A retrieval of the climatological ocean-atmosphere turbulent fluxes by using the properties of special type distribution

KONSTANTIN BELYAEV
Research Centre Bureau of Meteorology, AUSTRALIA and P.P. Shirshov Institute of Oceanology, RAS, MOSCOW, e-mail: K.Beliaev@bom.gov.au

SERGEY GULEV
P.P. Shirshov Institute of Oceanology, RAS, MOSCOW, e-mail: gul@sail.msk.ru

Abstract - In order to improve the accuracy of computation of climatological ocean-atmosphere turbulent fluxes a special type of distribution is proposed. Analysis of air-sea fluxes in the North Atlantic shows that the applied distribution effectively describes statistical properties of the ocean-atmosphere turbulent fluxes of heat and moisture. An optimal algorithm for the quantitative determination of the distribution parameters is derived along with the methodology for estimation of the confidence limits for the parameters. The distribution is used for the computation of statistical parameters of turbulent heat fluxes in the Gulfstream area of the North Atlantic. It is shown that the application leads to the improvement of the accuracy of climatological flux estimates and allows for estimation of extreme fluxes crucially important for quantification of the role of sea-air interaction in the ocean and atmosphere dynamics.

Key words: double-exponential distribution, maximum likelihood estimations, air-sea interaction

1 Introduction: motivation and formulation of the problem

Turbulent fluxes of sensitive (Q_h) and latent (Q_e) heat are responsible for the adiabatic heating of the lower atmosphere and, thus, for the generation of atmospheric cyclones [2-3]. These fluxes also force the ocean circulation and form the boundary conditions for the ocean general circulation models [1, 6] Knowledge of the fluxes is also very important for the adequate calculation of the parameters forming climate change forcing. Therefore, accurate estimation of climatological, monthly and seasonal fields of air-sea turbulent heat fluxes is one of today's burning issues of marine climatology. Of a special importance is the analysis of statistical characteristics of surface turbulent flux distributions and estimation of extreme air-sea flux values. Observations and diagnostic studies [7] particularly show that these are strongly localised in space and in time extreme fluxes of heat and evaporation what is responsible for the generation and explosive development of atmospheric cyclones over the ocean. Observations also show that deep convection in the Labrador Sea, steering the ocean meridional circulation is provided by locally very high surface turbulent fluxes, occurring on space scales of tens of kilometres and time scales of tens of hours.

Currently, numerical weather prediction (NWP) models in a data assimilation mode provide high resolution (6-hourly in time and 50 to 200 km in space) global fields of atmospheric variables, including surface fluxes for the last several decades, the so-called reanalyses. Although widely used, these flux products are only loosely connected to nature, being largely influenced by the NWP model used. Moreover, reanalyses due to relatively coarse resolution do not resolve effectively the extreme turbulent fluxes.

Global turbulent flux fields can be also derived from the observations of merchant ships in the Global Ocean. In these products surface turbulent fluxes are computed from the observed sea surface and air temperatures, humidity, wind speed and atmospheric pressure using the so-called bulk aerodynamic algorithms. Although the achievements of the last decades provide quite accurate (about 5%) estimations of the turbulent flux from individual observation [2], this does not yet guarantee reliable representation of statistical distribution and accurate estimation of climatological flux fields from individual estimates for the reason of high sampling uncertainties inherent in the flux products based on ship observations.

As a result, the statistical properties of the distribution of surface turbulent fluxes, especially in poorly sampled areas are still unknown. Moreover, a proper estimation of the impact of sampling on these characteristics can not be performed without explicit knowledge of the type of statistical distribution and its parametric estimates. Distributions of the other meteorological variables are well fitted by the known probability density functions (PDF), such as Gamma PDF for precipitation or Weibull PDF for the wind speed [10]. However, for surface turbulent fluxes such a distribution has not yet been found.

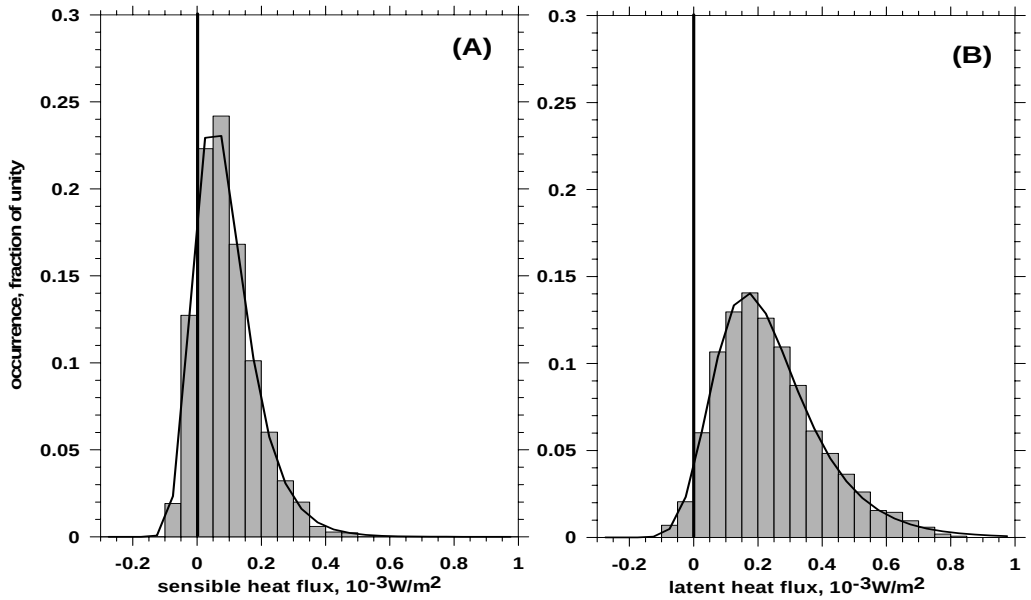

Figure 1. Occurrence histograms (grey bars) of the sensible heat flux (A) and latent heat flux (B) in the Gulfstream region for the winter season of 1988 and the double exponential PDFs fitted to the empirical data (bold lines).

Figure 1 shows empirical occurrence histograms of the surface turbulent fluxes for the winter season (JFM) of 1996 in the Gulf Stream region of the North Atlantic. These graphs were derived from 6-hourly high resolution surface fluxes diagnosed by the NWP model of European Centre for Medium Range Weather Forecasts (ECMWF) in a data assimilation mode. Despite being influenced by the model performance, model reanalysis is capable to simulate reliable synoptic variability of surface turbulent fluxes [5]. For both sensible and latent heat fluxes the occurrence histograms imply more heavy tails of the distributions in the range of high positive values, where positive fluxes are directed form the ocean to the atmosphere, and thus, strong asymmetry of the PDF. In different regions air-sea flux data can exhibit different skewness and kurtosis. The modal values and long-term means are typically positive, although in some areas they can be negative. The occurrence histograms in Figure 1 imply that the distribution of surface turbulent fluxes likely can be modelled by the family of the double exponential distributions (2ePDF).

The family of double exponential distributions has the density distribution function:

$$f_{\mu,\theta}(y) = a(\mu,\theta)\exp(-\psi(\mu)y)dG_n(y) \quad (1)$$

with the parameters θ, μ, n and arbitrary functions α, ψ, G. The constant $a(\mu,\theta,n)$ is chosen to make $\int_{-\infty}^{\infty} f(y)dG(y) = 1$. There are several studies of the properties of double exponential distributions and various aspects of its applications. In particular, this family has been used to generalize the exponential regression models [4-5] who applied the double exponential distribution to the two-way contingency tables. Some schemes for the estimation of the distribution parameters have been suggested in [9]. However, they proposed neither specific formulas for maximum likelihood estimation nor a numerical algorithm for the determination of parameters.

The aim of this study is to justify the application of the 2ePDF for the analysis of turbulent fluxes, to derive basic moments and estimates of statistical significance and to quantify the limits of the application of 2ePDF for this crucially important geophysical problem. The maximum likelihood estimators for the parameters will be derived and an effective algorithm solving the maximum likelihood equations will be proposed. The existence and the uniqueness of the solutions are proved. Additionally, the interval estimates are also derived as well as the confident limits for the distribution parameters. Finally, an example of the application of 2ePDF for the analysis of surface turbulent fluxes in the Atlantic is presented.

2 Mathematical model

In order to describe the probability distribution of the air-sea turbulent fluxes during continuous time (e.g. month or season) (Figure 1) the following distribution can be considered:

$$P(x) = -\alpha\beta\exp(\beta x)\exp(-\alpha\exp(\beta x)) \quad (2)$$

where $P(x)$ is a density function with variable x representing either Q_h or Q_e, α and β are the location and scale parameters respectively. Of these parameters α is assumed to be positive and β is negative. Obviously, the distribution $P(x)$ defined by (2) belongs to family (1). As mentioned above Weibull distribution is commonly used to describe the wind speed in a wide range of time scales. Wind speed along with temperature and humidity gradients in near-surface layer represents one of the key-parameters for estimation of surface turbulent fluxes using bulk formulae. Eq. (2) implies a similar type of distribution, however, it accounts for the well pronounced asymmetry with respect to zero flux, shown in Figure 1. One can note that the distribution (2) is obtained from a Weibull distribution by the replacement of variable $y = \ln x,\ x > 0$, leading to the equation $P(y) = -\alpha\beta y^{\beta} e^{-\alpha(y)^{\beta}}$, $y>0$, representing a Weibull distribution. Integration of (2) results in the following expressions for the mean and variance:

$$\bar{x} = \int_{-\infty}^{\infty} P(x)\cdot x dx = \frac{C + \ln\alpha}{-\beta},$$
$$\operatorname{var} x = \int_{-\infty}^{\infty} P(x)\cdot x^2 dx - \bar{x}^2 = \frac{\pi^2}{6\beta^2} \qquad (3)$$

where ***C*** is the Euler constant, appearing through the integration. If each member of sample $x_1, x_2, \ldots x_n$ is distributed according to (2), the likelihood function can be given as:

$$p(\vec{x},\alpha,\beta) = (-\alpha\beta)^n \exp(\beta\sum_{i=1}^{n} x_i)\cdot$$
$$\cdot \exp(-\alpha\sum_{i=1}^{n}\exp(\beta x_i)) \qquad (4)$$

where n is the number of observations, $\vec{x} = (x_1, \ldots x_n)$ denotes the vector of flux values and $p(\vec{x},\alpha,\beta)$ stands for the density of the joint distribution of the vector of flux values. In this case the natural logarithm of the likelihood function can be presented in the following form:

$$L(\bar{x},\alpha,\beta) = \log(p(\bar{x},\alpha,\beta)) =$$
$$= n\log(-\beta) + n\log(\alpha) + \qquad (5)$$
$$+ \beta\sum_{i=1}^{n} x_i - \alpha\sum_{i=1}^{n}\exp(\beta x_i)$$

and the maximum likelihood estimators should satisfy to the equations:

$$\frac{\partial L(\bar{x},\alpha,\beta)}{\partial\alpha} = 0, \qquad \frac{\partial L(\bar{x},\alpha,\beta)}{\partial\beta} = 0 \quad (6).$$

Equations (6) lead to the following system of equations:

$$\frac{n}{\alpha} = \sum_{i=1}^{n}\exp(\beta x_i)$$
$$\frac{n}{\beta} + \sum_{i=1}^{n} x_i = \alpha\sum_{i=1}^{n} x_i \exp(\beta x_i) \qquad (7, 8)$$

whose solution should further provide with the parameter estimators. From (7) and (8) the following statement is valid:

Theorem 1. The solution of the equations (7)-(8) provides the maximum likelihood estimation, i.e. the maximum of likelihood function (4). If all observed values are not equal in the vector $\vec{x} = (x_1, \ldots x_n)$, the solution of (7) - (8) will contain the root $\alpha_* > 0, \beta_* < 0$ and this root will be unique on the intervals $0 < \alpha < \infty$, $-\infty < \beta < 0$.

Proof. Actually, one needs to prove three statements. In order to prove the statement that the solution of equations (7)-(8) really provides the maximum of function (4), it is sufficient to show, that the matrix of the second derivatives

$$I = (i_{11,} i_{12}, i_{21}, i_{22})$$

where

$$i_{11} = \frac{\partial^2 L}{\partial\alpha^2}, i_{22} = \frac{\partial^2 L}{\partial\beta^2}, i_{12} = i_{21} = \frac{\partial^2 L}{\partial\alpha\partial\beta}, \qquad (9)$$

is negatively defined. Denoting for brevity $L(\bar{x},\alpha,\beta)$ through *L*, direct calculation gives

$$i_{11} = -\frac{n}{\alpha^2}, \quad i_{22} = -\frac{n}{\beta^2} - \alpha\sum_{j=1}^{n} x_j^2 e^{\beta x_j},$$
$$i_{12} = -\sum_{j=1}^{n} x_j e^{\beta x_j} \qquad . 10)$$

The negative (positive) definition of the matrix *I* is equivalent to the fact that the set of matrix eigenvalues is negative (positive). The latter results in the two statements:

(*a*) $SP(I) < 0$ ($SP(I) > 0$),
(*b*) $\det(I) > 0$,

where $SP(I)$ is the trace of matrix (sum of its diagonal elements) and det*(I)* is the matrix determinant. Indeed, *SP(I)* is the sum of eigenvalues and $\det(I)$ is their product. If conditions (*a*) and (*b*) hold, both eigenvalues are negative (positive). The condition (*a*) is obvious

(because α is positive) and the condition (*b*) is the consequence of the inequality

$$\sum_{j=1}^{n} x_j^2 e^{\beta x_j} \geq (\sum_{j=1}^{n} x_j e^{\beta x_j})^2 , \qquad (11)$$

representing a partial case of the Jensen's inequality. Thus, the first part of the statement is proved. In order to prove the existence and the uniqueness of the negative (positive) solution of (7)-(8), these equations can be rewritten in the following equivalent form:

$$\frac{1}{\beta} + n^{-1}\sum_{i=1}^{n} x_i = \frac{\sum_{i=1}^{n} x_i e^{\beta x_i}}{\sum_{i=1}^{n} e^{\beta x_i}} \qquad (12)$$

$$\alpha = \frac{n}{\sum_{i=1}^{n} e^{\beta x_i}} \qquad (13)$$

Clearly, Eq. (13) independently on β-parameter, requires α in (13) to be always positive and unique if the β is unique**.** Thus, one can further focus on the β parameter only. According to the *theorem 1* conditions, there are several different values in the observational sample. Without the loss of generality, let's assume the first *k* observations $x_1,...x_k$ ($k \leq n$) to be negative and x_k to be a maximum negative value, i.e. for any $i < k$, $|x_i| \leq |x_k|$. When observations $x_1,...x_n$ are fixed, both left hand and right hand sides of the relation (12) represent continuous functions of β. The left hand side varies from $S = n^{-1}\sum_{i=1}^{n} x_i$ to $-\infty$ with β going from $-\infty$ to zero, remaining negative. The right hand side of (12) can be represented as

$$\frac{1}{\sum_{i=1}^{n} e^{\beta x_i}} \sum_{i=1}^{n} x_i e^{\beta x_i} = \frac{1}{\sum_{i=1}^{k} e^{\beta x_i} + \sum_{i=k+1}^{n} e^{\beta x_i}} \cdot \cdot (\sum_{i=1}^{k} x_i e^{\beta x_i} + \sum_{i=k+1}^{n} x_i e^{\beta x_i}) \qquad (14)$$

In (14) the terms with indices greater then *k* vanish when β goes to -∞. For the terms with the indices less or equal *k* a simple mathematical transform gives:

$$\frac{1}{\sum_{i=1}^{k} e^{\beta x_i}} \sum_{i=1}^{k} x_i e^{\beta x_i} = \frac{e^{\beta x_k}(\sum_{i=1}^{k-1} x_i e^{\beta(x_i - x_k)} + x_k)}{e^{\beta x_k}(\sum_{i=1}^{k-1} e^{\beta(x_i - x_k)} + 1)} \qquad (15)$$

The last expression converges to x_k and since x_k is the minimum of sample, the inequality $S = n^{-1}\sum_{i=1}^{n} x_i \geq x_k$ holds. When β equals 0, the right hand side of (15) equals *S*. This means that the variability of the left hand side of the equality (12) overlaps the range of the variability of the right hand side and, hence, due to continuity, the root of (13) exists. Its uniqueness follows from the monotonic variability of the left hand side of the relation (13). Thus, the statement is proved. Formulas (12) and (13) define the pointed maximum likelihood estimators α_*, β_*. It is possible to obtain their asymptotic joint distribution when $n \to \infty$, *n* being the sample length.

Corollary. Function

$$F(\beta) = \frac{n}{\beta} + \sum_{i=1}^{n} x_i - \alpha\sum_{i=1}^{n} x_i \exp(\beta x_i) \qquad (16)$$

monotonically decreases when β varies from $-\infty$ *to 0* and α is fixed.

Theorem 2. The vector $\sqrt{n}(\alpha - \alpha_*, \beta - \beta_*)$ is distributed asymptotically as the two-dimensional Gaussian vector with zero mean and covariance matrix $I^{-1}(\alpha_*, \beta_*)$, where I is the informational matrix defined above and $I^{-1}(\alpha_*, \beta_*)$ is its inverted matrix when $(\alpha = \alpha_*, \beta = \beta_*)$.

Proof. The proof of this statement follows from the general statement [8] for regular distributions, and the fact, that the Weibull distribution is regular.

Corollary. The confident intervals for parameters with 95% significant level will be given as an internal domain of the ellipse

$$\frac{(\alpha - \alpha_*)^2}{\sigma_1^2} - \frac{2\rho}{\sigma_1\sigma_2}(\alpha - \alpha_*)(\beta - \beta_*) + \frac{(\beta - \beta_*)^2}{\sigma_2^2} = 6(1 - \rho^2) \qquad (17)$$

where σ_1^2, σ_2^2 are diagonal elements of I^{-1}, $\rho = \frac{i_{12}}{\sigma_1 \sigma_2}$ and i_{12} are non-diagonal elements of the matrix I^{-1}, taken at points $(\alpha = \alpha_*, \beta = \beta_*)$. This ellipse can be rewritten as

$$\begin{aligned} \alpha &= R\sigma_1(\sin\varphi + \cos\varphi) + \alpha_* \\ \beta &= R\sigma_2(\cos\varphi - \sin\varphi) + \beta_* \end{aligned} \qquad (18)$$

where $R = \sqrt{3(1+\rho)}$ and φ varies from 0 to 2π. Now the optimal algorithm of the computation of the roots of (7) and (8) is derived.

Theorem 3. Let α_0, β_0 be an initial iteration, lying within the interval $c_1 < \alpha_0 < c_2, c_1 < \beta_0 < 0$ where c_1, c_2 are some constants, lying within the interval $x_{\min} < c_1 < c_2 < x_{\max}$ and $x_{\min}$, $x_{\max}$ are minimum and maximum values of the sample respectively. Then the algorithm, realising the iterative scheme

$$\frac{n}{\alpha_{l+1}} = \sum_{i=1}^{n} \exp(\beta_{l+\frac{1}{2}} x_i) \qquad (19)$$

$$\frac{n}{\beta_{l+\frac{1}{2}}} + \sum_{i=1}^{n} x_i = \alpha_l \sum_{i=1}^{n} x_i \exp(\beta_{l+\frac{1}{2}} x_i) \qquad (20)$$

will converge to the equilibrium point (α_*, β_*), which is the unique solution of (7) and (8).

Proof. This proof follows from the corollary of theorem 1 and the fact that function $F(\beta)$ (16) is bounded.

3 Application of the double-exponential distribution to the computation of the averaged turbulent fluxes

We will now apply the derived double-exponential distribution for the evaluation of statistical parameters of surface fluxes in the Gulf Stream area characterized by the strongest mean values and the most intense variability of surface turbulent fluxes. Due to high intensity of synoptic variability poor sampling in this crucially important area may significantly affect the accuracy of climatological estimates of surface turbulent fluxes. Figure 1 shows approximations of empirical histograms of sensible and latent heat fluxes by the double-exponential distribution, presented above. This distribution fits well to the data and *k-s* test returns the probability of the distribution to be of the kind of (2) of higher than 96.7%.

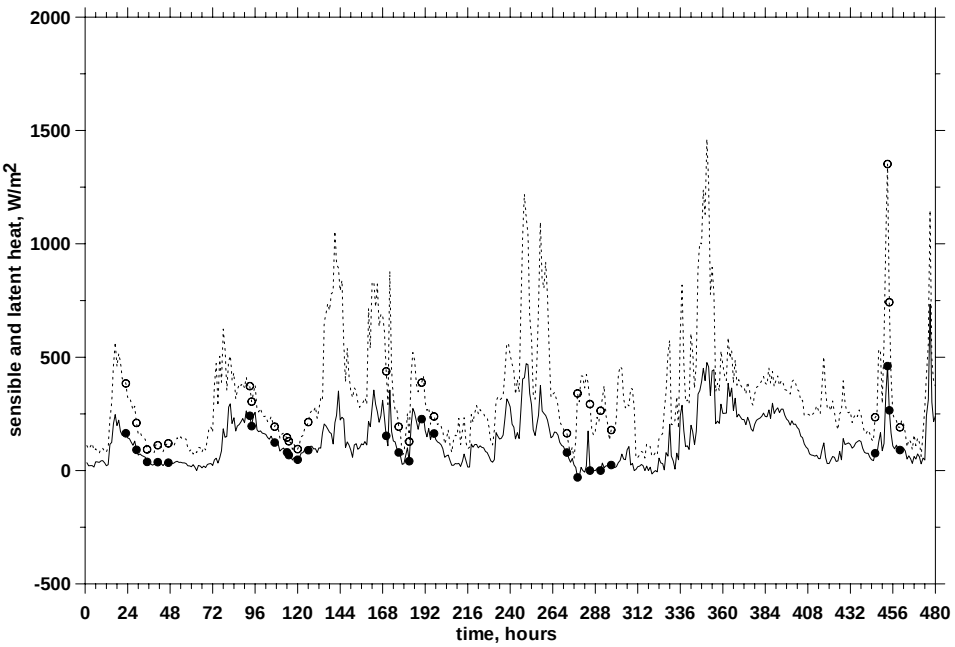

Figure 2. Estimates of sensible (solid line) and latent (dashed line) heat flux derived from 1-hourly observations during NEWFOUEX-88 experiment in March 1988 (484 observations) and sensible (closed circles) and latent (open circles) fluxes computed only for the moments when VOS observations were available (26 observations).

During the NEWFOUEX-88 experiment carried out in winter season of 1988 research vessels collected time series of hourly observations of surface meteorological parameters of exceptionally high quality. In Figure 2 we show estimates of the sensible and latent heat fluxes for the location 42°N, 44°W in Newfoundland basin for the period from 01.03.1988 to 21.03.1988, computed using the

COARE-3 algorithm [2] from the directly observed surface meteorological data. In the vicinity of this observational point for the same period contribution of the VOS data consisted of 26 reports sampled within the radius of 50 km from the research vessel location. To achieve comparability we subsampled the original time series at exactly the same UTC time instances as VOS reports were taken. The differences between the turbulent flux estimates derived from the VOS data and from the regular time series were within ±5 W/m^2. It is obvious, that VOS sampling is not adequate to reflect the actual variability of surface turbulent fluxes in the Gulfstream area.

Table 1. Example of estimation of turbulent heat fluxes in the Gulfstream area for the period 01.03-21.03.1988 using direct averaging and the double exponential distribution.

	Sensible heat		Latent heat	
	26 samples	484 samples	26 samples	484 samples
Mean (raw averaging), W/m^2	98	131	264	331
Std (raw averaging), W/m^2	86	102	156	234
Mean (2ePDF), W/m^2	112	129	292	322
Std (2ePDF), W/m^2	92	99	171	195
Location parameter	2.731	2.978	5.190	4.642
Scale parameter, W/m^2	-14.16	-12.91	-7.63	-6.55
95%-percentile, W/m^2	283	315	608	688
99%-percentile, W/m^2	401	441	824	937

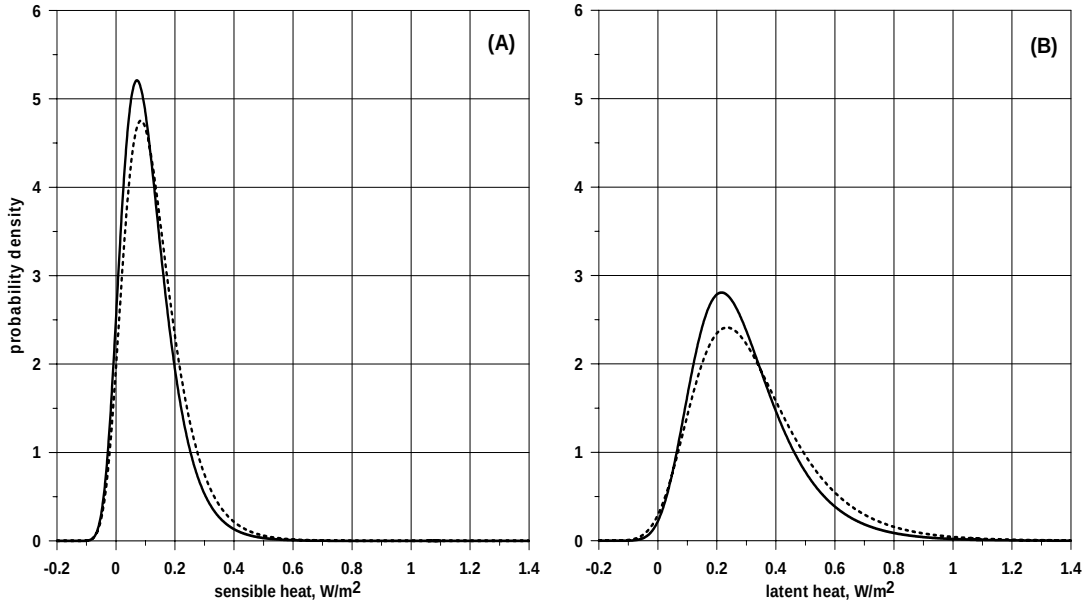

Figure 3. Double-exponential PDFs fitted to the estimates of sensible (A) and latent (B) heat fluxes in the Gulfstream region for the period 01-21 March 1988 using high resolution regularly sampled data (dashed line) and undersampled data (solid line).

The results of estimation of statistical characteristics of turbulent fluxes are given in Table 1. Direct averaging of the flux estimates for the 20-day period gives sensible and latent heat flux values, which are respectively 25% and 20% smaller than those derived from the high resolution data. Application of the double exponential distribution to the undersampled time series returns the mean values which are respectively 13% and 9% smaller than those derived using 2ePDF from the large sample. Of a special importance is that application of the double exponential distribution allows for the estimation of 95% and 99% percentiles of turbulent fluxes. These estimates are crucially important for the quantification of the role of extreme fluxes in the ocean and atmospheric dynamics because these are extreme fluxes what determines the generation of the most intense cyclones in the atmosphere and the most dramatic convection events in the ocean. These percentiles expose differences of about 10%, when estimated from the undersampled data compared to the regularly sampled data. However, it is important that the double exponential distribution allows for the estimation of extreme fluxes, while the raw data do not.

Figure 3 shows PDF of the sensible and latent heat fluxes, derived from the regularly sampled and undersampled data. These PDF were computed according to Eq. (2). Clearly, the PDF for the regularly sampled data imply higher mean and extreme fluxes compared to the undersampled time series. Figure 4 shows the ellipses implying the confident limits for the parameters estimated for sensible and latent heat fluxes from the regularly sampled and undersampled time series in the α, β – plane. Remarkably, the ellipses derived for the data with small sample.

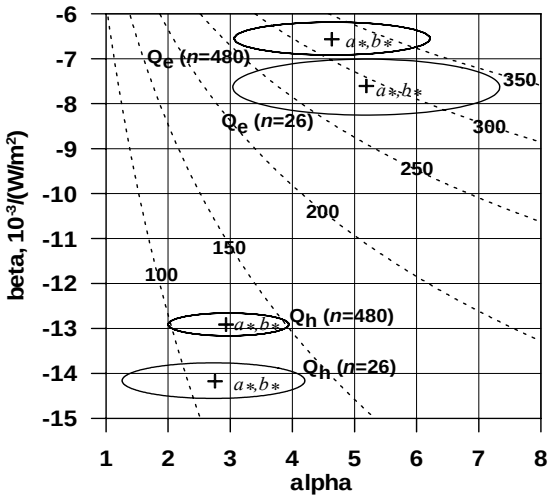

Figure 4. Ellipses, implying confident limits for the sensible and latent heat flux estimates for the period 01-21 March 1988 in the Gulfstream area. The graph is overploted with the flux values for given α and β (dashed lines).

Further application of the double-exponential distribution to the global ocean may help to derive the next generation monthly and seasonal of surface turbulent fluxes climatology. In contrast to the routinely averaged surface fluxes, such a climatology will minimize sampling errors, which are quite large in poorly sampled areas (e.g. subpolar latitudes of the Northern Hemisphere and Southern Ocean), where they are significantly higher than the other uncertainties inherent in flux computations [7]. This will provide more accurate estimates of surface turbulent fluxes and will considerably improve our understanding of their climate variability. Moreover, this climatology will provide a wide spectrum of the surface turbulent flux statistics which cannot be derived form the raw data. These statistics, first of all extreme fluxes will stay as imply considerably larger confident limits for α, and β in comparison to the confident intervals for the regularly sampled time series.

Acknowledgements. This study is supported by Russian Foundation for Basic Research (grant 05-65-439) and by Ministry of Education and Science of Russian Federation. We thank Jeff Kepert and Eric Schultz of BMRC (Melbourne) for helpful comments on the manuscript.

References

1. Barnier B. Forcing the ocean. *1998 Modeling and Parameterization*, *E.P.Chassignet and J.Verron,(eds). Kluwer Academic Publishers, The Netherlands, pp. 45-80.*

2. Brunke, M.A., C.W. Fairall, X. Zeng, L.Eymard,J. A.Curry. Which bulk aerodynamic algorithm are least problematic in computing ocean surface turbulent fluxes? *2003, J. of Climate,. 16, 619-635*

3. Curry, R.G., M.S. McCartney, and T.M. Joyce. Oceanic transport of subpolar climate signals in mid depth subtropical waters. *1998 Nature, 391, 575-577.*

4. Diaconis, P. and Efron, B. Testing for independence in a two-way table: new interpretations of the chi-square statistic (with discussion)**.** *1985, The Annals of Statistics*, *13, 845-874.*

5. Efron, B. Double exponential families and their use in generalized linear regression. *1986, J. of American Statistic Association, 81, 709-721.*

6. Gulev S. K., B. Barnier, Knochel H., J.-M. Molines, and M. Cottet. Water mass transformation in the North Atlantic and its impact on the meridional circulation: insights from on ocean model forced by NCEP/NCAR reanalysis surface fluxes. *2003 J. of Climate, 16, 3085-3110.*

7. Gulev, S.K., T. Jung, and E. Ruprecht Estimation of the impact of sampling errors in the VOS observations on air-sea fluxes. Part I. Uncertainties in climate means. *2006 J. of Climate, 19 (in print)*

8. Leman E. Verification of statistical hypotheses. *1982, Moscow, "Nauka", pp. 352 (in Russian).*

9. Pregibon, D. Review of generalized linear models by McCullagh and Nelder.

1984, The Annals of Statistics, 12, 1589-1596.

10. Wilks, D.S. Statistical methods in atmospheric science. *1995, Academic Press, London, pp. 467.*