New Methods of Infilling Southern African Raingauge Records Enhanced by Annual, Monthly and Daily Precipitation Estimates Tagged with Uncertainty

Report to the
Water Research Commission

by

GGS Pegram and and Scott Sinclair
Pegram and Associates (Pty) Ltd

&
András Bárdossy
Universität Stuttgart

WRC Report No. 2241/1/15
ISBN 978-1-4312-0758-9

March 2016

Obtainable from
Water Research Commission
Private Bag X03
Gezina, 0031

orders@wrc.org.za or download from www.wrc.org.za

Executive Summary

The available maps and tables describing Mean Annual Precipitation (MAP) and Mean Monthly Precipitation (MMP) over Southern Africa are over 10 years old and need updating. To do this as well as possible, and to add an important measure of reliability to these estimates, new mathematical and statistical tools for infilling gauge data and interpolation over space were developed and tested against existing methods and found to be a meaningful improvement. The MAP and MMP maps of areal estimates over quaternary catchments in the region were produced for the use of hydrological practitioners and are incorporated in an addendum. Daily, monthly and annual rainfall records were infilled where the inter-station distance was small enough to make the estimates meaningful and the technique developed was also tested successfully on daily records. The worth of the infilled data were indicated by computing the expected (mean) value, augmented by the median and the upper and lower deciles of the distributions of the estimates, as well as the probability of dry and of exceeding a pre-determined threshold.

In other work, the links between satellite estimates of rainfall, in particular TRMM, were compared against spatially interpolated raingauge data over the TRMM pixels. The result was a reasonable match over months, but a poor match at the daily scale. As an alternative approach, a novel proposal for quantile-quantile adjustment of TRMM (and its successor, NASA's GPM) was mooted. However, with a dwindling raingauge network, it is important (if expensive) to augment it with gauges not further apart than 25 km else there is poor correlation between them at the daily scale. Without a reasonably dense gauge network in the wetter regions, there is no meaningful way of ground referencing remotely sensed rainfall using satellites and radar, whose rainfall measurements we know to be biased. Finally, the repaired annual and monthly data, together with their error estimates, and the maps and algorithms developed and used, are available on a CD accompanying the report and summarised in Chapter 10.

The main product outlined in the first chapter of this report is an update of the MAP maps of observed and repaired rainfall records over the Republic of South Africa. It also offers Mean Monthly Precipitation (MMP) maps crafted from observed and infilled data. In contrast to previous studies, we chose to infill the missing data at the scale of interest (Annual or monthly, rather than infilling at daily and accumulating) to exploit the far stronger spatial correlation between gauges in order to provide more robust estimates of the missing values. In addition, the precision of these MAP and MMP maps is indicated by sets of estimates of low and high quantiles of the distributions of the data. The supporting material in later chapters describes the new mathematics and algorithms developed specifically for this purpose and subsidiary applications of these tools forms the remainder of the report.

The following figure is possibly the most important product of this project. It shows the high variability of rainfall in mountainous areas, in contrast to the gradual increase of rainfall from West to East of the country. Also evident is the sparseness of gauges in some areas, particularly the dry ones.

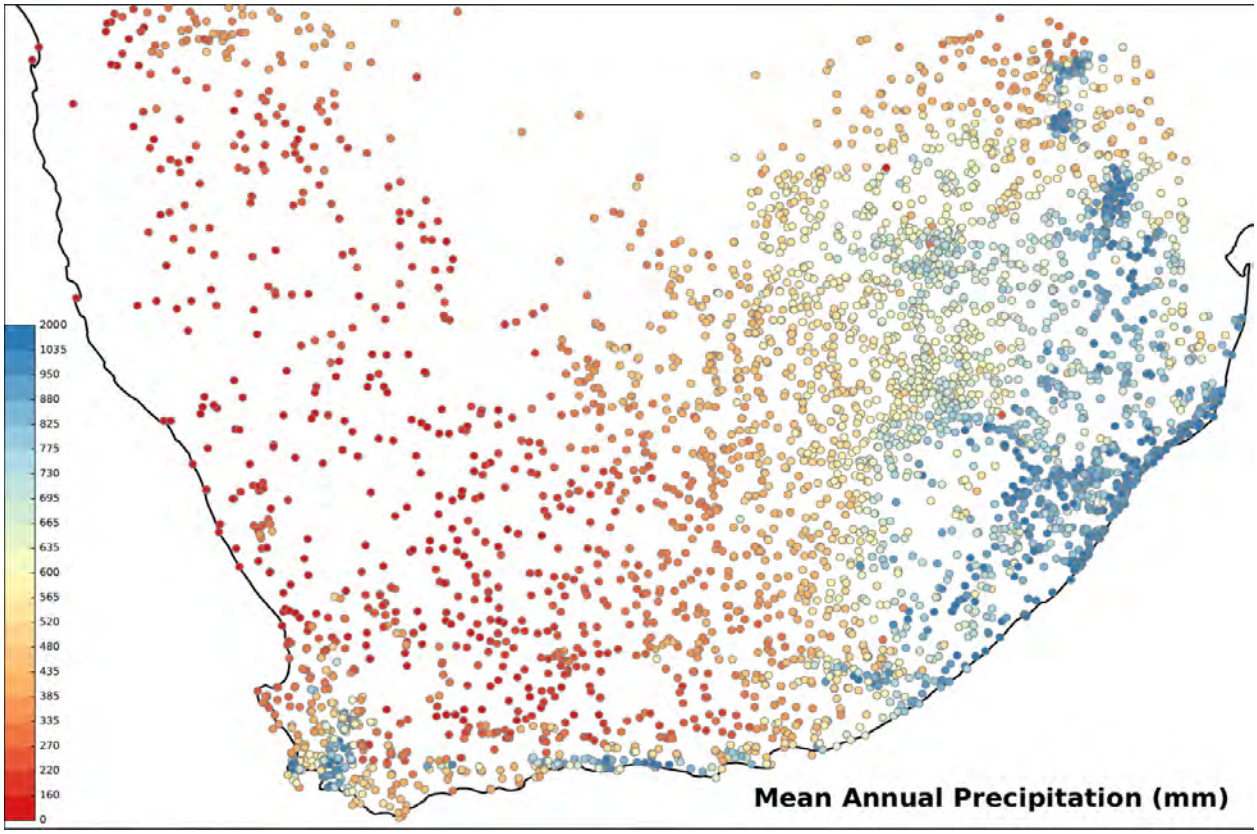


Figure ES.1 The repaired estimates of Mean Annual Precipitation (mm) at all infilled gauges in the daily rainfall data base held by CSAG

The principal endeavour undertaken in this report is to repair the daily, monthly and annual rainfall estimates over South Africa and not necessarily in that order. Not only do we give 'best' estimates of the repaired values, we also offer the distributions, and hence confidence limits, of the estimates. Once this has been done, we can spatially interpolate the point information into the intervening space. From these interpolations we can make maps.

The map of much interest in the practical implementation of our results is that of MAP over the quaternary catchments in the country. Using the same colour palette as the dotty plot above in Figure ES.1, Figure ES.2 is the interpolated map of MAP values of the set of repaired gauges in the region. Its colour legend is specifically designed in intervals to bracket the mean precipitation estimates of each quaternary catchment, so no inference is required in reading off the upper and lower limits of the MAP on the catchments.

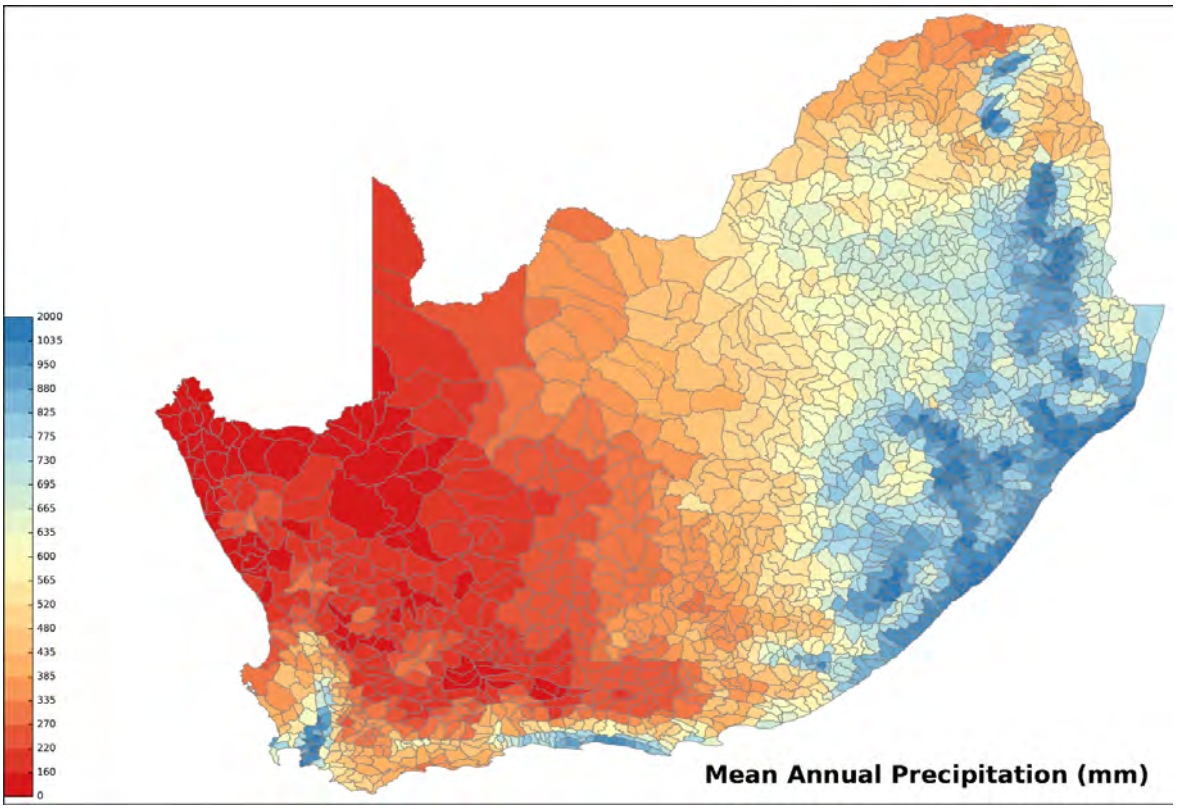


Figure ES.2. The MAP map of quaternaries, created by interpolating infilled gauges of Figure ES.1, using Gaussian copulas and quantile-quantile transforms of the data, as described in Chapters 3, 4 and 6.

At some daily raingauge locations, measurements are available for more than a hundred years. Unfortunately these measurements have been sporadically obtained at a few specific sites. For most applications, such as hydrological modelling or flood forecasting, the full spatial distribution of precipitation (including the values at unobserved locations) are also needed. Remote sensing provides useful additional information – radars and satellite observations offer an important insight to the spatial properties of precipitation, but they are limited to recent times and are biased relative to gauge measured rainfall. Nevertheless, in order to use these methods for hydrological applications in a meaningful way, they have to be calibrated to ground truth.

The different spatial (and temporal) scale of the ground measurements and remote sensing makes a direct comparison of point scale and remotely sensed spatial estimates of precipitation meaningless. Thus our first step is to repair the gauge records using appropriate methods, which have been recently devised by Bardossy and Pegram (2014). Once that is done we use interpolation methods devised by Bardossy and Pegram (2013) which offer precipitation estimates at relatively large spatial scales and provide reasonable uncertainty estimates for these. Precipitation interpolation presents a serious challenge if one wishes to derive a large number of realizations.

The specific problems that presented themselves include the following:

- due to the large number of realizations automatic procedures are needed
- precipitation has a mixed discrete continuous distribution with the number of zeros in the record depending on the aggregation scale – thus specific handling of the zeros is necessary
- precipitation is influenced subtly by external factors such as topography and possibly distance from the coast – this dependence should be determined and incorporated in the methodology where meaningful
- Knowledge of the uncertainty of the interpolation is vital for ground truthing of remote sensing products or for hydrological modelling – specifically the uncertainty at large spatial scales is of great importance
- Interpolation quality has to be assessed – not only in the sense of squared errors but also for possible bias.

Early interpolation methods were based on empirical assumptions and did not provide any error estimates. However, Geostatistical methods have been used for precipitation interpolation since the 1980s. They provide reasonable unbiased estimates of precipitation together with the estimation of variance over different scales. However the standard Kriging based methods have substantial deficiencies for our task, namely:

- they are usually based on fixed spatial stationarity assumptions which do not hold for precipitation, for example due to orographic effects
- they are based on covariances (variograms) which are based on a Gaussian (symmetrical) assumption of the raw data, which does not accord with observations, due to the physical mechanisms of rainfall generation which result in an asymmetrical correlation structure at daily scales
- they treat the mixed discrete continuous distribution of precipitation (dry areas represent a discrete distribution) as a single continuous distribution by default
- their uncertainty estimates are simply dependent on the spatially stationary covariance functions and the network geometry, and not on the measured precipitation values.

In the framework of the research reported by Bardossy and Pegram (2013), we developed a methodology for precipitation interpolation, simulation and uncertainty assessment using point observations and indirect measurements which we call Dynamic Copula Regression. In the first phase we concentrated on the problems related to point-observation-based interpolation and simulation. We focussed on an adequate representation of uncertainty at different spatial and temporal scales. The goal of this phase was to obtain a realistic ground truth for indirect measurements which could then be used in the second phase of the research as a basis for the combination of direct and indirect measurements. Quantification of uncertainty is crucial, because an optimal combination of the different measurements is only possible if their uncertainty is reasonably modelled.

In our recently accomplished research (ibid.), it was found that it was important:

- to find a good method to incorporate external (topographical or geographical) information into rainfall interpolation where necessary

- to perform rainfall interpolation using a copula based approach
- to improve the informative contribution of dry stations in the interpolation of precipitation
- to define and test new measures of precipitation interpolation quality
- to test copula based interpolation uncertainty estimates
- to develop simulation methodologies to estimate interpolation uncertainty on larger spatial scales
- to test the above methodologies on regions with different topography using rainfall depth accumulations in time intervals ranging from daily to annual
- to develop a copula based interpolation method for the possible incorporation of climatological information
- to separate interpolation error into a random error and a temporally correlated bias
- to apply the above developed models to improve indirect precipitation estimates using radar and satellite information precipitation accumulations where available.

To achieve meaningful infilling of missing data we draw from the paper by Bardossy and Pegram (2014), whose contribution became available only half-way through the project. Infilling missing data might be an unpleasant and tedious task, but is vital for analysis and water resources management, so should not be done in a lackadaisical manner. The important thing about the infilled values is that they need to be as good as possible, because poor infilling is likely to lead to poor decisions. Traditionally, a range of methods has been routinely employed, e.g. Nearest Neighbour substitution through to Kriging, but few methods attach a quality estimate to the infilled values. In the 2014 paper (ibid.), a new copula based method we choose to call Dynamic Copula Regression, which was developed for infilling missing daily, monthly and annual rain gauge data. The new method was compared with six other commonly used methods, in a semi-arid environment with a range of rain-rates and interstation distances, in the Southern Cape region of South Africa. For daily data it is clear that the copula- based methods are superior to the others in terms of point estimation and have the added benefit of providing an estimate of the precision of the interpolation, not provided by the others.

It was found that the addition of atmospheric Circulation Patterns (CPs) (Pegram et al., 2013 and Pegram and Bardossy, 2013) designed to add information for infilling, has a relatively small positive effect on the quality of the estimation. The main reason for this is that a small number of wet days does not allow a good estimation of the conditional distribution of precipitation amounts; note that the average probability of a dry day in the test region is 86%. A minor improvement of the estimate of the probability of a dry day was however observed if CPs were used as conditions. In other regions, with a higher number of wet days, a CP-based method might lead to further improvements. Using copula-based methods, the estimated probability of a dry day corresponded well with the observed frequency of dry days over the test data. The monthly data yield the same conclusion, with the qualification that the Expectation Maximisation (EM) algorithm (Pegram, 1989 & 1997) performs as well as the copula method. This is because of the low count of dry months in this region and because monthly data are much less skew than daily and do not violate the Gaussian assumption of the data used in regression. Its relative disadvantage compared with the copula based method is that it does not routinely offer as valuable a precision estimate.

As a result, in this project, the methodologies developed by Bardossy and Pegram (2013, 2014) are used to perform (i) the infilling of missing point data and (ii) the spatial interpolation over intervening areas. The interpolation of rainfall data in individual time intervals in Bardossy and Pegram (2013), ranging from a day to a year, was an inter-comparison of the skill of Ordinary Kriging, external Drift Kriging, Gaussian copulas and unsymmetrical v-copulas, with a variety of treatments of altitude as an exogenous variable. For time aggregations such as daily and pentads, zero precipitation occurrences were treated as censored variables. We note that in that German study the monthly and annual data reported no dry periods (unlike our Southern African data) so that for each selected time step the marginal distributions of precipitation amounts were modelled using non-parametric density estimators, while the dependence structures were estimated using a maximum likelihood methodology. Several measures of bias and error structure were used to assess the efficacy of the methods over a range of comparative split-sampling studies. The result was that the copula-based methods were far more informative than the other traditional ones, so that is the procedure we have adopted in this work for two main reasons. The first is that the dry periods were included in the infilling (pointwise) or interpolation (space-wise) through a Gaussian transformation accommodating the zeros meaningfully into the Dynamic Copula Regression. The second reason was that the Gaussian copula method yielded meaningful error structures that are not spatially uniform, but depend on the local precipitation intensity.

Having set the technical stage, this report condenses the advances that have been made in the 3 years of this project's life. The report deals with the following issues in 11 chapters:

- The first chapter contains the final product – the maps of MAP and MMP together with maps of their variability and a comparison with previous estimates
- In the second chapter we reported on work using Circulation Patterns associated with the rainfall regimes, based on the output of WRC project K5/1964 (Pegram et al., 2013), but with new regions based on SAWS criteria.
- Chapter 3 describes the cross-validation of Gauge data, with a view to selecting the best infilling procedure, by comparing several standard methods of infilling against the new Dynamic Copula Regression method, which is also outlined in this chapter.
- In Chapter 4 we explain how to visualise the worth of the infilled values, through pictorial explanation of the methods, complementing the previous chapter.
- In Chapter 5 we determine the value of the data and examine the results of the infilling.
- Chapter 6 is a summary of the methodology developed for spatial interpolation between the repaired gauges for the production of smooth maps and for estimating rainfall amounts over catchments.
- In Chapter 7, we describe a straightforward spatial Interpolation using the Fast Fourier Transform to produce radar-like random fields, as a possible alternative to the copula-based methods, developed in Chapter 6.
- Chapter 8 describes an attempt to develop a method to downscale TRMM rainfall data to block averaged daily read gauge rainfall data using regression
- Chapter 9 introduces a novel idea which is suggested for performing a valid quantile-quantile transform of TRMM to Block averaged gauge rainfall where there are records and then interpolating the methodology to ungauged locations. Although

not exploited in this study, the methodology needs to be recorded and used elsewhere.

- Chapter 10 describes the data and algorithms and indicates how these have been archived for access by practitioners.
- Chapter 11 contains a summary and conclusion to the report.

In memoriam

In spite of all the good new things we are providing in this project, the sad truth is that we are working with a dying resource. The SAWS raingauge network currently has approximately 1200 live gauges – about the same number as were recorded daily by weather stations and volunteers in the 1930s – down from a peak of near 3000 in the 1970s. Unless these are augmented soon, to allow us to infill the records ***backwards in time*** using the methods devised herein [to be done again once we have 10 years or so of new data], there will be insufficient gauged readings of actual rainfall at ground level against which to calibrate and adjust radar and satellite estimates of rainfall.

To illustrate the point, the next two images display the change in gauge density in Southern Africa over the last 160 years, indicating the total number of active daily recording rain gauges in each year.

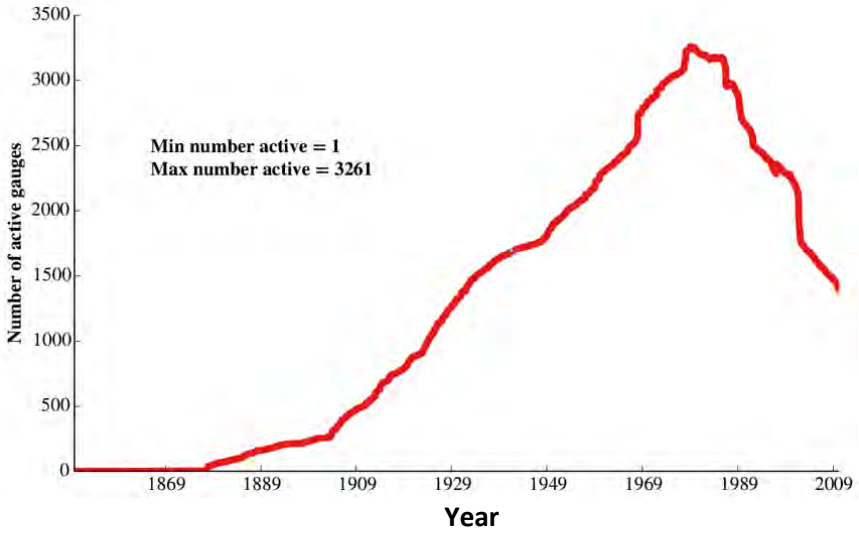


Figure ES.3. The total of active daily recording rain gauges in each year, from the daily rainfall data-base maintained by CSAG at the University of Cape Town.

We point out that the sudden drop off by approximately 500 gauges in the year 2000 is an artefact of the data-set. Prior to 2000 the database contained gauges from organizations other than SAWS, however only SAWS gauges are in the database after 2000. Nevertheless, there is still a very real and concerning drop-off in the number of gauges.

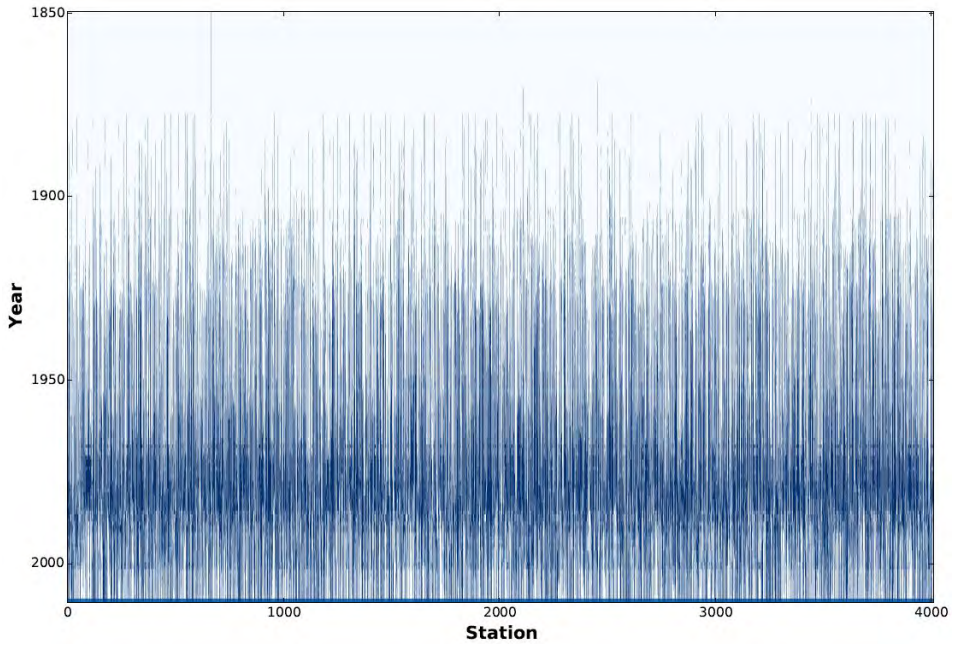


Figure ES.4. An alternative summary of gauge density by year. The blue vertical lines only appear when a year has data; the white space means there were no data recorded at that time.

In our opinion, the WRC and other concerned research-orientated bodies with strong voices should prevail upon the large institutions currently managing gauge networks (SAWS, ARC, DWS) to immediately deploy a complementary set of rain gauges, so that the average density over the wetter part of the country is increased to an inter-gauge spacing of closer than 25 km. If the gauges are sparser than 35 km then there is a very small spatial correlation link between them at the daily scale and they might as well be treated as independent, lonely gauges.

Acknowledgements

Our thanks go to the reference group for their valuable encouragement, suggestions and support over the 3 years:

Mr W Nomquphu	:	WRC (Chairperson)
Dr E de Coning	:	SAWS
Ms T Chetty	:	UKZN
Dr C Lennard	:	UCT
Dr E Kapangaziwiri	:	CSIR
Dr J Mwenge Kahinda	:	CSIR
Mr A Bailey	:	Royal Haskoning DHV
Prof J Ndiritu	:	WITS

In addition, we thank SAWS for the permission to access their daily rainfall records held by the CSAG group of the University of Cape Town, whom we also thank for their assistance. We would like to thank those who attended our two Workshops for their constructive criticism and suggestions supporting their expressed needs in practice.

Capacity Building

The organisations which have benefitted directly from this project have been those whose people have supported us as reference group members and represent SAWS, DWS, the CSIR, Universities and Civil Engineering consultants. The two workshops we held were very well attended by the personnel of the above organisations. It is hoped that this product will add value to the hydrologically focussed design and construction industry and those who need up-to-date information on rainfall data in its many forms.

There were two postgraduates who benefited from involvement in this project. The first was Mr Erik Becker of SAWS who completed his MSc in 2015. The second was Mr Malose Ngoepe who has worked as a Masters student on this project for the last three years and is soon to complete his degree.

This report records the development of a project which started as a venture into unknown mathematical territory, to meaningfully repair and spatially interpolate rainfall data. It is the sum of the work of a team whom I value highly and hereby express my gratitude to my companions and supporters in this endeavour – Dr Scott Sinclair, Prof András Bárdossy and Mr Malose Ngoepe. We have developed original, effective, and complex algorithms to deal properly with uncertainty of estimates of data repair and interpolation. We offer a product which is not only definitive, but vital, and which will grow in time as others add value and ideas to the methodology. Nevertheless, in spite of the novelty and efficacy of our work, we caution that there is no such thing as a fixed Mean Annual Precipitation (MAP) although, to our surprise, it has been remarkably stable over the last century.

Geoff Pegram

29 February 2016

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ACRONYMS

BAGD	Block Averaged Gauge Data
ccc	cross correlation coefficient
cdf	cumulative distribution function
cfdf	cumulative frequency distribution function
CP	Circulation pattern
CSAG	Climate System Analysis Group – University of Cape Town
CSIR	Council for Scientific and Industrial Research
DCR	Dynamic Copula Regression
DWA	Department of Water Affairs
DWS	Department of Water and Sanitation
ECDF	Empirical Cumulative Distribution Function
EDK	External Drift Kriging
EM	Expectation Maximisation
FFT	Fast Fourier Transform
GHCN	Global Historical Climate Network
GPM	Global Precipitation Measurement – NASA
GTS	Global Telecommunication System
hPa	hectopascal
KDE	Kernel Density Estimation
KZN	KwaZulu-Natal
MAP	Mean Annual Precipitation
MLR	Multiple Linear Regression
MMP	Mean Monthly Precipitation
MSc	Master of Science
N(0,1)	Normal probability – zero mean, unit standard deviation
NASA	National Aeronautics and Space Administration
NN	Nearest Neighbour
OK	Ordinary Kriging
QQ	Quantile- Quantile
R^2	Coefficient of determination
RCM	Regional Circulation Model
RMSE	Root mean square error
RSA	Republic of South Africa
SADC	Southern African Developing Community
SAST	South Africa Standard Time
SAWS	South African Weather Service
SRTM	Shuttle Radar Topography Mission
SW	South West
TRMM	Tropical Rainfall Measuring Mission
WMO	World Meteorological Organisation
WP	Wet proportion
WRC	Water Research Commission

Chapter 1. Annual and Monthly precipitation maps

This first important chapter contains a sample of the maps that might be useful to the practitioner, to complement those offered in the Executive Summary. The first set in Figure 1.1 are the Annual Precipitation quantiles of the infilled annual rainfall totals in the CSAG data-base, starting at 10% at the top, increasing to the median in the centre and 90% in the bottom panel, rendered as dotty plots. The legend ranges from 0 to 2000, equal count of gauges in each interval.

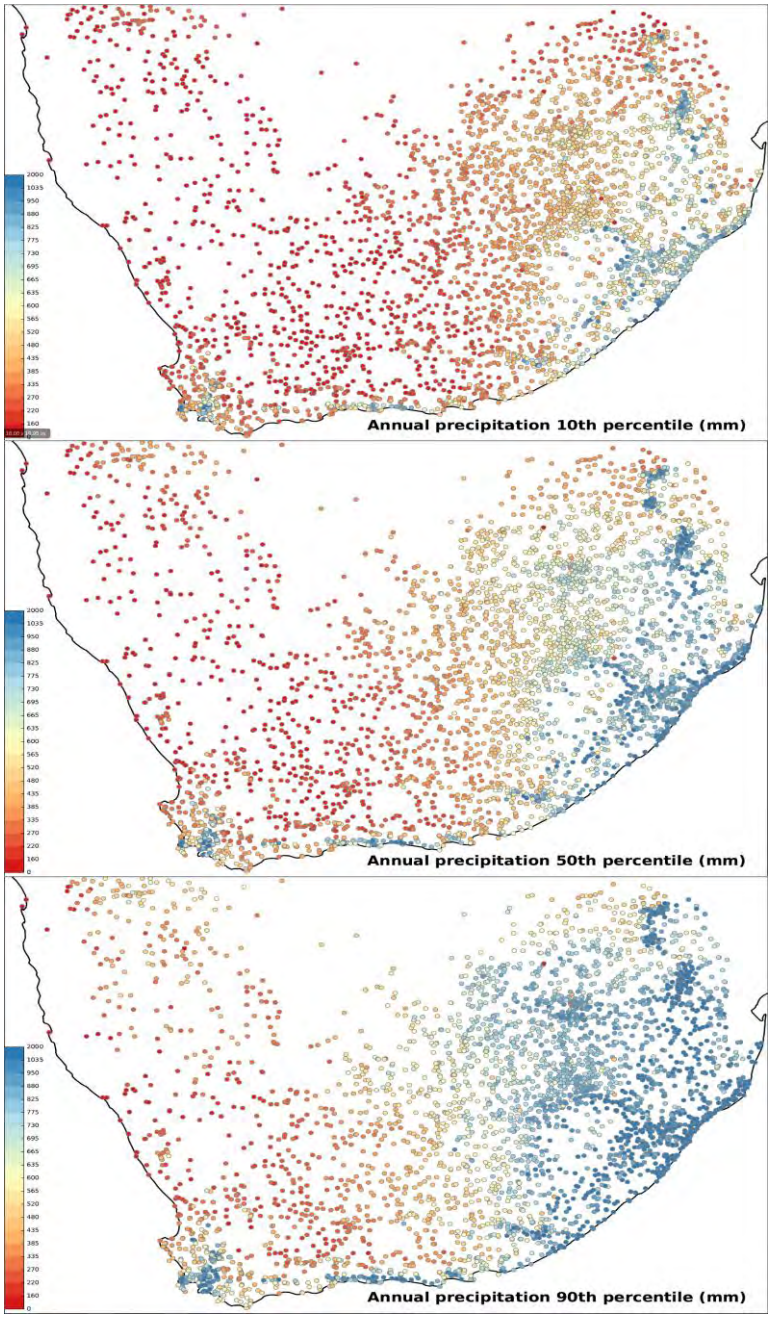


Figure 1.1. Annual rainfall totals at all infilled daily rain gauge sites in the CSAG data-base: from top to bottom 10th, 50th & 90th percentiles, whose individual values in the trio of maps are to be exceeded 9 years out of 10, 5 years out of 10 and 1 year out of 10 respectively.

We next turn to a selection of the monthly plots, at 3-month intervals, for January, April, July and October in the next set of four figures. These show repaired datasets, with quantiles of each infilled gauge at 10%, 50% and 90% rainfall at each site, like the annual values which were displayed in Figure 1.1. Note change of values for the colour scale, which ranges from 0 to 375 mm with equal numbers of gauges in each interval.

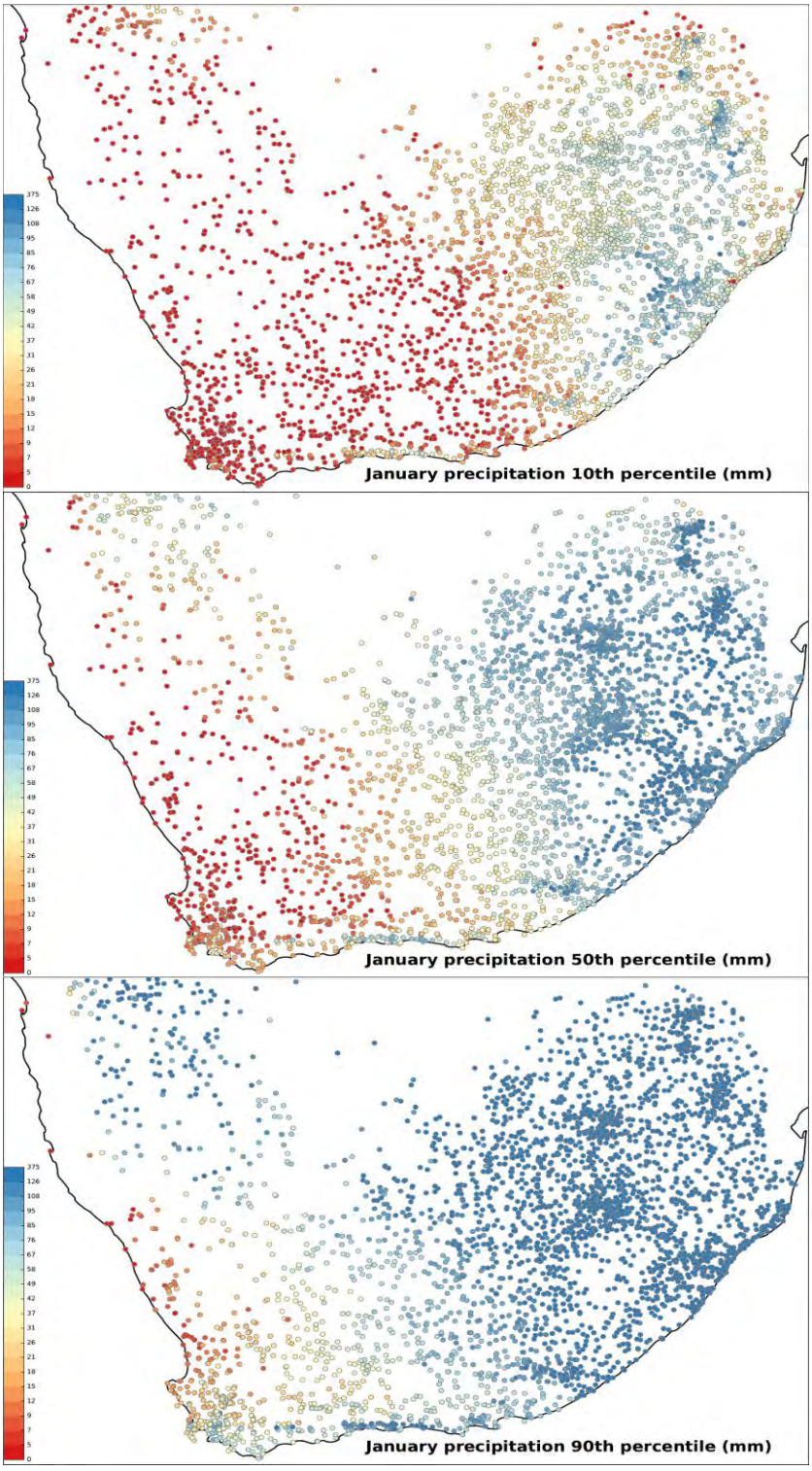


Figure 1.2. January: from top to bottom 10th, 50th & 90th percentiles

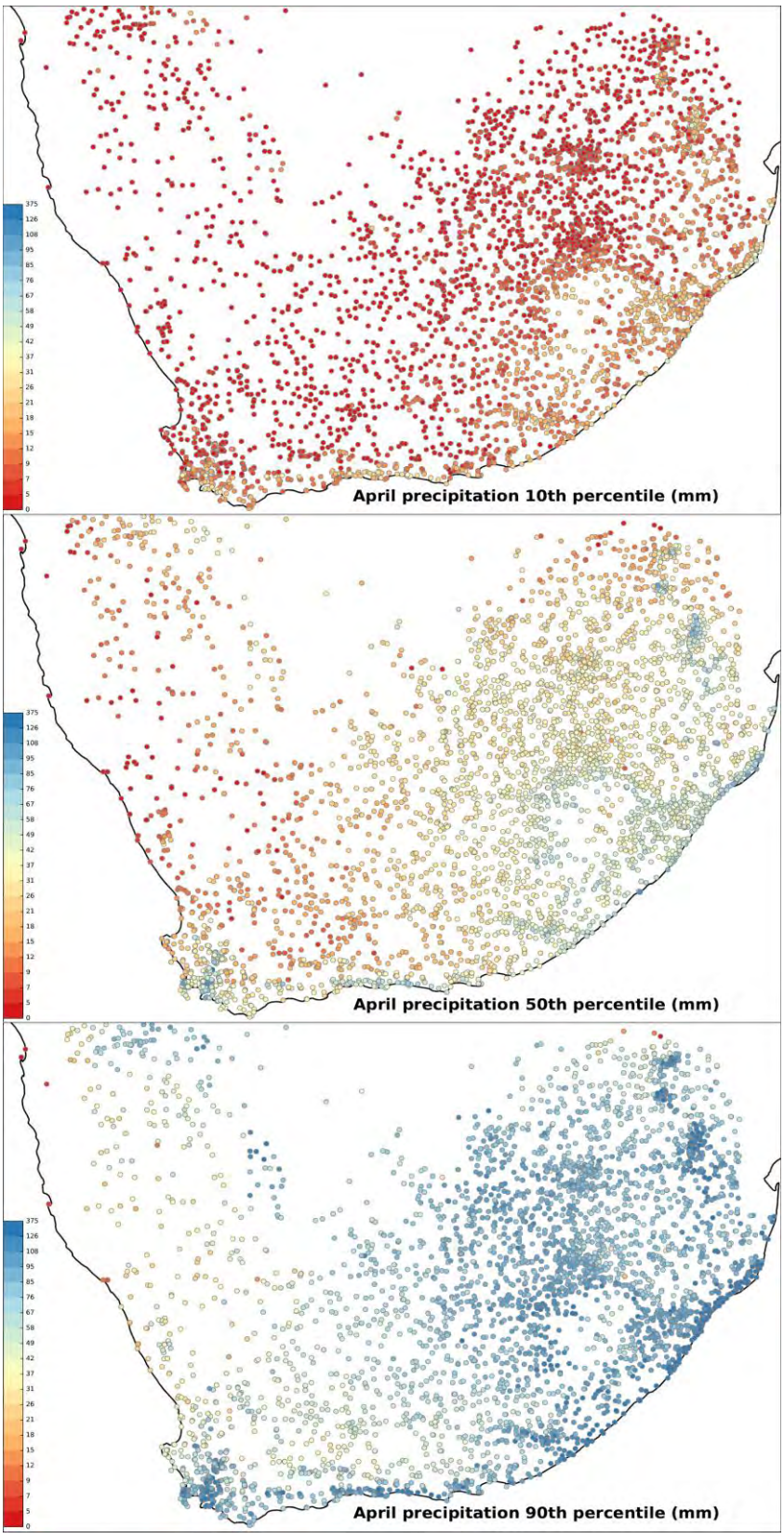


Figure 1.3. April: from top to bottom 10th, 50th & 90th percentiles

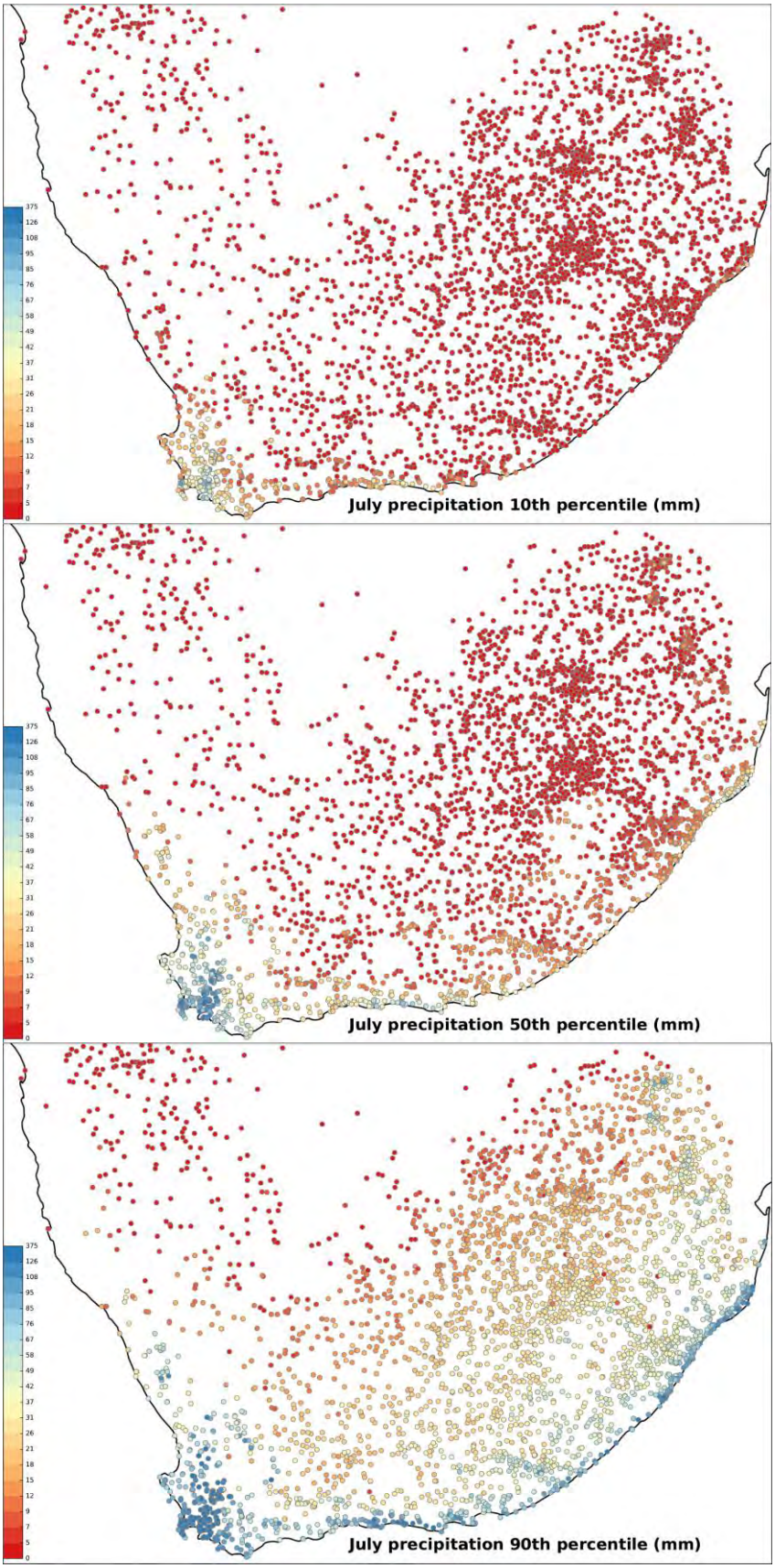


Figure 1.4. July: from top to bottom 10th, 50th & 90th percentiles

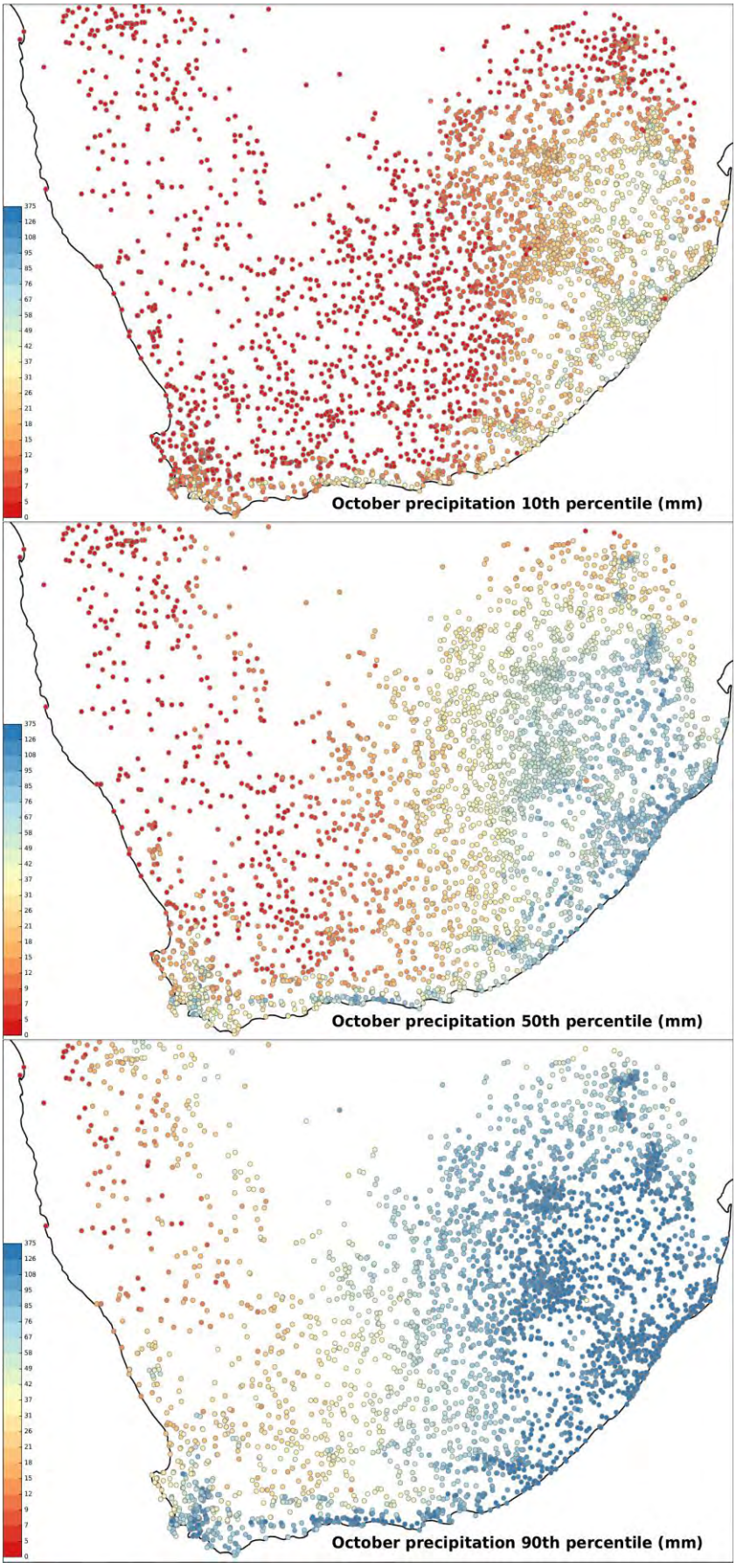


Figure 1.5. October: from top to bottom 10th, 50th & 90th percentiles

Figure 1.6 displays an ensemble of all monthly rainfall areal means on the quaternary catchments, obtained by interpolation of the infilled gauges displayed in quantiles in Figures 1.2 to 1.5.

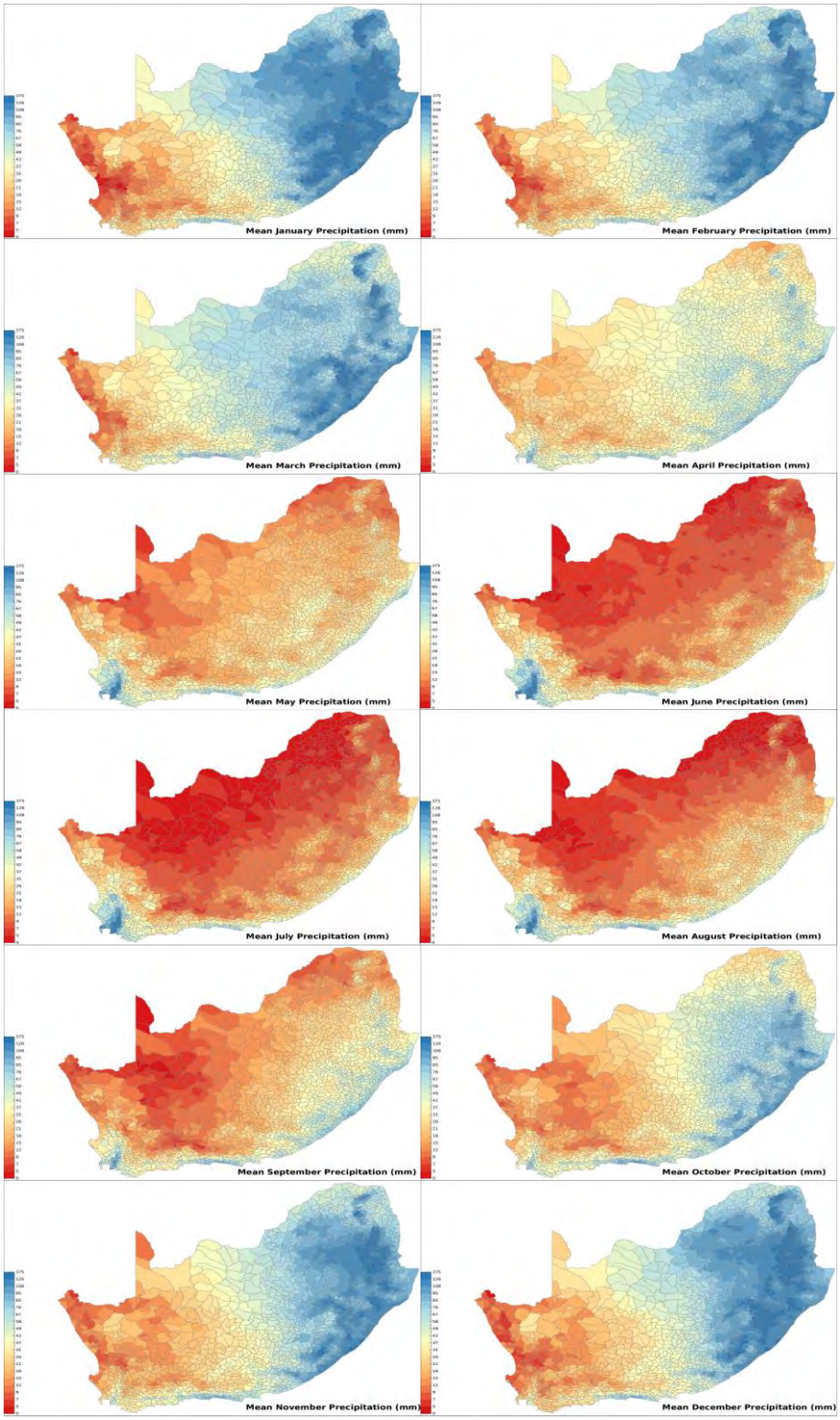


Figure 1.6. All monthly Means. The order of the maps is January and February in the first row to November and December in the last row; the range of the legends is 0-375 mm, with equal numbers of gauges in each interval.

We conclude this introductory Chapter with an intercomparison and a sequence of MAP maps. In Figure 1.7, the top image is our new MAP map for the region, interpolated from the infilled data shown in Figure ES.1 in the same colour scale; the range in the legend is 0-2000 mm, where the intervals have similar counts of gauges. The two lower maps of MAP in Figure 1.7, the one by Dent et al. (1987) the lower right by Lynch (2004) included for comparison, have a slightly different colour scale from ours. We draw attention to a zone of differences in the North East of RSA; our map shows a drier zone than the other two, confirmed precisely by the dotty plot in Figure ES.1. There is another change apparent in the Eastern part of Swaziland; it is clear from Figure ES.1 that there is high rainfall above the escarpment, not meaningfully captured by Lynch (2004) who used geographically weighted regression.

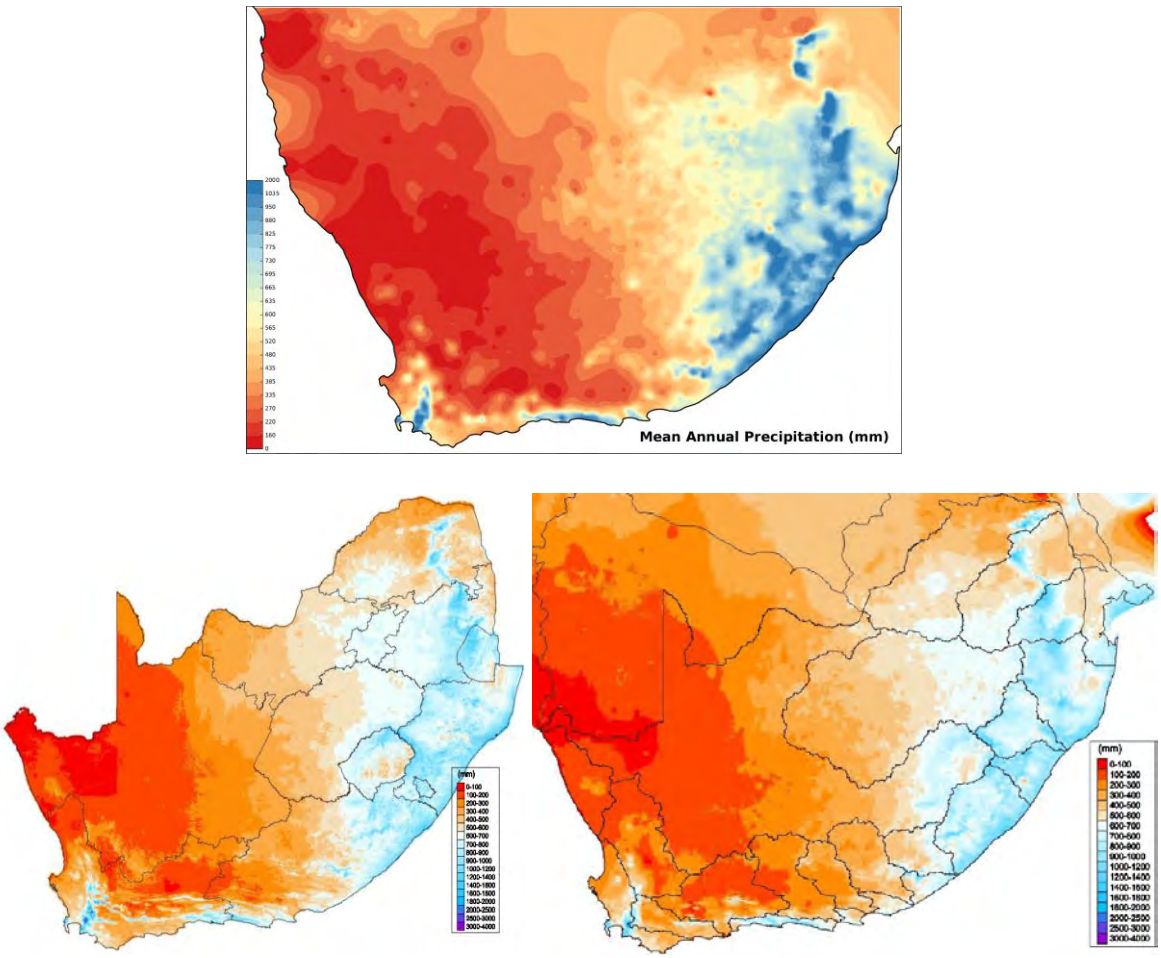


Figure 1.7. The top image in this Figure is our new MAP map for the region. This is in a different colour scale of the other two maps of MAP, the lower left by Dent et al. (1987) the lower right by Lynch (2004) for comparison.

The last set of images in this Introduction, making up Figure 1.8, is a sequence of eight separate 20-year periods of MAP starting in 1850 and finishing in 2010. These are assembled from the infilled data-set and show some interesting characteristics. The first thing that is immediately obvious, is the gradual spread of installed gauges from the Cape towards the Northeast of Southern Africa over about 80 years. The second thing is that the colours of the dots in the plots do not change by very much between the periods. This

observation leads to three interim conclusions: (i) the MAP has been remarkably stable over the 20th century (ii) the notorious and troublesome interdecadal variations are smoothed out by averaging over 20 year periods and (iii) we have more than 20 years of observations over Namibia, interrupted some time before 1990. This is relevant, as in our new Dynamic Copula Regression methodology we do not try and infill missing data at targets where the overlap with control gauges is less than 20 intervals, not necessarily contiguous. These Northwest observations anchor that area.

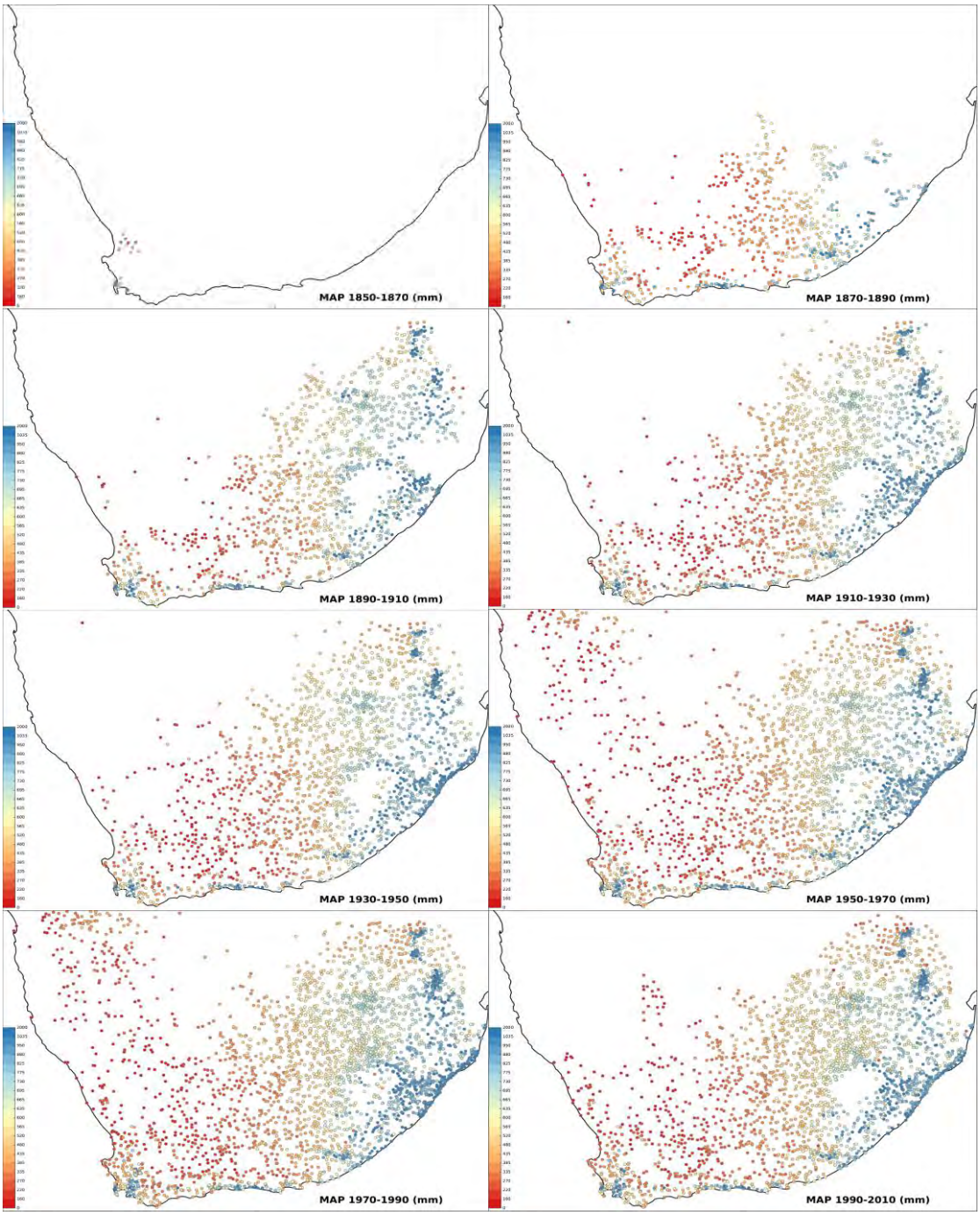


Figure 1.8. A sequence of eight separate 20-year periods of annually averaged recorded gauge rainfall starting in 1850 and finishing in 2010. All available data were used however short, hence the odd anomaly.

We will offer a selection of more maps in a later section of the report, but first we outline and justify some of the theory behind the infilling procedures that we created and adopted to perform the necessary tasks.

Chapter 2. Choosing Circulation Patterns by region conditioned on daily rainfall

Based on the work done in WRC project WRC_1964_CC_RAIN (Pegram et al., 2013) we started by choosing a set of climatically homogeneous regions and decided to sort the days by Circulation Patterns (CPs). To do this effectively, we had to choose the climate regions.

After the Workshop held in June 2013, where we derived valuable suggestions from the participants, we abandoned our original intention of compartmentalising the country into drainage regions for the purpose of data repair. Instead we adopted the 24 Climate Regions defined by Kruger (2004), slightly modified by concatenating some of the very small regions (mostly in dry areas) with larger ones. The Kruger map is as follows in Figure 2.1. A map of active rainfall stations during the years 1970-1980 follows in Figure 2.2.

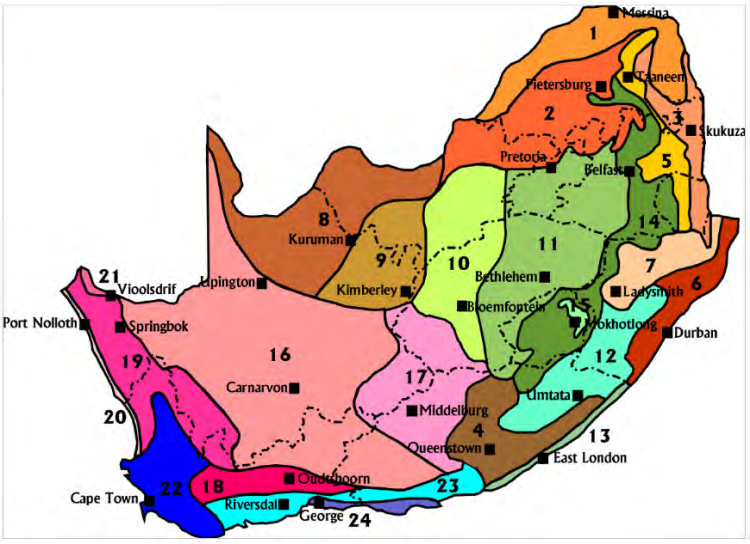


Figure 2.1. Climate regions after Kruger (2004) labelled:- 1-9: Savanah; 10-15: Grassland; 15-20: Karoo; 21: Desert; 22-23: Fynbos; 24: Forest.

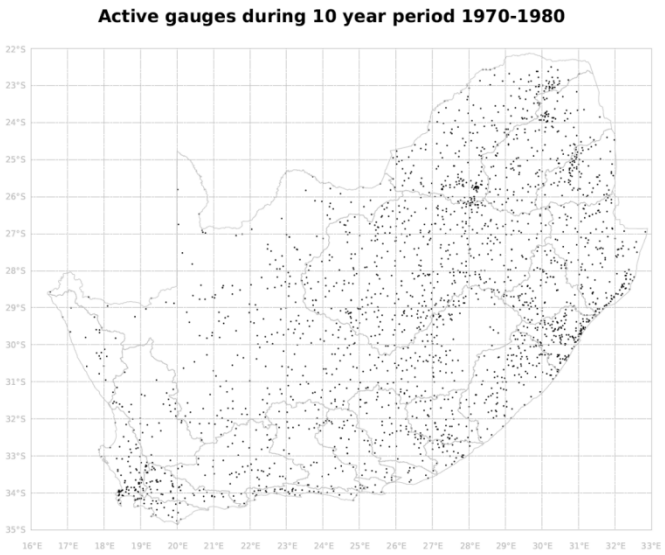


Figure 2.2. A map of South Africa showing the active SAWS gauges in the period 1970 to 1980; percentage of missing data is not taken into account in the figure.

To illustrate the CP-based methodology, we selected region 6 in Figure 2.1 then selected some gauges within the region, in different configurations, to classify the 700 hPa Circulation Pattern anomalies [CPs], which we found are associated with different types of rainfall over this region (Pegram et al., 2013).
Figure 2.3 displays two random choices of samples from the region to determine the sensitivity of the CP choices to different gauge patterns.

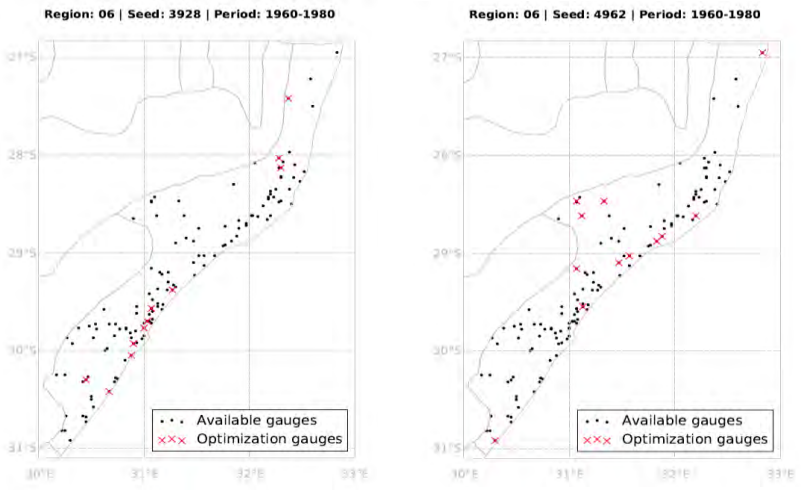


Figure 2.3. Region 6 of the map in Figure 2.1 with available gauges during the period 1960 to 1980 [black dots] and two subsets of randomly selected gauges [red dots] for conditioning 700 hPa fields into two sets of Circulation Pattern anomalies

We find we get similar CPs, with enough similarity of shape to pick a set, [note that the labelling within each set is random, so we match by correlation of shape, not label] and obtain Figure 2.4.

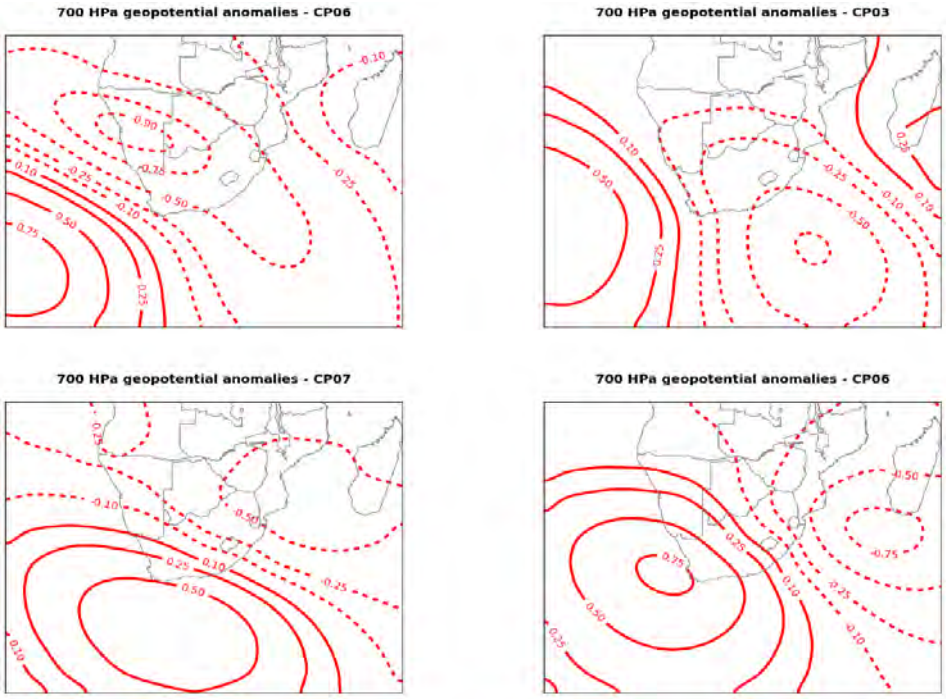


Figure 2.4. 2 pairs of CPs: top two similar to each other and the bottom two likewise, selected from the CPs chosen on the 2 sets of randomly selected gauges in Figure 2.3

This similarity allowed us to settle on one set of CPs per region [and season] because we have a robust method. Although interesting, this early work was superseded by the methodology summarised in chapter 3. We will work on the premise that correlations of rainfall between successive periods [day, month and year] are so low that the infilling can be usefully done at one interval at a time, thus making the use of CPs redundant in this study.

Chapter 3. Selecting a good infilling procedure using cross-validation

We needed to find the best available method of infilling missing data. In order to select the most viable among a range of methods, a good test is the method of cross-validation. This means that we take an intact data-set and pretend that a proportion of the data are missing. The infilled values are compared with the 'missing' data using several criteria and we then select the most effective method. For this comparative work, we chose a set of gauges in the Southern Cape whose intact records span 32 years. In the monthly data we found that an average of about 5 % of the months were dry, whereas in the daily data, the average proportion of dry days was approximately 80 %. The way that the cross-validation was done was that in 32 years, we left out 20 % at a time for each gauge in turn, modelling in 2 seasons. Thus each gauge has each estimated value individually compared against every one of the observed data. For the daily records about 140 000 comparative calculations were done; for the monthly, about 4 600. A map of the region follows in Figure 3.1 with the sites of the gauges indicated. This is the most southerly part of Africa, with Gansbaai in the South-West corner of the figure, with the Riviersonderend and Langeberg Mountains bordering the green plain where the gauges are sited.

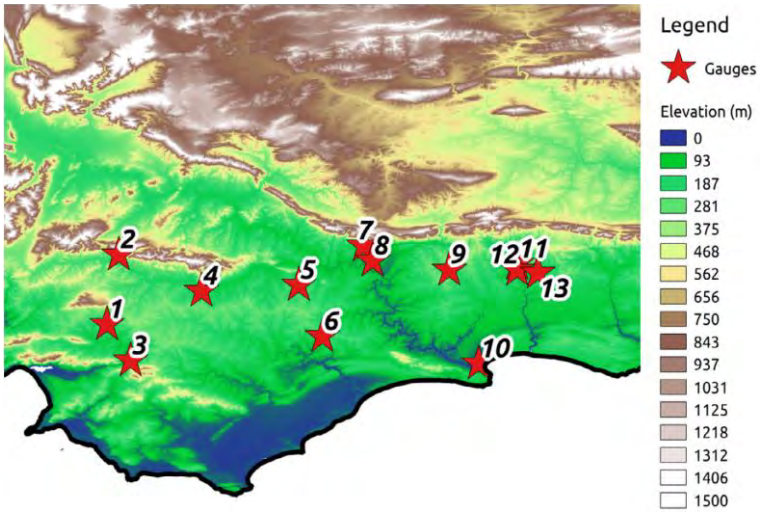


Figure 3.1. Locations of the 13 selected rain gauges used for this study.

The monthly distribution of mean rainfall at Station 1 follows and it is noted that the heaviest rainfall is in the second half of the year, justifying the choice of 2 seasons to explore.

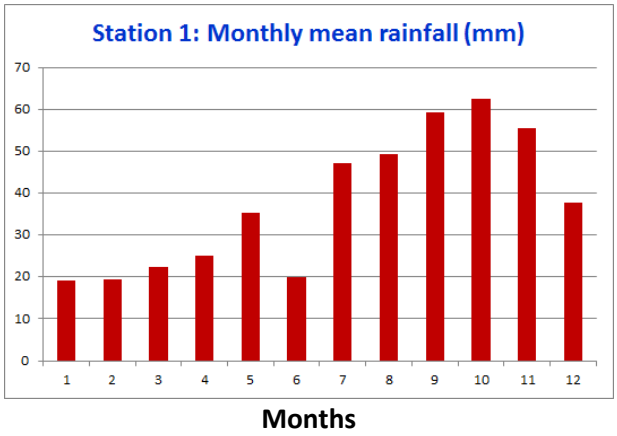


Figure 3.2. Histogram of mean monthly rainfall (mm) at Station 1 from 32 years of data

The statistics of the gauges follow in Tables 3.1 and 3.2. We note the wide variability between the different sites' characteristics, due to the topography, particularly station 7 nestled near the Langeberg Mountains, useful for testing the relative efficacy of the methods.

Table 3.1. Statistics of daily precipitation for the 13 selected stations in season 1.

Station	Mean (mm)	Stdev	Skewness	P[dry]
1	0.77	3.78	11.47	0.905
2	1.45	5.89	12.75	0.841
3	0.72	3.62	9.71	0.918
4	0.99	4.18	11.75	0.858
5	0.74	3.91	11.82	0.908
6	0.92	3.98	9.23	0.893
7	2.85	8.72	5.24	0.804
8	1.30	4.80	6.79	0.829
9	1.33	5.01	6.98	0.882
10	0.77	4.08	11.21	0.925
11	1.22	4.66	6.25	0.880
12	1.08	4.41	8.11	0.874
13	1.04	4.26	8.53	0.853

Table 3.2. Statistics of daily precipitation for the 13 selected stations in season 2.

Station	Mean (mm)	Stdev	Skewness	P[dry]
1	1.70	5.39	7.87	0.829
2	1.65	5.32	6.14	0.822
3	1.73	5.41	5.05	0.846
4	1.39	4.61	7.40	0.820
5	1.09	3.90	6.36	0.863
6	1.33	4.53	7.30	0.845
7	2.23	7.11	6.52	0.808
8	1.54	5.05	6.94	0.805
9	1.49	5.31	8.53	0.854
10	0.89	3.84	7.91	0.896
11	1.33	5.17	10.43	0.863
12	1.15	4.33	8.13	0.853
13	1.17	4.35	8.47	0.832

The daily and monthly distributions of Station 1 in Season 2 are shown in Figure 3.3. The daily values are plotted in the top left of the figure on a logarithmic axis and the monthly values below left on a linear axis for ease of visualisation. The Cumulative frequency distributions of daily and monthly values are plotted on the right of the figure, where it will be seen that the proportion of dry days is 83%, while the proportion of dry months is 8% for this station. These dry days and months will need special treatment when we come to infill neighbouring gauges' missing values.

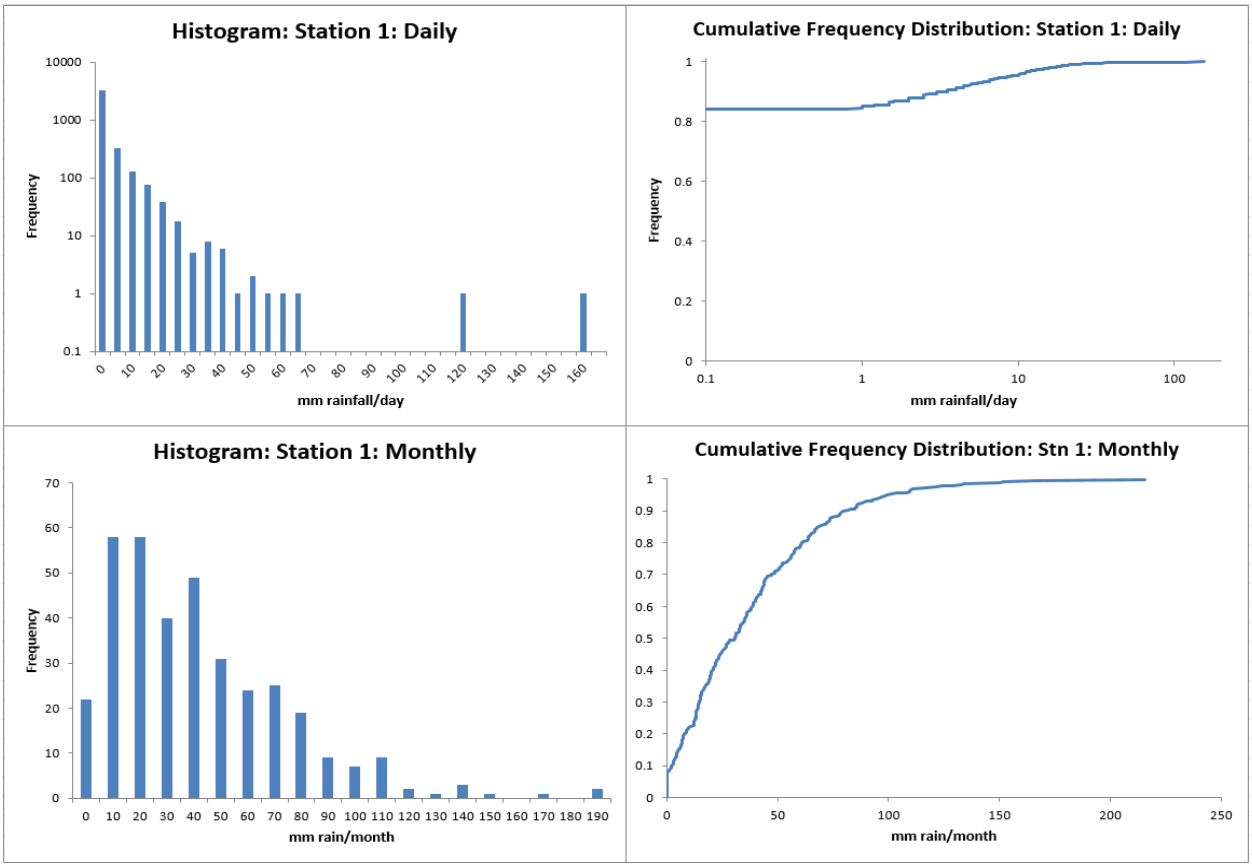


Figure 3.3. Daily and monthly distributions of Season 2 of Station 1; Histograms on the left and cumulative frequency distributions [cfds] on the right; daily above, monthly below.

3.1 The Infilling methods used in the intercomparison

We compared the following suite of methods of infilling missing data, many of which are well known. In the following subsections, we will start with a brief mathematical description of the well-known methods, followed by a summary of Dynamic Copula Regression in Section 3.2. The seven methods in the intercomparison are:

- Nearest Neighbour – based on proximity or correlation
- Nearest Neighbour – scaled by long term means
- Inverse distance weighting of some controls
- Linear regression using control with highest correlation coefficient
- Multiple linear regression
- Dynamic Copula Regression (Bardossy and Pegram, 2014) [no monthly CPs]
- The PATCHR algorithm (Pegram, 1997) [applied to monthly data only]

For all the methods, we describe how to infill the targets with information in the controls. We distinguish between the missing values $y_m(t)$ of the target and the observations $y_o(t)$ in the target's time series. For the epochs with missing values of the targets, where $\boldsymbol{x}(t) = (x_1(t), \ldots, x_o(t))$ is the vector of observations at the controls at each epoch t in the record, we define the vectors:

$\{(\boldsymbol{x}(t); y_m(t)),\ t = 1, 2, \ldots, T_1\}$

For the times with observed target gauge values $y_o(t)$ we have:

$\{(\boldsymbol{x}(t); y_o(t)),\ t = T_1+1, \ldots, T\}$.

3.1.1 Nearest neighbour

The simplest method to infill missing data is to take the nearest neighbouring station and to use its observation to substitute for the missing datum. The nearest neighbour can be selected geometrically or by taking the station with the highest correlation to the target location. The value can be:

- transferred directly without any change
- obtained by linear scaling (with the long term means)
- obtained by quantile-quantile [QQ] transformation

3.1.2 Linear regression

For simple linear regression one uses as control the series i from a set for which $x_i(t)$ and $y(t)$ have the highest correlation; then:

$$y(t) = a + bx_i(t) \qquad (3.1)$$

3.1.3 Inverse Distance Weighting

Another simple way of finding an interpolated value y at a given point in epoch t using Inverse Distance Weighting is an interpolating function:

$$y(t) = \Sigma_i\, w_i(t).x_i(t) \qquad (3.2)$$

where $w_i(t) = s_i / \Sigma_j s_j(t)$

where s_i is the [unsigned] distance between the target y and the controls x_i and the denominator $\Sigma_j s_j(t)$ ensures that the weights $s_i/\Sigma_j s_j(t)$ sum to unity. In a more general form s_i can be raised to some power, but we stayed with the linear model.

3.1.4 Multiple Linear regression

Instead of a single variable one can also use a few neighbouring observations for the estimation of the missing values. The form of the estimator in this case is:

$$y(t) = a_0 + \Sigma_i\, a_i(t).x_i(t) \qquad (3.3)$$

where the coefficients $\{a\}$ are derived using the linear set of Normal equations in the usual way. We note that it is important that the ring of surrounding data should not be more than 2 deep, else some of the $a_i(t)$ become strongly negative (Wesson and Pegram, 2004).

3.2 The infilling of daily gauge records using the new method of Dynamic Copula Regression

This section outlines the methods we devised for infilling missing data in rainfall records and can be applied to daily, monthly and annual data, without modification, except that the rather weak CP-conditioned infilling of daily data is replaced by monthly or seasonal conditioning in the monthly case. The annual infilling procedure is season-free, but we are sensitive to the need for accounting for any trend inherent in the data. We used this method exclusively for infilling work described in the remainder of the report. The simplified version of the mathematics is given in Section 3.5.1, drawn from Bardossy and Pegram (2013).

3.2.1 The Mathematics to deal with dry control stations when others are wet

This section's focus is on how to deal with the thorny problem of infilling missing data at a target when there are differing proportions of dry stations in the changing availability of controls over a range of chosen days or months; the task is also affected by the distances of such dry controls from the target.

For estimation using Dynamic Copula Regression, we assume that each one of a localised collection of gauges [12 to 20, say] has its own specific distribution of rainfall. Their membership of the set of controls is assessed at the chosen common observation time, based on which of the controls has observations and which have not. Once the controls are assembled [using a method of grouping them developed and described later in this section] the distribution function of the rain-depths at the ith control $X_i(t)$ is determined as $F_i(x)$. The distribution function of the target gauge $Y(t)$ we label $F_Y(y)$. In order to relate these observations we assume that the joint (zero truncated) copula of (X_1, …, X_k, Y) is $C(u_1, \ldots, u_k, v)$. The distribution of a missing observation at the target $Y_m(t_0)$ is given by the conditional distribution function:

$$P(Y_m(t_0) < y | X_1, \ldots, X_k) = C\{F_Y[y], F_1[x_1(t_0)], \ldots, F_k[x_k(t_0)]\} \,/\, C\{F_1[x_1(t_0)], \ldots, F_k[x_k(t_0)]\} \quad (3.4)$$

In order to use this formula one has to obtain the marginal distributions $F_i(x)$ and $F_Y(y)$ over a period when the target and controls have contemporaneous available data. Due to the discrete/continuous behaviour of precipitation (dry [discrete] and wet [continuous]) one writes all the stations' distributions as:

$$\begin{aligned} F_i(x) &= p_i && \text{if } x = 0 \\ &= p_i + (1 - p_i)G_i(x) && \text{if } x > 0 \end{aligned} \quad (3.5)$$

Here p_i is the probability of a dry record at location i based on the full available record. The continuous distribution $G_i(x)$ of the positive values has to be estimated. This is done either by fitting a parametric distribution [or a nonparametric distribution using appropriate kernel functions] using the observations available for the common time period.

Note that the assessment of the individual distributions using the observation of the common time period is important. In this way, possible nonstationary behaviour is compensated for, if all the observations are subject to the same direction of change during the period. The Gaussian copula (the copula of the multivariate Gaussian distribution) is used to describe the dependence.

The Gaussian copula is characterized by its correlation matrix Γ. The correlations are estimated either by using the maximum likelihood method [which can accommodate values below a threshold] or by substituting for the zero values a predefined normal value. Thus the positive Gaussianised copula values become:

$$U_i(t) = \Phi^{-1} F_i (X_i(t)) \quad \text{if} \quad X_i(t) > 0 \tag{3.6}$$

The zero precipitation amounts are given the probability p_i. For these locations, as a first choice, one might assign the condition:

$$U_i(t) < \Phi^{-1}(p_i) \quad \text{if} \quad X_i(t) = 0 \tag{3.7}$$

Using these conditions for all the zero values, the correlation matrix of the observation can be calculated using a maximum likelihood approach. However the latter inequality in (3.7) used to obtain U_i(t) would make the calculations relatively complicated and will likely cause difficulties with the correlation matrix (not being positive semidefinite).

A simplification which ensures a stable correlation matrix can be obtained by taking a fixed value instead of an interval [this is the copula value, an alternative to the Gaussian y_0 calculated in the text supporting Figure 4.1]:

$$U_i(\mathrm{t}) = \Phi^{-1} (p_i/2) \quad \text{if} \quad X_i(t) = 0 \tag{3.8}$$

The advantage of the copula based method is that it delivers the conditional distribution (conditioned on the available measurements at time t) of precipitation at the selected target. Based on this formulation, one can calculate point estimators (expected value, median and quantiles), and/or one can simulate a possible realization.

3.2.2 An extension to the method of Zero correction to cope with ranges of spatial dryness

Although the procedure described in Section 3.5.1 was adopted in Bárdossy and Pegram (2013) and Pegram and Bárdossy (2014), the approach has the problem that it assigns the same value $U_i(t)$ given in equation (3.8) to all dry days of site *i.* More awkwardly, it is complicated by those choices (i) on days where all but one of the stations was dry ranging through to (ii) those where only a single station was dry. The expected value of the constrained normal distribution will differ for these cases. Thus a more sensible approximation, given in Bardossy and Pegram (2016) and used in this study, is obtained by introducing $W(t)$, the wetted proportion of the n control stations on each day t:

$$W(t) = \#[X_i(t) > 0]/n \qquad (3.9)$$

then we suggest that a reasonable approximation of the Gaussian value corresponding to a zero precipitation value is obtained by taking the relative wetness into account replacing p_i by $p_i[1 + W(t)]$ to give:

$$U_i(t) = \Phi^{-1}(p_i[1 + W(t)]/2) \qquad \text{for all } X_i(t) = 0 \quad \text{on the day} \qquad (3.10)$$

The same transformation is applied to the target Y leading to the Gaussianised variable V. The argument of Φ^{-1} can range from $p_i/2$ through to p_i depending on the wetted proportion of the control stations.

The transformation of the zero precipitation values to quantiles with a given probability, has the advantage that the correlation matrix obtained is positive semidefinite; consequently the calculation of the conditional distribution is simple and is performed using multivariate densities.

Assuming a Normal copula for the dependence means that ($\mathbf{U}$,V) is Normally distributed where $\mathbf{U}(t) = (U_1(t), \ldots, U_n(t))$ is the vector of Gaussianised observation data (controls) and V is the target. Thus the conditional distributions are also Normal. For the estimation of $V^*(t)$ this means that its distribution is $N[\mu(t),\sigma(t)]$, with:

$$\mu(t) = \Gamma^T_{U,V}\Gamma^{-1}{}_U\mathbf{U}(t) \qquad (3.11a)$$

and

$$\sigma^2(t) = \sigma^2_V - \Gamma^T_{U,V}\Gamma^{-1}\Gamma_{U,V} \qquad (3.11b)$$

using the well-known conditional distribution of a subset of multinormal variables, given the other variables in the set. In (3.11a), $\Gamma^T_{U,V} = (\mathrm{Cov}(V; U_1), \ldots, \mathrm{Cov}(V; U_n))$ is the vector of covariances between the vectors of observed points and that of the point to be infilled and Γ_U is the covariance matrix of the observations/controls.

Two things that are important to notice are the following:

1. The expected value of the estimated target value $V^*(t)$ is $\mu(t)$ which is dependent on $\mathbf{U}(t)$, the Gaussianised transforms of the observations of the control stations at time t, so it can vary in time
2. The variance of V^*(t) is $\sigma^2(t)$ which, in contrast to $\mu(t)$, is independent of the time t, as it is constant over time, if the configuration of all the controls remains constant. This is a disadvantage inherited from the selection of the Normal copula. However in a practical situation, to overcome this problem, up to three of the available control stations are chosen in each time interval which have the highest cross correlations with the target, so Γ is modified where necessary to accommodate the missing data in the controls. The smallest number of controls used in any regression is allowed to be one. Therefore the precision of the infilled value will change with circumstances.

Finally, the distribution of the estimator for the unknown precipitation amount Y^*(t) is obtained by back transforming V^*(t) to rainfall space:

$$G(y) = P(Y^*(t) < y) = \Phi[\Phi^{-1}\{F_Y(y)) - \mu(t)\}/\sigma(t)] \quad \text{for} \quad y > 0$$

and

$$= P(Y^*(t) = 0) = \Phi[\Phi^{-1}\{p_{Y0} - \mu(t)\}/\sigma(t)] \quad \text{otherwise} \qquad (3.12)$$

These equations were programmed into the suite of infilling code for this project.

To help the reader grasp the flow of this technical section, we proffer the following passage in which the sequence of calculations follows in pseudo-code. Days are chosen by CP, months by season, years are not conditioned:

1. Pick a target station together with up to 20 controls surrounding it and assemble the subset for a given group in periods of days by CPs, months by season or years
2. Gaussianise the target, and then all the controls in turn, because the treatment of the zeros depends on the number of (i) missing values and (ii) the number of dry stations in each period, as indicated above in (3.10)
3. Do a preliminary cross correlation coefficient [ccc] calculation between all the target and control stations for the group using their full Gaussianised records, so that the cccs of the controls can be ranked relative to the target from highest to lowest ccc in each interval
4. Once target and controls are all Gaussianised, assemble them in a matrix with the target in the first column and the controls, ranked by their cccs with the target from highest to lowest, in the remaining columns; pick the three controls most highly correlated with the target
5. Call the Infilling routine to read the matrix, infill the target's missing data and output the repaired Gaussianised target's data, with mean and stdv [$\mu(t)$ and $\sigma(t)$] associated with each infilled element
6. Pass the repaired target vector, with $\mu(t)$ and $\sigma(t)$ values where appropriate [i.e. these are only associated with infilled values], to be reverse Gaussianised using the QQ transform to recover the estimated rainfall in mm, as well as the median and upper and lower quartiles of the infilled estimates [these will help to define the distribution of the infilled data]
7. Pick a new group for the given target and repeat steps 1 to 6 until all CPs are done, then go to 8
8. Pick a new target and go to 1, until all gauges in a region are infilled, when pick a new region and go to 1 until all regions are done then go to 9
9. Finish.

3.3 Infilling Model intercomparisons

Returning to the results published in Bardossy and Pegram (2014), we start the model intercomparison by cross-validating monthly values. The following three figures summarise the results, where the best Bias and Root Mean Square Error [RMSE] should be as small as possible and the Correlation [Corr] as large as possible.

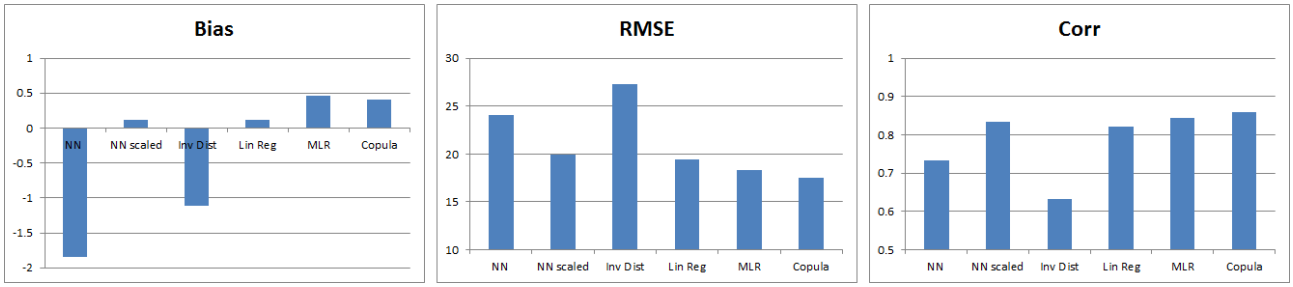


Figure 3.4. Histograms of averages of (i) Bias, (ii) Root Mean Square Error and (iii) Correlations between the estimated monthly values for all 13 stations, using the 6 methods without EM: Nearest Neighbour (NN), Nearest Neighbour scaled (NN scaled), Inverse distance weighting (Inv Dist), Linear regression (Lin Reg), Multiple linear regression (MLR) and Gaussian Copula.

When it comes to Bias, Nearest Neighbour and Linear regression show the smallest bias. Nevertheless, the copula-based method is the best of the remaining four methods and is marginally the best of all methods with respect to RMSE and Correlation. It has the added advantage of offering a meaningful error estimate with each estimate.

Because monthly records have been routinely infilled in South Africa using the Expectation Maximisation (EM) algorithm, as coded as PATCHR by Pegram (1987) for the Department of Water Affairs (DWA) and published in Pegram (1987), an independent comparison was done using this method. The following two figures show the results of the infilling; 20% of each of 8 of the gauges was infilled in 5 steps per gauge and the pooled result of the infilling appears in Figure 3.5, where for ease of visualisation, results for only 6 out of the 8 stations are displayed.

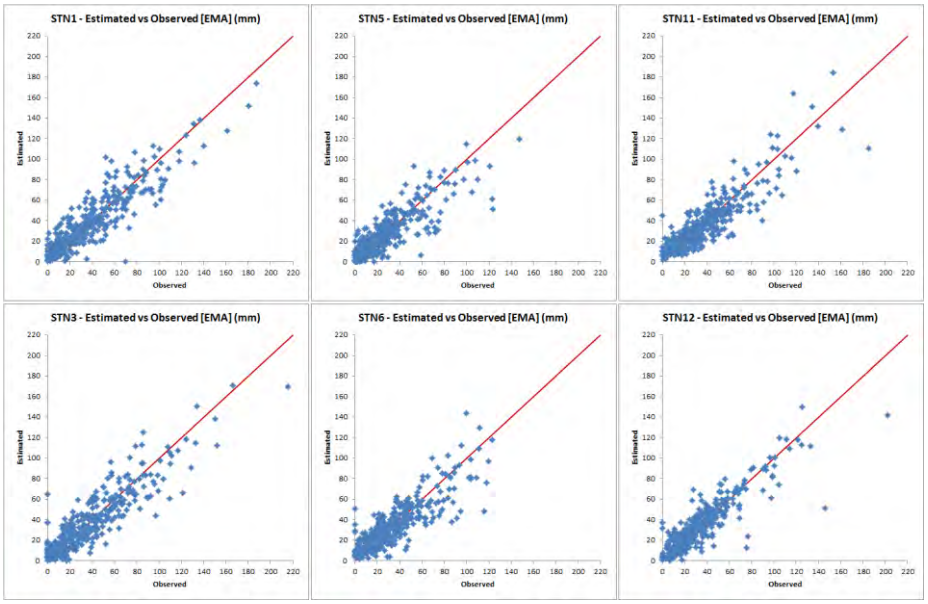


Figure 3.5. Estimates of censored monthly values compared with the observed censored values for 6 stations [Observed on horizontal and Estimated on vertical axes].

In Figure 3.6, we give a comparison of the EM algorithm (labelled EMA in the figure captions) and the Copula-based infilling of all 8 gauges. The four panels compare (i) the monthly means [all good and negligible difference] (ii) the Mean bias [EM has the edge, but the values are small] (iii) the RMSE [again EM has a slight edge] and (iv) the cross correlations [the copula method is better on balance]. The reason that the EM algorithm works so well compared to Dynamic Copula Regression on monthly data is that the wet months typically have low skewness, so they would benefit little by being Gaussianised. This does not apply to daily rainfall, so EM was not used for daily comparisons in this study.

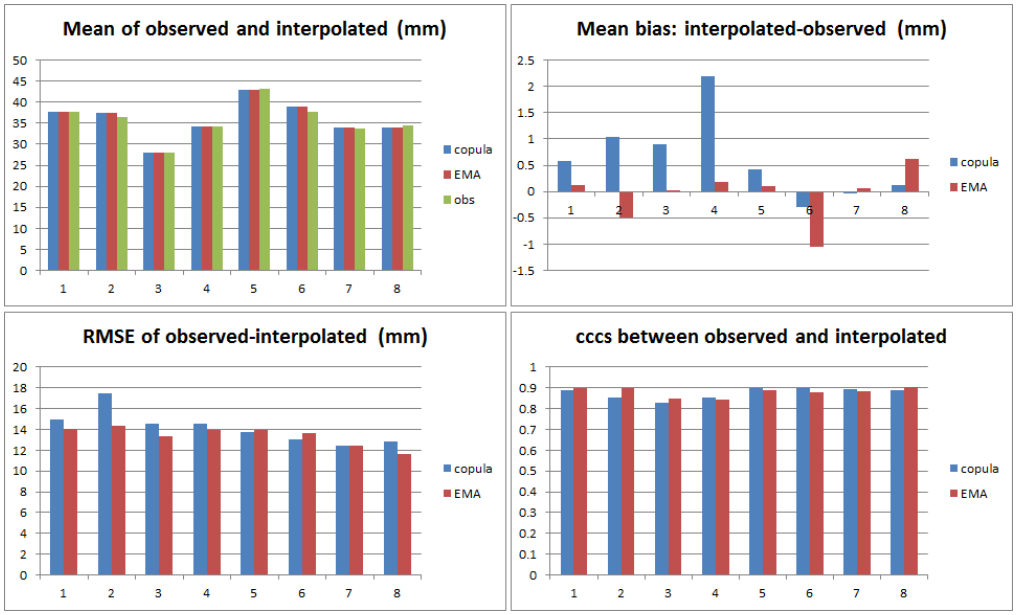


Figure 3.6. Comparison between the EM algorithm [EMA] and the Copula-based infilling of all 8 gauges. The four panels compare (i) the monthly means (ii) the bias (iii) the RMSE and (iv) the cross correlations. For best results, we would choose the method with the lowest score in the first three comparisons and the highest score in the fourth.

When it comes to daily values, we use the estimation methods as were used for monthly totals, but we drop the NN scaled method and in its place add the copula method conditioned on the CPs identified in Pegram et al. (2013). The results follow in Figure 3.7.

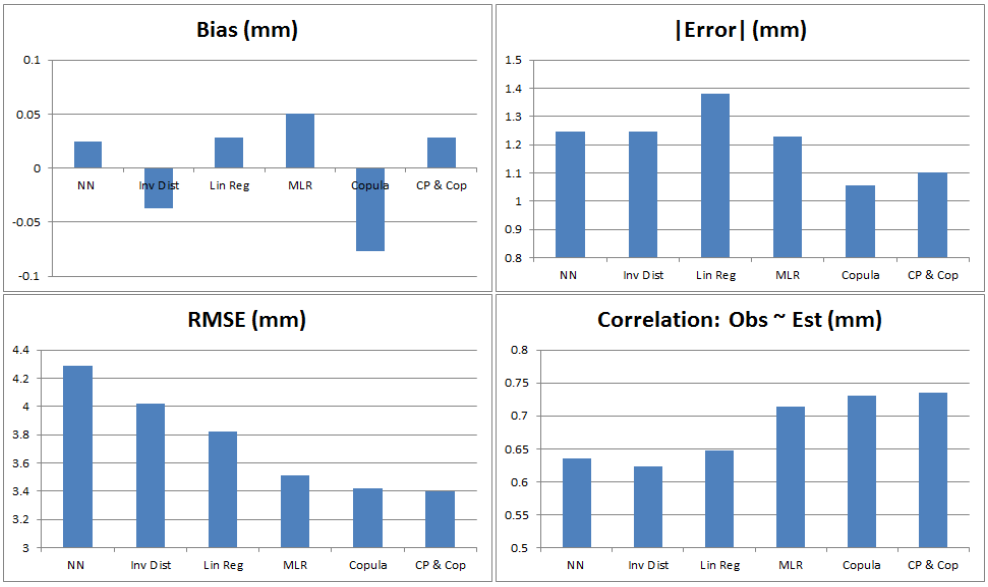


Figure 3.7. Comparison between six methods of infilling daily values at all 13 gauges. The four panels compare (i) the bias (ii) the average absolute difference (iii) the RMSE and (iv) the cross correlations of the estimated censored and observed values.

For best results, for daily data, we would choose the method with the lowest score in the first three comparisons and the highest score in the fourth. On balance, the CP-based copula method is better than the unconditioned copula, in all but the absolute error mean, and is better than all other methods in nearly all criteria; it marginally loses to Nearest Neighbour in the Bias comparison.

In summary, we choose the copula-based method as the one to use for data repair, not only because of its success in the above tests but because it can give a valuable additional product: the error structure of the interpolant, tailored to the local spatial distribution of the controls, as well as their rainfall data values. Although the CP dependent copula is a fraction better than the plain copula method for repairing the daily data, we decided that the extra effort is not worthwhile for infilling over the whole Southern African region, because deriving the CPs over a large region is a challenging and laborious task.

Chapter 4. Determining the value of the infilled values

4.1. How good is Dynamic Copula Regression for infilling?

To determine the value of the infilled values obtained by competing methods, we begin by using cross-validation and evaluation of simple error distributions to explain the methodology. In this section, we offer an informal set of images to describe the process. The full method is described in Bárdossy and Pegram (2014). The first step is to Gaussianise the data, using a quantile-quantile [QQ] transform by rank. In this explanation, which is an expansion of the methodology described in Section 3.2, the zero values get special treatment and are set to

$$y_0 = -\phi[\Phi^{-1}\{p_0\}]/p_0 \qquad (4.1)$$

which is the mean of the area below b, where the probability of dryness is $p_0 = \Phi\{b\}$ and where $b = p_i[1 + W(t)]/2$ is the N(0,1) variate corresponding to zero in the rainfall set; $\phi\{b\}$ and $\Phi\{b\}$ are respectively the density and cdf of the standard N(0,1) distribution at b. Figure 4.1 shows the procedure.

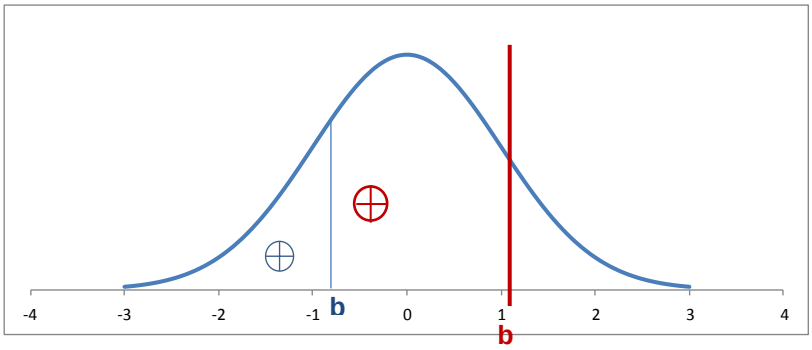


Figure 4.1. Computing the location of the average of the tail of an N(0,1) distribution below point [b]. y_0 at the blue cross and b are for a humid environment, whereas y_0 at the red cross is the centroid of a semi-arid environment with a dry probability P[0] = 0.8, typical of South African daily data, whose cutoff is at the red line marked by a red b.

We now present a series of cartoon images to explain the cross-validation procedure.

Choose a set of 1 [fictitious] rainfall data as shown in Figure 4.2 and label some of them as targets, leaving the others as controls. In the procedure, we will remove the red crosses we have labelled 'Target', then infill them and determine how good the infilled values are.

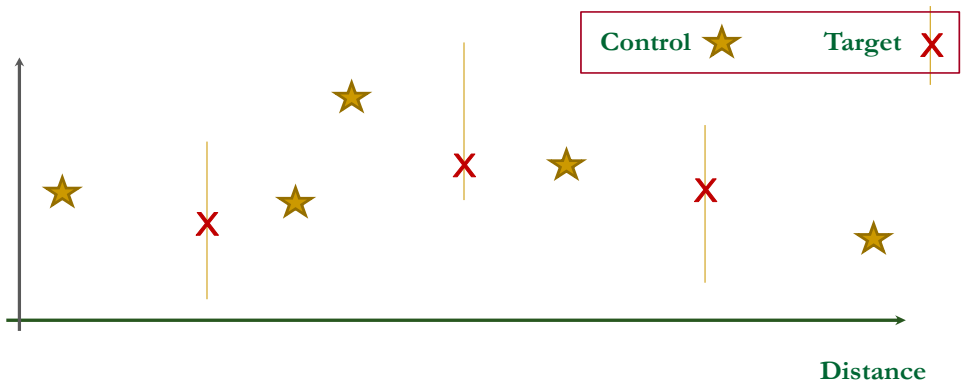


Figure 4.2. A 1-dimensional example describing cross-validation to find its value in copula space.

Fit a curve through the controls using an interpolator [e.g. Copula-based Kriging] to give the blue line shown in Figure 4.3. Estimate the values at the Targets [green Xs] and calculate standard deviation envelopes [the pair of amber curves]. These Kriged envelopes assume a Gaussian distribution of the error at each target – the standard deviations [indicated by the small gold horizontal lines] will vary with location. There will be no estimation error at the controls.

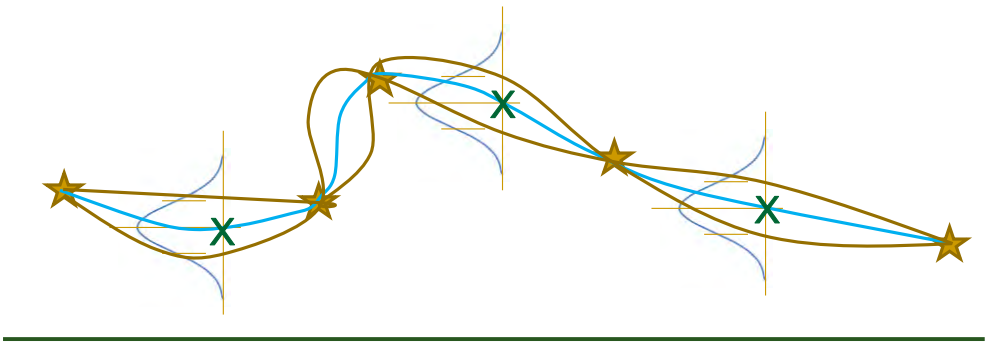


Figure 4.3. Use Kriging of the Gaussianised controls to obtain the best interpolator and the standard deviation of the estimates along the curve. Note the standard deviations at the sites of the targets, centred on the expected values of the estimates, depicted by horizontal gold lines. [An apology: the continuous gold curves showing the Kriging standard deviations were drawn free-hand in Powerpoint and the loops between controls 2 and 3 are too wide horizontally; they are technically infeasible because they are multivalued vertically in some parts of the segment. We claim artistic license!]

Next, replace the hidden target values superimposed over the target error distributions. The collection of hidden target values *should* be normally distributed in the target error distributions, by the intrinsic Kriging hypothesis. They are the red X's in Figure 4.4.

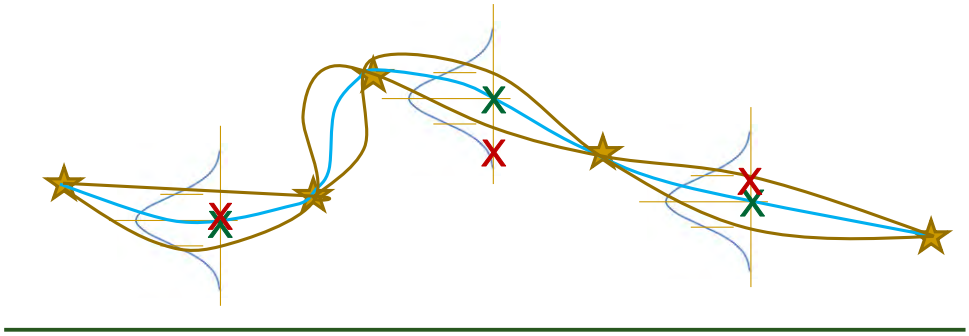


Figure 4.4. Superimpose the hidden target values on the target error distribution estimates

In Figure 4.5, by following the succession of transforms via the green arrows, combine the original targets scaled to an N(0,1) distribution [red crosses] on the left of the figure. Then rotate this distribution to the bottom figure, then finally plot the scaled original target values on a standard Gaussian distribution curve to obtain their cumulative probabilities F(x).

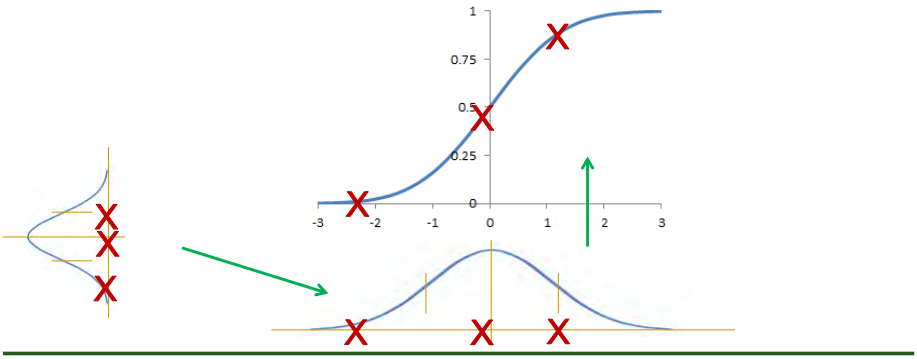


Figure 4.5. Rotate the assembled standardised original target data relative to the target distributions to the horizontal, following the sequence of green arrows; project the original targets onto a Normal cdf and note their $F(x)$ values.

In Figure 4.6, we plot the $F(x)$ values against their scaled ranks = $j/(n+1)$, where here n = 3.

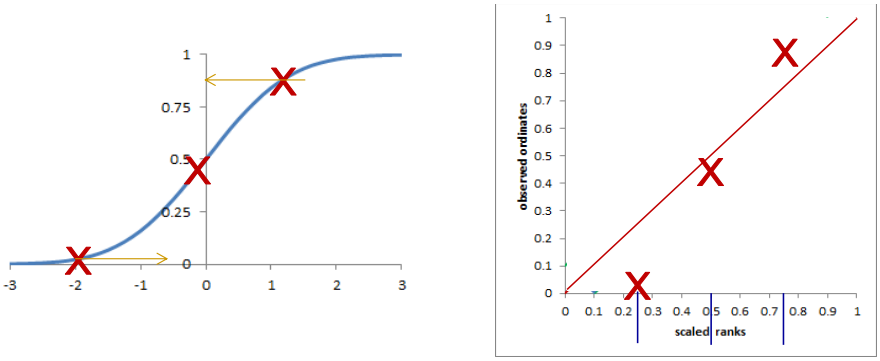


Figure 4.6. Plot of the $F(x)$ values [y-axis] against their scaled ranks = $j/(n+1)$ [x-axis].

If the procedure is good then the points should lie close to the diagonal [see a more realistic example below left in Figure 4.7]; if not, then the procedure is biased [see below right].

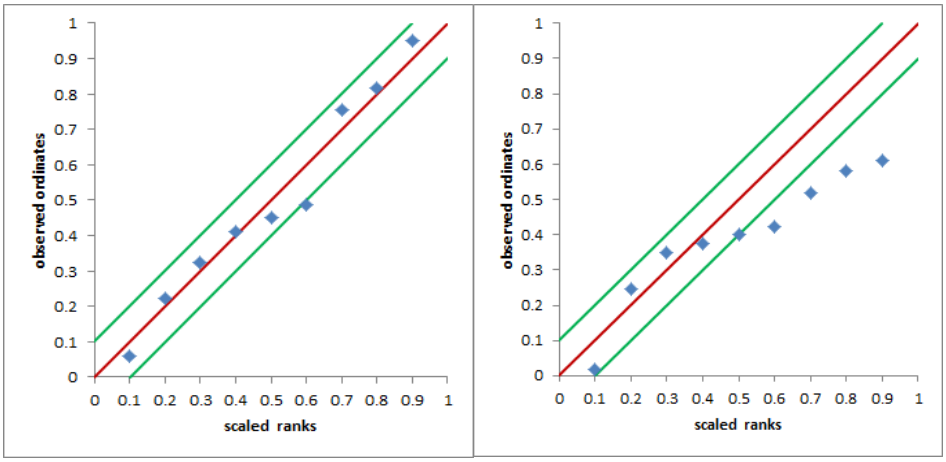


Figure 4.7. An acceptable [left] and unacceptable [right] plot of $F(x)$ values of the hidden target values against their ranks – a fictitious development of Figure 4.6.

Bardossy and Pegram (2014) applied the Kolmogorov-Smirnoff test for the infillings of Seasons 1 and 2 of the monthly Cape data, as shown in Figure 4.8. The infillings were chosen from the 26 monthly RSA data sets estimated by Copulas and subjected to Cross-validation as described above.

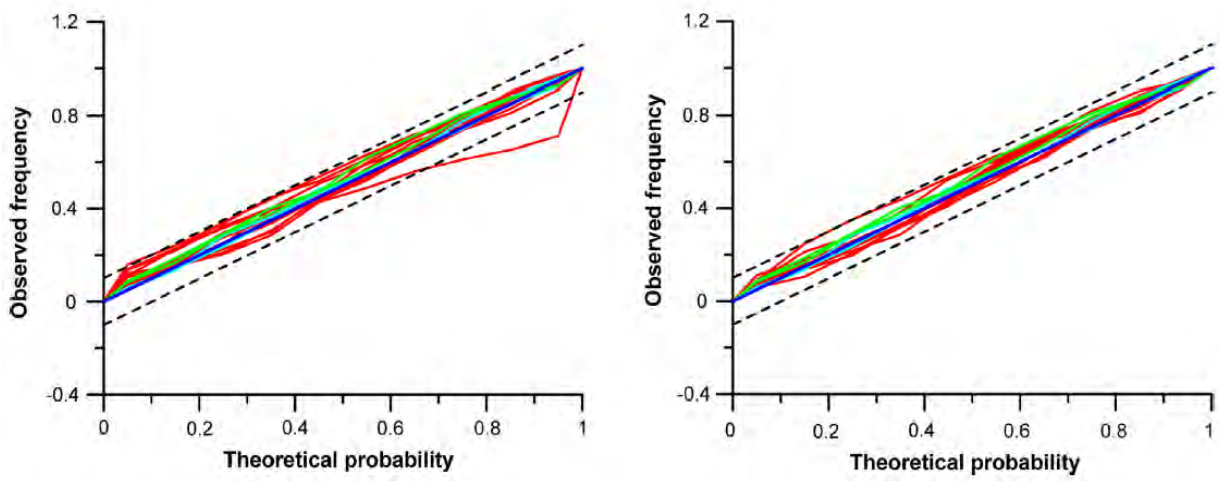


Figure 4.8. Based on the idea in Figure 4.7, the figure shows the infillings from the 26 monthly RSA data sets for the two 6-month seasons estimated by Copulas; the blue line is the 1:1 relationship, the green and turquoise lines are 4 of the 26 cumulative plots, to show their individual behaviour; the 22 red lines are the complement of the 26 plots.

In Figure 4.8, 25 out of 26 [= 96%] cumulative [Uniformly distributed] distribution functions lie inside the 95% Confidence Limits [indicated by the dashed lines], demonstrating that the Copula-based interpolation procedure is successful for the experiment with these 13 Cape stations. What is more, the method yields the precision of the interpolation, offering us a way to provide the uncertainty of the estimates.

4.2 Correlation Links and how to make them meaningful for Infilling and Interpolation

Consider the following fictitious example which demonstrates how careful one needs to be when estimating correlations between samples. Take two sets A & B of Normal (0, 0.5) random numbers, whose sample serial correlation coefficients are negligible -.08 and -.02 with a mutual cross correlation coefficient at a small 0.13, shown in Figure 4.9.

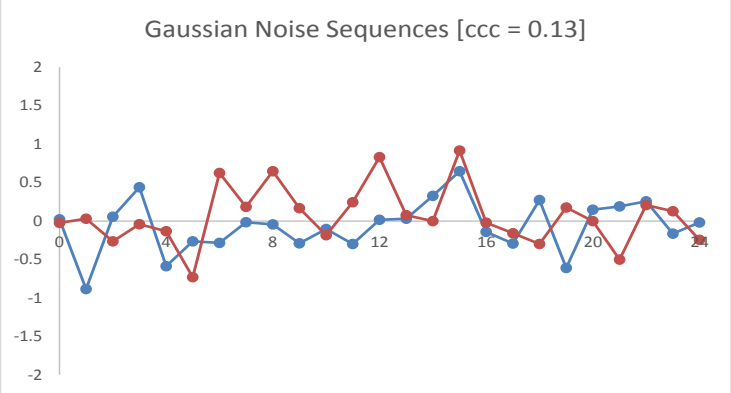


Figure 4.9. Sequences of Gaussian random noise

Take a sine wave of amplitude 1, superimposed on the noises, shown in Figure 4.10:

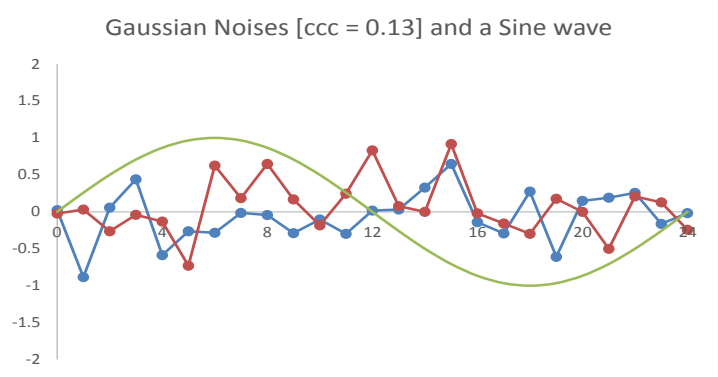


Figure 4.10. Sequences of Gaussian random noise and a sine wave of period 24 intervals

Add the sine wave to the two series A and B to get 2 new sequences C and D in Figure 4.11.

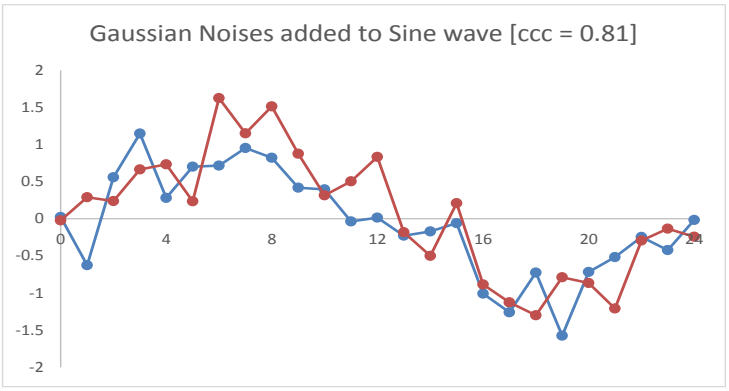


Figure 4.11. Sequences of Gaussian random noise added to the same sine wave of period 24 intervals, shown in Figure 4.9, to make series C & D

The result is an increase of the cross correlation coefficient (ccc) from 0.13 between A & B to 0.81 between C & D. This strong 'relationship' is due purely to the 'seasonality' underlying the two new sequences. The lesson learned is that, when working with sets of rainfall time series, we must deal with the seasonality sensibly, usually by standardisation and Gaussianising the result.

If we now take the sequences A & B, shown in Figure 1, exponentiate them to make them positive and raise them to the power 3 to introduce skew to yield two new series E and F. Their ccc becomes 0.52, purely due to the skewness and the fortuitous synchrony of two large values:

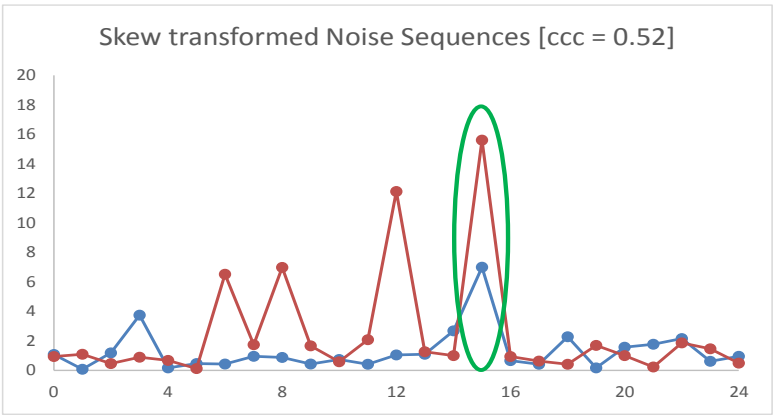


Figure 4.12. Artificially skewed sequences E & F

Sequences E and F (no seasonality added) have not changed their order relative to A & B, but the increase in ccc comes from the conjunction of two large values at 'time step' 15, ringed in green. The lesson learned here is either to use the Spearman rank correlation estimator, or to Gaussianise the series before estimating the ccc by Pearson. There are other benefits of Gaussianisation, to be described in the next section.

4.3. Spatial correlation in rainfields on individual days

To determine the spatial correlation links between gauges in a region, we chose Region 6 which mostly covers the KZN coastal areas. The period chosen with relatively well populated data-sets was 1965 to 1984, i.e. 7305 days of data over 131 stations. The sites, in context, are shown in Figure 4.13.

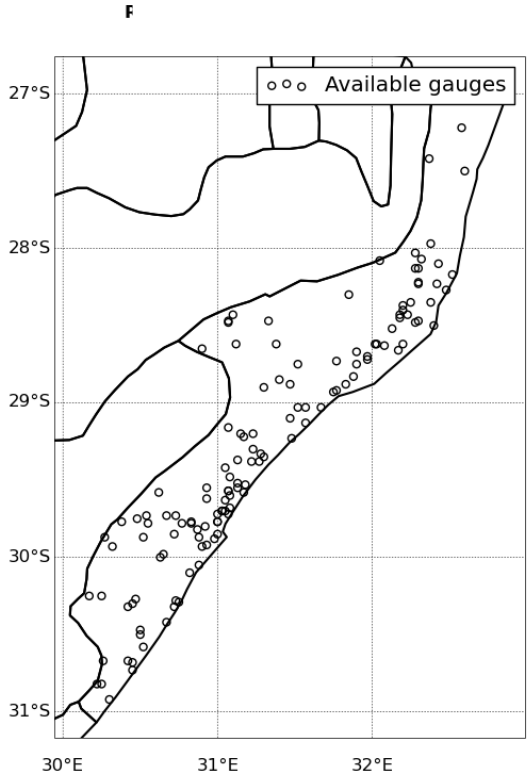


Figure 4.13. Map of Region 6 with active gauges at 131 stations during the period 1965 to 1984.

In Figure 4.14 we show the number of observations (a maximum of 131) on each of the 7305 days. These observations over the network include dry days (zeroes), and each day's count is marked by a small blue cross in the figure. There is an inexplicable fall off of high counts at either end of the period.

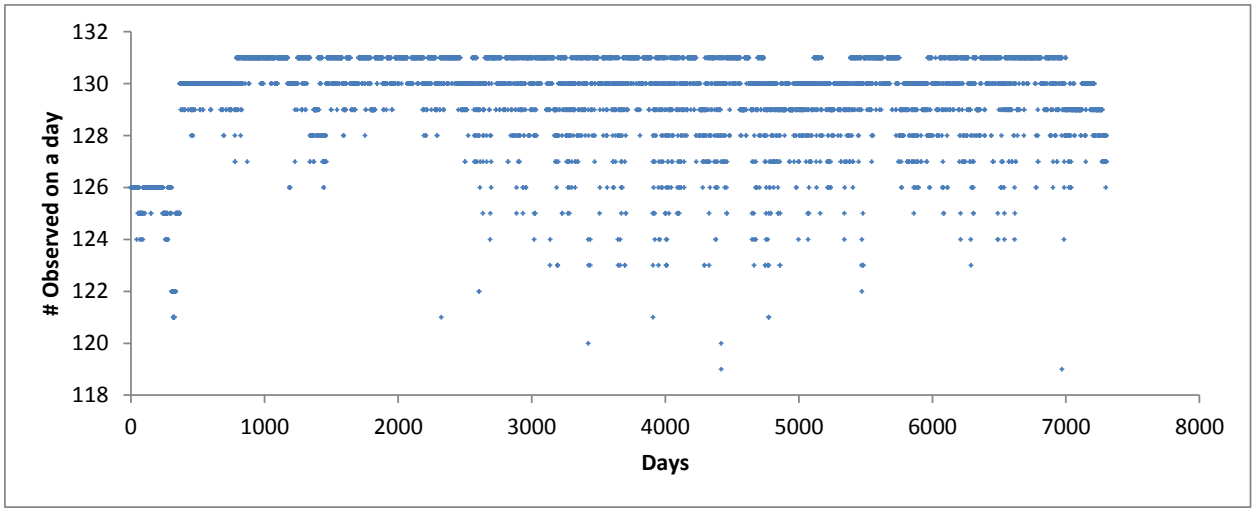


Figure 4.14. Number of intact observations per day over 20 years, 1965 to 1984 inclusive.

Figure 4.15 displays the number of wet gauges on each day of the same set by calendar day over the 20 years, as a wet proportion [WP] of the number of *active* gauges on each day recording more than 0.999 mm.

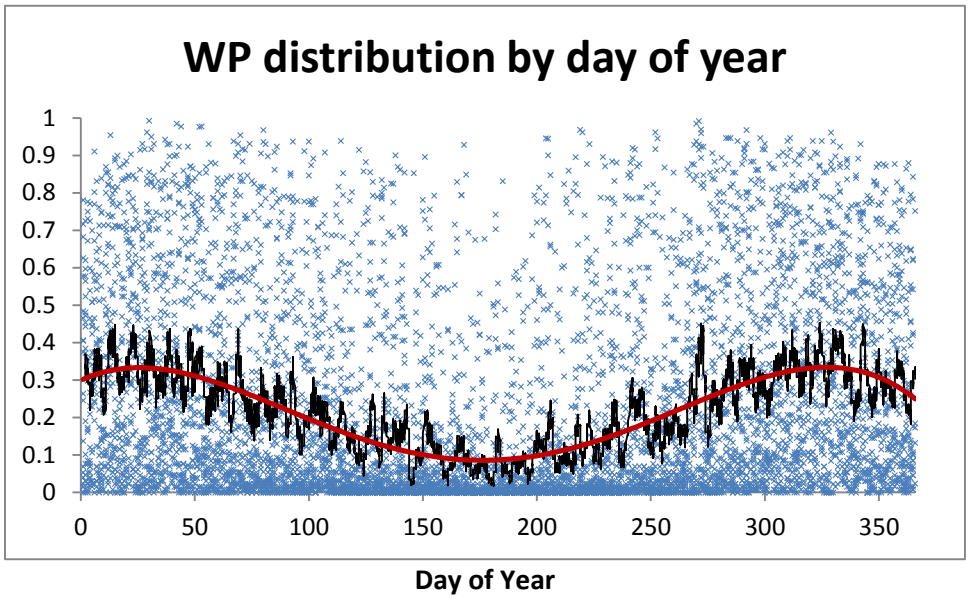


Figure 4.15. The wetness proportion [WP] of the number of days out of the active gauges on a day recording > 0.999 mm. The black line shows the mean WP for each calendar day, while the red line is a smoothed trend-line fitted to those means.

Clearly the Summer and Spring seasons [October to April] are wetter than the others. The fall-off at both ends of the year is inexplicable – we could not work out the reason. We decided to separate the data into classes to determine the links between them. The following Table 4.1 gives counts in each class determined by Wetness Proportion (WP).

Table 4.1. Counts of days experiencing different levels of wetness

CLASSES	1	2	3	4	5
WP	< 0.1	0.1 <	0.3	0.5	0.7
Count	3708	1482	773	674	668

We determined the cccs by WP using the following procedure. From the latitude y_i and longitude x_i values for each gauge in minutes, we get the interstation distance d_{ij} in minutes of arc as $d_{ij} = [(x_i-x_j)^2 + (y_i-y_j)^2]^{1/2}$ in order to build the correlogram for each wetness class. This distance d_{ij} is proportional to the true great circle distance for gauges that are contained within an area the size of region 6. 131 gauges yields 8515 deltas (interstation distances). We count the number of intact days, the number of wet days and the total depth of rain above 0.999 mm on each day, then

(1) find the WP for each day
(2) on each day separately to eliminate the effects of seasonality, Gaussianise the intact data, putting the zeroes to y_0, the mean of the lower tail of the N(0,1) distribution as $z_i(t)$ for station i on day t, following Figure 4.1
(3) then calculate the cccs, c_{ij}, over all days with a given WP and station distance d_{ij}

This procedure ensures that seasonality has been removed on each day by the Gaussianisation, so that we do not fall into the trap described in Section 4.1. The result is the following set of images in Figure 4.16, where we have neglected the driest class 1.

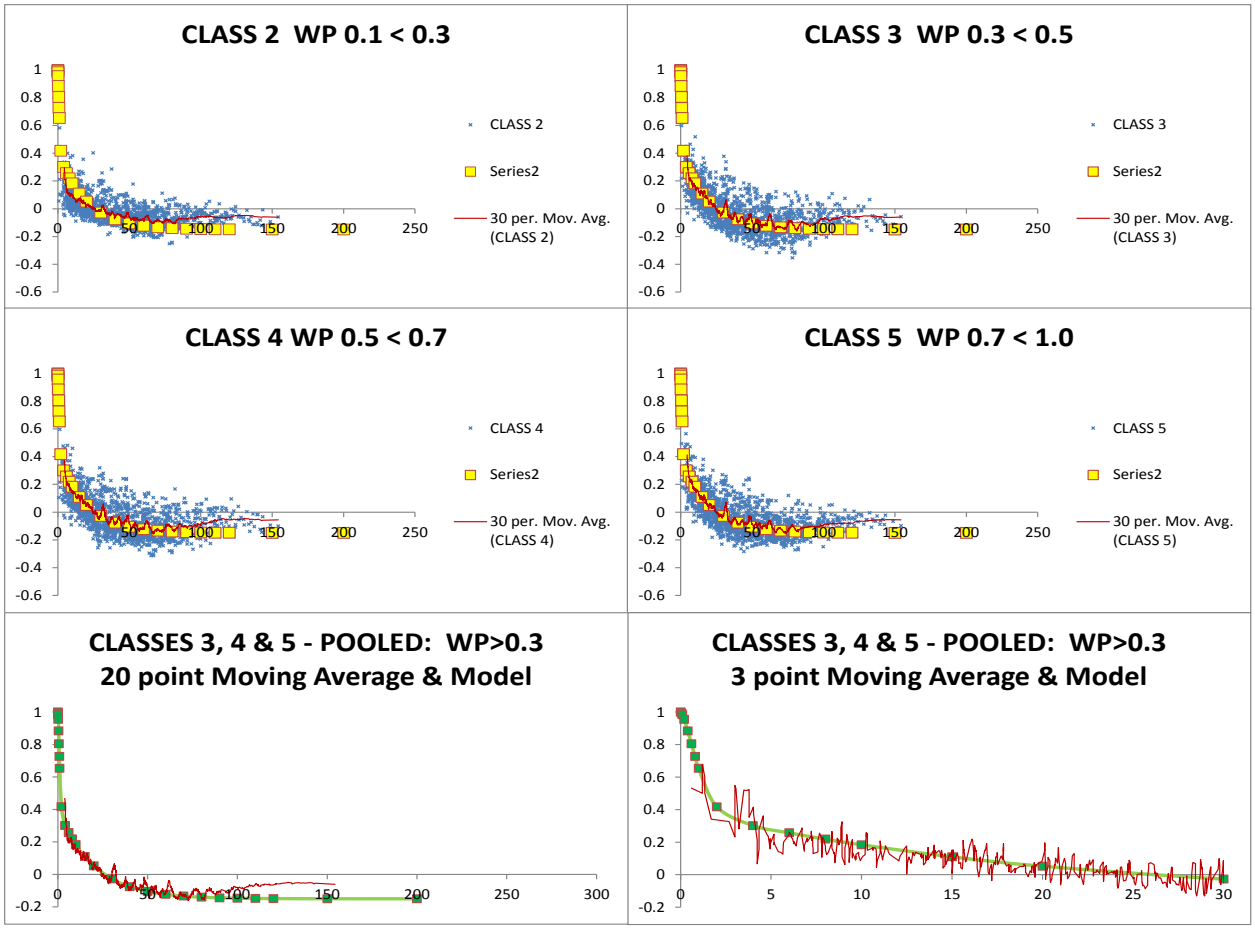


Figure 4.16. Individual spatial correlation coefficients [blue crosses] by interstation distance [horizontal axis]. The top 4 images show results for the upper 4 classes; the lower two images the results for the pooled classes 3, 4 & 5, with points removed and substituted by a moving average. This is given by the red wriggle, which is a moving average of various lengths in each class. The yellow and green markers indicate the fitted correlation model. The bottom right panel is a segment of the bottom left from the origin to 30 minutes of arc [about 48 km].

We note that the Wetness Proportion has very little effect on the correlation models obtained, hence pooling the upper 3 classes is a valid decision.

The fitted correlation model in Figure 4.19 is the same in all WP images (shown as the solid line with rectangular markers) and is a mixed hybrid exponential model, designed to capture the quick drop-off and the long tail which goes negative after 26 minutes of arc, then flattens out. Near the origin, the sharpness of the exponential is modified by the power *d* in the following formula constrained to be greater than 1 (the pure exponential model). If *d* = 2, then this component would be bell-shaped like the Gaussian.

The common spatial correlation model is:

$$ccc(h) = A.exp[-(h/L_1)^d] + B.exp[-h/L_2] - C \quad (4.2)$$

with A = 0.6, B = 0.55, C = 0.15, d = 1.5, L_1 = 1.2 and L_2 = 20

The following Figure 4.17 demonstrates what happens if the rainfall is not first Gaussianised day-by-day. The fitted correlation is unnaturally high because of seasonality, considerably higher than those in Figure 4.16, as explained in the section on Correlation Links above.

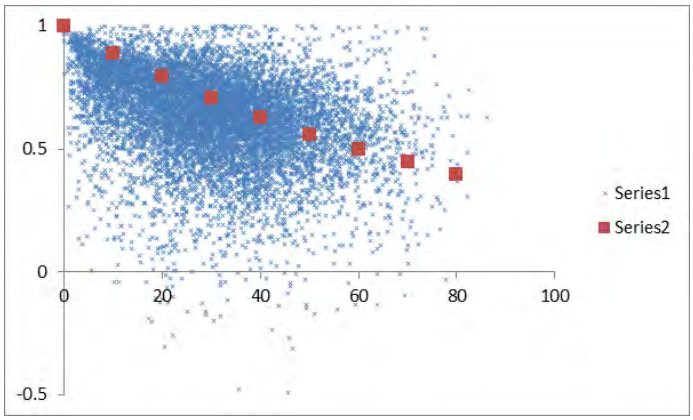


Figure 4.17. Correlation coefficient estimates obtained by temporal Gaussianisation (over the whole record), instead of performing the Gaussianisation of rainfalls each day individually, before computing the spatial correlations, as was done in Figure 4.16.

In summary, the lessons we can take from the investigations in this section are that we have strengthened and augmented the discussions in Chapter 3:

- Gaussian copulas are superior to other methods of infilling
- We must de-seasonalise data before computing cross correlation coefficients (cccs)
- Cross correlation coefficients of de-seasonalised spatial daily are independent of wetness

Chapter 5. A look at the data and the results of the infilling procedure

In this chapter we present some images of the data-set we worked with and demonstrate the results of the computed infilled missing data with estimates of error-bounds.

The first figure is a time series of the number of active gauges in the CSAG data-base, dramatizing the serious recent drop-off in numbers. This image appeared in the Executive Summary as Figure ES.2.

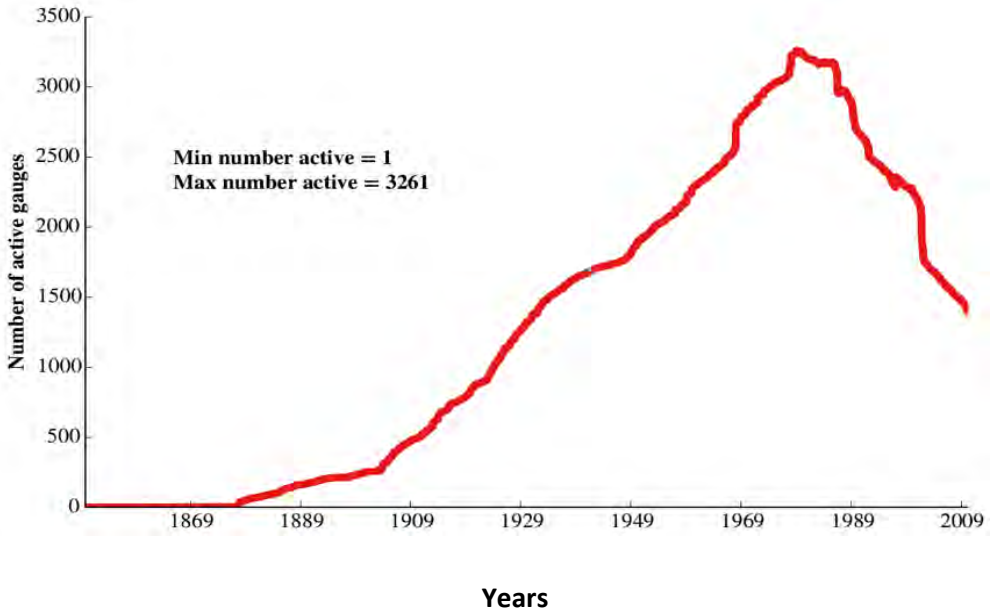


Figure 5.1. Count of active gauges in the CSAG data-base over the last 160 years

The next figure is a histogram of the frequency of gauges with altitude, noting that some of the CSAG gauges have no altitude recorded.

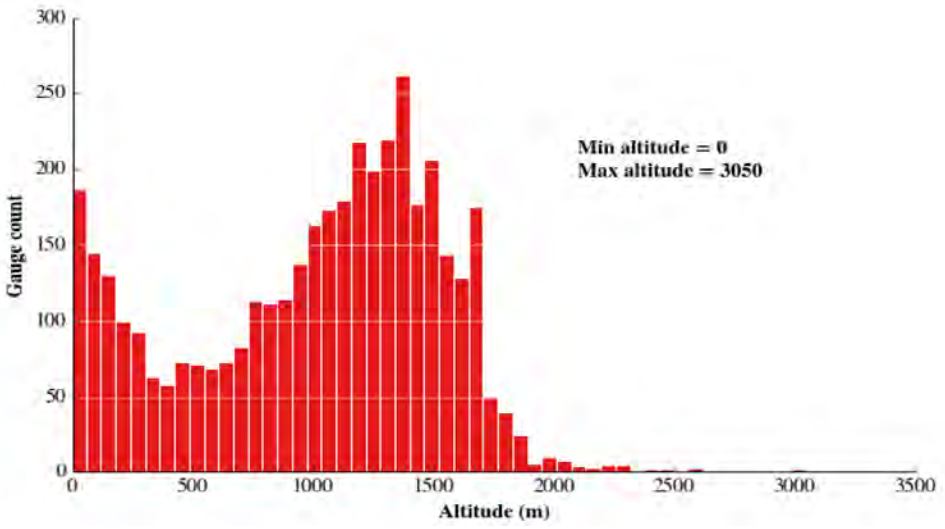


Figure 5.2. Count of gauges by altitude in the CSAG data-base

On this page and the next, we show a detail of the number of active gauges in our history. The first set in Figure 5.3 shows the history of years with intact data; Figure 5.4 gives some detail.

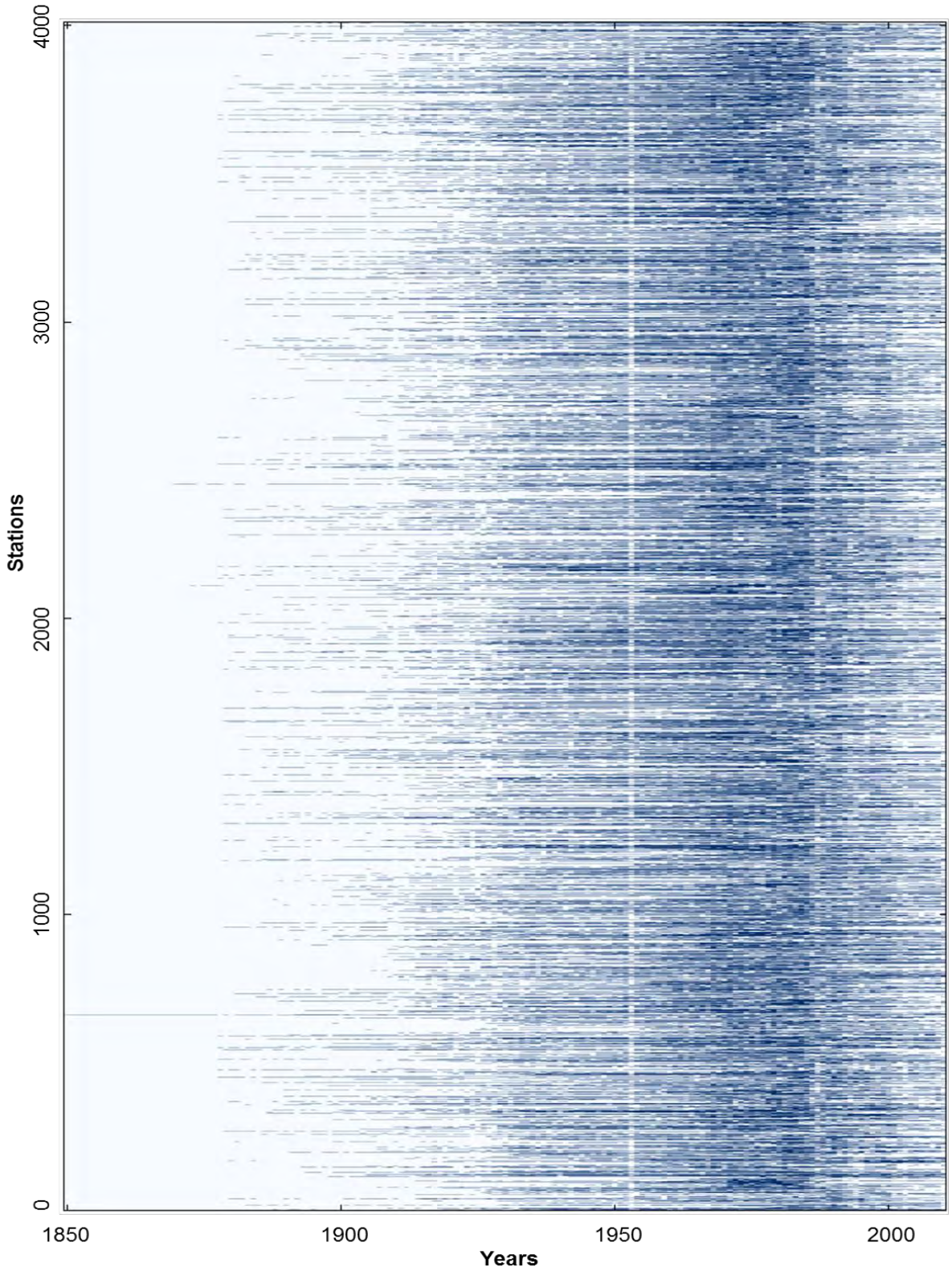


Figure 5.3. Image of the number of years of gauging with full data. We have no explanation for the pale period in the early 1950s.

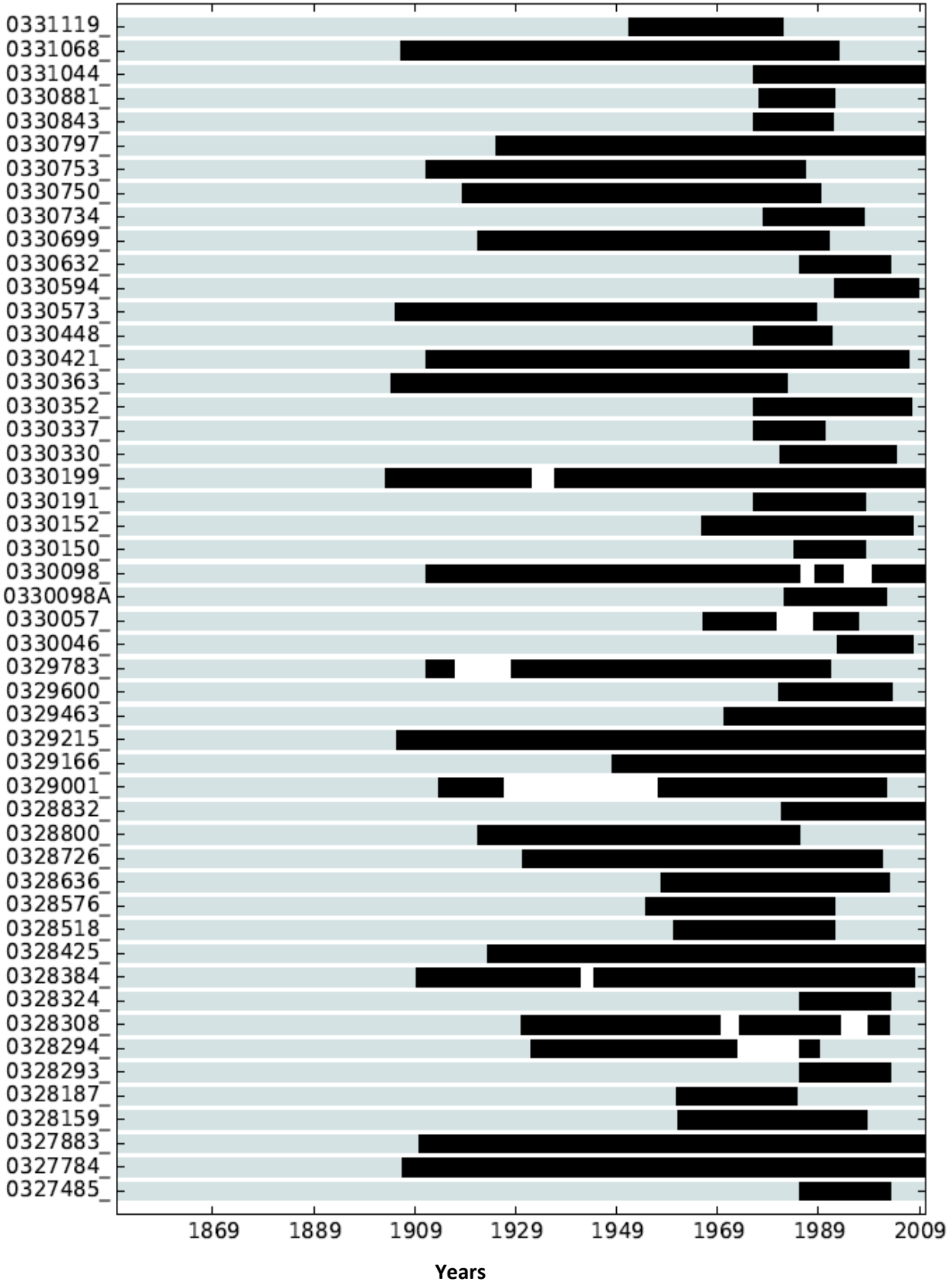


Figure 5.4. Gauges in the Eastern Free State alive in calendar years – SAWS gauge numbering.

5.1 The Infilling procedure described, following the algorithm of Section 3.2

Figure 5.5a shows the availability of 20 control gauges [blue] and a target gauge [red] in South-western KwaZulu-Natal. Their data availability over the period is shown in Figure 5.5.b as control gauges [black] and the target gauge [red]. The procedure described in pseudo-code described in Chapter 3 is used to infill the target with the highest correlated gauges available at any epoch. The steps are repeated and realised below.

Step 1. Pick a target station, together with up to 11 controls surrounding it and assemble the subset for a given CP-group in a season, as shown in Figure 5.5.

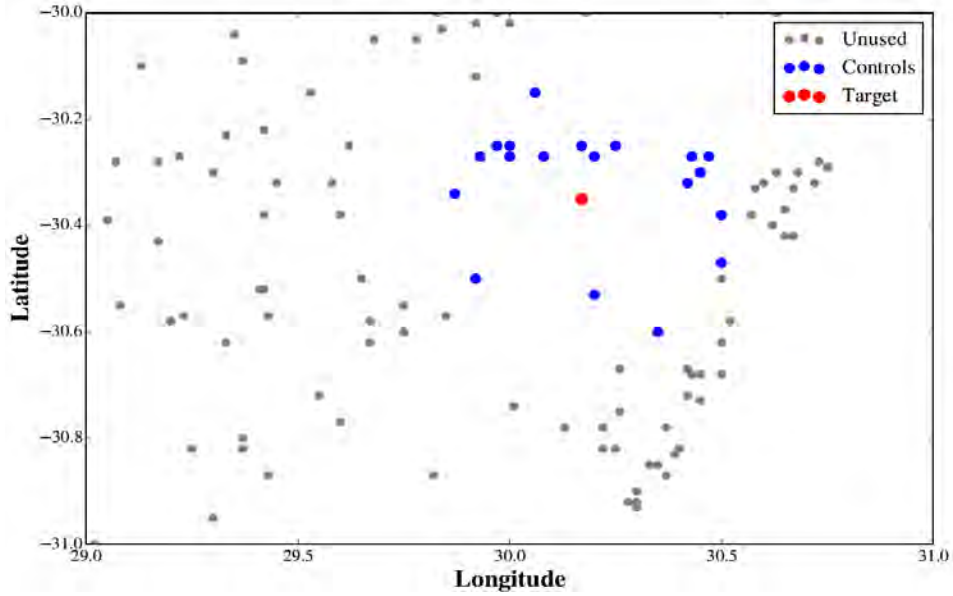


Figure 5.5a. An example of selected target and control stations

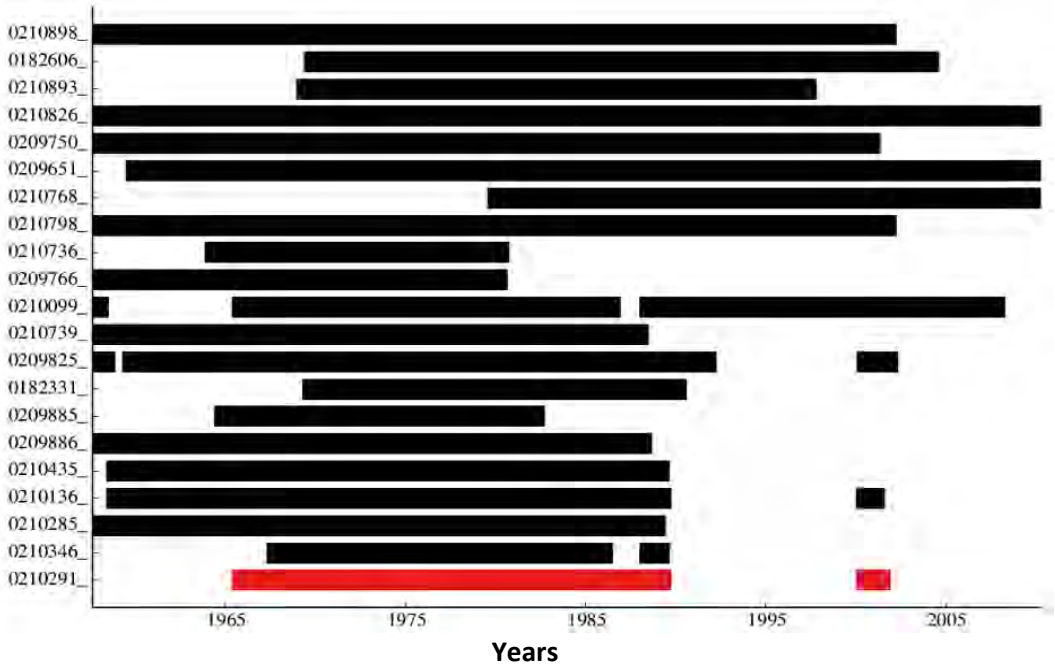


Figure 5.5b. Bar-chart of availability of data of 20 gauges (SAWS numbering) for infilling one gauge, starting from year 1957.

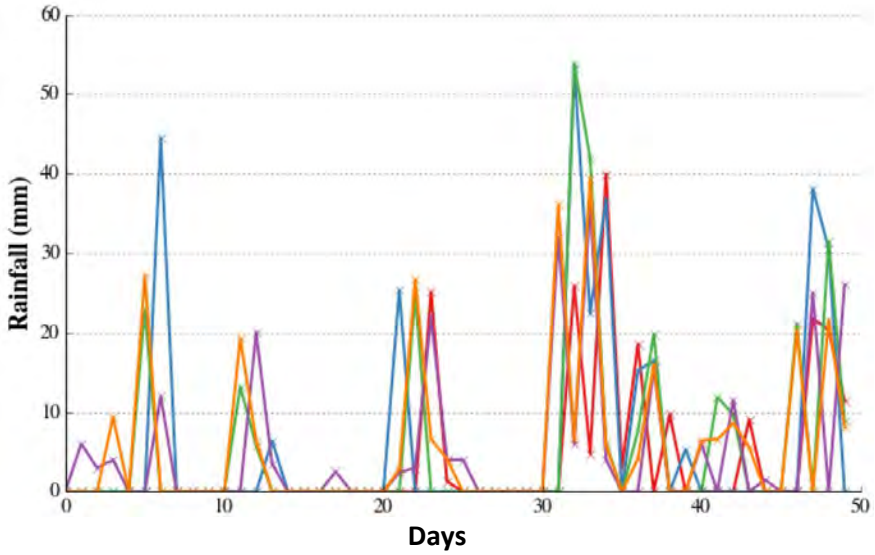


Figure 5.6. A 50-day period of recorded rainfall at 11 selected stations surrounding the target.

Step 2. Gaussianise the target, and then all the controls in turn, because the treatment of the zeros depends on the number of (i) missing values and (ii) the number of dry stations on each day, as indicated above.

The result is shown in Figure 5.7; note the different Gaussianised values of the zeros, each depending on the number of dry gauges on the day, using Equations (3.9) and (3.10).

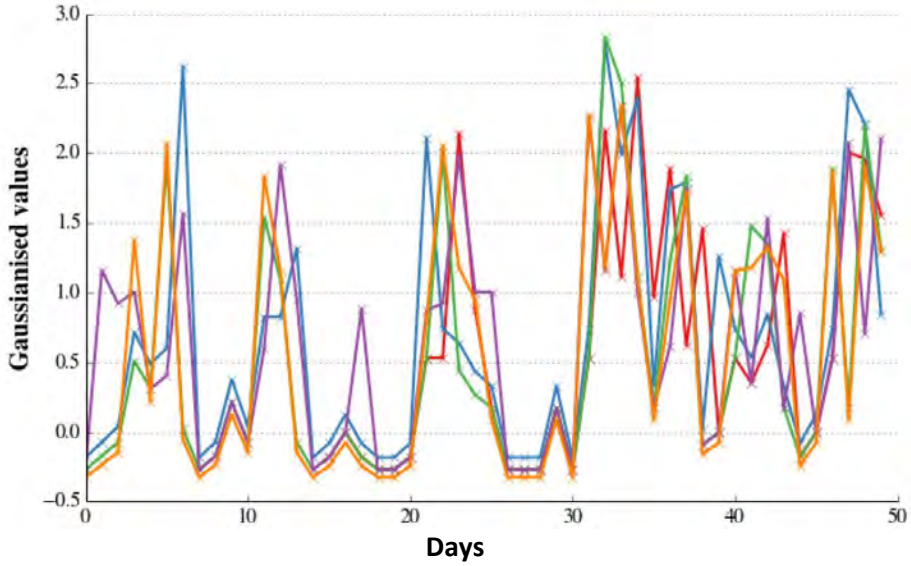


Figure 5.7. Gaussianised data from Figure 5.6

Step 3. Do a quick preliminary cross correlation coefficient [ccc] calculation between all stations for the CP-defined group using their full Gaussianised records, so that the controls can be ranked relative to the target from highest to lowest ccc

Step 4. Once target and controls are all Gaussianised, assemble them in a matrix with the target in the first column and the controls, ranked by their cccs with the target from highest to lowest in the remaining columns

Step 5. Call the Infilling routine to read the matrix, infill the target's missing data and output the repaired Gaussianised target's data conditioned on the CP, with mean and stdv [$\mu(t)$ and $\sigma(t)$] associated with each infilled element

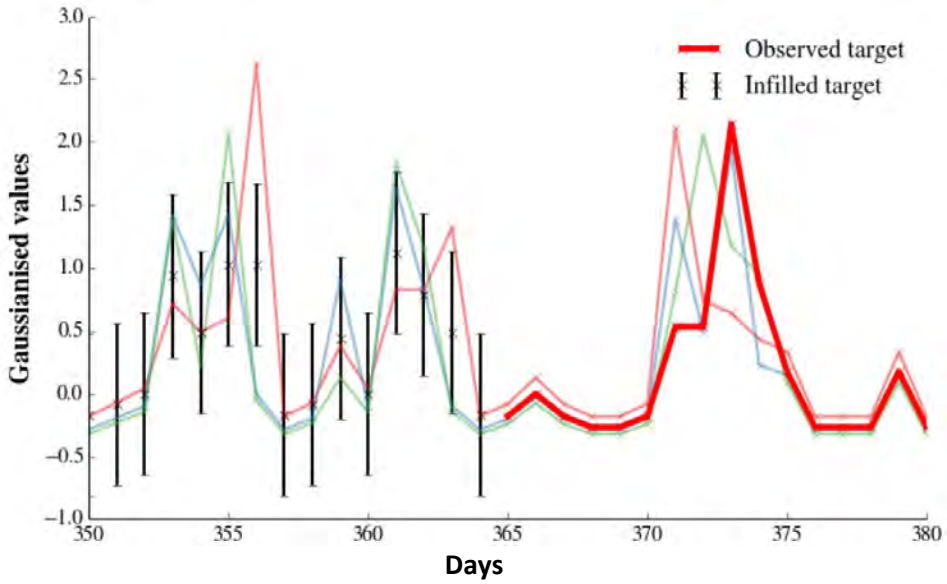


Figure 5.8. The infilling procedure, showing the expected value and the range of 2 standard deviations calculated in the Gaussian domain. 3 of the controls' time series are also shown; the period is different from Figure 5.7. Also shown is an intact part of the target's record.

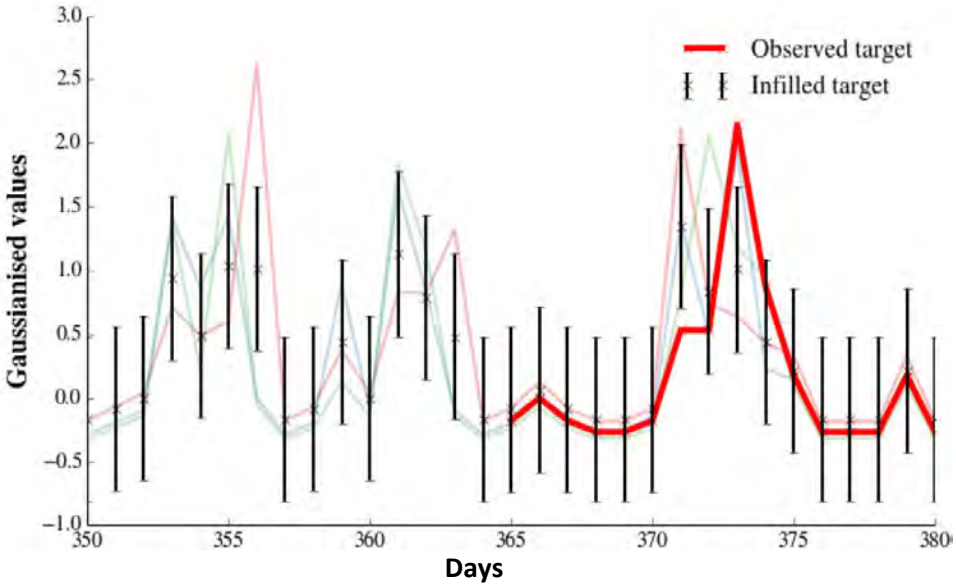


Figure 5.9. The same as Figure 5.8, except that the intact part of the target's record was hidden and then infilled, as a visual cross-validation exercise.

The temporal shift between target and the controls on days 370-4 is plausibly due to misread gauge data; nevertheless, only 2 of 14 [14%] of the estimates were outside the error-bars which contain 69% of the probability [complement 32%], indicating a successful infilling. The next step is to:

Step 6. Pass the CP-dependent repaired target vector, with $\mu(t)$ and $\sigma(t)$ values where appropriate [i.e. these are only associated with infilled values], to be reverse Gaussianised using the QQ transform to recover the estimated rainfall in mm, as well as the upper and lower quartiles of the infilled estimates.

To demonstrate the effect of reverse Gaussianisation of the infilled data, we turn from infilling daily data to some annual data to show the result of infilling long sequences of data. We offer two sets of targets and controls; one for a medium wet area and another for a wetter one. The 4 separate panels of Figures 5.10 to 5.13 show the observed marginal distributions of target and controls and then the associated infillings with error bars for two sets of annual data. The first set is from a moderately wet site (Figures 5.10 and 5.12, with MAP of 730 mm), the other set is from a much wetter site (Figures 5.11 and 5.13, with MAP of 1010 mm). Note the near Normality of all the raw data in Figure 5.10, with more skew in Figure 5.11.

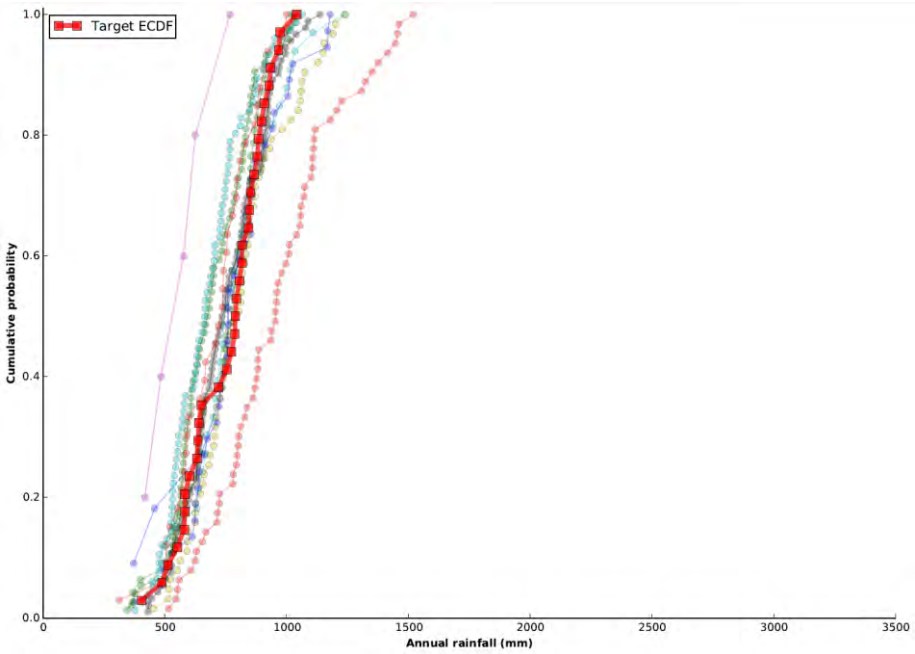


Figure 5.10. Marginal Distributions for set 1 of Target and Controls before infilling: Moderately wet. The result of the infilling of annual data with uncertainty; note the different counts of data for each gauge.

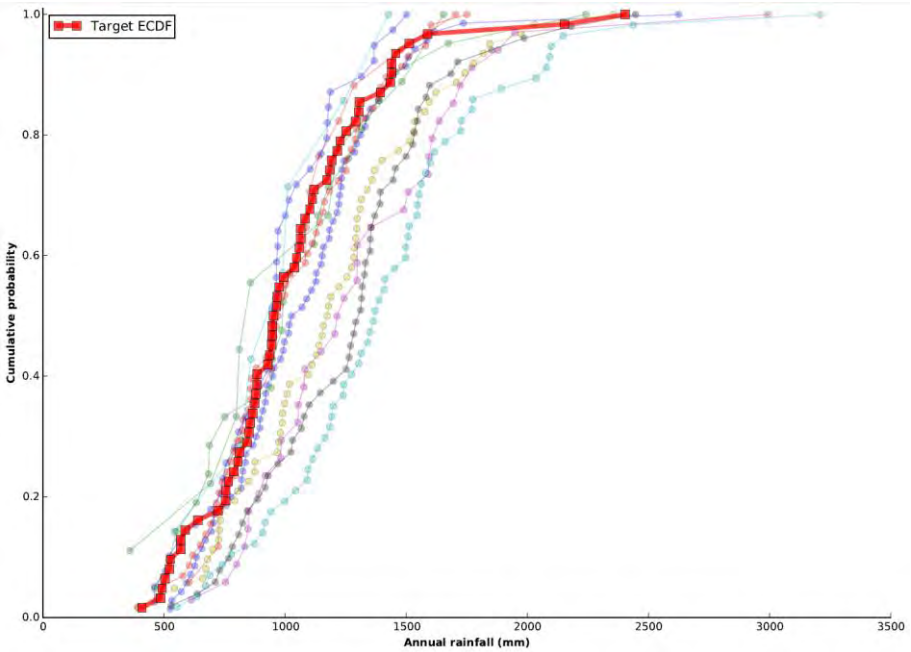


Figure 5.11. Marginal Distributions for set 2 of Target and Controls before infilling: Very wet

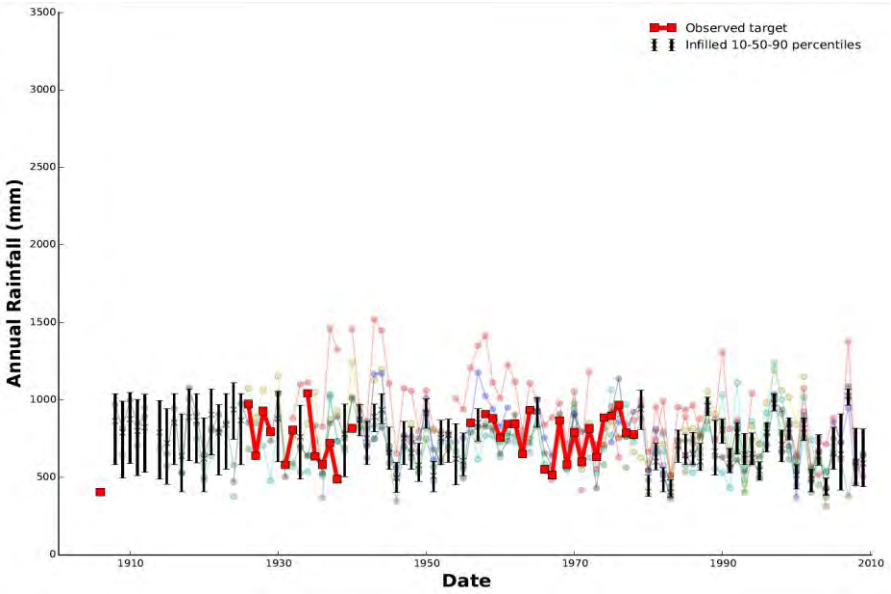


Figure 5.12. Infillings for set 1 of Target and Controls: Moderately wet, showing the corresponding intact data and the error bars (10^{th}, 50^{th} and 90^{th} percentiles) of the infilled values. The vertical axis has the same limits as that of Figure 5.13 for ease of comparison.

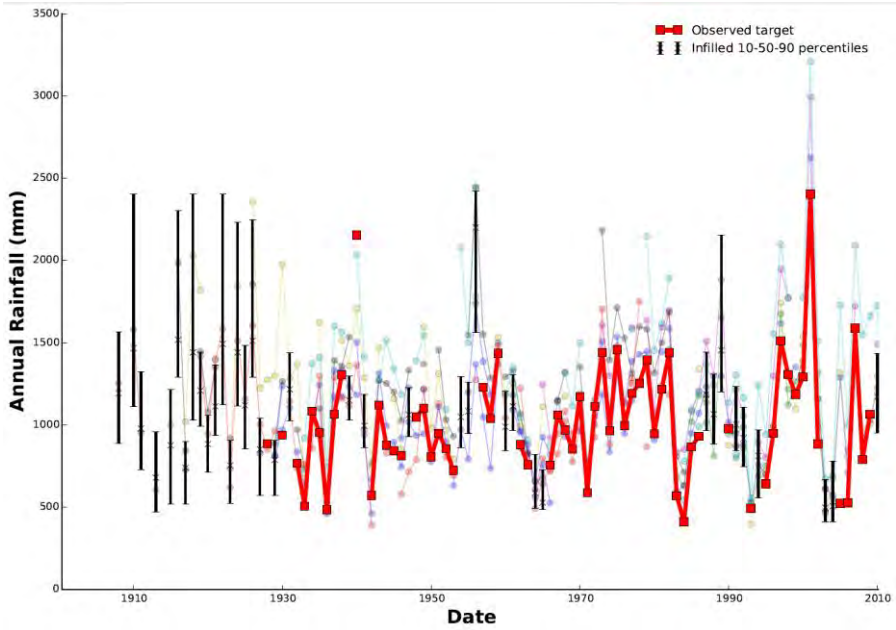


Figure 5.13. Infillings for set 2 of Target and Controls: Very wet, showing the corresponding intact data and the error bars (10th, 50th and 90th percentiles) of the infilled values.

In both Figures 5.12 and 5.13, note the variability of precision of the infilling depending on the amount of available data as indicated by the error bars. Where there are more controls, the percentile whisker plots are much narrower than where there are few controls which typically occur, but not always, in the earlier years of the 20th century. Also, the heavy upper tail of the near Normal distributions in Figure 5.11 leads to some comparatively large values in Figure 5.13.

If we take the average of the 80-percentile (or inter-decile) spread, including the zero values where there are observations, which we call Mean Annual Precision, and compare it to the calculated MAP, we obtain an indication of the precision of the infilling. This ratio, for all stations, comes at the end of this chapter in Figure 5.20, with a summary.

Step7. Choose another gauge to repair etc.

To achieve these transformations, some other housekeeping details have to be undertaken. In particular, to recover the infilled target rainfall values from their Gaussianised infilled values, we need a probability distribution of the target. The following is the procedure:

Fit a sample cdf to the available observed target data, as long as it has 20 or more observations. The sample cdf is fitted by a Gaussian Kernel Density Estimation function (KDE) as shown in Figure 5.14.

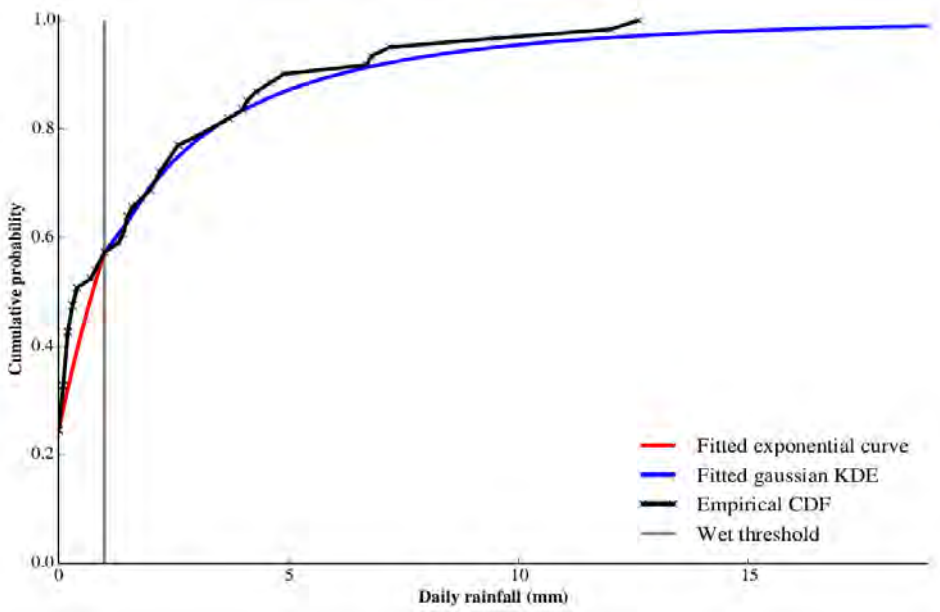


Figure 5.14. Sample and fitted cdfs of target gauge

Note that we define a wet threshold of minimum sensible measurement, here shown as 1 mm, and fit an exponential segment in red, between the P[0], here about 0.23, and the P[1] values. We generate many target values where there are no observations using Dynamic Copula Regression (DCR) then reverse transform these through the KDEs of Figures like 5.14. A typical ensemble of 100 trials is shown in Figure 5.15.

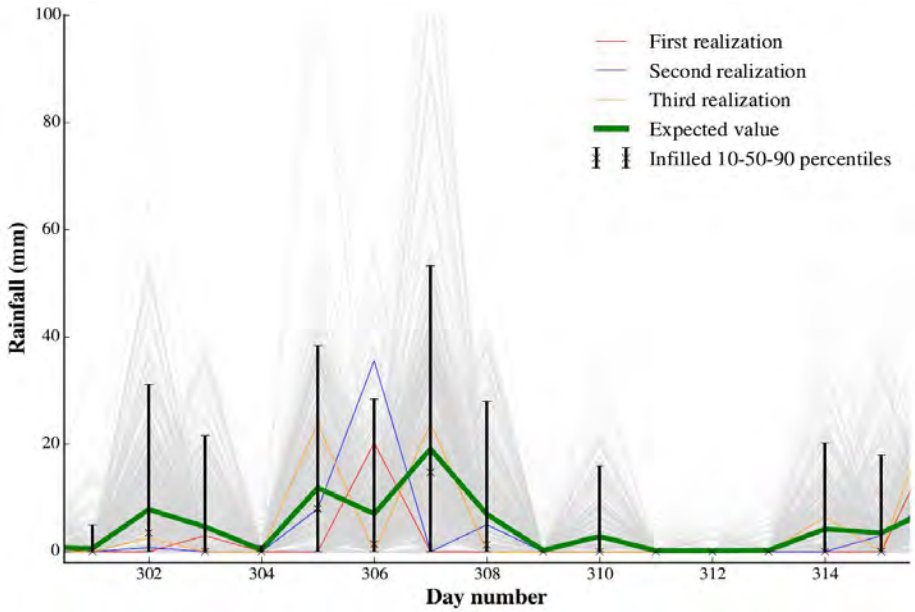


Figure 5.15. A reverse-transformed set of 100 infilled target estimates using DCR. Three single realizations are shown in red, blue and gold, while the remainder of 1000 realizations are shown as light grey traces.

In Figure 5.15, which treats daily data, we have summarised the many traces with whisker plots of 10th-, 50th- and 90th-percentiles. Note the asymmetry of the highly skewed data and infilled values. In the Figure 5.16 we give the equivalent plot for monthly data.

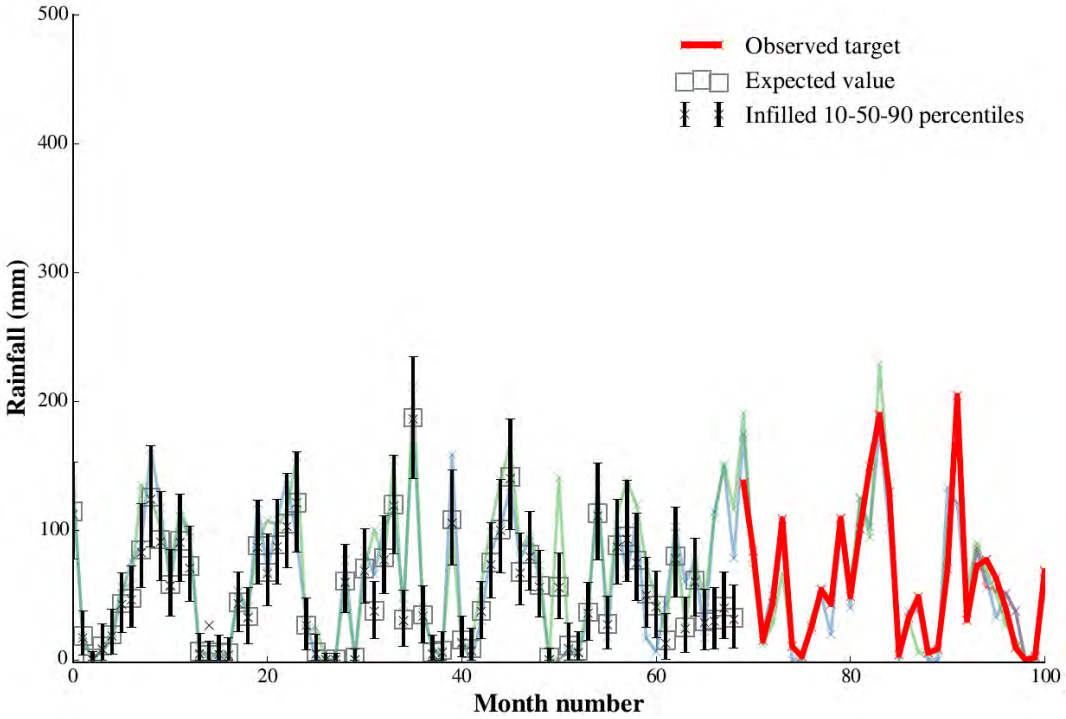


Figure 5.16. Infilled monthly data complementing a partly intact record.

We found that many monthly data are approximately Normal during the individual months, and in support of this observation note that the whisker plots of the infilled targets are nearly symmetrical about the medians. This symmetry is also seen in the whisker plots, which also vary in aperture with the precision of the estimate of the target.

In summary, the infilling procedure has been carefully (and we think effectively) done using as much of the available data as is meaningful and useful. We are going to have to live with the fact that we cannot exactly capture the past but can at least offer some understanding of its uncertainty.

5.2. Problems with Data that had to be overcome

We now turn to the thorny issue of the problems inherent in the data-sets, for which we tried to find automated solutions, but with some difficulty. The first is the duality of the numbering in some cases. In Figures 5.17 and 5.18 we show 2 records which, although parts are missing, are clones of each other. These hampered the computations, so that there was a need to perform some triage. Problems such as those illustrated here typically appear for stations that are closed and moved to nearby locations. This required developing software to weed out the offending gauges from the portion of the data-set used, else the infilling programs crashed.

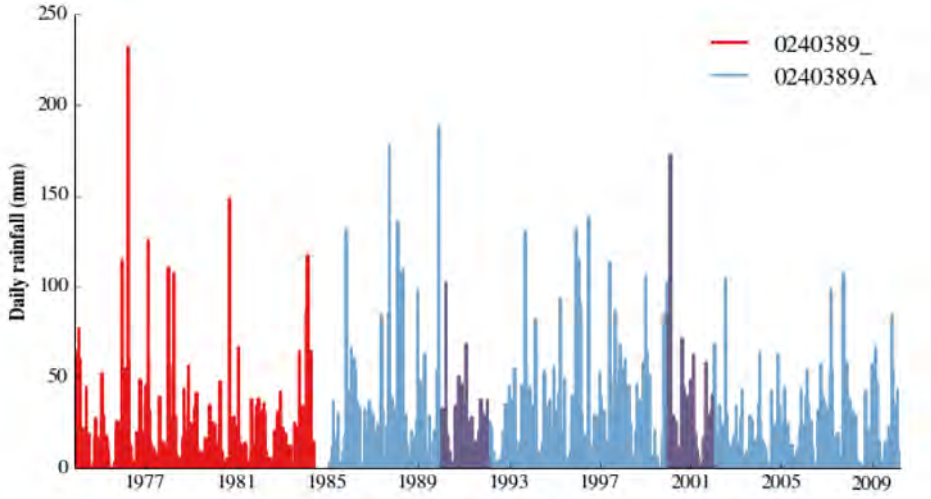


Figure 5.17. Two records [red and blue] which stop and start at different times, but where they overlap [purple] are identical.

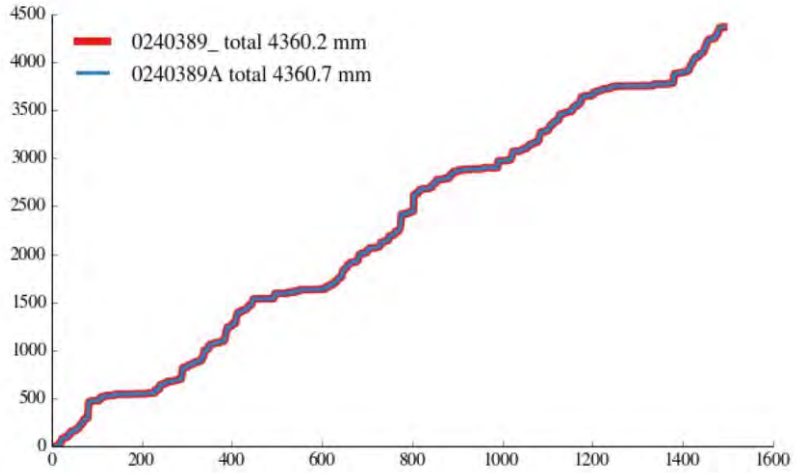


Figure 5.18. Cumulative plot of the combined overlapping periods of the records shown in purple in Figure 5.17.

In Figure 5.18, the total rainfall in the two records differs by 0.5 mm in 4360 mm. The next example shows mixed behaviour.

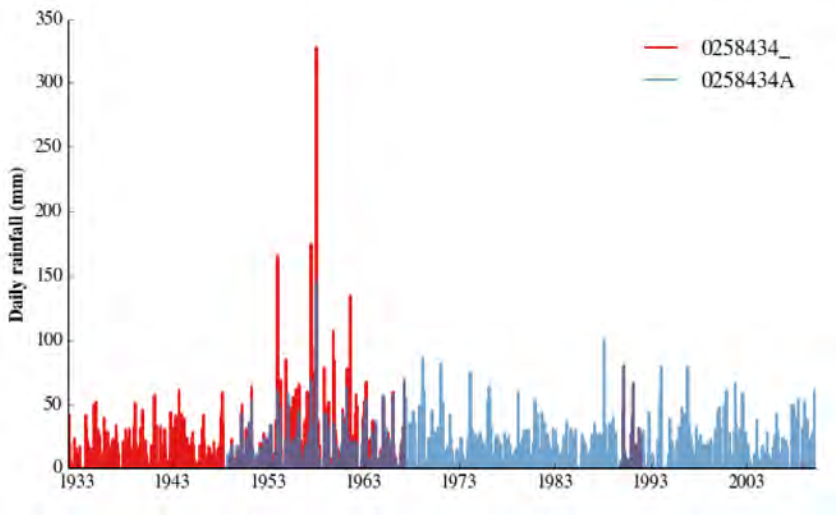


Figure 5.19. Two records, covering overlapping periods for similarly coded gauges. The red record overlaps the blue in two places: between 1950 and 1965, then 1990 to 1993.

In the first overlapping period [1958 to 1968] shown in Figure 5.19, there is little correspondence (cross-correlation for the period = 0.66], but in the second starting in 1990, the two are identical. This is a puzzle because they have the same code and are within a kilometre of each other.

We conclude this chapter on infilling with two figures in Figure 5.20 which illustrate the relative error associated with each infilled station. The first compares the infilled MAP values with the MAP of the observed data before infilling. The maximum ratio of spread to MAP is 8% and is mostly about 4%.

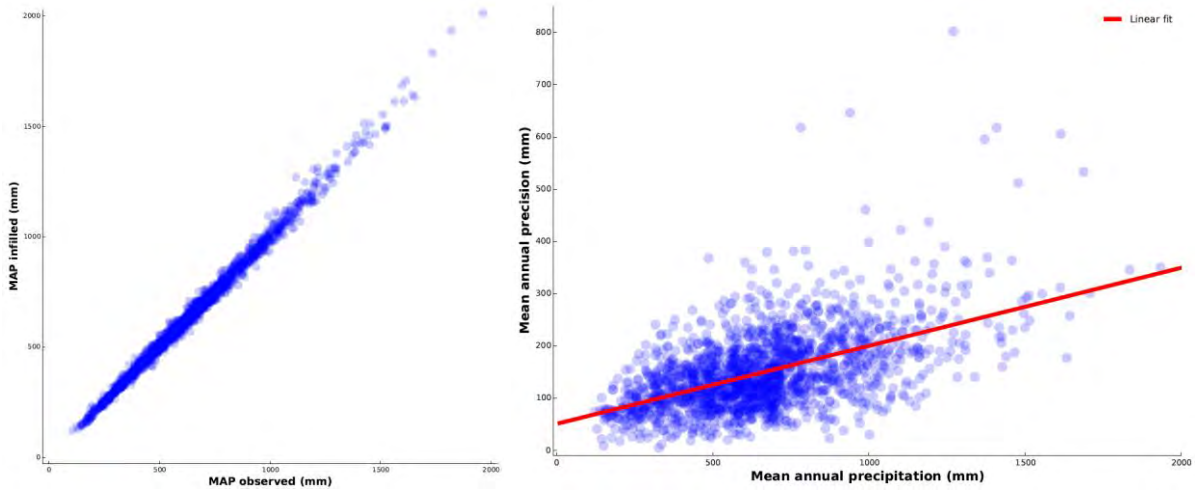


Figure 5.20. Left panel: the infilled MAP values plotted against the MAP of the observed data before infilling. Right panel: the Mean Annual Precision on the y-axis (average of the 80-percentiles of infilled interval estimates), plotted against MAP, for all filled stations.

The second image in Figure 5.20 shows the Mean Annual Precision, described in the passage following Figure 5.13, computed as the average range of the interquartile spread of 80% (90 percentile minus the 10 percentile) over the complete record, including intact and infilled

values of each station plotted on the vertical axis, compared to the final MAP estimate plotted on the horizontal axis.

It is clear that there is some uncertainty associated with the individual estimates, especially the infilled short records, as shown in the right panel. Nevertheless it is comforting to note that the MAP, calculated from the mean infilled values, averages out well when compared with the observed data before infilling, as shown in the left panel of Figure 5.20.

In summary, we have devised an original and good method of infilling missing data and because the records are not unique, they have had to be carefully examined in case we make a nonsense of the procedure if the data are inappropriate. We managed this problem with a simple censoring approach, which would have been impossible without high speed computers and intelligent software creation. If we found that (i) there was a portion of a control station's record which matched another so that it would cause a singularity in the regression matrix, or (ii) if the names were the same, we would choose the longer record and abandon the other. This difficulty with the data set caused us much grief and frustration, increasing the infilling time from what should have taken one month to four.

Chapter 6. A description of Interpolation after Infilling

6.1. Why we use ensembles of infilled values for spatial interpolation

Where data are missing, the traditional way of infilling (using regression-based methods) is for practitioners to estimate the expected values and ignore their sampling variability. That treatment typically increases the correlation beyond what would have been calculated if the data were intact, as shown in Section 4.2. By resampling the distribution of the errors for the purpose of infilling using ensembles, we are able to give a more likely set of fields reflecting the true spatial structure better than the simple approach using expected values only. True, it is more work, but the benefits outweigh the effort. The result is that, when we infill missing gauge data, as described in Chapters 4 and 5, we report not only the expected value, but also the probability of dryness, the median and the upper and lower deciles, from which information we can reconstruct the distributions of the estimates of the missing values at each gauge, on the day in question.

When we generate spatially interpolated fields, we fix the observed gauge values on the day (or month or year) and sample from the distributions of the missing gauge data estimates. It bears repeating that the scheme has three benefits: (i) we have a better estimate of the mean field (with error structures at each infilled pixel in the field); (ii) we can generate ensembles of possible spatial fields, matching the observed data, getting sharp estimates of the missing values and (iii) the ensembles can be used to determine the uncertainty of the fields.

The following 6 images in Figure 6.1 illustrate the points made above. We took 2 sets of 25 Normally distributed random numbers, we call *x* and *y*, and correlated them. They are plotted, as Figure 6.1 (a) through Figure 6.1 (e), reading from left to right and from top to bottom.

- Image (a), plotting *y* against *x*, shows that the $R^2 = 0.3931$.
- In (b), 5 values were removed at random from (a) and the survivors' (controls) regression equation was fitted as $0.7692x + 0.6018$, with a consequent drop in R^2 to 0.3576. The estimated (expected) values are shown in red, but are not included in the regression line fitted to the blue points in this image.
- (c) Shows the expected missing values included with the controls after the regression; note the increase of R^2 to 0.4529, exaggerating the dependence beyond the original.
- In contrast, (d) shows 20 sets of the 5 missing target values, calculated from the regression equation of (b) by including random errors and Figure (e) shows them combined before fitting a trend line
- (f), the bottom right panel, shows the trend-line fitted through the 'observed' set in (b) with the simulated data-sets in (d). The R^2 comes out to a more believable value of 0.362, which is closer to that of the original data (0.3931) than the value in (c) (0.4529).

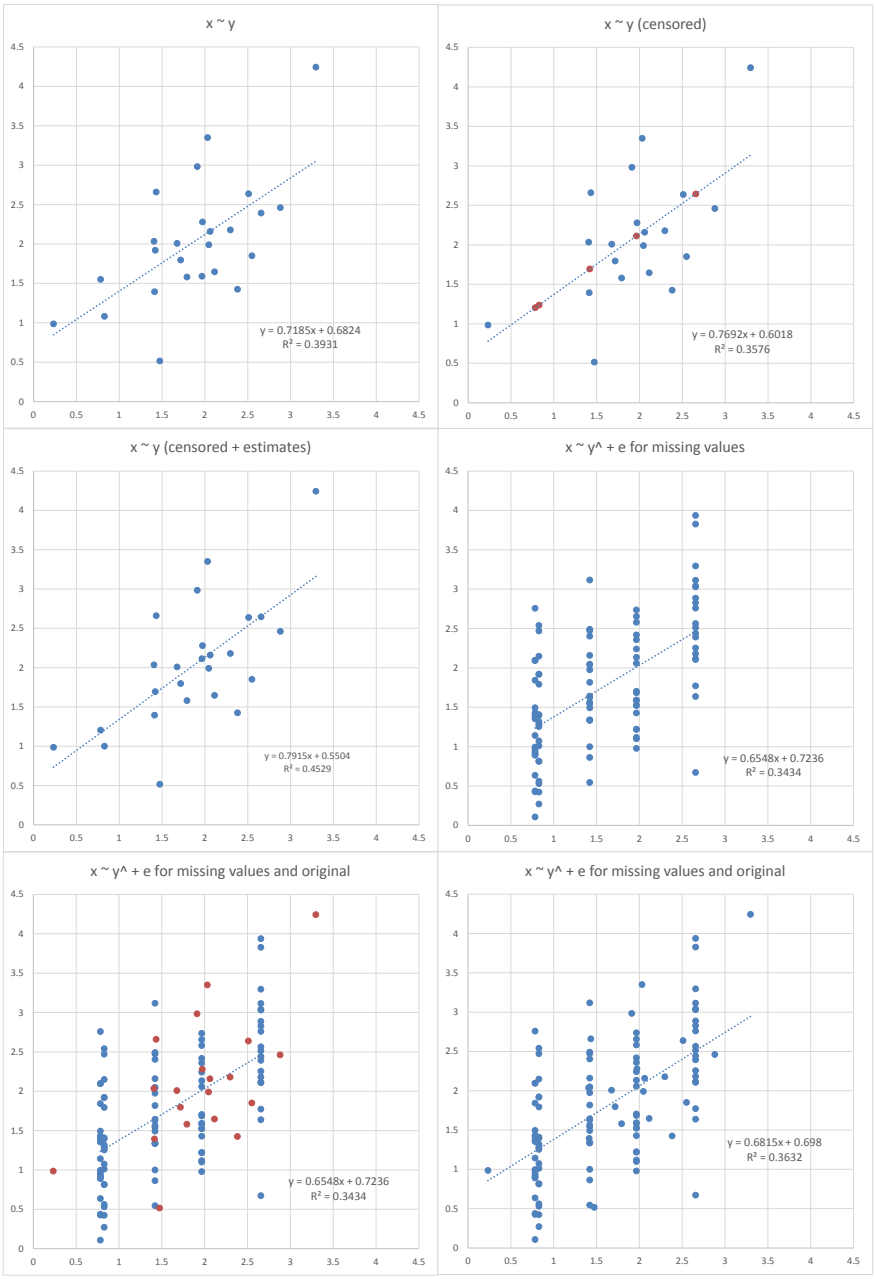


Figure 6.1(a) to (e). Regression example, comparing relationships of ensemble estimates with expected values.

This simple illustration is the justification for using the distributions (and not just the expected values) of the estimates of the missing data values in the gauge records for two reasons: (i) a realistic assessment of the value of the infilled data can be made and (ii) these distributions contribute realistic uncertainty in the estimation of generated spatially interpolated rainfields.

6.2 Why we use Gaussian copulas to interpolate spaces between gauges

Borrowing from our work in Germany (Bardossy and Pegram, 2013), the following images in Figures 6.2 and 6.3 show the estimates and associated errors determined by three

competing methods of spatial interpolation [the successor to, and extension of, infilling]. The methods are: Ordinary Kriging [OK], External Drift Kriging [EDK], which uses altitude as the exogenous variable, and Gaussian copula estimation.

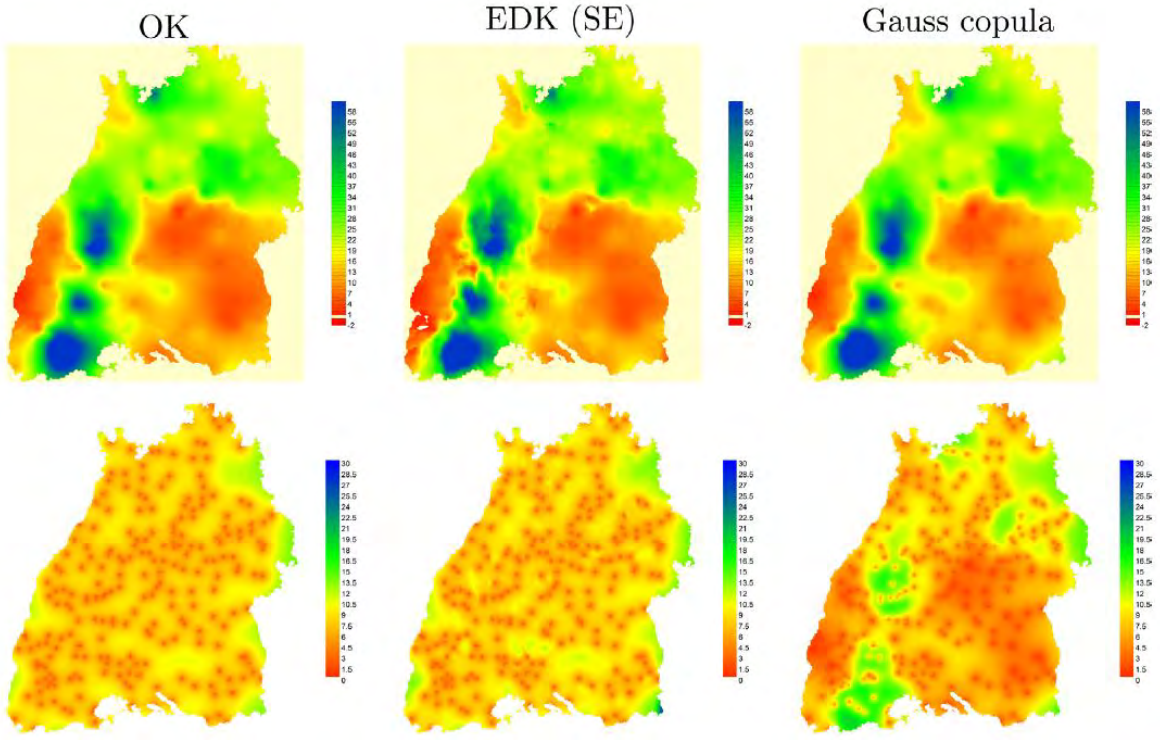


Figure 6.2. Results of a day-by-day interpolation: Baden-Würtemburg, December 18, 1993. The upper row shows the estimated mean field. The second row shows maps of the standard deviation of the interpolated values (Bardossy and Pegram, 2013). The scale of rainfall has a maximum of 60 mm; the standard deviations a maximum of 30 mm.

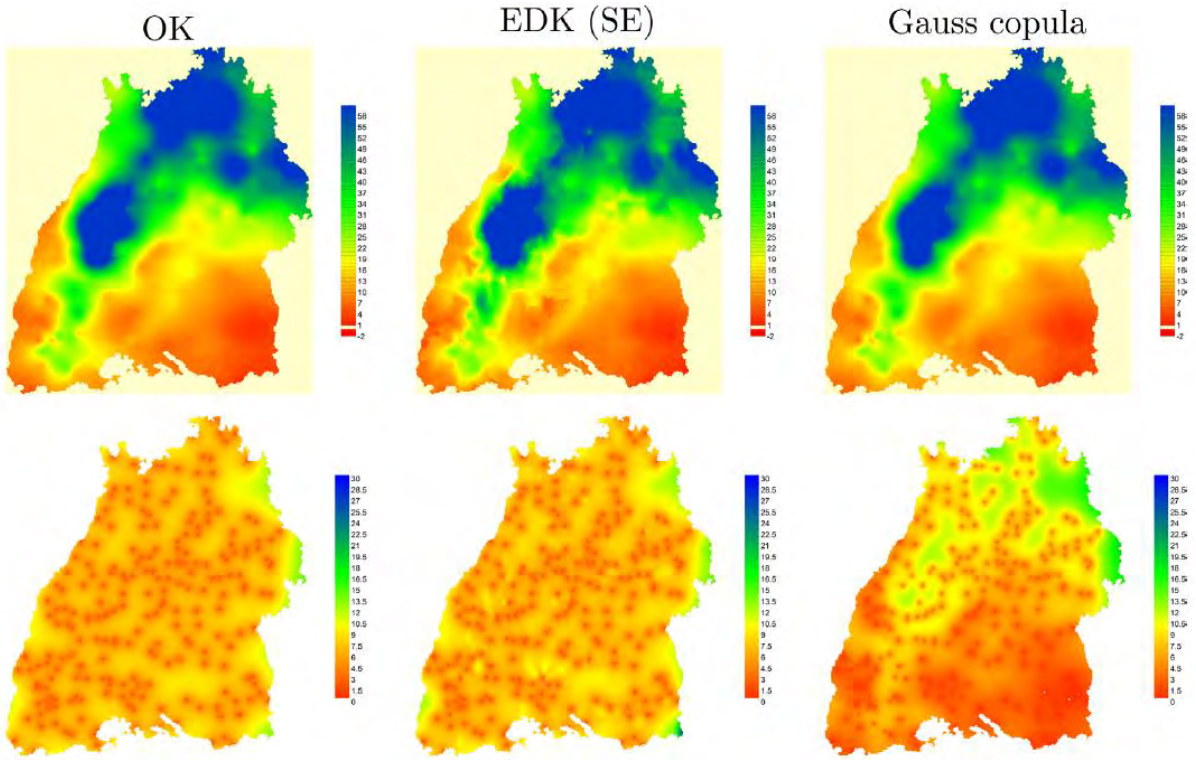


Figure 6.3. Results of a day-by-day interpolation: Baden-Würtemburg, December 19, 1993. The upper row shows the estimated mean field and the second row shows maps of the standard deviation of the interpolated values, as in Figure 6.2. The scale of rainfall has a maximum of 60 mm; the standard deviations a maximum of 30 mm.

In Figure 6.2, in all three images of rainfall interpolation in the upper row, there is not much difference between them except that the EDK (SE) which is using Smoothed Elevation as an exogenous variable, tends to follow the terrain more than the others. Note the spatially invariant error structures of the Kriging methods, independent of the amount of rainfall. The red points are the sites of the gauges and the error structure depends only on separation distance – the sparser regions show more green. On the other hand the standard error of the Copula-based interpolation varies not only with sparseness, but also with the rainfall *amount*. Where it is dry [Southeast of the region where there is low precipitation] the error is low and vice versa in the western region.

Figure 6.3 treats the same region on the following day. The same remarks apply to the rainfield as to the previous day in Figure 6.2; the wetter parts are less precisely estimated by the copula, but the Kriged fields have almost no change to their error structure compared to the patterns on the previous day. It is this ability to discriminate that makes the copula infilling and interpolation more truthfully attractive than standard Kriging methods.

In Figure 6.4, we summarise in an image the relevant information in Table 7 of Bardossy and Pegram, (2013). To summarise the results of the comparisons of the various methods' abilities to get the interpolations right in terms of their cumulative frequencies, Figure 6.4 makes it clear that the Gauss copula procedure proposed here is substantially superior to Kriging methods for infilling and interpolating rainfall values in periods of days, pentads, months and years, in that they achieve scores of around 80% when aiming for 95%,

considerably better than the Kriging methods, with or without External Drift, which can only achieve about 50%.

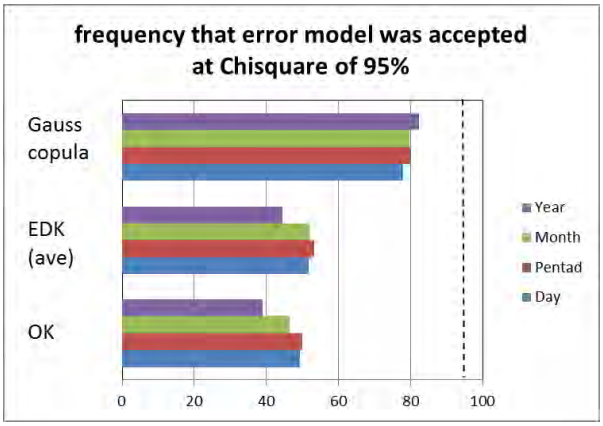


Figure 6.4. Histogram of the frequencies that the various error models were acceptable at the 95% confidence level, based analyses similar to those depicted in Figure 4.8

Now that we have defined the core methodology, we need to look at the data-sets and determine how to manage them

6.3 The new art of interpolation applied

A region on the Eastern Seaboard, shown in Figure 6.2 below was chosen for (i) testing the spatial interpolation procedure and (ii) assessing the effect of including altitude as an exogenous variable; it is a 3° square (approximately 300 km across). We have picked out the gauge sites in small red filled circles and indicated the altitude by colour (sea level is black and the highest mountain peak in Lesotho is white). This domain is to be used to demonstrate our experiments. The coordinates of the top left corner in degrees are (28°E, 28°S) and of the bottom right corner (31°E, 31°S). The sparse gauges in Lesotho are clearly seen in the middle left panel, which we will omit from our analysis. We number the panels 1 to 9 from bottom left to top right. The Umgeni catchment can be discerned in the right-most block in the middle row, so that Durban is just off the image. The orange rectangle is the area used for interpolation experiments.

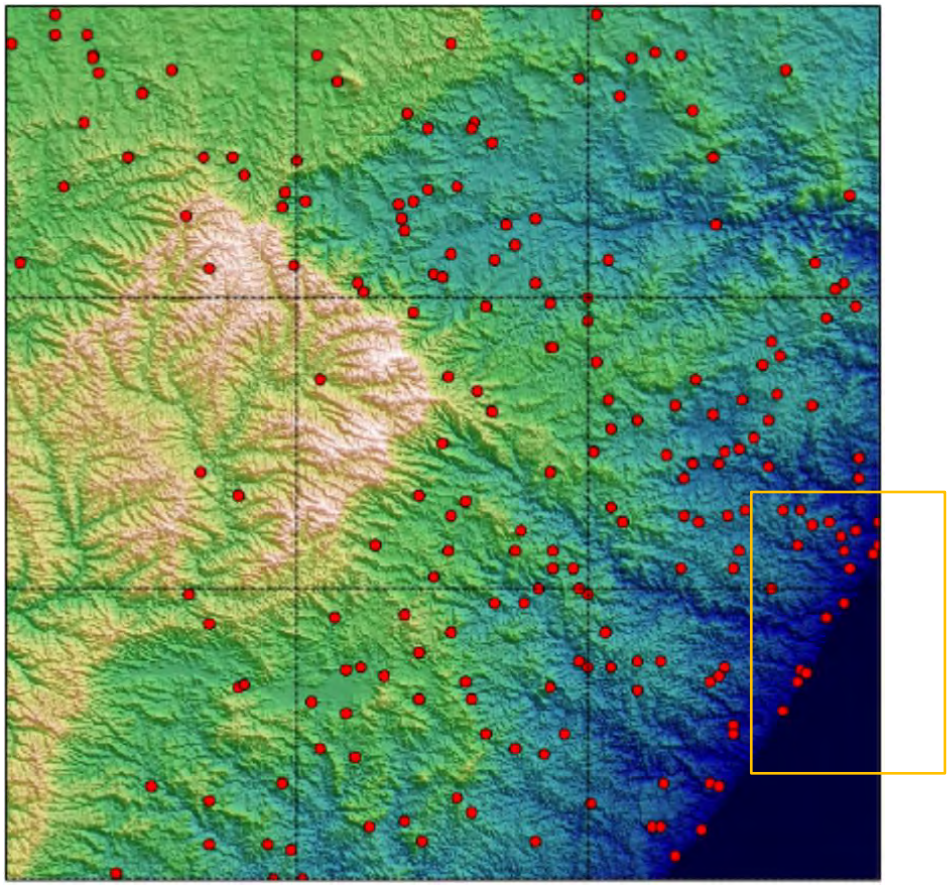


Figure 6.5. The gauges used for interpolation trials are shown against an elevation backdrop. The image, with coordinates at the top left corner (28°E, 28°S) and of the bottom right corner (31°E, 31°S), shows all gauges available in the CSAG database during the 1965-1985 time period. The orange rectangle is the area used for interpolation experiments.

The first step in the spatial interpolation procedure is to choose a day and determine the available gauge readings on that day. All the observations are assembled for Gaussianisation to (i) fit an empirical cumulative distribution function (cdf) to the data then (ii) fit a smoothed estimated probability distribution to (i). We will demonstrate this procedure on a day when there are no missing data.

In Figure 6.6 is a frequency distribution function of the recorded rainfalls on a certain day over a part of the region. Notice the awkward kink below the 1 mm value, ringed in red. We do not believe that recordings below 1 mm per day are accurate, so fit this range with an exponential function. Values above 1 mm are approximated by a Kernel Density Estimation procedure using Gaussian kernels and the hybrid cumulative distribution function (cdf) model is shown in Figure 6.7, a repeat of Figure 5.14.

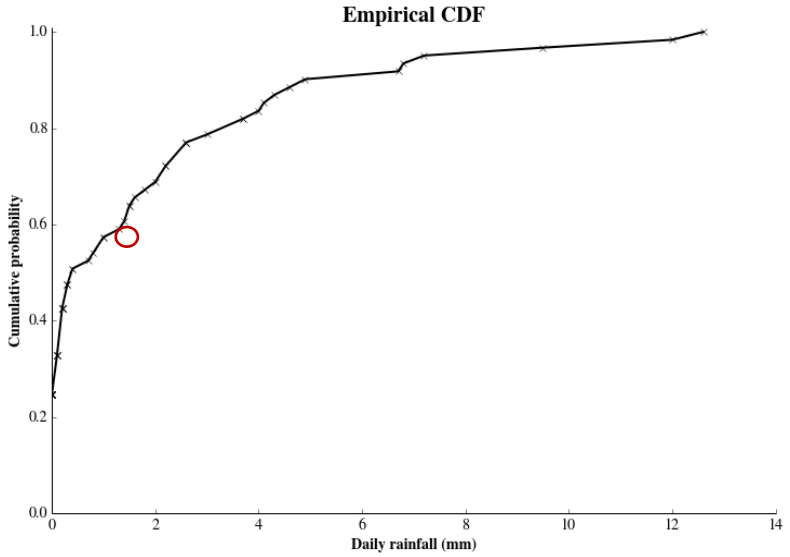


Figure 6.6. An empirical cumulative distribution function (cdf) obtained by ranking all observed rainfall records on a chosen day. Note the dry probability p_0 of 0.23, which indicates that this particular day is quite wet.

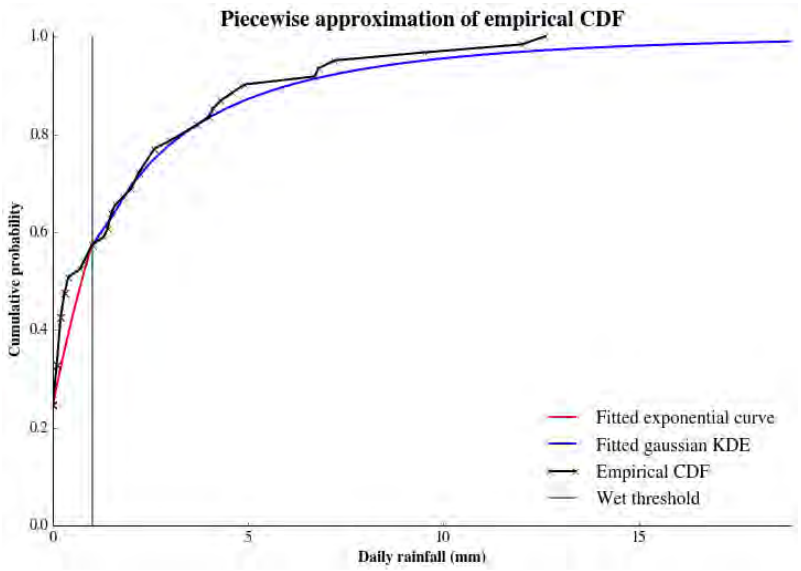


Figure 6.7. The combined piece-wise approximation of the empirical cdf in the rainfall domain, superimposed on the empirical cdf in Figure 6.6. The approximation is used to transform simulated fields in the Gaussian domain to rainfall on the given day.

6.4 Interpolation using missing data, covariates and TRMM: Methodology

This section outlines the methodology used to obtain the results in the sections that follow; the heavier mathematics have been omitted in an attempt to improve its readability.

6.4.1 Use of infilled data

As explained in the discussion leading to Figure 6.1, the use of infilled expected values in the gauges with missing values, without taking into account their precision of estimation, is not very informative and is likely to increase the spatial correlation spuriously. These estimated

expected values are often given the status of an observation by modellers, neglecting their uncertainty. The distribution of the expected infilled values is different from the distribution of the observations; they tend to have a positive value, which induces an incorrect proportion of wet locations, a reduced variance and a concomitant increase of spatial correlation of the set of infilled values. The formula for the marginal distribution used in Figure 6.7 is:

$$\begin{aligned} F_{y_i}(z) &= p_{y_i} && if\ z = 0 \\ &= p_{y_i} + \left(1 - e^{-z\theta_{y_i}}\right)\left(q_{y_i} - p_{y_i}\right) && if\ 0 < z \leq \delta \\ &= q_{y_i} + \left(1 - q_{y_i}\right) G_{y_i}(z) && if\ z > \delta \end{aligned} \qquad (6.1)$$

where

y_i is the location of a gauge
p_{y_i} is the probability of dryness on the day
q_{y_i} is the probability of the threshold δ not being exceeded
δ is a selected wet threshold (1 or 2 mm)
z is the rainfall amount in mm
θ_{y_i} is the exponent spread constant for the threshold data
$G_{y_i}(z)$ is the fitted Gaussian Kernel Distribution Estimate (KDE)

The dependence structure of precipitation for the given time interval and for any set of locations $\{x'_1, \dots, x'_l\}$ is described with the help of its copula leading to the multivariate distribution:

$$F_{x'_1,\dots,x'_m}(z_1, \dots, z_l) = C_{x'_1,\dots,x'_m}\left(F_{x'_1}(z_1), \dots, F_{x'_l}(z_l), \theta\right) \qquad (6.2)$$

where θ represents the parameters of the copula (correlation function parameters for the Gaussian copula).

In order to simulate or interpolate precipitation on a given time interval the following procedure is adopted:

1. Missing gauged precipitation values are first simulated using the marginals defined in (6.1) using the spatial copula introduced in Section 3.2.
2. The simulated infilled values and the observed gauge values are merged to an observation set
3. A spatial field of precipitation is simulated using the spatial copula conditioned on the extended observation dataset.
4. Steps 1-3 are repeated *K* times.

The procedure produces *K* simulated spatial precipitation fields. The following properties hold:

- For all fields the observed precipitation is reproduced at each observation location
- For all gauge points with infilled data the distribution is given by Eq. (6.1)
- The dependence of the spatial field follows the copula defined above.

6.4.2 The use of covariates

Covariates can be used with either local (neighbourhood) or global correlations. If the correlation to the normalized covariate is relatively small, then the correlation conditions can be directly included into the simulation. Higher correlations require a two-step procedure. There are two cases:

- The covariate is available for the whole domain *D* (for example, topography)
- The covariate is available at a certain number of locations – partly different from the precipitation observations – and/or in the form of block values (for example, TRMM).

In case one the fields used for conditioning are obtained by

$$X = rY_c + \sqrt{1 - r^2}\, U \qquad (6.3)$$

where Y_c is the Gauss transformed topography (or temperature over the domain) and *U* is the unconditional simulated rainfield.

In case two, formula (6.3) can be extended to an arbitrary number of covariates. A simultaneous simulation of the variable of interest and the covariates can be performed if local conditions apply to the covariates. This requires a good parameterisation of the Gauss-transformed cross correlation functions.

6.4.3 The use of TRMM (or altitude) as an exogenous variable

Spatial similarity
One might assume that the precipitation patterns in a given time interval estimated by TRMM are correlated to the patterns of the true rainfall. This means that for the block averaged precipitation and the TRMM estimates, the correlation is significantly greater than zero. Under these circumstances one can simulate spatial fields which combine point observations on the ground and are correlated to TRMM according to the estimated correlation.

If the correlation is relatively small $r < 0{:}4$ simulation can be performed directly with the simulation procedure (which is used for all simulations). For higher correlations a mixed procedure is needed:

1. Gauss transform the TRMM data with corrected block variance (according to support)
2. Simulate a spatial field *Y* conditioned on the above transformed block values using the spatial copula.
3. Add the above simulated *Y* field to the unconditional fields *X* used as a basis for conditioning with an appropriate weight
4. Simulate the spatial field of precipitation Z^k in the observation dataset and the above simulated *X* fields.

Steps 1-4 are repeated *K* times.

The procedure produces *K* simulated precipitation fields. The following properties hold:

- For all fields the observed precipitation is reproduced at each observation location

- For the simulated blocks the correlation with the TRMM values is the prescribed *r* (in the Gauss domain) from equation (6.3).
- The dependence of the spatial field follows the copula defined above

In order to have a good match between the patterns an *a priori* bias removal through a block by block QQ transform is beneficial. For this an interpolation of the block or point precipitation distributions (over time) can be required, due to the fact that some of the blocks do not have any ground observations. Furthermore, discharge can also be used for this purpose to obtain statistics at larger time scales.

Although a potentially useful suggestion, this methodology was not used in this project because the correlation between rainfall and TRMM (and altitude) at the daily scale is so low, as is shown in Section 6.6 of this chapter and in Chapter 8. A remedy is suggested in Chapter 9.

6.5 Interpolation over tiles by stitching, a realisation of the algorithms of Section 6.3

We want to be able to interpolate rainfall between gauges anywhere in the country. Because of the variability of rainfall and altitude (let alone seasonality) over the region, the spatial interpolation of the gauge rainfall information on any day should be done in a localised manner. This section demonstrates how we devised a means of interpolation over 1° squares at 0.01° precision, conditioned on those already interpolated, so that there is no evidence of discontinuity at the boundaries, which should be invisible in the final product. The method is rather like patchwork quilting, but ensuring that the 'pattern' is not discontinuous at the edges joining the panels/tiles, a choice of terminology which explains the section heading. In other words, we need to ensure statistical continuity over the lines of separation.

In Figure 6.8 we show two 1° square blocks, staggered North-South by 0.333°, for one day over region 6, the rainfall stations given in Figure 4.13. We have performed two independent sets of 100 spatial simulations over each square and averaged the sets.

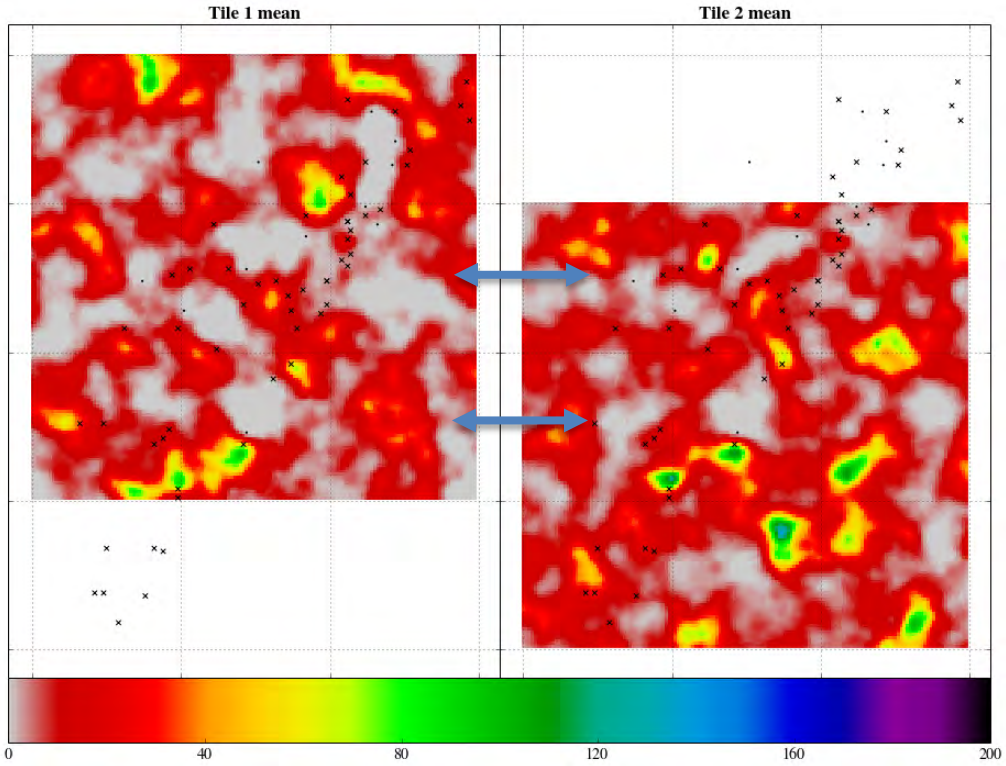


Figure 6.8. Averages of 100 sets of simulations of one day over a large area (1° – about 100 km), the 0.333° tiles in the two images covering overlapping regions. Common strips of the three central tiles are indicated by grey arrows. Note the rainfall stations in the background are to be found in Figure 4.13. Here, those experiencing rain on the day are shown as crosses and the dry stations as black dots in the dry grey areas. Each of these images is the spatial mean of two different sets of 100 simulations.

Careful inter-comparison shows that there are differences, particularly in the dry grey areas. In this case there is no stitching of the 0.333° blocks within either of the 1° regions. Each has been divided into 9 tiles taken together, discernible by vertical and horizontal dividing lines. The central tile in each panel is clipped out and assembled one above the other in the left panel of Figure 6.9.

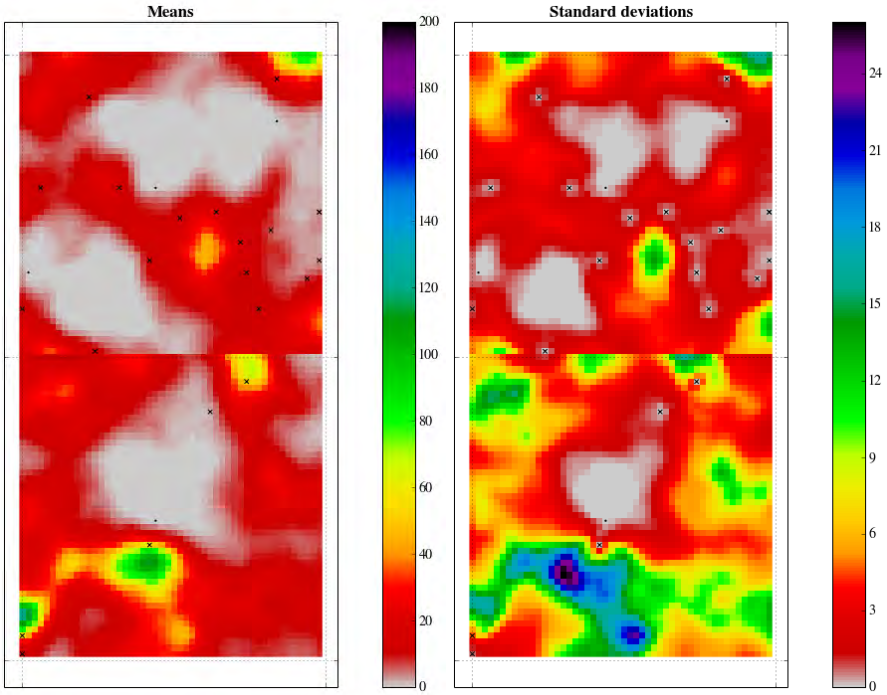


Figure 6.9. The central 0.333° tile of each panel in Figure 6.8 juxtaposed in their correct positions on the left and their standard deviations in the right panels of this figure. As noted in Figure 6.8's caption, these were assembled from 100 simulations. There is no stitching between the stacked tiles.

It is evident from Figure 6.9 that there is serious discontinuity at the common edge of the upper and lower tiles, much more so in the right panel showing the standard deviations over the field. This is likely caused by the fact that the lower tiles of the pair possess only 4 gauges and partly covers the Indian Ocean, so there is less restraint (due to sparser data) in the lower tile than in the upper. Figure 6.10 shows the efficacy of conditioning the lower tile on the simulation of the upper tile. There is now no discernible edge.

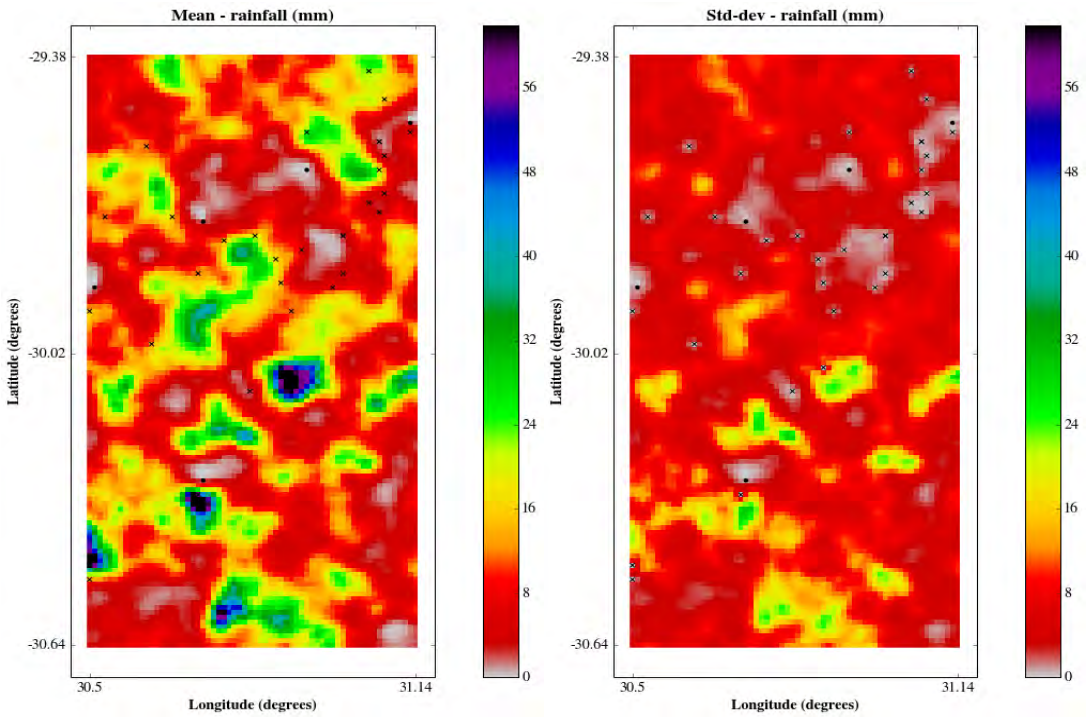


Figure 6.10. The pair of tiles in Figure 6.6 stitched together in 100 simulations: means and standard deviations of the resulting fields are shown in the left and right panels.

Figure 6.11 completes the set. The image is a single realisation of an infilled region of 4 connected tiles, spanning 1.25° square. The apparently seamless interpolation was achieved by starting with one tile and then successively stitching the remaining tiles together one-by-one. This method allows statistical variability over the macro region by using a parsimonious simulation methodology.

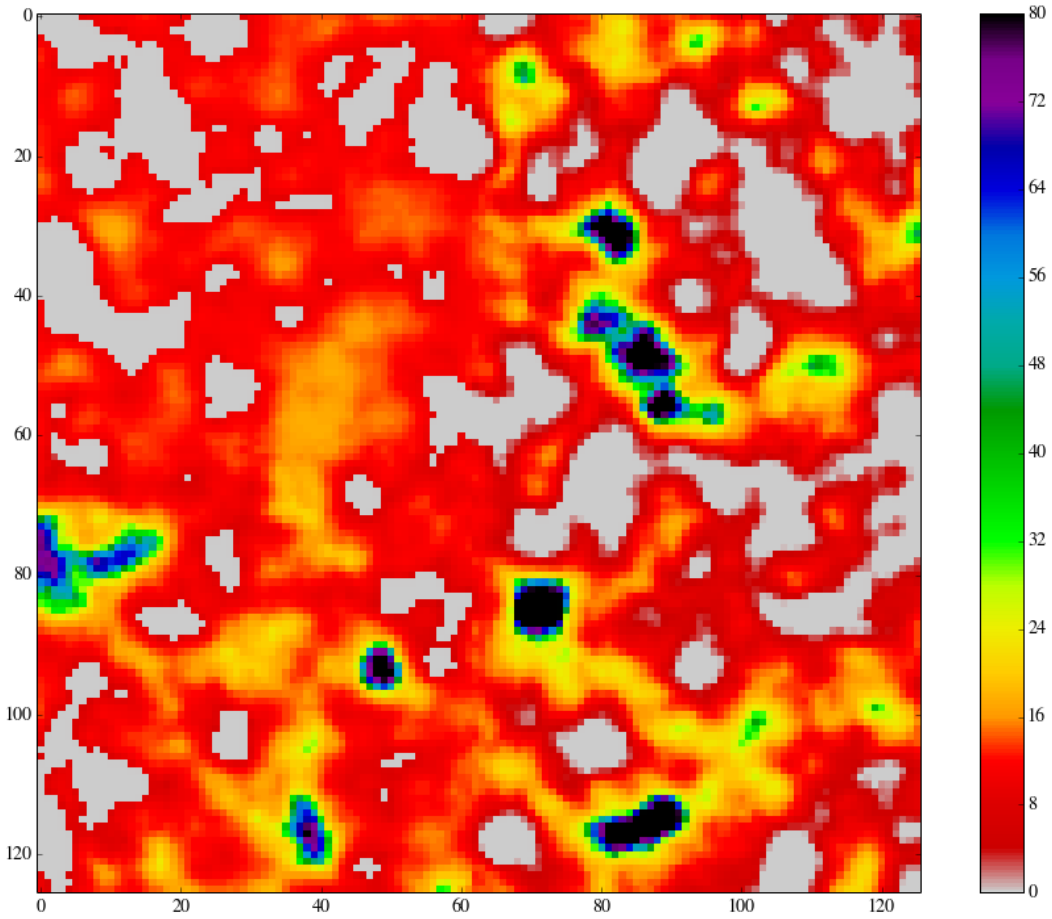


Figure 6.11. A 1.25° square region consisting of four tiles – a single conditioned realisation stitched together in sequence

The conclusion we can draw from the work of this section is that we have devised a meaningful, informative and relatively parsimonious method of generating ensembles of interpolated rainfields, potentially useful for hydrological applications, besides giving estimates of the spread of the interpolated fields.

6.6 Using exogenous variables to improve the interpolation product

We turn now to the determination of the links between rainfall and altitude and start with sample correlations between altitude and rainfall, day-by-day over the region.

Figure 6.12 presents the correlation coefficients estimated between gauged daily rainfall and altitude on each of the 7305 days of record over the 3° square of Figure 6.5. The range is approximately -0.5 to 0.4 over the years and there is evidence of strong annual variability.

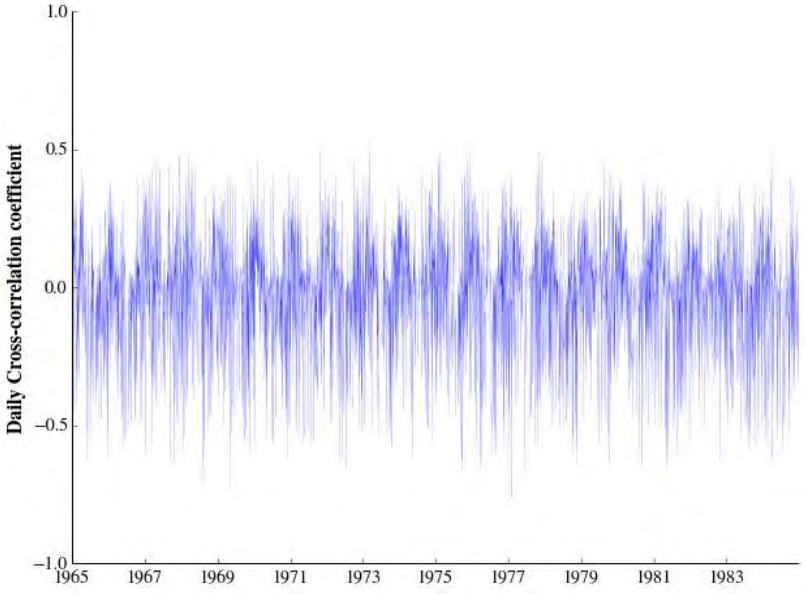


Figure. 6.12. A time-series of the cross-correlation coefficients computed between observed daily rainfall and elevation over region 6 for the analysis period. The station elevation reported in the CSAG database was used.

To get a better feeling for intra-annual variability each of the above data values were plotted in Figure 6.13 as a function of the day of the year and an average over the 365 days of the year superimposed. There is evidence of a weak variation of the cccs in this figure.

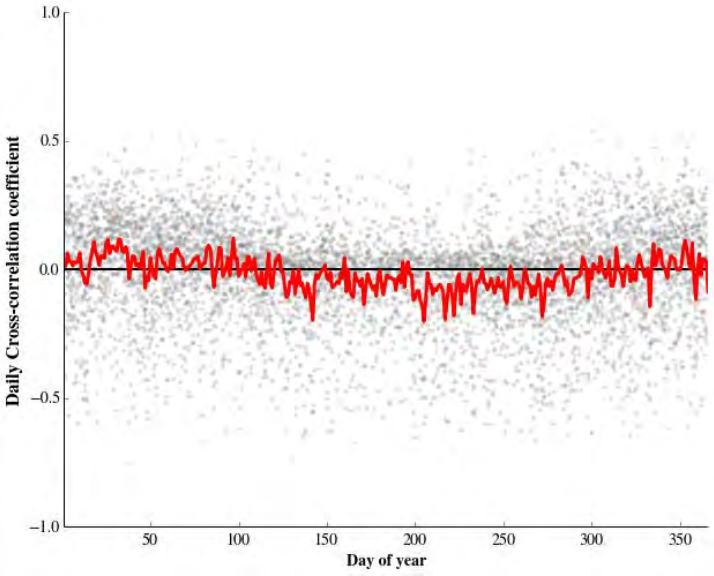
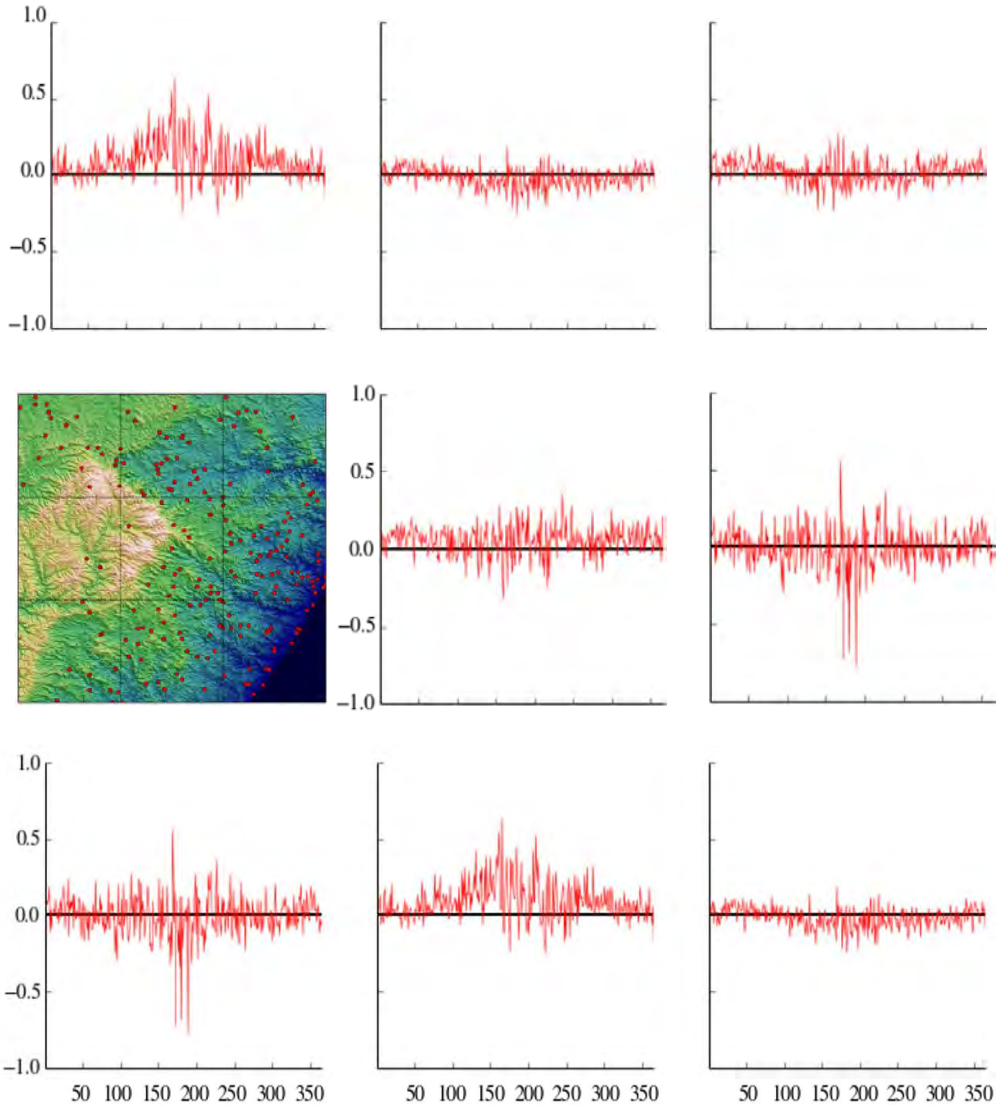


Figure. 6.13. An attempt to discern if there is any seasonality in the correlations. All correlations over the 3° square region in Figure 6.5 have been binned according to the day of the year (see grey scatter points) and the mean for each day computed (red line)

We decided to see if there was any spatial variability of the results, smoothed over the year by averaging all regions in Figure 6.13. Figure 6.14 shows the unpacked ccc values for 8 of

the 9° blocks of Figure 6.5 (all blocks except # 4, the left middle block of the figure covering part of Lesotho which contains only 2 gauges).



Numbering is from bottom left to top right as in Figure 6.2, omitting 4 – see panel below.

7	8	9
	5	6
1	2	3

Figure. 6.14. Day of year means (on each day over the 20 years) for eight of the nine 1° blocks shown in Figure 6.5 (repeated here as an insert in place of block 4). Compared with Figure 6.13, these blocks have been unpacked. Note that block 4 of Figure 6.5 has not been included due to a lack of sufficient observations (only 2).

It is evident from Figure 6.14 that the coastal and mountainous regions have different correlation linkages between rainfall and altitude over the year. The strongest correlations

are about 0.4 during the dry period of the year and there are some negative ones of about -0.2 (going down to -0.7 in block 1, the lower left degree block on mountainous slopes) in the winter. Also evident is the distinct variability of the correlations between the blocks, justifying the tessellation of the regions into smaller tiles within which to assume spatial homogeneity, to determine both gauge-gauge and gauge-altitude cross correlations, where meaningful. Overall, the correlation of altitude and raining days (not dry ones) is disappointingly low and will not be used in the sequel; location is more relevant.

6.7 Examples of spatial infilling by simulation

In this section we address the problem of deciding how many simulations are good enough to obtain an informative set of individual interpolations such as that displayed in Figure 6.11.

Investigate the stability of unobserved rainfall distributions

The conditional simulation procedure provides an estimate of the rainfall distribution at unobserved locations. A "sufficient" number of realizations is required to provide a good estimate of the unknown distributions. An attempt was made to find out how many realizations are "sufficient" (*K* in the algorithm passage of subsection 6.3.1).

The region chosen was contained in the orange rectangle at the lower right of Figure 6.5, which includes the coast-line near Durban. In this part of the investigation, we did not tile the region, as only 2 tiles are needed to make the point; we simulated over the subregion as an un-partitioned unit.

The variability between the median fields of the two simulated interpolations shown in Figure 6.15 is large, especially in sparsely gauged regions when there are few members of the ensemble. The large difference is mostly due to the limit of 10 simulations used to provide the median surfaces. We have started the comparisons with a sample of a small number of simulations on a day (which was chosen because it had a reasonable spread of rainfall) and increase the number progressively in the figures following this one. It will be seen in the sequence of figures that the differences between the two sets of quantiles (and means) of the multiple surfaces disappear as the number of simulations increases, until they look like smoothed Kriged fields.

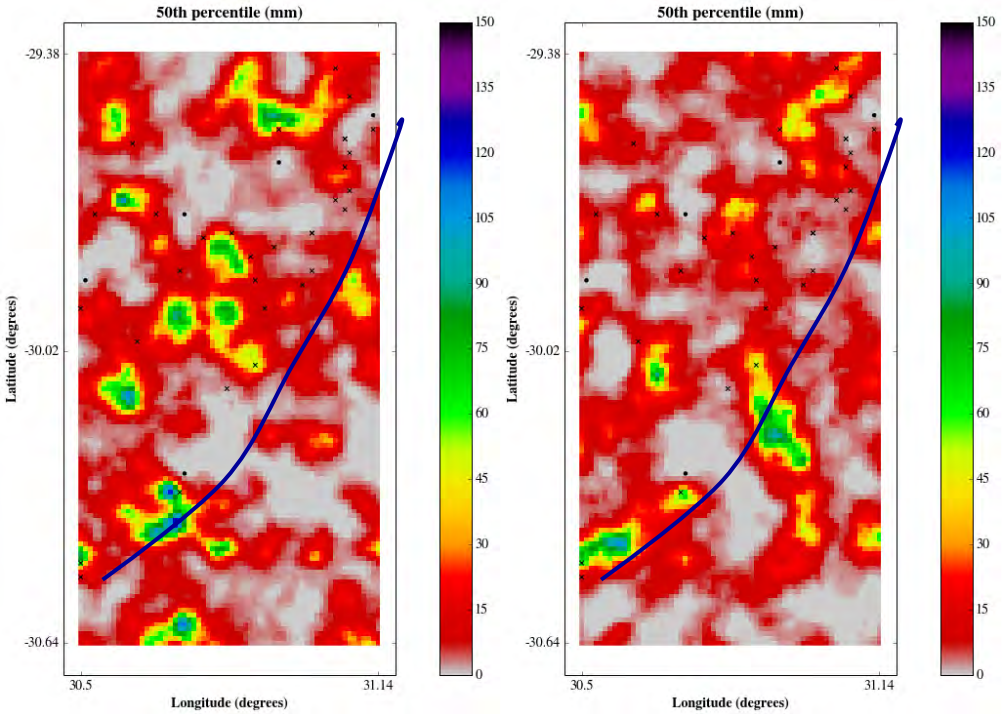


Figure 6.15. Comparison of the 50th percentile (median) at each 0.01° (about 1 km) square pixel based on two different conditional simulation runs of 10 realisations each. It is clear that the spatial patterns are very dissimilar, especially in the bottom right which contains no gauges, as it is over the sea – the coastline is shown by the blue curve. The dry gauges (dots) are surrounded by grey areas; the wet gauges are indicated by crosses.

In Figures 6.15 through 6.17, the small black crosses indicate wet rain-gauges on the day, whereas the black dots are the locations of gauges which were dry on the day, the latter being surrounded by grey areas signifying zero error. Figure 6.16 shows the result of 100 and Figure 6.17 1000 runs.

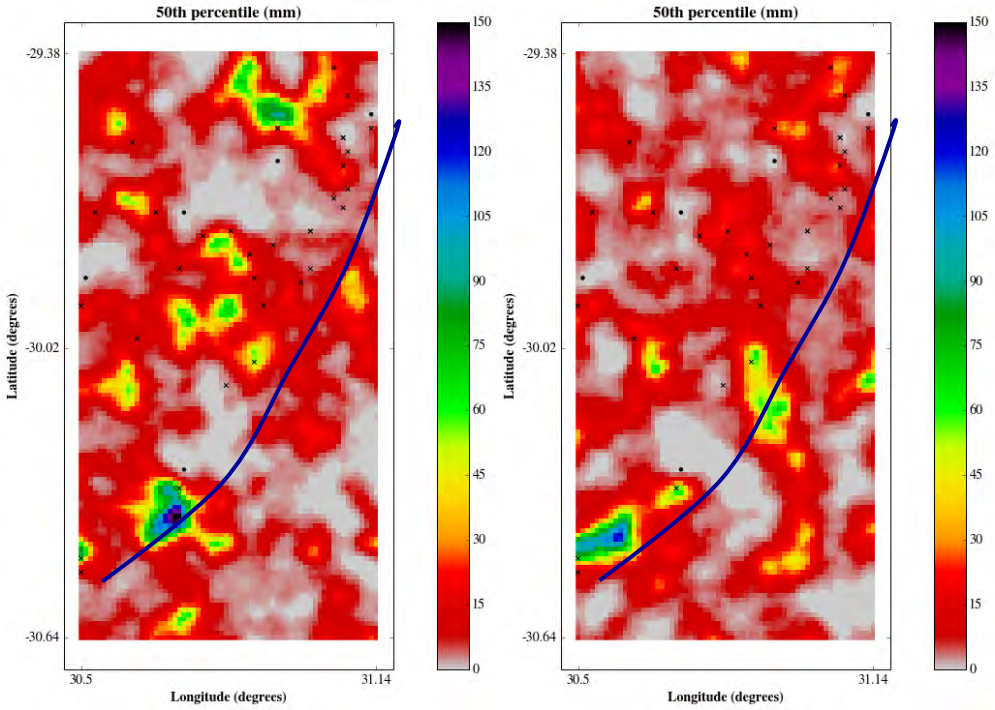


Figure 6.16. Comparison of the 50th percentile (median) at each 1 km square pixel based on two different conditional simulation runs of 100 realisations each. Note that the spatial patterns over land are less dissimilar, except for the sparsely gauged northwest region.

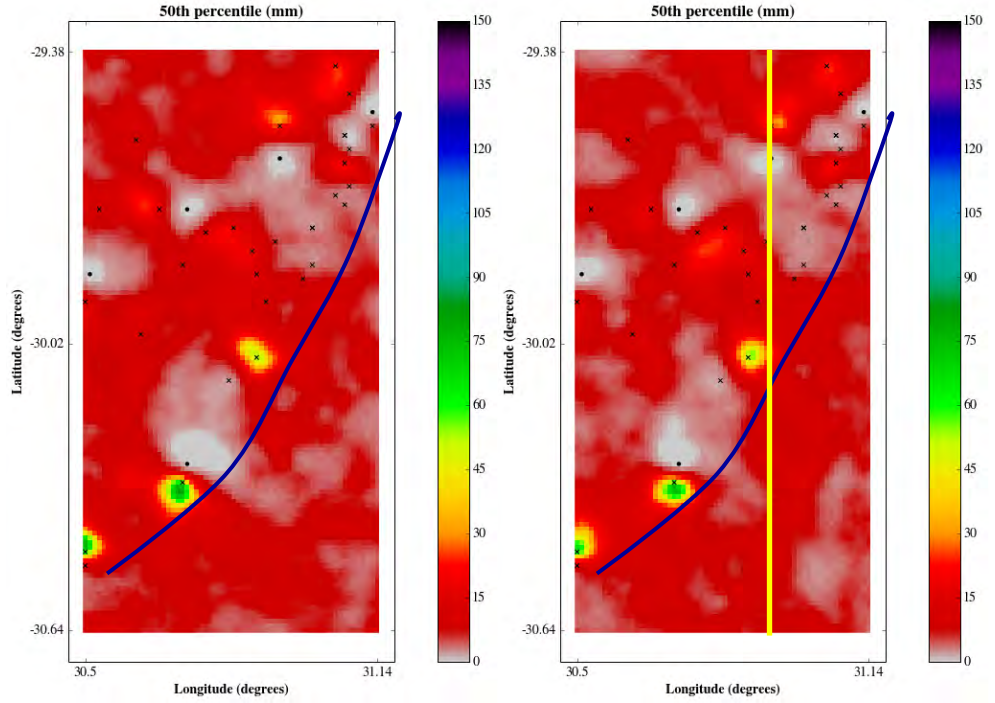


Figure 6.17. Comparison of the 50th percentile (median) at each 1 km square pixel based on two different conditional simulation runs of 1000 realisations each. Over land the images converge nicely except for the sparsely gauged regions. The yellow line indicates a transect through two gauges, which yields the 1 dimensional plots to follow in Figures 6.18 & 19.

In Figure 6.18, we show some transects through the sets of simulations to show the variation in the quantiles between samples. These 'slices' follow the yellow line which runs through 2 gauges as shown in the right panel of Figure 6.17.

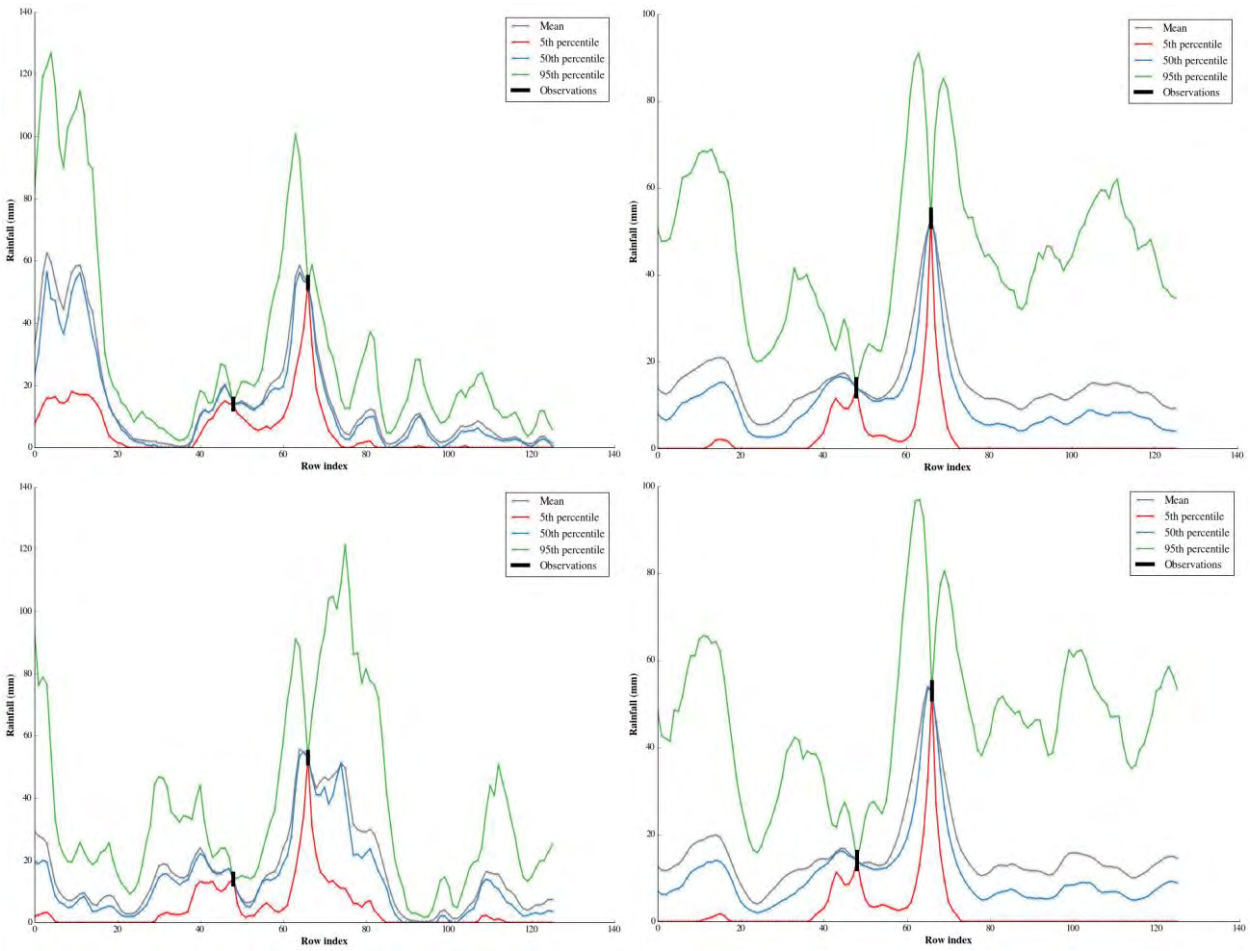


Figure 6.18. Transects through the two independent stacks of 100 (left column) and 1000 (right column) simulation images whose medians are shown in Figures 6.15 and 6.16, intersecting two gauge observations. The 4 different trajectories in each panel are the 5^{th} and 95^{th} percentiles and the median and mean.

The upper and lower images on the left of Figure 6.18 (100 simulations) are dissimilar in shape. After 1000 realisations they converge quite well, as shown in the right column. It takes an ensemble of 2500 simulated images to provide almost perfect convergence, nevertheless a modest number of simulations yields a fair measure of uncertainty. Note that the mean trace (grey) is above the median (blue) because of the skewness on the simulated rainfalls. Note also that there is no error at the gauge locations and that, far from the gauges, the 90 percentile range increases dramatically, indicating the imprecision of the estimates where networks are sparse.

6.8 Conditional Simulations with Altitude as an exogenous variable

In the next two sets of images in Figures 6.19 and 6.20, we demonstrate the effect of using altitude as an exogenous variable, which was not included in the above simulations. The

area is the coastal zone in the middle right panel, number 6, in Figures 6.5 and 6.14, and is shown in detail in Figure 6.21.

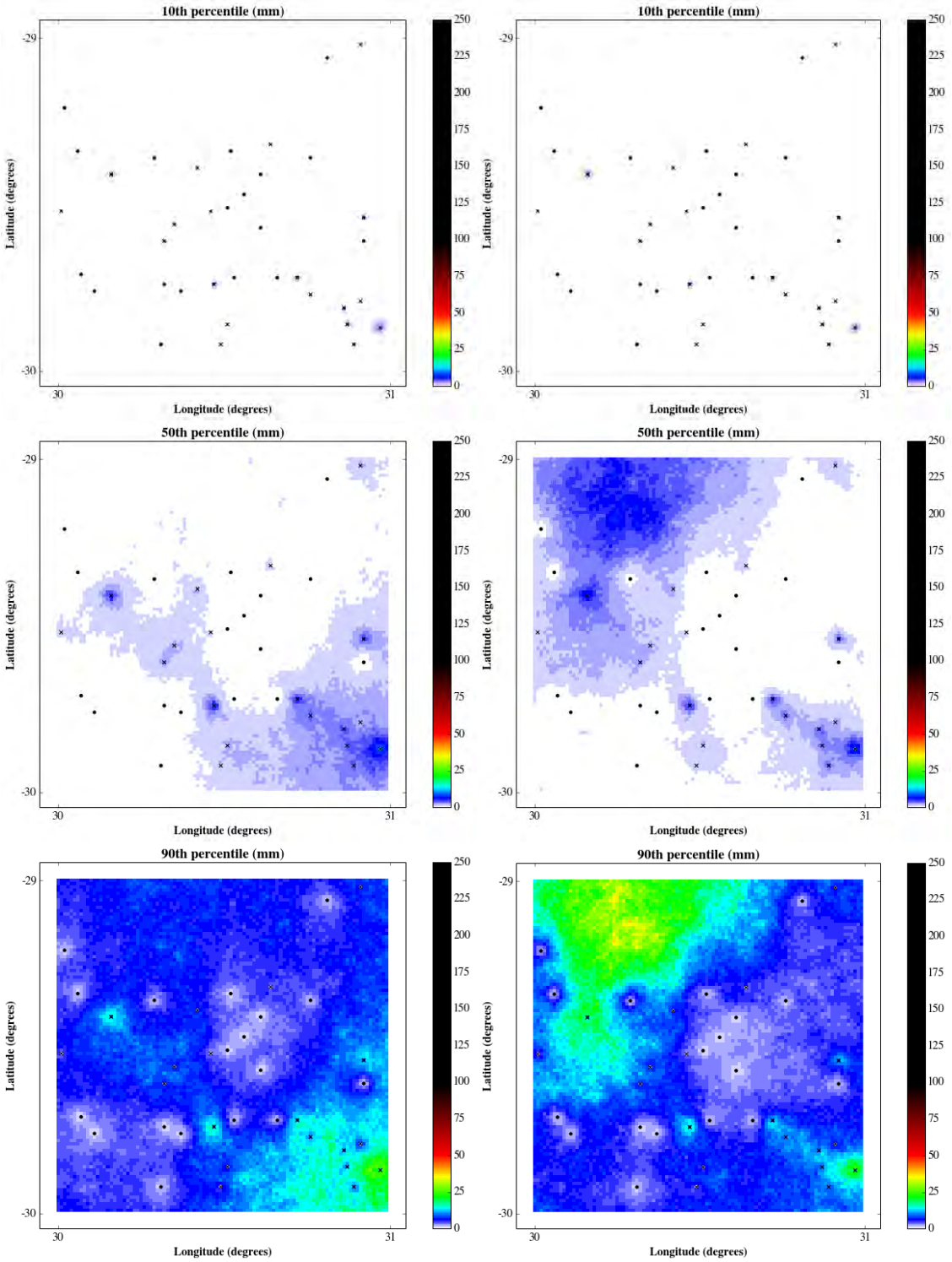


Figure 6.19. The effect of altitude. Top, middle and bottom rows of this figure are respectively 10th, 50th and 90th percentiles of 1000 simulations on block 6 (middle row right hand side of Figure 6.14). In the left column of this figure, correlation with altitude is not included in the simulation constraints. In the right column there is a 0.2 correlation between rainfall on the day and the altitude.

In the images above in Figure 6.19 and below in Figure 6.20, the influence of imposing a correlation between altitude and rainfall can be seen. Note that for zero correlation and median surface, in the middle panel on the left of Figure 6.19, the rain at the wet gauges is near the coast and is absent at the gauges in the high ground to the Northwest. However, even with a weak correlation of 0.2, rain appears where the gauges are dry in the higher ground, as is clear from the middle right panel in Figure 6.19. The rainfall intensity grows there as the correlation increases, so the altitude has a false influence on the spatial distribution of rainfall as it increases. All scales are the same in these images for ease of comparison.

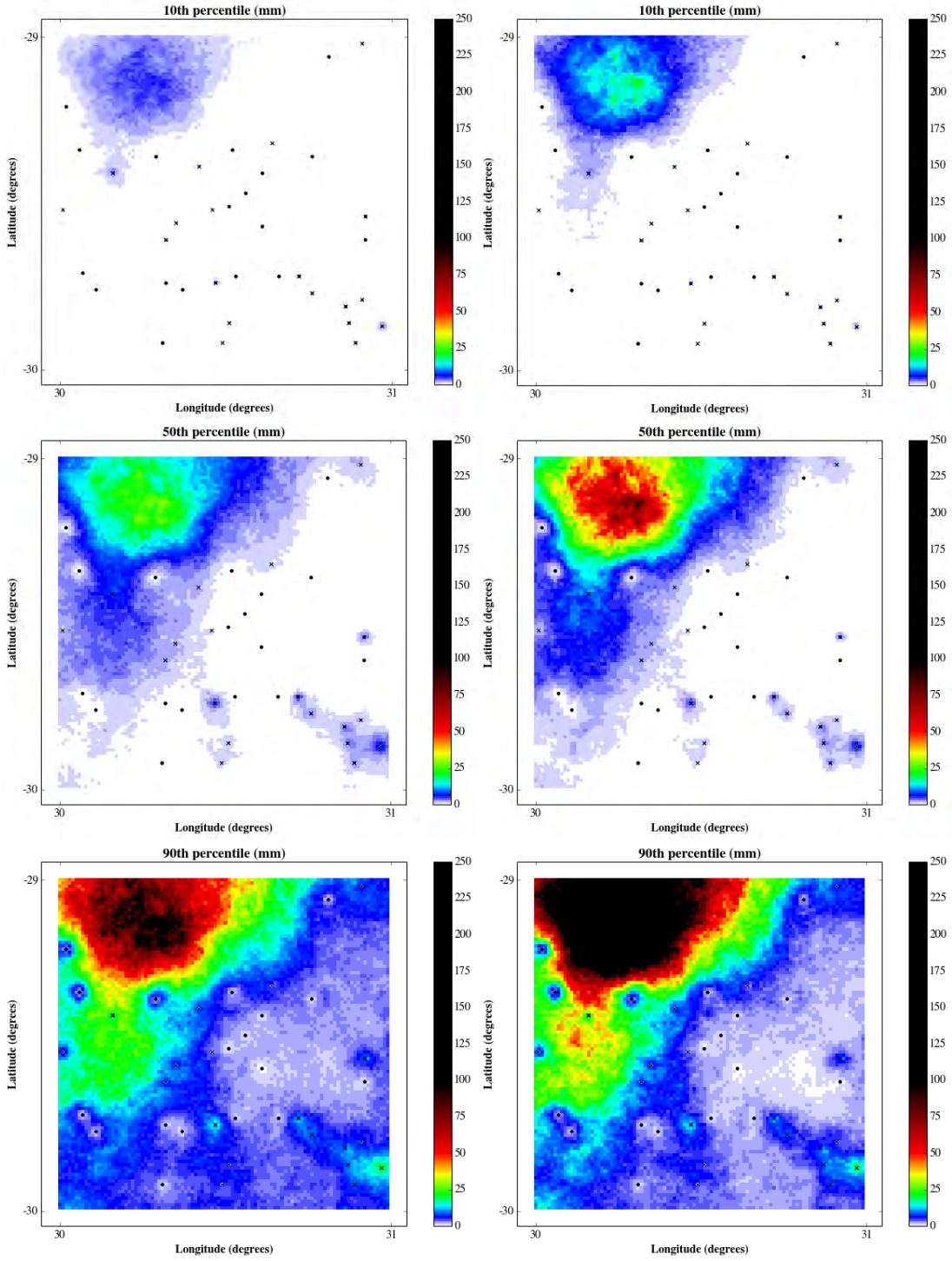


Figure 6.20. The effect of altitude. 10th, 50th and 90th percentiles of 1000 simulations on block 6 (middle row right hand side). Left column 0.5 correlation; right column 0.75 correlation between rainfall on the day and altitude, the latter shown in Figure 6.19.

The effect on the spatial distribution of rainfall due to high correlations between rainfall and altitude are clearly demonstrated in Figure 6.20, to an absurd degree as even the 10th percentile (upper panels) shows exaggerated rainfalls. The result is that we opted to ignore this spurious linkage in our spatial interpolations between gauges.

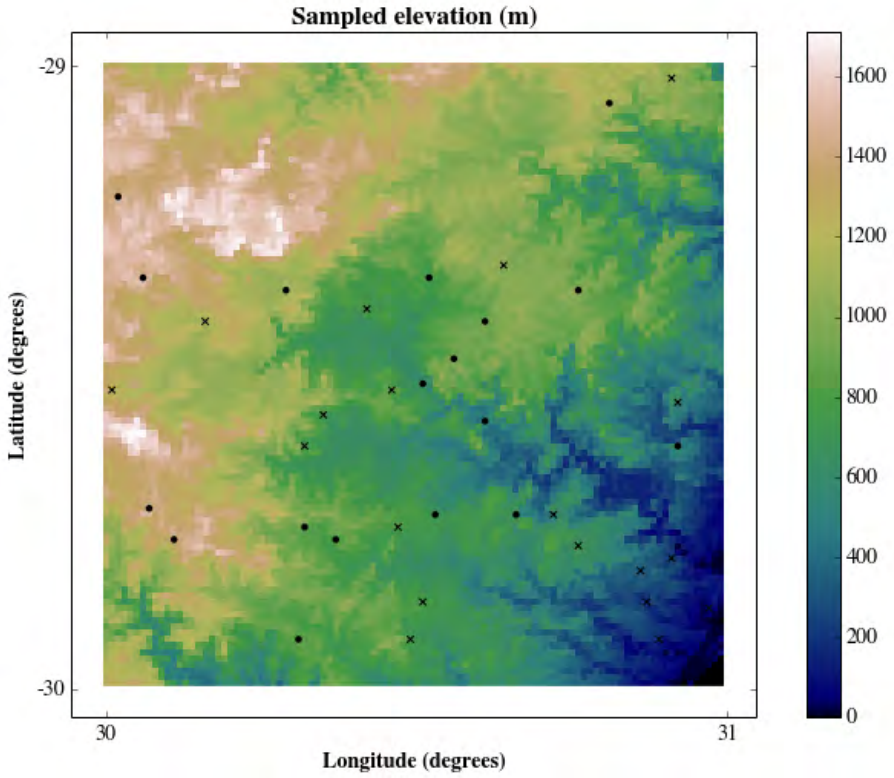


Figure 6.21. Elevation map sampled from the product of the Shuttle Radar Topography Mission – Jet Propulsion Laboratory. This map was used in the above analysis summarised in Figures 6.19 and 6.20. Note the dry gauges on the day are marked by black dots and gauges recording rain are marked by crosses.

We conclude this section noting that the relatively weak correlations of the observations with altitude as shown in Figures 6.13 and 6.14 might have some value in the interpolation. Nevertheless we are worried that using altitude as a surrogate variable might warp the rainfield untruthfully as demonstrated in Figures 6.19 and 6.20. To take account of mountainous areas, in the final interpolation of the space between the gauges, we will shorten the correlation length in rugged terrain but leave it longer in the flatter areas.

Chapter 7. Spatial Interpolation using simulated radar-like random fields

The commonly used technique used for spatial interpolation is Kriging in its many guises and is a precursor to the spatial simulation methodology introduced in Chapter 6. Spatial simulation is valuable where the rainfall characteristics are changing rapidly in space, however the method requires care and a relatively deep knowledge of the mathematics to enable a good estimate to be made. As an alternative, where spatial variability of the characteristics of rainfall is small, it is tempting to use Kriging, but with the condition that all the rainfall variables be Gaussianised as described in Chapters 3 and 4. Kriging offers a field of uncertainty described by the Kriging variance, which is not useful for raw daily data unless a Gaussianising transform is first performed because of the skewness of the data. Annual and perhaps monthly data can be Kriged directly because they are less skew than daily data, nevertheless Gaussianisation as introduced in Chapter 3 would be prudent in all cases.

In this chapter, a novel way of interpolating between raingauges on any day over a large area is introduced. The key to the method is to use Gaussian random fields, modelled on Gaussianised radar images. An ensemble of these fields provides not only a measure of uncertainty [median and quartiles at each location in the area] but also plausible rainfields for rainfall-runoff calculations through catchment modelling. The idea can be extended to areal rainfall simulation, by using a daily rainfall network model like PEGRAIN (Pegram, 2010) and merging random fields with the gauge values using the correct correlation structure for the field, as introduced by Sinclair and Pegram (2005). This correlation structure can be obtained by analysing many daily accumulations of measured radar fields. The advantage of this hybrid model is that plausible spatial daily rainfields [as big as desired, but typically 200 km across – say 40 000 sq km and as many fields as required] can be generated where there is reasonable spatial stationarity.

The computations and images in this section were modified and subsequently published in Gyasi-Agyei and Pegram (2014). The region treated here is in the Free State; and the set of gauges taken from WRC project K5/1964 [Pegram et al. (2013)], shown in Figure 7.1.

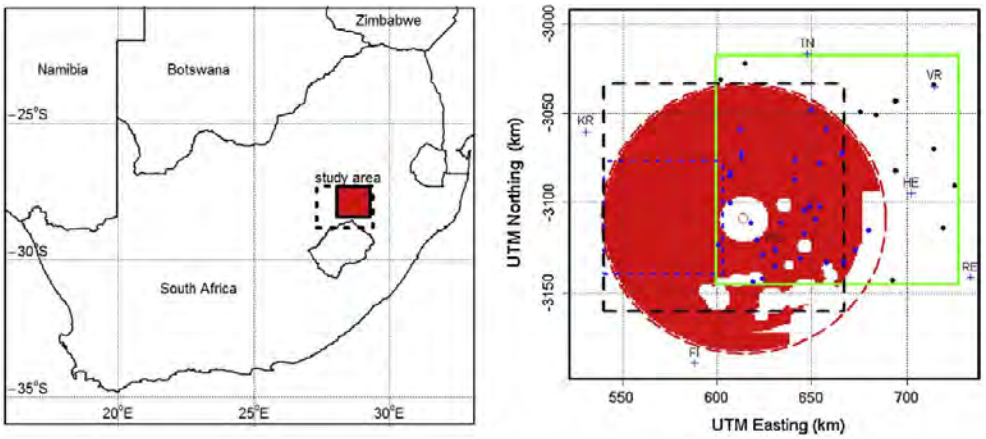


Figure 7.1. Left panel: Location of the study region; dashed area encloses the radar, wind and temperature stations. Right panel: the red dashed circle is 75 km radius of the radar coverage; red area is a radar mask; green square area is the red area in the left panel; dots are gauge locations; the black dashed square is Region 1; Region 2 is intersection of red radar circle and green boundary square; dashed blue square is Region 3.

There are 54 gauge sites in this region of 256 km square so that the mean interstation distance is 34 km, but we note the clustering of gauges in some parts.

In Figure 7.2 we have selected a relatively wet day and have left out 10 stations, whose amounts are coloured blue, located at the blue circles and can be thought of as targets. The control stations are sited at the locations of the red crosses. All rainfall amounts on the day [13 March 1991] are in millimetres at the crosses in the figure. The two sites ringed in maroon will be used in the cross-validation later in the section; all the blue sites will be omitted from the simulation and act as targets; those heavily ringed in blue will be used in Figure 7.9.

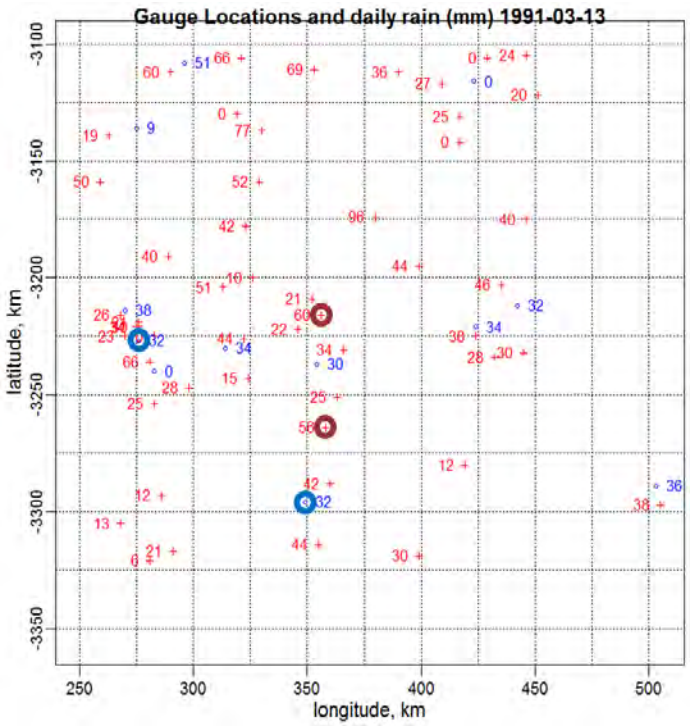


Figure 7.2. Gauge locations and amounts of rain on day 13 March 1991. The region is nearly square and covers SAWS 30' rainfall blocks numbered [230, 231, 232, 260, 261, 262, 292, 293 and 294] in the Eastern Free State.

The first step of the analysis is to determine the frequency distribution of rainfall at the gauges for the day in question. This appears in Figure 7.3 and it will be seen that p_0, the ratio of dry gauges, is 9%. The highest two readings are used to fit an exponential curve, asymptotic to probability 1, to permit extrapolation outside the range of the gauge readings; the curve is shown in red. As noted on the figure, p_e and y_e are the coordinates of the penultimate point and L_e is the correlation length of the exponential fit (red curve) through this and the last point.

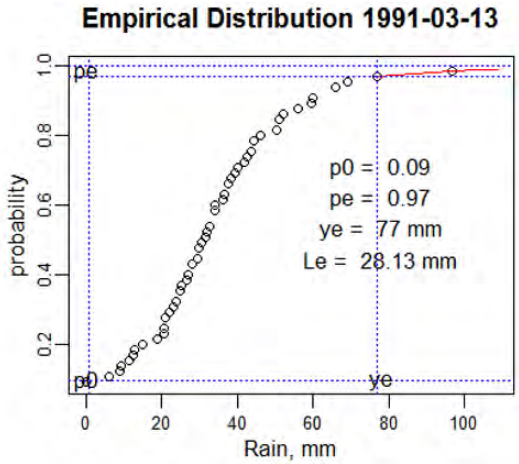


Figure 7.3. Cumulative frequency distribution of rainfall values derived from Figure 7.2.

Each of these points is then QQ transformed onto a Gaussian curve, with the dry readings given the value $y_0 = -\phi[\Phi^{-1}\{p_0\}]/p_0$ as described earlier in the discussion in Chapters 3 and 4 .

Thus, the drier the day, the lower will be the values of the dry points in the Gaussian domain.

Having Gaussianised the rainfall readings, the next step is to use Ordinary Kriging to determine the mean field through the gauge locations and determine the standard deviation at each point in the field. The correlogram is chosen to fit the one obtained from the power spectrum of the observed daily rainfall fields. These were estimated by the S-band radar at Bethlehem [Free State]; Dr Sinclair created 800 days of radar images from the set used by Clothier and Pegram (2001), from 5 minute MRL5 radar images, which were analysed to determine spatial structure. This correlation is much stronger and smoother than any computed from the gauges on the day, because the latter comprise only 54 points, whereas each radar field contains up to 50 000 estimates of rainfall values. Figure 7.4 shows the Sample Spectrum derived from the radar image [left panel] and transformed correlogram [right panel] calculated from the red line fitted to the red dots.

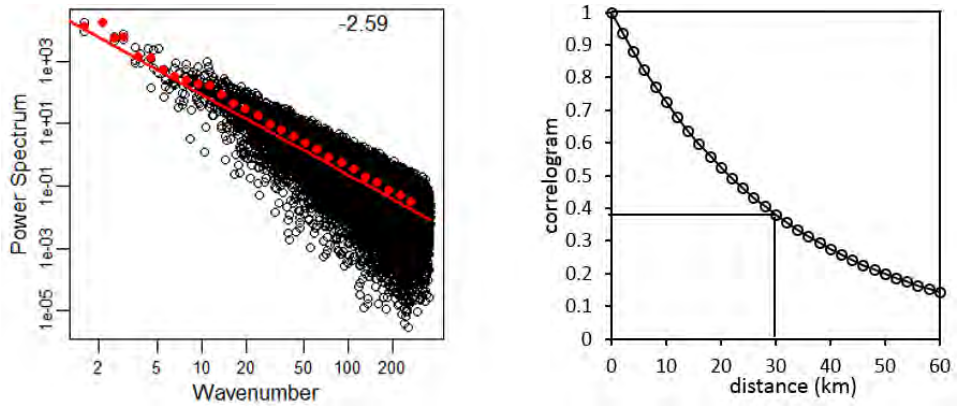


Figure 7.4. Sample Spectrum of radar image and transformed correlogram. The left panel shows the 2D power spectrum of a radar field plotted in 1D [black circles], the average of these in discrete bins [red dots] with the best linear fit in log-space [red line]. In the right panel, the Fourier transformed correlation [black circles from red line] is an exponential curve with a correlation distance of 29.7 km obtained where $1/e$ intersects the curve.

The mean slope of the power spectrum used to correlate [filter] the random field is chosen as $\beta = -2.6$, a value adopted from the WRC report on the String of Beads Model [SBM] (Pegram and Clothier, 1999 and Clothier and Pegram, 2002). This number is close to the slope of $\beta = -2.59$ estimated from the 1D power spectrum shown in the left panel of Figure 7.4.

The correlation length of the correlogram on the right is 29.7 km, just short of the interstation distance of 34 km between gauges, calculated above from the gauges. What this means is that stations near each other will allow the variability of the fitted fields between them to be smaller than those which are far apart. We note that the correlation length of an exponential correlation function is where the curve crosses the vertical axis at $1/e = 0.368$, a level explaining 14% of the variance. Thus, in locations at 60 km separation, only 2% of information is translated between them. What this drop-off of correlation with large spacing means is that, in sparse fields, the gauges are sharing little information with

each other, hence the need for maintaining raingauge networks at a minimum density of 30 km spacing. Figure 7.5 shows one of a set of 100 random Gaussian fields, filtered to have a correlation length of 30 km. Its size of 256 km square is chosen to exploit the computational economy of the Fast Fourier Transform (FFT).

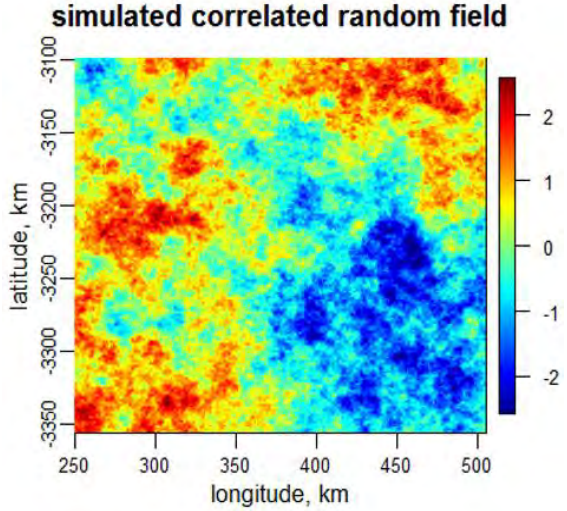


Figure 7.5. One of 100 random Gaussian fields, 256 km square, FFT filtered to have a correlation length of 30 km.

This field has structure, in the sense that it has clusters of high [red] regions and valleys of low [blue] regions. We need to statistically stitch this down onto the Kriged surface matching the Gaussianised gauge values, using the conditioning technique introduced by Sinclair and Pegram (2005). After we have interpolated the mean field between the gauges by Kriging, we obtain the Gaussian mean field image below left in Figure 7.6. The random field above in Figure 7.5, when merged with the mean field using the conditioning algorithm, gives us the combined field below right in Figure 7.6, after thresholding at the level corresponding to p_0. The white areas are where there is no rain, as the field has been thresholded at $y_0 = -\phi[\Phi^{-1}\{p_0\}]/p_0$. This is just one of the 100 merged Gaussian fields produced for the conditioning.

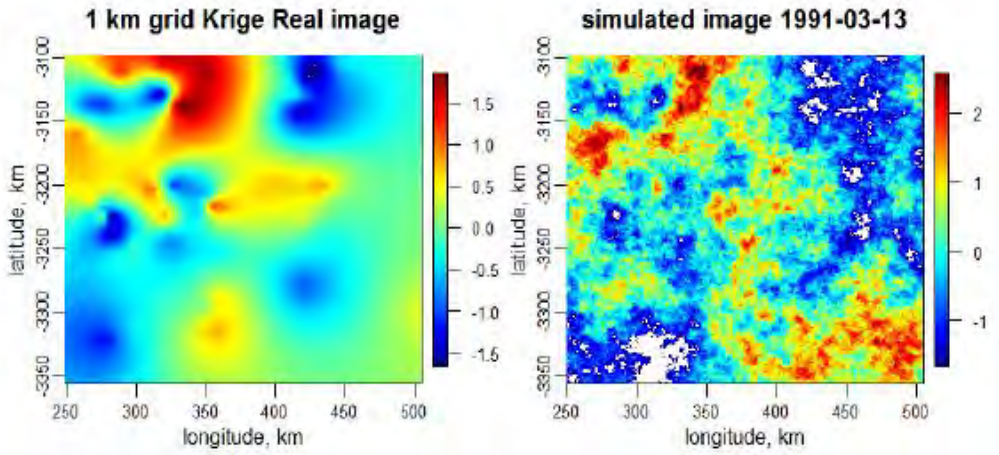


Figure 7.6. After we have interpolated the mean field between the gauges by Gaussian Ordinary Kriging, we obtain the image on the left. The random field in Figure 7.5, when merged with the mean field gives us the combined field on the right.

The following pair of images in Figure 7.7 shows (i) the mean field on the left, obtained by averaging the 100 simulated images and (ii) the median [Q50] field obtained by finding the median of the 100 values at each 1 km pixel.

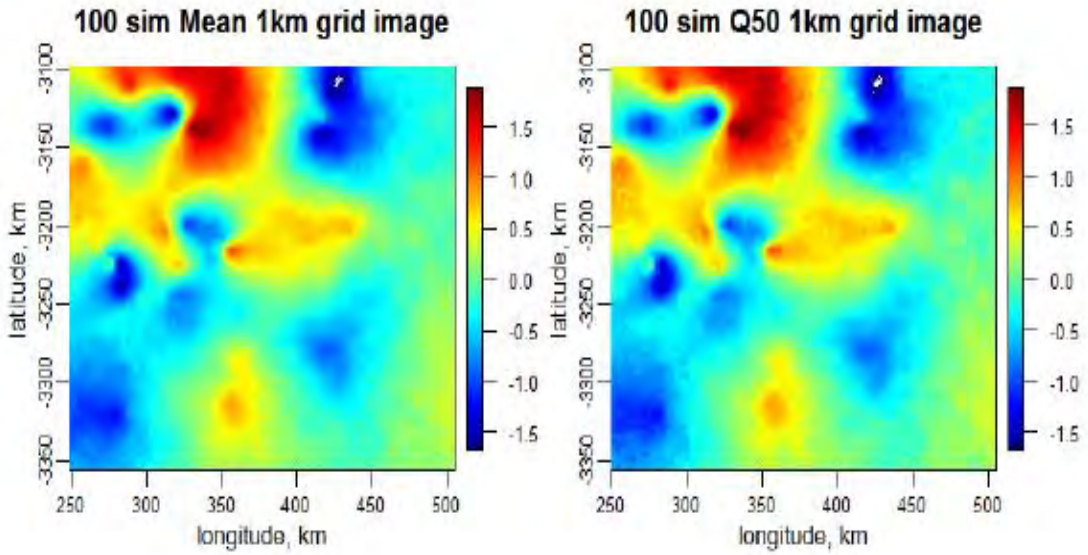


Figure 7.7. Combining of 100 simulations like that in Figure 7.6 [right] we obtain their sample mean field on the left, by averaging the 100 simulated images. The median [Q50] field on the right is obtained by finding the median of the 100 simulated values at each 1 km pixel.

There are minor textural differences between these two images in Figure 7.7 and the Kriged mean field in Figure 7.6, as expected. If we had generated very many more fields for the simulation, there would have been no detectable differences between the mean and median images because they coincide in the Gaussian domain. We will use the median and the quartile images to transform back to rainfall space, in order to obtain the local probability distributions in the rainfall domain, because although means and standard deviations change their quantiles in the QQ transformation process, the median and quartiles maintain theirs, being probabilities.

We next provide some examples of 1D vertical sections across [on latitudes] and down [on longitudes] through the 2D images to indicate the behaviour of the quantile surfaces and how well they interpolate values at missing points. The first set of four, of the six in Figure 7.8, collocate through one or more of the gauge sites, so there is no error at these locations, as we do not use a nugget in the variogram that would determine the 2D power spectrum. The longitude and latitude are marked on each image, as is the gauge number of the intersected gauge location, so they can be identified on Figure 7.2 as heavier maroon rings. The y_0 threshold is at -1.53 in the Gauss domain for these images, as they all come from the same day, where the p_0 is 9%.

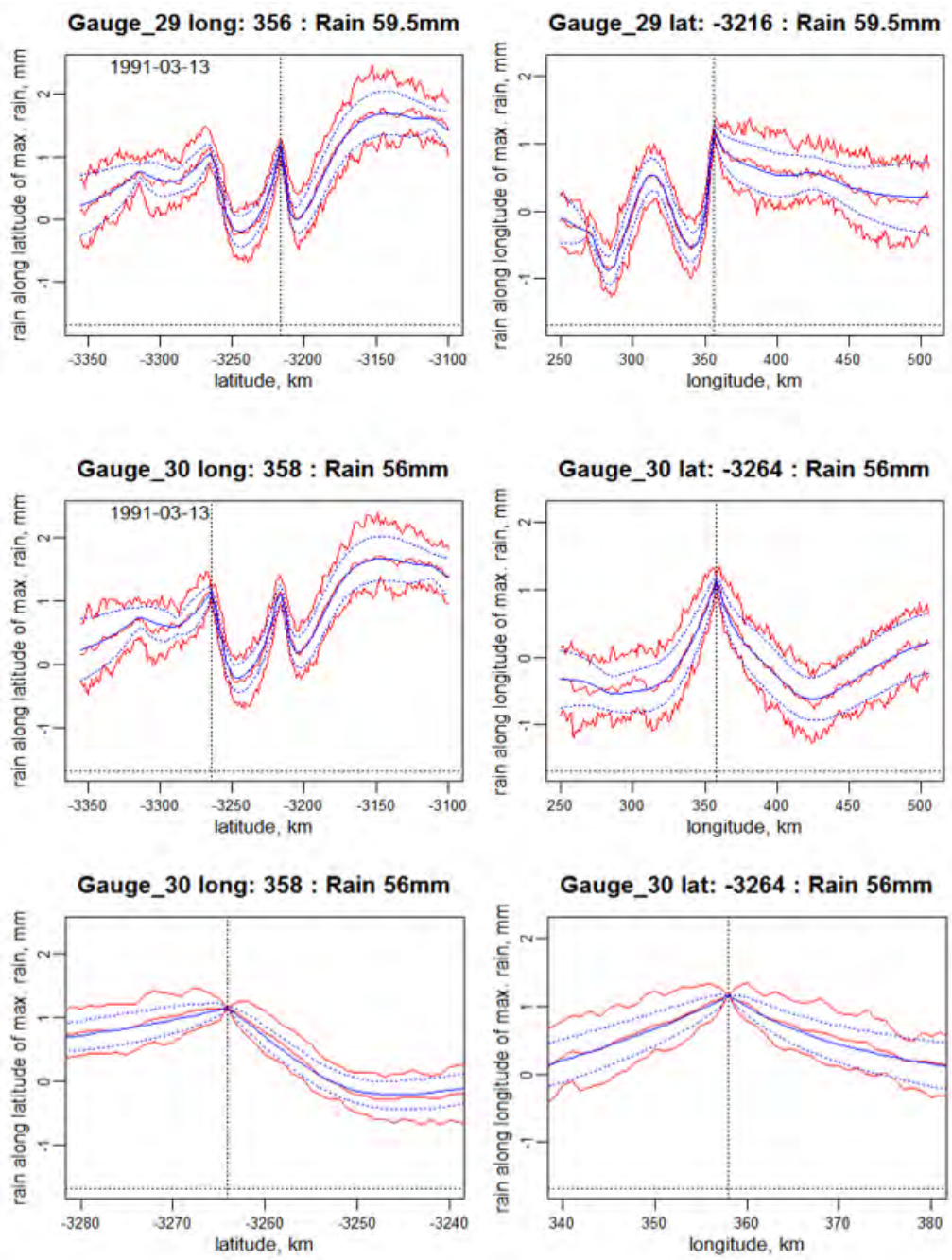


Figure 7.8. Upper 4 images, sections through gauges ringed in maroon in Figure 7.2, showing the mean and standard deviation fields of the Gaussian Kriging and FFT simulations. The lines are interpreted as follows: solid blue = Kriged mean; dotted blue = Kriged quartiles; inner wiggly red line = 50th percentile of 100 simulations; outer wiggly red lines = 25th and 75th percentiles of 100 simulations. Note their coincidence with the Kriged lines and the narrowing of the quartiles when the sections are near a gauge. The lowest pair is an expansion of the curves to give more detail.

The images in Figure 7.9 are of sections through hidden gauge sites, or targets, i.e. the ones not included in the merging procedure, and indicated by thick blue circles in Figure 7.2 and black dots here.

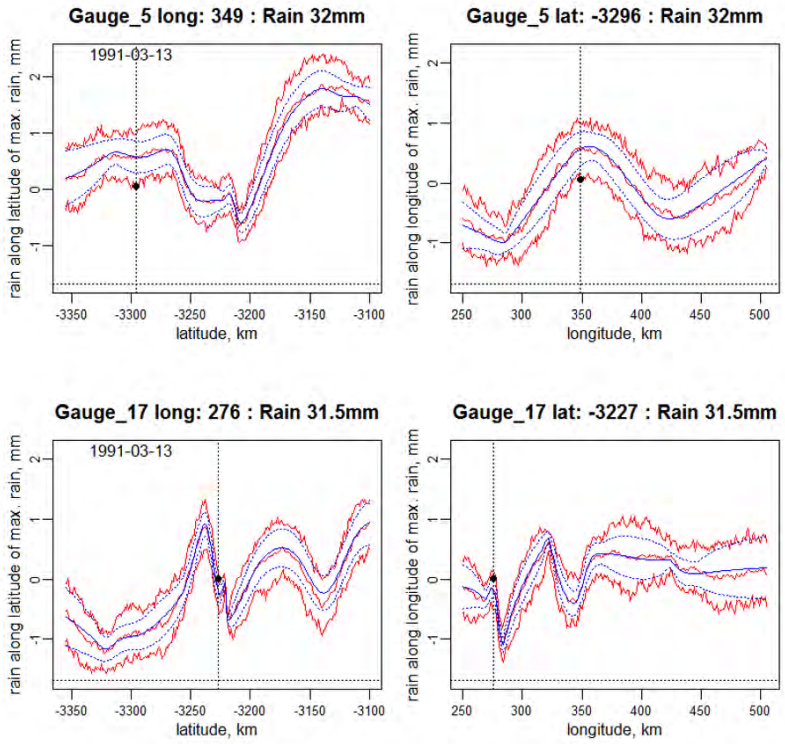


Figure 7.9. Same as Figure 7.8, but the sections are through sites with gauges removed from the computation. The vertical dashed lines in this figure, and the black dots, respectively indicate the location and value [Gaussianised] at the target. The narrowing of the quartiles is due to the presence of other nearby control gauges shrinking the gap and influencing the surface. These are representative images – others show better and worse results.

In summary, this section has demonstrated that interpolation using Gaussian Kriging combined with conditioned artificial random radar fields provides several things of value:

- Firstly we can obtain a reasonable median interpolation field in the rainfall domain [or mean field if we want to calculate it];
- secondly we can determine how good the estimate is through the calculation of the median and the quartiles [which can be used, with the mean to calculate the standard deviation – which are not very useful in a highly skewed distribution];
- thirdly we can provide some reasonable simulations of spatial fields modelled on the covariance structure of radar images, conditioned on the gauge readings, which yield the same uncertainty as the Gaussian Kriging – these can be used to determine the uncertainty of interpolating spatial rainfall from gauges over catchments in a convenient manner;
- fourthly, if the random field interpolations could be combined with a gauge network stochastic daily rainfall generator like PEGRAIN (Pegram, 2010), to make simulated radar fields using the above procedure which we call PEGRAD, then hydrologists can use this product as input to either distributed models like PyTOPKAPI (Sinclair and Pegram, 2005) or to semi-distributed models like ACRU (Schulze, 1975).

We consider that this section provides a novel and valuable methodology, which will enrich the arsenal of hydrologists, as long as the observed rainfield is spatially homogeneous from

the correlation point of view. We note that this work is one of those serendipitous pieces of research which adds value to the whole. In chapter 6, we have already seen how much better copula based interpolation and simulation refine this Gaussian Kriging methodology, by exploiting spatial inhomogeneity of the dependence structure, which the method described here ignores.

Chapter 8. Interpolate daily rainfall on 0.25° grid for TRMM comparison

8.1. Introduction

We have unfortunately discovered some things which have put a damper on this deliverable which we wrote four years ago in the proposal. The first is that TRMM finished its work in April 2015, to be replaced by the Global Precipitation Mission [GPM] satellites which are now in operation. The second is that the availability of raingauge data over RSA in 2010 was lower than we hoped – we are currently down to the data from just over 1000 SAWS sites. Thirdly, the raingauge data in the SADC countries has declined drastically since the turn of the millennium, a large proportion of the overlap time of TRMM and gauges. Therefore, if we are to make estimates of corrected TRMM data over data sparse regions, we are going to have to somehow export our quantile-quantile [QQ] transforms from relatively data-rich areas. Fourthly, the gauge density we can obtain in SADC is so sparse that the interstation distance is over 300 km in Zimbabwe and Mozambique, so data infilling is out of the question.

In this chapter we describe the work we have done to prepare data to bias correct TRMM using RSA data; the methodology should be able to be adapted to GPM. The method of interpolation chosen is one we developed for WRC contract K5/1964 (Pegram et al., 2013), elaborated on in the next section. This means that, instead of using the methods of infilling and interpolation reported in Chapters 3 and 4, we use Multiquadrics limiting the interpolation over the 0.25° squares where we have gauge data. We determine the block average rainfall on each day over 0.25° squares matching TRMM, which have active gauges on that day – we do not interpolate rainfall into empty squares for TRMM bias correction. Furthermore, it was not the task of this project to perform the actual bias correction of TRMM data; we contracted to assess feasibility of bias correcting TRMM 3B42RT 3-hourly rainfall estimates in two steps: via daily accumulations then disaggregation over RSA since 2000 and, if meaningful, extend to SADC.

Thus the purpose of the work reported in this chapter was to develop a set of daily rainfall totals over some of the 0.25° square pixels of the TRMM 3B42RT grid in blocks containing gauges, in order to provide a basis for investigating bias correction of TRMM. This chapter describes what we achieved and the methods used to do so.

8.2. Spatial average of daily data in each 0.25° block

We elected to compute an estimate of the gauged rainfall on each TRMM block containing gauges by using the Multiquadric interpolation code developed by Pegram and Pegram (1993). This FORTRAN code was wrapped in a Python package interface to make it more convenient to use in conjunction with our Python-based workflow. At the core of the Multiquadric approach is the calculation of weights to multiply each gauge value in a given block and thereby obtain the spatial average rainfall for the block on each day. The configuration of gauges on different days within a given block is sure to change. Since the gauge configuration defines the weights, it was necessary to check for active gauges and

then compute the possibly different weights for each TRMM block on each day of the 10 year analysis period – a procedure requiring some deft programming.

The final product of the Multiquadric analysis was a netCDF file containing a large three dimensional array of block averaged daily rainfall totals for each TRMM block and all 3682 days in the analysis period running from 2000-03-01 until 2010-03-31. This overlap period was chosen because the TRMM dataset runs from 2000-03-01 until April 2015, while the gauge dataset is for the period 1850-01-01 until 2010-03-31.

A similar dataset of daily rainfall accumulations was developed for the TRMM data, being careful to match the accumulation times of the TRMM in UTC to those of the gauge reporting periods in SAST (a 2 hour shift). It was important to ensure that the TRMM accumulations represented the 24 hour accumulation reported at 08:00 SAST.

We take as an explanatory example of gauge block averaging, the arrangement of gauges in the Mpumalanga region and present the following image in Figure 8.1 [was Figure 1]. This excerpt in quotes and indented is taken verbatim from WRC contract K5/1964, *so figure numbers are taken from that text*. The grid shown in Figure 1 in that passage was chosen to match the PRECIS RCM grid of 0.44°, so this explanation should be read with that in mind – the TRMM grid is 0.25°, but the method is identical in both cases.

-----////O\\\\-----

Start of quote.

> "The green dots are the SW corners of each of the blocks containing the gauges are offset by 0.01' so as not to dichotomise the gauges; the lilac crosses are the coordinates of the PRECIS grid where the RCM rainfall estimates are found; the blue diamonds are the sites of the gauges. The coordinates are in minutes East of the Prime Meridian and North of the Equator (hence the negative values on the vertical axis). The box labels count across and then up from the SW corner as shown in Table 1, following SAWS numbering convention, with the difference that their blocks are 30' square, whereas the PRECIS blocks are 26.4' square. This mismatch required some careful organisation of coordinates.
>
> "The gauge weights are all calculated by integrated Multiquadrics, restricted to those gauges which lie in each box, because inclusion of sites outside the box induces negative weights, which we want to avoid. Also, it seems sensible to restrict the box averages to the gauges inside the box, otherwise one is importing information from distance. The weights sum to 1.0. The method is fully described (with mathematical development) by Pegram and Pegram (1993), so the detail of that will not be repeated here, however the method is very close to Simple Kriging, but with the important addition of analytically computed weighting functions and the restriction to local interpolation. An example of the output of such a computation was performed on Box 6 (denoted by the red square) which contains 8 gauges. The gauge labels, coordinates and weights appear in Table 2. The closer the gauges cluster, the lower their weights, which is clear from the details of Tables 2 and 3."

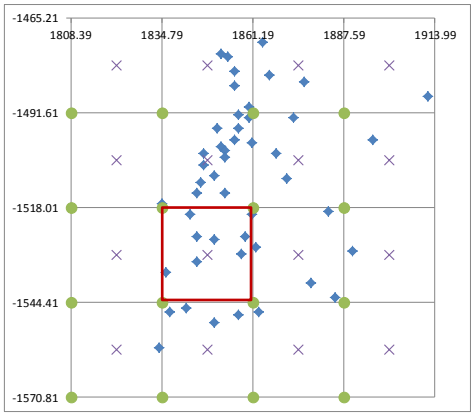


Figure 8.1. PRECIS grid and rain gauge sites – Mpumalanga. Red square is # 6, whose gauge weights appear in Table 2

"Table 1: labelling of PRECIS blocks in the square containing gauges

13	14	15	16
9	10	11	12
5	6	7	8
1	2	3	4

"The block occupations are:

BOX	1	2	3	0	0	6	7	8	0	10	11	12	0	14	15	16
# GAUGES	1	4	1	0	0	8	4	1	0	16	3	1	0	5	3	1

"Table 2. Labels, coordinates and calculated weights of gauges in Block 6 of Figure 1.

Gauge	X	Y	Weight
2	1836	-1536	0.2186
5	1845	-1533	0.2021
13	1843	-1520	0.1141
14	1845	-1526	0.0549
18	1850	-1527	0.0759
24	1858	-1531	0.2094
25	1859	-1526	0.0471
26	1861	-1520	0.0780

"Once all weights have been calculated for each box, they are arranged in a matrix relating gauge weights in each block with individual gauges. This matrix ensures that the correct information is collected in each column of box averages, the set for these Mpumalanga blocks appears in Table 3.

"Table 3. Weights of all 48 Mpumalanga gauges in the 12 occupied blocks of Figure 1.

		active blocks											
block	gauge	1	2	3	6	7	8	10	11	12	14	15	16
1	1	1	0	0	0	0	0	0	0	0	0	0	0
6	2	0	0	0	0.2186	0	0	0	0	0	0	0	0
2	3	0	0.3691	0	0	0	0	0	0	0	0	0	0
2	4	0	0.0093	0	0	0	0	0	0	0	0	0	0
6	5	0	0	0	0.2021	0	0	0	0	0	0	0	0
2	6	0	0.2805	0	0	0	0	0	0	0	0	0	0
2	7	0	0.3410	0	0	0	0	0	0	0	0	0	0
3	8	0	0	1	0	0	0	0	0	0	0	0	0
censored		0	0	0	0	0	0	0	0	0	0	0	0
7	9	0	0	0	0	0.3115	0	0	0	0	0	0	0
7	10	0	0	0	0	0.1117	0	0	0	0	0	0	0
10	11	0	0	0	0	0	0	0.1431	0	0	0	0	0
10	12	0	0	0	0	0	0	0.0745	0	0	0	0	0
6	13	0	0	0	0.1141	0	0	0	0	0	0	0	0
6	14	0	0	0	0.0549	0	0	0	0	0	0	0	0
10	15	0	0	0	0	0	0	0.0636	0	0	0	0	0
10	16	0	0	0	0	0	0	0.1482	0	0	0	0	0
10	17	0	0	0	0	0	0	0.0619	0	0	0	0	0
6	18	0	0	0	0.0759	0	0	0	0	0	0	0	0
10	19	0	0	0	0	0	0	0.0198	0	0	0	0	0
10	20	0	0	0	0	0	0	0.0122	0	0	0	0	0
10	21	0	0	0	0	0	0	0.0071	0	0	0	0	0
10	22	0	0	0	0	0	0	0.0385	0	0	0	0	0
10	23	0	0	0	0	0	0	0.1111	0	0	0	0	0
6	24	0	0	0	0.2094	0	0	0	0	0	0	0	0
6	25	0	0	0	0.0471	0	0	0	0	0	0	0	0
6	26	0	0	0	0.0780	0	0	0	0	0	0	0	0
7	27	0	0	0	0	0.3209	0	0	0	0	0	0	0
11	28	0	0	0	0	0	0	0	0.1604	0	0	0	0
11	29	0	0	0	0	0	0	0	0.4673	0	0	0	0
7	30	0	0	0	0	0.256	0	0	0	0	0	0	0
8	31	0	0	0	0	0	1	0	0	0	0	0	0
10	32	0	0	0	0	0	0	0.1462	0	0	0	0	0
14	33	0	0	0	0	0	0	0	0	0	0.5110	0	0
14	34	0	0	0	0	0	0	0	0	0	0.0551	0	0
14	35	0	0	0	0	0	0	0	0	0	0.019	0	0
14	36	0	0	0	0	0	0	0	0	0	0.1289	0	0
10	37	0	0	0	0	0	0	0.0177	0	0	0	0	0
10	38	0	0	0	0	0	0	0.0637	0	0	0	0	0
10	39	0	0	0	0	0	0	0.0019	0	0	0	0	0
14	40	0	0	0	0	0	0	0	0	0	0.2859	0	0
10	41	0	0	0	0	0	0	0.0298	0	0	0	0	0
10	42	0	0	0	0	0	0	0.0607	0	0	0	0	0
15	43	0	0	0	0	0	0	0	0	0	0	0.3365	0
15	44	0	0	0	0	0	0	0	0	0	0	0.1597	0
11	45	0	0	0	0	0	0	0	0.3723	0	0	0	0
15	46	0	0	0	0	0	0	0	0	0	0	0.5038	0
12	47	0	0	0	0	0	0	0	0	1	0	0	0
16	48	0	0	0	0	0	0	0	0	0	0	0	1

"To obtain the daily block averages of the observed gauge data, the column matrix of gauge observations (48 in the case of Mpumalanga) is post-multiplied by the weight table in Table 3."

-----\\\\O////-----

End of quote.

If the gauge population of a particular block changes, then its gauge weights are recalculated for the day in question, as previously mentioned.

8.3 Calculating daily rainfall averages on active Blocks

Figure 8.2 shows the region chosen to perform an initial analysis to determine the gauge density over a region and refine the application of the Multiquadric method outlined above. It will be seen that the range of MAP in the block is fairly representative of that over RSA as indicated by the gauge MAP.

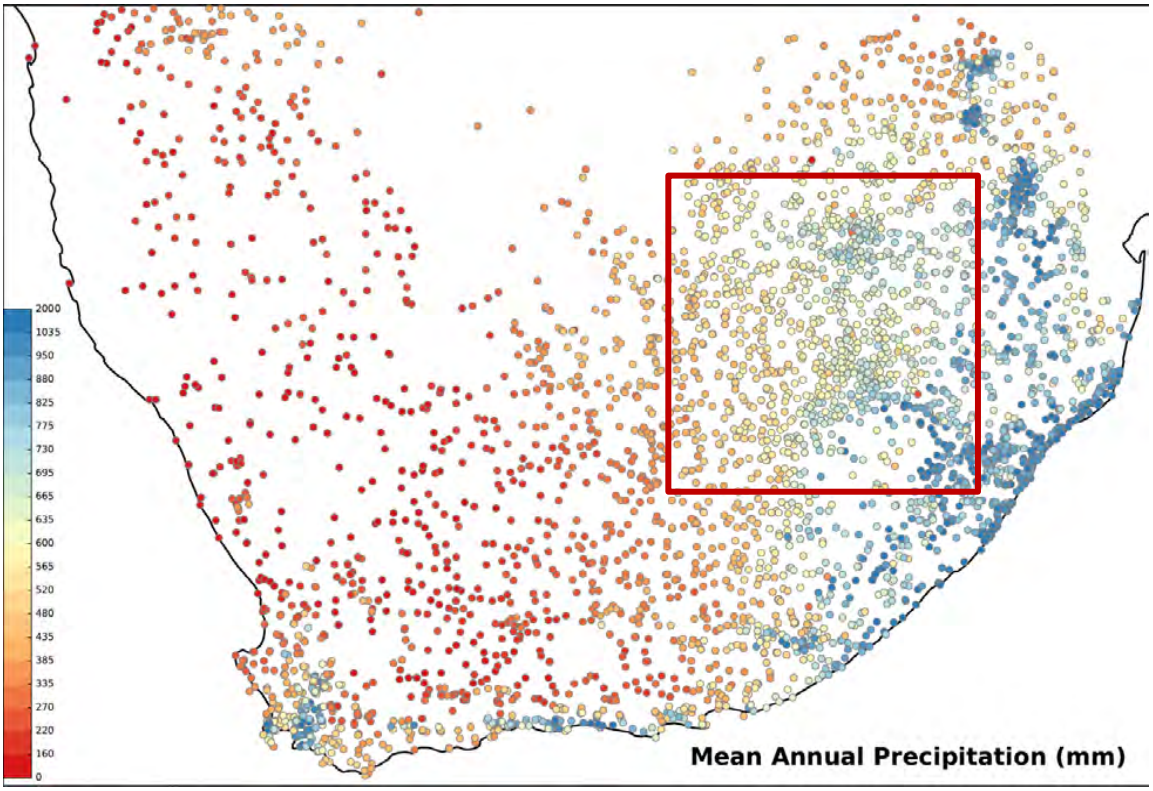


Figure 8.2. Location of the subregion of South Africa, chosen to bound Figures 8.3 and 8.4. The 5° by 5° region chosen [25°S to 30°S and 25°E to 30°E] is shown by the red square.

As mentioned in Section 8.2, the core of the Multiquadric approach is the calculation of weights by which to multiply each gauge value in a given block on a given day and thereby obtain the block average rainfall for the block. As illustrated, by comparing Figures 8.3 and 8.4, the configuration of gauges on a day within a given block may change. Since the gauge configuration defines the weights, it was necessary to check and possibly re-compute the weights for each TRMM block on each day of the 10 year analysis period.

The properties of the two datasets [gauge block averages and TRMM] are presented in Figures 8.5 to 8.11, illustrating the salient points drawn from the large data-base. In each case the caption provides a full description of the figure's contents; the story develops with the figures, so we do not provide supporting text in the body of the report.

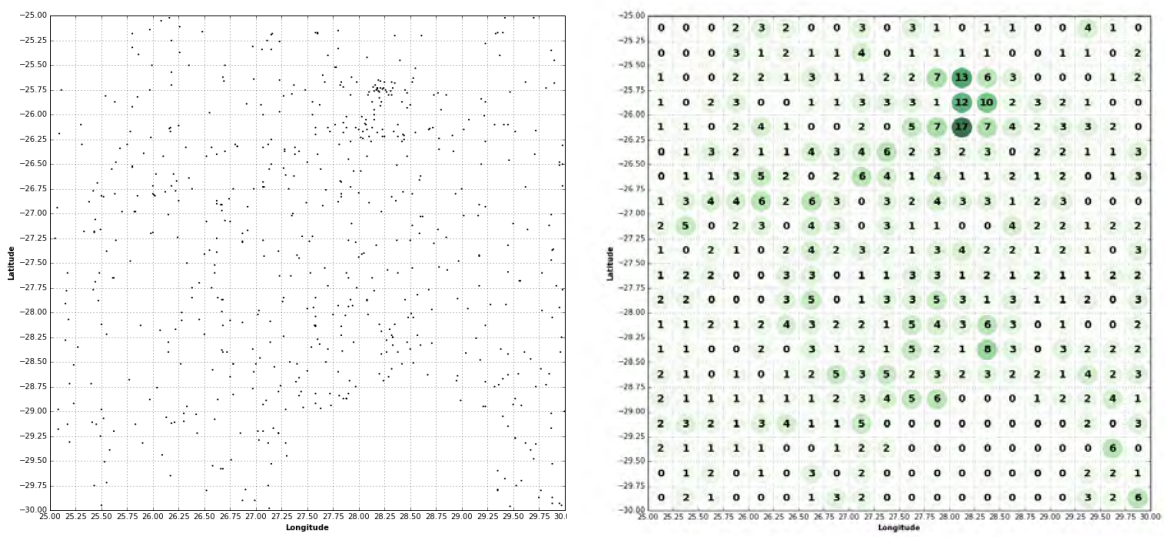


Figure 8.3. The 5° square subregion of South Africa indicated in Figure 8.2, illustrating the layout of rain gauges active within the period 2000-03-01 to 2010-03-31 and overlaid by the 0.25° TRMM grid (left panel). The right hand panel shows the total number of gauges in each grid block active at any time in the 121 month period.

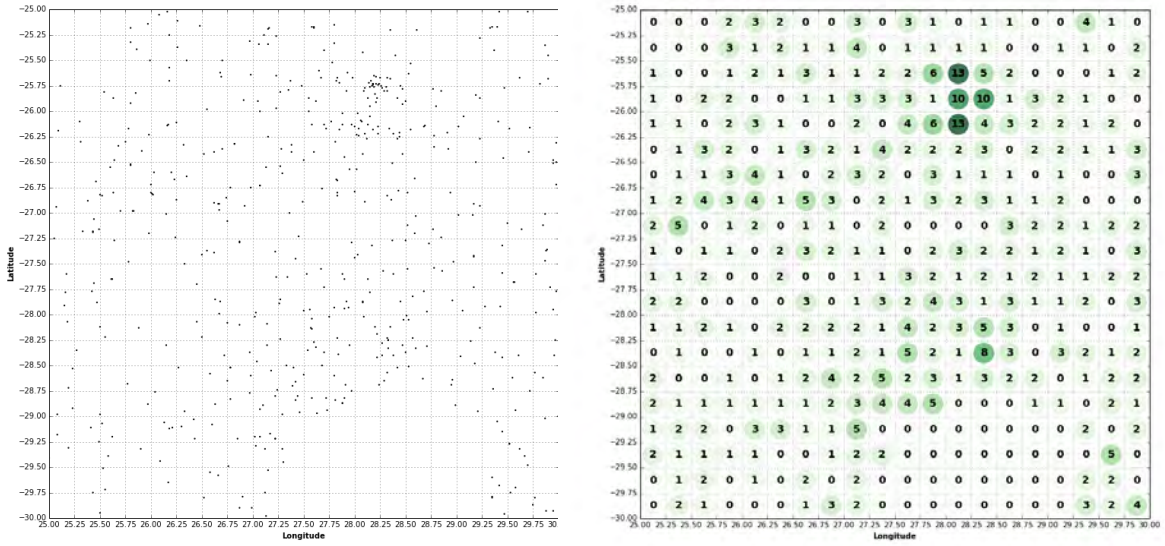


Figure 8.4. As for figure 8.3, but here showing gauges active on the first day of the overlapping data-sets: day (2000-03-01). Note the lower gauge counts in the dense cluster in the upper right corner when compared to Figure 8.3. The layout of active gauges is not constant throughout the period and this had to be accounted for in the analysis, by recalculating the weights, in each gauge-active block, on each day.

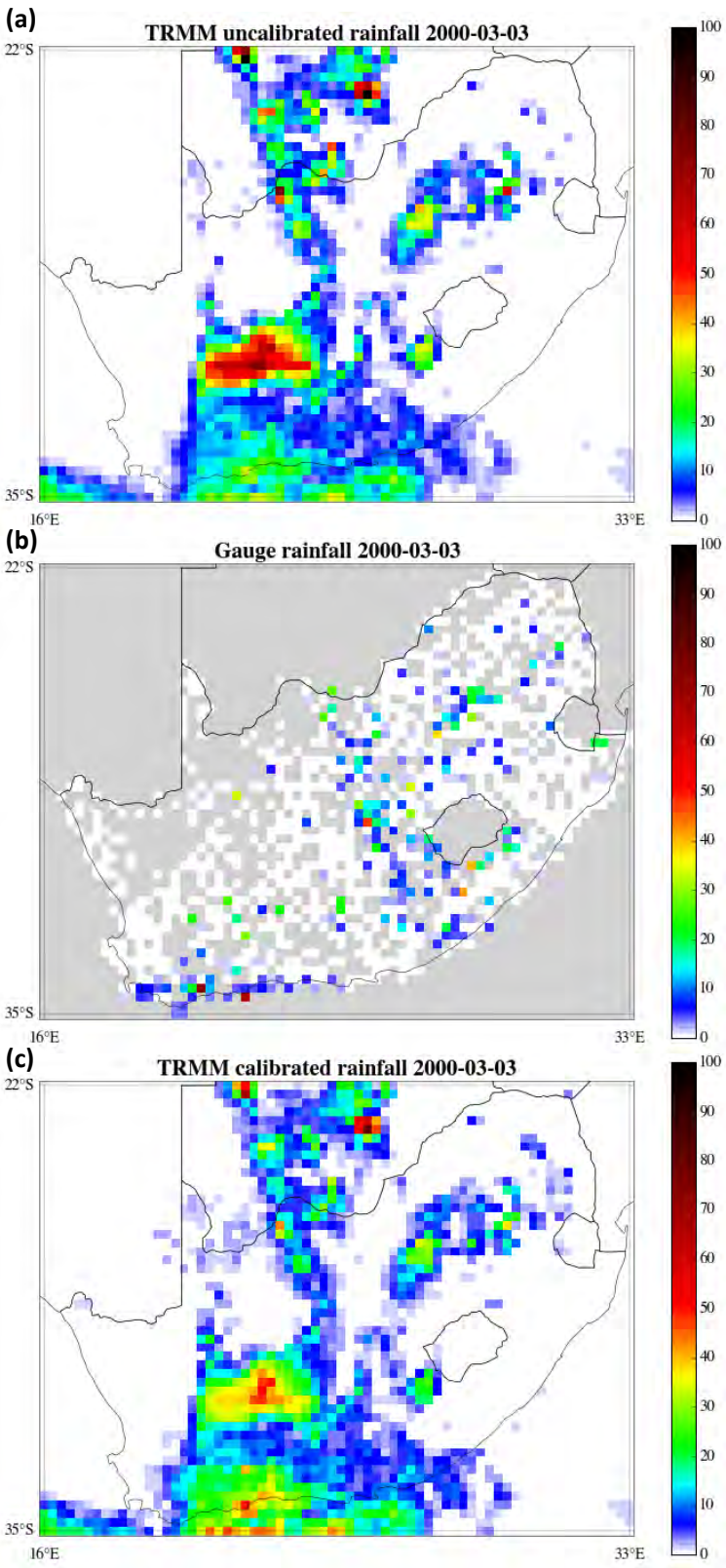


Figure 8.5. A comparison of daily totals from gauges and TRMM on 3 March 2000. Panel (a) shows the rainfall amount estimated by the uncalibrated TRMM algorithm – uncalibrated means the rainfall estimates are made using only satellite data and retrieval algorithms. Panel (b) shows the block averaged gauge rainfall recorded on the same day, with grid blocks containing no data coloured grey. Panel (c) shows the

calibrated TRMM estimate; this is the uncalibrated estimate of panel (a) adjusted via a quantile transform to match the gridded GPCP rainfall product (Huffman et al., 2010). Note the general agreement on raining areas, but with far more zeros in the gauge estimates (b) when compared to TRMM (a) and (c).

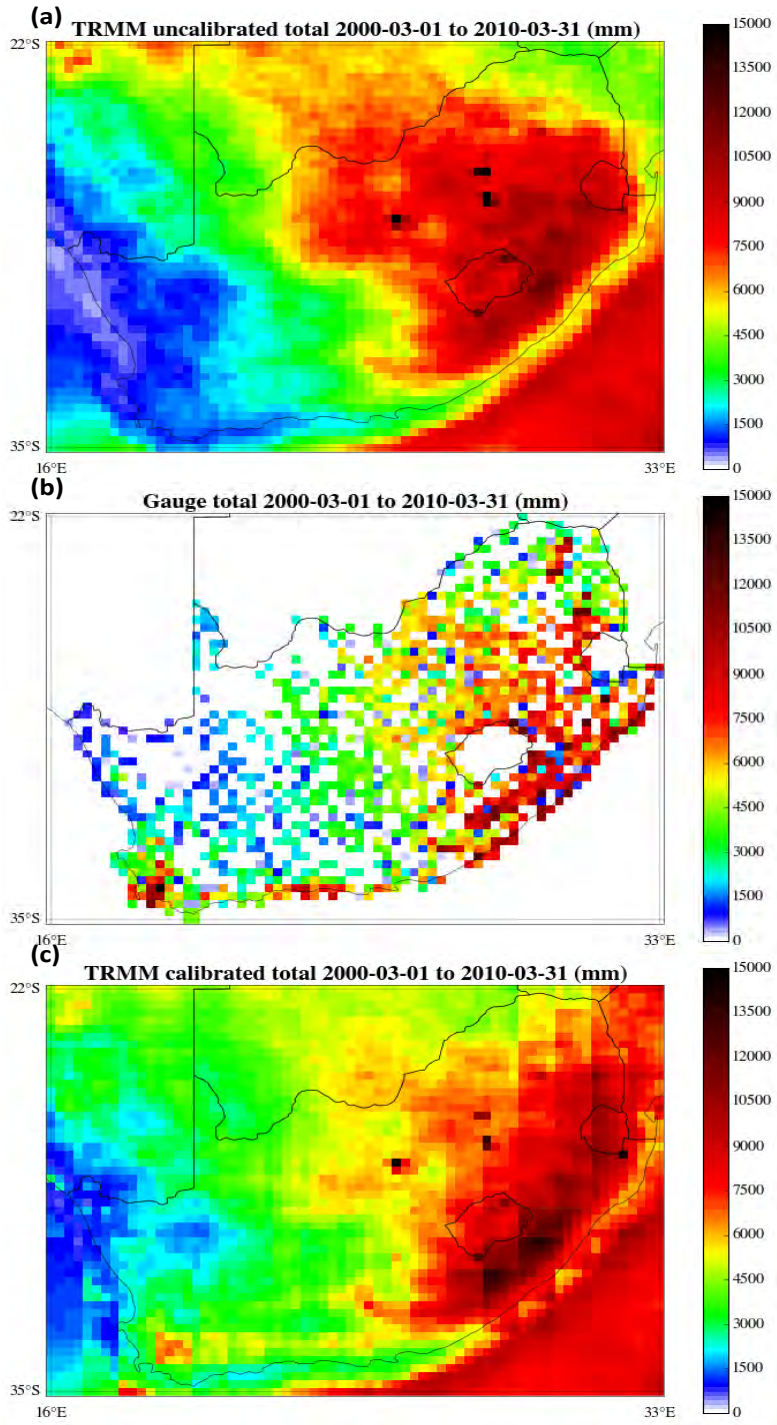


Figure 8.6. The total rainfall accumulations for the 10 year analysis period as estimated by each product. The general patterns and amounts show good agreement, but the gauge values show considerable noisy variation. This variation is explained by the variability in available record lengths which strongly affects the total (see Figure 8.7). In addition, note the artefacts in panel (c) from the calibration process, particularly in the Southern and Eastern parts, which are very 'blocky '. The Cape's annual rainfall is severely underestimated by TRMM. Even so, we will be wise to downscale the uncalibrated (a) rather than calibrated TRMM (c).

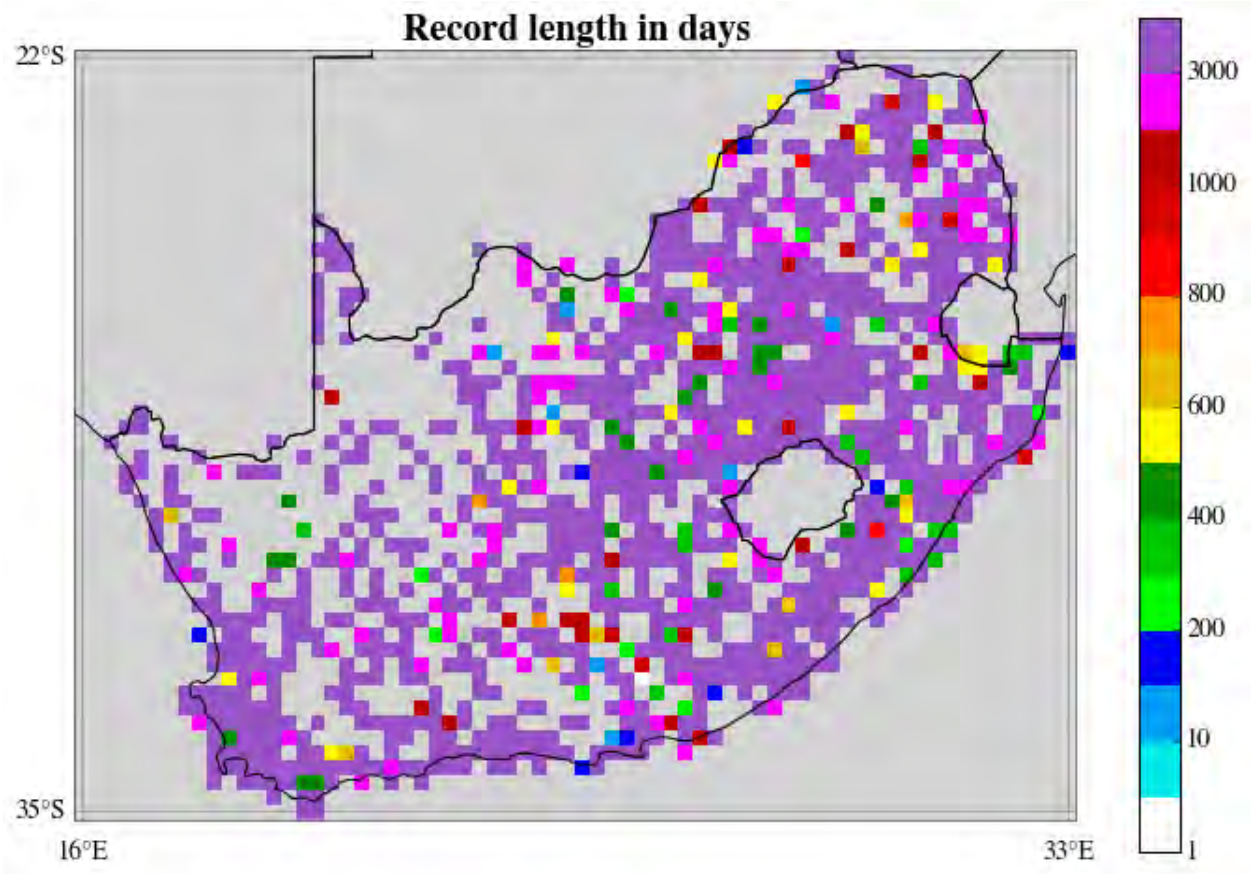


Figure 8.7. Length of the available gauge record in each block (in days). The total analysis period is 3682 days. Several blocks do not have a record spanning the entire period – this is usually the result of a block containing only a single gauge which is sporadic.

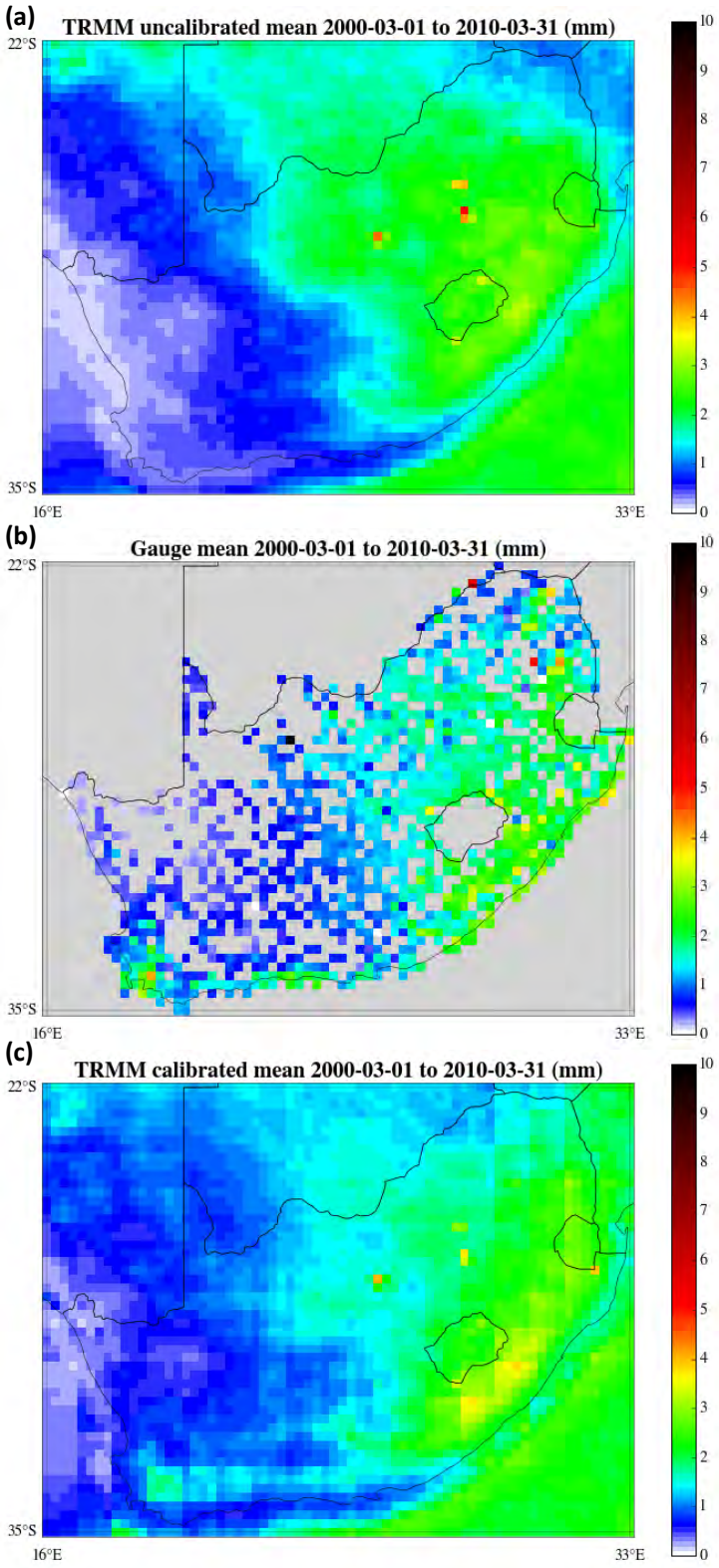
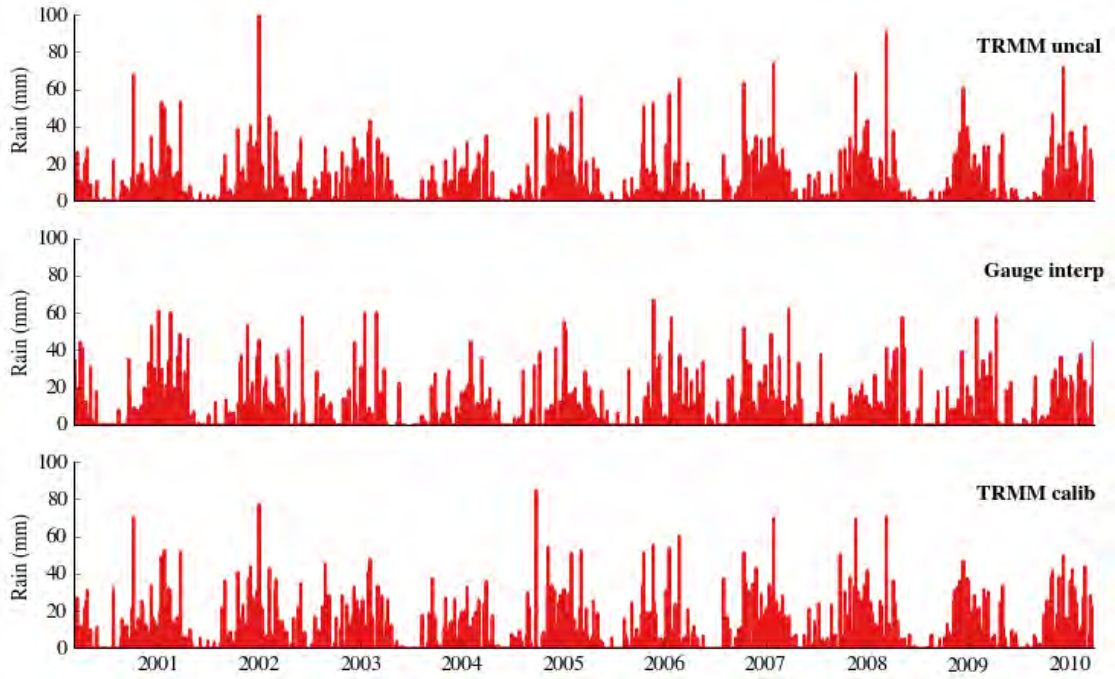
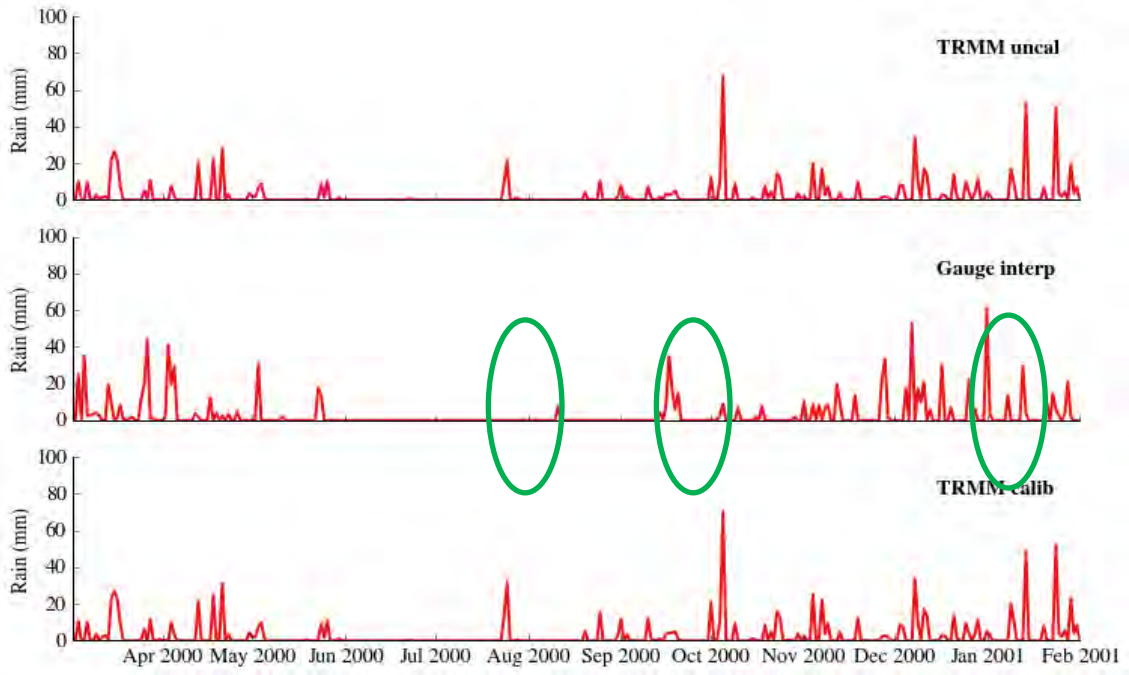


Figure 8.8. The mean rainfall values for the 10 year analysis period as estimated by each product. The general patterns and amounts show good agreement. The values are low, mostly due to the large proportion of zeros in the dataset (we have accounted for missing values). The gauge estimates (b) are smoother than the totals shown in Figure 8.6 since the length of record has a much smaller effect. Particularly noticeable in panel (a) are three isolated very high counts in small areas in Gauteng. They appear to be associated with large water-bodies.

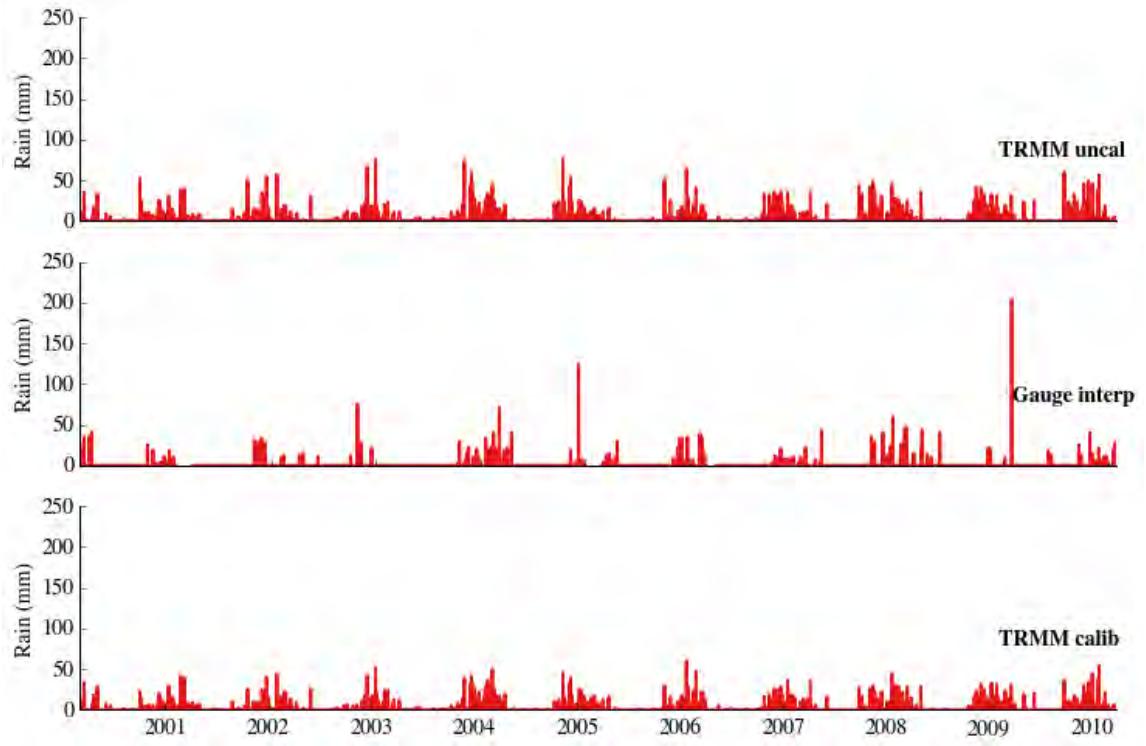


(a)

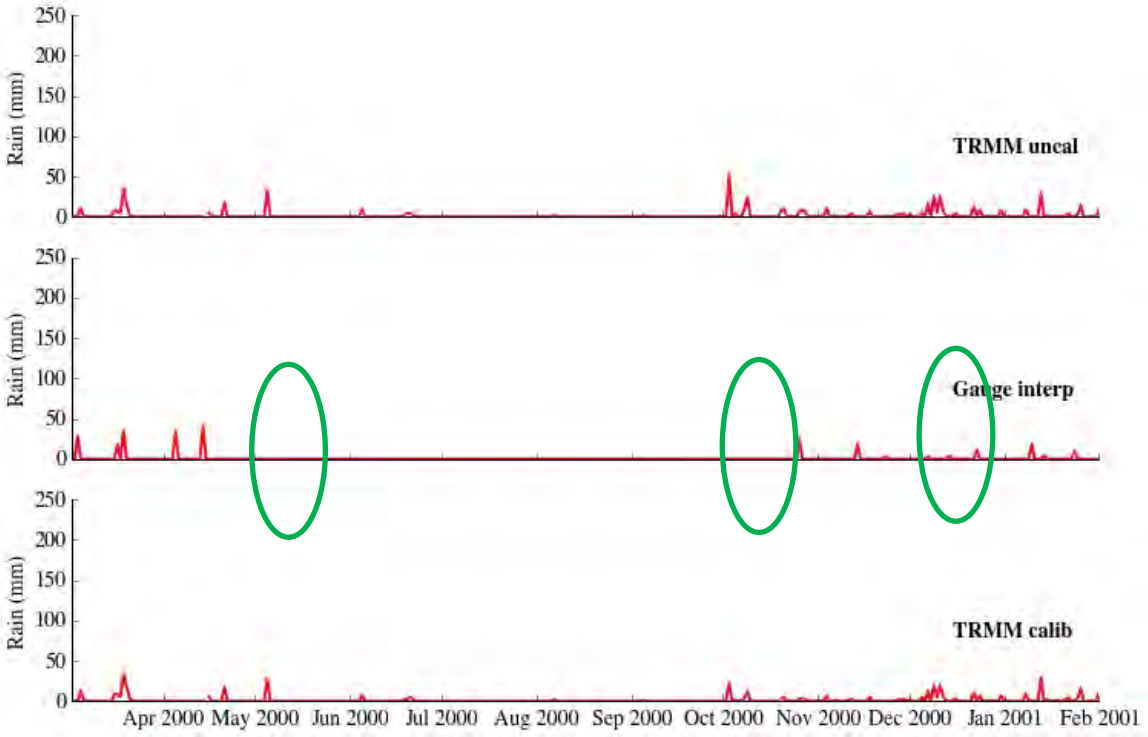


(b)

Figure 8.9. Comparison of time series for a single grid block centred on (30.87°5 S, 27.625°E). Panel (a) shows the comparative daily time series for the entire analysis period, while panel (b) shows the time series for a single year of data at the beginning of the period of comparison. There is good agreement on the wet and dry periods and the magnitudes of rainfall. However, there are many timing mismatches evident [three of them indicated by the green ovals] which reduce the correlation between these daily Gauge Block averages and TRMM time series to below 0.5.

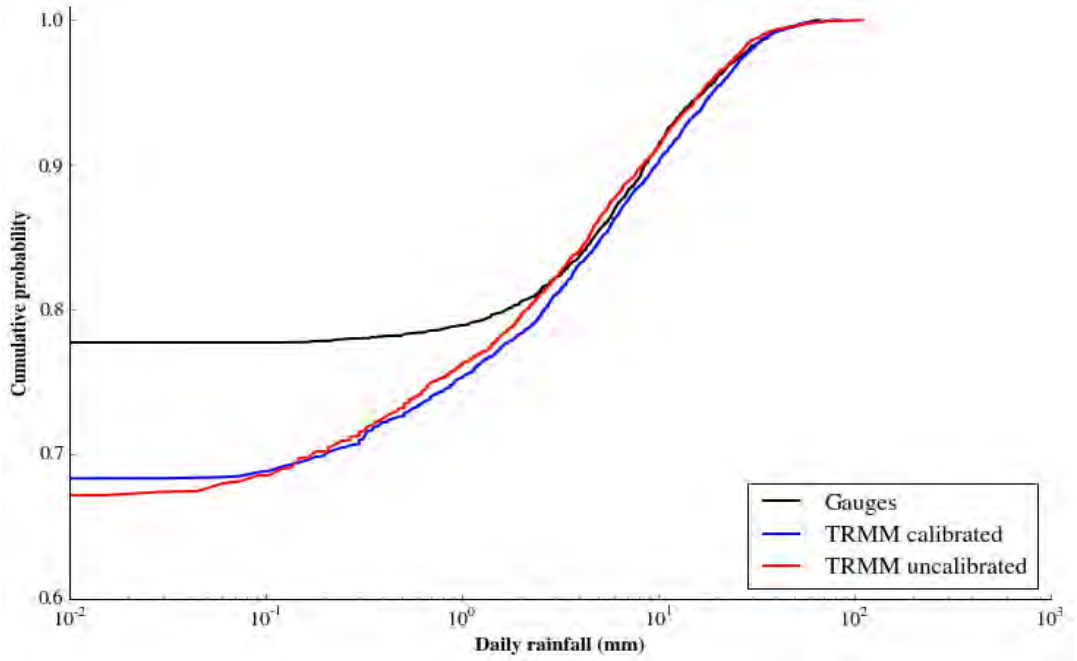


(a)

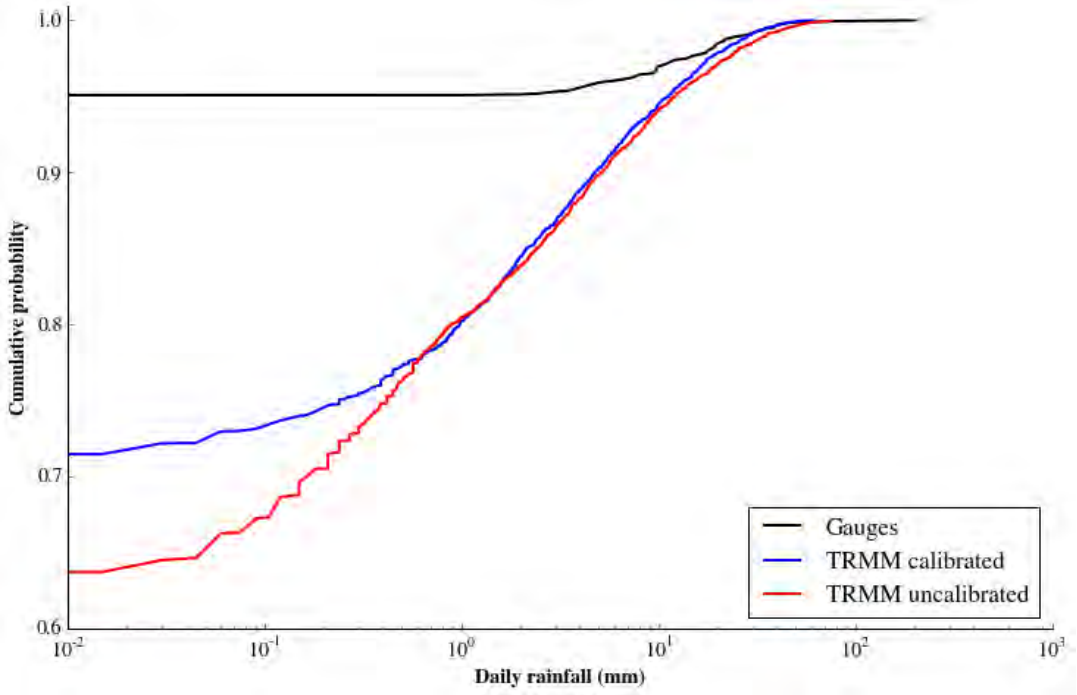


(b)

Figure 8.10. Comparison of time series for a single grid block centred on (24.375 S, 28.875 E). Panel (a) shows the comparative daily time series for the entire analysis period, while panel (b) shows the time series for a single year of data. There is good agreement on the wet and dry periods and the magnitudes of rainfall at the monthly scale. However, there are many mismatches evident at the daily scale [three of them indicated by the green ovals] which reduce the correlation between the time series.



(a)



(b)

Figure 8.11. The reason behind providing the previous Figures (8.9 and 8.10) is illustrated by comparing the Empirical Cumulative Distribution Functions (ECDFs) for the two different locations in this pair of distributions. In both cases the dry probabilities of the gauge block estimates are higher than the dry probabilities of the TRMM estimates. However, in the case of panel 8.11 (b), which matches the time relatively dry series shown in Figure 8.10, there is also a marked difference between the gauge and TRMM distributions for the higher rainfall amounts.

8.4 Possibility of extending downscaled TRMM gauge blocks to SADC

To date, the only rain gauge database we have discovered that covers the SADC region is the Global Historical Climate Network (GHCN) database (Menne et al., 2012). Figure 8.12 illustrates the gauge availability for the SADC region. Figure 8.12 (a) paints a promising picture with excellent gauge coverage over South Africa and Namibia, and reasonable coverage over several other countries. Unfortunately, the situation deteriorates after the late 1990's as shown in Figures 8.12 (8.1b) and 8.12 (c), which show very sparse coverage in this period. It seems unlikely that we will be able to extend the coverage outside of South Africa meaningfully, however we will make use of the limited gauge set that is currently available to us and pursue possible alternatives. Clearly the gauges do (or did) exist in SADC, however the difficult question is how one can obtain the data ... we may be forced to perform quantile transforms 'exported' and based on similar climate regions in South Africa. A suggested way forward is mooted in Chapter 9. We note, on looking back at Figures 8.5, 8.6 and 8.8, that we might be wise to downscale uncalibrated TRMM directly, because of the artefacts present in the calibrated version.

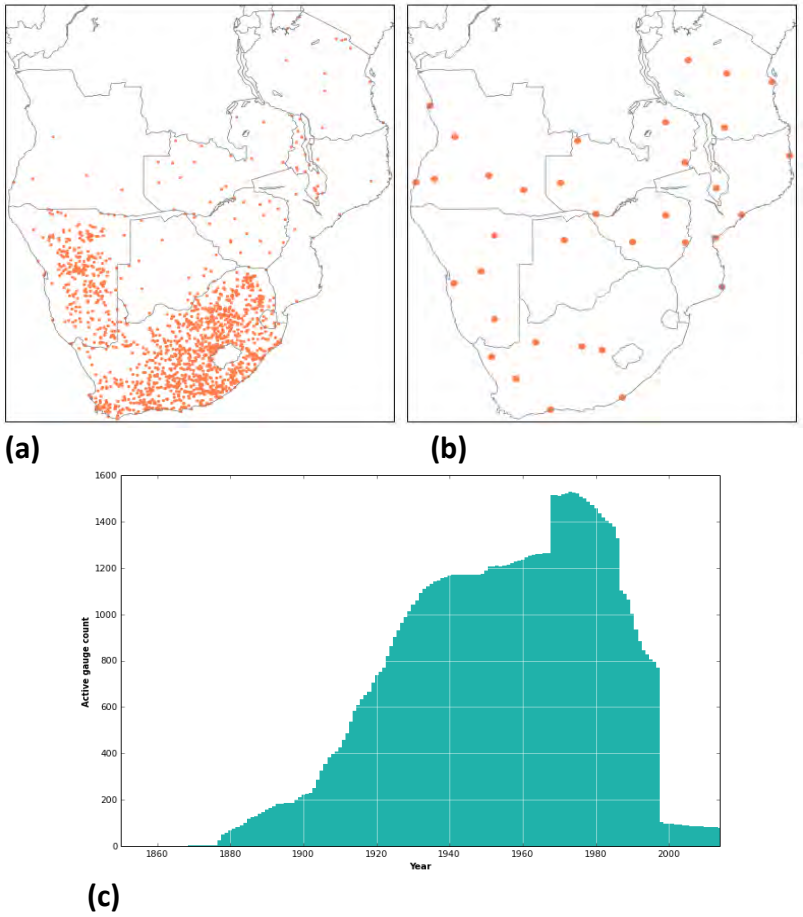


Figure 8.12. Rainfall gauges contained in the Global Historical Climate Network database (Menne et al., 2012). Panel (a) shows all available gauges in the database, while panel (b) shows the subset available during our analysis period. It is clear from panel (c), the record of active gauges in the region from 1850 to 2010, that there is a large die off from the late 1990's. This is most likely after a major collection effort was made, while after 1997 the updates to the database relied on the limited gauges of the WMO GTS network.

The regression methods outlined next in Section 8.5 are a start, however the ideas introduced in Chapter 9, composed only very recently, may go some way to ameliorate this difficult problem.

8.5 How can we Bias correct TRMM?

The key issues in the bias correction that we undertake here are that we need to correct the differences of the TRMM and gauge rainfall estimates in (i) the spatial scale and (ii) the temporal scale of the data sampling. The TRMM data are sampled as snapshots at three hour intervals, nearly uniformly in space over the region; the gauge estimates are daily accumulations at points which are randomly and unevenly scattered in space and are effectively limited to RSA. As for temporal concurrence of the two datasets, we note that we only have a 10-year overlap because the TRMM dataset runs from 2000-03-01 until April 2015, while the gauge dataset spans the period 1850-01-01 until 2010-03-31.

We proceed by doing the following. The TRMM daily rainfall estimates are clocked at 3-hour intervals from midnight, so they need to be carefully adjusted relative to the temporal pattern of raingauge readings, standardised at 8 am in RSA by SAWS. The 8 TRMM images on each day are then accumulated (taking care of the temporal mismatch) and their rain-rates converted to daily totals. The TRMM pixels are 0.25° square, so we need to spatially average the raingauge catch on each day over the TRMM pixels for comparison purposes, but of course, this can only be done on TRMM blocks where there are gauges.

Once these data have been matched in time and space, the next step is to find statistical linkages between the two sets. The aim is to bias correct the TRMM images in a series of steps. The first is to determine how well they match at the daily scale, then over different periods of accumulation: pentads, months and years; we choose pentads instead of weeks because there are exactly 73 per year except in a leap-year. Because of the strong seasonal signal, one needs to be careful how one calculates simple statistics like cross-correlation. As an example, if we were to take the ten years of monthly totals of gauge and TRMM block-averages and calculate the cross-correlation coefficients (cccs) the seasonality dominates the result, artificially increasing the cccs (as illustrated in section 4.2). We therefore need to 'de-seasonalise' the data to determine the true level of information transfer. This goes for months, pentads and days, but of course not years.

The purpose of determining the cccs is to ascertain whether we need to use a coarser period than a day to find a useful correlation linkage. Once we have settled on a period, then we can proceed to perform a bias correction of TRMM blocks to gauge-averaged blocks using a quantile-quantile (QQ) transformation of the individual periods using the products of regression. It turns out that pentads are usually much better correlated than days, so one might scale daily TRMM estimates using the appropriate parent pentad. We have yet to decide how to scale TRMM over areas with no gauge coverage – perhaps over geographically similar regions?

The product of the Multiquadric analysis of the block averaged daily gauge data was a netCDF file containing a three dimensional array of block averaged daily rainfall totals for each TRMM block and all 3682 days in the analysis period running from 2000-03-01 until

2010-03-31. This overlap period was chosen because the TRMM dataset runs from 2000-03-01 until April 2015, while the gauge dataset spans the period 1850-01-01 until 2010-03-31.

A similar dataset of daily rainfall accumulations was developed for the TRMM data, being careful to match the accumulation times of the TRMM in UTC to those of the gauge reporting periods in SAST (a 2 hour shift). It was important to ensure that the TRMM accumulations represented the 24 hour accumulation reported at 08:00 SAST.

It is important to check the cross-correlations between the uncalibrated TRMM and the block-averaged gauges. In 4 provinces, we chose areas with different climates – Gauteng, Kwazulu-Natal, Western Cape and Limpopo – to compare the TRMM and block averaged precipitation.

The sites are shown on a map of Southern Africa in Figure 8.13, where the darkness of the circles centred in the 0.25° blocks indicates the number of gauges available to be averaged.

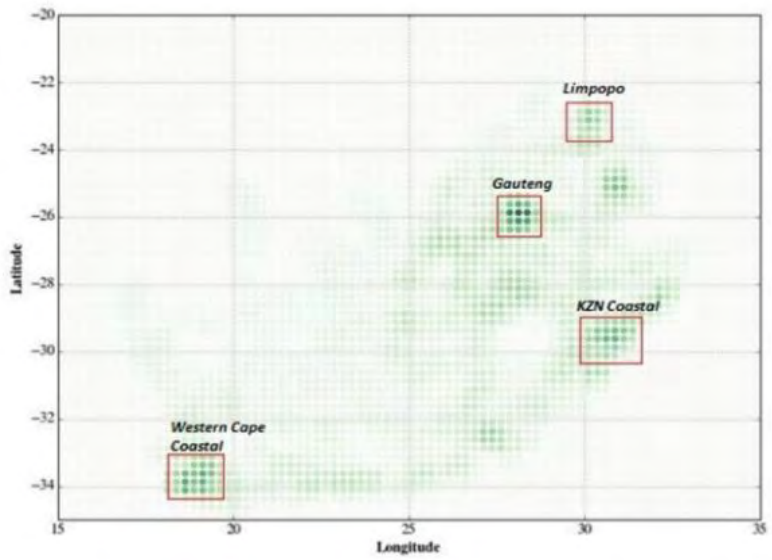


Figure 8.13. 4 areas in RSA with different climates in which to compare the TRMM and block averaged precipitation: from North to South, Limpopo, Gauteng, KZN coastal and Western Cape coastal.

To de-seasonalise the data, we needed to ascertain a smoothed mean and standard deviation of the daily, pentad and monthly accumulated data. As explained in the introduction, if this was not done then a spuriously high ccc would result, giving false hopes for reasonable bias correction using curves like that in Figure 8.11.

We found that the R^2 values between TRMM and Block Averaged Gauge Data (BAGD) ranged from 0.06 to 0.46 for daily totals, a disappointingly low result. In contrast the accumulations into pentads and monthly totals move the R^2 values up to 0.4 and 0.7 respectively. Figure 8.14 shows the scatter-plot for Block 5 in Gauteng, typical of daily data.

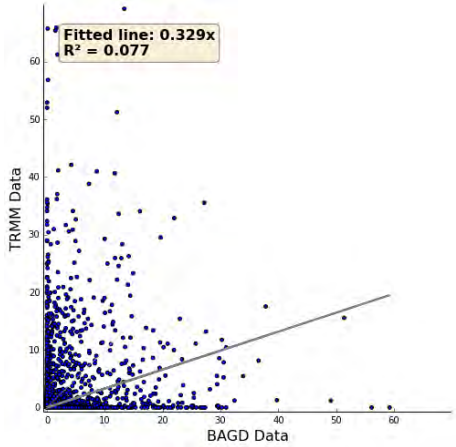


Figure 8.14. scatter-plot between TRMM and BAGD daily data for Block 5 in the Gauteng area.

By contrast the monthly accumulations are much improved as shown in Figure 8.15 for the same block; the bias has reduced and the R^2 increased.

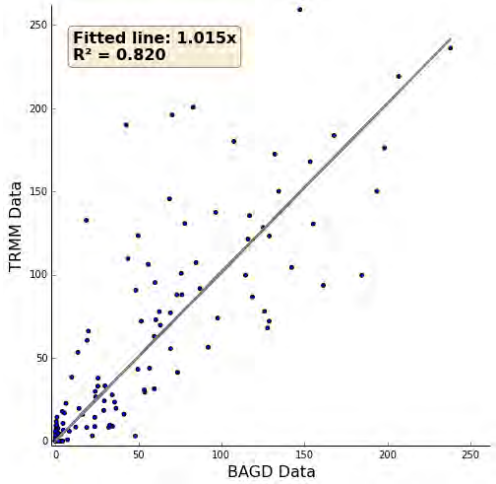


Figure 8.15. scatter-plot between TRMM and BAGD monthly data for Block 5 in Gauteng.

Unfortunately the rest of the results are not as good as that. For example, Figure 8.16 compares the daily totals of Block 2 and Block 7 in the Western Cape.

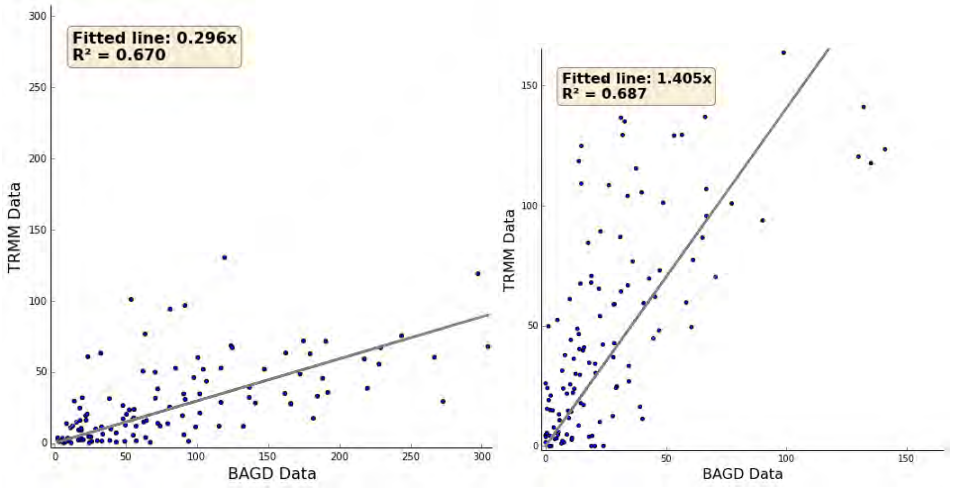


Figure 8.16. scatter-plot between TRMM and BAGD daily data for Blocks 2 and 7 in Western Cape.

The R^2 of 0.670 Block 2 is reasonably high, but the slope is 0.296, indicating an average bias of 2/3. To confound the problem, Block 7 has nearly the same R^2 at 0.687, but the slope is 1.41, the opposite of Block 2, as shown in Figure 8.16. This difference is evident in those already alluded to between TRMM and gauge data in the Western Cape, as noted in the caption of Figure 8.6. The relationship between TRMM and gauges is therefore very site-specific, likely due to topography and the number of gauges in each block.

Here follows a summary of the daily, pentad and monthly statistics, preceded by some examples of methods used to obtain smoothed estimates of means and standard deviations. Figure 8.17 shows the difficulties encountered in fitting smoothing functions where there is a high probability of a month in the data being dry. The curves fitted are Fourier Series (2 and 4 harmonics) and a triangular numerical filter of 3 weights summing to 1. In the figure, the months are augmented beyond 12 to emphasise the periodicity.

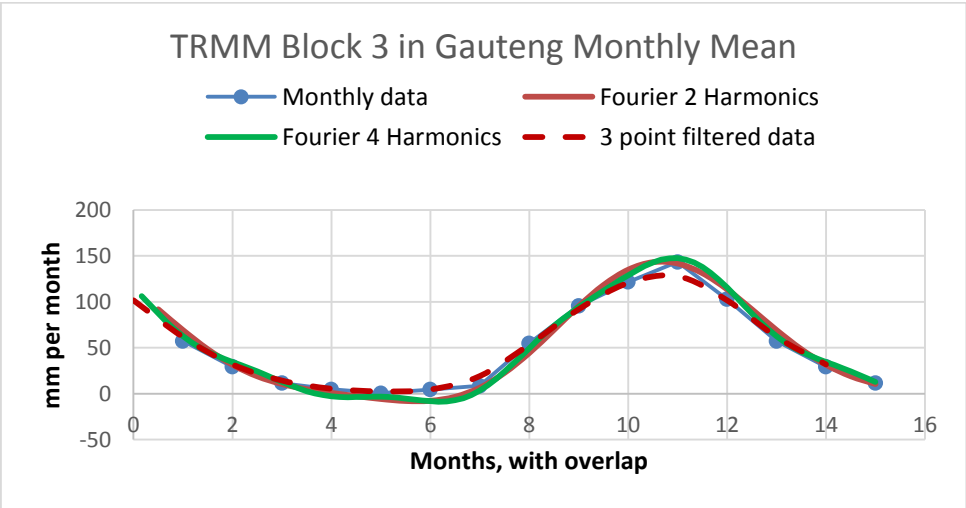


Figure 8.17. Example of fitting different functions to summary data. These are monthly means calculated from TRMM data obtained from Block 7 in Gauteng, using Fourier series and numerical filters.

We abandoned Fourier fitting because of the tendency to go negative as shown in Figure 8.17 and chose numerical filters instead. True, we could have used a logit transform to ensure positivity but decided to go a more direct route. Figure 8.18 shows the result of filtering standard deviations, each calculated from a given calendar day in the ten-year data-set. The smoother is a numerical filter of 31 days, which is longer than the largest gap of dry days in the record of observations.

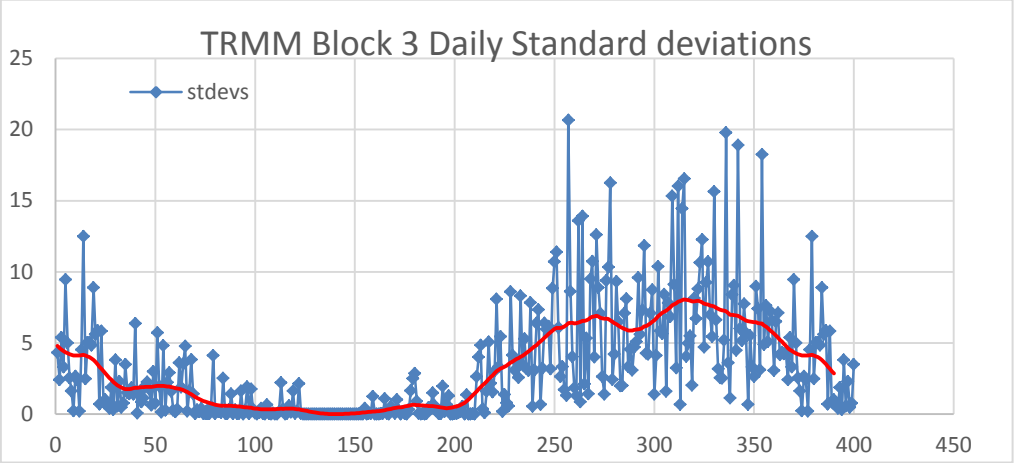


Figure 8.18. Example of fitting a triangular numerical filter to summary data. These are daily standard deviations calculated from TRMM data obtained from a block in Gauteng.

Figure 8.19 shows a part of the sequence of standardised daily data for year 1 of Block 3 in Gauteng, calculated using the filtered means and standard deviations of the individual blocks. This work was done before the idea of Gaussianisation was introduced as described in Chapters 3 and 4. Instead, not only were Pearson cccs calculated, we calculated Spearman rank correlations as well. It turns out that the correlations tend to be spuriously high because of the long dry periods of daily rainfall in South Africa.

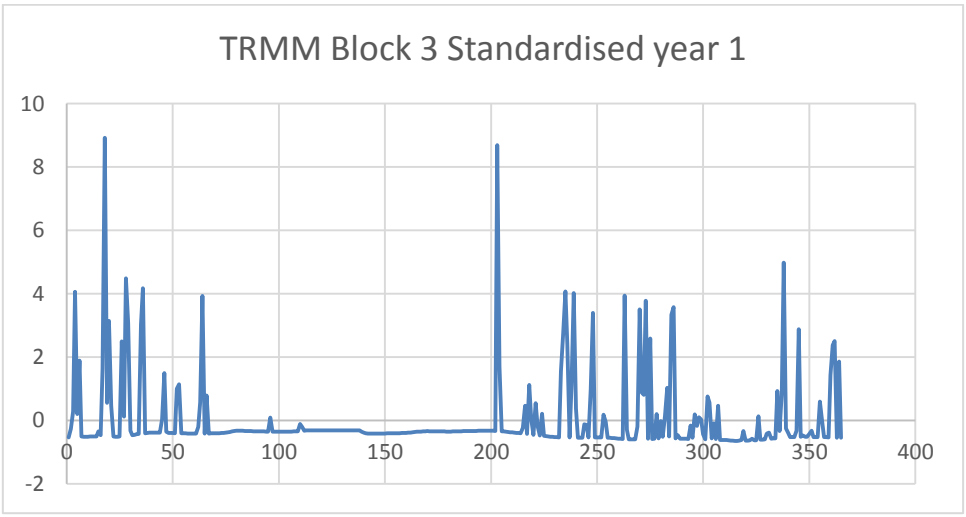


Figure 8.19. Year 1 of standardised daily data of Block 3 in Gauteng.

These standardised data were then used to calculate cross-correlations to determine a suitable grouping of daily data in order to inform us of the best way to proceed with downscaling.

In the following tables and figures we compare the Block Averaged Gauge Data (BAGD) with the TRMM data. One interesting observation was the number of dry days that were distributed over the 10 year period, as represented in the following table. As expected, TRMM is wetter than BAGD, recording much more light rain as indicated in Figure 8.11:

Table 8.1. Number of dry days per year in Block 3 in BAGD and TRMM data over the 10 years

number of zeros per year											
year	1	2	3	4	5	6	7	8	9	10	mean
BAGD	255	259	266	284	261	270	269	247	242	235	**258.8**
TRMM	199	203	204	221	217	217	215	234	228	243	**218.1**

From the corresponding standardised values, we calculated the year-by-year cccs and report them in Table 8.2.

Table 8.2. For block 3, cccs of the standardised daily data by year.

	Block 3 – Standardised Data Observation by year										
year	1	2	3	4	5	6	7	8	9	10	Mean
ccc	0.255	0.451	0.454	0.359	0.409	0.476	0.825	0.558	0.491	0.404	**0.468**

Notably, except for year 1, the coefficients are compatible. Year 7 is a surprise at 0.825 and a scatter-plot of that year 7 is given below in Figure 14. The high correlation is due to the 3 large standardised BAGD and TRMM values dominating the values of the lower TRMM values and the strong cluster of zeros, transformed as in the example shown in Figure 8.19.

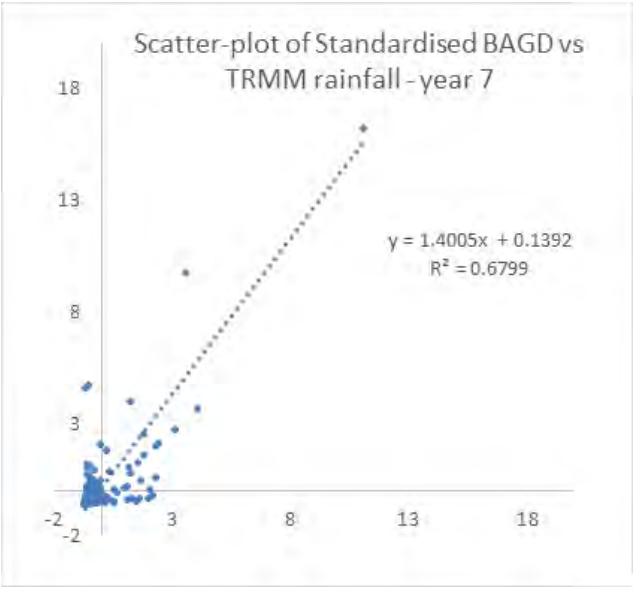


Figure 8.20. Plot of standardised daily data of Block 3 of the Gauteng group in year 7, r = 0.825.

Next, we summarise the cccs obtained from sets of data for the 4 regions as R^2 from the regressions of standardised data. They are the daily data for the period from March 1st, 2000 until February 28th, 2010. The total number of data points is 3 652 per dataset used for comparison.

Table 8.3: BAGD vs TRMM R^2 for all Standardised daily data

	Standardised Daily: Coefficient of Determination Results (R^2)			
Block Number	**Site 1: Gauteng**	**Site 2: KwaZulu-Natal**	**Site 3: Western Cape**	**Site 4: Limpopo**
Block 1	0.656	0.120	0.082	0.545
Block 2	0.638	0.079	0.189	0.333
Block 3	0.786	0.137	0.154	0.800
Block 4	0.784	0.193	0.093	0.664
Block 5	0.480	0.203	0.105	0.259
Block 6	0.686	0.197	0.132	0.548
Block 7	0.752	0.212	0.196	0.846
Block 8	0.507	0.194	0.021	0.514
Block 9	0.627	0.196	0.086	0.553
Average	**0.657**	**0.170**	**0.109**	**0.562**

The TRMM and BAGD datasets for all the blocks yield poor correlations between the datasets for the coastal regions (Western Cape and KwaZulu-Natal) but relatively good correlations for both Gauteng and Limpopo. We suspect that the long dry periods in the interior are the reason for this difference. The highlighted value of 0.189 for Block 2 is the highest for the Western Cape and the time series appears in part in Figure 8.23.

Here follow the means and standard deviations of daily data of Western Cape Block 2 in Figure 8.21, followed by the block (by month) standardisation of the daily data in Figure 8.22. This procedure overcomes the difficulties experienced with the filtered means and standard deviations. In Figure 8.23 is the plot of the standardised sets of BAGD and TRMM daily values.

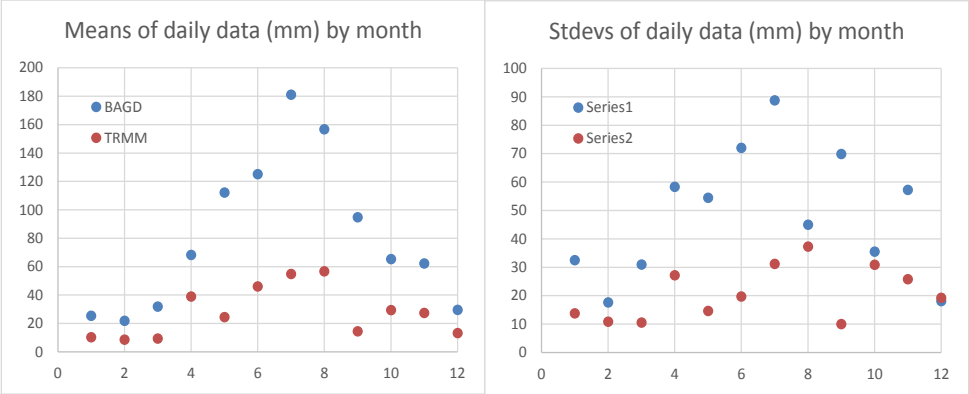


Figure 8.21. Means and Standard Deviations of daily data of Western Cape Block 2.

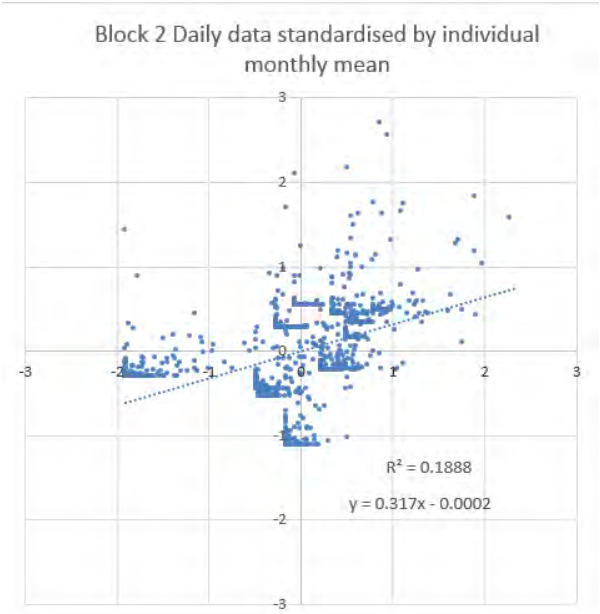


Figure 8.22. Standardised daily data of Western Cape Block 2.

The value of R^2=0.189 is reflected in Table 8.3, highlighted in yellow. Clearly there is a very tenuous link between the two sets as is exemplified by 1 year of their data plotted coaxially in Figure 8.23.

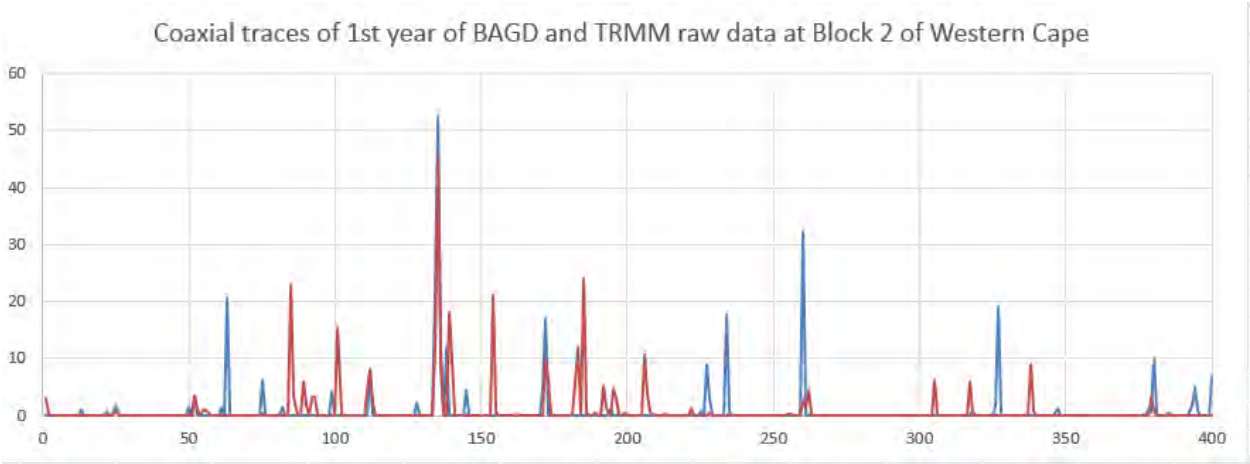


Figure 8.23. Coaxial traces of the 1st year of two sets of rainfall estimates for comparison.

Clearly there is a great deal of mistiming (as noted in Section 8.3) among the odd coincidental rainfall measurement as shown in Figure 8.23. This observation echoes the conclusions drawn in the discussion on Figures 8.9 and 8.10.

The data used to obtain the results in Table 8.4 are the pentad data for all 4 sites for the period from March 1st, 2000 until February 28th, 2010. The total number of data points is 730 per dataset used for comparison.

Table 8.4: BAGD vs TRMM R^2 for Standardised pentad data

Block Number	Standardised Pentads: Coefficient of Determination Results (R^2)			
	Site 1: Gauteng	Site 2: KwaZulu-Natal	Site 3: Western Cape	Site 4: Limpopo
Block 1	0.447	0.142	0.138	0.208
Block 2	0.489	0.170	0.263	0.283
Block 3	0.580	0.186	0.221	0.631
Block 4	0.554	0.258	0.171	0.473
Block 5	0.417	0.291	0.193	0.131
Block 6	0.542	0.372	0.202	0.213
Block 7	0.428	0.262	0.358	0.635
Block 8	0.419	0.247	0.060	0.477
Block 9	0.398	0.282	0.159	0.404
Average	**0.475**	**0.246**	**0.196**	**0.384**

The TRMM and BAGD datasets for the pentads in all the blocks except in Gauteng, show a poor correlation as measured by R^2, when compared to the daily data, particularly for Gauteng and Limpopo.

The data used below is the set of monthly data for the period from March 1st, 2000 until February 28th, 2010. The total number of data points is 120 per dataset used for comparison.

Table 8.5: BAGD and TRMM R^2 for Standardised monthly data

Block Number	Standardised Monthly: Coefficient of Determination Results (R^2)			
	Site 1: Gauteng	Site 2: KwaZulu-Natal	Site 3: Western Cape	Site 4: Limpopo
Block 1	0.075	0.127	0.137	0.275
Block 2	0.043	0.174	0.193	0.318
Block 3	0.230	0.228	0.288	0.394
Block 4	0.333	0.273	0.167	0.361
Block 5	0.276	0.238	0.138	0.096
Block 6	0.282	0.326	0.143	0.262
Block 7	0.351	0.288	0.339	0.187
Block 8	0.324	0.325	0.060	0.275
Block 9	0.222	0.346	0.131	0.333
Average	**0.237**	**0.258**	**0.177**	**0.278**

8.6 Summary

The bias, as measured by the regression lines of the raw (unstandardized) data is relatively low for the daily R^2 values (0.3 to 0.5) and is much improved for the monthly totals (0.7 to 0.9), no doubt due to the occasional large value. When it comes to the standardised values, the R^2 values improve in some areas but are rather inconsistent. This makes it difficult to suggest a global treatment of bias correction. In other words, the relationships are very site specific, so require great care in matching corrections to locations. In addition, the daily data do not yield helpful correction equations as the R^2 is low (see Figure 8.22 for example) so a plausible solution to the scaling problem is to use the monthly relationships to scale the daily values. Unfortunately, this ruse does not solve the problem on the mistiming of the daily TRMM totals when compared to the BAGD values – we are probably going to have to live with that.

Given the above, it is likely that TRMM data (and the output of its successor GPM) will be useful for large-scale hydrology and agriculture, particularly at the monthly scale, in contrast to daily. Thus crop monitoring and reservoir storage calculations will benefit, but not Flash Floods. The short conclusion is that TRMM is useful for hydrology in a coarse way, but poor in detail.

Chapter 9. A new idea for bias-correcting TRMM/GPM rainfall

9.1. Introduction

TRMM, and its successor GPM, yield measurements of daily rainfall which does not match well with ground-based raingauge estimates. There is therefore a felt need to bias-correct these spatially valuable products to be useable for hydrological purposes. Investigations reported in Chapter 8 have shown that there is a low correlation between daily block averaged gauge data [BAGD] rainfall and TRMM/GPM estimates [which for convenience we will call TRMM herein]. This fact implies that regression is not going to be valuable as a method of information transfer to enable bias-correction of TRMM, so we need to explore the usefulness of a direct Quantile-Quantile [QQ] procedure.

The QQ technique can provide a set of estimated cumulative distribution functions [cdfs] of the TRMM block rainfall values which closely match the cdfs of the BAGD estimates, but whose values will not necessarily be a good match on a day-by-day basis. The result is that the daily correlations between the TRMM and BAGD time series will likely remain low, at the same level as have already been determined, but we judge that with accumulations over several days, the amounts will match reasonably well. Where gauge rainfall data are available this QQ method should be straightforward to apply from location to location. The problem then arises: how does one bias-correct TRMM where there are no gauges in the TRMM block spaces? This problem is of particular concern in the SADC region outside our borders.

Turning to the problem of sparse raingauge data, especially in SADC countries, we need to develop a method of meaningfully interpolating rainfall estimates on the ground which are more realistic than the relatively biased TRMM estimates. We do not think that it is feasible to attempt to spatially infill rainfall fields on a given day from gauge data in data-sparse regions using regression-based methods like Kriging, as the daily rainfall spatial dependence structure has a correlation length of the order of 20 to 40 km depending on the character of the rainfall on the day – be it convective to stratiform. Thus the sparseness of gauge-sites in countries outside South Africa means that gauge data cannot be meaningfully interpolated because there is far too weak a spatial correlation to allow us to use standard interpolation methods if we use the amounts on a given day. There is a need to think differently.

The idea mooted herein is to interpolate the *parameters* of the gauge or BAGD rainfall probability distributions, rather than the *amounts*. The number of these parameters is 2 or 3, depending on the adopted distribution function that is used to fit the ranked historical wet amounts. The first parameter of importance is the local probability of a dry day, p_0. If an Exponential distribution is used for the wet amounts, then it is described by 1 parameter; a Weibull distribution has 2 parameters. The premise adopted here is that these parameters are likely to vary relatively slowly over a region with sharp changes in altitude so that where there is spatial discontinuity due to topography, we may be well advised to use external drift to incorporate the effects of altitude, but that might come later. At this stage, this document will address the idea of spatially interpolating the 2 or 3 parameters of fitted pdfs using Multiquadrics, i.e. Ordinary Kriging with a linear variogram.

9.2. The adopted rainfall pdf model

There are 2 candidate probability distribution functions we choose to fit to the wet amounts at a given location: Exponential and Weibull as mentioned above. These have been chosen because of their goodness of fit to these data when compared to other distributions including Lognormal, Gamma and Gumbel, which are popular for this purpose. These two distributions also come with the added benefit of ease of manipulation, as they are mathematically invertible using simple algebra. The cumulative distribution functions of $F(x) = p$ in terms of the variable x with parameters α and b, including the dry probability p_0, are:

$$\text{Exponential pdf: } F_1(x) = p = 1 - (1 - p_0)\, exp(-x/\alpha) \quad (9.1)$$

$$\text{Weibull pdf: } F_2(x) = p = 1 - (1 - p_0)\, exp(-x/\alpha)^b \quad (9.2)$$

with their inverses:

$$\text{Exponential pdf: } x = \alpha[ln\{(1-p_0)/(1-p)\}] \quad (9.3)$$

$$\text{Weibull pdf: } x = \alpha[ln\{(1-p_0)/(1-p)\}]^{1/b} \quad (9.4)$$

The consequence is that if we know p, the quantile determined from the TRMM estimate at a given block on a given day, we can immediately obtain x, the transformed rainfall value from one of equations (9.3) and (9.4). We first need to decide whether the more economical Exponential is a suitable choice and compare it with the Weibull distribution. For example, Figure 9.1 uses data from a BAGD site (no 9) in Gauteng.

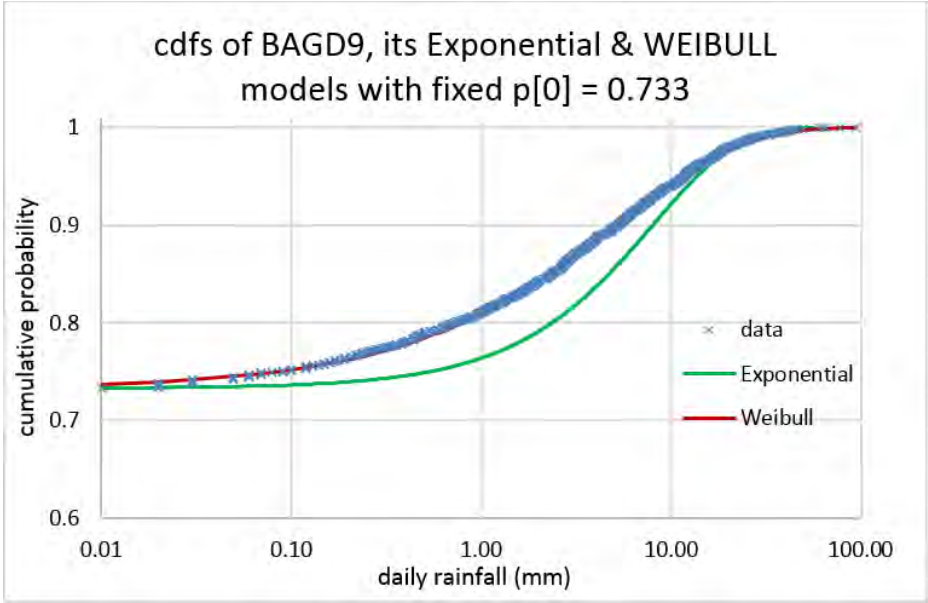


Figure 9.1. Cumulative Distribution Functions (cdfs) fitted to the daily data on Block 9 of the Gauteng group. Green line – Exponential model; Red line – Weibull model; Blue crosses [partially hiding the Weibull curve]: data.

There is no contest – in this case [as in others] the Weibull is worth the extra parameter to get an excellent fit – please note the divergence of the Exponential from the data in the vicinity of 1 to 10 mm of rainfall. The good fit of the Weibull can be seen in Figure 9.1 but

better in Figure 9.2 which shows the double quantile plot and the very good R^2. The parameters for the Weibull distribution fitted here are: a = 5.31 and b = 0.656; it is worth noting that the Weibull model's exponent b is significantly different from the Exponential model's default value of b = 1.

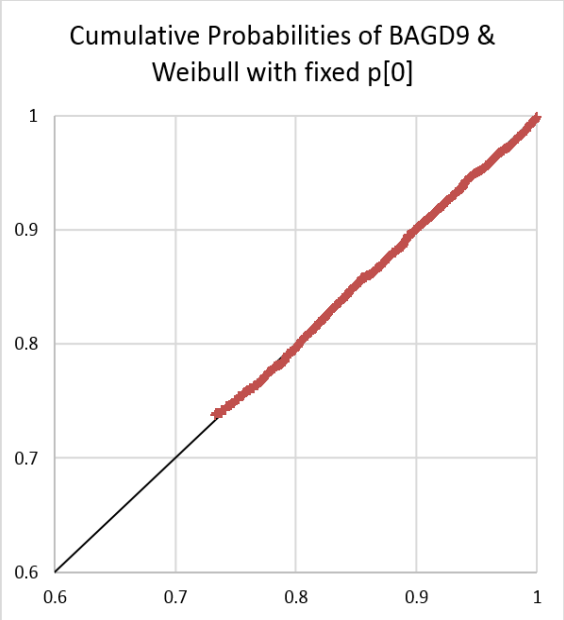


Figure 9.2. QQ plot of data and the fitted Weibull distribution shown in Figure 9.1

Figure 9.3 shows the cdf of the corresponding TRMM distribution of daily estimates on Gauteng's Block 9 and its fitted Weibull cdf; again a very good match.

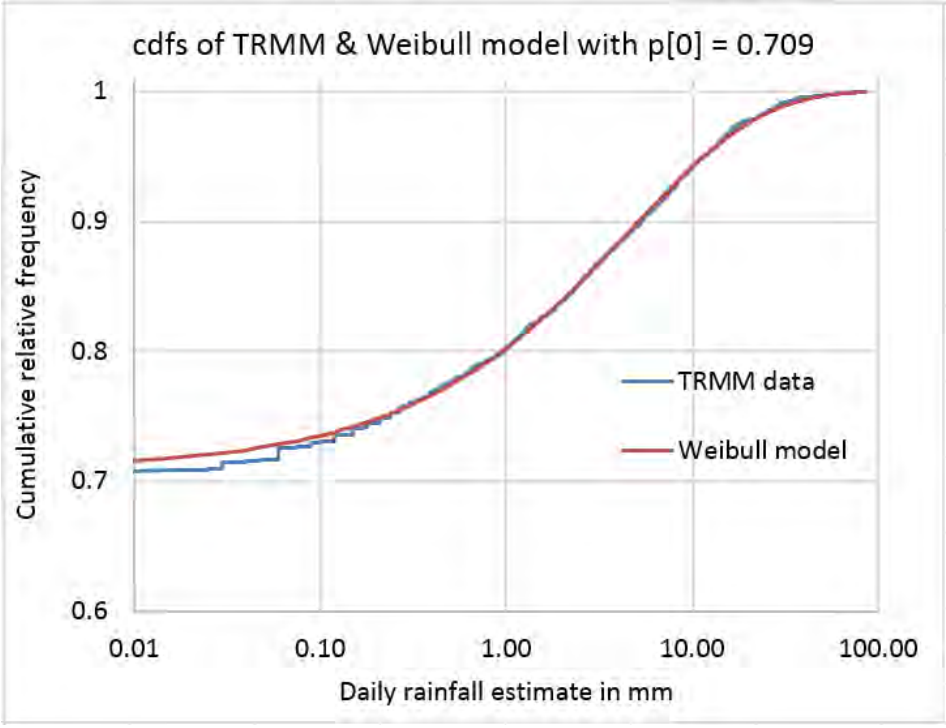


Figure 9.3. cdfs of the TRMM distribution of daily estimates on Gauteng's Block 9 and its fitted Weibull cdf

In Figure 9.4 we give a pictorial explanation of the QQ transform. The example uses the BAGD and TRMM daily values from Block 9 in Gauteng treated above.

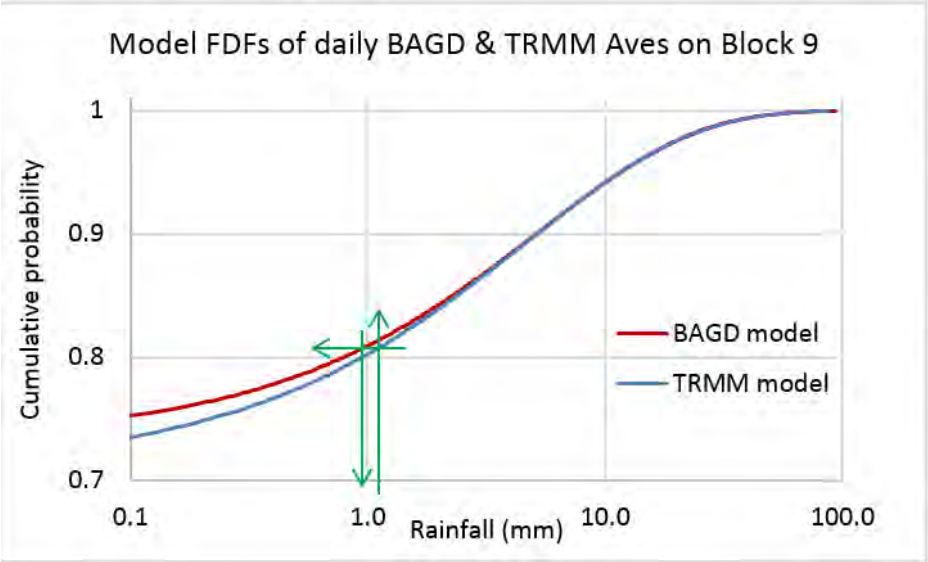


Figure 9.4. Sequence of calculations to perform a QQ transform of TRMM rainfall to Gauge. Blue curve: Weibull model fitted to TRMM as in Figure 9.3; Red curve: Weibull distribution fitted to the BAGD data.

The Blue line in Figure 9.4 is the Weibull distribution fitted to the TRMM data shown in Figure 9.3, but now truncated at a more meaningful 0.1 mm, with distribution parameters p_0 = 0.709, a = 4.66 and b = 0.619. The Red curve is the Weibull distribution fitted to the BAGD data with parameters p_0 = 0.733, a = 5.31 and b = 0.656, as fitted to the data in Figure 9.1.

There is a very close resemblance between the two curves, TRMM and BAGD, which is not always the case. Nevertheless, the QQ procedure is described as follows using Figure 9.4. If we wish to bias correct, via QQ transform, a value of 1.2 mm rainfall estimated using TRMM, the sequence of calculations follows the green arrows in the figure. The upward green arrow intersects the blue curve of the TRMM cdf at quantile 0.807 determined from (Eq. 9.2). Using this value of 0.807 in the Weibull pdf describing the red curve fitted to the BAGD data (Eq. 9.4) gives the transformed rainfall value as 0.97 mm, pictorially described by following the downward pointing green arrow.

That should work for an individual site where we have data from both sources. What to do where we do not have gauge data? The answer offered here is that we interpolate the three local distribution parameters of the BAGD data from the observed locations to the unobserved ones, using Multiquadrics, i.e. Kriging with a linear variogram.

9.3. Interpolation of parameters using Multiquadrics

Consider the set of BAGD data in the Limpopo region, summarised in the map in Figure 9.5. The maximum number of gauges in each 0.25° block is shown – there are many zeros indicating ungauged areas. In the numbered blocks, it is possible that there are less than

this number on any given day in the 11 years available for overlapping gauge and TRMM observations. The green coloured squares are where there are at least some days with observations; the heavy red border around the site with 6 gauges indicates a problem site which has a p_0 = 0.608, whereas the other coloured squares report p_0 values of about 0.8 and above. The example of interpolation conducted here is to determine what an interpolated value of p_0 would be at that square used as a target site.

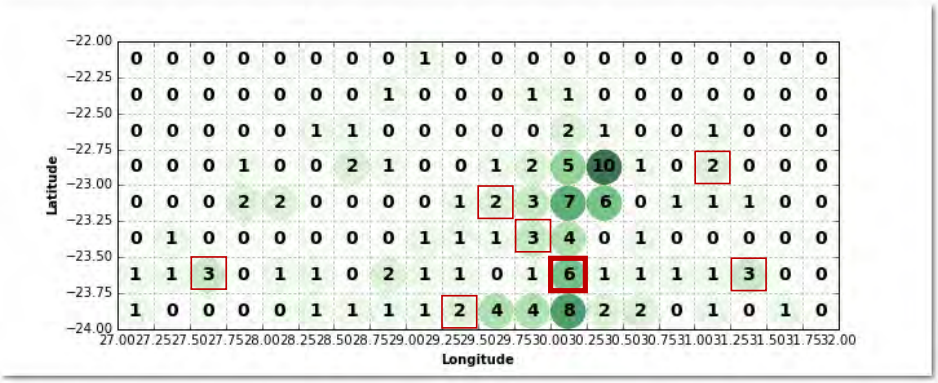


Figure 9.5. Number of active gauges in the Limpopo region from 2000 to 2010. The red squares indicate blocks used for the interpolation experiment. The thick red square includes the target block.

In the original gauge data-set, the 0.25° blocks [20 across by 8 down] were numbered from left to right starting with zero at the top left corner, so that the blocks in the far right column read (from the top) 19, 39, 59, 79, 99, 119, 139, 159. To compute distances between the blocks, we need a coordinate system, so we will number *columns* from left to right as 1, 2, ... 20 and the *rows* from the top down as 1, 2, ... 8. The *x* and *y* coordinates of the control squares with thin red borders (labelled 1 to 7, with the target square number 132 and labelled 5, highlighted yellow in Table 9.1) are those of the sequence number # counted from the top left corner, shown in Figure 9.5.

Table 9.1 Coordinates of control and target squares

	blocks chosen	from top left corner (1,1)	
Label	Sequence No. #	x	y
1	76	17	4
2	90	11	5
3	111	12	6
4	122	3	7
5	132	13	7
6	137	18	7
7	149	10	8

The mutual distances, in units of 1 block width, are given in the following matrix G in Table 9.2, highlighted in yellow. The control blocks are numbered in the first row and column, but omitting the target site, label 5 (or # 132). The final row and column of 1's and a zero, also part of the G-matrix, are highlighted in blue and are vectors of the Lagrange Multipliers ensuring that the derived interpolation coefficients sum to 1:

Table 9.2 Matrix G

	1	2	3	4	6	7	
1	0	6.083	5.385	14.318	3.162	8.062	1
2	6.083	0	1.414	8.246	7.280	3.162	1
3	5.385	1.414	0	9.055	6.083	2.828	1
4	14.318	8.246	9.055	0	15.000	7.071	1
6	3.162	7.280	6.083	15.000	0	8.062	1
7	8.062	3.162	2.828	7.071	8.062	0	1
	1	1	1	1	1	1	0

The inverse of the G matrix above, we call Ginv is given in Table 9.3:

Table 9.3 Matrix Ginv

-0.2038	0.0500	0.0387	0.0011	0.1401	-0.0261	0.2387
0.0500	-0.4123	0.3055	0.0371	-0.0307	0.0504	0.0220
0.0387	0.3055	-0.4947	-0.0160	0.0418	0.1248	-0.1026
0.0011	0.0371	-0.0160	-0.0750	-0.0046	0.0574	0.4862
0.1401	-0.0307	0.0418	-0.0046	-0.1795	0.0329	0.3293
-0.0261	0.0504	0.1248	0.0574	0.0329	-0.2395	0.0264
0.2387	0.0220	-0.1026	0.4862	0.3293	0.0264	-7.7965

The distances from the target block labelled 5 to the 6 control stations is given in vector $g(132)^T$, with a 1 in the 7th cell to match the Lagrange multiplier entries:

$$g(132)^T = [\ 5.000,\ 2.828,\ 1.414,\ 10.000,\ 5.000,\ 3.162,\ 1.000\ ]$$

Next we obtain the weights lambda(132) as a vector from the matrix product $g(132)^T$Ginv:

$$lambda(132)^T = [\ 0.0446, -0.0852, 0.6987, -0.0173, 0.1628, 0.1964, -0.0943\ ]$$

The two highlighted numbers which are slightly negative are due to the geometry of the gauge network. It is possible that these will produce negative rainfall if the corresponding gauges are wet, but the values are never large, so can be censored at run-time. In this application we are interpolating $p[0]$ values, all positive, so there is no problem. Dropping the last value -0.0943 in the above row, which relates to the Lagrange Multiplier, we use the vector of these 6 weights to vector multiply term by term with the observed values of the vector $\boldsymbol{p_0}$ at the 6 control points:

$$\boldsymbol{p_0} = [\ 0.8762, 0.8628, 0.7998, 0.8161, 0.8623, 0.8167]^T$$

Thus we finally obtain the estimate of p_0 at our target point at station 5 (# 132) as 0.811, quite different from its originally estimated value of 0.615.

On investigation, it turned out that one of the 6 rainfall gauges in the target block # 132 was acting quite differently from the others as shown in Figure 9.6, and it seems to have pulled down that block's p_0 value:

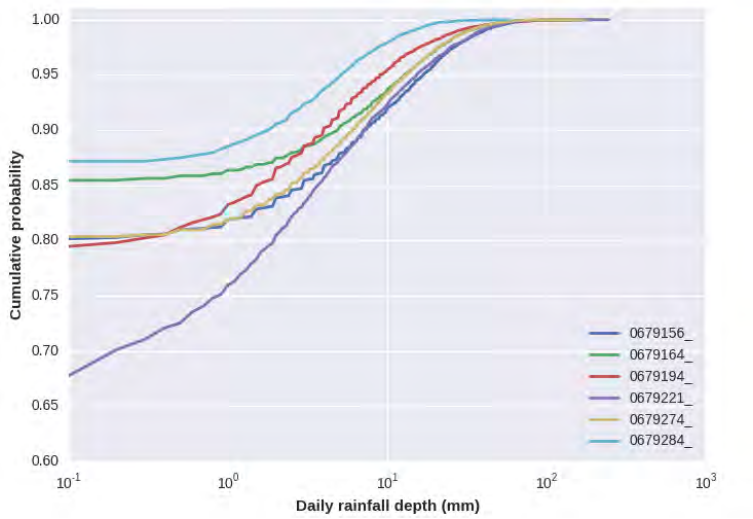


Figure 9.6. cdfs of the 6 individual rainfall stations active in the above target block.

We now try another example, where there are no data. We'll 'infill' the target's p_0 value, choosing site # 0 whose local coordinates are (1,1), i.e. at the top left corner of the Limpopo region shown in Figure 9.5. Its distance vector to the 6 control points, augmented by the 1 in cell 7, is:

$$g(0) = [16.28,\ 10.77,\ 12.08,\ 6.32,\ 18.03,\ 11.40,\ 1.000\ ]^T$$

and after pre-multiplying Ginv with this vector we get, removing the 7th entry:

$$lambda[0] = [\ 0.1622,\ 0.3415,\ -0.0830,\ 0.8078,\ -0.1059,\ -0.1225\ ]^T$$

and multiplying this, term-by-term, with the $\boldsymbol{p_0}$ vector above, we get: $p_0 = 0.8382$ at location [0], which is very close to the average p_0 value of the 6 control stations: 0.8390, which seems to make sense.

9.4. Interim summary

If all goes well, the procedure described above can be now used to fill in all the missing values of p_0 in the region. After that, using the same procedure for determining the p_0 values at the unobserved locations we developed in Section 9.3, we can estimate by interpolation the parameters *a* and *b* of the Weibull distribution in the places where we do not have data. This will enable us to build plausible surrogate distributions at these ungauged places, so that we can bias-correct TRMM/GPM data, as if we had rainfall data at these sites, in a meaningful way. This procedure is more sensible than using regression, which usually biases estimates downwards when the correlations are low.

It has to be admitted that this methodology has been applied above in a simple way to only two points – one with suspect data the other where there are no data – but this sets the stage for a more thorough extension of the method to the many blocks in the 4 regions, some as validation, the others as infilling.

To summarise the procedure:

(i) find the p_0 values and the 2 Weibull parameters of the BAGD data where they are available in gauged blocks over the region, then

(ii) cross-validate the parameters and see if the blocks with strange behaviour can be repaired, then

(iii) use the good data to infill the empty blocks with the 3 parameters [p_0, a & b], then

(iv) bias-correct TRMM in all the blocks by QQ transform.

Having proved its worth, the idea can then be used to estimate ground rainfall from TRMM (or GPM) using QQ transforms on a day-by-day basis, after interpolating Weibull distribution parameters over the ungauged regions of RSA and outside our borders.

9.5. A caution: reconsider the above procedure based on other information

During an October 2015 visit to Stuttgart with for research collaboration with Prof Andras Bardossy, Pegram found that things are not as simple as have been recorded above. Although in principle the methodology set out above is relatively sound, there are four corrective issues which need to be dealt with before carrying on with the work of TRMM bias correction.

9.5.1: The first important issue is that the two Weibull parameters are strongly spatially correlated with each other once they have been estimated at gauged locations. They cannot be meaningfully interpolated independently as suggested above. If this procedure was used without modification there would be some instances where the distributions would likely not make sense.

9.5.2: The second important point is that the spatial interpolation method of Multiquadrics turns out to be flawed in the way it was used until now to estimate the daily average spatial rainfall on the blocks from the gauges. The procedure used (based on Pegram and Pegram, 1993) did not include the constraint of the Lagrange multipliers described in the treatment in Section 9.3 above. This omission has the result that the averages calculated are often warped from the true block means. A good example of this is the estimation of the spatial average of the daily rainfall in Limpopo Block # 5, where it might be inferred in Figure 9.5 that the average p_0 value is above 0.7, whereas it was calculated from the block averaged data as 0.615.

9.5.3: The third point is that, sadly, it is not even as simple as described in the example given in the preceding sub-section 9.5.2, because on the very driest of wet days, only one gauge out of the 6 will be recording, probably gauge # 0679221. This gauge's pdf is still dropping in the region of 0.1 mm/day, whereas all the others have flattened off to dry, many below 0.5 mm/day, as shown in Figure 9.6. The take-home point here is that the more gauges there are in a block, it is more likely that the calculated p_0 will be lower. In the experiment above in Section 9.3, the block averages calculated at target block 5 were based on control blocks containing only from 2 or 3 gauges; thus the sparsely populated blocks were used to check the value of the one with 6 gauges. Therefore it is important to determine the behaviour of block averages of a range of numbers of gauges on a block. These are then to be extensively sampled from (and then have their block averages

compared to) a set of full spatially simulated daily surfaces of rainfall over the block, to be described in Section 9.6. This approach will then give us a sensible scaling procedure to correct the block estimates from 1 or more gauges inhabiting the block.

9.5.4: The fourth important issue, which springs from the first, described in sub-section 9.5.1, is to find the correct method of interpolation of the Weibull parameters. Figures 9.7 and 9.8 with the text describing them, are from a PhD thesis (supervised by Prof Bardossy) authored by Hans-Henning Lebrenz (2013), entitled 'Addressing the input uncertainty for hydrological modelling by a new geostatistical method'. In it, he models monthly spatial rainfall over a region of South Africa. His pilot study area is shown in Figure 9.7 and encompasses our Gauteng Block in Simon Ngoepe's MSc study, shown about middle left.

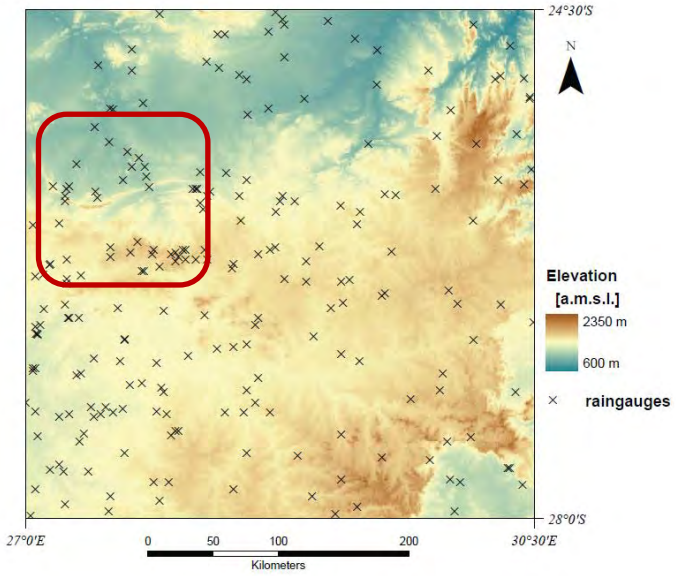


Figure 2.2: Details of the pilot area.

Figure 9.7. Lebrenz's Pilot Area for monthly interpolation of parameters, with Ngoepe's Gauteng study area.

Lebrenz showed that the mean and standard deviation, sampled monthly from 226 gauges over 22 years, have the structure shown in Figure 9.8 following. Although there is strong correlation between mean and standard deviation, there is no correlation between the data points in the space of the principle components *r* and *s* achieved by transformation of the data to these axes. Usefully, there is a unique linear relation between the pair of means and variances on the one hand and the principle components on the other.

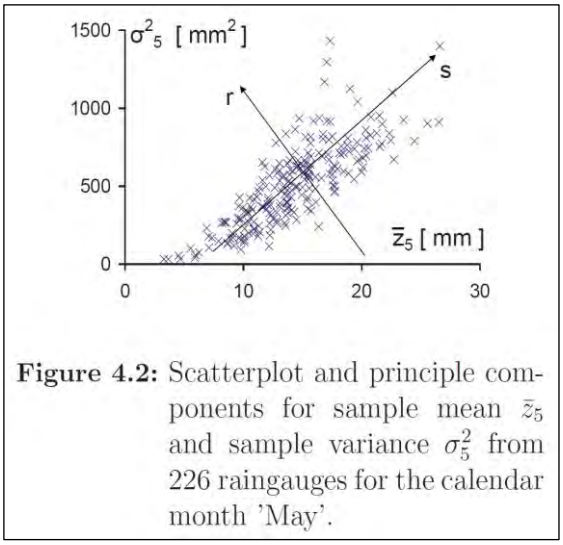


Figure 4.2: Scatterplot and principle components for sample mean $\bar{z}_5$ and sample variance σ^2_5 from 226 raingauges for the calendar month 'May'.

Figure 9.8. Lebrenz's Figure 4.2

Taking this idea further, the advantage of this procedure is that we can separately and independently interpolate the uncorrelated *r* and *s* values, determined by this transformation from the Weibull parameters, of the correctly estimated gauge block averages [see subsection 9.5.3 above]. Then at the interpolated locations, we can recover the Weibull parameters from *r* and *s* to obtain the necessary interdependence between them. To achieve the interpolation, we would then use exactly the same Multiquadrics procedure we proposed in Section 9.3, but with safety.

Turning to our daily rainfall data, for the purpose of discussion let us assume that we have determined the Weibull *a* and *b* parameters of 100 blocks and they plot as in Figure 9.9.

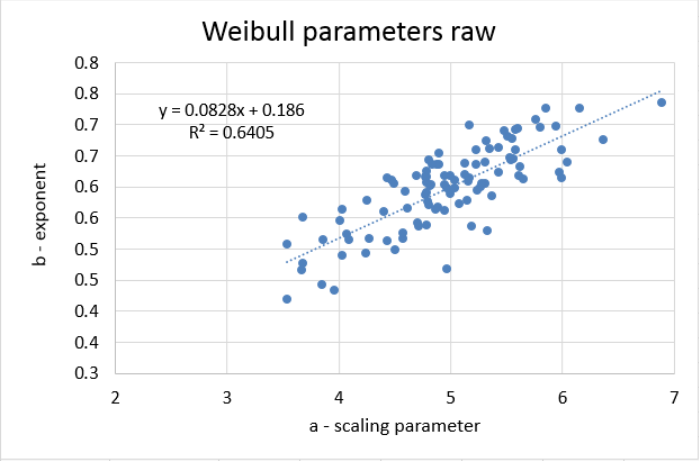


Figure 9.9. Plot of Weibull *b* versus *a* before transformation to uncorrelated *r* and *s*.

There is evident structure here and the cross correlation coefficient (ccc) between *a* and *b* is 0.8003. To transform these parameters to independent values like *r* and *s* we first need to standardise each sequence by subtracting its mean and scaling by its standard deviation given below in Table 9.4:

Table 9.4: Means and standard deviations of *a* and *b*.

	a-scaling	*b*-exponent
Means	4.977	0.598
stdevs	0.650	0.067

Hence, we obtain the plot of *a*1 and *b*1 in Figure 9.10, the standardised versions of *a* and *b*:

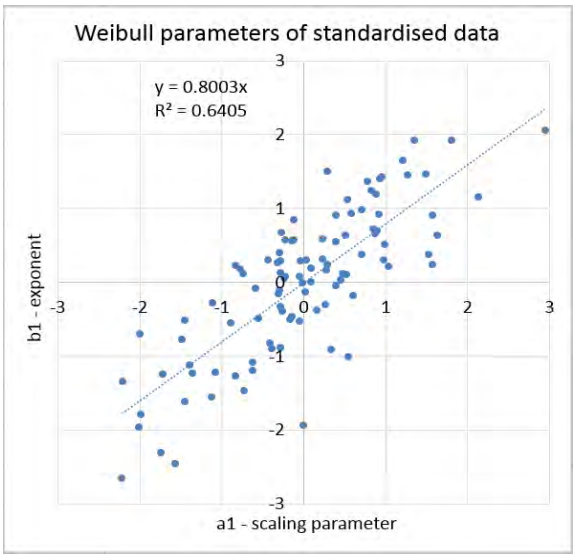


Figure 9.10. plot of standardised Weibull *b*1 versus *a*1

There is still a ccc of 0.8003 between the standardised variables, as there should be. We need to decorrelate this set of parameters *a*1 and *b*1 which form the 2 columns in the matrix *D*, which has 100 rows, to the set *a*2 and *b*2. Decorrelation is performed using the following procedure.

Let *R* be the 2 by 2 correlation matrix between the standardised parameters *a*1 and *b*1. It has 1s on the main diagonal and *r* = 0.8003 on the off-diagonal:

$$R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \qquad (9.5)$$

We need the inverse square root of this matrix to decorrelate the vectors of pairs of parameters. The sequence of calculations goes as follows. We first find the 'square root' matrix *P* in the following relationship:

$$R = P^2 = \begin{bmatrix} p & q \\ q & p \end{bmatrix} \begin{bmatrix} p & q \\ q & p \end{bmatrix} = \begin{bmatrix} p^2 + q^2 & 2pq \\ 2pq & p^2 + q^2 \end{bmatrix} \qquad (9.6)$$

By comparing the elements in the two versions of *R* in (9.5) and (9.6), we obtain

$$p^2 + q^2 = 1 \; and \; 2pq = r.$$

Substituting *r/2p* for *q* in the first of these equations, we get a quadratic in p^2:

$$p^2 + \frac{r^2}{[4p^2]} = 1$$

and solving p^2 we get

$$p = \left\{\left[1 \pm \sqrt{1 - r^2}\right]/2\right\}^{1/2}$$

So that we choose:

$$p = \left\{\left[1 + \sqrt{1 - r^2}\right]/2\right\}^{1/2} \quad (9.7)$$

and

$$q = \left\{\left[1 - \sqrt{1 - r^2}\right]/2\right\}^{1/2} \quad (9.8)$$

which define *P*, the square root matrix of *R*. For decorrelation, we want the inverse of the matrix *P,* which we will call $Q = P^{-1}$, which after a bit of manipulation can be obtained as:

$$Q = \begin{bmatrix} p & -q \\ -q & p \end{bmatrix} / \sqrt{1 - r^2}$$

This matrix is determined by *r*, after substituting for *p* and *q* from (9.7) and (9.8). In our case *Q*, turns out to be

$$Q = \begin{bmatrix} 1.4907 & -0.7454 \\ -0.7454 & 1.4907 \end{bmatrix}$$

If we now post-multiply the matrix *D* [formed by the pair of vectors of the standardised Weibull parameters shown in Figure 9.9] by *Q* we get 2 decorrelated vectors *a*2 and *b*2, whose pairs are plotted in Figure 9.11.

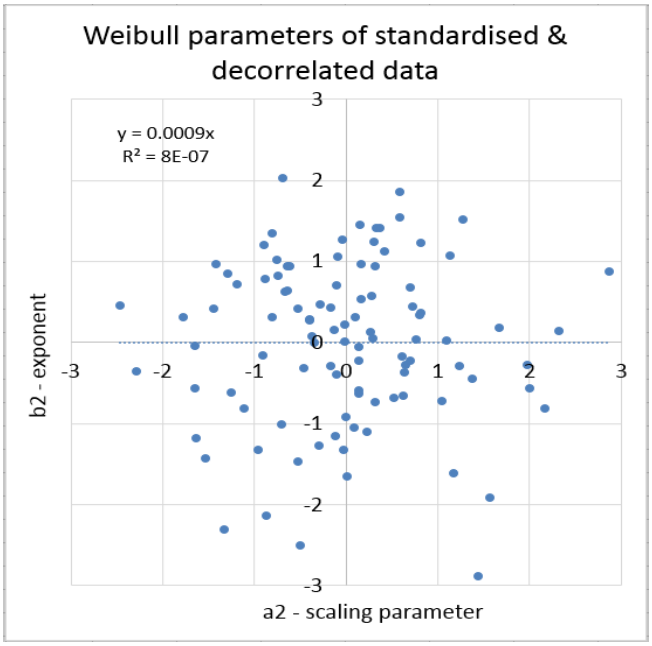


Figure 9.11. Decorrelated vectors *a*2 and *b*2 of standardised parameters of Figure 9.10

Note that the correlation between these vectors $a2$ and $b2$ is 0.0009, essentially indicating complete independence of these two variables.

This means that we can comfortably interpolate the transformed parameters one set at a time using $a2$ and $b2$, then reverse-transform each pair to mutually correctly paired Weibull parameters a and b at the target sites. This reverse transformation is performed by post-multiplying the pair of vectors $a2$ and $b2$ by P, the square root of R obtained above, using equations (9.7) and (9.8) to substitute for p and q. The newly interpolated parameters a and b are recovered by using the means and standard deviations in Table 9.4, by scaling using the standard deviations and adding back the means. These parameters are then to be used in the QQ transform of TRMM at the gauged and ungauged sites. However, there is more to do before completion of the QQ procedure.

In the discussion around Figure 9.6, it was suggested that one gauge had pulled down the p_0 value for the block from 0.8 to 0.6. It turns out that the answer is not that simple and the conclusion drawn was incorrect. First of all, the lowest cdf value of the supposed culprit at 0.1 mm is above 0.6 and it is clear that the average of the values of all 6 gauges at 0.1 mm is above 0.7. Thus there is another task to perform before we can complete the task of QQ-ing TRMM and that is to deal with the problem of the proper estimation of the distributions of interpolated gauge block means highlighted in sub-section 9.5.3. The problem is more subtle than we thought and needs to be dealt with in a more rigorous manner, as follows.

9.6. The point to area transform of daily gauge rainfall to properly QQ TRMM

The purpose of determining the gauge block averages of the rain gauges is to provide ground referencing to obtain QQ transforms of TRMM's block estimates of rainfall. It turns out that if we simply average the cdfs of 1, 2 or more gauges over the block, we are ignoring a subtle but important fact. The averages of a small number of gauges' daily catch can be very different from the true block averages of rainfall and our estimates need to take this fact into account.

9.6.1. An experiment to determine the link between a true rainfall field and averaged gauges.

In a numerical experiment conducted by Prof Bardossy [private communication], a set of ten thousand daily images of rainfall on an area 25 km square [the size of a TRMM pixel/block] were generated. Each 1 km pixel on the square was populated with properly spatially correlated 'rainfall', generated by a Fourier transform and using a fixed Exponential distribution for the amounts, for each set. The correlation length of the spatial variogram was set at 20 km, so the generated rainfields were relatively variable, similar to fields of convective rainfall.

The 'true' block average on each image was calculated by numerically averaging all the pixel values on each 'day'. Then a set of evenly spaced points was carefully selected from each field as if they were gauge locations; the numbers of sites per image chosen were 1, 2, 4, 8 and 16 and the location of each site chosen was kept the same for the set of 10 000 estimates ('days'). Numerically averaging the individual 'daily' gauge samples and using

these to compute their cfds yielded sets of gauge block average estimates to compare against the distribution of the 'true' spatial averages of the set of full fields. The numerical averaging of the gauge values was done by calculating their simple mean, but because they were equally spaced, their averages would match those of Multiquadrics. The following results are thus not only illustrative but very useful.

The cumulative frequency distributions (cfds) of the range of populations of gauge averages in the square shown in Figure 9.12 are somewhat surprising and a summary of these follows. Note that the lower bound of the vertical axis in the figure has been set at a probability of 0.4 [4000 days] to help in visualising the differences. We truncate the lower estimates of 'precipitation' at 0.1 mm. The obvious reason is that the value of p_0 at a threshold of 0.1 mm depends heavily on the number of gauges when they are few, such as 1, 2 or 3, as shown by the blue, green and brown curves. There is not much difference between cumulative cfdfs when 8 (red curve) or 16 gauges (black curve) are considered and the convergence of the latter to the 'true' (navy blue) curve derived from 625 sites is quite good.

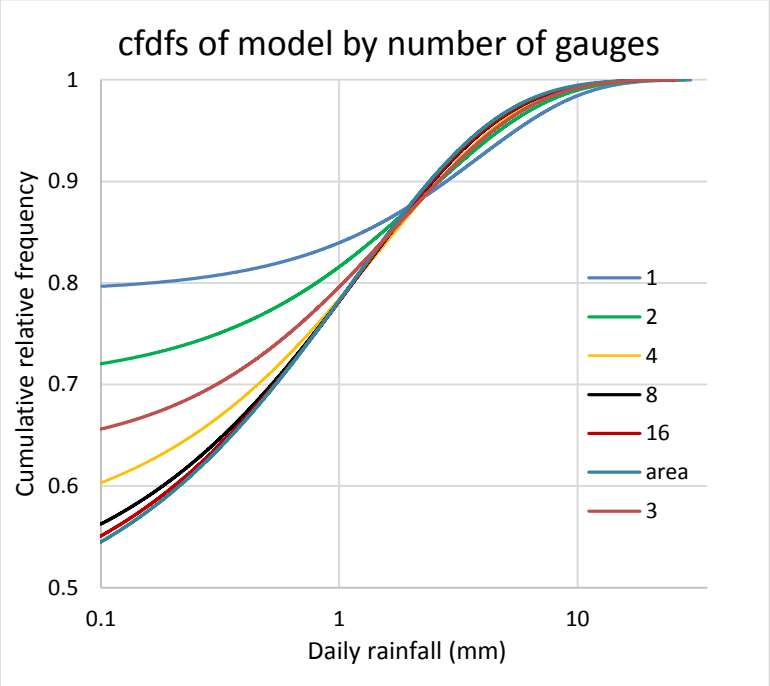


Figure 9.12. Cumulative frequency distributions of block averages of rainfall above 0.1 mm on gauges over a 25 by 25 pixel square in 10 000 days. Blue: 1 gauge; green: 2 gauges; brown:
3 gauges; yellow: 4 gauges; black: 8 gauges; magenta: 16 gauges; navy blue 625 sites (full square).

If we were to threshold the curves at 1 mm, as shown by the vertical axis (because below that measurement is technically a 'trace', which suggests that in that interval there is poor sampling of fine drizzle), then 4 gauges or more are quite adequate for a good areal estimation. A comparison of rainfall amounts in the vicinity of 0.95 (9500 days out of 10 000) shows that the 1-gauge curve reads 4.3 which overestimates the areal daily rainfall of 3.3 by about 1 mm, but that the curves of 4 gauges and above are quite faithful in their

estimates of the true values. It seems we only need to be concerned when there are between 1 and 3 gauges used to get average daily rainfall on a 625 km^2 area. A full justification of these procedures is given in Section 9.6.2.

In a sense, the paradox of Figure 9.6 and the estimation of the p_0 from the average has been resolved by Figure 9.12. The reasons can be summarised as follows.

- The mismatches of p_0 estimates previously obtained as described in vpp 9.2 are partly because of the comparison of results from a small number of gauges versus a larger number
- The threshold of the lowest block averaged precipitation on a day should be set to 1 mm which means that only in some limited cases (1, 2 or 3 gauges active in a block) do we need to perform a bias correction of the cfdfs of sparsely gauged blocks

9.6.2. Performing the bias correction of gauge readings to a gauge block average

The purpose of this sub-section is to provide a method of correcting areal averages obtained from a few gauges to what should be a true areal block average. It contains the detail of the procedure to produce Figure 9.12. As described in section 9.6.1, a set of 10000 'days' of correlated rainfall was generated over a 25 pixel-square area, which was sampled at 1, 2, 4, 8 & 16 locations as if these were gauges. For each of these sets scatter-plots were compiled. Figure 9.13 shows the results for 2 and 8 gauges per block.

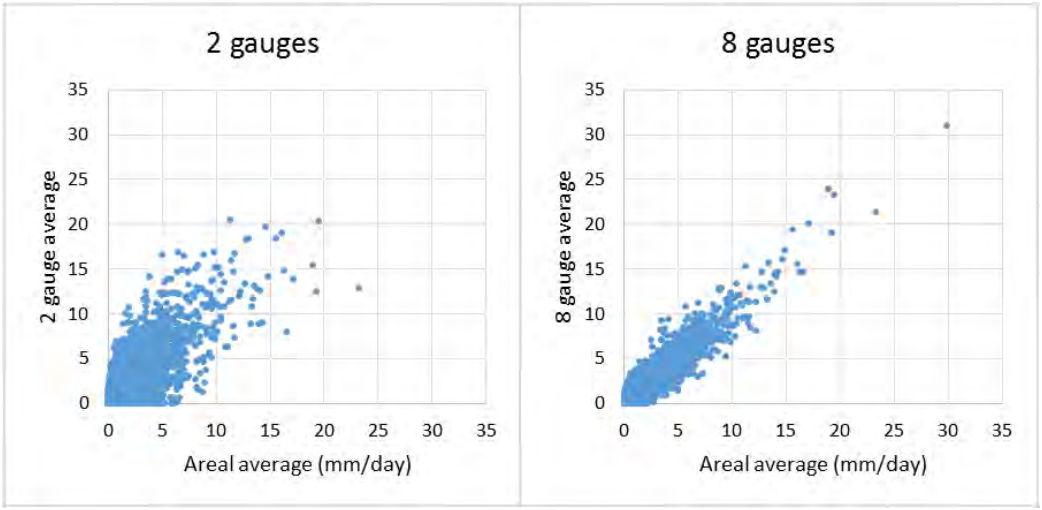


Figure 9.13. scatter-plots of gauge-averages and spatial areal averages of rainfall simulations over a 25 by 25 pixel square region.

Figure 9.14 shows the cfdfs of the gauge averages and the areal field averages plotted coaxially, for 2 and 8 gauges per block. The two orange curves are the gauge average cfdfs which differ and the black curves are the cfdfs of the 'true' areal averages which are the same in both panels. Note the improvement with more gauges recording in the block.

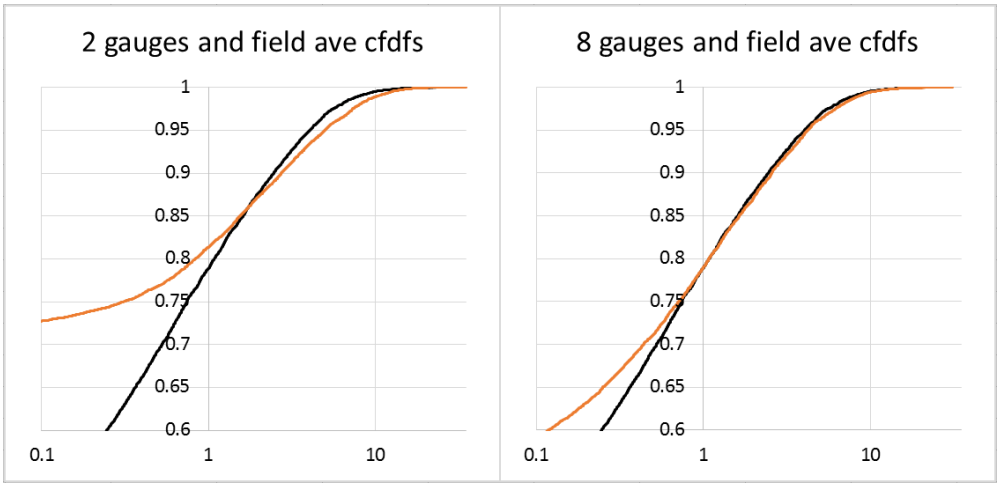


Figure 9.14. The cumulative frequency distribution functions of gauge averages [orange] over the 625 block square compared to the field's average [black].

We now fit a Weibull distribution model [Equation 9.2] to each of the gauge cfdfs in order to smooth the curves for comparison purposes. The fitted distributions match the samples relatively well, particularly above the 1 mm threshold, the 8 gauges better than the 2 gauges in the left panel of the figure. Figure 9.15 following shows the goodness of fit of the Weibull distribution to the generated samples.

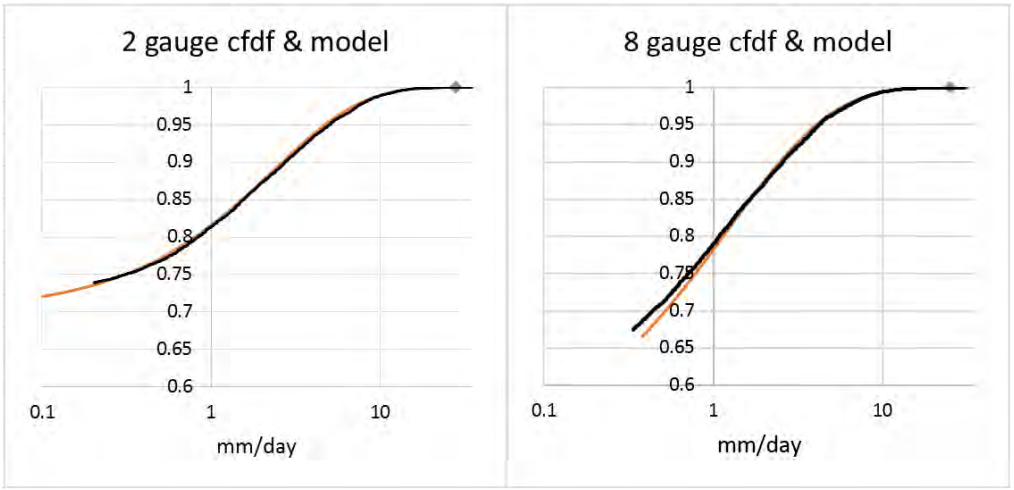


Figure 9.15: The sample and fitted Weibull distribution functions for 2 and 8 gauges. The black curves are the samples' cfdfs and the orange curve the fitted distributions. [Horizontal axis mm and vertical axis cumulative probability.]

The curve on the right of Figure 9.15 for 8 gauges, does not fit too well below 1 mm, an exception from the remainder of the curves, but the effect of that will be negligible in practice. The blue diamonds indicate the maximum values estimated by the fitted distributions.

Figure 9.16 is a plot of the Weibull distributions' parameters as the number of gauges increases. A value of b = 1 [on the grey curve] indicates that the Weibull simplifies to the

Exponential distribution, which happens only for the case of 1 gauge in the area, which confirms that the generated distribution is indeed exponential, by design. The averages of more than one gauge yield genuine Weibull distributions as observed in Figure 9.15 above and indicated by the grey curve of *b* in Figure 9.16.

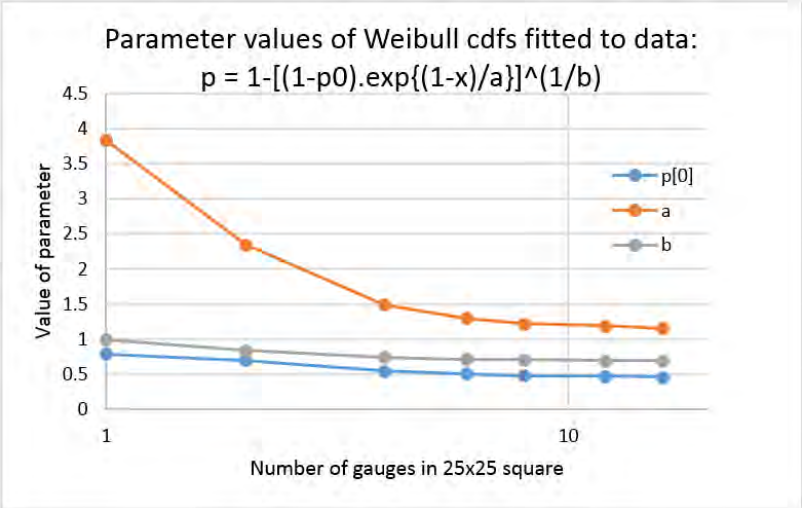


Figure 9.16. Parameter values of Weibull distribution functions fitted to sample curves as in Figure 9.15, for 1, 2, 4, 6, 8, 12, & 16 gauges. [blue: p_0; orange: a; grey: b]

The collection of modelled cfdfs obtained from the fitted Weibull distributions is given in Figure 9.17, using the parameters displayed in Figure 9.16.

In figure 9.12, it is clear that above 1 mm per day, there is no material distinction between the gauge block averages and the 'true' average, as long as there are 4 or more gauges active on a given day. We need to put some energy into transforming areal estimates from 1 and 2 [maybe 3] gauges. Figure 9.17 achieves this.

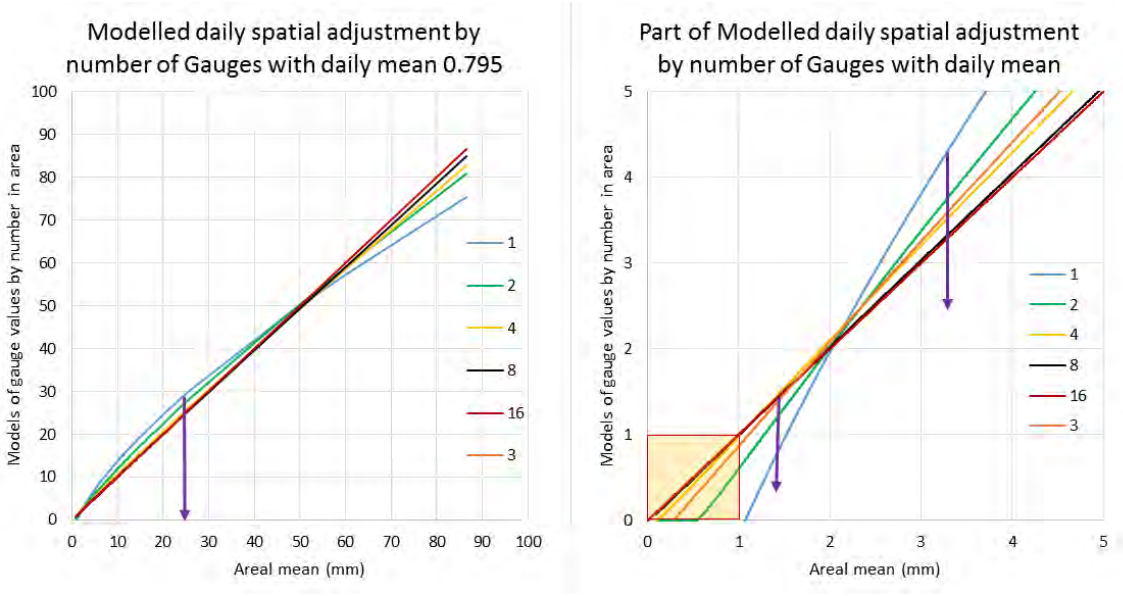


Figure 9:17. Ensemble of gauge-averaged distributions plotted against the areal average, for matching exceedance probabilities. The purposes of the purple arrows and the small orange rectangle are described in the following text.

The curves in Figure 9.17 are what are to be used to transform rainfall estimated from a few gauges on a 625 km^2 block to reasonable estimates of the true block averaged on each day. The purple arrows intersect all curves. If a single gauge is reading rainfall on a day, let us say 28 mm as shown in the left panel of Figure 9.17, then the gauge average is given where the arrow meets the horizontal axis at 24 mm. The image on the right is a blow-up of the lower values of the image on the left, showing the detail for 5 mm rainfall and less. The same rule applies to the purple arrows. If 2 gauges [green curve] average 1.2 mm as shown by the left arrow in the right panel, its value should be increased to 1.4 mm on the horizontal axis; the second purple arrow in the right image, a single gauge [blue curve] reading 4.3 should have its estimate reduced by 1 mm to 3.3 mm, confirming the remark following Figure 9.12.

The orange square in the bottom left of the right image in Figure 9.17 indicates the region where we propose there is no transform, but that the estimated areally averaged gauge rainfall for the day be set to zero. We consider it prudent that the lower limit of the transformed rainfall be set to 1 mm, which by default applies to all situations where only 1 gauge is estimating the areal average. However, if a trace is to be recorded, the transform might be set to a lower limit of 0.5mm, but not much less, because that will likely be nonsense. This would apply to all situations where 2 or more gauges are averaged.

For 4 gauges and above, adjustment appears not to be required and the Weibull parameters for the 3-gauge case can easily be read off Figure 9.16. In any case, (i) if the gauge average is x and the parameters of the gauge average are p_0, a and b and (ii) those of the areal average χ are π, α and β, then we can first calculate $p = 1 - (1-p_0)\, exp(-x/a)^b$ and then obtain the areal estimate as $\chi = \alpha[ln\{(1-\pi)/(1-p)\}]^{1/\beta}$.

Alternatively the areal average is calculated directly from x as:

$$\chi = \alpha\left[ln\{(1-\pi)/[(1-p_0)\, exp(-x/a)^b]\}\right]^{1/\beta}. \qquad (9.9)$$

9.6.3. QQ transforming TRMM observations to Gauge Block Distributions.

The above treatment achieves the first part of the transform, in that it yields the distribution functions of the block averaged gauge values, as if they were true areal averages. These distribution function parameters $\{\pi, \alpha$ and $\beta\}$ of the areal averages, where we have observations, must next be interpolated over the region where we do not have gauges. This is achieved by using the methodology developed in Section 9.5.

The aim is to determine the Weibull distribution parameters for all TRMM observations at all blocks in the region of interest, or alternatively create look-up tables of values and matching TRMM quantiles at each block. Whichever is more handy, on a given day, at a TRMM site where there has been rainfall, we want its quantile, determined as a cumulative probability from that site's historical distribution.

Therefore, at this stage of the TRMM bias correction exercise, we have devised procedures which are able to:

1. Estimate areally averaged rainfall at all blocks containing gauges.
2. Interpolate the Weibull distribution parameters of the gauge block averages $\{\pi, \alpha$ and $\beta\}$, which will have been established at each ungauged block using Multiquadrics after canonical decomposition as described in Sections 9.4 and 9.5.
3. Obtain the TRMM frequency distributions for each block in the region of interest.

Thus, on a given day, we can take a chosen TRMM block's rainfall estimate, then determine its quantile from the TRMM frequency distribution at the block in question. We next perform the QQ transform as in Figure 9.18, where the two probability distributions [Weibull] of the TRMM and areally averaged data are known. We can again apply Equation (9.9), but now with the appropriate parameters of TRMM quantile and gauge block average estimate. In this case in Equation 9.9 we make a new substitution of the symbols: the x is the TRMM value, the p_0, a and b are the TRMM cdf distribution parameters, the π, α and β are the gauge areal average Weibull parameters and χ is the transformed, rescaled TRMM rainfall estimate.

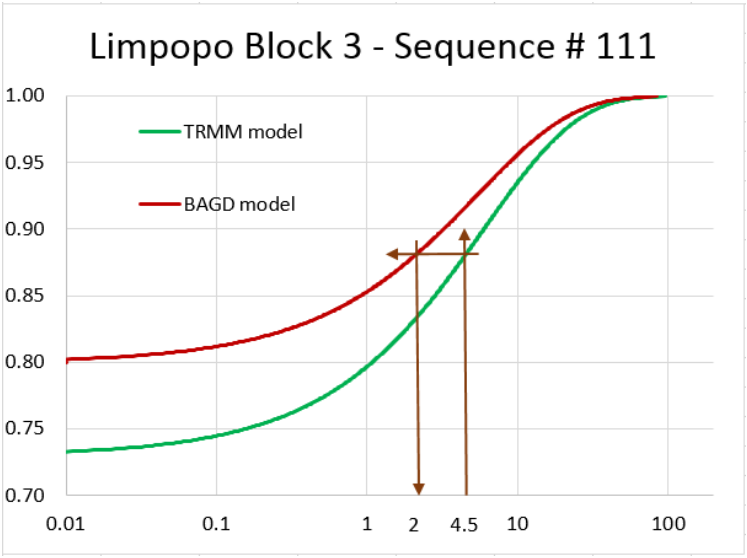


Figure 9.18. Sequence of calculations to perform a QQ transform of TRMM rainfall to a Gauge Block Average estimate in Block number 111 as listed in Table 9.1. TRMM value on the day is 4.5 mm. Reading the probability on the Green TRMM model curve gives a value of 0.88. The corresponding BAGD value for this probability is 2 mm. P[0] = 0.80 for this site.

9.7. An application of the QQ transform method to Limpopo block data – to be done in the future

It is clear that this chapter reports work in progress and was not part of the mandate of this project. The idea was included here so that (i) we could have the proposed methodology published and (ii) could be referred to in a parallel WRC project where it will prove useful: "K5/2312 EXSMET – Exporting PYTOPKAPI and HYLARSMET over SADC with EXtended spatial and computational capacity of Soil Moisture and EvapoTranspiration for flood and drought monitoring".

A forward looking summary:

1. For a sample area, as a proof of concept, choose 6 gauged sites as controls and 10 or so targets to keep it manageable, thence obtain the distributions of the gauge averages of the raw control data.
2. Derive the cdfs of the areal averages at the controls and obtain the Weibull parameters.
3. Determine the dependence structure between the *a* and *b* Weibull parameters at the controls as described in Section 6.5 and decorrelate them to *a*2 and *b*2.
4. Interpolate these parameters to the target blocks using Multiquadrics and recorrelate them to *a* and *b* sets.
5. Determine cfdfs of the TRMM at all 16 blocks, targets and controls.
6. QQ transform the TRMM rainfall estimates at all blocks to areal gauge estimates.
7. Check the distributions of the TRMM transformed values to cfdfs of the areal gauge block averages at the controls and evaluate the effectiveness of the methodology.

Chapter 10. Maps, Data handling and algorithms

In this chapter we outline the source rainfall data used, the data management procedures, the products produced and the software developed. All of these materials are provided on the accompanying DVD, with the exception of the source daily rainfall data-set, due to licensing restrictions. These procedures were discussed and decided upon at the final reference group meeting on 20 Jan 2016.

10.1. Description of source data-set

We obtained a source data-set of daily rainfall records from the UCT Climate Systems and Analysis Group (CSAG). This rainfall database was produced during a previous WRC project by Lennard et al. (2013). This data-set was selected on the basis that it not only included SAWS and ARC station data up to the year 2000 from the Lynch (2004) database and additional SAWS station data for the period 2000-2010, but also that extensive quality control had been done by Lennard et al. (2013) in order to remove many of the typical anomalies in rainfall station data-sets. The applied quality control procedures were based on those developed by Durre et al. (2010) for the Global Historical Climate Network (GHCN) project.

The CSAG data-set was therefore chosen as a suitably quality controlled collection of daily gauge records to serve as the core data-set for this project. However it should be noted that despite the hard work done by Lennard et al. (2013), there were still some issues such as those illustrated in Figures 10.1 and 10.2 (three images repeated from Figures 5.16, 5.17 and 5.18 and explained in the captions). These hampered the computations, so that there was a need to perform some triage. Problems such as those illustrated there typically appear for stations that are closed and moved to nearby locations. This required developing software to weed out the offending gauges from the portion of the data-set used, else the infilling programs crashed.

The CSAG database consists of a large number of station data files which are formatted *plaintext*, with included metadata. We developed Python code to read these files, and later converted them into a single database in the form of a NetCDF file to make processing more convenient.

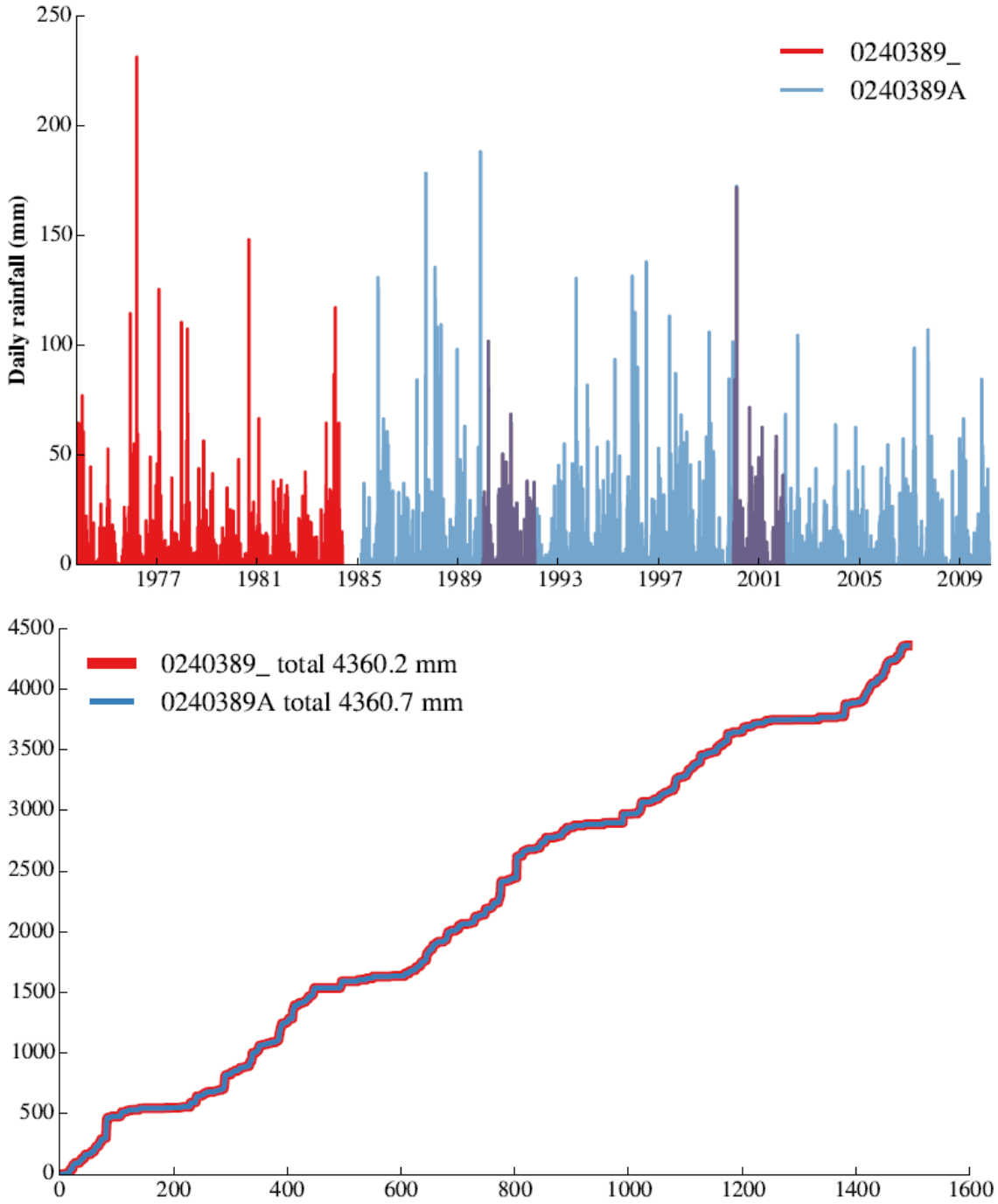


Figure 10.1. Two gauges located in the same SAWS 1-minute block, with different but partially over-lapping periods of record. The surprising thing in this case is that the rainfall cumulative sums during the overlapping period are identical, apart from a 0.5mm difference occurring on a single day. This despite the meta-data suggesting that the stations are at exactly the same position, with one replacing the other at some point.

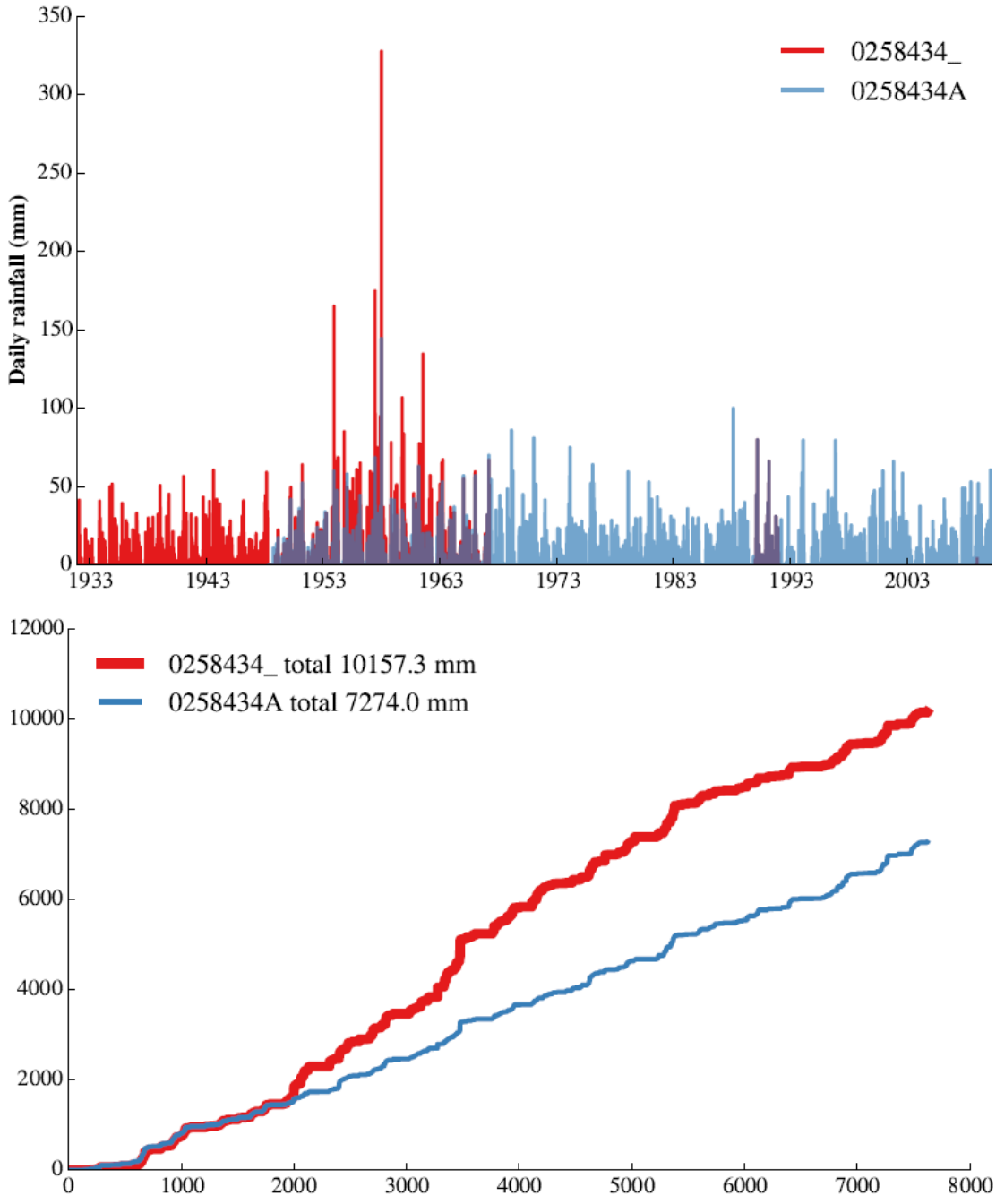


Figure 10.2. Two gauges located in the same SAWS 1-minute block, with different but partially overlapping periods of record. In this case the cumulative sums during the period of overlap start off following each other, but then begin to deviate significantly despite the meta-data suggesting that the stations are within 1 km of each other in the central Free State.

10.2. Analysis procedure

Here follows a brief high level outline of the processing steps/algorithms that were devised to make the computations required to perform the data repair and the spatial interpolation

1. Convert 4012 CSAG station files to single NetCDF database. This necessitated developing access speed improvements, a single file containing all data etc.
2. Compute monthly and annual totals from daily data, taking care of the missing data, count *nmissing* etc..
3. Apply the infilling algorithm to monthly and annual data. Infill all time periods with one or more missing days. This required developing procedures to handle the variable lengths of record, the availability of neighbours, limit the search radius for neighbours, generate ensembles, save distribution parameters etc.
4. Combine observed data and the expected values of infilled distributions to produce the "best" time-series possible, and compute means/percentiles of the time-series at each station. An example is given in Figure 10.1
5. Spatially interpolate the means of the infilled time-series onto a 0.1° grid as shown in Figure 10.2.
6. Bi-linearly interpolate 0.1° maps to 1 min of arc as shown in Figure 10.3.
7. Compute the average over each quaternary catchment for both annual and monthly maps as shown in Figure 10.4, for example.
8. Produce Figures
9. Produce raster data-sets for GIS
10. Produce shapefiles of quaternary averages for GIS

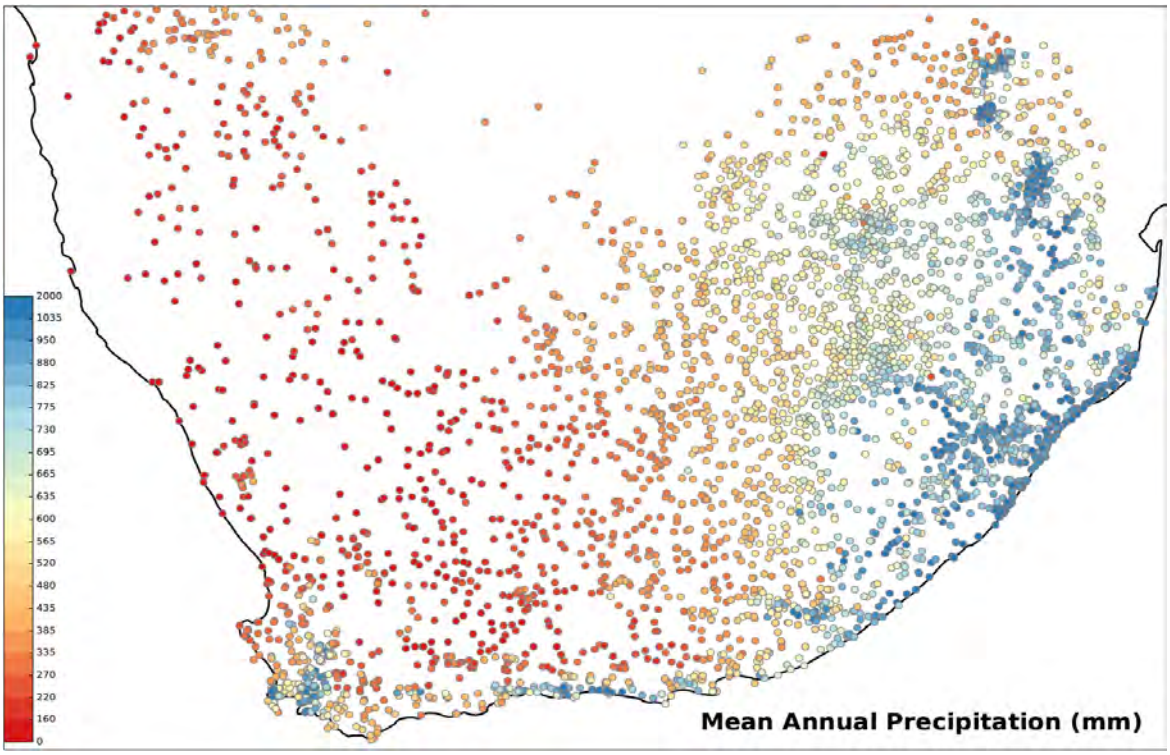


Figure 10.3. MAP at the stations calculated using both the observed and infilled data.

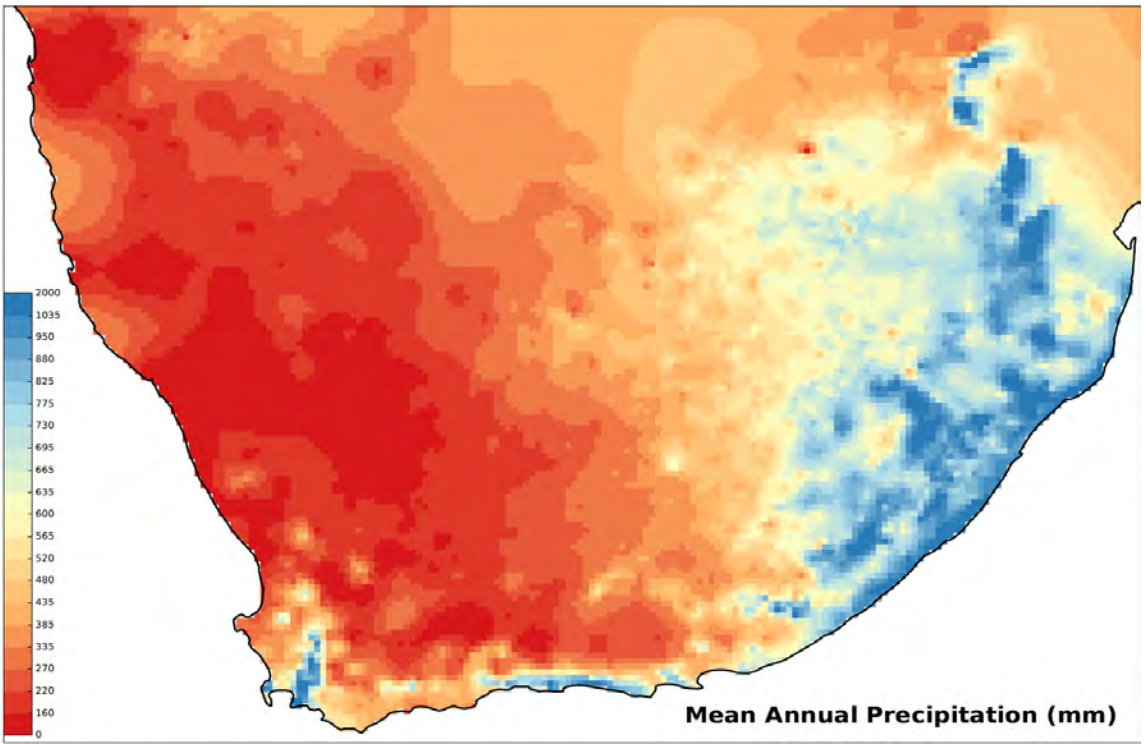


Figure 10.4. Station based MAP (Figure 10.3) interpolated onto a 0.1° grid, using an exponential Kriging variogram with correlation length 0.5° (monthly interpolations were done with 0.3° correlation length).

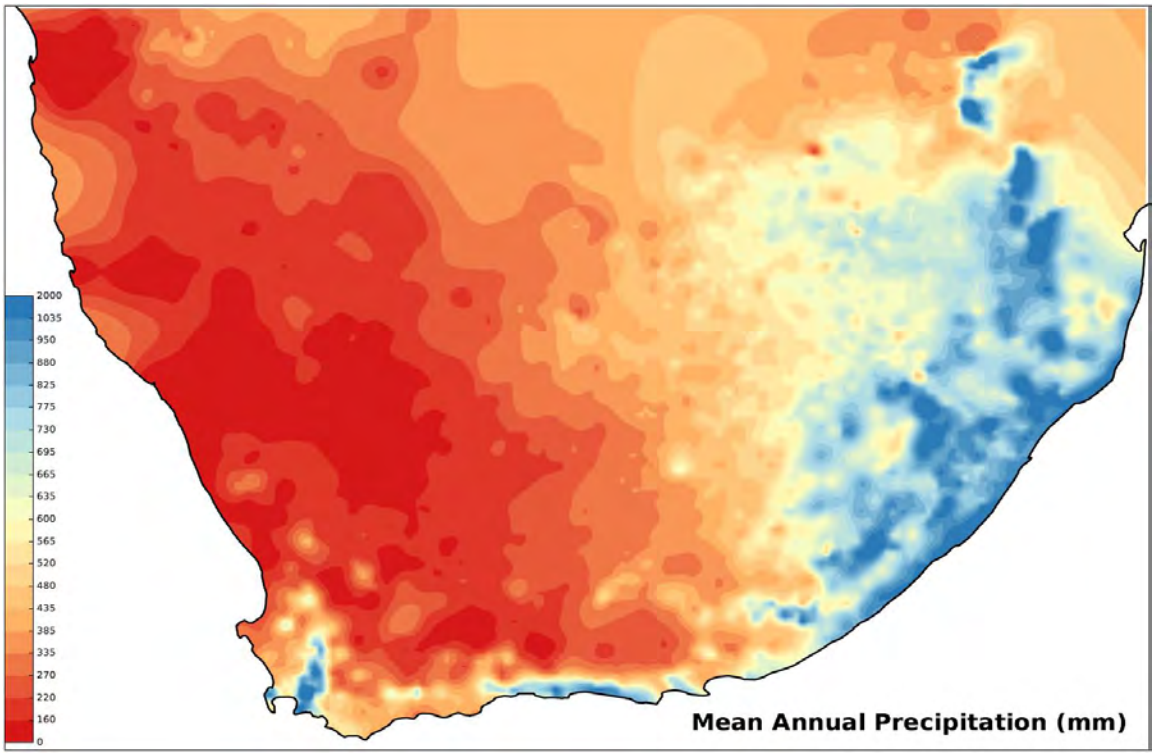


Figure 10.5. The gridded MAP from Figure 10.3, bi-linearly interpolated onto a finer 1 arc minute grid (0.0167°).

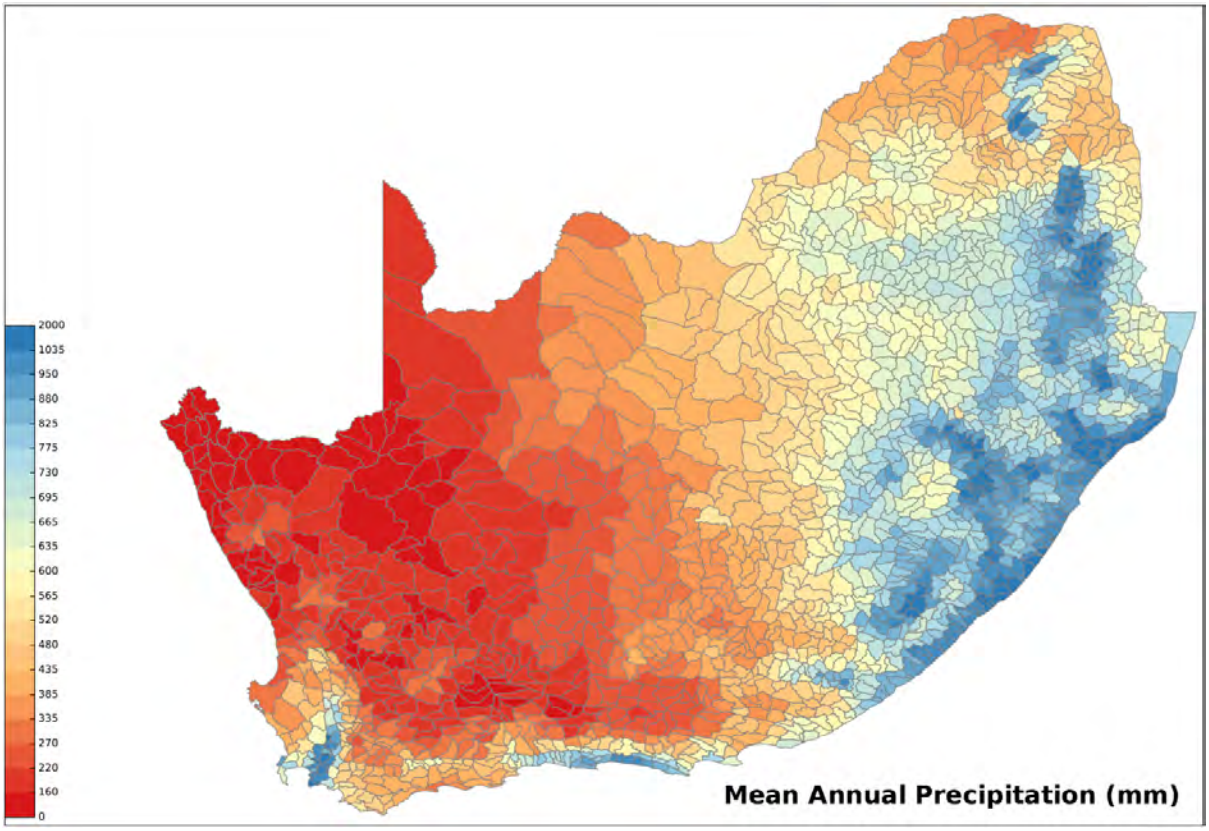


Figure 10.6. The gridded MAP from Figure 10.2 averaged over each of the 1946 quaternary catchments in the region.

10.3. Summary of DVD contents

A selection of the processed data, high quality PDF Figures and software developed during the course of the project are included on an attached DVD. These are summarized below:

Data products

- **CSAG station metadata**. A CSV file containing the station metadata for each station in the CSAG station archive. The file lists the following metadata:
  ID – the station identifier,
  ALTITUDE – Station height in metres above sea level,
  CLEANING – Cleaning level,
  *CO*UNTRY – Country code,
  CREATED – Date CSAG station file was created,
  END DATE – Date of the last observation,
  FORMAT – CSAG station file version,
  LATITUDE – Latitude in decimal degrees North,
  LONGITUDE – Longitude in decimal degrees East,
  NAME – Station name,
  START DATE – Date of the first observation,
  VARIABLE – Variable recorded (always precipitation),
  RECORD LENGTH – Length of the data record in days [File name csag-gauge-metadata.csv]

- **Annual totals.** A NetCDF file containing the annual totals for each gauge. The file has dimension variables station and time with variables *rain*, *nmissing, lat, lon* and *elev*. The station dimension variable contains the station identifier for each station while the time dimension variable gives the end date of each annual accumulation period (31 Dec each year). The *rain* variable contains the total of all observed daily rainfall for each year at each station (these are partial totals for years with missing daily data). The *nmissing* variable contains the count of missing observations of daily rainfall for each year at each station. The *lat* variable gives the latitude of each station. The *lon* variable gives the longitude of each station. The *elev* variable gives the elevation of each station. [File name *csag-station-database-annual.nc*]

- **Monthly totals.** A NetCDF file containing the monthly totals for each gauge. The file has dimension variables *station* and *time* with variables *rain, nmissing, lat, lon* and *elev*. The station dimension variable contains the station identifier for each station while the time dimension variable gives the end date of each monthly accumulation period (last day of each month). The *rain* variable contains the total of all observed daily rainfall for each month at each station (these are partial totals for month with missing daily data). The *nmissing* variable contains the count of missing observations of daily rainfall for each month at each station. The *lat* variable gives the latitude of each station. The *lon* variable gives the longitude of each station. The *elev* variable gives the elevation of each station. [File name *csag-station-database-monthly.nc*]

- **Annual infilled time-series.** A NetCDF file containing the infilled annual totals for each gauge. The file has dimension variables *station* and *time* with variables o*bs rain, p0, pt, mean, 10percentile, 50percentile* and *90percentile*. The *station* dimension variable contains the station identifier for each station while the *time* dimension variable gives the end date of each annual accumulation period (31 Dec each year). The *obs* rain variable contains the total of all observed daily rainfall for each year at each station (these totals are only for years without any missing daily data). The *p0* variable contains the probability of dry for each infilled year. The *pt* variable gives the probability of being below the threshold in an infilled year. The *mean* variable gives the expected value of the infilled distribution for each infilled year. The *10percentile* variable gives the 10th percentile value of the infilled distribution for each infilled year. The *50percentile* variable gives the 50th percentile (median) value of the infilled distribution for each infilled year. The *90percentile* variable gives the 90th percentile value of the infilled distribution for each infilled year. [File name *infilled-station-database-annual.nc*]

- **Monthly infilled time-series.** A NetCDF file containing the infilled monthly totals for each gauge. The file has dimension variables *station* and *time* with variables *obs rain, p0, pt, mean, 10percentile, 50percentile* and *90percentile*. The *station* dimension variable contains the station identifier for each station while the *time* dimension variable gives the end date of each monthly accumulation period (last day of each month). The *obs_rain* variable contains the total of all observed daily rainfall for each month at each station (these totals are only for years without any missing daily data). The *p0* variable contains the probability of dry for each infilled month. The *pt* variable gives the probability of being below the threshold in an infilled month. The

mean variable gives the expected value of the infilled distribution for each infilled month. The *10percentile* variable gives the 10th percentile value of the infilled distribution for each infilled month. The *50percentile* variable gives the 10th percentile (median) value of the infilled distribution for each infilled month. The *90percentile* variable gives the 90th percentile value of the infilled distribution for each infilled month. [File name *infilled-station-database-monthly.nc*]

- **MAP grids at 1 minute spatial resolution**. Interpolated Mean Annual Precipitation at a spatial resolution of 1 arc minute, as raster grids for use in a GIS package. The rasters are provided in two 32 bit floating point formats, GeoTIFF and Arc/Info ASCII Grid to facilitate access from a wide variety of GIS and related software packages. [File names – mean-annual-precip.tif (GeoTIFF) – mean-annual- precip.asc, mean-annual-precip.prj (Arc/Info ASCII Grid)]

- **MAP averaged over quaternary catchments.** A shapefile giving the average of the 1 arc minute Mean Annual Precipitation raster over each quaternary catchment. The attribute MAP is populated with the relevant MAP value. [File names quat-infilled-mean-annual-precip.shp, quat-infilled-mean-annual-precip.shx, quat-infilled-mean-annual-precip.dbf]

- **MMP grids at 1 minute spatial resolution.** Interpolated Mean Monthly Precipitation for each month at a spatial resolution of 1 arc minute, as raster grids for use in a GIS package. The rasters are provided in two 32 bit floating point formats, GeoTIFF and Arc/Info ASCII Grid to facilitate access from a wide variety of GIS and related software packages. [File names – mean-*-precip.tif (GeoTIFF) – mean-*-precip.asc, mean-*-precip.prj (Arc/Info ASCII Grid)]

- **MMP averaged over quaternary catchments.** A shapefile giving the averages of the 1 arc minute Mean Monthly Precipitation rasters over each quaternary catchment. The attribute MMP is populated with the relevant MMP value. [File names quat-infilled-mean-*-precip.shp, quat-infilled-mean-*-precip.shx, quat-infilled-mean-*-precip.dbf]

Here follow two examples of the contents of the files:

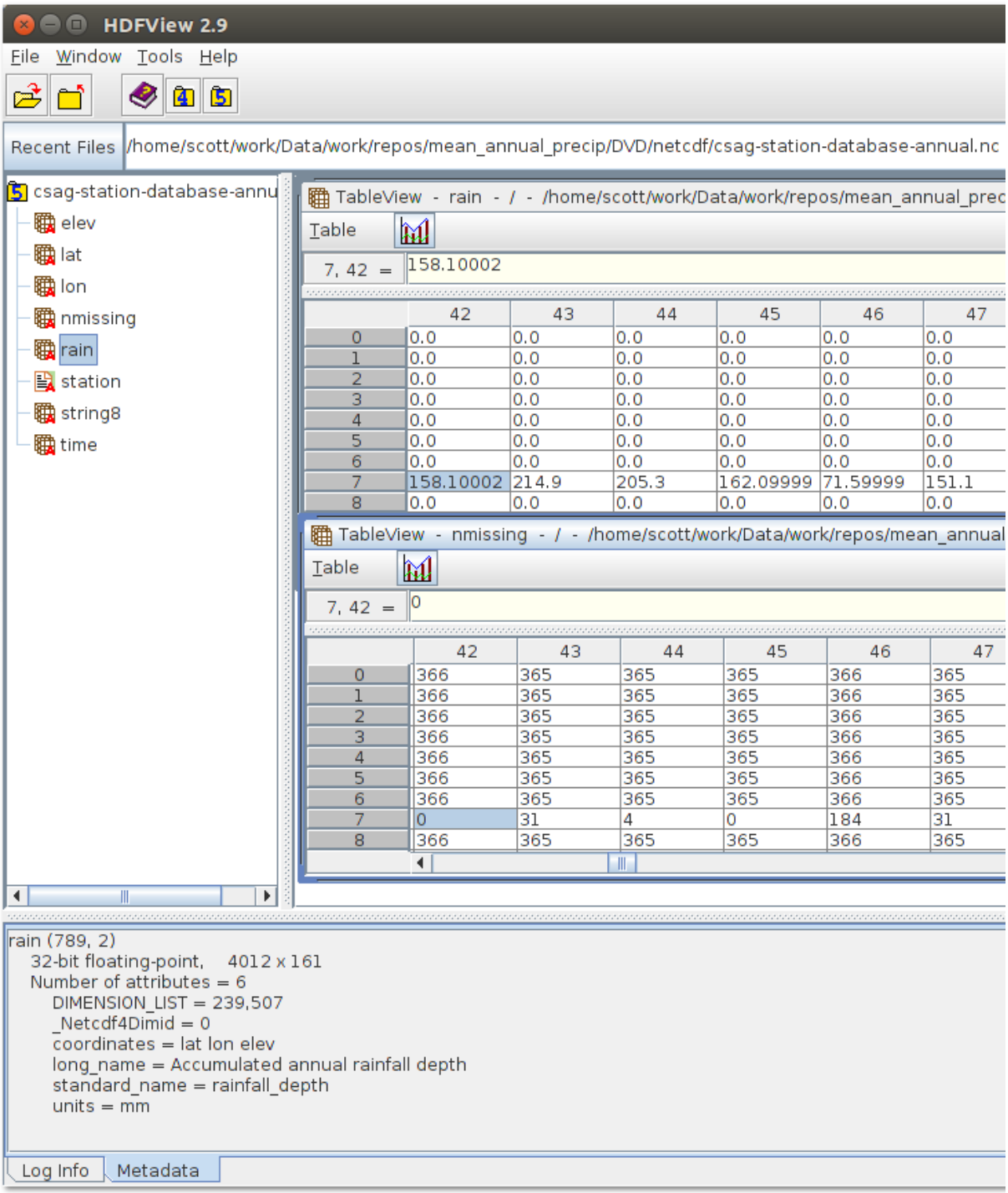


Figure 10.7. Annual accumulations stored in a NetCDF file, as viewed by HDFView. The left hand panel shows the variables in the file. In the right hand panel is a partial tabular view of the *rain* (observed annual rainfall total) and the *nmissing* (number of missing days) variables. The bottom panel shows the file metadata for the *rain* variable.

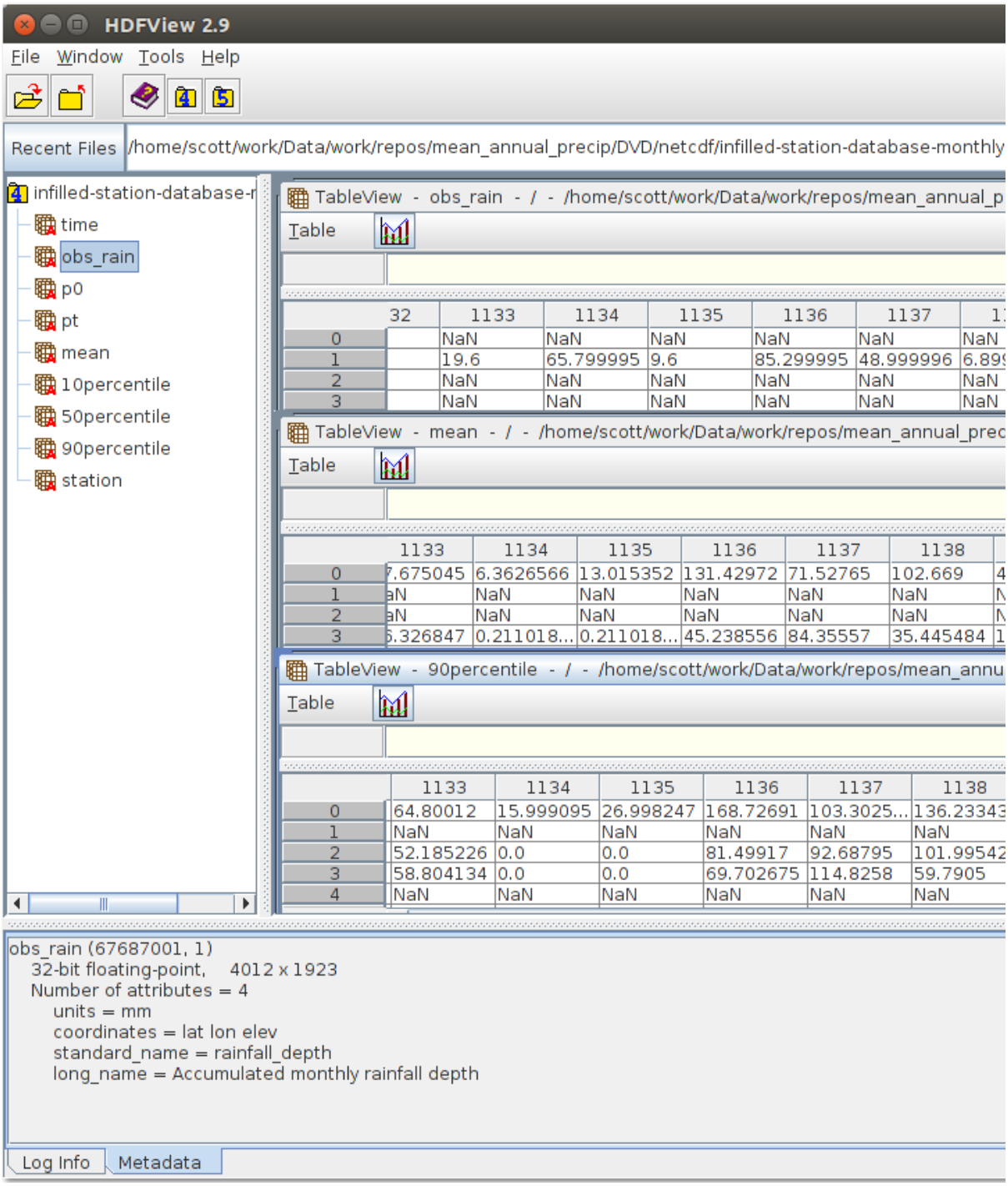


Figure 10.8. Monthly infilled accumulations stored in a NetCDF file, as viewed by HDFView. The left hand panel shows the variables in the file. In the right hand panel is a partial tabular view of the *obs_rain* (observed annual rainfall total), *mean* (infilled expected value) and the *90percentile* (90th percentile of the infilled distribution) variables. The bottom panel shows the file metadata for the *obs_rain* variable.

Figures

A selection of high quality PDF format figures are included on the DVD. These are concatenated into a single PDF document (*dvd-figures.pdf*) in the root directory. The figures are summarized below and most appear in various places throughout this report (where they are described in more detail).

- Selected percentiles (10, 50, 90) of infilled annual totals
- Selected percentiles (10, 50, 90) of infilled monthly totals
- Mean annual infilled precipitation as recorded at the gauges
- Mean annual infilled precipitation interpolated onto a 0.1_ (6 min) grid
- Mean annual infilled precipitation interpolated onto a 0.0167_ (1 min) grid
- Mean annual infilled precipitation averaged over quaternary catchments from the 1 min grid.
- Mean annual infilled precipitation at the gauges, calculated from consecutive 20 year periods
- Mean monthly infilled precipitation as recorded at the gauges
- Mean monthly infilled precipitation interpolated onto a 0.1_ (6 min) grid
- Mean monthly infilled precipitation interpolated onto a 0.0167_ (1 min) grid
- Mean monthly infilled precipitation averaged over quaternary catchments from the 1 min grid.

Software

- Python package for dynamic Copula infilling
- Selected data processing scripts
- Selected figure generation scripts

Caveats and recommendations for improvements

- The CSAG database still has anomalies to be resolved
- The CSAG database does not include non-SAWS/ARC gauges which are included in the Lynch data-base. These should be blended in. We realized this late and tried to effect the merge, but this was beyond our resources.
- A single coherent daily rainfall Data Base (easily updateable via the WR 2012 series etc.) would be a valuable project. However, getting the data providers blessing is extremely difficult and possibly very political.
- Apply infilling technique at daily scale. Code is there, it can work, but main concern is data licensing – it's pretty easy to infer the observed data from an infilled time-series.

Chapter 11. Summary and Conclusion

A large proportion of the work developed and recorded in this report is original. We draw attention to the new method of infilling and interpolating missing information in data we call the Dynamic Copula Regression (DCR) technique. This method exploits the treatment of zeros and Gaussianises the data to obtain not only the best estimate possible of the missing value but adds a meaningful error tag to the estimate. This approach follows Jaswinski's (1970) dictum: "An estimate is meaningless unless you know how good it is".

The maps we have produced are slightly different from their forbears, but we are confident that they are an honest representation of the rainfall record in Southern Africa. Remarkably, we found that the MAP is very stable over the last 150 years. We conclude with a summary of the messages of the Chapters.

The first Chapter contains samples of the final product – the maps of MAP and MMP together with maps of their variability and a comparison with previous estimates. It contains a sample of the maps that might be useful to the practitioner, to complement those offered in the Executive Summary. The last set of images in the Introduction, making up Figure 1.8, is a sequence of eight separate 20-year periods of MAP starting in 1850 and finishing in 2010. We noted that the MAP has been remarkably stable over the 20th century and that the notorious and troublesome interdecadal variations are smoothed out by averaging over 20 year periods.

In the second Chapter we reported on work using Circulation Patterns associated with the rainfall regimes, based on the output of WRC project K5/1964, but with new regions based on SAWS criteria. Although these are interesting from a local climate point of view, this early work was superseded by the methodology summarised in Chapter 3. We worked on the premise that correlations of rainfall between successive periods [day, month and year] are so low that the infilling can be usefully done at one interval at a time, thus making the use of CPs redundant in this study.

Chapter 3 describes the cross-validation of Gauge data, with a view to selecting the best infilling procedure, by comparing several standard methods of infilling missing data values against the new Dynamic Copula Regression (DCR) method, outlined in this Chapter. For this comparative work, we chose a set of gauges in the Southern Cape whose intact records span 32 years. In the monthly data we found that an average of about 5 % of the months were dry, whereas in the daily data, the average proportion of dry days was approximately 80 %. The way that the cross-validation was done was that, in 32 years we left out 20 % at a time for each gauge in turn, modelling in 2 seasons. We chose the copula-based method (DCR) as the one to use for data repair, not only because of its success in the above tests but because it can give a valuable additional product: the error structure of the interpolant, tailored to the local spatial distribution of the controls, as well as their rainfall data values. Although the CP dependent copula is a fraction better than the plain copula method for repairing the daily data, we decided that the extra effort is not worthwhile for infilling over the whole Southern African region.

In Chapter 4 we explain how to visualise the worth of the infilled values, through pictorial explanation of the methods, complementing Chapter 3. Again we found that Gaussian copulas used in DCR are superior to other methods of infilling. We also note that it is important to de-seasonalise data before computing cross correlation coefficients (cccs), so it was surprising to find that cccs of de-seasonalised spatial daily are independent of the area's wetness.

In Chapter 5 we determine the value of the data and examine the results of the infilling, introducing an important quality measure in the guise of quantile error bounds, rather than the not very useful mean and standard deviation. This applies particularly to the estimation of precision of the infilling whose distribution is not Gaussian, particularly in the case of Daily rainfall records. We also note the difficulty of coping with 'dirty' data – for example those which are supposed to be different but have identical periods of record. We offer a new idea called Mean Annual Precision, which is the average of the interquartile spread of 80% over the complete record including intact and infilled values of each station. The less the amount and the tighter the infilling, the smaller the Mean Annual Precision and vice versa.

Chapter 6 is a summary of a new methodology developed for spatial interpolation between the repaired gauges for the production of smooth maps and for estimating rainfall amounts over catchments, as an alternative to conventional Kriging methods. Here, we generate spatially interpolated fields, we fix the observed gauge values on the day (or month or year) and sample from the distributions of the missing gauge data estimates. The scheme has three benefits: (i) we have a better estimate of the mean field (with error structures at each infilled pixel in the field); (ii) we can generate ensembles of possible spatial fields, matching the observed data, getting sharp estimates of the missing values and (iii) the ensembles can be used to determine the uncertainty of the fields. The key is to use a spatial variant of Dynamic Copula Regression to perform the interpolation. Although we are confident that this is a valid procedure for interpolating over mesoscale areas, it is computationally demanding and we had to forgo the method in favour of the Gaussian Kriging procedure introduced in Chapter 3 when it came to interpolate the space between the 4000 or more gauges over the 1.22 million sq km region. As part of the work in this Chapter, we explored the concept of covariates and determined whether these add value to the interpolation. We had in mind altitude and TRMM and tried altitude, reserving TRMM for Chapter 9. We found there was a very weak link between daily gauge rainfall and altitude, and showed that, even if we used as small a ccc as 0.2, the covariate altitude made rain over places where the gauges were dry, so this idea was abandoned, in favour of univariate interpolation.

In Chapter 7, we describe a straightforward spatial Interpolation using the Fast Fourier Transform to produce radar-like random fields, as a possible alternative to the copula-based methods, developed in Chapter 6. The key to the method is to use Gaussian random fields, modelled on Gaussianised radar images. An ensemble of these fields provides not only a measure of uncertainty [median and quartiles at each location in the area] but also plausible rainfields for rainfall-runoff calculations through catchment modelling. The idea can be extended to areal rainfall simulation, by using a daily rainfall network model and merging

random fields with the gauge values using the correct correlation structure for the field. Although it was not used in this MAP map project, it was introduced here as a valid alternative to Dynamic Copula Kriging as long as the observed rainfield is spatially homogeneous from the correlation point of view.

Chapter 8 describes an attempt to develop a method to downscale TRMM rainfall data to block averaged daily read gauge rainfall data using regression. Unfortunately the timings and amounts of the TRMM daily rainfall estimates do not match with daily rainfall catches, so the correlations are generally too low to allow regressions to be useful. This is so even after measuring correlations using the Spearman formulation, which works on quantiles and not on distributions. Given the above, it is likely that TRMM data (and the output of its successor GPM) will be useful for large-scale hydrology and agriculture, particularly at the monthly scale, in contrast to daily. Thus crop monitoring and reservoir storage calculations will benefit, but not Flash Floods. The short conclusion is that TRMM is useful for hydrology in a coarse way, but poor in detail.

Chapter 9 introduces a novel idea which is suggested for performing a valid quantile-quantile transform of TRMM to Block averaged gauge rainfall where there are records and then interpolating the methodology to ungauged locations. This QQ transformation is done by using the appropriate and easy-to-manipulate Weibull probability distribution fitted to gauge data, where available, and to all TRMM data. Although not exploited in this study, the methodology needs to be recorded and used elsewhere. The Chapter reports work in progress and was not part of the mandate of this project, nevertheless, the idea was included here so that (i) we could have the proposed methodology published and (ii) it could be referred to in a parallel WRC project researched by this team, where it will prove useful.

Chapter 10 describes the data used and algorithms used and introduced in the project and indicates how these have been archived for access by practitioners. Here we outline the source rainfall data used, the data management procedures, the products and the software developed. A summary of the filing procedure is made of all of the products which are provided on the accompanying DVD, with the exception of the source daily rainfall data-set, due to licensing restrictions.

In conclusion, this was a challenging project, introducing and comparing a collection of modelling ideas, some new, that would produce a robust and workable methodology for infilling and interpolating daily raingauge data over the region at the three useful scales: daily, monthly and annual. We hope that the product becomes trusted and used for the common good.

...---<<<oOo>>>---...

References

1. Bárdossy, András, and Geoff Pegram (2013), Interpolation of precipitation under topographic influence at different time scales, *Water Resources Research*, Vol. 49, 1-21, doi:10.1002/wrcr.20307

2. Bárdossy, András and Geoff Pegram (2014), Infilling missing precipitation records – A comparison of a new copula-based method with other techniques, *Journal of Hydrology*, Vol. 519, pp. 1162-1170

3. Bárdossy, András, Geoffrey Pegram (2016), Space-time conditional disaggregation of precipitation at high resolution via simulation [Paper #2015WR018037RR], accepted for publication in Water Resources Research

4. Clothier AN and GGS Pegram (2002) Space-time Modelling of Rainfall using the String of Beads Model: Integration of Radar and Raingauge Data. WRC Report No. 1010/1/02, ISBN 1 86845 835 0, *Water Research Commission*, Pretoria.

5. Dent M. C., S. D. Lynch, Roland E. Schulze, (1987) Mapping Mean Annual and Other Rainfall Statistics Over Southern Africa, University of Natal

6. Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., and Vose, R. S. (2010). Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology*, 49:1615-1633.

7. Huffman, G. J., Adler, R. F., Bolvin, D. T., and Nelkin, E. J. (2010). *Satellite Rainfall Applications for Surface Hydrology – The TRMM Multi-Satellite Precipitation Analysis (TMPA)*. Springer Science and Business Media B.V.

8. Jazwinski AH (1970). *Stochastic Processes & Filtering Theory*, Academic Press.

9. Kruger AC (2004) Climate of South Africa: Climate regions. Report WS45, *South African Weather Service*, Pretoria.

10. Lennard, C., Coop, L., Morison, D., and Grandin, R. (2013). Extreme events: Past and future changes in the attributes of extreme rainfall and the dynamics of their driving processes. Technical report, WRC Report No. 1960/1/12, *Water Research Commission*, Pretoria.

11. Lynch, S. (2004). Development of a raster database of annual, monthly and daily rainfall for Southern Africa. Technical report, WRC Report No. 1156/1/04, *Water Research Commission*, Pretoria.

12. Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. (2012). An Overview of the Global Historical Climatology Network-Daily Database. *Journal of Atmospheric and Oceanic Technology*, 29:897-910.

13. Pegram, G.G.S. (1989). Patching Rainfall Data – A Guide. BKS report to the Department of Water Affairs. November.

14. Pegram GGS (1997). Patching rainfall data using regression methods. 3. Grouping, patching and outlier detection. *Journal of Hydrology*, 198, p319-334.

15. Pegram, Geoffrey GS (2010). PEGRAIN – a complete daily rainfall network simulator, Report to Department of Water Affairs & Forestry

16. Pegram, G. C. and Pegram, G. G. S. (1993). Integration of rainfall via multiquadric surfaces over polygons. *Jnl. Hydraulic Eng.*, ASCE, 119(2):151-163.

17. Pegram, Geoff and András Bárdossy (2013), Downscaling Regional Circulation Model rainfall to gauge sites using recorrelation and Circulation Pattern dependent quantile-quantile transforms for quantifying Climate Change, *Journal of Hydrology*, Vol. 504, pp 142-159

18. Pegram GGS and Clothier AN. (1999) High Resolution Space-Time Modelling of Rainfall: The "String of Beads" Model. WRC Report 752/1/99, *Water Research Commission*, Pretoria.

19. Pegram GGS, Scott Sinclair and András Bárdossy (2013). Modelling daily rain-gauge network measurement responses under changing climate scenarios, , WRC Research Report No. 1964/1/13 ISBN 978-1-4312-0476-2, *Water Research Commission*, Pretoria.

20. Sinclair S and GGS Pegram, (2005) Combining radar and rain gauge rainfall estimates using conditional merging, *Atmospheric Science Letters*, Volume 6, Issue 1, Pages 19-22

21. Schulze, R.E. (1975). Catchment Evapotranspiration Modelling in the Natal Drakensberg. Unpublished PhD thesis. University of Natal, Pietermaritzburg, RSA, Department of Geography. pp 244.

22. Wesson SM and GGS Pegram, (2004). Radar Rainfall Image Repair Techniques. *Hydrological and Earth Systems Sciences*, European Geophysical Society, 8(2), 220-234

23. Yeboah Gyasi-Agyei and Geoffrey Pegram (2014) Interpolation of daily rainfall networks using simulated radar fields for realistic hydrological modelling of spatial rain field ensembles, *Journal of Hydrology*, Vol. 519, pp. 777-791