IL NUOVO CIMENTO VOL. 12 C, N. 2 Marzo-Aprile 1989

Statistical Forecasting of Daily Precipitation.

S. MORELLI and R. SANTANGELO
Osservatorio Geofisico dell'Università - Via Campi 213/A, 41100 Modena

(ricevuto il 22 Febbraio 1988)

Summary. — From the time series of daily precipitation observed in Modena (Italy) since 1830, a model for the daily statistical forecasting has been built. The main tool of the model is an urn which contains balls labelled by «wet» and «dry». The daily extraction from the urn determines whether, on that day, there will be a precipitation, *i.e.* if that day will be «wet» or «dry». If the day is dry, the content of the urn is changed by adding other balls some labelled by dry and some other labelled by wet. The numbers of added balls depend on the day of the year (seasonal dependence). If the day is wet, the urn is reset to an initial condition, which depends on the day of the year (seasonal dependence). Moreover, if the day was wet, the likely quantity of precipitation is deduced from a gamma distribution with parameters which are seasonally dependent. All the seasonally dependent parameters in the statistical processes previously discussed can be expressed by Fourier expansions, having one year as fundamental period. The observed distributions are adequately fitted by expansions which do not exceed the second harmonic. Although the model has been tuned on the observations in Modena, it can presumably be extended to the entire region having same climate, *i.e.* the Po Plain.

PACS 92.60 – Meteorology.

1. – Introduction.

It is very useful to build a model capable of integrating both the deterministic and the stochastic variability of a time series. If adequate, a stochastic model summarizes all the information contained in the time series with only few parameters and allows a fairly simple comprehension of the principal features of

the phenomenon. The time series on daily precipitations are multivariate since each piece of information is given by several statistical variables: day labelled by presence or absence of precipitation, type of precipitation (rain, snow, hail, etc); quantity of precipitated water per unit surface area. A first simplification is to reduce the variables to the following: day labelled by presence or absence of precipitation and height of precipitated water.

Another simplification is to assume that the time series is cyclostationary with a fundamental period of one year. This assumption is a good approximation at least for the data observed in Modena and considered in this paper. The daily temperatures in Modena have shown that a long-term trend is present due to an urban greenhouse effect([1]). However this effect is small and should be less important for the precipitation. Rain clouds actually come from distant regions and precipitate mainly on account of effects exceeding the urban scale. In any case a seasonal behaviour can be considered quite reasonable for a first-generation model, and a further small trend can be easily added to the model presented in this paper.

The models reported in the literature are univariate models. They study the marginal probabilities, as a function of the day of the year, of: height of precipitated water, presence or absence of precipitation, number of consecutive days with or without precipitation([2-5]). For the data observed in Modena a bivariate model has been presented([6]). The statistical variables of the model are: height of precipitated water and number of consecutive dry days just preceeding the wet day. The two variables appeared to be independent([7]). However this statistical model, although accurate, has drawbacks. On a rainy day it forecasts the next rainy day, which may come several days later. Therefore the model is not capable to include some further information that has arrived on a day between the two rainy days, such as an updated forecast from a general circulation model.

The model presented in this paper improves the bivariate model quoted above, since it gives, for each day, the probability of precipitation for the next day, with the same or even better accuracy. The model can be combined with other models such as a general circulation model.

([1]) M. MARSEGUERRA, S. MORELLI, G. SALTINI and R. SANTANGELO: *Nuovo Cimento C*, 2, 499 (1979).

([2]) D. A. WOOLHISER and G. G. S. PEGRAM: *J. Appl. Meteorol.*, 18, 34 (1979).

([3]) A. BERGER and C. H. R. GOOSSENS: *J. Climatol.*, 3, 21 (1983).

([4]) J. ROLDAN and D. A. WOOLHISER: *Water Resour. Res.*, 18, 1451 (1982).

([5]) D. A. WOOLHISER and J. ROLDAN: *Water Resour. Res.*, 18, 1461 (1982).

([6]) S. MORELLI and R. SANTANGELO: *Statistical model for daily precipitation, in New Perspectives in Climate Modelling*, edited by A. L. BERGER and C. NICOLIS (Elsevier, Amsterdam, 1984), p. 69.

([7]) S. MORELLI and R. SANTANGELO: *Nuovo Cimento C*, 8, 743 (1985).

2. – Data.

Precipitation has been observed daily and accurately in Modena by the Osservatorio Geofisico of the University since March 1st, 1830. The minimum height of water measurable during the long period of observations was not always the same. However a careful analysis has shown that rain measurements are homogeneous provided the height of the precipitated water was not smaller than 0.1 mm. Therefore, for precipitation amounts smaller than 0.1 mm the precipitation is considered as absent. Old observations on snow, hail, etc. may have some systematic error since the equivalent height of precipitated water was not always obtained during the same day of precipitation. Therefore, in the development of the model, data on rain only have been used. However the conclusions obtained by the rain precipitations can be extended to all the hydric precipitations with some minor adjustment of the parameters, this was confirmed by a comparison with recent observations.

The reliability of the model can be estimated by comparing the expected and the observed distributions. For a χ^2 test the number of observations in each interval must be greater than a certain value, say 5 or 10, in order to have acceptable statistical errors. Therefore it is useful to lump the data of consecutive days together to have a better statistics. However integrating over a too long interval implies a loss of time resolution, *i.e.* a worse capability of the model to explain time variations. As reasonable compromise between these two conflicting requirements, the period of integration has been chosen to be 14 consecutive days. One year is given approximately by 26 groups of lumped data ($26 \times 14 = 364 \approx$ one year). Moreover in the assumption of cyclo-stationarity data of corresponding days in different years can be grouped together. Such grouping could be performed also if cyclo-stationarity is broken by a small trend.

In conclusion each day is labelled by: the number t' ($1 \leqslant t' \leqslant 365$) which orders the day within the year; an alphabetic flag with two positions: a, for days without precipitation; and b, for days with precipitation; the equivalent height h (in mm) of precipitated water.

The data have been further grouped for 14 consecutive days in 26 groups, by assigning to all t' of the same group a new number t ($1 \leqslant t \leqslant 26$) (*).

3. – Presence or absence of precipitation: model.

As discussed previously the statistical variables concerning presence or absence of precipitation and the equivalent height of precipitated water

(*) Data on February 29th in leap years and December 31th in each year have been ignored.

appeared independent, at least in the time series of the observations performed in Modena.

Therefore the overall probability is given by the product of the probabilities of the two processes which can be considered separately.

In this section the evaluation of the probability of presence or absence of hydric precipitation is discussed.

The position of the alphabetic flag, a, or b is obtained for each day by extraction of one ball from an urn, containing A balls of type a and B balls of type b.

The extraction of a ball of type b means that there will be precipitation in the day. In this case the content of the urn is reset so that

$$\frac{A}{A+B} = q\,, \tag{1}$$

where q is a periodic function of time with fundamental period of one year (cyclostationarity).

The extraction of a ball of type a means that there will be no precipitation on the day. In this case the content of the urn is changed by adding A' balls of type a and B' balls of type b to the urn, so that

$$\frac{A'}{A+B} = \alpha\,, \qquad \frac{B'}{A+B} = \beta\,, \tag{2}$$

where α and β are periodic functions of time with fundamental period of one year.

Assuming that the harmonics of the periodic functions q, α, β do not exceed the second, they are:

$$\begin{cases} q(t) = q_0 + q_1^c \cos 2\pi \dfrac{t}{26} + q_1^s \sin 2\pi \dfrac{t}{26} + q_2^c \cos 2\pi \dfrac{t}{13} + q_2^s \sin 2\pi \dfrac{t}{13}\,, \\ \alpha(t) = \alpha_0 + \alpha_1^c \cos 2\pi \dfrac{t}{26} + \alpha_1^s \sin 2\pi \dfrac{t}{26} + \alpha_2^c \cos 2\pi \dfrac{t}{13} + \alpha_2^s \sin 2\pi \dfrac{t}{13}\,, \\ \beta(t) = \beta_0 + \beta_1^c \cos 2\pi \dfrac{t}{26} + \beta_1^s \sin 2\pi \dfrac{t}{26} + \beta_2^c \cos 2\pi \dfrac{t}{13} + \beta_2^s \sin 2\pi \dfrac{t}{13}\,. \end{cases} \tag{3}$$

Clearly this process contains 15 unknown parameters to be determined (*).

(*) The number 26 is due to the grouping described in the previous section. In general the expressions of (3) should read:

$$\gamma(t') = \gamma_0 + \gamma_1^c \cos 2\pi \frac{t'}{T} + \gamma_1^s \sin 2\pi \frac{t'}{T} + \gamma_2^c \cos 2\pi \frac{t'}{T/2} + \gamma_2^s \sin 2\pi \frac{t'}{T/2}\,,$$

where T is the fundamental period of one year.

In conclusion the probability of a sequence of $n-1$ dry days, just following a wet day and ending with a wet day is:

$$p = q(t)\frac{q(t)+\alpha(t)}{1+[\alpha(t)+\beta(t)]}\frac{q(t)+2\alpha(t)}{1+2[\alpha(t)+\beta(t)]}\cdots\frac{1-q(t)+(n-1)\beta(t)}{1+(n-1)[\alpha(t)+\beta(t)]}. \tag{4}$$

4. – Presence or absence of precipitation: estimation of the parameters of the model.

All the sequences of days beginning by a day of type b (wet) and ending with a day of type b (wet) with only days of type a (dry) in between have been considered: *e.g.*, *bb*, *bab*, *baab*, *baaab*, ..., *baa* ... *ab*, ...

According to the number t' of the last day of the sequence (final b), any sequence is assigned to the appropriate group labelled by t. A short remark: the last day t' of the sequence must satisfy the requirements of note of sect. 2, that is a sequence with a final b as February 29th or December 31st is ignored, whereas a sequence having a day of type a on February 29th or December 31st is accepted. Any sequence is further characterized by the number d ($\geqslant 1$), which counts the number of the days of type a in the sequence, increased by one.

Therefore for any group t, a distribution is obtained which gives the number $n^{\text{obs}}(d)$ of observed sequences *vs.* d. The model described in the previous section allows one to evaluate the number $n^{\text{mod}}(d)$, of expected sequences *vs.* d, by normalizing to the total number of observed sequences belonging to the same group t.

The evaluation through the model implies, however, that the parameters of eqs. (3) are known. The parameters are estimated by fitting all the distributions of the 26 groups, *i.e.* by minimizing the χ^2 defined as follows:

$$\chi^2 = \sum_{t=1}^{26}\left(\sum_{d}^{N(t)}\frac{[n^{\text{obs}}(d)-n^{\text{mod}}(d)]^2}{n^{\text{obs}}(d)}\right) = \sum_{t=1}^{26}\chi^2(t). \tag{5}$$

The variance of each bin d has been assumed to be given by $n^{\text{obs}}(d)$ as expected if the observations in bin d follow a Poisson distribution having mean (and variance) $n^{\text{obs}}(d)$. However such estimate of the variance may have a low confidence level when $n^{\text{obs}}(d)$ is low. This implies an understimation of the true variance. As a partial attenuation of such effect a lumping procedure has been adopted. Especially for high values of d, $n^{\text{obs}}(d)$ is often smaller than 5. In these cases several close bins have been lumped to a unique bin, which therefore has a number of observed sequences greater or equal to 5. The total number of bins depends on t: $N(t)$.

Since it appears that q is not correlated with α and β, whereas α and β are

correlated, the minimization was performed by iteration in the parameters of q and then in the parameters of α and β.

Initially the second harmonic was neglected. In a second step it was included and retained only if the χ^2 was significantly improved. The estimated values are:

$$q_0 = 0.503\,, \quad q_1^c = -0.097\,, \quad q_1^s = -0.019\,, \quad q_2^c = 0.048\,, \quad q_2^s = 0.037\,,$$

$$\alpha_0 = 1.871\,, \quad \alpha_1^c = -0.875\,, \quad \alpha_1^s = -0.278\,, \quad \alpha_2^c = 0.438\,, \quad \alpha_2^s = 0.214\,,$$

$$\beta_0 = 0.353\,, \quad \beta_1^c = -0.210\,, \quad \beta_1^s = -0.010\,, \quad \beta_2^c = 0.111\,, \quad \beta_2^s = -0.067\,.$$

However, due to the many independent variables (*e.g.*, 10 for the two parameters α and β) the solution found must not be considered as the true minimum but only a relative minimum which gives a plausible solution.

Using the quoted values, the functions $q(t)$, $\alpha(t)$, $\beta(t)$ never become negative and $q(t)$ never exceeds 1 as expected by their definitions (1), (2).

TABLE I. – *Initial and final date of t, values of* $\chi^2(t)$ *and* $N(t)$ *as* (5).

t	$\chi^2(t)$	$N(t)$
1 (1/1 - 1/14)	6.3	16
2 (1/15 - 1/28)	17.5	16
3 (1/29 - 2/11)	12.9	17
4 (2/12 - 2/25)	16.6	18
5 (2/26 - 3/11)	15.4	16
6 (3/12 - 3/25)	20.3	18
7 (3/26 - 4/8)	17.7	17
8 (4/9 - 4/22)	18.2	16
9 (4/23 - 5/6)	26.8	16
10 (5/7 - 5/20)	13.8	16
11 (5/21 - 6/3)	10.0	14
12 (6/4 - 6/17)	23.2	17
13 (6/18 - 7/1)	10.9	16
14 (7/2 - 7/15)	35.7	17
15 (7/16 - 7/29)	42.7	20
16 (7/30 - 8/12)	21.6	19
17 (8/13 - 8/26)	21.8	20
18 (8/27 - 9/9)	9.8	19
19 (9/10 - 9/23)	15.6	19
20 (9/24 - 10/7)	11.2	18
21 (10/8 - 10/21)	16.3	18
22 (10/22 - 11/4)	13.0	17
23 (11/5 - 11/18)	10.1	14
24 (11/19 - 12/2)	15.0	16
25 (12/3 - 12/16)	8.5	14
26 (12/17 - 12/30)	14.0	17

With the parameters previously quoted the value of χ^2 is 444.9 to be compared with the expected mean of 426, obtained by the number of degrees of freedom. The result appears satisfactory.

In any case in table I are given the partial values of the $\chi^2(t)$ pertaining to the group t and the corresponding number of bins: $N(t)$.

It would be useful to show the 26 distributions observed and expected. However this would require too much space. Therefore only three typical distributions will be shown: in fig. 1*a*) the best case, corresponding to the lowest

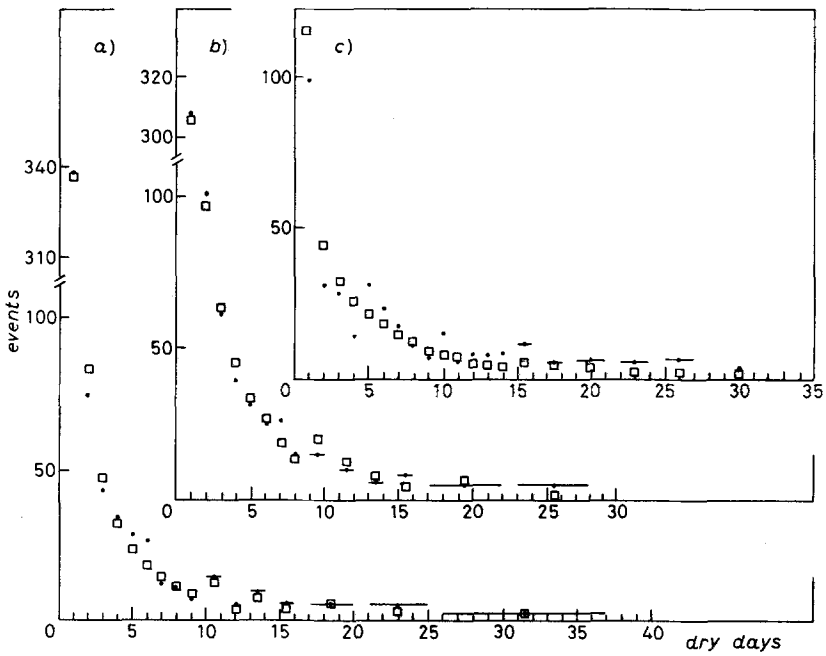

Fig. 1. – Observed distributions (•) and expected distributions (□) for the dry sequences: *a*) the best case, *b*) an average case, *c*) the worst case.

value of $\chi^2(t)$; in fig. 1*b*) an average case corresponding to a value of $\chi^2(t)$ as expected; in fig. 1*c*) the worst case corresponding to the highest value of $\chi^2(t)$. In any case the distribution of partial $\chi^2(t)$ as given in table I appears consistent and could even improve if the variances used in formula (5) were the true variances.

5. – Height of precipitated water. Model and estimation of the parameters.

When the previous model declares that the day is of type *a* (dry), the height of precipitated water is immediately set equal to zero. If the day happens to be of type *b* (wet), an independent statistical process is activated, which extracts from an urn the height of the precipitated water according to the model distribution.

The distribution of the height of precipitated water appeared to follow a gamma distribution ([6])

$$Q(h) = h^{l-1} \exp[-h/m] \frac{1}{\Gamma(l) \cdot m^l}, \qquad h > 0, \tag{6}$$

where l and m are assumed to be periodic functions of time with fundamental period of one year (cyclostationarity). The model distribution is given by a slight modification of formula (6) in order to take into account the minimum resolution of the equipment which measures the height of precipitated water:

$$Q'(h) = \frac{Q(h)}{1 - \int_0^{h_{\min}} Q(h)\,\mathrm{d}h}, \qquad h \geqslant h_{\min}, \; h_{\min} = 0.1\,\mathrm{mm}. \tag{7}$$

For the 14 days of the group t, a distribution of the number of observed heights between h and $(h + \Delta h)$: $n^{\mathrm{obs}}(h, h + \Delta h)$ is obtained. A corresponding

TABLE II. – *Initial and final date of t, values of* $\chi^2(t)$ *and* N (t) as (9).

t	$\chi^2(t)$	$N(t)$
1 (1/1 - 1/14)	5.8	16
2 (1/15 - 1/28)	7.3	15
3 (1/29 - 2/11)	29.1	15
4 (2/12 - 2/25)	23.8	16
5 (2/26 - 3/11)	13.4	16
6 (3/12 - 3/25)	4.8	15
7 (3/26 - 4/8)	19.2	17
8 (4/9 - 4/22)	9.1	16
9 (4/23 - 5/6)	8.8	16
10 (5/7 - 5/20)	15.1	18
11 (5/21 - 6/3)	23.1	18
12 (6/4 - 6/17)	13.5	17
13 (6/18 - 7/1)	17.1	17
14 (7/2 - 7/15)	12.8	16
15 (7/16 - 7/29)	21.9	15
16 (7/30 - 8/12)	14.7	16
17 (8/13 - 8/26)	15.6	16
18 (8/27 - 9/9)	27.2	16
19 (9/10 - 9/23)	24.1	19
20 (9/24 - 10/7)	14.2	18
21 (10/8 - 10/21)	28.3	21
22 (10/22 - 11/4)	20.0	22
23 (11/5 - 11/18)	17.4	19
24 (11/19 - 12/2)	12.9	17
25 (12/3 - 12/16)	25.0	17
26 (12/17 - 12/30)	18.3	16

number of model heights: $n^{\mathrm{mod}}(h, h+\Delta h)$ can be obtained by using the formula (6), (7) if the functions l, m are known. To evaluate l, m the harmonics higher than the second are assumed negligible. Nevertheless 10 parameters remain still unknown as shown by the following expressions:

$$(8)\qquad \begin{cases} l(t) = l_0 + l_1^c \cos 2\pi \dfrac{t}{26} + l_1^s \sin 2\pi \dfrac{t}{26} + l_2^c \cos 2\pi \dfrac{t}{13} + l_2^s \sin 2\pi \dfrac{t}{13}, \\ \\ m(t) = m_0 + m_1^c \cos 2\pi \dfrac{t}{26} + m_1^s \sin 2\pi \dfrac{t}{26} + m_2^c \cos 2\pi \dfrac{t}{13} + m_2^s \sin 2\pi \dfrac{t}{13}. \end{cases}$$

The unknown parameters have been estimated by minimizing a χ^2 expression:

$$(9)\qquad \chi^2 = \sum_{t=1}^{26} \left(\sum_{\Delta h}^{N(t)} \frac{[n^{\mathrm{obs}}(h, h+\Delta h) - n^{\mathrm{mod}}(h, h+\Delta h)]^2}{n^{\mathrm{obs}}(h, h+\Delta h)} \right) = \sum_{t=1}^{26} \chi^2(t).$$

The variance in each bin $(h, h+\Delta h)$ has been assumed to be given by $n^{\mathrm{obs}}(h, h+\Delta h)$. For high values of h, $n^{\mathrm{obs}}(h, h+\Delta h)$ is often smaller than 5. In these cases several close bins have been lumped to a unique bin, which therefore

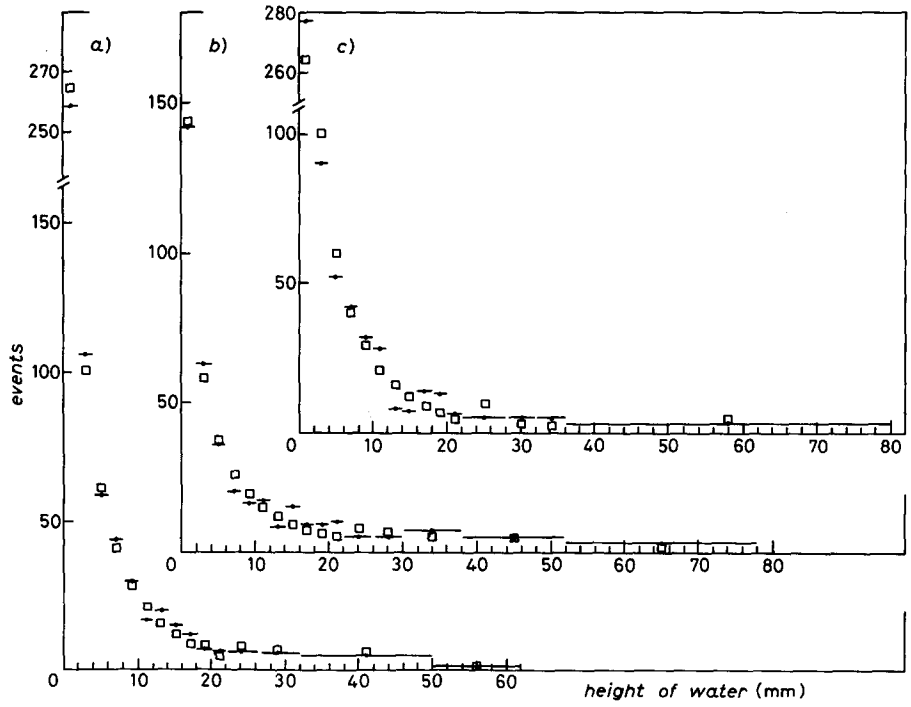

Fig. 2. – Observed distributions (•) and expected distributions (□) for the height of precipitated water: *a*) the best case, *b*) an average case, *c*) the worst case.

has a number of observed sequences greater or equal to 5. The total number of bins depends on t: $N(t)$.

The minimization procedure was initially on the parameters l_0, m_0 only. A successive minimization was performed starting from the previous values and adding l_1^c, l_1^s. No significant improvement of χ^2 was obtained. Therefore they have been neglected. A further successive minimization was performed by adding the parameters m_1^c, m_1^s. The χ^2 was somehow improved. No further improvement was obtained by adding the second harmonic. In conclusion the final parameters are:

$$l_0 = 0.41\,, \quad m_0 = 12.94\,, \quad m_1^c = -1.81\,, \quad m_1^s = -1.83\,.$$

The fit is good as it appears from the comparison between the value of χ^2: 442.8 and its expected value obtained by the degrees of freedom: 436.0. Also the partial values of χ^2: $\chi^2(t)$ show a good agreement, as appears from table II.

Also in this case three distributions are shown: in fig. 2*a*), the best case corresponding to the lowest value of $\chi^2(t)$; in fig. 2*b*) an average case corresponding to a value of $\chi^2(t)$ as expected; in fig. 2*c*) the worst case corresponding to the highest value of $\chi^2(t)$.

6. – Conclusion.

The time series of precipitation in Modena (Italy) can be expressed by a stochastic process involving two urns. An extraction from the first urn every day gives the statistical forecast about the type of the day: *i.e.* if that day is wet or dry. The extraction from the second urn is performed only if the day happened to be wet, and gives the statistical forecast of the amount of hydric precipitation.

The results appear satisfactory for a first-generation model. How well the quoted model forecasts the occurrence and the amount of rain will be presented in a next paper.

The model has been tuned on the time series of observations in one town situated within the Po Plain. Since the Plain can be considered a unique structure from a climatic point of view (coherence of observations), the model is likely to be valid for all the Plain.

The model gives statistical forecasts every day. Therefore it is suited to be combined with other models giving forecasts every day, in particular: general circulation models.

● RIASSUNTO

È presentato un modello per la previsione statistica giornaliera di precipitazione, messo a punto mediante la serie temporale delle osservazioni giornaliere a Modena (Italia) che

inizia dal 1830. Si tratta di un modello ad «urna» la quale contiene palline etichettate «pioggia» e «secco». L'estrazione giornaliera dall'urna determina se in quel giorno ci sarà precipitazione, ossia se il giorno è di pioggia» o «secco». Se il giorno è secco, il contenuto dell'urna è variato con l'aggiunta di altre palline etichettate da «secco» e «pioggia». Il numero di palline aggiunte dipende dal giorno dell'anno (dipendenza stagionale). Se il giorno è piovoso, il contenuto dell'urna è riportato ad una condizione iniziale, che dipende dal giorno dell'anno (dipendenza stagionale). Inoltre, nel caso di giorno piovoso, la quantità di pioggia è dedotta da una distribuzione gamma con parametri dipendenti dal tempo. Tutti i parametri con dipendenza stagionale del precedente processo statistico sono stati espressi mediante uno sviluppo in serie di Fourier, avente un periodo fondamentale di un anno. Le distribuzioni osservate sono adeguatamente rappresentate da sviluppi in serie che non superano la seconda armonica. Questo metodo di previsione statistica giornaliera può essere facilmente combinato con metodi di previsione giornaliera dinamica, poiché le previsioni (sia statistica che dinamica) sono effettuate ogni giorno. Sebbene il modello sia stato messo a punto mediante le osservazioni di precipitazione a Modena, presumibilmente esso può essere esteso all'intera regione climaticamente omogenea, ossia la Pianura Padana.

Статистическое прогнозирование суточного выпадения осадков.

Резюме (*). — Из временной последовательности суточных выпадений осадков, зарегистрированных в Модене (Италия) с 1830 г., построена модель для суточного статистического прогнозирования. Основной инструмент модели представляет урна, которая содержит шары с метками «мокрый» и «сухой». Суточное извлечение из урны определяет прогнозирование на этот день, т.е. будет ли этот день «мокрый» или «сухой». Если день «сухой», то содержание урны изменяется посредством добавления шаров, часть из которых помечена «сухими» и другая часть помечена «мокрыми». Число добавленных шаров зависит от дня в году (сезонная зависимость). Если день «мокрый», то урна возвращается в начальное состояние, которое зависит от дня в году (сезонная зависимость). Кроме того, если день «мокрый», то вероятность прогнозирования выводится из гамма-распределения с параметрами, которые зависят от времени года. Все зависящие от сезона параметры в статистических процессах, которые ранее обсуждались, могут быть выражены с помощью Фурье-разложений, которые имеют основной период, равный одному году. Наблюденные распределения адекватно описываются с помощью разложенний, которые не превышают вторых гармоник.

(*) *Переведено редакцией.*