Section 2.2

RAINFALL DATABASE

S.D. Lynch and R.E. Schulze

Background to the Rainfall Database

The rainfall database described in this Section derives from a WRC project the final report of which was titled “Development of a Database of Annual, Monthly and Daily Rainfall for South Africa” (Lynch, 2004). In addition to the development of revised spatial and temporal databases of rainfall and related rainfall statistics, the research in that project also provided for:

- quality control and provision of standardised daily rainfall datasets for input into hydrological simulation models,
- development and application of new infilling/data extension techniques for rainfall, and for
- support of the ACRU modelling system.

Point and Raster Rainfall Data

The rainfall data described in this Section fall into three categories, viz.

- point temporal rainfall data that are recorded at a site,
- infilled point temporal rainfall values that are estimated at a site, and
- spatial rainfall information that is stored in a raster.

The point rainfall data have generally been recorded at a daily time step, however, with some data available only as monthly. These point rainfall values are converted onto a rectangular grid, or raster, using various regression and interpolation techniques that are discussed later in more detail.

Rainfall Data for South Africa

The initial daily and monthly rainfall datasets used in the study by Lynch (2004) were acquired from the erstwhile Computing Centre for Water Research (CCWR) early in 2000. Those datasets had been developed for a WRC funded project titled “Mapping the Mean Annual Precipitation and Other Rainfall Statistics over Southern Africa” (Dent et al., 1989) and had been updated annually until the study by Lynch (2004) commenced in 1999.

The daily rainfall database reported on in this Section consists of data from the RSA, Lesotho and Swaziland. The data were obtained from a variety of organisations and individuals (Table 2.2.1) that include, inter alia:

- the South African Weather Service (SAWS), which also supplied the data for Lesotho and Swaziland,
- the Agricultural Research Council (ARC),
- the South African Sugarcane Research Institute (SASRI), and
- a number of municipalities, private companies and individuals (PVT).

Table 2.2.1 Number of rainfall stations per organisation for South African rainfall data

<table>
<thead>
<tr>
<th>Organisation</th>
<th>No. of Stations</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAWS</td>
<td>8 281</td>
</tr>
<tr>
<td>ARC</td>
<td>2 661</td>
</tr>
<tr>
<td>SASA</td>
<td>181</td>
</tr>
<tr>
<td>PVT</td>
<td>1 050</td>
</tr>
<tr>
<td>Total</td>
<td>12 153</td>
</tr>
</tbody>
</table>

The distribution of rainfall stations collected by Lynch (2004) for the rainfall database, including stations in neighbouring countries, is shown in Figure 2.2.1.

Decline of the South African Raingauge Network Over Time

The earliest systematic recording station was the Royal Observatory in Cape Town, with records dating back to 1850. By 1880 the region which now comprises South Africa had more than 100 active daily recording stations and this number increased to a maximum of 3 841 in 1938, with a steady decline in the number of rainfall stations since then, but particularly since 1960 (Figure 2.2.2).
Quality Control

The original rainfall database before any infilling was done, consisted of over 100 million observed values. A rainfall database of such a magnitude is bound to contain numerous errors. Some of the errors that are known to exist in the data only become apparent when the rainfall data are used in agrohydrological modelling exercises and the model results are tested against measured values. Two sources of error are described below.

First and the most common of these errors is the incorrect recording of the time and date at which the gauge is read, which in South Africa should be at 08:00 each morning for the previous day’s rainfall. Errors from incorrect dating result in adjacent raingauge recordings being out of phase by a day. Such phasing errors, while having little effect on the monthly and annual rainfall totals *per se*, are very significant for daily streamflow modelling from multiple interlinked catchments, however, where flows have to be accumulated in a correct historical sequence. Checking programs were therefore written to identify phasing problems and to correct for them.

The second type of error concerns extreme daily rainfall events. A rainfall amount of 597 mm recorded at St Lucia Lake on 31 January 1984 during cyclone Domoina is the largest verified daily total of rainfall recorded in South Africa to date. A quality control procedure that was performed on the daily rainfall data was to flag all rainfall amounts greater than 597 mm as suspect data. Surprisingly many such data were found, believed to be the result of incorrect keying in of data (e.g. 680 instead of 68.0 mm).

In addition to these two quality controls, many other sources of error were described by Lynch (2004).

The Infilled Daily Rainfall Data in the Database

Missing records can severely limit the use of rainfall data. In daily simulation modelling, for example, the models cannot function without a continuous daily dataset. A number of infilling algorithms were used in the research by Lynch (2004) and a complete list, with discussion on advantages and disadvantages of each, is contained in the full report.
Four infilling techniques were eventually selected and infilling of missing values more than doubled the size of the daily rainfall information base (Table 2.2.2). The rainfall information base consists of 105,753,218 daily observed values with 236,154,934 infilled values. The total size of the observed and infilled rainfall database is thus 341,908,152. The infilling process has also increased the size of the annual database considerably from an initial 5,118 stations with more than 15 years of complete record to 9,641 stations that have more than 15 years of record.

Table 2.2.2  Number of infilled daily rainfall values in the South African rainfall database (Lynch, 2004)

<table>
<thead>
<tr>
<th>Infilling Technique</th>
<th>Number of Daily Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expectation Maximisation Algorithm</td>
<td>113,869,517</td>
</tr>
<tr>
<td>Median Ratio Method</td>
<td>40,823,148</td>
</tr>
<tr>
<td>Inverse Distance Weighting</td>
<td>81,451,381</td>
</tr>
<tr>
<td>Monthly Infilling Technique for &lt; 2 mm</td>
<td>10,888</td>
</tr>
<tr>
<td>Total</td>
<td>236,154,934</td>
</tr>
</tbody>
</table>

Frequency Analysis of Rainfall Amounts in the Database

Approximately 80% of the daily rainfall database consists of zero rainfall on a day, while amounts of observed rainfall of less than 5 mm per day account for 10% of the database (Table 2.2.3). It appears from the results in Table 2.2.3 that infilling procedures generated many more 0 - 5 mm and 5 - 10 mm amounts than there were observations.

Table 2.2.3  Frequency analysis of the daily rainfall database (Lynch, 2004)

<table>
<thead>
<tr>
<th>Rainfall (mm)</th>
<th>% of Observed Values</th>
<th>% of Infilled Values</th>
<th>% of Observed and Infilled Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>zero</td>
<td>84.9093</td>
<td>77.4160</td>
<td>79.5348</td>
</tr>
<tr>
<td>0-5</td>
<td>6.7346</td>
<td>12.4999</td>
<td>10.7158</td>
</tr>
<tr>
<td>5-10</td>
<td>3.2739</td>
<td>5.9959</td>
<td>5.2536</td>
</tr>
<tr>
<td>10-25</td>
<td>3.5895</td>
<td>3.2651</td>
<td>3.4655</td>
</tr>
<tr>
<td>25-50</td>
<td>1.2073</td>
<td>0.6985</td>
<td>0.8559</td>
</tr>
<tr>
<td>50-100</td>
<td>0.2569</td>
<td>0.1129</td>
<td>0.1575</td>
</tr>
<tr>
<td>100-200</td>
<td>0.0261</td>
<td>0.0108</td>
<td>0.0155</td>
</tr>
<tr>
<td>200-300</td>
<td>0.0019</td>
<td>0.0007</td>
<td>0.0011</td>
</tr>
<tr>
<td>300-600</td>
<td>0.0005</td>
<td>0.0002</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Choice of Interpolation Technique for Estimation of Rainfall at Raster Points Throughout South Africa

Once missing rainfall station records had been infilled, and the station records extended by the various infilling techniques shown in Table 2.2.2, an interpolation technique had to be selected in order to estimate rainfall values at points (e.g. a raster) where there are no stations with observed data and infilled values.

It is, however, unwise to use any interpolation technique with the observed point data without carefully considering of how the results will be affected by the assumptions inherent in the interpolation method (Burrough, 1986). Inverse distance weighting (IDW) techniques are a method of choice when converting point data to a raster format, as they are easy to use and are not computer intensive. When the density of the point data is sufficient and the variation in the parameter to be converted to raster is not too complex, as in physiographically relatively uniform areas, then IDW would suffice. However, rainfall in areas of complex topography, and especially where a dense and representative network of stations is lacking there, is better converted to a raster using an approach that relies on additional explanatory variables using multiple regression techniques.

The area surrounding the Jonkershoek mountains near Stellenbosch in the Western Cape lends itself to a pictorial illustration of the main differences between a regression approach and an IDW approach in a physiographically complex area. The MAP values increase from 1,085 mm to 3,199 mm over a distance of approximately 8.5 km and the altitude increases from approximately 230 m to 1,300 m (Figure 2.2.3).

The area surrounding Jonkershoek has rainfall stations at varying altitudes (Figures 2.2.3 and 2.2.4) and the areas towards Franschoek also have a reasonable rainfall station network respect to altitude. The area further northeast, however, is covered by only two rainfall stations.

Two rainfall surfaces, created using IDW and multiple regression, have been draped onto a 3-dimension digital elevation model (DEM) to illustrate the differences.
When using an IDW approach the relationship of an increase in MAP with altitude around Jonkershoek is evident, but the mountainous area further northeast has a similar MAP value (Figure 2.2.5). The regression approach, on the other hand, has varying MAP values in the ungauged mountainous area northeast of Jonkershoek. This example highlights the fact that IDW relies on a dense gauge distribution in physiographically complex areas, whereas regression can utilise a relationship that was built up some distance away to estimate the MAP in an area of complex topography that also has a sparse raingauge network.

Lynch (2004) evaluated a wide range of regression and other statistically based techniques to interpolate rainfall values at points where no measurements are made. These included Geographically Weighted Regression (GWR); regionalised multiple linear regression as had been used by Dent et al. (1989) in their study of rainfall mapping over South Africa; a stochastic daily rainfall method previously developed for use over South Africa by McNeill et al. (1994); and interpolation by cokriging (Hughes et al., 2001). From the research it was concluded that the GWR approach should be used in a South African context (Hughes et al., 2001; Lynch, 2001).

The Geographically Weighted Regression Approach as an Interpolation Technique

Once the GWR approach had been decided upon, the decision needed to be made as to which set of explanatory variables was to be used. Different approaches were used to attempt to find which set of explanatory variables would produce the best, or most appropriate, spatial estimate of MAP. One approach was that of cross-validation. However, Hutchinson (1998) found that the cross-validation technique does not always represent a reliable estimate of model error, especially when short range correlation in the data is present. Another method of deriving the bandwidth and selection of the explanatory variables, which provides a trade-off between goodness-of-fit and degrees of freedom, is to minimise the so-called Akaike Information Criteria (AIC). The AIC has the advantage of being more general in application than cross-validation (Fotheringham et al., 2002).
A spatial database consisting of all the different MAP rasters, created using GWR, using more than 20 different combinations of explanatory variables was created to assist in the selection of the appropriate model. The idea behind this was to determine which areas have the same estimate of MAP, irrespective of the GWR model. The analysis yielded a one arc minute raster for which each pixel contains a number representing the number of times that the more than 20 rasters at that pixel have estimates that are within ±10% of the average of the values at that pixel (Figure 2.2.6). The larger part of South Africa comprises of areas that are not topographically complex and this is highlighted once again in Figure 2.2.6. The areas where less that half of the GWR models do not produce similar MAP estimates correspond well with areas that are topographically complex and where there is a large void in rainfall stations (cf. Figure 2.2.1) and these are the areas where more detailed analysis should be performed in future to determine which GWR is the most appropriate.
The Raster Surface of Mean Annual Precipitation Estimated Using GWR

The set of explanatory variables that generates the most plausible estimate of MAP is:

\[
\begin{align*}
ialtCV &= \text{the coefficient of variation (CV) of a 5 arc minute mask of altitude values,} \\
latlong &= \text{product of the latitude and longitude co-ordinates (degrees decimal) of the pixel,} \\
xplusy &= \text{sum of the latitude and longitude co-ordinates (degrees decimal) of the pixel,} \\
xx &= \text{square of the longitude co-ordinates (degrees decimal) of the pixel,} \\
slope &= \text{slope in degrees of the 8 pixels surrounding the pixel in question, and} \\
yint &= \text{y intercept term.}
\end{align*}
\]

The regression coefficients vary spatially over South Africa (Figure 2.2.7), and this emphasises the spatial non-stationarity of the data. The interactions of the coefficients at some pixels are akin to the introduction of a “new” explanatory variable, i.e. in certain areas the product of latitude and longitude (latlong) might, for example, have the same effect as if distance from the sea had been used at that location.

Adjustment of the GWR Raster of Mean Annual Precipitation

The majority of the regression procedures endeavour to minimise the residuals between observations and estimates at point where there are measurements. The aims of selecting different sets of explanatory variables is to produce a model that will fit the observed data the best, in other words, the model that produces the smallest residuals. These residuals, however, also imply that when a raster at the station pixel is queried, a different MAP value to that recorded in the point rainfall database may be obtained.

A technique that was used successfully in the Dent et al. (1989) rainfall mapping programme, and which uses these residuals to enhance the GWR raster, was again used in the development of the rainfall database of

![Figure 2.2.7](image) Spatial variations of the final GWR coefficients used to estimate MAP (Lynch, 2004)

Lynch’s (2004). In this technique, the residuals, i.e. the differences between the observed MAPs and the MAPs estimated using GWR, are interpolated onto a rectangular one arc minute raster using IDW. This raster is then added to the one arc minute MAP raster generated using GWR (Figure 2.2.8). This process adjusts the initial GWR surface \textit{locally}, to fit where there are observed values, as well as \textit{globally} at the ungauged pixels, using the residual information. In other words, the surface was not only adjusted to fit the MAP at the rainfall stations, but the surrounding areas were also adjusted according to the interpolated residual surface.
This process can also be explained by thinking of pulling a blanket up with one hand and pressing down on the bed with the other hand. The areas surrounding the hands are also adjusted and not only where the hands touch the blanket, which would happen if only the pixels, where observed data are available, were adjusted, and not the surrounding areas as well.

![Image]

**Figure 2.2.8** Technique for adjusting the regressed MAP surface using residuals (Lynch, 2004)

The final result of applying each of the above techniques was a surface of MAP as illustrated in **Figure 2.2.9**.

**Raster Surfaces of Mean and Median Monthly Precipitation**

The use of GWR to estimate monthly rainfall surfaces directly is not recommended owing to the fact that the monthly data are more variable than the MAP data and the process would be extremely time consuming as well as being computer intensive. Raster surfaces of monthly precipitation were therefore calculated using the mean and the median statistics of MAP. The technique used to create these surfaces is similar to that used by Dent *et al.* (1989) which consists, *inter alia*, of expressing the median or mean monthly values as a ratio of the MAP values (done using information at the existing network of raingauges). These ratios are then interpolated onto a rectangular raster, at a spatial resolution of one arc minute (**Figure 2.2.10**). This interpolated raster is then multiplied by the raster of estimated MAP values generated using GWR (**Figure 2.2.9**) and this is repeated for each month. This methodology has been widely accepted as a means of creating monthly surfaces of rainfall without using a regression approach.

![Image]
Section 2.2  Databases: Rainfall

Applications of the Rainfall Databases

Applications, in this Atlas, of the rainfall databases described above are at the levels of either:

- 50 year time series of continuous daily rainfalls at those stations selected to have representative values of the 1 946 Quaternary Catchments covering South Africa, Lesotho and Swaziland, e.g. used to “drive” the CERES model for maize yields and the ACRU model for hydrological, soil moisture, irrigation and crop yield related responses;
- 1’ x 1’ (~ 1.7 x 1.7 km) raster values of either mean or median monthly rainfalls where those are required, e.g. to derive rainfall seasonality and concentration or in the estimation of horticultural, agricultural, pasture and timber crop yields; and
- 1’ x 1’ raster values of mean annual rainfall, where those are required.

Figure 2.2.10  Procedure used to calculate the median or mean monthly rainfall surfaces (Lynch, 2004)

References (In the sequence in which they appear in this Section, with the full references given in Section 22)


Citing from this Section of the Atlas

When making reference to this Section of the Atlas, please cite as follows: