EFDA–JET–PR(13)13

A. Murari, F. Pisano, J. Vega, B. Cannas, A Fanni, S. Gonzalez, M. Gelfusa,
M. Grosso and JET EFDA contributors

A New Methodology for the Statistical Analysis of Plasma Instabilities with Application to ELMs on JET with a Carbon Wall

A New Methodology for the Statistical Analysis of Plasma Instabilities with Application to ELMs on JET with a Carbon Wall

A. Murari[1], F. Pisano[2], J. Vega[3], B. Cannas[2], A Fanni[2], S. Gonzalez[3],
M. Gelfusa[4], M. Grosso[5] and JET EFDA contributors*

JET-EFDA, Culham Science Centre, OX14 3DB, Abingdon, UK

[1]Consorzio RFX-Associazione EURATOM ENEA per la Fusione, I-35127 Padova, Italy
[2]Electrical and Electronic Engineering Department - University of Cagliari, Italy
[3]Asociación EURATOM/CIEMAT para Fusión. Avda. Complutense, 22. 28040 Madrid, Spain
[4]Associazione EURATOM-ENEA - University of Rome "Tor Vergata", Roma, Italy
[5]Department of Mechanical, Chemical and Material Engineering, University of Cagliari, Italy
** See annex of F. Romanelli et al, "Overview of JET Results",*
(24th IAEA Fusion Energy Conference, San Diego, USA (2012)).

Preprint of Paper to be submitted for publication in
Nuclear Fusion

ABSTRACT

Edge Localised Modes (ELMs) are bursts of instabilities which deteriorate the confinement of H mode plasmas and can cause damage to the divertor of next generation of devices. On JET individual discharges can exhibit hundreds of ELMs but typically in the literature, mainly due to the lack of automatic analysis tools, single papers investigate only the behaviour of tens of individual ELMs. In this paper, an original tool, the Universal Event Locator (UMEL), is applied to the problem of automatically indentifying the time location of ELMs. With this approach, databases of hundreds of thousands of ELMs can be built with reasonable effort. The analysis has then been focussed on the investigation of the statistical distribution of the inter ELM intervals at steady state for type-I ELMs. Numerous probability distributions have been tested to address the data analysis and different distributions provided a best fit for sets of data from different experiments. This result constitutes robust experimental evidence that Type-I ELMs are not all the same type of instability. Moreover, the most likely distributions are not memoryless meaning that the waiting time, from a particular instant until the next ELM, does depend on the time elapsed from the previous event. These properties pose important new constraints on the models aimed at describing the ELM dynamics. This work also demonstrates the widespread applicability of the UMEL tool.

1. THE RELEVANCE OF INSTABILITIES AND THEIR STATISTICAL BEHAVIOUR

The operational space accessible to a Tokamak is highly restricted by a large set of macroscopic instabilities, which can affect the confinement and sometimes even cause the unplanned extinction of the plasma [1]. Some of these instabilities have a periodic or quasi-periodic pattern which needs to be interpreted in order to understand the dynamical behaviour of the instabilities. Any satisfactory dynamical model of a macroscopic instability must be able to reproduce its time and spatial evolution. It is therefore crucial to determine the statistical properties of the instabilities, such as their period or the probability density functions of the intervals between successive occurrences, in the case of non periodic events.

During a Tokamak discharge, hundreds of these macroscopic instabilities can take place. Typically the identification and time location of these events is achieved by means of visual analysis of plasma signals (normally waveforms of amplitude versus time). Given the fact that on JET tens of gigabytes of raw data can be collected in each shot, the proper analysis of all this data is very heavy in terms of human resources (if not absolutely prohibitive). Experts have therefore to devote an enormous amount of time examining in great detail each waveform in order to determine the events (e.g. ELMs, sawteeth, disruptions etc.) and their temporal locations. In future devices, with pulse lengths at least 100 times longer than in JET, the manual analysis of signals for detection of events will become impossible and this task will have to be performed automatically. Recently, data mining and artificial intelligence tools have been applied to fusion databases. These methods provide fast and accurate results in locating and identifying plasma phenomena [2,3]. Moreover, they can predict dangerous plasma phenomena (such as disruptions) before they occur [4,5,6].

These automatic tools present several advantages with respect to the manual analysis of data. An important issue to take into consideration is the cost of the analysis. The typical visual analysis of the waveforms requires a lot of manpower and hence, it is expensive. Computer codes are much cheaper. Their price includes their design and maintenance / update but then, they can be run many times without additional costs. They therefore allow the analysis of large quantities of data on a low-cost basis. Another advantage of data mining techniques is their speed. Computer codes can be faster than any visual analysis carried out by experts. Visual analysis is time-intensive and the only way to speed up the process is increasing the number of experts and thus, increasing the costs. The deterministic behaviour of computer codes is also one of their benefits. The results of an analysis using a computer code are always the same, no matter the number of times that the program is executed. It is important to note that it does not imply that the codes are error-free; it does just mean that the same errors (if they exist) are made. This cannot be guaranteed by the analysis carried out by experts. For example the same event can be located by two experts at two slightly different times and even the same expert can determine two different times for the same event. The reasons of these differences can be, among others, a different detail level in the analysis (the higher the level of detail, the longer the analysis time) or just a mistake.

Once a sufficiently representative and reliable database has been built, adequate tools have to be deployed to determine the statistical properties of the events of interest. One of the typical tasks consists of determining the appropriate probability density function (pdf) for fitting the properties of the events. In [7] for example, starting from simple experimentally motivated assumptions, the authors derived a Weibull pdf for the waiting times between ELMs. To test this hypothesis they considered 84 datasets with a steady period of type I or type III ELMs that lasted for at least 3 seconds, finding that Weibull and Gaussian pdfs provided a similarly good fit to type I ELM data, but a clearly better fit to the type III ELM data.

The distribution parameters that make a distribution type best fit the available data can be determined in several ways. The most common technique to obtain the parameter values is known as maximum likelihood estimation (MLE). Once a model is specified with its parameters, goodness-of-fit allows evaluating how well the data are modeled by that distribution. In the present paper various tests, such as the Kolmogorov-Smirnov [8], the Anderson-Darling [9], the Cramer Von Mises [8] and proper model selection tools, such as the Akaike Information Criterion and Bayesian Information Criterion [10,11] have been applied. In order to assess the effects of the macroscopic plasma parameters on the dynamical behaviour of the instabilities, the events must be properly grouped. The technique typically used to decide whether the different samples belong to the same population is the analysis of variance (ANOVA) [13]. Unfortunately these techniques assume that all samples belong to a population having a normal distribution. In general this hypothesis cannot be met and therefore it is necessary to utilise more advanced methods such the Kruskal-Wallis test [14].

In this paper, a coherent methodology to assess the general statistical properties of quasi periodic instabilities is described in detail. As mentioned, it consists of the deployment of data mining tools, mainly the Universal Multi Event Locator (UMEL), for the automatic identification of events [15]. Then a series of statistically sound techniques is utilised to determine the statistical properties of the instability. To exemplify the potential of these techniques, they are applied to one of the most important macroscopic instabilities in Tokamaks, Edge Localised Modes or ELMs [16], which affect H-mode plasmas. The main results obtained with the new techniques indicate that the dynamics of the Type I ELMs is more involved than originally thought and that probably they comprise more than one simple type of instability.

When Tokamak plasmas reach the H-mode, the plasma confinement improves. The energy confinement increases typically by a factor of 2. This is due to a thin region of increased gradients at the edge known as the Edge Transport Barrier (ETB). Originally discovered on ASDEX [17], the H mode was reproduced by most of the other major international Tokamaks in the following years. The H-mode was reached in other tokamaks devices e.g, PDX in 1984 [18], DIII-D in 1986 [19], JET in 1987 [20] and even stellarator devices e.g. W 7-AS in 1993 [21], demonstrating that the H-mode is a generic feature of fusion by magnetic confinement [22]. The H mode is now a routine regime of operation of all major Tokamaks.

During the operation of ASDEX in H-mode in 1982, bursts in the H_α signal were detected [17]. These bursts or spikes were associated with MHD instabilities at the edge of the plasma and, in 1984, they were named Edge Localised Modes (ELMs) [23]. ELMs cause a reduction in density and temperature in the outer zone of the plasma (edge), resulting in a deterioration of the plasma confinement through the reduction of the ETB.

Originally, three different types of ELMs were identified in the DIII D tokamak in 1991 [22]:

- Type I, giant ELMs: the Type I ELMs are instabilities typically showing a ballooning behaviour and whose repetition frequency increases with power and drops with increasing current. They appear as large isolated sharp bursts on the emissivity signal (H_α or D_α). An example of Type I ELMs in JET is reported in Figure 1a. Type I ELMs are the most dangerous ones, given the large heat losses involved and the consequent unacceptable high heat load on the divertor.
- Type II, grassy ELMs: they are believed to appear when the plasma edge is in the connection regime between the first and the second stable ballooning regimes. They are irregular and low-amplitude ELMs (Figure 1b).
- Type III: they are medium amplitude ELMs whose repetition frequency decreases as the power is increased. The plasma edge pressure gradient is below the ideal ballooning limit. Figure 1c provides an example of Type III ELMs in JET.

Experiments have shown that it is possible to obtain ELM-free H mode phases. Unfortunately, the ELM-free H-mode is typically non stationary and ELMs are therefore an unavoidable aspect of this regime of operation. In the rest of the paper, the analysis will focus on Type I ELMs, since these are the most common in the present regimes of JET operation. They are also potentially the most dangerous and therefore the most urgent to understand and control.

With regard to the structure of the paper, the next section describes the basics of the data mining tools and statistical techniques utilised in the rest of the paper. Section 3 describes the results obtained in the localization of ELMs using UMEL. In section 4 the analysis is focussed on the steady state phases of the discharges to determine the pdf of the inter ELM intervals. The main implications of the identified pdf for the dynamics of the ELMs are discussed in section 5. The ELMs are then divided in coherent groups in section 6. The final conclusions of the statistical analysis with the implications for the formulation of dynamical models of the ELMs are discussed in the last section 7 of the paper.

2. INVESTIGATION OF ELMS WITH ADVANCED MACHINE LEARNING AND STATISTICAL TOOLS

In this section the mathematical background, on the data mining and statistical tools used in the rest of the paper, is provided. In subsection 2.1, Support Vector Regression and its use by UMEL is introduced. Section 2.2 describes the goodness-of-fit and model selection tools deployed to extract the most appropriate pdf to interpret the experimental data. Section 2.3 illustrates the method adopted to cluster the various examples in their proper groups.

2.1. Introduction to machine learning for identification of events: UMEL

This subsection describes UMEL, a technique to locate events in plasma waveforms and films. UMEL is a universal technique because it is independent of the type of the pattern sought (peaks, drops or slope changes) and the type of waveforms analysed (time domain or frequency domain). UMEL is based on Support Vector Regression (SVR) [24], a version of SVM [25] for function estimation. SVR fits the training data without depending on factors such as sampling rate or noise distribution. This technique computes a fitting function and, in addition, it retrieves a list of the points from the training set that become Support Vectors (SVs).

SVR uses the e-insensitive loss function, also called e-tube:

$$|\xi|_e = \begin{cases} 0 & \text{if } |\xi| \le e \\ |\xi| - e & \text{otherwise} \end{cases} \tag{1}$$

The goal of SVR is to find the flattest function that fits the training data within the e-tube. The errors lower than *e* are not taken into consideration (the value of the e-insensitive loss function is 0 in the region [ξ [–e;+e]) but the errors higher than *e* are minimised. It is therefore possible to define the e-tube in such a way that the normal variations in the signals, including noise, remain within it and the specific events to be detected fall outside this interval.

A representative example is fitting to data sampling the Mexican hat. The Mexican hat is the well-known function:

$$\psi(t) = \frac{2}{\sqrt{3\sigma}\pi^{\frac{1}{4}}}\left(1 - \frac{t^2}{\sigma^2}\right)e^{-\frac{t^2}{2\sigma^2}} \tag{2}$$

In this example, a data set of 30 points from the Mexican hat has been randomly chosen to illustrate the use of SVR. A white Gaussian noise has been added to the samples. Four different kernels have been tested: linear, polynomial, Radial Basis Function (RBF) and Gaussian. Figure 2 depicts the results obtained with the various kernels.

The regression function obtained by the linear kernel is shown in Figure 2a. The blue line represents the Mexican hat given by Eq. (2) with $\sigma = 2$. The blue crosses indicate the data points. The green solid line is the regression function and the green dashed lines are the bounds of the e-tube. Since it uses a linear kernel, the regression function computed by SVR is a straight line. The fit obtained with the polynomial kernel of degree 2 is depicted in Figure 2b. Figures 2c and 2d show the results of the RBF and Gaussian kernels respectively.

UMEL can be used as an exact locator of singular points within signals. To achieve this, UMEL gives a novel interpretation of the SVs. In SVM and SVR, the complexity of the model determines the number of SVs (the higher the complexity, the larger the number of SVs). The regression of complex data sets requires large numbers of SVs. In contrast, simple data sets require smaller numbers of SVs. But the number of SVs does not depend only on the complexity of the data set to regress. It also depends on the smoothness of the regression function. Smoother functions require fewer SVs than crispy functions. Using UMEL, not all the SVs have the same degree of relevance. The SVs that lie on or outside the e-tube are called External Support Vectors (ESVs). In contrast, the SVs within the e-tube are called Internal Support. They are defined by the relations:

$$\begin{array}{lll} ESV \subseteq SV & \forall i \in ESV, & \left|y_i - f(x_i)\right| \geq e \\ ISV \subseteq SV & \forall i \in ISV, & \left|y_i - f(x_i)\right| < e \end{array} \qquad (3)$$

ISVs are necessary samples for the regression estimation, but they do not provide the same degree of relevance that can be assigned to ESVs. UMEL is based a novel interpretation of ESVs: the SVs that become ESVs are the most difficult samples to regress (they cannot be fitted inside the e-tube) and these SVs provide essential information in the regression process. ESVs reveal the occurrence of special patterns inside a signal: peaks, high gradients or segments with different morphological structure in relation to the bulk of the signal.

Figure 3 shows two examples of UMEL using a step function and a sinusoidal function. The green dashed lines delimit the e-tube. Then, the SVs within these lines are ISVs (cyan squares) and the SVs outside the e-tube are ESVs (red circles). The ESVs are clearly the most difficult samples to regress. In the case of the step function (Figure 3a), the samples around the step become ESVs. The ISVs are found inside the e-tube. In the case of the sinusoidal function (Figure 3b), the ESVs appear at the beginning of the function and at the extremal points corresponding to the maximum and the minimum of the function.

Most plasma phenomena are characterised by high frequency components in the time domain (spikes, drops, rapid slope changes, etc.). For example, an ELM is recognised as a spike in the D_α signal accompanied by a drop in the diamagnetic energy as well as a drop of the plasma density at the plasma edge. A disruption is identified by a fast drop in the plasma current at the same time that a plasma loop voltage peaks [15]. Therefore, it is possible to apply UMEL to locate these events. All the mathematical details about UMEL can be found in [15]. More details about the application to ELM detection are given in section 3.

2.2 Criteria to assess the quality of pdf fit

In this paper, the maximum likelihood estimation (MLE) method is used to estimate the parameters of theoretical models directly on the basis of the available data. According to this approach, from the theoretical probability density function *f* (pdf), analytically known, it is possible to estimate the vector of distribution parameters θ according to the sample observations.

The maximum likelihood method consists of estimating θ so that the likelihood function

$$L\left(x_1, x_2, \ldots, x_n, \theta\right) = \prod_{i=1}^{n} f\left(x_i, \theta\right) \tag{4}$$

or equivalently, its logarithm, is maximized.

Then, after the model parameters are estimated, model validation and model selection, the assessment of theoretical models quality to interpret the observed data, is carried out using different statistical tests. In particular, the Kolmogorov–Smirnov (K–S) test, the Cramer-Von Mises test (C-VM), the Anderson-Darling (A-D) test, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) have been applied. The main properties of these criteria, relevant to the discussion in the rest of the paper, are given in the following.

The Kolmogorov–Smirnov (K–S) test [8] is a goodness-of-fit test for any statistical distribution, based on the comparison between the empirical cumulative distribution function $F_n(x)$ of n experimental data and the theoretical one $F(x)$. The K–S test consists on finding the K-S statistic,

$$K = \sup_x \left|F_n\left(x\right) - F\left(x\right)\right|, \tag{5}$$

that is the greatest discrepancy between the empirical and theoretical distribution function, and comparing it against the critical K-S statistic for that sample size. If K is greater than a critical value depending of the significance level α, then the null hypothesis H_0: $F(x) = F_n(x)$ is rejected at the significance level α.

The Cramer Von Mises test [10] is an alternative to the K-S test. The C-VM test consists on finding the statistic

$$W^2 = \int_{-\infty}^{+\infty} \left(F_n\left(x\right) - F\left(x\right)\right)^2 dF\left(x\right). \tag{6}$$

It quantifies the difference between two cumulative distribution functions by comparing these two functions over their entire range. A problem with the Cramér-von Mises statistic is that the difference between the empirical distribution function and the reference cumulative distribution functions tends to zero when x → ±∞. Consequently, the value of W^2 is rather insensitive to the precise positions of the observations in the tails of the distribution.

A modification of the statistic consists in granting more importance to these observations by introducing a ponderation function into the definition of the statistic so as to make these observations more influential in the outcome of the test. The most widely used ponderation function is that of the Anderson-Darling test [9], $[F(x)(1-F(x))]^{-1}$, which is minimal around the median of the distribution, and tends to infinity when x → ±∞. In fact, it can be shown that $F(x)(1 - F(x))$ is the variance of the empirical distribution function in x, so that the test statistic is now the integral of the squared standardized difference between the empirical distribution function and the reference cumulative distribution function.

The A–S test statistic is

$$A^2 = \int \frac{\left(F_n(x) - F(x)\right)^2}{\left(F(x)\, 1 - F(x)\right)} dF\left(x\right). \tag{7}$$

Critical values of the Anderson-Darling test statistic depend on the specific distribution being tested. The effectiveness of these statistics is that they are distribution-free as long as F is continuous, that is the probability distribution of this statistic is free of F. When the shape and the parameters of the theoretical distribution function are estimated from the data, the distribution of the test statistics under the null hypothesis depends on the tested distribution, and the critical values, have to be recalculated [26]. An alternative is the parametric bootstrap, a data-based Montecarlo method [27] that has been mathematically shown to give valid estimate of goodness of fit probabilities. Bootstrap generates a new bunch of statistics under the null hypothesis that really represents a random sample from the tested distribution with parameters determined by the data. The p-value of the goodness of fit test is given by the percentage of bootstrap samples for which the calculated statistic is higher than the one evaluated from the original data. Thus, if the p-value is greater than the confidence level α, the null hypothesis H_0: $F(x) = F_n(x)$ cannot be rejected.

The Akaike Information Criterion is a way of selecting a model from a set of models. It is defined as

$$AIC = -2\ln L + 2k \tag{8}$$

where L is the value of the likelihood function evaluated with the MLE method and k is the number of parameters of the model. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Hence AIC not only rewards relative goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty penalises overfitting.

The Bayesian Information Criterion is a criterion alternative to AIC. It is defined as

$$BIC = -2\ln L + k\ln n \qquad (9)$$

Also for *BIC*, given a set of candidate models for the data, the preferred model is the one with the minimum *BIC* value. The fitted model favoured by *BIC* ideally corresponds to the candidate model which is “a posterior” more probable.

2.3 The Kruskal-Wallis criterion

The technique typically used to decide whether different groups of samples belong to the same population is the analysis of variance (ANOVA). Using this approach, the variation between different sample means is used to estimate the variation between individual observations, assuming that the variation among the means reflects only random sampling from a population, in which individuals vary with a normal distribution, and that the variance of the means of random samples of size n_i is σ^2/n_i, where σ^2 is the population variance. In our application, the hypothesis that all samples belong to a population having a normal distribution cannot be met, because time intervals between ELMs are positive quantities, which cannot be described by a distribution defined also for negative values. For these reasons, it has been necessary to use a different method, the Kruskal-Wallis test [14]. This test makes no assumption about the distribution of samples, but requires using ranked data instead of the original observations. This can be achieved by listing all the observations in order of magnitude and replacing the smallest by 1, the next-to-smallest by 2, and so on; if there are ties (two or more equal observations), each observation is replaced by the mean of the ranks to which it is tied. The test statistic to be computed is:

$$H = \frac{\frac{12}{n(n+1)}\sum_{i=1}^{C}\frac{R_i^2}{n_i} - 3(n+1)}{1-\sum_{j=1}^{J} T_j \Big/ \left(n^3-n\right)}, \qquad (10)$$

where C is the number of groups, R_i is the sum of the ranks in the ith group. The sum in the denominator is over all groups of ties, $T_j = t_j^3 - t_j$ for the j-th group of ties, t_j being the number of tied observations in the j-th group. If the samples come from identical continuous populations, for high n_i, H is approximately distributed as χ^2 (chi-squared distribution) with $C - 1$ degrees of freedom. If the samples come from identical continuous populations and the n_i are not too small, H is distributed as χ^2 distribution with $C - 1$ degrees of freedom and the critical H statistic is given by

$$H_{cr}^{\alpha} = F^{-1}\left(1-\alpha\right), \qquad (11)$$

where α is the significance level and $F^{-1}(x)$ is the inverse cumulative distribution function of the χ^2 distribution with $C - 1$ degrees of freedom. If $H > H_{cr}^{a}$, then the null hypothesis H_0, that the samples come from the same population, is rejected at the significance level α.

3. AUTOMATIC LOCALISATION OF ELMS WITH UMEL

Most ELMs taking place in JET experiments are not indexed and thus it is not easy to study them using a large statistical base. Average ELM studies contain dozens of ELMs, a greatly reduced set of the total number of ELMs in the JET database. For example, JET pulses can contain more than one hundred ELMs each and the JET database currently contains more than 80,000 pulses (the majority of which, even if not all, presents ELMs). Previous work has developed an automatic ELM classification system [15]. This tool utilises the ELM times as an input and classifies them as Type I or Type III. It has been tested using a small set of 256 ELMs (122 training and 143 test) manually located by experts.

In order to apply this system to a wide range of discharges, it is necessary to provide an automatic tool to locate ELMs without human intervention. With one exception, all present codes to locate ELMs are completely dependent on waveform amplitudes and affected by noise. In those cases, if the amplitude or noise changes from one discharge to another (or even in a single discharge), the software must be retuned. The exception is Ref. [7], that uses the signal's average amplitude and standard deviation, calculated over a short time period prior to the time point in question to determine thresholds for the onset of an ELM. This allows their algorithm to be applied to signals that vary in real time, or from different experiments, without needed to change settings within the code. The method described here is different, but has the same benefits of being able to be applied to a wide range of plasma discharges without human intervention or modification of the code. This subsection describes the application of UMEL to the location of ELMs in plasma pulses [28]. The method described in the following can be applied to a wide range of plasma discharges without human intervention or modification of the code. It is made up of two steps: first the location of the temporal interval containing ELMs and then the location of individual ELMs within it. This method has been applied to a JET database of more than 1,200 JET pulses, locating more than 220,000 ELMs. The typical sequence of a plasma pulse starts with the plasma in L-mode. Then, when auxiliary power is injected to the plasma above a certain power threshold, the plasma accesses the H-mode. The plasma returns to L-mode after the injected power is switched off. Since ELMs occur only during the phase when the plasma is in H-mode, the location of the H-mode implies the determination of the region where ELMs appear and *vice versa*. The focus of this first step is to delimit the time interval in which ELMs appear rather than the location of every single ELM, so this phase is actually a gross H-mode locator system.

Three sequential tasks are carried out in this first step: D_α normalisation, dimensionality reduction and H-mode location.

1) **D_α normalisation**. In order to optimise the computation of the SVR regression and allow the use of the same UMEL parameters over a wide range of discharges, the D_α signal is normalised to between 0 and 1.
2) **Dimensionality reduction**. The computation time of the SVR regression can be shortened significantly by reducing the number of samples to regress. Since this step implements a gross

H-mode locator, a high decomposition level of the Wavelet transform is used to reduce the number of samples to regress. The approximation coefficients of the wavelet transform are used to model the D_α waveform. These coefficients retain the most relevant signal information in both the time and frequency domains. The wavelet decomposition level has been set to 5. For example, a signal with 150,000 samples is reduced to 4,688 samples.

3) **H-mode location**. The capability of UMEL to locate time segments with relevant behaviour has been applied to locate the H-mode time interval in a plasma discharge. Given a pulse number, a SVR regression with UMEL is computed using the wavelet approximation coefficients of the D_α signal. Then, a histogram of the ESVs in time windows of 0.1 s length is computed. This histogram defines the temporal segment that has been more difficult to regress, and therefore, the time interval where the D_α signal contains high frequency components (peaks). This temporal interval corresponds to the H-mode region and therefore, the time phase with ELM activity. The borders of this region are the first and the last bins with more ESVs than a certain threshold value. The outputs of this step are the boundaries of the time interval with ELMs.

Usually, individual ELMs are localised by means of visual analysis. The process consists of recognizing peaks in the H_α / D_α signals that are synchronous with a drop in the stored diamagnetic energy. The second phase of the UMEL analysis implements this process. Since the D_α signal typically has a better signal-to noise ratio, the ELM location process begins by searching peaks in this signal. Then, simultaneous drops in the diamagnetic energy are sought. This searching process is limited to the time interval determined in the previous step. The location of ELMs is carried out in four steps: D_α peak location, ESVs combination, diamagnetic energy division and combination of information.

1) **D_α peak location**. This step locates the peaks in the D_α signal that are candidates to be ELMs. It is important to note that not all these peaks are ELMs (it must be checked using the diamagnetic energy). The location of peaks in the Da waveform is performed using UMEL . Figure 4 shows the location of ELMs in JET Pulse No: 73337. The points of the waveform outside the e-tube become ESVs. Although the SVR fit can seem a straight line, it is adapted to the low frequency shape of the D_α waveform. The identification of peaks as the points above a certain threshold is not valid since ELMs have different amplitudes and the D_α amplitude can vary from one pulse to another. The main advantage of using UMEL resides in the fact that UMEL looks for samples that do not fit a smooth regression, independently of their amplitudes.
2) **ESVs combination**. As can be observed in Figure 4 left, more than one ESV appears on each peak of the D_α signal. This step concentrates all the ESVs of each peak in a single one. The selected point is the D_α sample with the highest amplitude. After this task, each peak is represented by just one ESV. Figure 4 right shows the result of this step using the peaks in Figure 4 left as inputs. After this step, only one ESV remains on each peak, located at the maximum of the D_α waveform.

3) **Diamagnetic energy division**. The diamagnetic energy waveform is divided into small segments around the time of each D_α peak located in the previous step. These segments must allow the identification of the drop in diamagnetic energy without possible confusion from signal noise. It has been empirically determined that a segment of 35ms is enough for a clear recognition of the diamagnetic energy drop. Figure 5 shows the time windows computed for the Da peaks located in the previous tasks.
4) **Combination of information**. The last step of this method locates the drops in the diamagnetic time windows determined in the previous step. UMEL is again used as event locator. It is not possible to set a simple threshold to determine the drop in the diamagnetic energy because the amplitude of the waveform changes from one pulse to another. On the one hand, if one or more ESVs are found in the diamagnetic energy at a maximum distance of 5ms from the D_α peak, it has been empirically demonstrated that the event can be reliably classified as an ELM. In this case, the time of the ELM is determined as coincident with the maximum value of the diamagnetic energy just before the drop (see figure 6). On the other hand, if no ESV is found in the diamagnetic energy within 5ms of the D_α peak, it is discarded.

The ELM location method was applied to a JET database of more than 1,200 pulses in the range of discharge numbers [73337; 78156], which correspond to the last campaigns with the carbon wall. 226,751 ELMs have been identified in these pulses. Due to the lack of a large validated ELM database for benchmark, the performance of the ELM location method has been tested comparing its results with the ELMs manually located by experts in 20 JET discharges from the above range. The method achieves a success rate of 95% in the location of ELMs. The main statistical properties of all the ELMs in the database are reported in table I.

It is probably worth mentioning that the use of additional signals, presenting a clear signature of the ELMs occurrence, would increase the reliability of the detection even further. On the other hand, at least on JET, in many cases the D_α alone provides sufficient information for a quite accurate determination of the ELM times. The method can therefore be used, with care, even if the diamagnetic energy signal is not available, as can be the case in some discharges

4. ANALYSIS OF A GLOBAL DATABASE AT STEADY STATE

4.1 DATABASE AT STEADY STATE

To study the type I ELM dynamics, the analysis has been restricted to time intervals in which the plasma conditions are stationary. To determine stationarity, plasma current, vacuum toroidal field at R=2.96, neutral beam input (NBI) power and lower triangularity of the equilibrium field have been considered as the crucial input quantities. The signal tolerances of ±4% for toroidal magnetic field, ±2% for plasma current, ±10% for NBI input power and lower triangularity, have been taken into account. Moreover, only cases in which the heating power of the plasma has just only two contributions, ohmic power and neutral beam power, have been used, since the greatest contribution to the variation of the input power is given by radiofrequency and LH-waves heating. In addition, all experiments dealing with ELM control and mitigation techniques have been excluded. Attention has also been paid to ensuring that good quality signals are available (mainly the diamagnetic energy time trace). In the end, a database of 60 shots (reported in Appendix 3) has been retained for a total of 3448 Type I ELM time intervals. In Table II, a list of the candidate experiments for full analysis has been reported.

Figure 7 shows the planar orthogonal projections of the input space for each pulse, to provide a visual overview of the database characteristics.

As shown in this figure, there is a noticeable correlation between the plasma current and the toroidal magnetic field. This does not result from a physical relationship, but from the choices made in the implementation of the experimental programme. As will be noticed, the different pulses have been taken under different plasma conditions. A statistical analysis has been performed first generally over the whole set of examples, then a more detailed analysis has been performed either for each individual discharge or for a suitable subset of discharges (see section 5).

4.2 EXPLORATORY ANALYSIS OF THE WHOLE DATABASE AT STEADY STATE

A first exploratory approach to the data has been graphical. By means of a histogram of the ELM time intervals T and a non-parametric estimation, based on a normal kernel, of the probability density function (fig. 8), it has been possible to draw some important conclusions about the statistics of the data. It is important to note that all timings of the signals are affected by the discreteness of the sampling, performed in this case with a time step of 0.1ms. It is also worth mentioning that in some cases post-cursor/weakly compound ELMs are present, as can be confirmed by their signatures in both the D-alpha and the bolometric signals. For these cases the post-cursor component is not counted as an ELM. Of course, in the future the same tools could be directly applied to investigate also the statistics of this type of secondary events.

Figure 8 suggests a probability distribution defined on non-negative real numbers, having a maximum at about 32 ms. In this context, an attempt has been made to find the theoretical probability distribution which best describes the experimental data. Thus, the search scope has been reduced to the distributions listed in table III.

The maximum likelihood method has been used to estimate the parameters characterizing these models on the basis of the available data. For each distribution, the parameters that maximize the likelihood function have been evaluated. Table III lists the values of these parameters. The detailed mathematical expression for each of the pdfs in Table III is reported in Appendix 1.

Table IV shows the results for the different goodness of fit and model selection tests for all the pdfs investigated.

The results of AIC and BIC criteria have been sorted in ascending order by assigning a rank to each distribution. In this way, the distribution with rank 1 is the one that according to the used criterion best fits the experimental data. The top three distributions according to the AIC and BIC criteria are, in rank order, the Burr XII, Pearson VI and Dagum distributions. The K-S, C-VM and

A-D tests lead to the rejection of the null hypothesis for all the distributions, with the exception of the Burr XII and Dagum distributions. In addition, the A-D test failed, for all other distributions, rejecting the hypothesis that the distributions belong to the same population with a significance level of 5%.

For the reader's convenience, the best two distributions are repeated here:

- Burr

$$f(x) = \alpha k \beta^{\alpha k} \frac{x^{\alpha-1}}{\left(x^{\alpha} + \beta^{\alpha}\right)^{k+1}} \quad k, \alpha, \beta \in \Re^{+} \tag{12}$$

- Dagum

$$f(x) = \alpha k \beta^{\alpha k} \frac{x^{\alpha-1}}{\left(x^{\alpha} + \beta^{\alpha}\right)^{k+1}} \quad k, \alpha, \beta \in \Re^{+} \tag{13}$$

The validity of the previous results can be confirmed by a graphical method, known as the quantile-quantile (Q-Q) plot, used for comparing the empirical and theoretical probability distributions. In a Q-Q plot, the quantiles of the sample are plotted against the quantiles of the theoretical distribution. If the two sets come from a population with the same distribution, the points should fall approximately along a 45-degree reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the sample set has come from a population with a different distribution. Indeed in our case, the Q-Q plots clearly confirm that the Burr distribution provides the best fit (see Appendix 2). The Q-Q plot for the Burr distribution is shown in figure 9, together with the pdf and the cumulative distribution function, to show the good quality of the fit obtained.

Of course a good fit of the entire database does not imply necessarily that the obtained pdf represents properly the physics in the data. This fact will be discussed in detail in section 6 devoted to a more detailed analysis of subsets of the database.

5. MEMORYLESSNESS PROPERTY

The analysis described in the previous section indicates that all distributions, which can be properly fitted to the data, are very far from an exponential distribution, which is well known for being the only pdf with the property of memorylessness. This clear statistical evidence motivates the investigation of the presence of memory between subsequent ELMs.

The memorylessness property [29] refers to the conditional behaviour of random variables related to the time intervals between two subsequent events. Let *T* be the time interval between two subsequent ELMs. Suppose we know in advance that the time *T* between two ELMs is greater than a fixed value τ. The conditional probability that we need to wait less than another $\Delta\tau$ seconds before the subsequent ELM, given that the following ELM has not yet happened after τ seconds, is given by

$$\Pr\left\{T \le \tau + \Delta\tau \mid T \ge \tau\right\} = \frac{\Pr\{\tau \le T \le \tau + \Delta\tau\}}{\Pr\{T \ge \tau\}} = \frac{F_T(\tau + \Delta\tau) - F_T(\tau)}{1 - F_T(\tau)} . \tag{14}$$

where $F_T(\tau)$ is the cumulative distribution function.

If the probability (14) is independent of τ, i.e. if

$$\Pr\left\{T \le \tau + \Delta\tau \mid T \ge \tau\right\} = \Pr\left\{T \le \Delta\tau\right\} = F_T\left(\Delta\tau\right), \tag{15}$$

then, the distribution is called memoryless.

The probability (14) has been evaluated on the entire database and its dependency on the variable τ has been used as an indicator of the presence of memory. Figure 10 shows the conditional probability for the model Burr distribution with respect to the variables τ and $\Delta\tau$.

As it can be noticed from the contour lines in fig. 10, the dependence on the variable τ is not negligible, thus the process is not memoryless. If α is the conditional probability (14), the contour line indicated by α in fig. 11 represents the values of τ and $\Delta\tau$ for which the conditional probability $\Pr\{T \le \tau + \Delta\tau | T \ge \tau\} = \alpha$. For example, the last contour line at the top is related to $\alpha = 99\%$. Thus, looking at the point (150, 250) in the graph, if after $\tau = 150$ms no ELM has occurred, then there is a 99% probability that the next ELM occurs within other $\Delta\tau = 250$ms. As it can be noticed from the contour lines, the dependence on the variable τ is not negligible, thus the process is not memoryless. In fact, if the process was memoryless, the contour lines would be parallel to the x-axis, indicating no dependence on τ. This result applies to the whole database, as demonstrated here, and also applies equally to the cliques (sets of three or more pulses) since they cannot be fitted with an exponential pdf either.

It is also worth mentioning that Eq.14, in the limit of $\Delta\tau$ sufficiently small, tends to the conditional probability $p(t)$ in [7]. Moreover, for type I ELMs, in the same paper it was found that the conditional probability $p(t)$ increases roughly linearly with time or faster, and this is consistent with our analysis.

6. GROUPING OF DIFFERENT SHOTS AT STEADY STATE

All the analyses and observations reported in the previous section have been made assuming that the inter ELM intervals of pulses all belong to the same population. But the question can be asked whether this assumption is correct or not, and whether the different pulse conditions affect the statistical behaviour of ELMs. In our case, as already mentioned in section 2.3, the hypothesis that all samples belong to a population having a normal distribution cannot be met, because time intervals are positive quantities which cannot be described by a distribution defined also for negative values. For these reasons, classical techniques of the ANOVA type cannot be applied and it is necessary to use a different method, the Kruskal-Wallis test, introduced in subsection 2.3. This test makes no assumption about the distribution of the samples.

The Kruskal-Wallis test has been applied first to the observations of ELM intervals of all pulses, giving a negative result. In fact, for the test samples, the H statistic gives H = 1981.4, which, compared with the χ^2 distribution of 59th degree, is quite large. Therefore, it is concluded that, in all probability, different samples, i.e. ELM intervals belonging to different pulses, do not belong to the same population.

Given the previous results, it has been decided to group together, when possible, pulses relative to similar inputs, taking into account the signals' typical uncertainties or tolerances (±4% for toroidal magnetic field, ±2% for plasma current, ±10% for NBI input power and lower triangularity), and to perform a more detailed analysis of the new groups. In particular, two input conditions have been considered similar if all the input parameters have fallen within the limits dictated by their tolerance. A set of pulses is classed as a group if each pair of pulses of the set is characterized by similar input conditions and if all pulses belong to the same experiment. It has been possible to distinguish between three different cases: single pulses, pairs of pulses and cliques (sets of three or more pulses). The algorithm used for finding cliques is clearly reported in [30].

By means of this division, from the initial 60 pulses, 24 groups have been created, including 10 of singles, 4 of pairs and 10 of cliques. The groups are listed in table V.

For each group, the number of samples, the input conditions, the identifier of the experiment and the result of the Kruskal-Wallis test with 5% confidence level are reported. With this kind of division, there are four groups (6, 10, 16, 19) for which the test gave a positive result.

With regard to the other groups, it has not been possible to say with good confidence that the intervals of a single group belong to the same distribution, therefore pulses belonging to these groups have been kept unpaired.

For the two cliques 16 and 19, a more detailed analysis has been performed, by repeating the same steps executed for the total database. Thus, this approach differs from that described in [7] where all pulses are individually analysed.

The maximum likelihood method has been used to estimate the parameters characterizing the distributions on the basis of the available data. For each distribution, the parameters that maximize the likelihood function have been evaluated. The verification of conformity of the observed data to the different theoretical models, i.e., the goodness of fit, has been carried out using five statistical tests: the Kolmogorov–Smirnov test, the Cramer-Von Mises test, the Anderson–Darling test, the Akaike Information Criterion and the Bayesian Information Criterion. Table VI shows the results for the different tests for clique 16.

Unlike the previous case, the K-S, C-VM and A-D tests give a positive result also for Log-Logistic distribution. The A-D test faile for all other distributions, rejecting the hypothesis that the distributions belong to the same population with a significance level of 5%. These results are confirmed by the AIC and BIC criteria, assessing that the best fitting distribution is the Dagum distribution. Given the small difference between AIC and BIC of the Burr XII and Dagum distributions, and since the log-logistic distribution is a subset and a link between the two above-mentioned distributions, the distribution fitting related to the clique 16 confirm the results obtained for the entire sample, although statistical conclusions are "weaker" because of the lower number of samples.

Different results have been obtained for clique 19. In that case, by means of a histogram of the ELM time intervals *T* and a non-parametric estimation, based on a normal kernel, of the probability density function (fig.11), it has been determined that the statistics that describes the data set has a bimodal distribution. Bimodal distributions have also been found in the six pulses analysed in [31].

This pdf cannot be reproduced by the distributions considered and will probably require further examination. In any case, it is a general conclusion that the cliques do not seem always to obey the same statistics as the total database. The same conclusion has been reached in [31] by applying delay plots to the measured inter ELM time intervals in only six similar plasmas in the JET tokamak.

CONCLUSIONS

In this paper, a technique, based on the event identifier UMEL, has been applied to the localisation of ELMs in a large JET database. The success rate of the technique is better than 95% allowing the building of large, good quality databases of ELM events. Such positive results should strongly motivate the analysis of large databases of ELMs from other experiments to improve the statistical basis on which to draw conclusions about their dynamics.

The analysis has then been particularised to Type I ELMs at steady state as a basis for the assessment of their dynamical behaviour. Overall the required conditions have been met by 60 shots, corresponding to 24 experimental conditions, for a total of 3448 Type I ELM time intervals. It has been demonstrated that by far the most widely applicable, even if not universally valid, pdfs to interpret the distribution of the inter ELM intervals for this set of ELMs are the Burr and the Dagum. Moreover, the memorylessness property has been investigated suggesting the presence of memory in the ELM time intervals for the type I ELMs considered here, in agreement with previous studies [7]. The presence of memory in these time series is relevant not only for the modelling of the ELMs but can also be significant for the optimisation of the tools to control them.

On the other hand, there are clear subsets of the database, which cannot be fitted with the Burr and Dagum distributions. Therefore, at least from a statistical point of view, Type I ELMs do not show the same behaviour in their inter-ELM periods. The main consequence is that the analysis should be restricted to a homogeneous set of shots and care must be taken in deriving global conclusions from large datasets without a detailed analysis of the statistical properties of the various subsets. In particular, the operational regimes will have to be analysed in detail to see what parameters of the discharge could explain the differences in the statistical behaviour. In any case, the evidence presented indicates quite strongly that the so called Type I ELMs could in reality comprise more than one single dynamical behaviour. Again this aspect can have implications also for the strategies of ELM controls.

With regard to future developments, the evidence that different subsets of shots present different inter-ELM time statistics requires redefining the analysis to understand the critical discharge parameters influencing this aspect of the dynamics. After that, it will be possible to perform a critical investigation of more sophisticated aspects of the ELM evolution, such as memory effects, determinism and presence of chaotic behaviour. The same methods could then also be applied to other ELM types, to identify the major differences and similarities between them. Moreover, since after the installation of the new ITER Like Wall, the ELM dynamics have changed significantly on

JET, it is planned to perform the same studies for the new campaigns to assess these differences in a sound statistical way. The tools presented in this paper are considered perfectly adequate to perform this investigation in a robust and efficient way and, being absolutely general, can also be deployed for the study of other types of instabilities.

ACKNOWLEDGEMENTS

This work was supported in part by the Euratom Communities under the contract of Association between EURATOM/CIEMAT/ENEA. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

REFERENCES

[1]. J. Wesson (1987), "Tokamaks", *Oxford University Press*. ISBN 0-198-50922-7.

[2]. S. González, J. Vega, A. Murari, A. Pereira, J. M. Ramírez, S. Dormido-Canto and JET-EFDA contributors (2010), "Support vector machine-based feature extractor for L/H transitions in JET", *Review of Scientific Instruments, Proceedings of the 18th Topical Conference on High Temperature Plasma Diagnostics (HTPD 2010), Wildwood, New Jersey, USA*, vol. 81 (10E123).

[3]. A. Murari, G. Vagliasindi, M. K. Zedda, R. Felton, C. Sammon, L. Fortuna, P. Arena and JET-EFDA contributors (2006), "Fuzzy logic and support vector machine approaches to regime identification in JET", *IEEE Transactions on Plasma Science*, vol. 34 (3), pp. 1013–1020.

[4]. A. Murari, G. Vagliasindi, P. Arena, L. Fortuna, O. Barana, M. Johnson and JET-EFDA contributors (2008), "Prototype of an adaptive disruption predictor for jet based on fuzzy logic and regression trees", *Nuclear Fusion*, vol. **48** (3), p. 035010.

[5]. G. A. Rattá, J. Vega, A. Murari and JET-EFDA Contributors (2012), "Improved feature selection based on genetic algorithms for real time disruption prediction on JET", *Fusion Engineering and Design*, vol. **87** (9), pp. 1670-1678.

[6]. G. A. Rattà, J. Vega, A. Murari, G. Vagliasindi, M. F. Johnson, P. C. de Vries and JET-EFDA contributors (2010), "An advanced disruption predictor for JET tested in a simulated real-time environment", *Nuclear Fusion*, vol. **50** (2), p. 025005.

[7]. A. J. Webster, R. O. Dendy and JET EFDA contributors (2013), "Statistical Characterisation and Classification of Edge Localised Plasma Instabilities", *Physical Review Letters*, vol. 110 (15), p. 155004.

[8]. G. W. Corder, D. I. Foreman (2009), "Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach", Wiley. ISBN 978-0-470-45461-9

[9]. T. W. Anderson, D. A. Darling (1954), "A test of goodness of fit", *Journal of the American Statistical Association*, vol. **49** (268), pp. 765-769.

[10]. D. A. Darling (1957), "The Kolmogorov-Smirnov, Cramer-von Mises Tests", *The Annals of Mathematical Statistics*, vol. **28** (4), pp. 823-838.

[11]. H. Akaike (1974), “A new look at the statistical model identification”, *IEEE Transactions on Automatic Control*, vol. **19** (6), pp. 716–723.

[12]. G. Schwarz (1978), “Estimating the dimension of a model”, *Annals of Statistics*, vol. **6** (2), pp. 461–464.

[13]. D. R. Cox (2006), “Principles of statistical inference”, Cambridge University Press, New York. ISBN 9780521685672.

[14]. C. Davis (2002), “Statistics and data analysis in geology”, John Wiley & Sons, Inc., New York.

[15]. J. Vega, A. Murari, S. González and JET-EFDA contributors (2010), “A universal support vector machines based method for automatic event location in waveforms and video-movies: applications to massive nuclear fusion databases”, *Review of Scientific Instruments*, vol. **81** (2), p. 023505.

[16]. F. Wagner *et al.* (1982), “Regime of improved confinement and high beta in neutral-beam-heated divertor discharges of the ASDEX Tokamak”, *Physical Review Letters*, vol. **49** (19), pp. 1408–1412.

[17]. F. Wagner *et al.* (1984), “Development of an edge transport barrier at the H-Mode transition of ASDEX”, *Physical Review Letters*, vol. **53** (15), pp. 1453–1456.

[18]. S. M. Kaye *et al.* (1984), “Attainment of high confinement in neutral beam heated divertor discharges in the PDX Tokamak”, *Journal of Nuclear Materials*, vol. **121**, pp. 115–125.

[19]. K. H. Burrell *et al.* (1987), “Observation of an improved energy-confinement regime in neutral-beam–heated divertor discharges in the DIII-D Tokamak”, *Physical Review Letters*, vol. **59** (13), pp. 1432–1435.

[20]. A. Tanga *et al.* (1987), “Magnetic separatrix experiments in JET”, *Nuclear Fusion*, vol. **27** (11), pp. 1877–1895.

[21]. V. Erckmann *et al.* (1993), “H mode of the W 7-AS stellarator”, *Physical Review Letters*, vol. **70** (14), pp. 2086–2089.

[22]. F. Wagner (2007), “A quarter-century of H-mode studies”, *Plasma Physics and Controlled Fusion*, vol. **49** (12B), pp. B1–B33.

[23]. M. Keilhacker *et al.* (1984), “Confinement studies in L and H-type ASDEX discharges”, *Plasma Physics and Controlled Fusion*, vol. **26** (1A), pp. 49–63.

[24]. A. J. Smola, B. Schölkopf (2004), “A tutorial on support vector regression”, *Statistics and Computing*, vol. **14** (3), pp. 199–222.

[25]. V. Vapnik (2006), “Estimation of Dependences Based on Empirical Data”, *Springer*. ISBN 978-0387-30865-4.

[26]. H. W. Lilliefors (1969), “On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown”, *Journal of the American Statistical Association*, vol. **64** (325), pp. 387–389.

[27]. R. D’Agostino, M. Stephens (1986), “Goodness-of-fit Techniques”, *Marcel Dekker Inc.*, New York.

[28]. S. González, J. Vega, A. Murari, A. Pereira, M. Beurskens and JET-EFDA contributors (2010), “Automatic ELM location in JET using a Universal Multi-Event Locator”, *Fusion Science and Technology*, vol. **58** (3), pp. 755–762.

[29]. W. Feller (1968), “An introduction to probability theory and its applications”, John Wiley & Sons, Inc., New York.

[30]. F. Harary, I. C. Ross (1957), “A procedure for clique detection using the group matrix”, *Sociometry*, vol. **20** (3), pp. 205-215.

[31]. F.A. Calderon, R.O. Dendy, S.C. Chapman, A.J. Webster, B. Alper, R.M. Nicol, “Identifying low-dimensional dynamics in Type-I edge-localised-mode processes in JET plasmas” *Physics of Plasmas* **20**, 042306 (2013).

Period (s)	Number of ELMs	Period (s)	Number of ELMs
< 0.01	3,200	0.04 to 0.05	19,809
0.01 to 0.02	56,550	0.05 to 0.06	11,409
0.02 to 0.03	61,373	0.06 to 0.07	6,800
0.03 to 0.04	30,100	0.07 to 0.08	4,000

Table I: Distribution of the periods of the located ELMs

Experiment		N° of pulses
E-1.1.1	Characterisation of divertor detachment	1
E-1.1.6	Massive gas injection	2
E-1.3.2	Carbon Migration – ITER Like Wall (ILW) reference scenarios	8
E-1.3.4	Disruption mitigation by massive gas injection	1
E-1.3.5	Fuel retention - ILW reference scenarios	1
E-2.4.1	Characterization of large/regular ELMs	36
H-1.1.1	ITER Like Antenna commissioning at 42MHz + High resolution Thomson scattering	1
HLC-1.1.4	Large ELMs (>0.5MJ)	1
HLC-9	Quarts Microbalance tests	1
S1-2.4.9	Pedestal identity with AUG & DIII-D + rho* scan	2
S1-2.4.12	Scaling of confinement and pedestal with rho* and beta	6

Table II: List of candidate experiments for full analysis. The left column reports the acronym of JET experiments the analyzed shots belong to.

Distribution	Parameters	Distribution	Parameters
Burr XII	$k = 1.8385$, $\alpha = 2.6276, \beta = 0.0507$	**Levy**	$\sigma = 0.0302$
Dagum	$k = 0.6225$, $\alpha = 3.7915, \beta = 0.0457$	**Log–Logistic**	$\alpha = 3.0902, \beta = 0.0370$
Fatigue Life	$\alpha = 0.6044, \beta = 0.0357$	**Log-Normal**	$\sigma = 0.5769, \mu = -3.3201$
Frechet	$\alpha = 1.5978, \beta = 0.0269$	**Pearson VI**	$\alpha_1 = 4.2871, \alpha_2 = 14.910$ $\beta = 0.1372$
Gamma	$\alpha = 3.3424, \beta = 0.0127$	**Rayleigh**	$\sigma = 0.0346$
Generalized Gamma	$k = 1.0475$, $\alpha = 3.2125, \beta = 0.0142$	**Weibull**	$\alpha = 1.8418, \beta = 0.0478$
Inverse Gamma	$\alpha = 2.9570, \beta = 0.0894$	**Weibull (3P)**	$\alpha = 1.7368, \beta = 0.0450$ $\gamma = 0.00237$
Inverse Gaussian	$\lambda = 0.1061$ $\mu = 0.0423$		

Table III: List of distributions and best fit parameters.

Distribution	Kolmogorov-Smirnov		Cramer-Von Mises		Anderson-Darling		AIC		BIC	
	Statistic	P-Value	Statistic	P-Value	Statistic	P-Value	Value	Rank	Value	Rank
Burr XII	0.0183	0.193	0.2053	0.276	1.5154	0.175	-10086	1	-16957	1
Dagum	0.0165	0.317	0.1709	0.34	1.5375	0.184	-10074	3	-16946	3
Fatigue Life	0.0579	0	3.4426	0	17.771	0	-9963	8	-16842	8
Frechet	0.1061	0	15.427	0	94.073	0	-9061	14	-15941	14
Gamma	0.0358	0	0.7053	0.01	4.5715	0.003	-10033	4	-16887	7
Gen. Gamma	0.0498	0	1.6358	0	9.4171	0	-10017	5	-16899	4
Inv. Gamma	0.0880	0	9.2138	0	51.409	0	-9567	13	-16447	12
Inv. Gaussian	0.0622	0	4.1252	0	21.409	0	-9927	9	-16756	9
Levy	0.3935	0	161.98	0	757.57	0	-5599	15	-12487	15
Log-Logistic	0.0324	0.003	0.8871	0.007	7.5503	0	-10015	6	-16888	5
Log-Normal	0.0443	0	2.1979	0	12.329	0	-10007	7	-16888	6
Pearson VI	0.0248	0.03	0.6678	0.014	3.8108	0.012	-10078	2	-16949	2
Rayleigh	0.0799	0	5.2845	0	28.971	0	-9716	12	-16595	11
Weibull	0.0629	0	3.5819	0	24.537	0	-9764	11	-16351	13
Weibull (3P)	0.0564	0	2.7866	0	18.432	0	-9862	10	-16734	10

Table IV: Results of tests for each theoretical model. Highlighted in grey distributions for which the K-S, C-VM and A-D tests gave a positive result.

G	Pulses	n_i	Inputs	Exp.	K-W test (5%)
1	73397, 73445, 73446, 73447, 73450	246	2.5MA, 2.4T, 12.5-14.5MW, 0.23-0.27	E-1.3.2	NO
2	73484	108	2.5MA, 2.4T, 14.8MW, 0.28	E-1.3.2	NO
3	74130	78	2.5MA, 2.4T, 13.6MW, 0.25	E-1.3.2	NO
4	74364	116	2MA, 1.8T, 15.5MW, 0.36	E-1.3.2	NO
5	74365, 74366, 74367, 74368, 74369, 74371, 74372, 74373, 74374, 74375	347	2.5MA, 2.5T, 15.5-18.5MW, 0.34-0.35	E-1.3.2	NO
6	74375, 74376	54	2.5MA, 2.5T, 15.1-15.5MW, 0.33-0.35	E-1.3.5	YES
7	74378, 75724, 75726, 75727, 75728, 75731, 75732, 76481	497	2MA, 2T, 11-12.8MW, 0.32-0.37	HLC-9	NO
8	74378, 75724, 75726, 75727, 75728, 75731, 75732, 76476	497	2MA, 2T, 12.1-14.5MW, 0.32-0.37	E-2.4.1	NO
9	74378, 75724, 75728, 75731, 75732, 76473, 76474, 76475, 76476, 76477, 76478, 76479	705	2MA, 2T, 12.5-15.2MW, 0.32-0.37	E-2.4.1	NO
10	74443, 74444	90	2.5MA, 2.7T, 14-15.8MW, 0.32	E-2.4.1	YES
11	74612, 74613, 77073	291	2.6MA, 2.3-2.4T, 16.7-17MW, 0.36	E-2.4.1	NO
12	74793, 74795	286	1.7MA, 1.6T, 9.1-10.5MW, 0.35-0.36	E-2.4.1	NO
13	74798	120	1.7MA, 1.6T, 16.8MW, 0.37	E-2.4.1	NO
14	75118	33	1.7MA, 1.8T, 9.4MW, 0.26	E-2.4.1	NO
15	75724, 75728, 76471, 76472, 76473, 76474, 76475, 76476, 76477,76478, 76479	439	2MA, 2T, 12.6-15.4MW, 0.32-0.37	E-2.4.1	NO
16	76428, 76430, 76431, 76437, 76438	197	2MA, 2T, 7.5MW, 0.35	E-2.4.1	YES
17	76440, 76443, 77192	126	2MA, 2T, 7.5-8.4MW, 0.31-0.36	E-2.4.1	NO
18	76470, 76480	45	2MA, 2T, 16.8-19.8MW, 0.37	E-2.4.1	NO
19	76470, 76471, 76472, 76473, 76474, 76475, 76476, 76477, 76478, 76479	363	2MA, 2T, 14.5-16.8MW, 0.35-0.37	E-2.4.1	YES
20	76812	60	2MA, 1.8T, 18.5MW, 0.37	E-2.4.1	NO
21	78448	19	2.5MA, 2.7T, 12.9MW, 0.4	S1-2.4.9	NO
22	78750	104	1.5MA, 1.8T, 17.6MW, 0.32	S1-2.4.9	NO
23	79389	132	2MA, 2T, 11.9MW, 0.32	S1-2.4.12	NO
24	79546	177	1.5MA, 1.8T, 17.7MW, 0.45	S1-2.4.12	NO

Table V: Groups with division by input conditions and belonging to the same experiment.

Distribution	Kolmogorov-Smirnov		Cramer-Von Mises		Anderson-Darling		AIC		BIC	
	Statistic	P-Value	Statistic	P-Value	Statistic	P-Value	Value	Rank	Value	Rank
Burr XII	0.0550	0.575	0.1147	0.526	0.8231	0.478	-473.9	2	-852.1	3
Dagum	0.0512	0.640	0.1042	0.557	0.7853	0.482	-478.5	1	-856.6	1
Fatigue Life	0.1596	0	1.4995	0	8.8388	0	-369.7	10	-753.1	11
Frechet	0.2724	0	5.0472	0	27.820	0	-203.8	14	-587.2	14
Gamma	0.0883	0.097	0.4549	0.056	2.9634	0.025	-445.2	5	-828.6	5
Gen. Gamma	0.0885	0.101	0.4245	0.076	2.7816	0.043	-452.0	4	-830.1	4
Inv. Gaussian	0.1601	0	1.5069	0	8.8837	0	-366.3	12	-749.7	12
Levy	0.5130	0	14.657	0	67.038	0	12.76	15	-376.0	15
Log-Logistic	0.0511	0.654	0.0991	0.580	0.7569	0.486	-471.6	3	-855.0	2
Log-Normal	0.1185	0.006	0.8216	0.007	5.1908	0.003	-407.5	9	-790.9	9
Inv. Gamma	0.1813	0	2.2259	0	13.173	0	-309.5	13	-692.9	13
Pearson VI	0.0886	0.071	0.4563	0.038	2.9715	0.022	-445.0	6	-823.2	6
Rayleigh	0.2572	0	3.9910	0	20.623	0	-368.3	11	-757.0	10
Weibull	0.1095	0.022	0.8282	0.006	5.2721	0.001	-436.1	8	-819.6	7
Weibull (3P)	0.1087	0.024	0.8205	0.006	5.2308	0.001	-436.2	7	-814.4	8

Table VI: Results of tests for each theoretical model for clique 16. Highlighted in grey distributions for which the K-S, C-VM and A-D tests gave a positive result.

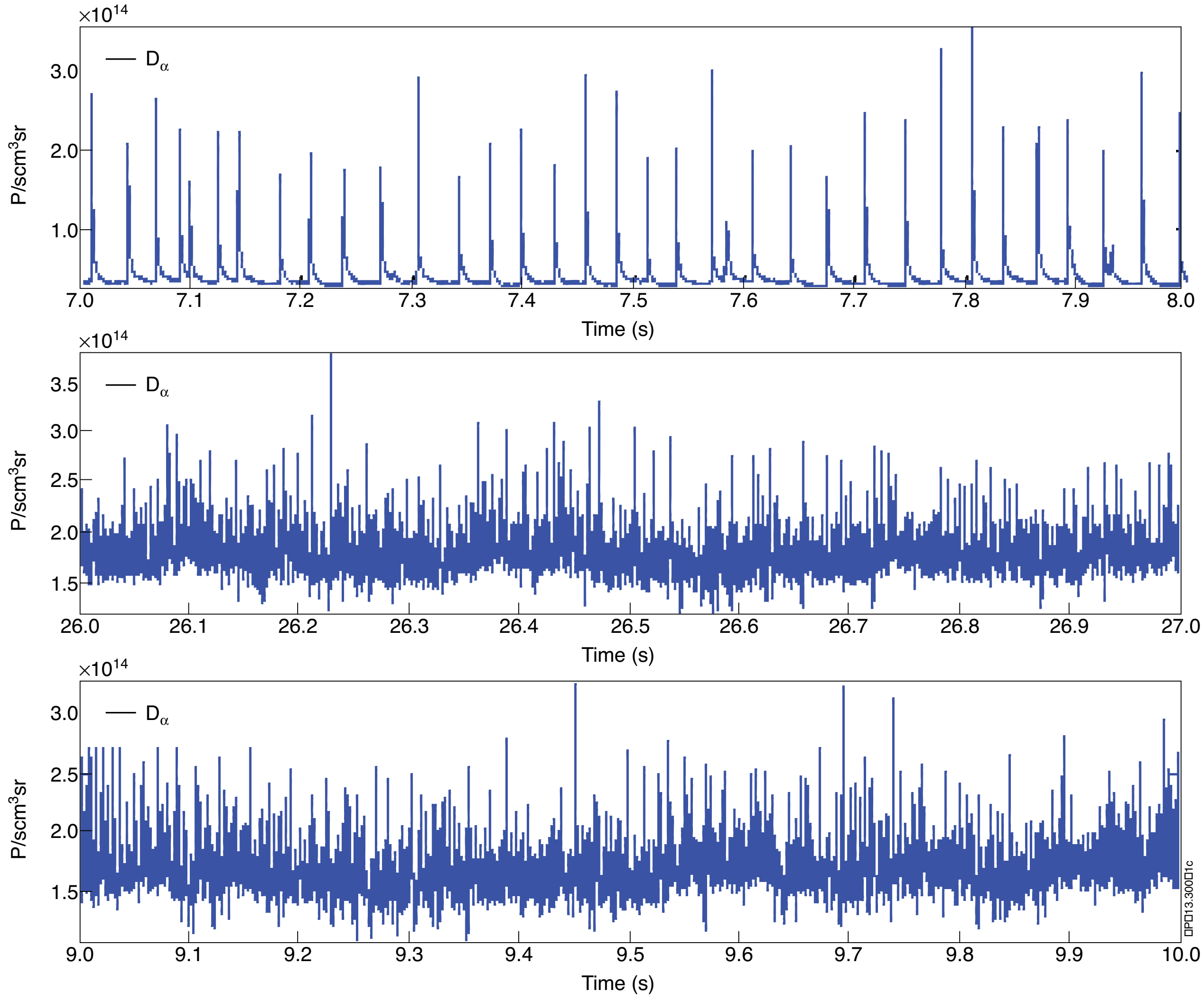
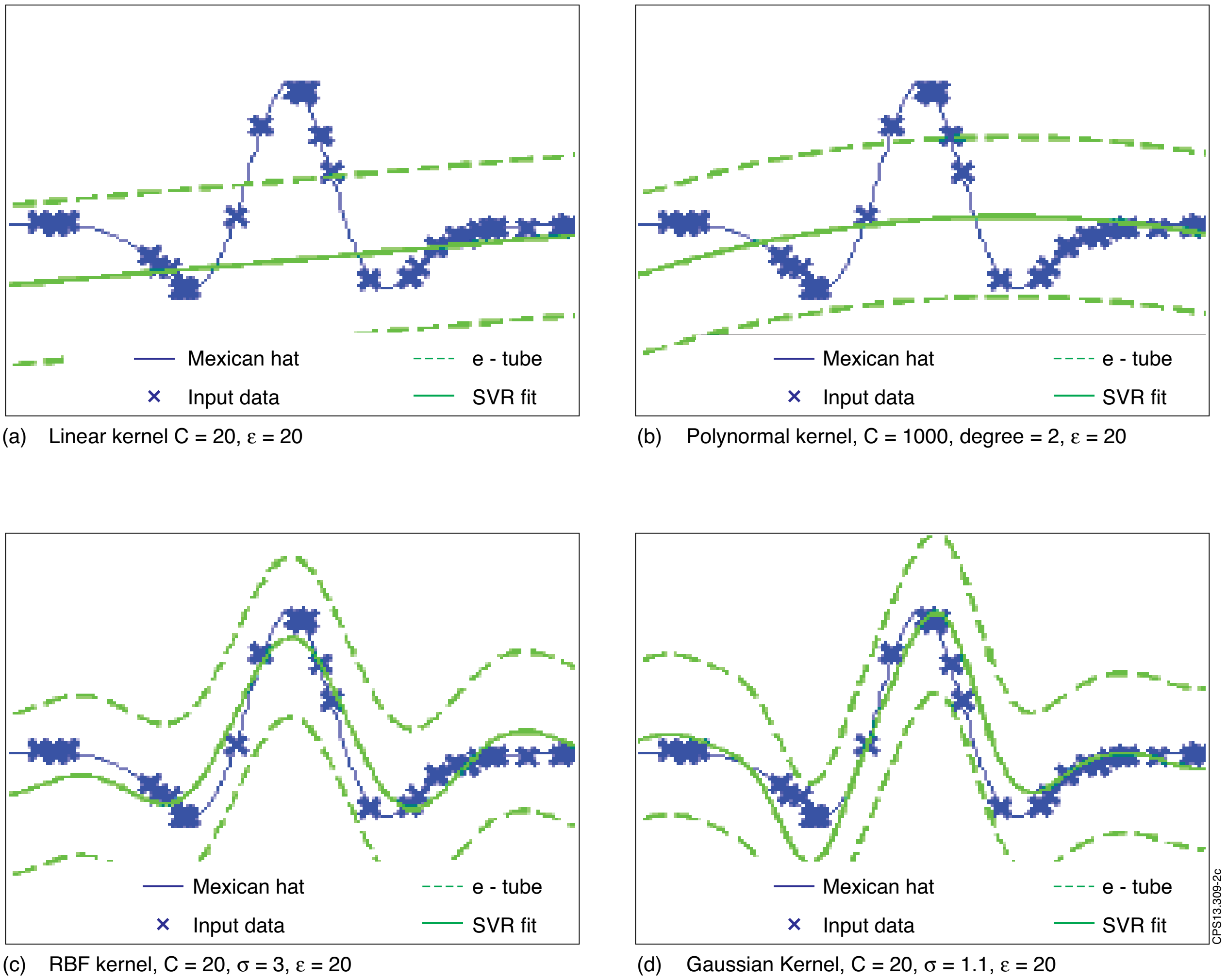


Figure 1: List of different bursts types as defined in [20]. Top: Type I ELM; Middle: Type II ELM; Bottom: Type III ELM.



(a) Linear kernel C = 20, ε = 20

(b) Polynormal kernel, C = 1000, degree = 2, ε = 20

(c) RBF kernel, C = 20, σ = 3, ε = 20

(d) Gaussian Kernel, C = 20, σ = 1.1, ε = 20

Figure 2: SVR of the Mexican hat function using four different kernels. a) linear kernel b) polynomial kernel of degree 2 c) RBF kernel d) Gaussian kernel.

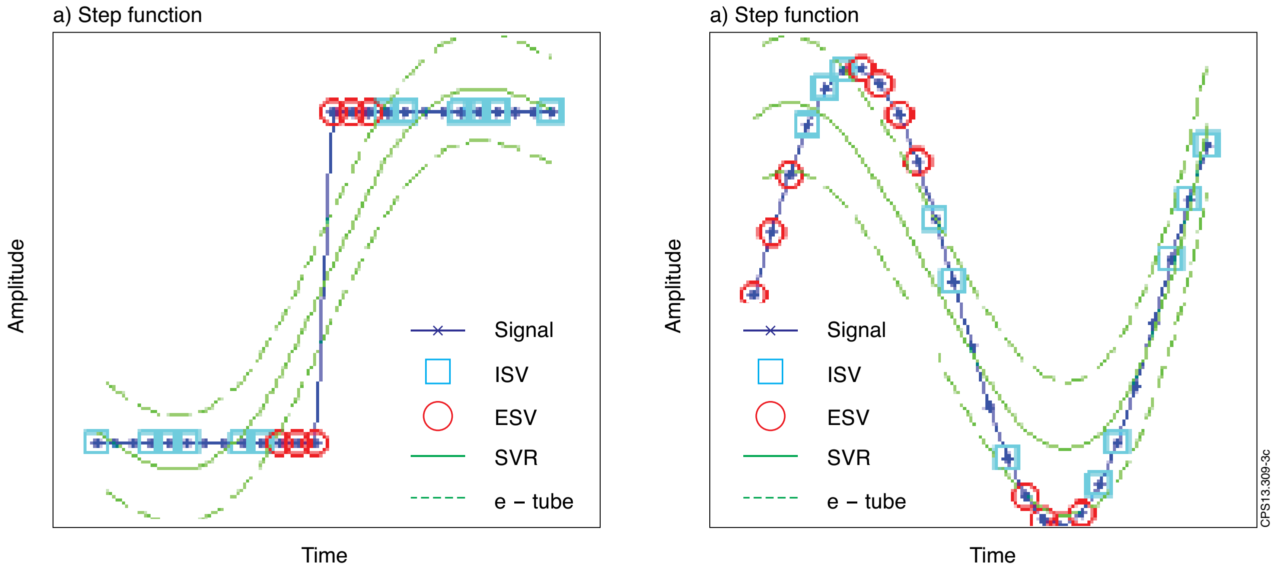


Figure 3: UMEL fit to a step and a sinusoidal function.

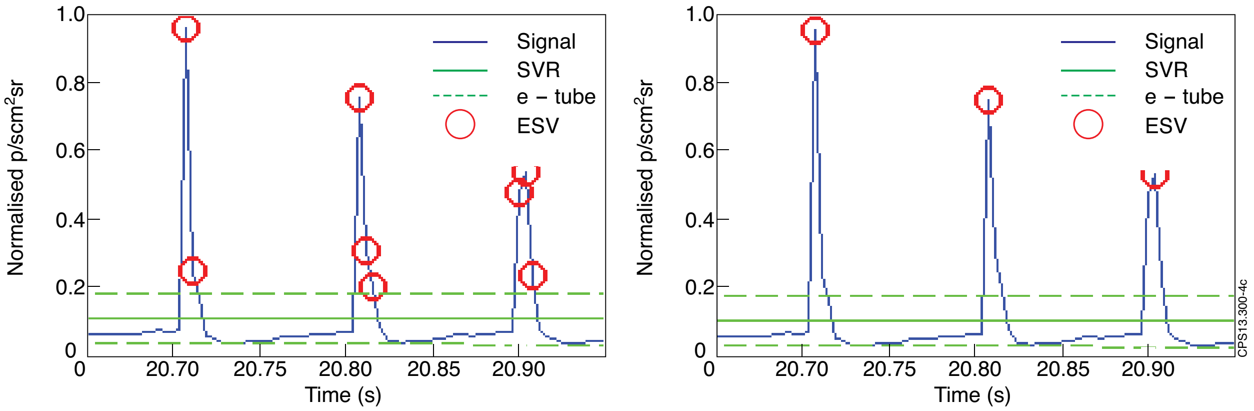


Figure 4: ELMs location step 2. Example of D_α peak location and ESVs combination, JET Pulse No: 73337

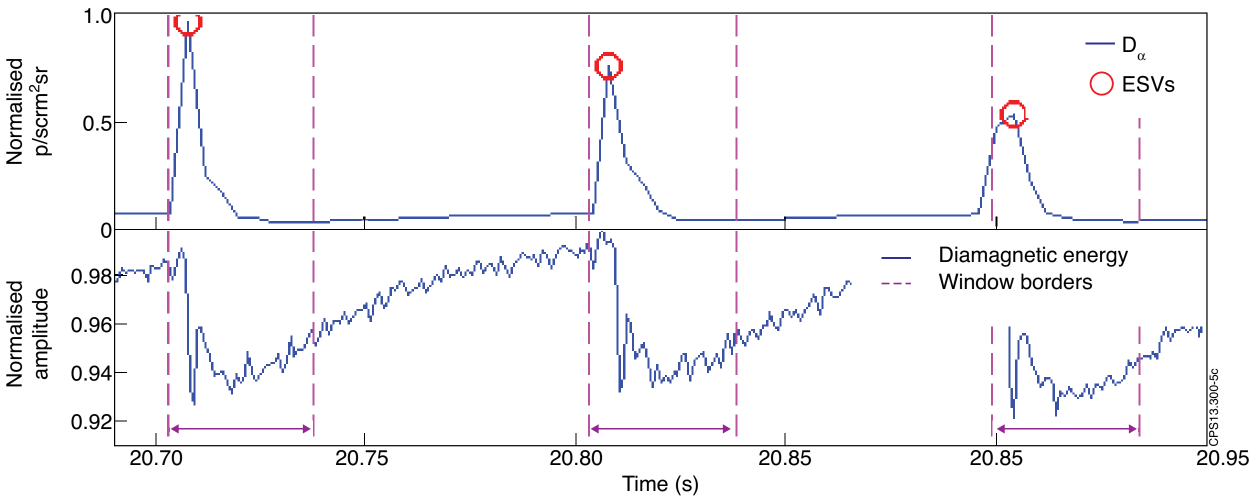


Figure 5: ELM location, step 2. Division of the diamagnetic energy example, JET Pulse No: 73337.

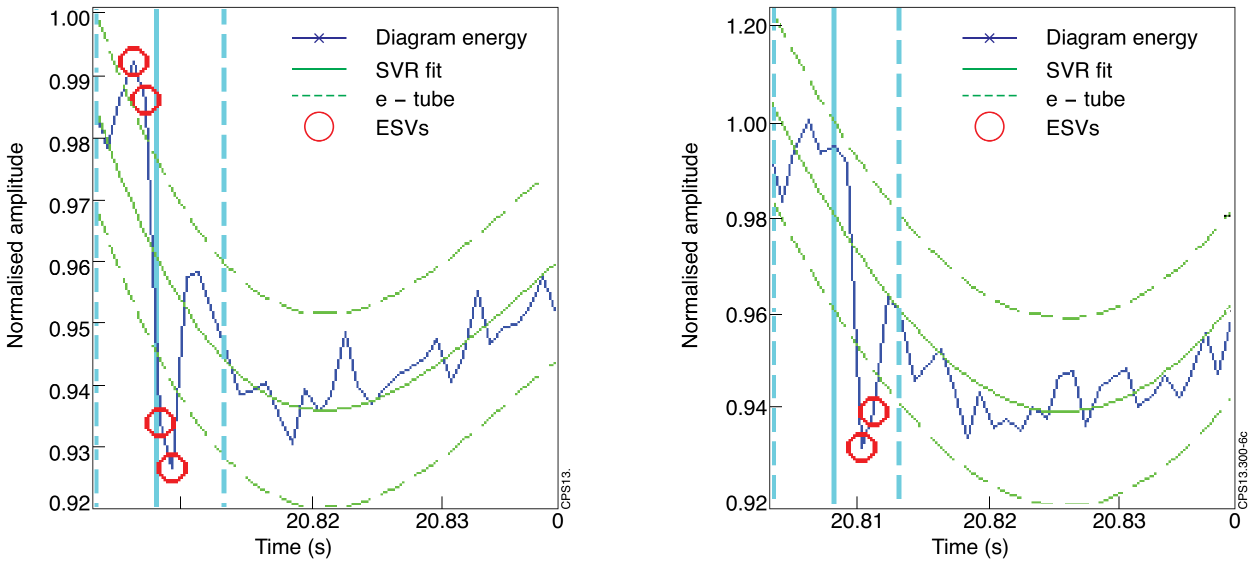


Figure 6: ELM location, step 2. Example of the diamagnetic energy drop location, JET Pulse No: 73337.

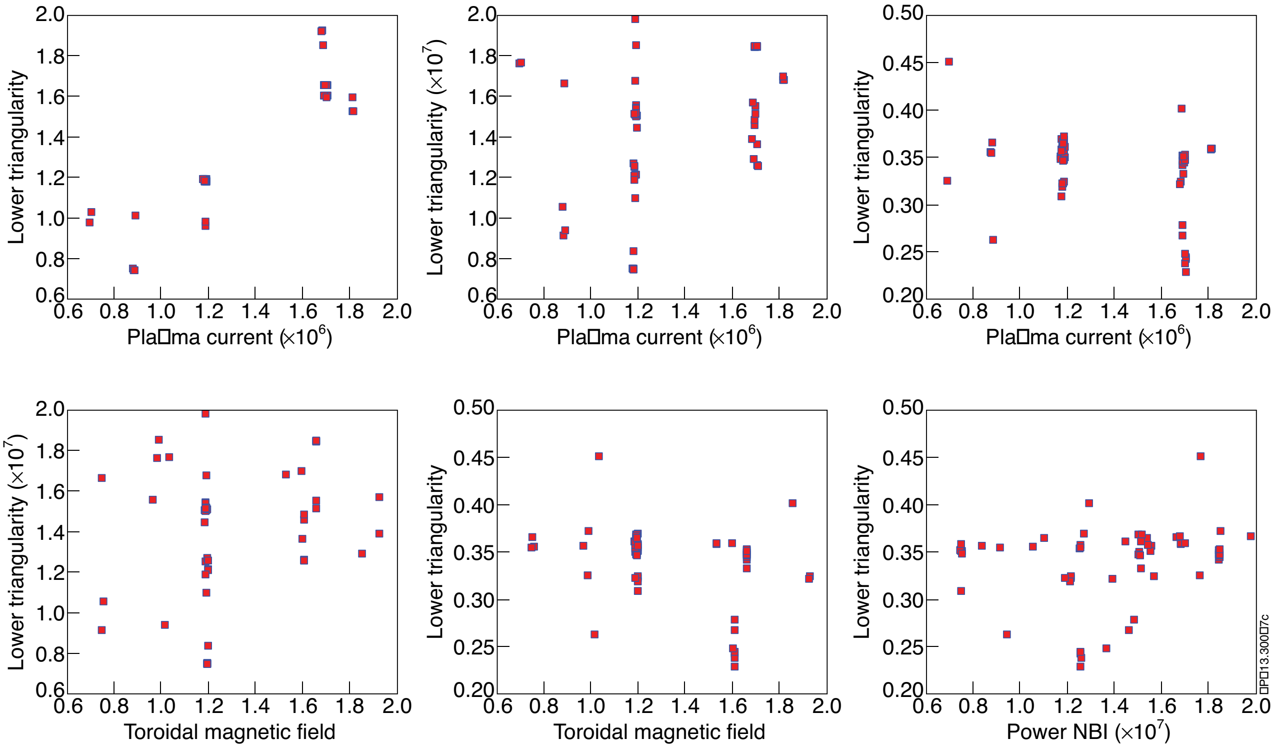


Figure 7: Planar orthogonal projections of the input space.

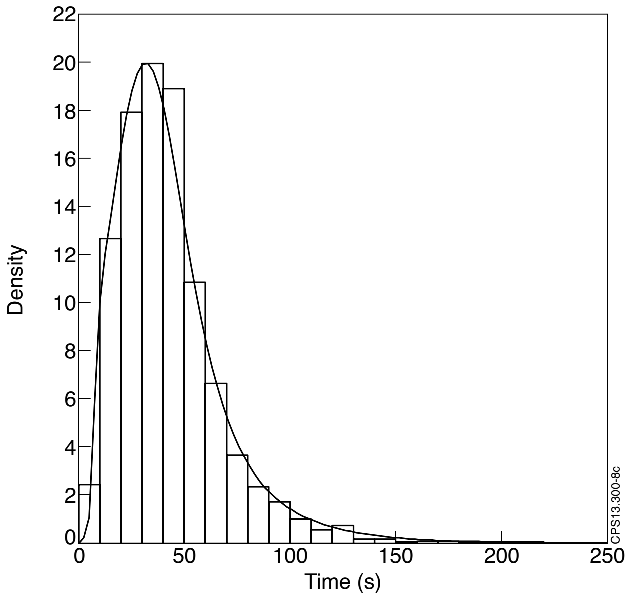


Figure 8: Data histogram and non-parametric estimation of the pdf.

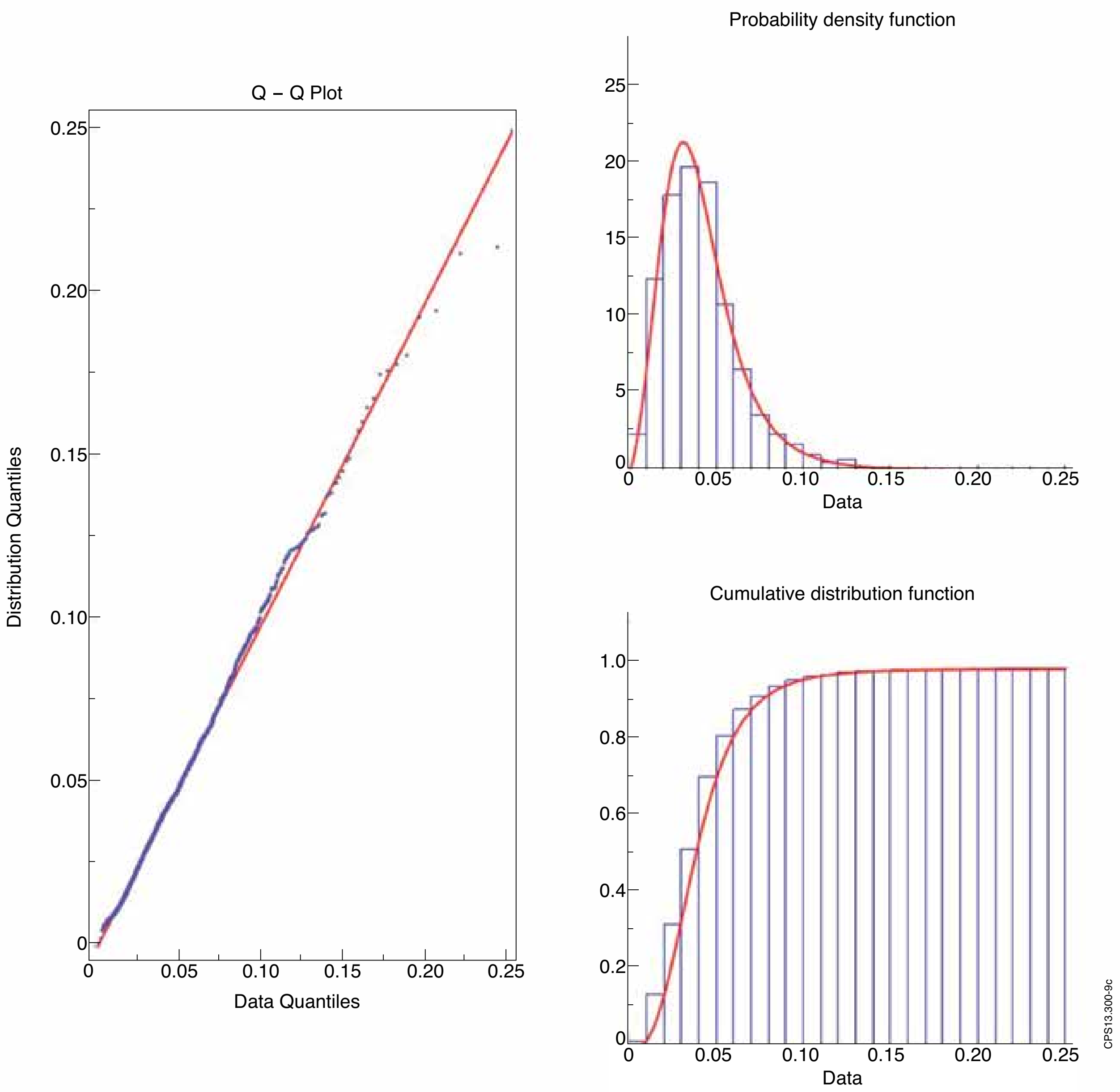


Figure 9: Q-Q plot, probability density function and cumulative distribution function for the Burr XII model (parameters: k = 1.8385, α = 2.6276, β = 0.0507). The blue lines are the experimental values and the red line the theoretical distribution.

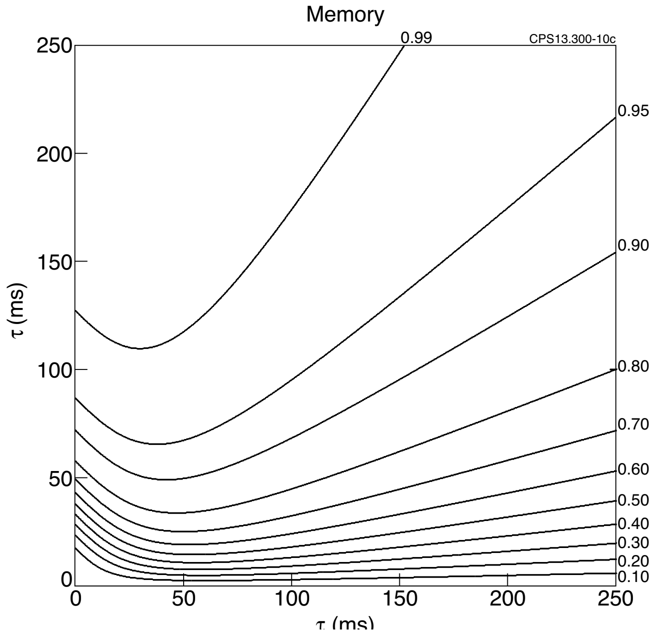


Figure 10: Contour plot of probability Pr{T≤τ+Δτ|T≥τ} with respect to τ and Δτ for the Burr pdf using the parameters fitting the experimental data (k = 1.8385, α = 2.6276, β = 0.0507).

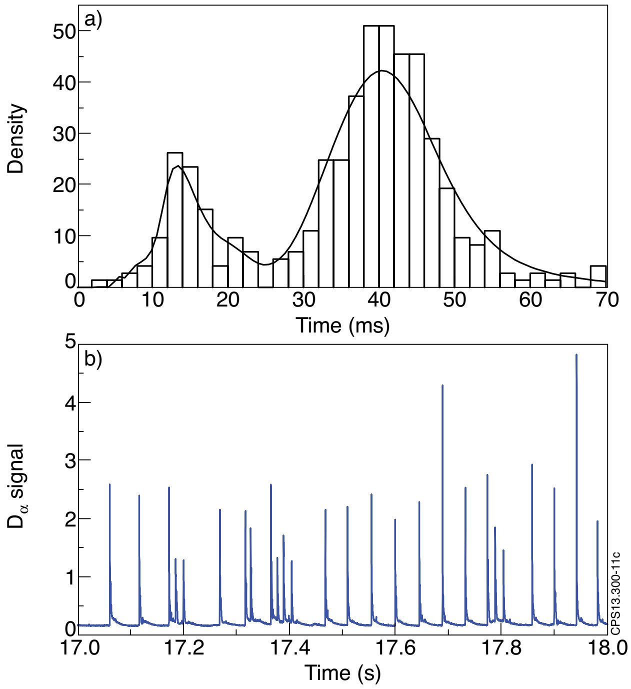


Figure 11 (a) Data histogram and non-parametric estimation of the probability density function for clique 19; (b) example of ELM time series taken from Pulse No: 76474 from clique 19.

APPENDIX 1

Mathematical expressions of the tested Probability Density Functions

Distribution	Probability density function	Distribution	Probability density function
Burr XII or Singh-Maddala	$\alpha k\beta^{\alpha k}\dfrac{x^{\alpha-1}}{\left(x^{\alpha}+\beta^{\alpha}\right)^{k+1}}$, $k,\alpha,\beta>0$	Levy	$\sqrt{\dfrac{\sigma}{2\pi x^{3}}}\exp\left(-\dfrac{\sigma}{2x}\right)$ $\sigma>0$
Dagum	$\alpha k\beta^{\alpha}\dfrac{x^{\alpha k-1}}{\left(x^{\alpha}+\beta^{\alpha}\right)^{k+1}}$, $k,\alpha,\beta>0$	Log–Logistic	$\alpha\beta^{\alpha}\dfrac{x^{\alpha-1}}{\left(x^{\alpha}+\beta^{\alpha}\right)^{2}}$, $\alpha,\beta>0$
Fatigue Life or Birnbaum-Saunders	$\dfrac{x+\beta}{2\alpha x\sqrt{2\pi\beta x}}\exp\left(-\dfrac{(x-\beta^{\ 2})}{2\beta\alpha^{2}x}\right)$, $\alpha,\beta>0$	Log–Normal	$\dfrac{1}{x\sqrt{2\pi\sigma^{2}}}\exp\left(-\dfrac{\ln x-\mu^{\ 2}}{2\sigma^{2}}\right)$ $\sigma>0,\mu\in\mathbf{R}$
Frechet	$\dfrac{\alpha\beta^{\alpha}}{x^{\alpha+1}}\exp\left(-\left(\dfrac{\beta}{x}\right)^{\alpha}\right)$, $\alpha,\beta>0$	Pearson VI	$\dfrac{\beta^{\alpha_2}x^{\alpha_1-1}}{B(\alpha_1,\alpha_2)\left(x+\beta\right)^{\alpha_1+\alpha_2}}$, $\alpha_1,\alpha_2,\beta>0$
Gamma	$\dfrac{x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}}\exp\left(-\dfrac{x}{\beta}\right)$, $\alpha,\beta>0$	Rayleigh	$\dfrac{x}{\sigma^{2}}\exp\left(-\dfrac{x^{2}}{2\sigma^{2}}\right)$ $\sigma>0$
Generalized Gamma	$\dfrac{kx^{k\alpha-1}}{\Gamma(\alpha)\beta^{k\alpha}}\exp\left(-\left(\dfrac{x}{\beta}\right)^{k}\right)$ $k,\alpha,\beta>0$	Weibull	$\alpha\beta^{-\alpha}x^{\alpha-1}\exp\left(-\left(\dfrac{x}{\beta}\right)^{\alpha}\right)$ $\alpha,\beta>0$
Inverse Gamma	$\dfrac{\beta^{\alpha}}{\Gamma(\alpha)x^{\alpha+1}}\exp\left(-\dfrac{\beta}{x}\right)$, $\alpha,\beta>0$	Weibull (3P)	$\alpha\beta^{-\alpha}\left(x-\gamma\right)^{\alpha-1}\exp\left(-\left(\dfrac{x-\gamma}{\beta}\right)^{\alpha}\right)$ $\alpha,\beta>0$
Inverse Gaussian	$\sqrt{\dfrac{\lambda}{2\pi x^{3}}}\exp\left(-\dfrac{\lambda(x-\mu)^{2}}{2\mu^{2}x}\right)$ $\lambda,\mu>0$		

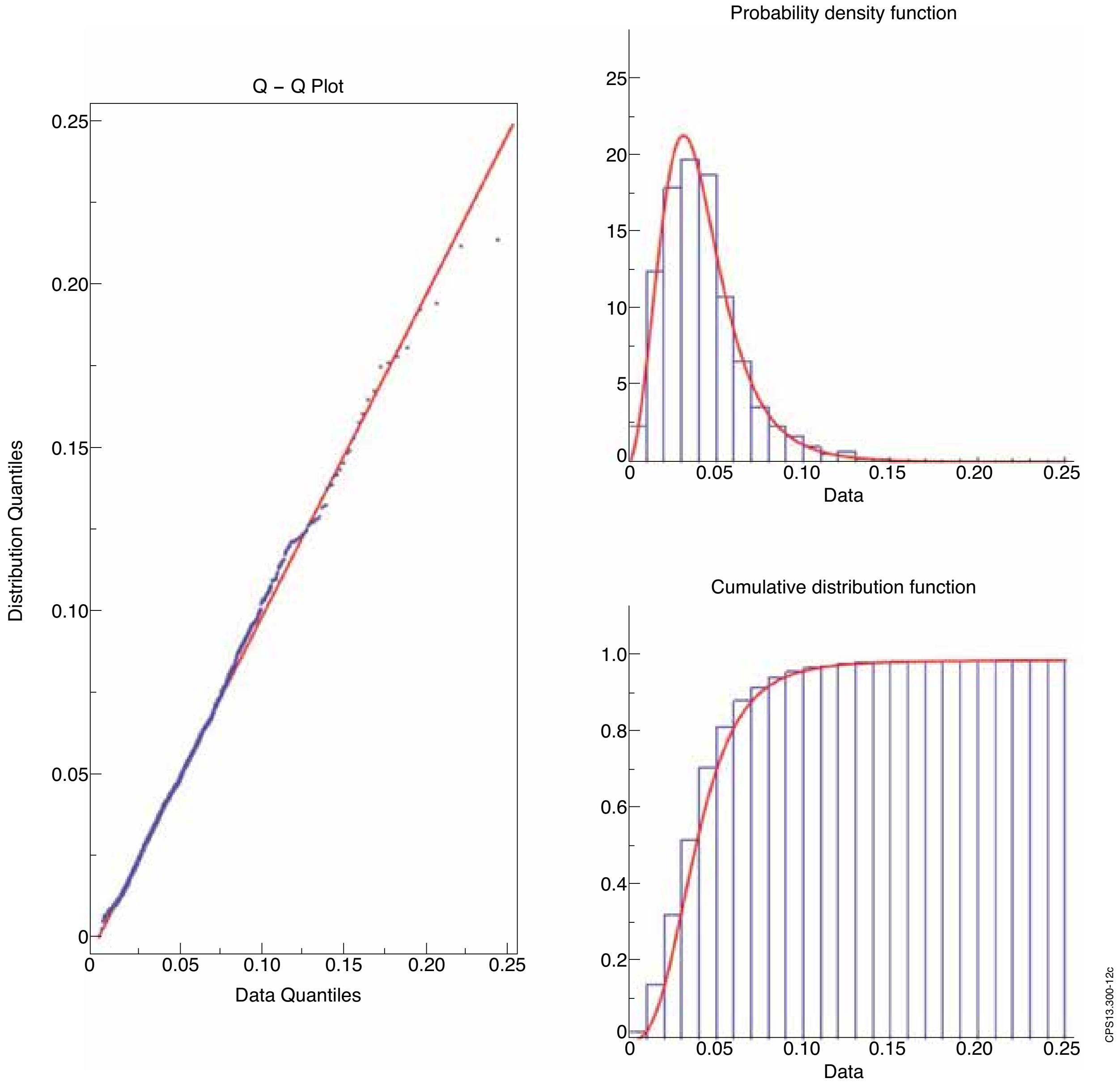
$\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$ is the Gamma function, $B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}dt$ is the Beta function and $I_0(z)=\sum_{k=0}^{\infty}\frac{(z/2)^{2k}}{(k!)^2}$ is the modified Bessel function of the first kind with order zero.

APPENDIX 2

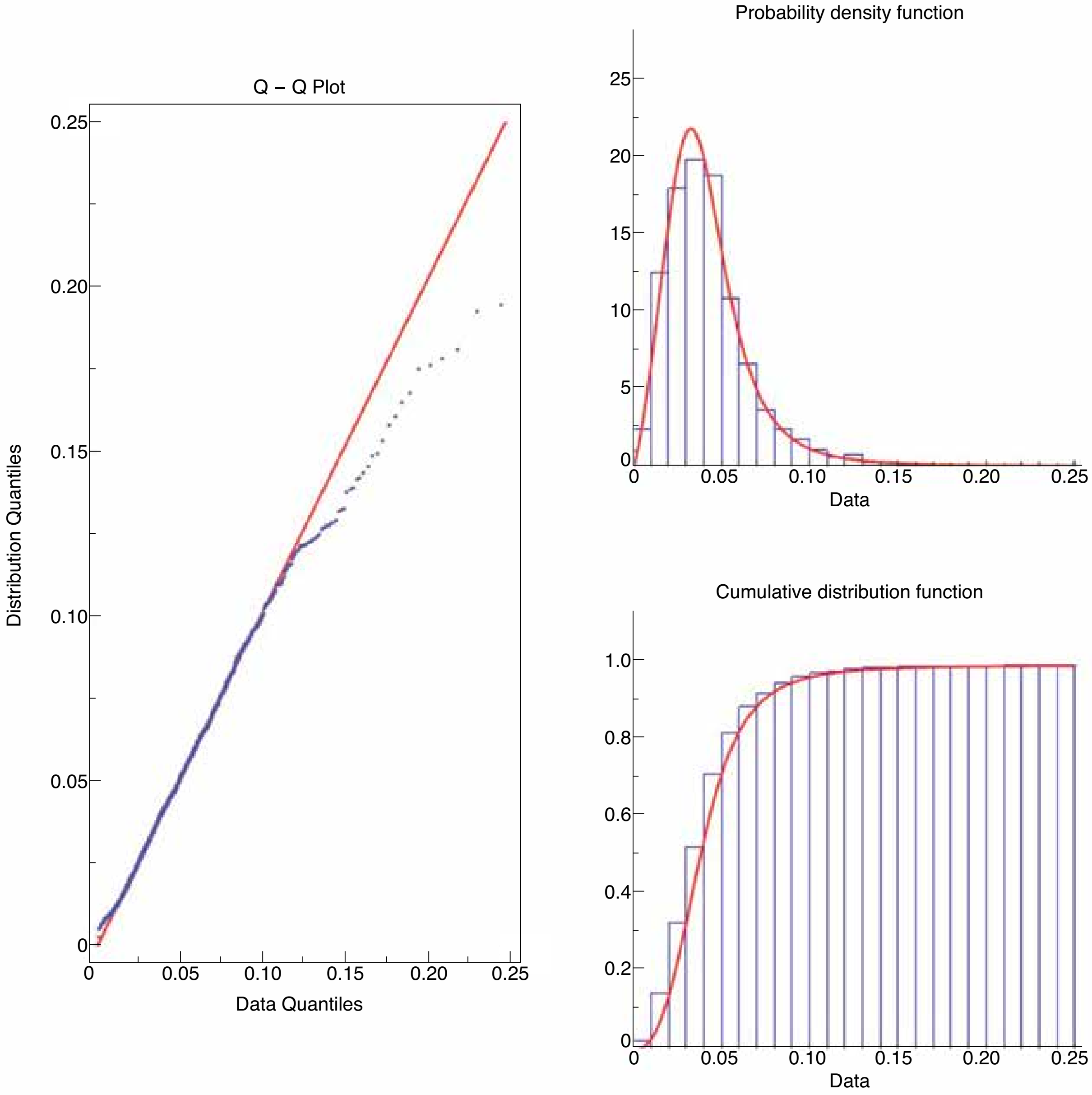
Q-Q plots and the probability and cumulative density functions graphs

The following figure shows, for each theoretical model, the Q-Q plots and the probability and cumulative density functions graphs. The Q-Q plot is a graphical method for comparing two probability distributions by plotting the quantiles of the first data set against the quantiles of the second data set. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions

Burr XII

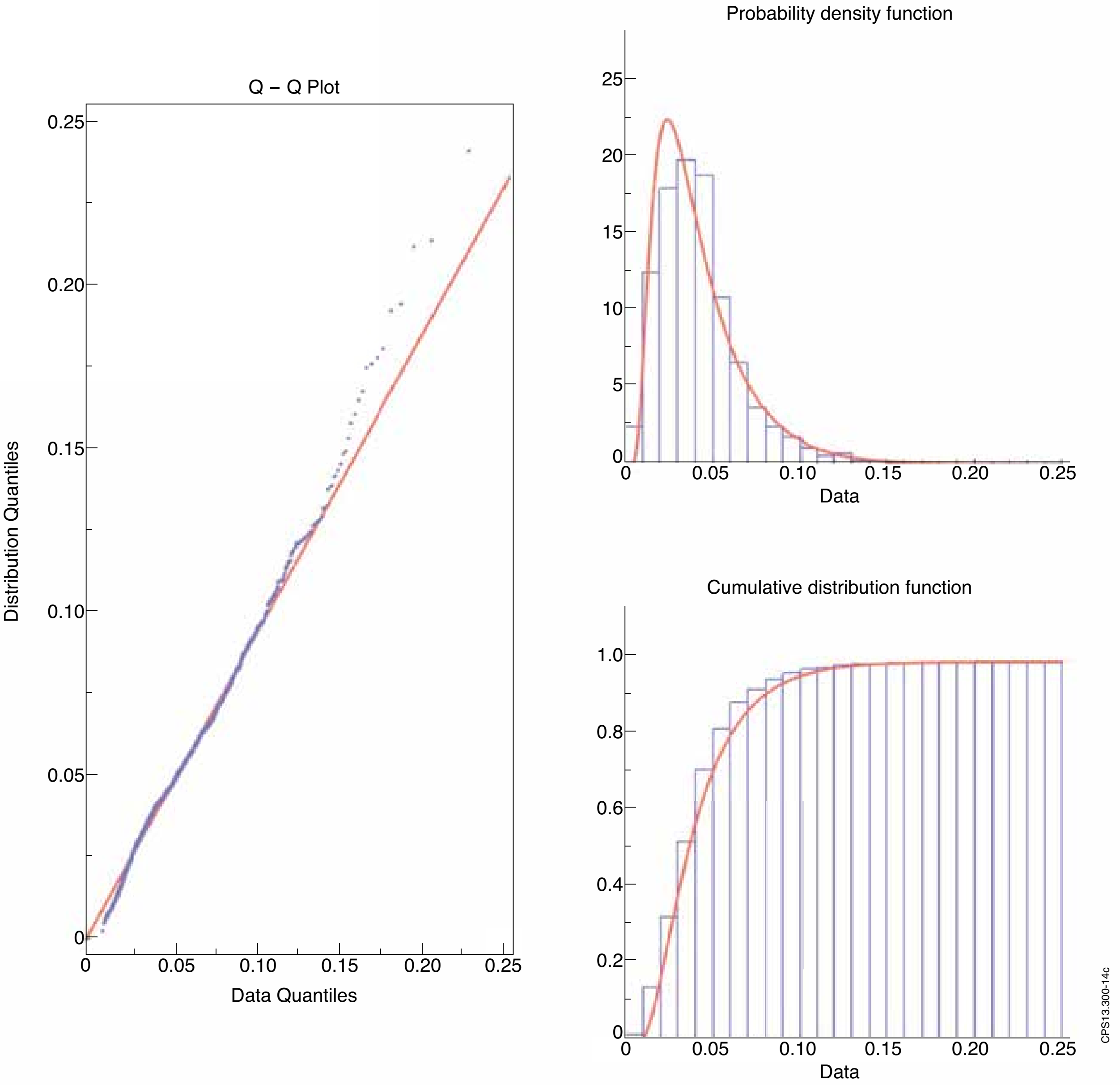


Dagum

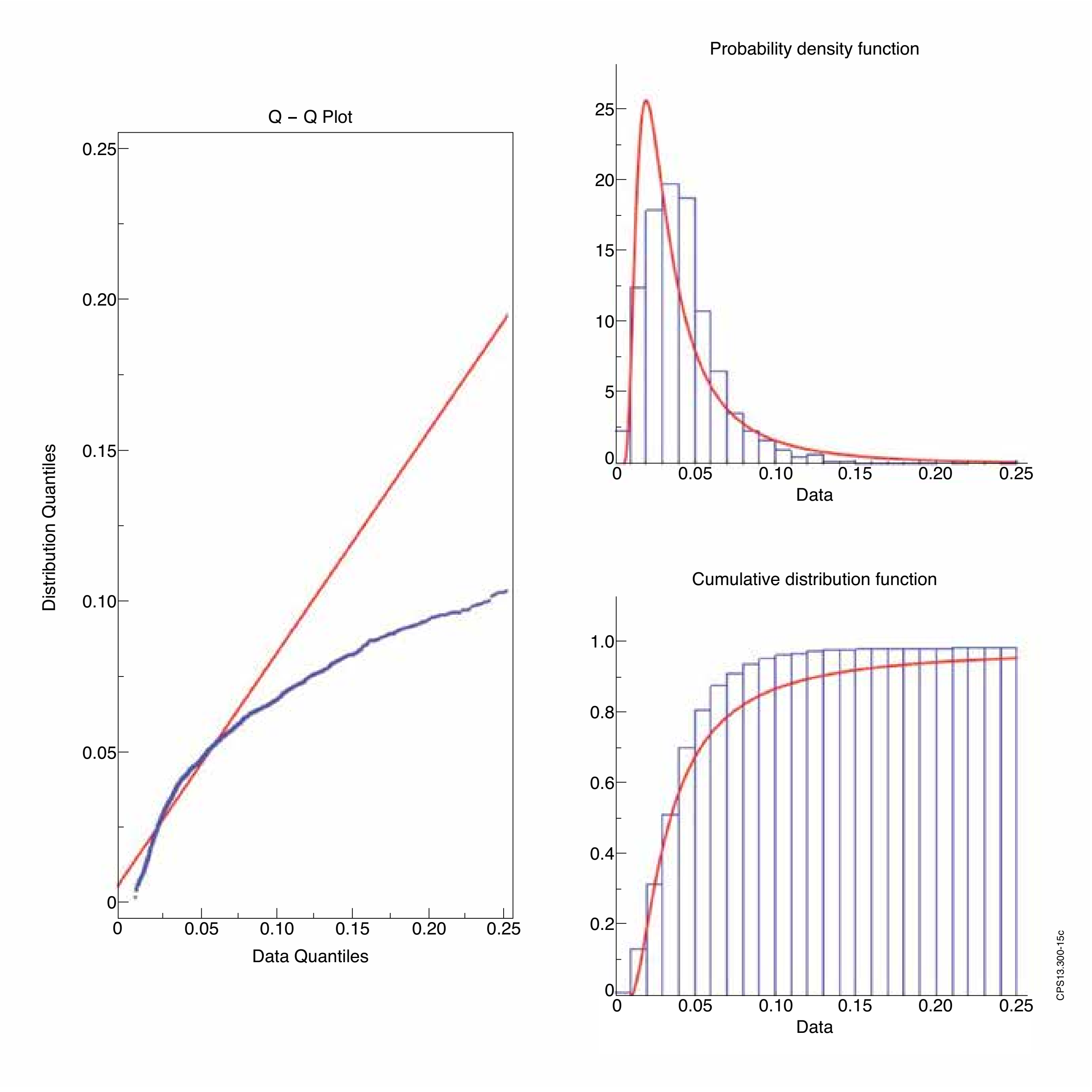


CPS13.300-13c

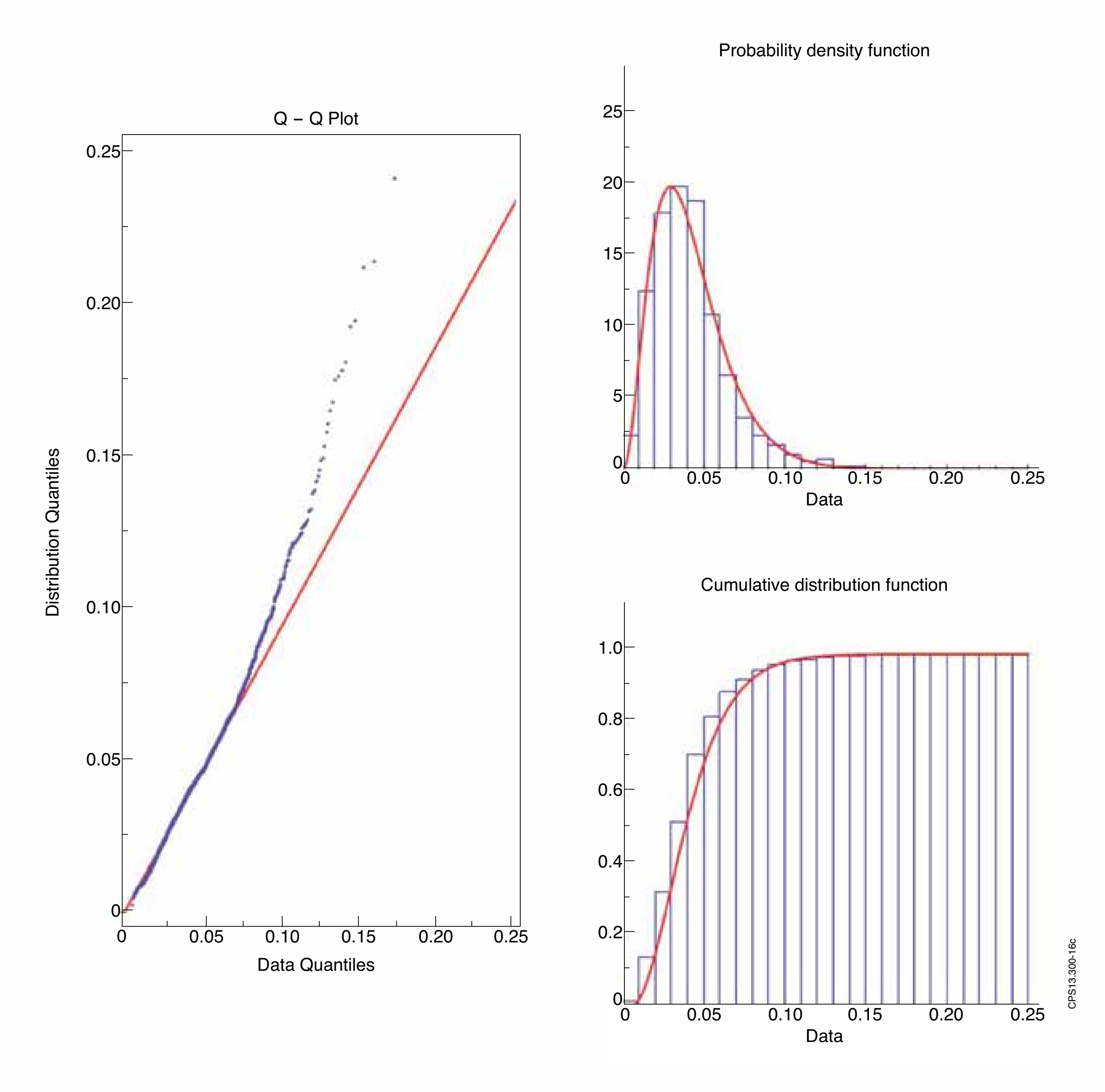
Fatigue Life



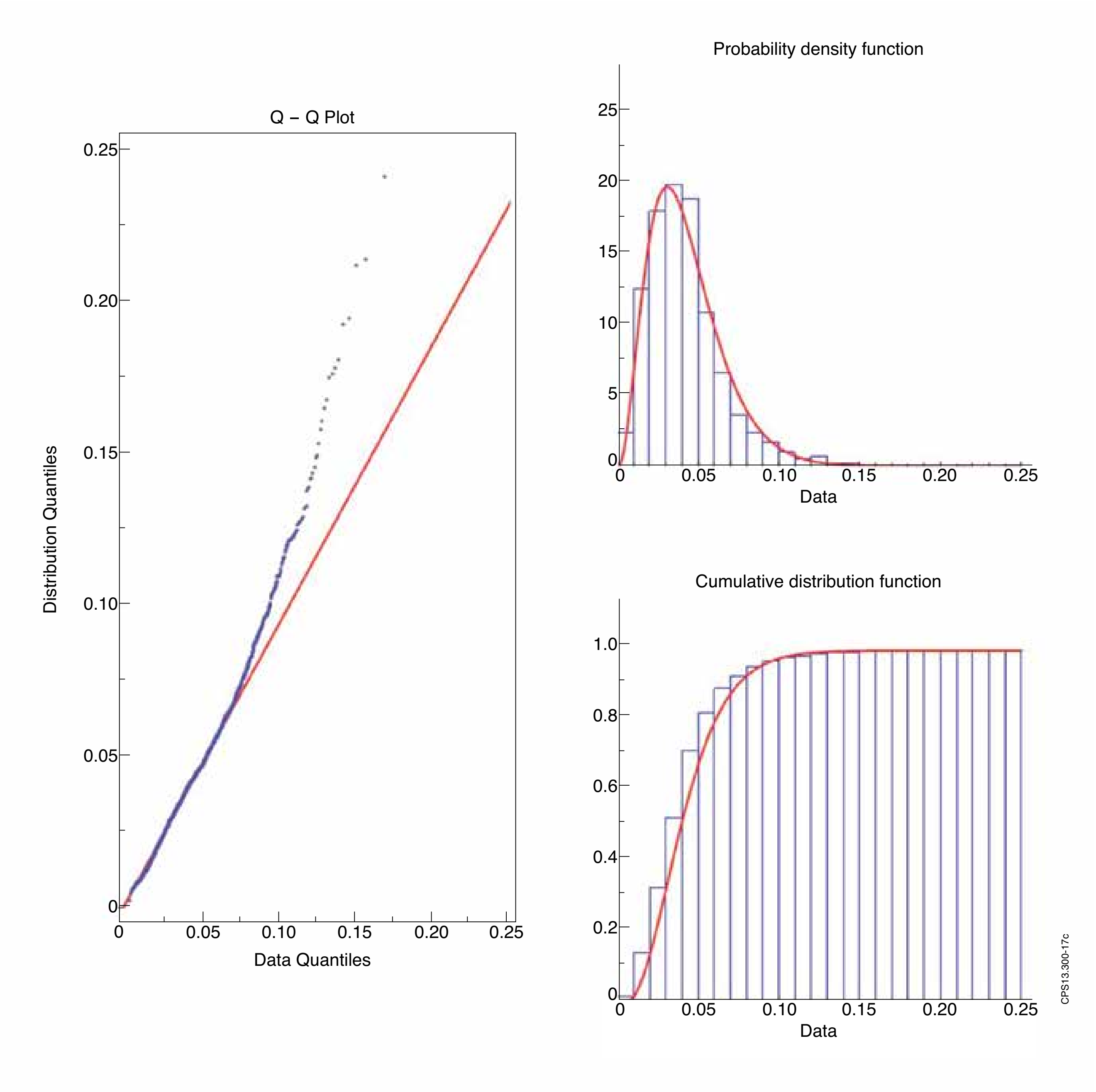
Frechet


Q – Q Plot
Distribution Quantiles
Data Quantiles
0
0.05
0.10
0.15
0.20
0.25
Probability density function
0
5
10
15
20
25
Data
Cumulative distribution function
0
0.2
0.4
0.6
0.8
1.0
CPS13.300-15c

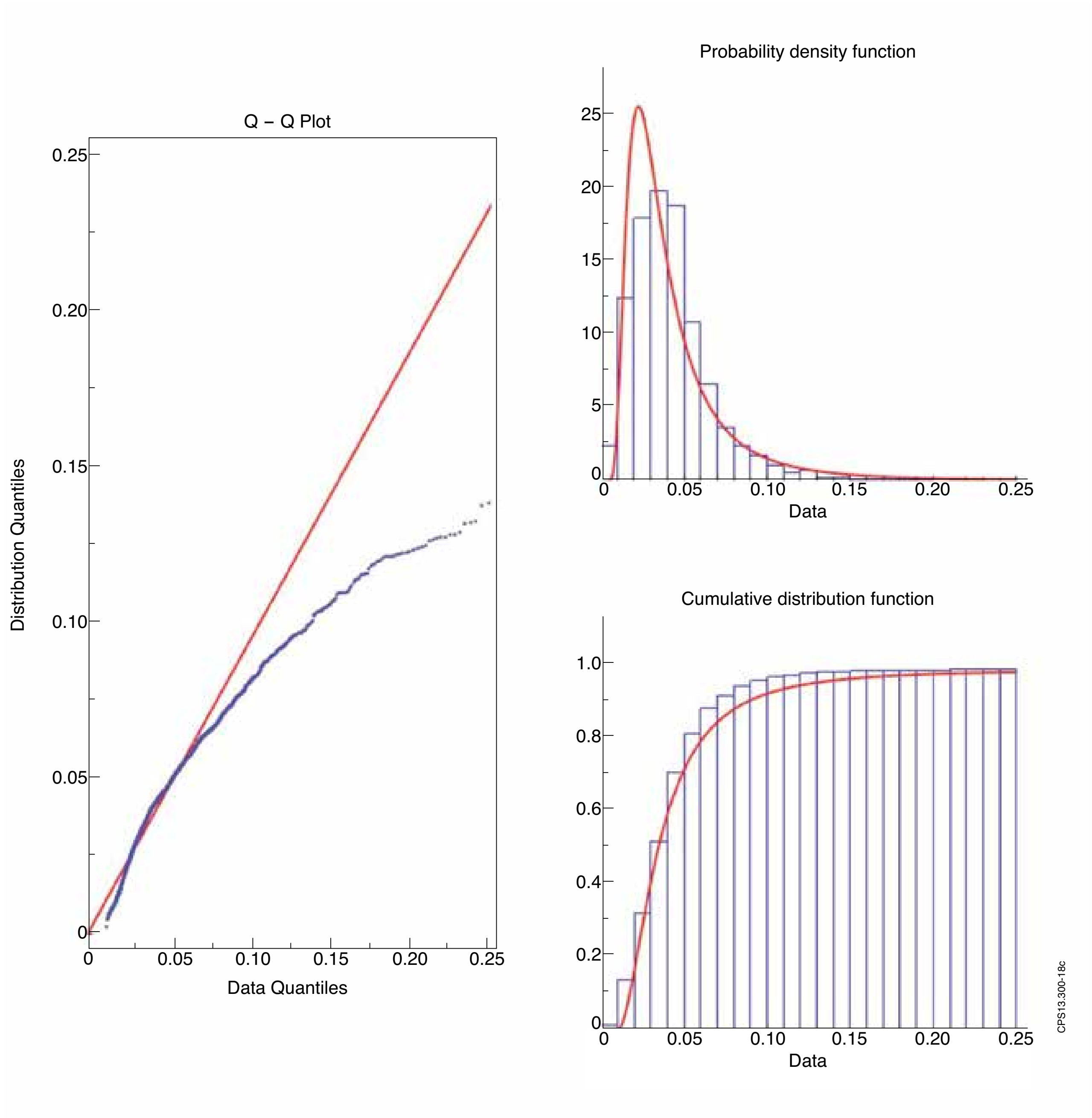
Gamma


Q – Q Plot
Distribution Quantiles
Data Quantiles
0
0.05
0.10
0.15
0.20
0.25
Probability density function
0
5
10
15
20
25
Data
Cumulative distribution function
0
0.2
0.4
0.6
0.8
1.0
Data
CPS13.300-16c

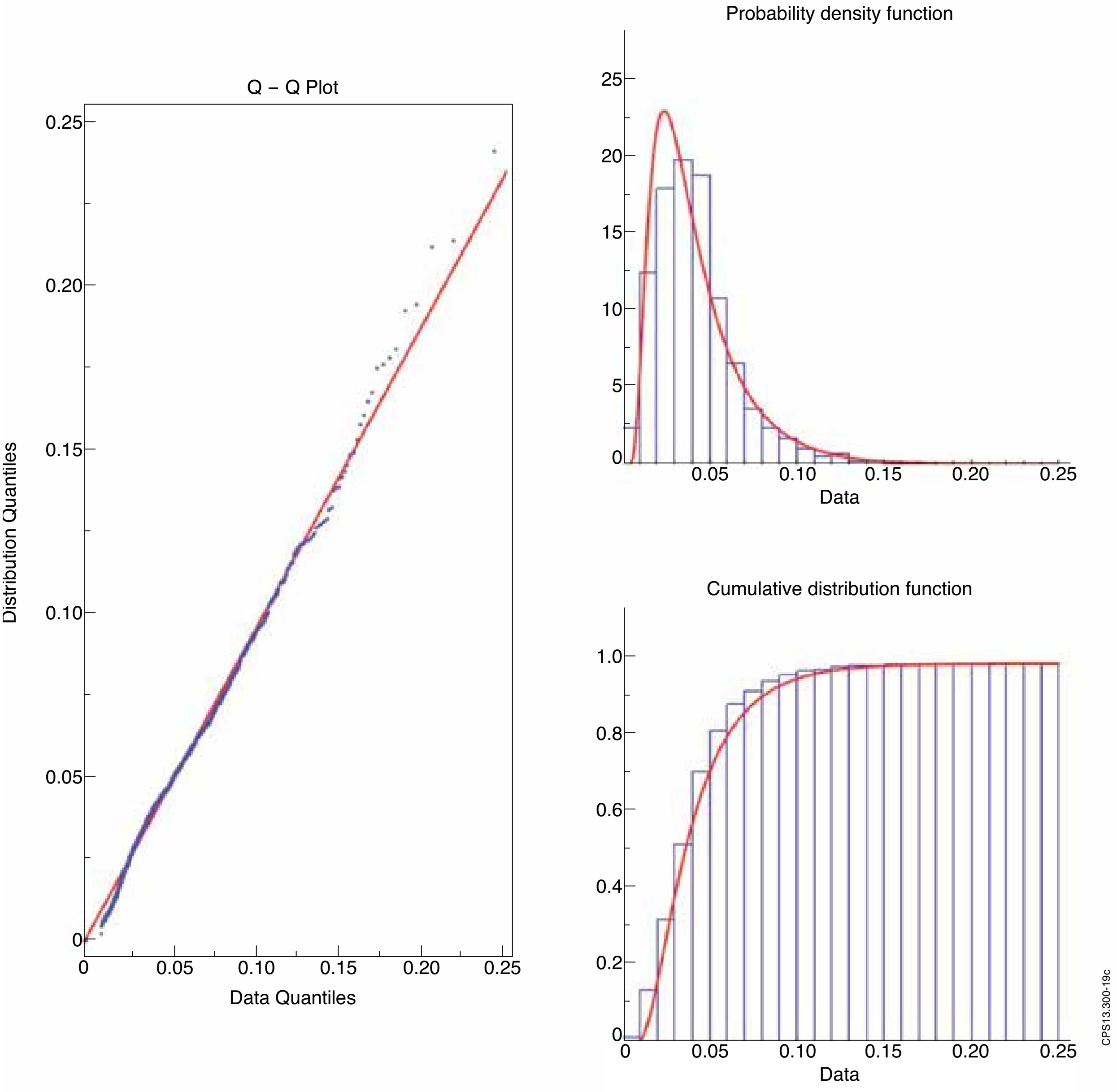
Gentralised Gamma



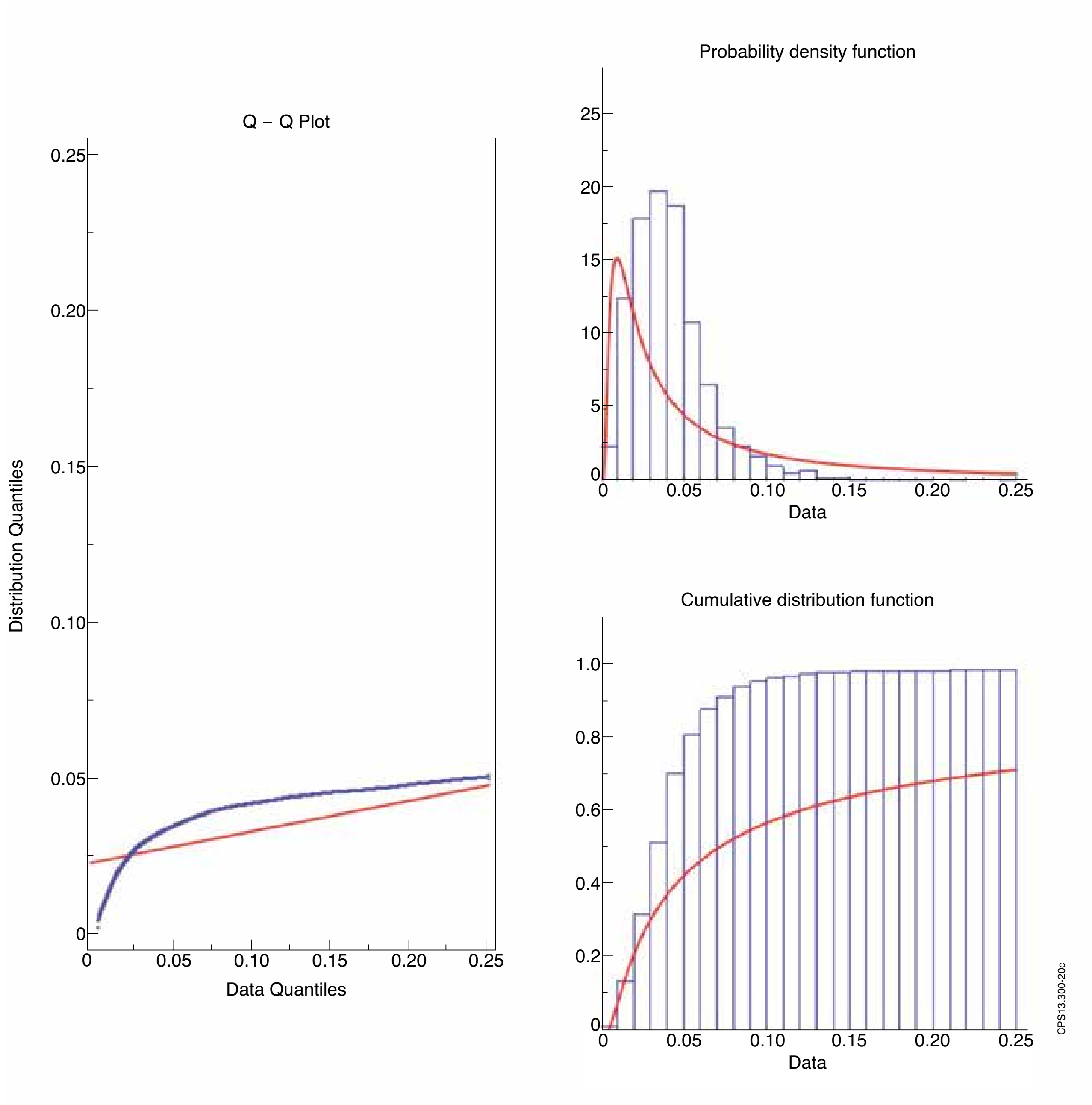
Inverse Gamma



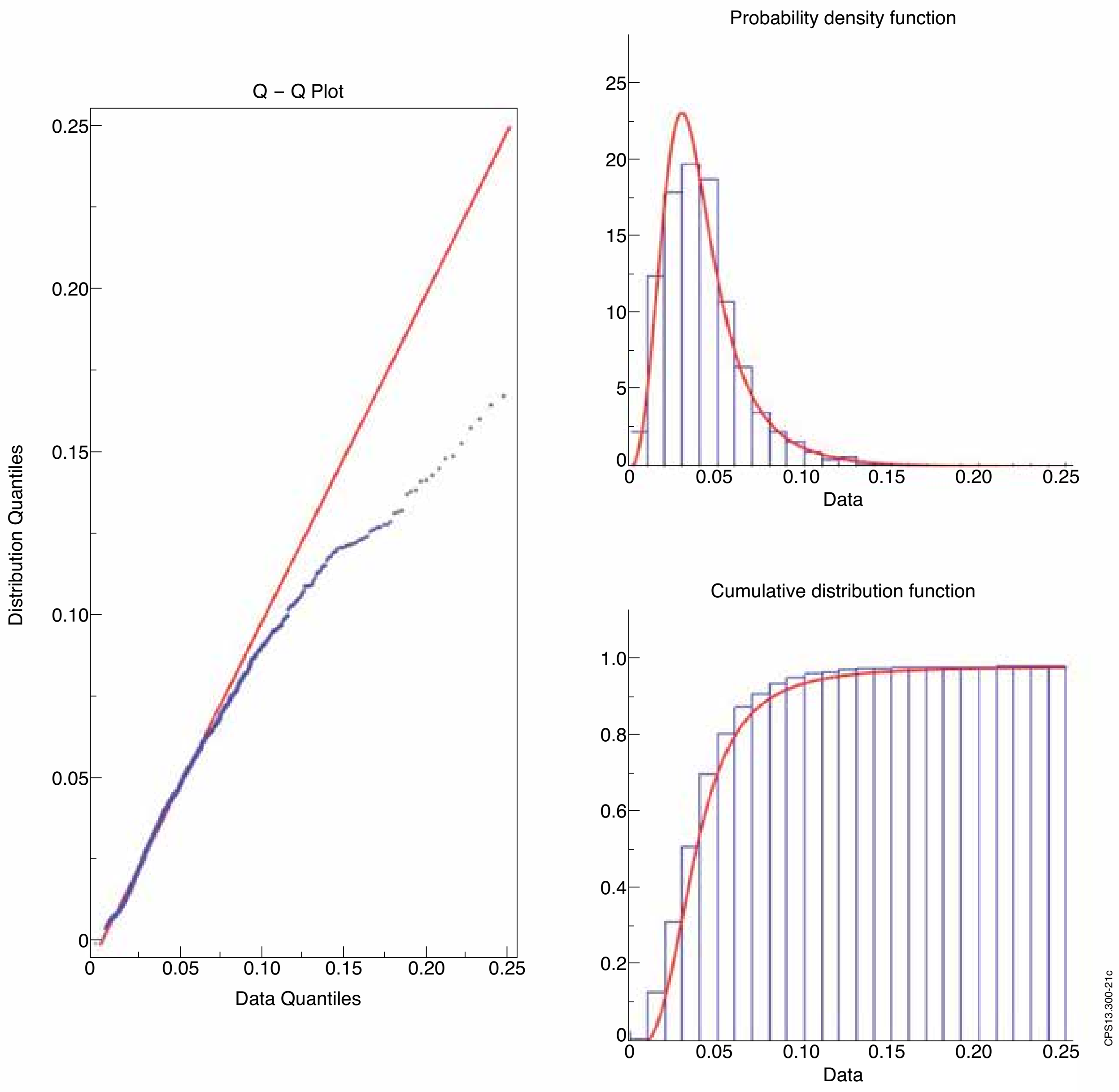
Inverse Gaussin



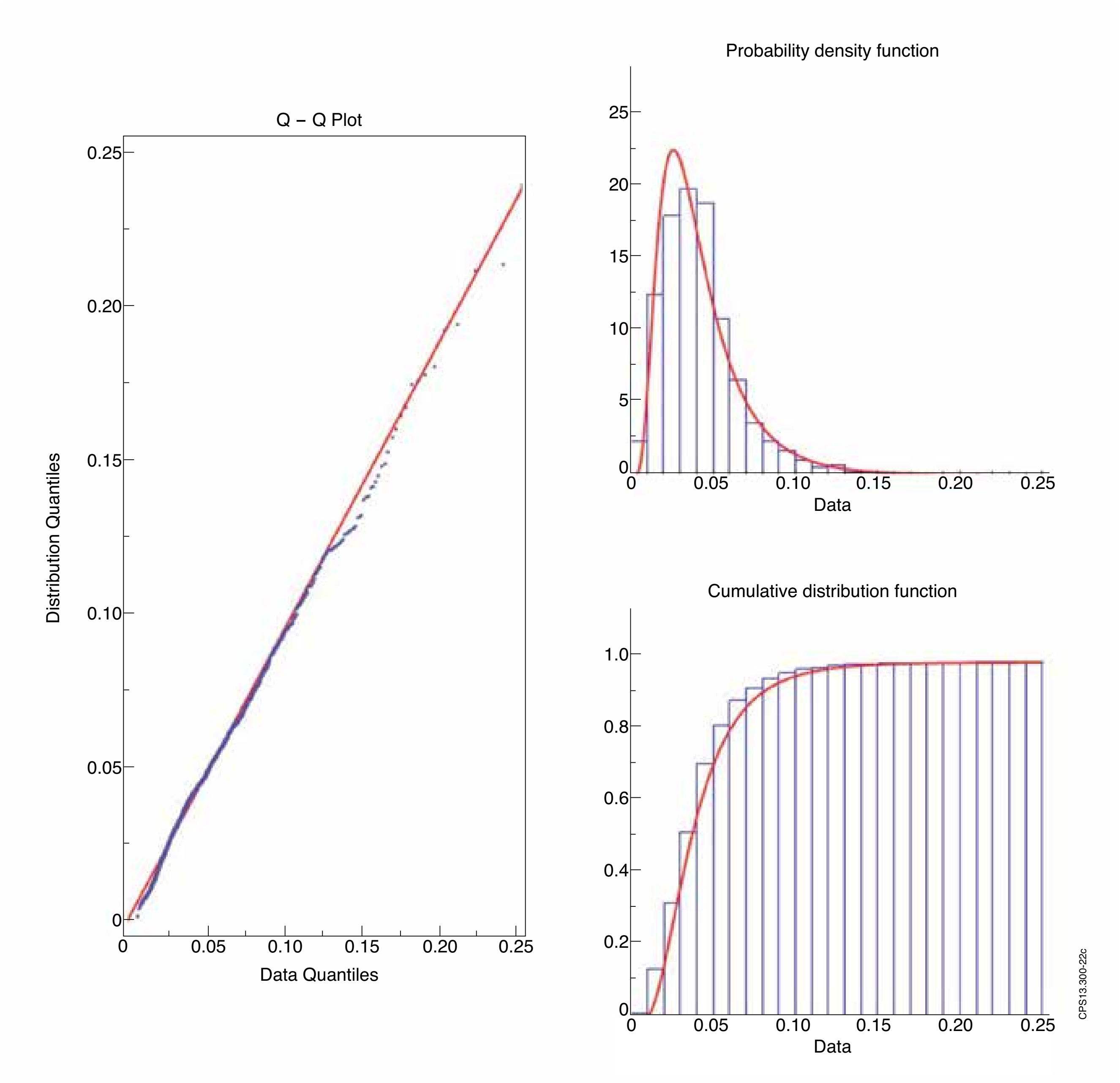
Levy


Q – Q Plot
Distribution Quantiles
0.25
0.20
0.15
0.10
0.05
0
0
0.05
0.10
0.15
0.20
0.25
Data Quantiles
Probability density function
25
20
15
10
5
0
0
0.05
0.10
0.15
0.20
0.25
Data
Cumulative distribution function
1.0
0.8
0.6
0.4
0.2
0
0
0.05
0.10
0.15
0.20
0.25
Data
CPS13.300-20c

Log-Logistic



Log-Normal

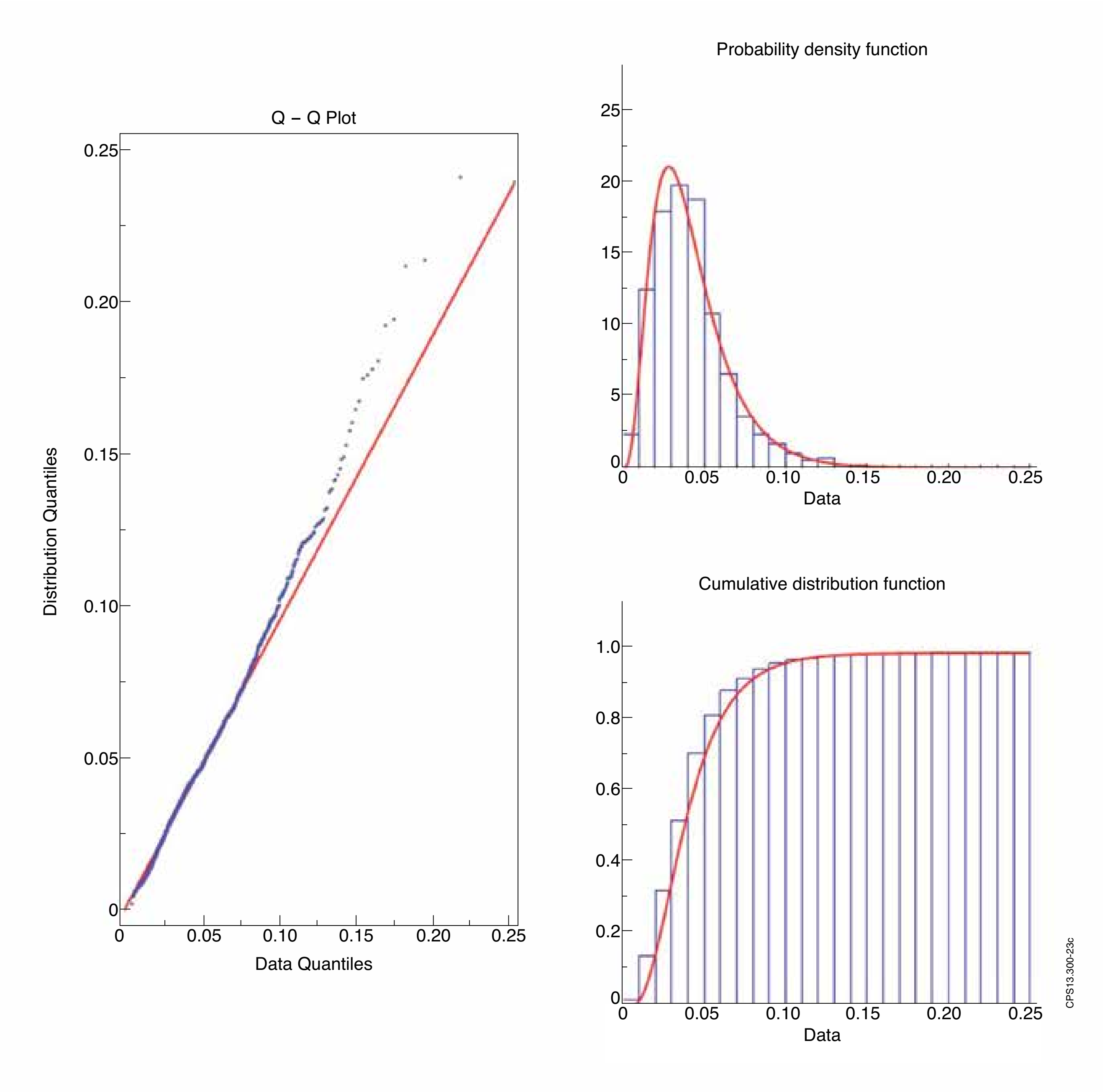


Q – Q Plot

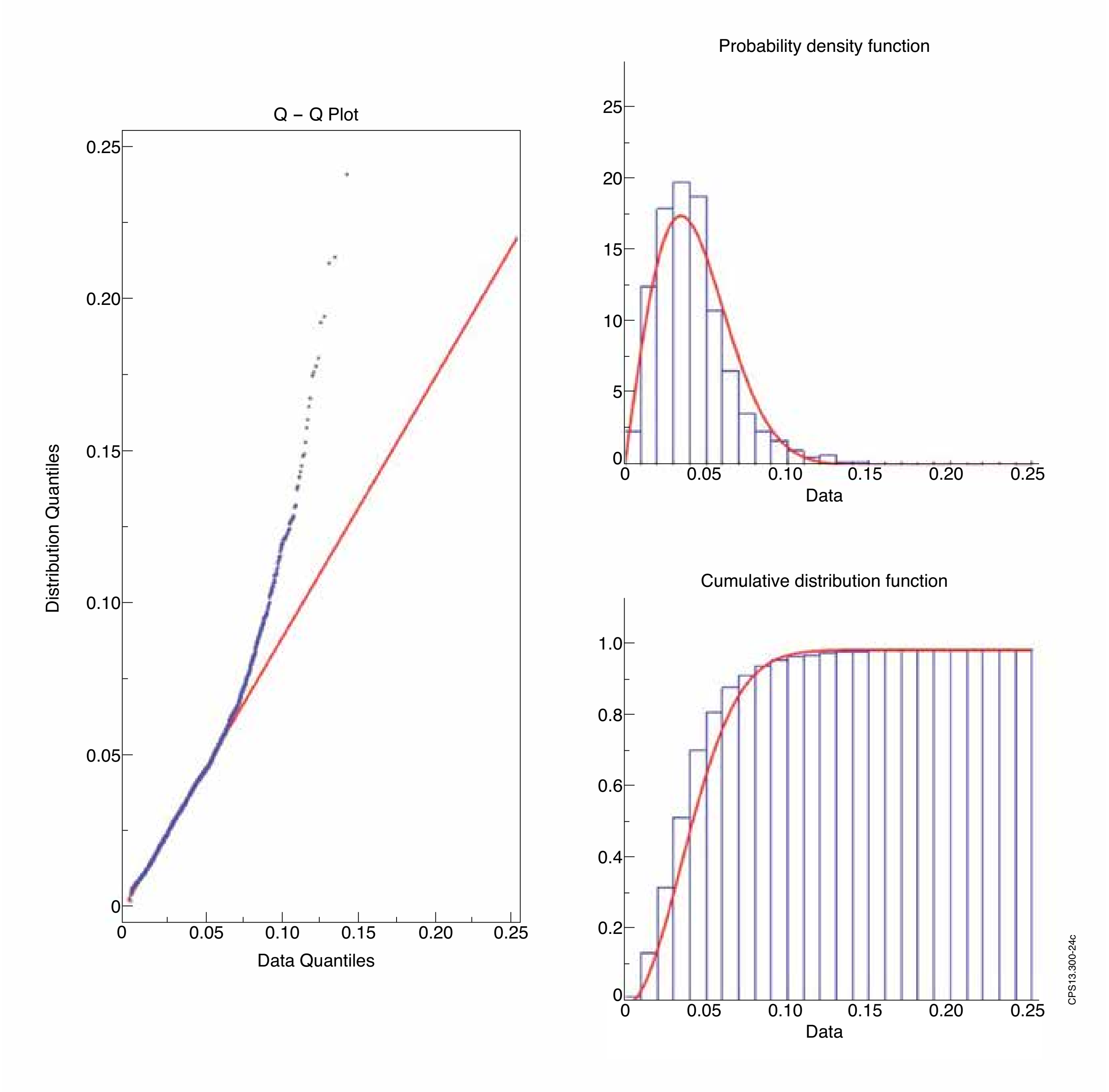
Probability density function

Cumulative distribution function

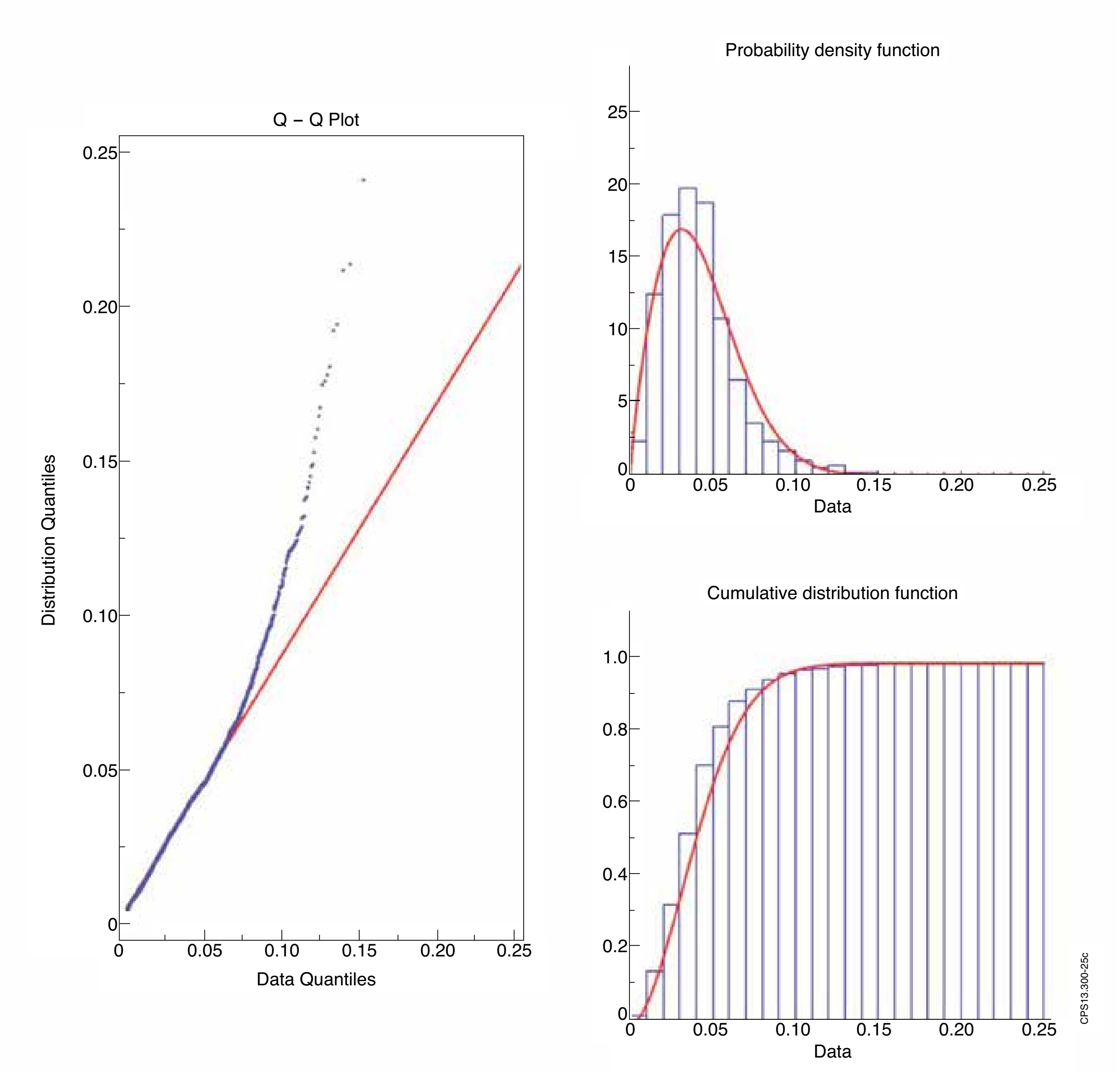
Pearson V1


Q − Q Plot
Distribution Quantiles
Data Quantiles
0.25
0.20
0.15
0.10
0.05
0
0
0.05
0.10
0.15
0.20
0.25
Probability density function
25
20
15
10
5
0
Data
Cumulative distribution function
1.0
0.8
0.6
0.4
0.2
0
Data
CPS13.300-23c

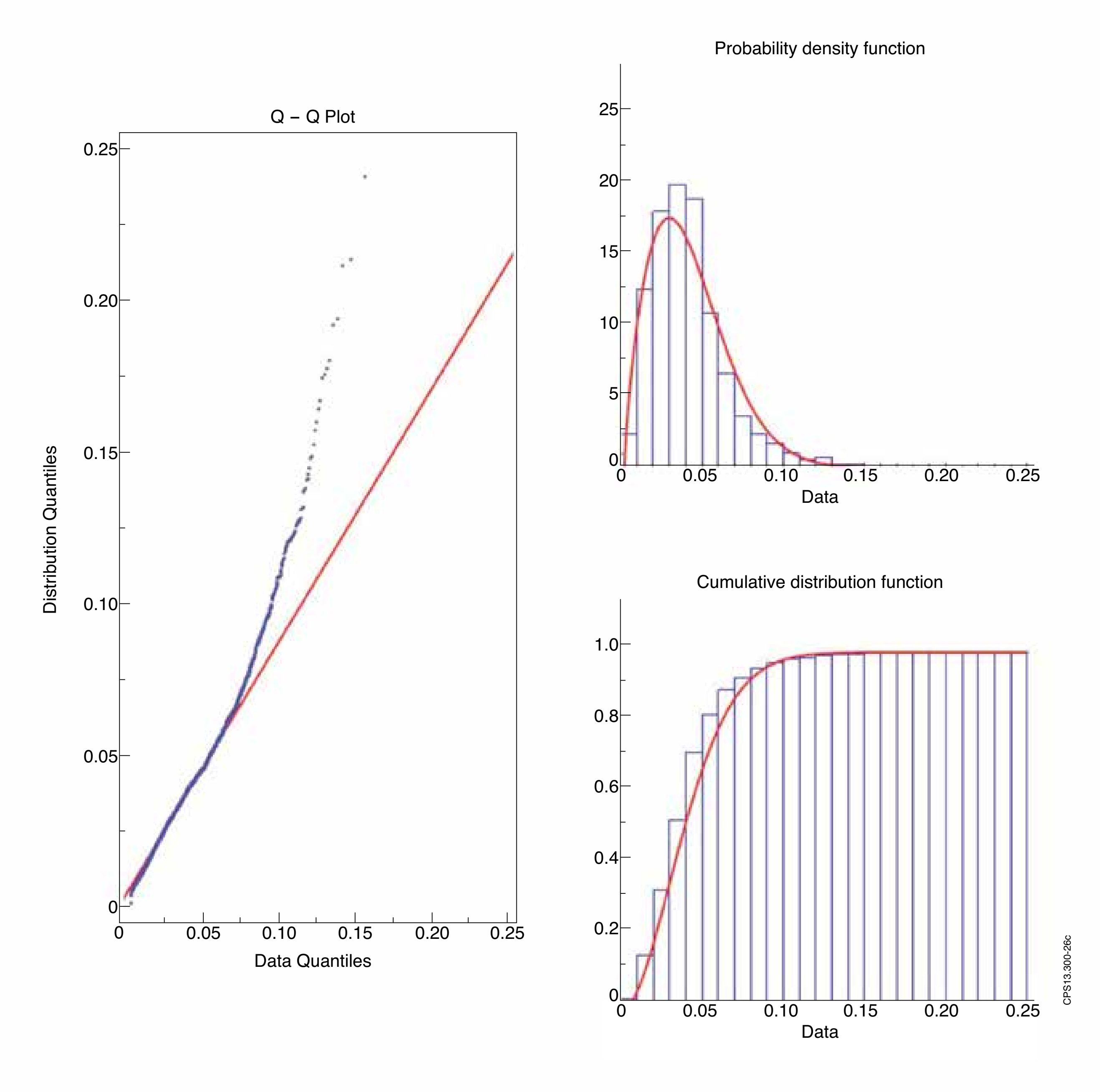
Rayyleigh


Q – Q Plot
Distribution Quantiles
Data Quantiles
0
0.05
0.10
0.15
0.20
0.25
Probability density function
0
5
10
15
20
25
Data
Cumulative distribution function
0.2
0.4
0.6
0.8
1.0
CPS13.300-24c

Weibull


Q – Q Plot
Distribution Quantiles
Data Quantiles
0
0.05
0.10
0.15
0.20
0.25
Probability density function
25
20
15
10
5
0
Data
Cumulative distribution function
1.0
0.8
0.6
0.4
0.2
Data
CPS13.300-25c

Weibull (3P)


Q – Q Plot
Distribution Quantiles
0.25
0.20
0.15
0.10
0.05
0
0
0.05
0.10
0.15
0.20
0.25
Data Quantiles
Probability density function
25
20
15
10
5
0
0
0.05
0.10
0.15
0.20
0.25
Data
Cumulative distribution function
1.0
0.8
0.6
0.4
0.2
0
0
0.05
0.10
0.15
0.20
0.25
Data
CPS13.300-26c

APPENDIX 3

List of Pulses for steady state analysis

Pulse	t_0 (s)	t_f (s)	Pulse	t_0 (s)	t_f (s)
73397	47.6	50.6	75727	62.1	63.5
73445	58.4	60.9	75728	62.1	65.3
73446	58.5	61.6	75731	62.2	66.3
73447	58.5	60.3	75732	64	65.9
73450	58.6	61.1	76428	63.2	67.9
73484	47.6	52.9	76430	65	67.9
74130	58.5	63	76431	63.3	66.9
74364	57.1	60.8	76437	63	67.7
74365	57.2	59.1	76438	64.8	67.7
74366	57.2	59.6	76440	63.7	67.9
74367	57.2	58.6	76443	64	67.7
74368	57.2	58.3	76470	60.1	61
74369	57.6	60	76471	57.7	58.4
74371	57.2	59.1	76472	57.8	59
74372	57.7	59.4	76473	58.2	59.4
74373	57.2	59.5	76474	57	58.7
74374	57.1	59.2	76475	57.6	60.2
74375	57.1	60.2	76476	56.6	58
74376	59.1	61.4	76477	60.5	61.3
74378	59.2	67.5	76478	58.5	61.2
74443	57.3	59	76479	57.7	58.2
74444	57.4	58.9	76480	57.9	58.6
74612	51.5	54.7	76481	59.1	60.9
74613	51.6	54.4	76812	62.2	64.5
74793	58.2	61.9	77073	54.4	57
74795	58.7	62.9	77192	63.7	67.9
74798	62.5	65.8	78448	54.4	57
75118	64.5	65.6	78750	61	63.7
75724	66.8	68.7	79389	57.6	59.8
75726	62.2	64	79546	63.3	67