Chapter 5: Creating the Empirical Model

5.1. Linking atmospheric variability with rainfall

The previous chapters of this thesis have contained a description of the creation of a set of variables for use in an empirical model. Chapter 3 (and section 4.2.3) detailed the production of six rainfall regions to be predicted, and Chapter 4 described the creation of thirty-seven three-dimensional atmospheric predictor variables. This chapter will recount the search for a suitable form of model to link the two sets of variables, and will describe the creation, execution and interpretation of that model.

One of the simplest, yet most powerful, methods of statistical analysis is linear regression; for a full description, see for example von Storch and Zwiers (1999, p150-168). Simple linear regression relates a response variable to a single predictor variable by calculating the 'line of best fit' between the two variables. The best fit is that which minimises the sum of squared errors. An extension, known as multiple linear regression (MLR), allows for multiple predictor variables.

Briefly, suppose we are given a variable **Y** that we wish to model. This variable has been observed a total of n times. We can consider each observation, y_i, as the realisation of the corresponding random variable, $\mathbf{Y}_i$. We also have a set of p predictor variables, $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_p$, with the ith observation of the jth predictor variable being x_{ij}. The model takes the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad (5.1)$$

where β_0, β_1, …, β_p, are coefficients estimated by least-squares (see the above reference), and ε_i is the modelling error, known as the residual error, in the ith observation. We can consider that ε_i is the realisation of a random variable $\mathbf{E}_i$. The model demands that each $\mathbf{E}_i$ (for i = 1, … p) is normally distributed with mean zero and with a common variance σ^2. Furthermore, each $\mathbf{E}_i$ must be independent.

Determining the βs involves inverting the matrix $\mathbf{X}^{\mathrm{T}}\ \mathbf{X}$, where **X**, known as the design matrix, is defined as:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{5.2}$$

Note that the first column of **X** is represents a dummy variable equal to one for all observations, and produces the intercept coefficient β_0. In the case where no intercept is required, this column in omitted.

If two of the predictor variables, and hence two columns of **X**, are highly correlated, then $\mathbf{X}^T\ \mathbf{X}$ is nearly uninvertible. As a result, the outputted coefficients are greatly affected by small changes in the measurements of the x_{ij}, and are therefore highly unstable. This problem, often referred to as multicollinearity, leads to large uncertainty in the estimation of the βs.

The problem is usually solved by omitting one or more of the predictor variables. However, an alternative approach is to use ridge regression (von Storch and Zwiers, 1999, p165-166). Ridge regression adds restraints to the model, which reduces the influence of the off-diagonal elements of $\mathbf{X}^T\ \mathbf{X}$, increasing its invertability. Thus the parameters can be estimated with less uncertainty, but at the cost of unbiased estimation.

Unfortunately, as noted above, the linear regression family of analyses all demand that the residual errors are independent, normally distributed and share a common variance. This makes the procedure unsuitable for variables with highly skewed distributions, such as daily rainfall. However, MLR is a special case of a modelling strategy known as the Generalised Linear Model (GLM), which allows for a wider range of distributions.

5.2. Generalised Linear Models (GLMs)

5.2.1. The formulation of GLMs

Section 5.1 introduced the idea of attempting to model $\mathbf{Y}_i$, the distribution of some response variable for an *i*th observation, based on a set of *p* predictor variables, using a multiple linear regression model. The Generalised Linear Model is based on assuming that all the $\mathbf{Y}_i$ come from some common distribution type from the exponential family. For example, we could make the assumption that each distribution is binomial. Many other choices exist, the most common including the Normal, Gamma or Poisson distributions.

Let μ_i represent the expected value of $\mathbf{Y}_i$ given the predictor variable values x_{i1}, x_{i2}, …, x_{ip}. Given these predictors, we assume each $\mathbf{Y}_i$ is independent. Sometimes we must assume they share a common 'dispersion' parameter, φ, which is typically related to the shape of the distribution. The GLM model then takes the form:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \quad (= \eta_i) \tag{5.3}$$

where $g(\cdot)$ is a monotonic, differentiable function known as the link function. The right hand side of the equation is known as the linear predictor, and, as shown, is often represented by the symbol η_i, (McCullagh and Nelder, 1989).

The MLR model of equation (5.1) is a special case of (5.3), where $\mathbf{Y}_i$ is normally distributed with mean μ_i and some variance σ, which in this case is the dispersion parameter φ that we must assume is constant for all observations. The link function g is the identity function (Fahrmeir and Tutz, 2001).

Note that models of form (5.3) do not try to predict a value of $\mathbf{Y}_i$; rather they attempt to model the distribution of $\mathbf{Y}_i$. The expected value of $\mathbf{Y}_i$, μ_i, is often used as a predicted value, but it also allows us to make probabilistic models. For example, if we let $y_i = 0$ if day *i* is dry, and 1 if it is wet, then we can model $\mathbf{Y}_i$ with a binomial distribution, and create a model that predicts the chance of rain given a set of predictor variables.

GLMs can be solved, like MLR, by mean likelihood estimation. However, for GLMs the likelihood equations produced are nonlinear, and have to be solved iteratively, see Fahrmeir and Tutz, (2001, p42) for details and possible solution schemes.

The choice of link function is left to the analyst, provided it is monotonic and differentiable, as noted above. However, each distribution is associated with a particular function, known as the canonical or natural link, chosen for mathematical elegance. Sometimes it is useful to use an alternative function. For example, when modelling strictly positive distributions, it is common to choose a link that forces all possible values of **Y** to be positive also, see section 5.3 for an example.

Fuller introductions to GLMs are provided by Dobson (2002) and McCullagh and Nelder (1989), which give a description of the form of exponential distributions allowed. Furthermore, for each distribution, they describe the nature of the dispersion parameter (which must be assumed constant), and describe the function providing the canonical link.

GLMs have not been used extensively in the field of climatology. However, the examples that exist demonstrate their power in modelling non-normally distributed variables, such as daily maximum wind speed (Yan et al., 2002) and daily rainfall (Coe and Stern, 1982; Stern and Coe, 1984; Chandler and Wheater, 2002).

5.2.2 Generalised Linear Model Output

The basic output of a GLM is the set of parameters estimated, namely the fitted model coefficients, β_0, β_1, …, β_p, and, where applicable, the estimated dispersion parameter, φ. However, as with a linear regression, other output is available which measures the quality and validity of the fitted model. Firstly, statistics measuring the goodness of fit are considered.

Most software packages output the deviance of the fitted model as an indication of goodness of fit. Deviance is a measure of the discrepancy between the model in question and the 'full model': the 'perfect' model in which each predicted mean is

equal to the observation in question (see McCullagh and Nelder, 1989, p33 for further details). The deviance, *D*, is defined as:

$$D = 2\hat{\varphi}\left(\ln L_F - \ln L\right) \tag{5.4}$$

where L_F is the likelihood of the full model, and L is the likelihood of the model in question. Deviance formulas for each of the major distributions can be calculated implicitly; see McCullagh and Nelder, (1989, p34) for a full list.

For GLMs based on the Normal distribution, the deviance is equal to the residual sum of squares. In other GLMs, it serves a similar purpose. Indeed, analysis of variance procedures carried out in linear regression can be extended to analysis of deviance for GLMs (McCullagh and Nelder, 1989, p35).

Deviance is a particularly useful statistic when comparing two 'nested' models, where the predictors in one model are a subset of the predictors in a second. If the larger model has *p* factors, and the smaller has *q*, then the difference in deviance between the two models can be modelled approximately with the χ^2 distribution with (*p*-*q*) degrees of freedom (see Dobson, 2002, p80-81).

Another available measure of fit is the generalised Pearson X^2 statistic, usually simply called the Pearson statistic, defined by McCullagh and Nelder (1989, p34) as:

$$X^2 = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\mu}_i\right)^2}{\mathrm{var}\left(\hat{\mu}_i\right)} \tag{5.5}$$

where var (μ) is the variance function of the assumed distribution which, by the definition of an exponential distribution, is a function of the mean. As one might expect by the name, this statistic is an extension of the original Pearson X^2 statistic, and indeed is equal to it for GLMs based on binomial or Poisson distributions. For GLMs based on the Normal distribution, X^2 is equal to the residual sum of squares, and hence also equal to the deviance.

Unfortunately, the deviance and Pearson statistics are not directly interpretable metrics. Chandler and Wheater (1998) define a measure "analogous to the R^2 of

standard regression, which measures the proportion of variability explained by the model". This measure is defined as:

$$\hat{R}^2 = 1 - \frac{\text{Mean square prediction error}}{\text{Variance of original observations}} \tag{5.6}$$

As pointed out in the cited reference, the numerator in the definition has to be altered from the variance of residual errors in the original definition of R^2, because in a GLM the mean of the residuals need not be zero. This altered R^2 is easily interpretable, but can be misleading, as unlike for linear regression models, it is not maximised for GLMs.

Another useful output of regression models is the residual errors, which allow for the identification of outliers, and validation of the fitted model. For example, if a fitted linear regression model is valid, then the residuals will be normally distributed. Any systematic error in the residuals indicates that the assumptions made are invalid.

Several different forms of residuals are output for GLMs, and are introduced here. All produce one residual for each of the *n* observations.

Raw Residuals:

$$r_i^R = y_i - \mu_i \tag{5.7}$$

The raw residual is the standard residual defined in linear regressions, and probably the least useful of the residuals, as it takes no account of the modelled distribution.

Deviance Residuals:

$$r_i^D = \text{sign}(y_i - \mu_i)\sqrt{d_i} \tag{5.8}$$

where d_i is the contribution of the given observation to the total deviance D, defined such that $\Sigma\ d_i = D$. The exact formula for d_i is dependent on the modelled distribution, but is easily derived; see Fahrmeir and Tutz (2001, p153).

The deviance residual measures the contribution of the individual observation to the total deviance, and hence can be used to identify any outliers that have a large effect on the fitted model.

Pearson Residuals:

$$r_i^P = \frac{y_i - \mu_i}{\sqrt{\text{var}(\mu_i)}} \tag{5.9}$$

The Pearson residual measures the contribution of the individual observation to the Pearson statistic. Furthermore, Chandler and Wheater (2002) note that the Pearson residuals can be used to check that systematic structure has been captured by the model. They note that if a fitted model is valid, then all the Pearson residuals should have mean 0, and variance 1. Furthermore, any subset of residuals should have a mean close to 0, and root mean square close to 1. Chandler and Wheater check the model for unexplained structure by examining carefully selected subsets of residuals, for example those of a given year, or from a given month. A systematic trend in residuals can suggest possible improvements to the model.

Anscombe Residuals:

The residuals above can be highly skewed for certain distributions. Anscombe residuals are specifically defined to be as normally distributed as possible, and are defined separately for each distribution (see McCullagh and Nelder, 1989, p38 for further details). For example, for a gamma-distributed predictand, the Anscombe residuals are defined by:

$$r_i^A = \frac{3\left(\sqrt[3]{y_i} - \sqrt[3]{\mu_i}\right)}{\sqrt[3]{\mu_i}} \tag{5.10}$$

Note that Yan et al. (2002) and Chandler and Wheater (2002) use a simpler formulation:

$$r_i^A = \left(\frac{y_i}{\mu_i} \right)^{1/3} \tag{5.11}$$

which provides identical results, save for a scaling factor. However, the studies presented here use residuals calculated by the software package used, MATLAB, using formula (5.10).

Whichever formula is used, if the assumption that the predictand can be modelled by a certain distribution is correct, then the Anscombe residuals should be normally distributed.

5.2.3 Model Selection

So far, methods of selecting a suitable set of covariates have not been mentioned. Often a large number of possible covariates are available, for example in this study 37 factors have been identified. Whilst we could build a model incorporating all these factors, it is usually inadvisable. The problem of multicollinearity was mentioned at the start of this chapter in the context of linear regression, but it also affects GLMs. Therefore, it is prudent to avoid selecting highly correlated predictors. Furthermore, selecting a large number of predictors can result in overfitting. Ideally the set of predictors should be parsimonious, that is, the selected set should be as simple and small as possible, yet explain a substantial proportion of total variation in the model. Furthermore, all covariates in the model should have a significant effect on the predicted variable.

The ideal approach is to consider all possible combinations of available covariates, known as the 'all-subsets' method (Fahrmeir and Tutz, 2001, p142). A model is fitted to each subset, and some statistic, Q, is calculated that measures the trade off between quality of fit and simplicity of the model. Atkinson (1981) notes that most variations of choice of Q are based on minimising the formula:

$$Q = D + \alpha p \varphi \tag{5.12}$$

where D is the deviance of the model, φ is the estimated dispersion parameter, p is the size of the covariate subset in the given model, and α is either a constant or a function of the number of observations. One of the most common choices, Akaike's Information Criteria (or the AIC) uses $\alpha = 2$ (McCullagh and Nelder, 1989, p91). Hence a penalty is added to the deviance, which increases with model size.

Unfortunately, an all-subsets search is not always possible. For example, in the case of this study with 37 potential covariates, there are 2^{37} possible models, which is approximately equal to 10^{11}. Obviously, it is computationally infeasible to consider all of these models.

Other possibilities are methods based on stepwise regression, often used in a multiple linear regression (von Storch and Zwiers, 1999, p166-7). One variety, a forward stepwise approach, is based on the following simple procedure:

1. Begin with an empty model, that is the model with only an intercept term.

2. Forward Step: Search for the factor not in the model which, when added, causes the largest decrease in deviance (and thus improves the fit most). Test to see if the decrease in deviance is significant. If so, add the factor to the model.

3. Backward Step: Search for the factor in the model which, when removed, causes the smallest increase in deviance (the least important factor in the model). Test to see if the increase in deviance is significant. If not, remove the factor from the model.

4. If steps 2-3 have altered the model, return to step 2. If not, finish the procedure, and use the current model.

A backward stepwise algorithm also exists, which begins with the full model: the model with all possible covariates included. The procedure is the same as for the forward approach, but with the order of steps 2 and 3 reversed. Unfortunately, the procedures used to calculate the maximum-likelihood estimates for GLMs are particularly prone to not converging when using large numbers of covariates (Fahrmeir and Tutz, 2001, p140 & 143). If the full model cannot be fitted, the

algorithm breaks down at the first stage. Furthermore, the time taken to fit the smaller models of a forward-stepping algorithm is typically much shorter than that taken to fit the large models in the backward method.

Therefore, the forward stepwise method seems most appropriate for this study. Note that the selected model cannot be considered optimal, but is merely a comparatively good choice. However, the optimal model may only be a marginal improvement, and the time taken to find it exorbitant.

5.3. Fitting the model to daily Sahelian rainfall

The previous sections have demonstrated that Generalised Linear Models provide a suitable framework for the modelling of daily Sahelian rainfall. This section will provide a description of how the model was fitted, and what steps were taken to ensure the model is valid.

The aim is to fit a GLM to daily rainfall in the six regions created in Section 4.2.3, illustrated in Figure 5.1. The predictor variables used will be the 37 leading components of atmospheric variability in the NCEP reanalysis in the region 0 - 20 °N, 30 °W – 60 °E, extracted using a rotated PCA, as described in Chapter 4. These predictor variables will be allowed to lead the modelled rainfall by up to five days, to allow for the study of atmospheric conditions prior to rainfall events.

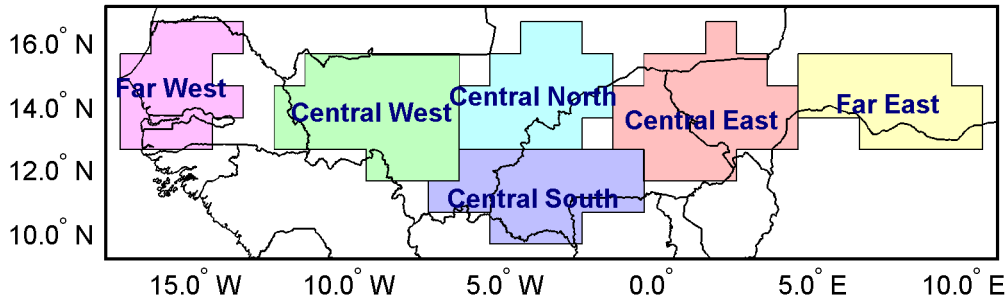


Figure 5.1. The six rainfall regions to be modelled using GLMs. For details of the creation of the regions see section 4.2.3.

This model will only be fitted to the period of interest: the four-month wet season from June to September. Furthermore, due to the concerns about data quality expressed in Section 4.1, only data for 1968 to 1997 will be used.

5.3.1. Details of the model

GLMs have been shown to be useful in the modelling of daily rainfall (Coe and Stern, 1982; Stern and Coe, 1984; Chandler and Wheater, 2002; Buishand et al., 2004; Abuarrea and Asín, 2005). These studies use a two-step approach. Firstly, logistic regression is used to determine whether a day is wet or dry. Secondly, gamma distributions are fitted to determine the amount of rain falling on wet days. The two-step approach is necessary as the gamma distribution is not defined for zero (or negative) values, and hence attempts to fit to data with zero values fail.

Fortunately, the processes of gridding and regionalising Sahelian rainfall described in Chapter 3 and Section 4.2.3 have removed the need for the first step in this study. The model will be based on the 3660 days of June to September, 1968-1997. During this period, the four central regions each had less than five days with zero rainfall (recall that the regional rainfall is given in units of standard deviations). The two extreme regions registered more zeros, probably due to the smaller number of stations available in the regions, with the Far West region recording 14 zeros, and the Far East recording 44.

Therefore, the chosen model was a simple one-step GLM based on the gamma distribution. On the very few dry days, rainfall in the regions was reset at a trace value (defined as 0.0001 standard deviations) to ensure a model could be fitted.

The other component of the model to be defined was the link function. The previous studies listed above all use the natural logarithm as a link function, although the canonical link for the gamma distribution is the reciprocal function. Hence the two obvious options were models defined by:

$$\log \text{link}: \quad \ln \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{5.13}$$

$$\text{reciprocal link:} \quad \frac{1}{\mu_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{5.14}$$

which can be rearranged to give mean rainfall in terms of the linear predictor (the right hand side of 5.13 and 5.14) as:

$$\text{log link:} \quad \mu_i = e^{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}} = e^{\beta_0} e^{\beta_1 x_{i1}} \ldots e^{\beta_p x_{ip}} \tag{5.15}$$

$$\text{reciprocal link:} \quad \mu_i = \frac{1}{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}} \tag{5.16}$$

Using the log link rather than the reciprocal gives two benefits. Firstly, note that values for μ_i using a log link are by definition positive, which is obviously useful when modelling rainfall. Secondly, note that as the linear predictor approaches zero, μ_i tends to one for the log link, but tends to infinity for the reciprocal link (or negative infinity if approaching from the negative side). Hence, a very small linear predictor with a reciprocal link will lead to a very large predicted mean. Consequently, modelling with a reciprocal link can give some surprisingly large prediction errors for a few observations.

For these reasons, the model selected was a gamma-distributed GLM with a log link function. Note that the expansion of the exponential function in equation 5.15 reveals that the model is now multiplicative.

5.3.2. Is the gamma model valid?

Before the model was fitted, a few tests needed to be carried out, to ensure that it would be valid. Firstly, the regional rainfall needed to be shown to be approximately gamma distributed. Secondly, section 5.2.1 introduced the dispersion parameter φ, which is assumed constant for all i. In the case of the gamma distribution, φ is the variance divided by the square of the mean (McCullagh and Nelder, 1989, p30). The validity of this assumption was tested by calculating the corresponding value of φ from the data for each day of the year. If the assumption holds true, the value should be equal for each day of the wet season. However, some degree of variability is

inevitable with only thirty years of data. Results of the two tests are displayed in Figure 5.2.

The plots suggest that the gamma distribution approximates rainfall in each region well, except for some extremely high values. Estimated dispersion parameters vary considerably at the start and the end of the year (partly due to the high number of zeros), but appear nearly constant in most of the wet season. Problems may occur at the start and end of the season, as the plots indicate the dispersion parameter may be slightly higher, particularly for the Far West and Far East plots.

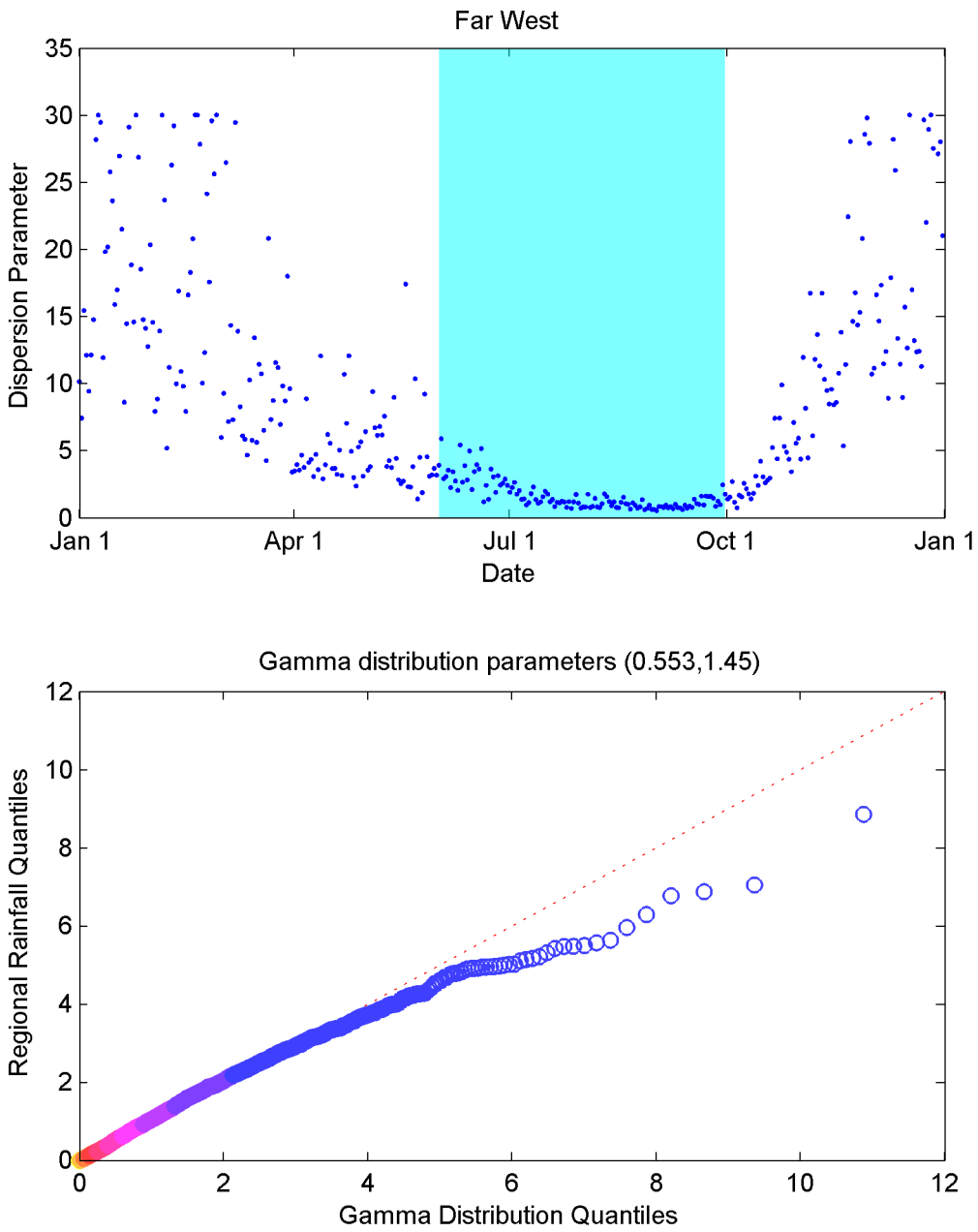
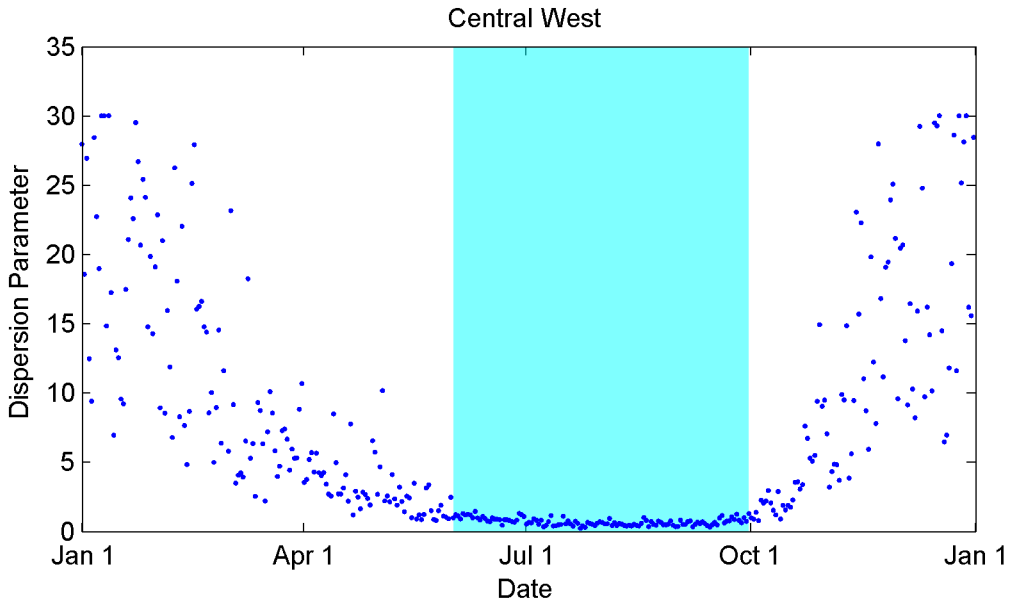


Figure 5.2(a). Validation plots for the Far West region. The top plot shows the estimated dispersion parameter for each calendar day, with the shaded region indicating the wet season to be modelled. The bottom plot shows a quantile-quantile plot comparing the distribution of regional rainfall with the best fitting gamma distribution. The different shades of circles indicate deciles of data (i.e. each shade represents 10% of total data), illustrating the clustering of data near the origin.



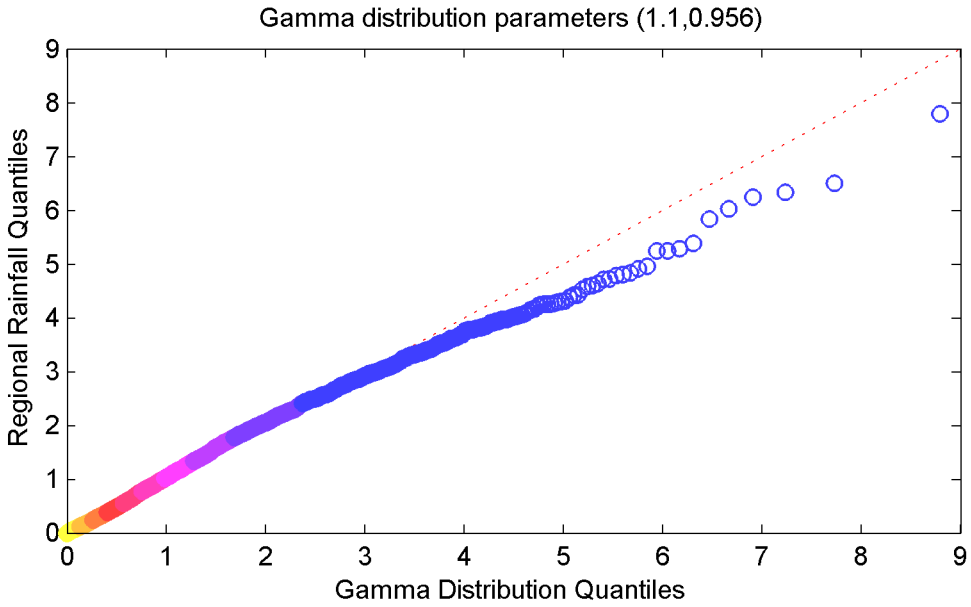


Figure 5.2(b). Validation plots for the Central West region. See Figure 5.2(a) for explanation.

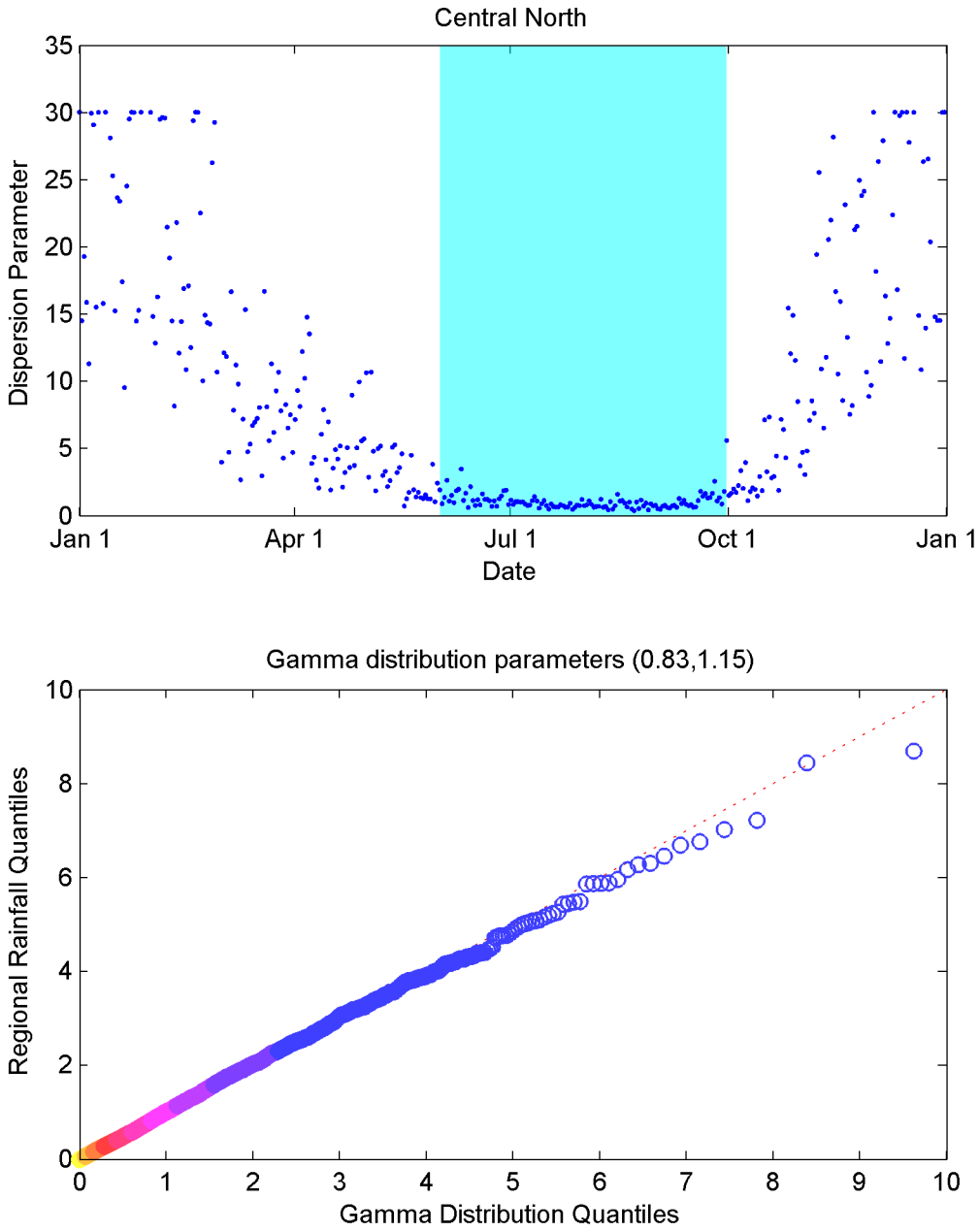
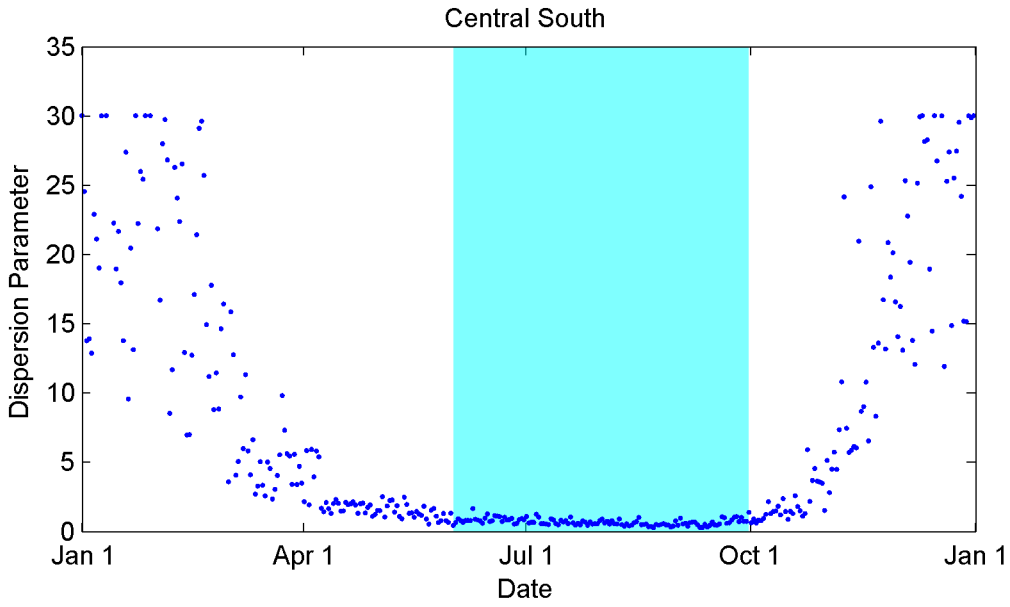


Figure 5.2(c). Validation plots for the Central North region. See Figure 5.2(a) for explanation.



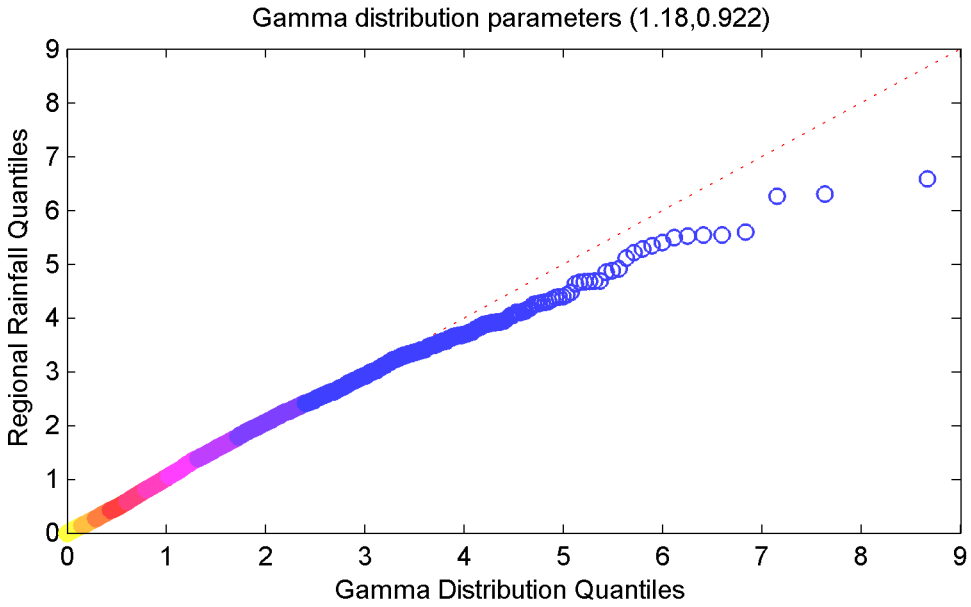
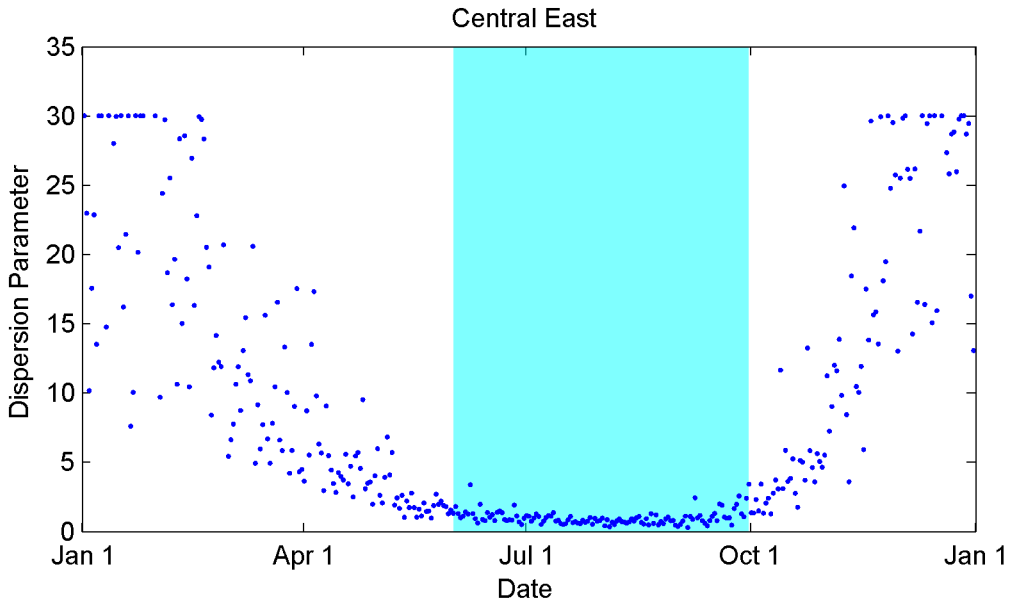


Figure 5.2(d). Validation plots for the Central South region. See Figure 5.2(a) for explanation.



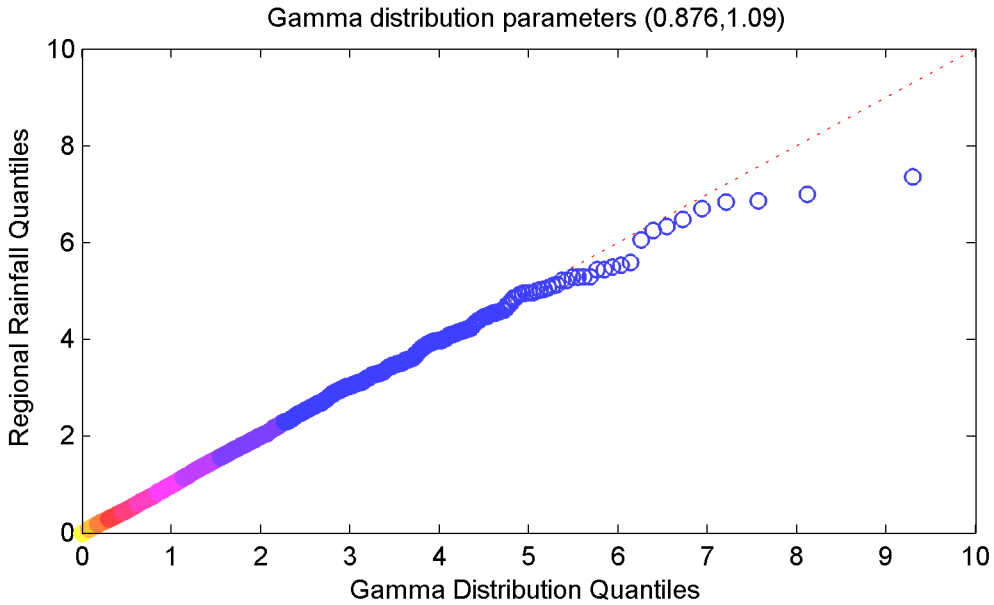


Figure 5.2(e). Validation plots for the Central East region. See Figure 5.2(a) for explanation.

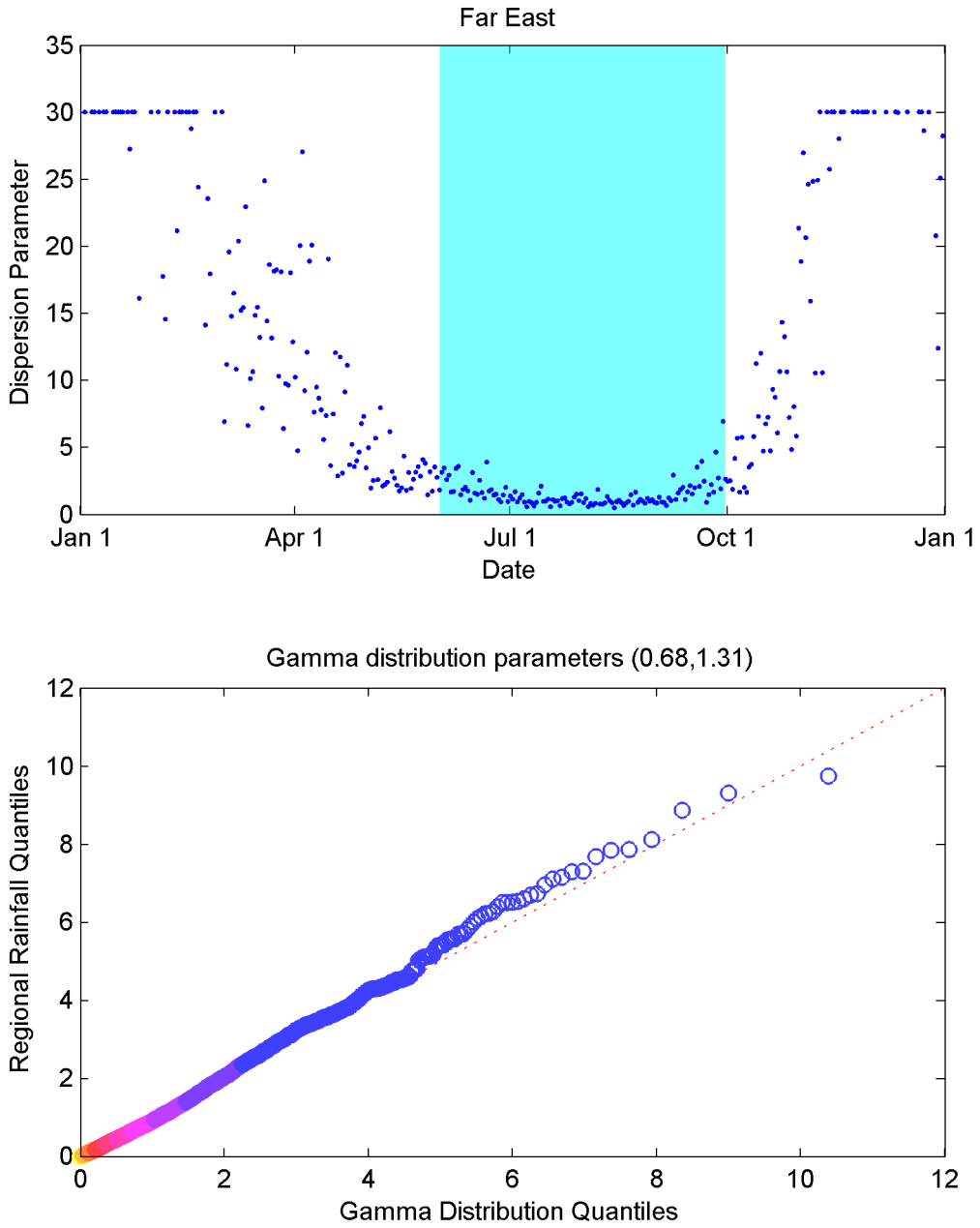


Figure 5.2(f). Validation plots for the Far East region. See Figure 5.2(a) for explanation.

A final assumption made is that each $\mathbf{Y}_i$ is independent. Hence, we are assuming that rain today is not dependent on rain yesterday, probably not a safe assumption. However, this can be rectified using the strategy of Chandler and Wheater (2002), who use the amount of rainfall on a previous day as a predictor (see Chapter 6 of Fahrmeir and Tutz, 2001, for a justification of this procedure). Note that the initial model (described in the next section) did not include this modification.

5.3.3. The Initial "Full" Model

Initial attempts involved fitting separate gamma-distributed GLMs with a log link function to each of the six rainfall regions. The 37 predictor variables were allowed to lead rainfall by up to five days, giving a total of 37 × 6 = 222 possible predictors. At this stage, the possible inconsistencies in the NCEP dataset prior to 1968 were not known, so the analysis was based on 1958-1997. Furthermore, the model did not include any previous values of rainfall as a predictor.

The predictors used were chosen using a forward stepwise procedure, with a 5% level of significance required to add or remove an element from the model. The procedure was carried out using MATLAB. Due to the large number of possible models, a completed procedure for each region took several hours.

Table 5.1 displays the number of factors in the model fitted for each region, and its goodness of fit as measured by the deviance and the adjusted R^2 of Chandler and Wheater (1998). Note that as the response variable is modelled by the gamma distribution, the deviance is given by (McCullagh and Nelder, 1989, p34):

$$\sum_{i=1}^{n}\left[-\ln\left(\frac{y_i}{\mu_i}\right)+\frac{y_i-\mu_i}{\mu_i}\right] \tag{5.17}$$

Each of the six fitted models contained at least 23 factors, with the two outlying regions requiring over 40. Such large models cause a number of problems. First, the models are almost certainly overfitted; they provide a good fit to the particular sample of data taken, but not necessarily to the underlying population. Second, even if the model were valid, a model of that size would be very hard to interpret.

Figure 5.3 shows the evolution of the two goodness of fit parameters for the Central West region during the forward stepwise variable selection procedure. Note that for this particular model, no factors were removed from the model during backward steps. This is an exceptional case; for all the other models between four and seven factors were removed. Hence, for the Central West region deviance decreases at each step of the procedure.

Region	Factors in Model	R^2	Deviance
Far East	43	0.218	6766
Central East	29	0.203	5298
Central West	25	0.256	3614
Central North	31	0.209	5304
Central South	23	0.180	3888
Far West	45	0.160	7118

Table 5.1. Size of model and goodness of fit statistics for the initial six models.

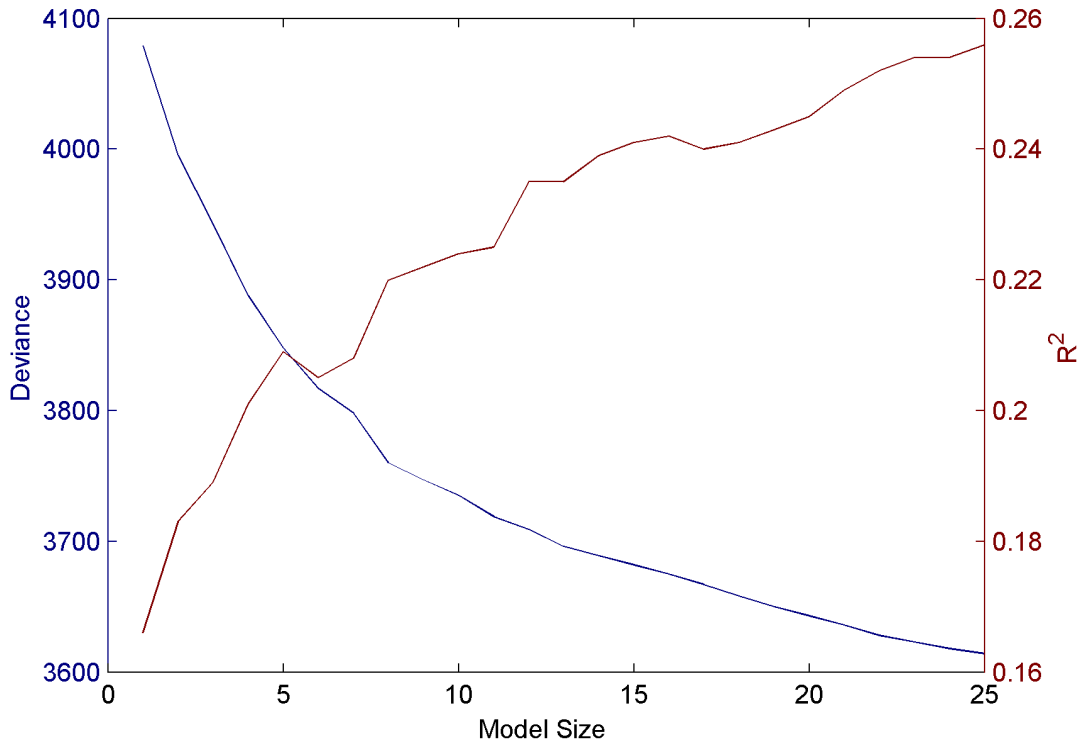


Figure 5.3. Evolution of the deviance and R^2 goodness of fit diagnostics during the stepwise variable selection procedure for the initial GLM fitted to the Central West region.

The decay of the deviance is approximately exponential, so the marginal benefit from adding each successive factor decreases. The same is largely true for the R^2, although note that occasionally R^2 decreases when a factor is added. This demonstrates that minimising deviance does not maximise R^2, and hence R^2 is not a

perfect indicator of fit for gamma-based GLMs. The marginal benefit in adding extra factors has to be balanced against the increase in complexity of the model, added risk of overfitting, and the considerable rise in time required to fit larger models.

The accusation of overfitting was investigated by bootstrapping the results of the final fitted model. Bootstrapping allows the investigation of stability of fitted parameters without making assumptions about their distributions, and is performed by the resampling of data cases with replacement (see Davison and Hinkley, 1997, Chapters 6 & 7).

The bootstrapping procedure was carried out as follows (adapted from Davison and Hinkley, 1997, p264):

For r = 1, …, 1000

- Sample $i^*_1,\ldots,i^*_n$ randomly from {1, 2, …, n}
- for j = 1, …, n, k = 1, …, p, set $x^*_{jk} = x_{i^*_j k}$, $y^*_j = y_{i^*_j}$
- Fit a GLM to (x^*_{11}, $x^*_{12,}$ …, x^*_{1p}, y^*_1), giving estimated for model coefficients and diagnostics (e.g. deviance and R^2)

where j represents the index relating to observation, and k the index relating to predictor variable.

Figure 5.4 shows the bootstrapped coefficients for the specific humidity variables in the GLM fitted to the Far East region. Note that many predictors appear at multiple lags. Also note the wide range in estimated coefficients, particularly in the 'shum1' variable. This instability is a typical indicator of collinearity, and suggests overfitting has occurred, particularly as there seems no physical reason why a large fluctuation in 'shum1' (humidity at low and mid-levels over the Gulf of Guinea) over three days should be associated with increased rainfall in western Niger.

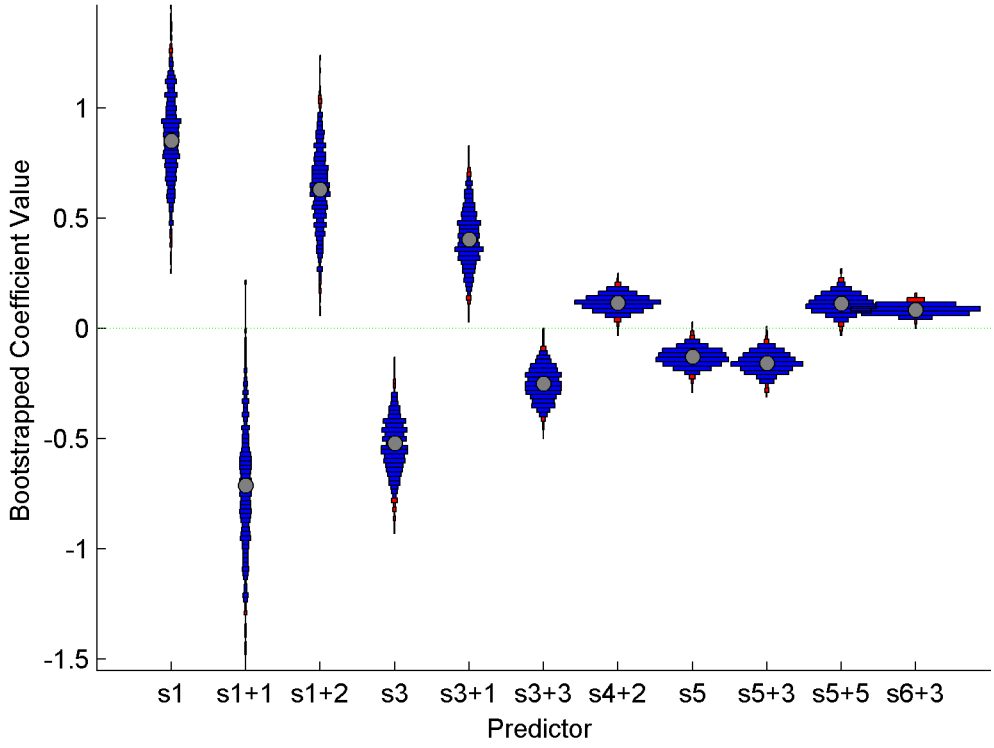


Figure 5.4. Bootstrapped coefficients of the specific humidity variables in the original GLM fitted to the Far East region. The width of bars indicate the proportion of the 1000 bootstraps fitting coefficients of that value. Blue bars indicate a 95% confidence interval for the true coefficient, red bars indicate values outside the interval. Grey circles indicate the value of the coefficient in the original GLM. Predictor names are represented by the first letter of the variable, number of the PC, and the lag of the predictor. Hence 's4+2' represents the fourth specific humidity component leading rainfall by two days.

Any theoretical problems with the initial models are moot, to a certain extent, as the large size of the model make them almost uninterpretable. Later models reduced model size by considering predictors at each lag separately, and increasing the level of significance required to change the model during the stepwise procedure.

5.3.4. Improving the model

The improved model was fitted considering each lag separately. Models were based on data from 1968-1997, ignoring the pre-1968 period when NCEP atmospheric data is problematic. In order to reduce model size, the level of significance required to add or remove a factor was increased to 1%. The issue of autocorrelation in the

rainfall time series was accounted for by adding the rainfall on the previous day as a predictor in the baseline model. Finally, models were also fitted to June-July data only, and to August-September data only, recalling the two-stage monsoon theory suggested, amongst others, by Nicholson and Palao (1993) and Lebel et al. (2003).

The coefficients, fit statistics and model size of the fitted models are displayed in Appendix B, figures B.1-42. Note that model sizes are now considerably smaller, containing between two and thirteen atmospheric factors. Most models explain between ten and twenty percent of regional rainfall. Deviance suggests that the models for the four central regions are the best fitting, despite typically containing fewer atmospheric predictors. Note deviance in the two-month models is roughly half the deviance in the four-month models. This is a predictable result: deviance is an equivalent measure to the residual sum of squares in a linear regression.

Of particular note is the apparent failure of the model to predict rainfall in the Far West region in June / July. Negative scores are recorded for R^2, suggesting the fitted gamma distribution mean for each day is a worse predictor than the overall mean of the rainfall series. This failure will be investigated later, but note that the previous day's rainfall ('Prior1') is a considerably more important in the Far West than in any other region.

Some care must be taken when comparing the value of coefficients in different models. Recall that all predictors (including prior rainfall) were standardised before being used, hence within a model the most influential factor will have the largest absolute coefficient. However, coefficients in a small model will tend to be larger than those in model with many factors. Also, note that if a factor is not included in a model (indicated by a value of 0 in the figures), it does not mean it has no effect on rainfall, but that any effect it does have can be better explained by other factors.

An initial scan of results shows that some factors are clearly more important than others. Air temperature and geopotential height components appear in very few models; indeed the second geopotential height component ('gph2') does not appear in any model. However, there are some components, particularly those representing wind speeds, which occur in many of the models.

Specific humidity also provides some important components, in particular 'shum1', for which a positive value represents enhanced humidity over equatorial Africa and the Gulf of Guinea, particularly at lower levels. 'shum1' appears in 77 of the 108 models, with a strong positive loading in all cases. This importance is unsurprising. The average daily time series of the factor (shown in Figure A.18) is similar to the typical daily rainfall cycle, and the factor represents humidity in roughly the right area, although slightly to the south of the Sahel.

Zonal wind also plays a critical role, particularly in August and September. The appearance of positive loadings for 'uwind4' at short lag times suggest enhanced low-level westerlies over the Gulf of Guinea are linked to increase rainfall, hence the factor seems to represent the monsoon flow. Interestingly, 'uwind4', westerly flow over the oceanic Gulf of Guinea, is selected in preference to 'uwind3', westerly flow over the continental Guinea Coast and Sahel. Thus, Sahel rainfall has a stronger link to offshore monsoon flow.

At longer lead times rainfall is associated with negative loadings of 'uwind5' (representing strengthened low and mid-level easterly flow centred over Northern Sudan) and 'uwind1' (easterlies at 200 hPa and over the Gulf of Guinea at 600 hPa contrasted by westerlies elsewhere, particularly over East Africa and the coastal Indian Ocean). This suggests that at higher lags, zonal flow east of the Sahel is more important. However, the zone of interest is further east than the area where easterly waves originate, so the link to rainfall is not obvious.

The meridional wind components provide some extremely interesting results. Two components, 'vwind5' and 'vwind6', are particularly notable. 'vwind6' seems particularly important in the first half of the wet season. At no lag, and lags of one day, positive scores are associated with decreased rainfall in the western regions (and vice versa). However, at a lag of one day, positive scores are also associated with increased rainfall in the east. Positive scores are also associated with increased rainfall in the central regions at two days, and at the western regions at three days. Hence, the positive link appears to 'drift' across the Sahel. 'vwind5' shows a similar, but reversed pattern, with negative loadings drifting across the Sahel.

These drifting motions from the east to west are rather reminiscent of the movement of easterly waves across the Sahel. Furthermore, both pattern loadings show a band of northerlies just to the west of a band of southerlies in the lower and mid troposphere, possibly suggesting a wave motion. The drift takes about four days to travel the width of the whole region (approximately 3000 km). This equates to a speed of 8.5 ms^{-1}, comparable to the wave speed of 8 ms^{-1} quoted in Reed et al. (1977). Furthermore, the gap between the positive and negative centres in the patterns is approximately 1500 km, suggesting a wavelength of 3000 km. Estimated wavelengths for African waves range from 2000 – 4000 km (Carlson, 1969; Burpee, 1972).

Prior rainfall proves to be a relatively unimportant (but significant) factor in all models, except those fitted to the Far West region. Investigations were carried out as to whether using rainfall two and three days prior to the day in question improved the model, factors referred to as 'Prior2' and 'Prior3'. The coefficients relating to atmospheric predictors changed little when these factors were included, hence they are not shown here, and this suggests that using only 'Prior1' is sufficient.

The 'Prior1' predictor is particularly important in the Far West region. This is a result of the much higher autocorrelation than in other region, with a lag 1 correlation of 0.41 for June-August, compared to a maximum of 0.23 in the other regions. 'Prior1' also indirectly results in the poor performance of the Far West model in June and July, when measured by the R^2 statistic. However, analysis suggests this may be due to a failure of the R^2 statistic to represent variability, rather than a failure in the model itself.

Models without the 'Prior1' predictor had been fitted to all regions. Table 5.2 shows the statistics for the two models fitted to the Far West region for June-July using predictors at a lag of five days; the right column uses 'Prior1', the left does not. The selected models and fitted coefficients (not shown) are similar; only three factors appear in exactly one of the models. The R^2 statistic suggests that the model without a previous rainfall predictor is reasonable, and the model with the predictor is abysmal. Note, however, that the deviance is lower for the model using 'Prior1'. Comparing deviances of non-nested models should be done with caution, but this

does suggest that the quality of the two models is nowhere near as distant as the R^2 statistic suggests.

Statistic	Model without 'Prior1'	Model with 'Prior1'
Atmospheric Factors in model	6	7
Deviance	4469	4305
R^2	0.10	-2.14

Table 5.2. Fit statistics for two models fitted to the Far West region for June-July using atmospheric predictors at a lag of five days. One model uses the 'Prior1' predictor (rainfall on the previous day); the other does not.

Figure 5.5 shows three histograms. The first represents the variable to be predicted, regional rainfall in the Far West region for June-July 1968-1997: units are in standard deviations. The second and third histograms represent the predicted rainfall for the models without and with 'Prior1' respectively. As can be seen, variability of the 'observed' rainfall is much higher than for either of the predicted models. However, recall that the predicted rainfall values are the mean of a distribution, the mean of the fitted distribution being the best prediction. If a direct comparison of distributions was required, then random values from gamma distributions using the fitted parameters should be created.

The top histogram indicates that an extreme outlier exists in the rainfall series, a value of 8.85, which occurred on the 29th July 1975. The extreme outlier on the predicted series using 'Prior1', with an astonishingly high value of 50.79, occurs for the 30th July 1975, when the actual rainfall was 2.67.

In the second model, the coefficient corresponding to the previous day's rainfall is 0.40. Hence, the contribution of 'Prior1' to the multiplicative model is given by $e^{(0.40 \times 8.85)} = 34.47$. As a comparison, the next highest value of rainfall, 5.50, would only contribute 9.03. Hence the extreme rainfall on 29th July 1975 caused an even more extreme predicted rainfall for 30th July 1975. Obviously, the model not using 'Prior1' as a predictor did not suffer from this problem.

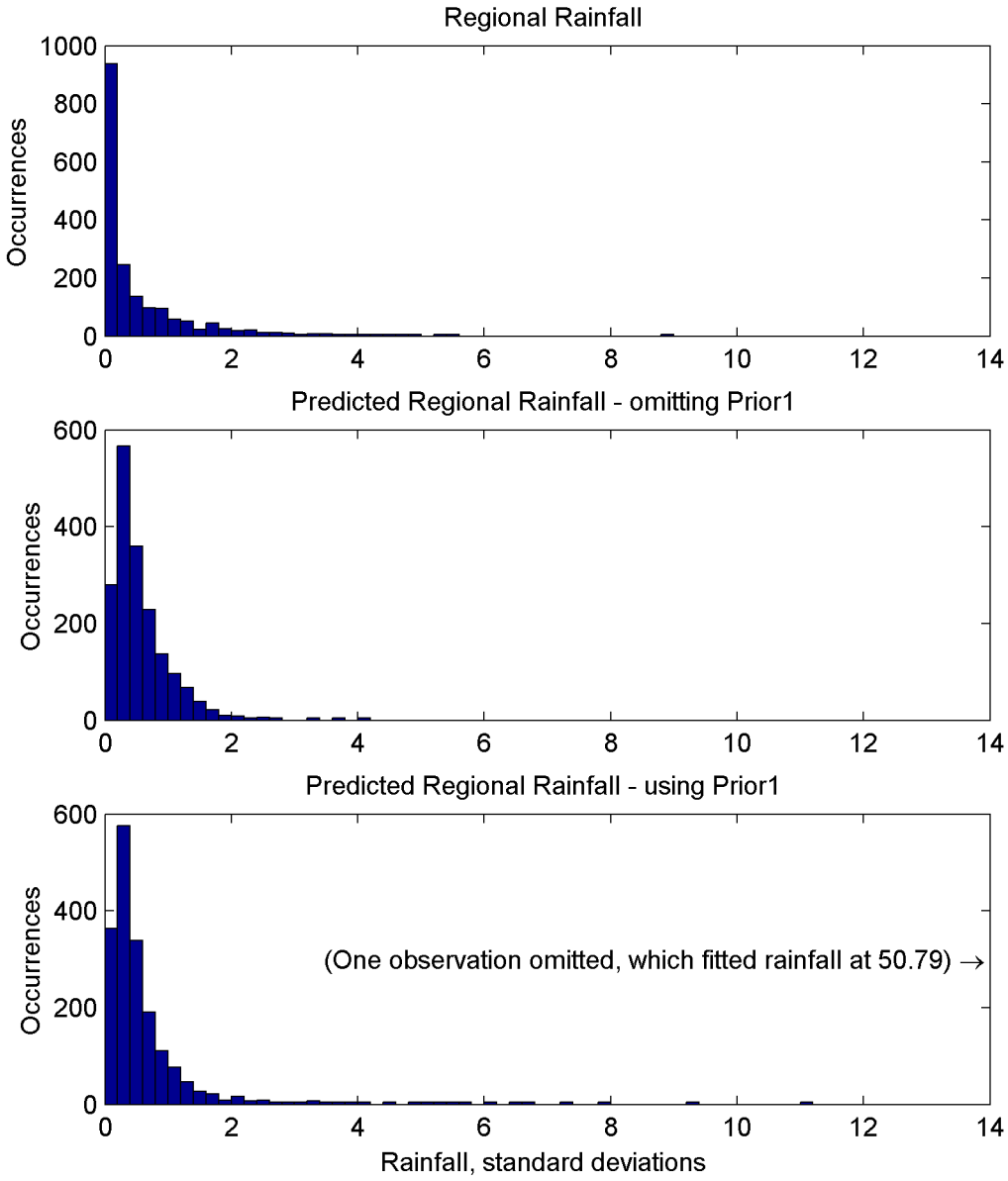


Figure 5.5. Histograms of actual and fitted regional rainfall for the Far West region. The bottom two plots use predicted data formed from the GLM using atmospheric predictors leading rainfall by five days, with the lower plot also using the previous day's rainfall as a predictor. The bin width of each bar on the x-axis is 0.2. Note the different scale on the y-axis in the top plot.

Note that even ignoring this extreme, the tail of the 'Prior1' model is much longer than the model without. It is likely that these outliers heavily influence the R^2 for the 'Prior1' model. This is a further indication of the caution that should be used in interpreting the R^2 statistic. Note that the deviance residual for the 30th July 1975 for the 'Prior1' model is –2.0. This seems rather small, for comparison see Figure 5.6, which shows histograms of the deviance residuals for the two fitted models in question. However, the appearance of fitted rainfall in both the denominators in

equation 5.17 suppresses the influence of vastly overestimated cases; a desirable property for distributions as highly skewed as daily rainfall. In terms of deviance, the second model is better fitting.

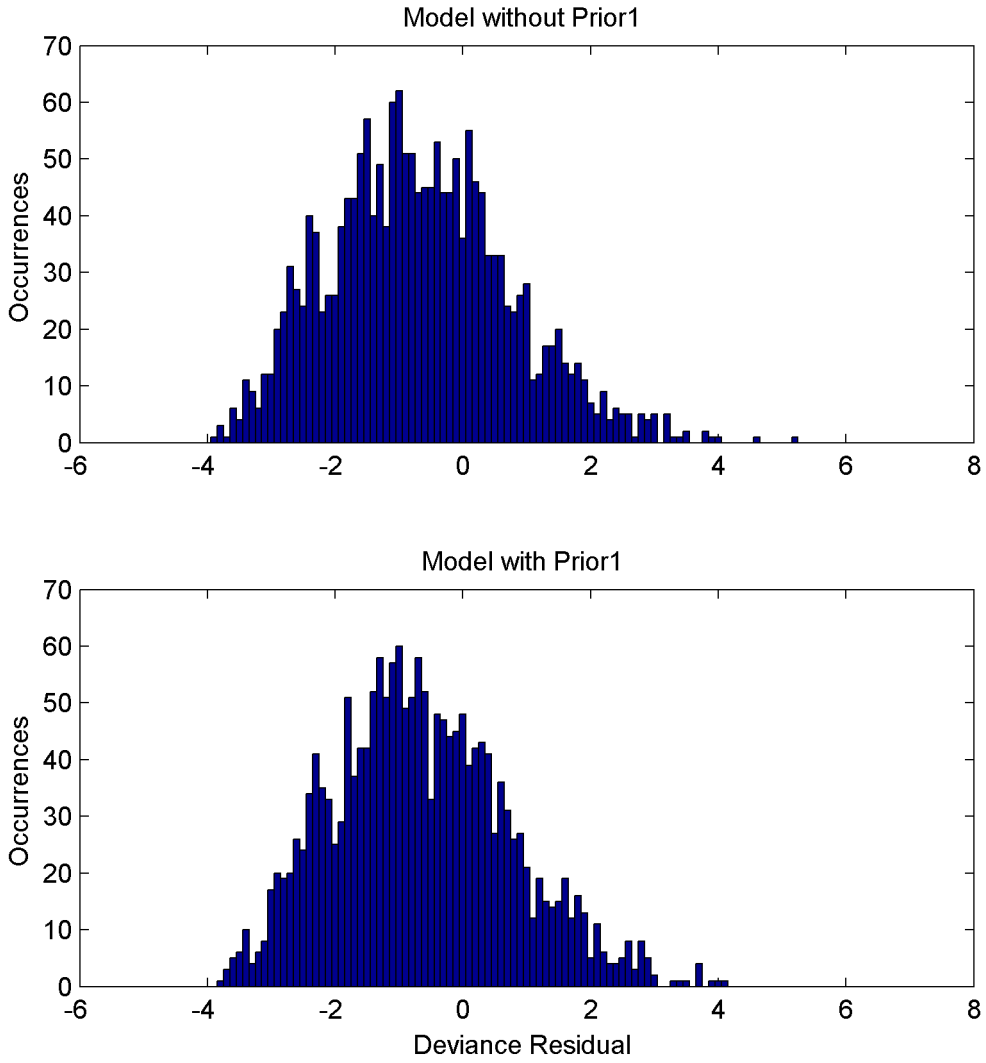


Figure 5.6. Histograms of deviance residuals for the two models fitted to the Far West region using atmospheric predictors leading rainfall by five days. The lower plot also using the previous day's rainfall as a predictor. The bin width of each bar on the x-axis is 0.1.

Remember that irrespective of quality of fit, the model not using 'Prior1' as a predictor should be viewed with suspicion, due to the autocorrelation in the rainfall series.

The validity of the models was checked by examining the distribution of Anscombe residuals and the mean of Pearson residuals, in the manner of Chandler and Wheater, (1998). If the assumption that rainfall is gamma-distributed holds, then Anscombe

residuals should be approximately normal. Systematic trends in residuals can be discovered by examining Pearson residuals: any subset should have a common mean (zero) and variance. Chandler and Wheater indicate that the standard error of any given subsample of Pearson residuals is given by:

$$\text{standard error} = \frac{1}{\sqrt{\hat{\varphi} N}} \tag{5.18}$$

where φ is the dispersion parameter and N is the size of the sub-sample.

A typical set of validation plots is produced in Figure 5.7. The top plot, analysing the distribution of Anscombe residuals, shows they are almost normal. Deviation from the 45° line only occurs at the ends of the distribution, and, in this case, only noticeably at the bottom tail. This deviation is typical when modelling rainfall amounts (see Chandler and Wheater, 2002), and occurs due to the Gamma distribution being unable to yield negative values, whilst the normal distribution does.

The bottom plots show the mean of Pearson residuals subsets divided in two ways, on the left by year, and on the right by calendar day. As can be seen, yearly averages lie well within the confidence limits for a zero mean residual. For monthly average, more points lie outside the limits than may be expected. However, there is no clear systematic trend to the outliers.

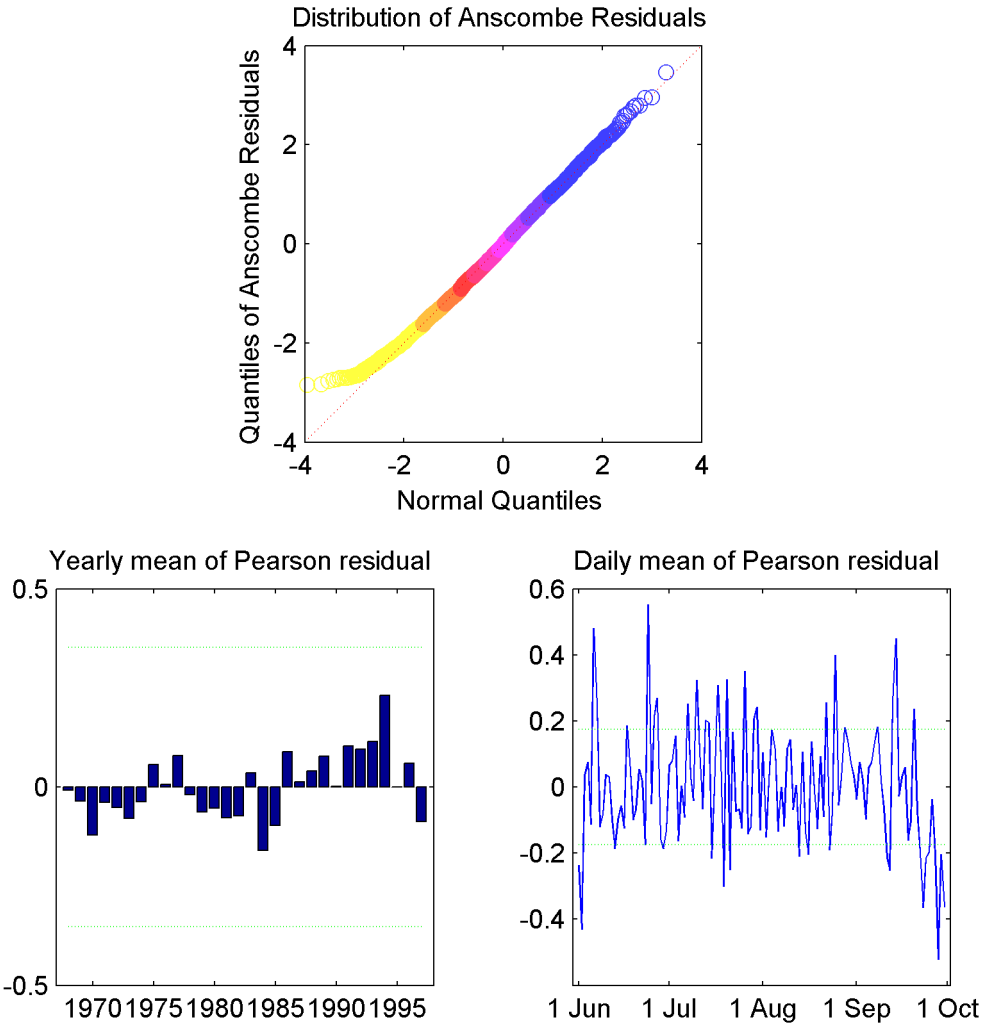


Figure 5.7. Verification plots for the GLM fitted to the Central North region for June-September, using atmospheric predictors at a lag of three days. The top plot is a quantile-quantile plot of Anscombe residuals against a normal distribution: as in Figure 5.2, each 10% of data is represented in a different colour. The bottom two plots represent the means of subsets of Pearson separated by two different techniques. The left plot considers yearly averages; the right plot splits residuals by day of the year. Dotted lines indicate 95% confidence limits for a zero mean.

Finally, the stability of the fitted coefficients was assessed by bootstrapping the fitted models, using the method outlined in Section 5.3.3. The bootstrapped coefficients for the model fitted to the Far East are again displayed, so results can be directly compared with Figure 5.4. The results, shown in Figure 5.8, consider the model fitted using predictors at a lag of five days, and for a four month wet season. In this case, all fitted coefficients are displayed.

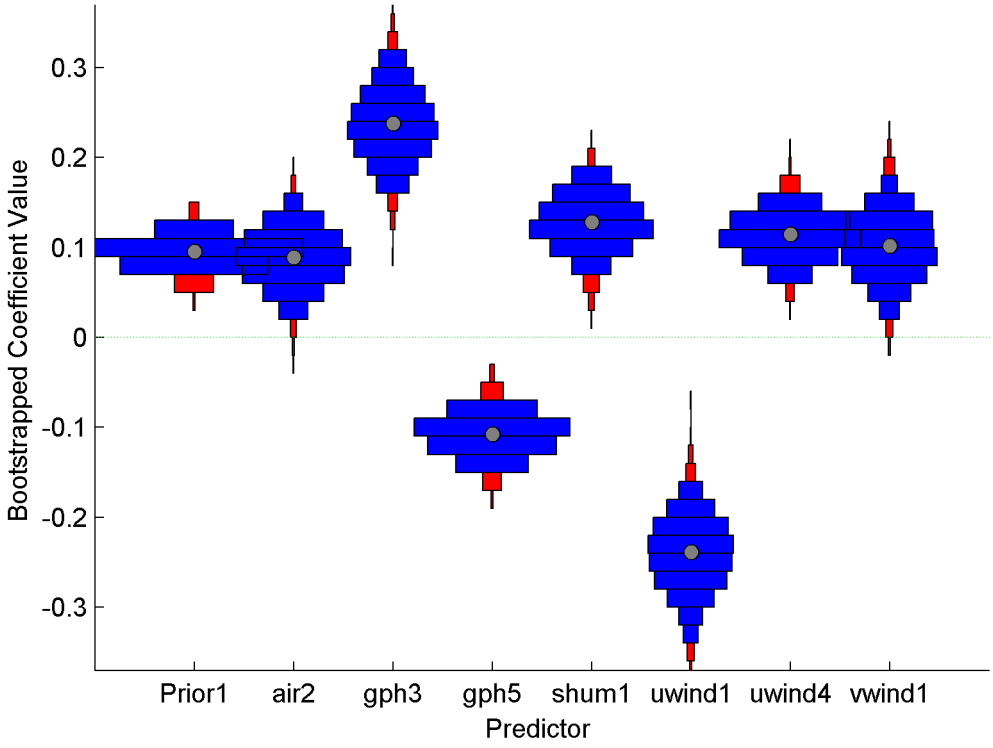


Figure 5.8. Bootstrapped coefficients of variables in the modified GLM fitted to the Far East region, for June-September, using predictors at a lag of five days. As for Figure 5.4, except predictor codes are shown in full.

The bootstrapped coefficients seen in Figure 5.8 are clearly more stable than those shown in Figure 5.4, presumably as a result of the reduced correlation between the underlying variables. Hence, the improvements made to the model have had the desired effect.

5.3.5. Fitting to the gridded rainfall series

The analyses described in the previous sections have created a set of models for rainfall in the six regions. The next step was to attempt to fit GLMs to the individual grid boxes. This would enhance spatial resolution in the analysis, which would increase the proportion of actual rainfall variability available to be modelled.

The criterion for selecting which grid boxes to use was the same as in section 4.2.3: only grid boxes with a rainfall stations reporting 50% of values for June-September 1958 – 1997 were used. Models were again fitted to 1968 – 1997, and for the three

different yearly periods (June/July only, August/September only, and June-September). The amount of rainfall occurring on the previous day was included as a predictor. Atmospheric predictors were again selected using a stepwise method. However, air temperature and geopotential height components were not considered, due to their apparent lack of importance in the regional GLMs. This decision significantly reduced the time required to fit the models.

Fitted coefficients and model statistics of these 'gridded GLMs' are shown in Appendix B, figures B.43-73. Each plot shows the rainfall regions in green to allow comparison with the previous analyses. Results are similar: the best fitting and simplest models are in the centre of the domain, particularly in the south.

The large negative values of R^2 fitted to the western region in June / July are reflected in the gridded GLMs. Indeed, the greater variability of the gridded rainfall enhances this problem: for example the worst fitting box (16°N, 16 °W) at zero lag has an R^2 of –10780, (note that this is a highly exceptional case; the second worse case is 'only' –11.3). This poor fit also extends to August / September for boxes in the north of the Far West region. This suggests all fits to the extreme northwest should be ignored.

The gridded results permit a more detailed evaluation of some patterns. For example, the gridded results for 'shum1' (shown in Figure B.48) suggests that the association between the factor and rainfall also drifts across the Sahel, in a similar manner to the wave-related meridional wind components. So, enhanced humidity along the Gulf of Guinea is associated with enhanced rainfall in the east of the region on the same day, and enhanced rainfall in the west several days later. This trend is particularly apparent in August and September.

Similar August / September drifts occur in some of the zonal wind components. The importance of 'uwind1' (associating rain with westerlies in the lower troposphere overlaid by easterlies at 200 hPa) drifts in from the east from a lag of 1 day, and is important for all central and eastern regions for five day lags. The drift of 'uwind4' (low level Guinea Coast westerlies) is almost an exact reverse; importance in all regions but the west at zero lag is diminished to relevance mainly in eastern regions

by day 5. Finally, the negative 'uwind5' pattern (implying rainfall is linked with low and mid level easterlies over Northern Sudan) seems to drift slowly to the west.

Whilst zonal wind predictors are principally useful in August and September, several meridional wind predictors dominate June and July. For example, 'vwind2' appears in many models throughout the wet season, but appears to exert a stronger influence in the first half. 'vwind2' is mainly represented in the model by negative loadings, especially in the western half of the domain. This factor links increased rainfall within the next three days to a band of low-level northerlies over Chad and Sudan, to a band of southerlies at 600 hPa just to the west, and to a band of surface southerlies over the Atlantic Ocean.

The drift of the 'vwind5' and 'vwind6' predictors, seemingly indicators of wave activity, is more apparent in the gridded GLMs. They too play an important role throughout the wet season, but are particularly influential in June and July. Furthermore, note that the fitted GLM coefficient patterns (Figures B.68 and B.69) for the two predictors in June and July seem to be 90° out of phase, with 'vwind5' lagging behind 'vwind6' by one day. This is consistent with a wave activity interpretation; the PC pattern for 'vwind6' is located to the west of 'vwind5' (Figures A.38 and A.39), thus the wave has travelled further west.

Other meridional wind patterns seem to have a greater influence in the later part of the wet season. For example, 'vwind3' suggests increased low-level southerly flow is associated with increased rainfall at short lags across Burkina Faso and the segment of Niger covered in the analysis. This seems a reasonable supposition; increased southerly flow would bring extra moisture into the Sahelian region.

Some links between predictors and rainfall are not so easy to explain. 'vwind10' suggests higher levels of rainfall occur in the east half of the region 2-5 days after the occurrence of strengthened northerly flow at the top of the troposphere. A physical explanation for this association is unclear. Similar difficulty surrounds the link between a positive score for 'vwind4' and decreased rainfall in the eastern half of the domain on the same day, followed by increased rainfall four and five days later across much of the area. The 'vwind4' PCA pattern links positive loadings with an

area of increased low and mid level southerlies to the east of the considered region, at about 30 °E.

One final observation concerns the difference between the two halves of the wet season. It is noticeable that the processes that seem to dominate in June and July, (namely: 'shum1, 'vwind2', 'vwind5' and 'vwind6') also seem to have an influence in August and September. This suggests that the two halves of the monsoon season are not controlled by completely different processes: rather, the second half has some additional influences.

5.4. Conclusions

This chapter has presented a set of Generalised Linear Models that have linked daily rainfall across the Sahel to wider atmospheric variability. Valid models have been fitted for all but the far-western regions of the Sahel, where attempts failed due to a higher level of autocorrelation in the rainfall series.

The loadings of the GLMs suggested that, of the atmospheric variables considered, changes in specific humidity and wind speed are most closely related to rainfall variability. Many of the factors in the models can be interpreted in terms of features of the West African rainfall cycle that are well known, such as westerly monsoon flow and easterly waves.

Some results seen in these analyses are straightforward. It is hardly surprisingly that an increase in the 'uwind4' factor, representative of monsoon flow, is linked with an increase in rainfall. The modelling of rainfall in several different regions, and for two different periods of the year, has allowed the identification of some subtle changes in the importance of factors. For example, the monsoon flow pattern 'uwind4' is more important in the east of the region, and in the second half of the wet season. Furthermore, the production of models for different lags has identified factors whose importance appears to 'drift' across the Sahel over several days.

Finally, these models have indicated that winds over East Africa may have an influence on rainfall in the Sahel, as indicated by 'uwind5' and 'vwind4'. However,

the complexity of the PCA and GLM patterns for these factors mean that the mechanism behind this link is not obvious. The most likely explanation is that these changes in wind somehow affect the genesis of easterly waves.