Turbulence and Diffusion Note No. 283

An ensemble approach to the simulation of fluctuating concentration time-series using a correlation-distortion technique

by

A.R. Jones and D.J. Thomson

22nd March 2002

Met Office
London Road
Bracknell
Berkshire
RG12 2SZ

An ensemble approach to the simulation of fluctuating concentration time-series using a correlation-distortion technique

A.R. Jones & D.J. Thomson

Turbulence and Diffusion Note 283, GMR, March 2002.

Abstract

Short-duration concentration fluctuations are important in assessing hazards from harmful substances released into the atmosphere. For example, a risk assessment model might include the effect of fluctuations when investigating toxicity or assessing the design of detection strategies for hazardous airborne materials. The driving force behind the current project is a requirement by DSTL for a methodology to model realistic concentration fluctuations with prescribed energy spectrum and probability density function by simulating realisations of concentration time-series. We do not aim to predict the detailed fluctuations as the concentration changes from one instant to the next, but rather to describe the correct 'climatology' of these fluctuations.

In our initial approach to this problem, we developed an iterative scheme for the simulation of fluctuating time-series based on a note by M. Nielsen at Risø. This technique is moderately successful at recreating realistic fluctuations but has its limitations. Thus an alternative approach has been sought.

The current report presents a second scheme for simulating realisations of fluctuating concentration time-series. An ensemble approach is adopted: the input data now prescribes the probability distribution and spectral structure describing an underlying ensemble of realisations although individual time-series in the ensemble do not satisfy identically the target data. The technique uses a correlation-distortion method. This generates time-series by transforming Gaussian time-series into series with the desired probability distribution. Here the transformation of the pdf distorts the auto-correlation function and spectrum, and so the spectrum of the initial Gaussian time-series is pre-calculated in order to match the target spectrum after this distortion.

This ensemble approach has several advantages over our earlier iterative scheme. It does not create statistical clones that all have an identical probability distribution, it produces less 'spiky' time-series and is better at representing the long interludes of zero concentration between individual 'bursts'. Finally, it is computationally efficient, having the capacity to rapidly generate a large ensemble of realisations.

Contents

1 Introduction

Concentration fluctuations in dispersing plumes have been studied experimentally in many field dispersion trials over the years. For instance, the time-series data set that we are using in this simulation research was collected in short-range tracer dispersion experiments conducted by C. Jones at MRU Cardington in the summer of 1998. However, specialised dispersion studies of this type are often costly and time-consuming; in particular, this approach would not be especially practical for routinely providing the large quantities of time-series data that might be required by risk assessment models. Thus an alternative strategy is needed – there is a requirement to model the fluctuations mathematically by generating realisations of concentration time-series using numerical simulation techniques. The interested reader is referred to Gurley et. al. [1] for an overview of some time-series simulation tools. In our work (see also [2, 3]), we aim to recreate realistic fluctuations given information on the expected amplitudes of the concentration values and their temporal structure.

The two principal characteristics describing a fluctuating concentration time-series are its probability distribution and energy spectrum. The basic concept underlying our numerical simulation scheme will be to use input information on these characteristic features to simulate realistic fluctuations with the prescribed structure. As such, the approach must actively involve both the probability distribution and spectrum in its scheme, such that it is able to accurately represent both features in any simulated realisations.

An example concentration time-series from the Cardington field trials data set is shown in Figure 1 below, along with its probability density function (pdf) and energy spectrum, see Figure 2. These observed concentration fluctuations were recorded during the J14 trial on data channel 3 (the propylene concentration measured by the second dual UVIC detector in the receptor array). The experiment was performed at 13:50 BST on 24th June 1998. See the previous note [2] for details of the experimental technique and further information on these field experiments. We will adopt this realisation as the test case for our simulation schemes later in this report.

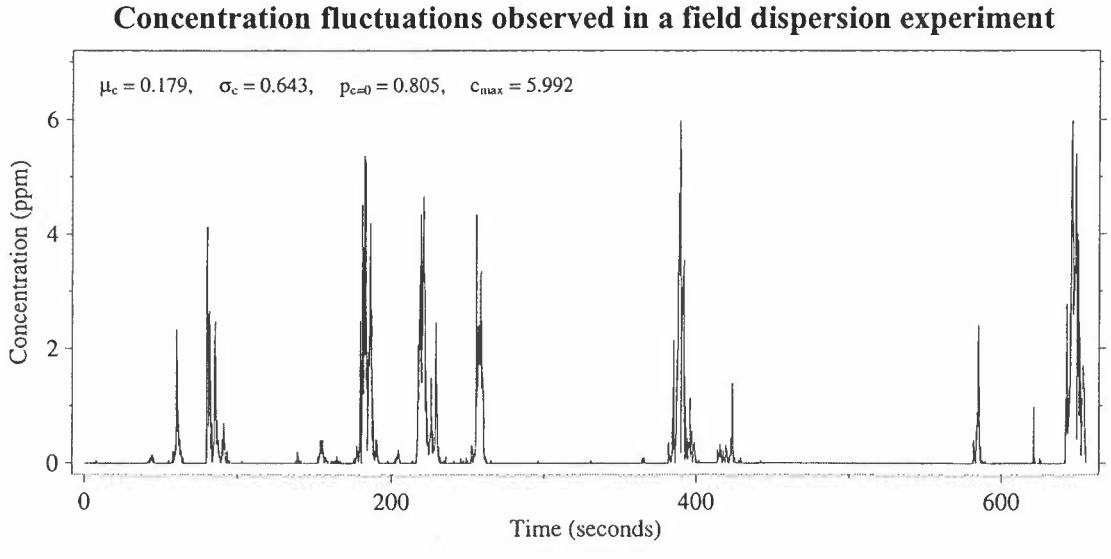


Figure 1: an example concentration time-series for the short-range dispersion of propylene gas in a Cardington field trial conducted by C.D. Jones in June 1998.

A probability density function and spectrum for concentration fluctuations

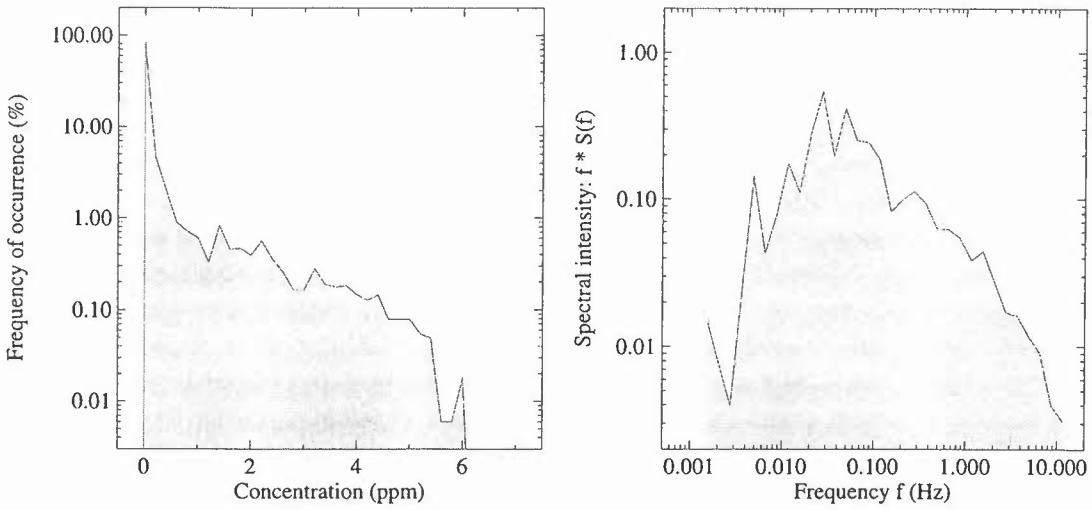


Figure 2: the probability density function and energy spectrum of the concentration fluctuations in the Cardington experiment displayed in Figure 1.

2 A summary of the iterative simulation scheme

A first scheme for simulating realisations of fluctuating time-series was developed during the summer of 1999. The approach adopts an iterative simulation technique, discussed in a note by M. Nielsen [4], to gradually evolve a time-series with the prescribed statistical and temporal structure. This method is discussed at some length in a previous report [2] on our simulation work. The discussion includes full details of the iterative algorithm, its implementation and its limitations. Here in the current note, we shall confine ourselves to a brief outline of the method, shown diagrammatically in Figure 3, and its ability to generate realistic fluctuations.

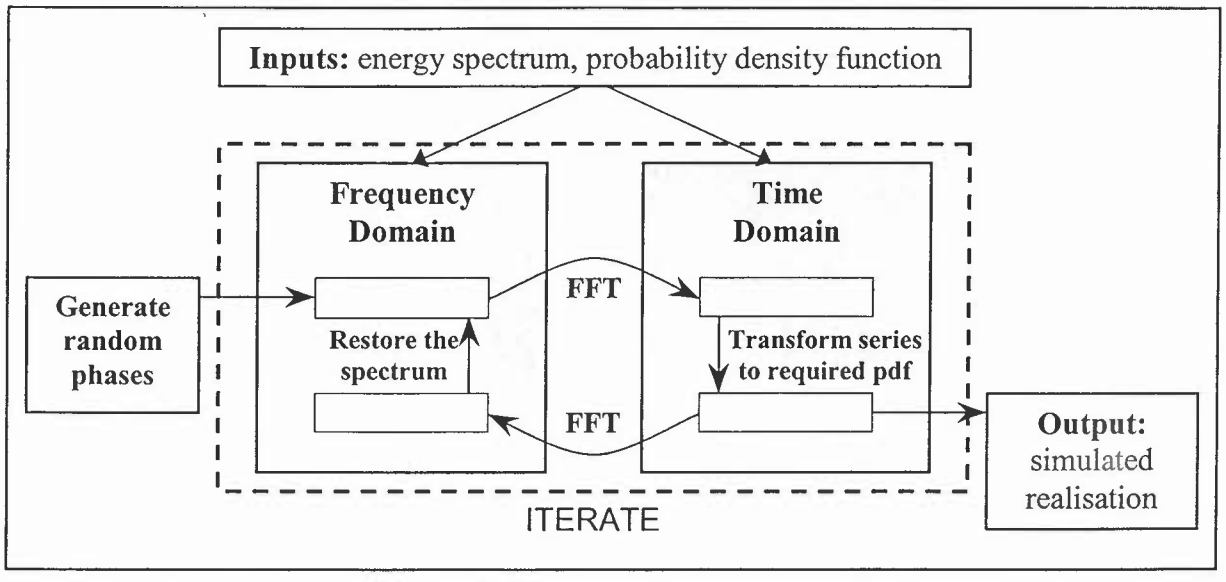


Figure 3: an overview of the iterative simulation scheme.

The scheme uses an iteration process to evolve a time-series realisation matching the given input pdf and energy spectrum. The main features of the algorithm are summarised below:

- **Initialisation**

  The algorithm is initialised with a Gaussian time-series having the prescribed spectrum. This time-series is constructed via its Fourier transform: the Fourier components are assigned with amplitudes matching the energy spectrum using randomly generated phase angles.

- **Iteration**

  The following procedure is repeated until a 'stable' realisation emerges. Note that Fast Fourier Transform (FFT) techniques are employed to transform between any time-series in the time domain and its Fourier representation in the frequency domain.

  1) *The first step transforms the time-series so that it has the required probability distribution, i.e. the amplitudes of the concentration values are monotonically adjusted to agree with the input pdf.*

  2) *The complementary step adjusts the amplitude of each Fourier mode to restore the prescribed energy spectrum. The phase angles are preserved in this step.*

  Thus each iteration is effectively nudging the initially random phases of the Fourier components towards a preferred solution.

- **Termination**

  The iterative procedure described above is repeated until it generates a time-series realisation displaying sufficient convergence to the required spectral structure. By exiting the algorithm routine after the first of the two steps, we ensure that the simulated realisation has the correct probability distribution. For instance, if we terminated the algorithm immediately after restoring the energy spectrum, it could potentially give negative values for the concentration!

The convergence behaviour of this algorithm has not been investigated in any detail. A reasonable fit to the input spectrum is usually obtained after just a few iterative loops, although the scheme is not convergent in the strict mathematical sense, i.e. it is not possible to reduce the spectral error below an arbitrary threshold simply by performing a greater number of iterations. Instead, after a sufficient time, the method will frequently find a 'fixed-point' time-series where the two iteration steps become reciprocal. The iteration effectively ceases at this stage, and there is no mechanism to further reduce any spectral error.

The final realisation is highly sensitive to the initialisation process where the random phases are selected for the initial time-series. A different selection of initial phases will give a different succession of time-series in the simulation sequence. Some initial states may produce fixed-point realisations with smaller spectral errors than those provided by other initialisations, although the output time-series tend to be remarkably similar in their general appearance. We found that it is very difficult to precisely match the target spectrum using this approach (clearly it must be possible if a consistent pdf and spectrum are supplied), but that any spectral errors are generally quite small.

An example time-series realisation generated by the iterative scheme is shown in Figure 4. On this occasion, the simulation engine was driven using the probability distribution and spectrum displayed in Figure 2 for the concentration fluctuations observed in the Cardington field trial. The realisation below should be compared with the experimental fluctuations plotted in Figure 1.

An example time-series realisation generated by the iterative scheme

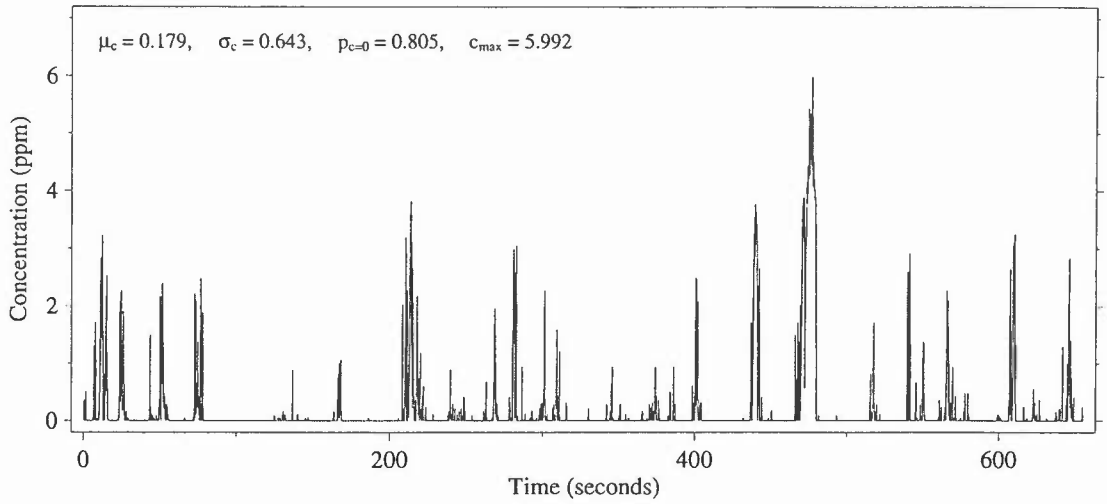


Figure 4: this realisation is an example of a concentration time-series produced by the iterative simulation engine driven using the field trial data shown in Fig 1.

This method for simulating concentration time-series generates realisations that capture the basic structure and appearance of the fluctuations reasonably well. However, the simulation is far from ideal with several limitations inherent in this approach. Firstly, examining the realisations in detail, the simulated time-series are rather 'spiky' in appearance and the long interludes of zero concentration between individual events are poorly represented. These quiet periods are difficult to reproduce because any perturbations in the Fourier amplitudes or phases will disrupt any such sequence of zero values.

Another drawback with this scheme is that each simulated realisation is a statistical clone of the input time-series data set; by forcing an exact match to the prescribed probability distribution, it gives no opportunity for any variability between separate realisations in statistical parameters such as mean, variance and intermittency. On a computational note, the iteration procedure can become rather expensive for simulations of long time-series because of the need to repeat FFT and sorting routines. Furthermore the scheme is only capable of generating one realisation at a time, and needs to be run successively if multiple time-series are required.

3 The second simulation method: an ensemble approach

In an effort to improve our simulation capability and overcome the limitations associated with the iterative scheme, a second technique has been developed for simulating realisations of concentration time-series. The method adopts an ensemble approach by generating a family of time-series realisations where the spectral properties and probability distribution are prescribed for an underlying ensemble rather than the individual realisations themselves. It uses a variation of the *correlation-distortion method* discussed by Gurley et. al. [1, 5].

This section presents the main details of the ensemble technique and discusses its implementation in a computational scheme. We assess the performance of the scheme at generating realistic concentration time-series in an efficient manner. Further details of the concepts used in this approach can be found in appendices after the main report.

3.1 An overview of the ensemble scheme

The approach aims to calculate the effect on the energy spectrum of a change in the probability distribution, so that each member of a simulated ensemble can be derived from a Gaussian time-series generated using the modified spectrum. The process is summarised below in Figure 5.

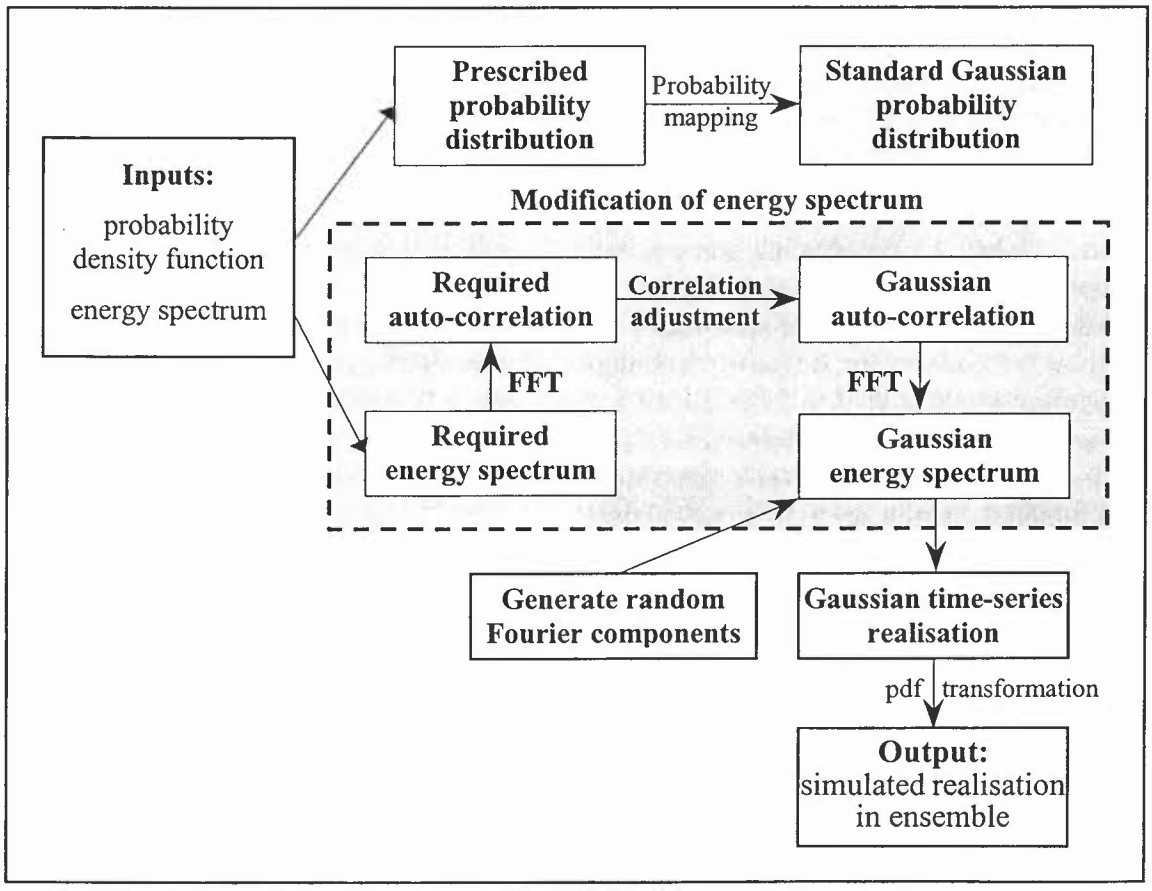


Figure 5: an overview of the ensemble simulation scheme.

At the heart of this technique lies the modification of the energy spectrum; this first step is a pre-requisite to the main simulation routine. Each output time-series is obtained by monotonically transforming a Gaussian time-series into a series with the

target probability distribution. The spectrum of the Gaussian process is first calculated such that the transformed time-series will have the desired spectrum. This is actually achieved by applying a mathematical transformation to the associated auto-correlation function; more details of this 'correlation adjustment' are given later in Section 3.2.

Once the adjusted spectrum has been computed, then the simulation engine can efficiently generate a large ensemble of time-series realisations. There are two steps in the simulation of each realisation:

1) *Generate a Gaussian time-series by selecting its Fourier components as independent Gaussian random variables with variances given by the modified energy spectrum.*

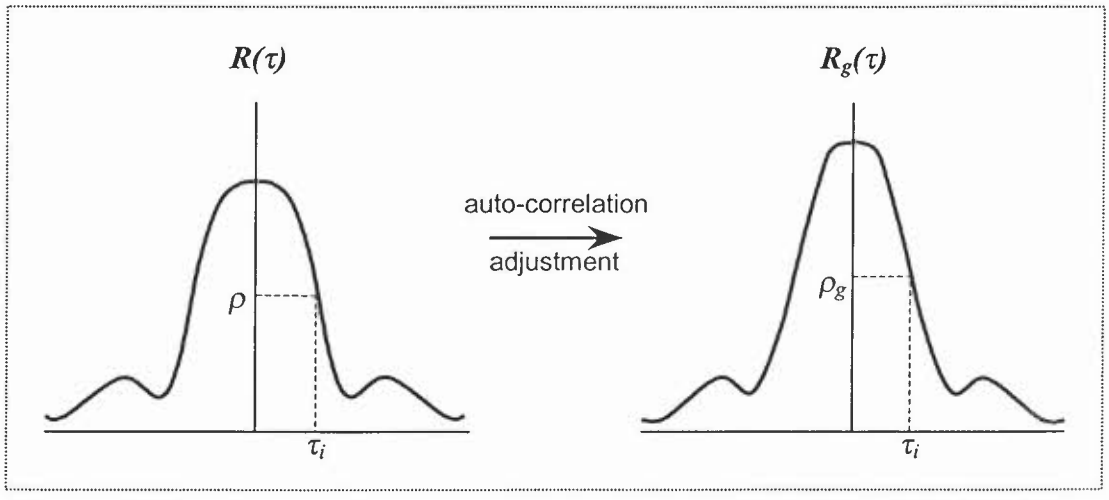
2) *The output realisation is produced by monotonically transforming the Gaussian time-series so that it has the prescribed probability distribution.*

Further details of the construction process are provided in Section 3.3.

3.2 The modification of the energy spectrum

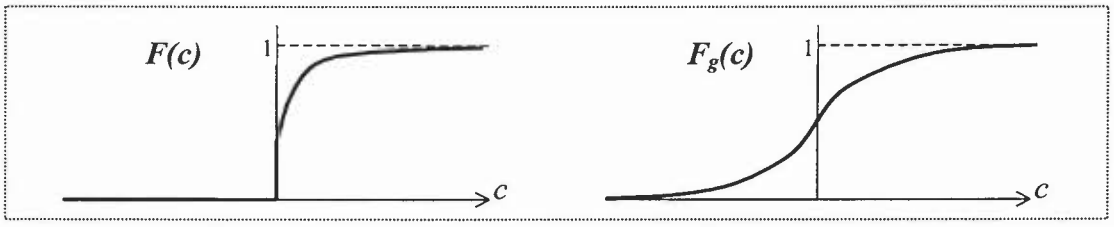
As a first step in the simulation scheme, it is necessary to compute the appropriate spectrum for the Gaussian process. Here the spectrum is constructed such that any Gaussian time-series with this modified spectrum transforms to a time-series with the original target spectrum when the prescribed probability distribution is restored. In fact, the problem can be translated into a more natural setting of auto-correlation functions by considering Fourier transforms instead. In this framework, the spectral modification procedure is described via an expression for the distortion in the auto-correlation function of the fluctuations. The correlation-distortion method involves calculating a certain double integral evaluated over a (two-dimensional) joint Gaussian probability distribution; the details of this calculation are presented below.

Let us begin by introducing some notation. Let $R(\tau)$ denote the auto-correlation function for the required concentration fluctuations (that is, the inverse Fourier transform of the input energy spectrum). Here, the parameter τ denotes the time-lag and may be positive or negative; although the symmetry of the auto-correlation function ensures that $R(\tau) = R(-\tau)$ for all τ and hence it is sufficient to consider only positive lag times $\tau \geq 0$. Similarly, let $R_g(\tau)$ be the auto-correlation function associated with the modified spectrum (for generating the Gaussian-distributed time-series); it is this function R_g which needs to be calculated.



In our discretised case, the auto-correlation functions R, R_g are defined at discrete lag times $\tau_i = i\Delta t$ ($i = 0,\ldots,N\text{-}1$). Here it is convenient to define auto-correlation without subtraction of the mean, so that $R(\tau_i) = \overline{c(t)\,c(t+\tau_i)}$ where the over-bar denotes an ensemble average. To simplify the notation, for any fixed value of the index i, put $\rho = R(\tau_i)$ and $\rho_g = R_g(\tau_i)$. The correlation distortion relating the values ρ and ρ_g will be presented shortly, but its description requires additional notation to be introduced.

Let F denote the *cumulative distribution function (cdf)* of the target probability distribution for the concentration fluctuations; that is, $F(c) = prob(conc \le c)$ for each concentration threshold value c. Similarly, let F_g be the cdf of the standard Gaussian distribution with zero mean and unit variance. These functions may typically have the following form.



Finally, denote the *probability density function (pdf)* of the two-dimensional joint standard Gaussian distribution by $g_2(r;\ x,\ y)$. This is the probability distribution satisfied by two jointly Gaussian random variables X, Y (each with mean zero and variance one) having a cross-correlation value $-1 \le r \le +1$; see Appendix A for more details. As a useful note to the reader, any *pdf* appearing in this report is written using *lowercase* notation, whereas any *cdf* appears with *UPPERCASE* notation.

Returning now to the main discussion on the calculation of the correlation distortion $\rho \to \rho_g$, the adjustment of the auto-correlation function satisfies the relation

$$\rho = \int_x \int_y \left(F^{-1}F_g(x)\right)\left(F^{-1}F_g(y)\right) g_2(\rho_g; x, y)\, dx\, dy. \qquad (3.1)$$

To explain this formula, note that ρ is the value of the auto-correlation function R at some fixed lag time τ_i. By definition, this is the expected value $\overline{c(t)\,c(t+\tau_i)}$ evaluated over the ensemble of time-series. However the right hand side of (3.1) is precisely this expected value when $c(t)$ and $c(t+\tau_i)$ are obtained by transforming random variables from a joint Gaussian probability distribution with suitable correlation ρ_g. To utilise the relation (3.1), it is necessary to invert it; that is, to compute the required value of ρ_g for each given value of ρ. The reader may find the illustration in Figure 6 useful in understanding the principle.

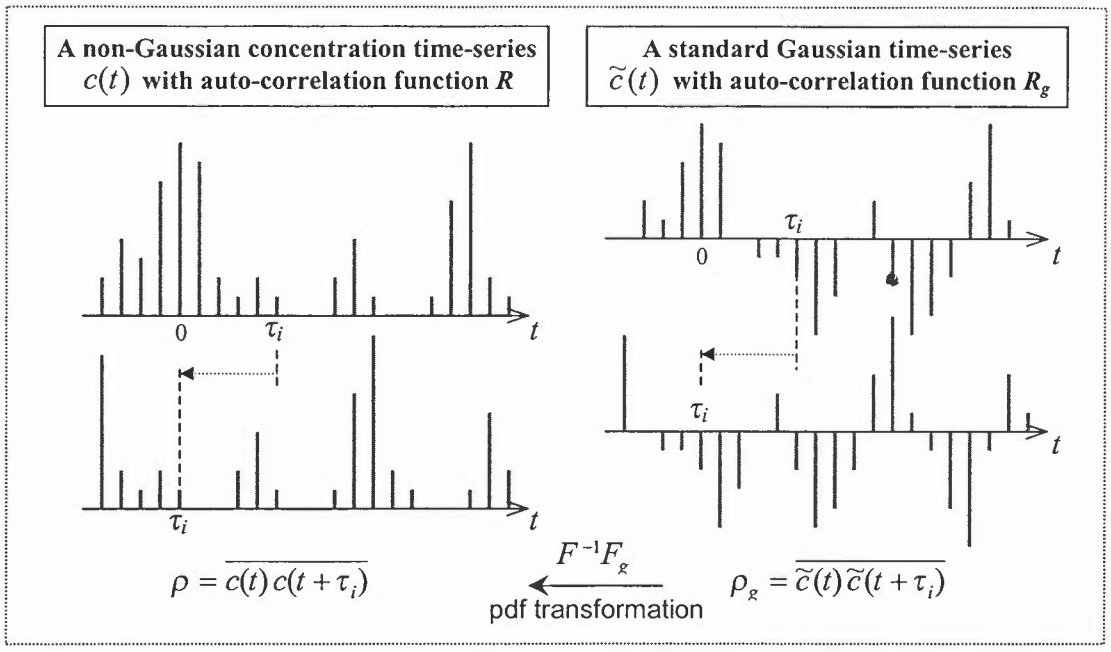


Figure 6: This diagram illustrates the basic principles for simulating a concentration time-series $c(t)$ via a standard Gaussian time-series $\tilde{c}(t)$ that has an adjusted auto-correlation function. The transformation in the probability distribution distorts the correlation structure according to the formula (3.1). By pre-calculating this correlation-distortion, it is possible to generate non-Gaussian time-series with a prescribed auto-correlation function.

From a mathematical viewpoint, the natural transformation for the distortion of the auto-correlation function is the 'forward' mapping $\rho_g \rightarrow \rho$ described in (3.1) as a double integration over a pair of correlated Gaussian random variables. In practice, this will be calculated by a numerical integration scheme, see Appendix B. The ensemble simulation approach actually requires the 'inverse' transformation $\rho \rightarrow \rho_g$; this is calculated indirectly using a numerical inversion scheme, see Appendix C.

3.3 The simulation of an ensemble of realisations

Once the modified energy spectrum has been calculated, it is a reasonably straight-forward process to generate time-series realisations for the simulated ensemble. There are two steps in the simulation of each concentration time-series: firstly, a Gaussian series is generated using the adjusted spectrum; then, this Gaussian realisation is transformed to obtain the prescribed probability distribution for the concentration fluctuations. A detailed description of the two stages in this procedure is now given.

Let us begin by describing our approach for simulating Gaussian time-series in an ensemble framework. Let the modified energy spectrum for the Gaussian simulation be written explicitly as $S_g(f_i)$. This is a discrete spectrum defined at the individual frequencies $f_i = i/T$ $(i=0,\ldots,N\text{-}1)$ where $T = N\Delta t$ is the total time of the time-series record (the fundamental period). The two-sided spectrum is defined here in such a way that the sum (from 1 to N-1) of its components equals the variance of the fluctuations; see [2] for further detail of the spectrum's definition. Note that the energy spectrum exhibits symmetry in its structure: $S_g(f_i) = S_g(f_{N-i})$ for each $i=1,\ldots,(N/2)\text{-}1$.

Recall that each component in the energy spectrum of a time-series describes the (square of the) amplitude of a particular Fourier mode in its spectral decomposition. Thus, a time-series can be constructed with any given spectrum by simply adding together Fourier modes of the required form. If the individual phases of these Fourier components are randomly chosen then the resulting time-series generated in this manner will approach a Gaussian distribution.

However, this technique will always generate a realisation having a spectrum that *precisely coincides* with the prescribed energy spectrum – there is no variability in the spectral structure between the individual members of the ensemble. This is an obvious limitation of the simple approach outlined above but one that is easily remedied by allowing greater freedom in the specification of the Fourier components. The aim is to match the prescribed spectrum in an *ensemble-average sense* while allowing some flexibility in the spectral structure of any individual realisation. This is achieved by the following construction for each member of an ensemble via its Fourier transform.

As a first step towards creating each ensemble member, generate N-1 independent random values $r_1,\ldots,r_{N-1}$ from a Gaussian distribution with zero mean and unit variance. Then define the Fourier transform $X(f_i)$ of a time-series realisation $x(t_i)$ by

$$
\begin{aligned}
&X(0) = 0; \\
&\left.\begin{aligned} X(f_i) &= r_i \sqrt{S_g(f_i)/2} + I\, r_{N-i} \sqrt{S_g(f_i)/2} \\ X(f_{N-i}) &= r_i \sqrt{S_g(f_i)/2} - I\, r_{N-i} \sqrt{S_g(f_i)/2} \end{aligned}\right\} \quad \text{for each } i = 1,\ldots,(N/2)-1; \\
&X(f_{N/2}) = r_{N/2} \sqrt{S_g(f_{N/2})}\,.
\end{aligned}
$$

Here the symbol I denotes the imaginary number $\sqrt{-1}$. The Fourier transform of a real sequence is an Hermitian sequence with complex-valued components, except for the zero-th and N/2-th components which are purely real. Note that the spectral energy is, on average, split equally between the real and imaginary parts of each Fourier component; this property ensures that there is no bias in the phase angles of the individual Fourier modes – the phases will be completely random with a uniform distribution. The scaling factor $1/\sqrt{2}$ is introduced so that the total contribution to the energy in the Fourier mode from both the real and imaginary parts conforms to the desired spectral characteristic.

It can be easily shown that, in the ensemble-average sense, this description for the Fourier transforms of the time-series realisations provides the required energy spectrum $S_g(f_i)$. In fact, ensemble-averaging gives $\overline{r_i^2} = 1$ for each index i, and it then follows immediately that

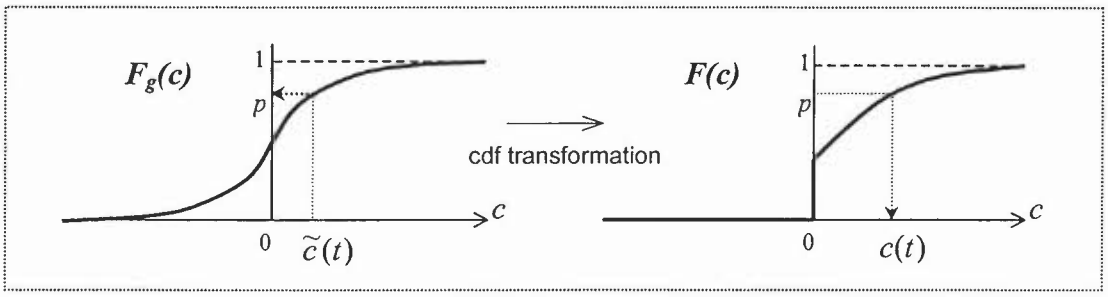
$$
\overline{|X(f_i)|^2} = S_g(f_i) \quad \text{for each } i = 1,\ldots,N-1.
$$

As a final step, each Gaussian time-series in the ensemble is then computed as the inverse discrete Fourier transform of such a sequence. Over a large ensemble, these simulated realisations approach a Gaussian distribution with zero mean and unit variance. Furthermore, their spectral structure will average to the prescribed spectrum $S_g(f_i)$ over a large ensemble of realisations, as required.

The second stage in the simulation procedure involves generating a non-Gaussian concentration time-series from each Gaussian realisation in the above ensemble. This is achieved by transforming the Gaussian series to obtain the required probability distribution. Specifically, each entry in the Gaussian time-series $\tilde{c}(t)$ is individually transformed to its matching value $c(t)$ in the prescribed cumulative distribution function F; that is,

$$c(t) = F^{-1}F_g(\tilde{c}(t))$$

for each discrete time t. Identifying the two probability distributions in this manner ensures that the time-series transformation is monotonic and so preserves the ordered structure of the fluctuations in the realisation: if $\tilde{c}(t_1) < \tilde{c}(t_2)$ for the Gaussian series at two times t_1, t_2 then $c(t_1) < c(t_2)$ for the output realisation.



Further details that aim to cover some computational aspects of the cumulative distribution functions F and F_g, including the calculation of the inverse cdf, are given in Appendices D and E respectively.

The family of concentration time-series realisations generated by this simulation scheme will have, in the ensemble-average sense, the desired energy spectrum and (single-time) probability distribution; this follows directly from the construction of the modified spectrum (via the auto-correlation distortion) that adopted an ensemble-average viewpoint. Of course, individual realisations within this family may not precisely match the given probability distribution and spectrum.

3.4 The performance of the ensemble simulation scheme

This final section presents an assessment of the performance of the ensemble simulation scheme, and compares it with the earlier iterative method for simulating concentration time-series. An example concentration time-series generated by the ensemble approach is shown in Figure 7. Here the simulation has been driven using the time-series realisation observed in the Cardington field trial (see Figure 1) to provide the input probability distribution and energy spectrum. One member of the simulated ensemble has been selected.

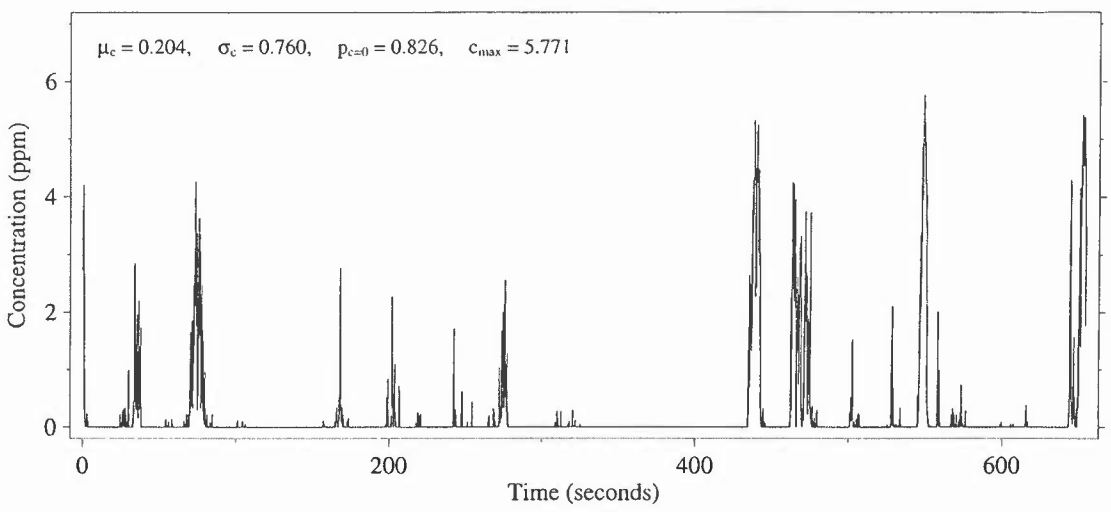


Figure 7: An example time-series realisation generated by the ensemble scheme. The simulation engine adopts the concentration time-series recorded in the Cardington field trial to prescribe the probability distribution and spectrum of the concentration fluctuations. The simulated realisation shown is just one member of a family of time-series realisations produced by this approach.

The concentration fluctuations modelled in the above realisation should be compared with the observed fluctuations shown in Figure 1 and against our earlier simulation, displayed in Figure 4, using the iterative method. The overall appearance of the simulated time-series created by the ensemble scheme is encouraging: the technique models the main features of the concentration fluctuations reasonably well on a broad scale. In particular, this second approach to the simulation of fluctuating time-series generally appears to be an improvement on the earlier iterative method. Enlarged sections of these three time-series are shown together in Figure 8. Here the simulations compare very favourably against the observed fluctuations, although it is apparent that the iterative scheme does produce time-series with a noisier appearance.

The iterative scheme produces rather spiky realisations, as previously noted in Section 2, and as a consequence does not properly capture the long sequences of zero concentration experienced when the plume drifts away from the detector's position. In contrast, the ensemble approach generates more realistic time-series that provide a better match to observed fluctuations experienced in atmospheric dispersion. There is an improvement in the representation of the quiet interludes with zero concentration, although the simulated realisations are still not perfect. This improvement over the iterative scheme does occur systematically for a range of cases.

Despite this improvement, we generally found that the ensemble scheme does not recreate the full extent of the quiet interludes; instead, they are interrupted by sporadic blips in concentration activity giving the time-series a slightly noisy appearance. This shortcoming in the ensemble scheme is perhaps a consequence of not being able to directly model phase structure in our approach. Although the discussion here has been centred on a single case study, it should be noted that the example case shown is fairly typical of the relative merits of the two simulation methods. Further examples using experimental time-series data support the above assessment; in each instance, the realisations have a similar appearance exhibiting these strengths and weaknesses of the two schemes.

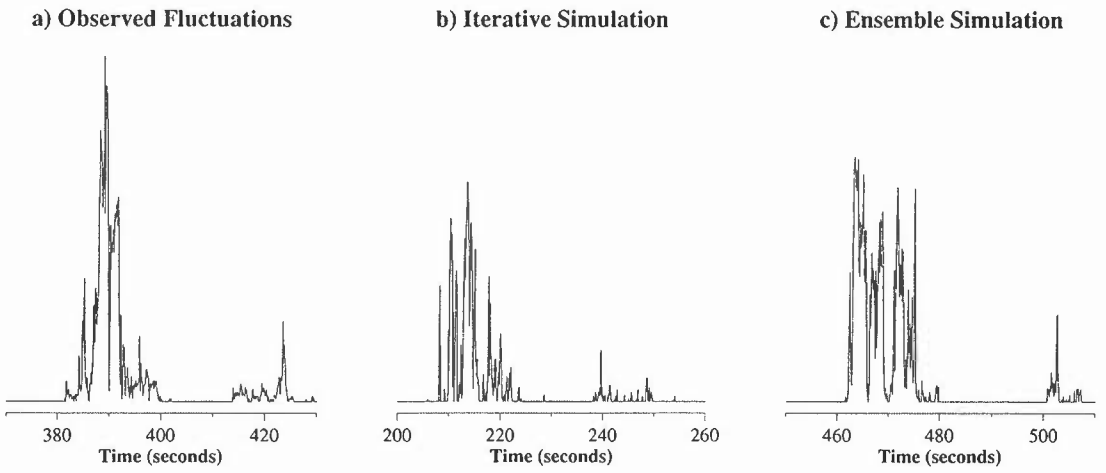


Figure 8: an enlarged section of each time-series realisation illustrating the detailed structure of the concentration fluctuations over short time-scales.

As a final note, it is appropriate here to comment on some potential difficulties with the correlation-distortion method that should be borne in mind for any practical implementation of the scheme. These problems concern the provision of a target energy spectrum and auto-correlation function that are physically meaningful and give a practical Gaussian spectrum S_g.

The first difficulty that can arise here is the appearance of negative components in the computed spectrum S_g, since there is no guarantee that all components of the Gaussian spectrum will be positive (or zero). As a corrective measure, we propose that the modified spectrum S_g should be truncated so that any negative spectral components are clipped to zero. This procedure removes negative energy and so increases the total energy (variance) of the fluctuations. Therefore a scaling of the physical spectrum is necessary to preserve the total spectral energy. This correction procedure overcomes any problems created by the existence of negative components in the energy spectrum. However it can have a dramatic impact on the structure of the concentration fluctuations if there are a large proportion of truncated components.

It is therefore prudent to use an input spectrum that does not produce a large occurrence of negative values in the distorted spectrum S_g. We found that an important factor contributing to the amount of truncation is the smoothness of the target spectrum. The truncation issue can become a significant problem if the input spectrum is not sufficiently smooth, but does not appear to give any real difficulties for smooth spectra. Note that the raw spectrum of a single time-series realisation is usually very noisy in appearance because spectral amplitudes at different frequencies are uncorrelated, and so there is a need to smooth any such spectra. We found that the extent of spectral truncation becomes less as greater smoothing is applied to the input spectrum.

Unfortunately, smoothing a raw spectrum obtained from a single realisation can create a second potential difficulty, so that some care needs to be taken in applying a suitable smoothing technique. Any smoothed spectrum should be physically realistic in the sense that its inverse discrete Fourier transform gives a sensible auto-correlation function. Here we are considering the auto-correlation function $R(\tau)$ as defined previously without subtraction of the mean value.

The auto-correlation must always be positive; in fact, it is bounded below by a minimum (positive) value ρ_{min} (see Appendix C). From a mathematical viewpoint, of course, it is quite possible that the Fourier transform calculation could generate some unrealistic correlation values below this ρ_{min} threshold from an apparently innocuous smoothed spectrum. If this problem occurs, a suggested solution is to truncate the auto-correlation function at the lower bound, clipping any lesser values to ρ_{min}, in a similar manner to our earlier treatment of the modified spectrum. Fortunately, this correction step is not usually necessary for any sensible smoothing scheme. The issue has been raised largely as a precautionary tale to the reader of the potential problems encountered by the ensemble simulation approach.

4 Summary

This note reports on the simulation of fluctuating concentration time-series from an ensemble viewpoint using a correlation-distortion technique.

Concentration fluctuations are created in a dispersing plume by the processes of turbulent diffusion acting in the atmospheric boundary layer. These fluctuations are especially pronounced at short ranges downwind from a source where they can have a significant impact on the severity of a hazard (toxicity, flammability, etc.). Therefore it is often essential to include the effects of fluctuations in risk assessment models.

Concentration fluctuations have been extensively studied in many field trials over the years; however, such dispersion experiments are a costly and time-consuming means of providing information on fluctuations to risk models. There is a requirement to model fluctuations by simulating realisations of concentration time-series using analytical methods. The aim is to generate realistic time-series given some basic information on the structure of the concentration fluctuations, such as their probability distribution and spectrum, that could be estimated for any particular circumstance.

In a previous report [2], an iterative method was given to generate a concentration time-series with prescribed characteristics. Although reasonably successful at recreating realistic fluctuations, the iterative scheme has some clear deficiencies and limitations. The simulated realisations are rather 'spiky' in appearance with a poor representation of the quiet interludes of zero concentration between individual concentration bursts. Furthermore, the scheme creates statistical clones of the input probability distribution and spectrum in the sense that these are matched by each realisation individually rather than just by the ensemble as a whole. Finally, on a computational note, it can be expensive for simulations involving long time-series.

An alternative simulation scheme for creating single concentration time-series is proposed here. This adopts an ensemble approach to the problem, where input data prescribes the probability distribution and energy spectrum in an ensemble-average sense – describing a family of time-series realisations rather than the individual time-series themselves. The ensemble approach uses a correlation-distortion technique to modify the input spectrum. Each time-series realisation in the ensemble is created by transforming a Gaussian time-series into one with the desired probability distribution. The transformation of the pdf distorts the correlation function and spectrum of the process, so that the spectrum of the initial Gaussian series needs to be pre-calculated to produce the target spectrum after distortion.

The ensemble scheme overcomes some of the limitations inherent in the earlier iterative approach, and is capable of generating a large ensemble of time-series realisations in an efficient manner. The new approach still has some difficulties in modelling the long interludes of zero concentration that occur when the plume meanders away from the detector, although it does capture this phenomenon better than the previous scheme. It is possible to generalise the ensemble approach to allow the simulation of multiple correlated time-series representing spatially-separated detectors. This extension of the ensemble method is the subject of the third TDN in this series [3].

On a final note, any application of the ensemble scheme will require input data to drive the simulation engine (*either* a sample concentration time-series *or* a prescribed probability distribution and spectrum). This information might be supplied from actual realisations observed in dispersion experiments, or ultimately by idealised functions based on relevant parameters. This 'parametrisation' problem would need to be addressed before an operational time-series simulation scheme can be fully established, perhaps, for example, by using concepts such as those in the ADMS fluctuations scheme [6, 7].

Acknowledgements

This research was funded jointly by *DSTL Porton* and the *Met Office Core Research Programme*.

References

[1] GURLEY K.R., TOGNARELLI M.A. AND KAREEM A., 1997; "Analysis and simulation tools for wind engineering"; *Prob. Engng. Mech.*, **12**, 9 – 31.

[2] JONES A.R. AND THOMSON D.J., 2002; "An iterative approach to the simulation of fluctuating concentration time-series"; *Turbulence and Diffusion Note No. 282*, Government Meteorological Research, Met Office.

[3] JONES A.R. AND THOMSON D.J., 2002; "An ensemble approach to the simulation of concentration time-series at two detectors using a correlation-distortion technique"; *Turbulence and Diffusion Note No. 284*, Government Meteorological Research, Met Office.

[4] NIELSEN M., 1999; "Simulating time-series with prescribed probability distribution and spectral distribution"; unpublished report.

[5] GURLEY K.R., KAREEM A. AND TOGNARELLI M.A., 1996; "Simulation of a class of non-normal random processes"; *Int. J. Non-Linear Mech.*, **31**, 601 – 617.

[6] THOMSON D.J., 1992; "The fluctuations module"; *ADMS Technical Specification paper P13/01E/92*, published by CERC.

[7] THOMSON D.J., 2000; "Concentration fluctuations in ADMS3, including fluctuations from anisotropic and multiple sources"; *ADMS Technical Specification paper P13/07D/00*, published by CERC.

[8] TONG Y.L., 1990; "The Multivariate Normal Distribution"; Springer-Verlag.

[9] ABRAMOWITZ M. AND STEGUN I.A., 1970; "Handbook of Mathematical Functions"; Dover Publications, New York.

Appendix A: The joint Gaussian distribution on two variables

Let x, y be two Gaussian random variables, each with mean zero and variance one, having a correlation value $-1 \le r \le 1$ such that their joint probability density function is given by

$$g_2(r;x,y) = \frac{1}{2\pi\sqrt{1-r^2}} e^{-\left[\frac{x^2 - 2rxy + y^2}{2(1-r^2)}\right]}.$$

This describes a two-dimensional joint Gaussian distribution on the pair (x, y). As an example, Figure 9 illustrates the typical form of the joint probability density function – in this case when $r = 0.4$. Note that the formula becomes undefined at $r = \pm 1$, where the distribution collapses to a one-dimensional form. As the correlation r approaches one (where the variables x and y become perfectly correlated), the probability density accumulates along the line $y = x$. Similarly as r decreases towards minus one (perfect anti-correlation), the density collects along the line $y = -x$. For uncorrelated variables, where $r = 0$, the density function is axially symmetric with a circular set of contours.

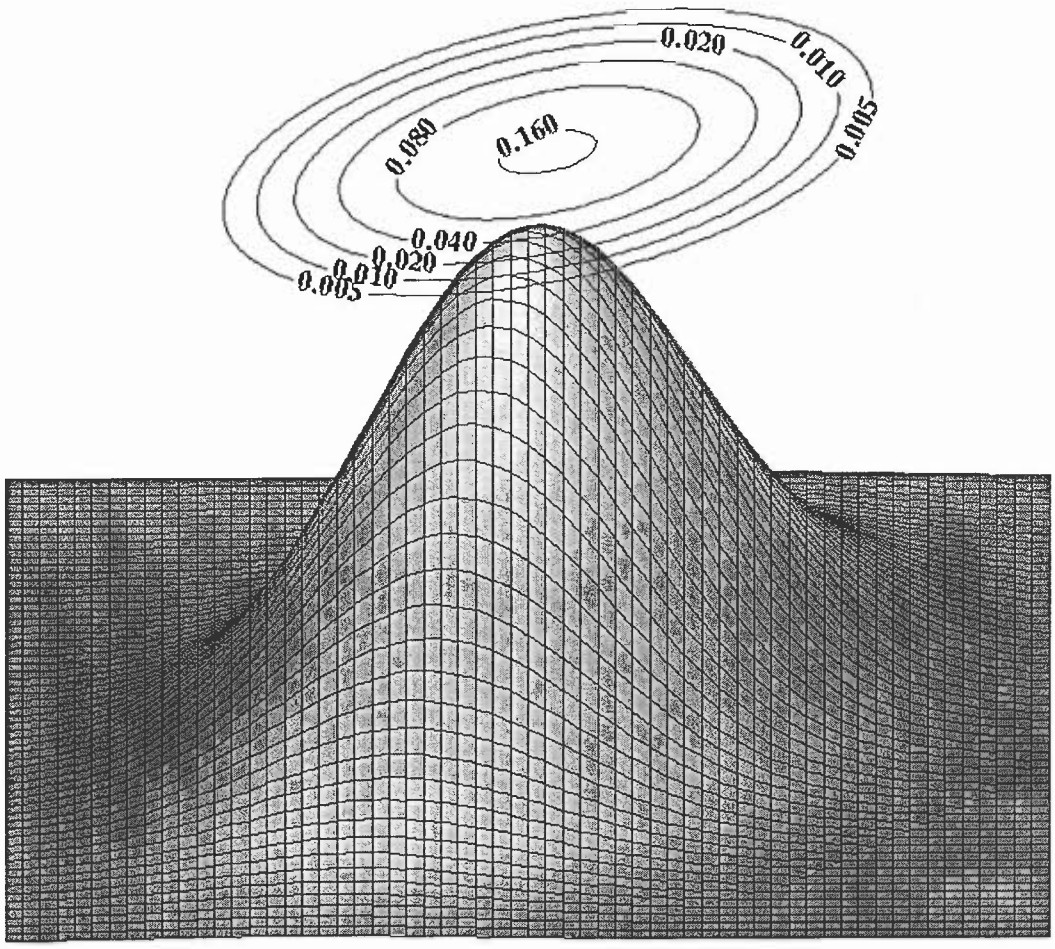


Figure 9: A graphical representation of the probability density function $g_2(r; x,y)$ for the two-dimensional joint Gaussian distribution with correlation $r = 0.40$. The lower part of the figure depicts the 2-d surface plotted above the (x, y) plane; the upper part overlays a contour plot of the values of the density function.

Appendix B: A dynamic numerical integration scheme

The correlation-distortion process requires us to evaluate the double integral

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(F^{-1}F_g(x)\right)\left(F^{-1}F_g(y)\right) g_2(r;x,y)\, dx\, dy$$

for many different values of the joint Gaussian correlation r in the interval (–1 , +1). It is probably not possible to give an exact analytic expression for this double integral, and so it has not been attempted here. Instead, an approximate value for the integral is calculated using a numerical integration method.

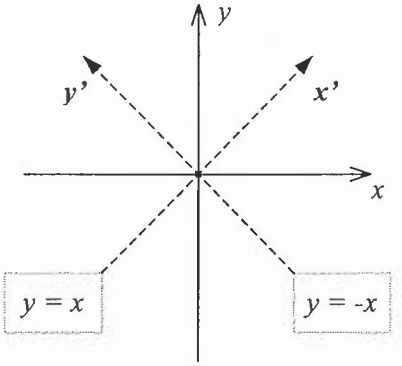
There are two requirements for the numerical integration. Firstly, it is necessary to have computational routines for the functions F^{-1}, F_g and g_2 involved in the integrand; these procedures are dealt with separately in Appendices A, D and E. Secondly, a numerical integration scheme is required using some appropriate discretisation of the (x, y)-plane. A uniform square grid was initially considered for the discretisation but it became apparent that an improved dynamic approach could be adopted. The proposed scheme adapts the discretisation grid to the particular value of the correlation r giving a more effective integration process.

In this study, we only consider a discretisation using uniform grid spacing in each direction. It is possible that a more elaborate non-uniform scheme, allowing variable size grid cells, may offer a more robust numerical scheme with better accuracy. However, for the current purposes, a uniform grid approach appears to be sufficient – giving reasonable accuracy in a scheme that is straightforward to implement. Here it is important to ensure that the finite grid over which the numerical integration is performed covers a sufficient proportion of the joint Gaussian distribution (typically to ±4 standard deviations in practice) and that it has the necessary resolution to yield a good approximation for the integral (in practice, a 64 × 64 grid provides sufficient resolution).

Dynamic numerical integration

An integration scheme is developed that utilises a dynamic discretisation grid dependent on the correlation r between the two Gaussian distributions. This technique adopts a rectangular grid aligned with the principal axes of the joint Gaussian probability density function. The length and width of the integration area required to cover a given fraction of the total distribution are easily described in terms of the correlation value r.

The principal axes of the joint Gaussian pdf $g_2(r; x, y)$ are $y = \pm x$. Consider the transformation to these natural co-ordinates (x', y'):



$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

When r is zero, the joint Gaussian distribution is axially symmetric. As $r \to 1$, the joint distribution collapses onto the x' axis (that is, $y = x$); whereas, as $r \to -1$, the distribution converges along the y' axis where $y = -x$.

Let us calculate the spread of the Gaussian distribution in the (x', y') frame of reference. Recall that both the x and y co-ordinates of the joint Gaussian profile are distributed with mean zero and variance one. In the (x', y') co-ordinate system, their mean values are again zero, and the variances are given by

$$\overline{x'^2} = \frac{1}{2}\overline{(y+x)^2}\,; \quad \overline{y'^2} = \frac{1}{2}\overline{(y-x)^2}$$

where the over-bar denotes the expectation (mean) value. Now expanding the terms in parentheses gives

$$\begin{aligned}\overline{(x \pm y)^2} &= \overline{x^2} + \overline{y^2} \pm \overline{2xy} \\ &= 1 + 1 \pm 2r \\ &= 2(1 \pm r).\end{aligned}$$

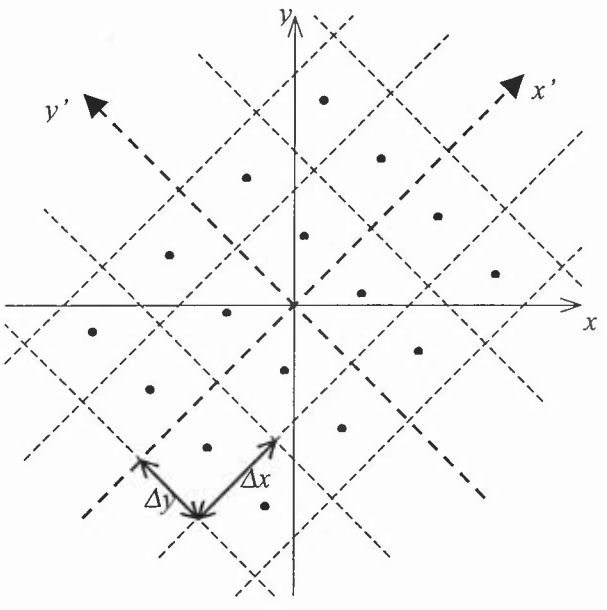
That is,

$$\overline{x'^2} = 1 + r\,, \quad \overline{y'^2} = 1 - r.$$

Hence the standard deviations of the distribution along the principal axes are $\sqrt{1 \pm r}$.

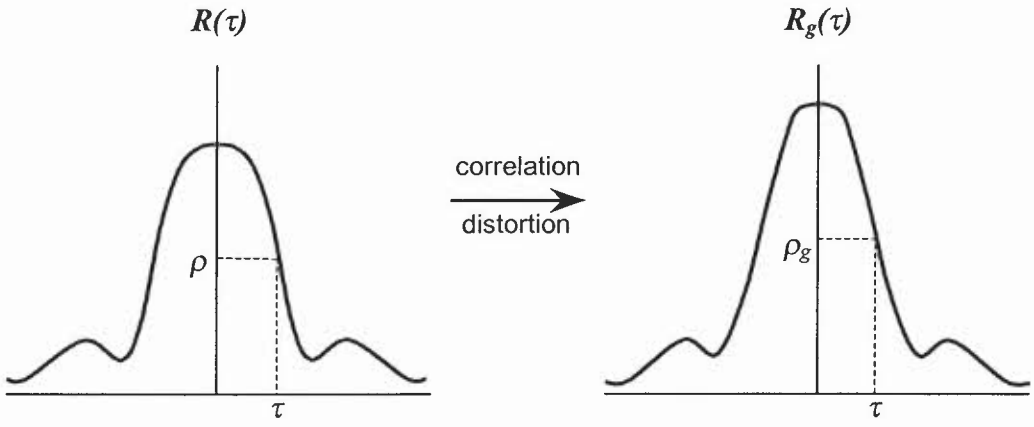
Thus consider discretising with a rectangular grid in the (x', y') co-ordinate system and then immediately converting back to the regular (x, y) co-ordinates for calculation of the joint Gaussian pdf values, etc. Note that the grid is no longer square (except for the special case when $r = 0$). We use numerical values evaluated at the *mid-point* of each grid cell. A large proportion of the joint Gaussian distribution will be considered by selecting the grid to cover the region up to ± 4 standard deviations along the principal axes. If we have (n_x, n_y) intervals in the discretisation then the required grid cell lengths $(\Delta x, \Delta y)$ are given by

$$\Delta x = \frac{8}{n_x}\sqrt{1+r}\,, \quad \Delta y = \frac{8}{n_y}\sqrt{1-r}\,.$$



Appendix C: The distortion of the auto-correlation function

This appendix discusses the calculation of the correlation-distortion at the heart of the ensemble simulation technique. Here the aim is to compute the auto-correlation function (or equivalently, energy spectrum) of a Gaussian process which will give the desired auto-correlation function after transforming the time-series to the target pdf.



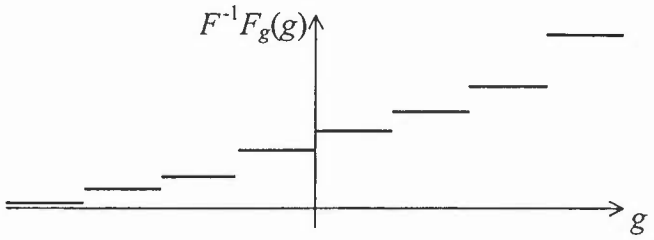
Objective: compute the correlation ρ_g for each given value ρ.

Recall that we consider an unnormalised auto-correlation function R without subtraction of the mean value; that is, $R(\tau) = \overline{c(t)\,c(t+\tau)}$. Suppose that the value of the input auto-correlation function R at some fixed lag time τ is $\rho = R(\tau)$. Then the required value $\rho_g = R_g(\tau)$ for the adjusted correlation at this lag time will satisfy the condition in (3.1), viz.

$$\rho \;=\; \int_x \int_y \big(F^{-1}F_g(x)\big)\big(F^{-1}F_g(y)\big) g_2(\rho_g; x, y)\, dx\, dy .$$

So the problem is to determine the particular value of the correlation ρ_g in the joint Gaussian distribution needed to produce the observed correlation ρ. On a theoretical note, the integral given above for the correlation ρ is a monotonic increasing function in ρ_g. A proof of this intuitive result will be outlined below. The fact that the mapping is monotonic implies that if a solution for ρ_g exists then it will be unique.

The mapping $F^{-1}F_g$ describes the transformation of the probability distribution (from a Gaussian distribution to the target one). This is a monotonic increasing (non-negative) function that can be approximated by a piecewise step function to arbitrary resolution. In fact, if we prescribe a target distribution based on a finite-length (discrete time) concentration time-series, it will directly have this stepped nature.



Then the integral for ρ can be regarded as the limiting value of a summation

$$\sum_{i,j} k_{i,j} \int_{x_i}^{\infty} \int_{y_j}^{\infty} g_2(\rho_g; x, y)\, dx\, dy ,$$

where the coefficients $k_{ij} \geq 0$ are the heights of steps in the resulting two-dimensional step function. Each component integral

$$\int_{x_i}^{\infty} \int_{y_j}^{\infty} g_2(\rho_g; x, y)\, dx\, dy$$

is evaluated over the upper-right quadrant from the point (x_i, y_j). Tong [8, Pg 191] states that this integral is an increasing function of the Gaussian correlation ρ_g. Therefore it follows that the correlation ρ is a monotonic increasing function in ρ_g.

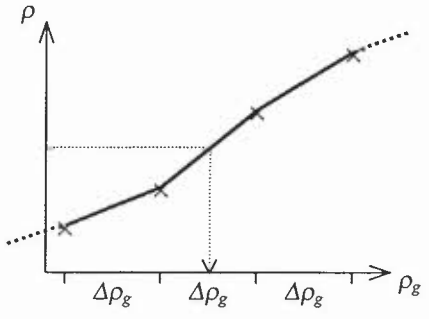
Computational routines are required for the correlation transformations $\rho \leftrightarrow \rho_g$. Here the natural transformation, from a mathematical viewpoint, is the 'forward' mapping $\rho_g \rightarrow \rho$ described in the above formula by a double integration over the joint Gaussian distribution. In practice, this integral can be approximately calculated by a numerical integration scheme, see Appendix B for details. However, it is the 'inverse' transformation $\rho \rightarrow \rho_g$ that is actually needed in the correlation-distortion process. The inverse mapping is calculated indirectly using a numerical inversion scheme.

Although it is fairly straightforward to compute the approximate correlation ρ for any given value of the Gaussian correlation $-1 < \rho_g < +1$, it is still necessary to have an effective strategy for addressing the inverse problem. One possible approach would be an iterative scheme that gradually homes-in on the required solution ρ_g. However, since a large collection of correlation values need to be inverted, it was decided that a global approach to the problem would be more appropriate. Thus an inversion method based on linear interpolation between sampling points has been developed.

The inversion scheme first calculates the forward transformation $\rho_g \rightarrow \rho$ at a collection of sample values for the Gaussian correlation $-1 < \rho_g < +1$ and stores the results in a 'look-up' table in preparation for the inversion process. The sampling points have been chosen at equally-spaced values of ρ_g extending from +1 to –1 such that the total number of bands is an integer power of two. The selection of 2^k bands is not absolutely essential here, but it does allow us to implement an efficient bisection search routine for isolating the band containing a given correlation value ρ.

ρ_g	$\rightarrow$	ρ
1.00		ρ_{max}
-1.00		ρ_{min}

A 'look-up' table of correlation values for the inversion process



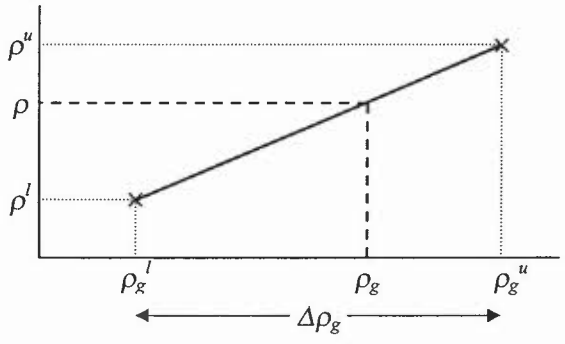
The inversion step uses linear interpolation within each band

The numerical inversion uses the stored array of sample points for the correlation mapping to evaluate an approximate solution of the equation (3.1) for any given value of the correlation ρ. Firstly, it searches through the look-up table to locate the band containing ρ (using the bisection search routine). If the value ρ appears as an entry in the array then we take the corresponding value for the Gaussian correlation ρ_g.

Otherwise, let us apply linear interpolation across the relevant band to calculate the approximate Gaussian correlation ρ_g. The interpolated estimate for ρ_g based on the two sample points at the boundaries of this band is

$$\rho_g \approx \rho_g^l + (\rho_g{}^u - \rho_g{}^l)\frac{(\rho - \rho^l)}{(\rho^u - \rho^l)}$$

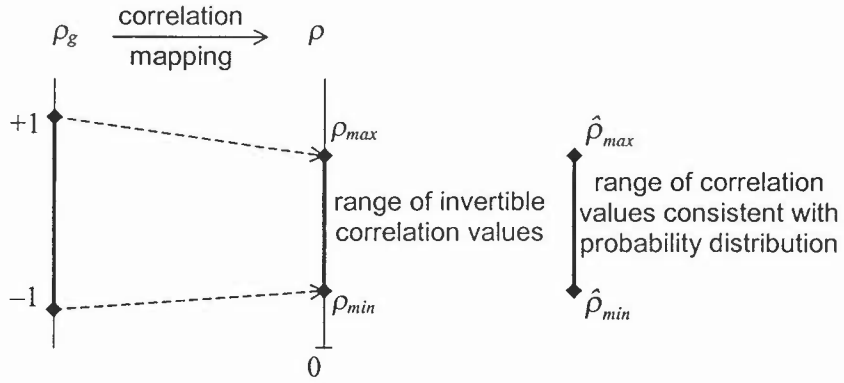
where $(\rho_g{}^l, \rho^l)$ and $(\rho_g{}^u, \rho^u)$ denote the co-ordinates of the lower and upper sampling points for the band respectively.



Note that this numerical inversion technique should yield a reasonably accurate approximation for the Gaussian correlation ρ_g provided that the correlation look-up table has a sufficiently fine resolution (that is, that there are an adequate number of sampling points for the forward correlation mapping). In practice, reasonable results are obtained using $2^6 = 64$ interpolation bands for the numerical inversion.

Next we consider several important issues concerning the correlation-distortion and subsequent calculation of the Gaussian spectrum. It is by no means obvious that equation (3.1) can always be solved for ρ_g; i.e. there may not exist a value $-1 \le \rho_g \le 1$ that gives an observed correlation value ρ. However, we show below that a solution is always possible whenever the auto-correlation function and probability distribution are consistent (in the sense that they are jointly realisable in a time-series).

Let ρ_{max} and ρ_{min} be the images of +1 and –1, respectively, under the correlation mapping $\rho_g \rightarrow \rho$ in (3.1). Similarly, let $\hat{\rho}_{max}$ and $\hat{\rho}_{min}$ denote the theoretical maximum and minimum correlation values, respectively, consistent with the input probability distribution. We prove that these theoretical bounds are attainable by our distortion approach. In particular, we will establish the equalities $\rho_{min} = \hat{\rho}_{min}$ and $\rho_{max} = \hat{\rho}_{max}$.



For example, let us consider the lower limit $\hat{\rho}_{min}$. The lowest value of $\overline{c_1 c_2}$ that is possible for two random variables c_1, c_2 with given marginal probability distributions occurs when c_1 is a monotonically decreasing (deterministic) function of c_2. This can be achieved by making c_1 and c_2 monotonically increasing functions of two perfectly anti-correlated Gaussian random variables g_1 and g_2 (that is, $g_1 = -g_2$). Hence the lowest value of $\overline{c_1 c_2}$ consistent with the probability distributions of c_1 and c_2 is achievable by our transform approach.

A similar argument can be applied to the upper limit $\hat{\rho}_{max}$ where c_1 and c_2 should now be monotonically increasing functions of two perfectly correlated Gaussian random variables g_1 and g_2 (that is, $g_1 = g_2$). A formal proof of these results is now presented.

Let us write $c(t)$ and $c(t+\tau)$ as c_1 and c_2, respectively. For the random variables c_1, c_2 with cumulative distribution function $F(c)$, the maximum correlation value $\hat{\rho}_{max}$ (that is, the maximum value of $\overline{c_1 c_2}$) occurs when $c_1 = c_2$. Then

$$\hat{\rho}_{max} = \int_{0^-}^{\infty} c^2 f(c)\, dc\,.$$

Similarly, the minimum value $\hat{\rho}_{min}$ (that is, the minimum value of $\overline{c_1 c_2}$) occurs when $c_1 = F^{-1}(1 - F(c_2))$; that is, c_1 is a monotonic decreasing function of c_2. Then

$$\hat{\rho}_{min} = \int_{0^-}^{\infty} c\big(F^{-1}(1 - F(c))\big) f(c)\, dc\,.$$

On the other hand, since the mapping $\rho_g \rightarrow \rho$ is monotonic increasing, the range of invertible correlation values can be obtained by directly evaluating the integral in (3.1) for the extreme values of Gaussian correlation $\rho_g = \pm 1$. Then

$$\begin{aligned}
\rho_{max} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \big(F^{-1}F_g(g_1)\big)\big(F^{-1}F_g(g_2)\big) f_g(g_1)\delta(g_1 - g_2)\, dg_1 dg_2 \\
&= \int_{-\infty}^{\infty} \big(F^{-1}F_g(g_1)\big)^2 f_g(g_1)\, dg_1 \\
&= \int_0^1 \big(F^{-1}(x)\big)^2\, dx \\
&= \int_{0^-}^{\infty} c^2 f(c)\, dc = \hat{\rho}_{max}
\end{aligned}$$

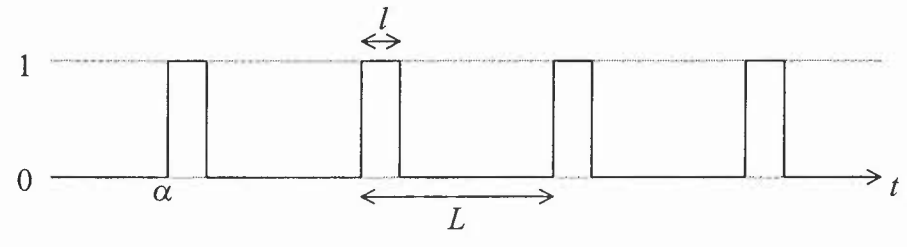
and

$$\begin{aligned}
\rho_{min} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \big(F^{-1}F_g(g_1)\big)\big(F^{-1}F_g(g_2)\big) f_g(g_1)\delta(g_1 + g_2)\, dg_1 dg_2 \\
&= \int_{-\infty}^{\infty} \big(F^{-1}F_g(g_1)\big)\big(F^{-1}F_g(-g_1)\big) f_g(g_1)\, dg_1 \\
&= \int_{-\infty}^{\infty} \big(F^{-1}F_g(g_1)\big)\big(F^{-1}(1 - F_g(g_1))\big) f_g(g_1)\, dg_1 \\
&= \int_0^1 \big(F^{-1}(x)\big)\big(F^{-1}(1-x)\big)\, dx \\
&= \int_{0^-}^{\infty} c\big(F^{-1}(1 - F(c))\big) f(c)\, dc = \hat{\rho}_{min}.
\end{aligned}$$

The above result shows that the correlation-distortion process is always possible provided that the target probability distribution and spectrum are consistent. In fact, this will be the case if a raw time-series realisation is used to prescribe these input functions. However, a raw spectrum is typically rather noisy in appearance and this can create its own problems, such as the spectral truncation issue mentioned in Section 3.4. Therefore smoothed or idealised spectra are frequently adopted, but then there is no guarantee that each value ρ of the auto-correlation function will lie in the invertible range $[\rho_{min}, \rho_{max}]$ such that (3.1) will always be solvable for ρ_g. There is no problem with the upper limit ρ_{max} provided that the input probability distribution and spectrum imply the same mean and variance (because $\rho_{max} = \overline{c^2}$). However, there is a potential problem with the lower limit ρ_{min}. The auto-correlation function will have to be truncated if necessary so that each value ρ occupies the invertible range $[\rho_{min}, \rho_{max}]$.

We have shown that it is always possible to solve equation (3.1) for ρ_g whenever the target probability distribution and energy spectrum are consistent. Unfortunately, even in this case, we cannot guarantee that all components in the associated Gaussian spectrum S_g will be non-negative. This is illustrated by the following counterexample. In practice, any negative spectral components in S_g can be clipped to zero. The overall effect on the simulation process is generally small provided that the truncation is not too severe.

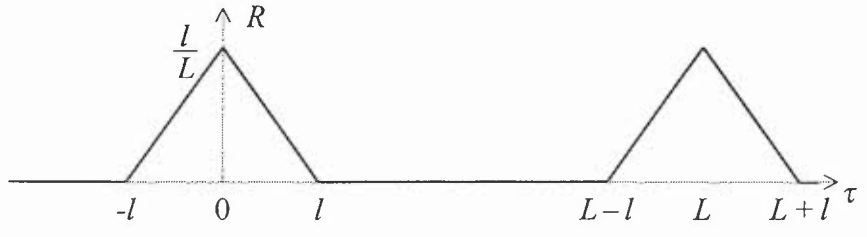
Consider an ensemble of periodic time-series $c_\alpha(t)$ of the form



where l, L are fixed parameters such that $L > 3l$, and α is a random displacement for each realisation in the ensemble. Here each time-series has two states (that is, it alternates between the values zero and one), and the probability distribution of the ensemble is described by

$$p(c=0) = 1 - l/L, \quad p(c=1) = l/L.$$

The auto-correlation function $R(\tau) = \overline{c_\alpha(t)\, c_\alpha(t+\tau)}$ (here the over-bar denotes ensemble average, or equivalently, average over t) has a saw-tooth form.



Let us assume that the concentration time-series $c_\alpha(t)$ can be constructed via a monotonic transformation of a Gaussian process $g_\alpha(t)$. Then

$$c_\alpha(t) = \begin{cases} 1, & \text{if } g_\alpha(t) > X; \\ 0, & \text{otherwise;} \end{cases}$$

where $F_g(X) = 1 - l/L$. Note that $X > 0$. For the prescribed probability distribution, the relation (3.1) becomes

$$\rho = \int_X^\infty \int_X^\infty g_2(\rho_g; x, y)\, dx\, dy.$$

It is clear that a correlation value $\rho = 0$ implies that $\rho_g = -1$ (otherwise the right hand side above would be a non-zero integral). In particular, this gives

$$R_g(\tau) = -1 \quad \text{for } \tau \in [l, L-l].$$

Now consider the Gaussian time-series at the three points $g_\alpha(0)$, $g_\alpha(l)$ and $g_\alpha(2l)$. By construction, $L > 3l$ so that $R_g(l) = R_g(2l) = -1$. Therefore these three points are mutually anti-correlated. This is a logical contradiction. Since the Gaussian correlation function is not valid, we conclude that it is not possible to generate $c_\alpha(t)$ by transforming a Gaussian time-series monotonically. Furthermore a standard result indicates that possible correlation functions are precisely those functions with non-negative Fourier transforms (i.e. spectra). The fact that our correlation function is not valid establishes that the Gaussian spectrum has negative components.

Appendix D: An approximation for the Gaussian cumulative distribution function F_g

The standard Gaussian distribution appears in a quite natural way in the time-series simulation process. In particular, a description of its cumulative distribution function F_g is required for the transformation of a Gaussian series to the prescribed probability distribution. This appendix presents some established approximations for the Gaussian cumulative distribution function F_g taken from Abramowitz and Stegun [9, Pg 932]. Two formulae are considered in the present report to provide approximations with different levels of accuracy.

Our operational scheme adopts the first formula (the product of a degree three polynomial with an exponential term) and combines good accuracy with computational efficiency. The second formula is a higher degree approximation that would give greater accuracy but at the expense of a slower calculation. It is presented here in the case that greater accuracy is required for the computational scheme. Two further approximations are available in [9] using rational expressions (no exponential terms).

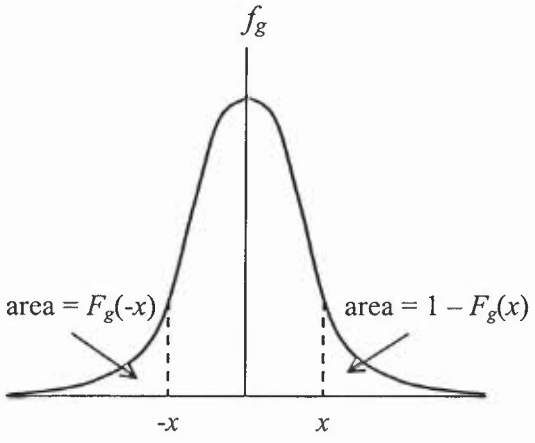
Consider a (standardised) Gaussian random variable X with zero mean and unit variance. The probability density function $f_g(x)$ is defined by

$$f_g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty$$

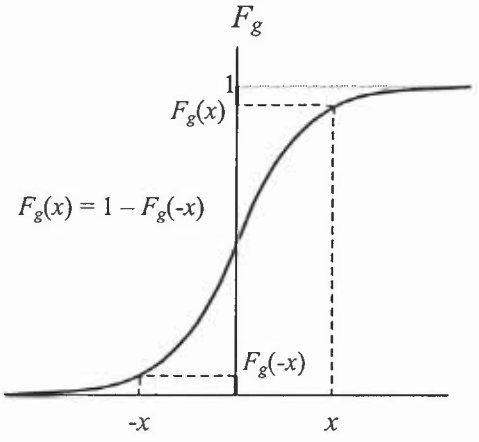
and the cumulative distribution function $F_g(x)$ is given by

$$F_g(x) = prob(X \le x)$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy, \quad -\infty < x < \infty.$$

These two functions have the following graphical form.



probability density function



cumulative distribution function

The Gaussian pdf $f_g(x)$ has an explicit form that can be easily calculated for any particular value of x. Each cumulative probability can be calculated by evaluating the appropriate definite integral. Numerical tables can be found in many reference books that tabulate approximate values for the Gaussian cdf.

Analytic approximations to the function F_g

Our simulation program adopts the following approximation for F_g given in [9]:

$$F_g(x) \approx \begin{cases} 1-P(x), & x \geq 0; \\ P(-x), & x < 0. \end{cases}$$

Here the function $P(x)$ is defined by

$$P(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \left(a_1 t + a_2 t^2 + a_3 t^3\right)$$

where $t = \frac{1}{1+px}$, $(x \geq 0)$ with numerical coefficients $p = 0.33267$ and $a_1 = 0.4361836$, $a_2 = -0.1201676$, $a_3 = 0.9372980$. The error in the approximation is less than 1.0×10^{-5}.

Greater accuracy could be obtained using the higher order formula

$$P(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \left(b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5\right)$$

where $t = \frac{1}{1+px}$, $(x \geq 0)$ with the coefficients $p = 0.2316419$ and $b_1 = 0.319381530$, $b_2 = -0.356563782$, $b_3 = 1.781477937$, $b_4 = -1.821255978$, $b_5 = 1.330274429$. Here the error is reduced to less than 7.5×10^{-8}.

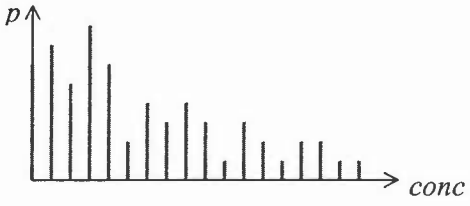
Appendix E: Computational procedures for the cumulative distribution function *F*

E1) Fitting a continuous cdf to a discrete probability distribution from field data

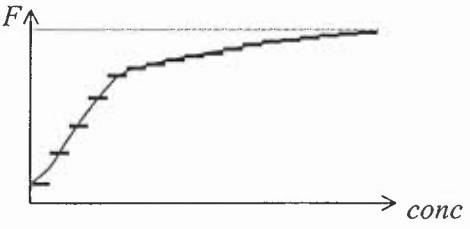
The ensemble simulation scheme requires a continuous probability distribution as one of its prescribed inputs. This section describes a methodology for obtaining a continuous cumulative distribution function F from a discrete time-series data set.

It is worth mentioning that the concentration time-series data recorded in our field experiments exhibits discrete quantisation levels produced by the 12-bit analogue-to-digital conversion used in the logging procedure. That is, the actual concentration recorded by the detector (represented as an analogue voltage signal) is approximated by the nearest quantisation level in the digital output. However, this rounding error is typically very small (at the level of detector noise) and has no significant implications for the quality of the output data.

The discretisation of the data at 12-bit resolution admits 4096 possible values for the recorded concentration at each time step. However, in practice, the technical need to maintain a positive offset voltage during the data-logging reduces this theoretical number of quantisation levels. Consequently the observed concentration time-series, rather than having a continuous distribution, actually has this discrete probability distribution with a large number of possible states.



By regarding each state as representing values distributed within its quantisation band, this discrete distribution approximates the actual probability distribution of the concentration fluctuations. Now the aim is to fit a cumulative distribution function F to this discrete probability distribution. One simple approach would be to adopt a step-wise function that increases in discrete steps at each quantisation level.



However, this gives the target data a slightly artificial appearance with sudden jumps occuring between quantisation levels, although the effect of quantisation of concentration values is not especially significant here. Using this discrete distribution in the simulation process would recreate the quantisation of concentration values in the simulated time-series. This constraint can be easily avoided by constructing a continuous function F interpolating linearly across each band. Since the concentration

values should have a near uniform distribution within each quantisation interval then it is the mid-point probability that is the most appropriate choice for the mid-point concentration (the quantisation value).

Note that the compatibility between the probability distribution and energy spectrum may be affected whenever either characteristic is prescribed by processed (that is, smoothed) experimental data or an idealised description. The averaging and interpolation of data in the probability distribution here, like smoothing of the energy spectrum, could produce inconsistent target functions with its consequences for the simulation process (this is potentially another reason why the auto-correlation function ρ may need truncating; see Appendix C).

The above procedure is reasonable, in principle, but the large number of values in the discrete distribution (the quantisation levels) is rather restrictive in practice. Firstly the data is rather noisy at this scale and some averaging would have a beneficial effect in obtaining a representative probability distribution. Secondly, a large number of discrete levels puts a high demand on the computational resources, so that a reduction in the resolution will improve the efficiency of the various computational routines. The following revised method provides an efficient mechanism for constructing a continuous cumulative distribution function F based on the input time-series from a dispersion experiment.

Our CDF scheme

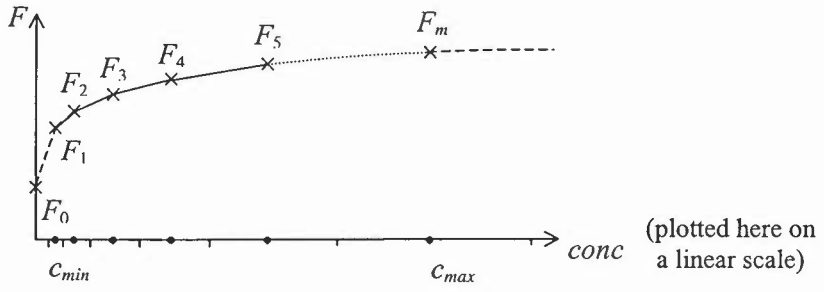
As a first step, let us construct a discrete probability distribution with a prescribed resolution by dividing the non-zero concentration range into contiguous bands and then counting the number of data points occurring in each interval. Suppose that there are m bands in this discretisation; for later convenience, we take $m = 2^k$ where k is an integer. Meanwhile the number of zero values, describing the intermittency of the time-series, is stored separately.

We only considered discretising the experimental data using equal-width intervals on a logarithmic scale, although clearly other options are possible here. Specifically, the approach in our implementation of the scheme constructs the required number of discretisation points over the observed range of concentration values $[c_{min}, c_{max}]$, where c_{min} and c_{max} denote the minimum and maximum values, respectively, of the non-zero concentration data. Here the data points are considered on a logarithmic scale with the mid-points of equal-width bands chosen as the discretisation points.

We found that $2^6 = 64$ bands in this construction gave a reasonable description of the cumulative distribution function F. The above resolution offers a good compromise between an appropriate level of smoothing of the input values and preserving some of the detailed structure of the probability distribution. Furthermore, the computational routine for evaluating the inverse function F^{-1} based on this description has good efficiency, supporting its frequent application in the simulation scheme.

Further discretisation techniques could be considered; for instance, adopting some variable-width bins to count the time-series data points provides an alternative scheme that might give better results. In particular, the bin sizes could then be adjusted to match the input data set. However, such an approach has not been considered here. In all cases, we should ensure that the discretisation bands have sufficient extent to cover the full range of observed concentration values.

The number of data points occupying each bin is counted and the proportion of data occurring in each band is then calculated by dividing this value by the total number of data points in the sample. As in the earlier approach, we adopt the mid-point value of the probability associated with each band in the description of the cumulative distribution function F. The values of the function F at the sample concentration points in our discretisation are held in a real array (F_0, F_1, …, F_m), where F_0 denotes the intermittency of the distribution (that is, the proportion of zeros) and F_i ($i = 1, \ldots, m$) are the calculated values of F at the discretisation mid-points.



Finally, use linear interpolation between these discrete values to give a continuous cumulative distribution function F defined from zero to the mid-point of the upper discretisation band. If $F_m = 1$ then this provides a complete description of the cdf since the distribution is bounded above by the upper band so that greater values for the concentration are not possible. However, to avoid this rather unnatural restriction on the concentration values, it is desirable to consider a discretisation for which $F_m \neq 1$ (since the mid-point probability of the upper band will be marginally less than one). This allows a probability distribution with an upper tail.

To construct the upper tail, we assume that the exceedence of any particular concentration threshold in the upper tail is governed by an exponential decay law. This simple approach appears to agree reasonably well with the upper tails observed in many of the Cardington experiments. Using the intermittency and peak value of the distribution, an inverse exponential tail can then be extrapolated by the formula

$$F(c) = 1 - a\,e^{-bc} \quad \text{for } c > c_{\max} \tag{E1.1}$$

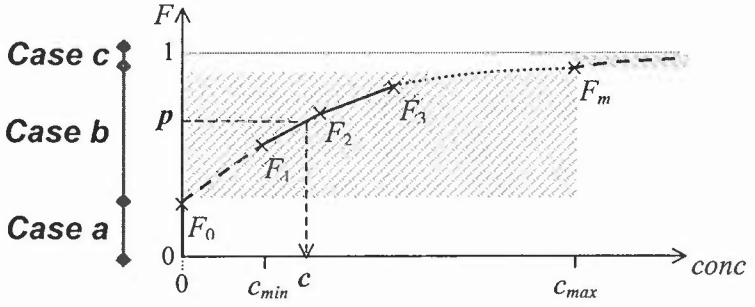
where the constants a, b are chosen to fit the profile to the two reference points $(0, F_0)$ and (c_{max}, F_m). Recall that $F_m = F(c_{max})$ for our discretisation. It can be shown that

$$a = 1 - F_0, \quad b = \frac{1}{c_{\max}} \log_e\left(\frac{1-F_0}{1-F_m}\right).$$

E2) Calculating the inverse cdf using a numerical inversion scheme

The first part of this appendix presented a method for producing a (continuous) cumulative distribution function F from any given (discrete) time-series data set. However, in our simulation scheme, we are actually interested in calculating the inverse function F^{-1}; that is, given any probability $0 < p < 1$, what is the concentration threshold c such that $p = F(c)$? In other words, we require a routine for calculating the value $c = F^{-1}(p)$.

The inverse function is evaluated by a numerical scheme that uses interpolation based on the stored array (F_0, F_1, …, F_m) of cdf values at the discretisation points. The method is essentially the same as the one outlined in Appendix C for the numerical inversion of the correlation-distortion transformation. One noticeable difference, however, is that the interpolation regime is now restricted to probabilities $F_0 \le p \le F_m$. Values of probability p outside this interval are also possible so that, as a prerequisite step, a given value $0 < p < 1$ should be inspected to determine the appropriate regime.



a) For any value of probability less than the intermittency (that is, $p < F_0$) then the routine should return the value zero.

b) For any probability $F_0 \le p \le F_m$ occupying the interpolation range, a numerical inversion technique provides a good approximation for the inverse value $F^{-1}(p)$. Here the same approach can be applied as that used in Appendix C for inverting the correlation mapping $\rho_g \rightarrow \rho$.

c) For any probability $p > F_m$ in the upper tail of the probability distribution, an analytic inversion of the extrapolation formula (E1.1) gives an expression for the concentration threshold $c = F^{-1}(p)$:

$$c = \frac{1}{b}\log_e\left(\frac{a}{1-p}\right)$$

$$= c_{max}\log_e\left(\frac{1-F_0}{1-p}\right)\Big/\log_e\left(\frac{1-F_0}{1-F_m}\right) \quad \text{for } p > F_m .$$