Computer Physics Communications 204 (2016) 152–158

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Conditions for generating synthetic data to investigate characteristics of fluctuating quantities

Jaewook Kim [a], M.F.J. Fox [b,c,d], A.R. Field [b], Y.U. Nam [e], Y.-c. Ghim [a,*]

[a] Department of Nuclear and Quantum Engineering, KAIST, Daejeon 34141, Republic of Korea
[b] EURATOM/CCFE Fusion Association, Culham Science Centre, Abingdon, OX14 3DB, UK
[c] Rudolf Peierls Centre for Theoretical Physics, University of Oxford, Oxford, OX1 3NP, UK
[d] Merton College, Oxford, OX1 4JD, UK
[e] National Fusion Research Institute, Daejeon, Republic of Korea

ARTICLE INFO

Article history:
Received 24 November 2015
Received in revised form
31 March 2016
Accepted 12 April 2016
Available online 19 April 2016

Keywords:
Synthetic data
Numerical domain
Turbulence
Statistical analysis

ABSTRACT

Synthetic data describing coherent random fluctuations have widely been used to validate numerical simulations against experimental observations or to examine the reliability of extracting statistical properties of plasma turbulence via correlation functions. Estimating correlation time or lengths based on correlation functions implicitly assumes that the observed data are *stationary* and *homogeneous*. It is, therefore, important that numerically generated synthetic data also satisfy the stationary process and homogeneous state. Based on the synthetic data with randomly generated moving Gaussian shaped fluctuations both in time and space, the correlation function depending on the size of averaging time window is analytically derived. Then, the smallest possible spatial window size of synthetic data satisfying the stationary process and homogeneous state is proposed, thereby reducing the computation time to generate proper synthetic data and providing a constraint on the minimum size of simulation domains when using synthetic diagnostics to compare with experiment. This window size is also numerically confirmed with 1D synthetic data with various parameter scans.

1. Introduction

Not being deterministic, turbulent structures must be studied based on the statistical grounds. Therefore, developing reliable statistical analyses to extract turbulence characteristics from the measured data is of paramount importance. For example, correlation functions can estimate correlation time and lengths of the turbulence, and the cross-correlation time delay method allows us to measure the velocity of pattern flows [1–3].

As numerical simulations and experimental diagnostics on plasma turbulence become more sophisticated, synthetic turbulence data generated from the simulations have been used to compare the results from simulations and experiments directly [4–6]. Turbulence synthetic data can also be used to examine the reliability of statistical techniques used to extract turbulence characteristics [1,7–9], i.e., turbulence characteristics extracted from the synthetic data using a statistical technique can be compared with the input parameters generating the synthetic data.

Validating and improving statistical analyses and physical model of turbulence using synthetic data may unveil many unknown factors associated with turbulence. For example, as the turbulence driven transport in a magnetically confined plasma exceeds the neoclassical transport level by at least an order of magnitude [10], it is desirable to suppress the turbulence. For this purpose, we wish to understand the basic characteristics of the turbulence such as decorrelation rate and correlation lengths, and to perceive how they are correlated with equilibrium quantities, how they react back to these equilibrium quantities, and hopefully how they might be controlled [11–18], and synthetic data is one of the tools that can enhance our physics understanding of turbulence.

The property of synthetic data themselves has not been thoroughly investigated so far. For instance, as estimating correlation time and lengths using correlation functions from the measured data implicitly assumes that the data are stationary and homogeneous, synthetic data must also comply with the conditions of stationary process and homogeneous state. Stationary process means

* Corresponding author.
E-mail addresses: ijwkim@kaist.ac.kr (J. Kim), ycghim@kaist.ac.kr (Y.-c. Ghim).

http://dx.doi.org/10.1016/j.cpc.2016.04.004

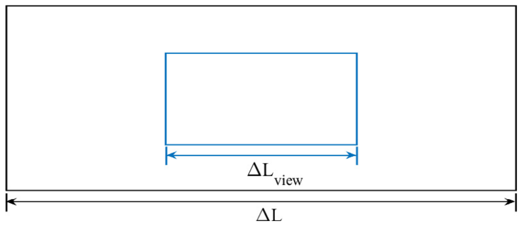

Fig. 1. A diagram depicting the total simulation domain ΔL and a smaller 'viewing' domain (domain of interest) ΔL_{view} where the generated synthetic data are stationary and homogeneous. Outside ΔL_{view} the synthetic data may become non-stationary and/or non-homogeneous depending on how they are generated.

that low moments of fluctuating data such as mean and variance do not vary with time; while if they are unchanged in space, then the data are said to be homogeneous. To generate 'true' stationary and homogeneous synthetic data, the simulation domain has to be infinitely large due to the finite correlation time and lengths of turbulent eddies. This is impractical. In practice, turbulent structures, or 'eddies', are generated within a finite spatial domain and temporal domain. Therefore, for eddies which have a finite spatial and temporal extent, there are no sources from outside of these domains that contribute to the response within the domain (assuming that the boundary conditions are not periodic). Hence, these cause a spatial (and/or temporal) variation that leads to an inhomogeneous (non-stationary) correlation function.

In this paper, we thus provide the minimal size of required simulation domain ΔL given the 'viewing' domain (domain of interest) ΔL_{view} upon where one would apply statistical analyses as shown in Fig. 1. This means that generated synthetic data within ΔL_{view} must be stationary and homogeneous, otherwise statistically calculated correlation functions may give us incorrect results. Of course, we wish to find the minimal ΔL so that we do not waste our computation resource. Or, for the case of local gyro-kinetic (GK) simulations where simulation domains ΔL are set, we provide the maximum possible ΔL_{view} where the synthetic data can be valid for direct comparisons with experimental observations.

We first describe the mathematical model of a fluctuating quantity, or 'eddy', such as density, temperature or potential in Section 2 and analytically derive correlation functions assuming that eddies are uniformly distributed in an infinitely large domain. In Section 3, we provide the condition on the total simulation domain ΔL as a function of the 'viewing' domain ΔL_{view} and the size of the turbulent eddies, based on the derived correlation function such that the generated synthetic data satisfy stationarity and homogeneity. This condition is verified numerically using the 1D (in space) fluctuating synthetic data with various parameter scans. Note that even though we use 1D synthetic data, our arguments can be generalized to 3D as long as the basis vectors are orthogonal to each other. Our conclusion follows in Section 4.

2. Correlation function of 'eddies'

2.1. Mathematical model of 'eddies'

In this section, we introduce a mathematical model describing real fluctuations as an ensemble of 'eddies' – its definition will follow soon – based on which we derive the correlation function and generate synthetic data [1,7,8]. For simplicity we model the fluctuations in a 1D spatial domain. We represent our data at the spatial location $x = x_a$ as a function of time as

$$S_a(t) = \sum_{i=1}^{N} S_{a_i}(t), \tag{1}$$

where $S_{a_i}(t)$ is the ith 'eddy', and N is the total number of eddies generated in the synthetic data.

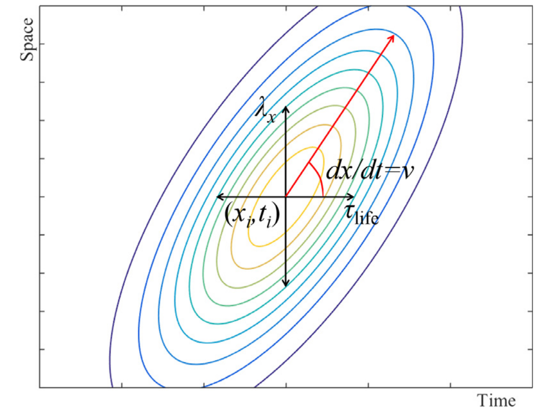

Fig. 2. An example of the contour of a single eddy in the space (ordinate) and time (abscissa) coordinate. The correlation length (λ_x) and time (τ_{life}) in Eq. (2) are also depicted. The slope of the red line is the velocity of the eddy. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We have many different possibilities on what mathematical form $S_{a_i}(t)$ would take. Inspired by the experimental observations on ion-scale density fluctuations [3,19], we model that eddies are Gaussian shaped in both time and space:

$$S_{a_i}(t) = A_i \exp\left[-\frac{(t-t_i)^2}{2\tau_{\text{life}}^2} - \frac{(x_a - v(t-t_i) - x_i)^2}{2\lambda_x^2}\right]. \tag{2}$$

Coherent properties of each eddy in space and time are parameterized by the characteristic spatial scale (λ_x) and the characteristic temporal scale (τ_{life}). The ith eddy has a maximum amplitude A_i at $x = x_i$ and $t = t_i$. Further, we allow an eddy to move with the velocity of v. Note that our model eddy does not contain the wave-like structures [1], and we justify it by arguing that we are primarily interested in the envelope of eddies. The envelopes can be readily obtained from the measured data by invoking Hilbert transform [20]. Here, A_i is selected from a normal distribution with zero mean and variance of A^2; whereas x_i and t_i are randomly selected from uniform distributions:

$$\begin{aligned}
P(t_i) &= \begin{cases} \dfrac{1}{\Delta T} & \text{if } -\dfrac{\Delta T}{2} \le t_i \le \dfrac{\Delta T}{2} \\ 0 & \text{otherwise} \end{cases} \\
P(x_i) &= \begin{cases} \dfrac{1}{\Delta L} & \text{if } -\dfrac{\Delta L}{2} \le x_i \le \dfrac{\Delta L}{2} \\ 0 & \text{otherwise} \end{cases} \\
P(A_i) &= \frac{1}{\sqrt{2\pi}A} \exp\left[-\frac{A_i^2}{2A^2}\right],
\end{aligned} \tag{3}$$

where $P(t_i)$, $P(x_i)$ and $P(A_i)$ are the probabilities of obtaining t_i, x_i and A_i, respectively. ΔT and ΔL are the total simulation domains in time and space, respectively (as in Fig. 1 for ΔL). Furthermore, to make sure that eddies do not occur too frequently or too rarely, we define a spatio-temporal filling factor F [1]. We determine the total number of eddies (N) generated in a set of synthetic data such that the following expression is satisfied:

$$F = N\left(\frac{\lambda_x}{\Delta L}\right)\left(\frac{\tau_{\text{life}}}{\Delta T}\right) \sim \mathcal{O}(1). \tag{4}$$

Fig. 2 shows an example of the contour of a generated eddy in the spatial and temporal coordinates.

2.2. Correlation function of stationary and homogeneous fluctuating data

As many kinds of statistical analyses are performed on the data based on the stationary and homogeneous assumptions, we