WORKING DOCUMENT

THIN PLATE SMOOTHING SPLINE INTERPOLATION OF PARAMETERS OF THE AR(1) ANNUAL RAINFALL MODEL ACROSS THE AUSTRALIAN CONTINENT

P. A. Hancock
M. F. Hutchinson

Working Document 02/7
December 2002

P. A. Hancock
M. F. Hutchinson

Thin Plate Smoothing Spline Interpolation of Parameters of the AR(1) Annual Rainfall Model Across the Australian Continent

Working Document 02/7

TABLE OF CONTENTS

Thin plate smoothing spline interpolation of parameters of the AR(1) annual rainfall model across the Australian continent

P.A. Hancock[1] and M.F. Hutchinson[2]

1. Centre for Resource and Environmental Studies, The Australian National University

Email: penlepen@cres.anu.edu.au

2. Centre for Resource and Environmental Studies, The Australian National University

Email:hutch@cres.anu.edu.au

1 Introduction

A first order autoregressive, or AR(1), model was found by Srikanthan and McMahon [20] to be appropriate for simulating annual rainfall amounts at a given location. The objective of this project was to develop thin plate smoothing surfaces that spatially interpolate the parameters of the AR(1) model across the Australian continent. The AR(1) parameters with which this study is concerned include the mean, the standard deviation, the skewness coefficient and the lag one autocorrelation coefficient.

Initially, surfaces were fitted to the BMRC high quality data set, consisting of 363 stations known to have high quality rainfall record. This preliminary analysis gave insights into the performance of thin plate smoothing spline models for sparse, noisy data. Statistics such as the generalised cross validation (GCV), the signal, and standard error estimates, along with testing using withheld data, were used to assess the accuracy of the spline models, and the ability of the spline surfaces to represent the processes measured by the data. It was concluded from this preliminary analysis that more data

were required to generate appropriate surfaces for the AR(1) model. Further analysis was therefore performed on a larger data set, consisting of 6334 stations. The thin plate smoothing spline techniques involved in this study, and the results of their application to the AR(1) parameters, are discussed in the following sections.

2 Background

2.1 Thin plate smoothing spline interpolation of climate

The thin plate smoothing spline method of spatial interpolation used in this study can be motivated by the following data model. Consider data observations $(z_i, x_{1i}, x_{2i}, \ldots, x_{di})$ measuring a dependent variable z and a set of d predictor variables $x_1, \ldots, x_d$. For example, climate is often well predicted using latitude, longtitude and elevation as independent variables. Assume that z has short range variation that is discontinuous and random, as well as continuous long range variation. We can then propose the following model

$$z_i = g(x_{1i}, \ldots, x_{di}) + \epsilon_i \tag{1}$$

where g is a slowly varying continuous function and ϵ is a discontinuous random variable, assumed to be independent with a mean of zero and a variance of σ^2. The errors ϵ_i are assumed to be due to measurement error, and short range microscale variation that occurs over a range smaller than the resolution of the data set. The microscale variation may be continuous, but the data is not spatially dense enough to identify it, so it is usually assumed to be discontinuous noise.

We aim to estimate the process g by choosing a function f to fit the data z_i. The function f must be able to separate the continuous signal from the noise ϵ_i. We can manufacture such a function by solving the following minimisation

problem

$$\text{Minimise } \frac{1}{n}\sum_{i=1}^{n}(z_i - f_i)^2 + \lambda J_m^d(f) \tag{2}$$

over suitably smooth functions f, where f_i is the value of the fitted function at i^{th} data point, λ is a fixed smoothing parameter, and $J_m^d(f)$ is a measure of the roughness of the function f in terms of partial derivatives. Calculation of J_m^d depends on m, the order of the partial derivatives, and the number of independent variables d. For example, if $m = 2$, which is a typical value, and $d = 2$, then

$$J_2(f) = \int_{-\infty}^{\infty} f_{x_1x_1}^2 + 2f_{x_1x_2}^2 + f_{x_2x_2}^2 dx_1 dx_2 \tag{3}$$

[22]. Expression (2) represents a trade off between fitting the data as closely as possible whilst maintaining smoothness. For sufficiently large λ this avoids highly oscillatory functions that are heavily reliant on individual data points. The effect is to 'smooth' the data, with the aim of removing the noise and representing the slowly varying, spatially continuous process.

The solution to this minimisation problem is well known to be a thin plate smoothing spline [19], [3], [15], [22]. Thin plate smoothing splines are slowly varying continuous functions with continuity of the first $2m - d - 1$ derivatives. Multivariate thin plate splines are not piecewise polynomial functions like the traditional univariate splines. They are termed 'splines' because the solution to (2) for the univariate case, with $m = 2$, is a natural cubic spline.

Consider the cases of temperature and precipitation. Thin plate smoothing spline functions of latitude, longtitude and elevation have been shown to accurately spatially predict long term annual and monthly mean temperature and precipitation [12], [5], [24], [7]. Temperature is the simpler of the two spatial processes, and has a roughly linear dependence on elevation that is independent of location [5]. Hutchinson [5] demonstrates that the following partial spline model is a sensible model for temperature

$$z_i = g(x_i, y_i) + \beta h_i + \epsilon_i \tag{4}$$

where x_i is the longtitude, y_i is the latitude and h_i is the elevation at data measurement location i. The resulting solution incorporates a bivariate thin plate smoothing spline $f(x, y)$ and a spatially constant linear trend on elevation with slope, or 'lapse rate' β.

Rainfall is a more complex and localised process than temperature, and it is not reasonable to assume a constant dependence on elevation throughout the study area. Rainfall can be accurately modelled by a thin plate spline with three independent variables, according to the following model

$$z_i = g(x_i, y_i, h_i) + \epsilon_i \tag{5}$$

[12, 5]. This allows for a spatially varying dependence on elevation. Other independent variables, such as slope and aspect, have been found to provide minimal additional explanatory power [7].

2.2 Optimisation using generalised cross validation

It can be seen from equation (2) that the thin plate spline solution to the minimisation problem will depend on the smoothing parameter λ. The next step is therefore to determine the value of λ that produces the best approximation by the thin plate spline to the actual continuous surface g that the spline is attempting to represent [22]. Craven and Wahba [1], in their analysis of the univariate case, argue that the ideal solution would be to minimise the true error, given in Craven and Wahba [1] as

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - g(x_i))^2 \tag{6}$$

It is shown in Craven and Wahba [1] that, using the principles of mathematical expectation and assuming that the errors are independent, an unbiased estimate of $R(\lambda)$ is given by

$$\hat{R}(\lambda) = \frac{1}{n}\|I - A(\lambda)\mathbf{z}\| - \frac{2\sigma^2}{n}trace(I - A(\lambda)) + \frac{\sigma^2}{n}traceA^2(\lambda) \tag{7}$$

where

$$E(\epsilon_i \epsilon_j) = \sigma^2 \delta_{ij}, \text{ and } E(\epsilon_i) = 0 \tag{8}$$

where σ^2 is the variance of the noise, δ_{ij} is the Kronecker delta and $A(\lambda)$ is an $n \times n$ matrix, known as the 'influence' matrix. The influence matrix takes the vector of data values to the vector of fitted values. It is thus defined by

$$\mathbf{f} = A(\lambda)\mathbf{z} \tag{9}$$

where $\mathbf{s}$ is a vector containing the values of the fitted spline at the data point locations. In order to minimise the expression in equation 7 one has to know σ^2. This value is difficult to ascertain, because it incorporates an interpolation error component that is unknown [1]. The optimal value of λ is therefore defined as the value of λ that minimises a function known as the generalised cross validation (GCV) [1], [22], [4].
The GCV is a measure of the predictive error of the fitted surface and is effectively calculated by removing each data point in turn and summing, with appropriate weighting, the square of the discrepancy of each omitted data point from a surface fitted to all other data points [7]. This is a relatively common concept in statistical analysis [14]. It is shown in [1] that, using the 'leaving out one' lemma [22] the generalised cross validation for the multidimensional case can be calculated implicitly and hence efficiently by

$$GCV(\lambda) = \frac{(\mathbf{z} - A(\lambda)\mathbf{z})^T(\mathbf{z} - A(\lambda)\mathbf{z})/n}{[trace(I - A(\lambda))/n]^2} \tag{10}$$

A theoretical justification for using the GCV to determine the optimum thin plate spline function is given in [1], where it is shown that, if $\hat{\lambda}$ is the minimiser of the true error and λ^* is the minimiser of the GCV, then

$$\lim_{n\to\infty} \frac{R(\hat{\lambda})}{R(\lambda^*)} = 1 \tag{11}$$

Thus, in theory, as the number of data points increases, the minimiser of the GCV approaches the minimiser of the true error.

2.3 Estimating the variance of the noise

According to Wahba [22], determining the optimum surface by minimising the GCV also yields an estimate of σ^2, the variance of the noise. The estimate is given by

$$\hat{\sigma}^2 = \frac{(\mathbf{z} - A(\lambda)\mathbf{z})^T(\mathbf{z} - A(\lambda)\mathbf{z})}{trace(I - A(\lambda))} \tag{12}$$

Hutchinson [6] explains that $trace(I - A)$ may be interpreted as the degrees of freedom of the residual sum of squares and thus equation 12 is analogous to the estimate of σ^2 obtained in linear regression [22], [18]. It follows that the effective number of parameters of the fitted model, known as the *signal*, is given by $trace(A)$ [13].

2.4 Interpretation of the signal

Hutchinson and Gessler [13] present evidence to show that the value of the signal is a useful diagnostic in its own right. They state that, in most applications, if the signal exceeds $n/2$, it is likely that the data are too sparse to adequately support spline interpolation. Exact interpolation corresponds to a signal equal to the number of data points. This implies that there is no measurement error and no microscale variation [13], which is generally an unrealistic assumption. It may indicate that the optimisation prcedures have failed due to insufficient data [13], short range correlation in the data values [11], or autocorrelation in the error structure that has been unaccounted for by the model [2]. On the other hand, when the signal reaches its minimum value, a number which depends on the number of independent variables and the order of the derivative [9], the fitted spline is equivalent to a least squares regression of the data [13]. This results in complete global smoothing of the data. Therefore, extreme signal values that are either at the top or at the bottom of the range of possible values can indicate a lack of spatial structure in the data.

2.5 Standard error estimates

According to Wahba [21], [6], it can be shown using the multivariate prior distribution which gives rise to splines that the posterior covariance of the vector of the fitted values is given by the symmetric matrix $A(\lambda)\sigma^2$. This result allows the estimation of the pointwise standard errors of the fitted spline estimate of g [6].

3 Application to the BMRC high quality data set

3.1 Data

Initially, surfaces were fitted to 363 stations known to have high accuracy of rainfall record. Annual rainfall was recorded at different time periods for each station, from as early as 1863 to as recent as 1998. The total time period considered was 135 years. There are two main sources of uncertainty in spatially interpolating summary statistics for these annual rainfall records. Firstly, given that the record for many stations is far from complete over this time period, there will be error in estimating the 135 year standard deviation, skewness coefficient and lag one autocorrelation for many of the stations. The number of years of record for the stations in the data network are shown in Figure 1. The bulk of the stations have relatively long records, of at least 80 years, which should minimise the effect of missing years on this analysis.
Secondly, it may not be valid to assume that rainfall patterns over the last 135 years have been stationary. Warner [23] present evidence of certain 'breaks' in rainfall regimes in New South Wales over the last 150 years, corresponding to times when the average rainfall amount and intensity changes significantly. Pittock [16], [17] show a particularly clear difference in rainfall trends between the periods 1881 - 1940 and 1940 - 1974.
Plots of the values of the AR(1) parameters are shown in Figures 2-5, along with spatial mappings in Figures 6-9 (the labeling of station 031043 will be explained in the following sections). The graphs for mean and standard devi-

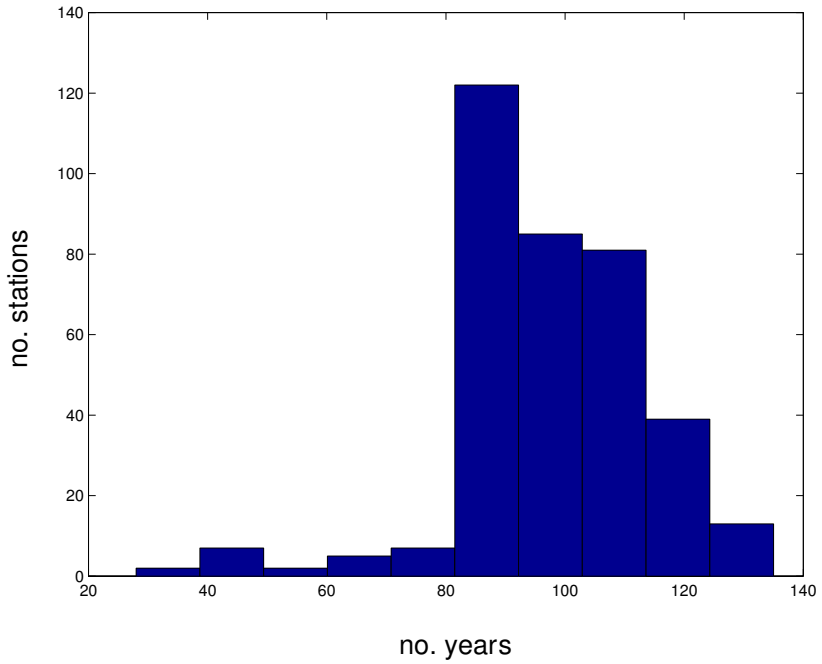


Figure 1: Histogram showing the number of years of record for the 363 stations

ation indicate greater trend and coherence than the graphs for the skewness and lag 1 correlation coefficients. This coherence is likely to be spatially driven given that the stations are ordered such that stations with similar location are close together on the horizontal axis. Trends are more difficult to see on the maps due to the limitations of using a linear colour scheme. A small number of high values on the east coast obscure the inland trends for mean and standard deviation. More variation is evident for the skewness and lag 1 correlation coefficients.

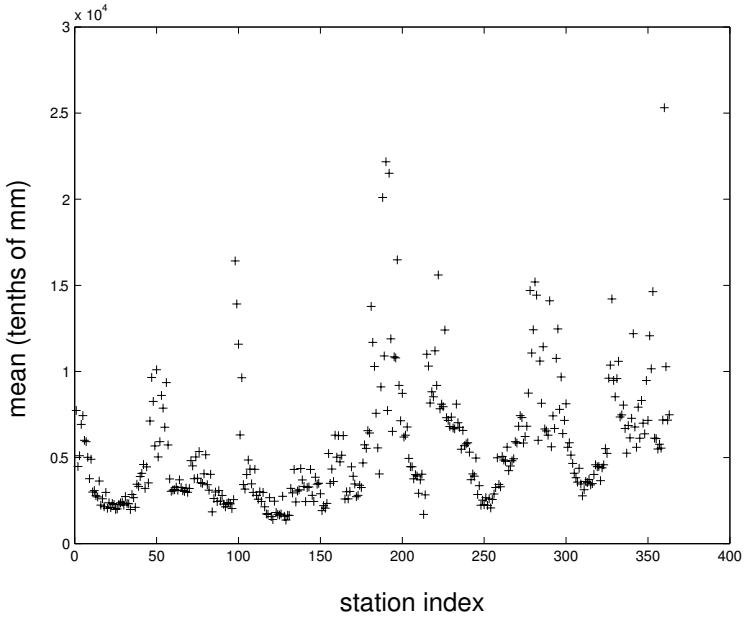


Figure 2: Station means

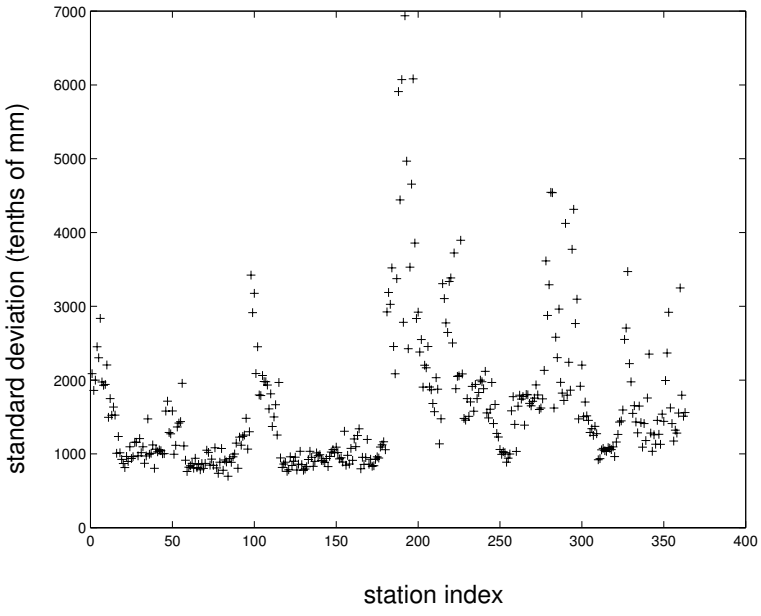


Figure 3: Station standard deviations

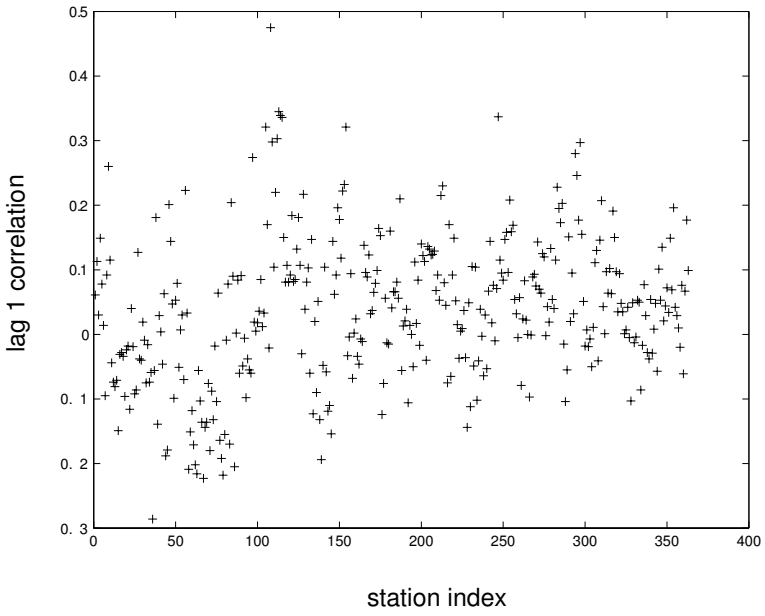


Figure 4: Station lag 1 correlations

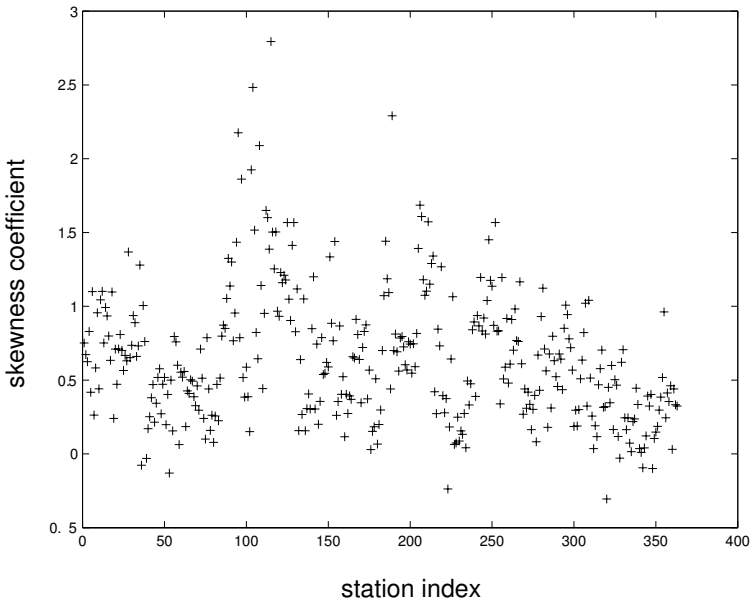


Figure 5: Station skewness coefficients

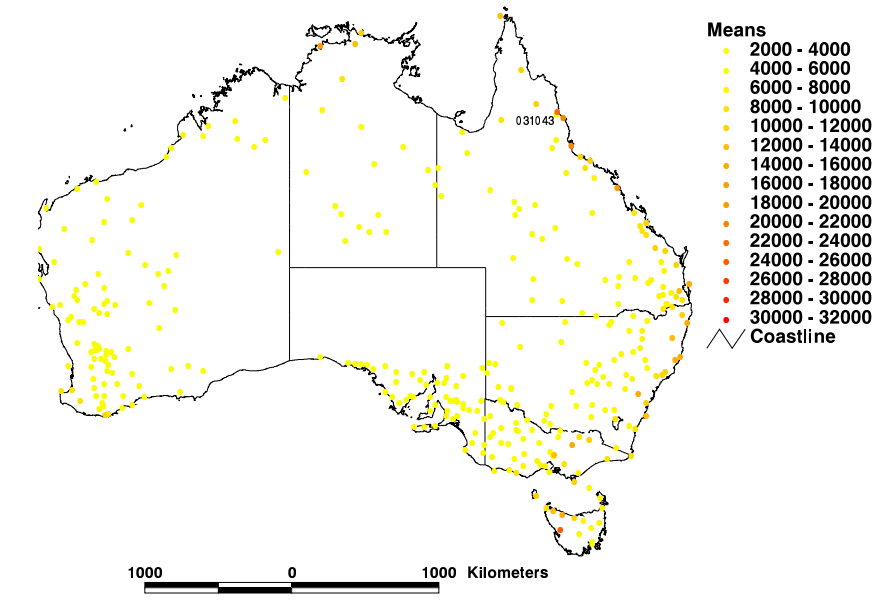


Figure 6: Station means

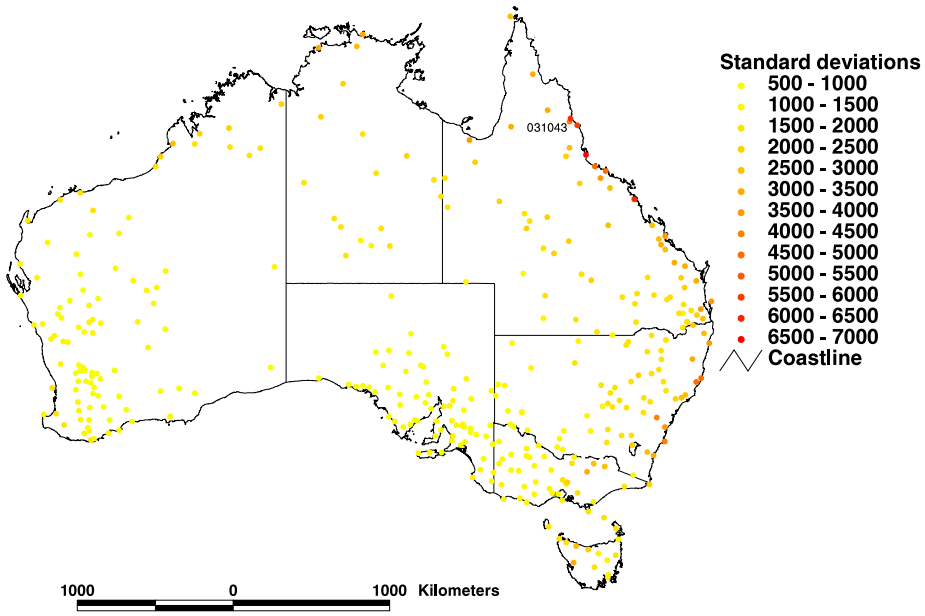


Figure 7: Station standard deviations

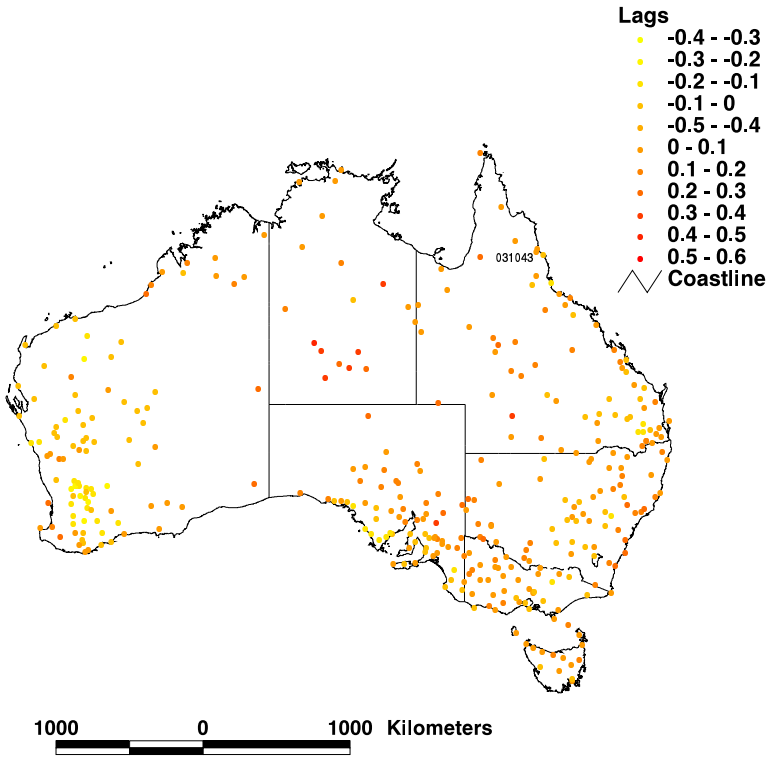


Figure 8: Station lag 1 correlations

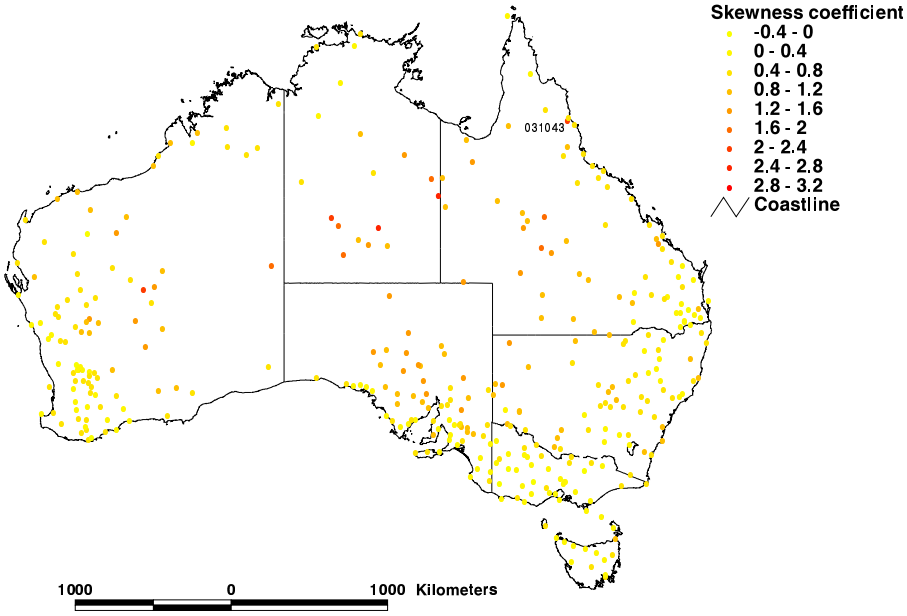


Figure 9: Station skewness coefficients

As a preliminary comment, less spatial coherence would be expected for the higher order moments, particularly skewness, as it is poorly defined without a large data set and therefore has a high noise component. There is no obvious reason why the skewness and the lag 1 correlations should be strongly linked to topography, though they may display broadscale spatial trends.

3.2 Thin plate smoothing spline interpolation

Three spline models were tested for spatially interpolating each of the AR(1) parameters, using the ANUSPLIN program SPLINA. These included trivariate spline models of latitude, longtitude and elevation, partial spline models of latitude and longtitude with a constant linear dependence on elevation, and bivariate spline models of latitude and longtitude. The trivariate spline model allows for a spatially varying dependence on location and elevation, though it requires more data to support construction of these more complex trends and may be unstable for highly variable, sparse data. The partial spline and bivariate spline models are more robust, although they have lower potential for accurate prediction. The relative merits of the different spline models were assessed by looking at the GCV, signal, and residuals of data from the fitted surface combined with the Bayesian standard error estimates as well as the ability to predict withheld data.

The square root transformation suggested by Hutchinson [9] was also tried. This transformation is designed to reduce positive skew in measured values, as can arise when fitting data that are naturally non-negative or positive. The square root transformation procedure is part of the ANUSPLIN package, which delivers transformed and untransformed output predictions and corresponding statistics. Hutchinson [10] has found that applying the square root transformation to daily rainfall data, before fitting a thin plate smoothing spline, could reduce interpolation error by about 10 percent.

The results for each spline model for each statistic are shown in Tables 1-4. According to the GCV values, the optimal spline models are a partial spline model with a square root transformation for annual mean rainfall and its standard deviation, and trivariate spline models for the skewness

spline model	signal	RMS residual of data values from fitted values (mm)	square root of GCV (mm)
bivariate	192.2	66.7	142
bivariate, square root transformation	191.8	46.7	99.2
partial	193.1	66.1	141
partial, square root transformation	195.4	45.1	97.8
trivariate	363	0.268×10^{-6}	103
trivariate, square root transformation	363	0.194×10^{-6}	74.5

Table 1: SPLINA results for annual mean rainfall

coefficient and the lag 1 correlation coefficient. Note that the signal values for the trivariate spline model for the annual mean and its standard deviation indicate that the data were exactly interpolated. This is a clear indication that the spline models are unstable, which implies that more data is required to explain the fine scale variability. However, the trivariate model did have the lowest GCV by a considerable amount, which is an indication that, with more data, this model could best represent the spatial trends in the rainfall process.

It is also interesting that the square root transormation resulted in significantly improved models for the mean and standard deviation, producing substantial reductions in the GCV. This affirms the suitability of the square root transformation for rainfall data. Clearly it could not be used for the skewness or lag 1 correlation coefficients as these statistics can take on negative values. There is, however, no reason why it should be suitable for these statistics.

Using the minimum GCV spline models, an error analysis was conducted to determine the error in the predictions obtained from these spline surfaces. Firstly, the residuals of each data point from its corresponding fitted estimate were plotted graphically, in Figures 10-13, and spatially, in Figures 14-19. Note that the relative residual is represented for the mean and standard deviation, and the absolute residual for the skewness and lag 1 correlation

spline model	signal	RMS residual of data values from fitted values (mm)	square root of GCV (mm)
bivariate	244.8	10.1	30.9
bivariate, square root transformation	230.1	8.89	24.3
partial	252.0	9.37	30.6
partial, square root transformation	244.6	7.75	23.8
trivariate	363	0.630×10^{-7}	24.1
trivariate, square root transformation	363	0.506×10^{-7}	19.4

Table 2: SPLINA results for the standard deviation of annual mean rainfall

spline model	signal	RMS residual of data values from fitted values	square root of GCV
bivariate	81.1	0.235	0.303
partial	82.8	0.234	0.303
trivariate	136.9	0.186	0.299

Table 3: SPLINA results for the skewness coefficient of annual mean rainfall

coefficients. There were too many small numbers to calculate relative residuals for these last two statistics. There is one obvious residual for the mean rainfall surface, as shown in Figure 10. This point has a prediction error of around 45%. This station was identified to be number 031043, and is marked on all maps. It is not surprising that the surface has trouble fitting trends in this area, as there is a cluster of three closely spaced points with very different rainfall values. Furthermore, looking at the elevations for these stations, station 031043 has an elevation of 389m, where as the other two stations are close to sea level. However, it has the lowest rainfall of the three

spline model	signal	RMS residual of data values from fitted values	square root of GCV
bivariate	159.2	0.0426	0.759×10^{-1}
partial	161.3	0.0422	0.760×10^{-1}
trivariate	264.3	0.0201	0.739×10^{-1}

Table 4: SPLINA results for the lag 1 correlation coefficient of annual mean rainfall

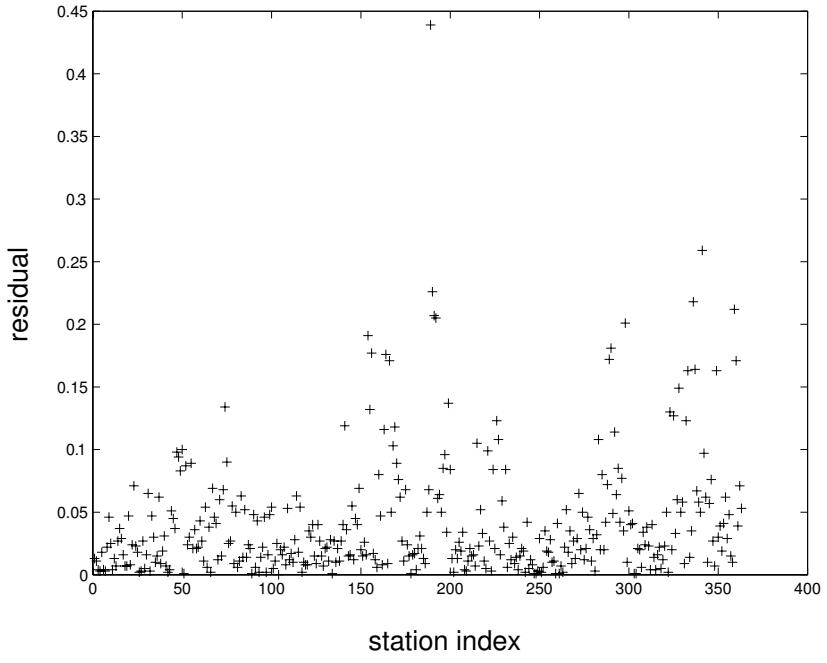


Figure 10: Relative residuals of data values from the fitted surface for the means

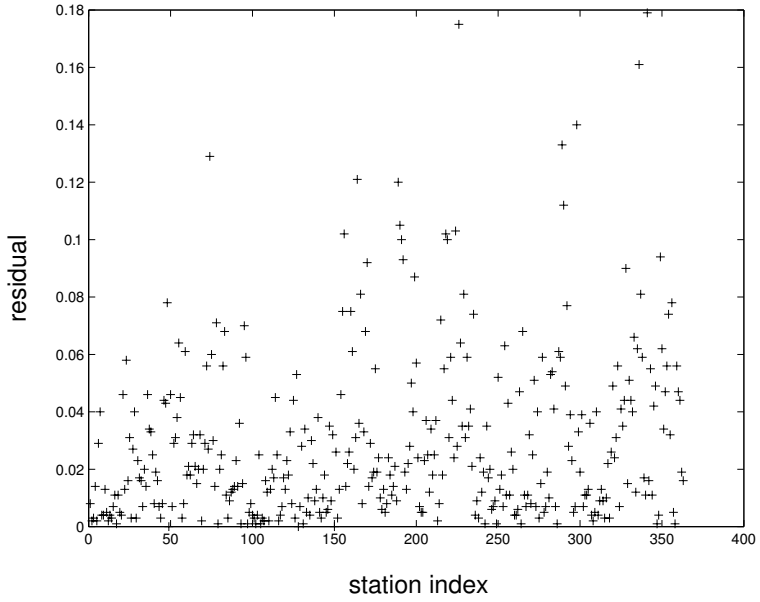


Figure 11: Relative residuals of data values from the fitted surface for the standard deviations

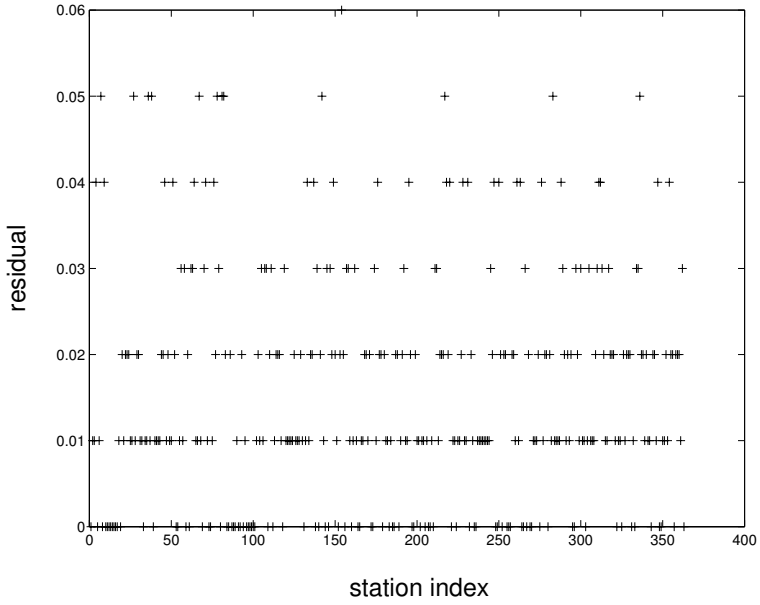


Figure 12: Residuals of data values from the fitted surface for the lag 1 correlations

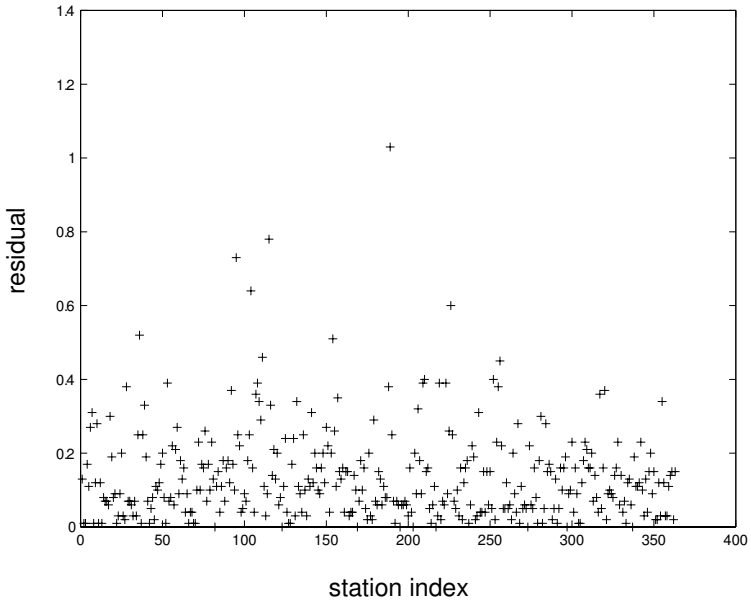


Figure 13: Residuals of data values from the fitted surface for the skewness coefficients

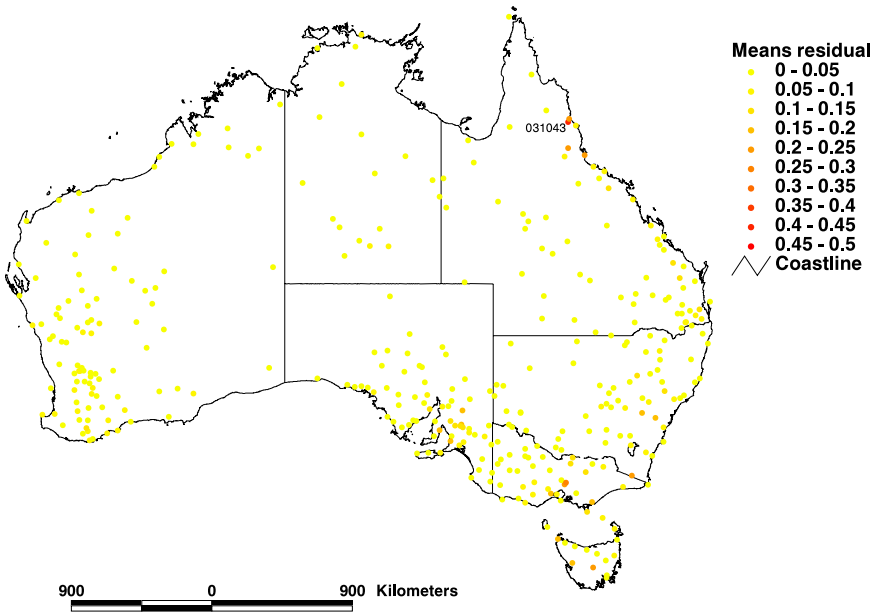


Figure 14: Relative residuals of data values from the fitted surface for the means

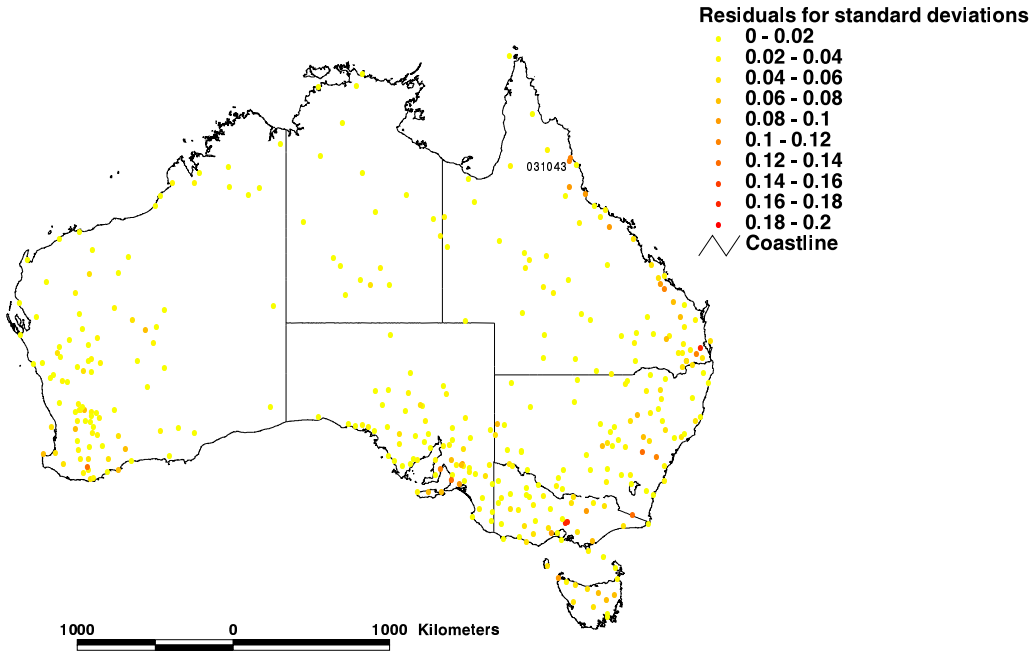


Figure 15: Relative residuals of data values from the fitted surface for the standard deviations

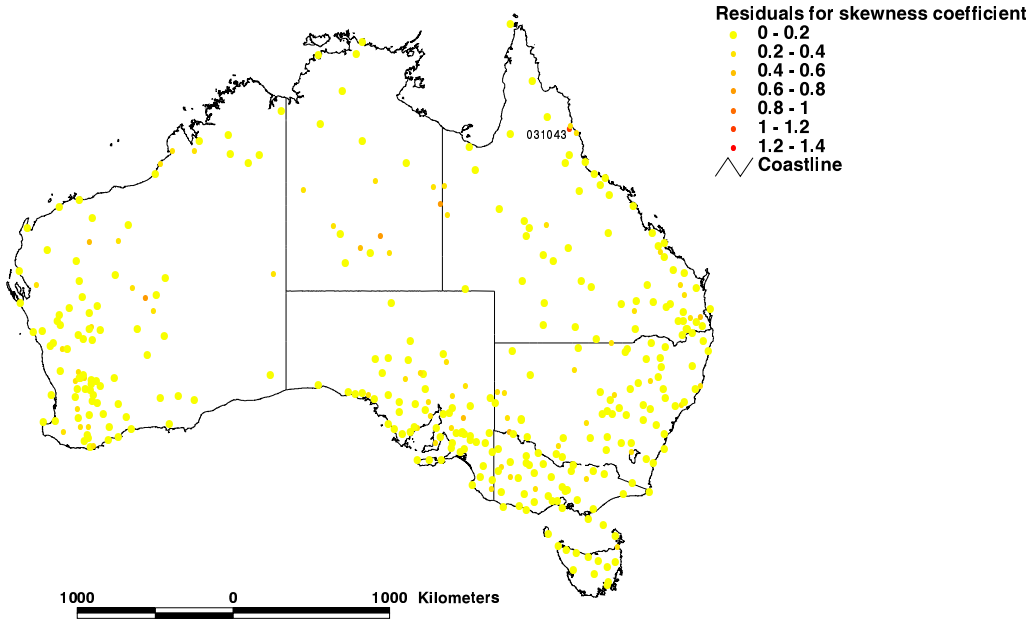


Figure 16: Residuals of data values from the fitted surface for the skewness coefficients

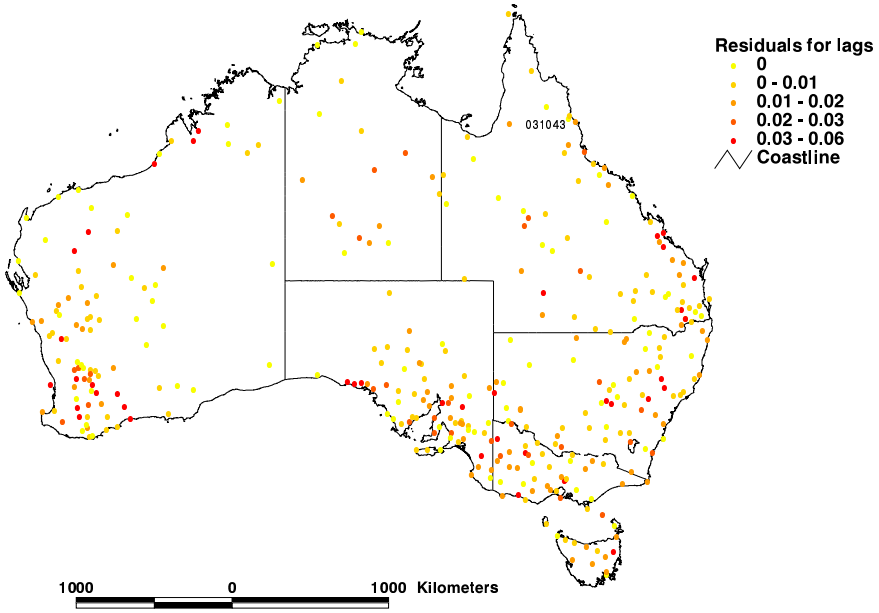


Figure 17: Residuals of data values from the fitted surface for the lag 1 correlations

stations. This local reverse elevation trend is certainly beyond the capacity of a partial spline to represent, and may indicate data error given that rainfall would be expected to increase with increasing elevation. The residual maps for standard deviation and skewness also show high error in this region. Similar analysis could be carried out for the other outstanding residuals shown on these maps.

Secondly, a spatially representative sample of 20 data points was withheld and the surfaces were refitted to see how well the withheld values were predicted. The sample was selected using the ANUSPLIN program SELNOT, which selects points by successively rejecting one point from the closest remaining pair of points in the independent spline variable space, until the specified number of points remain [9]. Summary results obtained after withholding the points are shown in Table 5. The RMS residual and the maximum of all the deviations of the withheld data values from the corresponding surface estimates were calculated by the SPLINA program. The maximum relative deviation of the witheld data was also calculated.

parameter	RMS residual	maximum deviation of withheld data values from fitted values	maximum relative deviation
mean	166	447	0.42
standard deviation	50.7	141	0.69
skewness coefficient	0.442	1.48	1.3
lag 1 correlation coefficient	0.08	0.158	6.0

Table 5: Validation results for withheld data

Table 5 gives an idea of the magnitude of error that can be expected in the predictions obtained for these surfaces. The relative error is considerably higher for the surfaces for the skewness and lag 1 correlation coefficients. This is due to the small values of these statistics for many stations. The incorporation of more data will reveal more about the ability of the spline model to spatially predict these statistics. Comparing the RMS residual of the withheld data with the minimum GCV values in Tables 1-4, the GCV values are smaller for all the AR(1) parameters, which indicates that the spline model has underestimated the amount of noise in the data. The incorporation of more data is likely to improve the reliability of these statistics. The model for lag 1 correlations may be an exception, as the difference between the GCV and the RMS residual of withheld data is quite small. This indicates that the spatial variability in this statistic was almost fully captured by the spline model.

Thirdly, surfaces, and standard error surfaces, were created for each of the statistics using the optimal spline models. The standard error surfaces represent the pointwise Bayesian standard errors in the fitted surface, as discussed in section 1.5. These surfaces are shown in Figures 18-28. The surface for the mean shows little variation in inland regions, and a sharp increase at the coast, particularly the east coast. More accurate fine scale trends are evident in the mean annual rainfall surface available at `http://cres.anu.edu.au/`

`outputs/software.html`, which used 12000 data points. The relative standard errors in the mean surface show a clear trend to higher standard errors in data sparse regions, particularly inland areas. The relative standard error values agree in magnitude with the relative errors of the withheld data in Table 5, which indicates that they are reliable.

For the standard deviation the patterns are similar, though the gradients are not as sharp. The relative standard errors in the standard deviation are also clearly higher where the data density is sparse. A comparison with the relative errors in the withheld data indicates that the standard error surface for the standard deviations may be underestimating the predictive error. Note, however, that comparison is being made with the highest prediction error for all of the withheld data, which is a conservative test.

For the skewness and lag 1 correlations, the standard error surfaces indicate that the Bayesian estimation of pointwise standard errors has failed for the trivariate analysis. As an example, the standard error surface for the lag 1 correlations is given in Figure 24. Although the range of standard errors is small, so the surface is practically flat, the errors peak at the data point locations, which is clearly not sensible. Looking again at the results in Tables 3 and 4, it appears that the minimum GCV model may not be be the best choice after all. For both skewness and lag 1 correlation the difference between the GCV values for all three spline models is negligable, where as the trivariate model requires a higher signal, especially for the lag 1 correlations. Given that the more complex model does not significantly improve the predictive capacity, it was decided to change to the simplest, bivariate spline model. This model is likely to be more robust, given the high noise level and the small data set.

Surfaces and standard error surfaces for the skewness coefficient and the lag 1 correlations, constructed using the bivariate spline model, are shown in Figures 25-28. The trends in the surfaces for skewness and lag 1 correlation are broadscale, with a clear inland gradient. This is particularly marked in the case of the skewness coefficient. The standard error surfaces are now sensible, with high standard errors in data sparse locations. Standard errors are notably high on the north-west Australian coast. However, according to

the results for the withheld data, these standard error surfaces also underestimate the predictive error.

3.3 Conclusions

The spline models developed for this data set give a preliminary indication of the broadscale spatial patterns of the AR(1) parameters, and their variability. Analysis of predictive capacity and reliability was performed using GCV statistics, withheld data, residual plots and standard error estimates. This determined that a partial spline model was optimal for interpolation of the mean and standard deviation, and a bivariate spline model was optimal for the interpolation of the skewness and lag 1 correlation coefficients. A square root transformation of the data values was found to significantly improve the predictive capacity of the fitted models for the mean and the standard deviation. The standard error surfaces corresponding to the spline surfaces indicate that predictive error is generally highest in inland areas where data is sparse.

There are, however, certain factors that imply that there is insufficient data to represent the spatial trends in the parameters of the AR(1) model, for this data set . Firstly, the RMS residual of the withheld data is consistently higher than the GCV values obtained from the spline models and the standard error estimates obtained from the standard error surfaces. There was, however, considerably greater agreement for the lag 1 correlations than for the other three statistics. Secondly, numerous past studies, including [5], [7], [10] and [11] indicate that mean annual rainfall interpolation requires a spline model that incorporates a spatially varying dependence on elevation. This data set was not large enough to support a trivariate model.

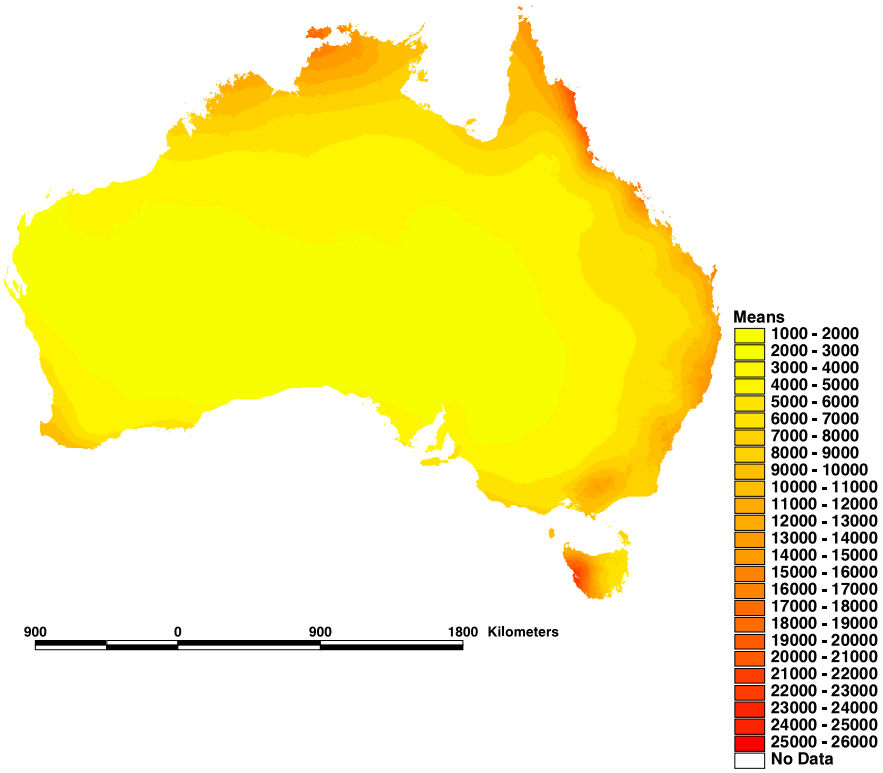


Figure 18: Fitted surface for the means

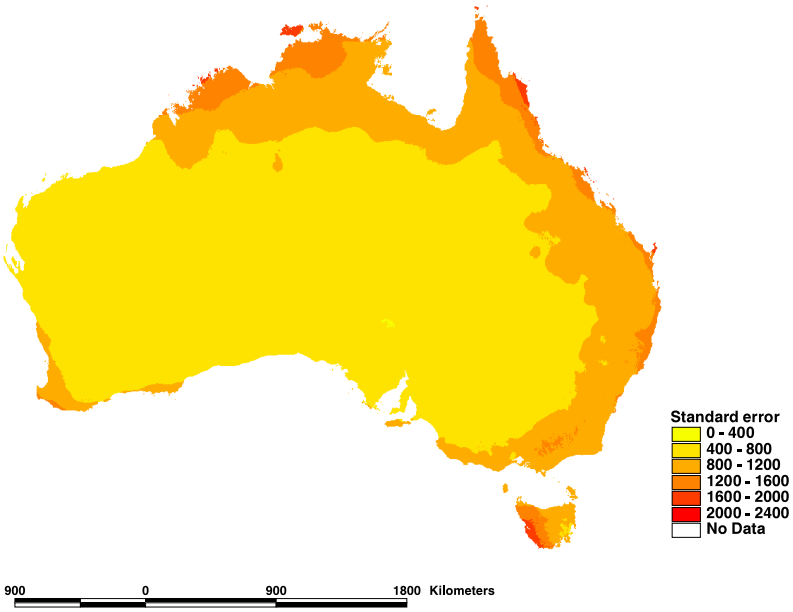


Figure 19: Standard error surface for the means

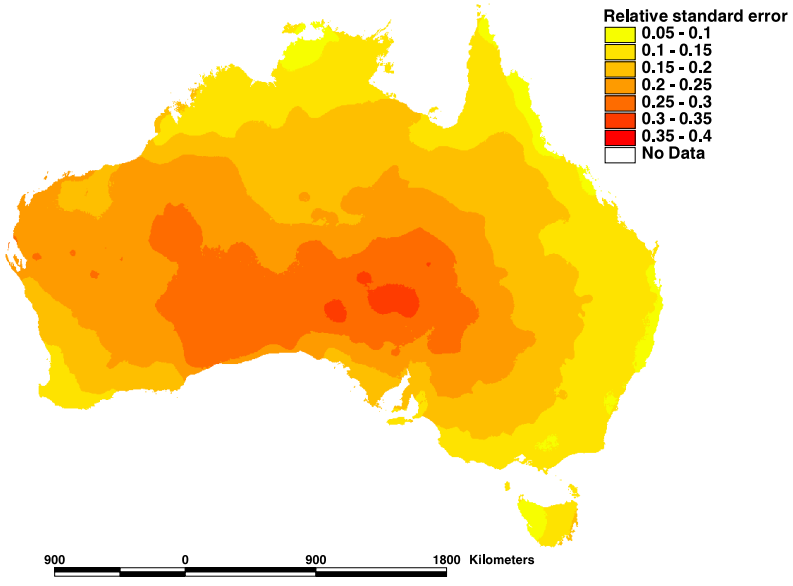


Figure 20: Relative standard error surface for the means

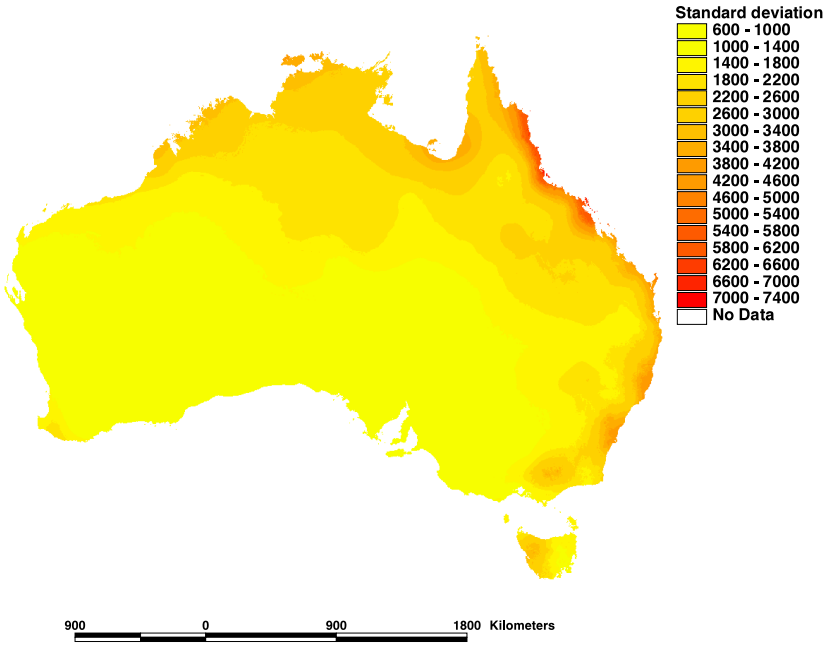


Figure 21: Fitted surface for the standard deviations

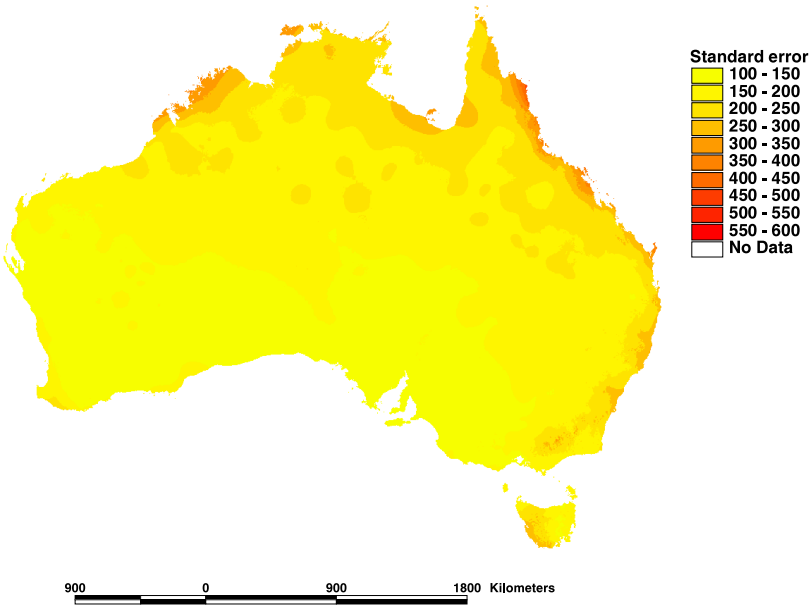


Figure 22: Standard error surface for the standard deviations

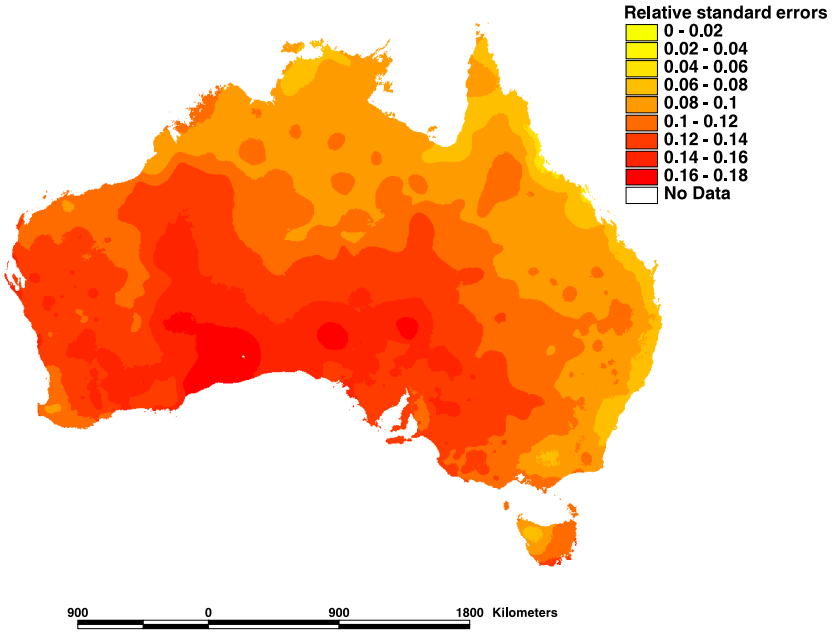


Figure 23: Relative standard error surface for the standard deviations

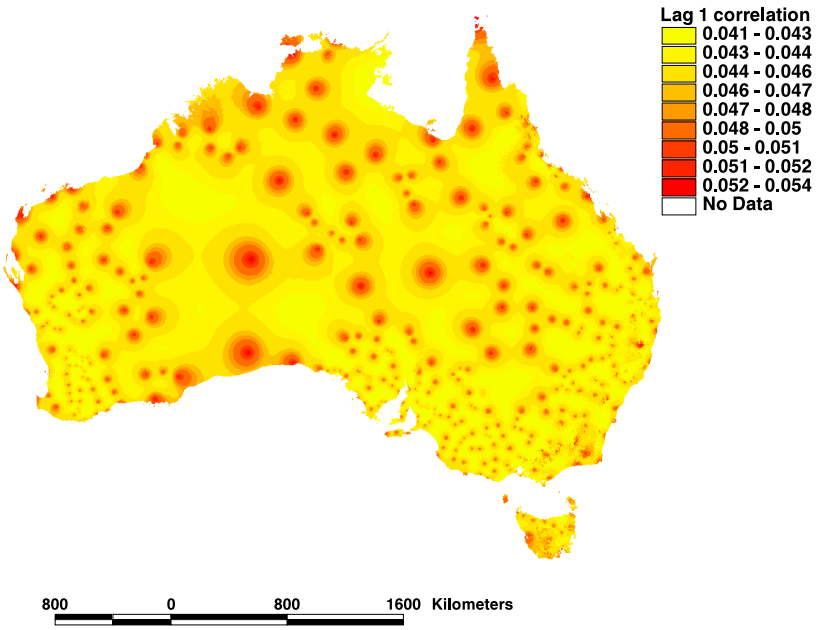


Figure 24: Standard error surface for lag 1 correlations, using the trivariate spline model

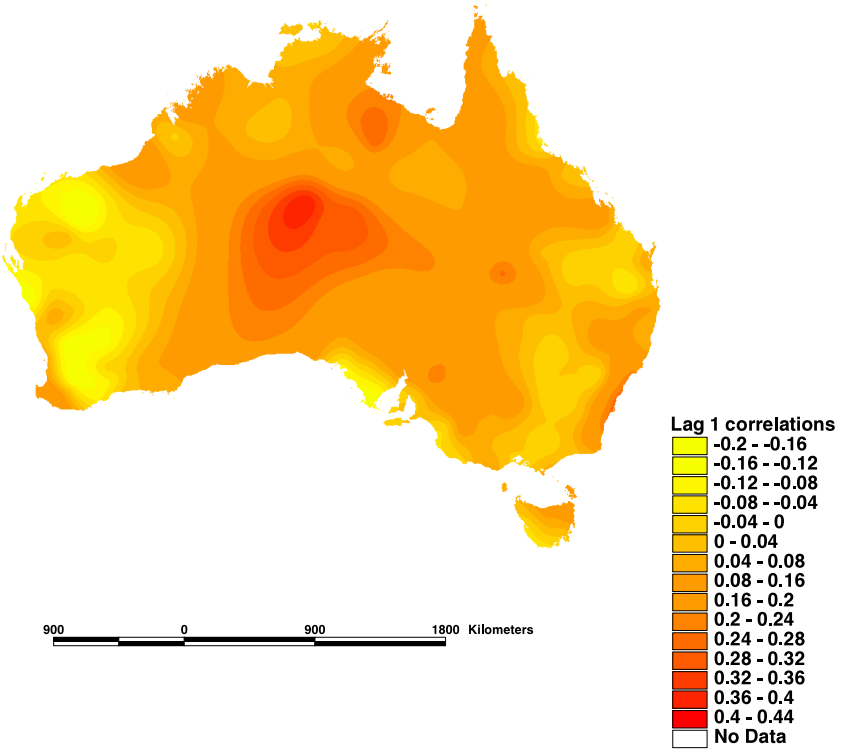


Figure 25: Fitted surface for the lag 1 correlations

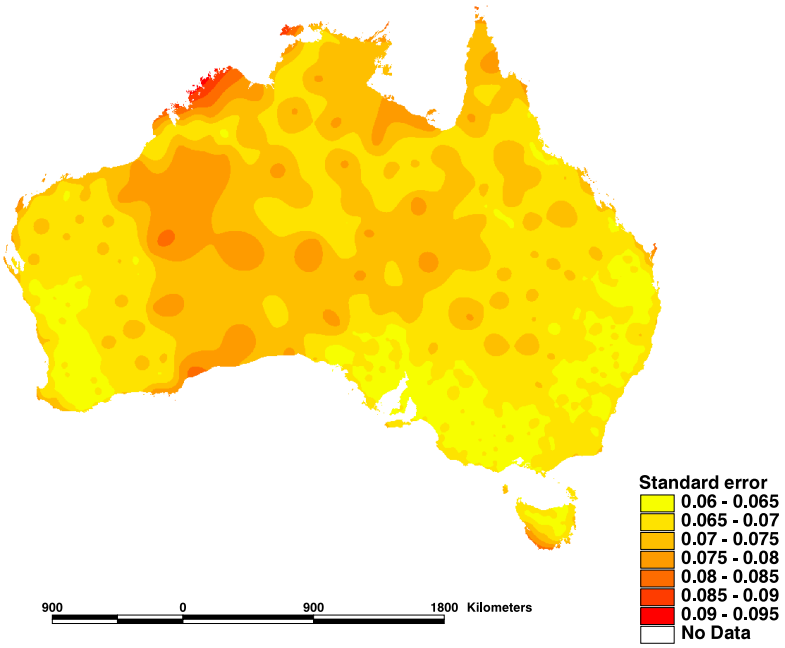


Figure 26: Standard error surface for the lag 1 correlations

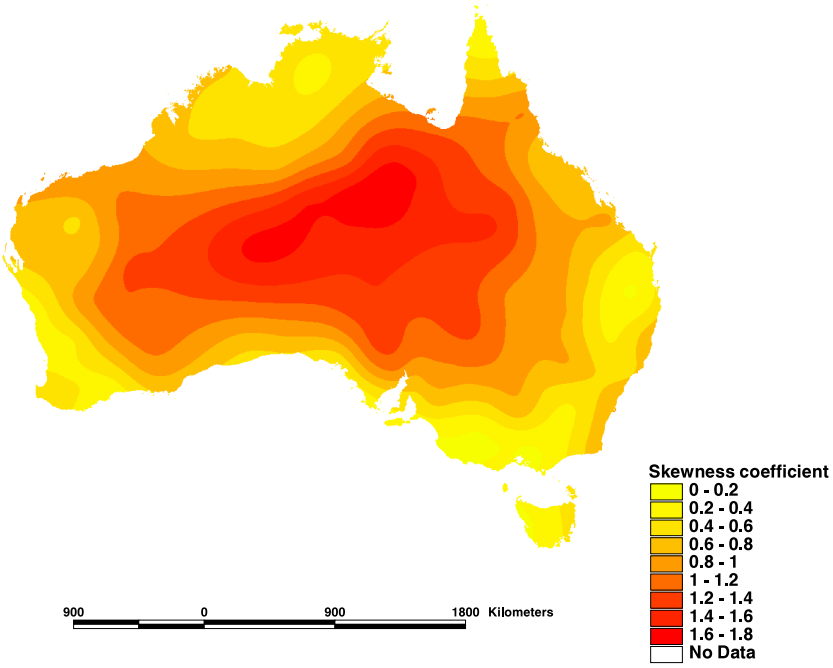


Figure 27: Fitted surface for the skewness coefficients

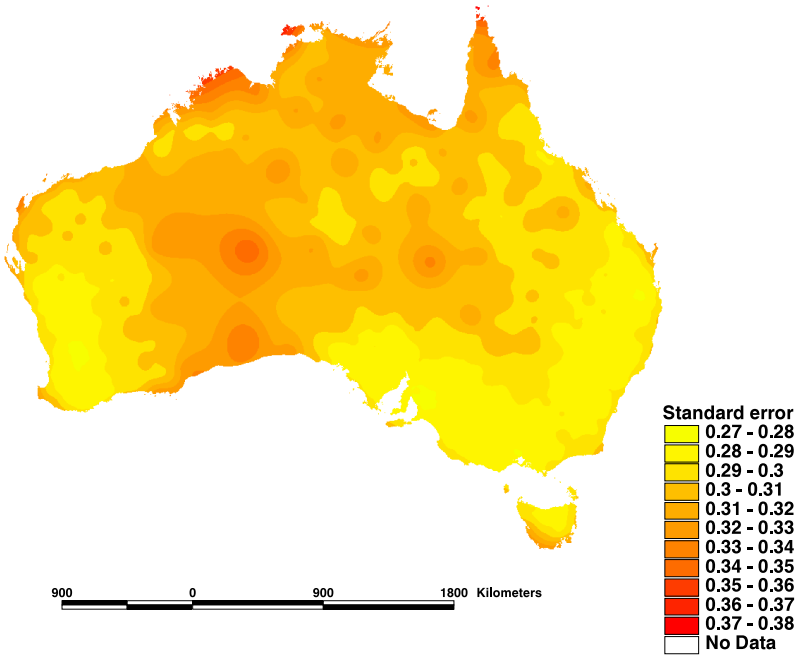


Figure 28: Standard error surface for the skewness coefficients

4 Application to the full data set

4.1 Data

Following the conclusions of the above analysis, thin plate smoothing spline surfaces were fitted to a larger data set consisting of 6334 observations of the parameters of the AR(1) model. Plots of the AR(1) parameters are shown in Figures 29-33. There is considerably more variability than was observed for the subset of 363 stations. It is possible that this larger data set contains higher data error than the smaller subset, given that the 363 points were known to be 'quality' stations. This may be a factor contributing to the higher variability in Figures 29-33. As with the previous analysis, there is still more coherence in the plots for mean and standard deviation than there is in the plots for the other 3 parameters. The skewness and correlation coefficients show particularly high noise. There is one clear outlier in the plots for mean and standard deviation. This was identified to be station 31141, which is marked on the maps in Figures 34 and 35. This station is on the top of Mt Belleden Ker, and has an elevation of 1555m.
The maps in Figures 34-38 show similar broad spatial patterns to those observed for the smaller data set. For the means and standard deviations, the majority of the variability occurs at the coastline. The coefficient of variation, correlations and skewness show a more gradual inland gradient. The stations that have been singled out in these maps relate to results from withheld data, and will be discussed in following paragraphs.

4.2 Thin plate smoothing spline interpolation

This plate smoothing splines were fitted to each of the parameters using the ANUSPLIN program SPLINB. SPLINB is used when the data set is too large to calculate surfaces using SPLINA. SPLINB generates an approximation to the thin plate smoothing spline solution, using knots [9]. For these surfaces, 1800 knots were selected using the ANUSPLIN program SELNOT. Once again, bivariate, partial and trivariate spline models were tested, to compare predictive capacity according to GCV values, and the amount of smoothing

Spline model	Square root of GCV (mm)	Signal
Bivariate, square root transformation	161.3	938.2
Partial, square root transformation	152.3	878.6
Trivariate, square root transformation	151.7	885.0

Table 6: Summary results for thin plate smoothing spline fits to annual mean rainfall.

as indicated by the signal values. The analysis for the smaller data set showed a significant improvement in the spline models for the mean and standard deviation when a square root transformation was performed on the data. This transformation was therefore used in this analysis. The results are shown in Tables 6-10.

For the means and standard deviations, the GCV shows that incorporation of elevation is advantageous. There is, however, little difference between the partial spline model and the trivariate model. This indicates that the data contains high noise level, so a less complicated, more robust model such as the partial spline model is comparable with the more spatially complex trivariate model. Signal values do not approach the number of knots, indicating that the spline models are reliable.

The trivariate spline model performs only marginally better than the partial spline model for the means and standard deviations. It could therefore be argued that the simpler model should be preferred, because there is no indication that the complex model provides a better explanation of the process. However, given that rainfall is widely known to display a spatially varying dependence on elevation, and the signal values and validation statistics show that the data do support the trivariate model, the trivariate model was chosen to produce surfaces for the mean and standard deviation.

Spline model	Square root of GCV (mm)	Signal
Bivariate, square root transformation	51.4	866.8
Partial, square root transformation	47.7	800.7
Trivariate, square root transformation	47.2	836.2

Table 7: Summary results for thin plate smoothing spline fits to the standard deviation of annual mean rainfall.

Spline model	Square root of GCV	Signal
Bivariate	0.127	572.5
Partial	0.127	573.0
Trivariate	0.127	619.8

Table 8: Summary results for thin plate smoothing spline fits to the lag 1 correlation coefficient of variation of annual mean rainfall.

Spline model	Square root of GCV	Signal
Bivariate	0.103	1337.4
Partial	0.103	1338.6
Trivariate	0.103	1323.5

Table 9: Summary results for thin plate smoothing spline fits to the coefficient of variation of annual mean rainfall.

Spline model	Square root of GCV	Signal
Bivariate	0.353	836.1
Partial	0.353	839.6
Trivariate	0.352	884.4

Table 10: Summary results for thin plate smoothing spline fits to the skewness coefficient of annual mean rainfall.

For the correlations, coefficient of variation and the skewness the difference in GCV is negligible for all 3 models, indicating that they do not contain a spatial dependence on elevation. The signal values are also similar, except in the case of the correlation coefficients, where the signal for the trivariate spline is higher, though still less than half the number of knot points. Bivariate models were chosen for these three statistics.

The resulting surfaces are shown in Figures 39-43. The patterns are similar to those obtained for the smaller data set. The means and standard deviations show extreme variability at coastal regions, and no visible changes inland. Trends for the coefficient of variation, skewness and correlations are more broadscale.

Residuals of the data points from the fitted surface are shown in Figures 44-48. For the mean and standard deviation surfaces, Mt Belleden Ker is the clear outstanding residual. The mean value for Mt Belleden Ker is much higher than that of surrounding stations, so it is not surprising that the fitted surface underestimated this withheld point. Apart from this residual, most of the residuals of the withheld data values from the fitted surface are below 1000mm for the mean surface, 200mm for the standard deviation surface, and 0.15 for the coefficient of variation.

In order to test the predictive abilities of these fitted surfaces, a sample of 100 data points was withheld and the surfaces were refitted. The sample was again selected using the ANUSPLIN program SELNOT, to select a spatially representative sample. Mt Belleden Ker was initially in the selection of withheld points, as would be expected, but it was returned to the data set so as not to bias the statistics calculated on the withheld data. Summary results are shown in Table 11. The maximum residuals tend to be high. This is a manifestation of choosing a spatially representative sample, which results in the selection of stations at high altitudes, and stations in areas where data are very sparse. Withholding these stations therefore withholds valuable information from the fitted surface. This is therefore a stringent test of predictive ability.

The statistics for the withheld data validate the spline model statistics in Tables 6-10. The RMS residual of the withheld data agrees well with the

GCV values for the optimal spline models in Tables 6-10. This demonstrates that the spline models used in this analysis have accurately estimated the noise in the data, and are likely to provide a reliable representation of the underlying broadscale trends. Notably, the RMS residual of the withheld data for the bivariate spline model for the mean was 176mm, larger than the value of 161mm obtained for the trivariate model. Thus the results of the withheld data analysis indicates higher accuracy for the trivariate spline model for annual mean rainfall, which is consistent with the GCV values for the different spline models, shown in Table 11.

Plots of the residuals of withheld data residuals are shown in Figures 21-25, with Mt Belleden Ker also withheld. These plots show why Mt Belleden Ker was returned to the data set. Its residual is very high, which would be expected given that there are insufficient data in this region to represent topographic effects on rainfall. For the skewness and correlation coefficients, the residuals are essentially homogeneous, which is further indication that these statistics are not strongly related to topography.

The locations of the maximum residuals for each parameter are marked on the maps in Figures 6-10. The colours show that the values of the outliers are very different to the values at surrounding stations, and that they are usually located in an area where patterns are changing steeply. The fact that the outliers are quite different from surrounding stations may indicate data error, or may indicate high spatial variability that cannot be captured by the model. In the case of Mt Belleden Ker, spatial variability is more likely to be responsible for the underestimation, given that its elevation is much higher than that of surrounding stations.

4.3 Conclusions

The statistics corresponding to the spline models generated for this data set indicate that the spline surfaces are reliably fitted, in that the signal values do not approach exact interpolation. The surfaces for this data set have considerably more fine scale structure than those generated for the smaller data set. The large outliers from the fitted surfaces for the mean and standard

Parameter	RMS of residuals of withheld data values from fitted values (mm)	Maximum residual of withheld values from fitted values (mm)	Station number for maximum for maximum residual
Mean	161	684	33658
Standard deviation	58.3	287	33658
Skewness coefficient	0.455	1.95	41243
Correlation coefficient	0.179	1.03	14806
Coefficient of variation	0.0539	0.163	44174

Table 11: Results of withholding a sample of 100 data points from the fitted surfaces for each parameter, for the chosen thin plate smoothing spline model.

deviation are a result of the high spatial variability of rainfall. The outliers were located in regions with complex topography and often very little data. The spline surfaces are unable to represent the localised rainfall patterns that occur in such areas. Apart from the small number of high outliers, the analysis of residuals from the fitted surfaces has indicated that surfaces predict the AR(1) parameters across the Australian continent with a useful degree of accuracy.

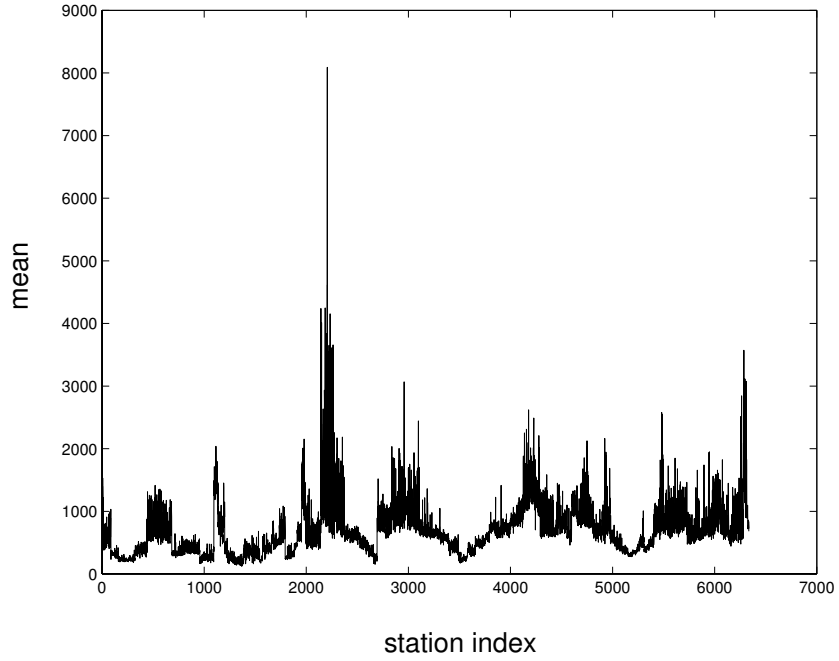


Figure 29: Annual mean rainfall for each station

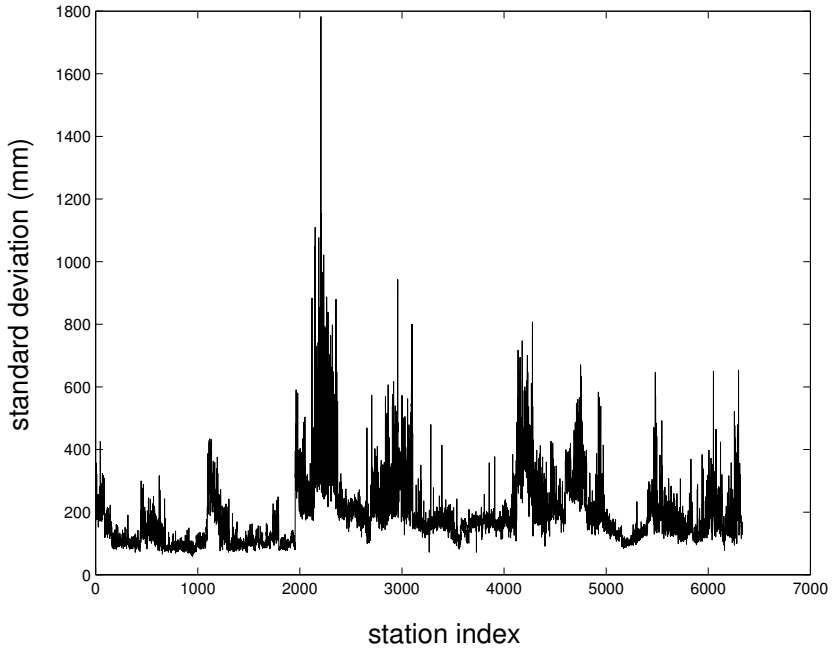


Figure 30: Standard deviation of annual mean rainfall for each station

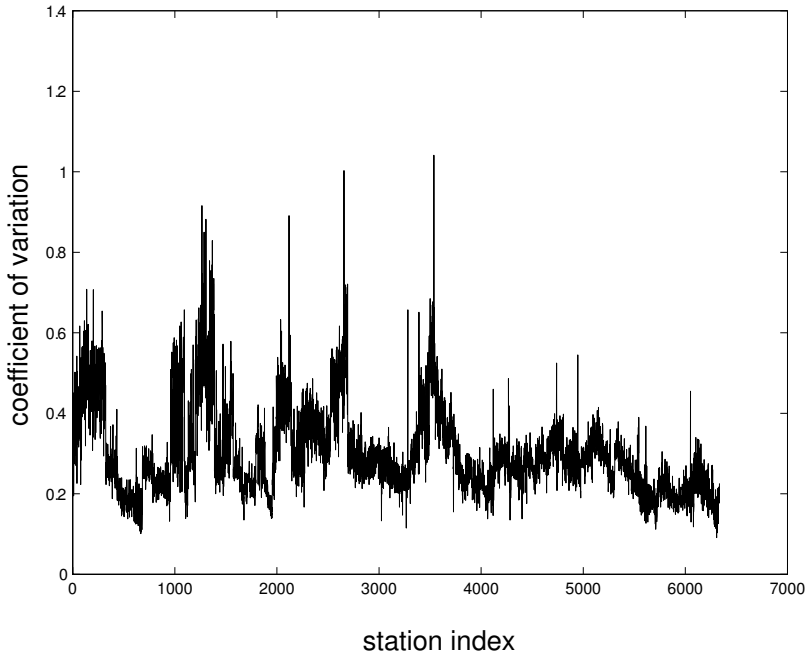


Figure 31: Coefficient of variation of annual mean rainfall for each station

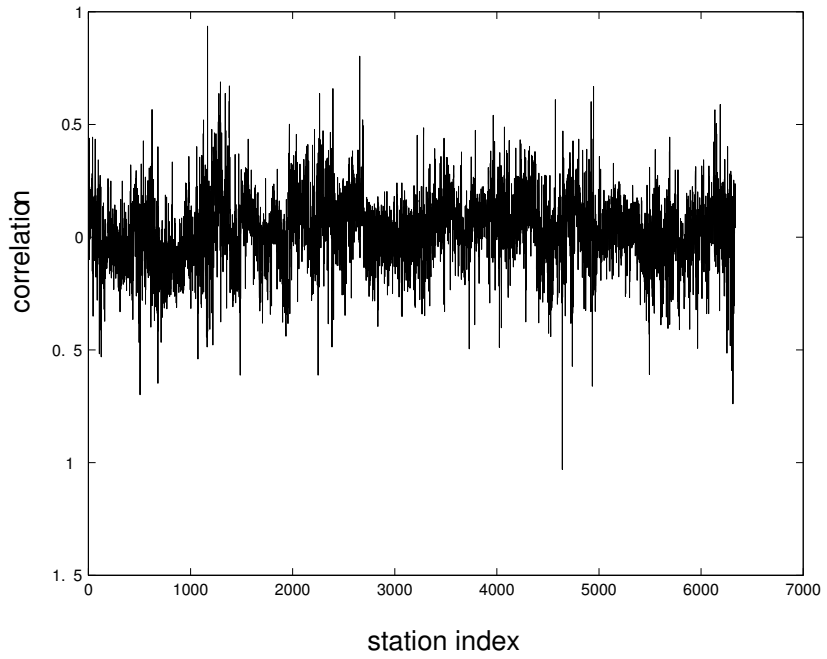


Figure 32: Correlation coefficient of annual mean rainfall for each station

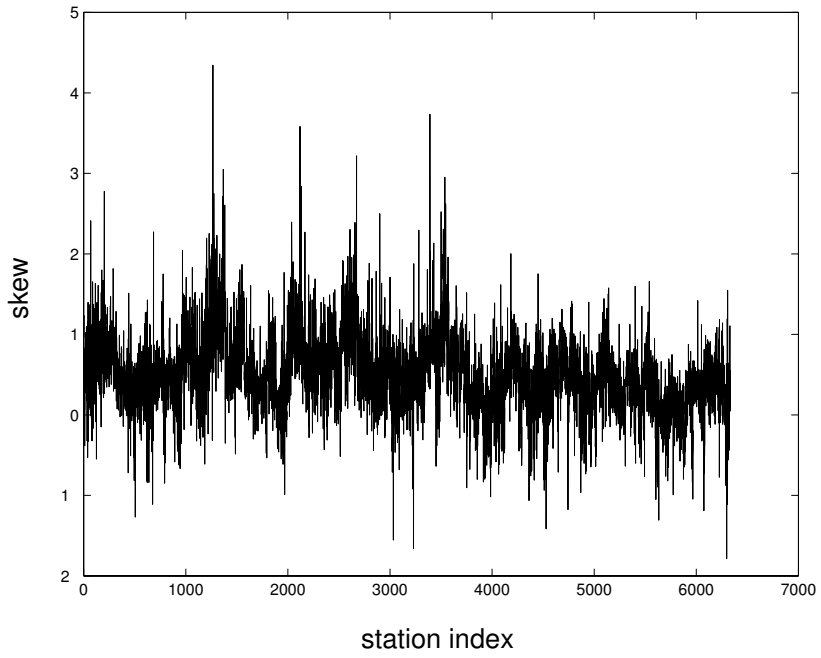


Figure 33: Skewness coefficient of annual mean rainfall for each station

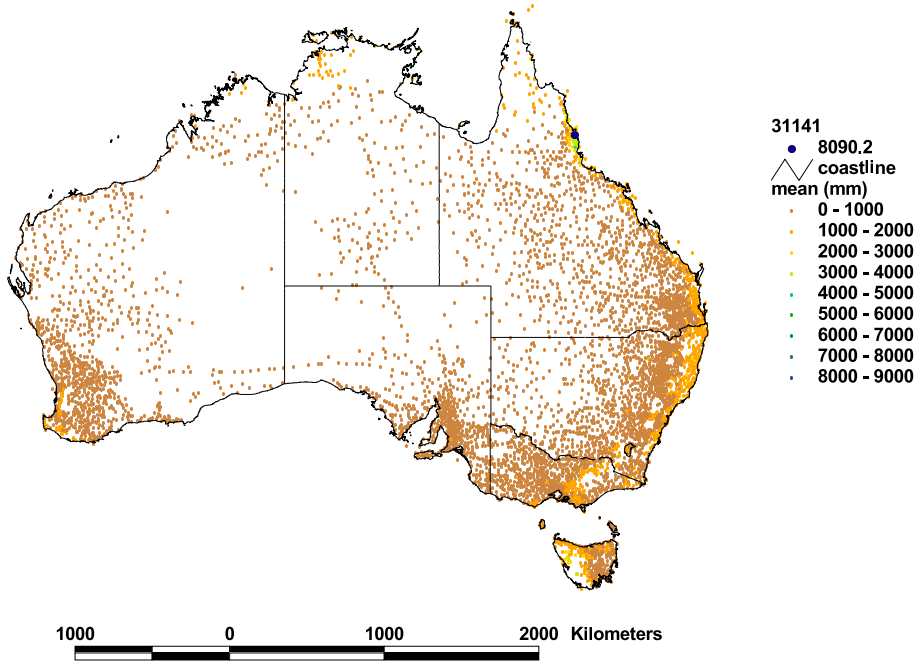


Figure 34: Annual mean rainfall

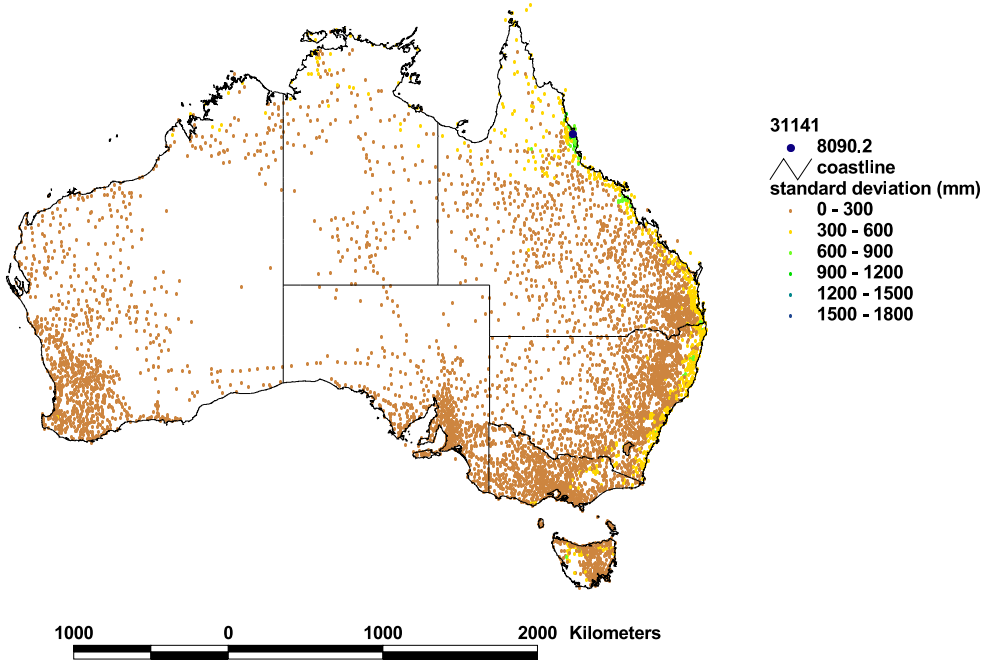


Figure 35: Standard deviation of annual mean rainfall

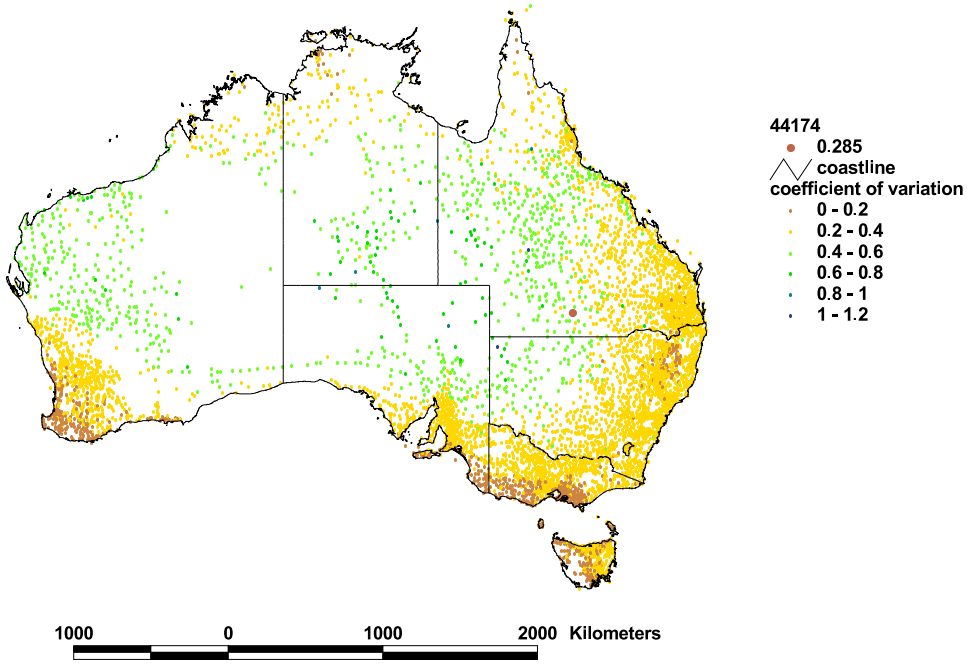


Figure 36: Coefficient of variation of annual mean rainfall

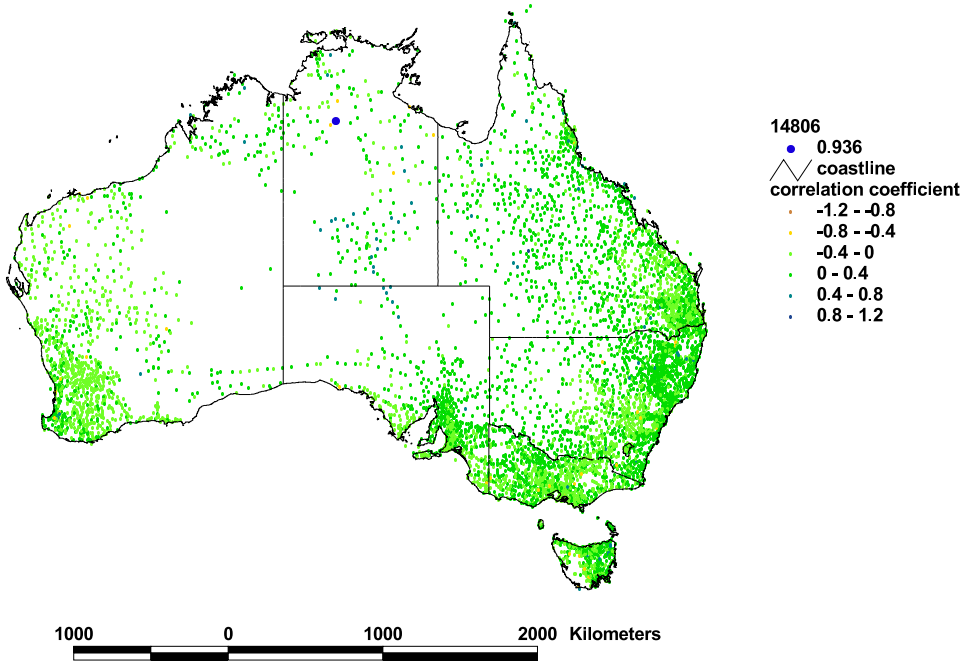


Figure 37: Correlation coefficient of annual mean rainfall

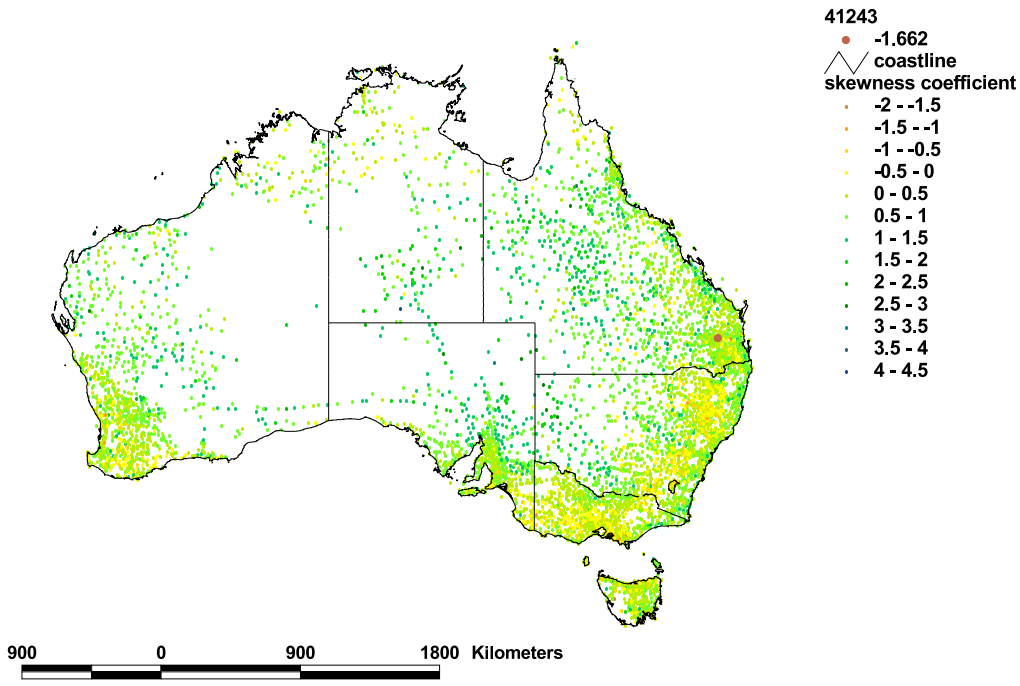


Figure 38: Skewness coefficient of annual mean rainfall

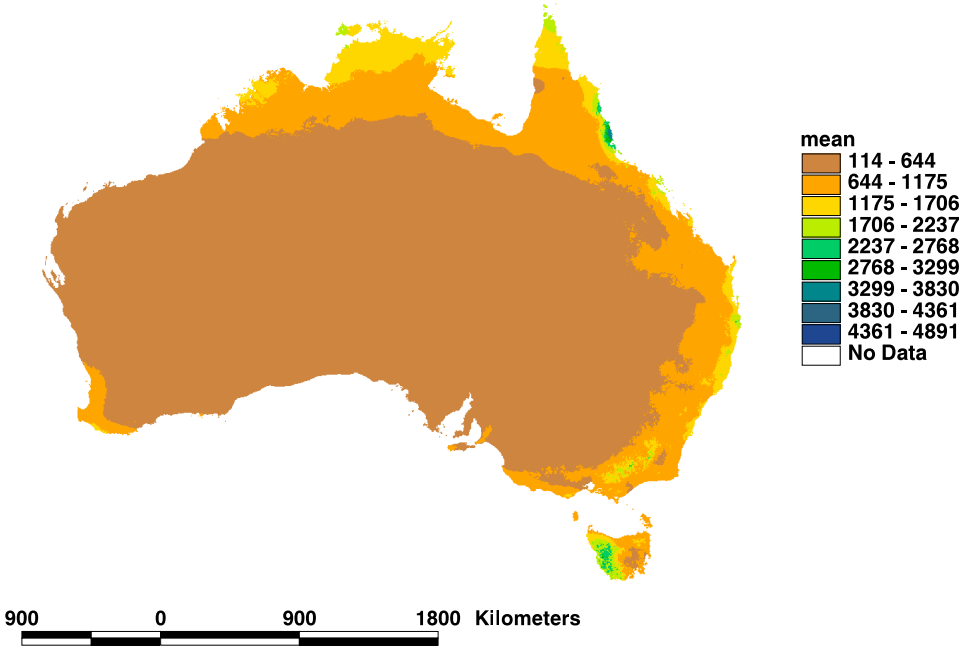


Figure 39: Annual mean rainfall surface

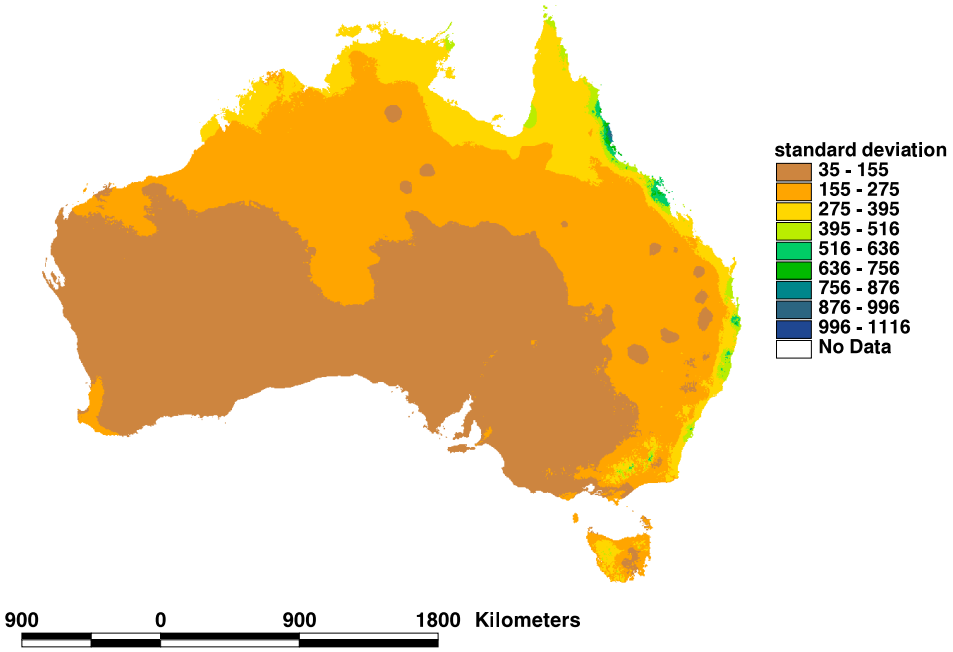


Figure 40: Surface for the standard deviation of annual mean rainfall

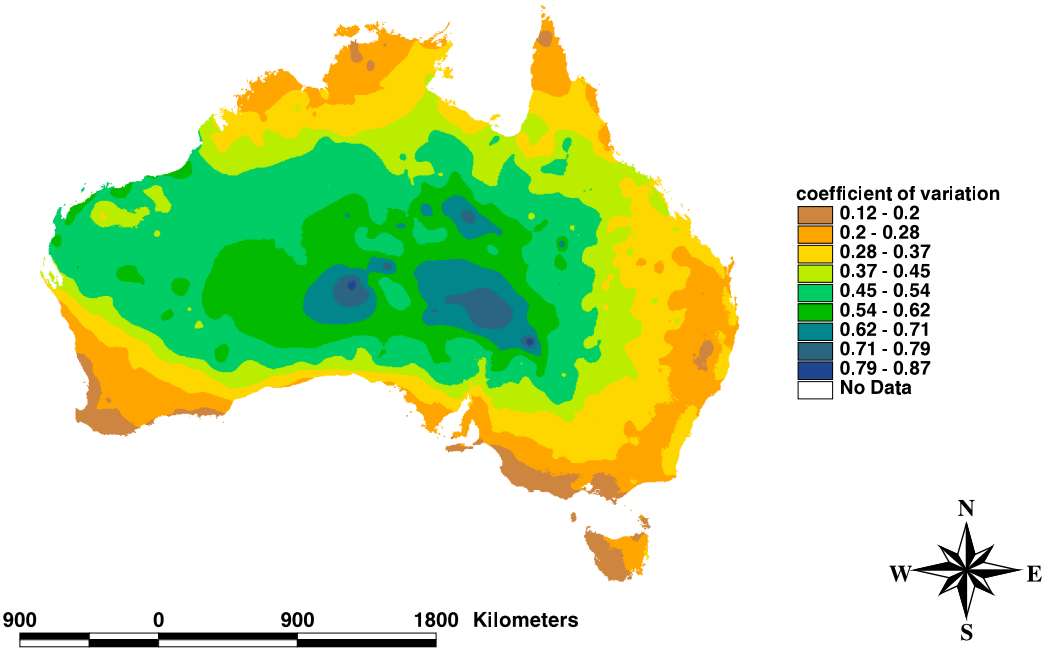


Figure 41: Surface for the coefficient of variation of annual mean rainfall

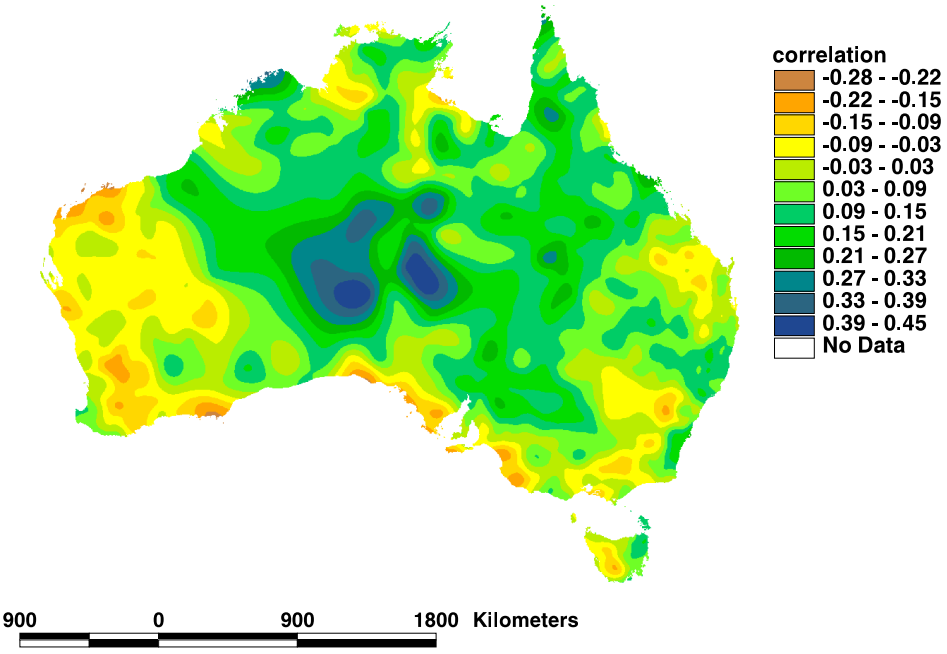


Figure 42: Surface for the correlation coefficient of annual mean rainfall

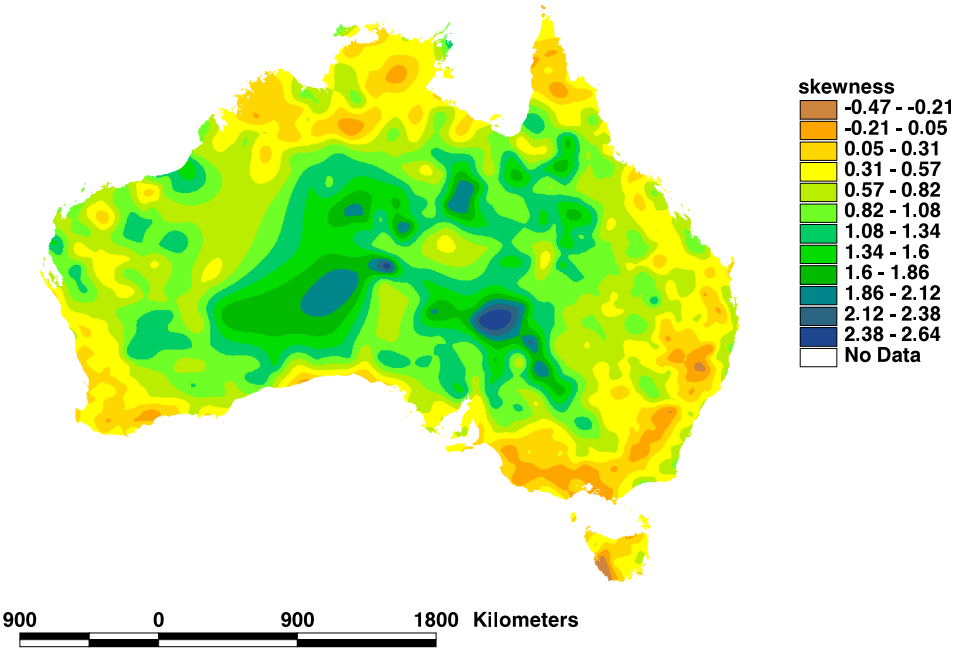


Figure 43: Surface for the skewness coefficient of annual mean rainfall

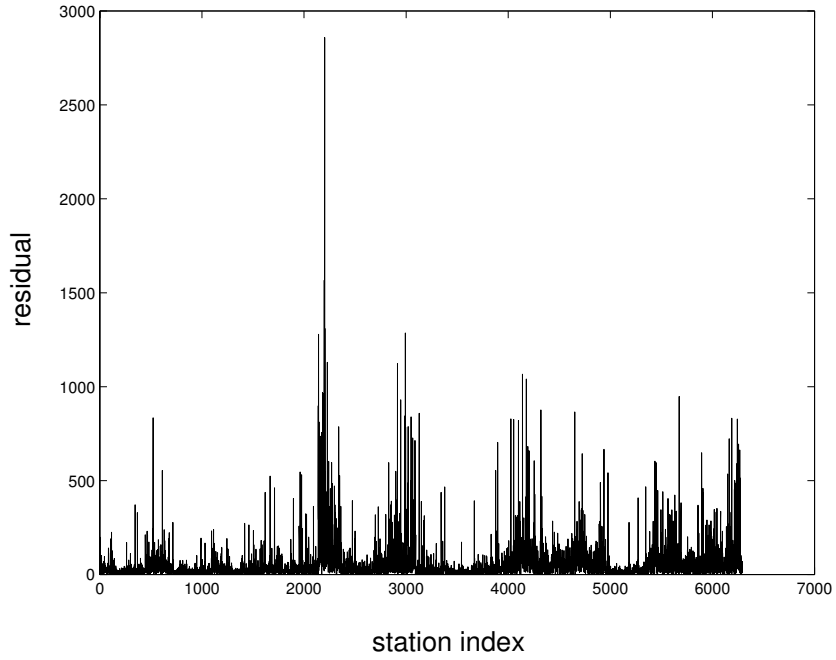


Figure 44: Residuals from the fitted surface for annual mean rainfall

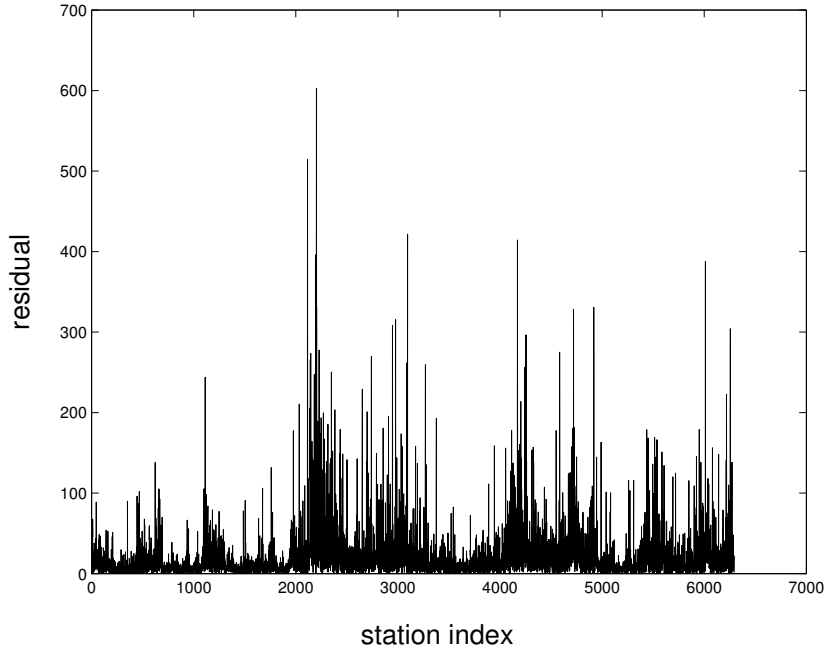


Figure 45: Residuals from the fitted surface for the standard deviation of annual mean rainfall

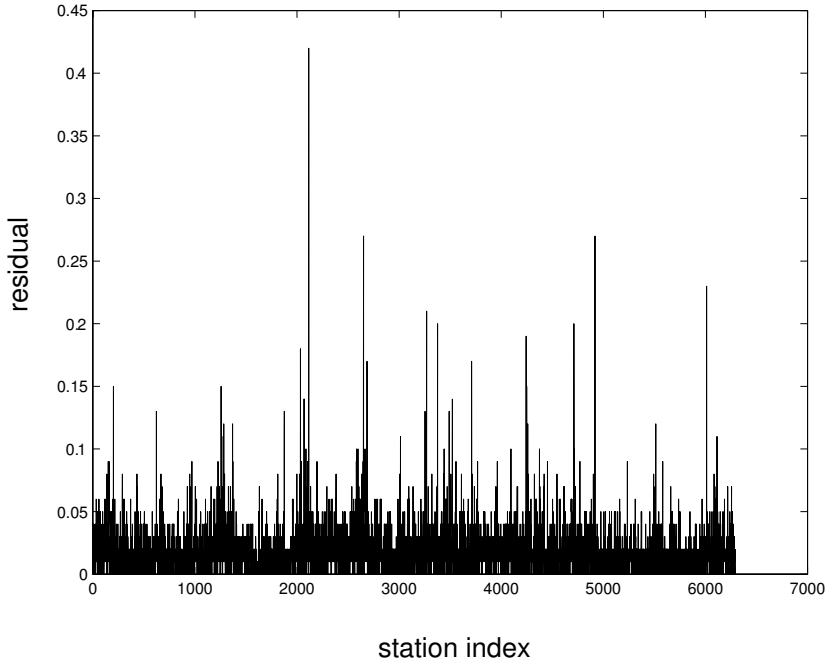


Figure 46: Residuals from the fitted surface for the coefficient of variation of annual mean rainfall

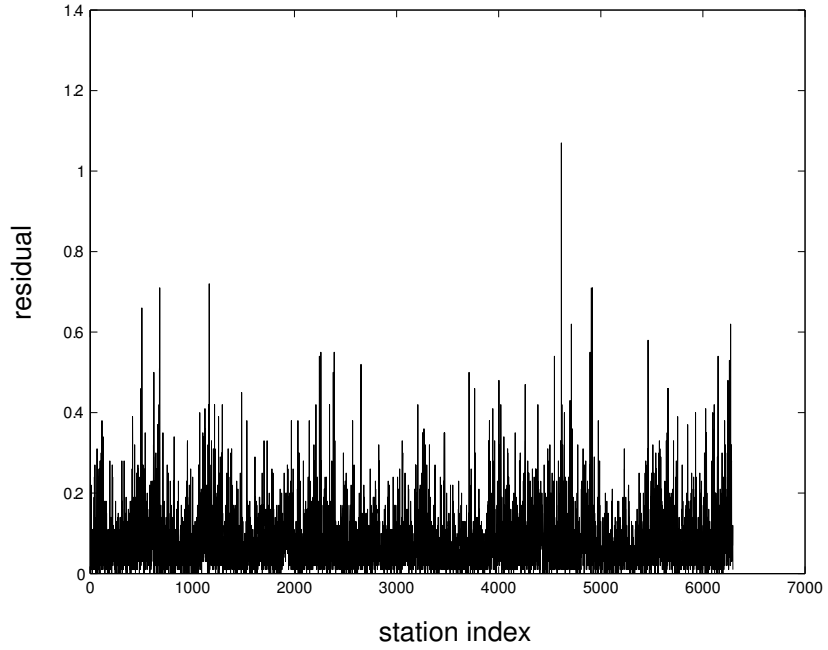


Figure 47: Residuals from the fitted surface for the correlation coefficient of annual mean rainfall

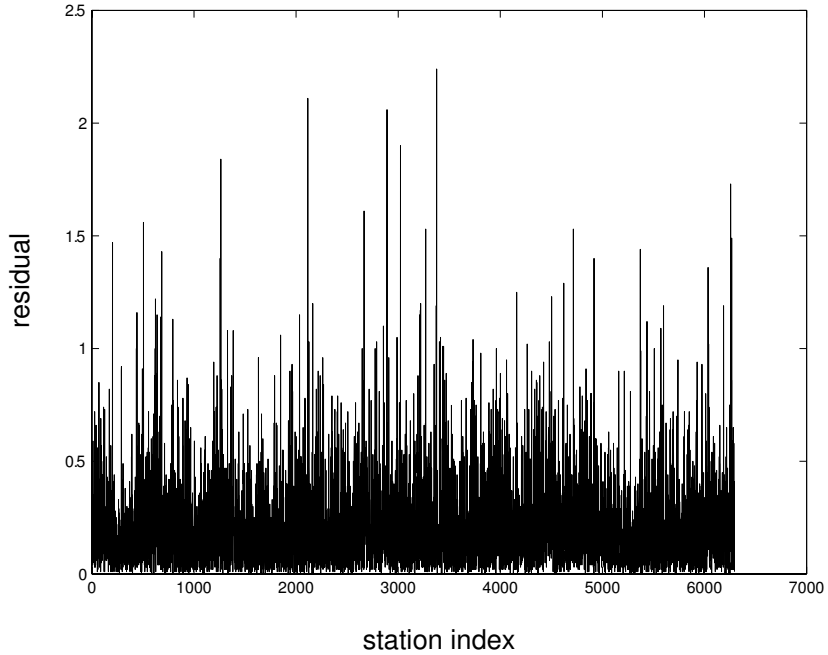


Figure 48: Residuals from the fitted surface for the skewness coefficient of annual mean rainfall

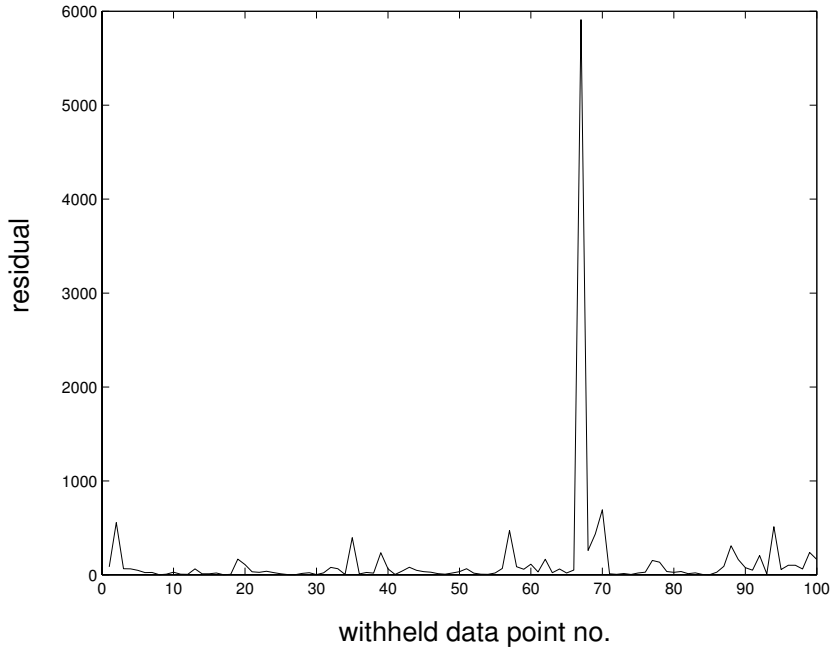


Figure 49: Withheld data residuals for annual mean rainfall

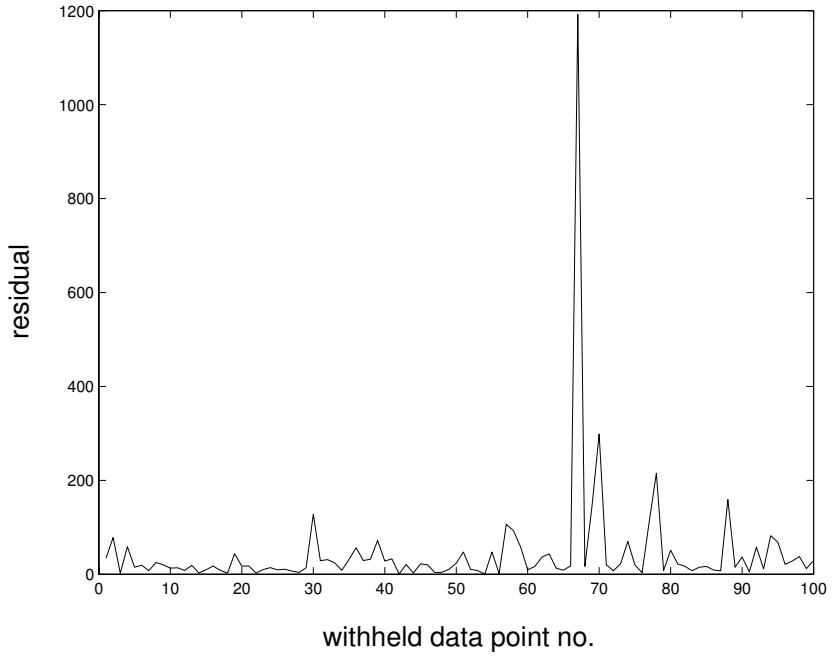


Figure 50: Withheld data residuals for the standard deviation of annual mean rainfall

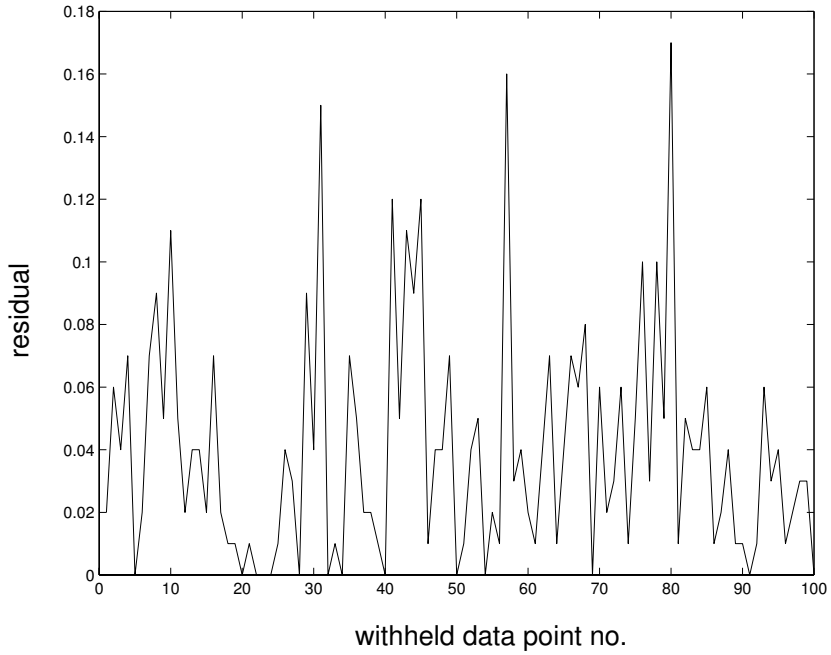


Figure 51: Withheld data residuals for the coefficient of variation of annual mean rainfall

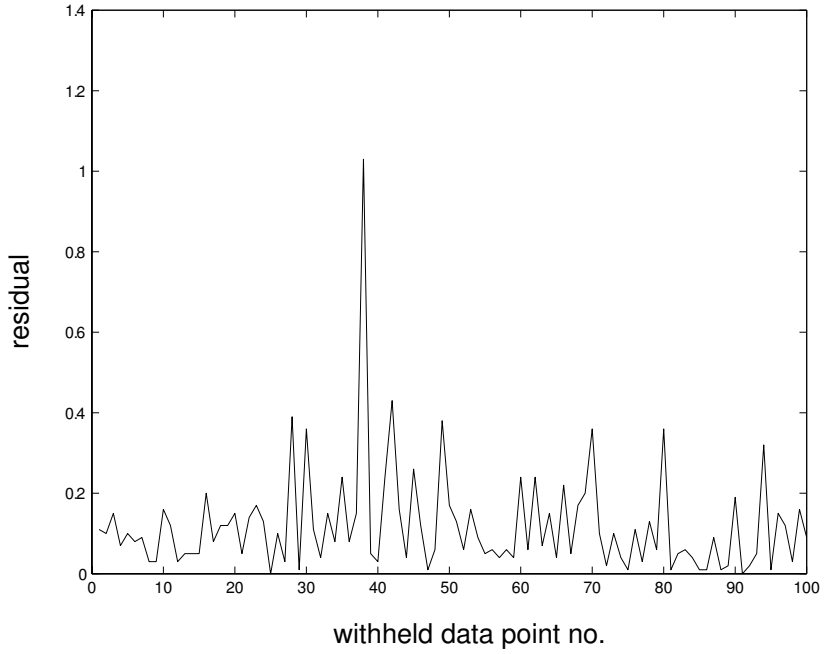


Figure 52: Withheld data residuals for the correlation coefficient of annual mean rainfall

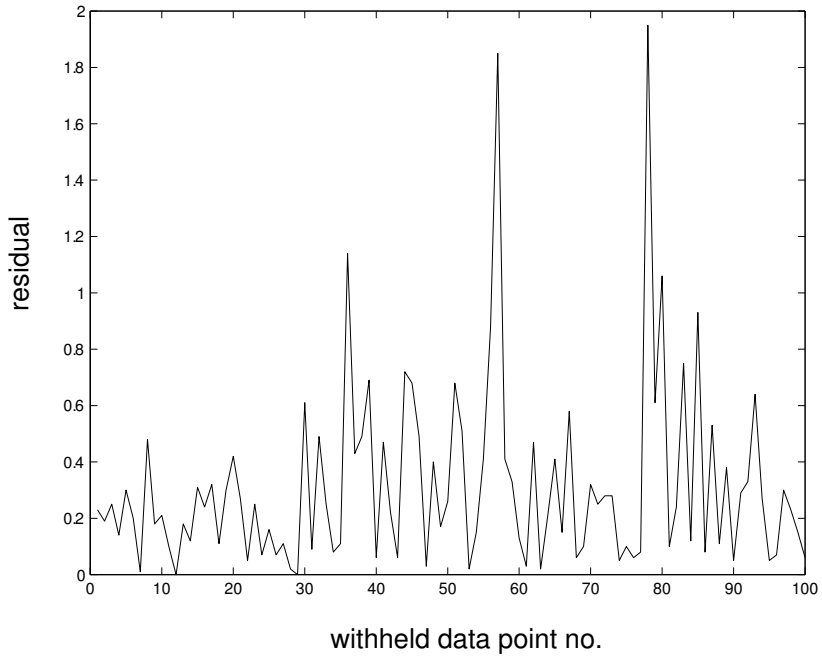


Figure 53: Withheld data residuals for the skewness coefficient of annual mean rainfall

References

[1] P. Craven and G. Wahba, Smoothing noisy data with spline functions, *Numerische Mathematik*, 31 (1979) 377–403.

[2] P.J. Diggle and M.F. Hutchinson, On spline smoothing with autocorrelated errors, *Australian Journal of Statistics*, 31 (1989) 166–182.

[3] J. Duchon, Splines minimizing rotation-invariant semi-norms in Sobelov spaces, In: *Constructive Theory of Functions of Several Variables*, Springer-Verlag, Berlin, (1977) 85–100.

[4] G.H. Golub and U.V. Matt, Generalized cross-validation for large scale problems: revised version, *Journal of Computational and Graphical Statistics*, 6 (1997) 1–34.

[5] M.F. Hutchinson, The application of thin plate smoothing splines to continent-wide data assimilation, In: J.D. Jasper (ed), *Data Assimila-*

tion Systems: BMRC Research Report No. 27, Bureau of Meterology, Melbourne, (1991) 104–113.

[6] M.F. Hutchinson, On thin plate splines and kriging, *Computing and Science in Statistics*, 25 (1993) 55–62.

[7] M.F. Hutchinson, Interpolating mean rainfall using thin plate smoothing splines, *International Journal of Geographic Information Systems*, 9(4) (1995) 385–403.

[8] M.F. Hutchinson, Stochastic space-time weather models from ground-based data, *Agricultural and Forest Meteorology*, 73 (1995) 237–264.

[9] M.F. Hutchinson, *ANUSPLIN version 4.2 User Guide*, http://cres.anu.edu.au/outputs/anusplin.html (2002)

[10] M.F. Hutchinson, Interpolation of rainfall data with thin plate smoothing splines:I. Two dimensional smoothing of data with short range correlation, *Journal of Geographic Information and Decision Analysis*, 2(2) (1998) 153–167.

[11] M.F. Hutchinson, Interpolation of rainfall data with thin plate smoothing splines: II. Analysis of topographic dependence, *Journal of Geographic Information and Decision Analysis*, 2(2) (1998) 168–185.

[12] M.F. Hutchinson and R.J. Bischof, A new method for estimating the spatial distribution of mean seasonal and annual rainfall applied to the Hunter Valley, New south Wales, *Australian Meteorology Magazine ?check*, 31 (1983) 179–184.

[13] M.F. Hutchinson and P.E. Gessler, Splines - more than just a smooth interpolator, *Geoderma*, 62 (1994) 45–67.

[14] E. Isaaks and R. Srivistava, *An Introduction to Applied Geostatistics*, Oxford University Press, Oxford (1989).

[15] J. Meinguet, Multivariate interpolation at arbitrary points made simple, *Journal of Applied Mathematical Physics (ZAMP)*, 30 (1979) 292–304.

[16] A.B. Pittock, Climatic Change and the Patterns of Variation in Australian Rainfall, *Search* 6(11-12) (1975) 498–504.

[17] A.B. Pittock, Recent climatic change in Australia: implications for a CO_2-warmed earth, *Climatic Change* 5 (1983) 231–340.

[18] J.A. Rice, *Mathematical Statistics and Data Analysis (second edition)*, Wadsworth Publishing Company, Belmont, California, (1995).

[19] I.J. Schoenberg, Spline functions and the problem of graduation, *Proceedings of the National Academy of Science* 52 (1964) 947–950.

[20] R. Srikanthan and T.A. McMahon, Stochastic generation of rainfall and evaporation data, *AWRC Technical Paper*, No. 84 (1985), 301pp.

[21] G. Wahba, Bayesian "confidence intervals" for the cross-validated smoothing spline, *Journal of the Royal Statistical Society Series B*, 45 (1983) 281–299.

[22] G. Wahba, Spline models for observational data, In: *CBMS-NSF Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics, Philadelphia, (1990).

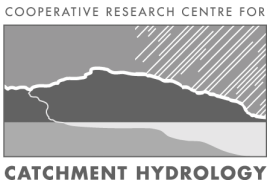
[23] R.F. Warner, The impacts of alternating flood-and-drought-dominated regimes on channel morphology at Penrith, New South Wales, Australia, *The Influence of Climate Change and Climatic Variability on the Hydrologic Regime and Water Resources* (Proceedings of the Vancouver Symposium, August 1987), IAHS Publication no. 168 (1987) 327–338.

[24] X. Zheng and R. Basher, Mapping rainfall fields and their ENSO variation in the data-sparse tropical south-west Pacific Ocean Region, *International Journal of Climatology*, 18 (1998) 237–251.

The Cooperative Research Centre for Catchment Hydrology
is a cooperative venture formed under
the Commonwealth CRC Program between:

- Brisbane City Council
- Bureau of Meteorology
- CSIRO Land and Water
- Department of Land and Water Conservation, NSW
- Department of Sustainability and Environment, Vic
- Goulburn-Murray Water
- Griffith University
- Melbourne Water
- Monash University
- Murray-Darling Basin Commission
- Natural Resources and Mines, Qld
- Southern Rural Water
- The University of Melbourne
- Wimmera Mallee Water

Associate:
Water Corporation of Western Australia



Centre Office
Department of Civil Engineering, PO Box 60, Monash University, Victoria, 3800 Australia.
Telephone: + 61 3 9905 2704 Facsimile: +61 3 9905 5033 Email: crcch@eng.monash.edu.au
http://www.catchment.crc.org.au