# Modelling weather data

D.J. Allcroft[1], C.A. Glasbey[1] & M. Durban[1]

Meteorological variables are essential inputs to many models in agriculture and hydrology. In particular, crop models need them because crop growth is substantially affected by weather. Frequently though, weather data will not be available in the required form. For instance, the available data may not be for exactly the right location, for the right time period or at the right scale, or simply not enough data may have been collected. However by developing appropriate mathematical models from the available data, we can simulate data of the right form and quantity.

Typical requirements for crop models are long series of daily data, but these are rarely available in the quantity needed. For example, 50 years of data might be required to encompass the range of weather patterns for model prediction. Alternatively, we might have data from a restricted network of weather stations, but want to simulate realistic data for the whole area containing the stations. Two examples are discussed below in some detail. In the first, we build a model to simulate time series of many weather variables simultaneously, taking account of the dependencies between variables. The second is a rainfall disaggregation problem, the aim being to produce a realistic pattern of rainfall at a finer spatial scale than that recorded. First we discuss some general issues about weather variables.

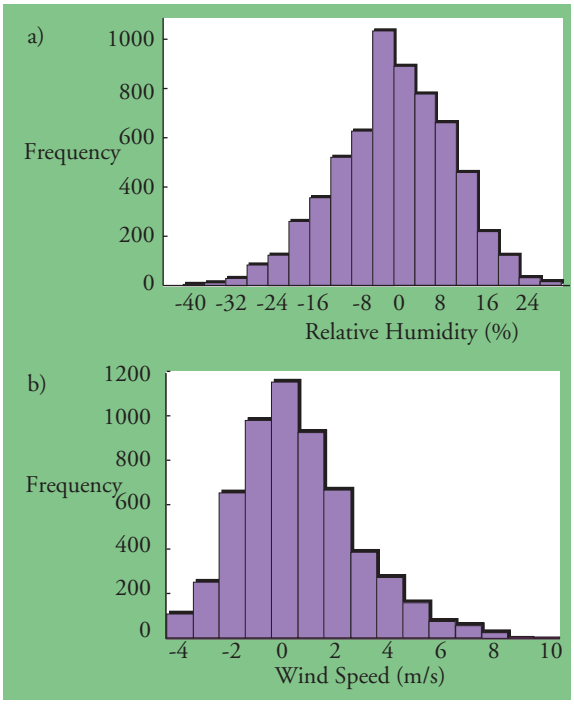


**Figure 1** An illustration of two weather variables that are approximately Gaussian distributed: histograms of a) relative humidity and b) wind speed, after subtraction of seasonal trend.

**Weather variables** Most weather variables can be assumed to be normally distributed, i.e. to have come from a Gaussian distribution. Consider, for example, daily wind speed and relative humidity. Figure 1 shows histograms of 17 years of daily data collected at Mylnefield (SCRI), after cyclic seasonal trends have been removed. Though some skewness is evident, the distributions are approximately Gaussian. Other weather variables that are approximately Gaussian include air pressure and daily maximum and mini-

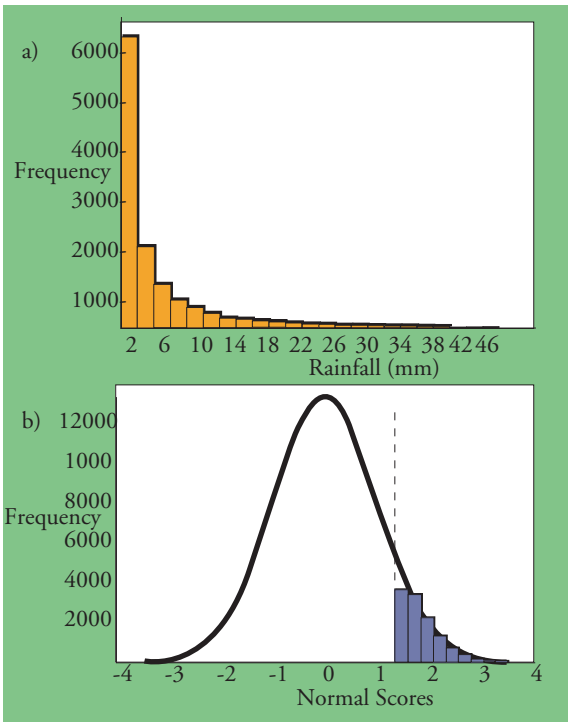


**Figure 2** An illustration that rainfall is not Gaussian distributed, but can be transformed to a censored Gaussian variable: a) histogram of hourly rainfall, with zero values omitted; b) histogram of rainfall after a normalising transformation, with superimposed Gaussian distribution (—) and zero rainfall threshold (- -).

[1]Biomathematics & Statistics Scotland, Edinburgh

mum temperatures. Rainfall, in contrast, is highly non-Gaussian, most periods having no rain at all and even restricting to wet periods, the distribution of rainfall is still very non-Gaussian. For example, Figure 2a shows the distribution for hourly rainfall in the Arkansas-Red Basin River area, USA. This histogram omits the 91% of hours that had no rain at all and shows that of the wet hours, most have little rainfall but the long tail of the distribution allows for the occasional hours which have very heavy rainfall.

For rainfall simulation, multi-stage approaches have previously been used, e.g. first simulating a rain/no rain process and then simulating the amount of rain for the wet periods from some skewed distribution. However, there are many advantages of working in a Gaussian framework. Most importantly, there is much well-developed theory for Gaussian processes, hence we can build models based on well-established methodology. A further advantage of Gaussian variables is that they are closed under scaling and addition. So for example, if the distribution of daily totals of a variable is Gaussian, then so is the distribution of weekly totals, and so is the distribution of differences between the daily totals and an overall daily average.

Hence for modelling rainfall, we seek a transformation to normality, enabling us to use established models for Gaussian processes. The method we have developed is to define a transformation such that for wet periods, rainfall values are converted to values distributed as the upper part of a Gaussian distribution, and for dry periods, zero rainfall corresponds to censored values from the lower part of the same distribution. Figure 2b shows how hourly rainfall values are transformed to match the upper tail of a Gaussian distribution, whereas the lower part of the distribution corresponds to the 91% of dry hours. Thus, a latent Gaussian process can be thought to have generated the rainfall data: for values below the threshold, the period is dry and for values above the threshold the period is wet, with a transformation applied to generate the actual rainfall value. Figure 3 illustrates this, showing both a realisation of the latent process (Fig. 3a) and the resulting rainfall (Fig. 3b).

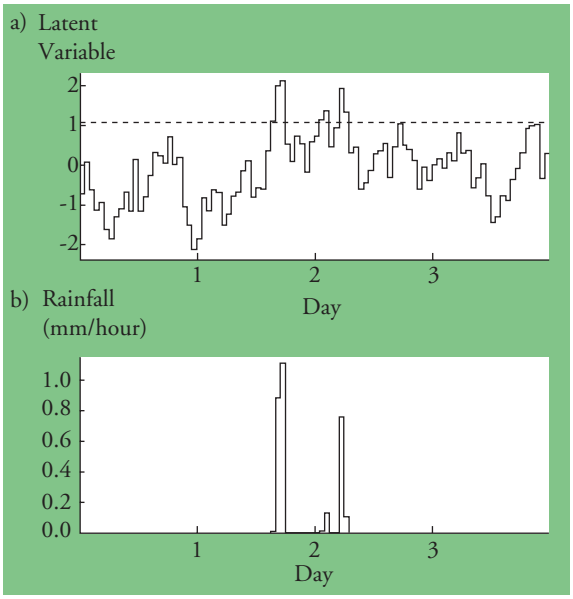


**Figure 3** An illustration of the relationship between the latent Gaussian process and rainfall: a) simulation of three days of hourly data from a Gaussian process (–), together with the zero-rainfall threshold (- - -); b) the resulting rainfall sequence.

### Example 1: A weather generator for several weather variables

The aim is to generate daily values at a single site for several weather variables simultaneously, namely minimum and maximum temperatures, radiation levels, humidity, wind speed and rainfall. Previous approaches have either first simulated rainfall and then conditionally simulated the other variables or, conversely, simulated the other variables and then conditionally simulated the rainfall. By using the approach described above to transform rainfall, and applying a simple log-transformation to the radiation level, all six variables can be assumed Gaussian. Hence we can use a standard multivariate Gaussian model to generate values for all variables simultaneously, taking into account the dependencies between the variables.

The data we model consist of 17 years of daily values of the six variables recorded at Mylnefield (SCRI). All exhibited annual cyclic patterns, which were accounted for by fitting finite Fourier series. This effectively removes the cyclic aspect of the data and we assume the resulting series are stationary, a feature assumed in all basic time series models.

ARMA (auto-regressive moving average) processes are well known statistical models for time series. They model the value of the variable at time $t$ as a function of the values of the variable at the previous few time points, plus a Gaussian-distributed random error. The steps in model fitting are usually: estimate the auto-correlations of the time series, i.e. the correlation of the variable with itself at given time-lags apart; from these identify the appropriate ARMA model to use; and finally estimate the ARMA parameters by

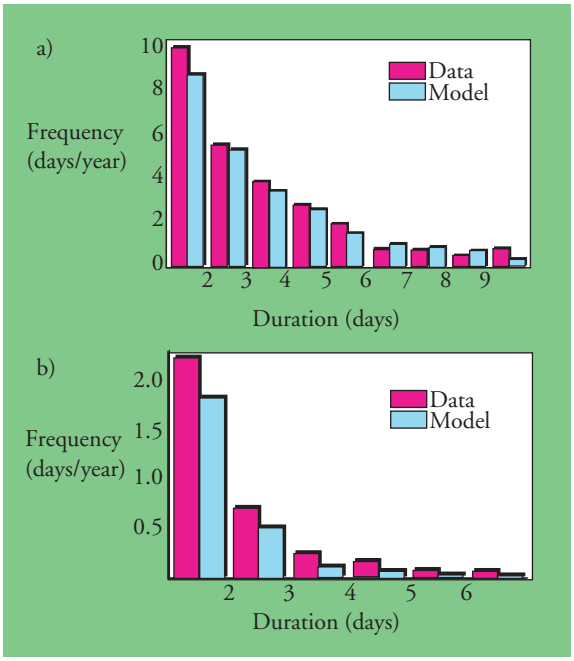


**Figure 4** An illustration of the agreement between Mylnefield data and simulations from our model for two summary statistics: a) frequency of wet periods; b) frequency of warm, humid periods with duration of two or more days.

maximum likelihood.

Simple ARMA processes model a single variable; VARMA (vector ARMA) processes model a vector of variables instead, i.e. model several variables simultaneously. To fit these models, as well as the auto-correlation of each variable, we need to estimate the cross-correlations between the variables, i.e. the correlation between one variable and another at a given time-lag apart. Because the transformed rainfall variable is censored, the model fitting procedure is slightly more complex than usual. Firstly, estimation of the auto-correlation for the rainfall variable, and for the cross-correlations involving the rainfall, have to take into account that the value of the latent Gaussian variable is unknown on dry days. Therefore, we must substitute instead an integral over the range within which it can fall. Secondly, in estimating the parameters, the usual maximum likelihood approach is unavailable due to the data being censored. Instead, we use a simple least squares approach, minimising the sum of squares of differences between the estimated correlations and those predicted by the model.

After fitting the model, we can then simulate from it. Here we simulated 100 runs of 17 years of data and calculated various statistics to compare with the observed data. These included monthly means of weather variables, numbers of wet days and total amount of rain per month. In addition, we compared run lengths of rainy days and of warm, humid days. Reasonable consistency was seen in all cases, hence this model can be used to make predictions about the frequency, duration, etc. of various types of weather conditions. For example, conditions that result in the onset of potato blight have been summarised as "a temperature in excess of 10°C and relative humidity above 90% for 11 or more hours in each of two or more consecutive days". Figure 4 shows how well the model captures the distribution of both this particular combination of conditions (Fig. 4b) and wet periods (Fig. 4a).

### Example 2: Rainfall disaggregation

Rainfall data are frequently collected at coarser spatial scales than required. Methods are, therefore, needed for simulation of realistic patterns of rainfall at finer scales. We use the same approach described above to transform the rainfall to a thresholded Gaussian variable, though now we are in a space-time framework and hence each measurement of rainfall corresponds to a given area over a given time.

We apply our model to 12 hours of hourly data from the Arkansas-Red Basin River. We model the data at 8km x 8km resolution, aggregate to 5 x 5 blocks and then disaggregate from here, so allowing an assessment of how well the disaggregation procedure works, since we can compare disaggregations to the original data at

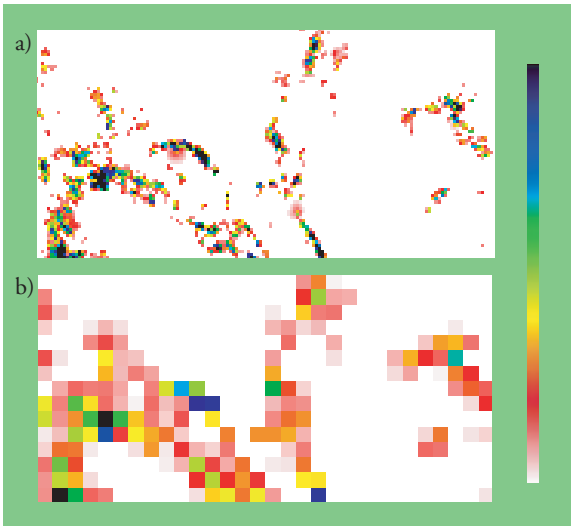


**Figure 5** The spatial distribution of rainfall in the Arkansas-Red Basin River area for one hour, at a) fine and b) coarse spatial scales. White indicates zero rainfall, through to black indicating the highest.

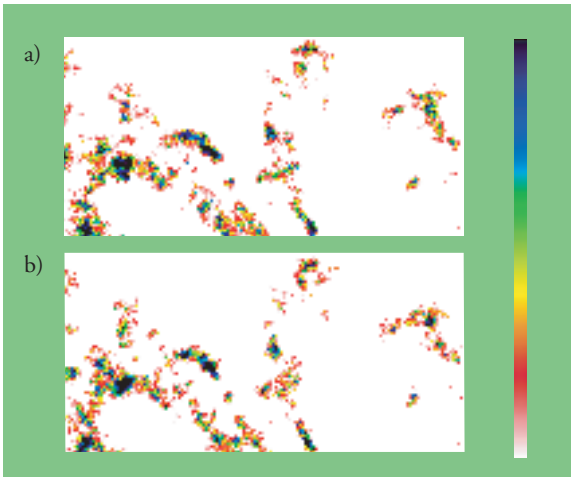


**Figure 6** An illustration of how a latent Gaussian Markov random field can be used to disaggregate the coarse-scale rainfall data in Figure 5b: two simulations of a disaggregation.

fine scale. Figure 5 shows one hour of rainfall data, both at the fine scale we wish to disaggregate to (Fig. 5a), and the coarse, aggregated scale (Fig. 5b).

As in the first example, we need to estimate the correlation of the underlying Gaussian variable, using methods which take into account that the values are censored at dry locations/times. As the problem here is spatio-temporal, we need to estimate the correlation over all combinations of lags in two-dimensional space and time, i.e. in three dimensions. The pattern of correlation was judged to be similar in all directions in space, i.e. the process is spatially isotropic.

We model the correlations as a Gaussian Markov random field (GMRF), 'random field' just being the term given to a random variable in several dimensions. A Markov process in time is one in which the observation at time $t$ depends only on the immediately preceding observation(s) and is conditionally independent of those occurring earlier. The Markov random field is the higher dimensional version of this, so an observation at a certain point in space and time depends on the values at a (small) neighbourhood of points around it, but is conditionally independent of values at locations further away, both in space and time. Parameters for the GMRF are estimated in a similar way as for the ARMA process in the previous example. Here weighted least squares are used to minimise the sum of squares of differences between the estimated correlations and those predicted by the fitted model.

Simulation of the fitted process is carried out using Gibbs sampling, the procedure being to start from some initial configuration, here the aggregated picture of Figure 5b, and then simulate updates for all the values, conditional on the values at neighbouring points. This is easily done, as the conditional distributions are multivariate normal, and computationally fast to simulate. We update 5 x 5 blocks in turn, constraining the total rainfall in the block to be consistent with the observed total for that block. The updating of all sites forms a complete update – many complete updates are run and, after each, statistics calculated in order to judge when the procedure has reached equilibrium. Once this is achieved, any of the realisations produced can be regarded as a candidate disaggregation.

Figure 6 shows two simulations of a disaggregation. Inspection of these and other realisations shows them to be visually more similar to the original data than has been achieved by previous methods, and comparison of quantities such as lag 1 auto-correlations in space and time and proportions of wet pixels also showed good agreement.

## Conclusion

We have seen, *via* two particular examples, that the modelling of rainfall using a latent Gaussian process is both mathematically convenient and effective. Simulations from the fitted models show that realisations are similar to the data and summary statistics generally show better agreement than previous, less elegant models.