INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

SMR.478 - 26

THIRD AUTUMN COURSE ON MATHEMATICAL ECOLOGY

(29 October - 16 November 1990)

"A Stochastic Decision Support Tool for Crop Production"

P. Racsko
Eotvos Lorand University
Computer Centre
Budapest 1117
Hungary

These are preliminary lecture notes, intended only for distribution to participants.

A STOCHASTIC DECISION SUPPORT TOOL

FOR CROP PRODUCTION

(PETER RACSKO)

(EOTVOS LORAND UNIVERSITY)

BUDAPEST

The following methods and models were developped by P. Racsko, M. Semenov, L. Szeidl /Budapest/

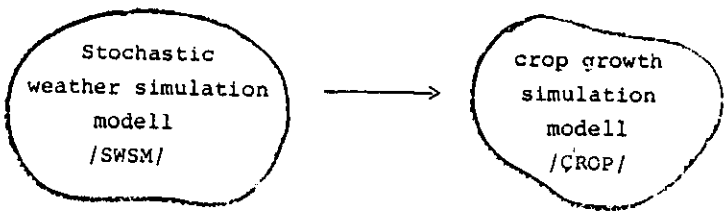
Introduction

Agricultural production in most countries heavily depends on climatic conditions. Even in those countries with relatively well developped agriculture as Hungary, extremely unfavourable but not improbable weather conditions may cause very serious damages. In Hungary due to the long dry period in July and August, 1990 the total yield of maize is 30% less then expected.

However, farmers usually apply various management practices /e.g. application of nitrogen fertilizer, irrigation, etc./ in order to reduse the damage.

Those practices mentioned above, of course, are expensive and reduce the farmer's profit. We thought it was necessary to develop a not too complex, easy to use decision sopport tool to help to analyze the data available and to eveluate the risk of the decisions.

The decision support mechanism consists of two subsystems - a stochastic weather summulation modell and a crop growth modell, both installed on a personal computer.



The two subsystems interact dynamically in the following way:

Suppose the parameters of both SWSM and CROP models are identified at a definite geographical location.

On day d we are given the identified models and two sets for empirical data, one of them is the time series of weather and other environmental parameters measured from the sowing date of the investigated crop plantation, while the other is the time series of the crop growth parameters, e.g. biomass of the roots, leaves, as a multidimensional stochastic process, represented by the multidimensional time series of weather parameters, e.g. daily precipitation, daily avarage temperature, etc. The crop growth is considered as a deterministic process, which depends on the stochastic weather process.

As the weather process is known until day d, the time series of the parameters, describing the crop growth process can be computed.

Thus, the variability of the crop production is determined by the stochastic weather process.

Suppose we estimate the parameters $A = A(a_1 \dots a_n)$ of the crop production. /e.g. final yield/ on day d on the basis of all empirical information we have.

If the estimation of A is repeated on day d^*, where $d^* > d$, more exact statistical values can be expected.

Now how to improve dynamically our estimations?

Let d_0 and d_T denote the beginning and the end of the vegetational period of a plant.

Let us generate at d_0 a sufficiently great number of possible weather processes with the SWSM, and use the generated weather prcesses as input for CROP, and collect the simulated data. Now we are given a set of sample data for each parameter of our interest, and we can make the statistical analysis required for the decision.

If the model is good enough, the simulation results are statistically identical to the empirical data set, collected for several years.

Now repeat the simulation on day d^*, $d_o < d^* < d_T$. As far as the weather process between d_o and d^* is known, obviously the variability of the generated weather trajectories between d^* and d_T is less, then between d_o and d_T. As a consequence, the estimation of A is improved.

What happens if the estimated final yield substantially differs from the value, expected at d_o?

The CROP model has control parameters, corresponding to agro-technological management practices.

Having generated the set of weather prcesses one can analyze how the management practices influence the final outcome and find - if possible - those agrotechnological operations which drive back the estimated outcome to the favourable area.

The above decision making scheme requires a weather simulation and a crop growth simulation model.

Two moderately complex but easy to use models are described below.

1. The weather simulation model

Weather simulation models have been developed from the 1950 - ies for various purposes. Good surveys are given in /Richardson, 1981/. Most models are built on 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. Then, other climatic parameters, as average temperature, quantity of the precipitation, solar radiation, etc. analyzed and conditional probability distributions are constructed, supposing two Markovian states of the system /Richardson, 1981/. Due to the annual periodicity of the weather the transition probabilities and the distributions depend on the period of the year.

This modelling philosophy was used for two sets of meteorological data in Hungary - Kompolt: 1951 - 1985, and Iregszemcse: 1951 - 1985 - with the purpose of constructing "weather generators". The transition probabilities of the Markovian chain, and the probability distributions of daily average temperature, solar radiation, and precipitation were determined for two-week preiods. Then a Fourier series was fitted to the parameter value in order to smooth the data. As it occurred, the total or average type data /average precipitation, average amount of wet and dry days during a given period, average daily temperature and total radiation/ derived from the model very well fit the observed data set.

However, some parameters particularly important for the plant growth and development can not be closely approximated from this model scheme in principle. One of those characteristics is the length of dry and wet series, i.e. series of days with no precipitation and days with significant precipitation without dry periods between. The probability of occurrence of long dry or wet series exponentially decreases with the length of series in Markovian chain /Feller, 1970/.

The observed relative frequency of long dry series is significantly higher than probability, derived from the exponential law /see Fig.1/. The probability of occurrence of a dry series longer the 19 days during a year is approximately 0.03 from the model's distribution, while the observed relative frequency of those series is greater the 0.5! This fact does not contradict the goodness of the average or total type output parameters of the model. For farming however, long dry series - drought - mean substantial loss in the production. Even if the probability of drought is not very high, nevertheless the consequences are significant, particularly in 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 estimates the parameters of 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 sizes decrease. That makes almost impossible to get a reliable approximation of the distribution parameters. Fortunately, after a careful analysis we came to the conclusion, that the type and parameters of the distribution of the weather factors statistically do not depend on the length of series. We have also concluded that the distribution do not depend on the position of the day within the series, except 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.

1.1. Statistical analysis of the weather processes in Hungary, serial approach.

The four dimensional weather process /daily average temperature, 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 temperature. It has occurred that temperature can "explain" more the 90% of the stochasticity of relative humidity. It was decided to keep daily average temperature as "independent" variable and ignore relative humidity in the further analysis. Thus we consider the weather process as three dimensional.

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

It is obvious that the observed data set is too small to make a good direct estimation of the probabilities of rare events. For day d there was selected a characteristic interval [d-R,d+R], where $R > 0$ integer, and the following hypothesis accepted. In the interval [d-R,d+R] the probability distributions of the length 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 hypotheses fails for large R-s. In our case R = 14 days was selected. Then, statistics were computed for the interval [d-R,d+R] and identified with the statistics for day d.

Let $N^W(d)$ denote the total quantity of wet series, associated with day d, i.e. the quantity of wet series in the interval [d-R,d+R] from 1951 to 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. Similar notation is used for dry series: $N^d(d)$, $N_i^d(d)$, $p_i^d(d)$. Figure 2. illustrates the graphs of $p_i^w(d)$ and $p_i^d(d)$ functions of i for selected values of d at the meteorological station Kompolt. The analysis of empirical distributions at the related meteorological station and all d values resulted in the following consequences.

The probability distribution of the length of wet series can be well approximated with 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 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 long series /longer than 8 days/.

The parameters of geometric distribution $\lambda(d)$ of wet series and dry series are estimated. The parameter $\lambda(d)$ 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 formula:

$$\tilde{\lambda}(d) = a_0/2 + \Sigma_{i=1}^{4}(a_i\cos(i\omega d) + b_i\sin(i\omega d))$$

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 long dry series on day d again were approximated by a finite Fourier series $\tilde{p}(d)$. Observed values and Fourier fit see on Fig.3.

<u>The daily precipitation quantity.</u>

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

Let us denote the length of a series by L and the index of the day within the series l. Figures 4 a,b give the idea of starting with an exponential distribution for a given pair of (L,l). For short series ($L < 4$) we obtained similar pictures for all d's. Significant difference between exponential distribution and observed data was obtained in the neighbourhood of 0, /the observed frequency of small precipitation quantities is significantly higher than it is predicted from the exponential distribution/ and at the tail of the distribution, where the data are obviously not distributed exponentially. A possible model of precipitation quantity might be the mixture of three distributions - one for little, one for medium and one for large precipitation,

At this point we met the following theoretical problem, connected not only with the modelling of precipitation, but also with other parameters. There must be made separate analysis and estimation for every 3-tuple (L,l,d). Obviously, the number of observations decreases very fast with the growth of L, it is impossible to make significant estimations for long series because of the lack of data. Thus it was decided to transfer the results of analysis of short series for long ones. The philosophy behind that was, that the weather parameters are more or less stable within a wet or dry series, they change significantly only at the end of series. The data do not contradict to 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) quantity per day. The probability of each group was estimated from the relative frequencies for each pair (L,l). The analysis of 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 the data. Fig.5 illustrates the Fourier curves.

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

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

The precipitation of the medium group is approximated by exponential distribution. The approximation was statistically tested and accepted for one, two, and three-day series. The parameter of exponential distribution $\lambda(d)$ was estimated by the average precipitation of 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 daily precipitation the serial autocorrelation of the precipitation time series was analyzed. As it occurred, there is only a very wear - 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. Fig.6 a,b show frequency histograms of the daily average temperature for various d. A normal distribution $N(M,\sigma)$ can be well fitted to the data. It is obious, that for wet and dry series

$M = M_t^x(d,L,l)$ and $\sigma = \sigma_t^x(d,L,l)$, where x = w for wet and x = d for dry series, index t stand for the temperature.

As it was predicted, both daily average temperatures and the stochastic behavior differ significantly during wet and dry series. In summer during the wet series the temperature decreases while during dry series increases with the increase of l. The daily average temperature neither for wet, nor for dry series depends on the length of series L, only on the position l in the series. Besides that, it was observed that average temperatures for l > 1 are indentical in both types of series. Consequently

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

Same holds for the variances.

The significant simplification described above made possible to compute the parameters of the distribution. Then, as usual, data were smoothed by a Fourier fitting. /Fig.7/. It is worth to mention that consequent days are highly correlated within the series /correlation coefficient 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 solar hours analysis, as for the average temperature - weak dependence on L and l. The difference from the temperature analysis occurs in the winter period for wet series, when 0 has a very high probability, and ruins normality.

Thus, solar hours were modelled using a mixed distribution - the 0-s were separated from other data with a given probability, while the remaining data described by a normal distribution with accumulation of negative values in the zero.

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

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

Weather process has to be treated as multidimensional stochastic process rather then a set of parallel independent processes. Thus further analysis was carried out to describe the interdependence of the three analyzed processes.

First, the two dimensional distribution of temperature and solar hours was examined. No surprise, the higher the temperature, the more the quantity of solar hours, and this ration is very well expressed for wet series.

The two dimensional distribution of precipitation and solar hours shows negative correlation, however not explicit.

The analysis of temperature vs. precipitation showed the independence of the two variables.

No 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 3-dimensional distribution make it impossible.

1.2. The stochastic weather model.

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

Let Pw(d) and Pd(d) denote the probability distribution of the length of wet and dry series on day d, respectively.

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

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

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

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

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

The distribution of precipitation depends on d, but do not depend on L or l. The distribution function is mixed:

$$P_p(d) = \begin{cases} \mathrm{UNI}(0,\ 0.3) \text{ with probability } p_s(d) \\ \mathrm{EXP}(\tilde{\lambda}(d)) \quad \dots \quad p_m(d) \\ M(d) \quad p_l(d) \end{cases}$$

where $p_s(d) + p_m(d) + p_l(d) = 1$ 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 temperature is modelled by a normal distribution with parameters $M_t^x(d,l)$ and $\sigma_t^x(d,l)$ where x stands for w - wet and d - dry series. The dispersion of the temperature is

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

where

$$R_t(d) = a^* R_{t-1}(d) + b^* F(0,1)$$

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

Solar hours are also modelled as normal random variable, with parameters, dependent on the type of the series on day d, and the position l within the series. Due to the wear correlation on consequent 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, the generated below zero numbers were replaced by 0.

1.3. Model analysis and testing.

The weather model was constructed and identified for two Hungarian meteorological stations, Kompolt and Iregszemcse. Then, several tests were carried out to find operational characteristics of the model.

First the average weather parameters were tested during the vegetation period of a plan /maize/ Fig.8. shows the average temperature sums from April to August, for observed and generated data. Temperature sum is a kind of biological time and plays a significant role in the plant growth. Fig.9 illustrates the measured and generated monthly precipitation, and the mean for the period April - August. Fig.10 shows the measured and generated solar hours.

In terms of sums of effective temperatures monthly precipitation and solar hours the weather generator very well reproduces the measured data.

Another group of tests was carried out for parameter values, critical for the plant growth. The most important one is the probability of occurrence of long dry series. On 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 critical for the plant. Both high and low temperatures may damage or slow down the plant's development. Model experiments gave also good results in this case.

The best of the quantity of "hot" days /days with average temperature above 25 C^o also was successful /Fig.11/.

2. For the crop growth simulation modell see /Racsko, Semenov, 1989/.

References

Racsko., Semenov V., 1989. Analysis of Mathematical Principles in Grop Growth Simulation Models. - Ecological Modelling, vol.47, p.291-302.

Richardson C.W., 1981. Stochastic Simulation of Daily Precipitation, Temperature, and Solar Radiation. - Water Resources Research, vol. 17, N.1. p.182-190.

Feller W., 1970. An Introduction to Probability Theory and Its Application. - J.Wiley & Sons, New York.

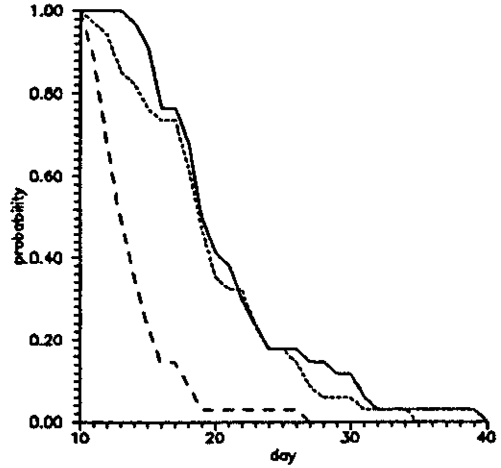


Fig. 1

Probability P(x) of the occurrence of dry series longer than x days during one year at Kompolt: solid line - measured data, dashed - Markovian chain model, dotted - model based on series.

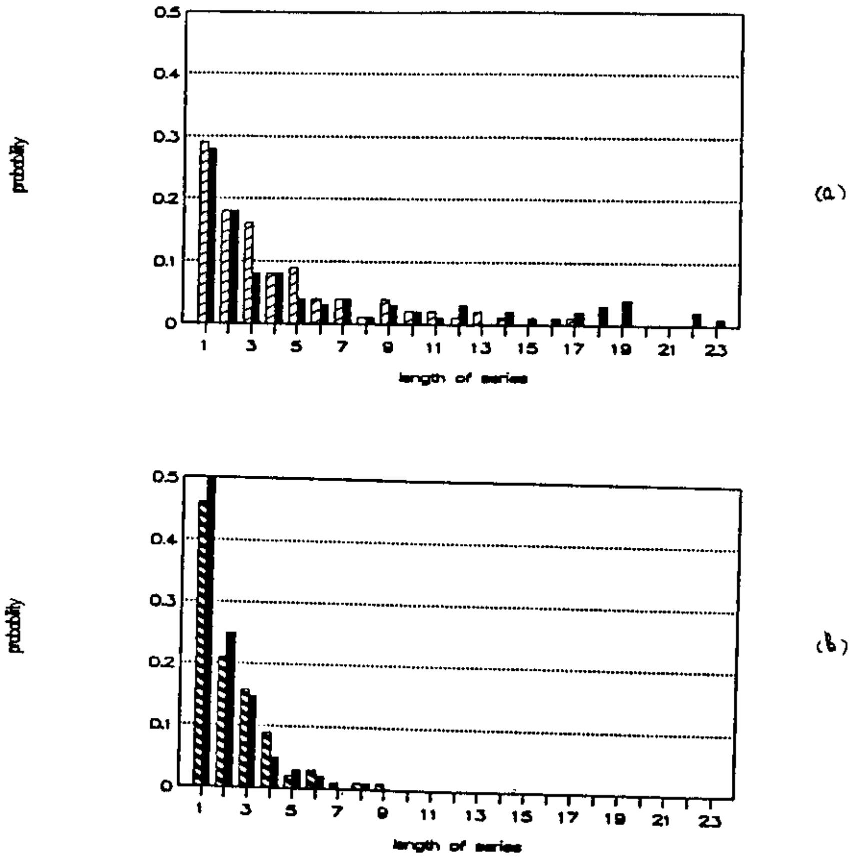


Fig.2 (a,b)

Probability of the occurrence of dry (a) and wet (b) series of varios length on a typical selected day in June and September.

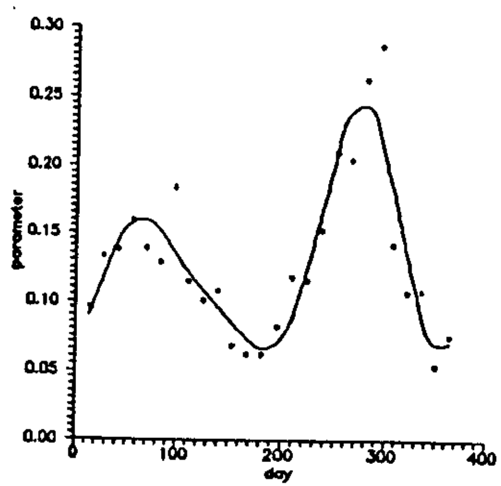
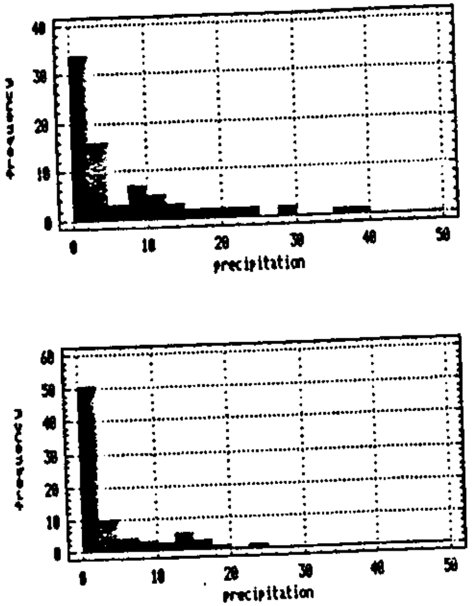


Fig. 3.

Probability of occurrence of dry series longer then 8 days in Kompolt. Solid line - Fourier fitting, points - empirical frequencies.



(a)

(b)

Fig. 4 (a, b)

Frequency histograms of the precipitation on 1-day long series in Kompolt: (a) on 105th day, (b) on the 135th day

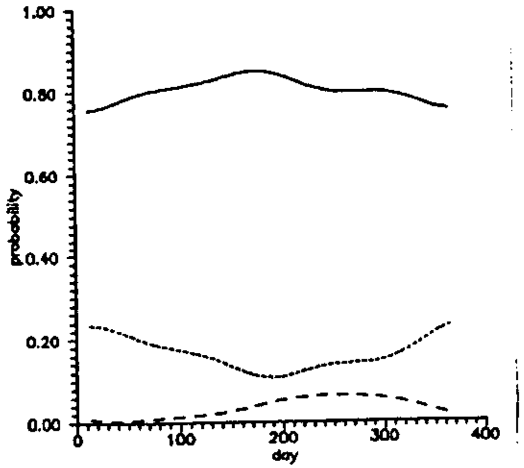
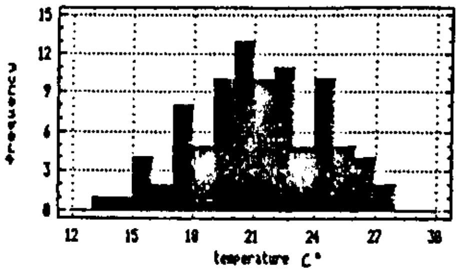
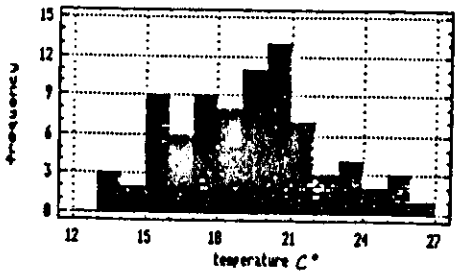


Fig.5

Fourier approximations of probability of the occurence of precipitation group: small (dotted), medium (solid) and large (dashed line).



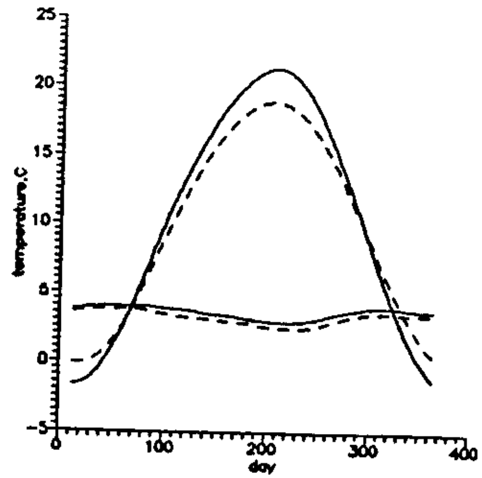
(a)



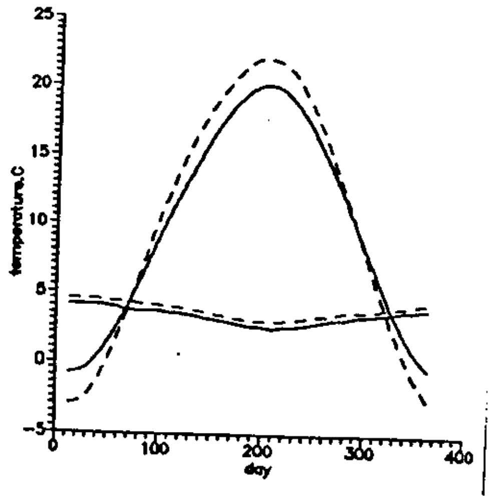
(b)

Fig. 6 (a, b)

Frequency histograms of the daily average temperature on 1-day long wet series in Kompolt: (a) on the 195th day, (b) on the 285th day.



(a)



(b)

Fig. 7 (a, b)

Fourier approximations of daily average temperature and std. deviation in Kompolt for the 1st day of series of any length (solid line) and for any other day of series (dashed line) (a) wet series, (b) dry series.

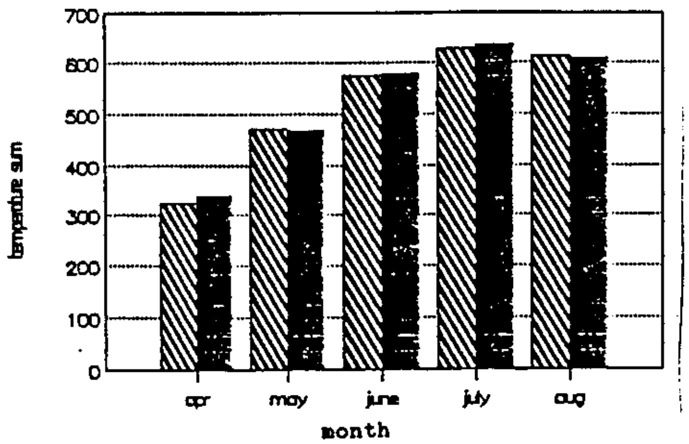


Fig.8

Monthly average of temperature sums in Kompolt.
Full bars - model data, striped bars - empirical data.

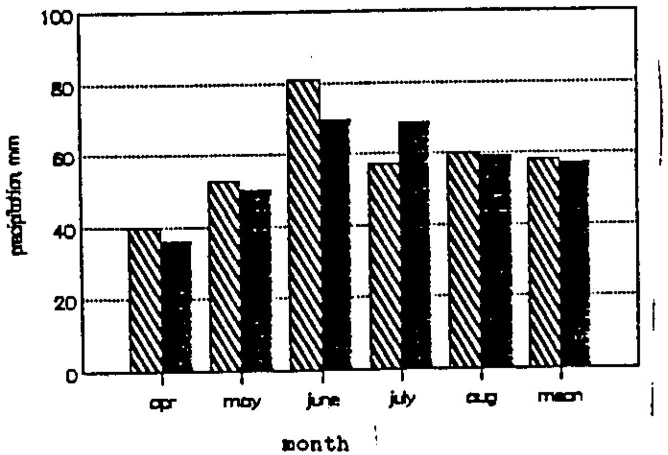


Fig.9

Monthly average of precipitation in Kompolt.
Full bars - model data, striped bars - empirical data.

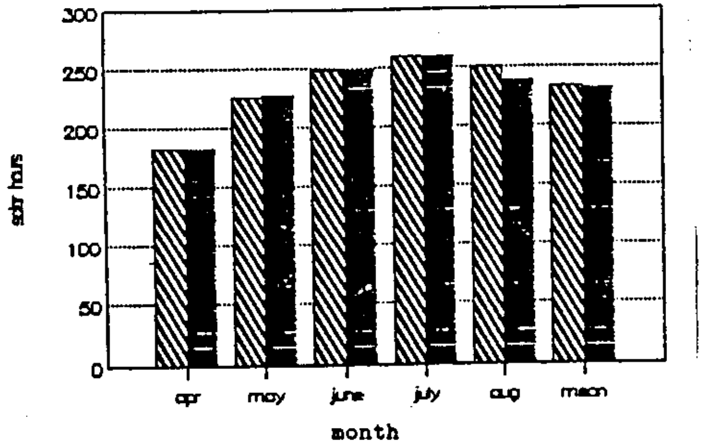


Fig.10

Monthly average of solar hours in Kompolt.
Full bars - model data, striped bars - empirical data.

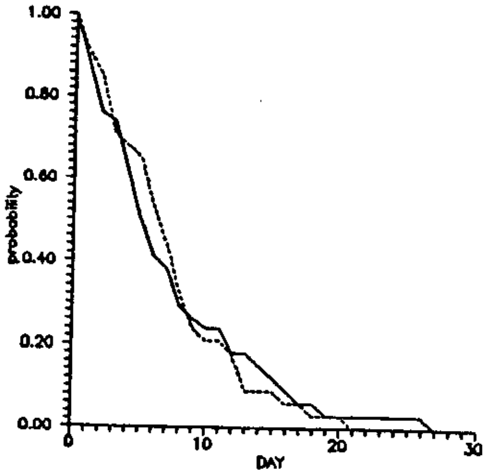


Fig.11

Probability P(x) of the event, that the quantity of days with average temperature beyond 25 C^{o} exceeds x in Kompolt, solid line - empirical data, dashed - model data.