Chapter 3

Estimating Dry-Day Probability for Areal Rainfall

Contents

3.1. Introduction

Previous attempts to evaluate the ability of GCMs to simulate precipitation with the appropriate level of daily variability have been hampered by the lack of areal rainfall observations, against which model simulated precipitation can be compared. For densely gauged regions, a simple or interpolated average of available stations can give a sufficiently reliable estimate of areal precipitation to be used for this purpose with relative confidence. However, there are many regions of the world for which the observation network is to too sparse to give a reliable estimate of areal precipitation, specifically with respect to the level of variability. Studies such as Osborn and Hulme (1997) are Booij (2002a, 2002b) are among the few which have attempted to resolve the differences in temporal variability in point and areal-average daily rainfall in order to make a quantitative assessment of model performance.

In the literature review (Chapter 2), it was noted that numerous studies have tackled the scaling of extreme values for point and areal rainfall (e.g. Omolayo, 1993; Coles and Tawn, 1996; Sivapalan and Bloschl, 1998) for use in other hydrological applications. Fowler *et al*. (2005), for example, made use of Areal Reduction Factors (ARFs) to evaluate the performance of HadRM3 with respect to extremes of daily rainfall over the UK. However, the scaling of characteristics of the non-extreme parent distribution, such as wet-or-dry-day frequency, mean intensity, standard deviation and parameters of the distribution with which the values can be modeled, have received less attention. Osborn and Hulme (1997) have addressed this issue and proposed an approach to estimating the standard deviation and wet-day frequency of areal rainfall for a grid box using the characteristics of a small number of stations in the region. This technique is described in 2.3.1.4. Osborn and Hulme (1998) go on to apply this in an evaluation of twelve Atmosphere-only GCMs.

The following two chapters attempt to bridge the gaps identified in the existing literature, building on the approach of Osborn and Hulme (1997). The Osborn and Hulme (1997) approach is improved and expanded in the following two chapters; the first (this chapter) proposes and tests an alternative to the estimates of areal rainfall wet-day frequency which addresses a caveat in their approach, the second (Chapter 4) goes on to expand the existing approaches by proposing a similar technique for estimating the parameters of the gamma distribution for areal rainfall.

In this chapter, an alternative approach to estimating dry-day probability in an areal average is proposed and tested. **Section 3.2** describes the three station rainfall datasets for the UK, China and Zimbabwe that are used to develop and test the techniques. **Section 3.3** works through the methodology proposed for estimating the dry-day probability for an average of n stations, describing the concept of the effective number of independent samples (n') and the use of a novel approach to estimating the spatial dependence of wet/dry day coincidences. The effectiveness of this approach is illustrated by applying it to data from three regions with different climatic regimes – UK, China and Zimbabwe, and the uncertainties associated with these estimates are explored. The expansion of this method to estimate the 'true' areal mean (i.e. the average of an infinite number of stations in a grid box) is demonstrated in **Section 3.4,** and additional uncertainties associated with applying the methodology to regions with sparse data coverage are explored. The methodology is summarized in the final section (**Section 3.5**) and its significance and potential applications are discussed.

3.2. Study Regions and Station Data

The relationships between point and spatially averaged rainfall are investigated using three station datasets from the UK, China and Zimbabwe. The stations within each dataset vary in their temporal coverage, but for the majority of the analysis in this thesis, only data from the period 1961 to 1990 is used. While this period does not necessarily have the most complete coverage from the available stations, it is consistent with the 'control' period used in climate model experiments and is widely used as the reference climatology for 'current climate' in studies of observed climate change. It is therefore convenient to retain this reference period.

The UK dataset consists of 170 stations over England, Scotland, Northern Ireland and Wales (Figure 3-1) that are available from the British Atmospheric Data Centre (BADC). These stations were selected for their length and completeness and have been used, and added to, for a series of studies of UK rainfall variability. A set of 110 of these stations were used by Osborn and Hulme (1997, 1998) in their study of areal and point rainfall variability (as detailed in Section 2.3.1.4) and also in Osborn *et al*. (2000) in a study of observed trends in UK precipitation intensity. An additional 36 stations were added before the set was used again by Osborn and Hulme (2002), improving the coverage to include the Channel Islands, Isles of Scilly and some of the Scottish Islands. Many stations from the original dataset were extended at this time to cover up to spring 2001. A further 24 stations have since been added.

The datasets for China and Zimbabwe were also used by Osborn and Hulme (1997). From China, 180 stations, obtained from the Oak Ridge National Laboratory in the United States, are used and span the period 1951-1982 (Figure 3-2). Finally, the Zimbabwean dataset contains 51 stations obtained from the Zimbabwean meteorological service, spanning 1951-1990 (Figure 3-3). The temporal coverage of all three datasets is shown in Figure 3-4.

These three datasets were selected partially for the following reasons;

(1) The data are available at daily temporal resolution
(2) The data have already been through a quality control procedure and are therefore relatively reliable. In this study the observations are not being used for trend detection, and therefore

54

the inhomogeneity often found in such datasets (due, for example, to changes is recording techniques or movement of stations) should not affect the data adversely for this application, and so further homogenisation procedures are not necessary.

(3) The three regions represent a range of precipitation regimes across the three regions Figure 3-5). This will allow any relationships or techniques used to be tested against a range of different climate types.
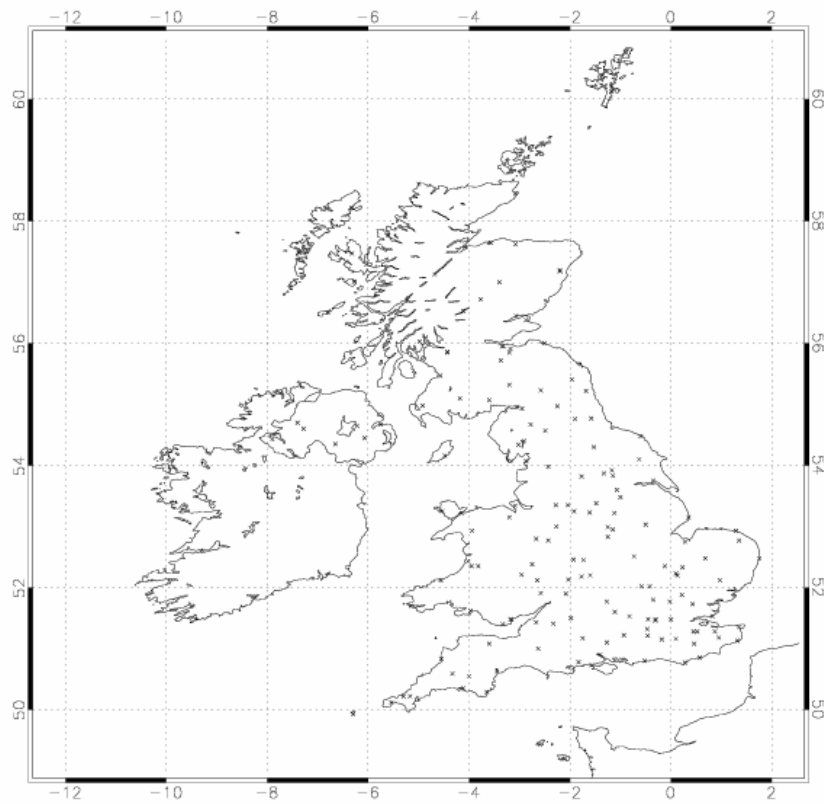

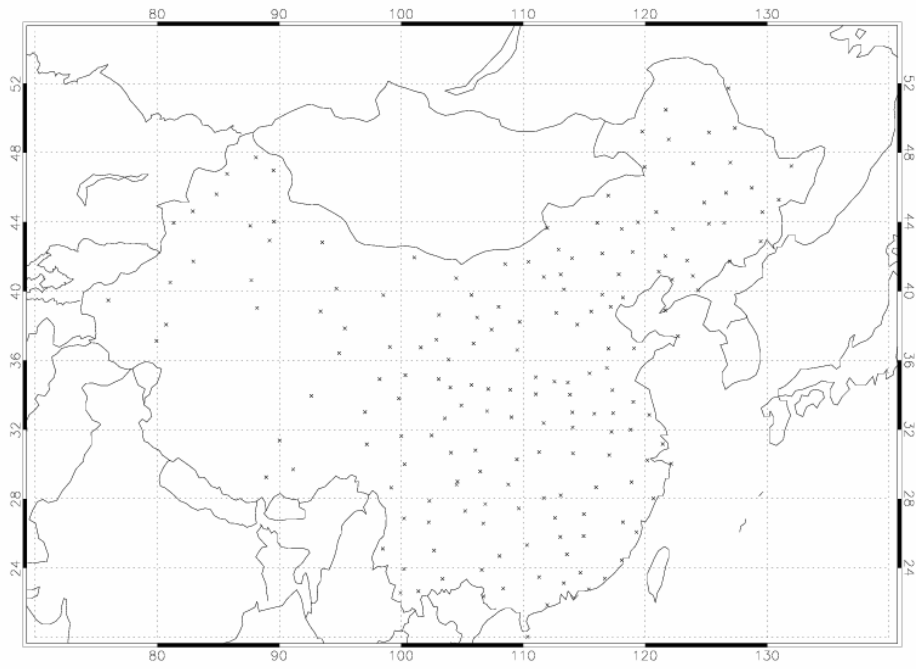**Figure 3-1: Locations of stations in UK dataset (170 stations)**

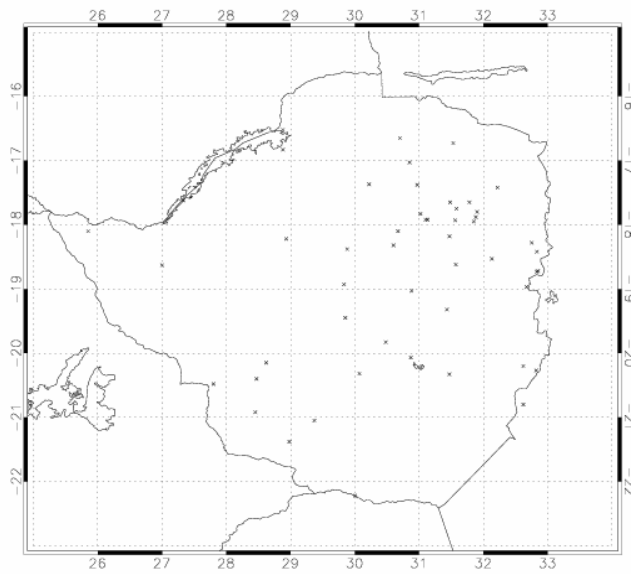**Figure 3-2: Locations of stations in Chinese dataset (180 stations)**

Figure 3-3: Locations of stations in Zimbabwean dataset (51 stations)
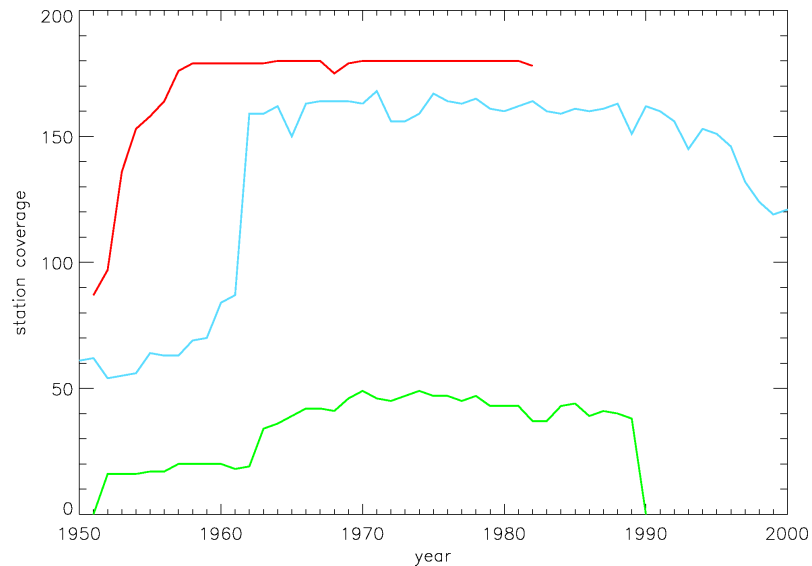
56

Figure 3-4: Number of stations with 70% or more complete data for the three regions, UK (blue), China (red) and Zimbabwe (green).

The simple monthly precipitation statistics presented in Figure 3-5 demonstrate the difference in precipitation regimes experiences in these three regions. The UK experiences a temperate climate which is strongly influenced by South Westerly prevailing winds, which bring a moisture, and thus precipitation, throughout the year. Zimbabwe and China experience more distinct dry and wet seasons in their climates. China's average rainfall shows a distinct wet season, caused largely by summer monsoon rainfalls, affecting the southern parts of the country. As China is a large country, it encompasses a range of topography and climate regimes. This is illustrated in Figure 3-6, in a precipitation climatology for China by Ye *et al*. (2004), which shows a clear contrast between the regimes of the dry, low-lying plains of north-west china which receive less than 50mm per year and the considerably wetter of the coastal south-east which receive more than 1500mm per year. Precipitation statistics at two of the Chinese stations (Figure 3-7) are shown in Figure 3-8 to highlight this contrast in regime in the different regions of the country.

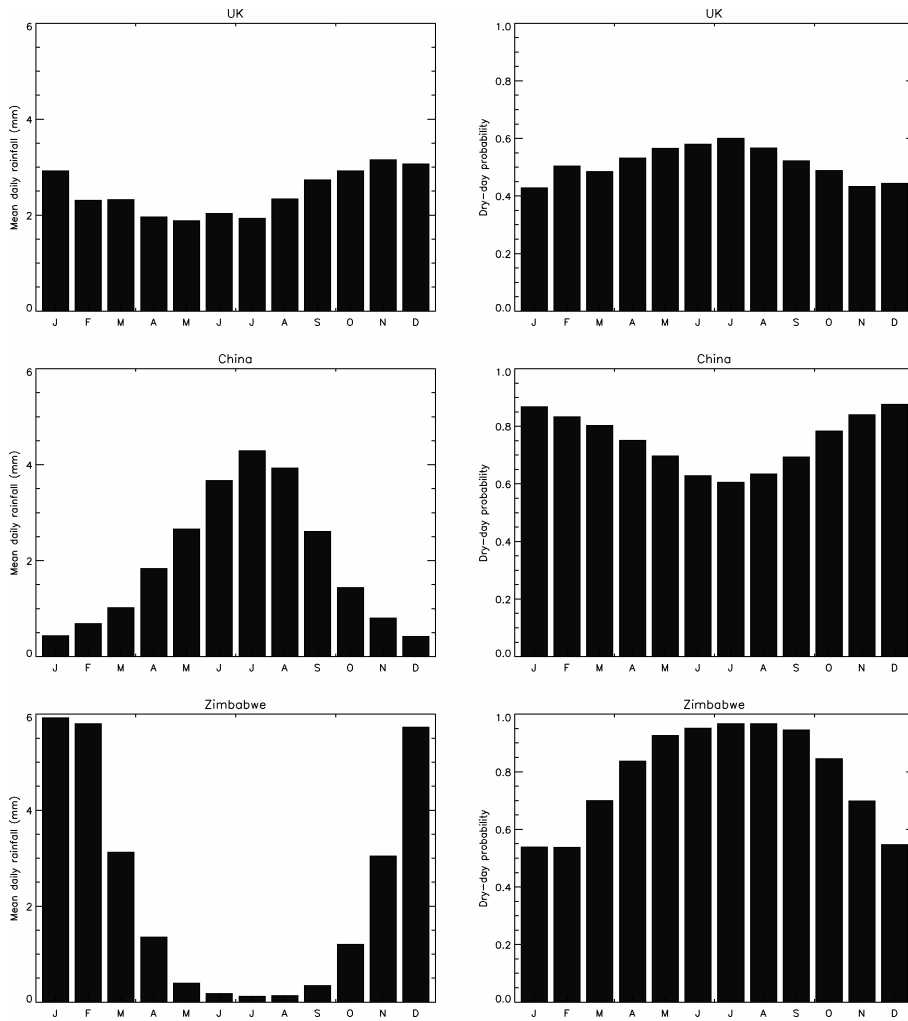**Figure 3-5: Mean daily rainfall and dry-day probability averaged over available stations for the 3 datasets (UK, China and Zimbabwe).**
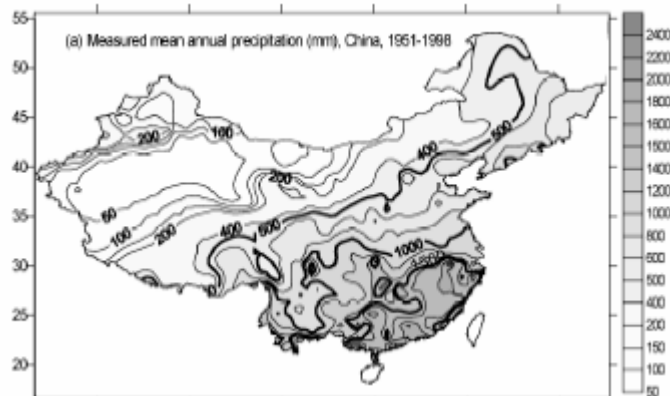

Figure 3-6: Contour map of mean annual precipitation over China, 1951-1998 (Ye *et al*., 2004)
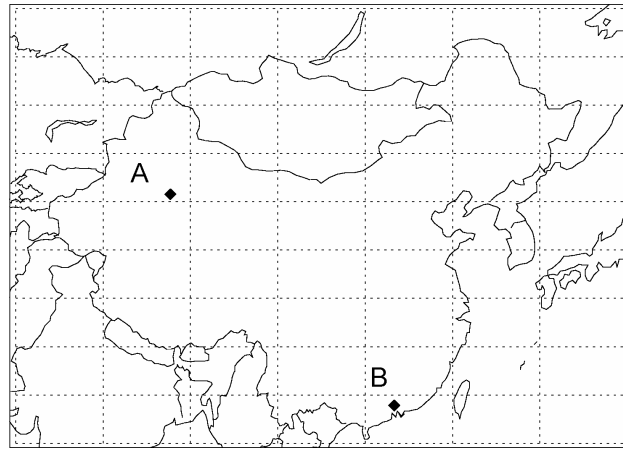
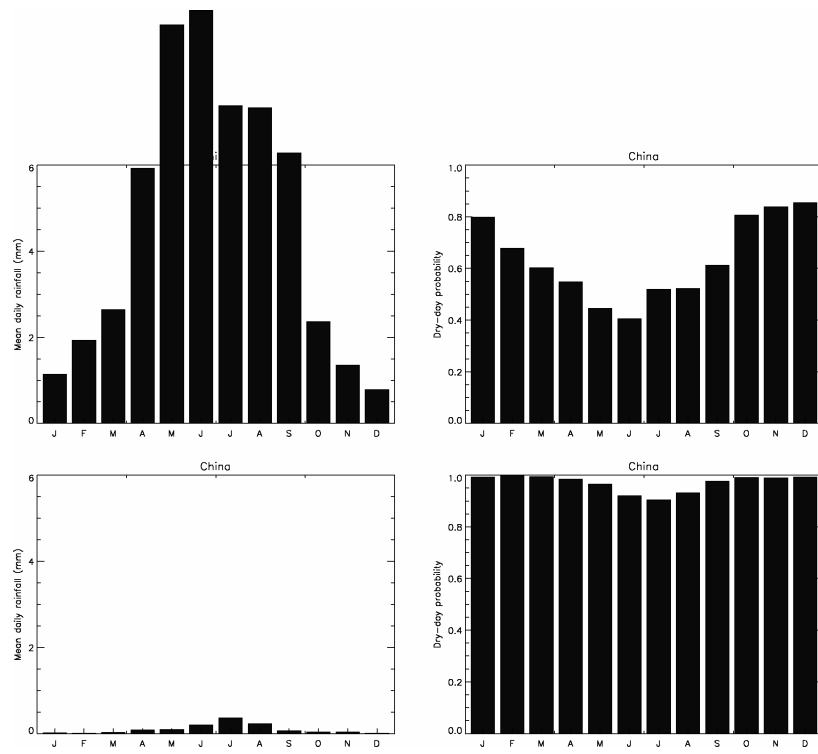**Figure 3-7: Locations of Chinese precipitation stations, A and B.**

Figure 3-8: Mean daily rainfall and dry-day probability at Chinese stations A and B.

3.3. Development of Methodology

3.3.1.   Selection of Wet-Day/Dry-Day Threshold

In selecting an appropriate wet/dry day threshold value, it is important to select a value which is low enough that only the minimal amount of recorded rainfall is lost, but also that the value is high enough to allow for the level of precision offered by rain gauges.

The observations of daily rainfall amounts given by the three station datasets are given to 0.1 mm. However, this precision may not reflect the level of precision offered by the measurement equipment. The UK data spans periods when rainfall has been measured in imperial and metric quantities. When measured in inches, the minimum recordable value is 0.1″, equivalent to 0.254 mm. The threshold of 0.3mm therefore encompasses the uncertainty introduced by the conversion between imperial and metric. Throughout this study a wet-day threshold of 0.3 is used, such that any value greater than or equal to 0.3 is classed as wet and any value of 0.2 or less is considered 'dry'.

Throughout this study, we refer to the frequency of dry-days, rather than the frequency of wet-days. These values are easily inter-converted, and whilst wet-day frequency is more commonly referred to in a climatological studies (e.g. Osborn and Hulme, 1997, 1998; New *et al.*, 1999; Frei *et al*., 2003; Mitchell and Jones, 2005), dry-day frequency is a more convenient quantity to deal with for the probability-based calculations which will be made in this study. In order to maintain consistency, the calculated dry-day probabilities are not converted to wet-day frequencies.

3.3.2. Dry-Day Probability in an *n*-Station-Mean Series and the Effective Number of Independent Stations (*n'*)

Every station rainfall record is a point observation of rainfall. At each point in space, it can be expected that there will always be some dry days (although in very wet regions/seasons there may be very few), and hence the dry-day probability, *P(d)*, will always be greater than zero. However, when rainfall over an area is considered, the average rainfall over that area will experience fewer dry-days than at each point within the area. The larger the area considered, the more likely it is that it will be raining somewhere in that area, and hence as spatial scale increases, the dry-day probability decreases. At very large spatial scales, such as global or hemispheric means, dry-day probability will approach zero, as it is highly unlikely that on any day it will not rain anywhere at all.

If the rainfall at every measured point , *i* to *n*, in an area of any size, were independent of the rainfall measured at any other point, the dry-day probability in the mean of all *n* measured points ($P(d)_{\overline{i,n}}$) would be:

$$P(d)_{\overline{i,n}} = P(d)_1 \times P(d)_2 \times P(d)_3 ... P(d)_n,$$

Equation 3-1

or:

$$P(d)_{\overline{i,n}} = \left[geometric\_mean(P(d)_{i,n}) \right]^n.$$

Equation 3-2

Using this algorithm, if every available point in a region were considered (*n=∞*), and all stations are entirely independent of one another, then the areal mean must theoretically have a dry-day probability of zero. Of course, in reality, precipitation is not spatially independent, at least not at the spatial scales at which GCMs operate. The atmosphere is fluid and circulating, such that the weather experienced at one site is always related in some physical way to that which is

61

experienced at nearby sites. Precipitation occurs in spatially coherent weather systems, such that stations that are very close together are likely to experience similar rainfall. This spatial dependence decreases the further apart the stations are positioned, and the relationship between stations separation distance and their level of dependence can be characterised using an exponential decay function, or 'correlation decay curve' (e.g. Osborn and Hulme, 1997; Wilby *et al.*, 2003). At large separation distances, the correlation can even become negative as it cannot be raining everywhere at one time because the atmospheric circulation requires areas of low pressure, where rainfall tends to occur, to be balanced by regions of high pressure, where rainfall is unlikely.

The dry-day probability in an areal average for a region is therefore determined not only by *P(d)* at the individual points within the area, but also by the level of spatial dependence between rainfall occurrence at points in the area. The degree of spatial dependence is determined by a number of factors, primarily the size of the area, but also by climate type, which varies geographically and seasonally, and with physical characteristics such as mean elevation, topography and orography.

The algorithm given above (Equation 3-2), for independent stations, does not necessarily become redundant when the dependence between points is considered. The *effective* number of independent stations can be calculated for any set of n stations, based on the mean inter-station correlation between available stations (Osborn *et al.*, 1997). The concept of the 'effective number of independent samples', '*effective n*', uses the theoretical relationship between variance, mean inter-station correlation ($\bar{r}$) and the number of observation points (n) to estimate the effective number of independent stations (n') that would give an average series equivalent to the available n stations. This approach has been used in a number of statistical applications where differences in sample size cause variability in sample averages to be incomparable. In tree-ring research, for example, varying sample size at different time periods in a chronology can cause deceptive trends in variability. Shiyatov *et al.* (1990) and Osborn *et al.* (1997) are examples of the application of the adjustment of variance in tree-ring chronologies according to sample size.

The basis for the calculation of n' is the theoretical relationship between variance, mean inter-station correlation ($\bar{r}$) and the sample size (n) in a set of n point observations,

$$S_x^2 = \overline{s^2}\left[\frac{1+(n-1)\bar{r}}{n}\right],$$

Equation 3-3

where $\overline{s^2}$ is the mean of the variances of the individual timeseries and $\bar{r}$ is the mean correlation between all pairs of observations (Osborn *et al.*, 1997).

If $\bar{r} = 0$ (i.e. all n series are independent of each other and are termed n'), this equation reduces to:

$$S_X{}^2 = \frac{\overline{s^2}}{n'},$$

Equation 3-4

which can be substituted back into Equation 3-2 and reduced to give :

$$n' = \frac{n}{1+(n-1)\bar{r}}\,.$$

Equation 3-5

In words, therefore, n' is equal to the number of independent series (stations) that would reduce the variance of the average series by the same extent that it is reduced by averaging the n correlated series.

This is a very useful concept in the present context, if proved valid, because it would allow the use of theoretical probabilistic relationships between independent time series in Equation 3-2.

Thus, when n' is substituted for n, Equation 3-2 becomes:

$$P(d)_n = (\overline{P(d)_{i,n}})^{n'}$$

63

Equation 3-6

where, $\overline{P(d)}_{i,n}$ is the geometric mean of the dry-day probabilities for stations i to n. This geometric mean can be approximated by the arithmetic mean providing the dry-day probability does not differ significantly between stations.

3.3.3. Spatial Dependence Between Dry-Day Occurrences

The application of n' to the variance of average time series is based on the concept that the decline in variance in an n-station average series compared to the variance in the point observations is determined by the inter-station Pearson correlation co-efficient ($\bar{r}$). In applying the 'effective n' concept to the estimation of dry-day probability of an n-station average precipitation series, it is assumed that the decline in dry-day probability is also dependent on some measure of the inter-station precipitation dependence. This assumption is tested in this chapter by attempting to derive functions that relate $P(d)_n$ to $\overline{P(d)}_{i,n}$, n and a measure of inter-station dependence. If such a relationship can be identified then this assumption is supported.

The inter-station correlation, which quantifies how similarly the rainfall *amounts* vary, may not be the most appropriate measure of inter-station dependence for this application. Instead a measure of the similarity in how wet/dry day *occurrences* vary is needed. It may be expected that, in many cases, the level of dependence between stations is likely to be higher if only wet-day/dry-day occurrence is considered than if the actual amount is considered. For example, if on a single day one of two stations (with otherwise similar daily rainfall) receives a very high rainfall, say 80mm, but another receives only 5mm, then the overall correlation might be skewed towards a lower value, because the Pearson correlation co-efficient is particularly sensitive to the agreement or disagreement of the largest anomalies in the time series. However, these days would both be counted as 'wet' despite the fact that one is considerably wetter than the other, and would contribute to a higher value of a measure of wet/dry day occurrence towards a higher value.

What is required, therefore, is a measure of the spatial dependence which quantifies the strength of the association between wet and dry day occurrence on a scale of -1 to 1 such that this value can be substituted for *r* in the calculation of 'effective *n*'.

Osborn and Hulme (1997) used a 'probability decay curve', in a similar way to the correlation decay curve, in order to quantify the spatial dependence in wet-or-dry-day occurrence between stations. Osborn and Hulme (1997) identified that for any pair of stations, a maximum and minimum *P(d)* for a two-station average series can be determined given their individual *P(d)* values.

The minimum possible *P(d)* value for an average of two stations occurs if it is assumed that dry-or-wet-day occurrences at the two stations are totally independent of one another.

$$Minimum\ P(d)_2 = [P(d)_1]^2$$

The maximum possible *P(d)* value for an average of 2 stations occurs if the wet-or-dry-day occurrence is identical at the two stations (i.e that the stations are totally dependent).

$$Maximum\ P(d)_2 = P(d)_1$$

These maximum and minimum values can be used to place the actual *P(d)* for the average of those two stations, *P(d)₂*, on a scale of 0 to 1 according to the relationship between the single and average station *P(d)* values.

$$'r(w/d)' = \frac{Actual - Minimum}{Maximum - Minimum}$$

or

$$'r(w/d)' = \frac{P(d)_2 - \left(\overline{P(d)_1}\right)^2}{P(d)_1 - \left(\overline{P(d)_1}\right)^2}$$

Equation 3-7

This value is termed '*r(w/d)*', as it quantifies the degree of dependence in dry-day occurrences at a pair of stations, and will be used in place of *r*, Pearson's product-moment correlation co-efficient, which quantifies the degree to which the precipitation amounts are dependent between two stations.

The minimum and maximum values used assume that the *P(d)* is equal at both stations. If the stations are truly totally dependent, then this is necessarily the case, but for two stations which are only partially dependent, this may not be so. For this reason, the mean of the *P(d)* values for the two stations is used as an approximation, but where a significant difference occurs between these values, negative values of *r(w/d)* can arise. Negative values can also arise if the actual number of coincident dry-days is smaller than the 'minimum', reflecting an anti-correlation between stations.

The values of *r(w/d)* compared to *r* (when *r* is based on all values in both series, including dry days) are compared for random station pairs from all three datasets in Figure 3-9, Figure 3-10 and Figure 3-11 . In most seasons the *r(w/d)* values tend to be a little higher than the *r* values, particularly in summer. However, these comparisons show that *r(w/d)* values are only higher for cases where correlation is less than 0.7. For more strongly correlated station pairs, the *r(w/d)* correlation is lower than standard *r*.

The reasons for these patterns can be explored by considering the different correlation that arises in specific examples. Hypothetically, a pair of stations might record values in Table 3-1. For most station pairs and seasons, *r(w/d)* is greater than *r* because two stations often record values of very different magnitudes, e.g 5mm and 50 mm (Case D in Table 3-1), which will both register 'wet', and therefore contribute to a higher '*r(w/d)*' than '*r*'.

However, in some situations, Case C will occur more often. In this situation, two stations can record very similar values, for example, 0.2mm and 0.5mm, but one will register 'dry' and the other 'wet' such that the values contribute towards a lower '*r(w/d)*' than '*r*'. This situation is most likely to arise frequently between station pairs that are close together, and normally correlate highly, particularly for dry regions and seasons.

	Station 1	Station 2	*r*	*r(w/d)*
Case A	0.5mm (Wet)	0.8mm (Wet)	High	High
Case B	0.2mm (Dry)	20mm (Wet)	Low	Low
Case C	0.2mm (Dry)	0.5mm (Wet)	High	Low
Case D	5mm (Wet)	50mm (Wet)	Low	High

Table 3-1: Hypothetical correlative properties of station pairs for different daily rainfall measurements, when different measures of correlation are used.

The exceptions to this pattern are the winter seasons (DJF in UK and China, JJA in Zimbabwe). This may be because rainfall in winter is caused more often by large-scale, synoptic weather systems. This means that fewer incidences of Case D or Case C occur as rainfall events tend to blanket a larger geographical area with relatively heavy rainfall.
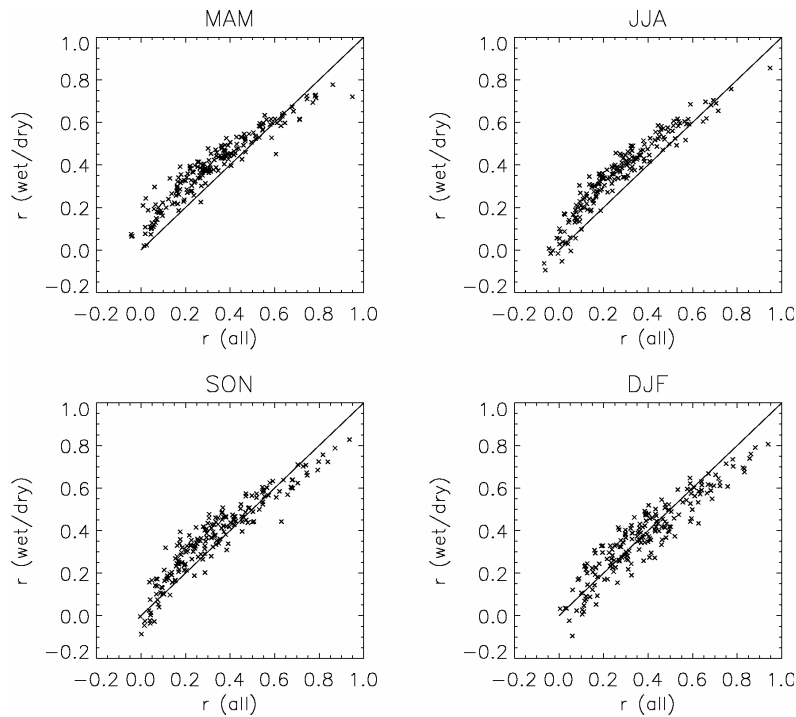

Figure 3-9: Comparison of *r* with *r(w/d)* for randomly selected station pairs from the UK data set.
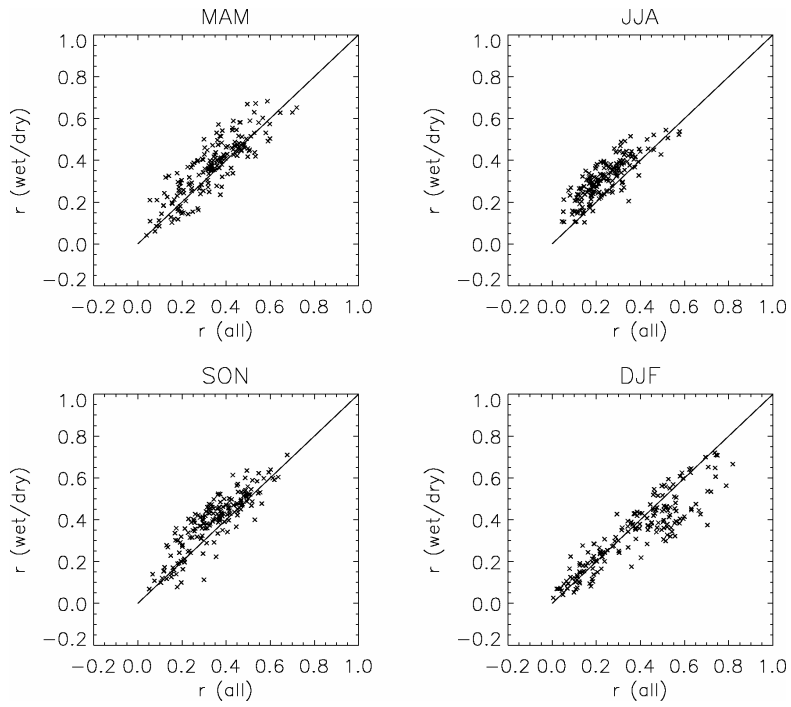
Figure 3-10: Comparison of *r* with *r(w/d)* for randomly selected station pairs from the Chinese data set.
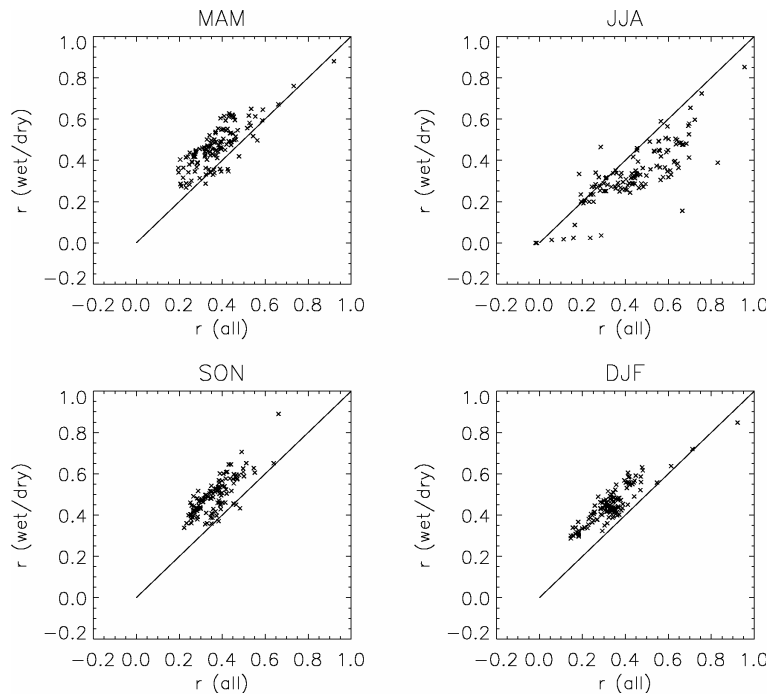

Figure 3-11: Comparison of r with r(w/d) for randomly selected station pairs from the Zimbabwean data set.

68

3.4. Application to Data

3.4.1.   Estimating Dry-Day Probability in an Average of *n* Stations

It is therefore proposed that the dry-day probability for an *n* station mean series can be estimated by:

$$P(d)_n = (\overline{P(d)_{i,n}})^{n'}$$

Equation 3-8

where: $n' = \dfrac{n}{1+(n-1)\overline{r(w/d)}}$

Equation 3-9

and: $r(w/d) = \dfrac{P(d)_2 - \left(\overline{P(d)_1}\right)^2}{P(d)_1 - \left(\overline{P(d)_1}\right)^2}$

Equation 3-10

This proposed methodology is tested against data from UK, Chinese and Zimbabwean datasets. Random combinations of *n* stations are chosen from the datasets by selecting grid boxes of random position and size from each dataset. For those grid boxes which contain 3 or more stations, we calculate an averages series using the stations within that box, and calculate the $P(d)_n$ directly from the average series [*Actual P(d)n*] for each season.

In order to ensure that the *n*-station average series is a truly an *n*-station average, only those days where all stations have recorded values are included, thus any missing data days which arise in any of the stations become missing data days in the mean series also. While this means that in some cases, the useable data is limited to a short period of total coverage, this avoids inconsistency in the number of stations included on any day. An arbitrary value of 30 days is used as the minimum required for any season in order to calculate a value of *r(w/d)*,

The mean $r(w/d)$ is then calculated for the n-stations, using all combinations of station pairs, and is used to calculate n' for those stations, and estimate $P(d)_n$, for each season [*Estimated P(d)$_n$*]. The results are shown in Figure 3-12 .
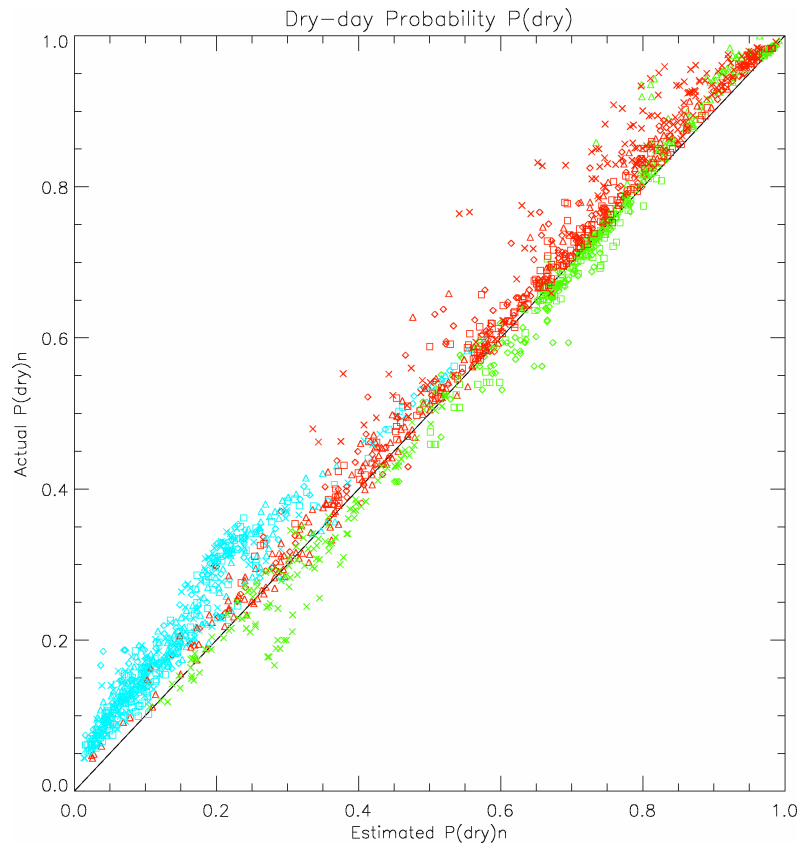

Figure 3-12: Estimated *P(d)$_n$* versus Actual *P(d)$_n$* for random n-station mean series from UK (Blue), Zimbabwe (Green) and Chinese (Red) station data.

The estimates are reasonable, particularly for the Chinese and Zimbabwean data, whilst for the UK, there is a systematic underestimation of *P(d)* of up to 0.1 compared to the actual value in the *n*-station average. Overall, this relatively good performance suggests that the assumption that

$\overline{P(d)_{i,n}}$ and $P(d)_n$ are related via a measure of the spatial dependence between stations, and that

the measure of wet/dry day coincidence, described in 3.3.3, is appropriate measure of spatial dependence in this application.

3.4.2.   Uncertainty in Estimates of Dry-Day Probability for an *n*-Station Average

As the analysis is limited to only three regions of the world, it is difficult to apply absolute confidence limits to the estimates made using the above methodology. However, as a guide to the level of accuracy that this approach offers, uncertainty bounds have been applied which encompass 95% of the points (Figure 3-13). These bounds are positioned at **+0.10** and **-0.03** relative to the line *x=y*. Repeating the random selection of grid boxes points three times indicates that this is a robust estimate for these datasets. It cannot be guaranteed that these bounds would not alter if data from an additional region were added, hence they offer only a guide to the level of certainty afforded by this technique.
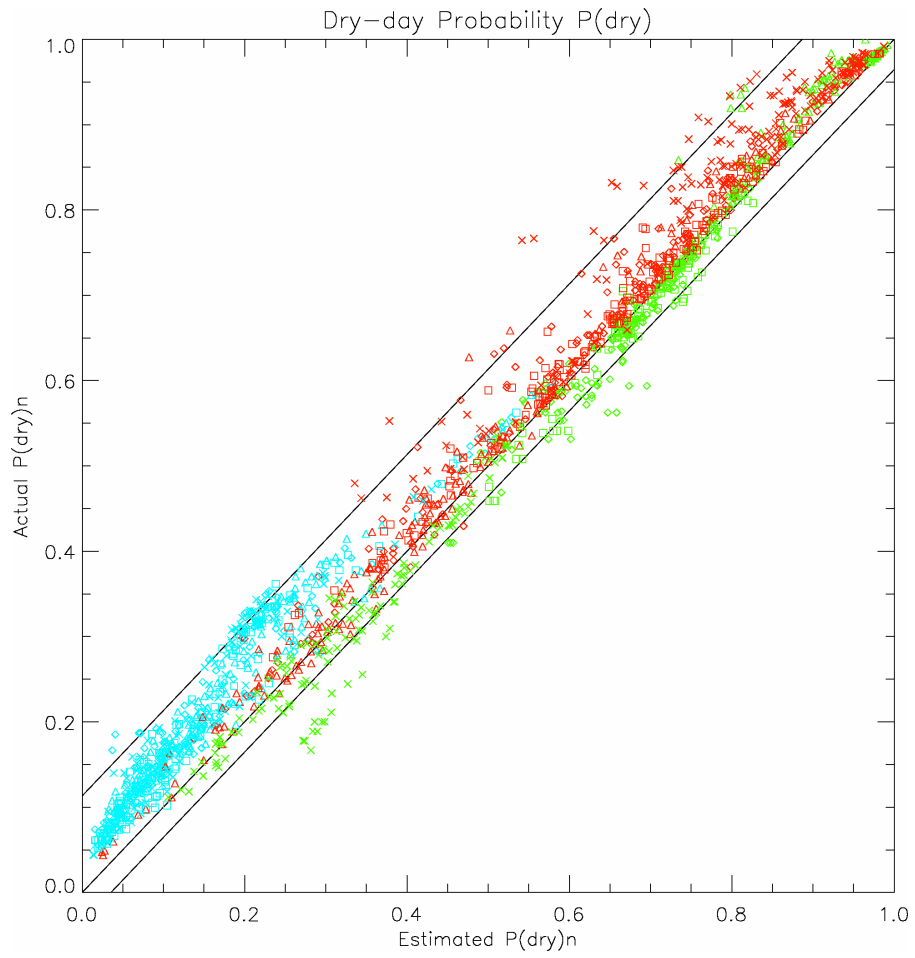
Figure 3-13: Estimated P(d) versus Actual P(d) for random n-station mean series from UK (Blue), Zimbabwe (Green) and Chinese (Red) station data with 95% confidence limits.

3.5. Application of Methodology to the Estimation of Dry-Day Probability in a 'True' Grid-Box Mean

It has been demonstrated in Section 3.3 that *n'* and *r(w/d)* can be used to estimate the dry-day probability in an average of a number of daily rainfall timeseries. This section details the expansion of this approach to estimate dry-day probability for a 'true' grid-box mean, using the available stations in a grid box (*n* stations) to estimate the dry-day probability which would be expected in the grid-box areal mean if an infinite number of stations (*N* stations) from the grid box were averaged. The method is illustrated using a sample grid box from the UK and the additional uncertainty introduced by this extension of the method is investigated.

The example grid box from the UK, used to illustrate the method, is selected with dimensions that are realistic for GCM resolution (2.5° x 3.75°) and a location that maximizes station coverage (Figure 3-14). This region of dense station coverage (58 stations) allows exploration of how much the estimates vary when smaller subsets of *n* stations are used to estimate $P(d)_N$, and thus gives some guidance as to how reliable estimates might be when based on smaller numbers of available stations (*n*).
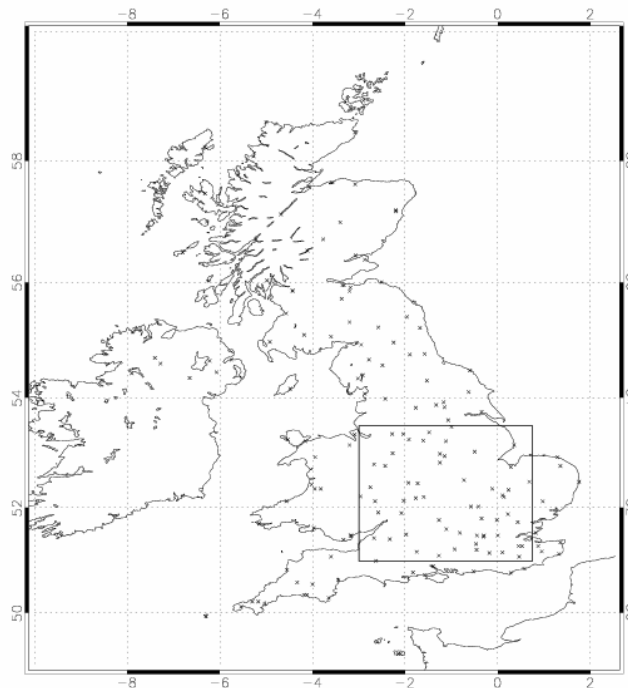

Figure 3-14: Example UK grid box: 3.75 x 2.5 degree, selected to give maximum station coverage.

3.5.1. Extension of Methodology to Estimation of Dry-Day Probability in the 'True' Grid-Box Mean (*N* stations)

If the value *n*, the number of real stations available, is replaced by *N* (any number of stations that is greater than the *n* available stations), Equations 4.8 and 4.9 become:

$$P(d)_N = [\overline{P(d)_{i,N}}]^{N'}$$

Equation 3-11

and $N' = \dfrac{N}{1+(N-1)\overline{r(w/d)}}$

Equation 3-12

In this case, the value of *N* for which we wish to make estimates for is infinity, as this number of stations give complete areal coverage and thus when averaged would give the true areal mean for a grid box. The temporal variability in an *n*-station average timeseries decreases hyperbolically as a greater number of series', *n*, are included, such that it levels off at a value which is representative of the 'true areal mean' at some value of *n* which is smaller than infinity (Osborn and Hulme, 1997). This means that *P(d)* estimated for a value of *N* considerably smaller than infinity could be used, e.g. 1000, to give a realistic areal mean value. However, infinity is actually a convenient value to use in this case because in the case that $n = \infty$, Equation 3-5 reduces to:

$$n' = \dfrac{1}{\overline{r(w/d)}}$$

Equation 3-13

When dealing with values of *n*, the associated values of $\overline{r(w/d)}$ and $\overline{P(d)}_{i,n}$ are computed directly from the available *n* stations. When applied to the case of *N*, it is now necessary to estimate the values $\overline{r(w/d)}$ and $\overline{P(d)}_{i,N}$ for all the hypothetical *N* stations, for which no real observations exist. These values, therefore, have to be estimated based on the characteristics of the available *n* stations, which imposes an additional source of uncertainty of the estimates of $P(d)_N$ instead of $P(d)_n$. The number, *n*, of available stations, and their distribution within the grid box determines how representative they are of the whole grid box, and therefore how reliable those estimates will be.

The following sections address the estimation of the values $\overline{r(w/d)}$ and $\overline{P(d)}_{i,N}$ for *N* stations, and explores the additional uncertainty which relates to the particular sample of *n* stations that is used in areal $P(d)_N$ estimation. This uncertainty is explored using the sample grid box (Figure 3-14) which is relatively densely gauged. This provides the best available estimates of 'true' grid-box inter-station correlation, station dry-day probability and hence 'true' grid-box dry-day probability. This region is therefore used to experiment with smaller subsets of *n* stations, in order to explore the additional uncertainty that other grid boxes may be subject to if a smaller number of stations are available for analysis. Assuming that the 58 available stations provide a reliable 'true areal mean', several samples of *n* stations can then be randomly selected to demonstrate the range of $P(d)_N$ values that might have resulted if only those *n* stations were available

It should be stressed that the estimates of uncertainty derived here should be considered as *indicative* of the real uncertainty because this analysis is limited to regions where station data is sufficient to allow a number of station sub-sets to be sampled. The uncertainty that can be investigated and even quantified for this grid box may not be appropriate for another region which displays different characteristics. However, this is still an important stage in assessing the reliability of the estimates which are made and can give an indication of the magnitude of additional uncertainty that might be encountered in other regions.

75

4.1.1.1. Estimating Mean Inter-Station 'Correlation' for N Stations

The average of *r(w/d)* from the available *n* stations will differ from the required average over all *N* unobserved points, particularly if they are clustered and/or too few in number to sample the grid box adequately. For the most part, these sources of error cannot be removed (though their typical magnitude can be estimated, as here). One source of error that can be, at least partially, avoided is that which arises from non-uniform station distribution and the consequent bias in the estimate of $\overline{r(w/d)}$. This bias is partially predictable due to the dependence of correlation on spatial separation distance. This spatial dependence can be empirically modeled by fitting a correlation decay function to the real stations and then can be used to estimate $\overline{r(w/d)}$ for all pairs of hypothetical *N* stations.

Correlation decay curves have been used to model spatial dependence in a number of applications, allowing estimates of mean inter-station correlation to be made (e.g. Jones *et al.*, 1997; Osborn and Hulme, 1997; Booij, 2002a, 2002b). The method makes use of the relationship between separation distance and correlation for pairs of points, and uses the expected distribution of separation distances with a given area to estimate 'true' mean inter-station correlation. An advantage of this approach is that stations from outside of a given grid box can be used to fit the correlation decay curve.

A correlation decay curve is first determined for each station by fitting a chosen function to the scatter plot of *r(w/d)* 'correlation' versus separation distance, *d*, between that station and every other station in the dataset. The estimated parameters of the correlation decay curves obtained for each station contained within the grid box are then averaged to obtain parameters of the grid-box correlation decay curve that are appropriate for the grid box as a whole.

Each station within a grid box is fitted with a decay curve for each season separately, using the following exponential function.

r(w/d)=ae^{-bd}

Equation 3-14

A curve is fitted to each season separately in order to account for seasonal differences in the spatial scale of rainfall events. Unlike the correlation decay curves fitted by Osborn and Hulme (1997) and Booij (2002a, 2002b), the decay curve does not start at $r(w/d)=1$ for separation distances of 0. This discontinuity in the correlation decay function is referred to in geostatistics as the 'nugget' and occurs because even two very close stations will experience spatially uncorrelated 'noise' or measurement error which causes slight differences in their records that are sufficient to reduce the agreement between records (Webster and Oliver, 2001).

The 'nugget' which occurs in curves for $r(w/d)$ correlation decay are relatively large compared to those found in traditional geostatistical applications. The 'nugget' seen here occurs because of the low $r(w/d)$, compared to r, that is seen even in closely located station pairs, as discussed in 3.3.3 and seen in Figure 3-9, Figure 3-10 and Figure 3-11. It could be argued that at zero separation distance, the correlation would still be 1.0, but then drop off very steeply to the nugget value before settling into the fitted curve function. However, for the present purposes, the single function is sufficient for estimation of the mean grid-box inter-station correlation; given that it is only the integral under the decay curve that is required for this analysis.

The decay function (Equation 3-14) is fitted to the data using a least-squares approach. In order to improve the fit of the curve at the lower end of the *x*-axis (i.e. for small separation distances), and give a more reliable indication of the 'nugget' value, the closest stations (where $d < 100$km) are weighted by a factor of 2 relative to the more distant stations. This weighting results in a better fit at smaller separation distances, where otherwise $r(w/d)$ values tended to be under-estimated. It is less important, in the present application, to fit the tail of the curve accurately, because only values up to the maximum distance possible within a GCM model grid box will be used, typically 400km or less.
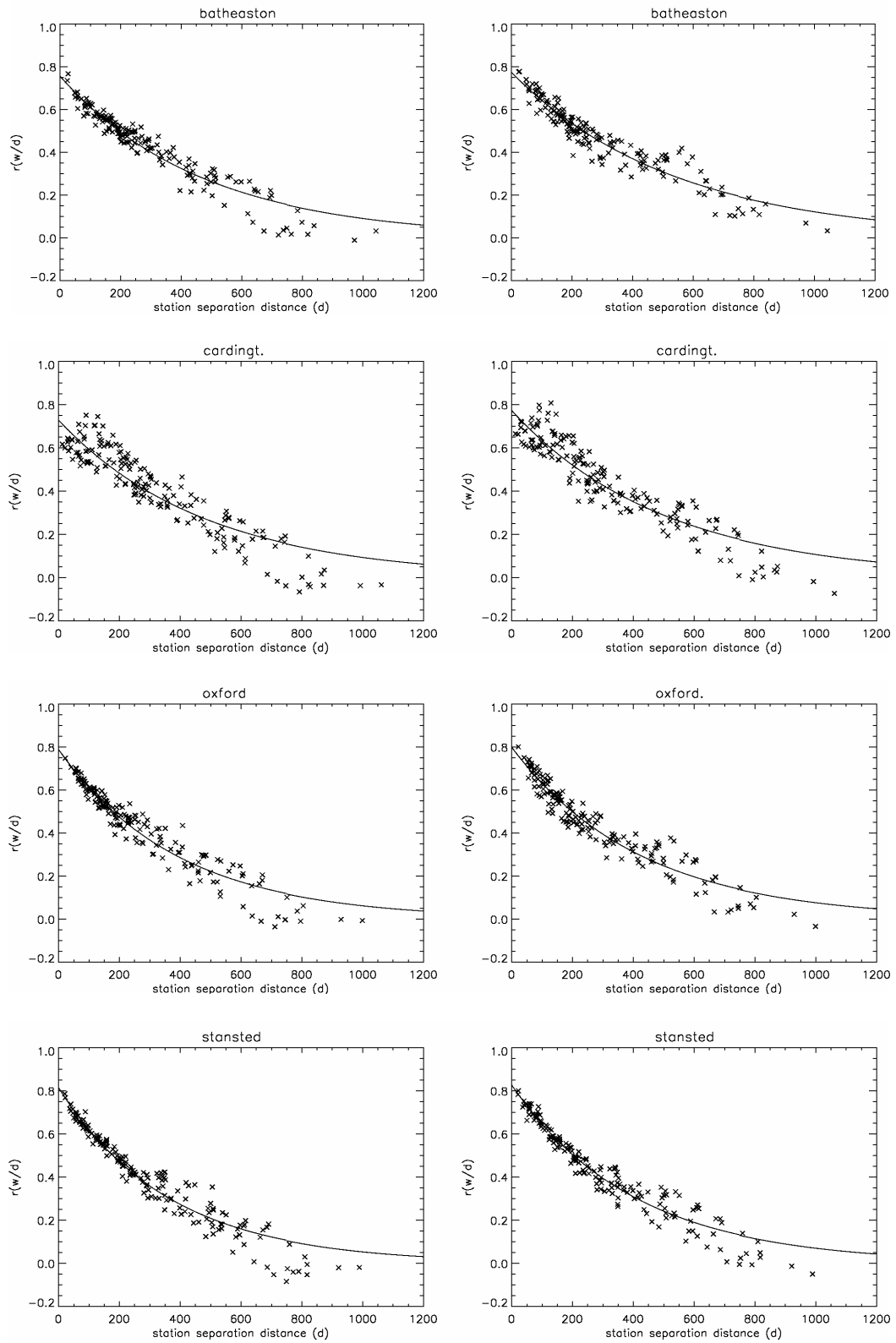
Figure 3-15: Correlation (*r(w/d)*) decay curves for 4 example stations (Batheaston, Cardington, Oxford and Stansted, for JJA (left) and DJF (right).

The curves for all *n* available stations in each grid box are averaged to give a grid-box correlation decay curve. The mean value of the parameter *a,* for the grid box, differs very slightly depending on whether the value is averaged as *a* or *1/a*, but this difference is negligible compared to the level of accuracy of the fitted function.

The grid-box average correlation decay curve is then used to estimate the mean inter-station correlation via a distribution of separation distances within the grid box. This distribution is estimated by selecting 5000 random pairs of points from the grid box and using the grid-box decay curve to estimate each of their expected 'correlations', *r(w/d)*, from which the arithmetic mean gives a mean inter-station *r(w/d)* for the grid box.
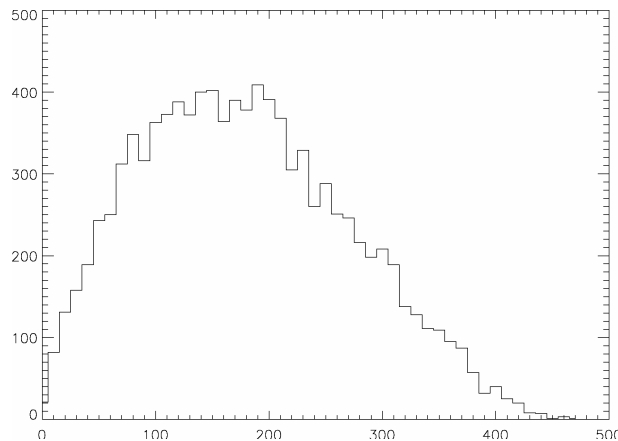

Figure 3-16: The distribution of separation distances between randomly selected pairs of points in a grid box (based on 10000 station pairs).

The mean inter-station correlations for the example grid box, which are estimated using the correlation decay curves, are given in Table 3-2, together with the value of 'effective n' for the 58 stations. The values of mean *r(w/d)* are similar throughout the year, except for Summer, when they are slightly lower at 0.55, reflecting the more convective nature of summer precipitation. The bracketed values are the *r(w/d)* and *n'* values calculated directly from all station pair combinations for the *n* available stations. These values are all similar to the values estimated for the grid box

average, because *n* here is relatively high, and those available stations are therefore a good representation of a fully sampled grid box.

	MAM	JJA	SON	DJF
r(w/d) for N stations	**0.58**	**0.55**	**0.58**	**0.58**
(mean r(w/d) from available n stations)	(0.53)	(0.56)	(0.56)	(0.60)
N'	**1.72**	**1.81**	**1.72**	**1.72**
(n')	(1.90)	(1.79)	(1.79)	(1.67)

Table 3-2: 'True' grid box mean *r(w/d)* and *N'* for an example UK grid box. The same values calculated, and given in brackets, directly from the n available stations for comparison.

In order to investigate the uncertainty range introduced by basing the estimates on smaller samples of available stations, up to 30 random combinations of *n* stations are sampled from the 58 stations available in the example grid box, for values of *n* between 3 and 50. For each sample, the mean inter-station correlation is then calculated by the same process: the two parameters of the curve from each of the *n* stations' correlation decay curves are averaged, and this curve used to estimate *r(w/d)* values for each of 5000 separation distances again. The resulting values are shown in (Figure 3-17) and span a range of around 0.05 above and below the 'best guess' value for the grid box for very small values of *n* (3-5), and this range falls to around +/- 0.02 for large values of *n*.
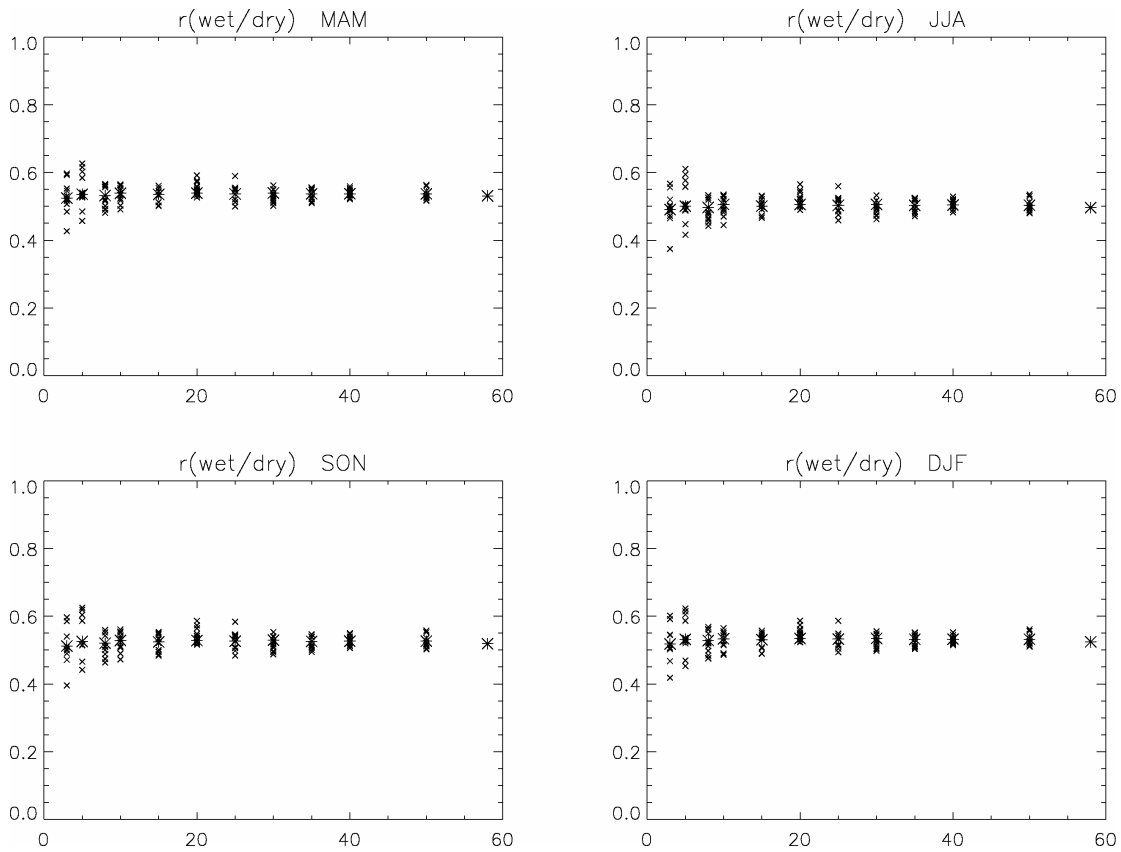
Figure 3-17: Variations in grid-box mean *r(w/d)* when based on random samples of *n* stations out of the available 58.

4.1.1.2. Estimating Dry-Day Probability for a 'True' Areal Mean

If the available stations are evenly distributed within a grid box, then it is reasonable to expect that $\overline{P(d)_{i,n}}$ at those stations may be close to the 'true' mean $\overline{P(d)_{i,N}}$ for the grid box. In the present application, it is assumed that this is the case.

It is important to bear in mind that, for some regions, coverage may well be biased towards lower lying and/or populated regions within a grid box. In areas where a substantial heterogeneity might exist across a grid box, it would be more appropriate to either:

- use a more advanced spatial averaging technique to estimate the mean station *P(d)* for the grid box (e.g. New *et al.*, 1999) interpolated station wet-day probability taking elevation differences into account via an empirically derived lapse rate),

or

- use only a subset of the available stations, selected to give a more even coverage.

The mean station *P(d)* from subsets of *n* stations are shown in Figure 3-18. There is little variation between mean station *P(d)* values for different samples, even for small values of *n* the *P(d)* values range over only around 0.05 (Figure 3-18). This error range may not be so narrow for other grid boxes if they have a greater spatial variation in dry-day probability across the grid box – a region with more varied elevation may well be more heterogeneous and therefore small samples would be more easily biased.
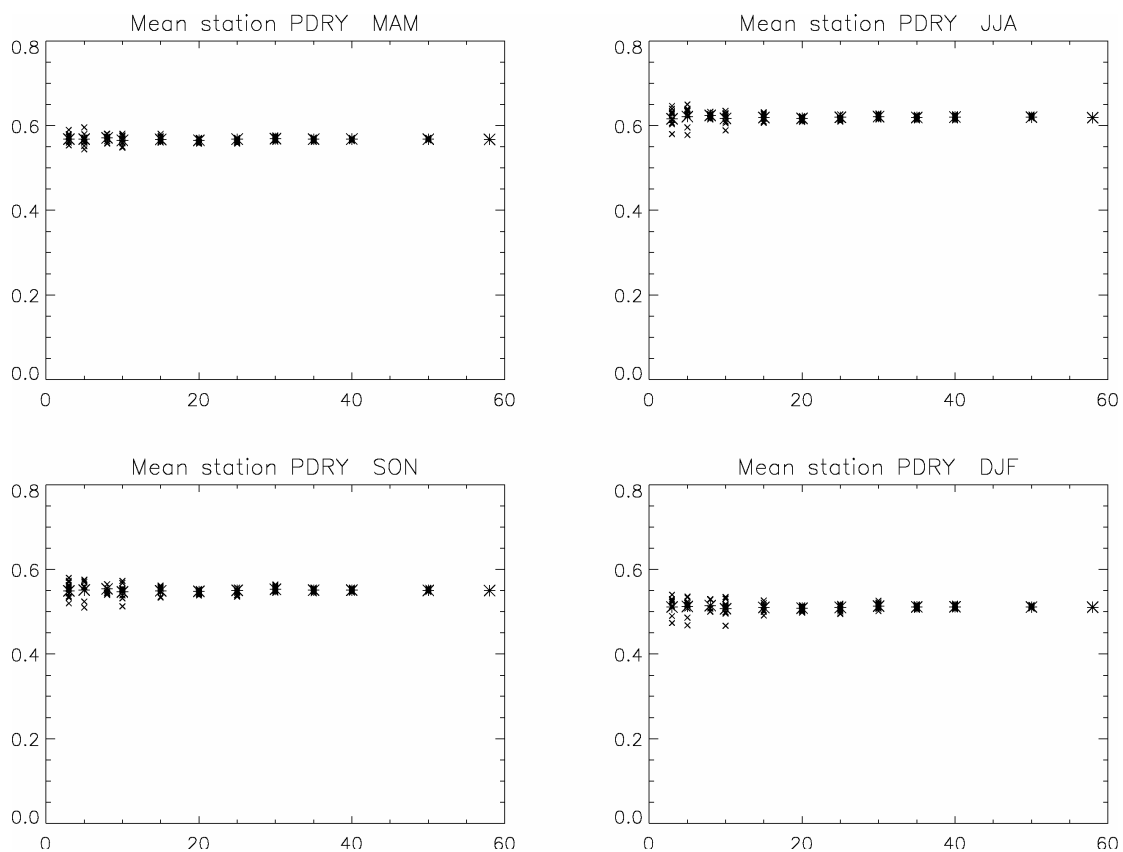

Figure 3-18: Variations in grid-box mean point *P(d)* when based on random samples of *n* stations out of the available 58.

82

3.5.2. Uncertainty in Estimates of Dry-Day Probability for a 'True' Areal Mean

The estimated values of $\overline{P(d)}_{i,N}$ and $\overline{r(w/d)}$ can now be used to make 'best guess' estimates of areal dry-day probability for the grid box, $P(d)_N$, (Equation 3-11 and Equation 3-13). Again, the example grid box is used here to investigate uncertainty associated with smaller numbers of available stations and this is shown in Figure 3-19. The span of resulting $P(d)_N$ values range from +/- 0.07 for n values less than 5, to +/-0.03 for n values up to 30. After n reaches around 30, there is little improvement in the accuracy of the estimates as n is increased. These ranges are generalised in Figure 3-19 using the boxes to indicate an uncertainty range of +/- 0.06 for n values of 3-10, and +/- 0.03 for n values 10-30. Values of n higher than thirty are all subject to a +/- 0.02 uncertainty range.
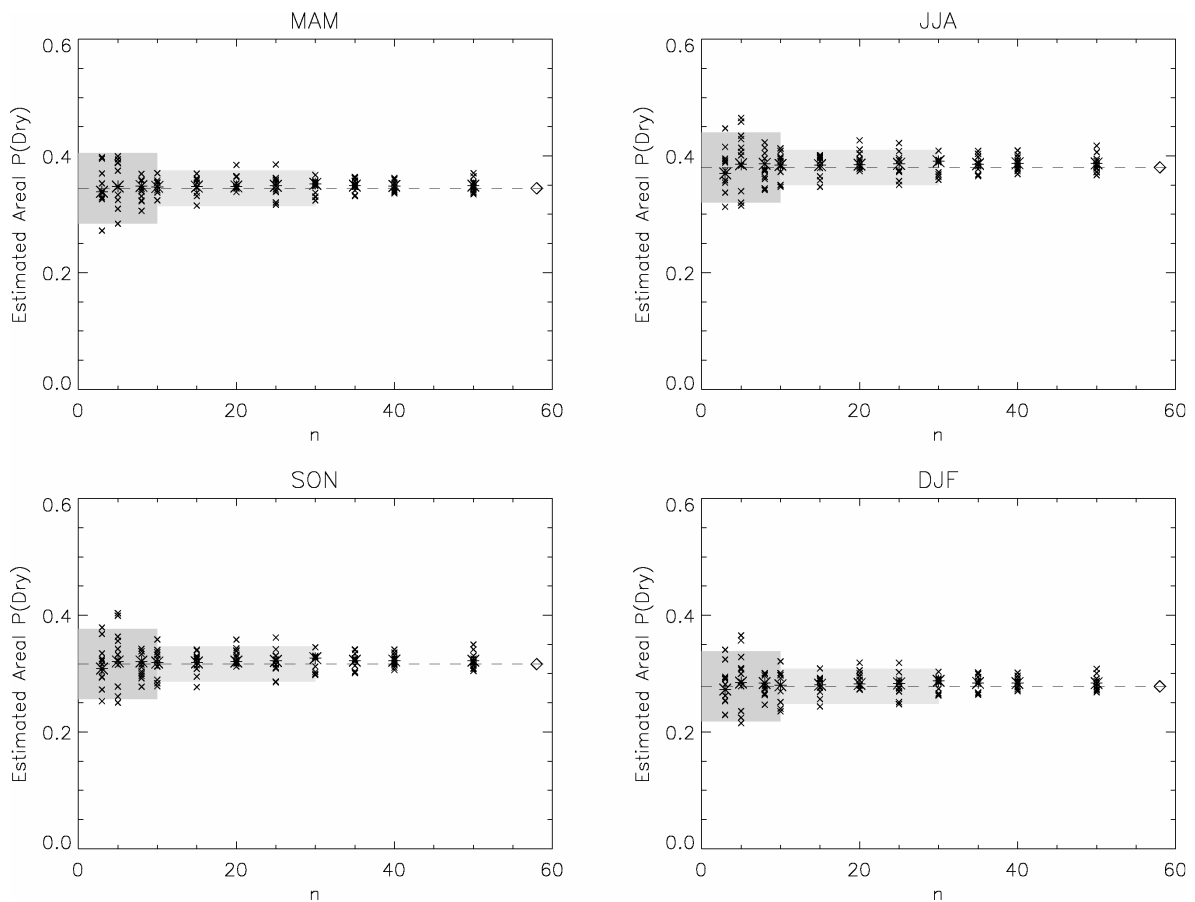

Figure 3-19: Variations in estimated 'true' areal mean *P(d)* when based on random samples of *n* stations out of the available 58.

Figure 3-20 illustrates the benefits of applying this methodology to estimate areal mean *P(d)* by showing the seasonal cycle of estimated 'true areal mean' dry-day probability for this UK grid box in comparison with values obtained from individual station values, or from analyzing an areal estimate that is calculated as the arithmetic mean of the available *n* stations. The 95% uncertainty bounds for the 'true' areal-mean estimate are indicated by the pale-blue shading indicate the level of uncertainty quantified in section 4.3.5 (Figure 3-13).
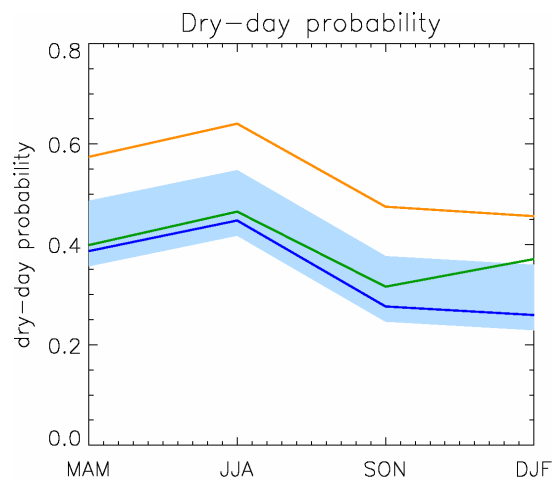

Figure 3-20: Seasonal dry-day probability for an example UK grid box calculated by 3 different methods: a) mean *P(d)* of 58 available stations in the region (yellow), (b) *P(d)* in areal average series constructed using the arithmetic mean of 58 stations (green), and (c) estimated true areal *P(d)* (blue) and the 95% confidence limits of this estimate.

The estimated areal values are around 0.2 lower than the equivalent station value (yellow) that might be used. The arithmetic areal mean of the 58 available stations (green) lies within the uncertainty bounds of the 'true' mean best guess (blue) for all seasons except winter, but still gives higher dry-day probabilities than the areal best-guess. In winter, the 58 station average appears to give the least good representation of the 'true' areal mean

3.6. Discussion and Conclusions

The method proposed here provides an approach to estimating dry-day probability of areal rainfall based on the dry-day probability and spatial dependence of available stations in the grid box. This approach uses the spatial dependence between the available stations to estimate the characteristics of a 'true' areal mean (an infinitely sampled grid box) and offers the advantage over using existing gridded datasets that a high density station network is not required to give an areal average which is representative of the 'true' areal mean in terms of its dry-day probability. The technique therefore provides a tool for estimating dry-day probability in a grid-box average for regions or time periods where only a sparse station network is available.

While the relative simplicity of this approach means that it can be employed in a range of climatological or hydrological applications, it is important to be aware of, and make the best effort to quantify, the uncertainty associated with those estimates. The estimation of dry-day probability in an *n*-station average is tested on three geographical regions – UK, China and Zimbabwe, and a 95% confidence limits of +0.1 and -0.03 for the estimates of dry-day probability in a mean of *n* stations were determined from the accuracy of estimates for those regions. Whilst these three regions span a range of different climates, which gives some confidence in their applicability to other regions of the world, these three regions do not account for all possible climate regimes and therefore any other region to which the methodology is applied may alter the uncertainty bounds on these estimates, and they should therefore be used with this in mind.

When applied in the estimation of the 'true' areal mean, additional uncertainties are added as the grid-box values mean values of the station dry-day probability ($\overline{P(d)_{i,N}}$) and wet/dry day correlation ($\overline{r(w/d)}$). These additional uncertainties are partially dependent upon the number of stations that are available on which to base estimates of these values. Exploration using one grid box from the UK has led to a suggested additional uncertainty of +/- 0.06 when 10 stations or fewer are available, and +/- 0.03 where 10-30 stations are available. This gives some broad indications of reliability of the estimates, but the quality of the estimate will depend not only on the number of stations but also on:

- the distribution of those *n* stations, as 20 stations clustered in one part of the grid box may result in a less representative estimate than 10 which are evenly distributed;
- the characteristics of the grid box – for example, a grid box covering a region of varying elevation, such as a mountainous regions is likely to have a less homogeneous climatic regime, and may require a larger number of stations to be representative. This may be a particular problem if the gauge distribution in the region is biased towards lower-lying, more accessible, locations.

For regions where either or both of these issues might pose a problem, there are additional steps which may be taken to reduce any bias introduced. A more advanced method for calculating mean station dry-day probability ($\overline{P(d)_{i,N}}$), such as a weighted area average, or information from an existing climatology, might be used to give a more reliable estimate of this value.

Despite these uncertainties, the approach described in this chapter provides a useful tool for estimating the dry-day probability in an areal average rainfall series, and this estimate gives a more appropriate value with which to compare grid-box climate model data than an arithmetic mean of the available station data, particularly where available stations are few, or unevenly distributed in space.

This approach offers several advantages over an earlier, similar methodology proposed by Osborn and Hulme (1997). Firstly, the approach presented here does not rely on an empirically determined relationship, as Osborn and Hulme's does, which means that it can be expected to be more robust when applied to other regions or climates. Secondly, the approach of Osborn and Hulme (1997), as discussed in Section 2.3.1.4, relies on the values of conditional probability between pairs of station (the conditional probability that the second station is dry on any day, given that that the first station is dry) as a measure of the spatial dependence between stations for wet/dry day occurrences. This value, however, reflects in part the climate regime of the region, as the conditional probability that the second station is dry depends partly on the dependence between the two stations, and partly on its own individual dry-day probability. This means that the empirical relationships which are based on this value may not be as robust as they appear. In the new approach presented here, the probability of coincident dry-days is scaled according to the dry-day probability at the individual stations (Section 3.3.3), which means that it is a measure of the spatial dependence only. This not only increases the confidence we can have in the estimates based on this value, but also improves the flexibility that the technique offers. The separation of

the values which represent how climate regime, and those which represent the spatial dependence between stations means that it is possible to investigate changes to either of these values independently from one another. This benefit is demonstrated in later chapters when the application of these relationships to future climate is investigated (Chapters 6 and 7).