Theory

In essence, there are two parts to MarkSim. One is a reliable stochastic rainfall generator to drive a weather simulation model. This is all very well when the user has the required parameters to generate synthetic weather records. But what about the situation (normal) when one does not? The second part of MarkSim is a set of surfaces of parameters that can be sampled by the user. More correctly, the parameters of the weather generator are not stored themselves, but rather an "intermediate" set of parameters is stored that can be used to reconstitute a full set of weather generator parameters. The reasons for this intermediate set of parameters are primarily to save space and to enhance efficiency. More details on the methods used in MarkSim can be found in Jones and Thornton (1993, 1997, 1999 and 2000). We summarise these below.

The Rainfall Model

Rainfall is modelled using a two-stage third-order Markov chain. First, it is determined whether any particular day is wet; this depends on whether there was any rainfall on the three previous days. If so, then the amount of rainfall is determined.

Probability of a Wet Day

The probability of day i being wet is defined as

$$P(W/D_1D_2D_3) = \Phi^{-1}(b_i + a_{i-1}d_1 + a_{i-2}d_2 + a_{i-3}d_3) \qquad (1)$$

where Φ^{-1} is the inverse of the normal probability (probit) function, b_i is the monthly baseline probit of a wet day following three consecutive dry days, a_m are binary coefficients for rain (1) or no rain (0) on day m, and d_m are lag constants. Thus the probability of a wet day following three dry days is $\Phi^{-1}(b_i)$, and the probability of a wet day following three wet days is $\Phi^{-1}(b_i + d_1 + d_2 + d_3)$, for example. This part of the model is thus specified by 15 parameters: the baseline probabilities, b_i, derived for each month, and three lag constants, d_1, d_2 and d_3, that are unchanging from month to month.

The model uses a binomial error term and a probit link function. The occurrences of rain on day i-1, day i-2 and day i-3 are treated as the independent variables and the monthly total as another variable. This allows us to test the significance of the lag constants by using a chi-squared statistic. The results showed conclusively that a third-order Markov rainfall model was necessary, because the chi-squared statistic related to the inclusion of the third-order lag in the model was highly significant for 92 percent of the tropical locations that we have studied. This method of fitting the model also allowed us to test the significance of any interaction between the lag constants and the probabilities for the 12 months. Although certain datasets did show small interaction effects, this was generally not the rule, and it was concluded that under a probit transform the lag effects could be considered additive to the monthly effects (see equation (1)). The residual deviance, tested as a chi-squared statistic, was insignificant in almost all cases.

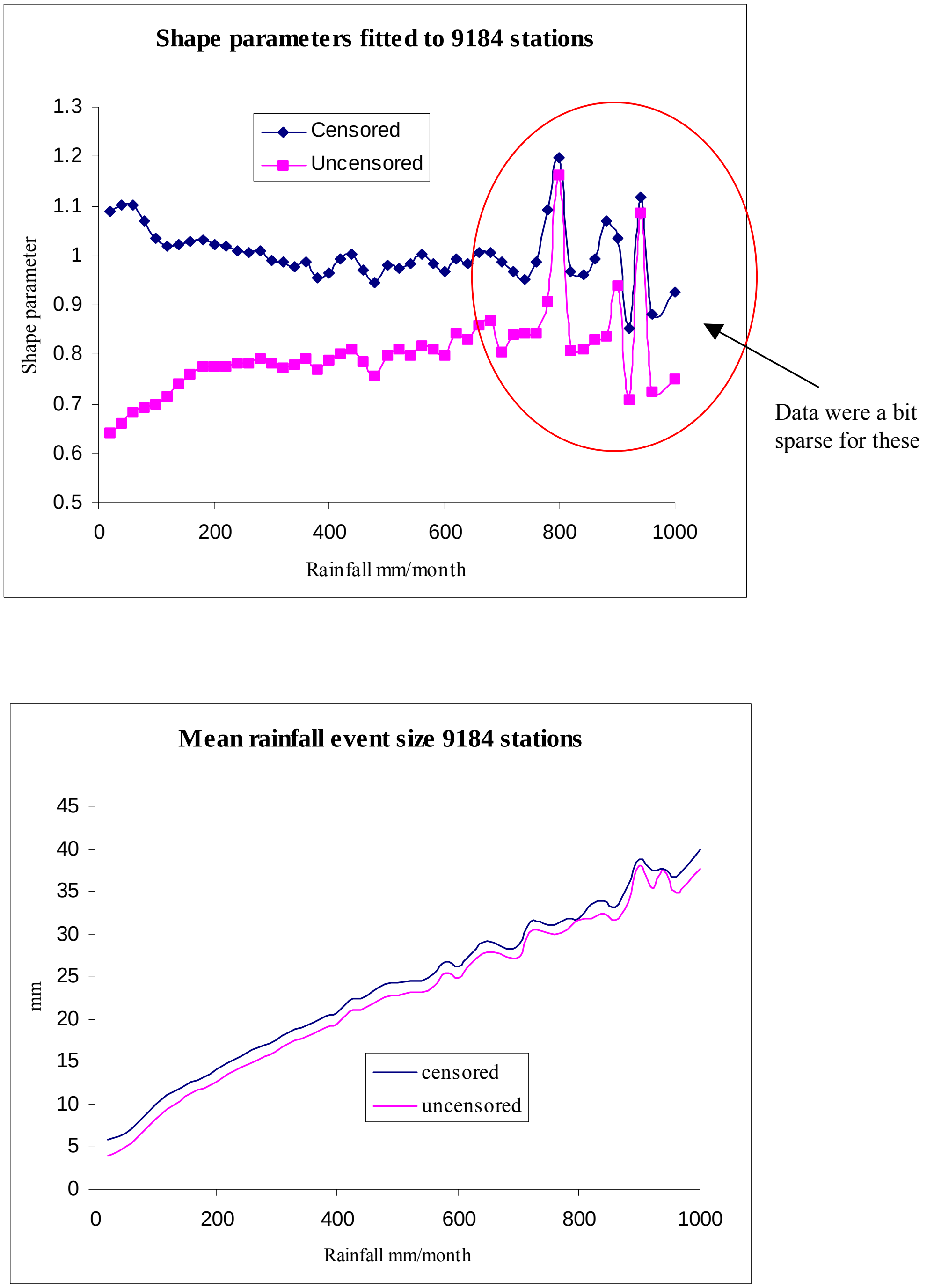
Rainfall on a Wet Day

Rainfall is modelled by using the censored gamma distribution, restricted below 1 mm, to determine daily rainfall amounts on those days that rainfall is experienced (Sterne and Coe, 1982). The method of maximum likelihood is used to estimate the mean and shape parameters of this distribution for each calendar month, thus giving rise to 24 additional model parameters.

The censoring of the gamma distribution means truncating the lower part of the distribution. This is especially important in the case of the gamma distribution because if the shape parameter is low there are a large proportion of small values (small rainfall events). Differences in the rainfall measurements or reporting mean that these small events are reported differently in different data sets. Sterne and Coe (1982) used a censoring at 0.1 mm, all values including trace records were discarded. They used a series of data where measurements greater than 0.1 mm were all reported more or less the same. Unfortunately the widely differing sets of data from all parts of the globe that we have used in MarkSim (almost 11,000 station records) means that there is different reporting with the data not uncommonly being truncated below 1mm. It is a great shame to lose the well reported data that go below this level but in the interests of consistency we had to eliminate them.

This is rather high for a censoring level and we were worried that it might have a large effect on the fitted gamma distribution models. We therefore took data from just over 9000 stations and fitted the gamma distribution to both censored and uncensored data.

The results showed clearly (see Fig 3.1) that, although there was not too much of a shift in mean rainfall size, there was indeed a large effect on the gamma shape parameter.


Shape parameters fitted to 9184 stations

Data were a bit sparse for these

Mean rainfall event size 9184 stations

Figure 3.1 The effects of censoring on the gamma distribution parameters actual data from stations throughout the world.

We therefore needed a way to correct for the effect of censoring as we could not disregard it. We ran 182 (14 x 13) Monte Carlo simulations producing 100,000 samples from each of the gamma populations on the intersections of the rectangular (lower) grid in figure 3.2. We calculated the mean and shape factor for each simulation to check the sampling. We had to use a censoring to 0.000001mm to avoid taking logs of very small numbers. For some of the populations typically one sample in 100,000 was rejected due to this and the sample parameters matched the population parameters within 0.001 for the shape parameter and about 0.02 for the mean. We then censored the sample data to 1mm and recalculated the parameters.

The distorted (upper) grid shows the distortion introduced by the censoring. As can be seen it is a monotonic distortion of the plane, like a map projection. The arrows at the corners show the movement needed at those points to undo the distortion. We can therefore correct for it by working out the projection functions.

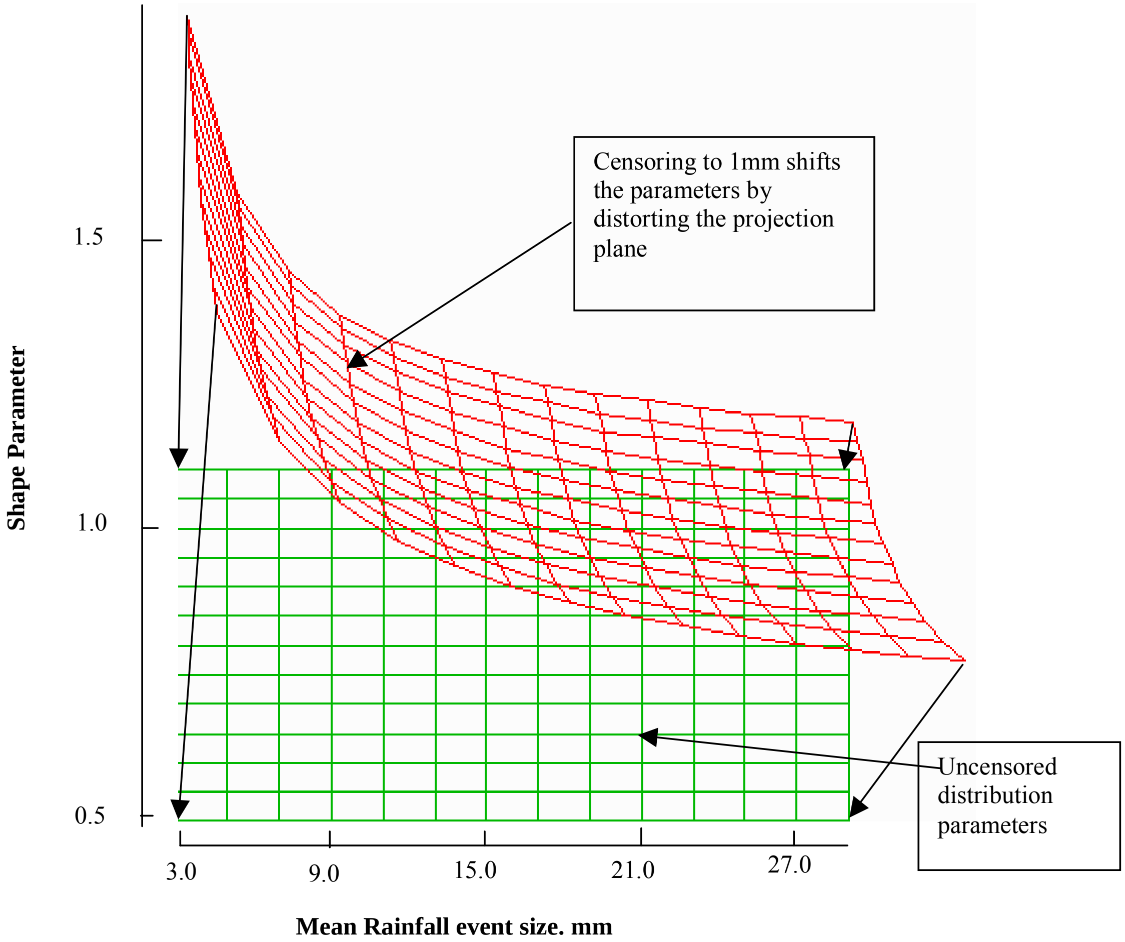


Figure 3.2 Distortion produced in the gamma distribution parameters by censoring to 1 mm.

We used Genstat to fit stepwise regressions to the complete set of 6th order polynomial variables. These are x … x^6, y… y^6 plus all cross products. The fitted functions are shown in Appendix II. Now, we know that by censoring the rainfall we've eliminated all the events with less than 1mm rainfall so we have to adjust the frequency of events as well (if we just use the corrections above the overall amount of rain per month will fall in the model).

The answer is to divide the rainfall probability by the probability of gamma(p,av) exceeding 1mm after reconstituting the probabilities from the probits.

Interpolating Back to Daily Data

In generating rainfall records, the monthly baseline probabilities (the probability of rain after no rain for three successive days) are interpolated to daily probabilities by using the 12-point Fourier transform described in Jones (1987). The lag effects are then added to each day's probit transform of the baseline probability to produce a matrix of 365 or 366 days by eight states (wet or dry conditions on three successive days). The inverse probit transform is then used to transform this matrix to normal probabilities. Similarly, the monthly mean and shape parameters of the gamma distribution of rainfall amounts are interpolated to daily values by using the 12-point Fourier transform.

Annual Variance and the Variability of Parameters

The parameters of the model, being simply estimates obtained from sometimes short data sets, have associated standard errors. To introduce sufficient variability into the model, any random sampling should be based on the uncertainty of the parameter estimates themselves. The 12 monthly baseline probabilities, b_i, are autocorrelated because of the yearly progression of weather, even in the tropics; thus, a resampling scheme must take these correlations into account. This is done by randomly sampling from a 12-variate normal distribution. The resampling scheme can be represented by

$$b^*_i = s_i \, RN_i + b_i, \; i=1,12 \qquad (2)$$

where b^*_i is the sampled value of b_i, the baseline probability of rain; s_i is the standard deviation of b_i; and RN_i is a random normal number. The resampling algorithm involves the Cholesky square root decomposition of the correlation matrix of monthly rainfall. The correct correlation matrix to use would be that of the baseline probabilities in their probit transform. In practice, however, this is difficult to calculate with short data sets. We thus assumed a surrogate correlation matrix and used the standard errors per year obtained in the original GLIM analysis multiplied by the square root of n-1, where n is the number of years.

The pseudo-random normal number generator of Marsaglia and Bray (1964) is used for rapid resampling of the 12 monthly baseline probabilities in their probit transform. The algorithm then adds in the lag constants and produces a new matrix of 365 or 366 days by eight states for each year for which rainfall records are required.

Problems With the Probit Transform

In the course of testing the model with random resampling, we found that it did not work well when the rainfall probabilities were very low. Subsequent analysis showed that the use of the probit transform produces a systematic bias. When resampling is used, low probabilities are overestimated and high probabilities are underestimated after retransformation. Simulations of completely random numbers were used to evaluate the empirical relationship of the standard error to the overall probability level. Probits, produced from runs of up to 200 years, were summed to monthly means and retransformed to probabilities. The variances of the retransformed monthly mean probabilities were then compared with the actual variances introduced in the simulations. The bias in the monthly probabilities was found to be related completely (explaining 100 percent of the variance) and simply, although empirically, to the probability level and the standard deviation. In the algorithm for the rainfall model with sampling, this relationship is used to correct the monthly baseline probabilities by adding to them the correction factor D_i, defined as

$$D_i = b_i(0.55228\ s^2 - 0.26154\ s^3), \qquad (3)$$

where for month i, b_i is the baseline probability of a wet day following three dry days, and s_i is the standard deviation of the baseline probability.

Simulating Temperatures and Solar Radiation

MarkSim uses the DSSAT weather generator [Pickering *et al*. (1994), based on routines of Richardson (1985) and Geng *et al*. (1988)] to generate daily values of maximum and minimum temperatures based on whether the day is wet or dry. The parameters for generating these variables are the long-term monthly means stored in the CLX site file. The original code was part of the WGEN weather estimator (Richardson and Wright, 1984), and this was modified for DSSAT version 3 (Tsuji et al., 1994). The DSSAT modifications use standard deviations rather than coefficients of variation, which make the estimator more stable than the original version. If monthly climate parameters are used as input, the routines use a combination of the regression equations in SIMMETEO (Geng et al., 1988) and Pickering et al. (1988) to compute the standard deviations.

Solar radiation data are generated from monthly mean values for daily solar radiation (or from sunshine hour means, if these exist in the CLI site file). MarkSim uses the routines in the DSSAT generator, which are again based on the equations in Geng et al. (1988) and Pickering et al. (1988). The monthly values of solar radiation are generated from the temperature normals using the model of Donatelli and Campbell (1997), which is a modification and improvement of the earlier model of Bristow and Campbell (1984). Briefly, this model calculates daily solar radiation at the earth's surface as the product of potential radiation and an estimate of the atmospheric solar radiation transmissivity coefficient (the ratio of the value of solar radiation outside the earth's atmosphere and its value at the earth's surface). Potential radiation outside the earth's atmosphere is estimated as a function of the declination, the half day length, a factor accounting for the distance to the sun, the day of year, and the latitude. Potential solar radiation is then modified by the transmissivity to produce an estimate of radiation at the earth's surface. The transmissivity is estimated as a function of clear sky transmissivity, daily maximum and minimum air temperatures, and two empirical parameters.

The Climate Surfaces

Spatially interpolated climate surfaces are now available for many areas. These usually handle long-term climate normals interpolated over a DEM by various methods (Hutchinson, 1997; Jones, 1991). Pixel size depends on the underlying elevation model. It may be as little as 90 m (Jones, 1996), which results in a massive dataset, or 10 minutes of arc (about 18 km), which is as large as is practicable in many instances. In the latter case, the normal elevation model is the National Oceanographic and Atmospheric Administration (NOAA) TGPO006 (NOAA, 1984). We have produced interpolated datasets at CIAT for Latin America and Africa using data from about 10,000 stations for Latin America, 7000 for Africa and 4500 for Asia. Each set of surfaces consists of the monthly rainfall totals, monthly average temperatures, and monthly average diurnal temperature range. This makes 36 climate variates in three groups of 12.

We use a simple interpolation algorithm based on the inverse square of the distance between the station and the interpolated point. For each interpolated pixel we find the five nearest stations. Then the inverse distance weights are calculated and applied to each monthly value of the data type being interpolated. Thus, for five stations with data values *x* and distances from the pixel distance d:

$$x_{pixel} = \sum_{1=1}^{5} d_i^{-2} \times \sum_{1=1}^{5} \frac{x_i}{d_i^{-2}} \qquad \textbf{(1)}$$

Temperature data are standardized to the elevation of the pixel in the DEM using a lapse rate model (Jones 1991). Using this simple interpolation has various advantages. First, it is the fastest of all the common methods. Second, it puts the interpolated surface exactly through each station point, because the weight *1/(d(I)**2)* becomes infinite as *d* approaches zero. Third, the interpolation is highly stable in areas of sparse data. It approaches the mean of the nearest stations while they all become equally distant. Fourth, it is relatively stable against errors in station elevation; only the local region of that station is affected. On the other hand, laplacian spline techniques and co-Kriging both propagate these errors more extensively. This is one advantage of using a proven lapse rate model instead of fitting a local one, as do both of these latter techniques.

The method has two small disadvantages. First, the derivative of the surface becomes zero as it passes through the station point. In other words, each station is on a small plateau or step in the interpolated surface. This is usually much smaller than the pixel size and hence is not noticeable. Second, a (usually small) step occurs in the fitted surface as stations come into or drop out of the fitting window. Where the station density is high with respect to the pixel size, this is almost impossible to see. Where the stations are not so dense, it can produce unsightly straight lines or smooth arcs in the fitted rainfall data, which are not tied to elevation. Inspection of the surface's profile usually shows that these are negligible artifacts, but they are unsightly and can undermine confidence in the surface maps.

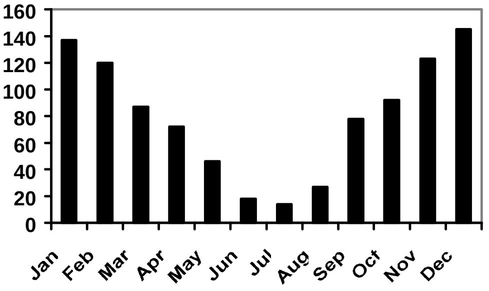
Climate Date Standardization (Rotation)

The climatic events that occur through the year, such as summer/winter and start/finish of the rainy season, are of prime importance when comparing one climate with another. Unfortunately, they occur at different dates in many climate types. The most obvious case is where climates are compared between points in the Northern and Southern Hemispheres, but more subtle differences can be seen in climate event timing throughout the tropics. What we need is a method of eliminating these differences to allow us to make comparisons free of these annual timing effects.

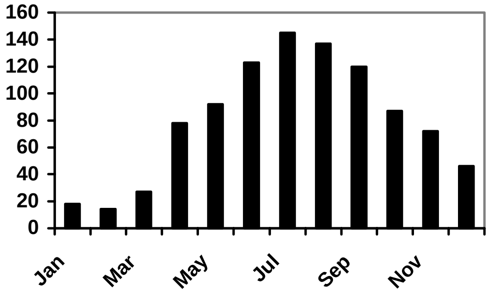
Let us look at two hypothetical climate stations. They are in a typical Mediterranean climate—warm wet winters, hot dry summers. Northville could be somewhere in California, and Southville might be in Chile. The August rainfall in Southville is received in January in Northville. If we plot these rainfalls in polar coordinates, we can readily see that to compare them we need to rotate them to a standard time.

Monthly rainfalls for Northville and Southville.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Northville	137	120	87	72	46	18	14	27	78	92	123	145
Southville	18	14	27	78	92	123	145	137	120	87	72	46



Northville monthly rainfall



Southville monthly rainfall

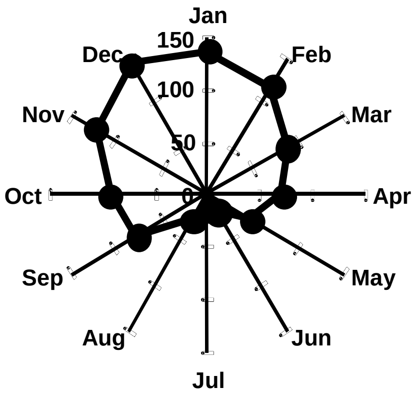
How do we do this automatically? The answer is the 12-point Fourier transform. This is fortunately the simplest of all the possible Fourier transform algorithms. It is highly computationally efficient and fast. In fact, it is the basis of nearly all Fast Fourier transform algorithms that break the problem down sequentially into the simple 12-point case. It takes the 12 monthly values and converts them to a series of sine and cosine functions. The one used in FloraMap has a modification to make it conserve the monthly total values (Jones 1987). The equation produced is:

$$r = a_0 + \sum_{i=1}^{6} a_i \sin(ix) + b_i(ix)$$

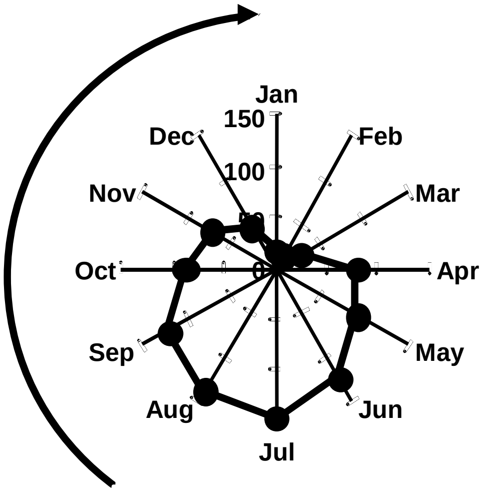
This can be rewritten as a series of frequency vectors, each with an amplitude α_i and a phase angle, $\theta_{i:}$

$$\alpha_i = \sqrt{\left(a_i^2 + b_i^2\right)} \qquad\qquad \theta_i = \sin\left(\frac{b_i}{\alpha_i}\right) = \cos\left(\frac{a_i}{\alpha_i}\right)$$

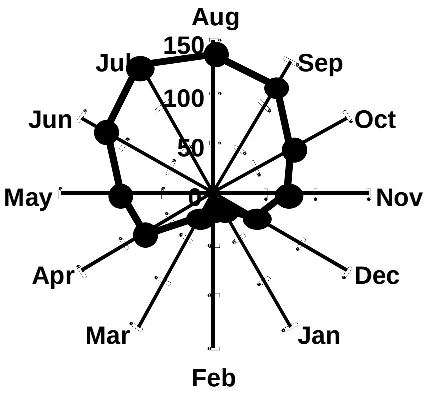
If we subtract the first phase angle from all the other vectors in the set then we have produced a rigid rotation of the vectors. This is the rotation that we are seeking. It puts the maximum of the first frequency at a phase angle of zero and places the rest in positions equivalent to their angular separation in the original data. We then use the first phase angle for rainfall to rotate the data for temperature and diurnal temperature range, and these variates are rigidly rotated along with the rainfall.



Northville



Southville



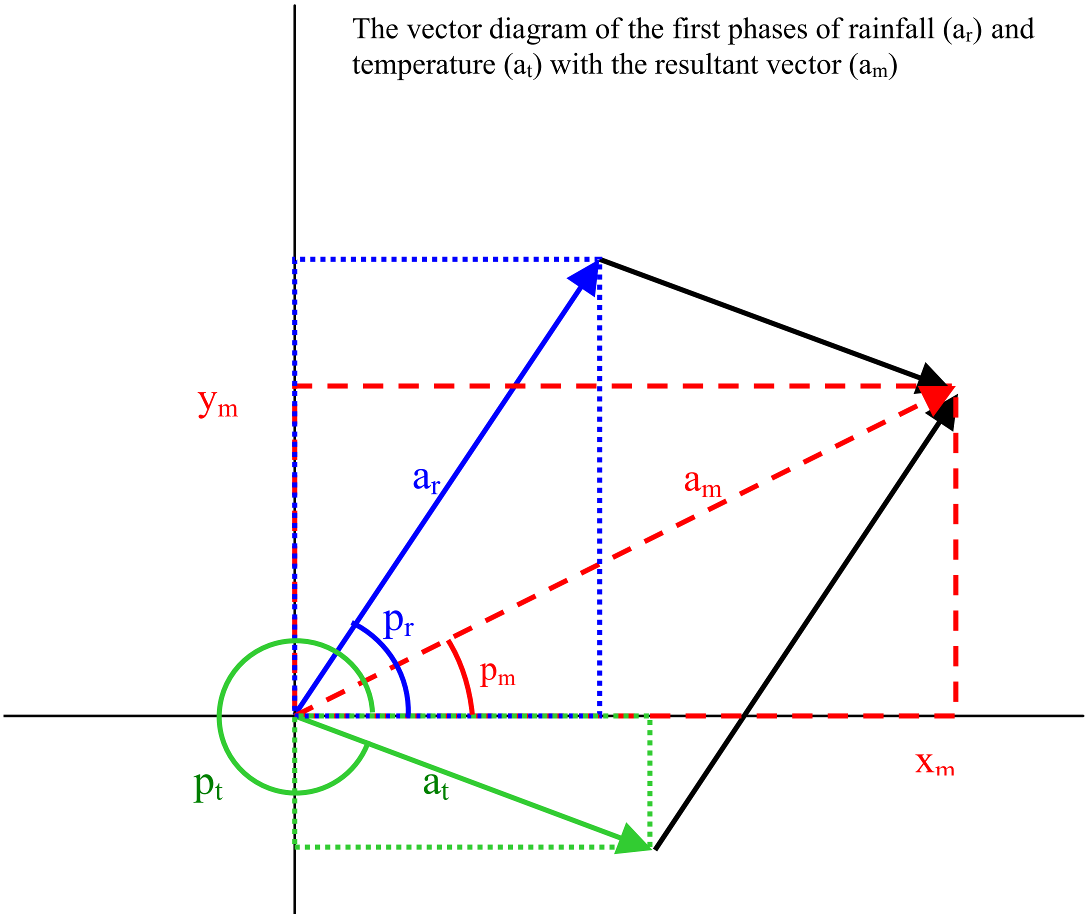
Southville rainfall
rotated to coincide with
timing of Northville

This explanation works well for the tropics. There was a small chance of the procedure going off the rails if an accession set was fitted to a model in latitudes high enough to exhibit Mediterranean climates (as used in the example above). In the case when some of the accessions fall in the winter rainfall areas and some in strongly summer rainfall (non-Mediterranean) areas, the resulting model could have a very poor fit. Because this is botanically unlikely, it probably has not yet been observed in practice; although the case has been happened on when running an artificial test set across the Andes in Chile/Argentina.

The beta release of MarkSim went out with this type of rotation algorithm, as did the first release of FloraMap. When the climate grids of the latter were extended to Europe the case came up where annual climate pattern was dominated by temperature and not rainfall.

We therefore have the possibility of rotating on rainfall or temperature, but when to decide which is the dominant? We tried many combinations of rules, but unfortunately came to the conclusion that none were acceptable. They all resulted in a hard line across the map at some point where the rotation basis changed. This led to climates that should have been grading imperceptibly from one type to another suddenly jumping at a discontinuity. This would have given the users serious problems when fitting models in these areas.

The best solution found is to use BOTH the rainfall and the temperature in calculating the rotation phase angle. Thus:



The vector diagram of the first phases of rainfall (a_r) and temperature (a_t) with the resultant vector (a_m)

The resultant phase angle and amplitude are then:

$$y_m = a_r \cos p_r + a_t \cos p_t$$
$$x_m = a_r \sin p_r + a_t \sin p_t$$
$$a_m = \sqrt{y_m^2 + x_m^2}$$
$$p_m = \mathrm{angle}\left({}^{x_m}\!/_{a_m}, {}^{y_m}\!/_{a_m} \right)$$

Unfortunately, this does not completely solve the problem of fitting a model to climates with different weather determinants. However, the vast majority of climates in the world are either:

(1) Rainfall determined where temperature is not an important seasonal effect (large areas of the tropics and subtropics),
(2) Temperature determined where rainfall is even throughout the year (most of the rest of the tropics and some temperate climates), or
(3) Rainfall and temperature determined when the two variates are highly correlated (summer rains - most of the rest of the world).

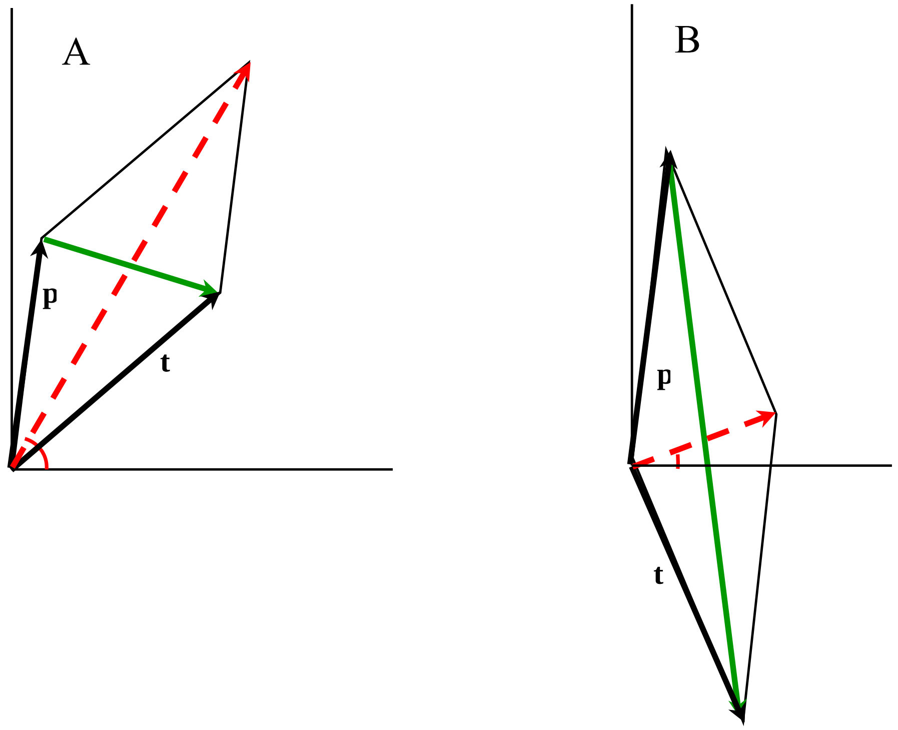
The Odd Man Out is:

(4) Winter Rains and Hot Dry Summers (almost only Mediterranean climates).

Luckily, the Mediterranean climates are at moderately high latitudes and we can afford to have the rotation dominated by temperature without losing generality in the rotations and comparisons. We therefore need to increase the weighting for the temperature vector smoothly as we approach the Mediterranean climates (in order to avoid a sudden swing).

The following weightings were found to work well:

p = rainfall mm
t = temperature x 2 x abs(latitude)



There is a potential trap when the two vectors almost cancel each other. This could result in wild swings of the rotation angle for small changes in the rainfall and temperature vectors. This becomes more likely as the situation passes from that in A (above) to B and beyond. The dashed arrows are the rotation vectors as before, but calculated on the weighted rainfall and temperature vectors.

Where the rotation vector is the vector sum **r** + **t,** the counter-diagonal vector is the difference **r – t**. It can be readily seen that the dangerous areas will be when **r – t** is much greater than **r + t.** We can therefore use a handy index of stability, s.

$$s = \arctan\left(\frac{|\mathbf{r}-\mathbf{t}|}{|\mathbf{r}+\mathbf{t}|}\right)$$

This will be zero for stable states where the rotation angle is dominated by rainfall, by temperature, or by both acting in concert. It will approach π/2 as the vectors tend towards cancelling their effects. Because we can map this index we can check for areas where this indeterminate rotation might occur. Areas of relatively high s (potential instability) occur on the USA Pacific Coast, in Chile, northeastern Brazil, Sri Lanka, and through some areas of Central Africa. However, in no area does the index reach 80 degrees. Although this appears high, the phase angles are rotated correctly and in fact there is little chance of a spurious rotation.

To save computing time, the whole climate surface is rotated according to these rules and all operations in MarkSim are done in the rotated phase space.

☞ *The only exception to this is when the user requests a climate diagram for a climate surface point.*

Surface Interpolation

As noted above, the rainfall model requires an extensive set of parameters: 12-monthly baseline probits (termed $\boldsymbol{\beta}$) and monthly mean (**av**) and shape (**ps**) parameters for the rainfall event gamma distribution. Twelve monthly standard deviations and the 66 off-diagonal elements of the 12 X 12-correlation matrix for $\boldsymbol{\beta}$ are also required. Three lag parameters (**d**) allow us to calculate a 12 X 8 probit transition matrix.

Interpolated climate surfaces commonly hold only climatic normals for monthly rainfall and maximum and minimum temperatures. We therefore need some help to get 117 parameters from 36 monthly values. This help comes from the structure that is inherent in the Markov process and similarities in climate processes within climate types that, although not included explicitly in the model, affect the model parameters in consistent ways.

To produce the surfaces, the first step consisted of clustering the available historical station data. We used the rotated data in a two-pass leader cluster algorithm analysis. The first pass allocated stations as cluster leaders whenever they exceeded a minimum cluster distance. The second pass reallocated the stations to their respective cluster leaders. The distance measure was the euclidean distance in the 36-dimensioned climate space. We tested various exponential transformations on the rainfall data and chose the exponent 0.5 (square root), based subjectively on the evenness of cluster sizes. Cluster sizes varied from 1 to 307 stations with a mean of 13.9 stations per cluster.

To calculate the expected parameter values of the model for any pixel in the interpolated climate surface, we first need to know to which cluster the pixel belongs and second, how the climate normals of the pixel adjust the parameter values within each cluster relative to the cluster mean values. We use the cluster seed as the type climate for each cluster and calculate the euclidean distance in climate space for each pixel. The pixel is then associated with the closest cluster seed. This need not be geographically close. For each of the parameter types, we fitted a regression sub-model within each cluster to trim the parameters estimated for the pixel to the best estimate we could make from the limited data recorded for each pixel of the climate surface. We dealt separately with two of the parameter types; rainfall event averages (***av***) and correlation matrices (see below).

Derivation of parameter estimates

The parameters for which we need regression sub-models fall into two classes. $\boldsymbol{\beta}$, ***ps*** and ***se*** have 12 monthly values. The lag parameters ***d*** are single valued for each station or pixel. We therefore created two sets of independent variates for their estimation. The sets were derived from the basic station information and scaled as follows:

$\boldsymbol{\beta}$, ***ps*** and ***se***

rm = monthly rainfall/200

tm = (monthly temperature - 15)/10

dm = (monthly diurnal temp. range -11)/4

srm = ***sqrt***(monthly rainfall)/14

tmsq = ***tm***^{2}

rmsq = ***rm***^{2}

dmsq = dm^2

lat = station latitude /90

elev = (Ln(station elevation+10)-5)/3

$\boldsymbol{d_1}$ $\boldsymbol{d_2}$ $\boldsymbol{d_3}$

ra = annual rainfall/200

ta = (annual temperature -15)/10

da = (annual diurnal temp. range - 11)/4

rar = (annual range rainfall)/200

tar = (annual range temp. - 15)/10

dar = (annual range diurnal temp -11)/4

rasq, tasq, dasq = ra^2***, ta***^{2}***, da***^{2}

rarsq, tarsq, darsq = rar^2***, tar***^{2}***, dar***^{2}

lat = station latitude /90

elev = (Ln(station elevation+10)-5)/3

sra = ***sqrt***(***ra***)

The scaling was designed to place regression parameter estimates within a reasonable range for the subsequent selection process.

We ran a five-stage stepwise regression for each cluster for $\boldsymbol{\beta}$, ***ps*** and ***se*** and a six-stage stepwise regression for the ***d*** lag parameters. Inspection of the results showed that correlations between the independent variates often resulted in large regression coefficients as a result of differential effects of the variates. Although the effects of fitting both terms were often statistically significant, their inclusion would have led to an undesirable instability of the regression as predictor when we present new data with slightly different values. Because we know the bounds of the clusters, we did not want a model predicting values outside these bounds. Inspection of each cluster for each of the parameters would have been far too time consuming. We therefore compiled a list of the independent variates ordered by the number of times that they occurred in each parameter set of parameter regressions. We then fitted the maximal model for each parameter and progressively eliminated variates until none showed a regression coefficient that would force a prediction out of the cluster bounds. Details of the regression analyses can be found in Jones and Thornton (1999).

Rainfall event averages

If we were to have fitted climate surfaces to rain days per month, the ***av*** parameters could be easily calculated as the monthly rainfall total divided by the rain days. Unfortunately, the main sources of monthly climate data used in the interpolated climate surfaces rarely contain the number of rain days. We therefore have to estimate these from the model. The probability coefficients used in the model are transition probabilities. They are the probability of the system passing from one triad state to another. The probabilities that we need to calculate the rain days per month are the state or stationary probabilities which, except for the calibration stations, we do not have.

As a fortunate consequence of some structural redundancies in the model these can be calculated from the monthly average rainfall and the estimates of $\boldsymbol{\beta}$. As noted above, the model works in two parts: one decides whether today will be a rain day, the other decides how much rain should fall. The two parts have a hidden link. A triad is a binary form of three digits denoting rain on each of 3 days. Thus triad $\boldsymbol{t}$ = 101 means it rained yesterday, it did not rain the day before yesterday, but it did rain 3 days ago. Within the model, there are two classes of probability. One, the transition probability $\boldsymbol{p}$(t), shows the probability of rain today given that the system is in triad state $\boldsymbol{t}$. The other, the state probability **s**(t), shows the probability of the system being in a certain triad state. The model calculates the transition probabilities as probits. Thus the transition probability for a given triad $\boldsymbol{t}$ in month $\boldsymbol{m}$ is:

$$P_{t,m} = \Phi^{-1}\left(\beta_m + \sum_{i=1}^{3} t_i d_i\right)$$

where $\mathbf{\Phi}^{-1}$ transforms from the probit form to a probability. We can write a transition matrix that governs the relationship between these two types of probabilities. Because we can calculate the **p**(t) from the equation above, we can use the transition matrix to calculate **s**(t).

$$S_{-1}\begin{pmatrix} 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{pmatrix} \times \begin{pmatrix} 1-p_{000} & p_{000} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1-p_{001} & p_{001} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1-p_{010} & p_{010} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1-p_{011} & p_{011} \\ 1-p_{100} & p_{100} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1-p_{101} & p_{101} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1-p_{110} & p_{110} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1-p_{111} & p_{111} \end{pmatrix} = S_1^T \begin{pmatrix} 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{pmatrix}$$

Unfortunately, this matrix is singular. However, the frequency of $s_{110} = s_{011}$ and that of $s_{100} = s_{001}$. The proof of this is simple. Any rainfall sequence longer than 1 day must start with the triad 011 and finish with the triad 110. Thus, in any sequence, the frequencies must be equal if we discount a possible difference of one depending on the starting condition. That is to say, if the sequence starts with a rain period and finishes with a dry period there will be exactly one more 110 than 011, irrespective of the length of the sequence. The same argument holds for triads 001 and 100 where dry days rather than rain days are counted. The state probabilities sum to unity as do the transition probabilities and the state outcomes. Adding alternative rows of the matrix eliminates four rows. We can therefore apply these restrictions by adding in four rows to the matrix. This then becomes positive definite and has a viable inverse.

$$\begin{pmatrix} 1 & 1 & 1 & 1+p_{011} & 1 & 1 & 1 & p_{111} \\ 1 & 1 & 1+p_{010} & 1 & 1 & 0 & 1+p_{100} & 1 \\ 1 & 1+p_{001} & 1 & 0 & 1 & 1+p_{101} & 1 & 1 \\ 1+p_{000} & 0 & 1 & 1 & 1+p_{100} & 1 & 1 & 1 \\ 2 & 2 & 2 & 3 & 2 & 2 & 1 & 2 \\ 2 & 2 & 3 & 2 & 1 & 1 & 3 & 2 \\ 2 & 3 & 1 & 1 & 2 & 3 & 2 & 2 \\ 2 & 1 & 2 & 2 & 3 & 2 & 2 & 2 \end{pmatrix} \times \begin{pmatrix} 2 \\ 2 \\ 2 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = S$$

We thus have a reliable algorithm to pass from transfer probabilities to state probabilities. Calculating the average rainfall event (***av***) now requires only the baseline probabilities, the lag parameters and the monthly rainfall normals. The rain-day probabilities are found by summing s_{001}, s_{011}, s_{101} and s_{111} and are divided into the monthly rainfall normals. This eliminates 12 unwanted degrees of freedom and we have constrained the model to simulate actual long-term monthly rainfall normals.

Correlation matrices

As noted in Jones and Thornton (1997), we can see distinct patterns in the correlation matrices of many climate clusters. These patterns can, however, be highly complex. We therefore decided not to try to refine the estimate of the correlation matrices by fitting sub-models within climate clusters, but to accept the correlation matrix calculated from the pooled variance/covariance matrices of the cluster members as being representative of all pixels allocated to that cluster.

References

Bristow, K.L., and G.S. Campbell. 1984. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agricultural and Forest Meteorology 31:159-166.

Donatelli, M., and G.S. Campbell. 1997. A simple model to estimate global solar radiation. PANDA Project, Subproject 1, Series 1, paper 26, ISCI, Bologna, Italy. 3 pp.

Geng, S., J. Auburn, E. Brandsletter, and B. Li. 1988. A program to simulate meteorological variables: documentation for SIMMETEO. Agronomy Report No. 204, University of California, Crop Extension, Davis, California.

Jones P G (1987). Current availability and deficiencies in data relevant to agro-ecological studies in the geographic area covered by the IARCS. In: A H Bunting, (editor), Agricultural Environments, 69-83. CAB International, Wallingford, UK.

Jones, P.G. 1991. The CIAT Climate Database Version 3.41. Machine readable dataset. Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia.

Jones P.G., and P.K. Thornton. 1993. A rainfall generator for agricultural applications in the tropics. Agricultural and Forest Meteorology 63, 1-19.

Jones, P.G., and P.K.Thornton. 1997. Spatial and temporal variability of rainfall related to a third-order Markov model. Agricultural and Forest Meteorology 86, 127-138.

Jones, P.G., and P.K. Thornton. 1999. Fitting a third-order Markov rainfall model to interpolated climate surfaces. Agricultural and Forest Meteorology 97, 213-231

Marsaglia G and Bray T A (1964). A convenient method for generating normal variables. SIAM Review 6 (3), 260-264.

NOAA (National Oceanographic and Atmospheric Administration). 1984. TGP-OO6 D. Computer compatible tape. NOAA, Boulder, CO.

Pickering N.B., J.R. Stedinger and D.A. Haith. 1988. Weather input for nonpoint-source pollution models. J Drain. Eng. 114 (4), 674-690.

Pickering, N.B., J.W. Hansen, C.M. Wells, V.K.Chan, and D.C. Godwin. 1994. WeatherMan: a utility for managing and generating daily weather data. Agron. J. 86:332-337.

Richardson C W (1985). Weather simulation for crop management models. Transactions of the ASAE 28 (5), 1602-1606.

Richardson C W and Wright D A (1984). WGEN: A Model for Generating Daily Weather Variables. USDA, Agricultural Research Service, ARS-8, 83p.

Sterne R D and Coe R (1982). The use of rainfall models in agricultural planning. Agricultural Meteorology 26, 35-50.

Tsuji, G.Y., G. Uehara, and S. Balas, editors. 1994. DSSAT Version 3. University of Hawaii, Honolulu, USA.