A generalized method for construction of linear regression and its application to the development of single-parameter aerosol extinction models

N.N. Shchelkanov

Institute of Atmospheric Optics,
Siberian Branch of the Russian Academy of Sciences, Tomsk

Received November 12, 2004

A generalized equation for determination of the regression coefficients of the linear equation $Y = K_0 + K_1X$ is presented for the general case, when the point spread in the correlation between X and Y is caused both by random measurement errors and by uncontrollable physical factors. All the known equations for the regression coefficients appeared to be particular cases of the equation obtained. Methods are proposed and equations are presented for estimation of the rms errors of measured parameters, entering into the equation for calculation of the regression coefficients, from experimental data.

Introduction

Working with the experimental data often calls for determination of the coefficients of linear regression between two physical parameters. In the majority of cases, the regression coefficients have specific physical meaning and for correct interpretation of the obtained results, it is very important to determine their values in a proper way. There are few formulas for determination of the regression coefficients,[1–3] but no general idea exists, which formula should be used in each specific case. Now there is no unified approach to determination of the linear regression coefficient in the general case, i.e., when the scatter of points in the correlation diagram between two values has been caused by both their random measurement errors and uncontrolled physical factors.

The purpose of the present paper is to determine a generalized formula for calculation of the linear regression coefficients and to apply it to construction of single-parameter regional models of the aerosol extinction.

1. Statement of the problem

Let us consider two physical parameters X_0 and Y_0 that correlate. Let us suppose that the correlation can be described by a linear dependence

$$Y_0 = K_0 + K_1 X_0, \qquad (1)$$

and the task is to determine the regression coefficients K_0 and K_1, which describe the physical relation between them in the best way.

As X_0 and Y_0 are measured with random errors, in practice we deal with the values X and Y, for which the linear regression equation is written in the form

$$Y = K_0 + K_1 X. \qquad (2)$$

The form of formulas (1) and (2) with equal regression coefficients is an evidence of the fact that the last should not depend on the random measurement errors in X and Y. Then let us consider determination of only the regression coefficient K_1, because K_0 is calculated after determining K_1 by the known formula

$$K_0 = \overline{Y} - K_1\overline{X}, \qquad (3)$$

where $\overline{X}$ and $\overline{Y}$ are the mean values of X and Y.

2. New approach

New approach to determination of the regression coefficient K_1 is based on two suppositions.

1. The values X and Y are normalized to the quantities $\sqrt{\delta_X^2 + \delta_{X_0}^2}$ and $\sqrt{\delta_Y^2 + \delta_{Y_0}^2}$, respectively.

2. The orthogonal mean-square regression is used for determination of K_1, i.e., the sum of squares of deviations perpendicular to the sought straight line is minimized.

The values δ_X and δ_Y are the random measurement errors in X and Y for the data array considered; δ_{X_0} and δ_{Y_0} are some values, which characterize the scatter of points in the correlation of the physical values X_0 and Y_0 due to uncontrolled physical parameters. Then the regression equation can be written in the form

$$\frac{Y}{\sqrt{\delta_Y^2 + \delta_{Y_0}^2}} = K_0' + K_1' \frac{X}{\sqrt{\delta_X^2 + \delta_{X_0}^2}} \qquad (4)$$

The values δ_{X_0} and δ_{Y_0} are determined here from solution of the system of two equations.

To obtain the first equation, let us use the known formula[1]:

$$\rho_{XY}\sigma_X\sigma_Y = \rho_{X_0Y_0}\sigma_{X_0}\sigma_{Y_0}, \qquad (5)$$

0235-6880 / 05 / 01–02 77-05 $02.00

where ρ_{XY} is the coefficient of correlation between X and Y; σ_X and σ_Y are the rms deviations of X and Y; $\rho_{X_0Y_0}$ is the coefficient of correlation between X_0 and Y_0; σ_{X_0} and σ_{Y_0} are the rms deviations of X_0 and Y_0.

Then let us write the first equation in the form

$$|\rho_{X_0Y_0}|\sigma_{X_0}\sigma_{Y_0} = \sqrt{\sigma_{X_0}^2 - \delta_{X_0}^2}\sqrt{\sigma_{Y_0}^2 - \delta_{Y_0}^2}, \quad (6)$$

where

$$\sigma_{X_0} = \sqrt{\sigma_X^2 - \delta_X^2};\ \sigma_{Y_0} = \sqrt{\sigma_Y^2 - \delta_Y^2}.$$

Then let us write the second equation in the form

$$\delta_{X_0} / \sigma_{X_0} = \delta_{Y_0} / \sigma_{Y_0} \quad (7)$$

and call it the condition of proportionality of the values δ_{X_0}, δ_{Y_0} and σ_{X_0}, σ_{Y_0}. Introduction of the values δ_{X_0}, δ_{Y_0} and formulation of the condition (7) are the key moments in this paper, because this has allowed us to obtain the generalized solution of Eq. (2) for the linear regression coefficients.

3. Results

Upon solving the system of equations (6) and (7) we obtain

$$\delta_{X_0} = \sigma_X\sqrt{\left(1-\frac{\delta_X^2}{\sigma_X^2}\right)\left(1-\frac{|\rho_{XY}|}{\sqrt{\left(1-\delta_X^2/\sigma_X^2\right)\left(1-\delta_Y^2/\sigma_Y^2\right)}}\right)}, \quad (8)$$

$$\delta_{Y_0} = \sigma_Y\sqrt{\left(1-\frac{\delta_Y^2}{\sigma_Y^2}\right)\left(1-\frac{|\rho_{XY}|}{\sqrt{\left(1-\delta_X^2/\sigma_X^2\right)\left(1-\delta_Y^2/\sigma_Y^2\right)}}\right)}. \quad (9)$$

Taking into account expressions (8) and (9), let us determine the values $\sqrt{\delta_X^2 + \delta_{X_0}^2}$ and $\sqrt{\delta_Y^2 + \delta_{Y_0}^2}$ in the following form:

$$\sqrt{\delta_X^2 + \delta_{X_0}^2} = \sigma_X A, \quad (10)$$

$$\sqrt{\delta_Y^2 + \delta_{Y_0}^2} = \sigma_Y B, \quad (11)$$

where

$$A = \sqrt{1-|\rho_{X_0Y_0}|\left(1-\frac{\delta_X^2}{\sigma_X^2}\right)} = \sqrt{1-|\rho_{XY}|\sqrt{\frac{1-\delta_X^2/\sigma_X^2}{1-\delta_Y^2/\sigma_Y^2}}}, \quad (12)$$

$$B = \sqrt{1-|\rho_{X_0Y_0}|\left(1-\frac{\delta_Y^2}{\sigma_Y^2}\right)} = \sqrt{1-|\rho_{XY}|\sqrt{\frac{1-\delta_Y^2/\sigma_Y^2}{1-\delta_X^2/\sigma_X^2}}}. \quad (13)$$

Taking into account Eqs. (10) and (11), let us write the linear regression equation (4) in the form

$$\frac{Y}{\sigma_Y B} = K_0' + K_1'\frac{X}{\sigma_X A}. \quad (14)$$

It is easy to reduce Eq. (14) to the form (2):

$$Y = K_0'\sigma_Y B + K_1'\frac{\sigma_Y B}{\sigma_X A}X = K_0 + K_1 X, \quad (15)$$

where

$$K_0 = K_0' A\sigma_Y B, \quad (16)$$

$$K_1 = K_1'\frac{\sigma_Y B}{\sigma_X A}. \quad (17)$$

Applying the orthogonal mean-square regression to Eq. (14) and using the relationship (17), we obtain the formula for the regression coefficient sought:

$$K_1 = \frac{\sigma_Y}{\sigma_X}\frac{B}{A}\frac{1}{2\rho_{XY}}\left\{\left(\frac{A}{B}-\frac{B}{A}\right)+\sqrt{\left(\frac{A}{B}-\frac{B}{A}\right)^2+4\rho_{XY}^2}\right\}, \quad (18)$$

where A and B are determined by formulas (12) and (13). Formula (18) was presented for the first time in Ref. 4 and described in detail in Ref. 5.

4. Analysis

The formula (18) allows one to establish an unambiguous relation between X and Y and to determine the conditions for applying the known types of linear regression.

Let us show that all analytical formulas known for the regression coefficient K_1 of Eq. (2) are the particular cases of Eq. (18).

4.1. In the case when the scatter of points in correlation between X and Y is caused only by their random errors, i.e., $\rho_{X_0Y_0} = 1$, we obtain the known formula for the regression coefficient K_1 presented in Ref. 1.

$$K_1 = \frac{\delta_Y}{\delta_X}\frac{1}{2\rho_{XY}}\left\{\left(\frac{\sigma_Y}{\sigma_X}\frac{\delta_X}{\delta_Y}-\frac{\sigma_X}{\sigma_Y}\frac{\delta_Y}{\delta_X}\right)+\sqrt{\left(\frac{\sigma_Y}{\sigma_X}\frac{\delta_X}{\delta_Y}-\frac{\sigma_X}{\sigma_Y}\frac{\delta_Y}{\delta_X}\right)^2+4\rho_{XY}^2}\right\}. \quad (19)$$

4.1.1. At $\delta_X = 0$ and $\delta_Y \neq 0$ we have

$$K_1 = \lim_{\delta_X\to 0}\frac{\delta_Y}{\delta_X}\frac{1}{2\rho_{XY}}\left\{\left(\frac{\sigma_Y}{\sigma_X}\frac{\delta_X}{\delta_Y}-\frac{\sigma_X}{\sigma_Y}\frac{\delta_Y}{\delta_X}\right)+\left(-\frac{\sigma_Y}{\sigma_X}\frac{\delta_X}{\delta_Y}+\frac{\sigma_X}{\sigma_Y}\frac{\delta_Y}{\delta_X}\right)\sqrt{1+4\rho_{XY}^2\left(\frac{\sigma_Y}{\sigma_X}\frac{\delta_X}{\delta_Y}\right)^2}\right\}.$$

Expanding the expression under the square root sign into the Maclaurin series[6] and taking only two first terms, we obtain

$$K_1 = \lim_{\delta_X\to 0}\frac{\delta_Y}{\delta_X}\frac{1}{2\rho_{XY}}\times$$

$$\times\left\{\left(\frac{\sigma_Y}{\sigma_X}\frac{\delta_X}{\delta_Y}-\frac{\sigma_X}{\sigma_Y}\frac{\delta_Y}{\delta_X}\right)+\left(-\frac{\sigma_Y}{\sigma_X}\frac{\delta_X}{\delta_Y}+\frac{\sigma_X}{\sigma_Y}\frac{\delta_Y}{\delta_X}\right)\times\right.$$

$$\left.\times\left[1+2\rho_{XY}^2\left(\frac{\sigma_Y}{\sigma_X}\frac{\delta_X}{\delta_Y}\right)^2\right]\right\} = \frac{\sigma_Y}{\sigma_X}\rho_{XY}. \quad (20)$$

It is the well-known formula for the coefficient K_1 of the direct regression $Y = K_0 + K_1 X$, which was

obtained by means of minimization of the sum of squares of deviations along Y axis from the sought straight line.[2]

4.1.2. At $\delta_Y = 0$ and $\delta_X \neq 0$ we have

$$K_1 = \lim_{\delta_Y \to 0} \frac{\delta_Y}{\delta_X} \frac{1}{2\rho_{XY}} \left\{ \left(\frac{\sigma_Y}{\sigma_X} \frac{\delta_X}{\delta_Y} - \frac{\sigma_X}{\sigma_Y} \frac{\delta_Y}{\delta_X} \right) + \right.$$

$$\left. + \left(\frac{\sigma_Y}{\sigma_X} \frac{\delta_X}{\delta_Y} - \frac{\sigma_X}{\sigma_Y} \frac{\delta_Y}{\delta_X} \right) \sqrt{1 + 4\rho_{XY}^2 \left(\frac{\sigma_X}{\sigma_Y} \frac{\delta_Y}{\delta_X} \right)^2} \right\}.$$

Expanding the expression under the square root sign into the Maclaurin series[6] and taking only two first terms, we obtain

$$K_1 = \lim_{\delta_Y \to 0} \frac{\delta_Y}{\delta_X} \frac{1}{2\rho_{XY}} \left\{ \left(\frac{\sigma_Y}{\sigma_X} \frac{\delta_X}{\delta_Y} - \frac{\sigma_X}{\sigma_Y} \frac{\delta_Y}{\delta_X} \right) + \right.$$

$$\left. + \left(\frac{\sigma_Y}{\sigma_X} \frac{\delta_X}{\delta_Y} - \frac{\sigma_X}{\sigma_Y} \frac{\delta_Y}{\delta_X} \right) \left[1 + 2\rho_{XY}^2 \left(\frac{\sigma_X}{\sigma_Y} \frac{\delta_Y}{\delta_X} \right)^2 \right] \right\} = \frac{\sigma_Y}{\sigma_X} \frac{1}{\rho_{XY}}. \quad (21)$$

Formula (21) also is the well known formula for the coefficient $1/K_1^*$ of the inverse regression equation $X = K_0^* + K_1^* Y$, which was obtained by means of minimization of the sum of squares of deviations along X axis from the sought straight line.[2]

4.1.3. At $\delta_X = \delta_Y \neq 0$ we obtain the known formula

$$K_1 = \frac{1}{2\rho_{XY}} \left\{ \left(\frac{\sigma_Y}{\sigma_X} - \frac{\sigma_X}{\sigma_Y} \right) + \sqrt{\left(\frac{\sigma_Y}{\sigma_X} - \frac{\sigma_X}{\sigma_Y} \right)^2 + 4\rho_{XY}^2} \right\} \quad (22)$$

for the coefficient K_1 of the orthogonal regression equation $Y = K_0 + K_1 X$, which was obtained by means of minimization of the sum of squares of deviations perpendicular to the sought straight line.[3]

4.2. If $\delta_X/\sigma_X = \delta_Y/\sigma_Y$, the formula for the regression coefficient

$$K_1 = \sigma_Y/\sigma_X. \quad (23)$$

follows from Eq. (18). Let us note that Eq. (23) is the geometric mean of Eqs. (20) and (21).

5. The range of variability of the regression coefficient

In the case when the scatter of points in correlation between the values X and Y has been caused only by their random errors, i.e., $\rho_{X_0Y_0} = 1$, the regression coefficient varies within the following limits:

$$\frac{\sigma_Y}{\sigma_X} |\rho_{XY}| \le |K_1| \le \frac{\sigma_Y}{\sigma_X} \frac{1}{|\rho_{XY}|}, \quad (24)$$

and at $\rho_{X_0Y_0} < 1$

$$\frac{\sigma_Y}{\sigma_X} |\rho_{XY}| < |K_1| < \frac{\sigma_Y}{\sigma_X} \frac{1}{|\rho_{XY}|}. \quad (25)$$

As is seen from Eqs. (24) and (25), the coefficients of direct and inverse regressions take their minimum and maximum values, respectively.

6. Calculation of random errors from the experimental data

Formula (18) is valuable only when the random errors δ_X and δ_Y have been known. The more accurate the random errors are determined, the less is the errors in calculating the regression coefficient K_1.

In some cases, the values δ_X and δ_Y can be determined from the experimental data. Two such arrays can be obtained for X and Y. Then, using Eq. (5), we can determine the sought errors:

$$\delta_X = \sigma_X \sqrt{1 - |\rho_{XX}|}, \quad (26)$$

$$\delta_Y = \sigma_Y \sqrt{1 - |\rho_{YY}|}, \quad (27)$$

where ρ_{XX} and ρ_{YY} are the normalized autocorrelation coefficients of X and Y, respectively.

If no initial data have been available, one can use the formulas for approximate estimates in order to determine the errors by the precise formulas (26) and (27). To do this, let us select the values X and X', Y and Y', insignificantly differing from each other.

Assuming in Eq. (5) $\delta_X = \delta_{X'}$, $\delta_Y = \delta_{Y'}$, $\rho_{X_0X_0'} = 1$, $\rho_{Y_0Y_0'} = 1$, we obtain the upper estimates of δ_X and δ_Y:

$$\delta_X = \sqrt{\frac{\sigma_X^2 + \sigma_{X'}^2}{2} - \sqrt{(\frac{\sigma_X^2 - \sigma_{X'}^2}{2})^2 + \rho_{XX'}^2 \sigma_X^2 \sigma_{X'}^2}}, \quad (28)$$

$$\delta_Y = \sqrt{\frac{\sigma_Y^2 + \sigma_{Y'}^2}{2} - \sqrt{(\frac{\sigma_Y^2 - \sigma_{Y'}^2}{2})^2 + \rho_{YY'}^2 \sigma_Y^2 \sigma_{Y'}^2}}. \quad (29)$$

Let us note that Eqs. (26) and (27) are the particular cases of Eqs. (28) and (29) under condition that $\sigma_X = \sigma_{X'}$ and $\sigma_Y = \sigma_{Y'}$, respectively.

If one of the errors (δ_X or δ_Y) has been known, and the scatter of points in the sought dependence has been caused only by random errors (i.e., $\rho_{X_0Y_0} = 1$), then, according to Eq. (5), the values of other errors are calculated by the following formulas:

$$\delta_X = \sigma_X \sqrt{1 - \rho_{XY}^2 \frac{\sigma_Y^2}{\sigma_Y^2 - \delta_Y^2}} \quad (30)$$

or

$$\delta_Y = \sigma_Y \sqrt{1 - \rho_{XY}^2 \frac{\sigma_X^2}{\sigma_X^2 - \delta_X^2}}. \quad (31)$$

If one of the errors (δ_X or δ_Y) has been equal to zero, then, as follows from Eqs. (30) and (31), the values of other errors are calculated by the following formulas:

$$\delta_X = \sigma_X \sqrt{1 - \rho_{XY}^2} \quad (32)$$

or

$$\delta_Y = \sigma_Y \sqrt{1 - \rho_{XY}^2}. \quad (33)$$

Thus, the use of the formulas (26)–(33) under different conditions makes it possible to estimate the random errors of the measured parameters directly from the experimental data.

7. Application of the generalized formula to construction of a single-parameter model of the aerosol extinction

Let us demonstrate the capability and efficiency of applying Eq. (18) by considering the case of constructing a model of aerosol extinction, which allows one to calculate the aerosol extinction in the wavelength range 1.06 μm (Y) from the data on the aerosol extinction coefficient in the wavelength range 0.48 μm (X). To do this, let us use the experimental data obtained in arid zone of Kazakhstan in summer.[7] Correlation between the aerosol extinction coefficients at the wavelengths of 0.48 and 1.06 μm is shown in Fig. 1. Let us divide the entire array into three subarrays I, II, and III, as is shown in the figure by dotted line, and then let us form four different arrays of their combinations.

To calculate the random errors δ_X and δ_Y, let us use formulas (28) and (29), and take the arrays of the aerosol extinction coefficients in the nearest wavelength ranges 0.55 and 0.87 μm as the missing values X' and Y'. Calculating the statistical characteristics of the aerosol extinction coefficients at the wavelengths of 0.48, 0.55, 0.87, and 1.06 μm for each array, we determine the values of the random errors.

The numbers of the formed subarrays are shown in Table 1, as well as their dimension, composition of the values of the statistical characteristics, which are necessary for calculation of the regression coefficients of the linear equation $\alpha(1.06) = K_0 + K_1\alpha(0.48)$ by the generalized formula (18) and by formulas (19), (20), and (22), where $\alpha(1.06) = Y$, and $\alpha(0.48) = X$.

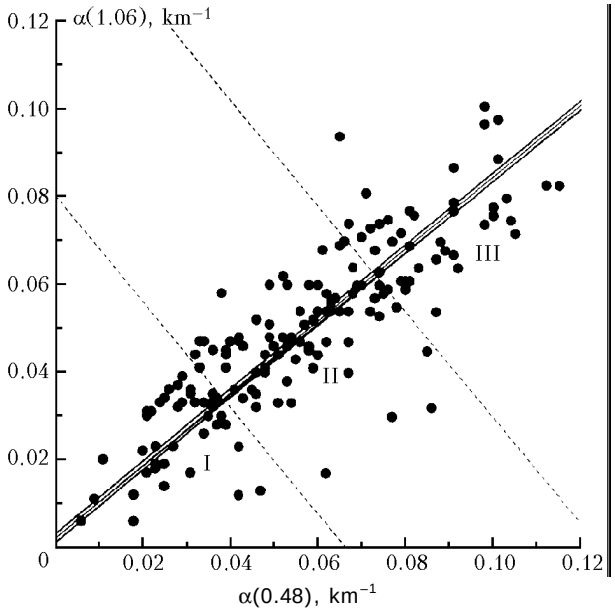

Fig. 1.

Parameters of the models of the aerosol extinction $\alpha(1.06) = K_0 + K_1\alpha(0.48)$ calculated by formulas (18)–(20) and (22) are presented in Table 2. The fact attracts one's attention that the models offering the regression Y on X are obtained for different arrays by Eq. (20) and are characterized by unstable values of the regression coefficients.

The models obtained from Eq. (22) for the orthogonal mean-square regression and by Eq. (19) for the structure ratio[1] provide more stable values of the regression coefficients, but also are unstable, this is well seen from the array No. 4, as an example.

Table 1. Numbers, dimension, and statistical characteristics of the formed arrays

Array			ρ_{XY}	$\overline{X}$, km^{-1}	σ_X, km^{-1}	δ_X, km^{-1}	$\overline{Y}$, km^{-1}	σ_Y, km^{-1}	δ_Y, km^{-1}
No.	dimension	composition							
1	160	I + II + III	0.84	0.057	0.0241	0.0063	0.049	0.0200	0.0050
2	120	I + II	0.71	0.047	0.0177	0.0058	0.041	0.0146	0.0051
3	120	II + III	0.72	0.067	0.0194	0.0066	0.057	0.0160	0.0049
4	80	II	0.33	0.056	0.0127	0.0071	0.049	0.0100	0.0049

Table 2. Parameters of the aerosol extinction models $\alpha(1.06) = K_0 + K_1\alpha(0.48)$

No. of the array	Formula							
	(20)		(22)		(19)		(18)	
	K_1	K_0, km^{-1}	K_1	K_0, km^{-1}	K_1	K_0, km^{-1}	K_1	K_0, km^{-1}
1	0.70	0.009	0.80	0.004	0.84	0.001	0.83	0.002
2	0.59	0.013	0.76	0.005	0.80	0.003	0.82	0.002
3	0.59	0.017	0.77	0.006	0.86	–0.001	0.83	0.001
4	0.26	0.034	0.50	0.020	1.00	–0.007	0.83	0.003

At the same time, the models obtained for different arrays by the generalized formula (18) are less different from each other. These four models are shown in Fig. 1 by solid lines. Thus, Eq. (18) makes it possible to obtain stable and reliable regional models, practically independent of the dimension of the array and the correlation coefficient ρ_{XY}. Besides, the models obtained by Eq. (18) are physically correct. The Kazakhstan arid zone in summer is characterized by practically quasi-neutral spectral behavior of the aerosol extinction coefficients.[7] So, the regression coefficient K_0 of the linear equation $\alpha(1.06) = K_0 + K_1\alpha(0.48)$ should be close to zero. As is seen in Table 2, the generalized formula (18) gives the closest to zero values K_0 for all four arrays. Taking into account stability of the obtained regression coefficients and physical correctness of the results, Eq. (18) is preferable for construction of linear regression models of the aerosol extinction.

Conclusions

Let us briefly formulate the principal results.

1. The generalized formula has been obtained, which makes it possible to determine the regression coefficients of the linear equation $Y = K_0 + K_1 X$ for the general case when the scatter of points in correlation between the parameters X and Y has been caused by both their random measurement errors and uncontrolled physical factors.

2. All known relationships for the regression coefficients are particular cases of the obtained formula.

3. The ways are proposed and the formulas are presented for estimation of the random rms errors of the measured values entering the formula for calculation of the regression coefficients from the experimental data.

4. The generalized formula makes it possible to construct stable reliable and physically correct single-parameter models.

The obtained formula is interesting for specialists in processing the experimental data and can be used for correct physical interpretation independent of the research field.

Acknowledgments

In conclusion, the author would like to thank Prof. A.A. Mitsel for useful remarks.

References

1. M.A. Kendall and A. Stuart, *Inference and Relationship* (C. Griffin, London), Vol. 2.
2. A.N. Zaidel, *Errors in Measuring Physical Parameters* (Nauka, Leningrad, 1985), 112 pp.
3. T. Cramer. *Mathematical Methods of Statistics* [Russian translation] (Mir, Moscow, 1975), 648 pp.
4. N.N. Shchelkanov, in: *Abstracts of Reports at II Workshop on Siberian Aerosols*, Tomsk (1995), p. 16.
5. N.N. Shchelkanov, "*A generalized method for construction of linear regression between two physical parameters taking into account their random errors*," Preprint No. 2, Institute of Atmospheric Optics, Tomsk (2003), 15 pp.
6. V.A. Kudryavtsev and B.P. Demidovich, *Brief Course of High Mathematics* (Nauka, Moscow, 1975), 624 pp.
7. Yu.A. Pkhalagov, V.N. Uzhegov, and N.N. Shchelkanov, Atmos. Oceanic Opt. 7, No. 10, 714–720 (1994).