REGRESSION MODEL FOR CLOUD-TOP HEIGHT DISTRIBUTION OVER CONTINENTS, BASED ON DATA OF FGGE INTERNATIONAL EXPERIMENT

O.I. Aldoshina, V.V. Bacherikov, E.E. Limar, and V.A. Fabrikov

All-Union Scientific-Research Institute of Optophysical Measurements, Moscow
Received October 30, 1989

Algorithms and computer programs are developed for calculating statistical moments, densities, and distribution functions as well as correlation matrices and regressions for statistical analysis of the FGGE (the First GARP Global Experiment) cloud data. Seasonal-geographical distributions of cloud coverage and top heights over continents are constructed. Cloud top heights and cloud coverage over continents are found to permit approximation by the log-normal law. Such approximations are made following Kolmogorov's criterion at a significance level of 0.95. Results from this study can be used to interpret data of satellite remote sensing in the optical wavelength range.

Understanding of the properties of Earth cloud cover is most important for an adequate interpretation of satellite optical remote sensing data. Understanding and forecasting such characteristics are vital for the adequacy and accuracy of target identification from space. According to published references, cloud top height estimates had, until lately, remained the one characteristic described by the poorest a most contradictory data. Such a situation was explained by insufficient available data, by differences in observational means, techniques and strategies, by scale disagreements, by the ambiguity of error sources, and by the local character of the observations. Cloud studies from spacecraft (SC) have made possible a more rigorous statement of that problem. However, having only one SC, cloudiness can only be monitored at the periodicity with which that SC orbits the Earth. Lack of global coverage and of temporal continuity rendered an objective picture of the Earth cloud cover impossible.

The FGGE experiment[1] produced measurement data on cloud top height and total coverage at 3-h time intervals, starting from 00:00 GMT, December 1, 1978 to November 30, 1979 at 262,144 points of the polar stereographic grid with a 50-km resolution. Below, an attempt is made to construct a model of the total cloud coverage and cloud top heights over continents based on those data.

To study these characteristics quantitatively the statistical analysis techniques were applied to 5×5° pixels of the geodetic grid, both on a monthly and seasonal basis. The computed values included mathematical expectations M_x, variances D_x, and correlation coefficients ρ.[2] Choosing a 5×5° grid reduced the processed data set to approximately 184,300 points, and monthly compiling of data — to a set of more than 44 million such points, which may be considered sufficient for reaching statistically significant conclusions.

Initially lacking distribution laws for the studied variables, we based our study on the fact that the first derivative of the logarithm of the similarity function L displays an asymptotically normal distribution with zero mean in case the distribution itself is regular and

$$D\left\{\frac{\partial \log L}{\partial \Theta}\right\} = M\left\{\left[\frac{\partial \log L}{\partial \Theta}\right]^2\right\} = -M\left\{\frac{\partial^2 \log L}{\partial \Theta^2}\right\}.$$

Here D and M are, respectively, the variance and the mathematical expectation of the function Θ. Then denoting $\Psi = \dfrac{\partial \log L}{\partial \Theta} \Big/ \left[M\left\{\left(\dfrac{\partial \log L}{\partial \Theta}\right)^2\right\}\right]^{1/2}$, we have that the function Ψ is a standard normal random variable for large data series.[3] Since Ψ is a monotonic function of Θ, its normal integral yields the confidence intervals for the confidence coefficient $1 - \alpha$. Such intervals were found for both the cloud coverage and cloud top heights at those pixels where large data series were available. If, on the other hand, a pixel was mostly occupied by a water surface, so that it was omitted from the above processing, except for its smaller land surface part (and the data series for the latter was small), the respective confidence intervals were constructed using the well-known Student's criterion. Converting from Θ to $Z = (\bar{\Theta} - M / S)$ (here S^2 is the sample variance) and recalling that the latter variable is distributed in accordance with Student's law: $dF = \dfrac{kdZ}{(1+Z^2)^{N/2}}$, one can find for a prescribed confidence coefficient $1 - \alpha$ variables Z_0 and Z_1 such that

$$\int_{-\infty}^{-Z_1} dF = \int_{Z_0}^{\infty} dF = \frac{\alpha}{2}$$

It follows from this that $P = (-Z_1 \le Z \le Z_0) = 1 - \alpha$, which is equivalent to the relationship $P(\bar{\Theta} - SZ \le M \le \Theta + SZ_1) = 1 - \alpha$.

0235-6880/90/04 366-10 $02.00

The confidence intervals for small series were constructed as $\bar{\Theta} - SZ_0$ and $\bar{\Theta} + SZ_1$, with a confidence coefficient of $1 - \alpha$. For illustration purposes Fig. 1 presents the monthly mathematical expectations and selective confidence intervals at an $M \pm \sigma$ level for both the total cloud coverage and cloud top height. These are plotted for the territories of the USA (Figs. 1a and c) and China (Figs. 1b and d). Pixel to pixel mathematical expectation histograms were smoothed using the relationship

$$M_1 = \frac{1}{4} M_{1-1} + \frac{1}{2} M_{1+1}.$$

The maps of the USA (Figs. 2a and b, left columns) and China (Figs. 2a and b, right columns) present for illustrative purposes the average monthly values of cloud cover (Fig. 2a) and cloud top height (Fig. 2b) in 5×5° pixels of the Mercator geodesic coordinates. The notations in Fig. 2a represent the cloud coverage from 0 to 9 points at one-point intervals, and those in Fig. 2b represent the cloud top heights from 0 (cloudless) to 12 km at 1-km intervals.

Figure 3 shows the monthly cloud coverage and cloud top height distribution densities for the USA and China, as obtained by the above technique.

If we assume that there are longer term trends in these characteristics besides the monthly and the annual ones, our distributions should also be assumed to have biased estimates of the momenta. However, one can obtain a mathematical expression of the distributions from these momenta, which would satisfactorily describe the obtained sample series. Computation of various distribution laws, including the normal, the log-normal, the power, the γ- and β-distributlons, the Weibull, Fisher, and Tippett laws (the first and second types of the latter)[4,5,6] was carried out using the Kolmogorov criterion

$$\max |F_{theor}(\Theta) - F_{emp}(\Theta)| = \Delta m,$$

where $F_{teor}(\Theta)$, $F_{emp}(\Theta)$ are, respectively, the theoretical and the empirical distributions; Δm is the discrepancy measure.

At attempt to describe empirical distributions by a normal law resulted in a considerable discrepancy at the 95% significance level: $\Delta m = 17\%$. Other drawbacks of such an approximation consist of its symmetrical shape and an infinite distribution spread; in depending of the variance, such an approximation can attribute too much weight to unrealistic values of the variables sought.

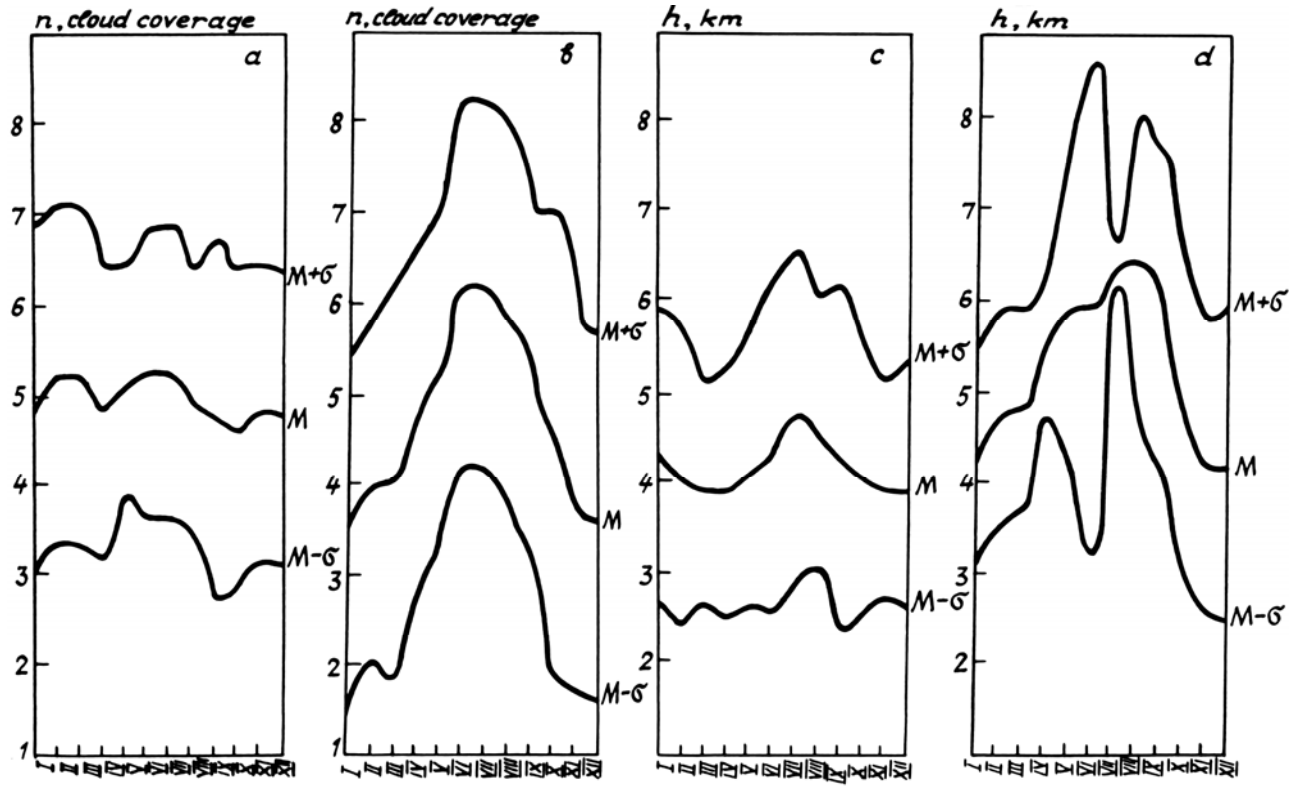


FIG. 1. Confidence intervals for the total cloud cover (n) and cloud top height (h) above the USA (a and c) and China (b and d). Roman numerals represent months of the year.

Let us consider more closely the possibility of applying a β-distribution to our data.

Its parameters are as follows:

$$p = \frac{m'q}{1 - m'}; \quad q = \frac{1 - m'}{(\sigma')^2}[m'(1 - m') - (\sigma')^2],$$

where $m' = \frac{m}{x_{\max}}$, $\sigma' = \frac{\sigma}{x_{\max}}$ are, respectively, the normalized empirical mathematical expectation and the standard deviation, and $x_{\max}$ is the maximum value of the random variable x.

Furthermore, the initial a_1 and central m_1 (i = 1, 2, 3, 4) moments of respective orders were computed (up to the fourth, inclusive) in accordance with Refs. 2 and 7:

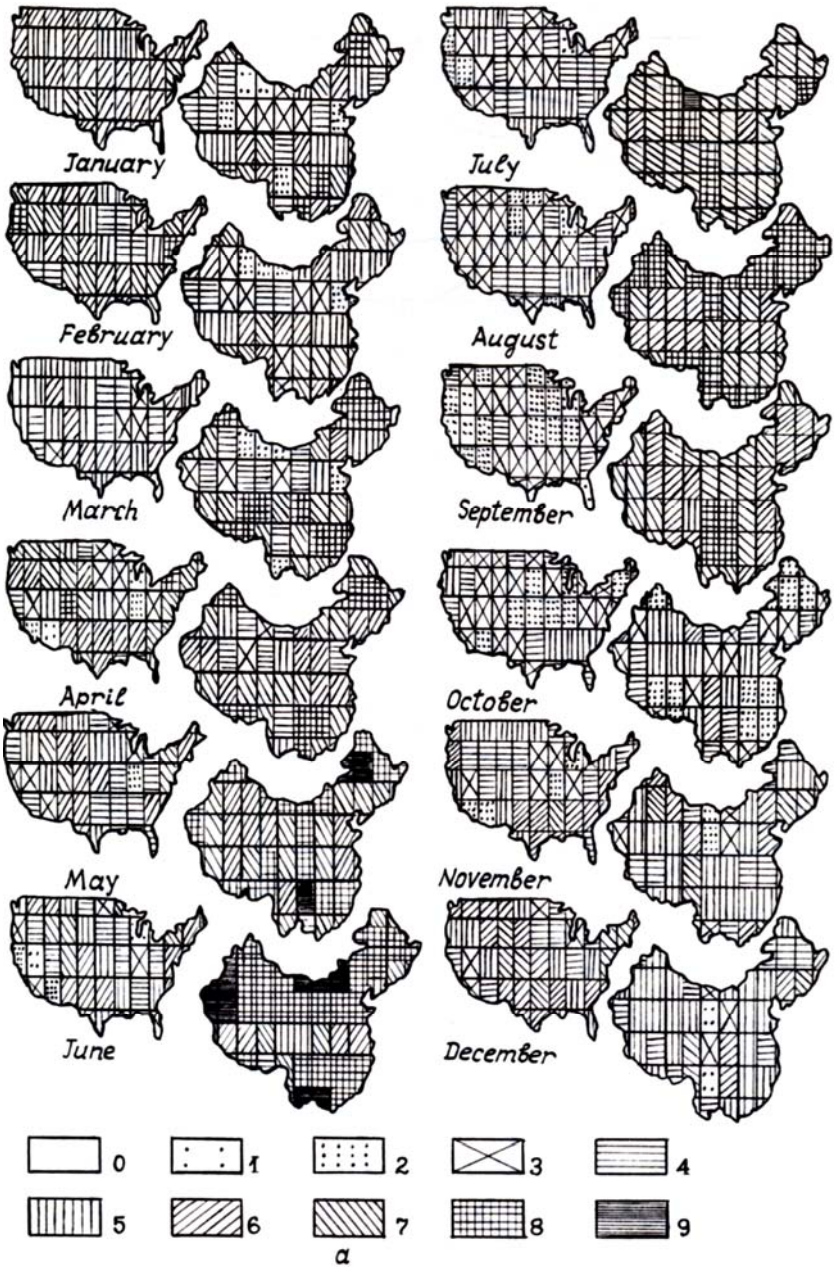
$$a_1 = \frac{\Gamma(p + 1)\Gamma(p + q)}{\Gamma(p)\Gamma(p + q + 1)}; \quad m_1 = 0;$$

$$a_2 = \frac{\Gamma(p + 2)\Gamma(p + q)}{\Gamma(p)\Gamma(p + q + 2)}; \quad m_2 = a_2 - a_1^2;$$

$$a_3 = \frac{\Gamma(p + 3)\Gamma(p + q)}{\Gamma(p)\Gamma(p + q + 3)}; \quad m_3 = a_3 - 3a_1a_2 + 2a_1^3;$$

$$a_4 = \frac{\Gamma(p + 4)\Gamma(p + q)}{\Gamma(p)\Gamma(p + q + 4)}; \quad m_4 = a_4 - 4a_1a_3 + 6a_1^2a_2 - 3a_1^4,$$

where $\Gamma(p)$, $\Gamma(q)$, $\Gamma(p + q)$, $\Gamma(p + i)$, $\Gamma(p + q + i)$ are the values of the Γ-function at the corresponding points. For $x > 2$ the following property of the Γ-function was employed: $\Gamma(x + 1) = x \cdot \Gamma(x)$.



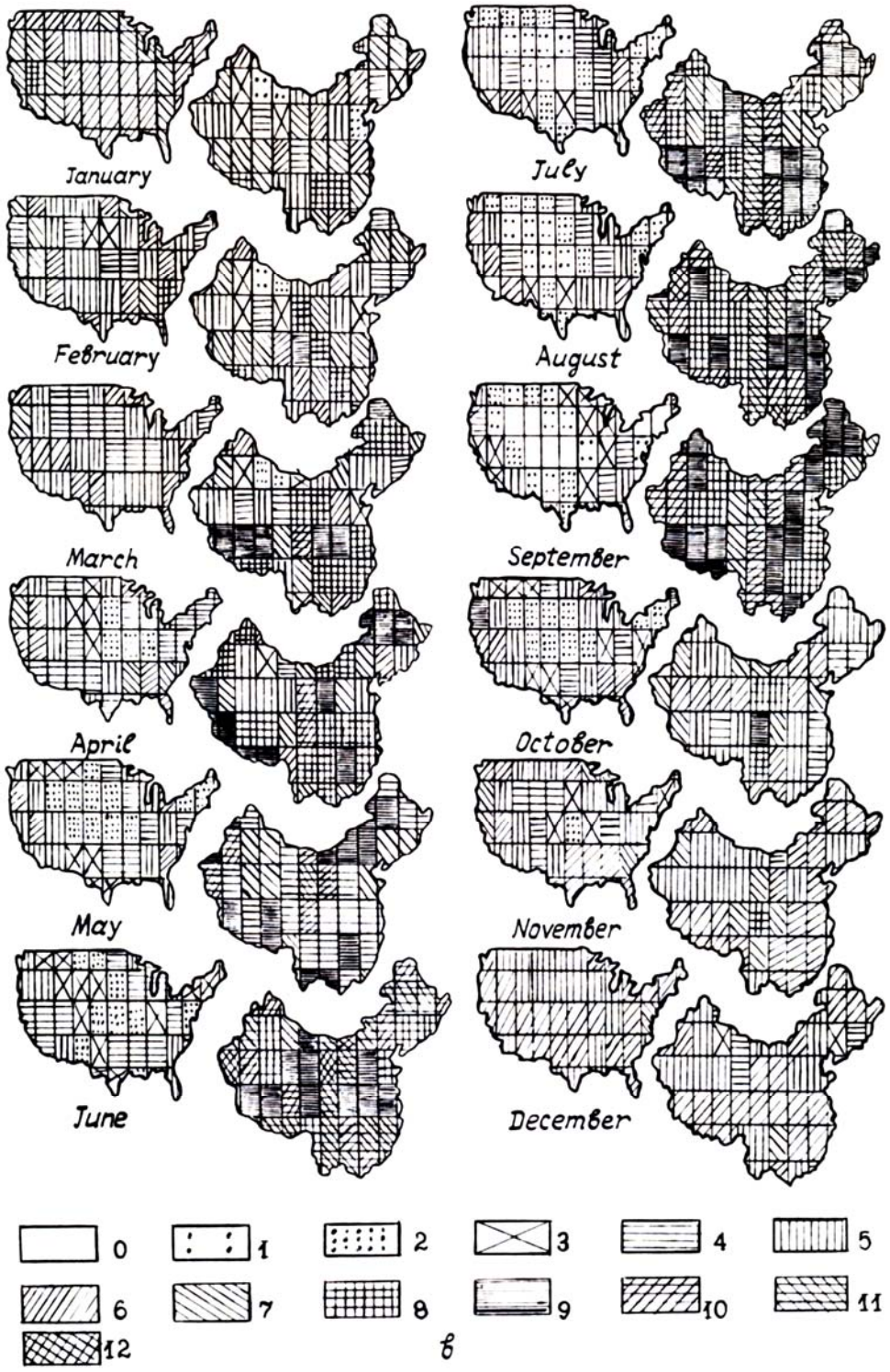


FIG. 2. Total cloud coverage (a) and cloud top height (b).

At large x ($x > 10$) the Stirling relationship was employed: $\Gamma(x+1) \simeq \left(\frac{x}{e}\right)^x \cdot \sqrt{2\pi x}$ (its error is below 1%).

The applicability criterion for the β-distribution[2] is the following:

$$\kappa = \frac{\beta_1(\beta_2+3)^2}{4(2\beta_2-3\beta_1-6)(4\beta_2-3\beta_1)} < 0;$$

where $\beta_1 = \frac{m_3^2}{m_2^3}$; $\beta_2 = \frac{m_4}{m_2^2}$.

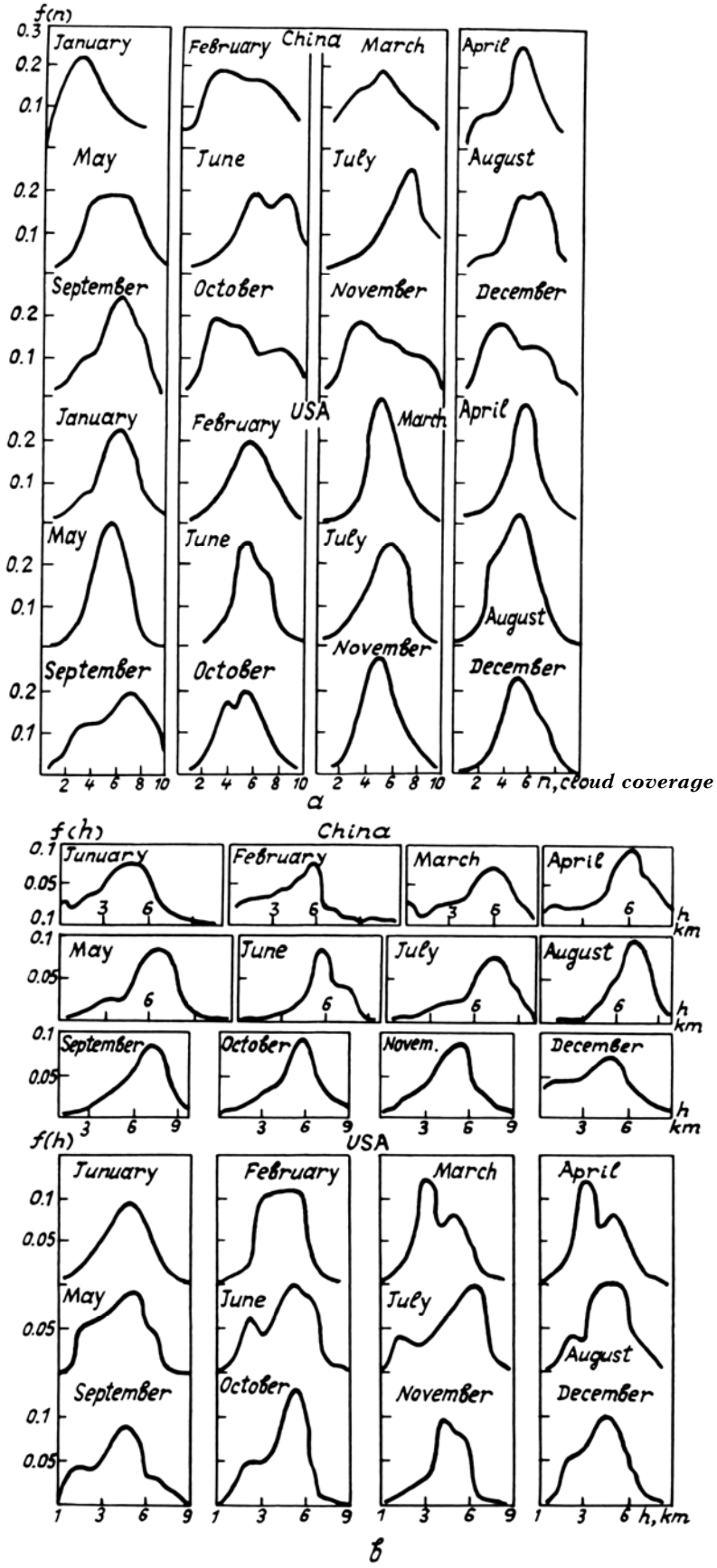


Fig. 3. Monthly distribution densities for the total cloud coverage (a) and cloud top height (b).

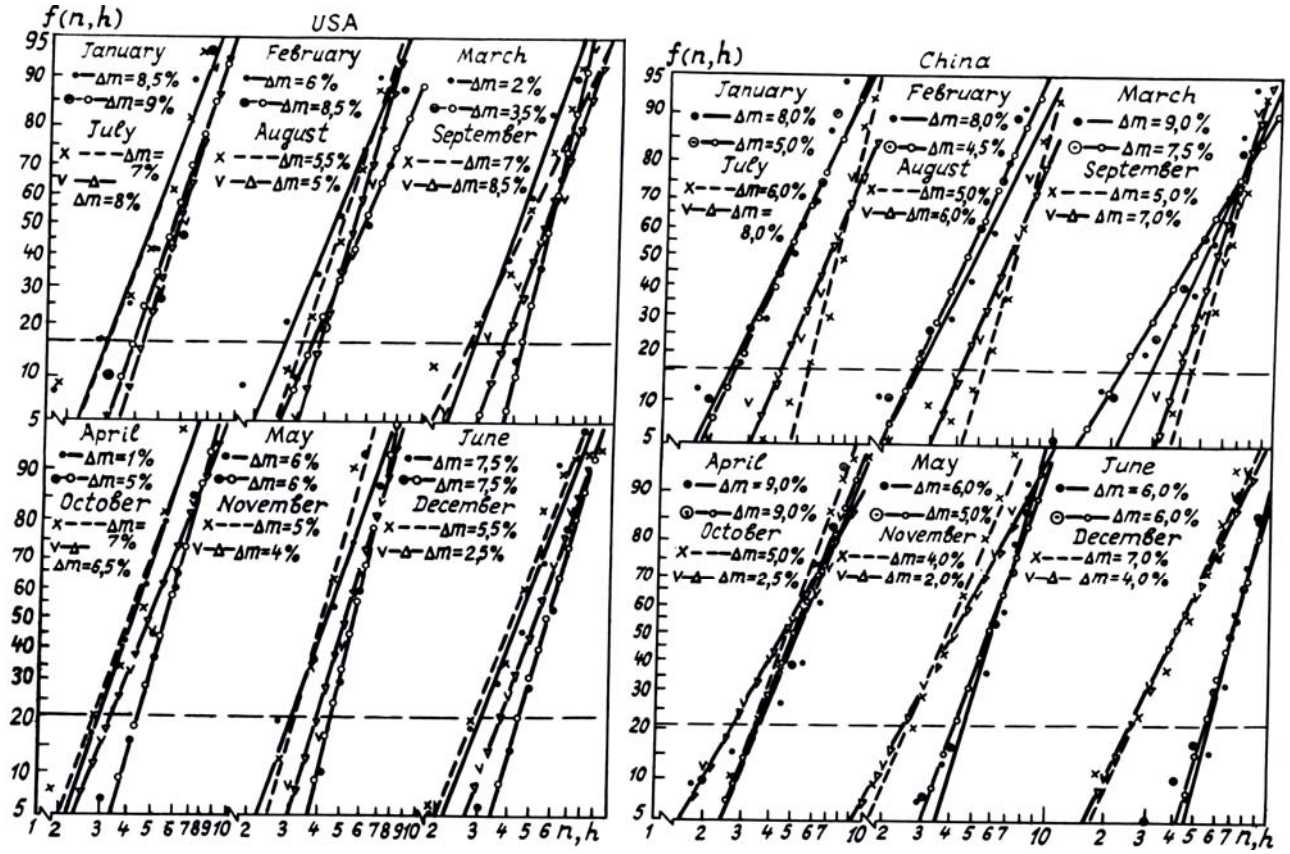


Fig. 4. Total cloud coverage (–Δ–, –o–) *and cloud top height* (– – – –) *distribution functions*

Computations demonstrated that all the values of x were less than zero. Hence the Pearson distribution of the first kind is applicable (the β-distrlbution)

$$F(x,\ p,\ q) = \begin{cases} 0 \text{ for } x \le 0 \\ \dfrac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \displaystyle\int_0^1 x^{p-1}(1-x)^{q-1}dx \text{ for } 0<x<1; \\ 1 \text{ for } x \ge 1 \end{cases}$$

Theoretical graphs of the log-normal function were plotted for the same empirical dependences of the total cloud coverage and cloud top height in logarithmic probability coordinates (Fig. 4).

Following the principle of minimizing the maximum difference between the empirical and the respective theoretical values of Δm we have

$$F(x,\ x_0,\ \sigma_0) = \begin{cases} 0 \text{ for } x \le 0 \\ \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{u} e^{-t^2/2}dt \quad \text{for } x > 0, \end{cases}$$

where

$$u = \frac{\lg x - \lg x_0}{\lg \sigma_0};\ k = \frac{1}{\lg e} = 2.303.$$

The value of $m_y = \log x_0$ is the mathematical expectation of the random variable (RV) $Y = \log x$, and $\sigma_y = \log \sigma_0$ is its rms error. The values of x_0 and σ_0 are determined graphically:[8] $x_0 = x_{50\%}$, (the median of the RV x), and $\sigma = x_{50\%}/x_{15.9\%}$.

The mathematical expectation m_x and the rms error σ_x of a random variable x are equal to[9]

$$m_x = x_0 e^{k^2 \lg^2 \sigma_0 / 2} = x_0 e^{2.651 \lg \sigma_0};$$

$$\sigma_x = m_x \sqrt{\left(\frac{m_x}{x_0}\right)^2 - 1}.$$

The analysis of the data obtained demonstrates that the maximum value of Δm for the β-distribution amounts τ_0 13.8%, while for the log-normal distribution it is only (–9)%. In that case the log-normal distribution happens to describe the empirical dependences better. To test the difference of the correlation coefficients from zero, the R. Fisher[4] confidence intervals were found. For two series taken from a normal distribution having a length N, total correlation coefficient ρ, and selective correlation coefficient r, its variable $\delta = \frac{1}{2}\log\frac{1+r}{1-r}$ is approximately normally distributed even for small r with a mean

$\bar{\delta} = \frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(N-1)}$ and a variance of $1/(N-3)$. Using the distribution for δ, the respective 95% confidence intervals were found for r on the assumption that $\rho = 0$. In this way the noncorrelated zones were identified in the overall distribution.

It was interesting to find the seasonal regressions for zones of high correlation coefficients. We attempted to choose a multiple linear regression of seasonal total cloud coverage employing the least-squares technique. To that end, residuals $(y - \hat{y})^2$ were found for a model $y = a + b_1x + b_2x + \ldots + \varepsilon_1$, where y is the dependent variable (the total cloud coverage in our case), x are the months of the year, b_1 are regression coefficients, a is the free term, ε is the measurement error with a zero mean, $\hat{b}_1$ is the estimate of the regression coefficient, and $\hat{y}$ is the predicted value, where $\hat{y} = \hat{a} + \hat{b}_1x_1 + \hat{b}_2x_2 + \ldots + \hat{b}_px_p$. Since the variance of the total cloud coverage changed from realization to realization, we employed weighted least squares. The realization weight was introduced as the inverse variance. Serial correlation if the residues was calculated from

$$\sum_{j=2}^{N} \frac{(W_jW_{j-1})^{1/2}(y_j - \hat{y}_j)(y_{j-1} - \hat{y}_{j-1})}{\left\{\sum_{j=2}^{N} W_j(y_j - \hat{y}_j)^2 \sum_{j=2}^{N} W_{j-1}(y_{j-1} - \hat{y}_{j-1})^2\right\}^{1/2}},$$

where W_j is the weight of the "j-th" realization. Application of the principle of minimizing the sum of regression squares $\sum(y - \hat{y})^2$ pointed to the best convergence in the case of a curvilinear (or polynomial) dependence of y on x. Selecting our coefficients from the Biometrike Tables[6] we were able to fit quite well an orthogonal polynomial regression of the third and fourth orders: $y = a + \hat{b}_1x + \hat{b}_2x^2 + \hat{b}_3x^3 + \hat{b}_4x^4$. The values of a and b_1 for cloud coverage over 81% of the area of China are presented below as an illustration. The mean square error amounts 0.0053626.

a	b_1	b_2
5.4249201	-1.1451431	0.2431453

b_3	b_4
-0.0091621	0.0000001

CONCLUSIONS

1. Algorithms and a set of programs were designed to compute mathematical expectations, variances, correlation matrices, regressions, densities, and distribution laws for certain cloud characteristics from the FGGE data.

2. Seasonal geographic distributions of the cloud top height were plotted for the continents. These results can be used for interpreting the data of satellite remote sensing of the Earth.

3. It is found that distributions of the cloud coverage and cloud top height can be approximated by a log-normal law. The approximation is performed using the Kolmogorov criterion at a 95% confidence level.

REFERENCES

1. FGGE Data Catalogue, World Data CENTER-B, Moscow, (1982).
2. M.G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 1, Distribution Theory, Griffin, London. 1958.
3. M.G. Kendall and A. Stuart, *The Advance Theory of Statistics*, Vol. 2, Statistical Inference and Statistical Relationship, 4-th edition, Hafner, 1979.
4. O.A. Avaste, O.Yu. Kyarner, K.S. Lamden, and K.S. Shifrin, *Optics of the Atm. and Ocean Statistical Characteristics of Cloudiness and Global Radiation above the Various Global Oceans* [in Russian, Moscow (1981)].
5. L.W. Falles, J. Geophys. Res. **79**, No. 9, 1261 (1974).
6. L.T. Matveev [Ed.], *Global Cloudiness* (Gidrometeoizdat, Moscow, 1986).
7. G.A. Korn and T.M. Korn, *Mathematical Handbook for Scientists and Engineers*, McGraw-Hill, New York, 1961.
8. P. Reist, *Aerosols. Introduction to Theory* [in Russian, Mir, Moscow (1987)].
9. G.T. Abezgauz, A.P. Tron', Yu.P. Konenkin, and I.A. Korovina, *Reference Book for Probability Computations* (Voenizdat, Moscow, 1970).
10. H. Cramer, *Mathematical Methods of Statistics*, Princeton Univ. Press, Princeton **5**, 1946.