



HAL
open science

CNTb, a set of scripts for batch processing and statistical analysis of photon correlation spectroscopy data via CONTIN inversion

Ivan Echavarri Franco, Jérôme Combet, François Schosseler

► **To cite this version:**

Ivan Echavarri Franco, Jérôme Combet, François Schosseler. CNTb, a set of scripts for batch processing and statistical analysis of photon correlation spectroscopy data via CONTIN inversion. 2011. hal-00665367

HAL Id: hal-00665367

<https://hal.science/hal-00665367>

Preprint submitted on 1 Feb 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CNTb, a set of scripts for batch processing and statistical analysis of photon correlation spectroscopy data via CONTIN inversion

I. Echavarri Franco, J. Combet, and F. Schosseler

Institut Charles Sadron, 23 rue du Loess, 67034 Strasbourg Cedex 2, France

Abstract:

We introduce CNTb, a set of scripts allowing a batch inversion of series of photon correlation spectroscopy data files via CONTIN with a single command line. In addition CNTb allows the fitting of the resulting distribution of relaxation times by a set of log-normal distributions and calculates mean values and standard deviations for their parameters over the data set. CNTb proves to be a very useful tool when a large number of experiments is needed to test the reproducibility and the reliability of CONTIN output in difficult experimental conditions, e.g., when the signal to noise ratio is low.

1. Introduction

Since the early developments of quasi-elastic light scattering nearly fifty years ago (Berne & Pecora, 1976), this technique has become an ubiquitous tool for the characterization of colloids and macromolecules in solution (Brown, 1996). In its modern version, it is based on the analysis of the temporal fluctuations of the light intensity scattered by the solution. These fluctuations are ideally dominated by the fluctuations in solute concentration generated by the brownian motion of scatterers in the solution. It is therefore possible to relate the temporal decay of the intensity-intensity correlation function (ICF) to the diffusion coefficient of the solutes and thus to their mean hydrodynamic size. In the case of solutes with a well-defined size, the correlation function exhibit a single exponential decay that yields easily the corresponding size. In cases where the scatterers are characterized by a size distribution, the analysis is however more intricate since the presence of noise in the signal makes the data inversion an ill-posed problem. Thirty years ago, Provencher proposed a constrained regularization method to analyze the noisy correlation functions measured in photon correlation spectroscopy (PCS) experiments: the so-called CONTIN program, written in FORTRAN language, was offered to the scientific community as a valuable tool for the analysis of the experimental ICFs (Provencher, 1978, 1979, 1982a, 1982b). Despite many alternative propositions (see, e.g., Zhu *et al.*, 2010), it remains one of the most popular methods for the inversion of PCS data and is included as a standard tool built in most proprietary data acquisition and analysis softwares provided with modern commercial correlator boards.

However several drawbacks are associated with the implementation of CONTIN analysis in these commercial softwares: i) data treatment must be done with the computer that contains the correlator board, which prevents access to the experimental setup for other users, ii) parameters for the data analysis are entered manually via keyboard typing and mouse clicking, which become cumbersome when the analysis of many data files is needed; iii) many useful output informations provided by CONTIN are not available and the analysis process tends to become a black box for many users.

The purpose of this short paper is to introduce CNTb (short name for CONTIN batch), a set of scripts that allows a batch processing of series of PCS data files via CONTIN with a

single command line. This allows one to overcome the drawbacks mentioned above. In addition, for series of data files measured in the same conditions, a statistical analysis is performed: mean values and standard deviations are calculated from the output values obtained for each data file and compared to the results obtained for the ensemble-averaged ICFs to check consistency. CNTb is particularly useful when the experimental noise results in large variations in CONTIN analysis outputs and the experimentalist is lost in choosing how to analyze the data.

The paper is organized as follows. We give first some technical details about CNTb, then we introduce a typical difficult system, give the experimental details and describe the statistical analysis of a set of 30 repetitive measurements performed with CNTb.

2. Technical details

CNTb is a combination of UNIX shell scripts that loops for CONTIN execution, GNUPLOT data plotting and fitting, and PERL scripts execution. The latter are used to read the original ASCII data files generated by the correlator board, to format them in the appropriate FORTRAN form required to feed CONTIN, and to generate POSTSCRIPT code that groups similar GNUPLOT PS outputs in a single PDF document: this graphical output allows a comparison of the set of experiments in a single glance. The input parameters for CONTIN analysis are typed in a text file placed in the same directory as the processed data files. A typical command line to launch the batch process is then "cnt directory_name a b", where the parameters "a" and "b" specify how the data are processed. Sets of directories can be processed as well with two levels of nesting.

Since all the steps are performed with open source softwares freely available for UNIX, OS X (used here) and Windows operating systems, CNTb should run smoothly independently of the operating system. However, Windows users need to translate UNIX commands in appropriate Windows command lines: this can be done at no cost via an interface software. CNTb is open source and can be downloaded free of charge as a ZIP archive at <http://www-ics.u-strasbg.fr/CNTb>.

The original CONTIN software is highly versatile and can process in a number of customized

ways ICFs, $g^{(2)}(t)$,

$$g^{(2)}(t) = \frac{\langle I(t')I(t+t') \rangle}{\langle I(t') \rangle^2}, \quad (1)$$

as well as field-field correlation functions $g^{(1)}(t)$,

$$g^{(1)}(t) = \frac{\langle E(t')E^*(t+t') \rangle}{\langle I(t') \rangle}, \quad (2)$$

which are related by Siegert relationship,

$$g^{(2)}(t) = 1 + \beta^2 |g^{(1)}(t)|^2, \quad (3)$$

In these expressions $E(t)$ is the electromagnetic field scattered at time t , $I(t) = E(t)E^*(t)$ the corresponding intensity, β^2 a numerical factor depending on the geometry of detection and $\langle \dots \rangle$ denotes a time average. In the case of polydisperse systems, the user is typically searching for the distribution of relaxation times $R(\tau)$ defined through:

$$g^{(2)}(t) = \int_0^{+\infty} R(\tau) \exp\left(-\frac{t}{\tau}\right) d\tau, \quad (4)$$

and CONTIN then solves for the discretized pseudo-distribution of relaxation times (RTD) $P(\tau_i)$, on a number of N_g gridpoint values in a chosen interval:

$$g^{(2)}(t) = \sum_{i=1}^{N_g} P(\tau_i) \exp\left(-\frac{t}{\tau_i}\right), \quad (5)$$

using a regularization that includes a priori knowledge on the solution and parsimony principle (Provencher, 1978, 1979, 1982a, 1982b). Alternately, CONTIN can also solve for a distribution of relaxation rates or a distribution of radii. Here, for the convenience of plotting data and $P(\tau_i)$ values on the same graphs, we chose to use the time domain representation for the sample polydispersity.

The standard procedure for data analysis is the following. After importation of the ASCII data files where needed, the analysis is run once the usual parameters have been specified in the parameters file: first and last experimental points to consider, interval of relaxation

times to be used, number of grid points in this interval, fixed or free baseline, type of data weighting in the sum of squared residuals to minimize. The user specifies also which initial data preprocessing should be done, e.g., if the analysis is performed on $g^{(2)}(t)$, $g^{(2)}(t)-1$, or $(g^{(2)}(t)-1)^{1/2}$. Moreover the full versatility of the original CONTIN is restored since any of the input parameters controlling how CONTIN executes (Provencher, 1982a, 1982b) can be set to a non-default value in the formatted input files by adding a single line code in the appropriate PERL script.

After a few seconds, all data files have been processed and a PDF file containing the results is created (see example in Supplementary Information). The first page contains composite plots displaying the original data points, the fitted data points and the resulting CONTIN distribution on semi-logarithmic scales, the second page contains the same information plotted on log-log scales, the third page groups the traces of the light intensities recorded during the experiments, the fourth page displays the RTDs as well as the corresponding fits using log-normal distributions, and the fifth one the repartition of weighted residuals.

Most importantly, CNTb calculates an ensemble-averaged ICF from the N data files as (Rouf-George *et al.*, 1997):

$$g_{av}^{(2)}(t) = \frac{\sum_{i=1}^N \langle I_i(t) \rangle^2 g_i^{(2)}(t)}{\sum_{i=1}^N \langle I_i(t) \rangle^2}, \quad (6)$$

and performs on it the same analysis as on the individual data files. CONTIN results obtained from this average correlation function are less sensitive to the experimental noise and help to discriminate which experiments should be discarded in the next step. They also help the user to decide how many log-normal distributions should be used to describe the RTD. In the second step, discarded data files are removed from the directory and CNTb is executed again after the user has specified in an additional text file the initial trial values for the position, amplitude and width of each log-normal distribution. Then averaged value and standard deviation are calculated for each parameter and saved in a text file. The fit of the RTD by log-normal distributions is done by the nonlinear least-squares Marquardt-Levenberg algorithm implemented in GNU PLOT, which is quite efficient in deconvoluting overlapping

peaks for well chosen trial values as is illustrated below.

2. Example

The purpose of our study was to follow the complexation of conjugated polyelectrolytes by oppositely charged surfactants in very dilute solutions. Here we will give only the results obtained on the pure polyelectrolyte solutions to illustrate the performance of the method.

Experimental ICFs were measured at a fixed scattering angle $\theta = 90^\circ$ with the ALV/DLS/SLS-5020F experimental setup (ALV Laser Vertriebsgesellschaft mbH, Langen, Germany), consisting in a He-Ne laser (22 mW, $\lambda_0 = 632.8\text{nm}$), a compact ALV/CGS-8 goniometer system, and an ALV-5000/E multi-tau correlator. The optical detection of this instrument is optimized for a nearly ideal detection on one coherence area and the β^2 factor in Siegert relation is close to 1 when measurements are performed on a standard latex solution. Temperature was regulated to $25.2 \pm 0.1^\circ\text{C}$ and 30 ICFs were accumulated overnight with an acquisition time of 600s for each one.

The polymer sample is a non-regioregular poly(3-thiophene acetic acid) dissolved in a buffered aqueous solution ($\text{NaHCO}_3/\text{Na}_2(\text{CO}_3)_2$, pH= 9.2, ionic strength 72 mM). In these conditions, the polyelectrolyte chains reach their maximum charge density fixed by Oosawa-Manning condensation (Oosawa, 1968; Manning, 1969) and the interchain electrostatic interactions are efficiently screened at the very small polymer concentration used (0.176 mM repeating units). The polymer chains have a mean polymerization degree about 125 with a size polydispersity about 2.2 (Vallat *et al.*, 2006). The resulting scattering intensity is very low, about 2.75 ± 0.03 kHz, and very close to the intensity scattered from the buffer solvent, about 1.5 ± 0.1 kHz. The dark noise of the photodiode is about 120 Hz. These are very challenging conditions for PCS experiments since the intensity fluctuations linked to the density fluctuations in the solvent give a strong contribution to the ICF decay in a time domain smaller than 0.2 microsecond that is not measurable with the correlator.

Figure 1 displays composite plots for two typical individual and for the ensemble-averaged ICFs. Due to the very low intensity scattered from the polymer chains, the amplitude of the

ICFs is far from the theoretical value that could be expected with our optical detection, about 1.2 instead of 1.9. For the same reason, the statistical noise on the data points at short times does not decrease for longer acquisition times and the results obtained from CONTIN can be rather sensitive to this noise. Our batch processing allows us to explore easily the influence of input parameters feeded to CONTIN on the resulting RTDs (Echavarri Franco, 2011). Here we show results obtained when executing CONTIN on $(g^{(2)}(t)-1)^{1/2}$ data with optimized input parameters as follows: first experimental point at 1 microsecond, last experimental point at 1s, 200 grid points equally spaced on a logarithmic scale between 1 microsecond and 0.01 s, free baseline, unweighted residuals.

At this point it is necessary to comment on the first peak in the RTDs appearing for delay times shorter than 10 microseconds. In our experimental conditions, a relaxation time of 1 microsecond corresponds to the diffusive motion of a scatterer with an hydrodynamic radius about 0.1 nm. Therefore this peak arises from the initial fast decay of the ICFs due to the solvent contribution mentioned above. Although it cannot be resolved satisfactorily due to the limited range of time delays available with the correlator, it is critical to allow for relaxation times in this domain for CONTIN analysis. We checked that better results were obtained this way rather than by starting the analysis at time delays larger than 10 microseconds where the diffusive motion of polymer chains contribute more significantly to the decay of the ICFs.

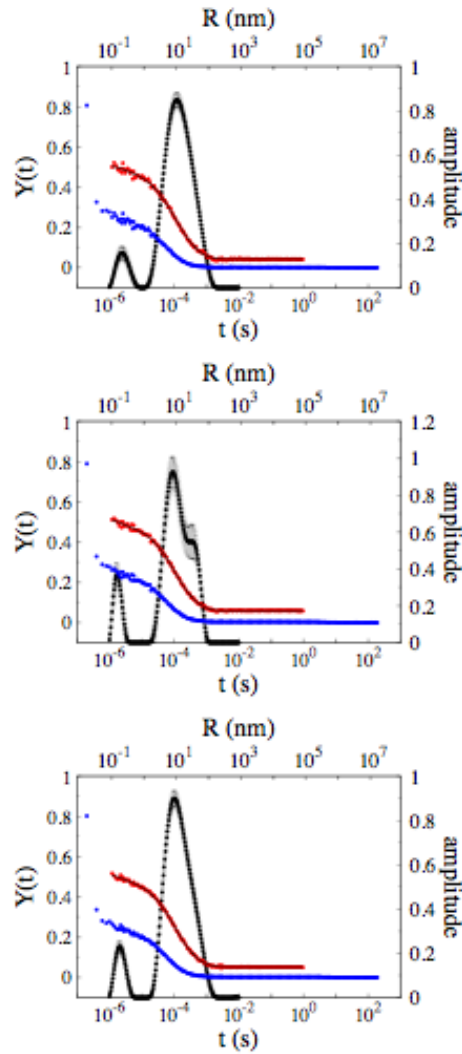


Figure 1: Typical composite plots for two individual measurements (top and middle) and for the ensemble-averaged ICF showing: $g^{(2)}(t) - 1$ initial values (blue), $(g^{(2)}(t) - 1)^{1/2}$ values used for the analysis by CONTIN (red), resulting RTD values (black) and fitting curve $(g^{(2)}(t) - 1)^{1/2}$ (solid line). The upper x-axis gives the equivalent hydrodynamic radius for hard spheres with the same diffusion time in the same experimental conditions. The error bars are those calculated by CONTIN. This information is usually lost by the proprietary softwares associated with commercial correlators.

Apart from the peak linked to the solvent contribution, individual RTDs produced by CONTIN appear either as a broad single symmetric peak or as a superposition of two more or less separated peaks. This variation shows the critical influence of the statistical experimental noise on the resulting RTDs. On the other hand, the RTD obtained for the ensemble-averaged correlation function shows that the most likely solution is a broad asymmetric peak with a shoulder on the larger times side. Therefore this distribution should be better described as the superposition of two log-normal distributions and CNTb is executed a second time while

specifying trial values for these two peaks. Figure 2 shows the resulting fits of the RTDs presented in Fig.1 by the sum of two log-normal distributions with amplitude A_i , position τ_i and width W_i ($i=1,2$).

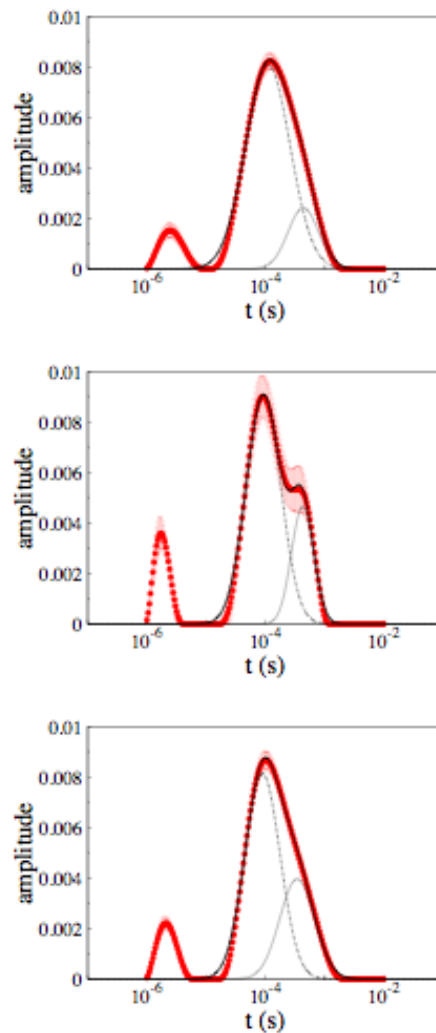


Figure 2: Same RTDs (red) as in Fig. 1 fitted by the sum (solid line) of two log-normal distributions (dotted lines).

As shown in the supplementary material, six out of the thirty RTDs differ qualitatively from the RTD obtained with the ensemble-averaged ICF, in particular because the respective amplitudes of the two fitted log-normal distributions are inversed. Therefore it does not make much sense to include them in a statistical analysis. At this step CNTb allows the user to remove very easily those atypical results from the statistical analysis. In Table 1 are listed the mean value and standard deviation for S_i et τ_i , calculated on the set of 24 typical individual values together with the corresponding values obtained for the ensemble-averaged ICF. Here S_i is the relative contribution of each log-normal distribution calculated from A_i ,

τ_i and W_i , normalized with their sum equal to unity.

S_1	τ_1 (μs)	S_2	τ_2 (μs)
0.76 ± 0.09 (0.675)	101 ± 7 (89)	0.24 ± 0.09 (0.325)	364 ± 49 (344)

Table 1: Mean values and standard deviations for the relative contributions S_i and relaxation times τ_i of the two log-normal distributions fitted to the RTDs. In parens are given the values obtained for the ensemble-averaged ICF.

The two types of values agree reasonably well taking into account the very difficult deconvolution of the RTDs into two log-normal distributions. For this particular example, it may still remain unclear whether this deconvolution makes sense. However further results (Echavarri Franco, 2011) showed that the two peaks separate continuously upon the addition of an oppositely charged surfactant in the solution and convey a physical significance. In this context, the informations brought by the use of CNTb allow us to adopt conservative estimates, e.g., $\tau_1 = 95 \pm 10 \mu\text{s}$, for this specific experimental condition.

Conclusions

To conclude we have described CNTb, a freely available set of scripts designed to allow the batch processing of PCS data by CONTIN. Batch CONTIN processing offers significant improvements compared to the on-line processing available with most commercial correlators in the case where many identical experiments are needed to solve difficult experimental problems, e.g., when the signal to noise ratio is small. In particular:

- i) The processing of many files with the same input parameters is fast and convenient, making it easy to vary these parameters in order to find their most appropriate values.
- ii) An ensemble-averaged ICF is calculated from the individual experiments. It provides safe hints for the influence of the statistical noise in the data: atypical results due to statistical noise can be discriminated with a single glance to the graphical output and discarded from further analysis. Reproducible qualitative features can be identified as well.
- iii) The RTD can be fitted with a set of log-normal distributions and the parameters of the latter can be used to compute mean values and standard deviations on the set of

experiments, which provides an estimation of error bars often sadly missed with on-line processing of a few measurements.

iv) The batch processing can be executed on a computer different from the one containing the correlator board, thus making the equipment available for further experiments.

Moreover the use of CNTb restores the full versatility of CONTIN, making it possible to deal easily with other problems involving the resolution of Fredholm's equations of the first kind like, e.g., in shear relaxation modulus measurements after a strain step.

References

Berne, B. J. & Pecora, R. (1976). *Dynamic Light Scattering with Applications to Biology, Chemistry and Physics*. New-York: Wiley & Sons.

Brown, W (1996). Editor. *Light Scattering: Principles and Developments*. Oxford: Oxford University Press.

Echavarri Franco, I. (2011). PhD thesis, University of Strasbourg, France.

Manning, G. (1969). *J. Chem. Phys.* **51**, 924-933.

Oosawa, F. (1968). *Biopolymers* **6**, 135-144.

Provencher, S.W. (1978). *J. Chem. Phys.* **69**, 4273-4276.

Provencher, S. W. (1979). *Makromol. Chem.* **180**, 201-209.

Provencher, S. W. (1982a). *Comp. Phys. Comm.* **27**, 213-227.

Provencher, S. W. (1982b). *Comp. Phys. Comm.* **27**, 229-242.

Rouf-George, C., Munch, J.-P., Schosseler, F., Pouchelon, A., Beinert, G., Boué, F. & Bastide, J. (1997). *Macromolecules* **30**, 8344-8359.

Vallat, P., Catala, J.-M., Rawiso, M. & Schosseler, F. (2006). *Macromolecules* **40**, 3779-3783.

Zhu, X., Shen, J., Liu, W., Sun, X. & Wang, Y. (2010). *Appl. Optics* **49**, 6591-6596 and references therein.

Supplementary material: Standard graphical output and statistics produced by CNTb.

In each panel the plots are labeled with the name of the corresponding data file. The label starting with "avrg" indicates the plot corresponding to the ensemble-averaged ICF. The PDF format of the graphical output allows the user to zoom in without losing resolution.

Panel 1: Composite plots showing: $g^{(2)}(t)-1$ initial values (blue) and $(g^{(2)}(t)-1)^{1/2}$ values used for the analysis by CONTIN (red) (left y-axis); resulting RTD values (black) and fitting curve $(g^{(2)}(t)-1)^{1/2}$ (solid line) (right y-axis). The upper x-axis gives the equivalent hydrodynamic radius for hard spheres with the same diffusion time in the same experimental conditions.

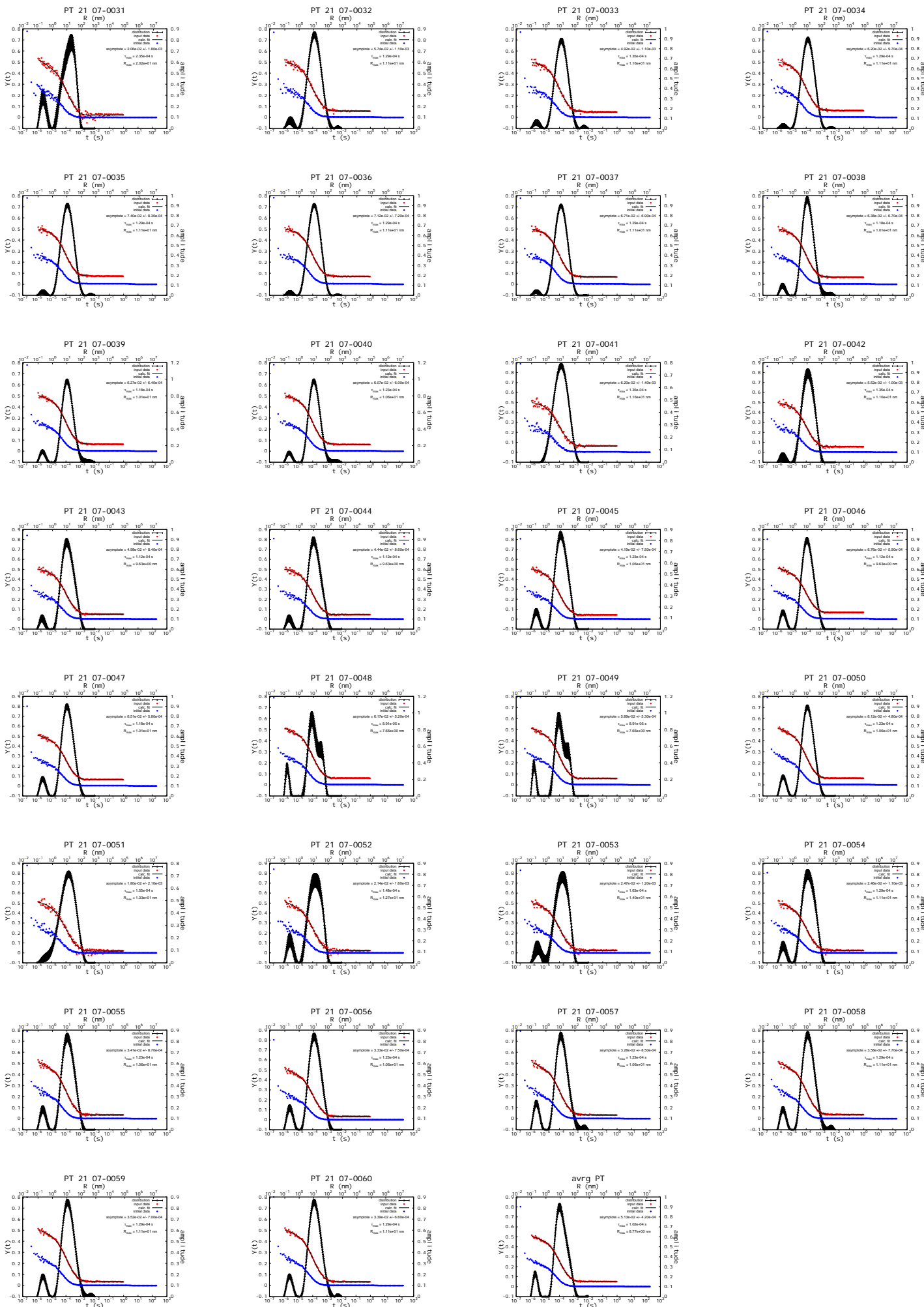
Panel 2: Same plots but with a logarithmic scale for the left y-axis.

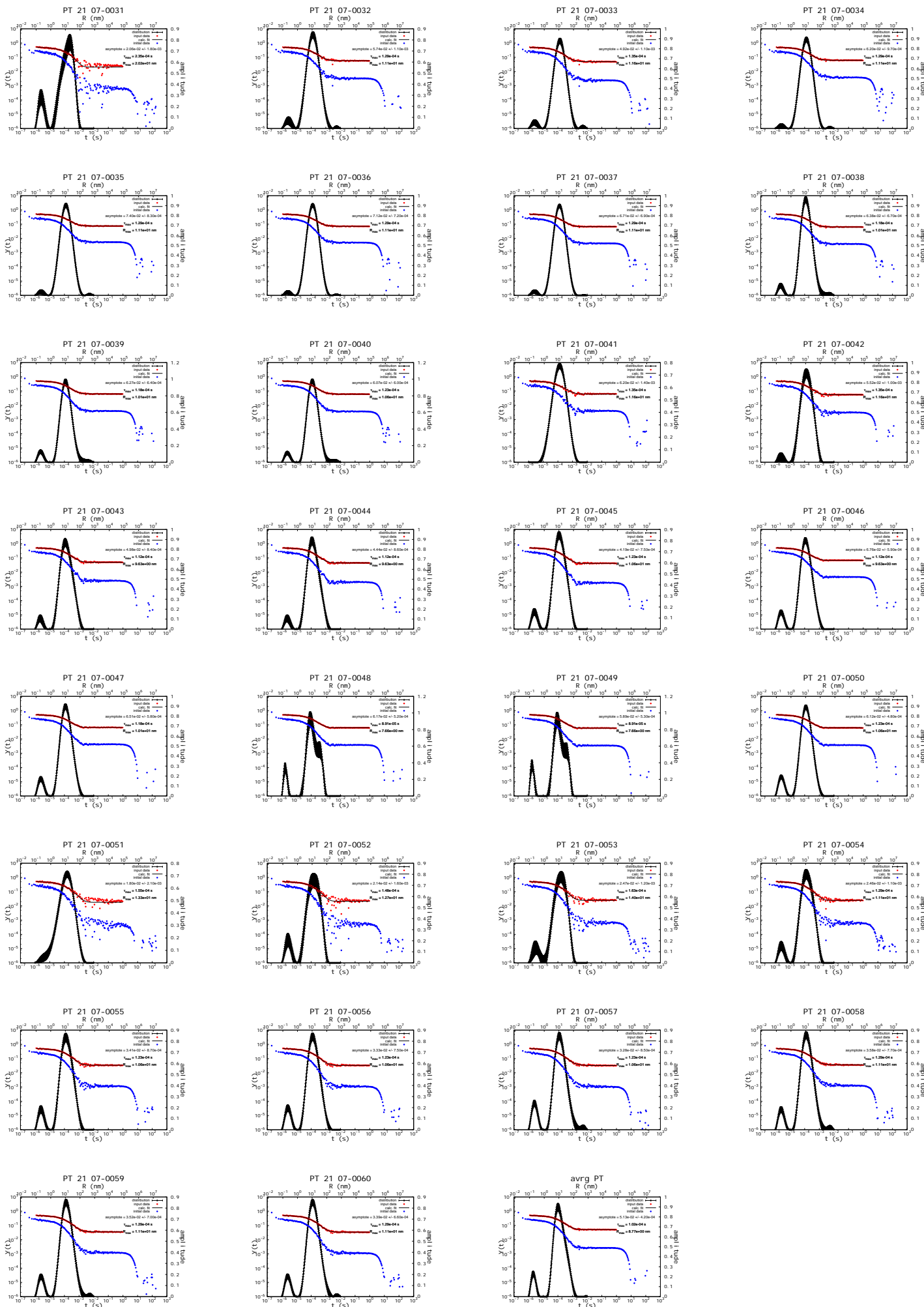
Panel 3: Traces of the scattering intensity recorded during the accumulation of the ICFs. The trace for the ensemble-averaged sample is randomly generated from the mean values for the scattering intensity and its standard deviation.

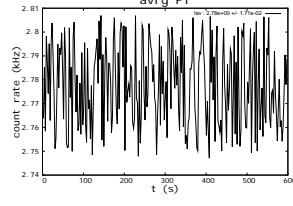
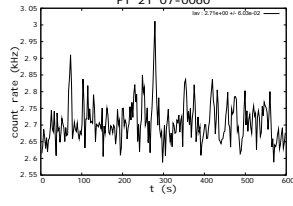
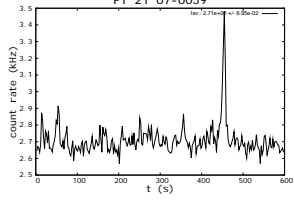
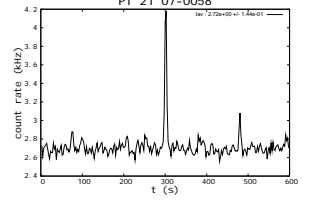
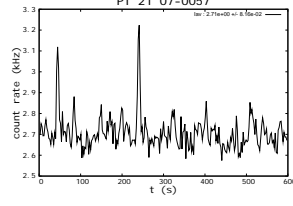
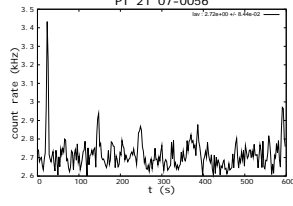
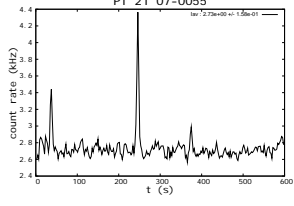
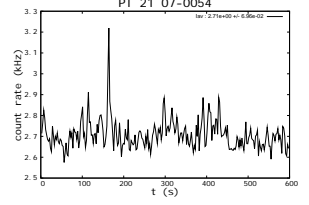
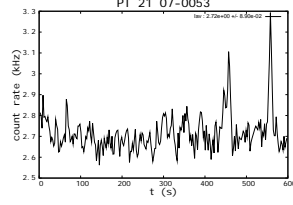
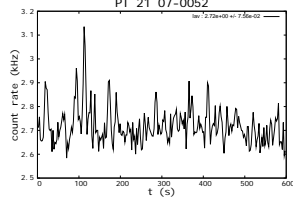
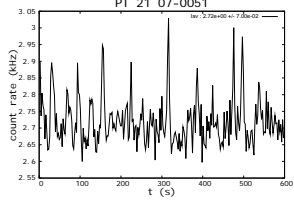
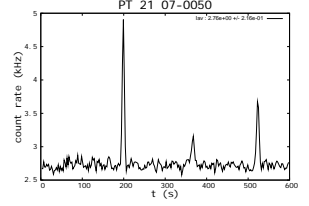
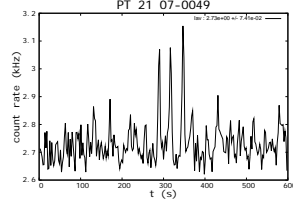
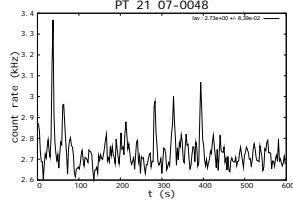
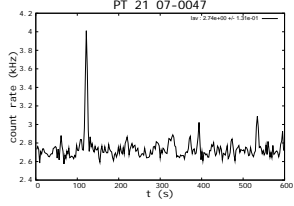
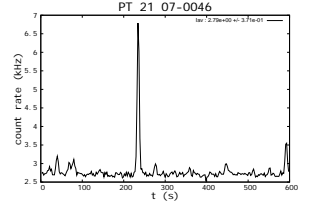
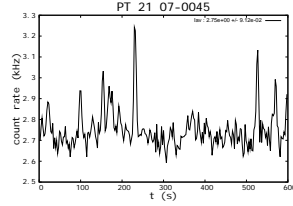
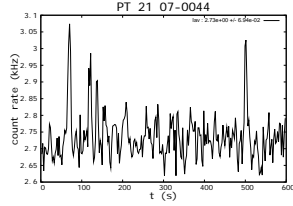
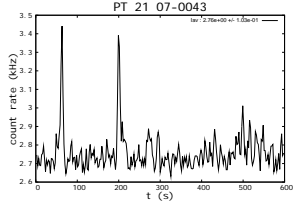
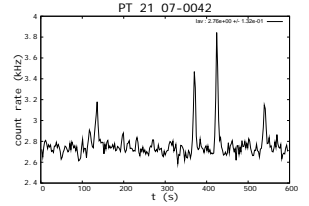
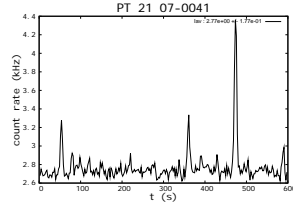
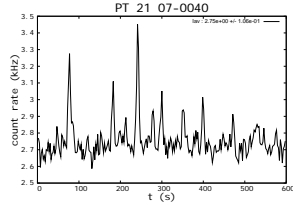
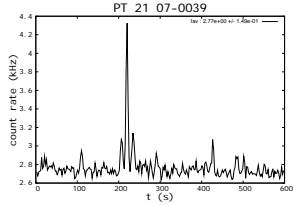
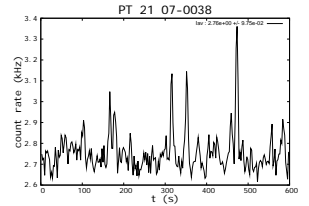
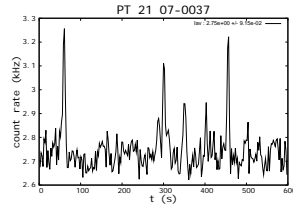
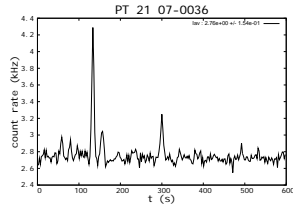
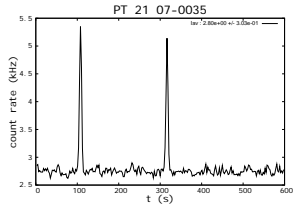
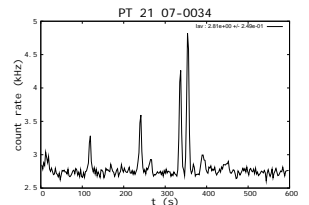
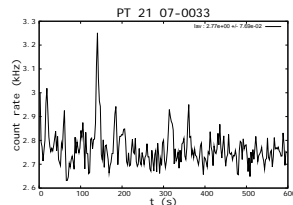
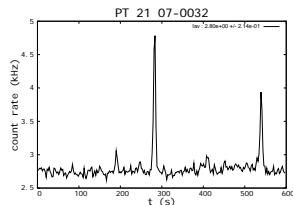
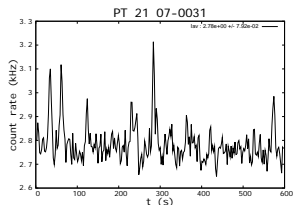
Panel 4: Fits of the RTDs (red) by a sum (solid line) of log-normal components (dotted lines). The error bars are those calculated by CONTIN. This information is usually lost by the proprietary softwares associated with commercial correlators.

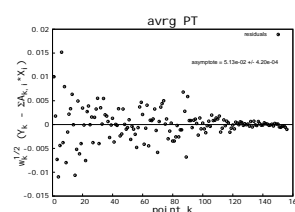
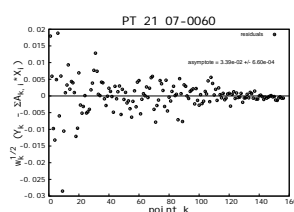
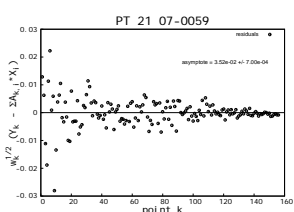
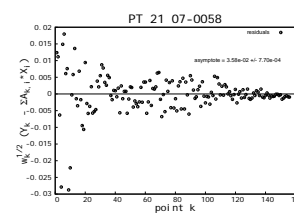
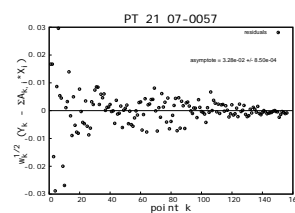
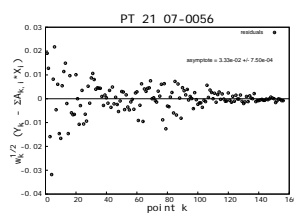
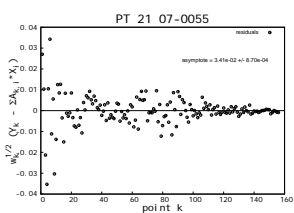
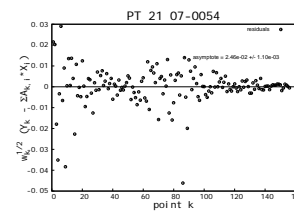
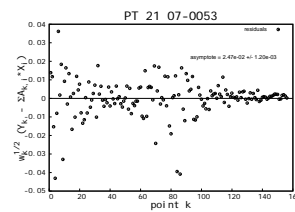
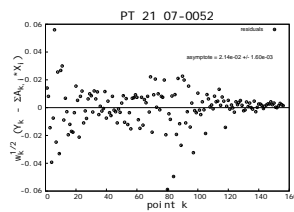
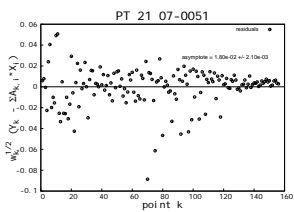
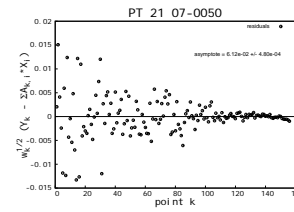
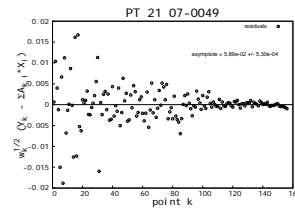
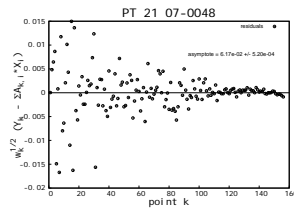
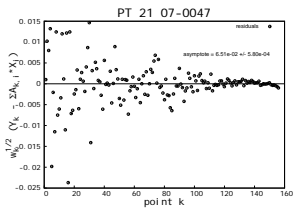
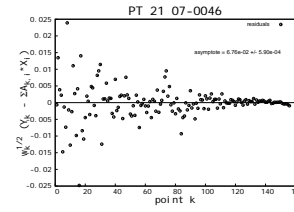
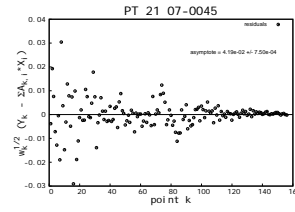
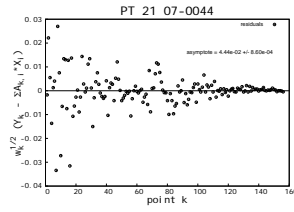
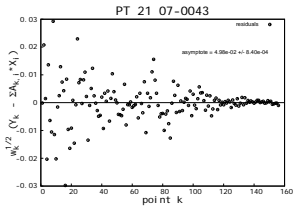
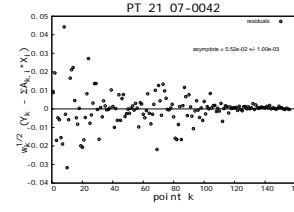
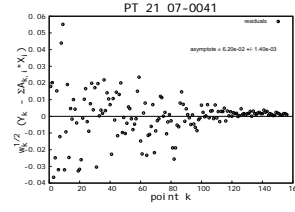
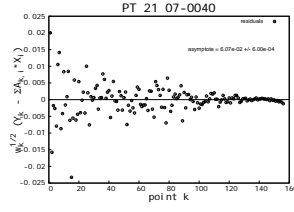
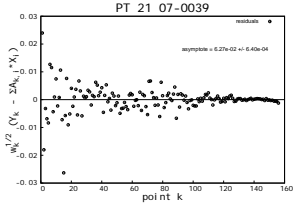
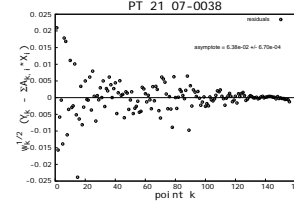
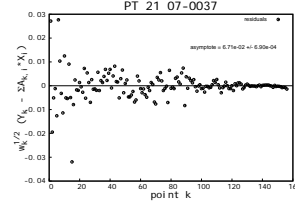
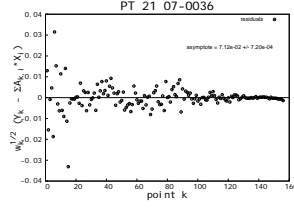
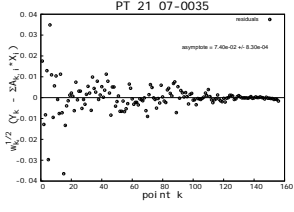
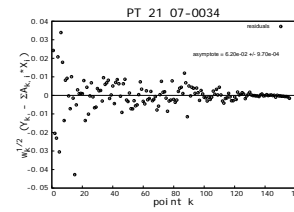
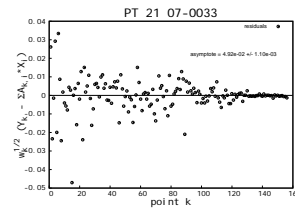
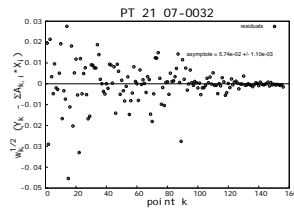
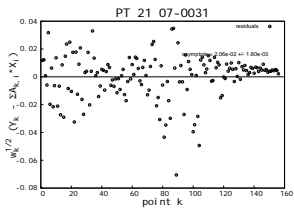
Panel 5: Weighted residuals computed from the difference between experimental correlation functions and their fitting curve.

Panel 6: Text file containing the statistics. Here the six data files with numbers 38-41, 51 and 57 were not considered for the statistics since their fits by the two log-normal distributions differ qualitatively from the one corresponding to the ensemble-averaged ICF.









Batch analysis of 24 files in 600/PT_21_07_600

PT_21_07-0031 ; PT_21_07-0032 ; PT_21_07-0033 ; PT_21_07-0034 ; PT_21_07-0035 ; PT_21_07-0036 ;
PT_21_07-0037 ; PT_21_07-0042 ; PT_21_07-0043 ; PT_21_07-0044 ; PT_21_07-0045 ; PT_21_07-0046 ;
PT_21_07-0047 ; PT_21_07-0048 ; PT_21_07-0049 ; PT_21_07-0050 ; PT_21_07-0052 ; PT_21_07-0053 ;
PT_21_07-0054 ; PT_21_07-0055 ; PT_21_07-0056 ; PT_21_07-0058 ; PT_21_07-0059 ; PT_21_07-0060 ;

Parameters for CONTIN and fit :

1.000e-06 < time [s] < 1.000e+00

1.000e-06 < Tau [s] < 1.000e-02

Free baseline

Unweighted residuals

Number of grid points : 200

3.162e-05 < 2 peaks < 1.000e-02

Average values on the file set

Peak 1

Amplitude : 7.52e-03 +/- 7.31e-04

Position [s] : 1.01e-04 +/- 7.37e-06

Position [nm] : 8.72e+00 +/- 6.34e-01

Width : 5.14e-01 +/- 5.09e-02

Relative area : 7.57e-01 +/- 9.10e-02

Peak 2

Amplitude : 3.21e-03 +/- 1.08e-03

Position [s] : 3.64e-04 +/- 4.94e-05

Position [nm] : 3.13e+01 +/- 4.24e+00

Width : 3.90e-01 +/- 5.49e-02

Relative area : 2.43e-01 +/- 9.10e-02

Intensity : 2.75e+00 +/- 3.11e-02

Intensity fluctuations : 1.34e-01 +/- 8.15e-02

g(0) : 0.474 +/- 0.026

Asymptote : 4.79e-02 +/- 1.69e-02

Average correlation function :

Peak 1

Amplitude : 8.13e-03

Position [s] : 8.94e-05

Position [nm] : 7.69e+00

Width : 4.20e-01

Relative area : 6.75e-01

Peak 2

Amplitude : 3.97e-03

Position [s] : 3.44e-04

Position [nm] : 2.96e+01

Width : 4.15e-01

Relative area : 3.25e-01

g(0) : 0.479

Asymptote : 5.13e-02