Ecological Modelling, 57 (1991) 27–41
Elsevier Science Publishers B.V., Amsterdam

A serial approach to local stochastic weather models

P. Racsko and L. Szeidl

Eotvos Lorand University, Computing Center 1117, Bogdanfy u. 10 / B, Budapest, Hungary

M. Semenov

Laboratory of Mathematical Ecology, Institute of Atmospheric Physics of USSR Academy of Sciences, Pyzhevski, 3, Moscow, 109017, USSR

(Accepted 16 January 1991)

ABSTRACT

Racsko, P., Szeidl, L. and Semenov, M., 1991. A serial approach to local stochastic weather models. *Ecol. Modelling*, 57: 27–41.

Crop growth simulation models have been studied and constructed by many authors, including the authors of this paper. It was realized that the available local time series of the weather parameters are not numerous or long enough for a good statistical identification of the model's parameters, and they are too short if one aims at generating the multidimensional probability distribution of the output parameters for the stochastic input variables. Thus it was decided to construct a stochastic weather 'generator' that provides as long a time series as is necessary and as many repetitions as the simulation experiments require. The weather 'generator' must, of course, be statistically identical to the observed time series. The paper gives a description of the weather generator developed by the authors. The generator contains a stochastic weather model based on a new approach to weather data analysis and also a computer program package that carries out the tuning of the model on real time series and generates the random weather processes. The model was identified and applied to two geographical locations in Hungary.

INTRODUCTION

Weather simulation models have been developed from the 1950's for various purposes. A good survey is given in Richardson (1981). Most models are constructed from the following simple concepts. The weather process is considered as a Markovian chain with two states – wet days with measurable precipitation and dry days with no or insignificant precipitation. The transition probabilities are derived from the local statistical data.

0304-3800/91/$03.50

Then, other climatic parameters, such as the average temperature, the quantity of the precipitation, the solar radiation, etc. are analyzed and conditional probability distributions are constructed, supposing the two states of the system are Markovian (Richardson, 1981). Because of the annual periodicity of the weather the transition probabilities and the distributions depend on the time of the year.

This modelling philosophy was used for two sets of meteorological data in Hungary – Kompolt from 1951–1985, and Iregszemcse from 1951–1985 – with the purpose of constructing 'weather generators'. The transition probability of the Markovian chain, and the probability distributions of the average daily temperature, the solar radiation, and the precipitation were determined for periods of two weeks, then a Fourier series was fitted to the parameter values in order to smooth the data. As it occurred, the total or average data (average precipitation, average amount of wet and dry days during a given period, average daily temperature and total radiation) derived from the model fitted the observed data set very well.

However, some parameters that are particularly important for the plant growth and development can not be closely approximated from this model scheme in principle. One of these characteristics is the length of dry and wet series, that is series of days with no precipitation and days with significant precipitation without dry periods between them, respectively. The probability of occurrence of long dry or wet series decreases exponentially with the length of the series in a Markovian chain (Feller, 1970). The observed relative frequency of long dry series is significantly higher than the probability, derived from the exponential law (see Fig. 1). The probability of occurrence of a dry series longer then 19 days during a year is approximately 0.05 from this model's distribution, while the observed relative frequency of these series is greater then 0.5! This fact does not contradict the validity of the average or total-type output parameters of the model. For farming, however, long dry series – drought – mean a substantial loss of production. Even if the probability of drought is not very high, the consequences are nevertheless significant, particularly in the Hungarian case study.

Thus for modelling weather sequences we developed a new approach. The basis of our model is the sequence of dry and wet series of days, and other weather parameters like precipitation and temperature are modelled as dependent on the wet or dry series. The main problem we had to solve is to select the type and estimate the parameters of the distributions of weather parameters as they depend on the length and type of the series and position within the series. It is obvious that as the length of the series increases the sample size decrease. That makes it almost impossible to get a reliable approximation of the distribution parameters. Fortunately, after

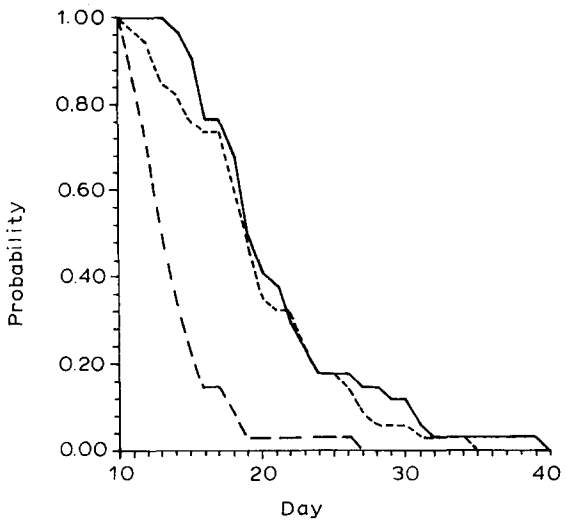

Fig. 1. Probability $P(x)$ of the occurrence of a dry series longer then x days during one year at Kompolt: solid line, measured data; dashed line, Markovian chain model, dotted line, model based on series.

a careful analysis we came to the conclusion that the type and parameters of the distributions of the weather factors do not depend statistically on the length of the series. We have also concluded that the distributions do not depend on the position of the day within the series except for the first day when one type of series is replaced by another. Thus it was satisfying to find the distributions for the first day and the following days of the series separately.

STATISTICAL ANALYSIS OF THE WEATHER PROCESSES IN HUNGARY, SERIAL APPROACH

The four-dimensional weather process (average daily temperature, number of solar hours, precipitation, relative humidity of the atmosphere) measured between 1951 and 1985 was first reduced to a three-dimensional subprocess because of the very high negative correlation between the relative humidity and the temperature. The temperature can 'explain' more then 90% of the stochasticity of the relative humidity. It was decided to retain the average daily temperature as the 'independent' variable and ignore the relative humidity in the further analysis. Thus we consider the weather process as being three dimensional.

Wet series were defined as maximum continuous series of days with precipitation of not less then 0.1 mm, which was the minimal precipitation registered per day.

It is obvious that the observed data set is too small to make a good direct estimate of the probabilities of rare events, for example, the probability of 5 July being a member of a 25 day long dry series. For day d there was selected a characteristic interval $[d-R, d+R]$, where $R>0$ is an integer, and the following hypothesis accepted. In the interval $[d-R, d+R]$ the probability distributions of the lengths of dry and wet series do not change. R must be as large as possible to provide as many statistical data as possible, but it must not be too large, because the hypothesis fails for large values of R. In our case data $R=14$ days was selected. Statistics were then computed for the interval $[d-r, d+R]$ and identified with the statistics for day d.

Let $N^{w}(d)$ denote the total number of wet series, associated with day d, that is the number of wet series in the interval $[d-R, d+R]$ from 1951 to

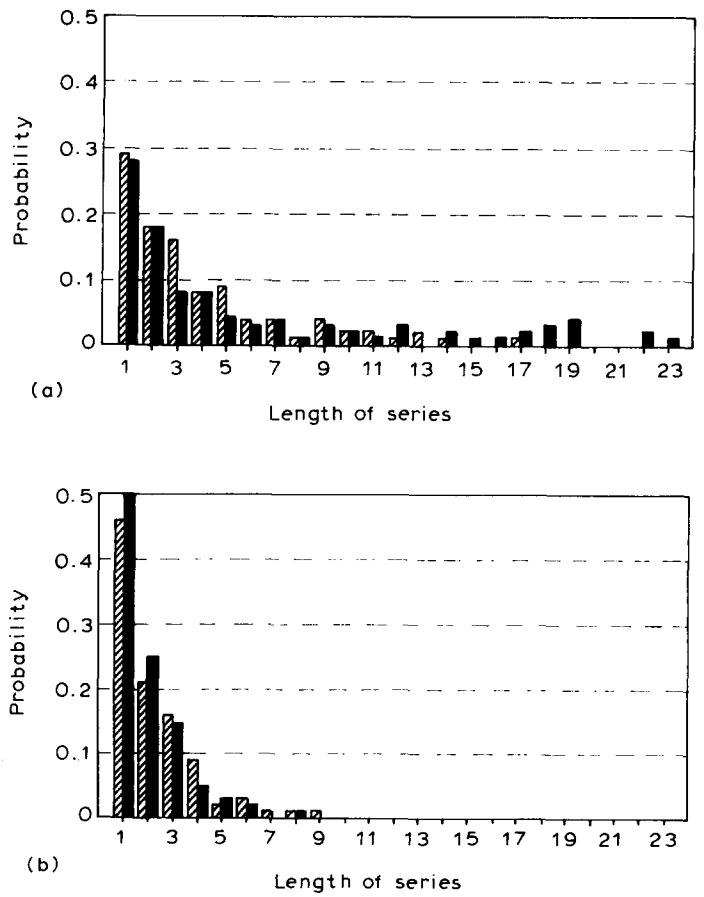

Fig. 2. Probability of the occurrence of (a) dry and (b) wet series of various lengths on a typical selected day in June (striped bars) and September (full bars).

1985. Let $N_i^w(d)$ denote the quantity of wet series of length i, associated with day d. Then, $p_i^w(d) = N_i^w(d)/N^w(d)$ is the probability of occurrence of a wet series of length. i. A similar notation is used for a dry series: $N^d(d)$, $N_i^d(d)$, $P_i^d(d)$. Figure 2 illustrates the graphs of $p_i^w(d)$ and $p_i^d(d)$, which are functions of i, for selected values of d at the meteorological station at Kompolt. The analysis of the empirical distributions at the related meteorological station and all the d values resulted in the following consequences.

The probability distribution of the length of wet series can be well approximated by a geometric distribution. Both χ^2 and Kolmogoroff–Smirnoff tests show a high significance level of the hypothesis about the geometric distribution with the parameter obtained by the maximum-likelihood method.

The probability distribution of the length of a dry series is totally different. It may be approximated by mixing two geometric distributions, with probability $1-p$ for the short series and probability p for the long series (longer then eight days).

The parameters of the geometric distribution $\lambda(d)$ of wet series and dry series were estimated. Then the parameters $\lambda(d)$ was approximated by a finite Fourier series. The Fourier approximation is a general technique when parameters change periodically, which is the case in weather processes.

For day d the parameter $\tilde{\lambda}(d)$ of the geometric distribution is given by the following formula:

$$\tilde{\lambda}(d) = \tfrac{1}{2}a_0 + \sum_{i=1}^{4} \left[a_i \cos(i\omega d) + b_i \sin(i\omega d)\right]$$

where $\omega = 2\pi/T$, $T = 365$ and a_i and b_i are the Fourier coefficients of the function $\lambda(d)$.

The probability $p(d)$ of occurrence of a long dry series on day d was again approximated by a finite Fourier series $\tilde{p}(d)$. The observed values and the Fourier fit are shown in Fig. 3.

Quantity of daily precipitation

The most difficult task in our weather analysis was the modelling of the quantity of precipitation, as it is necessary to start with developing appropriate models for series of different lengths, and for different days within the series.

Let us denote the length of a series by L and the index of the day within the series by l. Figures 4a and b shows the idea of starting with an exponential distribution for a given pair of (L, l). For a short series $(L < 4)$

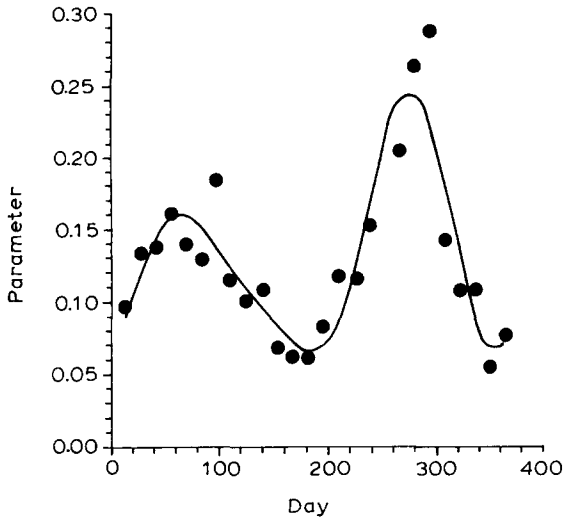

Fig. 3. Probability of occurrence of dry series longer then eight days at Kompolt. Solid line, Fourier fitting; points, empirical frequencies.

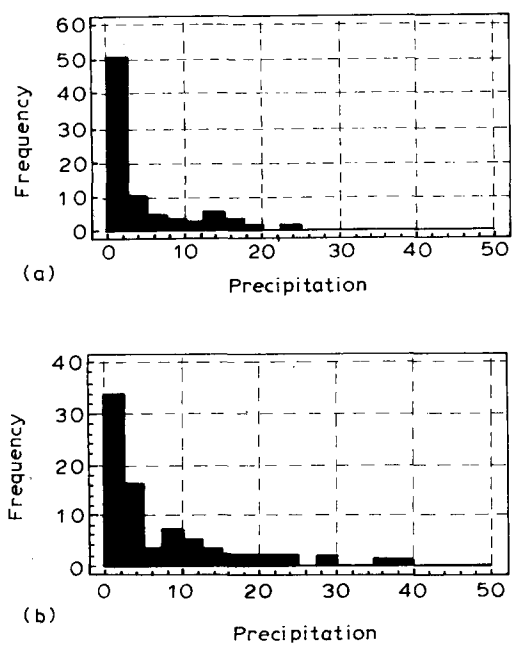

Fig. 4. Frequency histogram of the precipitation on a one day long series at Kompolt: (a) on the 105th day; (b) on the 135th day.

we obtained similar pictures for all days d. A significant difference between the exponential distribution and the observed data was obtained in the neighborhood of zero, (the observed frequency of small quantities of precipitation is significantly higher than that predicted from the exponential distribution) and at the tail of the distribution, where the data are obviously not distributed exponentially. A possible model of the quantity of precipitation might be a mixture of three distributions – one for a small amount of one a medium amount of and one for large amount of precipitation.

At this point we met the following theoretical problem, connected not only with the modelling of precipitation, but also with the other parameters. There must be a separate analysis and estimate made for every 3-tuple (L, l, d). Obviously, as the number of observations decreases very fast with the growth of L, it is impossible to make significant estimates for long series because of the lack of data. Thus it was decided to transfer the results of the analysis of short series to long ones. The philosophy behind this was that the weather parameters are more or less stable within a wet or dry series, they change significantly only at the end of the series. The data do not contradict our hypothesis.

The quantity of precipitation was analyzed for one-, two-, and three-day series. The precipitation was clustered into three groups – small (< 0.4 mm), medium, (0.4–20 mm) and large (> 20 mm) quantities per day. The probability of each group was estimated from the relative frequencies for each pair (L, l). The analysis of the probabilities showed that the dependence on L and l is very weak, they depend only on d. Thus, the probability of three groups were estimated independently of L and l, then a Fourier series fitting was made to smooth out the data. Figure 5 illustrates the Fourier curves.

The distribution of the small precipitation group is very close to the uniform, independently of L and l.

Large amounts of precipitation are very rare in Hungary, in fact there are not enough data to find an appropriate distribution. Large precipitation was modelled by its average value.

The precipitation of the medium group was approximated by an exponential distribution. The approximation was tested statistically and accepted for one-, two-, and three-day series. The parameter of the exponential distribution $\lambda(d)$ was estimated by the average precipitation in this group, $\lambda(d) = 1/pr_m(d)$. $Pr_m(d)$ weakly depends on L and l, thus this dependence was ignored.

After having finished the modelling of the daily precipitation the serial autocorrelation of the precipitation time series was analyzed. In this case

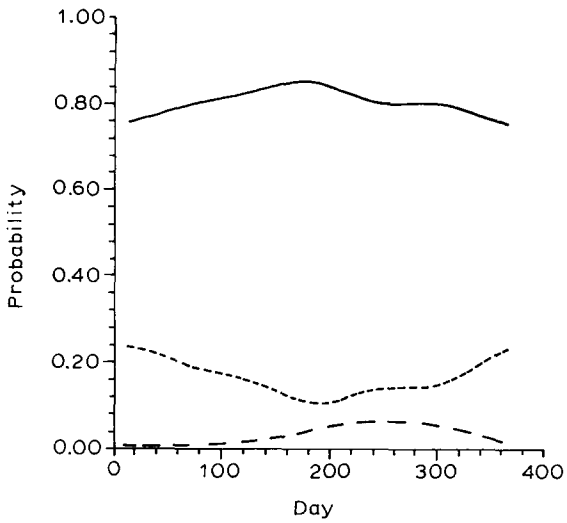

Fig. 5. Fourier approximation of probability of the occurrence of the precipitation group: small, dotted line; medium, solid line; large, dashed line.

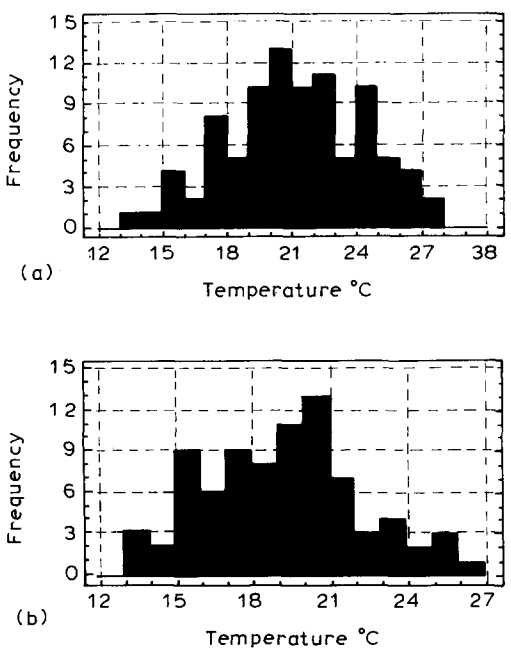

Fig. 6. Frequency histogram of the daily average temperature on a one-day long wet series at Kompolt: (a) on the 195th day; (b) on the 285th day.

there is only a very weak – ignorable – autocorrelation in the precipitation time series.

Daily average temperature

The analysis of daily average temperature was carried out for dry and wet series separately. Figures 6a and b show frequency histograms of the daily average temperature for various d, for all L and l. A normal distribution $\mathcal{N}(M, \sigma)$ is a good fit to the data. It is obvious that for wet and dry series $M = M_t^x(d, L, l)$ and $\sigma = \sigma_t^x(d, L, l)$, where $x = \mathrm{w}$ for a wet series and $x = \mathrm{d}$ for a dry series, and index t stands for the temperature.

As was predicted, both the average daily temperatures and the stochastic behavior differ significantly during wet and dry series. For example, in summer during a wet series the temperature decreases while during a dry series it increases as l increases. The average daily temperature does not depend on the length of the series L either for a wet or a dry series, only on the position in the series l. In addition to this it was observed that average temperatures for $l > 1$ are identical in both types of series and consequently

$$M_t^x(d, L, l) = \begin{cases} M_t^x(d, 1), \text{ if } l = 1 \\ M_t^x(d, 2), \text{ if } l > 1 \end{cases}$$

The same holds true for the variances.

The significant simplification described above made possible the computation of the parameters of the distribution. Then, as usual, the data were smoothed by a Fourier fitting. (Fig. 7). It is worth mentioning that consecutive days are highly correlated within the series (the correlation coefficient is 0.8).

Solar hours

As the radiation measured in solar hours can be directly transformed into physical units, the statistical analysis was carried out for the original data set.

We obtained the same basic results for the solar hours analysis as for the average temperatures – weak dependences on L and l. The difference from the temperature analysis occurs in the winter period for wet series, when zero has a very high probability, and ruins the normality.

Thus, the solar hours were modelled using a mixed distribution – the zeros were separated from the other data with a given probability, while the remaining data were described by a normal distribution, or a normal distribution with an accumulation of negative values in the zero.

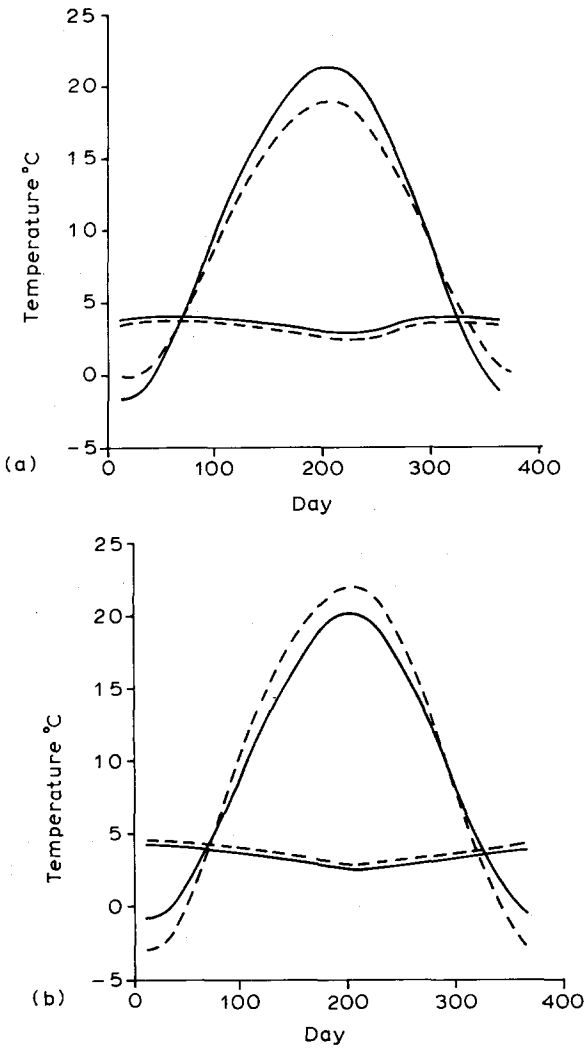

Fig. 7. Fourier approximation of daily average temperatures (bell-shaped lines) and standard deviations (level lines) at Kompolt for the first day of a series of any length (solid line) and for any other day of series (dashed line); (a) wet series; (b) dry series.

The autocorrelation found in the time series was very low, having its maximum of about 0.25 for consecutive days.

Analysis of the common distribution of precipitation, daily average temperature and solar hours

The weather process has to be treated as a multidimensional stochastic process rather then a set of parallel independent processes. Thus further

analysis was carried out to describe the interdependence of the three processes analyzed.

First, the two-dimensional distribution of temperature and solar hours was examined. As expected, the higher the temperature, the greater the number of solar hours, and this ratio is very definite for wet series.

The two distributions – precipitation and solar hours – show a negative correlation, which is, however, not very significant.

The analysis of temperature as a function of precipitation showed the independence of the two variables within the wet series.

Without doubt, the best weather model would be a three-dimensional stochastic process, and not a 3-tuple of independent variables. Unfortunately, the relatively short time series and the non-standard type of the three-dimensional distribution make that impossible.

STOCHASTIC WEATHER MODEL

Now we summarize the model for the daily precipitation, the average temperature and the number of solar hours.

Let $P_w(d)$ and $P_d(d)$ denote the probability distribution of the lengths of wet and dry series on day d, respectively.

At $d = 1$ the status of the system is generated first (a wet or a dry series), then the length of the series, for example, n_w. Then, the period $[d, d + n_w]$ is considered to be wet. According to the definition of the series, each wet series is followed by a dry one. Now we generate and value from the distribution $P_d(d)$, and the period $[d + n_w, d + n_w + n_d]$ is considered dry. This process is repeated until the end of the year.

For a wet series $P_w(d) = \text{Geom}(\tilde{\lambda}(d))$, while for a dry series:

$$P_d(d) = \begin{cases} \text{Geom}\left(\tilde{\lambda}_s^d(d)\right) & \text{with probability } 1-p \\ \text{Geom}\left(\tilde{\lambda}_l^d(d)\right) & \text{with probability } p \end{cases}$$

where Geom( ) means a geometric distribution for a short series, $\tilde{\lambda}$( ) is a fitted parameter of the distribution; w, wet series; d, dry series; s, short series; l, long series.

After we have generated the wet–dry layout of the year, the precipitation is modelled.

The distribution of the precipitation P_p( ) depends on d, but does not depend on L or l. The distribution function is mixed:

$$P_p(d) = \begin{cases} \text{UNI}(0, 0.3) & \text{with probability} \quad q_s(d) \\ \text{EXP}\left(\tilde{\lambda}(d)\right) & \dots \qquad\qquad\quad q_m(d) \\ M(d) & \dots \qquad\qquad\quad q_1(d) \end{cases}$$

where $q_s(d) + q_m(d) + q_l(d) = 1$ for each d, the probabilities of occurrence of 'small', 'medium' and 'large, precipitation; UNI, uniform distribution of small precipitation; EXP, exponential distribution of 'medium' precipitation; $M(d)$, average 'large' precipitation.

The distribution of temperature $P_t(\ )$ is modelled by a normal distribution with parameters $M_t^x(d, l)$ and $\sigma_t^x(d, l)$ where x stands for w, a wet series, and d, a dry series. The distribution of the temperature is:

$$P_t(d) = M_t^x(d, l) + \sigma_t^x(d, l)R_t(d)$$

where

$$R_t(d) = aR_{t-1}(d) + bF(0, 1)$$

Here R_t is the correlation coefficient between two consecutive days, $F(0, 1)$ is the Gauss function with parameters 0 and 1, a, b, with $a^2 + b^2 = 1$ are parameters providing the standard normal distribution for R_t.

The number of solar hours $P_h(\ )$ are also modelled as a normal random variable, with parameters that are dependent on the type of the series of day d, and the position l within the series. Because of the weak correlation on consecutive days, their dependence was ignored:

$$P_h(d) = M_h^x(d, l) + \sigma_h^x(d, l)F(0, 1)$$

where M_h^x and σ_h^x are parameters of the normal distribution of solar hours.

As 'solar hours' is always a non-negative value, any numbers generated below zero were replaced by zero.

MODEL ANALYSIS AND TESTING

Weather models were constructed and identified for two Hungarian meteorological stations, Kompolt and Iregszemcse. Then, several tests were carried out to determine the operational characteristics of the model. Those output parameters, mostly influencing the development of the plants were most carefully tested.

The first group of tests was computing the average weather parameters during the vegetation period of a plant, for example, maize. Figure 8 shows the average temperature sums from April to August, for observed and generated data. Temperature sums are a kind of biological time unit and play a significant role in plant growth. Figure 9 illustrates the measured and generated monthly precipitation, and the mean for the period April–August. Figure 10 shows the measured and generated numbers of solar hours.

In terms of the temperature sums, monthly precipitation and number of solar hours the weather generator reproduces the measured data very well.

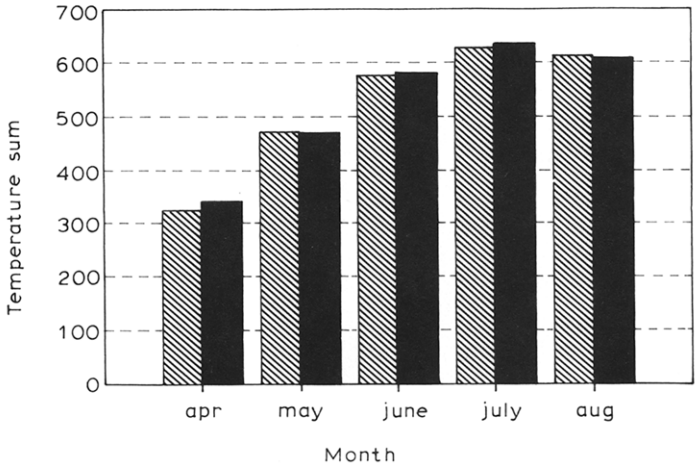

Fig. 8. Monthly average of temperature sums at Kompolt: full bars, model data; striped bars, empirical data.

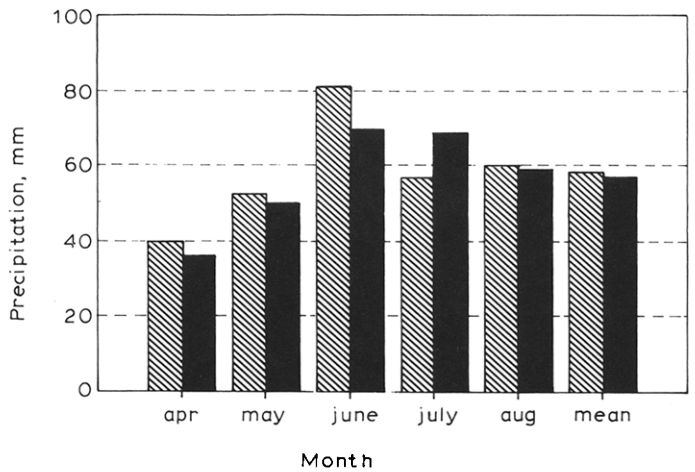

Fig. 9. Monthly average of precipitation at Kompolt: full bars, model data; striped bars, empirical data.

Another group of tests was carried out for the parameter values that were critical for the plant growth. The most important of these is the probability of occurrence of a long dry series. In Fig. 1 the graph $P(x)$ shows the probability of occurrence of a dry series longer then x.

The maximum and minimum temperatures during the vegetational period are critical for the plant. Both high and low temperatures may damage or slow down the plant's development. The model experiment gave good

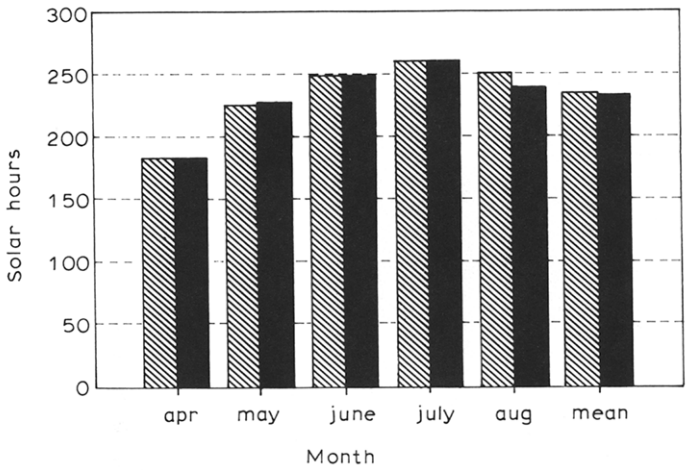

Fig. 10. Monthly average of solar hours at Kompolt. full bars, model data; striped bars, empirical data.

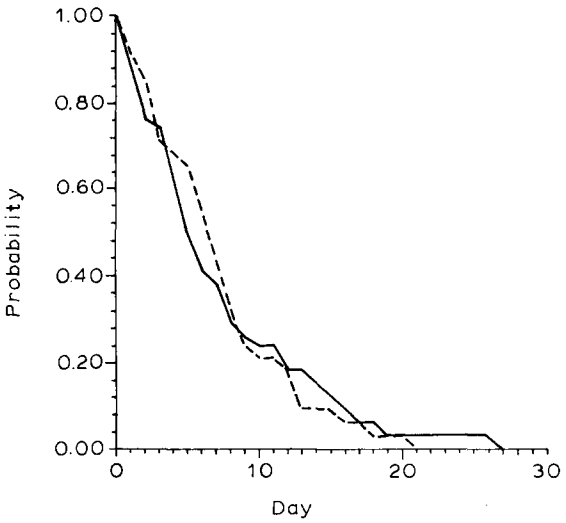

Fig. 11. Probability $P(x)$ of the event that the number of days with average temperature above 25 °C exceeds x at Kompolt; solid line, empirical data; dashed line, model data.

results in this case as well. The test of the quantity of 'hot' days (days with average temperature above 25 °C) was also successful (Fig. 11).

CONCLUSIONS

The stochastic weather model was constructed first of all for the risk analysis of crop production, where the risk is associated with the stochastic

weather conditions. The model gives a statistically adequate weather prognosis for any time interval, which in turn can be applied to the plant growth model CROP (Racsko and Semenov, 1989). The coupled weather and CROP models provide the tool of the building probability distribution of crop production at different agrotechnologies and at any time during (or before) the vegetational period. The distribution of the weather parameters can be refined as time goes an and the exact information about the past values is available. The approach described above might be called an 'adaptive risk assessment'.

The stochastic weather model is realized as a user-friendly computer software package on an IBM compatible PC. The software helps in the analysis of the weather sequences and in the estimation of distribution parameters at any particular geographical area supposing that appropriate time series of the measurements are available. The program also calculated the Fourier coefficients and creates the parameters file for that particular geographical location. Then, the user can start generating weather sequences, statistically identical to the measured time series. The program also provides statistical testing facilities for the simulated data.

REFERENCES

Racsko, P. and Semenov, M., 1989. Analysis of mathematical principles in crop growth simulation models. Ecol. Modelling, 47: 291–302.

Richardson, C.W., 1981. Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resour. Res., 17: 182–190.

Feller, W., 1970. An Introduction to Probability Theory and Its Applications. Wiley, New York.