Chapter 4: Formulating the Predictor Variables

Chapter 3 focused on the construction of a gridded daily rainfall dataset, which will form the response variable in the statistical model that is the goal of this thesis. This chapter will focus on the creation of a suite of predictor variables.

As noted in Chapter 2, the suite will be made up of atmospheric variables, as they can have a direct effect on rainfall in the Sahel. Whilst the review of past studies in Chapter 2 demonstrated that variables such as sea surface temperature or measures of land surface condition undoubtedly have an influence over Sahel rainfall, they are not included here. This is principally because variations take place over a much longer time scale than daily, and because any change in these variables should be reflected in changes in atmospheric composition and dynamics.

This chapter includes a description of the underlying data to be used for prediction, a review of the techniques used to transform it into useful predictor variables, and analyses to identify a suitable domain size.

4.1. The NCEP / NCAR reanalysis dataset

One of the biggest challengers facing researchers in analysing the West African climate is the comparative lack of physical observations. For example, Chapter 2 reported that Grist and Nicholson (2001) attempted to use radiosonde and pibal data to analyse the link between atmospheric variability and rainfall, but were severely hampered by a lack of data. Therefore, a common approach is to use reanalysis data, which is observations data from various sources processed using a General Circulation Model (GCM) to make the data physically consistent and fill missing values.

The most frequently used reanalysis data comes from the National Centers for Environmental Prediction / National Center for Atmospheric Research reanalysis project, hereafter referred to as the NCEP reanalysis (Kalnay et al., 1996). The aim of the project was to provide a 'frozen state of the art analysis / forecast system', as use of previous forecasts in climatological analysis produced inhomogeneities

whenever the forecast system was improved. Originally data were produced for 1957-1996, but now the record goes back to 1948, and is continuously updated.

The NCEP reanalysis provides data for 28 vertical levels at a $2.5° \times 2.5°$ resolution for a wide range of variables. Many data are reported for every six hours, but this study uses the daily averages, to match the gridded rainfall data. NCEP variables are grouped into four classes to illustrate whether they are primarily influenced by observations or by the model. Class A are described as 'strongly influenced by observed data', whereas whilst class B are directly influenced by observations, the model also has a strong effect upon them. Kalnay et al. suggest class C variables, such as rainfall, should be used with caution, as they are completely derived from the model. Finally, class D variables, such as albedo, are obtained purely from climatological values, and hence are independent of the model.

Other reanalyses projects have been performed. In particular, the ERA40 reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) has been released recently, unfortunately too late for it to be used in this project.

This study uses four atmospheric levels from the NCEP reanalysis: 1000, 850, 600 and 200 hPa. Together these levels cover the main atmospheric processes throughout the troposphere which affect rainfall at the surface. The 1000 and 850 hPa levels contain the main monsoon flow, whereas the 600 and 200 hPa levels contain the two jet streams described in Chapter 2.

This study uses six of the available NCEP variables: geopotential height, air temperature, specific humidity, the zonal and meridional components of wind, and vertical velocity. NCEP classify all of these as class A variables, apart from vertical velocity and specific humidity, which are class B. Specific humidity is not available at 200 hPa, so data from the highest available level, 300 hPa, was used instead.

Originally, the intention was to use the NCEP data over the same period as the rainfall data, that is 1958-1997. Unfortunately, studies by Camberlin et al. (2001) and Janicot et al. (2001) compared NCEP data with observations over West Africa, and found problems with the reanalysis data prior to 1968. This was demonstrated most clearly by Janicot and Sultan (2001), who used the West African Monsoon Index (WAMI), defined by Fontaine et al. (1995), as an observational dataset. The

WAMI is calculated directly from rawinsonde data, and compares the monthly mean standardised anomaly of wind speed at 900 hPa over the Sahel with the zonal wind component at 200 hPa. The authors compared this index with a WAMI calculated from the NCEP reanalysis. The two indices were very similar post-1968, but differed previously. However, when the data were subjected to a high-frequency filter, the indices were similar over the whole period. This suggests a long-term bias in the NCEP data. This and other studies indicate that the problem results from a change in the number of observations prior to 1968; particularly in land surface and rawinsonde observations (Poccard et al., 2000), and ship reports (Camberlin et al., 2001).

Because of these inhomogeneities, Camberlin et al. (2001) and Janicot and Sultan (2001) recommend not using NCEP data for years before 1968. Unfortunately, these studies were not discovered until some of the analyses in this thesis were carried out, so some pre-1968 data has been used. However, where this has been the case, further analyses have been carried out to ensure that results have not been significantly affected.

4.2. An Introduction to Principal Component Analysis

Having identified a suitable dataset for further analysis, the next step was to choose a method to extract the most important information from the dataset, and thus create the suite of variables to be used in the final model. The next section introduces Principal Component Analysis (PCA), an established technique for the efficient extraction of ranked patterns of variance from a dataset. This section explains why PCA was chosen, examine some of the main problems encountered when using the method, and consider possible extensions and alternatives.

4.2.1. The Purpose of Principal Component Analysis

Principal Component Analysis (hereafter referred to as PCA) is a multivariate statistical technique that has been widely used in climatological studies[3]. It aims to 'reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set' (Jolliffe, 2002), p1. In the climatological field, this enables the identification of dominant patterns of simultaneous variation of a given statistical field. Hence, in the context of this thesis, the complexities of variation in several variables at thousands of grid points measured at several heights could be simplified to just a few factors.

PCA is dogged by confusions over terminology. PCA is often referred to as Empirical Orthogonal Function (EOF) analysis, when used in a climatological context. Some authors (such as Jolliffe, 2002), suggest that the terms can be used interchangeably, whereas others (Richman, 1986) insist that the terms distinguish between two variants of the technique.

Similarly, the notation used to define PCA varies considerably; those used here are based on Jolliffe (2002). Given a vector of p random variables, let us call it $\mathbf{x}$, we aim to find a linear combination of these variables with maximum variance, let us call it z_1. The linear combination can be expressed mathematically as:

$$z_1 = \boldsymbol{\alpha}_1'\mathbf{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1p}x_p \qquad (4.1)$$

where $\boldsymbol{\alpha}_1$ is a vector of p coefficients, and $\boldsymbol{\alpha}_1$' signifies the transpose of $\boldsymbol{\alpha}_1$. Note that the choice of sign of $\boldsymbol{\alpha}_1$ is arbitrary, as the variances of $\boldsymbol{\alpha}_1$ and $-\boldsymbol{\alpha}_1$ are equal.

So, we have found z_1 with maximum variance. The next step is to find a second function, z_2, or $\boldsymbol{\alpha}_1'\mathbf{x}$, that is uncorrelated with z_1 but with maximum variance. This process continues, at each step finding z_k that maximises variance but is uncorrelated with z_1, z_2, …, z_{k-1}. In theory p components could be found, but typically an often

[3] Jolliffe (2002) reports that nearly 25% of articles appearing in the *International Journal of Climatology* in 1999 and 2000 used some form of PCA (p71).

small subset can account for most of the total variance in the system, so only the first m are calculated, where $m << p$.

Methods of calculation have been described in detail elsewhere (for example see Jolliffe, 2002 or von Storch and Zwiers, 1999). The coefficients of $\boldsymbol{\alpha}_k$ are the elements of the eigenvector of $\boldsymbol{\Sigma}$ corresponding to the kth largest eigenvalue, λ_k, where $\boldsymbol{\Sigma}$ is the covariance matrix formed from $\mathbf{x}$. Again, much confusion can be caused by the wide range of terminology for the $\boldsymbol{\alpha}_k$ and the z_k. Here the $\boldsymbol{\alpha}_k$ will be referred to as the principal component (or PC) loadings or patterns, and the z_k the principal component scores or time series.

So at each stage we wish to maximise the variance of z_k. However, in order to make sense of this problem some constraint must be imposed, to create some upper bound for the coefficients. This constraint is referred to as a normalisation constraint. A common choice is:

$$\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = 1 \tag{4.3}$$

which forces each eigenvector to have unit length. Furthermore, the variance of z_k is given by λ_k^2. It is this method which Richman (1986) refers to as EOFs. He defines PCA as the method that uses another common normalisation constraint:

$$\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = \lambda_k \tag{4.4}$$

Another useful choice (see Jolliffe, 1990) is:

$$\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = \frac{1}{\lambda_k} \tag{4.5}$$

which gives the variance of $z_k = 1$ for all k. Thus the choice of normalisation constraint determines whether information on the size of the eigenvalue λ_k is contained in the PC scores (case 4.3), the PC loadings (case 4.5) or both (case 4.4).

Note that as the PC loadings are eigenvectors, they are orthogonal (so for $i \neq j$, $\boldsymbol{\alpha}_i'\boldsymbol{\alpha}_j$ is zero). So the PC loadings are orthogonal, and the PC scores are uncorrelated.

These properties are sometimes referred to as the orthogonality properties of the components (Mestas-Nunez, 2000).

When moving from a theoretical to a practical formation of a PCA, where N observations are taken for each of the p variables, it is common to express formula 4.1 in full matrix form as follows:

$$\mathbf{Z} = \mathbf{XA} \tag{4.6}$$

where $\mathbf{X}$ is a $N \times p$ matrix of data observation, whose (i,j)th element is the ith observation of the jth variable, $\mathbf{A}$ is a $p \times m$ loading matrix, whose jth column is the vector of loadings for the jth principal component, and $\mathbf{Z}$ is the $N \times m$ score matrix, whose (i,j)th element is the score of the jth component relating to the ith observation. Principal component loadings are estimated using the sample covariance matrix, often referred to as $\mathbf{S}$, whose (i,j)th element is the covariance between variable i and variable j when $i \neq j$, and whose diagonal (j,j)th element is the variance of variable j.

By definition, the matrix $\mathbf{A}$ is orthogonal, so its inverse is equal to its transpose. Hence (4.7) is often expressed as:

$$\mathbf{X} = \mathbf{ZA}' \tag{4.7}$$

Preisendorfer (1988, p30) refers to (4.6) as the analysis formula and (4.7) as the synthesis formula. The synthesis formula indicates that we can 'decompose' the data matrix into two parts: a score matrix and a loading matrix.

A common practice is to standardise the variables, that is the columns of $\mathbf{X}$, before proceeding with the analysis. This ensures the variance of each variable is one, and hence prevents bias toward variables with greater variance. Standardisation of variables is equivalent to finding the eigenvectors and eigenvalues of the correlation matrix, rather than the covariance matrix.

Complications occur if two or more eigenvalues are equal. Together, the q eigenvectors corresponding to the equal eigenvalues span a q-dimensional subspace of the whole sample space, but they cannot be uniquely defined (Jolliffe, 2002, p27). Any of the infinite number of q mutually orthogonal vectors that span the given

subspace can be used: in effect, the q patterns cannot be separated. The q components are said to be degenerate (von Storch and Zwiers, 1999, p296).

Degeneracy is extremely rare in practice, as sampled eigenvalues are almost never equal. Nevertheless, the theory has some important ramifications. It is quite common to have some eigenvalues that are nearly equal. North et al. (1982) demonstrated that when this occurs, high sampling errors can cause the patterns to form an 'effectively degenerate multiplet'. To put it more simply, the patterns can become mixed. The study also showed that larger sample sizes increase the chance that patterns can be separated, although separation can never be guaranteed (see the case study of Richman, 1986).

North et al. (1982) formed a 'rule of thumb' to estimate where patterns may become mixed by deriving the following estimate of the typical error $\Delta\lambda_i$ for any given eigenvalue λ_i:

$$\Delta\lambda_i \approx \sqrt{\frac{2}{n}}\lambda_i \tag{4.8}$$

where n is the number of independent samples. If $\Delta\lambda_i$ is greater than the distance to a neighbouring eigenvalue, then it is possible that the patterns may be mixed. Unfortunately, in a climatological setting, samples can hardly ever be considered independent, and hence finding a value for n can be problematic.

Other problems can occur from zero or near-zero eigenvalues. This can result in an ill-conditioned sample correlation or covariance matrix, and hence unstable results, if sample size is too small. A common recommendation is to have more independent observations than variables (Richman, 1993).

One issue yet to be discussed is how to select a value for m, the number of principal components to be calculated. Many techniques exist, varying dramatically in their sophistication, but all have to make a trade-off between a small value of m, which provides for the greatest reduction of dimensionality of the data set, and a large m, which retains the greatest proportion of variance.

The simplest techniques retain principal components that collectively explain a given proportion of total variance. So, for example, the first *m* components may be chosen so that together they contain 95% of the total variance. This simple rule is surprisingly effective, as typically the first few components contain a large majority of the total variance. However, the selection of what percentage of total variance to retain is arbitrary.

Another simple method is to retain any individual component with a variance above a given cut-off point. A special case of this method, known as Kaiser's Rule, is to retain all components with a corresponding eigenvalue greater than one. Thus only components that contain more variance than the average variable (or any variable in the case of a correlation based PCA) are retained (Tabachnick and Fidell, 2001).

Other popular methods are based on a plot of successive eigenvalues, often referred to as a scree graph. Named by Cattell (1966), although already widely used at the time, the 'scree test' typically bases the choice of number of components to retain on the point when the plot moves from being steep to shallow, a point often referred to as the 'elbow' of the graph. Whilst useful in many cases, interpretation is not always straightforward. An example scree plot is displayed later in this chapter, as Figure 4.4. However, this plot could be interpreted as having two elbows, one at the fourth component, and one at the sixth. Furthermore, there is disagreement as to whether the cut-off should include the component at the elbow (Jolliffe 2002, p115-7; Tabachnick and Fidell 2001, p621).

Cattell (1966) preferred a slightly different formulation for a scree test. Instead of basing a choice on the elbow of the plot, Cattell recommends looking for the point beyond which the graph first becomes more-or-less a straight line, and taking to cut-off to the right of this point (Jolliffe 2002, p117). Hence, Cattell's method would retain six components for Figure 4.4, although again, this result is open to interpretation.

Many other criteria for selecting the number of components to retain exist; the textbooks by Preisendorfer (1988) and Jolliffe (2002) both have chapters devoted to different methods. Some of the more complex attempt to use statistical justification for a cut-off point. However, Jolliffe suggests that, at present, such methods seem to

"offer little advantage over the simpler rules in most circumstances". Similarly, von Storch and Zwiers (1999) advise against using complicated selection rules.

In climatological studies, we typically have three 'entities' to consider; time, location (i.e. a station or grid point) and meteorological fields (such as temperature, pressure or precipitation), one of which must be held constant. Richman (1986) divides climatological PCAs into six 'operation modes', based on the modes of Cattell (1952). Each of these modes is defined by which of the entities we hold constant, which is represented in the columns of the data matrix *X*, that is the variables, and which is represented by the rows of *X*, the cases. These six modes are summarised in Table 4.1. Note that Cattell described two extra modes: S^2 and T^2, which have the same entity in the rows of *X* as the columns.

PC mode	'Variables': Columns of *X* denote	'Cases' Rows of *X* denote	Fixed entity
O	time	field	location
P	field	time	location
Q	location	field	time
R	field	location	time
S	location	time	field
T	time	location	field

Table 4.1. The six modes of decomposition for climatological studies. Adapted from (Richman, 1986).

After analysis, the entity described in the columns of *X* is associated with the PC loadings, and the entity described in the rows of *X* is associated with the PC scores. Many climatological analyses are carried out in S-mode, hence the use of the terms 'patterns' and 'time series' to describe the loadings and scores respectively.

Note that the location entity usually covers more than one dimension. Analysis points are sometimes represented by a range of stations scattered across a map, usually at different heights. Alternatively the may be a series of points on a two (or three) dimensional grid. In this case, usual practice is to 'unfold' these dimensions

into one, thus ignoring any structure in the distributions of locations across the analysis region. Hence, any clustering of locations in a particular area will result in a bias toward that region. Thus, the use of regularly spaced gridded data is preferable to irregularly scattered station data.

Note that further modes can be created by 'unfolding' the field entity in a similar fashion. For example, supposing that we measured several fields at several locations on many separate occasions. We could perform a PCA using the occasions as rows, and using each field at each location as a separate variable. This approach is sometimes called Extended Empirical Orthogonal Functions (EEOF) (von Storch, 1999, p298).

4.2.2. Rotation of Principal Components

Once the initial stage of the PCA has been carried out, we are left with m PCs that geometrically span the most informative m-dimensional subspace of the p-dimensional space spanned by the original variables (Jolliffe, 1993). However, they do not uniquely define this subspace. Furthermore, typically the complexity of the structure of each consecutive PC pattern increases, for example see Richman (1986) and the analyses later in this chapter. Therefore, a common practice is to find an alternative set of vectors spanning the given subspace which have a much simpler structure. This process is referred to as rotation.

Algebraically, this amounts to finding a suitable $m \times m$ matrix **T**, which is used to construct a new loading matrix **B**:

$$\mathbf{B} = \mathbf{AT} \tag{4.9}$$

Once a new rotated loading matrix has been found, new rotated scores can be calculated using equation (4.6), but replacing **A** with **B**. Often **T** is orthogonal, referred to as an orthogonal rotation. However, non-orthogonal transformations, known as oblique rotations, are available, see Nicholson and Palao (1993) for an example. Again, terminology can cause great confusion; an orthogonal

transformation will not necessarily produce a set of orthogonal patterns, as discussed below.

There are many different techniques used to select **T**, and all differ in their definition of 'simple structure', see Richman (1986), for a comprehensive list. Most operate by attempting to obtain loadings or scores that either have a large absolute value, or are near to zero. The most commonly used orthogonal transformation, varimax, maximises the variance of loadings within each pattern. Another useful choice, the quartimax criterion, maximises the variance of one variable across all the components (Tabachnick & Fidell 2001, p614). Jolliffe (2002) suggests that the choice of rotation criteria is usually less important than the choice of how many variables to rotate (that is the choice of *m*). However, he cites some studies where a different choices of rotation criteria result in significantly different results.

Before rotation, the PCA exhibits two orthogonality properties: scores are uncorrelated and patterns are mutually orthogonal. The cost of rotation is the sacrifice of at least one of these properties. The choice of normalisation constraint will decide which properties are lost. For orthogonal rotation, Mestas-Nunez (2000) demonstrates that the use of normalisation constraint (4.3), referred to by Richman (1986) as an EOF, preserves orthogonality of patterns but results in correlation between scores. Constraint (4.5) results in uncorrelated scores, but loses orthogonality of patterns. Constraint (4.4), Richman's PCA, sacrifices both orthogonality properties. Oblique rotation, by definition, sacrifices orthogonality of patterns.

In addition to providing a simpler structure, and hence greater interpretability, rotation has been suggested to have a number of other benefits. Firstly, Richman (1986) notes that rotated patterns have a smaller sampling error than unrotated patterns; hence when eigenvalues are similar, rotated pattern can be separated more easily. Jolliffe (1987) argues that there is no reason why all *m* components need to be rotated, suggesting that effectively degenerate multiplets could be identified using North's rule of thumb, and then each multiplet could be rotated. Thus, the simplest solution for each multiplet would be used to represent the variance contained by its components.

In order to illustrate another benefit of rotation, it is necessary to introduce the concept of Buell patterns. In a series of studies, Buell discovered that principal components of two-dimensional fields often exhibit similar patterns, and these patterns are heavily influenced by the shape of the domain in which the analysis is carried out (Buell, 1975; Buell, 1979). Figure 4.1 illustrates a series of grid points forming a 9×9 grid, perhaps representing grid points on which climatological data have been measured. Usually, there is a definite spatial structure to data observed on such a grid, for example the temperature at a given point is likely to have strongest correlations with points immediately surrounding it, with correlations weakening as distance increases.

In Figure 4.1, the left panel illustrates typical strength of correlation with the black grid point at location (5,5), whereas the right figure illustrates correlation strength with the grid point at (1,1). Strongest correlations are represented by red circles, then as distance increases and correlations reduce, colours fade to orange, yellow then to white. Notice that on the left hand grid only a few points are yellow, whereas many are in the right hand grid.

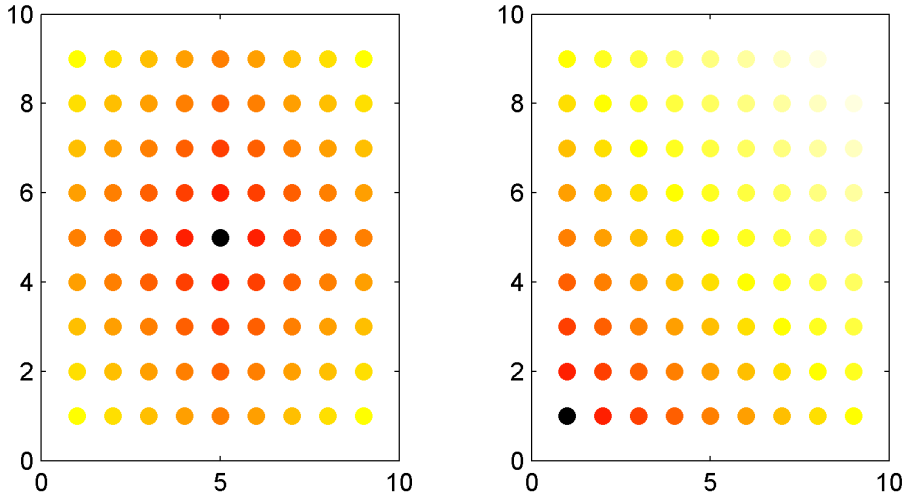

Figure 4.1. Examples of underlying structure of correlation in a typical gridded data set. The depth of colour in each plot represents strength of correlation with the black grid point. See text for a full explanation.

This illustrates a potential weakness in spatially based PCAs. As the central grid points are, on average, likely to be more strongly correlated with other grid points, they exert far more influence on the analysis than points nearer the edge. Hence, in many PCAs the first pattern is a global pattern, with most grid points having the same sign, but with loadings being greatest in the centre.

The second PCA represents the main axis of variance, once the variance of the first PCA has been removed. As the first pattern focused on the centre, most remaining variance is located at the edges of the domain. Hence, the second pattern tends to be one edge versus the opposite edge (such as left versus right), and the third pattern tends to be a 90° rotation of the second (up versus down). Buell calculated predicted patterns for square, rectangular and triangular domains. Figure 4.2 illustrates one of Buell's results for a 6×6 square (Buell, 1975), where the correlation r between two points is given by:

$$r = \exp[-\left(x^2 + y^2\right)^{\frac{1}{2}}] \tag{4.10}$$

where (x,y) represents the coordinate separation between the two points concerned. All 36 components are illustrated.

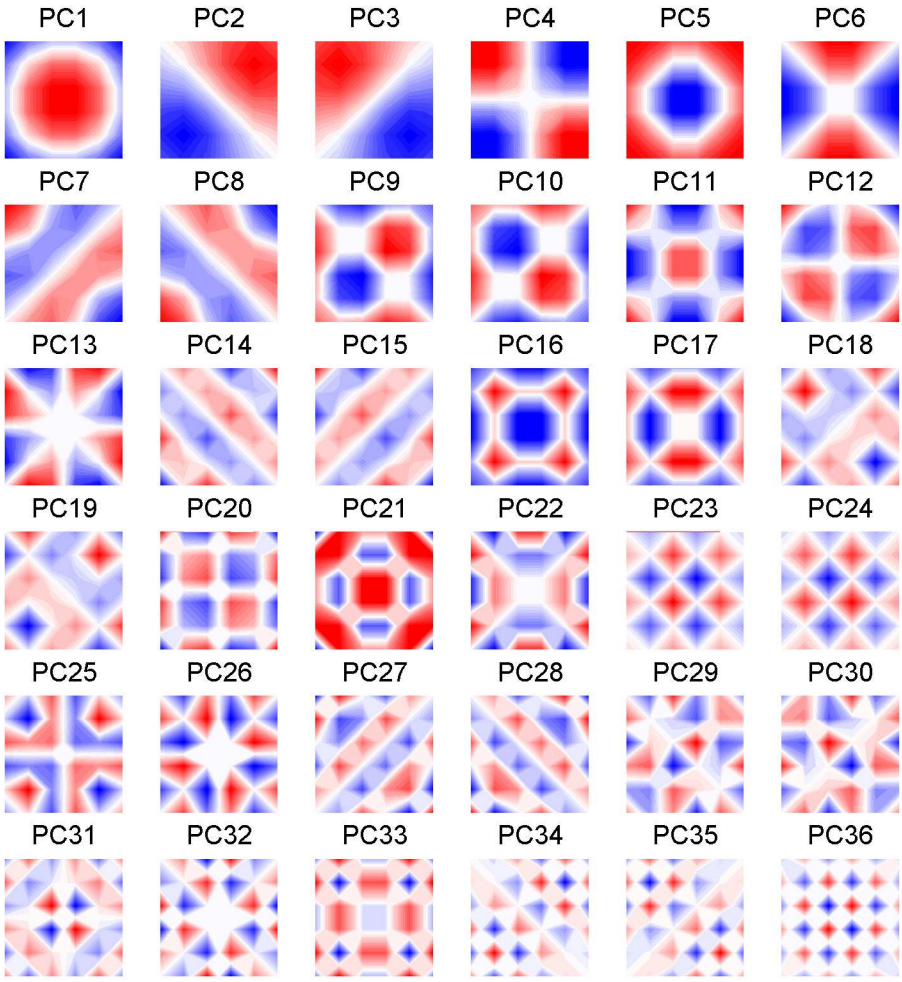

Figure 4.2. Example of Buell patterns for a 6 × 6 square, after (Buell, 1975). See text for explanation.

Note that in the example given in Figure 4.2 there are many pairs of equal eigenvalues, such as the second and third eigenvalues. Hence the order of many pairs is arbitrary, explaining why the ordering is slightly different to those shown in Figure 2 of Buell (1975). Furthermore, where eigenvalues are equal, the patterns are degenerate. Hence components two and three could be represented in an infinite number of ways. For example, instead of bottom-left versus top-right and bottom-right versus top-left; they could be represented as top versus bottom and left versus right. Also, note that in a climatological or meteorological context, a physical interpretation of all but the first few patterns would be impossible.

As rotation tends to force loadings to high values or zero, it often causes rotated patterns to be localised. This is clearly indicated by Figure 4.3, which shows the result of performing a varimax rotation on the first ten PCs, chosen according to

Kaiser's rule (use all components with an eigenvalue greater than one). All ten patterns show a small area in opposition to the rest of the domain, physically far more interpretable than the unrotated solution.

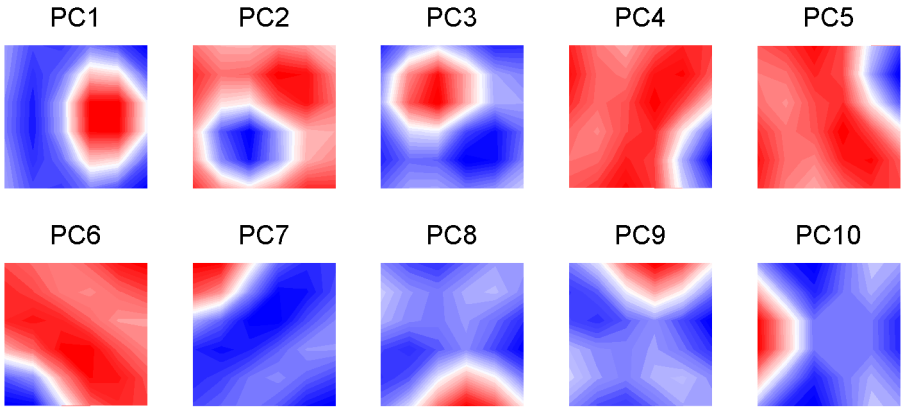

Figure 4.3. The varimax rotation of the first ten components from Figure 4.2.

Buell's results suggest the shape of the domain has a significant effect on principal component patterns. However, this does not render them useless. As Jolliffe (1987) indicates, even if the second PC is predictably a one side versus the opposite, it is not always predictable how it will be oriented. Furthermore, the result indicates the 'direction of maximum variation of the climatological variable of interest'. Jolliffe also claims that the strong spatial correlation of variables is the main reason for predictable patterns in climatological studies, with domain shape being a significant secondary factor.

The exchange between Legates (1991, 1993) and Richman (1993) shows the level of controversy the issue can cause. Legates produces two examples of a global PCAs on monthly data, one on temperature, one on rainfall, which he claims do not exhibit Buell patterns. Despite the deficiencies in Legates' analysis that Richman identifies, including the exceptionally small sample size, it does illustrate some important points. The rainfall analysis demonstrates that variables without the kind of strong domain-wide correlation illustrated in Figure 4.1 need not produce classic Buell patterns, although this does not mean they are unaffected by domain shape. The temperature example demonstrates that different authors can come to wildly different conclusions from the same analysis; Legates claims they do not represent Buell patterns, whilst Richman, supported by Smith et al. (1990), claims they do.

Buell patterns concern data that shows a high level of spatial correlation. Similar problems can occur for data with high levels of temporal correlation, although these seem to have been subjected to less scrutiny (Buell, 1979; Jolliffe, 2002, p298).

The issue of physical interpretation of PCAs has also been the cause of much discussion. Richman (1986) in particular seems to be of the view that the weaknesses of unrotated PCA noted above mean that they can cause severe, possibly terminal, problems in interpretation. Indeed, Richman (1987) notes that the original proponents of PCA never intended it to be used to discover physical patterns. Furthermore, Richman (1986) notes that often unrotated physical modes have no basis in reality, citing an example of PCA on 3-day precipitation over the United States. The first unrotated PCA represents the expected domain-wide variability. Richman argues this makes no physical sense, as it never rains (or is dry) simultaneously across the whole domain. Conversely, Dommenget and Latif (2002) (followed by discussion by Behera et al., 2003; Dommenget and Latif, 2003 and Jolliffe, 2003) suggest that rotating a PCA can sometimes hinder physical interpretation, as they produce patterns that 'have very little to do with climate physics'. Dommenget and Latif use a combination of real life and artificially created examples to illustrate that rotated PCAs can create 'artificial' patterns that have no basis in reality, although these claims are disputed by the following discussion papers.

The issue of the failure of PCA, whether rotated or not, to produce physically meaningful patterns is tackled by Jolliffe (2002, p296-8). He suggests that the proposed criticisms of PCA listed above are somewhat unfair; the aim of PCA is to produce orthogonal, uncorrelated components that successively maximise variance. If the physical modes do not exhibit these characteristics, then PCA will not uncover them. Indeed, Buell (1979) compares the interpretation of improperly conceived PCs (in the context of his analysis of the effect of domain shapes) with children seeing castles in the clouds.

Perhaps the most serious drawback of rotation is the effect of different flavours of PCA on the final rotated output. In a rotated PCA, we must make a number of choices: whether to use the covariance or correlation matrix, what normalisation constraint to use, how many components to retain, and what rotation criterion to use.

Each of these choices can have a significant effect on the rotated output, although some choices have a greater effect than others.

As noted earlier, the most important choice is how many components to retain before rotating (Jolliffe, 2002, p271). In an attempt to avoid this dilemma, Jolliffe et al. (2002) suggest a number of alternatives to the two-stage "PCA then rotation" approach. These approaches rely on applying an extra constraint when finding each consecutive PC. For example, in the SCoTLASS technique, when finding $\boldsymbol{\alpha}_k'\mathbf{x}$ subject to the normalisation constraint $\boldsymbol{\alpha}_k'\boldsymbol{\alpha}_k = 1$, we apply the extra constraint to the coefficients:

$$\sum_{j=1}^{p} |\alpha_{kj}| \leq t \tag{4.11}$$

for some tuning parameter t. For $t \geq \sqrt{p}$ the solution is the same as for PCA, whereas if $t = 1$, there will be exactly one non-zero coefficient for each loading vector. Selecting a value somewhere between these extremes results in patterns where some coefficients are forced to zero, with solutions showing greater simplicity as t is decreased (Jolliffe, 2002, p288). Unfortunately, this elegant approach is penalised by a severe increase in computational complexity. At present, optimal routines have yet to be developed to make this option feasible for data sets as large as those used in this thesis (Jolliffe et al., 2002).

Rotation should not be viewed as a panacea for all of the problems with PCA. Nevertheless, Richman (1986 and 1993) demonstrates that rotated solutions are less affected by domain dependence and sampling errors, and typically more interpretable. However, these benefits have a cost. In addition to uncertainty about choice of m, at least one of the orthogonality constraints will have to be sacrificed. This could be undesirable, for example if the scores are to be used in a regression analysis (Richman, 1986).

4.2.3. An example of PCA on the Sahelian gridded rainfall data set.

To give an example of how PCA operates, this section presents an analysis on the gridded rainfall data set created in Chapter 3. The ultimate aim of this thesis is to create an empirical model of Sahel climate relating atmospheric variability to rainfall. However, the high dimensionality of the gridded rainfall data set makes it unwieldy. The entire domain contains 230 grid boxes. Hence, mathematically, the data set can be said to be 230-dimensional. By using PCA, it may be possible to reduce the dimensionality of the data set, and hence reduce the number of factors we need to predict in the final empirical model.

The first thing to recognise is that the quality of each of the 230 grid boxes is not equal, as demonstrated in Chapter 3. Therefore, only grid boxes that contain a station used in the gridding procedure are used in this stage of the analysis. This eliminates all the border boxes that were created by extrapolation, and leaves us with 114 boxes, each with 14,610 observations, one for each day of the period 1958-1997.

As rainfall, and hence rainfall variability, is greater in the southern regions, performing a PCA on the covariance matrix would introduce bias against the northern Sahel. Therefore, each grid box time series was standardised prior to analysis, a technique equivalent to performing the PCA on the correlation matrix. The cost of this move is that the component time series are now dimensionless, expressed in terms of standard deviations, rather than millimetres.

Another consideration was the strong seasonal cycle in Sahelian rainfall. A PCA performed on raw data would undoubtedly output this as the leading principal component. This would not be new information; we already know the whole region is dominated by the seasonal cycle, as documented in Chapter 2. However, it would also force all following components to be orthogonal to it. Therefore, it is a common practice to deseasonalise data before analysis, to remove the predictable annual cycle. So, for example, each of the 40 observations made on the 17th August for the grid box 12 °N, 1 °W has the mean of these 40 values subtracted from it. Similar subtractions are performed for each calendar day in each grid box.

The PCA was carried out using the normalisation constraint (4.4), referred to by Richman (1986) as PCA. Rotated results for constraints (4.3) and (4.5) were also

calculated, and show very similar patterns. The analysis was carried out on a desktop PC using MATLAB.

Figure 4.4 and Table 4.2 display information about the first ten eigenvalues of the correlation matrix. Figure 4.4 displays the scree plot, whereas the table gives the eigenvalues, the percentage of total variance explained by each component, and the cumulative percentage of total variance explained by all components up to and including that component. It also gives two estimates of error from North's rule of thumb (4.8), one optimistic (all observations are independent, n = 14610), one pessimistic (only every fifth observation is independent, n=2922). The pessimistic value of n is used purely to illustrate the effect autocorrelation in the rainfall time series would have on the estimated error, the choice of n itself was arbitrary.

Component	Eigenvalue λ	Optimistic $\Delta\lambda$	Pessimistic $\Delta\lambda$	% of total variance	Cumulative %
1	19.48	0.23	0.51	17.09	17.09
2	12.26	0.14	0.32	10.75	27.84
3	8.90	0.10	0.23	7.80	35.65
4	6.76	0.08	0.18	5.93	41.58
5	6.32	0.07	0.17	5.54	47.12
6	4.22	0.05	0.11	3.71	50.82
7	3.81	0.04	0.10	3.34	54.16
8	2.91	0.03	0.08	2.55	56.72
9	2.89	0.03	0.08	2.54	59.26
10	2.54	0.03	0.07	2.23	61.49

Table 4.2. Statistics for the first ten eigenvalues of the PCA carried out on the daily gridded Sahel rainfall data set. See text for further explanation.

The estimated errors in the eigenvalues, whether optimistic or pessimistic, suggest that all components are well separated, with the exception of components eight and nine. Despite the clarity of the leading components, it is clear that they do not explain an overwhelming amount of the total variance in the system. This indicates the complexity of Sahelian rainfall, particularly when it is remembered that the gridding procedure itself removed a large amount of variance.

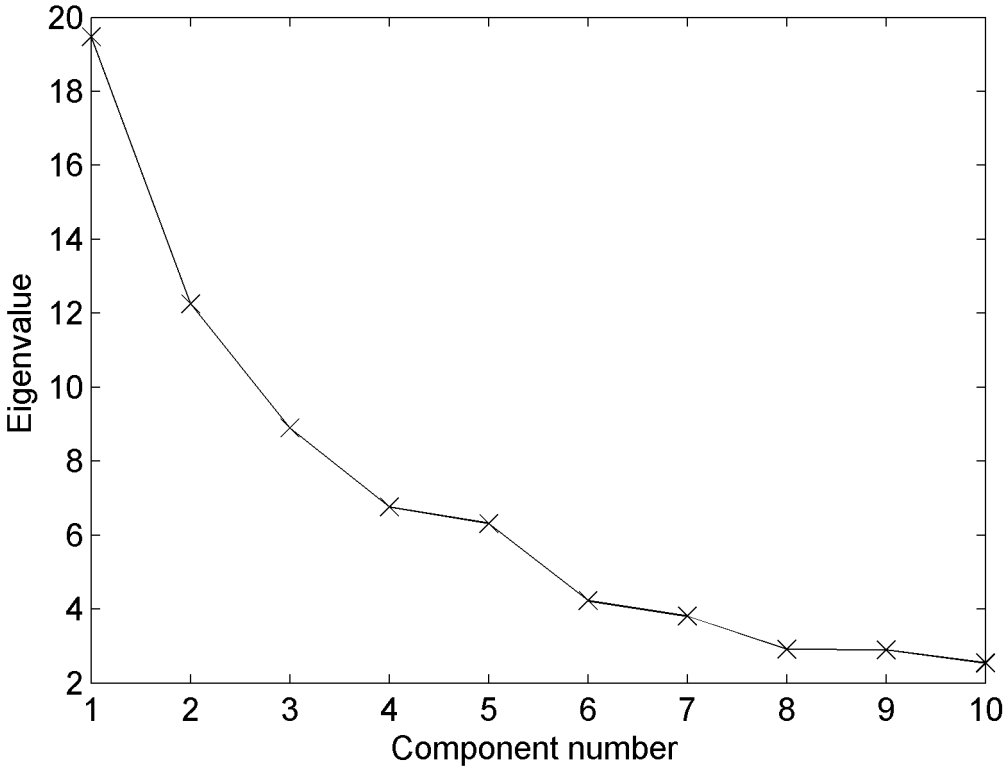

Figure 4.4. Scree plot of the first ten eigenvalues from the PCA of daily gridded Sahel rainfall data.

As a result of this complexity, most selection rules would choose a very high number of components to retain. For example, Kaiser's rule (retain all eigenvalues greater than one) would suggest retaining 22 components. However, this is still a considerable number of factors to predict in a final model. Therefore, it was decided to retain components that would explain a, somewhat arbitrary, value of 50% of total variance. The unrotated component loadings are displayed in Figure 4.5. However, this choice is not entirely arbitrary; as noted earlier in this chapter, Cattell's (1966) formulation of the scree test would retain six components.

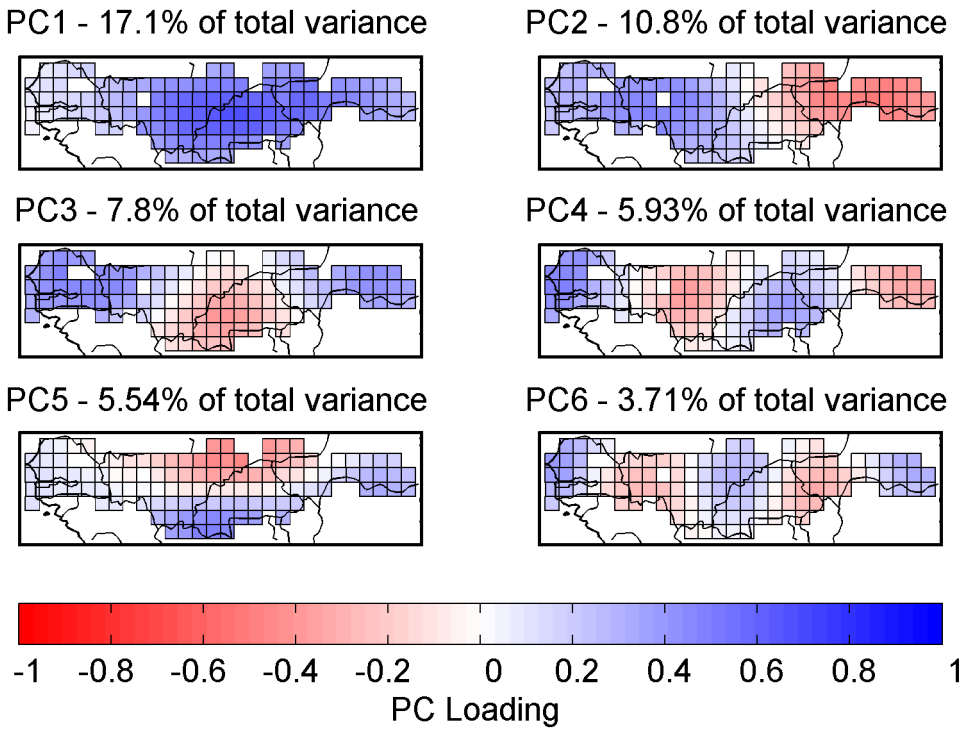

Figure 4.5. Unrotated PC loadings for the first six components of the PCA of daily gridded Sahel rainfall data.

Despite the strange domain shape, the unrotated patterns do seem reminiscent of Buell patterns. In particular, loadings four and six bear little physical relevance to rainfall.

Figure 4.6 illustrates the loading patterns after a varimax rotation is carried out. These results are, as might be expected, much more localised. Negative loadings are dwarfed by the positive loadings. Each pattern describes the rainfall variability in one part of the domain. Therefore, these patterns are used to define coherent regions of rainfall in the final empirical model.

Each of the regions was defined using one of the rotated components, but only considering squares where the loading was greater than 0.5. An exception was given for the blank space that can be seen in the middle of PC2; blank because the square was omitted from the PCA, as it did not contain a station. This square was reintroduced as boxes containing stations surround it, and hence the interpolated estimates of rainfall were comparatively accurate. The regions are illustrated in the next chapter, in Figure 5.1.

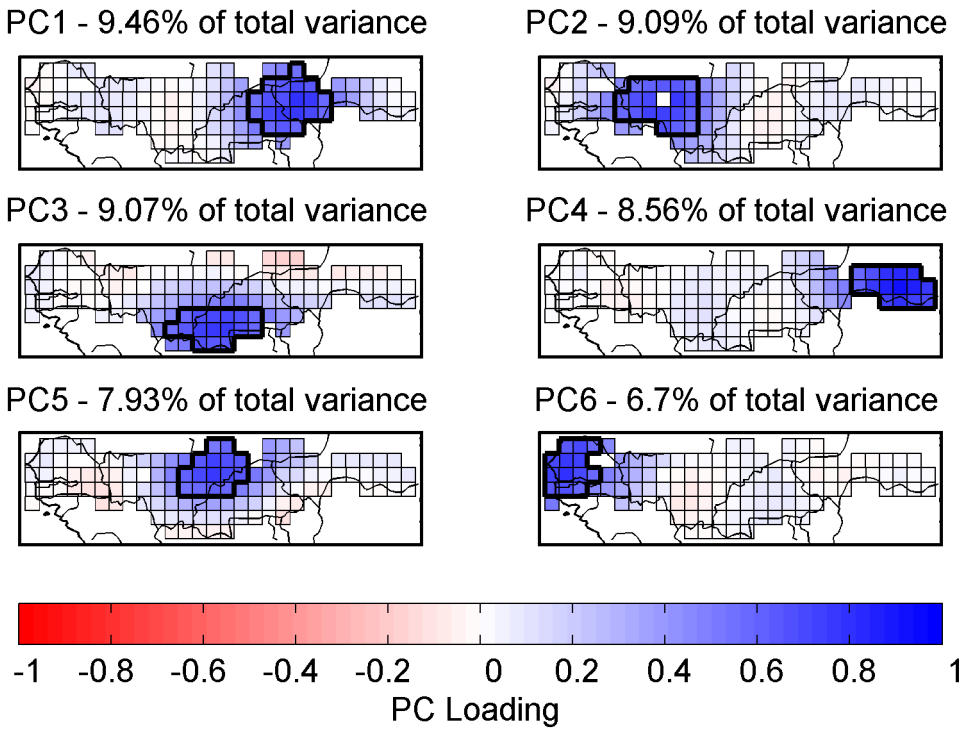

Figure 4.6. Rotated PC loadings from the PCA of daily gridded Sahel rainfall data. Six components were rotated using the varimax criterion and normalisation constraint (4.4). The black regions in each plot indicate areas where loadings are greater than 0.5; used to define the six rainfall regions used in the remainder of the study.

Time series were recalculated for each of the newly defined regions based on the raw data divided by its standard deviation, to prevent bias in the wetter, southern boxes. The PC time series could have been used, but this would have meant sacrificing all the variance in the system not represented by the first six components (almost 50%).

It is important to recognise that, had a different number of components been retained, different regions would have been formed that could have been just as representative as those shown here are. However, these regions conveniently separate out the somewhat dubious eastern and western regions (PCs 4 and 6 respectively), and divide the dependable central regions into north, south, east and west.

4.2.4. Complex extensions of PCA

Some interesting extensions of PCA exist for considering complex data. The term 'complex PCA' (or complex EOF) is used to describe several of them, causing yet more terminological confusion. Some, such as the technique von Storch and Zwiers (1999) refer to as 'Hilbert EOFs', take the complex part of the input variable as a transform of the real part, thus allowing analysis of patterns propagating in space and time. Another, which is here referred to as 'Vector PCA', refer to PCA carried out on vector (as opposed to scalar) variables. Vector PCA is formulated identically to the PCA described above, providing the conjugate transpose operator is used.

In a climatological context, the most obvious use of a vector PCA is in analysing wind data. Given **u**, a set of *p* random zonal wind variables, and **v**, a set of *p* random meridional wind variables, where u_i and v_i relate to observations at some location *i*, we can construct a set of complex random variables **x** using:

$$\mathbf{x} = \mathbf{u} + i\mathbf{v} \tag{4.12}$$

where *i* is the imaginary square root of −1. An example of a vector PCA is given by Klink and Willmott (1989).

Interpretation of vector data is more complex than the usual scalar case, but still feasible. Eigenvalues are still real numbers, but the PC patterns and time series are both complex. The patterns represent the main variability in direction of the wind, the time series the strength of the flow in the direction represented by the patterns.

In a normal scalar PCA pattern, each variable is represented by a scalar coefficient. The sign of each pattern is arbitrary, so the sign of every coefficient can be flipped if desired. So, in effect, each PC has two solutions, one 'positive' and one 'negative'. In a vector PCA, each variable is represented by a vector coefficient. The set of coefficients exhibit a similar arbitrary nature, but instead of flipping signs, each vector may be rotated by the same amount. Hence, there are an infinite number of possible solutions, although a sensible solution would be to maximise the fit between the data and the patterns. Note that this 'rotation' should not be confused with the rotation techniques described in section 4.2.2.

Interpretation of the time series is more confusing. The real part of the series represents the strength of the flow in the direction of the pattern, and the imaginary part represents the strength of the flow at 90° to the pattern. Interpretability is usually increased by converting each number the time series into a magnitude and an angle of the relevant vector. Then, the magnitude represents the strength of the flow, and the angle indicates that direction of the flow relative to the pattern.

Vector PCAs are an elegant extension of the usual method, however their rare use means that they have not been extensively investigated. Most arguments about PCA can be extended to the complex case; there are probably Buell-type patterns for vector PCAs, and the solutions can probably be rotated in the sense of section 4.2.2. However, no research into these areas could be found. Furthermore, the intention of the PCA is to produce a set of predictors for the final model. The relevant variables outputted from a vector PCA, the PC time series, are complex, and would be difficult to use is a regression type analysis. Therefore, vector PCA seems unsuitable for the main aim of this project.

4.2.5. Multiway extensions of PCA

The description of the PCA given in Section 4.2.1, when used in a climatological context, treats each location as a variable. Consequently, the data matrix does not contain information about the geographical position of each site, other than that implicit in the correlation (or covariance) matrix. This form of PCA is essentially a two-dimensional process (considering time and space); however the data usually exists in at least three dimensions (time, latitude and longitude). Extensions of PCA are available that can consider this structure, however at the time of writing no examples relating to climate science could be found, although examples of use can be found in the field of food science (Pravdova et al., 2001; Pravdova et al., 2002).

Recalling Preisendorfer's synthesis formula for PCA given earlier:

$$\mathbf{X} = \mathbf{Z}\mathbf{A}' \tag{4.7}$$

Formula 4.7 allows for the decomposition of the data matrix into two variables, representing time (**Z**, the score matrix) and space (**A**, the loading matrix). Extensions decompose the data into more variables. For example, when considering a data set on a regular grid, the PARAFAC (Parallel Factor) method approximates the observation x_{ijk} using a model of the form:

$$\tilde{x}_{ijl} = \sum_{k=1}^{m} a_{ik} b_{jk} z_{lk} \tag{4.13}$$

where x_{ijl} is the observed value at latitude i, longitude j for observation l, where $i = 1, \ldots, I$; $j = 1, \ldots, J$ and $l = 1, \ldots, n$. As for PCA, m components are extracted and each component k is represented by a time series, made up of the z_{lk}. However, instead of being represented by one loading pattern, the component k is represented by two loading vectors, one (the a_{ik}) relating to latitude, the other (the b_{jk}) relating to longitude.

The PARAFAC model is a special case of the general Tucker3 model, where x_{ijk} is approximated using a model of the form:

$$\tilde{x}_{ijl} = \sum_{l=1}^{w_1} \sum_{m=1}^{w_2} \sum_{n=1}^{w_3} a_{il} b_{jm} z_{kn} g_{lmn} \tag{4.14}$$

Here, instead of extracting k components, we extract a different number of factors for each 'dimension'. So in the example concerning gridded data, we extract w_1 factors relating to latitude, w_2 factors relating to longitude and w_3 factors relating to time. The matrix **G**, made up of the coefficients g_{lmn}, is known as the core matrix, and indicates the importance of each interaction. If the value of g_{lmn} is comparatively high, then the interaction between latitude mode l, longitude mode m and time mode n is comparatively important. Extra dimensions can be added to create a higher order Tucker model.

As can be imagined, interpretation of a PARAFAC or Tucker model is far from simple. Furthermore, their lack of use when compared to PCA means that they are less well understood. For example, do equivalents to Buell patterns exist for these multiway methods? For these reasons, the methods were considered, but eventually

deemed unsuitable for this thesis, but may prove to be a fruitful avenue for future studies.

4.2.6. An Assessment of PCA

A review of the literature has helped to assess the value of PCA. The need to reduce the dimensionality of increasingly higher resolution climatological data has stimulated interest in PCA, resulting in a thorough study of its application and implications. The strength and weaknesses of PCA in the context of climate analysis, therefore, are well known. It is a comparatively simple method to apply, can be solved numerically, and is computationally efficient.

However, a number of important decisions need to be made during the analysis. Should the covariance or correlation matrix be analysed? How many components should be used? Should the solution be rotated, and if so, how? Not all of these decisions are straightforward, and analytical approaches to devise optimal answers have generally failed. Often it is left to the analyst to make arbitrary, if informed, choices.

Furthermore, it must be remembered that PCA as represented here assumes a linear process. Non-linear extensions do exist; see Chapter 14 of (Jolliffe, 2002). However the simplicity of linear PCA, both in terms of complexity of calculation and interpretability, make it a powerful technique suitable for many circumstances, including producing a set of predictors for an empirical model of Sahel climate.

4.3. The selection of a domain for the PCA – Analysis of the NCEP reanalysis fields

Before carrying out a PCA, it was necessary to identify a suitable domain for the analysis. The NCEP dataset spans the whole globe at a $2.5\,^{\circ} \times 2.5\,^{\circ}$ resolution for multiple levels. Hence, the number of possible variables is far too great for the purposes of this study. Furthermore, many variables are likely to be irrelevant. For example, data points a great distance from the Sahel are unlikely to affect the rainfall

there. This section focuses on the identification of a suitable domain for the PCA. Part of this identification was performed using a priori knowledge, part by examining correlation between NCEP variables and Sahel rainfall.

Six variables from the NCEP dataset were to be analysed: geopotential height (representing air pressure, and abbreviated by gph), air temperature (air), specific humidity (shum), vertical velocity (omega), zonal wind (uwind) and meridional wind (vwind). Each variable was obtained for four atmospheric levels: 1000, 850, 600 and 200 hPa, with the exception of specific humidity, which was not available for the top level, so the 300 hPa level was used instead. These cover the surface level, the top of the boundary layer, and the mid and upper troposphere. Furthermore, the two upper levels contain the actions of the major jet streams, the African Easterly Jet and the Tropical Easterly Jet, which are suspected to have a significant influence on Sahelian rainfall (see Chapter 2).

Monthly data were obtained for each grid point between 50 °N and 50 °S; higher latitude regions were omitted, as they are very unlikely to affect Sahelian rainfall. Correlations were calculated between each grid point at each level for each variable and all six of the rainfall regions created in section 4.2.3.

An analysis of the results demonstrated that the seasonal cycle dominated most of the correlation patterns. Therefore all data series were deseasoned (by subtracting long-term monthly means) and correlations recalculated. Figures 4.7 to 4.12 show the result, based on 1958-1997, and only considering the Sahelian wet season (here defined as June to September). Only the strongest correlation with any of the six rainfall regions is displayed. The figures illustrate that the strongest correlations, regardless of which variable is considered, occur either over West Africa, or just offshore. Whilst absolute correlations over 0.3 (the minimum required to appear in the figures) do exist in other areas, particularly over the oceans, they are weaker.

The results suggest that rainfall is associated with cooler conditions over the south of the Sahel and the Guinea coast in the lower troposphere, overlaid by warmer conditions and ascending air in the mid troposphere. It is also associated with low pressures in the lower and middle troposphere just off the West African coast at about 20 °N. Just south of this low pressure pattern are a pattern of correlation with

westerlies which extend across to the Horn of Africa, overlaid by easterlies at the top of the troposphere. Southerly, low level winds across the Sahel, extending northeast toward the Middle East, are also associated with increased rainfall. Finally, and perhaps least surprisingly, is the association with high humidity across the Sahel at all levels, again extending toward the Middle East. Notable, however, is the lack of correlation with moisture in the Gulf of Guinea, indeed rain is correlated with dry conditions at lower levels. This is reminiscent of the findings of Fontaine et al. (2003) and Long et al. (2000) noted in Chapter 2, which suggest most moisture which falls as rain in the Sahel comes from the east.

Most of the variables show coherent spatial patterns of correlations. The exception is the plot for vertical velocity, in which patterns are very patchy. This might suggest the field would be of little use, although one of the few notable coherent patches of ascent occurs over the Sahel at the 600 hPa level.

Ideally, the final domain would be large enough to consider all areas that are thought to be connected to Sahelian climate. However, various considerations impose limits on possible domain size. Recall the suggestion that sample size should be larger than the number of variables in a PCA. As the analysis was to be carried out for 1958 to 1997, there are a total of 14610 days. However, ideally this analysis should only take place in the wet season, as this is the period of interest. Furthermore, the stability of the PCs was to be investigated by performing two separate PCAs on the two halves of the time period: 1958 to 1977, and 1978 to 1997, halving the acceptable number of variables. If the duration of the wet season were to be considered as June to September, this would set an upper maximum of 2435. Note this is only an advisory maximum; a much lower number would be preferable as the sampling error in the PCs would be less, and hence the results would probably be more stable. Furthermore, lowering the number of variables substantially decreases the time taken to compute results.

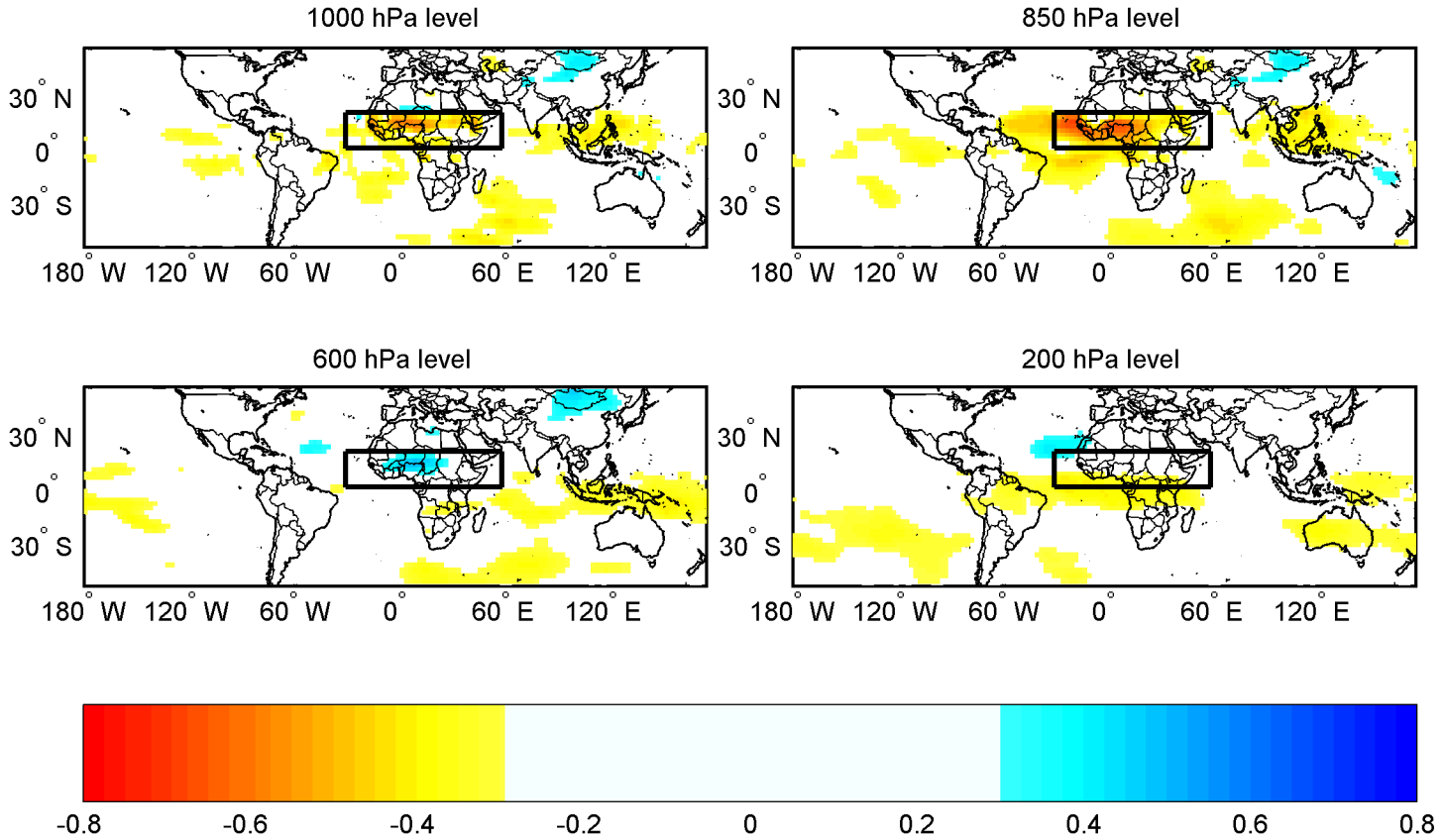

Figure 4.7. Correlation coefficient between monthly anomalies in air temperature and Sahel rainfall, June to September, 1958-1997. Correlations were calculated for all six rainfall regions, only the largest absolute correlation is displayed. The black box indicates the domain used for the PCA described in section 4.4

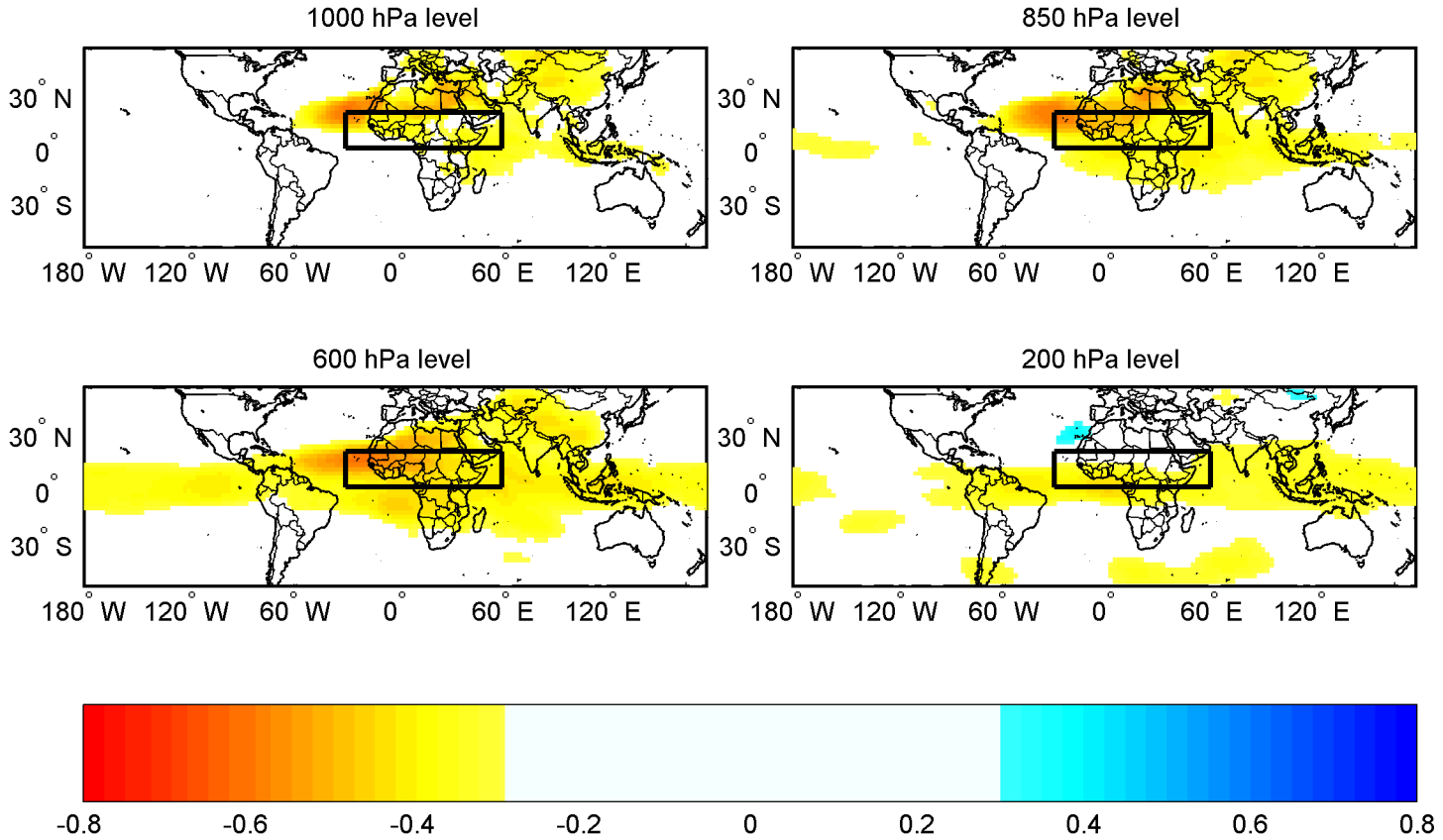

Figure 4.8. Correlation coefficient between monthly anomalies in geopotential height and Sahel rainfall, June to September, 1958-1997. Correlations were calculated for all six rainfall regions, only the largest absolute correlation is displayed. The black box indicates the domain used for the PCA described in section 4.4

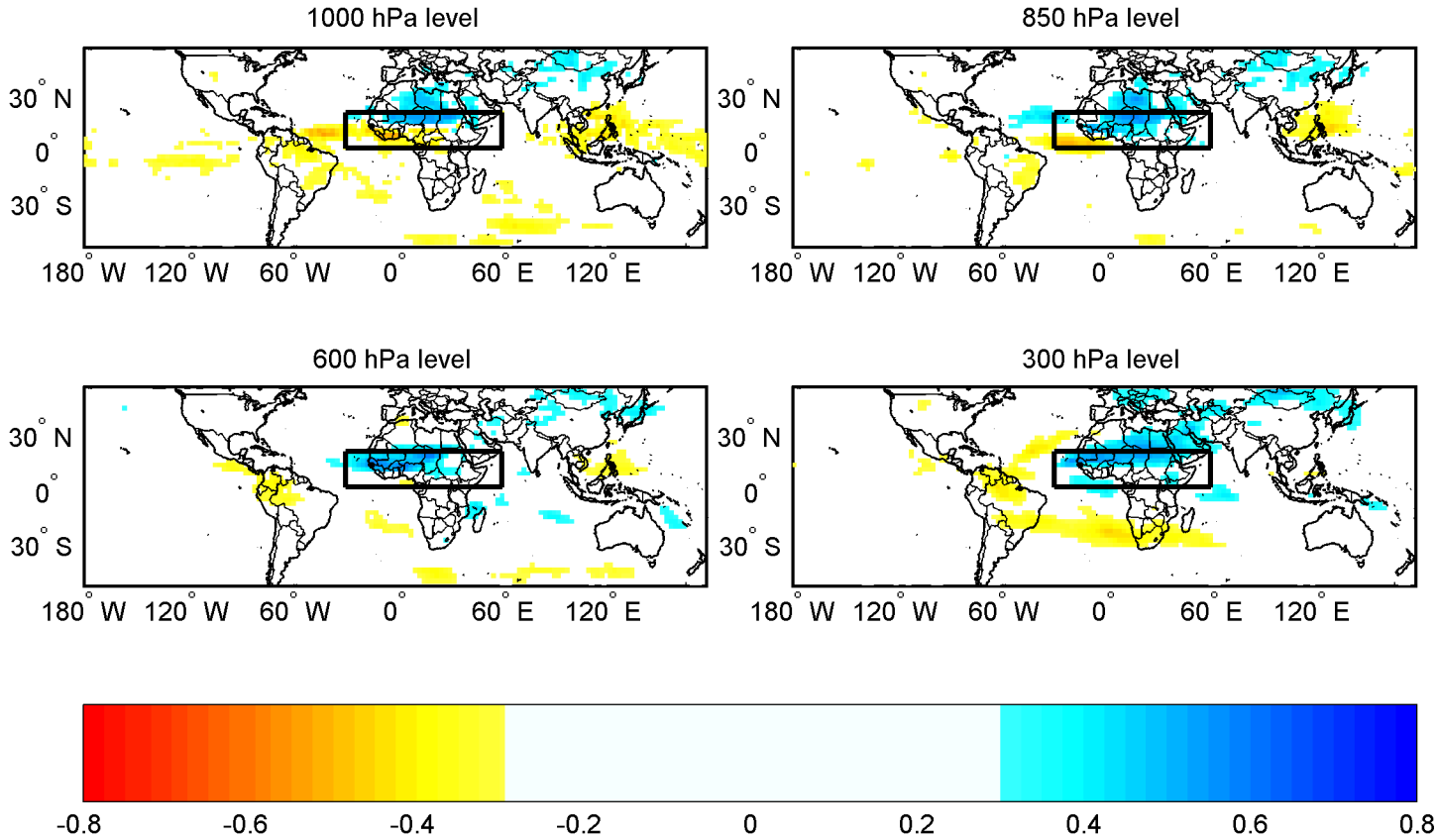

Figure 4.9. Correlation coefficient between monthly anomalies in specific humidity and Sahel rainfall, June to September, 1958-1997. Correlations were calculated for all six rainfall regions, only the largest absolute correlation is displayed. The black box indicates the domain used for the PCA described in section 4.4

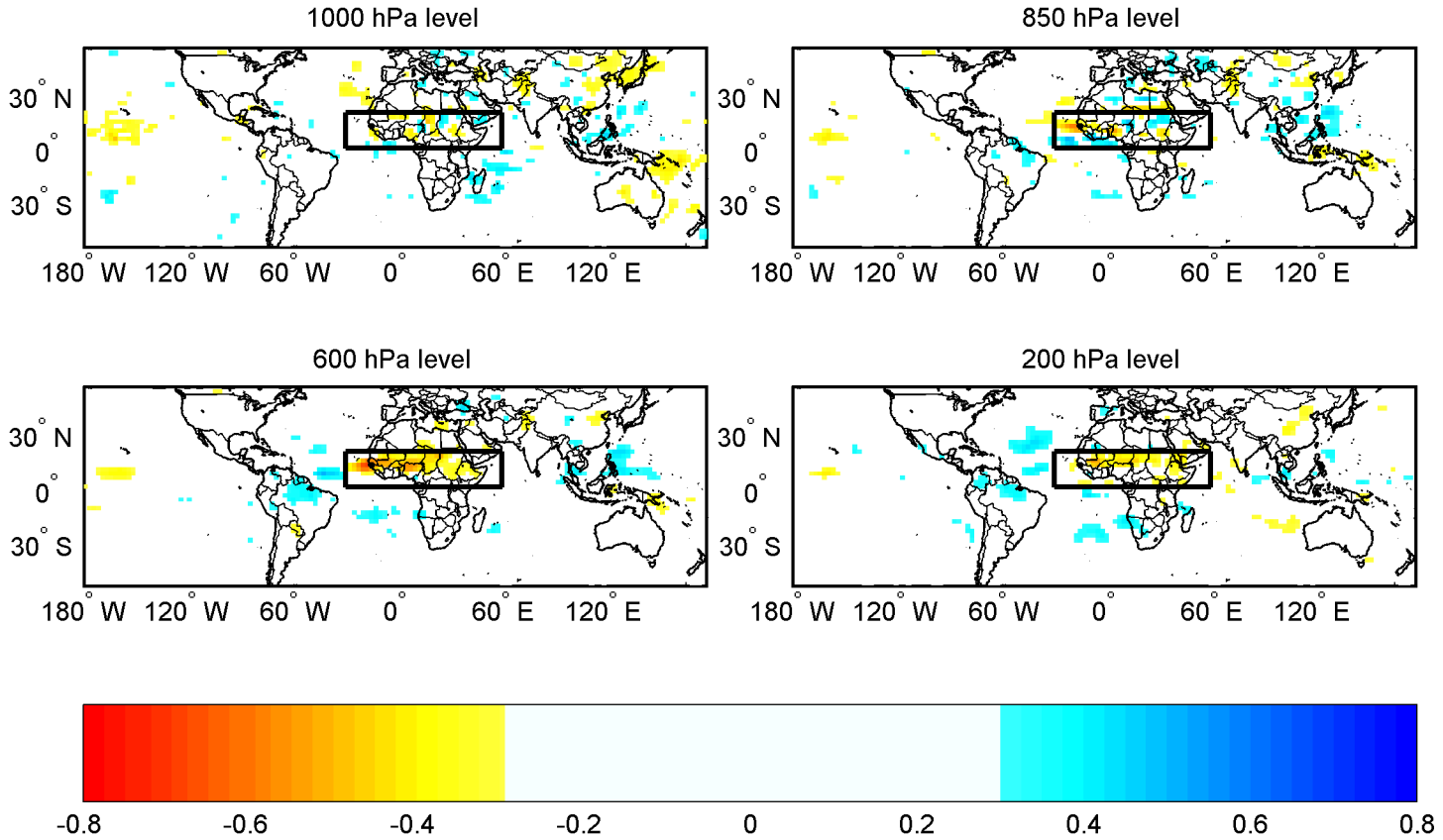

Figure 4.10. Correlation coefficient between monthly anomalies in vertical velocity and Sahel rainfall, June to September, 1958-1997. Correlations were calculated for all six rainfall regions, only the largest absolute correlation is displayed. Positive correlations indicate a link between greater rainfall and descending motion. The black box indicates the domain used for the PCA described in section 4.4

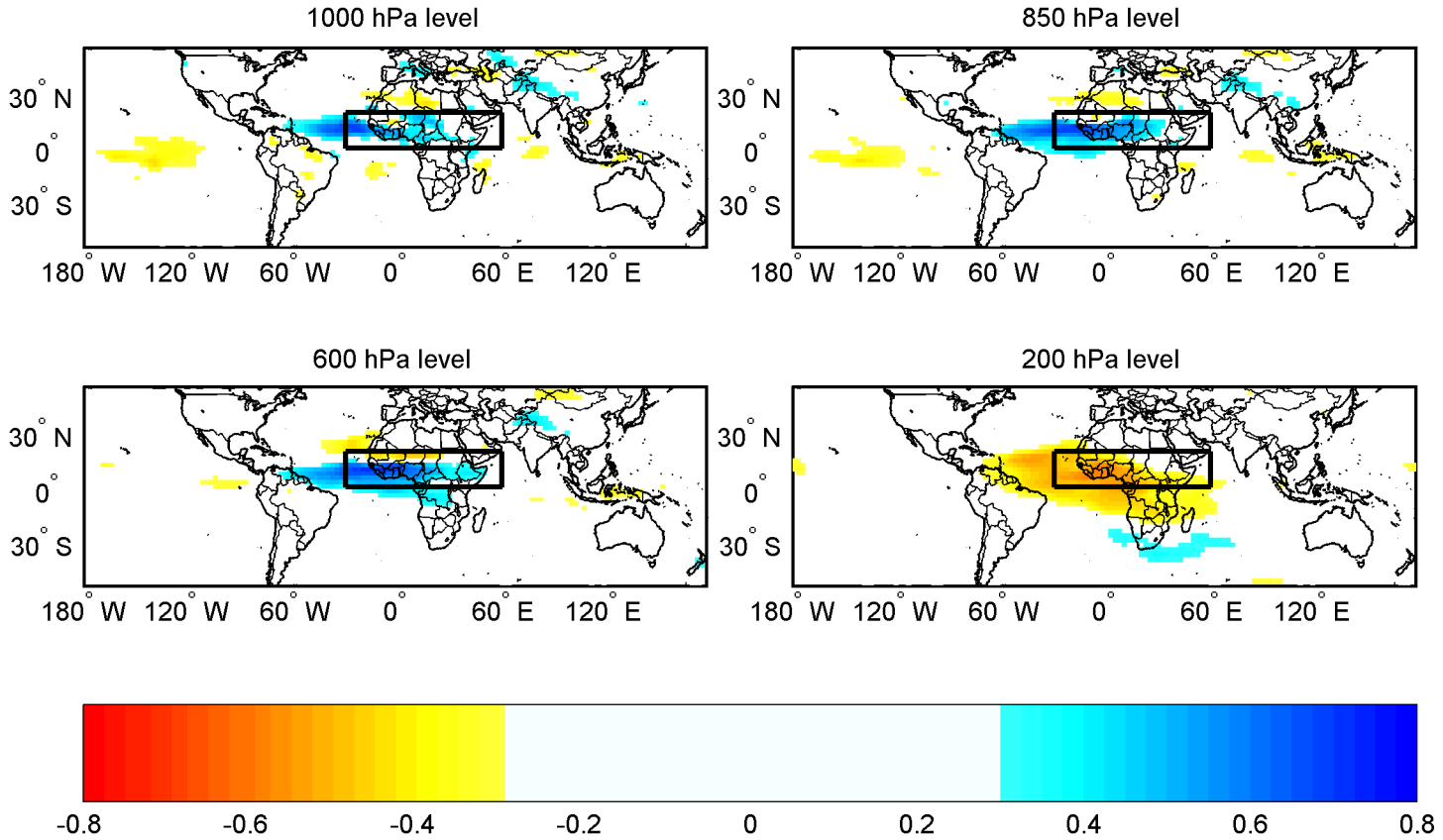

Figure 4.11. Correlation coefficient between monthly anomalies in zonal wind and Sahel rainfall, June to September, 1958-1997. Correlations were calculated for all six rainfall regions, only the largest absolute correlation is displayed. Positive correlations indicate a link between greater rainfall and westerly motion. The black box indicates the domain used for the PCA described in section 4.4

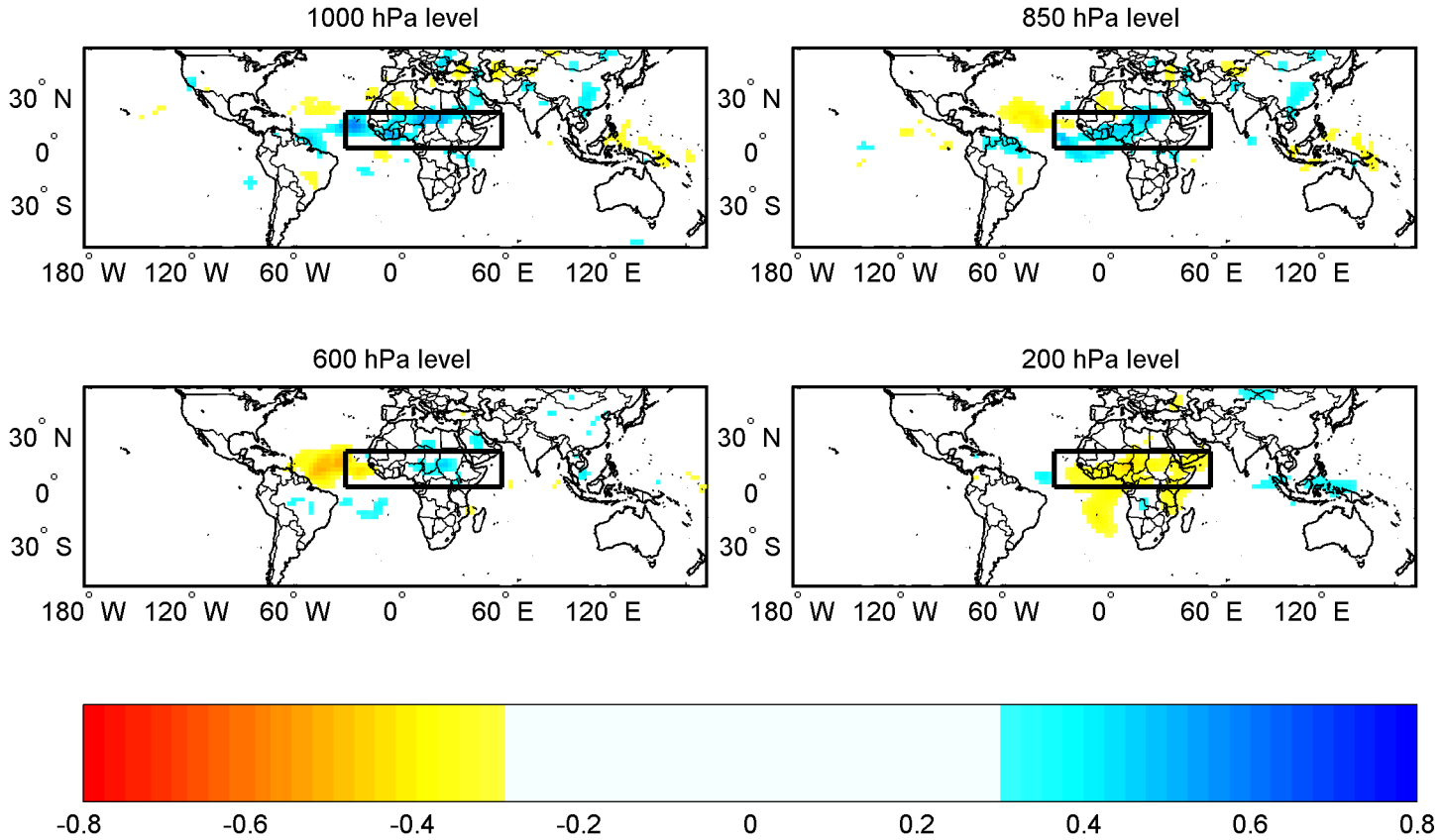

Figure 4.12. Correlation coefficient between monthly anomalies in meridional wind and Sahel rainfall, June to September, 1958-1997. Correlations were calculated for all six rainfall regions, only the largest absolute correlation is displayed. Positive correlations indicate a link between greater rainfall and southerly motion. The black box indicates the domain used for the PCA described in section 4.4

A final domain size was set at 0 – 20 °N, 30 °W – 60 °E. This encloses most of the areas of importance identified by the correlation analyses (see Figure 4.7 to 4.12), but contains only 333 grid points per level, giving a grand total of 1332 grid points. Furthermore, it was decided to extend the period analysed to May to October, to further increase the case to variables ratio; rainfall still occurs frequently in this period. This decision proved to have very little effect on results, indeed a PCA carried out on the whole year (not shown here) gave very similar patterns.

4.4. Execution and results of the Principal Component Analysis

The PCA was executed by performing an analysis on each field in turn. Theoretically, it would be possible to enter all six fields into one analysis, an EEOF, but this would make interpretation of results far more difficult. Each PC pattern would have to be represented by 24 maps (each of the six variables at four levels), and would result in the PC time series being a composite of six physical quantities. Furthermore, the number of variables would be increased by a factor of six, which could have a substantial impact on the stability of the solution.

Therefore, it was decided to analyse the different levels of each field in one PCA. It is common practice to 'unfold' a two-dimensional field prior to PCA and consider each grid point as a variable (see Pravdova et al., 2001, for a non-climatological example). It seems reasonable to extend this practice to three-dimensions. As a result, loading patterns are three-dimensional, and hence harder to interpret, but it does permit the analysis of variability between levels.

The analysis was planned to be carried out using MATLAB, but the large size of the data set prevented this: MATLAB had insufficient virtual memory. Hence, the correlation matrix for the data set was calculated in MATLAB and exported into a Fortran routine, where the PCA was carried out. The results were interpreted into MATLAB for post-analysis processing, such as the calculation of PC time series and examination of eigenvalues.

Table 4.3 shows the first ten eigenvalues for each of the six PCAs, and the percentage of total variance they each explain. It also indicates the number of

eigenvalues greater than one: Kaiser's criterion. As can be seen, Kaiser's criterion would retain an exceptionally large number of components for all six PCs, far too many to use in an interpretable regression analysis. Hence, an arbitrary decision was made to retain all components that explained at least 2% of total variance. The bottom two rows of Table 4.3 show the number of components retained for each PCA, and what proportion of total variance they explain collectively. Across the six PCAs, a total of 37 patterns were retained.

Eig. No.	air		gph		shum		omega		uwind		vwind	
	λ	%	λ	%	λ	%	λ	%	λ	%	λ	%
1	393.2	29.5	588.9	44.2	263.7	19.8	121.4	9.1	455.7	34.2	168.8	12.7
2	200.4	15.0	230.5	17.3	99.1	7.4	59.4	4.5	110.7	8.3	104.7	7.9
3	98.4	7.4	205.9	15.5	62.3	4.7	52.7	4.0	70.8	5.3	65.8	4.9
4	57.8	4.3	56.6	4.3	42.4	3.2	29.2	2.2	41.8	3.1	44.7	3.4
5	31.5	2.4	34.5	2.6	33.0	2.5	24.1	1.8	32.2	2.4	44.3	3.3
6	29.4	2.2	24.2	1.8	29.0	2.2	21.3	1.6	30.2	2.3	41.6	3.1
7	26.4	1.9	20.5	1.5	25.5	1.9	19.9	1.5	25.2	1.9	36.4	2.7
8	25.7	1.9	17.9	1.3	22.9	1.7	18.1	1.4	23.4	1.8	34.8	2.6
9	23.2	1.7	12.9	1.0	18.8	1.4	17.7	1.3	22.6	1.7	32.1	2.4
10	22.7	1.7	11.2	0.8	17.8	1.3	17.1	1.3	21.5	1.6	27.8	2.1
11	19.9	1.5	9.6	0.7	16.6	1.2	15.5	1.2	18.7	1.4	26.1	1.9
12	16.6	1.2	8.0	0.6	15.8	1.2	14.4	1.1	18.1	1.4	22.5	1.7
λ>1	98		39		175		243		111		145	
Components extracted	6		5		6		4		6		10	
% of total var. explained	60.9		83.8		39.7		19.7		55.7		45.1	

Table 4.3. Details of eigenvalues and percentage of total percentage of variance explained for each of the six PCAs. The bottom three rows indicate the number of eigenvalues greater than or equal to one, the number of components eventually extracted, and the percentage of total variance they explained collectively.

The thirty-seven component loading patterns are illustrated in Appendix A, figures A1-6. The original intention was to use unrotated PCs as the predictor variables in the final model, but the outputted loading patterns suggest this would be unwise for a number of reasons. First, many of the patterns bear little physical sense. For example, later meridional wind components (referred to as 'vwind') tend to involve multiple vertical strips of northerly winds opposed to strips of southerly winds.

Furthermore, although correlation structures are far from simple, Buell-type patterns seem to occur. This is probably most striking in the air temperature ('air') PCA, see

Figure A.1. The first component is an almost all-global loading pattern, the second counters northern regions with southern (although with a bias at the 200 hPa level), and the third is a high-level vs. low-level pattern. These are the type of patterns one might expect to see in a three-dimensional extension of Buell's analyses.

The primary benefit of not rotating the PCA was that time series would not be correlated, and hence much easier to use in a regression model. Unfortunately, the analysis of each field individually means that components from different fields can be highly correlated. For example, the leading component for air temperature is highly correlated with the second geopotential height component ('air1' and 'gph2' respectively); the coefficient of the correlation between the two corresponding time series being 0.922. Therefore, the cost of introducing correlation between time series from the same field is less problematic, as these strong across field correlations should be reduced.

As a result, a decision was made to rotate the extracted components, using the varimax criterion. The rotated patterns are also illustrated in Appendix A, figures A7-A43. The time series for these patterns were calculated for the whole year, and are presented in the form of a yearly average and a 'daily average': the average of the PC time series on each calendar day. This daily average indicates the typical seasonal cycle of the given component. Note that each PC time series was normalised to give it a standard deviation of one. This was done to ensure that when used in the final model, the regression coefficient allocated to each component would be directly comparable. However, at this stage it also allows for easier comparison between the series.

Table 4.4 contains a summary of the main physical processes involved in each component:

PC name	**Description of the main processes involved**
air1	High temperatures in low atmosphere, expect for low temperatures over Atlantic at surface
air2	High temperature at all levels in atmosphere over equator and Horn of Africa.
air3	High temperatures at top of atmosphere north of 5°N

PC name	Description of the main processes involved
air4	High temperatures at 600 hPa between 0-30°E
air5	High temperatures at 600 hPa, west of 0°
air6	High temperatures at 600 hPa over Somalia versus Low temperatures in lower atmosphere over Middle East Gulf.
gph1	High pressures in low to mid atmosphere over Sahel and Atlantic
gph2	High pressures in upper atmosphere over whole region
gph3	High pressure in low atmosphere over East Africa & Indian Ocean vs low pressure over equator at top of atmosphere
gph4	High pressures over Sudan/Ethiopia/Middle East in low atmosphere
gph5	High pressures over equatorial Gulf of Guinea just off African coast in low to mid atmosphere
shum1	High humidity over equatorial Africa and Gulf of Guinea, especially in lower levels
shum2	High humidity over Indian Ocean and Ethiopia at all levels, and to a lesser extent the Gulf of Guinea
shum3	High humidity over Atlantic (0-10°N), especially in lower levels
shum4	High humidity over North Atlantic and North Africa, especially in lower levels
shum5	High humidity over North East Africa and Middle East at all levels
shum6	High humidity over inland sub-Saharan North Africa at all levels
omega1	Noisy at surface, strong ascent in lower atmosphere over and off coasts of Ethiopia and Somalia, descent in middle atmosphere to south of this region, descent in upper atmosphere to the north of this region.
omega2	Noisy, but ascent over Guinea coast and descent over Sahel at 850 Pa. Descent pattern moves southward as height increases, by 200 hPa is over Guinea coast.
omega3	Noisy, but descent centred over southern Sudan in upper atmosphere.
omega4	Noisy, but broad descent over equatorial West Africa at all levels (Cameroon, Gabon and Republic of Congo)
uwind1	Westerlies at top of atmosphere and over equatorial Atlantic to middle atmosphere, Easterlies elsewhere. Cyclonic flow over the Sahel / Guinea Coast at the 600 hPa level.

PC name	Description of the main processes involved
uwind2	Westerlies centred over Sahel in lower & middle atmosphere. Extend over Atlantic. Strongest at 600 hPa.
uwind3	Westerlies in lower atmosphere centred over Nigeria, extend from Ghana to Chad. Centred around 10°N, do not extend more than 5° north or south.
uwind4	Westerlies over Gulf of Guinea and Cameroon in lower atmosphere. Easterlies lie to the north of this band at 600 hPa.
uwind5	Westerlies centred over Northern Sudan in lower and middle atmosphere.
uwind6	Westerlies at 600hPa centred at southern tip of Sudan. Strong pattern extends to coast of east Africa, but loadings weaker over Atlantic coast.
vwind1	Southerlies in lower atmosphere over Indian Ocean, north Somalia and Middle East. Noisy lesser patterns of Southerlies elsewhere in lower atmosphere (Atlantic Coast, southern Sudan), and lesser pattern of Northerlies from surface to mid atmosphere over Ethiopia, south Somalia and Kenya.
vwind2	Southerlies in lower atmosphere centred. over Chad and Sudan. Northerlies overlaying at 600 hPa, but with centre just to Southwest (over northern Nigeria/ northern Cameroon / southern Chad borders) and Northerlies at surface over Atlantic north of 10°N.
vwind3	'Backslash' type slanted pattern of Southerlies in the lower atmosphere, extending from West Coast equatorial Africa to Democratic Republic of Congo, extends northwards along coast. Leaves coast past Gulf of Guinea and at northernmost extent it extends from eastern Mali to Chad.
vwind4	Vertically slanted pattern of Southerlies in lower and middle atmosphere. Centred over Uganda and Kenya at surface but the pattern moves northwards as height increases. By 600 hPa centred over Southern Sudan.
vwind5	'Dipole backslash' structure in lower and middle atmosphere, especially 850 hPa. Southerlies along West African coast opposed by Northerlies inland (from Mali to Ghana). Cyclonic flow over the West Coast.

PC name	Description of the main processes involved
vwind6	'Dipole backslash' structure very similar to vwind5 but more westerly. Lower & middle atmosphere, again especially 850 hPa, Northerlies along West African coast opposed by Southerlies over Atlantic. Cyclonic flow off the West Coast.
vwind7	Top of atmosphere double dipole. Northerlies centred over Cameroon / Central African Republic border 'sandwiched' by two patterns of Southerlies centred over Gulf of Guinea and Horn of Africa.
vwind8	Top of atmosphere dipole. Southerlies over Atlantic opposed by Northerlies centred over Guinea Coast, but extending from Gulf of Guinea up to eastern Mali. High level cyclonic flow over West Africa.
vwind9	Southerlies in lower and middle atmosphere. Over Horn of Africa at surface, by 600 hPa in centred over Gulf of Aden.
vwind10	Northerlies at top of atmosphere centred over southern Sudan / Uganda. Flanked to east and west by weaker patterns of Southerlies.

Table 4.4. Main physical processes involved in each of the principal component patterns.

Many of the components are interrelated. For example, the three geopotential height components 'gph1', 'gph4' and 'gph5' are all correlated with coefficients exceeding 0.9. This is an unsurprising result, as each component represents low to mid-level pressure in a different part of the domain. Furthermore, it is noticeable that the daily and yearly average time series are very similar for all three components.

Other sets of components are highly correlated, but are slightly harder to explain. The correlation coefficient between the 'shum1' and 'uwind2' time series in the wet season is 0.90. The corresponding patterns show an increased low and mid-level westerly flow centred over the Sahel is associated with increased humidity over the Gulf of Guinea and the neighbouring coast. These may be related through the monsoon flow, but the westerlies seem located slightly too far to the north. Similarly, the correlation between 'shum3' and 'omega2' is –0.94. So low to mid-level ascent over the Sahel at the surface, but to the south at higher levels, is associated with increased humidity over the equatorial Atlantic and the Guinea Coast.

Finally, a set of four components are all cross-correlated with absolute correlations exceeding 0.9: with positive values of 'omega1' and 'uwind1' associated with negative values of 'vwind1' and 'shum2'. Together, the patterns suggest low-level meridional convergence over East Africa, with accompanying ascending air. Relative humidity in the area is decreased, perhaps as a result of rainfall; recall that results are based on daily averages, so at a zero lag, rainfall can affect the variables. Finally, low-level westerlies are overlaid by high-level easterlies.

4.5. A validation of the PCA

As mentioned in section 4.1, doubts have been cast over the quality of NCEP data over the Sahel from before 1968. Unfortunately, the PCA was carried out without knowledge of these doubts, so used data stretching back to 1958. Further analyses would be based only on data from after 1968, but recalculating the PCs would be very time consuming. If possible, the ideal solution was to use the PCs defined over the period 1958-1997, but only use the PC time series from 1968 onwards in the final empirical model. This should not pose a problem, providing the modes of variability are similar throughout the full period.

In order to check whether the full period PCAs represented variability after 1968 well, two additional PCAs were carried out for each variable. The first only considered data from the period 1958-1977; the second considered the period 1978-1997. Providing the outputted patterns were similar, the original PCs could be used in the final empirical model. A secondary benefit of this analysis would be to test the stability of the PCA solutions. Vastly different patterns would either be a result of different processes occurring in the two periods, or instability of the PCs. Either way, the usefulness of the results would be called into question.

Even in an ideal situation, one would expect minor differences in the leading patterns. Furthermore, PCAs insistence on orthogonality would lead to greater differences in the second patterns, and greater still in later patterns. Hence, only the first four patterns are illustrated. Also, results were not rotated. Patterns that were similar prior to rotation would remain similar after rotation. Results are shown in appendix A, figures A.48 to A.67.

The results suggest the main modes of spatial variability of the atmosphere are similar for the two periods. Patterns for air temperature, geopotential height and specific humidity are nearly identical, although the order of PC3 and PC4 seemed to have swapped for specific humidity. Results for the three velocity variables are less definitive. In all three cases, the first two PCs show a good, although far from exact match. As a result, differences in the third and fourth PCs are considerable. However, the broad structure often remains the same. For example, the zonal wind analyses have the same areas of major variability: positive loadings in the low to mid atmosphere over the Gulf of Guinea and central Africa for PC3, negative loadings in the low to mid atmosphere of Sudan and the Middle East for PC4. For meridional wind, PC4 for the first period looks similar to PC3 for the first period. Vertical velocity is the most problematic, as its noisy nature make it hard to interpret patterns visually.

The similarities between the two periods suggest that whilst the principal components defined over the period 1958-1997 may not be the optimal representation of atmospheric variability after 1968, they do represent it well. Therefore, at the end of this chapter we are left with 37 variables with which we will predict daily rainfall in the final empirical model.