Modern Environmental Science and Engineering (ISSN 2333-2581)
April 2018, Volume 4, No. 4, pp. 350-357
Doi: 10.15341/mese(2333-2581)/04.04.2018/007
Academic Star Publishing Company, 2018
www.academicstar.us

Statistical Analysis of Hourly Clearness Index and Diffuse Fraction Data Using Beta Probability Density Functions

Mathias Bashahu, and J. Ntirandekura

Department of Physics and Technology, Institute of Applied Pedagogy, University of Burundi, Burundi

Abstract: An analysis of hourly clearness index (k_{th}) and diffuse fraction (k_{dh}) data is presented in this paper. Those data have been extracted from a 1-year period records of the hourly global and diffuse solar irradiation on a horizontal surface from a subtropical African station, namely Dakar-Hann (Senegal). The monthly frequency distributions of those data have been firstly constructed and their main characteristics have been derived. Then, the theoretical beta probability functions (β-PDFs) fitting the observed frequency distributions have been implemented. Three complementary statistical tests, i.e., the mean bias error (MBE), the root mean square error (RMSE) and the t-statistics, have been performed to check the effectiveness of the fitting procedure. For any of the twelve monthly sets of k_{th} and k_{dh} data, the MBE has been found equal (or close) to zero as expected. The RMSE was much lower than the mean value (of the k_{th} or k_{dh} data), and the t-statistics was almost equal to zero or much lower than its critical value (t_c = 3.250). All the obtained β-PDFs are therefore good fits of the corresponding observed monthly frequency distributions of k_{th} and k_{dh} data. They should be used consequently as input data in any project of solar energy conversion or storage systems at the considered station.

Key words: β-PDFs, computation procedure for the basic data, hourly clearness index, hourly diffuse fraction, relative frequency distributions, statistical tests.

1. Introduction

The solar global irradiation on a horizontal surface is a key input parameter required when planning solar energy conversion and storage systems. Its hourly and/or daily data (G_h and/or G, respectively) are commonly recorded at a large number of meteorological stations. Hourly and/or daily data of the solar diffuse irradiation on a horizontal surface (D_h and/or D, respectively) are also sometimes recorded at many stations. However, the corresponding records of the solar direct (or beam) irradiation on a horizontal surface (I_h and/or I, respectively) are rather scare. Statistical properties of those quantities are often needed when looking for the long-term mean performance of the above-mentioned systems. Those properties can be extracted from the analysis of monthly frequency distributions of actual data on those quantities. Nevertheless, it is more convenient to handle suitable analytical expressions playing the role of probability density functions (PDFs), which represent those empirical frequency distributions.

The beta probability density functions (β-PDFs) make up one kind of such expressions. As a matter of fact, they have been demonstrated to successfully fit monthly observed frequency distributions of some dimensionless physical quantities having values which range randomly from zero to one. Those are notably the daily relative sunshine duration (s_r) [1, 2], the daily clearness index (k_t) [3], and the hourly relative wind speed (v_r) [4]. The random dimensionless variables to be considered in this paper are the hourly clearness index (k_{th}) and diffuse fraction (k_{dh}). The objective of

Corresponding author: Mathias Bashahu, Professor; research area/interest: renewable energy. E-mail: bashahuma@yahoo.fr.

this study is two-fold. First, to present the main characteristics of the monthly frequency distributions of the observed k_{th} and k_{dh} data for a selected station. Second, to fit those frequency distributions by suitable β-PDFs and to assess the effectiveness of the fitting procedure.

2. Materials

2.1 Selected Station

The data of the hourly solar global and diffuse irradiation on a horizontal surface (G_h and D_h, respectively) used in this study refer to a 1-year period records at the station of Dakar-Hann (Senegal: $L_e = -27.43°$, Φ = 14.72°, z = 10 m). L_e, Φ and z are the longitude, latitude and altitude of the station, respectively. The only criterion used for that selection was the simultaneous availability of data on the two physical quantities (G_h and D_h).

2.2 Computation Procedure for the Basic Data

The hourly clearness index (k_{th}) and diffuse fraction (k_{dh}) data to be processed have been obtained by using the following ratios:

$$k_{th} = \frac{G_h}{G_{oh}} \tag{1}$$

$$k_{dh} = \frac{D_h}{G_{oh}} \tag{2}$$

Values of the extra-terrestrial hourly solar irradiation on a horizontal surface (G_{oh}) have been computed (in $Wm^{-2}h^{-1}$) by the means of the next relationship:

$$G_{oh} = E_0\, I_{sc}\,(3600)\left[sin\delta\, sin\Phi + \frac{24}{\pi}\sin\left(\frac{\pi}{24}\right) cos\delta\, cos\Phi\, \cos\omega_{si}\right] \tag{3}$$

In Eqs. (1) and (2), the quantities G_h and D_h have been expressed in the same units as G_{oh}. In the Eq. (3), the solar constant I_{sc} has been taken equal to 1367 Wm^{-2} [5]. The parameters I_{sc} and **Φ** were constant for the considered station and period. Values of the eccentricity correction factor (E_o) and the solar declination (δ) depend on the order number (J) of the day in the year. They were extracted from relevant tables [5]. The quantity ω_{si} in Eq. (3) is the solar angle at the middle solar instant t_i of a given i-th hourly interval of the day. Its values (in degrees when t_i is in hours) were determined using the following common expression:

$$\omega_{si} = 15(t_i - 12) \tag{4}$$

According to Eq. (4), ω_{si} is positive after (and negative before) solar noon. At its turn, the solar time t_i was related to the standard local time t_i' (both in hours) as following for the considered station:

$$t_i = t_i' + \frac{4}{60}(L_s - L_e) + E_t \tag{5}$$

In Eq. (5), the values of t_i' are the round hours from 6:00 to 19:00 (every day). L_s is the longitude of the standard meridian closest to the station (i.e., L_S = 30.00°), so that the longitude correction term is here positive [5] and remains constant (i.e. equal to 0.1713 h) all the time for that station. At their turn, values (in hours) of the equation of time (E_t) have been extracted from published relevant data (in minutes), e.g., from [5].

Moreover, in the computation of the quantities t_i and ω_{si} (from Eqs. (4) and (5)), negative values of G_{oh} (from Eq. (3)) have been avoided by bounding the every day's t_i values between their lowest and highest limits, i.e. the sunrise (t_{sr}) and sunset (t_{ss}) instants, respectively. These two parameters are related to the corresponding solar angles (ω_{sr} and ω_{ss}, respectively) by the following common expressions:

$$\omega_{ss} = cos^{-1}(-tg\delta\; tg\Phi) \tag{6}$$

$$\omega_{sr} = -\,\omega_{ss} \tag{7}$$

$$t_{sr} = 12 + \frac{\omega_{sr}}{15} \tag{8}$$

$$t_{ss} = 12 + \frac{\omega_{ss}}{15} \tag{9}$$

where ω_{ss} is positive, the angles are in degrees and the instants are in hours. Finally, in the results of those computations, values of k_{th} and k_{dh} higher than one have been removed.

3. Methodology

3.1 Characteristics of the Observed Frequency Distributions

The previous computation procedure has led to twelve monthly sets of both k_{th} and k_{dh} data. The data of each set have been firstly arranged into the following ten intervals of the same width: [0; 0.10], [0.10; 0.20], … , [0.90; 1.00]. Then, the relevant relative frequency distribution has been implemented, together with its average (x_m), variance (σ_m^2), variation coefficient (c_v) and mean deviation of the data about their average ($\overline{\Delta x}$), which are defined as follows:

$$x_m = \sum_{i=1}^{n} x_i f_{i,exp} \quad (10)$$

$$\sigma_m^2 = \sum_{i=1}^{n} (x_i - x_m)^2 f_{i,exp} \quad (11)$$

$$c_v = \frac{\sigma_m}{x_m} \quad (12)$$

$$\overline{\Delta x} = \sum_{i=1}^{n} (x_i - x_m) f_{i,exp} \quad (13)$$

In Eqs. (10) to (13), n=10 and the x_i are the middle points of the above-mentioned intervals.

3.2 Fitting Procedure

Similarly to the procedure used in earlier studies related to the dimensionless quantities s_r, k_t and v_r [1-4], the k_{th} or k_{dh} data of each monthly set are represented by a random variable X, which values x range continuously from 0 to 1. Moreover, those x values are assumed to follow a β-PDF in terms of an Eulerian integral of the first kind, which is expressed as:

$$f(x) = x^{p-1}(1-x)^{q-1}/\beta \quad (14)$$

where $\beta(p, q)$ is the beta distribution with parameters p and q, defined as

$$\beta(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)} \quad (15)$$

with p >0, q > 0 and the gamma function is given by :

$$\Gamma(m) = \int_0^\infty t^{m-1}e^{-t}dt, \quad m > 0, t > 0 \quad (16)$$

Here, m is equal to p, q or p+q. The mean ($\bar{x}$) and the variance (σ^2) of that theoretical distribution are expressed as:

$$\bar{x} = \frac{p}{p+q} \quad (17)$$

$$\sigma^2 = \frac{pq}{(p+q)^2(p+q+1)} \quad (18)$$

For a given k_{th} or k_{dh} data set, the mean ($\bar{x}$) and the variance (σ^2) of the β-PDF to be constructed are assumed to be identical to the mean (x_m) and the variance (σ_m^2) of the observed relative frequency distribution of that data set. In practice, one firstly computes the quantities x_m and σ_m^2 for the data set under analysis. Secondly, the quantities $\bar{x}$ and σ^2 in Eqs. (17) and (18), are replaced by x_m and σ_m^2, respectively. Then, solving both equations for p and q, one obtains the following relationships:

$$p = Ax_m \quad (19)$$

$$q = A(1-x_m) \quad (20)$$

$$A = [x_m(1-x_m)/\sigma_m^2] - 1 \quad (21)$$

In the next step, the obtained values of the parameters p and q are used in Eq. (14) to construct the wished β-PDF.

Then, the curve of that function is plotted on the same system of co-ordinates axes as the observed relative frequency curve.

3.3 Test of the Efficiency of the Obtained β-PDFs

The effectiveness of the previous fitting technique has been checked through the implementation of the next three complementary statistical tests, which usefulness has been shown in various studies [6-9]:

(1) the mean bias error (MBE):

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(f_{i,th} - f_{i,exp}) \quad (22)$$

where n = 10 here, while $f_{i,th}$ and $f_{i,exp}$ are the theoretical and experimental relative frequency distributions (of a given k_{th} or k_{dh} data set), respectively;

(ii) the root mean square error (RMSE):

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}\left(f_{i,th} - f_{i,exp}\right)^2\right]^{1/2} \quad (23)$$

(iii) and the t-statistics:

$$t = \left[\frac{(n-1)(MBE)^2}{(RMSE)^2-(MBE)^2}\right]^{1/2} \quad (24)$$

An extended discussion on those statistical tests can be found in the previous references. In particular, the t-statistics indicates whether or not the model's estimates are statistically significant at a given confidence level. Standard statistical tables (e.g., in [10]) allow one to read the critical t value (i.e. $t_c = t$ (n-1, α)) at a particular level of significance (α) and n-1 degrees of freedom. The model's estimates are judged statistically significant at a confidence level γ = 1 – α, when the next relationship is satisfied:

$$t < t_c \quad (25)$$

4. Results and Discussion

4.1 On Hourly Clearness Index Data

4.1.1 Observed Frequency Distributions of k_{th} Data

The values of the characteristics x_m, σ_m, c_v and $\overline{\Delta x}$ (defined in section 3.1) of the observed monthly frequency distributions of k_{th} data are exhibited in Table 1. N is the total number of k_{th} data in each monthly set. The values of the parameters of the corresponding β-PDFs and statistical tests are also shown in that table. The observed frequency distributions of k_{th} data are presented in blue or green lines for some monthly sets in Fig. 1. An oversight on the results of Table 1 and the curves in blue or green lines allows one to point out the following features. Maximums of the observed distributions are noticed at relatively high values of x = k_{th}. Those maximums occur at x = 0.85 for two months (April and May), x = 0.75 for five months (January, February, March, June and October), x = 0.65 for four months (July, September, November, and December), and x = 0.55 for one month (August) at Dakar-Hann. A large scatter of the k_{th} data about their mean value is observed in any of the monthly distributions. This is in accordance with the relatively high values of the variation coefficient (c_v ranges from 33.1% to 48.4%). Nevertheless, the average difference between k_{th} data and their mean value in each monthly set is found very low as predicted by the theory. Moreover, the lowest mean values of k_{th} are noticed during the rainy season (from July to September), while high mean values of k_{th} (from 0.52 to 0.63) are observed during the remaining period of the year at Dakar-Hann. An annual average of k_{th} data, $\overline{x}_m = 0.56$, is obtained for that station (in the year 1982). This is in accordance with results from earlier works [8, 11] and indicates a quite sunny sky over the station.

4.1.2 β-PDFs Fitting the Observed Frequency Distributions of k_{th} Data

The values of the parameters p and q calculated according to the content of section 3.2 and used to construct the β-PDFs fitting the observed monthly frequency distributions of k_{th} data, are indicated in Table 1. That table also shows the results of the statistical quantities MBE, RMSE and t-statistics implemented to evaluate the degree at which the obtained β-PDFs fit the corresponding experimental frequency distributions of k_{th} data. Curves of some of the twelve β-PDFs are represented in red or orange

Table 1 Results on the quantities x_m, σ_m, c_v and $\overline{\Delta x}$ of the monthly frequency distributions of k_{th} data, together with the parameters p and q of the related β-PDFs and the variables MBE, RMSE and t of the statistical tests; Dakar-Hann (1982).

Months → Quantities ↓	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
N	324	308	340	348	363	360	372	341	329	312	298	312
x_m	0.60	0.59	0.63	0.61	0.62	0.58	0.50	0.44	0.51	0.55	0.52	0.52
$\overline{\Delta x}$	-0.0022	+0.0162	-0.0012	-0.0034	-0.0187	-0.0034	+0.0046	-0.0002	-0.0016	+0.0022	+0.0038	-0.0031
σ_m	0.2562	0.2210	0.2083	0.2116	0.2755	0.2507	0.2253	0.1931	0.2136	0.2610	0.2513	0.2519
c_v (%)	42.7	37.5	33.1	34.7	44.4	43.2	45.1	43.9	41.9	47.5	48.3	48.4
p	1.5918	2.3286	2.7537	2.6293	1.3045	1.6662	1.9607	2.4666	2.2850	1.4489	1.5337	1.5272
q	1.0612	1.6182	1.6173	1.6810	0.7996	1.2066	1.9607	3.1393	2.1953	1.1855	1.4157	1.4097
MBE	-0.0000	-0.0000	+0.0000	-0.0000	-0.0030	-0.0103	+0.0000	+0.0000	-0.0001	-0.0000	+0.0003	+0.0000
RMSE	0.0580	0.0580	0.0495	0.0480	0.0610	0.0610	0.0652	0.0270	0.0288	0.0413	0.0344	0.0403
t	0.000	0.000	0.000	0.000	0.514	0.514	0.000	0.000	0.010	0.000	0.026	0.000

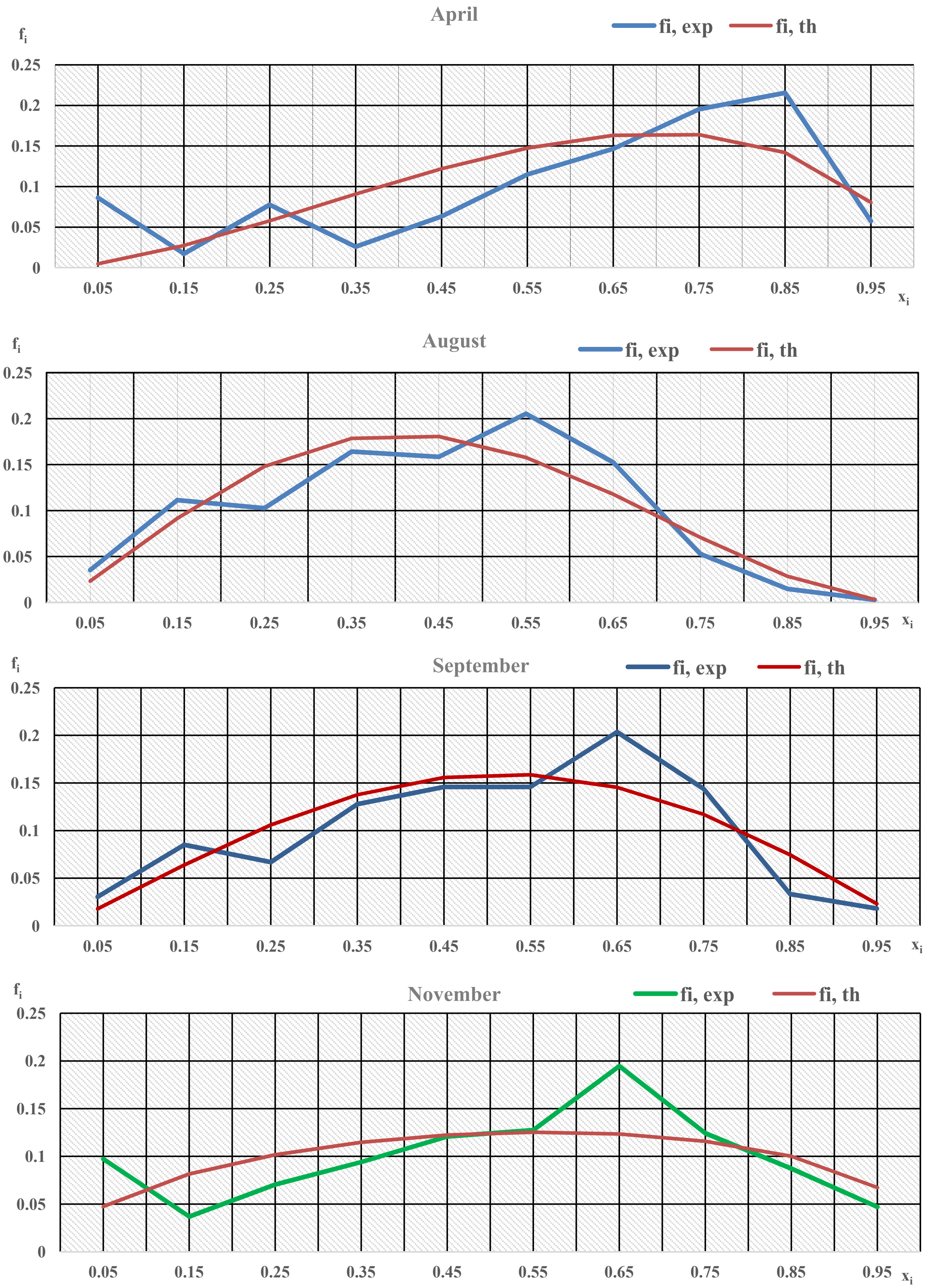

Fig. 1 Curves of the frequency distributions of k_{th} data for some monthly sets, Dakar-Hann (1982): (a) experimental data ($f_{i, exp}$, blue or green lines); (b) theoretical data ($f_{i, th}$, red or orange lines).

lines on Fig. 1. The following features can be pointed out from those results. For any of the twelve k_{th} data sets, as expected, the MBE (between $f_{i,th}$ and $f_{i,exp}$ values) is found either equal or close to zero. The RMSE is far lower than the k_{th} mean value (x_m). In comparison with its critical value t_c = t (n-1=9; γ = 99.5%) which is equal to 3.250 here, the t-statistics is: (i) almost equal to zero for 7 months, (ii) far lower than t_c (with the ratio t_c/t ranging from 64 to 235) for 3 months, and (iii) lower than t_c (with the ratio t_c/t ranging from 6 to 24) for 2 months. All the twelve theoretical β-PDFs implemented for the k_{th} data of this study are therefore good fits of the corresponding observed monthly frequency distributions. Moreover, the best degrees of fitting (of $f_{i, exp}$ data by $f_{i, th}$ ones) are noticed for the monthly sets exhibiting the lowest RMSE values consorted with MBE and t values equal (or close) to zero. August and september k_{th} data sets are two examples of such a case.

4.2 On Hourly Diffuse Fraction Data

4.2.1 Observed Frequency Distributions of k_{dh}Data

Different characteristics of the observed frequency distributions of hourly diffuse fraction (k_{dh}) data for Dakar-Hann (1982) are shown in Table 2. Here also, N in the total number of k_{dh} data in each monthly set. The results of the determination of the parameters p and q of the related β-PDFs, together with the statistical variables MBE, RMSE and t-statistics, are also exhibited in that table. Some of the above-mentioned frequency distributions of k_{dh} data are presented in blue lines in Fig. 2. The following main features can be pointed out from an analysis of those distributions. The maximums of those distributions are observed at relatively low values of x = k_{dh}. As a matter of fact, they occur at x = 0.25 for eight months (from February to September) and at x = 0.15 for the remaining four months of the year. Moreover, the maximums of all the observed frequency distributions are higher for k_{dh} data sets (where they range from 0.35 to 0.53) than for the corresponding k_{th} data sets (where they range from 0.19 to 0.30) for the station and period of this study. Accordingly, the observed frequency values ($f_{i, exp}$) of k_{dh} data lay out on shorter ranges of x values than the corresponding frequency values of k_{th} data. As an example, it is noticed that most of the time, i.e. for 9 months out of 12, the $f_{i, exp}$ values of k_{dh} data are equal to zero for x = k_{dh} larger than 0.5, 0.6 or 0.7. As a resulting statement, the monthly mean values (x_m) of k_{dh} data are lower than the corresponding mean values of k_{th} data. They range from 0.19 to 0.28, with an annual average $\bar{x}_m = 0.23$ for the station and period of this analysis. This is once again an index for a quite sunny sky over that station.

Table 2 Results on the quantities x_m, σ_m, c_v and $\overline{\Delta x}$ of the monthly frequency distributions of k_{dh} data, together with the parameters p and q of the related β-PDFs and the variables MBE, RMSE and t of the statistical tests; Dakar-Hann (1982).

Months → Quantities ↓	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
N	341	308	341	360	372	360	372	372	360	341	317	334
x_m	0.19	0.24	0.24	0.23	0.26	0.28	0.26	0.20	0.22	0.23	0.23	0.22
$\overline{\Delta x}$	+0.0010	-0.0008	-0.0023	+0.0033	+0.0005	-0.0003	+0.0018	+0.0043	-0.0042	+0.0023	-0.0002	+0.0049
σ_m	0.1083	0.1051	0.0820	0.1055	0.1137	0.1180	0.1015	0.0891	0.1056	0.1651	0.1633	0.1499
c_v (%)	56.5	43.8	34.2	45.9	43.3	42.1	39.0	44.6	48.0	71.8	71.0	68.1
p	2.3092	3.7396	6.2937	3.6697	3.6178	3.7810	4.5967	3.8506	3.1507	1.2675	1.2956	1.4579
q	9.8446	11.8422	19.9302	12.2854	10.2969	9.7226	13.0829	15.4026	11.1707	4.2475	4.3374	5.1688
MBE	+0.0000	-0.0000	-0.0000	+0.0000	-0.0000	+0.0000	-0.0000	+0.0000	+0.0000	-0.0000	+0.0000	-0.0000
RMSE	0.0602	0.0030	0.0141	0.0250	0.0538	0.0639	0.0729	0.0597	0.0506	0.0876	0.0893	0.0803
t	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

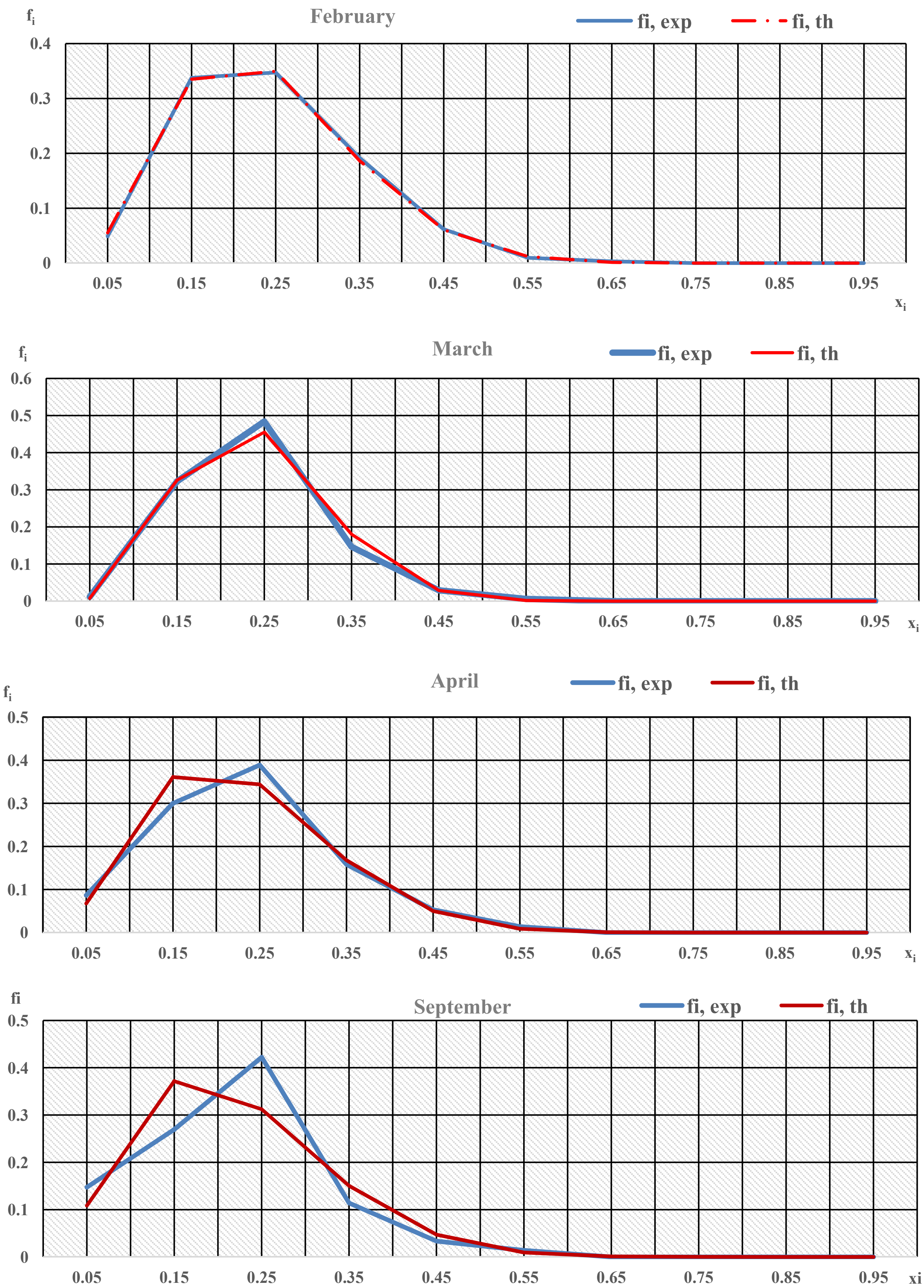

Fig. 2 Curves of the frequency distributions of k_{dh} data for some monthly sets, Dakar-Hann (1982): (a) experimental data ($f_{i,exp}$, blue lines); (b) theoretical data ($f_{i,th}$, red or orange lines).

4.2.2 β-PDFs Fitting the Observed Frequency Distributions of k_{dh} Data

The β-PDFs fitting the experimental frequency distributions of k_{dh} data have been constructed by using the p and q values of Table 2 and as indicated in section 3.2. Their curves are represented in red or orange lines on Fig. 2 for some monthly sets. The comparison between the observed and the theoretical frequency values has led to the following results for any of the twelve k_{dh} data sets. The MBE is almost equal to zero as expected. Consequently, the t-statistical is also almost equal to zero and is therefore lower than its critical value, $t_c = 3.250$. The RMSE is at its turn lower than the mean value (x_m) of k_{dh} data. All the twelve β-PDFs obtained so far are therefore good fits of the corresponding experimental monthly frequency distributions of k_{dh} data. Moreover, almost perfect fits are noticed for monthly sets with high ratios between x_m and RMSE. This is for instance the case for the February, March and April sets where those ratios are equal to 80.0, 17.0 and 9.2, respectively.

5. Conclusions

The monthly frequency distributions of hourly clearness index (k_{th}) and diffuse fraction (k_{dh}) data from Dakar-Hann and a 1-year period of observations have been firstly presented in this study and their main characteristics have been extracted. Then, the theoretical β-PDFs fitting those distributions have been constructed. The effectiveness of the fitting procedure has been tested through the computation of three complementary statistical variables, i.e., the MBE, the RMSE and the t-statistics. The MBE has been found equal (or close) to zero as expected for any of k_{th} and k_{dh} monthly sets. The RMSE was much lower than the mean value of k_{th} or k_{dh}. At its turn, the t-statistics was almost equal to zero or much lower than its critical value ($t_c = 3.250$). Therefore, even if the k_{th} and k_{dh} data used are related to a quite short period of observations, the β-PDFs obtained in this study are good fits of their experimental counterparts. Consequently, they should be very useful input data when planning any solar energy conversion and storage system at the station of this analysis.

References

[1] F.Y. Ettoumi, A. Mefti, A. Adane and M.Y. Bouroubi, Statistical analysis of solar measurements in Algeria using beta distributions, *Renewable Energy* 26 (2002) 47-67.

[2] M. Bashahu, J.C. Nsabimana and F. Manirakiza, Beta probability density functions and reference distributions of relative sunshine duration data from Burundian stations, *ISESCO Science and Technology Vision* 2 (2006) (1) 30-34.

[3] M. Bashahu and B. Butisi, Statistical analysis of daily solar global irradiation data using beta distributions at some Burundian stations, *International Journal of Physical Science* 2 (2010) (3) 35-41.

[4] M. Bashahu and M. Buseke, Statistical analysis of hourly wind speed data from Burundian stations using beta probability density functions, *Modern Environmental Science and Engineering* 2 (2016) (11) 740-746.

[5] M. Iqbal, *An Introduction to Solar Radiation*, Toronto, New York,Tokyo: Academic Press, 1983, pp.3-27, 59-69.

[6] C.C.Y. Ma and M. Iqbal, Statistical comparison of solar radiation correlations, *Solar Energy* 33 (1984) (2) 143-148.

[7] I. T. Toğrul and H. Toğrul, Global solar radiation over Turkey: Comparison of predicted and measured data, *Renewable Energy* 25 (2003) 55-67.

[8] M. Bashahu, Statistical comparison of models estimating the monthly average daily diffuse radiation at a subtropical African site, *Solar Energy* 75 (2003) 43-51.

[9] M. Bashahu and D. Mpanzimana, Estimation of the monthly average daily solar global irradiation with climatological parameters at some Burundian stations, *International Review of Physics* 3 (2009) (4) 237-243.

[10] J. Honerkamp, *Stochastic Dynamical Systems: Concepts, Numerical Methods, Data Analysis*, New York, Weinheim, Cambridge: VCH Publishers, 1999, pp. 511-512.

[11] M. Bashahu and P. Nkundabakura, Analysis of daily global irradiation data for five sites in Rwanda and one in Senegal, *Renewable Energy* 4 (1994) (4) 425-435.