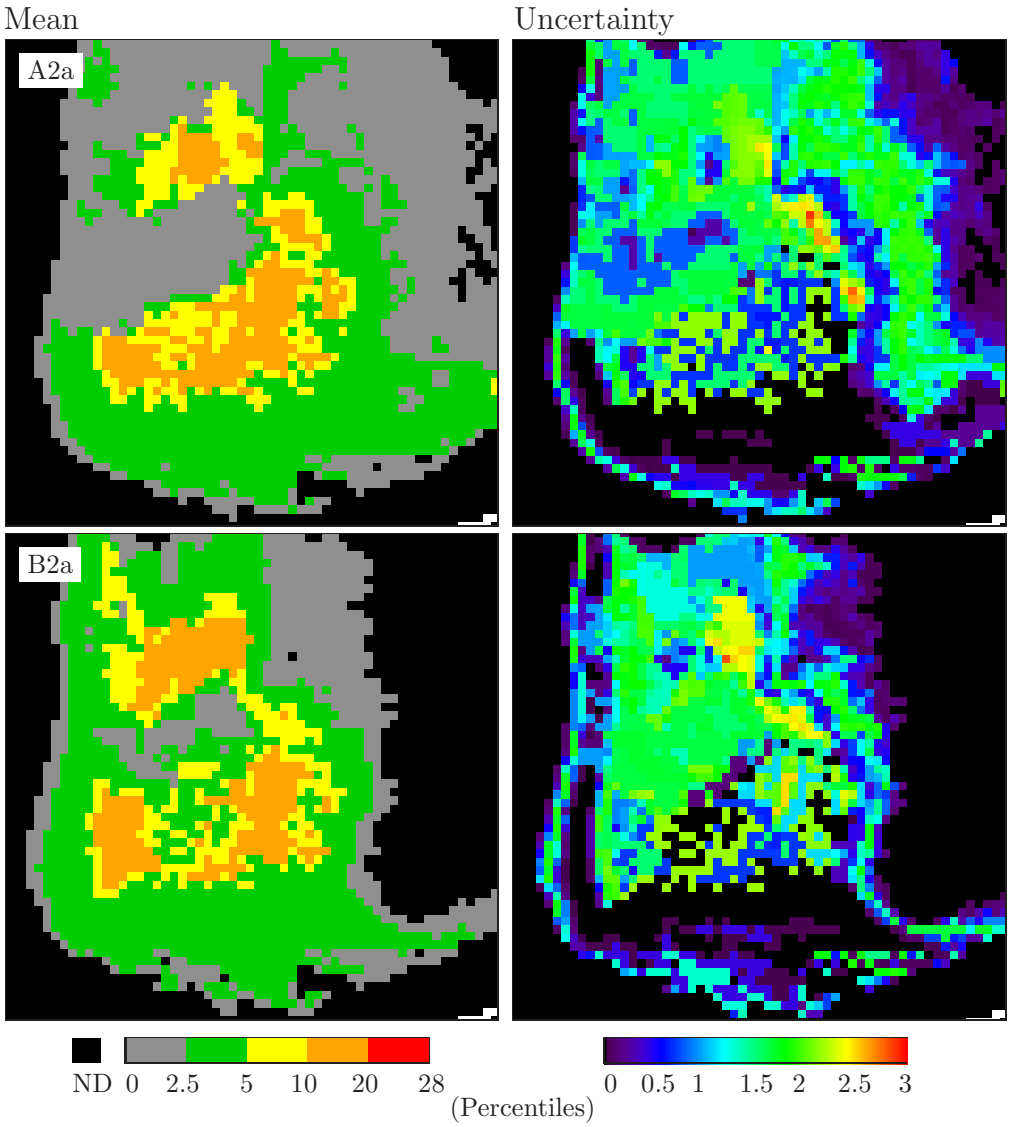


Figure 8.20: Mean and Uncertainty, BIOCLIM Future CSIRO A2a and B2a model predictions.

(coloured red and blue respectively), but less clear above those values (mirroring the higher prediction pattern in the present scenario).

Also as occurs in the present prediction, the highest uncertainty occurs where the highest prediction occurs. However, as previously discussed, the maximum uncertainty reached is considerably lower. Also notable is the relative lack of a "cloud" of uncertainty values above the mean versus uncertainty arcs.

8.3.2 Future: CSIRO A2a and B2a Scenarios with Skewed Uncertainty

A non-Gaussian distribution of the BIOCLIM climate inputs produces two consistent patterns in the predictions. More specifically, when compared to the Gaussian prediction, for both A2a and B2a, a positive skew in the uncertainty produces a prediction which is in slightly higher agreement with the default future prediction. The opposite is the case for B2a. This can be seen in Figure 8.22 and in the correlation between the mean and default predictions (see Table 8.6). In short, the greatest difference from the default predictions, for both future scenarios, occurs when the uncertainty in climate inputs are negatively skewed, with the greatest agreement occurring when the inputs are positively skewed (but there is not a large difference between uncertainty in the Gaussian model and positively skewed model).

The mean and default prediction results, when the inputs are positively skewed, is shown in Figure 8.23. In these plots, the closer the agreement between predictions, the lower the variation in the predicted results. As expected from the correlation results, this is most clearly seen in the positively skewed A2a

	normal	+ skew	- skew
A2a	0.976619	0.980089	0.947932
B2a	0.965691	0.977722	0.941394

Table 8.6: Correlation of Future Predictions, default (no uncertainty) to Simulated Inputs (GHCN uncertainty added)

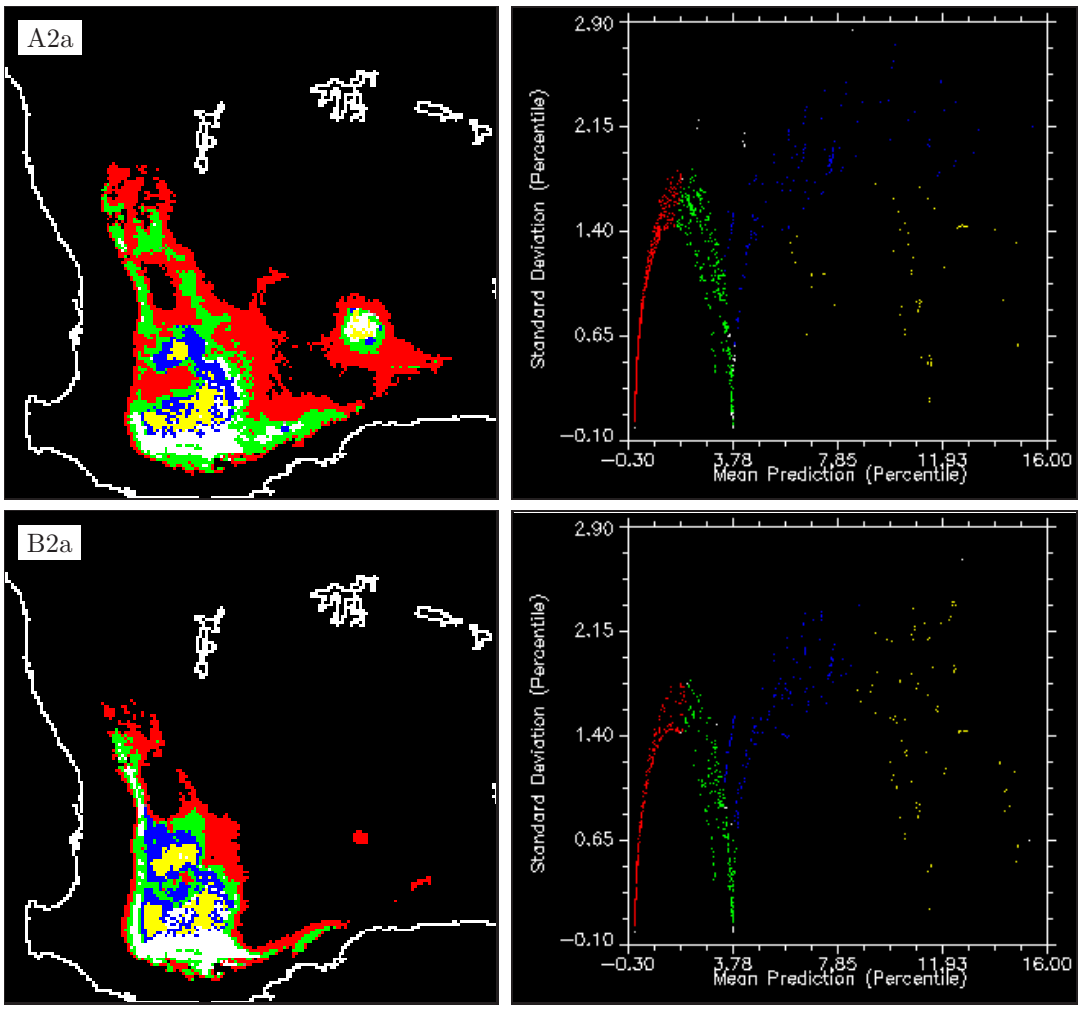


Figure 8.21: Mean versus Uncertainty. A2a and B2a Future Predictions, GHCN Gaussian Uncertainty distribution.

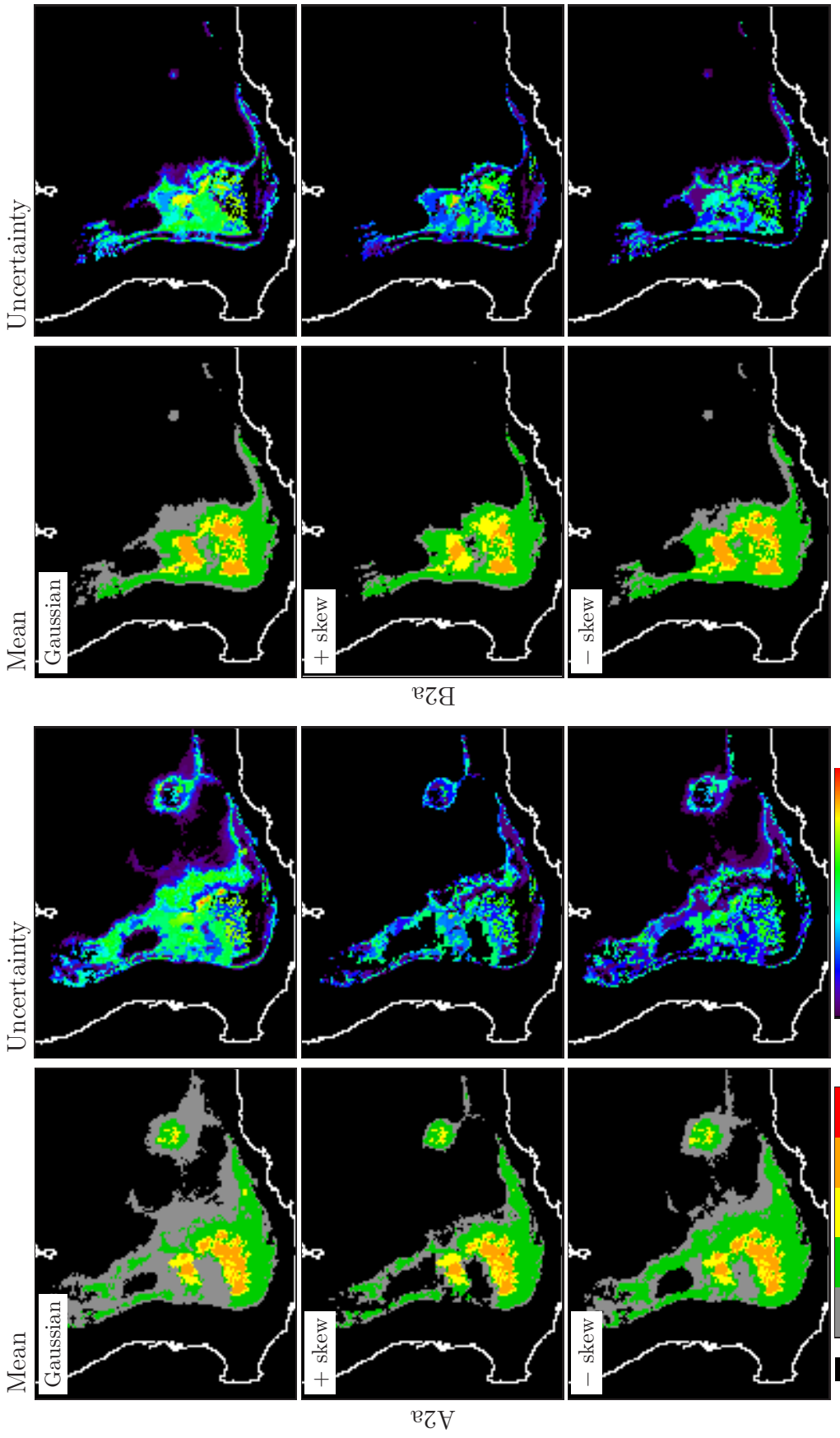


Figure 8.22: BIOCLIM future CSIRO A2a and B2a Predictions: Gaussian, positively and negatively skewed uncertainty distributions.

prediction, which is more clearly linear and has a higher maximum prediction. But, the variation does appear significantly higher than when the same analysis was applied to the present predictions (Section 8.2). Therefore, for future predictions, the input uncertainty's distribution (per grid cell) has a stronger influence on the value of the prediction made than on the uncertainty of this prediction.

As expected, the levels of agreement between the prediction layers is reflected in the uncertainty results (Figure 8.22 and Table 8.7). This is most notable in the A2a prediction (correlation for A2a: normal - positive skew, 0.7959; normal - negative skew, 0.8451; positive to negative skew, 0.5420), where the "positive skew" result clearly covers less of a spatial area and this area is close to the area covered by the default prediction. In general, the spatial patterns are reasonably consistent (such as the lower uncertainty on the north eastern border and the high uncertainty in areas where the prediction is higher). As expected, the same is observed when the inputs are positively skewed in the future B2a prediction, with the spatial area it covers being less.

The mean - uncertainty relationship in these scenarios is shown in Figure 8.24 and appears to follow the same "arc" like relationship observed in other predictions. Similar to the uncertainty of the normally distributed future predictions, the size of the uncertainty is much less and does not have the "cloud" of higher uncertainty.

8.4 Summary and Conclusions

A summary of BIOCLIM's sensitivity to uncertainty in the climate inputs is as follows:

Present predictions

Spatially, the uncertainty in the BIOCLIM prediction varies considerably, as does the prediction. The uncertainty does not have a clear relationship with the prediction, either spatially or with the predictions size. At some grid points, the uncertainty in the prediction can be very high relative to prediction, clearly

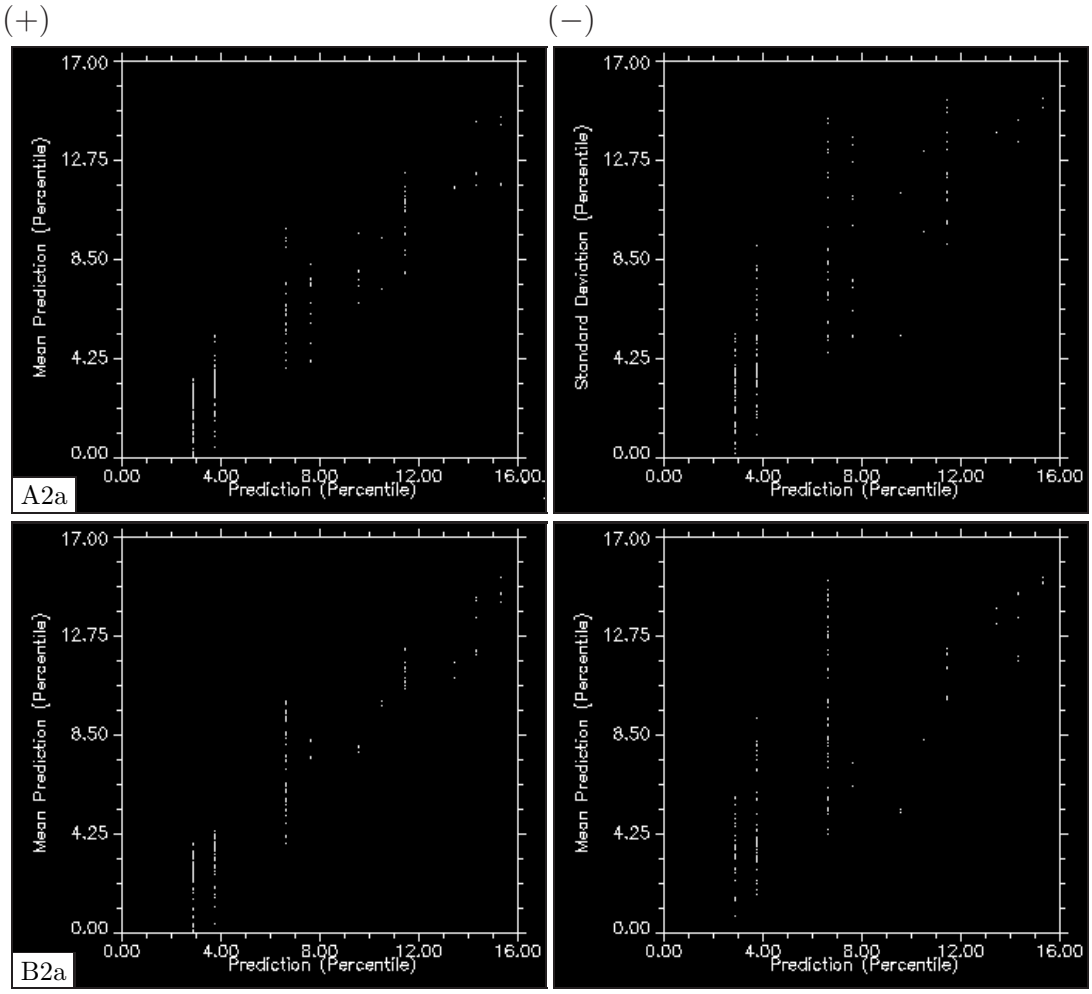


Figure 8.23: A comparison of Default and A2a or B2a Future Predictions. Positive and negative skewed uncertainty inputs.

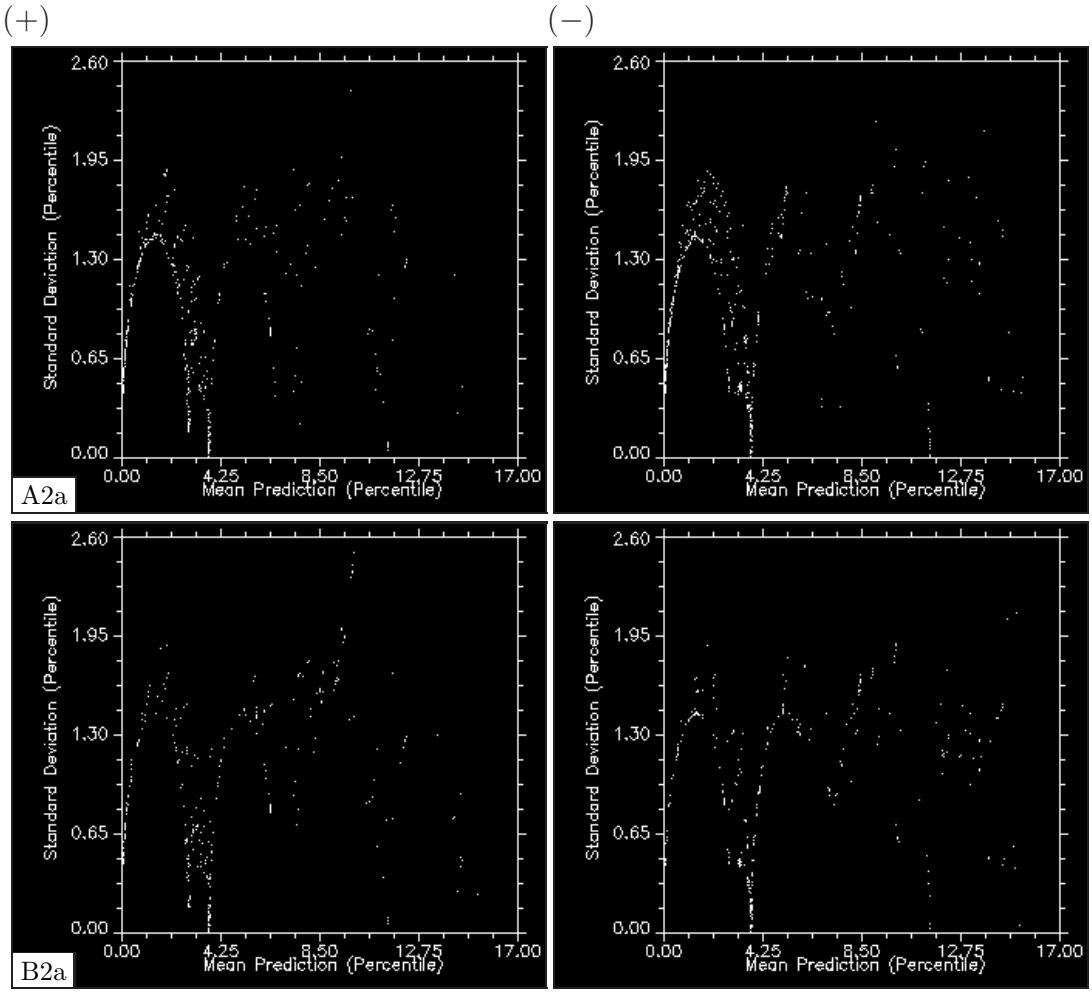


Figure 8.24: Mean Prediction versus Uncertainty, A2a and B2a future predictions. Positively and negatively skewed uncertainty.

	GHCN (normal)	GHCN (skew +2)	GHCN (skew -2)
Present			
GHCN (normal)	1.000000	0.955554	0.944184
GHCN (skew +2)	0.955554	1.000000	0.955979
GHCN (skew -2)	0.944184	0.955979	1.000000
Future A2a			
GHCN (normal)	1.000000	0.795861	0.845098
GHCN (skew +2)	0.795861	1.000000	0.541974
GHCN (skew -2)	0.845098	0.541974	1.000000
Future B2a			
GHCN (normal)	1.000000	0.874284	0.895760
GHCN (skew +2)	0.874284	1.000000	0.707967
GHCN (skew -2)	0.895760	0.707967	1.000000

Table 8.7: Correlation of Uncertainty: Future Predictions, Normal and Skewed Input Uncertainty.

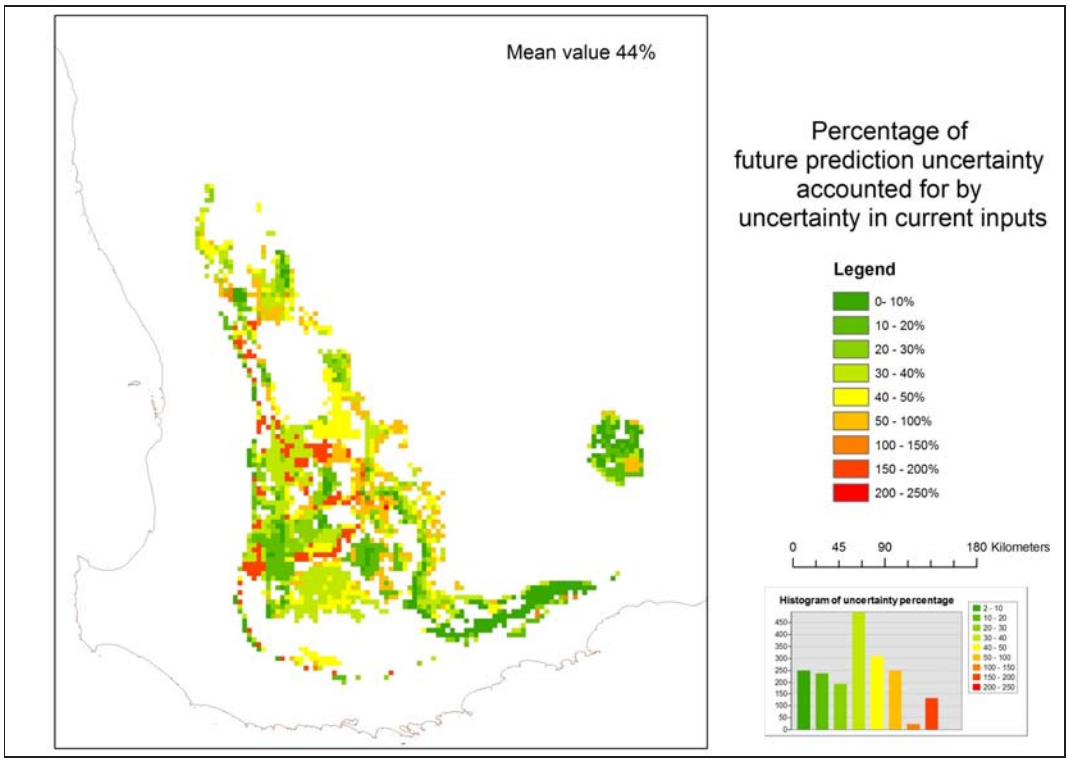


Figure 8.25: Percentage of future uncertainty accounted for by uncertainty in current inputs, as discussed by Corner and Marinelli (2008).

showing the low validity of the predictions. On the other hand, at grid points relatively close to these, the uncertainty can drop significantly as the prediction increases. This is most likely due to predictions crossing a threshold after which the uncertainty is significantly less. It seems reasonable to suggest that this is mostly caused by BIOCLIM's boxing of predictions, as the changes in the uncertainty grids are relatively smooth, so variation in them alone can not explain this observation. However, relative to the prediction, the highest uncertainty occurs in grid cells where the prediction is lowest, such as the north east boundary.

Skewed non-normal uncertainty reduces the uncertainty in the prediction as well as where this occurs. Also, the uncertainty in the prediction at each grid cell is reduced, regardless of the direction of the skew.

Future predictions

All the future predictions fall into a narrower range, 0 to 20^{th} Percentile. This is expected given the reduced rainfall predicted for the south of Western Australia.

Uncertainty in the present climate grids of the future models does change the prediction, but not to the same extent as occurs in the present climate model. This suggests that the predictions of the grid cells in the simulated bioclimate layers fall mostly within a narrower or more evenly distributed range.

The mean versus uncertainty relationship is similar to what occurs in the present BIOCLIM model. More specifically, (1) the $\surd$ mean versus uncertainty relationship, (2) the size of this uncertainty can be significant relative to the size of the prediction and (3) the size of the uncertainty appears to change randomly in many areas while remaining constant across others.

The difference between the two future climate predictions is significant and easily exceeds any difference that occurs when uncertainty is added to the present condition climate grids. Therefore, if BIOCLIM is used to predict the ecological niches for a range of different future climate conditions, the accuracy of the model used to calculate the future climate grids is more critical.

This final point does not suggest that uncertainty in the present condition's

inputs is not important if comparing the two future condition predictions, only that it may be less important depending on what an end user is interested in knowing. For example, Figure 8.25 illustrates the percentage of the difference between the two future predictions that can be accounted for by the uncertainty in the present climate condition grids (Corner & Marinelli 2008). For the majority of the prediction it is under 50% (which may be condidered low) and in some areas very high. Therefore, the difference between the two future scenarios is large and so minimising the uncertainty in the input climate grids is not going to improve this to a significant degree. However, as future climate predictions continue to converge with the development of future climate models, the importance of quantifying climate grid uncertainty propagation becomes important.

The similarity in the mean - uncertainty relationship in all the scenarios discussed suggests that the reason for this is mostly related to the models structure. Therefore, further investigation of why the observed relationships are occurring will be only on the present prediction, normally distributed uncertainty (Chapter 9)

Chapter 9

Uncertainty Propagation Path Analysis

As discussed in Chapter 8, the presence of a normally (or Gaussian) distributed uncertainty within the temperature and precipitation grids has a strong influence on the uncertainty of the *present* BIOCLIM Model prediction (hereafter this type of model is referred to as the *GaussianPresent* model). The uncertainty − mean relationship varies spatially, is clearly more sensitive at the borders between the default prediction "regions of interest" and has some characteristic patterns (such as the skewed $\surd$ shaped decrease $\rightarrow$ increase in uncertainty, as the prediction value increases). As discussed in Section 5.3.1, the total BIOCLIM model has three sections, which (1) calculates the the bioclimate grids and then (2) calculates the frequency distribution of the bioclimate values at the grid cells where the species are known to occur. These distributions are then used to (3) calculate where the species should occur in the other grid cells being studied.

As the frequency distribution and hence the final prediction is a result of the values in the Bioclimate grids, it is clear that understanding how uncertainty in the Bioclimate grids can influence a prediction is important. For example, if there is high variability in the precipitation of the warmest months, this is not intuitively expected to influence the prediction of crops not grown in this period

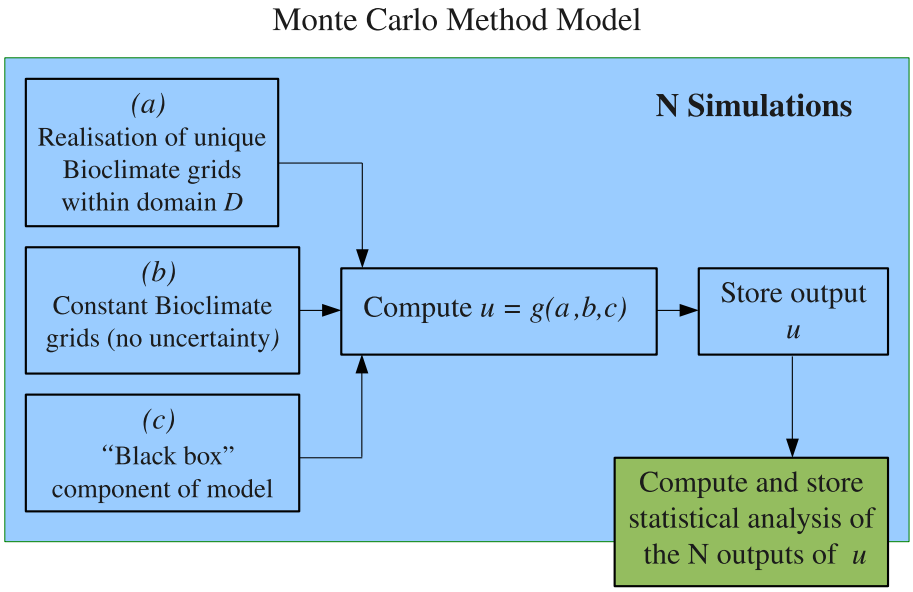


Figure 9.1: Unique uncertainty propagation sensitivity analysis models.

of the year[1]. However, this model can not exclude the climate of these months from its analysis so the climate of the warmest months is a part of the prediction model and so, this uncertainty propagation path must also be analysed.

To do this, uncertainty propagation through the 19 Bioclimate grids was individually analysed. That is, the Monte Carlo simulation model was repeated 19 times where in each case, the uncertainty propagation was limited to one Bioclimate grid pathway. These 19 models will be referred to as the *single-biogrid models*. In each simulation, 18 bioclimate grids are constant as are the frequency histograms generated from them. For this study, in these 19 Single-Biogrid models, only the Gaussian Present model was investigated.

The statistical analyses of these models were compared for similarities in their prediction-uncertainty relationship, with particular interest being if the $\surd$ shaped mean $-$ uncertainty relationship was visible, or components of it. From this it was possible to group the Bioclimate algorithms and investigate the resulting models to see the cumulative effect of having more than one uncertainty propagation path (see Figure 9.1 for a simple representation of such a model).

[1]Unless this period's rainfall is known to improve the moisture content in the soil significantly.

These simulations are discussed in Section 9.1 and its results, combined with other known information (such as when the crop is planted), are then used to help draw some final conclusions on the sensitivity of the model to the spatial patterns of the known input uncertainty (Section 9.2).

In this chapter the known uncertainty propagation paths are discussed since their structure is known and their importance in the prediction's accuracy is high. However, there is a likelihood that other paths within the "black box" component of the model may be contributing to the results of this study. But, as these algorithms are known in general but not known in detail, they cannot be further investigated or immediate questions raised (by knowing their mathematical or statistical algorithm). However, possible conclusions on their effect will be hypothesised.

9.1 Prediction Statistics and Spatial Patterns

The grid cells sensitive to uncertainty in the input climate grids, when propagating through only 1 of the 19 Bioclimate algorithms, is shown in Figure 9.3(A1) - (A19). The number of grid cells affected and the highest uncertainty in those grid cells is written in Table 9.1. From these, it can be seen that:

1. the number of grid cells affected, and the maximum uncertainty which occurs, varies significantly between each single-biogrid model.

2. the locations where the prediction has some uncertainty varies significantly across the area where the predictions are positive for the default model.

The overlap in the areas covered by the uncertainty-present grid cells is illustrated in Figure 9.3(B1). The theoretical maximum is 19 (per grid cell), but the maximum which actually occurred was 14, with most having 6 or less. So, if there is a *cumulative* [2] effect on the Gaussian model's uncertainty, it appears that it will not be as a result of the uncertainty in all 19 Bioclimate layers, at

[2]Cumulative refers to the combined effect of uncertainty propagation paths.

Single-Biogrid Model	Description	Number of Grid Cells where Uncertainty > 0	Maximum Uncertainty
1	Annual Mean Temperature	211	2.98
2	Mean Diurnal Range	350	2.58
3	Isothermality	1994	4.26
4	Temperature Seasonality	1151	2.90
5	Max Temperature of Warmest Month	721	3.06
6	Min Temperature of Coldest Month	1415	5.21
7	Temperature Annual Range	4012	2.75
8	Mean Temperature of Wettest Quarter	1418	7.21
9	Mean Temperature of Driest Quarter	1199	9.10
10	Mean Temperature of Warmest Quarter	348	2.98
11	Mean Temperature of Coldest Quarter	612	4.02
12	Annual Precipitation	728	4.29
13	Precipitation of Wettest Month	1243	6.23
14	Precipitation of Driest Month	658	5.54
15	Precipitation Seasonality	910	4.00
16	Precipitation of Wettest Quarter	808	5.01
17	Precipitation of Driest Quarter	794	3.01
18	Precipitation of Warmest Quarter	1601	4.06
19	Precipitation of Coldest Quarter	688	5.12

Table 9.1: No of Grid Cells where the uncertainty is greater than 0 in each *single-biogrid models*. The grid cells and where they occur are illustrated in Figure 9.3. The highest uncertainty value that occurs in each group of grid cells is in the last column. Figure 9.6 to 9.24 shows the mean to uncertainty relationship in these regions grid cells, as defined in the regions of interest discussed in Section 8.1.1.

any one grid cell. However, given how BIOCLIM calculates its predictions, the uncertainty will also be dependent on the uncertainty at the climate grid cells at which the Field Pea trial sites are located, as this introduces an uncertainty to the frequency histograms. This results in the prediction being influenced by multiple uncertainty propagation paths. An example of this is graphically illustrated in Figure 9.2, where uncertainty propagation paths a and b are the Bioclimate grid generation and histogram generation paths respectively. The section of the model which calculates the histograms is in the black box component of the model. But, as the histograms are calculated from cells in the Bioclimate grid, these two paths are not independent. Also, this relationship will vary across the area studied.

The agreement between the uncertainty in the full Gaussian model and the uncertainty in the single-Biogrid models is shown in Table 9.2. The comparisons of all grid cells in the study area had a low correlation of less than 0.5 which is due, in part, to most of the single-biogrid models having no uncertainty at most grid cells. When the grid cells compared are limited to those where uncertainty is present in both models compared, the correlation can be lower or higher. The most notable difference is in 3, 5, 12 (lower) and 7, 9, 13, 14, 16 (higher). The high correlation suggests that, in these models, the uncertainty is mostly (but not only) due to the uncertainty propagating in both models through the same path(s). For example, this may be the case for the single-biogrid model 13, which has a comparatively high correlation of 0.77 with the Gaussian present model. On the other hand, if the correlation is low, the uncertainty in the Gaussian present model cannot be explained in this way. Therefore, it is most likely due to the cumulative effect of multiple uncertainty propagation paths. That is, for the uncertainty at these grid cells, there is a stronger cross correlation component, or relationship, between the uncertainty at grids cells at differing locations in the Gaussian Present model.

The greater number of cells with a positive prediction, that occurs when uncertainty propagation is modeled, was discussed in Section 8.1. Therefore, the

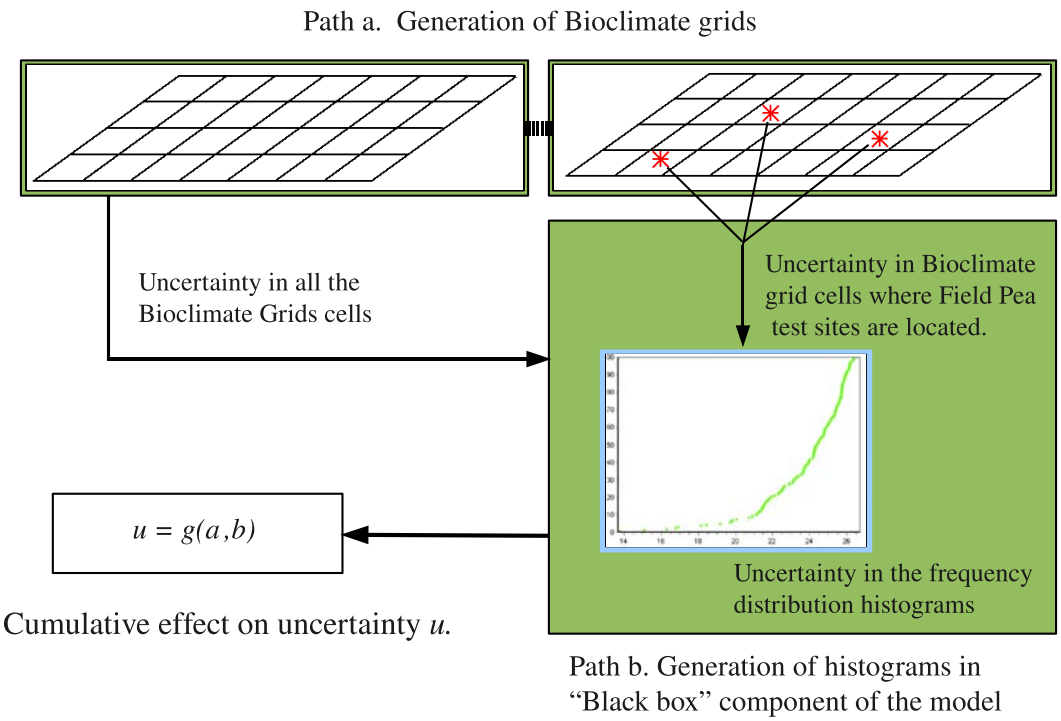


Figure 9.2: Two uncertainty propagation paths in the BIOCLIM model. These calculate (a) the Bioclimate grids and (b) the 19 frequency histograms.

Single-Biogrid Model	Description	Mean	Uncertainty	
			all	where <> 0
1	Annual Mean Temperature	0.9758	0.1345	0.0688
2	Mean Diurnal Range	0.9759	0.1514	0.1858
3	Isothermality	0.9782	0.3668	0.2209
4	Temperature Seasonality	0.9770	0.2544	0.2226
5	Max Temperature of Warmest Month	0.9763	0.2708	-0.1265
6	Min Temperature of Coldest Month	0.9768	0.3483	0.4372
7	Temperature Annual Range	0.9763	0.2407	0.5646
8	Mean Temperature of Wettest Quarter	0.9784	0.5100	0.5853
9	Mean Temperature of Driest Quarter	0.9782	0.4043	0.6628
10	Mean Temperature of Warmest Quarter	0.9759	0.1539	0.1002
11	Mean Temperature of Coldest Quarter	0.9777	0.3884	0.3043
12	Annual Precipitation	0.9767	0.2668	0.0678
13	Precipitation of Wettest Month	0.9803	0.5336	0.7127
14	Precipitation of Driest Month	0.9759	0.1690	0.3632
15	Precipitation Seasonality	0.9759	0.2620	0.3099
16	Precipitation of Wettest Quarter	0.9768	0.4912	0.6604
17	Precipitation of Driest Quarter	0.9765	0.2058	0.1754
18	Precipitation of Warmest Quarter	0.9766	0.3381	0.3387
19	Precipitation of Coldest Quarter	0.9766	0.4478	0.4822

Table 9.2: Correlation between Mean and Uncertainty of Predictions two different Models, the: single-biogrid models and the Gaussian model with uncertainty propagating through all 19 Bioclimate pathways. See Section 5.3.1 for a description of the Bioclimate Grids.

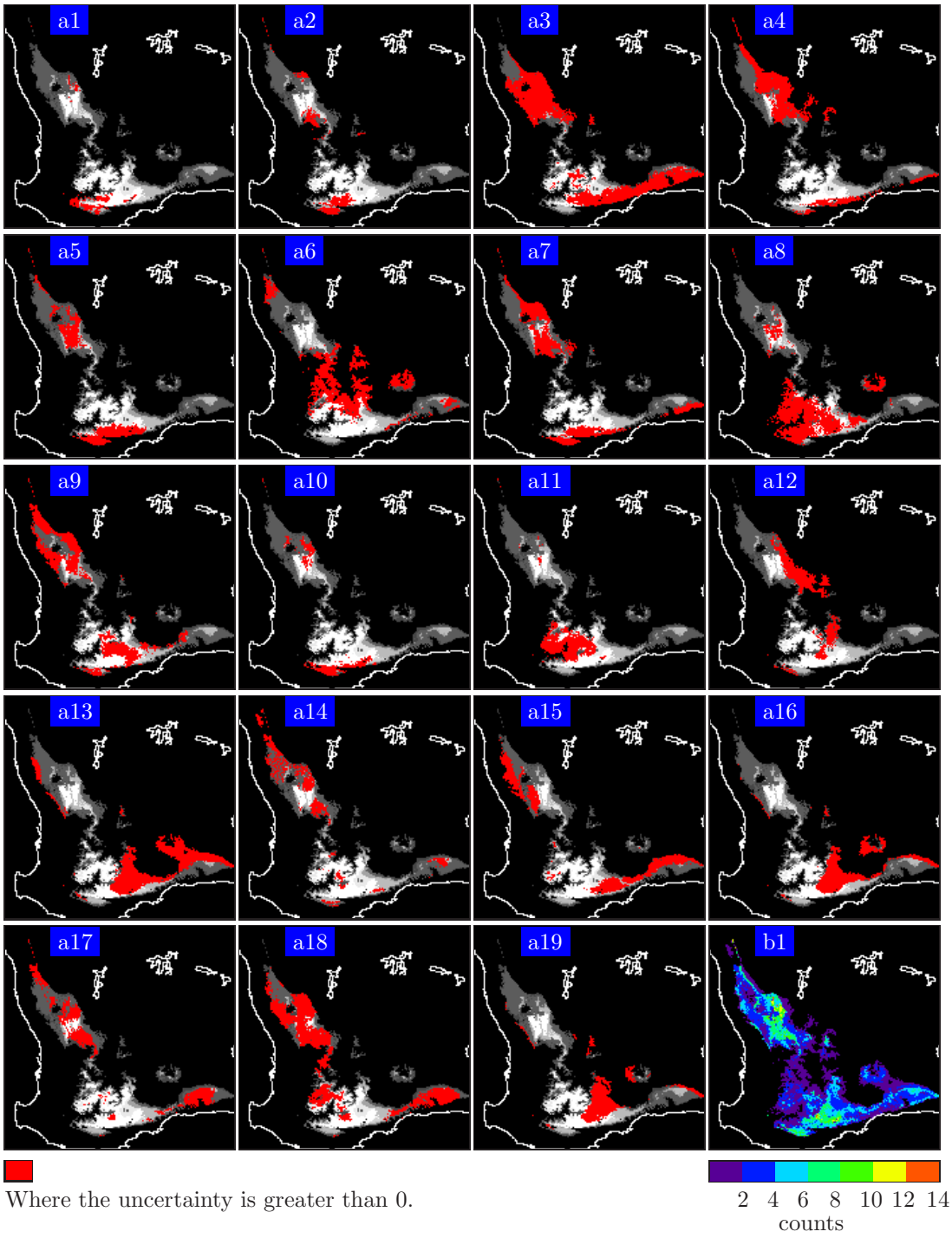


Figure 9.3: (a1 ... a19) Grid cells where the Single-Biogrid prediction has uncertainty. (b1) Frequency of occurrence.

cumulative relationship not only affects the magnitude of the uncertainty at a grid cell, but the number of grid cells with a positive prediction and uncertainty greater than 0. This is clearly seen in the comparison of 3 BIOCLIM model outputs (Figure 9.4); (a) the default model and (b) the Gaussian Present model discussed in Section 8.1; (c) mean and (d) standard deviation (at each grid cell) of the 19 single-biogrid models. In (d), the maximum value is 1.8 Percentile, with 95% of the grid cells having standard deviations greater than 0 and less than 0.7 Percentile. Therefore, the single-biogrid predictions are in very close agreement, which in turn is reflected in their similar correlations with the Gaussian Present model prediction (Table 9.2, all > 0.97).

However, the mean of these 19 single-biogrid model predictions at each grid cell; (c) have a higher agreement with the default model (a) than with the Gaussian Present model (b) (correlation of 0.99 and 0.97 respectively). In the later, the lower correlation is reflected in the lower maximum prediction and larger number of grids cells where the prediction increases from null (most notable in the circled areas of Figure 9.4, but also evident on the border areas highlighted in Section 8.1, such as the north-east of the Wheat Belt). This clearly shows the cumulative spatial effect of multiple uncertainty paths, through the 19 Bioclimate algorithms, that are present in the Guassian Present model.

To better understand which Bioclimate algorithms, and their correlative relationship, are most likely to be contributing to the changes in uncertainty (size and spatial extent), the mean - uncertainty relationship in each single-biogrid model was examined for similarities with the relationship discussed in Section 8.1.1. Of particular interest is whether the $\surd$ relationship was a product of all Bioclimate uncertainty paths, or is a result of a lesser number of propagation paths. Initial results showed two clear relationship types:

1. A reduction in the uncertainty as the prediction increased.

2. A reduction and then increase as the prediction increased,

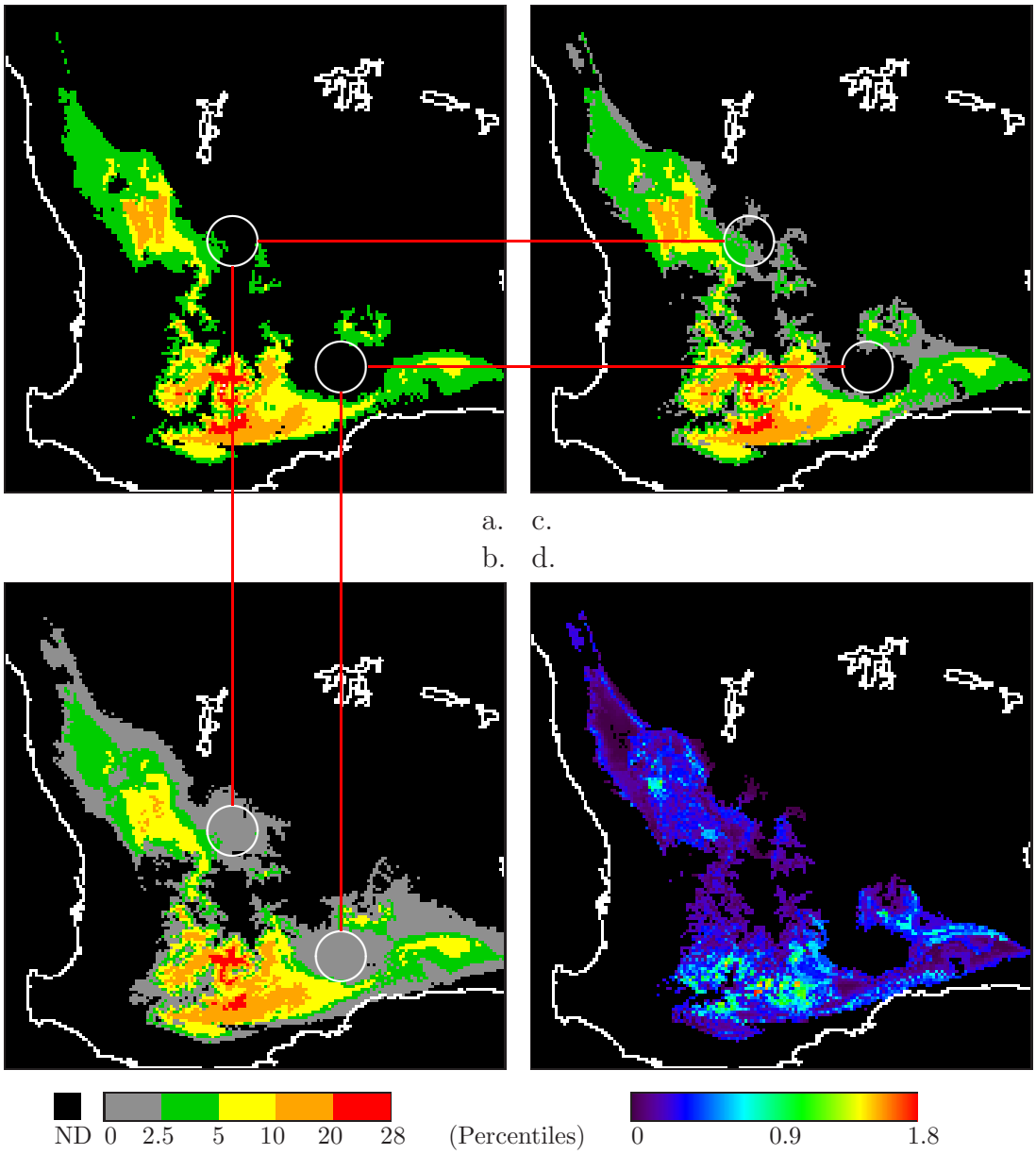


Figure 9.4: Comparison of Models (Percentile). (a) Default model, (b) Gaussian Present model, (c) mean and (d) standard deviation of the 19 Single-Biogrid models. Circled areas are examples of where the predictions are higher in the Gaussian and Single-Biogrid models.

as shown in Figure 9.6 to 9.24. This is discussed in the following three Sections, with the Groups of Regions of Interest investigated in Chapter 8.1.2 being further grouped into three "super groups" (Table 9.3): Super Group A − Group 2 and 3, Super Group B − Group 4 and Super Group C − Group 5 and 6. Group 1 is excluded. The colour labels assigned to each ROI in each Super Group is shown in Table 9.4.

9.1.1 Super Group A

In Super Group A, the differences in the mean versus uncertainty relationship are clearly visible. Firstly, when uncertainty propagates through one of the six Bioclimate grids - 1, 3, 7, 9, 14, or 19 [3]; the uncertainty reduces as the prediction approaches the default prediction value and, for most of the grid cells, the prediction is lower than or equal to the default prediction. Therefore, it can be concluded that the influence of uncertainty, when propagated through any one of these bioclimate grids, will (a) cause the prediction to be lower at most grid cells and (b) the uncertainty in the mean prediction decreases as that prediction approaches the default prediction. When, in a BIOCLIM Monte Carlo model, uncertainty simultaneously propagates through all six of these Bioclimate grids (referred to as the *Bioclimate-Group-1* model), there is a clear cumulative effect on the prediction's uncertainty, as the size of the uncertainty is higher. Also, the reduction to 0 as the mean prediction approaches the default value is still clearly visible (Figure 9.25(a)).

When uncertainty propagates through one of the 13 Bioclimate grids - 2, 4,

[3]The names of the Bioclimate Grid are in Table 9.2; see Section 5.3.1 for a detailed description of the Bioclimate Grids.

Super Group	Groups
A	2 and 3
B	4
C	5 and 6

Table 9.3: Grouping of Groups 2 to 6

5, 6, 8, 10, 11, 12 , 13, 15, 16 , 17 or 18; the $\surd$ shaped relationship is seen to differing degrees and the uncertainty of the prediction at each grid cell is mostly larger. When uncertainty simultaneously propagates through all of these grids (referred to as the *Bioclimate-Group-2* model), this relationship is much more clearly visible and the uncertainty is much higher (Figure 9.25(b)).

For the Bioclimate-Group-1 and Bioclimate-Group-2 models, the number of grid cells where the prediction equals the default model prediction and which have an uncertainty of 0 is shown in Table 9.4 and mapped in Figure 9.5. It is significant that the number of prediction grid cells influenced by uncertainty is considerably higher in the Bioclimate-Group-2 model.

Also, the large number of grid cells that are not influenced by uncertainty in the Bioclimate-Group-1 model are mostly in the south west of the area studied. At these grid cells the prediction is the same as the default prediction from which it can be concluded that at some grid cells the propagation of uncertainty, through some of the bioclimate grids, does not influence the prediction. This suggests that with each simulation the Bioclimate values at (a) these grid cells and (b) the Field Pea test site grid cells (keeping in mind that these two groups of grid cells are not mutually exclusive) does not change the prediction assigned. Therefore, it can be concluded that the changes that do occur in these Bioclimate grids and the 19 frequency histograms (with each simulation) must not be sufficiently large to change where the predictions are boxed.

This also occurs in the Bioclimate-Group-2 model, but to a much smaller extent. What is more significant is that it occurs at different grid cells. Therefore, the various bioclimate algorithms influence both the uncertainty and its location in the prediction, as discussed in Section 9.1.

9.1.2 Super Group B

In Super Group B, for all Bioclimate layers, the uncertainty decreases as the prediction approaches the default prediction, with the exception of Bioclimate grid 6 and 11. This suggests that the uncertainty in these 2 Bioclimate grids

Group	Region of Interest	Default Prediction	Bioclimate-Group-1	Bioclimate-Group-2	Colour Label
2	2.9	867	383	15	red
	3.8	1280	86	58	green
3	6.7	433	163	1	blue
	7.7	381	13	0	yellow
	8.7	31	0	0	cyan
	9.6	246	85	0	magenta
	10.6	100	10	0	maroon
	11.5	171	4	0	sea grass
4	13.5	77	51	0	red
	14.4	56	0	0	green
	15.4	77	34	0	blue
	16.3	1	0	0	yellow
	17.3	124	12	0	cyan
	18.3	34	0	0	magenta
	19.2	8	0	0	maroon
5	21.2	14	0	0	red
	22.1	48	0	0	green
	23.1	29	0	0	blue
	24.0	11	0	0	yellow
	25.0	2	0	0	cyan
	26.0	8	0	0	magenta
6	28.8	2	0	0	maroon

Table 9.4: Number of grid cells where default prediction occurs in Default Prediction, Bioclimate-Group-1 and Bioclimate-Group-2 models. The colour labels assigned to each Region of Interest in each Group is shown.

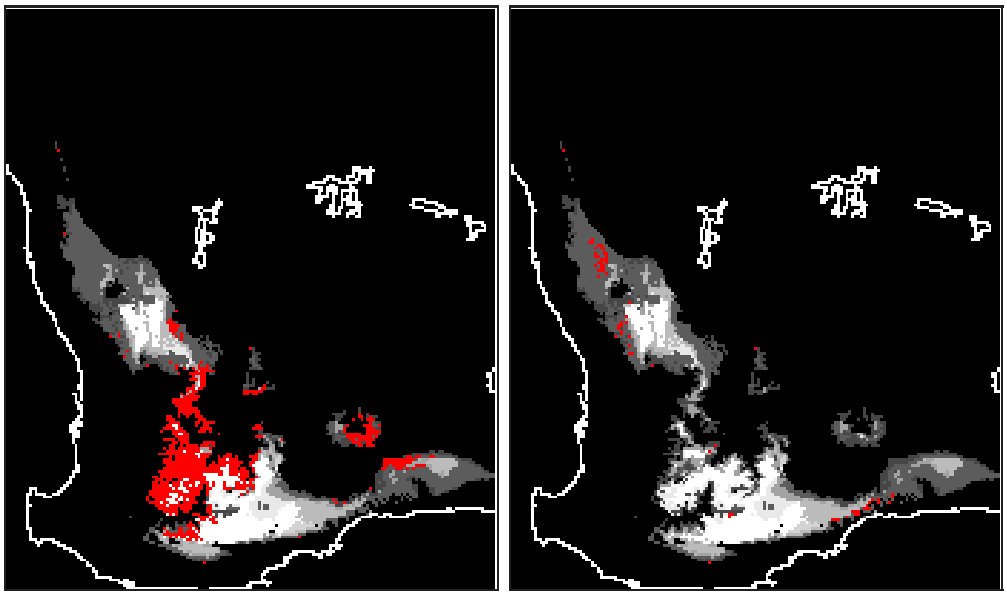


Figure 9.5: The red coloured grid cells are where the prediction is not changed by uncertainty in the Bioclimate grids: Bioclimate-Group-1 and Bioclimate-Group-2.

has a significant influence on the uncertainty in the grid cells of Group 4; as the uncertainty propagation is clearly greater in Bioclimate-Group-2 than in Bioclimate-Group-1, with more grid cells having higher uncertainty results and none having the value of 0.

9.1.3 Super Group C

In Super Group C, the only clearly visible $\surd$ shape is in the blue region. Also, it is more notable in the Bioclimate-Group-1 model than in the Bioclimate-Group-2 model. This is possibly due to the propagation of the uncertainty through the Bioclimate 9 grid (Figure 9.14). However, as discussed in Section 8.1.2, the low number of grid cells in Group 4 and 5 makes it difficult to draw a definite conclusion.

9.2 Discussion

This Section aims to clarify how the models structure influences the uncertainty in the prediction, expanding on the analysis in the previous Section (9.1). As discussed, it is clear that these Bioclimate algorithms can be grouped into two subsets of the default BIOCLIM model (Bioclimate-Group-1 and Bioclimate-Group-2). The following points describe why these differences are occurring by (a) considering the structure of the algorithms and (b) what the algorithm calculates and its relevance to the crop of interest being predicted. In drawing a conclusion from both of these, it is important to consider how uncertainty related changes in the Bioclimate grids can tip the prediction from one sub-group to another. Therefore, the effect of the discreet method of prediction, in the "black box" component of the BIOCLIM model, should be considered.

Not all the Bioclimate grids are discussed as there are similarities in their sensitivity to uncertainty propagation. Instead, those discussed here most clearly influence the uncertainty in the prediction (away from 0).

Bioclimate-Group-1

1. The uncertainty in the Bioclimate-Group-1 model prediction is either non-existent, or exists in grid cells where the prediction has been reduced (with the introduction of climate uncertainty into the model). In the first case, this suggests that the change in the Bioclimate grids and the BIOCLIM frequency histogram distributions is not sufficient to change the prediction (as discussed in Section 9.1.1). In the second case, the uncertainty of the mean prediction can start relatively high before decreasing as the mean prediction tends towards the default value. Also, and of significance, is that the prediction does not ever exceed the default prediction.

2. The reason for this mean to uncertainty relationship is most likely due to: (a) The simple mathematical and statistical algorithms in the known uncertainty paths of this model (such as Bioclimate 1, annual mean temp) and (b) that these Bioclimate algorithms do not assess the temperature or

precipitation climates across the temporal periods that are important to the species being modeled - in this case the Field Pea - so the prediction does not change.

3. The exceptions to this (in the Bioclimate-Group-1 model) are the Bioclimate 9 (Mean temperature of the driest quarter) and 19 (Precipitation of the coldest quarter) grids. These do produce notably higher uncertainty in predictions at some grid cells, especially in Super Group A (where lower predictions occur). This is expected as the grids where lowest predictions occur are the most sensitive to introduced uncertainty, as discussed in Section 8.1.1.

4. The structure of the Bioclimate 9 and 19 climate grid's algorithms is not significantly different in complexity or structure from the other Bioclimate algorithms, so it can be concluded that these algorithms structures is not the only cause of the higher sensitivity of the predictions. Instead, it is likely to be the result of the influence that these climates, have on the prediction of a certain crop or species. For example, the precipitation of the coldest quarter is important to the Field Pea as it is a winter crop in Western Australia. Also, but less obvious, is that the mean temperature of the driest quarter is also important as a hot summer will affect the amount of moisture in the soil.

Bioclimate-Group-2

1. The uncertainty in the Bioclimate-Group-2 model prediction is (a) present in a much larger number of grid cells and (b) has a high value in a larger number of grid cells. Therefore, in a larger number of grid cells than occurs in Bioclimate-Group-1, the change in the Bioclimate grids and the BIOCLIM frequency histogram distributions *is* sufficient to change the prediction in a larger number of grid cells. Also, the uncertainty in the predictions is significantly higher in a greater number of grid cells where higher predictions occur.

2. Also, unlike what occurs in the Bioclimate-Group-1 model, at a significant number of grid cells the prediction *does* exceed the default prediction. This is seen in the skewed $\surd$ mean versus uncertainty relationship (at each grid cell). From this, it can be concluded that the uncertainty paths in the Bioclimate-Group-2 model are the most likely contributors to the non linear prediction−uncertainty relationship discussed in earlier sections of this chapter and Chapter 8.

3. This relationship occurs to differing degrees at different grid cell locations, especially in Super Group A. It is also clear that (a) the size of the uncertainty varies significantly and (b) the skewed $\surd$ relationship is most visible when uncertainty propagates through the Bioclimate 11 and 12 grids and to a lesser extent the Bioclimate 13 grid (the purple grid cells).

4. When uncertainty propagates through Bioclimate grid 6 (minimum temperature in coldest month) the mean to uncertainty relationship is notably different in that the uncertainty only increases with increasing prediction. These increases are large and clearly binned, most notably in Super Group B and C. This discreetised output of uncertainty reflects the discreetised decision structure of the BIOCLIM model.

5. In Super Group A, when uncertainty propagates through the Bioclimate grids which represent quarterly (three month) climate information, the $\surd$ prediction to uncertainty relationship is almost always present. For example, when uncertainty is present in the Bioclimate grid 11 (Mean Temperature of Coldest Quarter) grids. This analysis of shorter time periods adds detail to the climate analysis of the region, so uncertainy in these bioclimate grids are more likely to influence the accuracy of the prediction.

6. The sensitivity of the prediction, to uncertainty propagation through Bioclimate grid 11, is possibly due to the fact that the temperature represented in these grid cells is below the minimum required for flowering and podding

(the Field Pea is particularly sensitive to late frosts during its flowering and podding (Moore 1998)). This possible conclusion is supported by the Bioclimate grid 6 analysis, which improves the spatial classification of where the coldest periods of the year occur. The effect on the prediction's uncertainty, by the uncertainty propagation through these bioclimate grids (6 and 11), is most likely to be cumulative as its influence mostly occurs at the same grid locations and in the same time of the year. However, these locations mostly border the areas of higher productivity. From this it can also be concluded that, in the grid cells where the prediction is highest, uncertainty in these bioclimate grids has less (if any) effect on the uncertainty in the prediction.

7. Also in Super Group A, the $\surd$ relationship it very clear when uncertainty propagates through the Bioclimate 12 grid (Annual Precipitation), which investigates a long time period. What is significant is that the uncertainty in the prediction is present in a large number of grid cells, but not where the highest prediction occurs. This is expected as the higher predictions occur closest to where where the crops were trialled.

8. As in Bioclimate-Group-1 model, the bioclimate algorithm's complexity alone can not explain the greater sensitivity of this model. So, determining the climate related reasons that have created the uncertainty propagation paths of significance is important in understanding what causes the uncertainty observed at each grid cell.

9.3 Conclusion

To conclude, this analysis can be used to determine which components of the model are most sensitive to propagation of uncertainty. This knowledge can be used to improve the accuracy of the model's results (by further model development) or, to better understand the validity of the model for a specific study.

From this Thesis' results it is clear that minimising uncertainty propagation in ecological niche models could be most easily achieved if (a) the uncertainty in the climate grids could be reduced and (b) the models algorithm(s) could be modified to minimise uncertainty propagation without effecting overall predictions. Finally (c), using the mapped statistical results and knowledge of the species being studied, modify the model to exclude uncertainty-sensitive Bioclimate grids.

The first of these may be difficult due to a number of reasons such as lack of quality tested data. The second option may be possible if it is not constrained by factors such as time (are results needed now?) and money (is funding available for future model development?). The final option would need to be tested by the model user specific to their requirements. For example, in this study the highest uncertainty is occurring in grid cells where the Field Pea prediction is low, so the crop will not be grown there regardless of the uncertainty of the prediction. Alternatively, if the prediction is high then its uncertainty will not influence the decision of where to plant the crop. Finally, if the model is known to have an anomaly that causes the uncertainty to be higher at mapped grid cells that border where the default predictions changed, then these results can be excluded in the final decision making stage. However, this would require a robust method of analysis and mapping of where this is occurring, similar to the methodology discussed in Chapter 8.

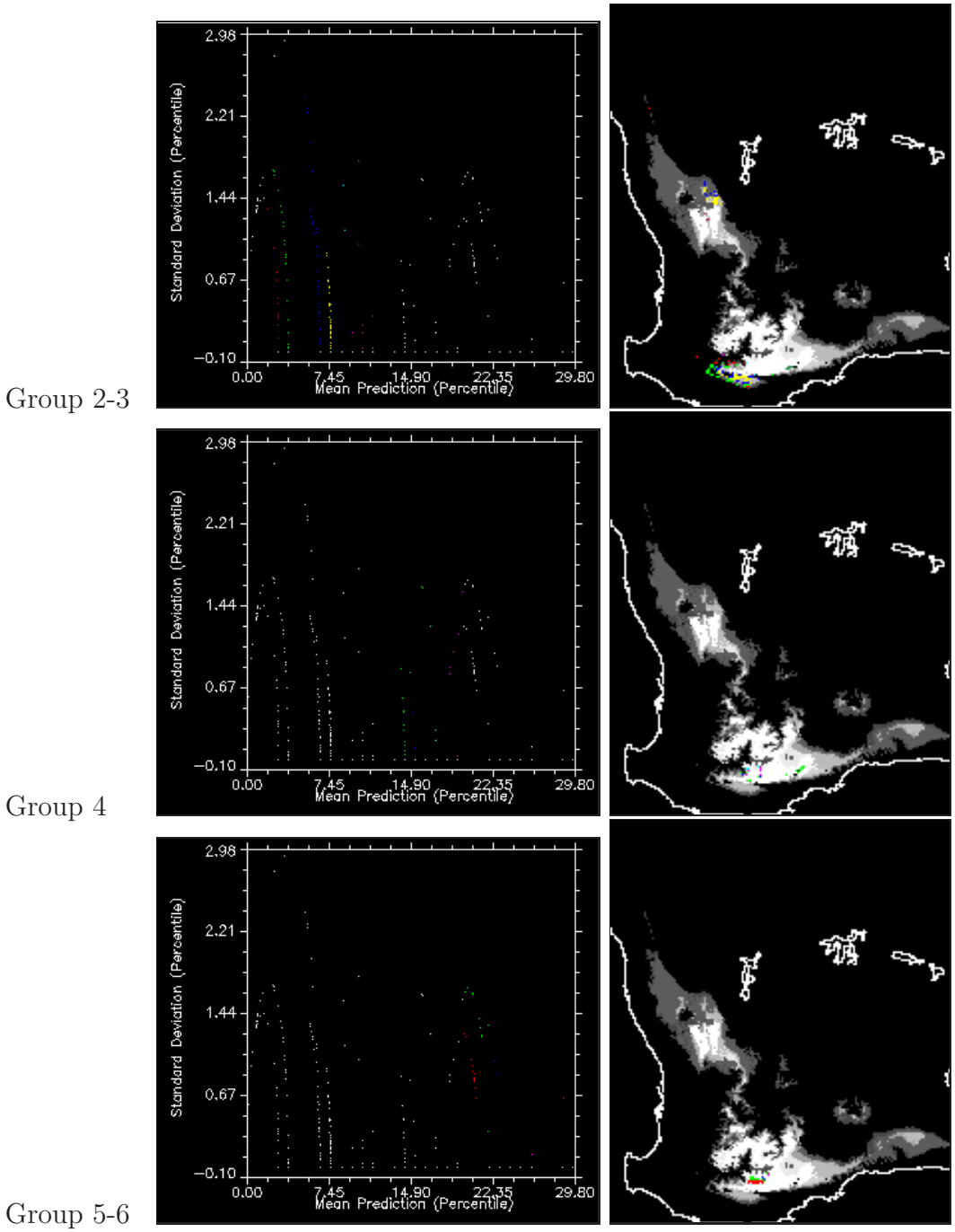


Figure 9.6: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 1. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

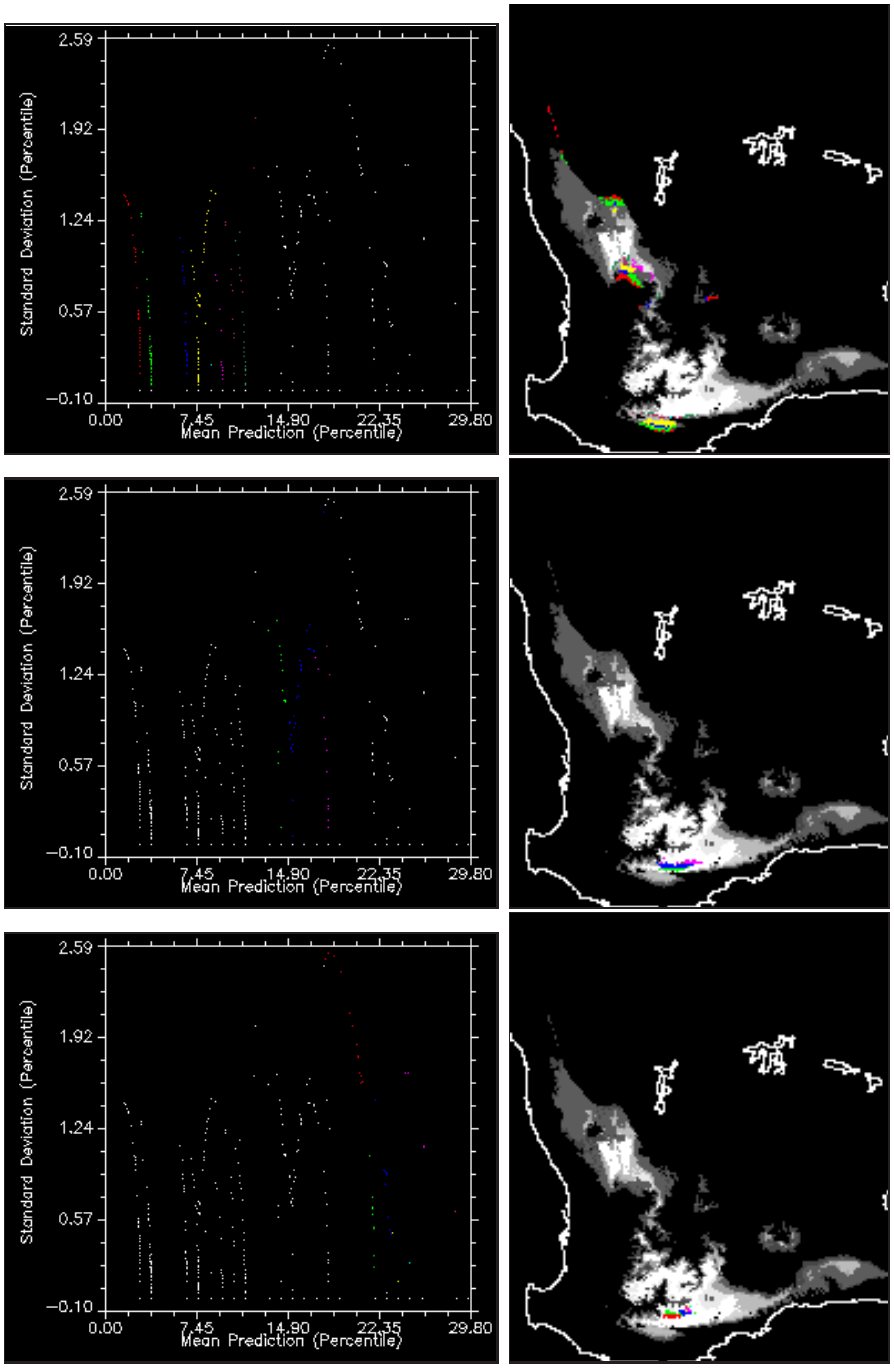


Figure 9.7: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 2. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

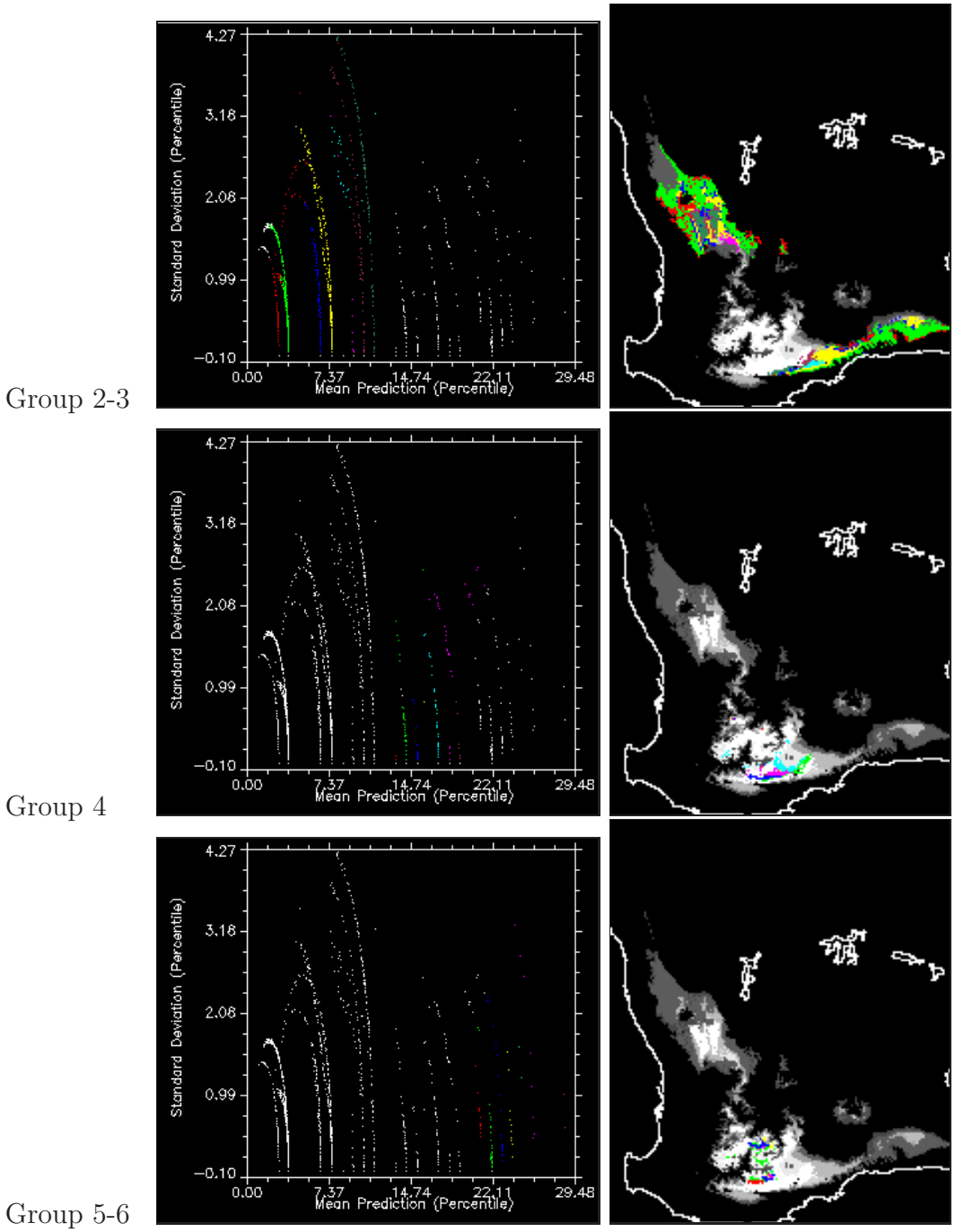


Figure 9.8: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 3. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

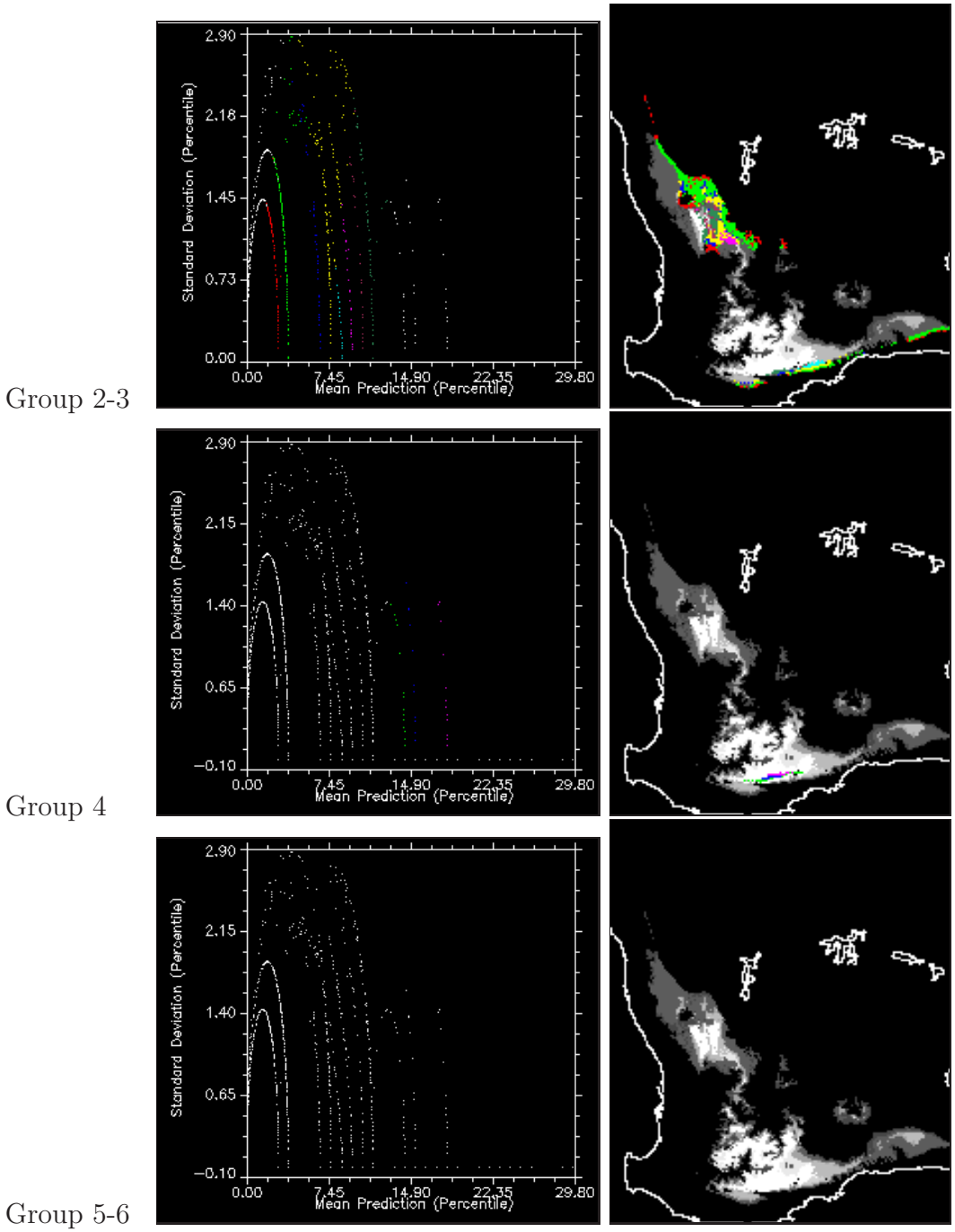


Figure 9.9: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 4. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

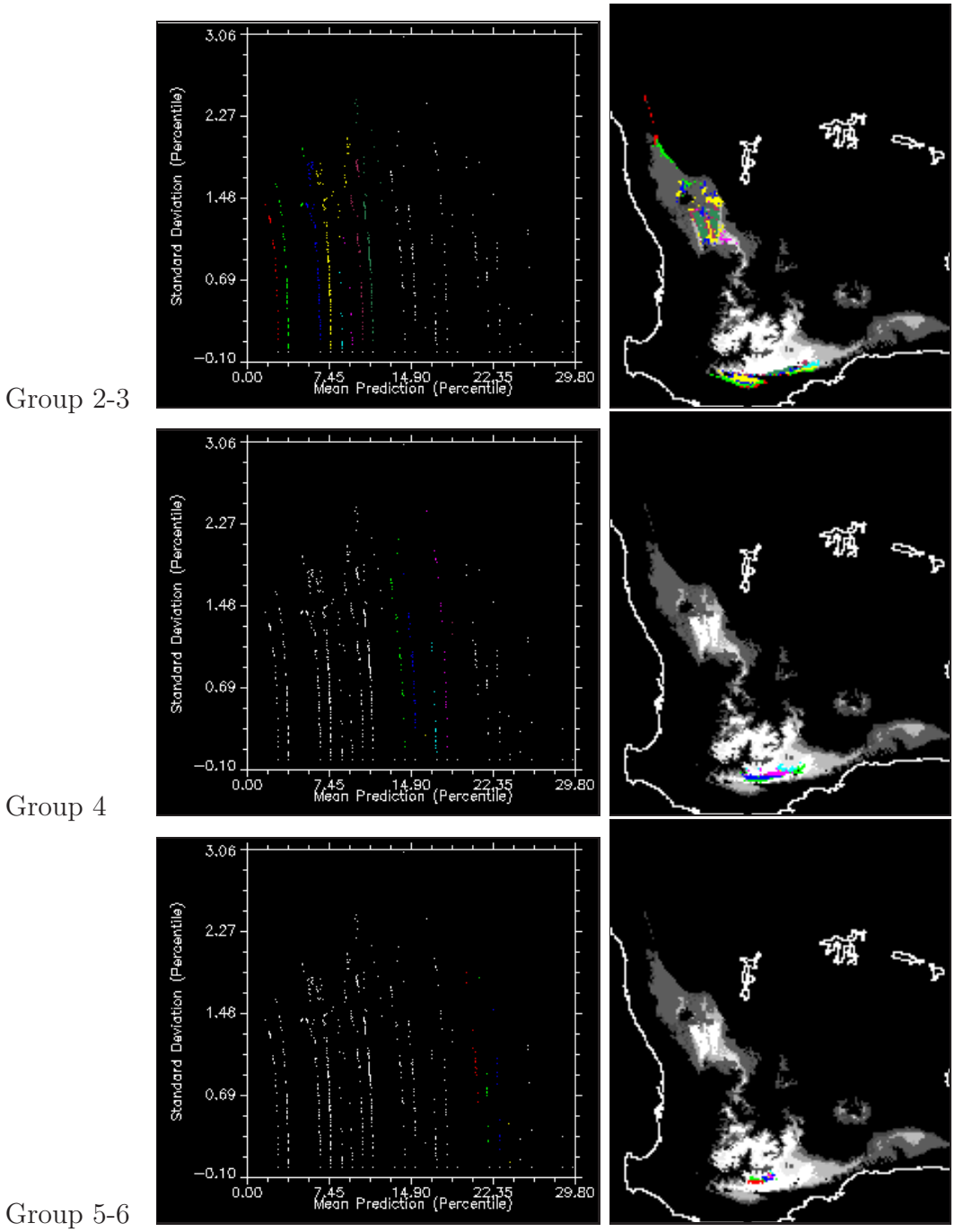


Figure 9.10: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 5. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

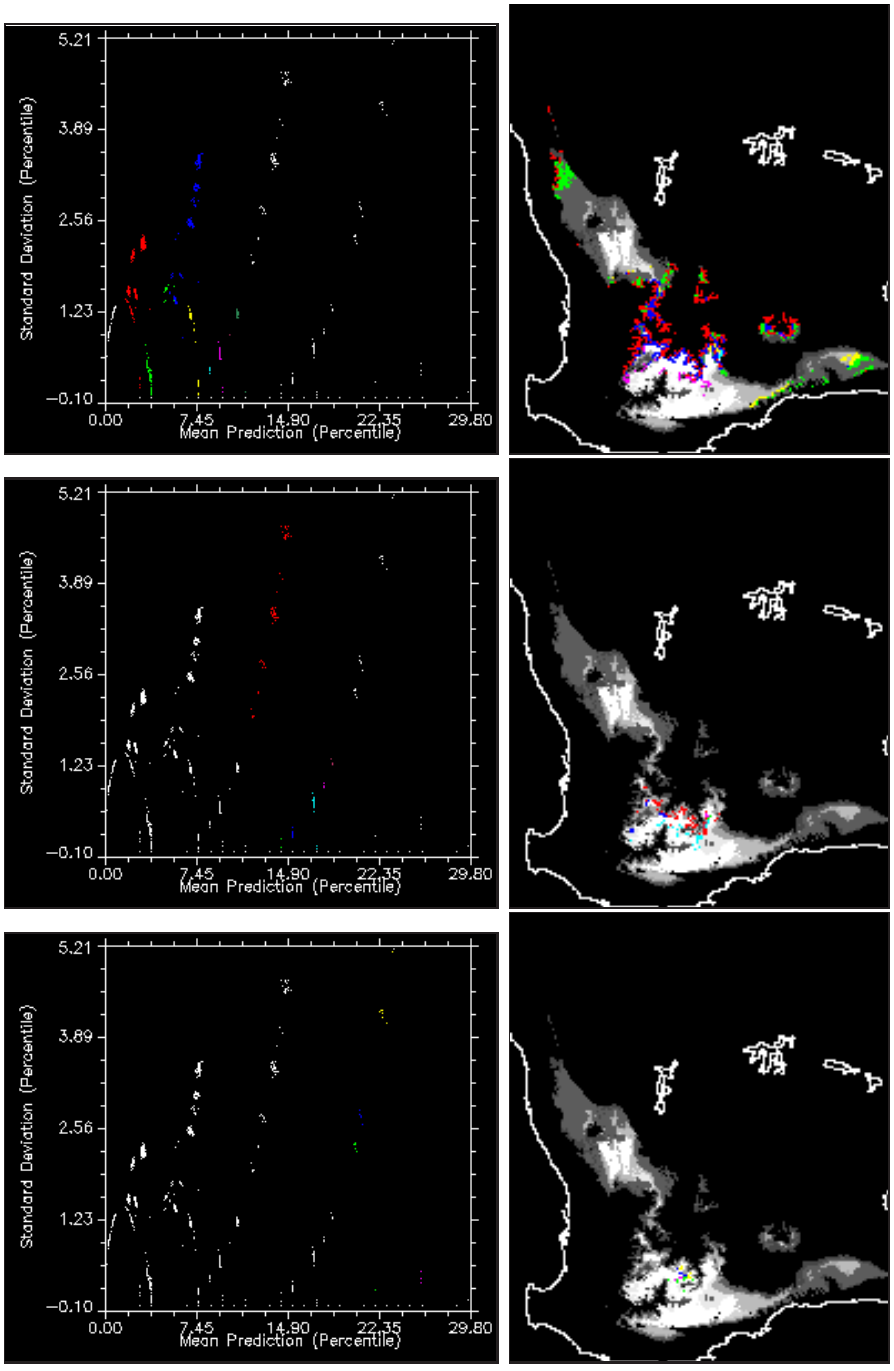


Figure 9.11: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 6. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

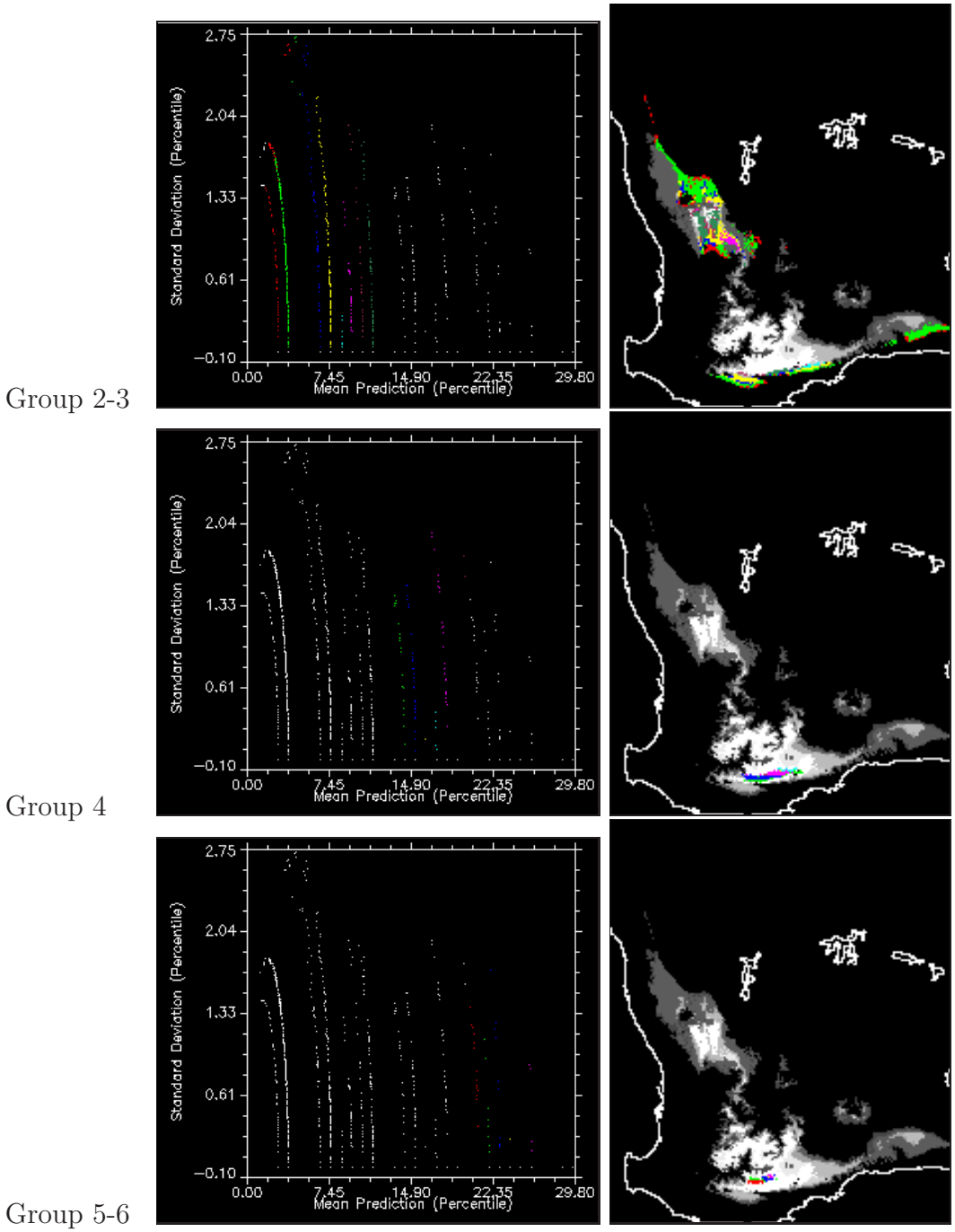


Figure 9.12: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 7. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

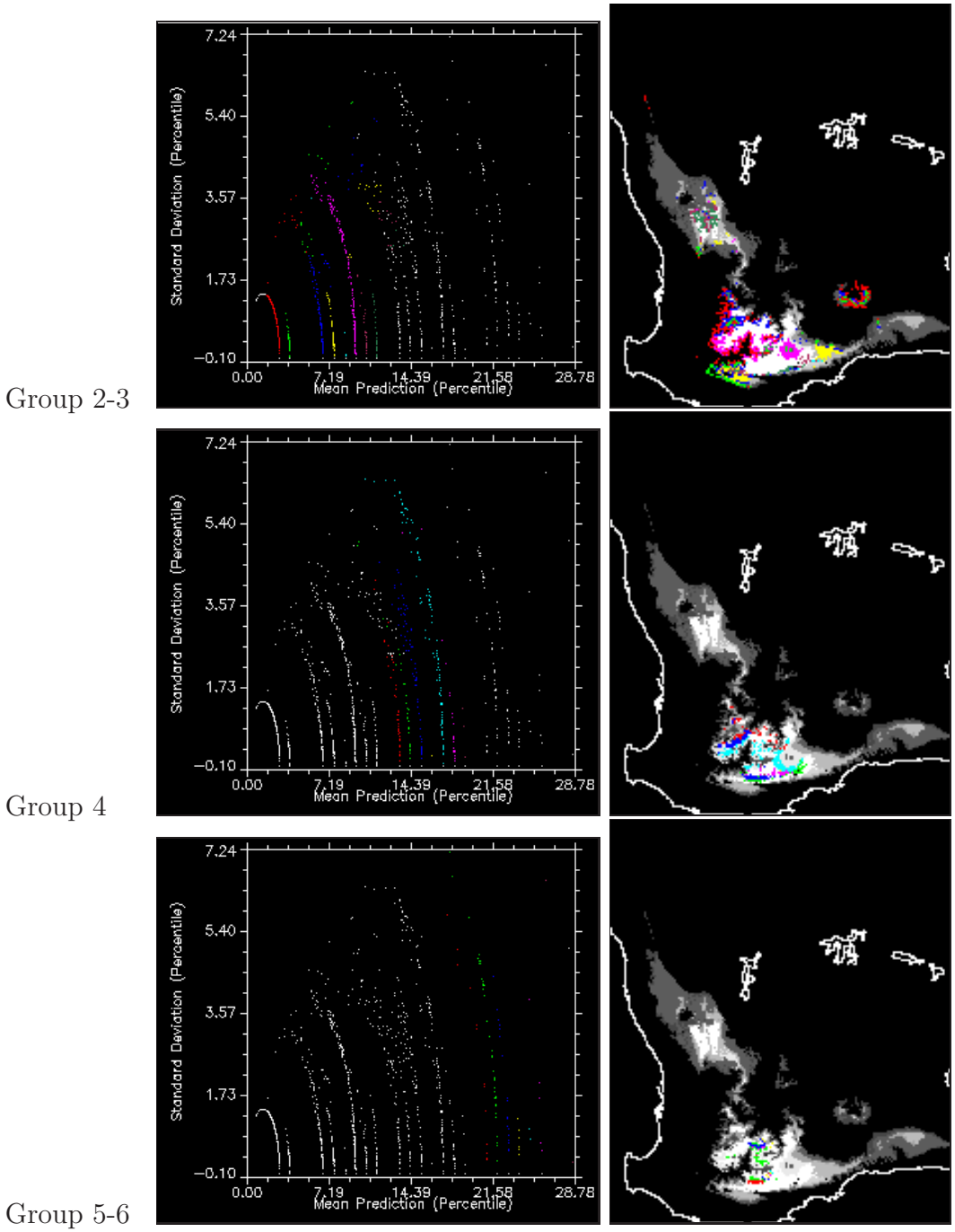


Figure 9.13: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 8. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

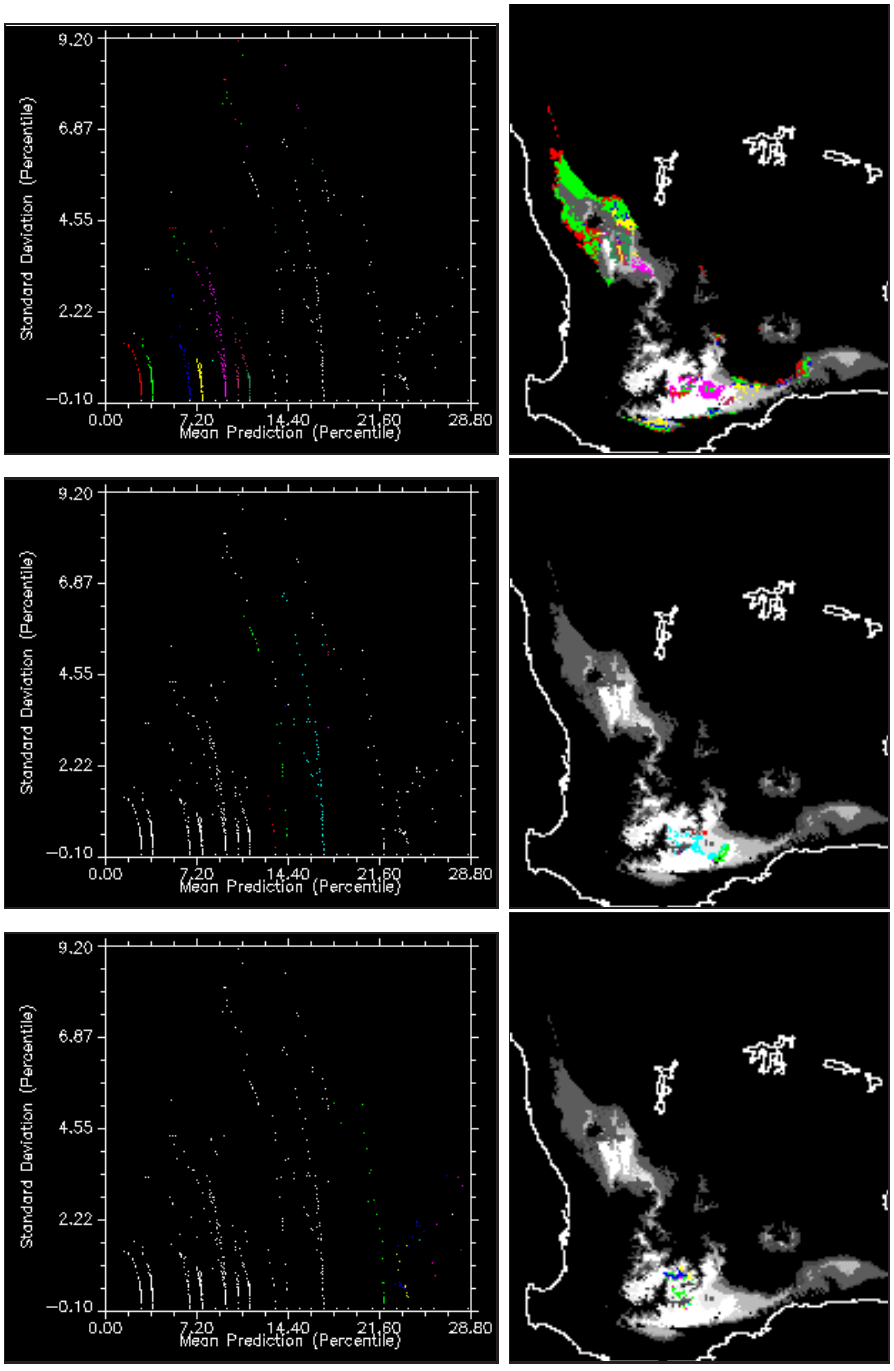


Figure 9.14: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 9. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

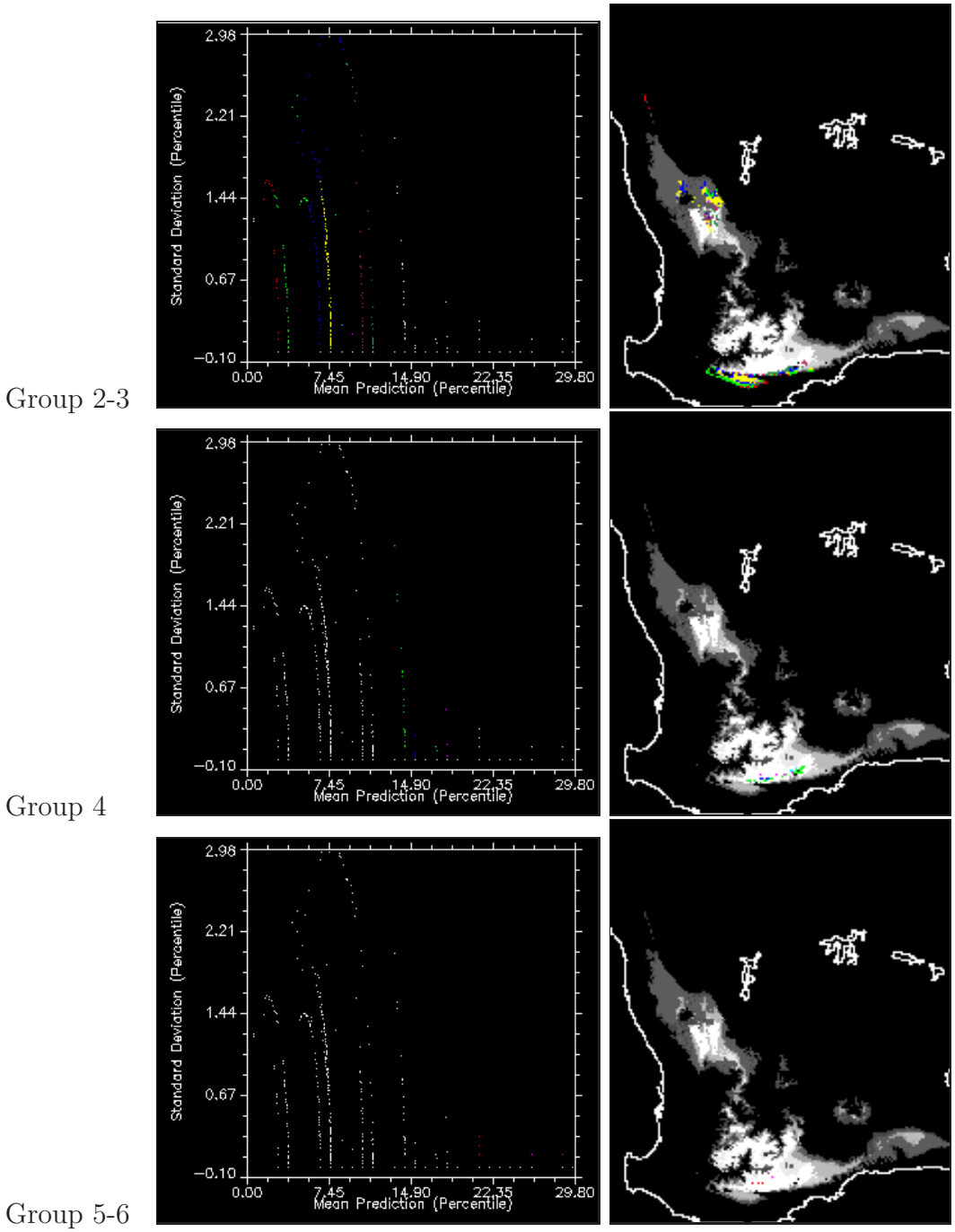


Figure 9.15: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 10. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

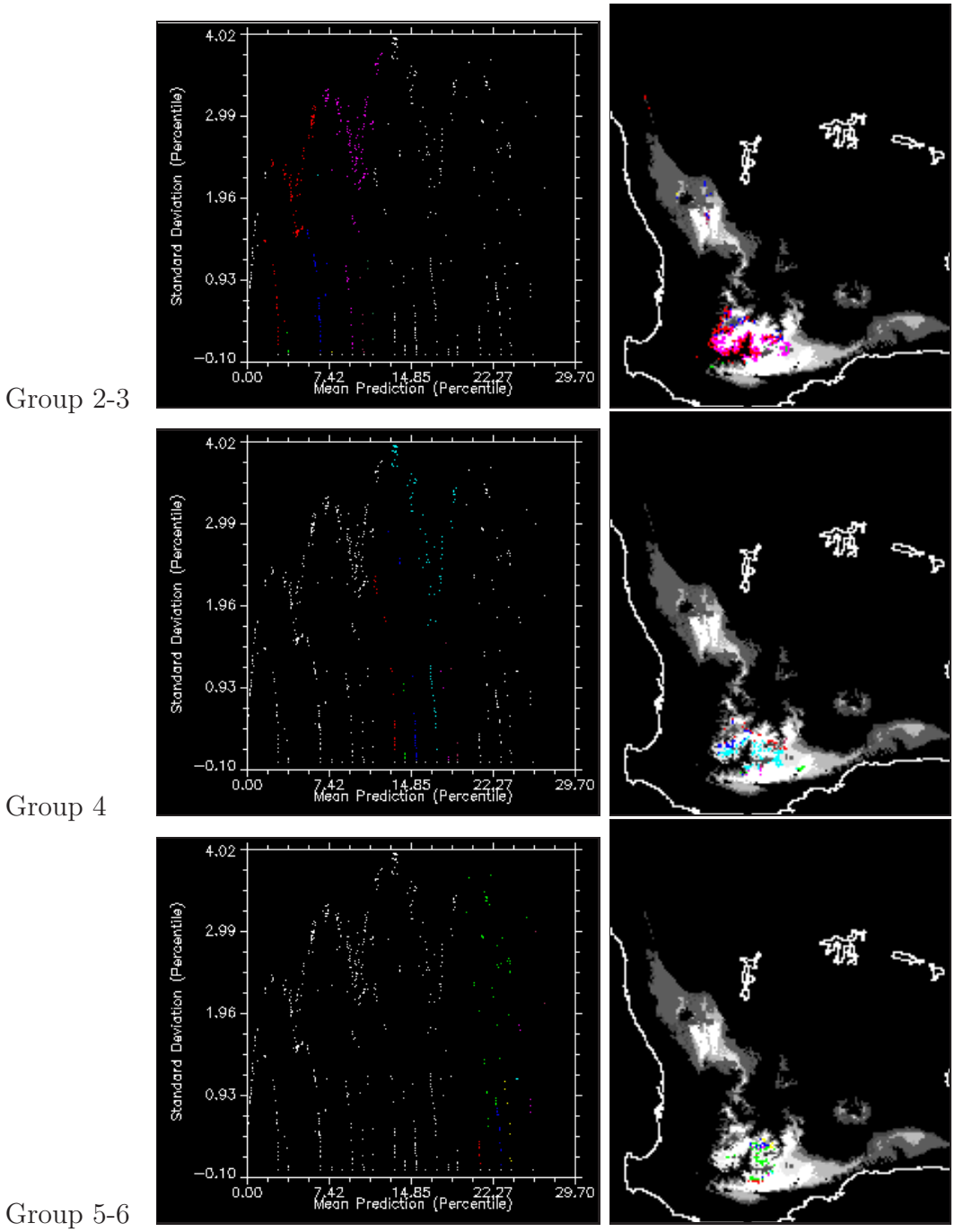


Figure 9.16: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 11. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

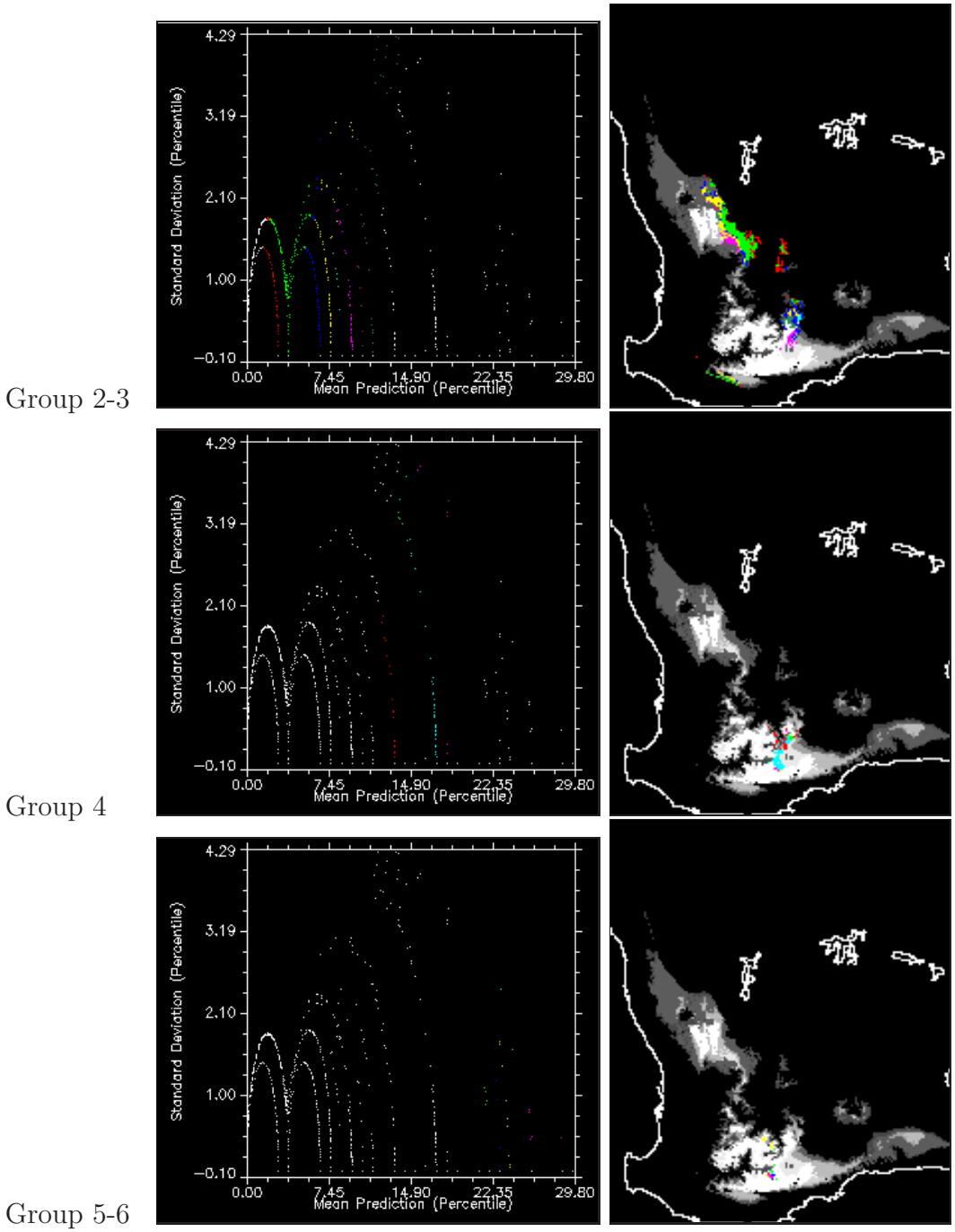


Figure 9.17: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 12. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

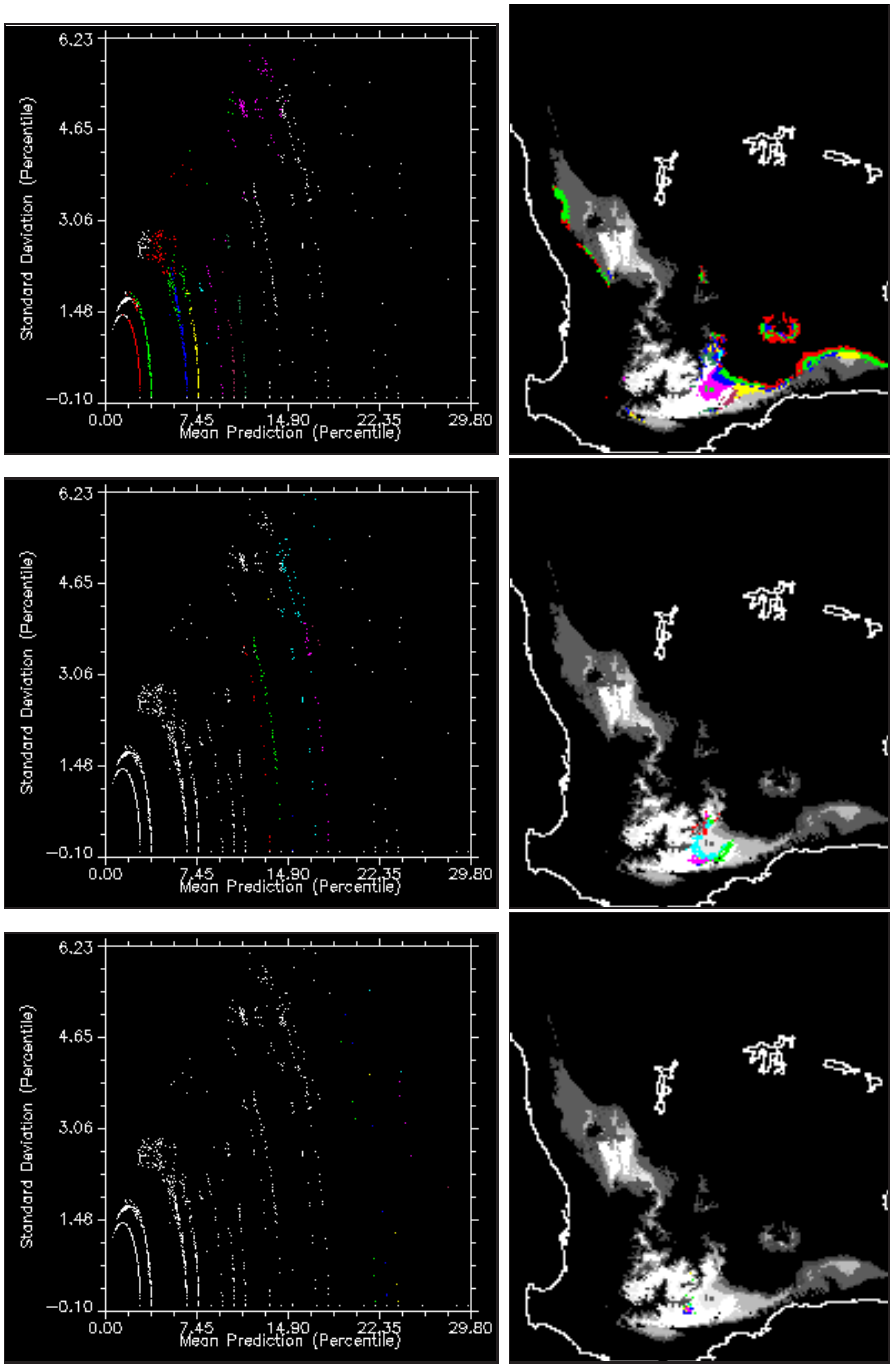


Figure 9.18: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 13. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

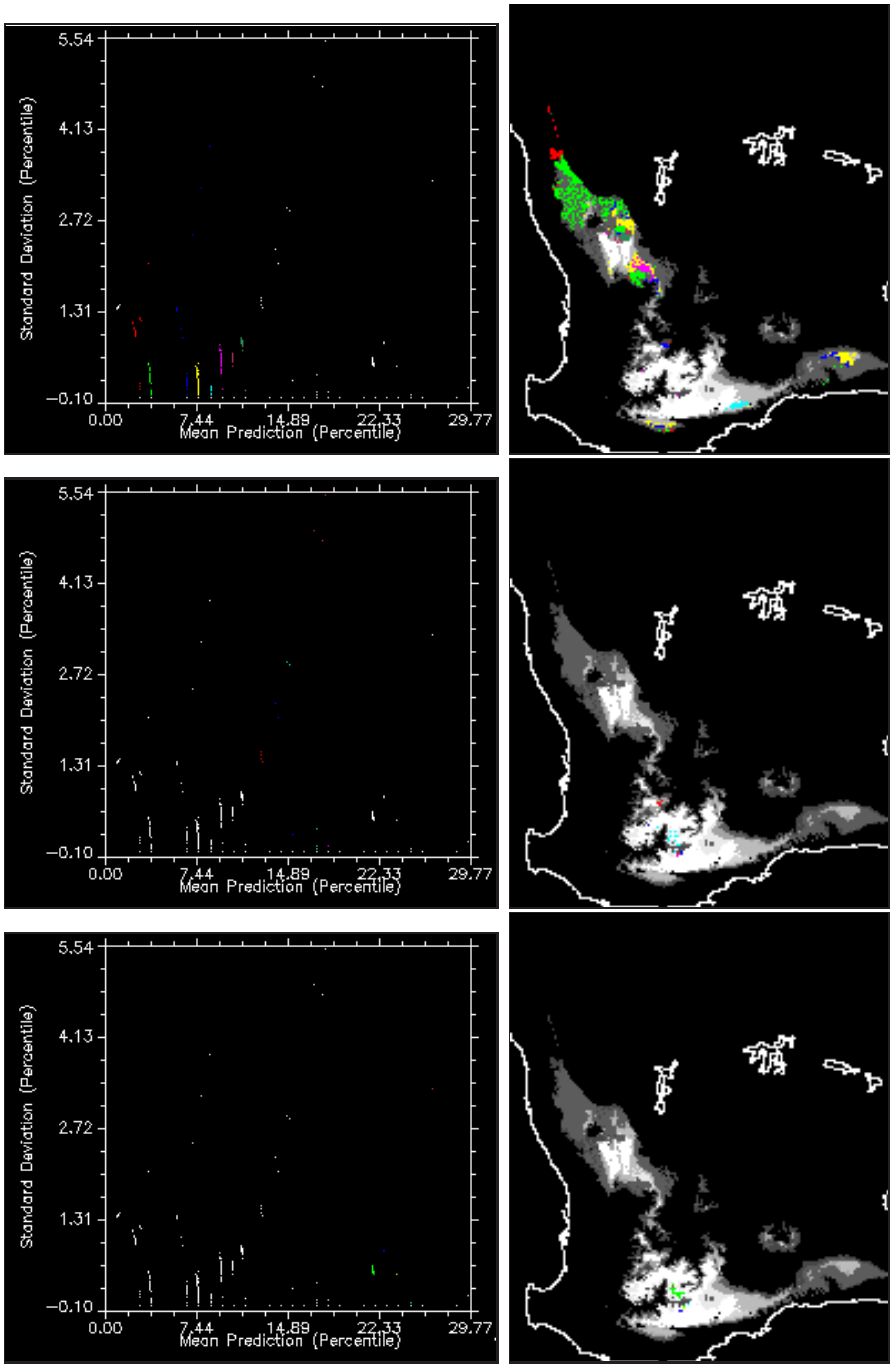


Figure 9.19: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 14. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

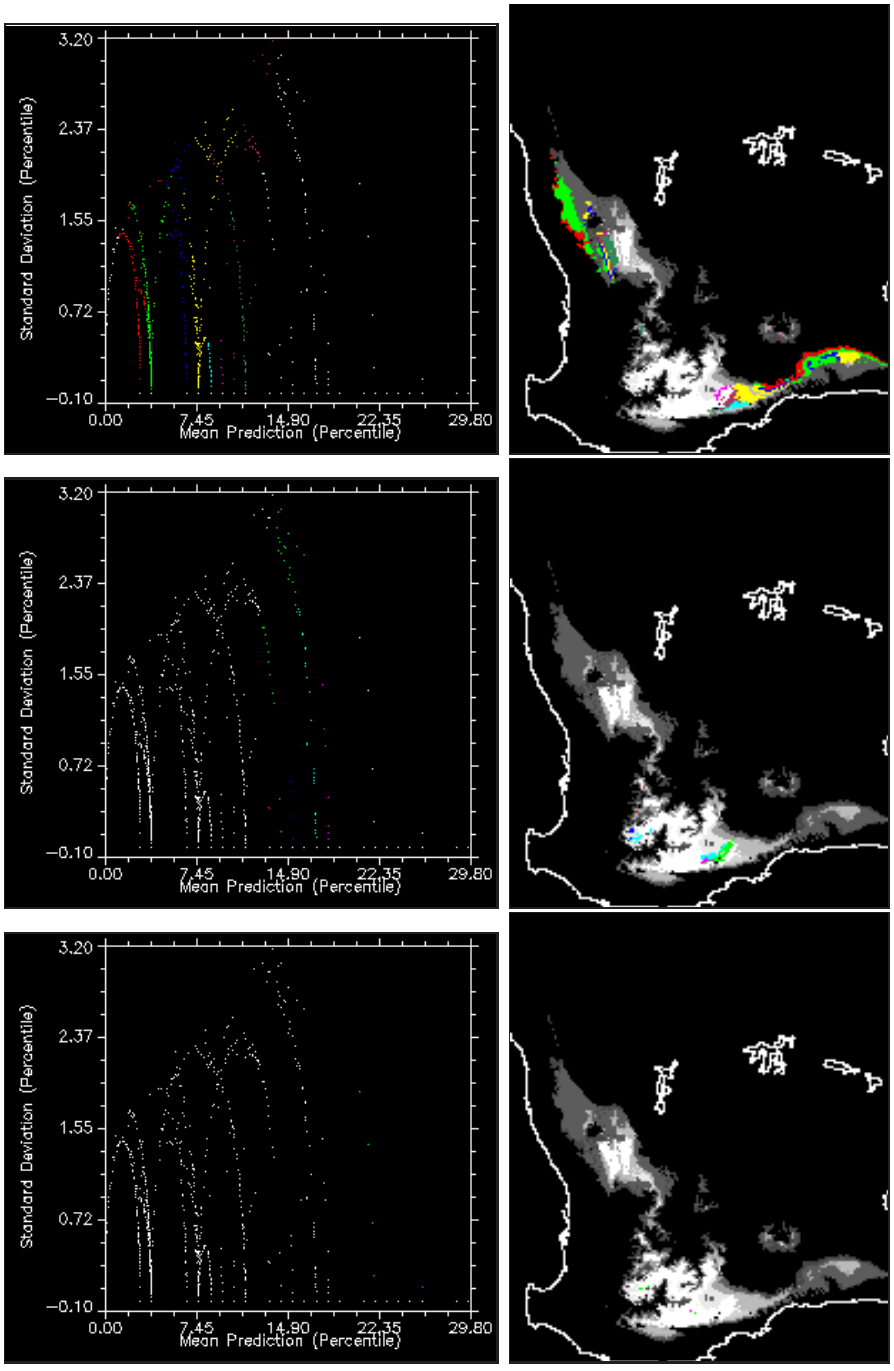


Figure 9.20: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 15. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

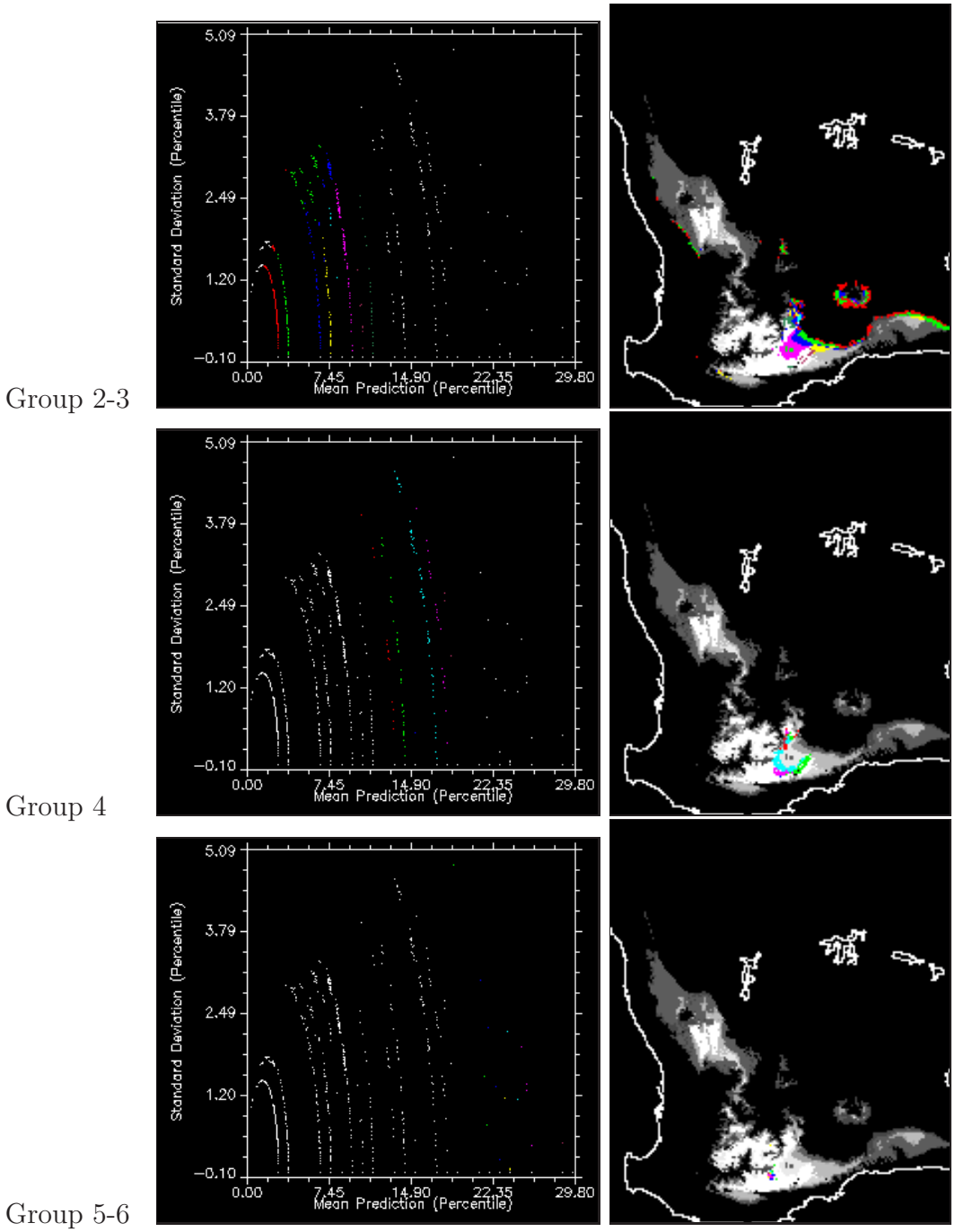


Figure 9.21: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 16. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

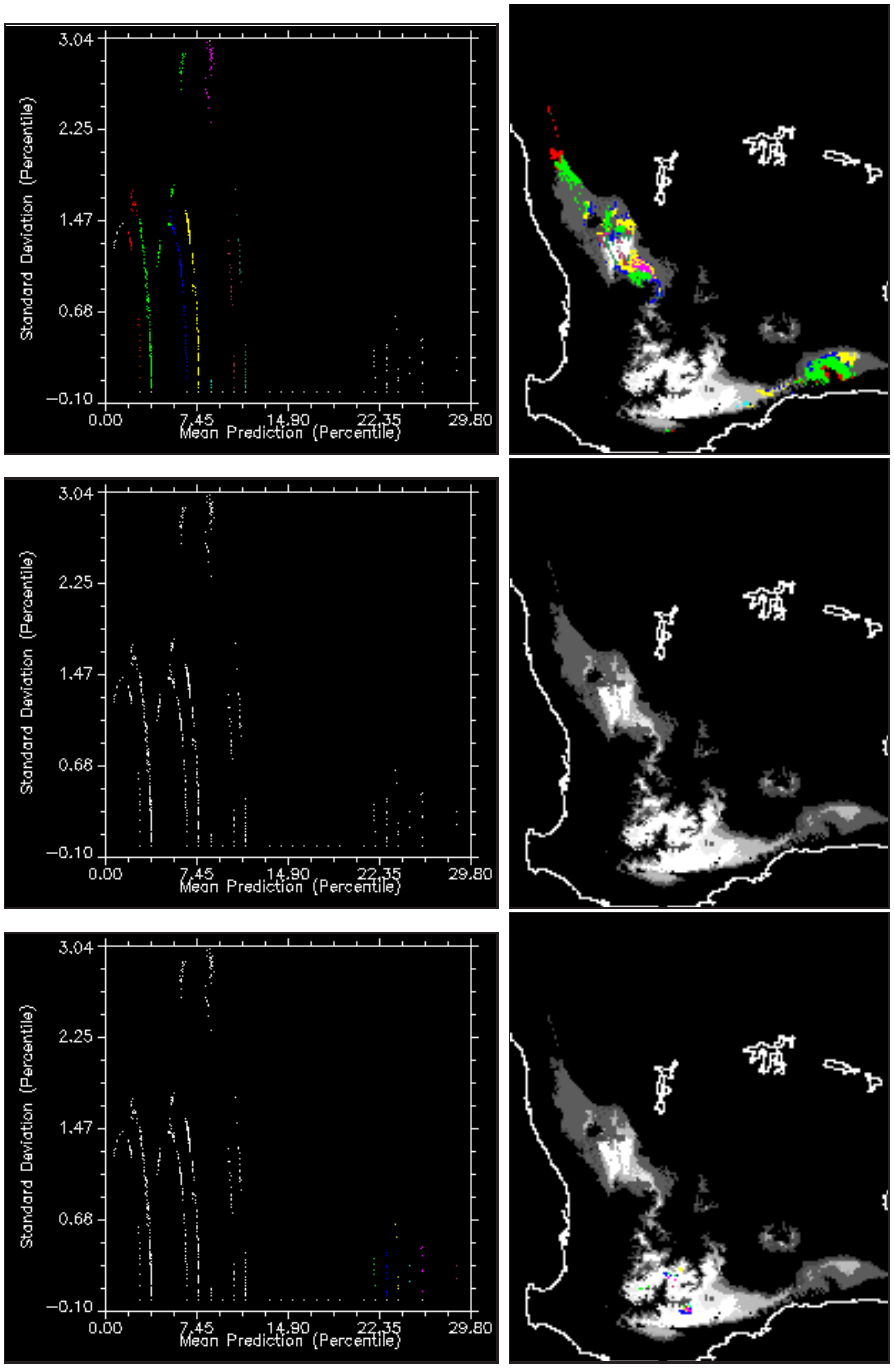


Figure 9.22: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 17. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

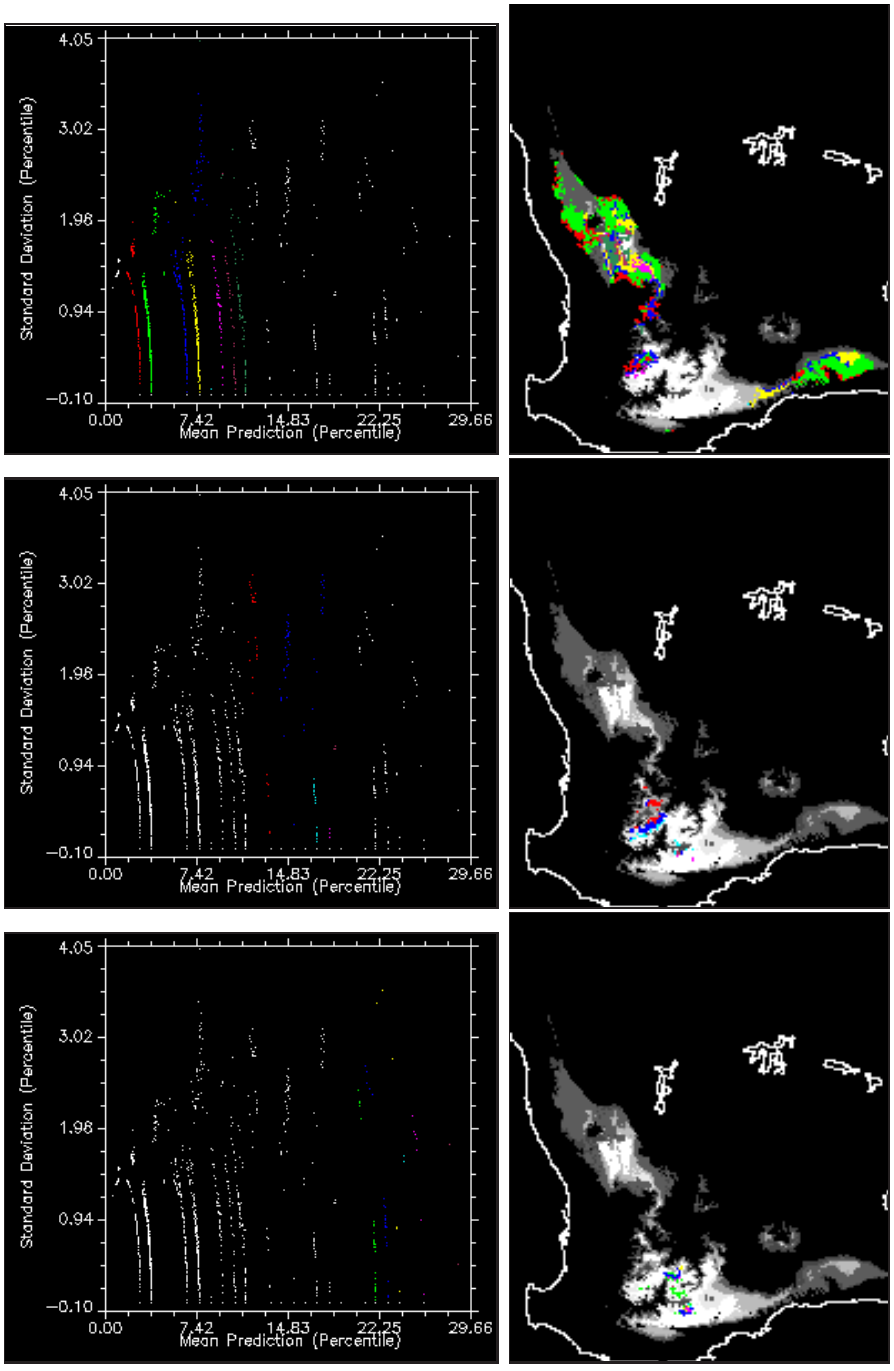


Figure 9.23: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 18. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

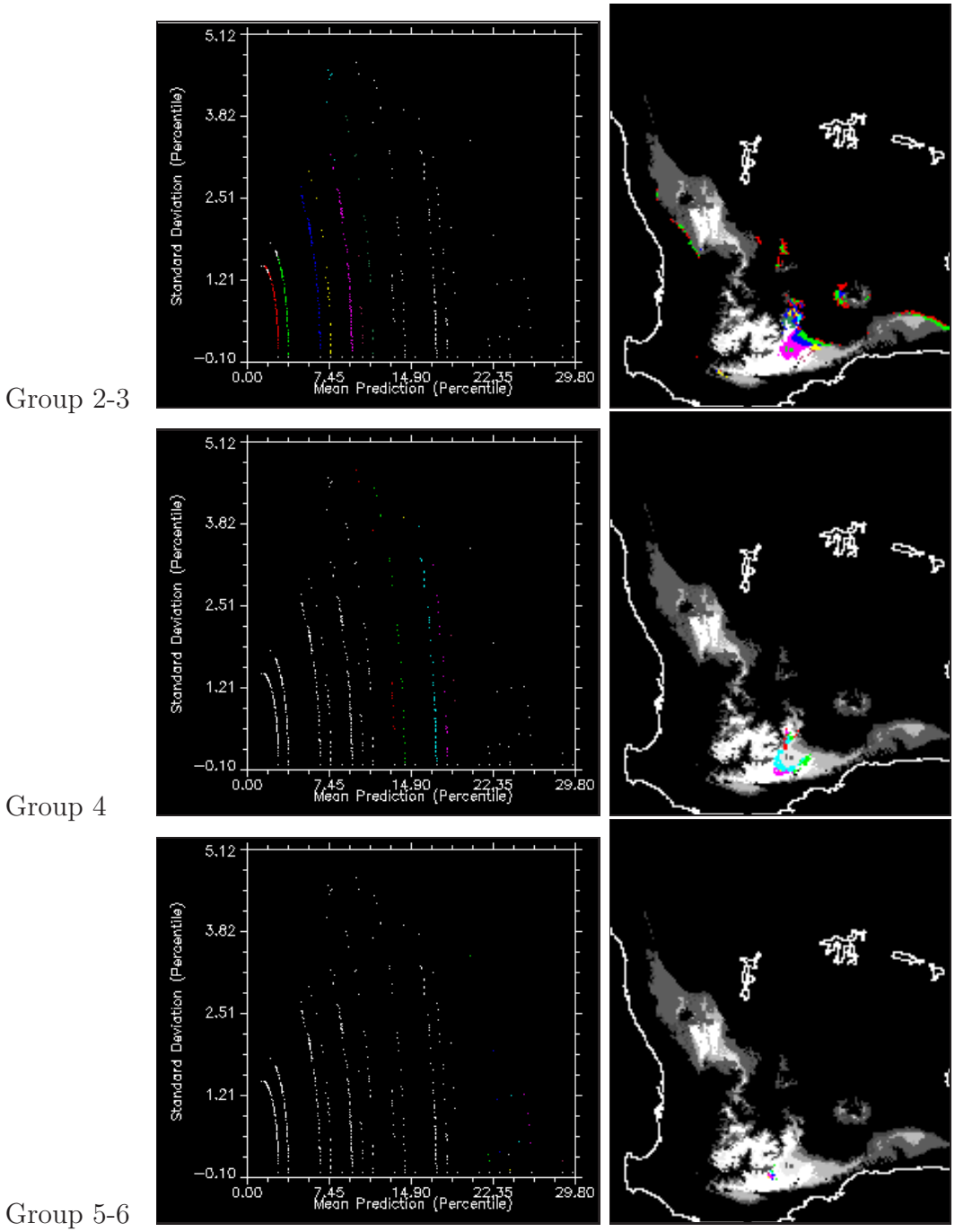


Figure 9.24: Mean Prediction versus Uncertainty at each grid cell. Single-Biogrid Model 19. For each region of interest, only where the uncertainty is greater than 0, is coloured. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

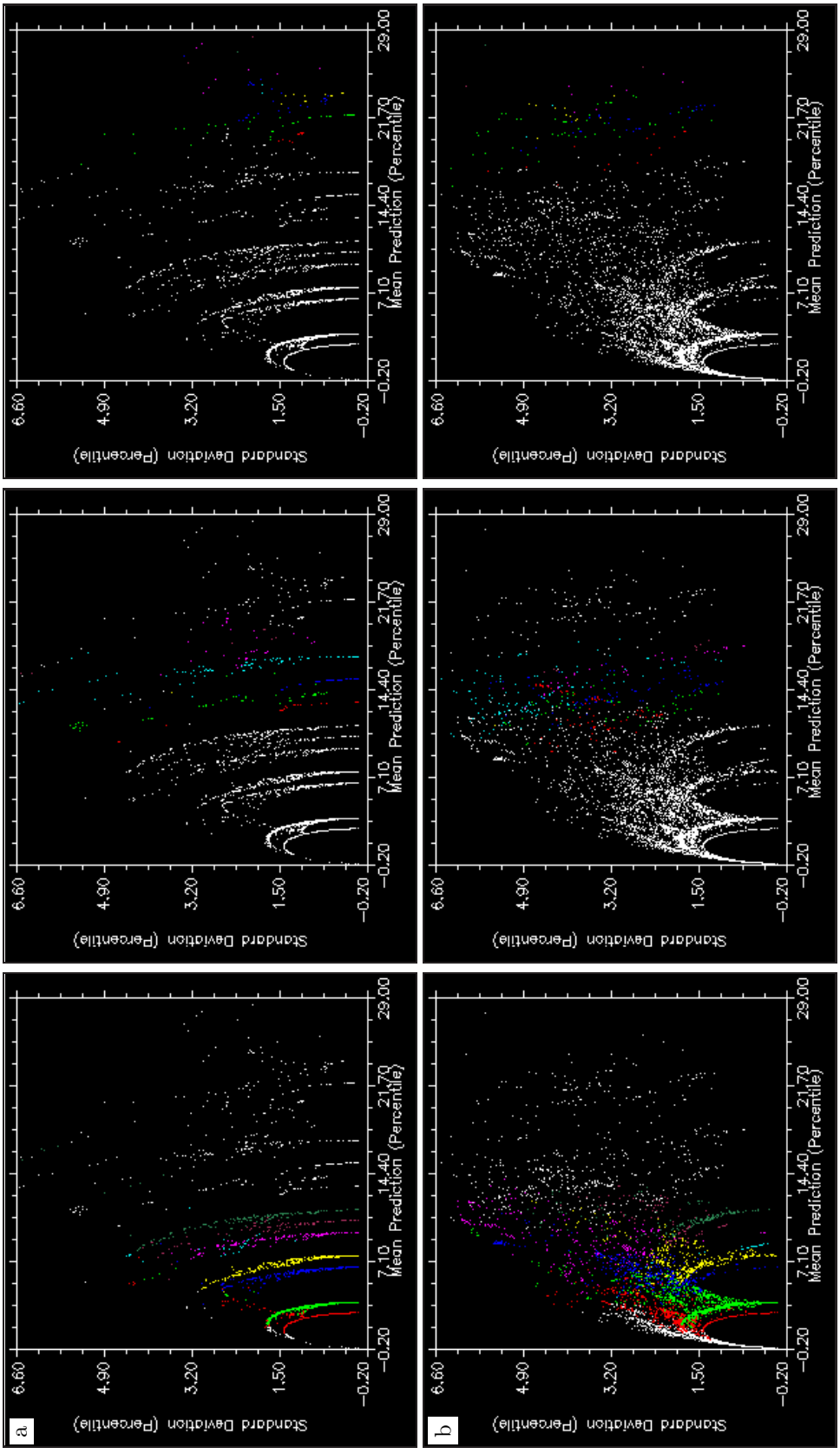


Figure 9.25: Mean to Uncertainty relationship, (a) Bioclimate-Group-1 and (b) Bioclimate-Group-2 models. Where the uncertainty is greater than 0, for each Region Of Interest, is illustrated in Figure 9.26. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

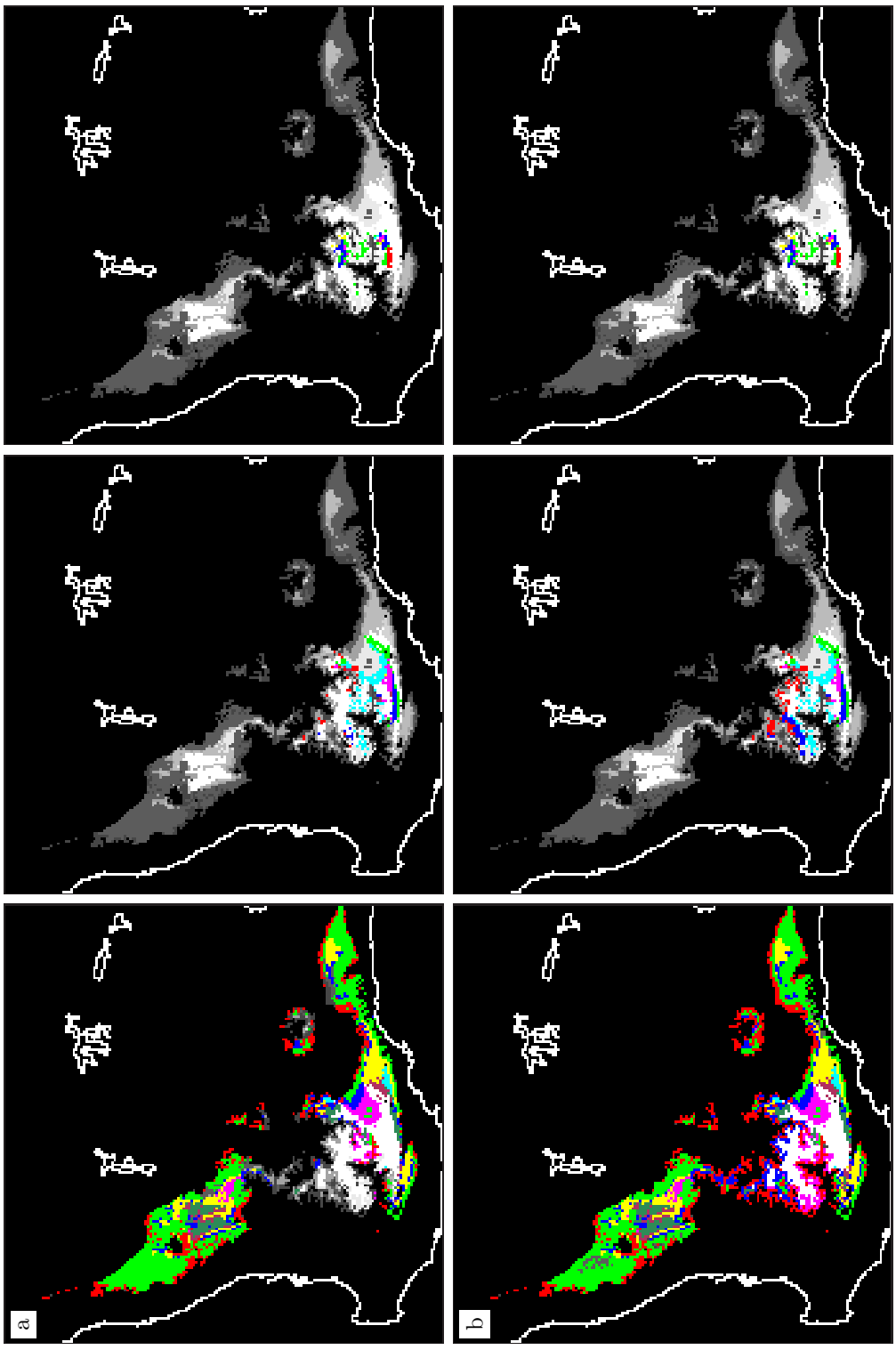


Figure 9.26: The grid cells where the uncertainty is greater than 0 in each Region of Interest: (a) Bioclimate-Group-1 and (b) Bioclimate-Group-2 models. The colour labels assigned to each Region Of Interest in each Group is shown in Table 9.4.

Chapter 10

Conclusion

The aims of this study are outlined in Chapter 1, Section 1.1. To conclude, the following outcomes have been achieved:

Computer programs have been developed to allow the analysis of uncertainty propagation in the Models of interest. Those analysed are Precision Agriculture and Ecological Niche models. The uncertainty propagated was, in all model's, in the model inputs.

The sensitivity of the models to uncertainty propagation has been quantified using two commonly used methods, Taylor Series and Monte Carlo Modeling. The results of this sensitivity analysis showed the degree to which the accuracy of the model was influenced. Also, the difference between these tool's results was investigated. These results were used to understand which components of the model's, in which way, most influenced the uncertainty in the model's product. This will be further discussed in the following two sections.

The programs developed were largely designed to suit each model being analysed. For each analysis method, the basic structure of the program is the same, so the major differences are in the unique functions or procedures written specifically for each model.

10.1 Precision Agriculture Models

For the Precision Agriculture models, both analysis methods were used because both model's algorithms were differentiable and continuous in the domain of interest. Therefore, two analysis products were produced.

As expected for the linear N-availability model, both methods showed the sensitivity of the prediction to input uncertainty to be low. Also, both methods showed this relationship to be linear.

For the more complex non-linear Mitscherlich equation, most results from both analysis methods were very similar when the uncertainty distribution of the Monte Carlo model was Gaussian; except where the predictions are greater than approximately 100 R. Further analysis showed this only occurred were the predictions were invalid, due to a limitation of the model itself. In the domain where the Mitscherlich equation is valid the uncertainty does grow linearly but the variation of this uncertainty is high. This clearly showed that the sensitivity to uncertainty in complex non linear models is higher. The combination of these analyses are useful in illustrating the accuracy of the model and the need to understand its limitations.

Using the Monte Carlo method, the distribution of the input uncertainty was shown to have a significant influence on the result. The prediction did not change to a great degree, but the uncertainty of that prediction very clearly did. The direction of the skew also produced differing results.

10.2 The BIOCLIM Model

The gridded results of a single prediction of the BIOCLIM model across the area of interest fall into clear prediction bins. With the introduction of normally distributed uncertainty into the model inputs, this clarity is lost so, the BIOCLIM model is clearly sensitive to uncertainty propagation. However, drawing an single conclusion on this sensitivity across the area studied is not possible.

Where further analysis of the climate functions was possible (the known al-

gorithms), it was concluded that their influence on the uncertainty-prediction relationship fell clearly into two categories. In the first the prediction was reduced and its uncertainty increased, by the presence of uncertainty in the input - but not at all the grid cells. In the second case, the area affected was significantly larger, the prediction was both lower and higher, and not evenly distributed (in size and spatial extent). Analysis of the functions in this second group showed that they produced a more exact "mapping" of the climate envelope of the studied species, which in part explains these results.

The size of the uncertainty domain about each grid cell value is clearly a critical component of the model. But, its size and spatial distribution is not reflected in the final model result for most of the area. This was caused by the combination of uncertainty propagation through the known functions of the model and possibly by the algorithms in the black box component of the model.

Skewing the input uncertainty distribution does not change the BIOCLIM predictions to any large degree, but it does change the uncertainties of the prediction. As in the Mitscherlich Model, the direction of the skew also produces a different sensitivity.

The uncertainty in the future predictions is low. This is expected as there is a significant difference between the present and future climate grids. Therefore, the uncertainty in the present grid is less likely to change where the BIOCLIM model boxes the future predictions.

Limiting the parts of the model through which the uncertainty could propagate allowed the most sensitive parts of the model to be isolated. In the BIOCLIM present model, the algorithms have two quite clear mean prediction versus uncertainty relationships, showing which algorithms had the greatest negative input on the validity of the prediction. Therefore, in aiming to minimise this, the initial conclusions are to:

1. Minimise the propagation of uncertainty through these algorithms. This is, in theory, most easily achieved if the uncertainty in the climate grids could be reduced and the model's algorithm(s) could be modified to minimise

uncertainty propagation.

2. Remove these algorithms. As this would result in the prediction changing in some grid locations, this is only possible if the changes occurring are small or do not occur in geographical areas of interest. Communicating this to an end user would be critical.

10.3 Future Research

The aim of this study was to quantify the uncertainty propagation in three different models, using two commonly used methods. These results were then used to investigate which components of the model's may have caused the uncertainty sensitivity observed. How the application of this analysis' results might be used in either the use and/or improvement of the models was also discussed.

Research groups interested in either the models studied or in the improvement of model accuracy, may have an interest in applying either the results and methodologies of this thesis in their research. For example,

1. BIOCLIM is being used to make a future prediction and the researcher wants to quantify the effects of uncertainty in future climate grids. This would require adding an uncertainty component to the future climate grids, such as those in the A2a projection. Unfortunately, there is no quantified uncertainty for the A2a grids produced by the Hadley, CSIRO or Canadian research group's coupled climate models. However, the difference between these three A2a projections does give an estimate of their uncertainty domain, or more specifically, the model's uncertainty. From these fields, a new prediction and associated uncertainty input could be calculated. If this methodology was applied it would be important to classify the new prediction as an estimate made from multiple model projections - not a projection itself. But, it still has value as using its uncertainty may give a valid domain across which to test the effect of uncertainty on future predictions.

2. Applying a rigorous statistical method that could map the relationship between the input uncertainty and prediction uncertainty. For example, principal component analysis is often used to quantify the relationships between fields generated in the steps of a GIS analysis.

3. Also, in implementing this analysis method, programs could be developed for use with other GIS packages and Models. Modules and filters specifically designed for each model will need to be developed. Ideally, a package could be written to import these modules as required, the analysis performed and then the results displayed. Alternatively, the analysis modules could be designed to be called by a popular Geospatial Information Package.

Appendix A

Taylor Series Anaylsis of Models

This appendix contains the 1st and 2nd order partial derivatives of the Nitrogen Availability Model and the fertiliser component of the Mitscherlich equation. If the second order derivatives is equal to 0 it is not shown.

A.1 Nitrogen Availability

.

The Nitrogen availability model is

$$N(available) = (RON \times RONDep(T-1) \times RONEff) + 10000 \times (OC \times (1 - GravProp) \times SONEff) + (15 \times FerTeff)$$

where the input data layers are the residual organic nitrogen (RON),organic carbon in the soil (OC) and the gravel proportion in the soil ($GravProp$). The other four parameters are the $RONDep$ depletion coefficient and three efficiency coefficients $RONEff$, $SonEff$ and $FertEff$. These coefficients are constants that were determined by productivity trials. The known uncertainties of the input layers and the coefficients were obtained from discussion with experts in soil testing. The other variable is time (T) in years, since the last lupin crop.

The N-available id in Kg/Ha.

A.1.1 Residual Organic Nitrogen

First Order.

$$\frac{\partial N}{\partial RON} = (0.3^{(T-1)}) \times RONEff \tag{A.1}$$

Second Order.

$$\frac{\partial^2 N}{\partial RONEff\ \partial RON} = 0.3^{(T-1)} \tag{A.2}$$

A.1.2 RONEff

First Order.

$$\frac{\partial N}{\partial RONEff} = RON \times (0.3^{(T-1)}) \tag{A.3}$$

Second Order.

$$\frac{\partial N}{\partial RON\ \partial RONEff} = 0.3^{(T-1)} \tag{A.4}$$

A.1.3 OC

First order.

$$\frac{\partial N}{\partial OC} = 10000.0 \times (1.0 - GravProp) \times SonEff \tag{A.5}$$

Second Order.

$$\frac{\partial^2 N}{\partial GravProp\ \partial OC} = (-10000.0) \times SonEff \tag{A.6}$$

$$\frac{\partial^2 N}{\partial SonEff\ \partial OC} = SonEff = 10000.0 \times (1.0 - GravProp) \tag{A.7}$$

A.1.4 GravProp

First order.

$$\frac{\partial N}{\partial GravProp} = (-10000.0) \times OC \times SonEff \tag{A.8}$$

Second Order.

$$\frac{\partial^2 N}{\partial OC\ \partial GravProp} = OC = (-10000) \times SonEff \tag{A.9}$$

$$\frac{\partial^2 N}{\partial SonEff\ \partial GravProp} = SonEff = (-10000) \times OC \tag{A.10}$$

A.1.5 SONEff

First order.

$$\frac{\partial N}{\partial SONEff} = 10000.0 \times (OC * (1.0 - GravProp)) \tag{A.11}$$

Second Order.

$$\frac{\partial^2 N}{\partial OC\ \partial SONEff} = OC = 10000.0 \times (1.0 - GravProp) \tag{A.12}$$

$$\frac{\partial^2 N}{\partial GravProp\ \partial SONEff} = GravProp = (-10000) \times OC \tag{A.13}$$

A.1.6 FerTeff

First order.

$$\frac{\partial N}{\partial FerTeff} = 15 \tag{A.14}$$

A.2 The Mitscherlich model.

The fertiliser component of the Mitscherlich model is

$$Y = A - B^{-CR} \tag{A.15}$$

where Y is the yield in Tonnes per Hectare; A is the maximum achievable yield with no other limitations; B is the response to potassium; C is a curvature parameter; and R is the rate of applied fertiliser.

It has been shown (Edwards 1997) that the response, B, to potassium fertiliser for a range of paddocks in the Australian wheat belt may be determined by Equation 3.3,

$$B = A(0.95 + 2.6 \times e^{-0.095 \times K_0}) \tag{A.16}$$

where K_0 is the soil potassium level. Substituting Equation 3.3 into Equation 3.2 and inverting provides a means of calculating the potassium requirements for any location with any given soil potassium value. This is shown in Equation 3.4,

$$R = \frac{-1}{C} \times \ln[\frac{Y_t - A}{-A(0.95 + 2.6e^{(-0.095K_0)})}] \tag{A.17}$$

where R is the fertiliser requirement (Kg/Ha) to achieve a target yield of Y_t Tonnes per Hectare.

A.2.1 Curvature Parameter (C)

First order.

$$\frac{\partial R}{\partial C} = \frac{1}{C^2} \times \ln(\frac{Y_t - A}{-1 \times B}) \tag{A.18}$$

Second order.

$$\frac{\partial^2 R}{\partial A\ \partial C} = \frac{1}{C^2} \times [\frac{Y_t}{A \times (Y_t - A)}] \tag{A.19}$$

$$\frac{\partial^2 R}{\partial K\ \partial C} = \frac{1}{C^2} \times [\frac{0.247 \times \ln(-0.095 \times K_0)}{0.05 + (2.6 \times \ln(-0.095 \times K_0))}] \tag{A.20}$$

$$\frac{\partial^2 R}{\partial^2 C} = \frac{-2}{C^3} \times \ln(\frac{Y_t - A}{-1 \times B}) \tag{A.21}$$

A.2.2 Maximum Achieveable Yield (A)

First order.

$$\frac{\partial R}{\partial A} = \frac{-1}{C} \times [\frac{Yt}{A \times (Y_t - A)}] \tag{A.22}$$

Second order.

$$\frac{\partial^2 R}{\partial K\ \partial A} = \frac{1}{C} \times [\frac{1}{(Y_t - A)^2}] \tag{A.23}$$

$$\frac{\partial^2 R}{\partial C\ \partial A} = \frac{1}{C^2} \times [\frac{Y_t}{A \times (Y_t - A)}] \tag{A.24}$$

$$\frac{\partial^2 R}{\partial^2 A} = \frac{-1 \times Y_t}{C} \times [\frac{(2 \times A) - Y_t}{A^2 \times (Y_t - A)^2}] \tag{A.25}$$

A.2.3 Soil Potassium Level K_0

First order.

$$\frac{\partial R}{\partial K_0} = \frac{1}{C} \times \frac{0.247 \ln((-0.095) \times K_0)}{0.05 + (2.6 \ln((-0.095) \times K_0))} \tag{A.26}$$

Second Order.

$$\frac{\partial^2 R}{\partial C\ \partial K_0} = \frac{-1}{C^2} \times \frac{0.247 \ln((-0.095) \times K_0)}{0.05 + (2.6 \ln((-0.095) \times K_0))} \tag{A.27}$$

$$\frac{\partial^2 R}{\partial^2 K_0} = \frac{1}{C} \times (X + Y) \tag{A.28}$$

where

$$X = \frac{0.023465 \ln(-0.095 \times K_0)}{0.05 + (2.6 \ln(-0.095 \times K_0))} \tag{A.29}$$

and

$$Y = -1 \times [\frac{(-0.247 \ln(-0.095 \times K_0)) \times (0.247 \ln(-0.095 \times K_0))}{(0.05 + (2.6 \ln(-0.095 \times K_0)))^2}] \tag{A.30}$$

Appendix B

AML BIOCLIM Variables Code

The 19 BIO layers used in this study and how they are calculated, is shown in this appendix. It is a copy of the original documentation from the DIVA-GIS website.

```
/* MkBCvars.AML
/* /*
/* Author Robert Hijmans
/* January 2006
/* rhijmans@uclink.berkeley.edu
/*
/* Version 2.3
/*
/* This AML creates the 19 BIOCLIM variables from
/* monthly Tmin, Tmax, and Precipitation grids
/* The results are rounded where integers would become reals
/* (I assume that input values were multiplied by 10
/* and stored as Integers to begin with)
/* P2 is first multiplied by 10
/* CVs are first multiplied by 100
/*
/* rounding of "x" is done with "int(floor(x + 0.5))"
/* because "int(x+0.5)" as suggested by ESRI (see INT in Arc Help), does not
/* round negative numbers correctly (-2.6 -> -2 intstead of -3).
/*
/* You must change the first four lines (input files and output directory)
/* If you do not have average temperature, create it with the lines that follow
/*
/* Also note that the AML removes some temporary grids if they exist
/* (the first "&do i = 0 &to 15" bit)
```

```
/* Please make sure that you do not have files
/* with those names that you want to keep.
/*
/* BIO1 = Annual Mean Temperature
/* BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
/* BIO3 = Isothermality (P2/P7) (* 100)
/* BIO4 = Temperature Seasonality (standard deviation *100)
/* BIO5 = Max Temperature of Warmest Month
/* BIO6 = Min Temperature of Coldest Month
/* BIO7 = Temperature Annual Range (P5-P6)
/* BIO8 = Mean Temperature of Wettest Quarter
/* BIO9 = Mean Temperature of Driest Quarter
/* BIO10 = Mean Temperature of Warmest Quarter
/* BIO11 = Mean Temperature of Coldest Quarter
/* BIO12 = Annual Precipitation
/* BIO13 = Precipitation of Wettest Month
/* BIO14 = Precipitation of Driest Month
/* BIO15 = Precipitation Seasonality (Coefficient of Variation)
/* BIO16 = Precipitation of Wettest Quarter
/* BIO17 = Precipitation of Driest Quarter
/* BIO18 = Precipitation of Warmest Quarter
/* BIO19 = Precipitation of Coldest Quarter
/*
/* These summary Bioclimatic variables are after:
/* Nix, 1986. A biogeographic analysis of Australian elapid snakes. In: R. Longmore (ed.).
/* Atlas of elapid snakes of Australia. Australian Flora and Fauna Series 7.
/* Australian Government Publishing Service, Canberra.
/*
/* and Expanded following the ANUCLIM manual
/*
/*
/* Temperature data is in units of oC * 10 because that allows me to store the data as Integer values,
/* (with 0.1oC precision) which is more efficient than storing the data as Real values.
/* However, you will want to report the data in oC. Precipitation data is in mm.
/*
/*

&TERMINAL 9999

&s program [locase [show program]]
&if %program% ^= grid &then grid
```

```
&sv tn = tmin\tmin_
&sv tx = tmax\tmax_
&sv ta = tmean\tmean_
&sv pt = prec\prec_

/* if TAVG does not exist
&do j = 1 &to 12
    &if [EXISTS %ta%%j% -grid] &then &type %ta%%j%
    &else %ta%%j% = (%tn%%j% + %tx%%j%) / 2
&end

&do i = 0 &to 20
/* &if [exists BIO%i% -grid] &then kill P%i%
/* &if [exists P%i% -grid] &then kill P%i%
/* &if [exists tmp%i% -grid] &then kill tmp%i%
/* &if [exists x%i% -grid] &then kill x%i%
/* &if [exists q%i% -grid] &then kill q%i%
/* &if [exists t%i% -grid] &then kill t%i%
/* &if [exists mnt%i% -grid] &then kill mnt%i%
/* &if [exists dry%i% -grid] &then kill dry%i%
/* &if [exists wet%i% -grid] &then kill wet%i%
/* &if [exists hot%i% -grid] &then kill hot%i%
/* &if [exists cld%i% -grid] &then kill cld%i%
/* &if [exists x%i% -grid] &then kill x%i%
/* &if [exists y%i% -grid] &then kill y%i%
/* &if [exists rg%i% -grid] &then kill rg%i%
&end

&if [exists drym -grid] &then kill drym
&if [exists wetm -grid] &then kill wetm

&sv TAvar = %ta%1
&sv TXvar = %tx%1
&sv TNvar = %tn%1
&sv PTvar = %pt%1

&do j = 2 &to 12
    &sv tavar = %tavar%,%ta%%j%
    &sv txvar = %txvar%,%tx%%j%
    &sv tnvar = %tnvar%,%tn%%j%
    &sv ptvar = %ptvar%,%pt%%j%
&end

/* P1. Annual Mean Temperature
&if [exists p1 -grid] &then &type P1 exists
```

```
&else
&do
    P1 = int(floor(mean(%tavar%) + 0.5))
    &type P1 done
&end

/* P4. Temperature Seasonality (standard deviation)
&if [exists p4 -grid] &then &type P4 exists
&else
&do
    P4 = int(floor(100 * std(%tavar%) + 0.5))
    &type P4 done
&end

/* P5. Max Temperature of Warmest Period
&if [exists p5 -grid] &then &type P5 exists
&else
&do
    P5 = max(%txvar%)
    &type P5 done
&end

/* P6. Min Temperature of Coldest Period
&if [exists p6 -grid] &then &type P6 exists
&else
&do
    P6 = min(%tnvar%)
    &type P6 done
&end

/* P7. Temperature Annual Range (P5-P6)
&if [exists p7 -grid] &then &type P7 exists
&else
&do
    P7 = P5 - P6
    &type P7 done
&end

/* P12. Annual Precipitation
&if [exists p12 -grid] &then &type P12 exists
&else
&do
    P12 = sum(%ptvar%)
    &type P12 done
&end
```

```
/* P13. Precipitation of Wettest Period
&if [exists p13 -grid] &then &type P13 exists
&else
&do
    P13 = max(%ptvar%)
    &type P13 done
&end

/* P14. Precipitation of Driest Period
&if [exists p14 -grid] &then &type P14 exists
&else
&do
    P14 = min(%ptvar%)
    &type P14 done
&end

/* P15. Precipitation Seasonality(Coefficient of Variation)
/* the "1 +" is to avoid strange CVs for areas where mean rainfaill is < 1)
&if [exists p15 -grid] &then &type P15 exists
&else
&do
    P15 = int(floor(100 * std(%ptvar%) / (1 + P12 / 12) + 0.5))
    &type P15 done
&end

&do i = 1 &to 12
    &if [exists rg%i% -grid] &then &type rg%i% exists
    &else rg%i% = %tx%%i% - %tn%%i%
&end

/* P2. Mean Diurnal Range(Mean(period max-min))
&if [exists p2 -grid] &then &type P2 exists
&else
&do
    P2 = int(floor(mean(rg1,rg2,rg3,rg4,rg5,rg6,rg7,rg8,rg9,rg10,rg11,rg12) + 0.5))
    &type P2 done
&end

/* P3. Isothermality (P2 / P7)
&if [exists p3 -grid] &then &type P3 exists
&else
&do
    P3 = int(floor(100 * P2 / P7) + 0.5)
    &type P3 done
```

```
&end

&do i = 1 &to 12
kill rg%i%
&end

&do i = 1 &to 12
    &sv j = %i%
    &sv k = [calc %i% + 1]
    &sv l = [calc %i% + 2]
    &if %k% > 12 &then &sv k = [calc %k% - 12]
    &if %l% > 12 &then &sv l = [calc %l% - 12]
    q%i% = %pt%%j% + %pt%%k% + %pt%%l%
    t%i% = %ta%%j% + %ta%%k% + %ta%%l%
&end

mnt0 = con(isnull(q1),0,100)
mnt1 = setnull(mnt0 ¡ 1, 1)
wet1 = q1

&do i = 1 &to 11
    &sv j = [calc %i% + 1]
/* &type i = %i% and j = %j%
    mnt%j% = con(q%j% > wet%i%, [calc %j%], mnt%i%)
    wet%j% = con(q%j% > wet%i%, q%j%, wet%i%)
&end
wetm = mnt12

/* P16. Precipitation of Wettest Quarter
&if [exists p16 -grid] &then &type P16 exists
&else
&do
    P16 = wet12
    &type P16 done
&end

&do i = 1 &to 12
    kill mnt%i%
    kill wet%i%
&end

mnt1 = setnull(mnt0 < 1, 1)
dry1 = q1
&do i = 1 &to 11
    &sv j = [calc %i% + 1]
```

```
    mnt%j% = con(q%j% < dry%i%, [calc %j%], mnt%i%)
    dry%j% = con(q%j% < dry%i%, q%j%, dry%i%)
&end
drym = mnt12

/* P17. Precipitation of Driest Quarter
&if [exists p17 -grid] &then &type P17 exists
&else
&do
    P17 = dry12
    &type P17 done
&end

&do i = 1 &to 12
    kill mnt%i%
    kill dry%i%
&end
kill mnt0

&do i = 1 &to 12
    x%i% = con(wetm == %i%, t%i%, -9999)
    y%i% = con(drym == %i%, t%i%, -9999)
&end

/* tmp1 = max(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12)
/* tmp2 = tmp1 / 3
/*P8 = int(floor(tmp2 + 0.5))

/* P8. Mean Temperature of Wettest Quarter
&if [exists p8 -grid] &then &type P8 exists
&else
&do
    P8 = int(floor(max(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12) / 3 + 0.5))
    &type P8 done
&end
/* tmp3 = max(y1,y2,y3,y4,y5,y6,y7,y8,y9,y10,y11,y12)
/* tmp4 = tmp3 / 3
/* P9 = int(floor(tmp4 + 0.5))

/* P9. Mean Temperature of Driest Quarter
P9 = int(floor(max(y1,y2,y3,y4,y5,y6,y7,y8,y9,y10,y11,y12) / 3 + 0.5))
&type P9 done

&do i = 1 &to 12
    kill x%i%
```

```
    kill y%i%
&end
&do i = 1 &to 4
&if [exists tmp%i% -grid] &then kill tmp%i%
&end

mnt0 = con(isnull(t1),0,100)
mnt1 = setnull(mnt0 < 1, 1)
hot1 = t1
&do i = 1 &to 11
    &sv j = [calc %i% + 1]
    mnt%j% = con(t%j% > hot%i%, [calc %j%], mnt%i%)
    hot%j% = con(t%j% > hot%i%, t%j%, hot%i%)
&end
hotm = mnt12

/* P10 Mean Temperature of Warmest Quarter
&if [exists p10 -grid] &then &type P10 exists
&else
&do
    P10 = int(floor(hot12 / 3 + 0.5))
    &type P10 done
&end

&do i = 1 &to 12
    kill mnt%i%
    kill hot%i%
&end

mnt1 = setnull(mnt0 < 1, 1)
cld1 = t1

&do i = 1 &to 11
    &sv j = [calc %i% + 1]
    mnt%j% = con(t%j% < cld%i%, [calc %j%], mnt%i%)
    cld%j% = con(t%j% < cld%i%, t%j%, cld%i%)
&end
cldm = mnt12

/* P11 Mean Temperature of Coldest Quarter
&if [exists p11 -grid] &then &type P11 exists
&else
&do
    P11 = int(floor(cld12 / 3 + 0.5))
    &type P11 done
```

```
&end

&do i = 1 &to 12
    kill mnt%i%
    kill cld%i%
&end
kill mnt0

&do i = 1 &to 12
    x%i% = con(hotm == %i%, q%i%, -9999)
    y%i% = con(cldm == %i%, q%i%, -9999)
&end

tmp1 = max(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12)

/* P18. Precipitation of Warmest Quarter
&if [exists p18 -grid] &then &type P18 exists
&else
&do
    P18 = int(floor(tmp1 + 0.5))
    &type P18 done
&end

tmp2 = max(y1,y2,y3,y4,y5,y6,y7,y8,y9,y10,y11,y12)

/* P19. Precipitation of Coldest Quarter
&if [exists p19 -grid] &then &type P19 exists
&else
&do
    P19 = int(floor(tmp2 + 0.5))
    &type P19 done
&end

&do i = 1 &to 12
    kill x%i%
    kill y%i%
&end

kill hotm
kill cldm
kill drym
kill wetm

kill tmp1 kill tmp2
```

```
&do i = 1 &to 12
    kill q%i%
    kill t%i%
&end

&do i = 1 &to 19
    rename p%i% bio_%i%
&end

&type Done!
&return
```

Appendix C

Annual Mean Temperature Header

[General]
Creator=DIVA-GIS
Created=20081012
Title=BIO1

[GeoReference]
Projection=
Datum=
Mapunits=
Columns=171
Rows=203
MinX=114.66668
MaxX=121.79168
MinY=-35.000004
MaxY=-26.541670
ResolutionX=0.041666668
ResolutionY=0.041666668

[Data]
DataType=FLT4BYTES
MinValue=140.000
MaxValue=225.000
NoDataValue=-3.4E38
Transparent=0
Units=

[Application]
Opt0=Procedure: Climate data to map

Opt1=Climate: frmClim2Grid
Opt2=Variable: Annual mean temperature range [1]

[ContLegend]
Count=2
Color1=16711680
Value1=140.000
Label1=
Color2=255
Value2=225.000
Label2=

[Legend]
Count=5
Color1=255
Min1=140
Max1=157
Label1=140.0 - 157.0
Color2=65450
Min2=157
Max2=174
Label2=157.0 - 174.0
Color3=65280
Min3=174
Max3=191
Label3=174.0 - 191.0
Color4=16755200
Min4=191
Max4=208
Label4=191.0 - 208.0
Color5=16711680
Min5=208
Max5=225
Label5=208.0 - 225.0
Transparent=0
NoDataColor=0
NoDataLabel=No Data
isContinuous=0
SpacedByColor=0

Bibliography

Adams, M. L., Cook, S. & Bowden, J. W. (2000), 'Using Yield Maps and Intensive Soil Sampling to Improve Nitrogen Fertiliser Recommendations from a Deterministic Model in the Western Australian Wheatbelt', *Australian Journal of Experimental Agricultural* **40**(7), 959–968.

Adams, M. L., Cook, S. & Corner, R. (2000), 'Managing Uncertainty in Site-Specific Management: What is the Best Model', *Precision Agriculture* **2**, 39–54.

Addiscott, T. & Tuck, G. (1996), Sensitivity analysis for regional-scale solute transport modeling, *in* D. Corwin & K. Loague, eds, 'Applications of GIS to the modeling of non-point source pollutants in the vadose zone', Soil Science Society of America Inc., Madison, U.S.A., pp. 153–162.

Aguilar, F. J., Aguilar, M. A. & Aguera, F. (2007), 'Accuracy assessment of digital elevation models using a non-parametric approach', *International Journal of Geographical Information Science* **21**(6), 667–686.

Agumya, A. & Hunter, G. J. (2002), 'Responding to the consequences of uncertainty in geographical data', *International Journal of Geographical Information Science* **16**(5), 405–417.

Anton, H. (1984), *Calculus*, John Wiley and Sons, New York.

Araújo, M. B., Pearson, R. G., Thuiller, W. & Erhard, M. (2005), 'Validation of speciesclimate impact models under climate change', *Global Change Biology* **11**, 15041513.

ARCMAP 9.2 (2004). ESRI.

Arnell, N. W. (1999), 'Climate change and global water resources', *Global Environmental Change* **9**, S31–S49.

Arras, K. A. (1998), An Introduction To Error Propagation: Derivation, Meaning and Examples of Equation $c_y = f_x c_x f_x^T$, Technical Report No. EPFL-ASL-TR-98-01 R3, Swiss Federal Institute of Technology, Lausanne.

Atkinson, P. M. (2005), 'Spatial Prediction and Surface Modeling', *Geographical Analysis* **37**, 113–123.

Austin, M. P. (1992), 'Modelling the environmental niche of plants − implications for plant community response to elevated *co*2 levels', *Experimental Methods* **40**, 615–630.

Austin, M. P. (2002), 'Spatial prediction of species distribution: an interface between ecological theory and statistical modelling', *Experimental Methods* **157**, 101–118.

Austin, M. P. & Heylingers, P. C. (1989), 'Vegetation survey design for conservation: gradsect sampling of forests in north-eastern New South Wales', *Biological Conservation* **50**, 13–32.

Austin, M. P., Nicholls, A. O. & Margules, C. R. (1990), 'Measurement of the realized qualitative niche: environmental niches of five *eucalyptus* species', *Experimental Methods* **60**, 161–177.

Bailey, T. C. & Gatrell, A. C. (1995), *Interactive Spatial Data Analysis*, Pearson, Harlow, England.

Bakkenes, M., Alkemade, J. R. M., Ihle, F., Leemans, R. & Latour, J. B. (2002), 'Assessing effects of forecasted climate change on the diversity and distribution of european higher plants for 2050', *Global Change Biology* **8**, 390–407.

Barry, S. & Elith, J. (2006), 'Error and uncertainty in habitat models', *Journal of Applied Ecology* **43**, 413–423.

Beaumont, L. J., Hughes, L. & Poulsen, M. (2005), 'Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions', *Experimental Methods* **186**, 250–269.

Beerling, D. J., Huntley, B. & Bailey, J. P. (1995), 'Climate and the distribution of $Fallopiajaponica$: use of an introduced species to test the predictive capacity of response surfaces', *Journal of Vegetation Science* **6**, 269–282.

Bosma, W., Marinussen, M. & van der Zee, S. (1994), 'Simulation and areal interpolation of reactive solute transport', *Geoderma* **62**, 217–231.

Boyer, T. & Levitus, S. (1994), Quality Control and Processing of Historical Oceanographic Temperature, Salinity, and Oxygen Data, Technical Report NESDIS 81, NOAA, Washington D.C.

Brennan, R. F. & Bolland, D. A. (2003), 'Soil properties as predictors of yeild response of clover (Trifolium subterraneum L.) to added P in soils of varying P sorption capacity', *Australian Journal of Soil Research* **41**, 653–663.

Brown, J. F., Loveland, T. R., Merchant, J. W., Reed, B. C. & Ohlen, D. O. (1993), 'Using multisource data in global land-cover characterization: concepts, requirements, and methods', *Photogrammetric Engineering & Remote Sensing* **59**(6), 977–978.

Burrough, P. A. & McDonnell, R. A. (1998), *Principles of Geographical Information Systems*, Oxford University Press, Oxford.

Busby, J. R. (1991), BIOCLIM − a bioclimate analysis and prediction system, *in* C. R. Margules & M. P. Austin, eds, 'Nature Conservation: Cost Effective Biological Surveys and Data Analysis', CSIRO, Canberra, Australia.

Carey, P. (1996), '*disperse*: A cellular automaton for predicting the distribution of species in a changed climate', *Global Ecology and Biogeography Letters* **5**, 217–226.

Carpenter, G., Gillison, A. N. & Winter, J. (1993), 'DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals', *Biodiversity and Conservation* **2**, 667–680.

Cawsey, E., Austin, M. & Baker, B. (2002), 'Regional vegetation mapping in australia: a case study in the practical use of statistical modelling', *Biodiversity and Conservation* **11**, 2239–2274.

Climate Change 2001, IPCC Third Assesment Report (2001), `http://http://www.grida.no/publications/other/ipcc_tar/`.

Collingham, Y. C., Hill, M. O. & Huntley, B. (1996), 'The migration of sessile organisms: A simulation model with measurable parameters', *Journal of Vegetation Science* **7**, 831–846.

Corner, R. J. (1997), Remote Sensing and Geographic Information Systems for Precision Agriculture, *in* 'Precision Agriculture, What can it offer the Australian Sugar Industry?', CSIRO Land and Water Division, Townsville, Australia, pp. 51–59.

Corner, R. J., Hickey, R. J. & Cook, S. E. (2002), 'Knowledge Based Soil Attribute Mapping In GIS: The Expector Method', *Transactions in GIS* **6**(4), 383–402.

Corner, R. & Marinelli, M. (2008), Modelling the Effects of Data Uncertainty on Agricultural and Environmental Models Under Global Change Conditions, Digital Information Summit on Geoinformatics, Wissenschaftspark Albert Einstein, Potsdam, Germany. Conference paper available from the Department of Spatial Science, Curtin University of Technology. Bentley. Westerm Australia.
URL: *http://www.isde-summit-2008.org*

Darwin, R. & Kennedy, D. (2000), 'Economic effects of CO2 fertilization of crops: transforming changes in yield into changes in supply', *Environmental Modelling and Assessment* **5**, 157–168.

Davis, A. J., Jenkinson, L. S., Lawton, J. H., Shorrocks, B. & Wood, S. (1998), 'Making mistakes when predicting shifts in species range in response to global warning', *Nature* **391**, 783–786.

Davis, J. C. (2002), *Statistics and Data Analysis in Geology*, John Wiley & Sons, New York.

Davis, M. B. & Shaw, R. G. (2001), 'Range Shifts and Adaptive Responses to Quaternary Climate Change', *Science* **292**, 673–679.

DIVA-GIS (2005), Exercise 2. Modelling the range of wild peanuts, Technical report, CIP (International Potato Center, Peru). DIVA-GIS tutorial.
URL: *http://www.diva-gis.org/documentation*

DIVA-GIS (2005). DIVA-GIS geographic information system (GIS).
URL: *http://www.diva-gis.org/*

Dobran, F. (1995), A risk assesment methodology at Vesuvius based on the global volcanic simulation., *in* T. Horlick-Jones, ed., 'Natural Risk and Civil Protection', E & FN spon, London, U.K., pp. 131–136.

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G. & Vose, R. S. (2010), 'Comprehensive Automated Quality Assurance of Daily Surface Observations', *Journal of Applied Meteorology and Climatology* **49**(8), 1615–1633.

Easterling, D. R. & Peterson, T. C. (1995), 'A new method for detecting and adjusting for undocumented discontinuities in climatological time series', *International Journal of Climatology* **15**, 369–377.

Easterling, D. R., Peterson, T. C. & Karl, T. R. (1996), 'On the Development and Use of Homogenised Climate Datasets', *Journal of Climate* **9**, 1429–1434.

Edwards, N. K. (1997), Potassium fertiliser improves wheat yield and grain quality on duplex soils, *in* 'Proceedings of the 1st workshop on potassium in Australian agriculture', UWA Press, Perth.

Eischeid, J., Baker, C. B., Karl, T. & Diaz, H. F. (1995), 'The Quality Control of Long−Term Climatological Data Using Objective Data Analysis', *Journal of Applied Meteorology* **34**, 2787–2795.

Elith, J., Graham, C. H., Anderson, R. P., Dudk, 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.

Elith, J. & Leathwick, J. (2010), 'Species Distribution Models: Ecological Explanation and Prediction Across Space and Time', *Annual Review, Ecological Evolutionary Systems* **40**, 677–697.

ENVI 4.4 (2007). ITT Visual Information Systems.

Errico, R. M. (1997), 'What Is an Adjoint Model', *Bulletin of the American Meteorogical Society* **78**(11), 2577–2591.

Etterson, J. R. & Shaw, R. G. (2001), 'Constraint to Adaptive Evolution in Response to Global Warming', *Science* **294**, 151–154.

Farber, O. & Kadmon, R. (2003), 'Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance', *Experimental Methods* **160**, 115–130.

Ferrier, S., Drielsma, M., Manion, G. & Watson, G. (2002), 'Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling', *Biological Conservation* **11**, 2309–2338.

Ferrier, S., Watson, G., Pearce, J. & Michael (2002), 'Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling', *Biological Conservation* **11**, 22752307.

Flato, G., Boer, G., Lee, W., Mcfarlane, N., Ramsden, D., Reader, M. & Weaver, A. (2000), 'The Canadian Centre for Climate Modelling and Analysis global coupled model and its climate', *Climate Dynamics* **16**, 451–467.

French, R. J. & Schultz, J. E. (1984), 'Water Use Efficiency of Wheat in a Mediterranean−type Environment. The Relation between Yeild, Water Use and Climate', *Australian Journal of Agricultural Research* **35**, 743–764.

Gabert, P., Papes, M. & Peterson, A. T. (2006), 'Natural history collections and the conservation of poorly known taxa: Ecological niche modeling in central African rainforest genets (Genetta spp.)', *Biological Conservation* **130**, 106–117.

Geoscience Australia (2011).
URL: *http://www.ga.gov.au/*

Gordon, C. C., Senior, C., Banks, H., Gregory, J., Johns, T., Mitchell, J. & Wood, R. (2000), 'The simulation of SST, sea ice extents and ocean heat transports in a version of the hadley centre coupled model without flux adjustments.', *Climate Dynamics* **16**, 147–168.

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 occurence data used in distribution models', *Journal of Applied Ecology* **45**, 239–247.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A. T. (2004), 'New developments in museum-based informatics and applications in biodiversity analysis', *Trends in Ecology and Evolution* **19**, 497–503.

Groisman, P. Y., Karl, T. R. & Knight, R. W. (1994), 'Observed Impact of Snow Cover on the Heat Balance and the Rise of Continental Spring Temperatures', *Science* **263**, 198–200.

Guisan, A., Edwards, T. C. & Hastie, T. (2002), 'Generalized linear and generalized additive models in studies of species distributions: setting the scene', *Experimental Methods* **157**, 89–100.

Guisan, A., Theurillat, J. & Kienast, . (1998), 'Predicting the potential distribution of plant species in an Alpine environment', *Journal of Vegetation Science* **9**, 65–74.

Guisan, A. & Theurillat, J. P. (2000), 'Equilibrium modelling of alpine plant distribution: how far can we go?', *Phytocoenologia* **30**, 353–384.

Guisan, A. & Thuiller, W. (2005), 'Predicting species distribution: offering more than simple habitat models', *Ecology Letters* **8**, 993–1009.

Guisan, A. & Zimmermann, N. E. (2000), 'Predictive habitat distribution models in ecology', *Experimental Methods* **135**, 147–186.

Guttman, N. B. & Quayle, R. G. (1990), 'A Review of Cooperative Temperature Data Validation', *Journal of Atmospheric and Oceanic Technology* **7**, 334–339.

Handcock, R. N. & Csillag, F. (2004), 'Spatio−temporal analysis using a multi−scale hierarchical ecoregionalization', *Photogrammetric Engineering & Remote Sensing* **70**(1), 1–10.

Hartkamp, A. D., Beurs, K. D., Stein, A. & White, J. W. (1999), *Interpolation Techniques for Climate Variables, NRG-GIS systems Series 99-01*, CIMMYT, Mexico.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2001), *The Elements of Statistical Learning : Data Mining, Inference and Prediction*, Springer-Verlag, New York.

Haxeltine, A. & Prentice, I. C. (1996), 'BIOME3: An equilibrium terrestrial biosphere model based on ecophysiological constraints, resource availability, and competition among plant functional types', *Global Biogeochemical Cycles* **10**, 693–709.

Hernandez, P. A., Graham, C. H., Master, L. L. & Albert, D. L. (2006), 'A comparison of the performance of species distribution models methods using a range of species' occurrences', *Ecography* **29**, 773–785.

Heuvelink, G. B. M. (1998), *Error Propogation in Environmental Modelling with GIS*, Taylor & Francis, London.

Heuvelink, G. B. M. & Burrough, P. A. (2002), 'Developments in statistical approaches to spatial uncertainty and its propagation', *International Journal of Geographical Information Science* **16**(2), 111–113.

Heuvelink, G. B. M., Burrough, P. & Stein, A. (1989), 'Propogation of errors in spatial modelling with GIS', *International Journal of Geographical Information Science* **3**, 302–322.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. (2005), 'Very High Resolution Interpolated Climate Surfaces for Global Land Areas', *International Journal of Climatology* **25**, 1965–1978.

Hijmans, R. J. & Graham, C. H. (2006), 'The ability of climate envelope models to predict the effect of climate change on species distributions', *Global Change Biology* **12**, 2272–2281.

Hijmans, R. J., Guarino, L., Jarvis, A., OBrien, R., Mathur, P., Bussink, C., Cruz, M., Barrantes, I. & Rojas, E. (2005), *DIVA-GIS Manual*, 5.2 edn. **URL:** *http://www.diva-gis.org/documentation*

Hijmans, R. J. & Spooner, D. M. (2001), 'Geographic Distribution of Wild Potato Species', *American Journal of Biology* **88**(11), 2101–2112.

Horne, B. V. (1983), 'Density as a Misleading Indicator of Habitat Quality', *Journal of Wildlife Management* **47**, 893–901.

Hulme, M., Mitchell, J., Ingram, W., Lowe, J., Johns, T., New, M. & Viner, D. (1999), 'Climate change scenarios for global impacts studies', *Global Environmental Change* **9**, S3–S19.

Huntley, B., Bartlein, P. J. & Prentice, I. C. (1989), 'Climatic control of the distribution and abundance of beech (Fagus l.) in Europe and North America', *Journal of Biogeography* **16**, 551–560.

Huntley, B., Berry, P. M., Cramer, W. & Mcdonald, A. P. (1995), 'Modelling present and potential future ranges of some european higher plants using climate response surfaces', *Journal of Biogeography* **22**, 967–1001.

Hutchinson, G. E. (1957), 'Concluding remarks', *Cold Spring Harbor Symposium on Quantitative Biology* **22**, 415–457.

Hutchinson, M. F. (1995), 'Interpolating mean rainfall using thin plate smoothing splines', *International Journal of Geographical Information Systems* **9**(4), 385–403.

Hutchinson, M. F. (2004), 'Anusplin Version 4.3. Centre for Research and Environmental Studies'. The Australian National University: Canberra, Australia.

Hutchinson, M. F. & Gessler, P. (1994), 'Splines - more than just a smooth interpolator', *Geoderma* **62**, 45–67.

Jarvisa, A., Lanec, A. & Hijmans, R. J. (2008), 'The effect of climate change on crop wild relatives', *Agricultural, Ecostems & Environment* **126**, 13–23.

Johns, T. C., Gregory, J. M., Ingram, W. J., Johnson, C. E., Jones, A., Lowe, J. A., Mitchell, J. F. B., Roberts, D. L., Sexton, D. M. H., Stevenson, D. S., Tett, S. F. B. & Woodage, M. J. (2003), 'Anthropogenic climate change for 1860 to 2100 simulated with the HadCM3 model under updated emissions scenarios', *Climate Dynamics* **20**(6), 583–612.

Jones, P. D. (1994), 'Hemispheric Surface Air Temperature Variations: A reanalysis and an update to 1993', *Journal of Climate* **7**, 1794–1802.

Jones, P. D. (1995), 'Land surface temperatures — is the network good enough?', *Climate Change* **31**, 545–558.

Kerry, R. & Oliver, M. A. (2007*a*), 'Determining the effect of asymmetric data on the variogram. I. Underlying asymmetry', *Computers & Geosciences* **33**, 1212–1232.

Kerry, R. & Oliver, M. A. (2007*b*), 'Determining the effect of asymmetric data on the variogram. II. Outliers', *Computers & Geosciences* **33**, 1233–1260.

Kershaw, A. P. (1997), 'A bioclimatic analysis of early to Middle Miocene brown coal floras, Latrobe Valley, south-eastern Australia.', *American Journal of Biology* **45**, 373–387.

Lanzante, J. R. (1996), 'Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applicaitions to historical radiosonde station data', *International Journal of Climatology* **16**, 1197–1226.

Laslett, G. M. (1994), 'Kriging and Slines: An Empirical Comparison of Their Predictive Performance in Some Applications', *Journal of the American Statistical Association* **89**(426), 391–400.

Leathwick, J. R. & Austin, M. P. (2001), 'Competitive interactions between tree species in New Zealand's old growth indigenous forests', *Ecology* **82**, 2560–2573.

Leathwick, J. R., Rowe, D., Richardson, J., Elith, J. & Hastie, T. (2005), 'Using multivariate adaptive regression splines to predict the distributions of New Zealands freshwater diadromous fish', *Freshwater Biology* **50**, 2034–2052.

Leff, B., Ramankutty, N. & Foley, J. A. (2004), 'Geographic distribution of major crops across the world', *Global Biogeochemical Cycles* **18**.

Legates, D. R. & Wilmott, C. J. (1990), 'Mean seasonal and spatial variability in global surface air temperature', *Theoretical and Applied Climatology* **41**, 11–21.

Leibold, M. A. (1995), 'The niche concept revisited: mechanistic models and community context', *Ecology* **76**, 1371–1382.

levin, S. A. (1992), 'The Problem of Pattern and Scale in Ecology', *Ecology* **73**, 1943–1967.

Loiselle, B. A., Howell, C. A., Graham, C. H., Goerck, J. M., Brooks, T., Smith, K. G. & Williams, P. H. (2003), 'Avoiding pitfalls of using species distribution models in conservation planning', *Conservation Biology* **17**, 1591–1600.

Lucas, C. (2010), 'On developing a historical fire weather data-set for Australia', *Australian Meteorological and Oceanographic Journal* **60**, 1–14.

Marinelli, M., Corner, R. & Wright, G. (2009), Error Propagation Analysis Techniques Applied to Precision Agriculture and Environmental Models, *in* A. Stein, W. Shi & W. Bijker, eds, 'Quality Aspects in Spatial Data Mining', CRC Press, Boca Raton, Florida, pp. 131–145.

McBratney, A. B., Whelan, B. M., Walvoort, D. J. J. & Minasny, B. (1999), A purposive sampling scheme for precision agriculture, *in* J. V. Stafford, ed., 'Precision Agriculture 99 (Part 1). Papers presented at the 2nd European Conference on Precision Agriculture, Odense Congress Centre, Denmark, 11-15 July 1999.', Sheffield Academic Press, Sheffield, U.K., pp. 101–110.

McPherson, J. M., Jetz, J. & Rogers, D. J. (2004), 'The effects of species'range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact?', *Journal of Applied Ecology* **41**, 811–823.

Metropolis, N. & Ulam, S. (1949), 'The monte carlo method', *Journal of the American Statistical Association* **44**(247), 335–.

Midgleya, G., Hannah, L., Millara, D., Thuiller, W. & Booth, A. (2003), 'Developing regional and species-level assessments of climate change impacts on biodiversity in the Cape Floristic Region', *Biological Conservation* **112**, 87–97.

Moore, G. (1998), A Handbook for Understanding and Managing Agricultural Soils, Technical report, Natural Resource Management Services Agriculture Western Australia.

New, M., Hulme, M. & Jones, P. (1999), 'Representing Twentieth-Century SpaceTime Climate Variability. Part i: Development of a 196190 mean monthly terrestrial climatology', *Journal of Climate* **12**, 829–856.

New, M., Lister, D., Hulme, M. & Makin, I. (2002), 'A high-resolution data set of surface climate over global land areas', *Climate Research* **21**, 1–25.

Nix, H. A. (1986), 'A biogeographic analysis of Australian Elapid snakes', *Atlas of Australian Elapid Snakes. Australian Flora and Fauna Series* **8**, 4–15.

NOAA (2010*a*), 'Ghcn-monthly version 2 - introduction'. `http://www.ncdc.noaa.gov/oa/climate/ghcn-monthly/index.php`.

NOAA (2010*b*), 'Ghcn-monthly version 2 - precipitation link'. `http://www.ncdc.noaa.gov/oa/climate/ghcn-monthly/index.php`.

NOAA (2010*c*), 'Ghcn-monthly version 2 - temperature link'. `http://www.ncdc.noaa.gov/oa/climate/ghcn-monthly/index.php`.

Olden, J. D. (2003), 'A species-specific approach to modeling biological communities and its potential for conservation', *Conservation Biology* **17**, 854–863.

Ott, R. L. & Longnecker, M. (2001), *An Introduction to Statistical Methods and Data Analysis*, Elsiever Publishing Company, Pacific Grove, California.

Parry, M. L., Rosenzweig, C., Iglesias, A., Livermore, M. & Fischer, G. (2004), 'Effects of climate change on global food production under SRES emissions and socio-economic scenarios', *Global Environmental Change* **14**, 53–67.

Parry, M., Rosenzweig, C., Iglesias, A., Fischer, G. & Livermore, M. (1999), 'Climate change and world food security: a new assessment', *Global Environmental Change* **9**, S51–S67.

Pearson, R. G., Dawson, T. E. & Lui, C. (2003), 'Modelling species distributions in britain: a hierarchical integration of climate and land-cover data', *Ecography* **27**, 285–298.

Pearson, R. G. & Dawson, T. P. (2003), 'Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful?', *Global Ecology and Biogeography* **12**, 361–371.

Pearson, R. G., Dawson, T. P. & P. M. Berry, P. A. H. (2002), 'SPECIES: A Spatial Evaluation of Climate Impact on the Envelope of Species', *Experimental Methods* **154**, 289–300.

Peterson, A. T. (2003), 'Predicting the geography of species' invasions via ecological niche modeling', *Quarterly Review of Biology* **78**, 419–433.

Peterson, A. T., Sanchez-Cordero, V., Sobero'n, J., Bartley, J., Buddemeier, R. W. & Navarro-Siguenza, A. G. (2001), 'Effects of global climate change on geographic distributions of Mexican Cracidae', *Experimental Methods* **144**, 21–30.

Peterson, A. T., Soberon, J. & Sanchez-Cordero, V. (1999), 'Conservatism of ecological niches in evolutionary time', *Science* **285**, 1265–1267.

Peterson, T. C. & Easterling, D. R. (1994), 'Creation of Homogeneous Composite Climatological Reference Series', *International Journal of Climatology* **14**, 671–679.

Peterson, T. C. & Vose, R. S. (1997), 'An overview of the Global Historical Climatology Network temperature data base', *Bulletin of the American Meteorogical Society* **78**, 2837–2849.

Peterson, T. C., Vose, R., Schmoyer, R. & Razuvaev, V. (1998), 'Global Historical Climatology Network GHCN Quality Control of Monthly Temperature Data', *International Journal of Climatology* **18**, 1169–1179.

Prentice, I. C., Cramer, W., Harrison, S. P., Leemans, R. & abd A. M. Solomon, R. A. M. (1992), 'A global biome model based on plant physiology and dominance, soil properties and climate', *Journal of Biogeography* **19**, 117–134.

Pu, Z., Kalnay, E., Derber, J. C. & Sela, J. G. (1997), 'Using forecast sensitivity patterns to improve future forecast skill', *Quarterly Journal of the Royal Meteorological Society* **123**, 1035–1053.

Pulliam, H. R. (2000), 'On the relationship between niche and distribution', *Ecology Letters* **3**, 349–361.

Purnomo, D., Corner, R. J. & Adams, M. L. (2003), Error Propagation in Agricultural Models, 4th Biennial European Conference on Precision Agriculture, Berlin, Germany.

Rayment, G. E. & Higginson, F. R. (1992), *Australian laboratory handbook of soil and water chemical methods*, Inkata Press, Melbourne.

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L. & Vanrolleghem, P. A. (2007), 'Uncertainty in the environmental modelling process - A framework and guidance', **22**, 1543–1556.

Rhoades, D. A. & Salinger, M. J. (1993), 'Adjustment of temperature and rainfall records for site changes', *International Journal of Climatology* **13**, 899–913.

Rissler, L. J., Hijmans, R. J., Graham, C. H., Moritz, C. & Wake, D. B. (2006), 'Phylogeographic lineages and species comparisons in conservation analyses: A case study of California herpetofauna', *The American Naturalist* **167**, 655–666.

Ritchie, J. T. (1972), 'Model for Predicting Evaporation from a Row Crop with Incomplete Cover', *Water Resources Research* **8**(5), 1204–1213.

Rodder, D. (2009), How to predict the future? On niches and potential distributions of amphibians and reptiles in a changing climate, PhD thesis, School of Mathematics and Science, Univerity of Bonn.
URL: *hss.ulb.uni-bonn.de:90/2009/1948/1948.pdf*

Rosenzweig, C. & Parry, M. L. (1994), 'Potential inpact of climate change on world food supply', *Nature* **367**, 133–138.

Rowe, R. J. (2005), 'Elevation gradient analyses and teh use of historical museum specimens: a cautionary tale', *Journal of Biogeography* **32**, 1883–1897.

Rowe, W. D. (1994), 'Understanding Uncertainty', *Risk Analysis* **14**(7), 743–750.

Segers, H., Branquart, E., Caudron, A. & Tack, J. (2001), Scientific Tools for Biodiversity Conservation: Monitoring, Modelling and Experiments. Web Version Part 3, Technical report, European Platform for Biodiversity Research Strategy.

Shao, G. F. & Haplin, P. N. (1995), 'Climatic controls of eastern north american coastal tree and shrub distributions.', *Journal of Biogeography* **22**, 1083–1089.

Skidmore, A. K., Gauld, A. & Walker, P. (1996), 'Classification of kangaroo habitat distribution using three GIS models.', *International Journal of Geographical Information Systems* **10**, 441–454.

Smith, R. C., Wallace, J. F., Hick, P. T., Gilmore, R. F., Belford, R. K., Portmann, P. A., Regan, K. L. & Turner, N. C. (1994), 'Potential of using Field Spectoscopy During Early Growth for Ranking Biomass in Cereal Breeding Trials', *Australian Journal of Agricultural Research* **44**, 1713–1730.

Stein, A., Staritsky, I., Bouma, I., van Eijnsbergen, A. & Bregt, A. (1991), 'Simulation of moisture deficits and areal interpolation by universal cokriging', *Water Resources Research* **27**, 1963–1973.

Stephens, D. J. & Lyons, T. J. (1998), 'Rainfall−yeild relationships across the Australian wheatbelt', *Australian Journal of Agricultural Research* **49**, 211–223.

Stewart, J. (2003), *Calculus*, Thomson and Brooks/Cole, Belmont, California. USA.

Sykes, M. T., Prentice, I. C. & Cramer, W. (1996), 'A bioclimatic model for the potential distributions of north European tree species under present and future climates', *Journal of Biogeography* **23**, 203–233.

Thomas, C. D., Bodsworth, E. J., Wilson, R. J., Simmons, A. D., Davies, Z. G., Musche, M. & Conradt, L. (2001), 'Ecological and evolutionary processes at expanding range margins', *Nature* **411**, 577–581.

Thuiller, W. (2003), 'BIOMOD optimizing predictions of species distributions and projecting potential future shifts under global change', *Global Change Biology* **9**, 1353–1362.

Tsutsui, A. V. S. N. D. (2004), 'The Value of Museum Collections for Research and Society', *Bioscience* **54**, 66–74.

Tyre, A. J., Possingham, H. P. & Lindenmayer, D. B. (2001), 'Inferring process from pattern: can territory occupancy provide information about life history parameters?', *Ecology Applications* **11**(6), 1722–1737.

Vitousek, P. M., Naylor, R., Crews, T., David, M. B., Drinkwater, L. E., Holland, E., Johnes, P. J., Katzenberger, J., Martinelli, L. A., Matson, P. A., Nziguheba, G., Ojima, D., Palm, C. A., Robertson, G. P., Sanchez, P. A., Townsend, A. R. & Zhang, F. S. (2009), 'Nutrient imbalances in agricultural development', *Science* **324**, 1519–1520.

Vose, R., Schmoyer, R. L., Steurer, P. M., Peterson, T. C., Heim, R., Karl, T. R. & Eischeid, J. K. (1992), 'Global Historical Climatology Network: long-term monthly temperature, precipitation, sea level pressure, and station pressure data'.
URL: *http://cdiac.ornl.gov/epubs/ndp/ndp041/ndp041.html*

Walker, P. A. & Cocks, K. D. (1991), 'HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species', *Global Ecology and Biogeography Letters* **1**, 108–118.

Wang, M., Chen, H., Wu, Y., Feng, Y. & Pu, Q. (2010), 'New Techniques for the Detection and Adjustment of Shifts in Daily Daily Precipitation Data Series', *Journal of Applied Meteorology and Climatology* **49**(12), 2416–2437.

Welsh, A. H. (1996), *Aspects of Statistical Interference*, John Wiley & Sons, New York, USA.

White, A., Cannell, M. G. R. & Friend, A. D. (1999), 'Climate change impacts on ecosystems and the terrestrial carbon sink: a new assessment', *Global Environmental Change* **9**, S21–S30.

Wieczorek, J., Guo, Q. & Hijmans, R. (2004), 'The point-radius method for georeferencing locality descriptions and calculating associated uncertainty', *International Journal of Geographical Information Science* **8**(18), 745–767.

Willis, K. J. & Whittaker, R. J. (2002), 'Species diversity − scale matters', *Science* **295**, 1245–1248.

Wong, M. T. F., Corner, R. J. & Cook, S. E. (2001), 'A decision support system for mapping the site−specific potassium requirement of wheat in the field', *Australian Journal of Experimental Agricultural* **41**, 655–661.

Woodward, F. I. (1990), 'The impact of low temperatures in controlling the geographical distribution of plants', *Philosophical Transactions of the Royal Society of London. Series B* **326**, 585–593.

Worldclim - Global Climate Data (2009), `http://www.worldclim.org/`.

Wratt, D. S., Tait, A., Griffiths, G., Epsie, P., Jessen, M., keys, J., ladd, M., Lew, D., Lowther, W., Mitchell, N., Morton, J., reid, J., Reid, S., Richardson, A., Sansom, J. & Shankar, U. (2006), 'Climate for crops: integrating climate data with information about soils and crop requirements to reduce risks in agricultural decision-making', *Meteorological Applications* **13**, 305–315.

Yee, T. W. & Mitchell, N. D. (1991), 'Generalized additive models in plant ecology', *Journal of Vegetation Science* **2**(5), 587–602.