Chapter 3: Construction of a Rainfall Dataset

This chapter will focus on the creation of a rainfall dataset suitable for use as a predictand or response variable in an empirical model. The aim is to create a high-resolution 'state-of-the-art' daily Sahel rainfall dataset that will capture the major daily variability across the Sahel, but also allow for more local variability.

The chapter begins with a description of the raw data obtained for the study, data validation, and the identification of a suitable temporal and spatial domain for the experiment. Then, the necessity of using a gridded dataset will be explained and the gridding process described and analysed. Finally, the new gridded dataset will be assessed to see if it is an accurate reflection of observed rainfall.

3.1. The Raw Rainfall Datasets

The rainfall dataset is based on daily rainfall data obtained from two sources. Daily rainfall station data was obtained from Peter Lamb of the University of Oklahoma (hereafter referred to as the Lamb dataset), and Daouda Z. Diarra of the Mali Meteorological Service (referred to as the Diarra dataset).

The Lamb dataset consists of 542 stations from Burkina Faso, Mali, Niger and Senegal. Regularly formatted data were only available for up to 1990. Additional records up to 1997 for some stations were in a variety of often unwieldy formats. Two stations were unnamed, and some lacked location data.

The Diarra dataset consists of 308 stations from Mali. Data extend into 2000 for some stations. Only 179 stations were named and given locations. Most of the unnamed stations, identifiable just by code, contain only a few years of data, so were discarded. The stations available in the Diarra dataset were largely duplicates of those in the Lamb dataset; only four new named stations could be added. However, the duplicate stations often had more data than in the Lamb dataset, allowing the Lamb stations ending in 1990 to be updated to the end of the century.

Combining these datasets resulted in a composite containing 523 stations that could be named and located. For reasons of duplication and data quality, this number of useful stations was later reduced to 519 (see section 3.3.3).

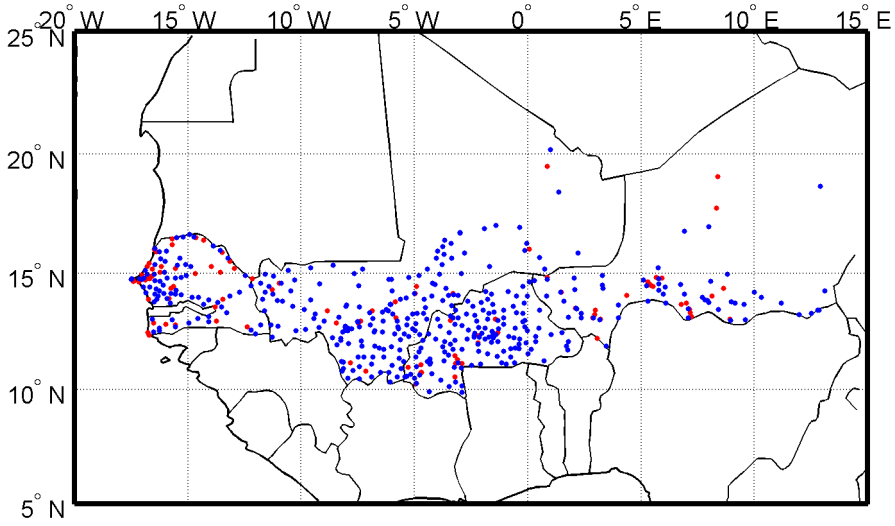

Figure 3.1. Location of the 523 stations. Stations with enough data to be used in the gridding procedure are coloured blue; those that do not are coloured red.

Figure 3.1 shows the location of the 523 stations. As can be seen, most stations are concentrated in the south of the region, where population density is greater. In Senegal, most stations are located along the coast, with significant gaps in eastern areas. There is also no data available for The Gambia. Very few data are available in the northern, desert areas of Mali and Niger.

Figures 3.2 and 3.3 illustrate the change in the number of available stations with time. Figure 3.2 is a plot of the number of complete yearly records and wet season (July to September) records over the twentieth century. Figure 3.3 shows the variability of complete monthly records for each country for 1958-1997. As can be seen from the seasonal cycles, there is a tendency for stations to only report rainfall in the wet season. This tendency is particularly pronounced in Senegal and Niger, and accounts for most of the difference between the two lines of Figure 3.2. It should be noted that rainfall in the dry season is rare, but not negligible.

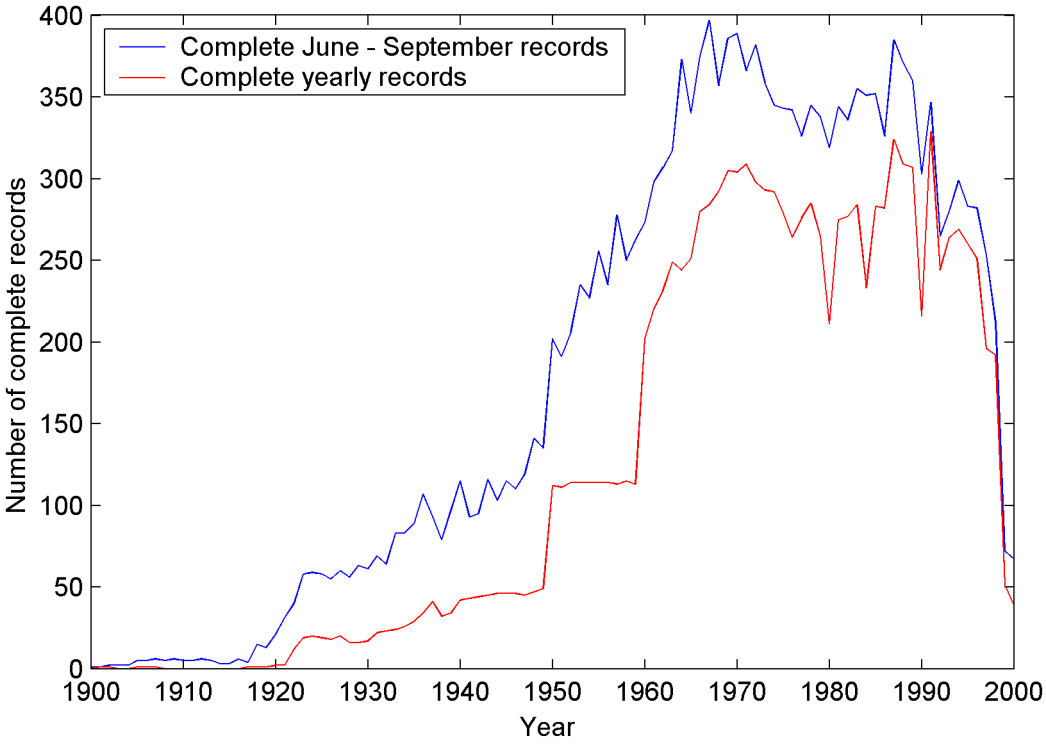

Figure 3.2. Number of complete yearly and wet season (June to September) records for the whole data set.

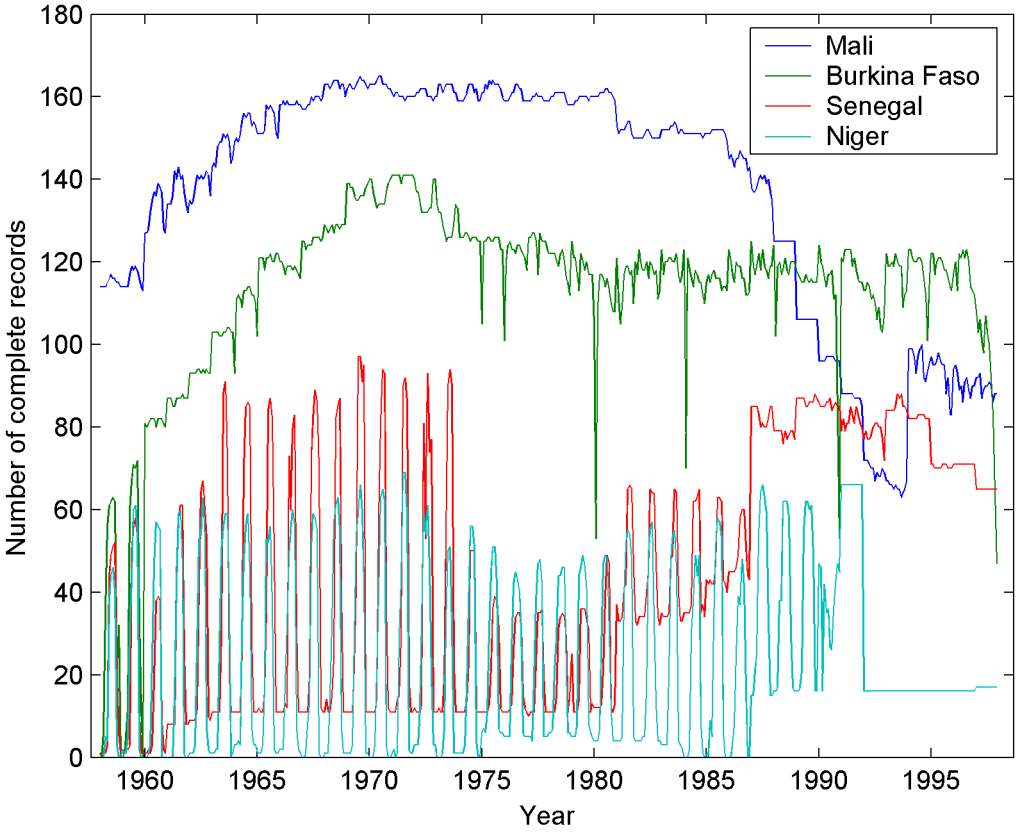

Figure 3.3. Number of complete monthly records per country, January 1958 to December 2000.

The lengths of the station records vary considerably. Whereas some records, such as the station at Kayes in Mali (14.43 °N, 11.43 °W), cover eighty years others extend over only a few years. The shorter records are of little value and were not used in the final analysis. The criteria by which stations were selected are described in the following sections.

Under ideal circumstances, extensive tests (such as those reviewed in Peterson et al., 1998) should be carried out to ensure station data are homogeneous. However, combination of the high spatial and interdecadal variability of Sahelian data and the considerable amounts of missing data in many series make identification of unnatural discontinuities in West African data problematical. For example, Tarhule and Tarhule-Lips (2001) performed a Normal Homogeneity Test on Sahelian station rainfall data, and identified discontinuities in 33 of the 55 series analysed. The discontinuities typically occurred at times marking the transitions between Sahelian wet and dry periods, which dominate the interdecadal variability of the series. Thus, any unnatural discontinuities tend to be masked. As a result, no attempt was made to test the validity of the station data using other stations. However, the verification of the gridded dataset, described in section 3.3, did identify some obvious errors in the station data.

3.2. The Creation of a Gridded Dataset

3.2.1. The Selection of Suitable Station Data

The principal aim of this thesis is to create an empirical model linking atmospheric variability to rainfall at the daily scale, but considering the different areas of the Sahel separately. Spatial analyses such as Principal Component Analysis (described in Chapter 4), or attempts to create spatially averaged series, could be biased by the uneven distribution of stations. Figure 3.1 shows that there is a greater concentration of stations in the south of the region, along the Senegal coastline and around major cities. Localised variability in these regions would dominate the large-scale pattern unless the bias is removed prior to analysis.

One method of removing bias is to grid the data, that is, to interpolate the station data onto a regular grid. The rest of this chapter focuses on the method used to create such a gridded dataset, and compares the variability of the gridded dataset with the variability of the station rainfall.

The first steps in the gridding process were:

- decide which stations should be used in the gridding procedure
- define a suitable temporal and spatial domain
- decide what is the highest spatial resolution that can be justified from the density of rainfall stations

Figure 3.2 indicates that station availability is at its greatest in the 1960-1990s, but falls rapidly at the end of the century. Another factor influencing the decision was that the atmospheric variables were to be provided from the NCEP reanalysis, and some major problems exist in its data prior to 1958. Therefore, a temporal domain of 1958 to 1997 was selected.

Initially, it was proposed to only allow stations with at least twenty years worth of data to be used in the gridding process, so all stations would have at least 50% of days present. However, the seasonal cycle of reporting in Senegal and Niger, demonstrated in Figure 3.3, results in a large amount of potentially useful information was being rejected. Therefore, it was decided to allow stations with 50% of June to September values to enter the analysis. These 416 stations are represented as blue dots in Figure 3.1. The remaining stations, coloured red in Figure 3.1, were reserved for use in validating the final dataset.

3.2.2. Interpolation using Continuous Global Surfaces

The aim of the gridding procedure is to produce an unbiased gridded daily rainfall dataset that contains as much of the variance in the underlying station dataset as possible. A common approach to such a problem is the fit a surface to the data, then calculate the value of the surface at each required grid point. In the case of Sahel

rainfall data, the aim is to fit a two-dimensional surface to three-dimensional data for each day of rainfall, with latitude, longitude and rainfall representing the x, y and z-axes respectively.

Billings et al (2002a and 2002b) review the most common approaches to fitting such as surface, which they call continuous global surfaces (CGSs). In the case of the Sahelian study, Billings et al's formulation of the problem can be reduced to considering a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, a function that maps a latitude-longitude pair (hereafter referred to as $\mathbf{x}_n$) onto a daily rainfall value, $f(\mathbf{x}_n)$. This function is to be approximated by a surface $s : \mathbb{R}^2 \rightarrow \mathbb{R}$, given the observed data values $\{f_n = f(\mathbf{x}_n) : n = 1, \ldots, N\}$, where N is the number of stations available for gridding. All of Billings et al's techniques fit a surface of the form:

$$s(\mathbf{x}) = \sum_{n=1}^{N} \lambda_n \Phi(\mathbf{x} - \mathbf{x}_n) + p(x) \tag{3.1}$$

where the λ_n are a set of weights, Φ is a fixed function mapping $\mathbb{R}^2 \rightarrow \mathbb{R}$, and $p(\mathbf{x})$ is an optional polynomial of a low degree k. A strict interpolation requires the condition:

$$s(\mathbf{x}_n) = f_n \text{ for } n = 1, \ldots, N \tag{3.2}$$

to be fulfilled, which forces the fitted surface to pass through each data point. A further condition required to remove additional degrees of freedom is:

$$\sum_{n=1}^{N} \lambda_n q(\mathbf{x}_n) = 0, \text{ for all } q \in \pi_k^d \tag{3.3}$$

where π_k^d is the space of all polynomials of at most degree k in d variables.

The differences in the techniques depend on the selection of the function Φ, which is usually chosen to minimise some statistic $J(s)$, known as the penalty function. So the aim is to solve:

$$\min J(s) \text{ subject to } s(\mathbf{x}_n) = f_n : n = 1, \ldots, N \tag{3.4}$$

For example, one of the most popular techniques, the thin-plate spline, aims to minimise the roughness of the interpolated surface (Hutchinson and Gessler, 1994). Thus, the thin-plate spline uses the penalty function:

$$J(s) = \iint_{\mathbb{R}^2} \left(\left(\frac{\partial^2 s}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 s}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 s}{\partial y^2} \right)^2 \right) \partial x \partial y \tag{3.5}$$

Duchon (1976) demonstrates that this condition is fulfilled by using the function:

$$\Phi(\|x\|) = \|x\|^2 \log(\|x\|) \tag{3.6}$$

3.2.3. Smoothing Continuous Global Surfaces

In many cases we may not desire the strict interpolation condition (3.2) to stand. For example, suppose we are gridding daily Sahelian rainfall to a 1º resolution, with grid points at the intersection of each line of longitude and latitude. A strict interpolation will introduce bias towards any stations that happen to lie on a grid point, as the surface will be forced to pass through this point. Conversely, stations that lie halfway between grid points will have much less influence.

A strict interpolation is useful when we wish to represent what is happening at the grid points. By relaxing condition (3.2), we can produce a 'smoothing' surface, which will represent what, on average, is occurring in the surrounding grid box. Bearing in mind the high spatial variability of Sahelian rainfall, this areal average is much more desirable than a point value as it will represent more of the underlying variability. This issue is discussed further in section 3.3.2.

Billings et al. (2002b) formulate the smoothing CGS by decomposing f_n into a signal component y_n and a noise component ε_n:

$$f_n = y_n + \varepsilon_n \tag{3.7}$$

In strict interpolation the aim was to solve problem (3.4), that is to minimise the penalty function $J(s)$. When smoothing, the problem usually becomes:

$$\min \sum_{n=1}^{N} \left[f_n - s(\mathbf{x}_n) \right]^2 + \nu J(s) \tag{3.8}$$

Thus, the first term represents the mean-square error in fitting the data. The second term represents the desired property from the interpolation, smoothness in the case of the thin-plate spline. The problem becomes selecting a value for the parameter v, which represents the trade-off between the two terms, in the case of the thin-plate spline the trade-off between goodness of fit and smoothness of the fitted surface. Billings et al. (2002b) describe techniques to choose a value for v.

The most commonly used CGS based smoothing procedures are the thin-plate spline and kriging (Hutchinson and Gessler, 1994). As noted above, thin-plate splines operate by minimising the roughness of the fitted surface. Kriging, however, aim to minimise the variance of the error of estimation. However, whilst thin-plate splines can be optimised easily, requiring the trade-off parameter v to be fitted, kriging depends on the fitting of a variogram with three parameters (Hutchinson, 1998a). For this reason, Hutchinson and Gessler (1994) suggest that a smoothing thin-plate spline approach is less model dependent and more robust than kriging. For this reason, in this study smoothing thin-plate splines are used to interpolate a surface from station data.

3.2.4. The Gridding Procedure: Smoothing Thin-Plate Splines

Smoothing thin-plate splines have been regularly used for the creation of global and local climatic data sets (e.g. Hutchinson 1998a and 1998b, New et al. 1999, Jeffrey et al. 2001). The governing equation for the smoothing thin-plate spline, formed by combining equations (3.5) and (3.8), is:

$$\min \left(\sum_{n=1}^{N} \left[f_n - s(\mathbf{x}_n) \right]^2 + \nu \left[\iint_{\mathbb{R}^2} \left(\left(\frac{\partial^2 s}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 s}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 s}{\partial y^2} \right)^2 \right) \partial x \partial y \right] \right) \tag{3.9}$$

As noted earlier, the first term in this equation represents the 'goodness of fit' of the fitted surface, the second term represents its smoothness, and the parameter v controls the trade-off between the two properties. The fitted surface, formed by combining (3.1) and (3.6), and dropping the polynomial term, is of the form:

$$s(\mathbf{x}) = \sum_{n=1}^{N} \lambda_n \left(\|\mathbf{x} - \mathbf{x}_n\|^2 \log \|\mathbf{x} - \mathbf{x}_n\| \right) \tag{3.10}$$

Thus the value of the surface at a desired location **x** is calculated by a weighted sum of a function of the Euclidean distance between the location of the required grid point and the location of each station. This can be calculated for any location, hence there is no theoretical limit on domain size or resolution. Furthermore, the selection of domain size and resolution will have no effect on the final fitted surface, providing the same set of stations is used. A large surface simply extrapolates outward from a smaller surface.

Two issues influenced the choice of a domain size. First, the shape of the four countries for which data was available meant that selecting a rectangular domain was infeasible. Data in eastern Mali extends northward of 17 °N, where no data are available to the west of the region, as this area lies in another country – Mauritania. Similarly, whilst data exist in the centre of the region as far south as 10°N, no data exist in the west (in Guinea) or in the east (Benin and Nigeria). Hence a simple rectangular domain would either be too small, and cut out useful data, or too large, and attempt to interpolate into areas where no data were available.

Second, the use of all the available station data would significantly increase the computational time needed to fit a surface. Furthermore, the further from a given grid point a station is, the less influence it will have on a fitted surface[1]. For

[1] This may seem contrary to equation (3.10), which suggests that influence *increases* with distance, as the value of the in brackets, Φ in equation (3.6), will increase with distance. However, (Billings et al., 2002a) notes that whilst Φ is unbounded as the norm of $\mathbf{x}$-$\mathbf{x}_n$ approaches infinity, 'suitable finite combinations of shifts of Φ can be shown to form highly peaked kernels with rapid decay at infinity'. In other words, certain suitable choices of λ will result in a high influence at a short distance, which decays as distance increases.

example, the value of a station in Niger will have very little influence on the gridded value in western Senegal.

For these reasons, the domain is composed of three separate regions, illustrated in Figure 3.4. The western region, extending over Senegal and western Mali, covers 11 – 17 °N, 18 – 8 °W. The central region, which included eastern Mali and Burkina Faso, covers 9 – 17 °N, 10 °W – 3 °E. The eastern region, comprising mainly of Niger, covers 12 – 17 °N, 1 – 11 °E. Each region overlaps slightly with its neighbour, which allows validation of the data set by comparing the results at the borders regions. Furthermore, each region was given a deliberately large border area to allow analysis of what occurs at the edges of the regions.

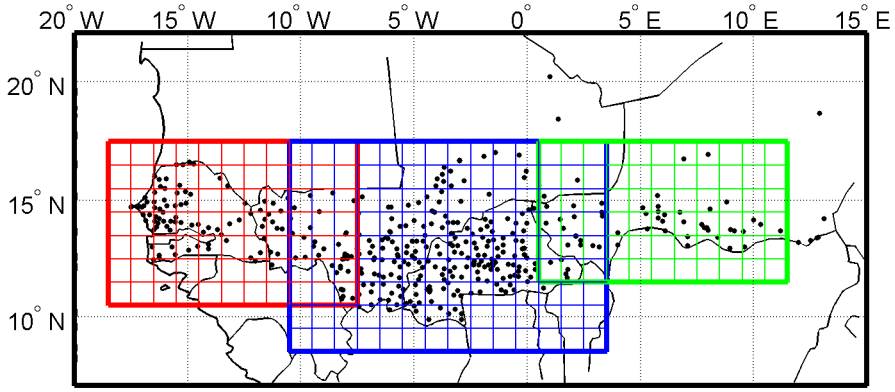

Figure 3.4. The three interpolation regions used in the gridding of the daily rainfall dataset. The black dots represent the rainfall stations used in the procedure.

A resolution of 1° × 1° was selected for the final data set. Greater resolution could have been justified in some areas, such as the Mali – Burkina Faso border. However, greater resolution would increase the computational time and require more disk space to store the grid. Furthermore, higher resolution data would greatly increase the computational power required to perform analyses on the data (such as a principal component analysis – see Chapter 4).

A potential problem in the use of latitude – longitude co-ordinates in gridding data sets occurs because of the curvature of the earth. The interpolation method assumes the data inhabits a flat plane, as illustrated in the idealised plot of Figure 3.4. In the idealised figure, each grid square has dimensions of one unit length by one unit height. Therefore, the distance between neighbouring grid points remains constant throughout the domain.

However, a true representation of Figure 3.4 could only be produced on the surface of a sphere, where the grid would be slightly warped. The result of this is that the grid squares are not truly squares; they are slightly trapezoidal. Furthermore, the further from the equator one travels, the more distorted they become. A grid square at the equator has equal *x* and *y* dimensions. At 9 °N, *y* is approximately 1.3% bigger than *x*, and so on. The difference between *x* and *y* becomes 4.6% at 17 °N.[2]

Similarly, the distance between neighbouring grid points varies. The distance between any grid point and its neighbour to the north (or south) is always 111.2 km. However, the distance between any point and its east / west neighbour varies. At 17 °N it is 106.3 km, whereas at 9 °N it is 109.8 km.

However, for simplicity it was assumed that the grid was truly square. Whilst this warping occurs, it is only slight at low latitudes. Moreover, the degree of complexity added to the problem to correct for a spherical earth is considerable. However, in regions further from the equator this warping would have to be taken into consideration. For example, at 57 °N the distance between east / west neighbours is 60.6 km, whereas at 49 °N it is 73.0 km, 20.5% greater.

A further technical issue arose over the use of a square root transform. Hutchinson (1998a) suggests the skewness of rainfall data can be reduced by taking the square root of rainfall values before gridding. Therefore the analysis was performed twice; once using raw data, and once using the transform, with results squared after gridding to produce comparable results. The effect of the transform is analysed in the next section.

The gridding was carried out using the MATLAB software package and code available from Steven Billings' website (http://www.geop.ubc.ca/~sbilling/cgs.html at the time of writing). One surface had to be fitted to each region for each day of the analysis from 1958-1997. All good stations within 2° of a region boundary were used to fit each surface. On days where no rain was recorded at any station, all grid

[2] Calculations in this and the following two paragraphs are based on the assumption that the earth is a perfect sphere.

points were also given a value of zero. Occasionally, small negative gridded rainfall values were outputted. Clearly, this is has no physical meaning, so all negative values were set to zero.

As well as interpolated data values for each grid point, several diagnostics were output:

- The value of v for each surface (see equation 3.9), which documents the trade-off between surface smoothness and goodness-of-fit
- The number of stations used to fit each surface
- The residual error of each surface at each station used in the gridding procedure. (The difference between actual rainfall and the interpolated rainfall at the station location)
- An indication of the number of stations in the vicinity of each grid box for each day. This was defined as all stations within 1 ° 'Euclidian distance' (i.e. within a circle of 1 units of degrees in the idealised 'flat' figure 3.4) of the centre of each grid box.

Each of the six runs (each of the three regions with and without the square root transform) took several hours to perform on a desktop PC.

3.3. Evaluation of the Gridded Dataset

3.3.1. The Number of Stations Used in the Gridding Procedure

Before the gridded output of the above method can be used in any analysis, it must be examined to ensure that it adequately represents the characteristics of Sahelian rainfall observations. This is done by:

- considering the effect of the gridding method on the raw data
- examining the diagnostic output of the procedure

- comparing statistics of the gridded dataset with statistics from raw rainfall data over a variety of different timescales.

First, the number of stations used in the gridding procedure will be examined in greater detail.

The number of stations used to create the gridded dataset for each region is displayed in Figure 3.5. Over the forty-year gridding period, the western region had 60-165 stations contributing to each day (with a mean of 118), the central region had 108-306 (mean 254) and the east 6-99 (mean 61). The eastern region low of 6 was only found during winter at the start of the record. After 1960, it never falls below 23. These statistics are predictable given Figure 3.3, as the central countries of Mali and Burkina Faso had the greatest number of available stations, and Niger, situated to the east of the region, had the least.

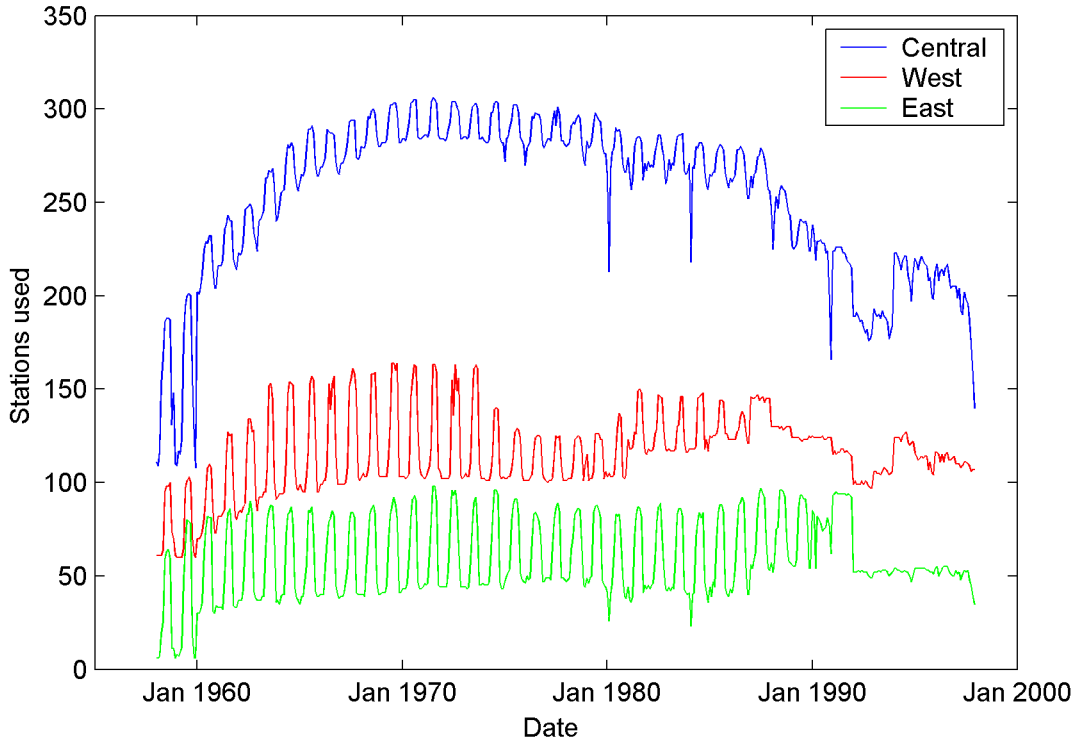

Figure 3.5. Number of stations used to create gridded dataset for each gridding region, 1958-1997.

Figure 3.6 demonstrates the regional distribution of station density by displaying the average value of the diagnostic describing the number of stations in the vicinity of each grid box (as defined at the end of the previous section). The most data-rich

areas are centred on major cities, particularly in the strip along 12.5 °N stretching from southwestern Mali to eastern Burkina Faso. Black squares contain no stations, and so values for these grid boxes are purely the result of interpolation or extrapolation. The lack of stations in certain regions, particularly in Niger, will undoubtedly compromise the quality of the gridded data set.

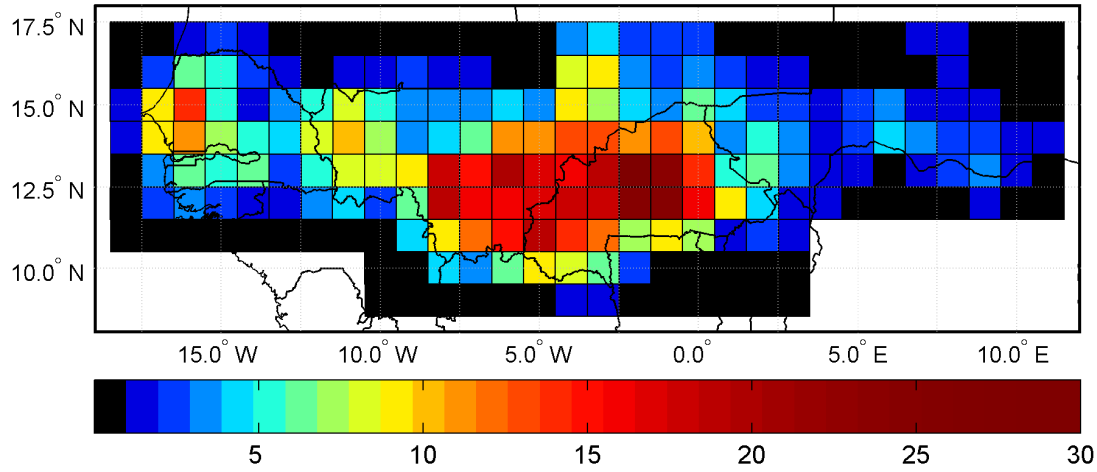

Figure 3.6. Number of stations in the vicinity of the centre of each grid box – averaged over the period 1958-1997. See text for a definition of 'in the vicinity of'.

3.3.2. Examples of Fitted Surfaces

In order to evaluate fully the effect of using thin-plate splines, this section considers several examples of the gridding procedure. It is important to note that some of the examples have been selected to illustrate occasions when the procedure performs poorly, and should not be considered as typical. Often the chosen days had exceptionally high rainfall across the region in question. But first, the justification for using a smoothing spline, as opposed to a purely interpolating spline, will be illustrated.

Section 3.2.3 described the argument for using a smoothing surface. This argument is illustrated by Figure 3.7, which shows two surfaces fitted to the central region on the 10th August 1996, firstly using an interpolating spline, then using a smoothing spline. This example illustrates an occasion of high spatial variability in rainfall.

Several stations across the region reported high rainfall, often surrounded by stations reporting no or little rainfall. Values for the gridded dataset occur where lines of unit latitude cross lines of unit longitude. As noted before, the interpolating spline will introduce bias toward those stations that happen to lie near these points.

Furthermore, problems occur when two stations close to each other report very different rainfall values. This occurred for the two stations near Kolokani in Mali (roughly 13.5 °N, 8 °W). The northern station reported a value of 52.4 mm; the southern station reported 1.2 mm. The interpolating spline extrapolates these values northward, causing an artificial 'shadow' to fall to the north. This results in a fitted value of 127.0 mm at 14 °N, 8 °W, an extreme value for which there is little justification. No such shadow occurs in the smoothing spline, and a value of 17.9 mm is fitted at 14 °N, 8 °W.

The second example demonstrates the trade-off between smoothness of the fitted surface and goodness of fit. This trade off is controlled by the value of ν in equation 3.9; a high value of ν indicates a smooth surface, a low value of ν indicates a good fit. Figure 3.8 shows the surfaces fitted by the smoothing spline in the central Sahelian region for the 17th and 19th August 1970. On both days a considerable number of stations reported high rainfall. On the 17th these extremes were situated in coherent cluster in the centre and east of the region, and the fitted surface has a ν value of 0.01. On the 19th, high rainfall stations were scattered across the region and usually surrounded by stations with low rainfall, and $\nu = 5.37$. On the first occasion, the fitted surface is able to reproduce the broad variation in rainfall across the region. On the second occasion the high spatial variability means a well-fitting surface would be very rough, so the chosen technique resorts to fitting an almost flat plane across the region, sloping from dry conditions in the northeast to wet conditions in the southwest.

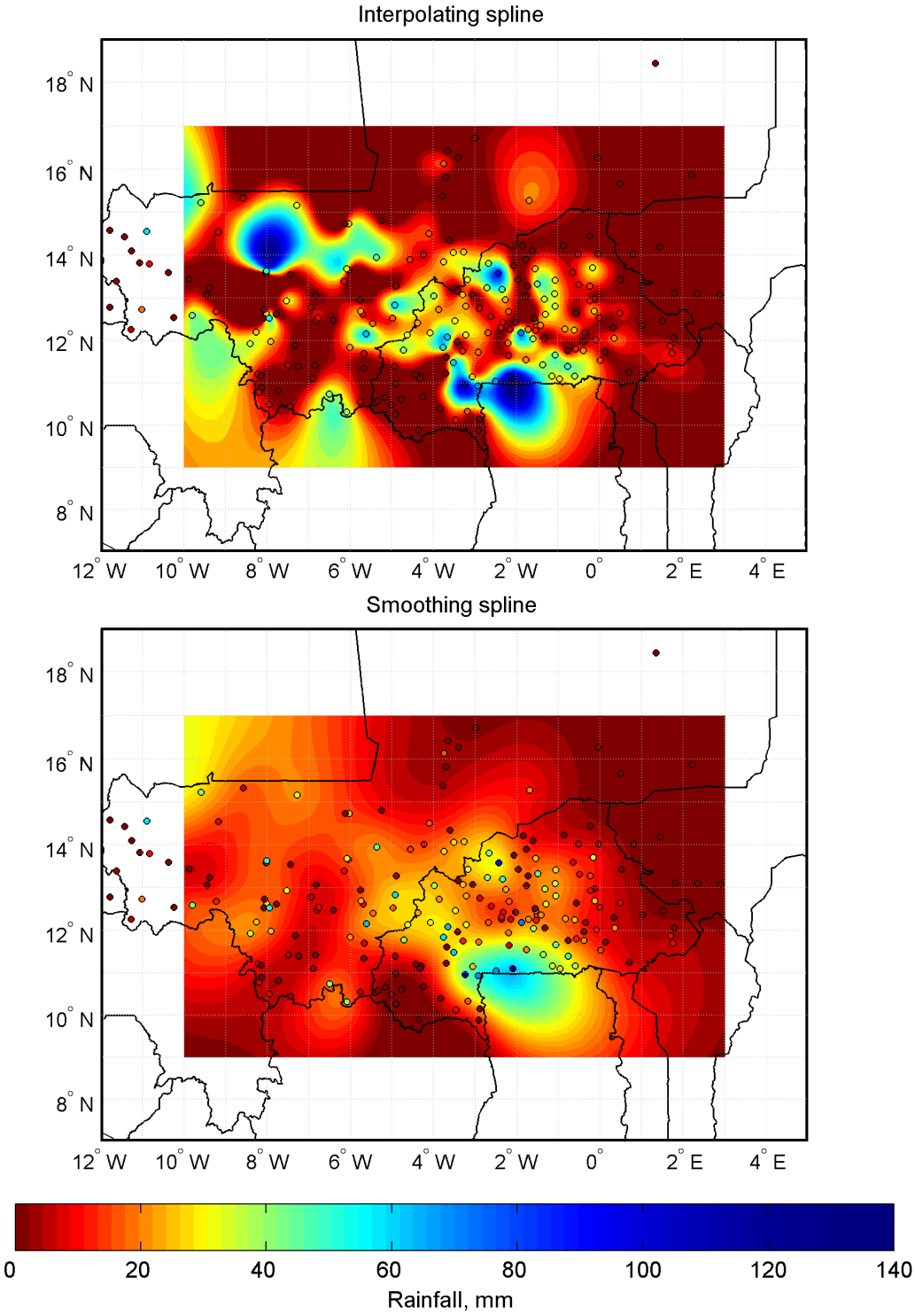

Figure 3.7. Surfaces fitted to station rainfall in the central region on 10th August 1996. The top plot uses an interpolating thin-plate spline; the bottom plot uses a smoothing thin plate spline. Coloured dots indicate rainfall values at stations; the underlying contours show the fitted surface.

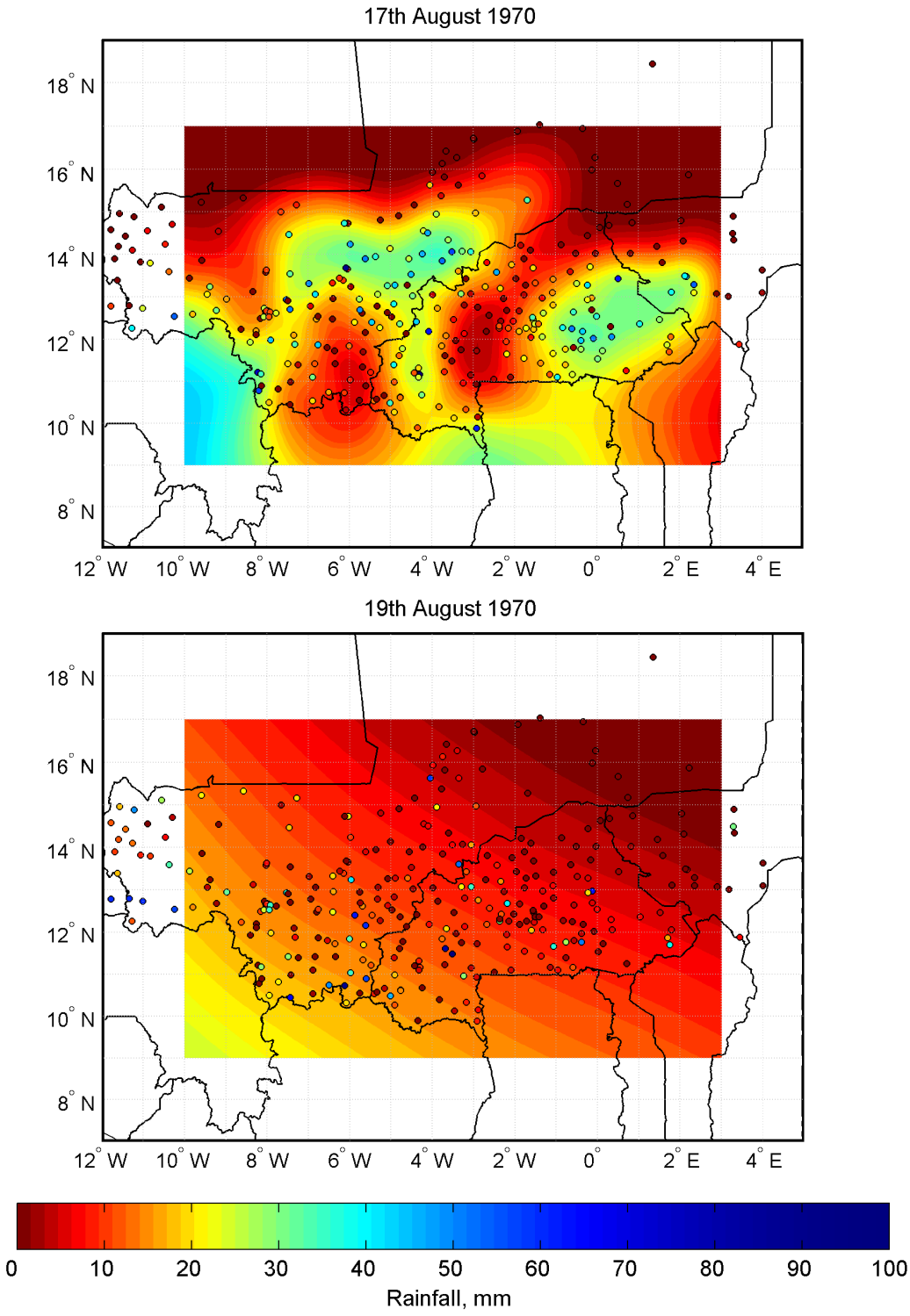

Figure 3.8. Smoothing thin-plate splines fitted to daily rainfall for the central region for 17th and 19th August 1970. For the 17th August ν=0.01, prioritising goodness of fit. For the 19th August ν=5.37, prioritising smoothness of the fitted surface.

Both fitted surfaces provide some information on the spatial variability of rainfall on that day. However, whilst the first provides information of fairly high resolution, the second only gives a very general overall picture: on the 19th it was dry in the northeast. In the first case a, resolution of 1° seems justifiable, in the second case it does not.

Fortunately, the second case is comparatively rare. Table 3.1 shows every 10th percentile of the distribution of ν between June and September for each of the six datasets (each of the three regions were gridded twice: once using untransformed data, referred to in Table 3.1 as RAW, and once using a square root transform, referred to as SQRT in Table 3.1). So, for example, 70 % of the fitted June to September surfaces in the Western region, using untransformed data, have a value of ν less than or equal to 0.785. Days outside the wet season are omitted, as rain is rare. As a result, surfaces tend to be naturally smoother and the distribution of ν is significantly different.

An upper bound for ν had to be selected as part of the gridding procedure, and an arbitrary value of 100000 was chosen. However, once ν starts to become large, its magnitude becomes somewhat irrelevant. For example, the surface in the lower half of Figure 3.8, for which ν is 'only' 5.37, is almost flat. A higher value of ν would result in an even flatter surface, but little change in its overall character or fitted values. Therefore, selection of a different value would have very little effect on the procedure.

Note that a value of zero either represents a ν so small it is rounded to zero or a case when no rain occurs anywhere in the region. The latter case is exceptionally rare, only occurring seven times in the western region, fifty times in the eastern region and never in the central region, over the course of the 40 wet seasons.

	Region / Transform used					
Percentile	West RAW	West SQRT	Central RAW	Central SQRT	East RAW	East SQRT
0	0	0	0	0	0	0
10	0	0	0	0	0.01	0.02
20	0.01	0.01	0.01	0.01	0.11	0.08
30	0.03	0.02	0.01	0.01	0.44	0.22
40	0.06	0.05	0.03	0.02	1.42	0.53
50	0.13	0.08	0.05	0.03	4.3	1.42
60	0.27	0.15	0.08	0.05	27.095	4.51
70	0.785	0.3	0.16	0.08	100000	100000
80	5.56	1.02	0.51	0.17	100000	100000
90	100000	100000	8.53	0.82	100000	100000
100	100000	100000	100000	100000	100000	100000

Table 3.1. Percentiles of ν for surfaces fitted to June to September days for each of the six gridded datasets.

As can be seen from Table 3.1, ν tends to be smallest in the central region and largest in the east. Furthermore, ν tends to be lower in the datasets formed using a square root transform. It should also be stressed that the example of 19th August is a very extreme case; rainfall was exceptionally high in a few stations scattered across the Sahel, but low in the stations surrounding them. A more typical high-ν example would see a handful of scattered stations reporting high rainfall and little rain elsewhere, and a very flat low rainfall surface fitted.

The final example, Figure 3.9, shows the effect of using the square root transform before computing the smoothing spline. This example is taken from the 9th June 1966. The top plot shows the fitted spline when using raw rainfall values (RAW), whilst the bottom plot shows the surface fitted when a square root transform is used (SQRT). The value of ν for the upper plot is 0.12, and for the lower plot is 0.06. Hence more smoothing is required for the untransformed data. This is unsurprising, as the effect of the transform is to smooth extreme values. Indeed, some transformed extremes are completely ignored by the smoothed spline, such as the two stations in Southeast Burkina Faso, at roughly 11.5 °N, 0 °W. The reduction in extremes by the transformation also leads to a smaller fitted value across the whole domain for SQRT.

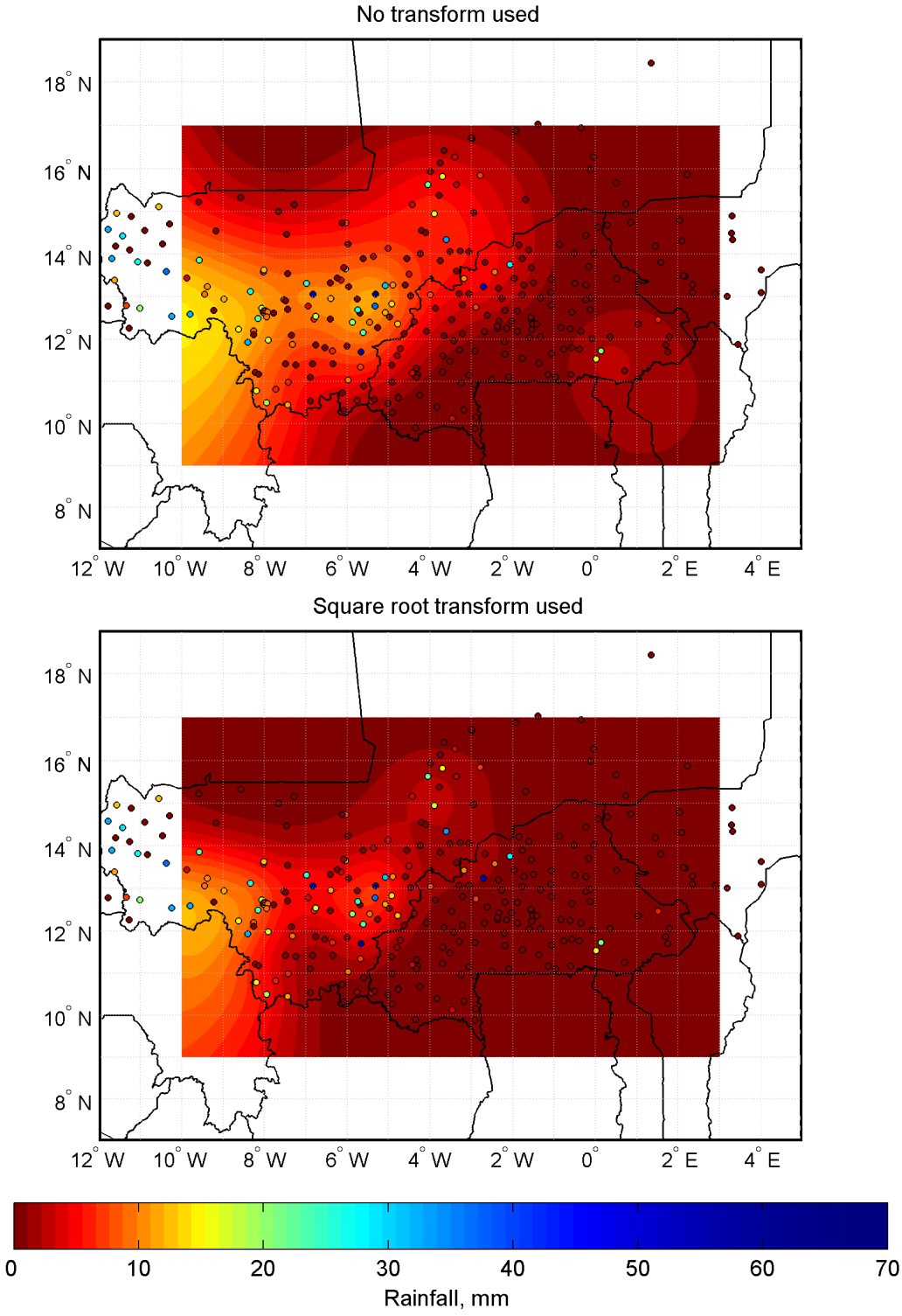

Figure 3.9. Smoothing thin-plate splines fitted to daily rainfall for the central region for 9th June 1966, with and without the square root transform. The top plot applies a spline to the raw data; the fitted surface has smoothing parameter v=0.12. The bottom plot takes the square root of the raw data, fits the surface, and takes the square of the output; the fitted surface has smoothing parameter v=0.06.

3.3.3. A Comparison of Station and Gridded Data

The extent to which the gridded dataset represents the underlying station dataset can be examined by considering the residual errors of the fitted surfaces at the station locations. Figures 3.10 and 3.11 illustrate the root-mean-square error (RMSE) at each station location. Figure 3.10 shows the error from daily residuals, Figure 3.11 the error from yearly totals. Both figures consider data formed without and with the square root transform.

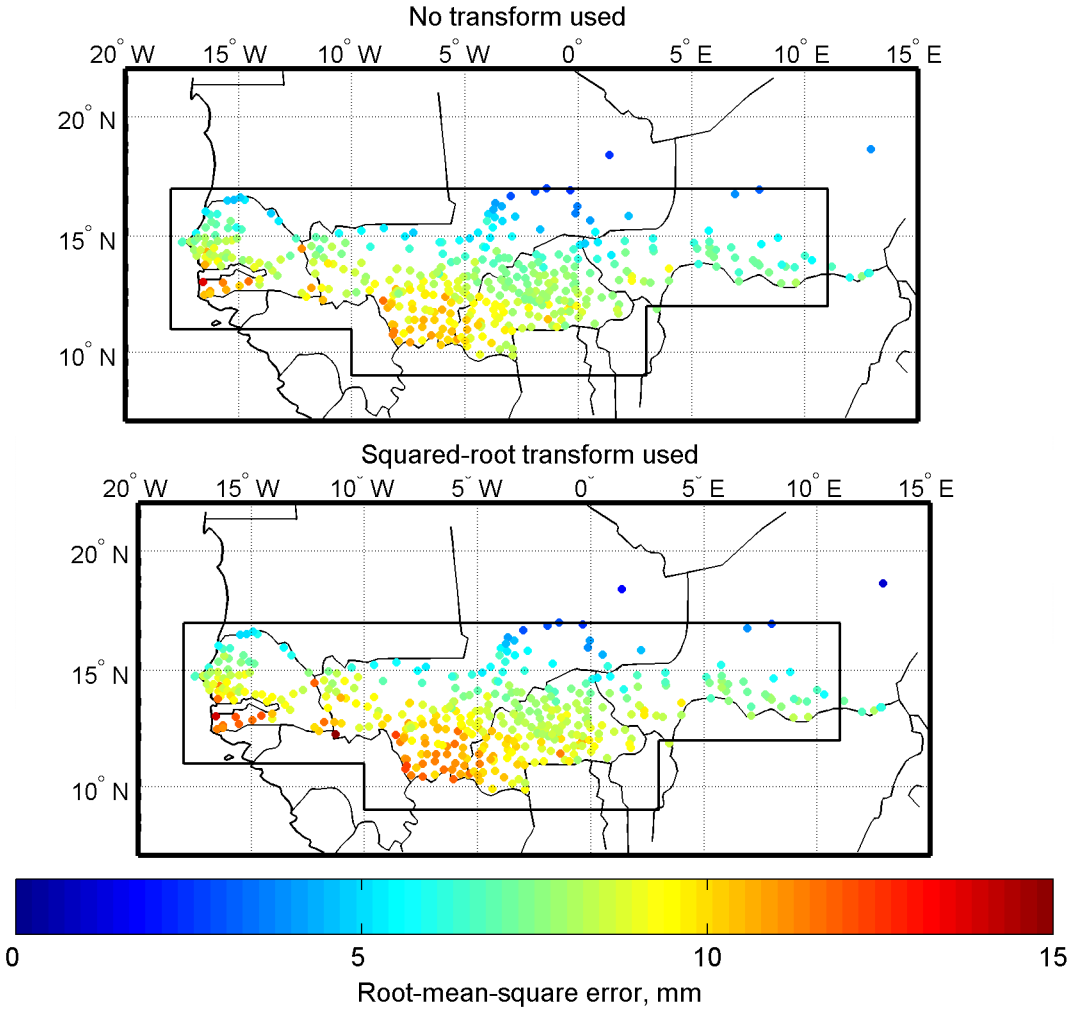

Figure 3.10. Daily root-mean-square error of the gridded dataset when compared to original station values, calculated for each station used in the gridding process. For June-September only.

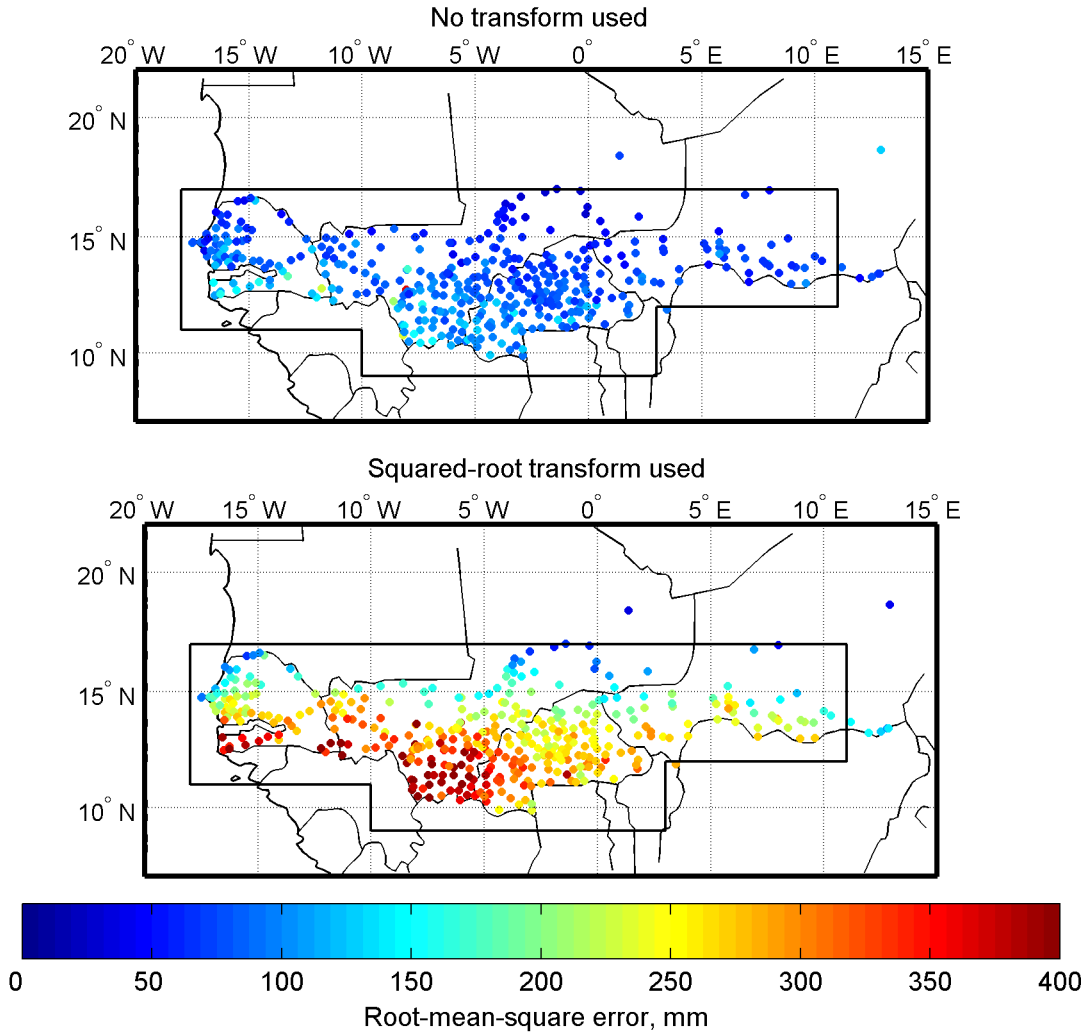

Figure 3.11. Root-mean-square error of yearly wet season (June-September) totals of the gridded dataset when compared to original station values.

For both daily and annual data, RMSE is greater for more the southerly stations. This result is predictable: as rainfall increases, so will absolute measures of error in modelling it. The RAW and SQRT dataset perform similarly for daily data, with errors for SQRT typically being slightly higher. However, at a yearly scale, error for SQRT is much higher.

The reason for SQRT performing poorly at longer timescales can be illustrated using quantile-quantile plots (referred to hereafter as qq-plots). These are a form of scatter plot where the two data series are sorted before plotting, so the *n*th largest *x*-value is plotted against the *n*th largest *y*-value. If the plotted values lie along a straight line, the two series come from the same family of distributions. If they lie along a 45° line, the two series come from identical distributions. Figure 3.12 illustrates the qq-

plot for daily data in Ouagadougou (12.35 °N, 1.52 °W). Other stations produce similar results.

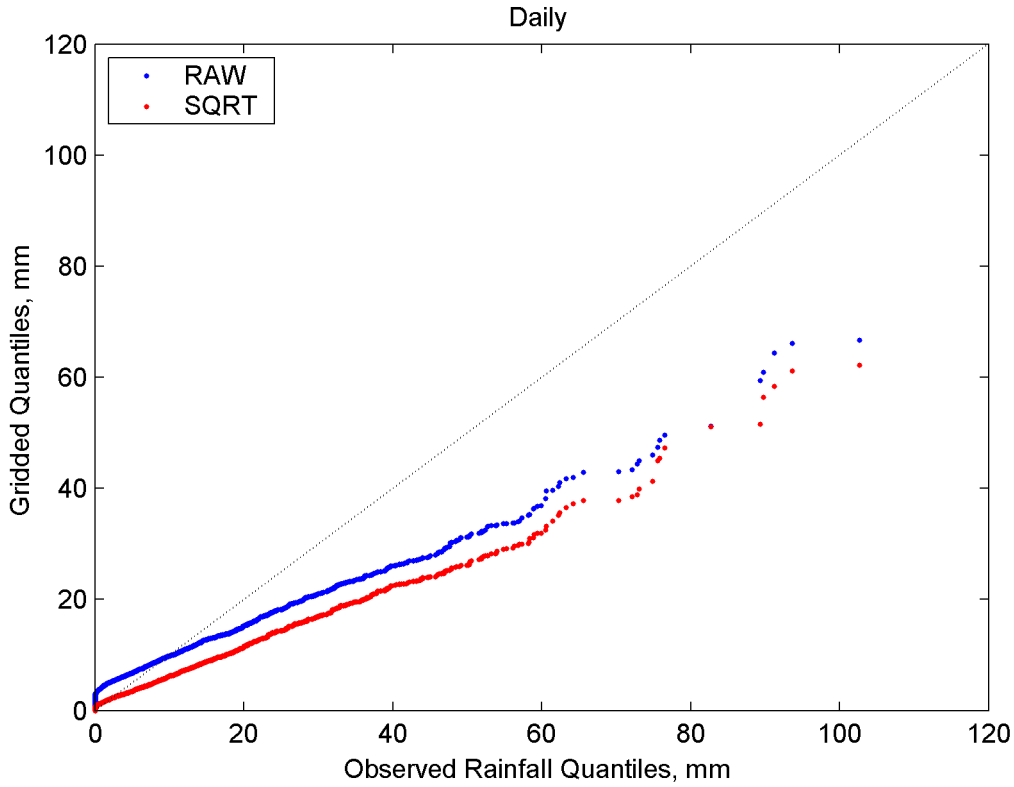

Figure 3.12. Quantile-quantile plot of daily station rainfall vs. gridded rainfall at Ouagadougou (12.35 °N, 1.52 °W) – for June to October data only. RAW dataset is displayed in blue, SQRT dataset in red.

The qq-plot illustrates the inability of either gridded dataset to reproduce extreme values, although this fault is more pronounced in the SQRT data. Also, as noted in the discussion of Figure 3.9, the use of a square-root transform typically lowers all values. Indeed, the RAW dataset tends to over-predict low values, and has very few zero values. This over-prediction of low values compensates for the underprediction of high values, leading to roughly correct long-term values. This is illustrated by Figure 3.13, a scatter plot of annual wet season rainfall at Ouagadougou. As seen, RAW dataset values are closer to the observed than SQRT. Thus, the gridded dataset seems to represent yearly data at a given station well, but daily data poorly.

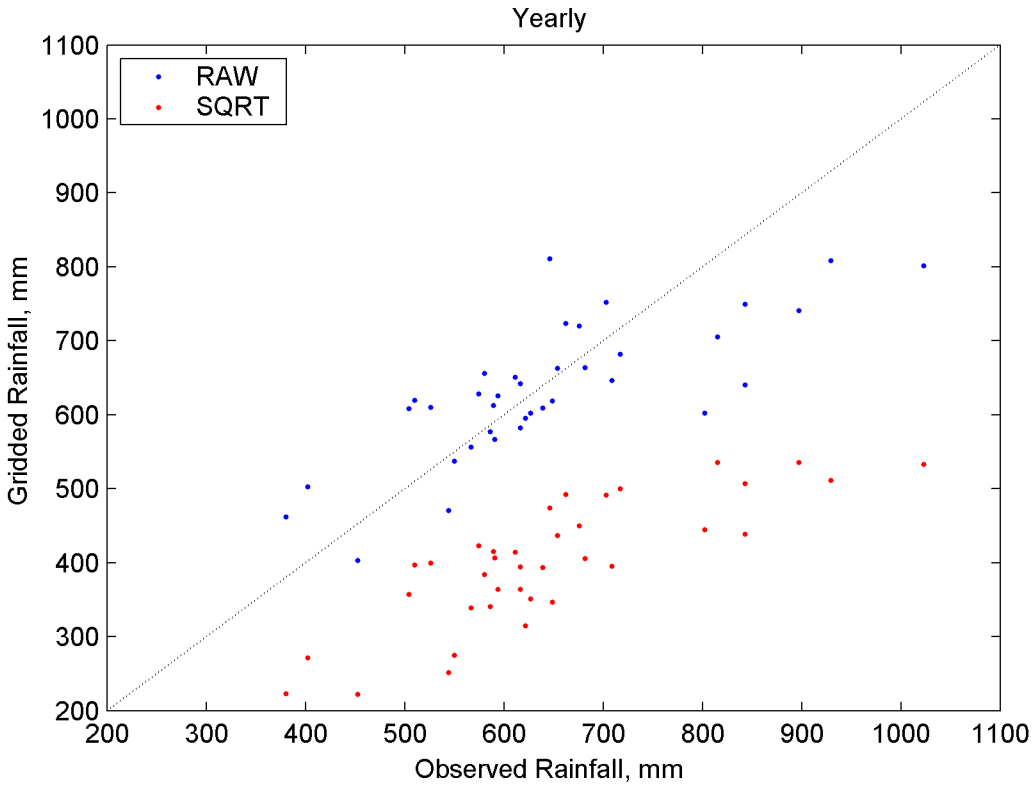

Figure 3.13. Scatter plots of June-September total rainfall vs. gridded rainfall at Ouagadougou, for RAW and SQRT dataset.

This pattern of good long-term but poor short-term representation is not surprising; in fact, it is desirable. As noted before, the gridded dataset does not represent what is happening at a given data point, it represents the average of what is happening in the vicinity of the point. Thus, as an average, the variability of the gridded dataset will be lower than that of point data. However, in the long term, variability at a point will also average out, hence the similarity in yearly values.

This theory can be tested by considering a test case. The grid box 15 ºN, 17 ºW contains the station at Dakar, but also contains ten stations withheld from the original analysis because of lack of data. Figure 3.14 shows the two qq-plots for the gridded wet season rainfall series: firstly against the Dakar series only, secondly against the mean of whatever data exists for the ten stations. The plot of gridded data versus the Dakar series shows the typical overprediction of low values and underprediction of high values. However, when compared to the average of the ten stations, the data series is closer to the line, especially at the start of the line where most of the values are clustered, with roughly 95% of days recording less 20 mm of rainfall averaged over the stations.

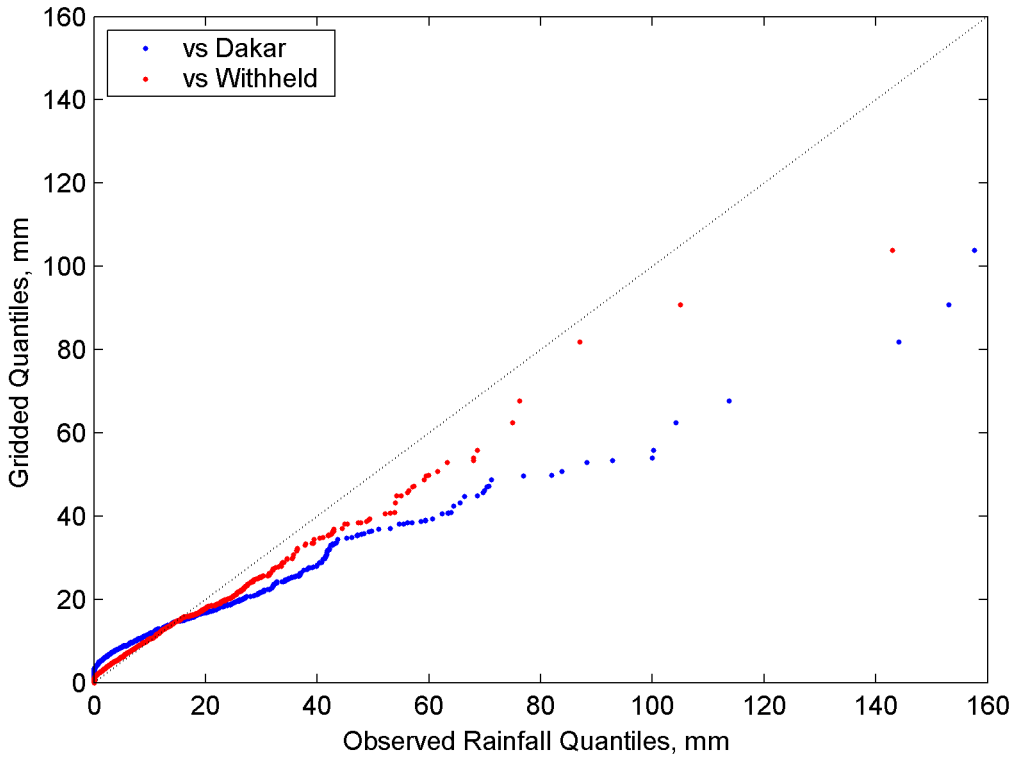

Figure 3.14. Quantile-quantile plot of daily station rainfall vs. gridded rainfall for the Dakar area, Senegal. Blue points compare gridded rainfall to the station in Dakar itself. Red points compare gridded rainfall to the mean of ten stations in the same grid box (15 ºN, 18 ºW) that were withheld from the gridding process. For June-September only.

Unfortunately, qq-plots can only display information about one location. In order to consider the quality of data across the whole area, plots similar to Figure 3.15 were created. Figure 3.15 illustrates the quartiles of the June to September monthly gridded data set for each grid box. It also displays the quartiles of monthly station data obtained from CRU (Climatic Research Unit, University of East Anglia), consisting of 282 stations in the area. Furthermore, the CRU dataset includes data from all countries in the area, thus covering some of the 'gaps' found in the gridded dataset, for example in northern Nigeria. As can be seen, the quartiles of the two datasets are similar, although the minima and maxima of the station dataset are much more dependent on local extreme values.

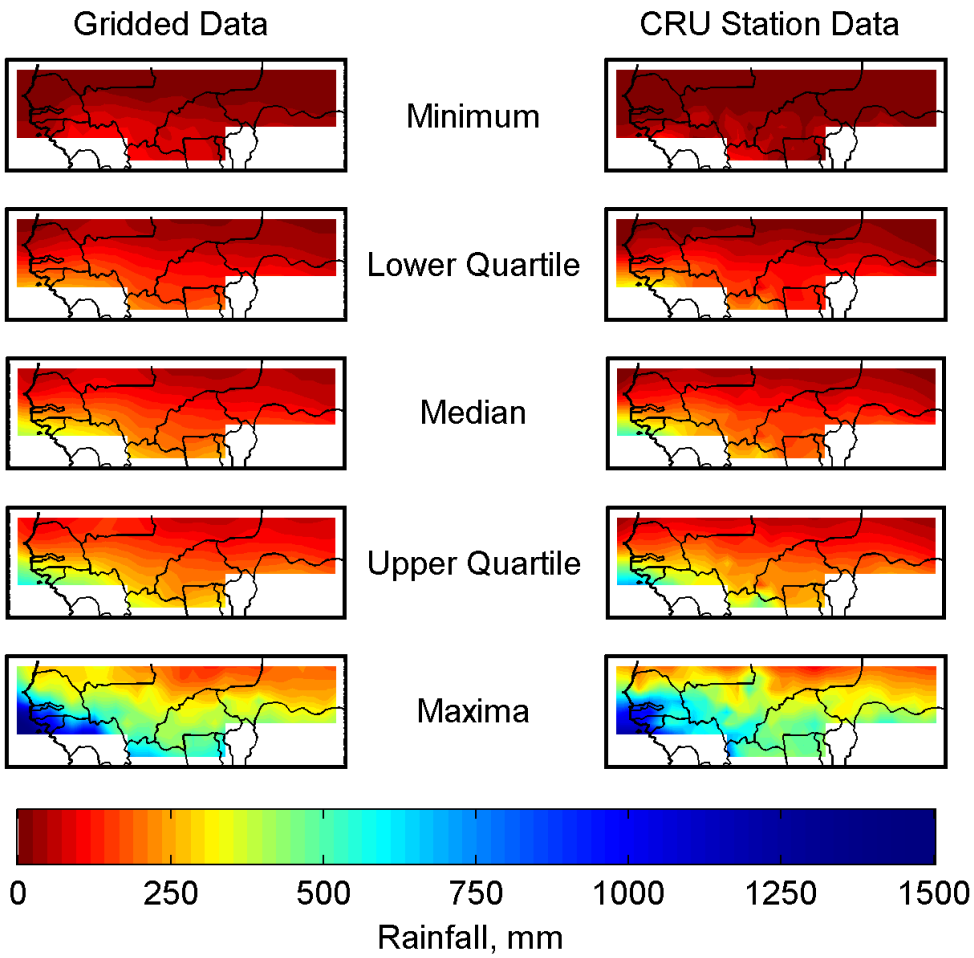

Figure 3.15. Comparison of the quartiles of monthly rainfall for the gridded (RAW) and CRU station data sets. For June to September, 1958 to 1998. CRU data quartiles have been interpolated to enable better comparison between the data sets.

Plots similar to Figure 3.15 were produced for each month of the year. As well as permitting comparison between the two datasets, they allowed for identification of possible errors in the data. For example, the plot for maximum rainfall in September showed an extreme maximum value for several grid boxes situated in northern Senegal. These were identified as being the result of extreme values from one station situated in Salde, 16.17 °N, 13.88 °W, which reported a rainfall of 380.2 mm on the 22nd September 1973. This is almost certainly an error; the CRU data represented in Figure 3.15 suggests a *monthly* maximum of 308.2 mm for the Salde grid box. Furthermore, the next highest amount of rainfall reported in Senegal on that day was 8 mm, from by a station in the far southwest (Ziguinchor, 12.55 °N, 16.27 °W). Other suspicious data was found in the Salde record, so the location was removed from the analysis and the gridded data recalculated. Interestingly, the SQRT dataset

produced rainfall of less than 1 mm in all Senegal grid boxes on the day in question, suggesting the SQRT data is less prone to spurious errors.

Whilst the analyses listed above are not exhaustive, they do indicate that the gridded dataset usually provides a reasonable representation of rainfall across the Sahel on a daily, monthly and yearly basis. However, lack of data seriously affects this representation at the edges of the data set, particularly in the eastern region. However, this thesis will only analyse grid boxes that contain station data, so these border areas are ignored.

Finally, it was decided to use the RAW dataset in later analyses. The tendency of the SQRT method to suppress extremes does mean it deals well with outliers. However, this is also its downfall, as it is unable to produce sufficient extremes to adequately represent long term rainfall. Conversely, the RAW data can represent average daily conditions in a grid box on most occasions, only failing on the very wettest days.

3.3.4. Representation of the Drought in the Gridded Dataset

A recent paper by Chappell and Agnew (2004) caused controversy by suggesting that the recent drying trend in the Sahel may not exist. Chappell and Agnew noted that over the past seventy years the number of stations recording rainfall in the west of the Sahel has decreased, whereas the number of recording stations in the east of the Sahel has increased. The west of the region is typically wetter than the east, as displayed in Figure 3.15. Chappell and Agnew claimed that typical Sahelian rainfall indices did not account for this change in station location distribution, and hence introduced an artificial trend into the index. This suggestion was strongly refuted by Dai et al. (2004), who criticised Chappell and Agnew's method and demonstrated a clear drying trend across the Sahel.

The gridded dataset produced in this chapter should enable the Chappell and Agnew hypothesis to be studied further. As the thin-plate spline method should be largely unaffected by the distribution of station locations, then, if Chappell and Agnew's claim is correct, the gridded data should show no clear drying trend.

This claim has been investigated by fitting a linear trend to the annual rainfall time series for each grid box (Figure 3.16). The top plot illustrates the slope of the fitted trend of each grid box, and the bottom plot shows this slope relative to the mean annual rainfall.

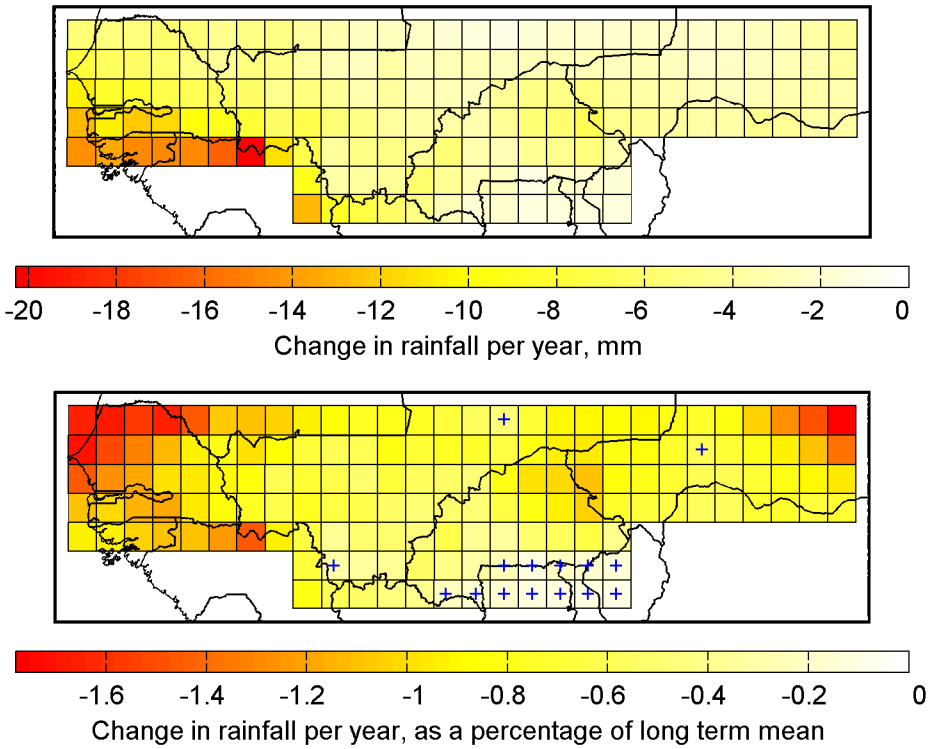

Figure 3.16. Trend of each grid box in the annual gridded rainfall, 1958-1997. The top plot represents the absolute trend; the bottom plot represents the trend as a percentage of the box annual mean rainfall. Trends that are not *significant at a 5% level are indicated by a '+' in the bottom plot.*

The dataset exhibits a significant drying trend across almost the whole Sahel. The relative trends seem to be strongest in the west. Results for the edges, and in particular the corners of the region should be treated with caution, as they typically coincide with areas with little or no data, see Figure 3.6. In particular, the region with insignificant trend is located over northern Ghana, Togo and Benin, where a major gap in data occurs.

Whilst this analysis is fairly naïve, as the drying trend is clearly non-linear, it indicates that across the Sahel, rainfall toward the end of the twentieth century was, on average, less than in the mid twentieth century. Hence, the gridded data provides further evidence that Chappell and Agnew's hypothesis is incorrect.