Simulation and downscaling models for potential evaporation

C. Yang, R.E. Chandler, V.S. Isham and C. Annoni
(Department of Statistical Science, University College London)
H.S. Wheater
(Department of Civil and Environmental Engineering,
Imperial College London)

July 11, 2003

Abstract

As hydrologists begin to explore the potential of modern computers to simulate hydrological systems over long periods of time, there is an increasing need for high-quality tools that can generate appropriate inputs to these systems. This paper presents some statistical models for generating sequences of potential evaporation, possibly conditioned on rainfall, applied to data from southern England. There have been significant trends in potential evaporation in this area over the period 1961–1993; moreover, the variance of the series shows strong seasonal structure. The methodology presented here provides an intuitively simple means of reproducing such features. Moreover, as well as allowing the generation of simulated sequences, it can be used to produce sequences at a fine timescale from coarse-resolution data.

1 Overview

In simple terms, many hydrological systems can be reduced conceptually to a water balance equation:

\[ \text{Precipitation} - \text{Evapotranspiration} = \text{Runoff} + \text{Change in storage} \]  

(1)

Most hydrological rainfall-runoff models are designed to solve this equation, for a given precipitation sequence (and possibly the associated evapotranspiration).

There is currently a substantial amount of research interest in the use of rainfall-runoff models to simulate hydrological systems over long periods of time. The hope is that this use of ‘continuous simulation’ will yield more insight into the behaviour of a system than
simpler approaches based, for example, on analysis of individual historical events. Continuous simulation does, however, require an ability to simulate realistic input sequences at an appropriate timescale. Consequently there is a substantial body of literature devoted to the development of precipitation models for this purpose. In comparison, attempts to simulate evapotranspiration sequences, which may also be required by rainfall-runoff models, are in their infancy.

The processes governing evapotranspiration are complex, and it can usually be measured only indirectly (Shuttleworth, 1993). For this reason, many runoff models work with potential evaporation (PE) rather than actual evapotranspiration. In the UK, since 1978 the primary source of PE data for hydrological applications has been the Meteorological Office Rainfall and Evaporation Calculation System (MORECS). This uses a modified version of the Penman-Monteith equation (Monteith, 1965) to estimate PE on a $40 \times 40$ km$^2$ grid covering the whole of the UK. For details, see Thompson et al. (1981). The calculations are based on daily values of meteorological variables, along with information on land usage. However, the published output is at a weekly timescale, which is unsatisfactory in those applications for which daily or subdaily values are required. To deal with this problem, the pragmatic approach of assuming PE to be constant throughout the week is widespread in current practice (E. Stewart, personal communication). As a zero-th order approximation this is probably reasonable, since evaporation varies much more smoothly in time than does precipitation, for example. However, the effects of ignoring fine-scale variability in PE sequences are not known.

In the work reported here, models have been developed for both weekly and daily PE sequences. These models are effectively heteroscedastic autoregressions. They have been constructed in such a way as to allow simulation of weekly or daily sequences, and also to simulate realistic daily sequences that are consistent with a given weekly sequence. They are therefore suitable both for use in continuous simulation studies, and for the derivation of realistic daily historical sequences from weekly data.

The methodology is illustrated using data from the south of England. The next section provides a summary of these data. Sections 3 and 4 describe the development of models for weekly and daily PE, as well as analyses of the relationship between precipitation and PE. In Section 5, these models are used to develop a downscaling procedure that is appropriate for generating daily sequences subject to fixed weekly totals; and the work is summarised in Section 6.

2 Data and preliminary analysis

Three separate sources of data have been used in this study:

1. Weekly MORECS PE data (in mm) from a single $40 \times 40$ grid-square in Surrey,
Southern England. The data span the period from January 1961 to March 1993, and are a subset of the Thames catchment data used in Jolley and Wheater (1996). The MORECS output includes PE calculations for a homogeneous grass surface, and for the estimated real land usage. For current purposes, the choice of PE measure is immaterial; the modelling methodology below can be applied in any situation, and we anticipate that the conclusions are unlikely to be sensitive to land use changes. Here, we focus on the ‘grass’ data since these are routinely used as a reference land use (Penman, 1948).

2. Daily rainfall and evaporation data (both in mm) from an experimental station at Silwood Park in Berkshire, which is located just outside the MORECS grid square. These run from July 1989 to August 1994. The data were collected as part of a set of lysimeter experiments carried out by Imperial College (Burne et al., 1994). The rainfall data were gathered using a 0.2mm tipping-bucket gauge; the PE calculations used the MORECS form of the Penman-Monteith equation, with a winter wheat crop.

3. Daily rainfall data (in mm) from a gauge located at Rotherfield Park in Surrey, near the centre of the MORECS grid square. The record spans the period 1923–2000; 1557 of the daily values (around 0.5%) are missing. These data are used to investigate the relationship between PE and rainfall.

The data have been checked for possible errors and outliers; their quality appears reasonable.

Preliminary analyses of the PE sequences revealed that evapotranspiration is higher, and substantially more variable, in summer than in winter. Also, the weekly MORECS sequence shows apparent increasing trends in almost all months of the year (Figure 1). There is, as expected, a relationship between rainfall and PE, although this relationship is weak: the largest cross-correlation between weekly MORECS PE and weekly rainfall at Rotherfield Park is −0.148, at zero lag (this decreases to −0.068 if seasonality is first removed from both sequences). These are preliminary indications only, however: we now proceed to a formal investigation using statistical models.

3 A model for weekly potential evaporation

As indicated above, the primary motivation for this work is the provision of PE sequences, for use in continuous simulation studies. Such studies require bivariate sequences of rainfall and evapotranspiration, incorporating any dependence between the two. Let \( \mathbf{R} = (R_1 \ldots R_T)' \) be a vector of random variables representing a rainfall sequence, and let \( \mathbf{Y} = (Y_1 \ldots Y_T)' \) be the corresponding PE values. Noting that the joint distribution of \( \mathbf{R} \) and \( \mathbf{Y} \) can be
factorised as
\[ f_{RY} (y, r) = f_{Y|R} (y | R = r) f_R (r), \] (2)
a bivariate sequence can be obtained in two stages:

1. Simulate a rainfall sequence \( r \) from a suitable rainfall model.
2. Simulate a PE sequence \( y \) using a model for \( f_{Y|R} (y | R = r) \), the conditional distribution of PE given the rainfall sequence \( r \).

It could be argued that, since PE is calculated deterministically from various meteorological variables, it is not necessary to develop a statistical model for its relationship with rainfall. However, in many applications the required meteorological variables will not available. A statistical model is therefore a convenient device for generating the required sequences.

3.1 Methodology

The PE models developed here are regression-based, and can be written as
\[ Y_t = x_t^T \beta + \varepsilon_t \quad (t = 1, \ldots, T) \] (3)
where $Y_t$ is the potential evaporation at time $t$, $x_t$ is a corresponding vector of covariates, $\beta$ is a vector of regression coefficients and $\varepsilon_1, \ldots, \varepsilon_T$ is an uncorrelated sequence of zero mean errors. Such a model can equivalently be written as

$$\mu_t = E(Y_t|x_t) = x_t'\beta,$$

where $\mu_t$ is the modelled mean of $Y_t$.

The standard theory of linear regression requires that the errors $\{\varepsilon_t\}$ all have the same variance. The PE series do not meet this requirement, because the variance is higher in summer than in winter. This should be reflected in any model designed to simulate PE sequences. An elegant solution is possible when the elements of $\varepsilon$ are normally distributed. Suppose, for example, that $\varepsilon_t \sim N(0, \sigma_t^2)$, where the variance $\sigma_t^2$ depends on the time of year (and possibly on other factors as well). In this case, $\varepsilon_t^2/\sigma_t^2$ has a chi-squared distribution with 1 degree of freedom, which is the same as a gamma distribution with mean 1 and shape parameter 0.5. From standard properties of the gamma distribution, therefore, $\varepsilon_t^2$ has a gamma distribution with mean $\sigma_t^2$ and shape parameter 0.5. This allows changes in $\sigma_t^2$ to be modelled via a generalized linear model (GLM — see McCullagh and Nelder 1989; Dobson 2001):

$$h(\sigma_t^2) = \xi_t'\gamma,$$

where $\xi_t$ is another vector of covariates, $\gamma$ is another vector of parameters and $h(.)$ is a monotonic link function. In the work reported below we take $h(.)$ to be a square root transformation, thus representing the standard deviation $\sigma_t$ as a linear function of predictors. This choice of link is natural from a consideration of measurement units.

We refer to (4) and (5) as the ‘mean model’ and ‘variance model’ respectively, and to the model obtained by combining them as the ‘joint model’. A maximum likelihood fit of the joint model can be obtained using an iterative procedure (McCullagh and Nelder, 1989, Section 10.2), alternating the following two steps to convergence:

1. Fit the mean model using weighted least squares, with weights inversely proportional to the variances given by the variance model. For the first iteration, all the variances are assumed equal.

2. (Re-)estimate the parameters of the variance model, by fitting a gamma GLM to the squared residuals from the mean model.

In the work reported below, convergence was deemed to have occurred when the maximum percentage change in any variance model parameter was less than 0.01.

In the context of time series analysis, this class of mean-variance models can be regarded as an extension of the (G)ARCH family (Fahrmeir and Tutz, 1994, Section 6.1). Chandler (2003) illustrates the application of joint mean-variance models to temperature time
series, and discusses some of the statistical issues that arise when modelling time series using regression methods. The bottom line is that, so long as the covariate vectors \( \mathbf{x}_t \) and \( \xi_t \) incorporate an adequate representation of time series dependence structure, models can be fitted and compared as if the observations were independent. Comparisons can be based on log-likelihoods and proportions of variance explained \( (R^2) \). \( R^2 \) is a familiar measure in regression modelling, but must be applied to the mean and variance components of a model separately — and for the variance component, the concept of ‘explained variance in the squared residuals from the mean model’ is not particularly intuitive. The log-likelihood, on the other hand, accounts for both mean and variance structure simultaneously and therefore provides a single overall measure of model fit. If a model specifies a probability distribution with density \( f_t(y|\mathcal{H}_t) \) for \( Y_t \), conditional on the history \( \mathcal{H}_t \) of the process up to time \( t \), then the log-likelihood, given a sequence of observations \( y_1, \ldots, y_T \), is

\[
L = \sum_{t=1}^{T} \ln f_t(y_t|\mathcal{H}_t) \ .
\]

In the current context, \( f_t \) is the density of a normal distribution with mean \( \mu_t \) and variance \( \sigma_t^2 \); hence (6) is easily computed for any model.

### 3.2 Preliminary modelling

We now describe the development of a model for the weekly MORECS PE sequence. To simplify the modelling task a preliminary linear model, of the form (3) and ignoring heteroscedasticity, was used to suggest an appropriate set of covariates for the mean model. The heteroscedasticity has little effect on the standard \( F \) test for comparing different models in this situation, since there are roughly equal numbers of observations in each month (Montgomery, 1997, Section 3.4).

In our models, we represent seasonal structure using four trigonometric covariates:

\[
\cos \left( \frac{2\pi j \times \text{day of year}}{366} \right) \quad \text{and} \quad \sin \left( \frac{2\pi j \times \text{day of year}}{366} \right) \quad (j = 1, 2) \ .
\]

where ‘day of year’ relates to the end of the week under consideration. Autocorrelation is incorporated by including lagged PE values as covariates. Other possible covariates are rainfall, along with representations of the apparent trends in Figure 1.

The preliminary modelling exercise indicated that all four of the seasonal covariates (7) are required, and that autocorrelation in the PE series can be represented by a first-order autoregressive structure. The effect of rainfall was investigated using various different parametrisations (for example, using simultaneous and lagged rainfall amounts at both weekly and daily timescales). In all cases, the regression coefficients relating to rainfall were non-significant. Moreover, the rainfall added very little to the models’ predictive power. For
example, when rainfall was added to a model containing 4 seasonal predictors and a first-order autoregressive term, the root mean squared error (RMSE) decreased from 2.618mm to 2.614mm. Hence, although rainfall is known to have some role in controlling potential evaporation, for practical purposes this role appears negligible at a weekly timescale. Thus, in developing a joint model for both the mean and variance of potential evaporation, rainfall can be excluded as a potential covariate.

3.3 Development of joint mean-variance model

The above analysis suggests that as a starting point in the development of a joint mean-variance model for the weekly series, the mean of each week’s distribution can be represented as a linear combination of four seasonal covariates and the previous week’s PE. This represents a ‘baseline’ against which to judge the significance of the trends suggested in Figure 1. Three representations of these trends have been investigated here. In the first, a single term was included, implying a common linear trend in all months. In the second, 12 separate terms were included, representing different trends in different months. This representation is conceptually straightforward, but does not allow for the fact that any seasonal variation in the trends is likely to be smooth — for example, we might reasonably expect the trend coefficients themselves to follow a seasonal cycle. Mathematically, this can be achieved by adding a set of interaction terms to the model: these are extra covariates formed by multiplying the trend by each of the seasonal terms. In the final trend model considered here, seasonal variation is represented via such interactions.

In this application, model choice is complicated by the need to select appropriate covariates for both the mean and the variance. To keep things manageable, a reasonable strategy is to consider only models that use the same covariates in each component. This has the potential disadvantage of being parameter-intensive since, for each covariate, two parameters must be estimated. A more parsimonious alternative is to seek relationships between the estimated parameters in the mean and variance components of the models, and to exploit these relationships. For example, all of the models we have fitted show an apparent linear relationship between mean and variance parameters. Therefore, a simple alternative is to represent the standard deviation as a linear function of the mean; this reduces the number of parameters in the variance model to 2.

3.4 Results

The various options set out above give rise to a choice of 8 models: two possible variance structures for each of the four trend scenarios (including the baseline model with no trend). Table 1 summarises the performance of each model in log-likelihood terms, along with $R^2$ values for the mean and variance components separately. The log-likelihoods indicate substantial differences between the models in terms of fit. On the basis of likelihood ratio
<table>
<thead>
<tr>
<th>Trend scenario</th>
<th>Variance structure</th>
<th>$R^2$ for model components</th>
<th>Total number of parameters</th>
<th>Log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>0.8803 0.1187</td>
<td>12</td>
<td>-3595.0</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>0.8817 0.1197</td>
<td>14</td>
<td>-3574.1</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>0.8843 0.1092</td>
<td>14</td>
<td>-3554.2</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>0.8825 0.1324</td>
<td>22</td>
<td>-3559.0</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>0.8806 0.1018</td>
<td>8</td>
<td>-3615.3</td>
</tr>
<tr>
<td>2</td>
<td>B</td>
<td>0.8821 0.1027</td>
<td>9</td>
<td>-3596.5</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>0.8847 0.0926</td>
<td>18</td>
<td>-3584.0</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>0.8829 0.1324</td>
<td>13</td>
<td>-3589.8</td>
</tr>
</tbody>
</table>

Table 1: Performance measures for weekly PE models. Trend scenarios are as follows: 1 = no trend, 2 = same trend in all months, 3 = separate trends in each month, 4 = smooth seasonal variation in trends. Variance structures are: A = standard deviation linearly related to covariates in mean model, B = standard deviation linearly related to mean. For the mean components, $R^2$ is computed from the residuals $\{y_t - \mu_t\}$ in the usual way; for the variance components, it is computed from the variance model residuals $\{\sigma_t^2 - (y_t - \mu_t)^2\}$.

tests (see Chandler and Wheater 2002 for example), the best model is number 4A, in which trends vary smoothly with season and the standard deviation is explicitly represented as a linear combination of the covariates in the mean model. Although model 3A has a higher log-likelihood (-3554.2, compared with -3559.0), for the addition of 14 extra parameters this increase is not significant (the p-value for the resulting likelihood ratio test is 0.796). The log-likelihoods also indicate the necessity of representing the standard deviation explicitly as a linear function of covariates, rather than via dependence on the mean. For example, model 4A offers a log-likelihood increase of 30.8 compared with model 4B — for 9 extra parameters, this is a hugely significant increase (the p-value for this test is around $6 \times 10^{-10}$).

To check the structure of model 4A, Figure 2 presents some residual plots. These are based on the standardised residuals

$$e_t = \frac{y_t - \mu_t}{\sigma_t},$$

which should all have a standard normal distribution if the model is correct. The quantile-quantile plot indicates that the assumption of normality is reasonable except in the upper tail of the distribution where there are 7 outliers; however, as these represent about 0.4% of the sample, for many purposes this can be considered adequate. The second plot, of standardised residuals versus fitted values, reveals no systematic structure, which confirms the adequacy of the assumed linear relationship in the mean model (4).
Figure 2: Residual analysis for model 4A in Table 1. Clockwise from top left: normal quantile-quantile plot of standardised residuals; plot of standardised residuals against fitted values; boxplots showing distributions of standardised residuals by year; boxplots showing distributions of standardised residuals by month.
The monthly and annual boxplots in Figure 2 are intended to check the representation of seasonality and trends in both the mean and variance components of the model. The lack of systematic structure in these boxplots suggest that the model has done a good job. The boxplots also show that the 7 outliers in the quantile-quantile plot occurred predominantly in the early summer, during the mid-1970s.

On the basis of these results, model 4A could be used to simulate realistic sequences of potential evaporation in this part of southern England. The simulated sequences would incorporate heteroscedasticity as well as long-term, seasonally-varying trends. Of course, there is no guarantee that exactly the same results will hold in very different locations. However, the modelling framework is flexible and can be adapted as required; the analyses above provide some guidance regarding a general model-building strategy.

4 Daily evaporation modelling

Models such as those developed above can be used to simulate potential evaporation sequences at a weekly timescale. But, since runoff models require inputs at a different timescale, in this section daily potential evaporation data will be used to develop a downscaling procedure. Of course, if daily data were widely available then they could be modelled and simulated directly in the same way as weekly data. However, in the UK at least, daily PE data are scarce and are generally insufficient, for example, to identify long-term trends. We therefore propose to make use of limited daily data to study sub-weekly structure, and to use this information to downscale weekly sequences. In this way the dual objectives of downscaling weekly data and simulating daily PE sequences can both be achieved, since daily sequences can be simulated by generating a weekly sequence and then downscaling it.

4.1 Daily data analysis

The downscaling procedure developed below is based on the assumption that daily values can be regarded as coming from a multivariate normal distribution. This assumption places some restrictions on the construction of a daily model. The practical consequences of these restrictions have been investigated using the 5-year daily rainfall and PE sequences from Silwood Park (see Section 2). As before, there is a clear seasonal cycle in the variance of daily potential evaporation so that a joint mean-variance model is required; again, linear regression models were fitted first to give a preliminary indication of structure. By and large, the conclusions were the same as in the weekly case except that this time, a highly significant relationship (with p-value less than $2 \times 10^{-16}$) was found between daily rainfall and daily PE. However, in practical terms this relationship has a negligible effect — if rainfall is omitted from the resulting model, the root mean squared error rises from 0.59mm to 0.61mm, and the percentage of explained variance drops from 72.2 to 70.4. Since rainfall has such a tiny effect,
and since the downscaling procedure would be considerably complicated by its inclusion in the daily model, it has not been considered further.

4.2 Joint model fitting

With only about 5 years of daily PE data, it is not possible to identify long-term trends. Thus there are fewer modelling options in this case — after excluding rainfall from the analysis, the only covariates to be considered are the seasonal terms (7), along with previous days’ values. As with the weekly data, all four seasonal terms are required, and autocorrelation is adequately represented by a first-order autoregression. It remains to choose a model for the variance structure.

Table 2 shows three different joint mean-variance models that have been fitted to the daily sequence. In model 1, the standard deviation is linearly related to the mean (variance structure B from the weekly modelling), and in model 2, the standard deviation is a linear combination of covariates in the mean model (variance structure A). However, neither of these models implies a multivariate normal distribution for collections of daily PE values. The reason is the presence of $Y_{t-1}$ in the calculation of $\sigma_t$ (in model 1, this is implicit rather than explicit, since $\sigma_t$ depends on $\mu_t$ here, and $\mu_t$ depends on $Y_{t-1}$). A multivariate normal distribution can, however, be recovered by dropping the $Y_{t-1}$ term from the variance component of model 2; this yields model 3.

The log-likelihoods for the three models are -1121.7, -1113.5 and -1132.7 respectively. On this basis, likelihood ratio tests indicate that model 2 has the best performance and model 3 the worst. However, there is little to distinguish between the models as far as other performance measures are concerned: all 3 models yield an $R^2$ of 70.4% for the mean model component, for example. Moreover, the residual plots for the three models are almost indistinguishable. The plots for model 3 are shown in Figure 3. There are again a few outliers in the upper tail of the quantile-quantile plot; the boxplots show that these occurred primarily in the winter months, which probably explains why they are associated with small fitted values in the ‘residuals vs fitted’ plots. Therefore although model 3 is, in statistical terms, significantly worse than either of the other two, in practical terms it is comparable. We therefore adopt this model, and will use it for downscaling.

5 Downscaling procedure

The preceding analysis indicates that, for practical purposes, any collection of daily PE values can be regarded as coming from a multivariate normal (MVN) distribution according to model 3 in Table 2. We now develop a methodology that uses properties of this distribution to downscale a given weekly PE sequence to a daily timescale. The procedure makes use of a standard result for the MVN distribution (e.g. Krzanowski 1988, Section 7.2). Suppose
Figure 3: Residual analysis for model 3 in Table 2. See caption to Figure 2 for an explanation of the plots.
<table>
<thead>
<tr>
<th>Predictors in model</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mu_t )</td>
<td>( \sigma_t )</td>
<td>( \mu_t )</td>
</tr>
<tr>
<td>Constant</td>
<td>0.8009</td>
<td>0.0786</td>
<td>0.7916</td>
</tr>
<tr>
<td>( \cos \left( \frac{2\pi \times \text{day of year}}{366} \right) )</td>
<td>-0.7692</td>
<td>-0.7606</td>
<td>-0.3140</td>
</tr>
<tr>
<td>( \sin \left( \frac{2\pi \times \text{day of year}}{366} \right) )</td>
<td>0.0954</td>
<td>0.0942</td>
<td>0.0619</td>
</tr>
<tr>
<td>( \cos \left( \frac{4\pi \times \text{day of year}}{366} \right) )</td>
<td>0.0986</td>
<td>0.1029</td>
<td>0.0342</td>
</tr>
<tr>
<td>( \sin \left( \frac{4\pi \times \text{day of year}}{366} \right) )</td>
<td>0.0040</td>
<td>0.0048</td>
<td>-0.0080</td>
</tr>
<tr>
<td>( Y_{t-1} )</td>
<td>0.3914</td>
<td>0.3981</td>
<td>0.0871</td>
</tr>
<tr>
<td>Conditional mean (( \mu_t ))</td>
<td></td>
<td>0.3462</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Coefficient estimates in three joint models for daily PE. All models are fitted to 1882 daily observations.

A random vector \( Y \), distributed as \( \text{MVN}(\mu, \Sigma) \), is partitioned into two subvectors \( Y_1 \) and \( Y_2 \). Then the conditional distribution of \( Y_1 \), given \( Y_2 = y_2 \), is

\[
\text{MVN} \left( \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (y_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12} \right),
\]

where \( \mu_1 \) and \( \mu_2 \) are the unconditional mean vectors of \( Y_1 \) and \( Y_2 \), \( \Sigma_{11} \) and \( \Sigma_{22} \) are their unconditional covariance matrices and \( \Sigma_{12} \) is the cross-covariance matrix between them. Therefore, once the mean and covariance matrix of \( Y \) are known, the conditional distribution (9) can be calculated.

The aim of the current exercise is to construct the distribution of seven daily PE values (one for each day of the week), given a sequence of weekly totals (and possibly daily values for previous weeks that have already been generated by downscaling). In the notation of equation (9), \( Y_1 \) represents the daily values for the current week and \( y_2 \) contains the information that will be used to predict \( Y_1 \):

1. The last daily value from the previous week, since the daily PE model involves a first-order autoregressive term;
2. The total evaporation for the current week; and
3. The total evaporation for the next week, since the autoregressive structure of the daily model implies a correlation between the last day of the current week and the first day of the next week.

To calculate the required conditional distribution, let $Y_t$ denote the PE value on day $t$ and let $Y_{w-1}^+$ denote the PE total for week $w$. Define the vector

$$
\mathbf{Y} = \left( Y_{7w-6} \ Y_{7w-5} \ \ldots \ Y_7 \ Y_{7(w-1)} \ Y_w^+ \ Y_{w+1}^+ \right)'.
$$

The first 7 elements are the daily values in week $w$; the next element is the final day’s value from week $w-1$, and the remaining 2 elements are the weekly totals for weeks $w$ and $w+1$. When downscaling for week $w$, the last 3 elements of $\mathbf{Y}$ will be known (except at the start of a sequence i.e. when $w = 1$, in which case $Y_{7(w-1)}$ will generally be unknown and should be excluded from the definition of $\mathbf{Y}$). Based on this information, we wish to calculate the conditional distribution of the subvector

$$
\mathbf{Y}_w = \left( Y_{7w-6} \ Y_{7w-5} \ \ldots \ Y_7 \right)'.
$$

Using (9), this conditional distribution can be derived from the mean $\mathbf{\mu}$ and covariance matrix $\mathbf{\Sigma}$ of $\mathbf{Y}$. To obtain these, write

$$
\mathbf{Y} = \mathbf{A} \tilde{\mathbf{Y}},
$$

where $\tilde{\mathbf{Y}} = \left( Y_{7(w-1)} \ Y_{7w-6} \ Y_{7w-5} \ \ldots \ Y_7 \right)'$ is a vector of 15 daily values and $\mathbf{A}$ is the matrix

$$
\mathbf{A} = 
\begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}.
$$

Standard multivariate results then yield

$$
\mathbf{\mu} = \mathbf{A} \tilde{\mathbf{\mu}} \quad \text{and} \quad \mathbf{\Sigma} = \mathbf{A} \tilde{\mathbf{\Sigma}} \mathbf{A}',
$$

where $\tilde{\mathbf{\mu}}$ and $\tilde{\mathbf{\Sigma}}$ are the mean and covariance matrix of $\tilde{\mathbf{Y}}$. These properties of the daily PE sequence can be derived for model 3 of the previous section. It is convenient to reparameterise the model at this point, by combining the sine and cosine terms for each seasonal component.
into a single cosine term with a phase shift. For example, the mean component of the model can be written as

\[ \beta_0 + \sum_{j=1}^{2} \beta_j \cos [2\pi (\omega_j t - \phi_j)] + \beta_3 y_{t-1} , \quad \text{where } \omega_j = j/366 \]  

(15)

and (from Table 2) \( \beta_0 = 0.783, \beta_1 = \sqrt{(-0.752)^2 + 0.099^2} = 0.758, \beta_2 = \sqrt{0.108^2 + 0.004^2} = 0.108, \phi_1 = (2\pi)^{-1} \arctan(-0.099/0.752) = 0.479, \phi_2 = (2\pi)^{-1} \arctan(0.004/0.108) = 0.005 \) and \( \beta_3 = 0.404 \) (the \( \phi \) calculations here assume that \( t = 1 \) falls on January 1st). The standard deviation can be rewritten similarly as

\[ \sigma_t = \gamma_0 + \sum_{j=1}^{2} \gamma_j \cos [2\pi (\omega_j t - \psi_j)] , \]  

(16)

where the frequencies \( \omega_1 \) and \( \omega_2 \) are the same as those in (15). In this parameterisation it can be shown (see Appendix) that the expected value of \( Y_t \) is

\[ \mathbb{E}(Y_t) = \frac{\beta_0}{1 - \beta_3} + \sum_{j=1}^{2} \beta_j \left\{ \frac{\cos [2\pi (\omega_j t - \phi_j)] - \beta_3 \cos [2\pi (\omega_j (t + 1) - \phi_j)]}{1 - 2\beta_3 \cos (2\pi \omega_j) + \beta_3^2} \right\} , \]  

(17)

and the covariance between \( Y_t \) and \( Y_{t-k} \) is

\[ \text{Cov}(Y_t, Y_{t-k}) = \beta_3^6 \sum_{p=0}^{6} \gamma_p^* \left\{ \frac{\cos [2\pi (\omega_p^*(t - k) - \psi_p^*)] - \beta_3^2 \cos [2\pi (\omega_p^*(t - k + 1) - \psi_p^*)]}{1 - 2\beta_3^2 \cos (2\pi \omega_p^*) + \beta_3^4} \right\} , \]  

(18)

where

\[ \begin{align*}
\omega_0^* &= \psi_0^* = 0 \\
\omega_1^* &= 2 \gamma_0^* \gamma_1, \quad \omega_1^* = \omega_1, \quad \psi_1^* = \psi_1 \\
\omega_2^* &= \gamma_1^2/2, \quad \omega_2^* = 2\omega_1, \quad \psi_2^* = 2\psi_1 \\
\omega_3^* &= \gamma_2^2/2, \quad \omega_3^* = 2\omega_1, \quad \psi_3^* = 2\psi_1 \\
\omega_5^* &= \gamma_1 \gamma_2, \quad \omega_5^* = \omega_1 + \omega_2, \quad \psi_5^* = \psi_1 + \psi_2 \quad \text{and} \quad \gamma_0^* = \gamma_1^2 + \gamma_2^2 / 2
\end{align*} \]

(19)

The variance of \( Y_t \) is obtained by putting \( k = 0 \) in (18).

Combining all of these results: in a straightforward series of computations, (17)–(19) can be used to calculate the mean and covariance matrix of \( Y_t \), while (14) determines the mean and covariance matrix of \( Y_w \). Hence the required conditional distribution of \( Y_w \) can be derived using result (9). Standard algorithms are available for simulating from such a distribution.

The downscaling scheme has been tested using the daily Silwood Park data from Section 4. The data were first aggregated to a weekly series, which was then downscaled without reference to the original data. Figure 4 shows specimen results for two short periods, in winter and summer respectively. The downscaled sequences here represent single simulations —
Figure 4: Illustration of downscaling scheme performance for two 2-week periods (top: summer, bottom: winter). Open circles represent weekly averages, computed from the observed daily sequences (solid lines). The dotted line in each plot represents a single simulated realisation. Crosses represent the mean of the conditional distribution of daily values, and dashed lines are 95% confidence bands obtained from this conditional distribution.
the aim is not to reproduce the observations exactly, but rather to provide sequences that have a realistic level of sub-weekly variability. In these examples, the calculated conditional distributions for each week have almost constant means — which provides some justification for the common practice (see Section 1) of assuming evaporation itself to be constant. However, this common practice ignores daily variability and, from this perspective, the simulated sequences in Figure 4 are much more realistic. The confidence bands show good agreement between the observed variability and that expected under the downscaling scheme. Note also that the scheme effectively captures seasonal differences in this variability.

These results indicate that the proposed scheme is both viable and successful. It can therefore be used either to downscale historical PE data to a daily timescale or, in conjunction with a weekly simulation model such as that developed in Section 3, to simulate daily sequences.

6 Conclusions

Observed evapotranspiration sequences are characterised by strong seasonal cycles in both mean and variance. These cycles can be described well using joint mean-variance models, which also allow for the inclusion of other systematic structure such as trends. Moreover, this class of models is easy to simulate. Although the methodology has been illustrated using a single dataset, it is completely general and is expected to apply (with suitable modifications to the exact models used) at a wide variety of locations.

A particularly interesting feature to emerge from this work is the existence of significant increasing trends in the studied MORECS evaporation sequence. The magnitude of these trends varies with season. In relative terms, their effect is substantial in winter due to the typically low values of PE during the winter months. In January, for example, the trend corresponds to an increase of around 30% in mean weekly evaporation over the 33-year record; this could have significant hydrological implications. In July however, the increase is only around 3%.

The primary source of potential evaporation data in the UK at present is the MORECS system, which provides weekly values. For applications that require inputs at a finer timescale, a new downscaling methodology has been developed here. This uses the theoretical properties of a joint mean-variance model, to define a multivariate normal distribution for daily potential evaporation given weekly totals. Sampling from this distribution yields daily sequences with the right totals, incorporating a realistic level of sub-weekly variability. The multivariate normal assumption imposes some restrictions on the form of the daily model — for example, relating to the inclusion of rainfall as a covariate, and to the incorporation of lagged observations in the variance component. In principle, these restrictions could be lifted by noting: first, that the method is effectively an application of the Kalman Filter (KF) to recover the elements of an unobserved ‘state vector’ of daily values from an ‘observation
vector’ of weekly values; and second, that approximate versions of the KF can be obtained for general mean-variance models (Harvey, 1989, Section 3.7). However, the practical consequences of the restrictions appear negligible, and the advantages of lifting them would be greatly outweighed by the additional complication involved.

The models developed here also provide the capability of simulating daily evapotranspiration sequences incorporating trends, when weekly data are available but daily data are scarce. This can be achieved by simulating from a weekly model that incorporates trends, and then downscaling the simulated weekly values. In such applications, it may be necessary to use a daily model from a different location, which in turn requires that the daily model is transferrable between locations. For downscaling purposes, the mean structure of this model is largely irrelevant since the daily values will adapt to the weekly totals. The most critical features are those that determine the smoothness of realisations within a week i.e. the seasonal cycle in the variance and the autoregressive coefficient \( \beta_3 \) in (15). These features are unlikely to vary rapidly in space. To demonstrate this, we have fitted model 3 in Table 2 to an 8-year potential evaporation sequence from Moel Cynedd in central Wales. This is approximately 250km from Silwood Park, with an annual precipitation around two and a half times that at Silwood. The estimate of \( \beta_3 \) here was 0.36 (compared with 0.40 at Silwood), and the fitted variance model was almost identical to that in Table 2. In applications, this degree of spatial extrapolation is not to be recommended; nonetheless, the result suggests that the downscaling procedure should be robust, in the sense that parameters obtained from a short data sequence in one location can be regarded as representative of a fairly large geographical area.

Acknowledgements

This research was partially funded by the Department for Environment Food and Rural Affairs, project FD2105.

Appendix: derivation of daily model properties

In this appendix, we derive expressions (17) and (18) for the mean and covariance structure of the selected model for daily PE sequences. The model can be written as

\[
Y_t = \beta_0 + \sum_{j=1}^{2} \beta_j \cos[2\pi (\omega_j t - \phi_j)] + \beta_3 Y_{t-1} + \varepsilon_t ,
\]

(20)

where \( \varepsilon_t \) is distributed as \( N(0, \sigma_t^2) \) independently (but not identically) for each \( t \), and \( \sigma_t \) is given by (16). The parameterisation in (20) is the same as that in (15).
The model is essentially a first-order autoregression, superimposed onto a deterministic seasonal cycle and with the added complication of a deterministic cycle in the variance structure. This (nonstationary) model has some similarities to a stationary process with a mixed spectrum (Priestley, 1981, Chapter 8), whence it would be surprising if the time domain properties of such a process had not been studied previously. However, we have not found any relevant references in the modern time series literature.

To derive the properties of model (20) it is useful to introduce the backshift operator, $B$ (see Priestley, 1981, p.123, for example). The effect of the operator is to shift time backwards one unit: $B(Y_t) = Y_{t-1}$, $B^2(Y_t) = Y_{t-2}$ and so on. An appealing property is that $B$ follows algebraic rules so that series expansions can be carried out. In particular, it is legitimate to write $(1-aB)^{-1} = \sum_{r=0}^{\infty} (aB)^r$ if $|a| < 1$.

In backshift notation, model (20) can be written as

$$Y_t - \beta_3 Y_{t-1} = (1 - \beta_3 B) Y_t = \beta_0 + \sum_{j=1}^{2} \beta_j \cos [2\pi (\omega_j t - \phi_j)] + \varepsilon_t$$

$$= \beta_0 + \mathcal{R} \left\{ \sum_{j=1}^{2} \beta_j \exp [2\pi i (\omega_j t - \phi_j)] \right\} + \varepsilon_t,$$

where $\mathcal{R}(z)$ denotes the real part of a complex number $z$ and $i = \sqrt{-1}$. Multiplying both sides by $(1 - \beta_3 B)^{-1}$ and applying the series expansion given above (which is legitimate providing $|\beta_3| < 1$), we obtain

$$Y_t = (1 - \beta_3 B)^{-1} \left[ \beta_0 + \mathcal{R} \left\{ \sum_{j=1}^{2} \beta_j \exp [2\pi i (\omega_j t - \phi_j)] \right\} + \varepsilon_t \right]$$

$$= \sum_{r=0}^{\infty} \beta_3^r B^r \left[ \beta_0 + \mathcal{R} \left\{ \sum_{j=1}^{2} \beta_j \exp [2\pi i (\omega_j t - \phi_j)] \right\} + \varepsilon_t \right]$$

$$= \beta_0 \sum_{r=0}^{\infty} \beta_3^r + \mathcal{R} \left\{ \sum_{j=1}^{2} \beta_j \exp [2\pi i (\omega_j t - \phi_j)] \sum_{r=0}^{\infty} \beta_3^r \exp [-2\pi i \omega_j r] \right\} + \sum_{r=0}^{\infty} \beta_3^r \varepsilon_{t-r}.$$

The first two terms involve infinite sums of geometric series, so that the equation can be rewritten as

$$Y_t = \frac{\beta_0}{1 - \beta_3} + \mathcal{R} \left\{ \sum_{j=1}^{2} \beta_j \exp [2\pi i (\omega_j t - \phi_j)] (1 - \beta_3 \exp [-2\pi i \omega_j])^{-1} \right\} + \sum_{r=0}^{\infty} \beta_3^r \varepsilon_{t-r}.$$

Simplifying the central expression, and taking the real part, yields

$$Y_t = \frac{\beta_0}{1 - \beta_3} + \sum_{j=1}^{2} \beta_j \left\{ \frac{\cos [2\pi (\omega_j t - \phi_j)] - \beta_3 \cos [2\pi (\omega_j (t+1) - \phi_j)]}{1 - 2\beta_3 \cos (2\pi \omega_j) + \beta_3^2} \right\} + \sum_{r=0}^{\infty} \beta_3^r \varepsilon_{t-r}. \quad (21)$$
Here, $Y_t$ is represented as the sum of a deterministic function and a sequence of uncorrelated terms (the $\varepsilon$s). Its properties can now be derived straightforwardly. For example, taking expectations of both sides of (21) yields (17) directly, since the $\varepsilon$s all have zero mean.

To derive the covariance structure of the model, notice that the only random variables in (21) are the $\varepsilon$s, with $\text{Cov} (\varepsilon_t, \varepsilon_s) = \sigma_i^2$ if $t = s$, and zero otherwise. Hence

$$\text{Cov} (Y_t, Y_{t-k}) = \text{Cov} \left[ \sum_{r=0}^{\infty} \beta_3^r \varepsilon_{t-r}, \sum_{r=0}^{\infty} \beta_3^r \varepsilon_{t-k-r} \right] = \sum_{r=0}^{\infty} \beta_3^{2r-k} \sigma_i^2. \quad (22)$$

Using parameterisation (16) for $\sigma_i$, together with standard trigonometric identities, we have

$$\sigma_i^2 = \left( \gamma_0 + \sum_{j=1}^{2} \gamma_j \cos \left[ 2\pi \left( \omega_j t - \psi_j \right) \right] \right)^2 = \gamma_0^* + \sum_{p=1}^{6} \gamma_p^* \cos \left[ 2\pi \left( \omega_p^* t - \psi_p^* \right) \right],$$

with notation defined in (19). Moving back to complex numbers, and substituting into (22), we obtain

$$\text{Cov} (Y_t, Y_{t-k}) = \Re \left\{ \sum_{r=k}^{\infty} \beta_3^{2r-k} \left[ \gamma_0^* + \sum_{p=1}^{6} \gamma_p^* \exp \left[ 2\pi i \left( \omega_p^* (t-r) - \psi_p^* \right) \right] \right] \right\}$$

$$= \gamma_0^* \sum_{r=k}^{\infty} \beta_3^{2r-k} + \Re \left\{ \sum_{p=1}^{6} \gamma_p^* \exp \left[ 2\pi i \left( \omega_p^* t - \psi_p^* \right) \right] \sum_{r=k}^{\infty} \beta_3^{2r-k} \exp \left[ -2\pi i \omega_p^* r \right] \right\}. \quad (23)$$

As before, the infinite sums here relate to geometric series; the resulting expressions can be simplified, and the real part taken to yield (18).

This derivation illustrates the need to exclude rainfall from the daily evaporation model. Specifically, if rainfall was included then it would have to be treated as a deterministic component of (20); the subsequent manipulation would then lead to an infinite sum of past rainfall values as well as past $\varepsilon$s in (21), which may lead to practical difficulties in calculation.

**References**


