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, 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*

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.

Keywords: Synthetic data; Numerical domain; Turbulence; Statistical analysis

1. Introduction

As the turbulence driven transport in a magnetically confined plasma exceeds the neoclassical transport level by at least an order of magnitude [1], 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 [2–9]. 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 [10–12].

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 [13–15]. Turbulence synthetic data can also be used to examine the reliability of statistical techniques used to extract turbulence characteristics [10, 16–18], i.e., turbulence characteristics extracted from the synthetic data using a statistical technique can be compared with the input parameters generating the synthetic data.

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 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 analy-

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

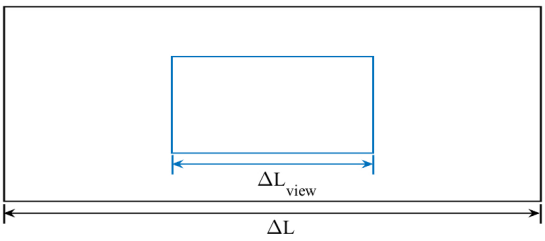

Figure 1: A diagram depicting the total simulation domain ΔL and a smaller 'viewing' domain (domain of interest) $\Delta L_{\rm view}$ where the generated synthetic data are stationary and homogeneous. Outside $\Delta L_{\rm view}$ the synthetic data may become non-stationary and/or non-homogeneous depending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

ses as shown in Fig. 1. This means that generated synthetic data within $\Delta L_{\rm view}$ 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 $\Delta L_{\rm view}$ 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 Sec. 2 and analytically derive correlation functions assuming that eddies are uniformly distributed in an infinitely large domain. In Sec. 3, we provide the condition on the total simulation domain ΔL as a function of the 'viewing' domain $\Delta L_{\rm view}$ 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 Sec. 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 [10, 16, 17]. 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 $i^{\rm th}$ 'eddy', and N is the total number of eddies generated in the synthetic data.

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 [12, 19], we model that eddies are Gaussian shaped in both time and space:

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

Coherent properties of each eddy in space and time are parameterized by the characteristic spatial scale (λ_x) and the characteristic temporal scale ($\tau_{\rm life}$). The $i^{\rm th}$ 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 [10], and we justify it by arguing that we are primarily interested in the envelope of eddies. 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:

$$P(t_i) = \begin{cases} \dfrac{1}{\Delta T} & \text{if } -\dfrac{\Delta T}{2} \le t_i \le \dfrac{\Delta T}{2} \\ 0 & \text{otherwise} \end{cases}$$

$$P(x_i) = \begin{cases} \dfrac{1}{\Delta L} & \text{if } -\dfrac{\Delta L}{2} \le x_i \le \dfrac{\Delta L}{2} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$P(A_i) = \frac{1}{\sqrt{2\pi}A} \exp\left[-\frac{A_i^2}{2A^2}\right],$$

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 [10]. We determine the total number of eddies (N) generated in a set of synthetic data such that the following expression is satisfied:

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

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 let ΔT and ΔL to be infinite to make sure that our model data are stationary and homogeneous. To analytically calculate the correlation function following Tal et al. [17] between two spatial positions, x_a and x_b, as a function of time delay τ, we average the signals $S_a(t)$ and $S_b(t)$ over the 'subtime' window $\Delta T_{\rm sub}$:

$$\begin{aligned} C_{a,b}(\tau) =& \overline{(S_a(t) - \overline{S_a})(S_b(t+\tau) - \overline{S_b})} \\ \approx& \overline{S_a(t) S_b(t+\tau)} - \overline{S_a}\,\overline{S_b}, \end{aligned} \tag{5}$$

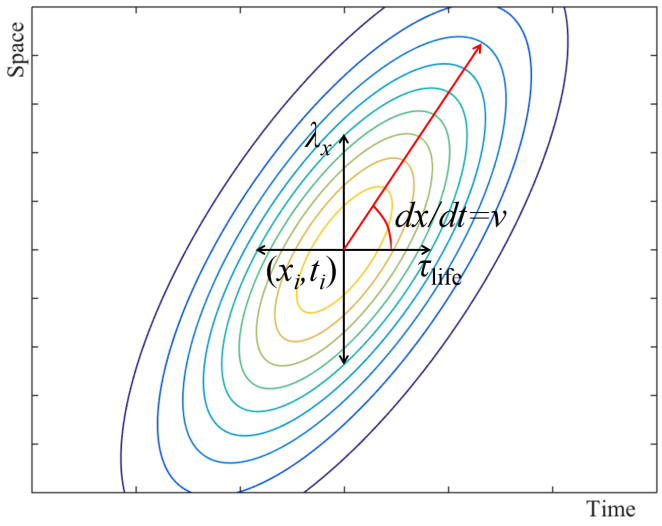

Figure 2: An example of the contour of a single eddy in the space (ordinate) and time (abscissa) coordinate. The correlation length (λ_x) and time ($\tau_{\rm life}$) in Eq. (2) are also depicted. The slope of the red line is the velocity of the eddy.

where the approximation is allowed because the data are stationary [17, 20]. The overline means the time average of the signal over the subtime window $\Delta T_{\rm sub}$:

$$\begin{aligned}\overline{S_{a_i}(t)S_{b_j}(t+\tau)} =& \frac{1}{\Delta T_{\rm sub}}\int_{-\frac{\Delta T_{\rm sub}}{2}+T_{\rm mid}}^{\frac{\Delta T_{\rm sub}}{2}+T_{\rm mid}} S_{a_i}(t)S_{b_j}(t+\tau)dt \\ =& C_{a_i,b_j}(\tau),\end{aligned} \tag{6}$$

where $T_{\rm mid}$ is the time at the middle of the selected subtime window. Note that $C_{a_i,b_j}(\tau) - \overline{S}_{a_i}\,\overline{S}_{b_j}$ is the correlation function between the $i^{\rm th}$ eddy at $x = x_a$ and the $j^{\rm th}$ eddy at $x = x_b$.

As a set of total fluctuation data is the sum of all eddies as in Eq. (1), we can expand $C_{a,b}$ as

$$C_{a,b} = \sum_i C_{a_i,b_i} + \sum_i \sum_{j\neq i} C_{a_i,b_j} - \sum_i \overline{S}_{a_i}\,\overline{S}_{b_i} - \sum_i \sum_{j\neq i} \overline{S}_{a_i}\,\overline{S}_{b_j}. \tag{7}$$

Finally, by averaging correlation functions estimated from many subtime windows, we get the ensemble averaged correlation function as

$$\begin{aligned}\langle C_{a,b}\rangle =& N\langle C_{a_i,b_i}\rangle + N(N-1)\langle C_{a_i,b_j}\rangle \\ & - N\langle \overline{S}_{a_i}\overline{S}_{b_i}\rangle - N(N-1)\langle \overline{S}_{a_i}\rangle\langle \overline{S}_{b_j}\rangle,\end{aligned} \tag{8}$$

where the second and the fourth terms on the right-hand-side cancel out. Furthermore, these two terms are independently zero if the mean of A_i in Eq. (2) is zero.

Eq. (9) shows C_{a_i,b_i} at the time delay $\tau = 0$ as one would do to attain the correlation length from the ensemble averaged correlation function.

$$\begin{aligned}C_{a_i,b_i} =& \frac{1}{\Delta T_{\rm sub}}\int_{-\frac{\Delta T_{\rm sub}}{2}+T_{\rm mid}}^{\frac{\Delta T_{\rm sub}}{2}+T_{\rm mid}} S_{a_i}(t)S_{b_i}(t)dt \\ =& A_i^2\frac{\sqrt{\pi}}{2}\frac{\tau_{\rm ac}}{\Delta T_{\rm sub}}\exp\left[-\frac{\tau_{\rm ac}^2}{2\lambda_x^2/v^2}\frac{(x_a-x_b)^2}{2\lambda_x^2}\right. \\ & \left. -\frac{\tau_{\rm ac}^2}{\tau_{\rm life}^2}\frac{(x_a-x_i)^2+(x_b-x_i)^2}{2\lambda_x^2}\right]\mathcal{W}_{\Delta T_{\rm sub}}(t_i) \\ \approx& \begin{cases} A_i^2\sqrt{\pi}\dfrac{\tau_{\rm ac}}{\Delta T_{\rm sub}}\exp\left[-\dfrac{\tau_{\rm ac}^2}{2\lambda_x^2/v^2}\dfrac{(x_a-x_b)^2}{2\lambda_x^2}\right. \\ \qquad\left. -\dfrac{\tau_{\rm ac}^2}{\tau_{\rm life}^2}\dfrac{(x_a-x_i)^2+(x_b-x_i)^2}{2\lambda_x^2}\right] \\ \quad \text{for } -\dfrac{\Delta T_{\rm sub}}{2}+T_{\rm mid}+\gamma_i \le t_i \le \dfrac{\Delta T_{\rm sub}}{2}+T_{\rm mid}+\gamma_i \\ \quad 0 \qquad\qquad \text{otherwise,}\end{cases}\end{aligned} \tag{9}$$

where $\tau_{\rm ac}$ is the usual auto-correlation time of eddies in the lab frame defined as [16]

$$\tau_{\rm ac} = \frac{\lambda_x\tau_{\rm life}}{\sqrt{\lambda_x^2+\tau_{\rm life}^2 v^2}}, \tag{10}$$

and γ_i is

$$\gamma_i = \frac{\tau_{\rm ac}^2}{\lambda_x^2}v\left[x_i - \frac{x_a+x_b}{2}\right]. \tag{11}$$

Here, $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$ which acts like a weighting factor is a function containing the error function Erf () defined as

$$\begin{aligned}\mathcal{W}_{\Delta T_{\rm sub}}(t_i) =& {\rm Erf}\left[\frac{1}{\tau_{\rm ac}}\left(T_{\rm mid}+\frac{\Delta T_{\rm sub}}{2}-t_i+\gamma_i\right)\right] - \\ & {\rm Erf}\left[\frac{1}{\tau_{\rm ac}}\left(T_{\rm mid}-\frac{\Delta T_{\rm sub}}{2}-t_i+\gamma_i\right)\right].\end{aligned} \tag{12}$$

The approximation in the last step in Eq. (9) is taken by assuming large $\Delta T_{\rm sub}$ such that $\mathcal{W}_{\Delta T_{\rm sub}}(t_i) \approx 2$.

Once we have C_{a_i,b_i}, we calculate $\langle C_{a_i,b_i}\rangle$ by taking an ensemble average with the probability density functions defined in Eq. (3):

$$\begin{aligned}\langle C_{a_ib_i}\rangle =& \int_{-\infty}^{\infty}dA_i\int_{-\infty}^{\infty}dt_i\int_{-\infty}^{\infty}dx_i P(x_i)P(t_i)P(A_i)C_{a_ib_i} \\ =& \left(A^2\frac{\pi}{2}\frac{\tau_{\rm life}}{\Delta T}\frac{\lambda_x}{\Delta L}\right)\exp\left[-\frac{(x_a-x_b)^2}{2\left(\sqrt{2}\lambda_x\right)^2}\right]\mathcal{W}_{\Delta L} \\ \approx& \left(A^2\pi\frac{\tau_{\rm life}}{\Delta T}\frac{\lambda_x}{\Delta L}\right)\exp\left[-\frac{(x_a-x_b)^2}{2\left(\sqrt{2}\lambda_x\right)^2}\right],\end{aligned} \tag{13}$$

where $\mathcal{W}_{\Delta L}$ plays the similar role as $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$ did for Eq. (9) containing the error function Erf() defined as

$$\mathcal{W}_{\Delta L} = \mathrm{Erf}\left[\frac{x_a + x_b + \Delta L}{2\sqrt{\lambda_x^2 + \tau_{\rm life}^2 v^2}}\right] - \mathrm{Erf}\left[\frac{x_a + x_b - \Delta L}{2\sqrt{\lambda_x^2 + \tau_{\rm life}^2 v^2}}\right]. \tag{14}$$

Similar to what we did for $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$, we approximate $\mathcal{W}_{\Delta L} \approx 2$ in Eq. (13) by assuming infinitely large ΔL. Note that we also have large ΔT because $\Delta T \gg \Delta T_{\rm sub}$, and $\Delta T_{\rm sub}$ is assumed to be large from Eq. (9). After another lengthy algebraic calculation, we find that $\left\langle \overline{S}_{a_i} \overline{S}_{b_i} \right\rangle$ in Eq. (8) is

$$\left\langle \overline{S}_{a_i} \overline{S}_{b_i} \right\rangle \approx \left(A^2 \pi \frac{\tau_{\rm life}}{\Delta T} \frac{\lambda_x}{\Delta L}\right)\left(2\sqrt{\pi}\frac{\tau_{\rm ac}}{\Delta T_{sub}}\right)\exp\left[-\frac{(x_a - x_b)^2}{4(\lambda_x^2 + \tau_{life}^2 v^2)}\right], \tag{15}$$

by applying the same assumptions, i.e., large ΔL, ΔT and $\Delta T_{\rm sub}$.

Collecting Eq. (13) and Eq. (15), we finally obtain the ensemble averaged correlation value or the expected correlation value between the signals at $x = x_a$ and x_b at the time delay $\tau = 0$:

$$\begin{aligned}\langle C_{a,b} \rangle \approx & A^2 \pi N \frac{\lambda_x}{\Delta L}\frac{\tau_{\rm life}}{\Delta T}\left(\exp\left[-\frac{(x_a - x_b)^2}{2\left(\sqrt{2}\lambda_x\right)^2}\right]\right. \\ & \left. -2\sqrt{\pi}\frac{\tau_{\rm ac}}{\Delta T_{\rm sub}}\exp\left[-\frac{(x_a - x_b)^2}{4\left(\lambda_x^2 + \tau_{\rm life}^2 v^2\right)}\right]\right) \\ \approx & A^2 \pi \left(\exp\left[-\frac{(x_a - x_b)^2}{2\left(\sqrt{2}\lambda_x\right)^2}\right]\right. \\ & \left. -2\sqrt{\pi}\frac{\tau_{\rm ac}}{\Delta T_{\rm sub}}\exp\left[-\frac{(x_a - x_b)^2}{4\left(\lambda_x^2 + \tau_{\rm life}^2 v^2\right)}\right]\right),\end{aligned} \tag{16}$$

where we use Eq. (4) to get the last line. This is our final form of the ensemble averaged correlation function at the time delay $\tau = 0$ with the assumptions of infinitely large domains and is used for numerical comparisons in Sec. 3.

The first term has a Gaussian form as the shape of an individual eddy is set to be Gaussian (see Eq. (2)). Note that the shape of a correlation function can be determined via a convolution of an individual eddy function with itself. In general, a correlation length is estimated by fitting correlation values $\langle C_{a,b} \rangle$ to a Gaussian function with the knowledge of $x_a - x_b$. Such a fitting procedure, thus, implicitly ignores the second term in Eq. (16) originated from $\left\langle \overline{S}_{a_i} \overline{S}_{b_i} \right\rangle$. Discussing the finite effect of the second term in estimating λ_x is not within the scope of this paper. However, we briefly mention that ignoring the second term can be justified for $\tau_{\rm life}^2/\Delta T_{\rm sub}^2 \ll 1$ given $\lambda_x^2 \gg \tau_{\rm life}^2 v^2$, or for $(\lambda_x/v)^2/\Delta T_{\rm sub}^2 \ll 1$ given $\lambda_x^2 \ll \tau_{\rm life}^2 v^2$. For the case of $\lambda_x^2 \sim \tau_{\rm life}^2 v^2$, the situation becomes a bit more complicated, but large enough $\Delta T_{\rm sub}$ allows one to ignore the second term effect as well. Furthermore, an astute reader may realize that estimating the correlation length by fitting a Gaussian function to the first term overestimates the characteristic spatial scale λ_x by a factor of $\sqrt{2}$. Nonetheless, such an overestimation may not become problematic if one is interested in the 'scaling' of the correlation lengths rather than absolute quantities. In fact, as we do not have the first principle argument on what form of the correlation function, i.e., exponential function, Gaussian function, power-law function, etc, should be fitted to the experimental turbulence data, speaking of an absolute correlation length from the fitting must be done with great care. Perhaps, it is worth to consider the Gaussian process [21] to fit the data since we do not have well defined a prior knowledge on a plasma turbulence model function.

3. Synthetic data

3.1. Conditions for generating stationary and homogeneous synthetic data

In Sec. 2, we have derived the correlation function assuming that arguments inside the Erf() are large enough so that Erf() returns ± 1 depending on the sign of arguments. It is important to realize that this assumption is not used for a mere simplification of equations, rather it is the consequence of data with finite coherent structures, i.e., $\tau_{\rm life} \neq 0$ or $\lambda_x \neq 0$, being homogeneous and stationary. To be quantitative, we argue that $|\mathrm{Erf}(x)| \geq 0.995$ (or $|x| \geq 2$) is the condition satisfying the assumption we have made, i.e., $|\mathrm{Erf}(x)| \to 1$.

First, we find the condition on how large $\Delta T_{\rm sub}$ has to be by referring to Eq. (12) as the $\Delta T_{\rm sub}$ appears inside the Erf():

$$\frac{\left|t_{i,\pm}^* - t_i\right|}{\tau_{\rm ac}} \geq 2, \tag{17}$$

where $t_{i,\pm}^*$ is defined as

$$t_{i,\pm}^* = T_{\rm mid} \pm \frac{\Delta T_{\rm sub}}{2} + \gamma_i, \tag{18}$$

and by definition we have $\Delta T_{\rm sub} = t_{i,+}^* - t_{i,-}^*$. Notice that at $t_i = t_{i,\pm}^*$, we have $\mathcal{W}_{\Delta T_{\rm sub}}(t_i) \approx 1$ as shown in Fig. 3 where the exact $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$ (blue) and its approximation (red), i.e., $\mathcal{W}_{\Delta T_{\rm sub}}(t_i) \approx 2$, are plotted. Considering the exact and approximated C_{a_i,b_i} in Eq. (9), we find that the value of approximated C_{a_i,b_i} underestimates the true value in the green shaded region; while it overestimates in the yellow shaded region in Fig. 3. The width of each shaded region is approximately $2\tau_{\rm ac}$. Thus, to obtain the correct correlation values, it is necessary to have $\Delta T_{\rm sub}$ much larger than $\tau_{\rm ac}$ such that the fraction of under- or over-estimated regions are small. Casting this idea into a quantitative form, it states that

$$\varepsilon \equiv \frac{2\tau_{\rm ac}}{\Delta T_{\rm sub}} \ll 1. \tag{19}$$

It is well known that the size of averaging time window $\Delta T_{\rm sub}$ must be much larger than the auto-correlation time $\tau_{\rm ac}$ to obtain correct correlation functions, and here we have provided a quantitative rationale behind such a criterion. Note that this criterion does not guarantee the stationary process of the data.

It is obvious that the condition of $\Delta T \gg \Delta T_{\rm sub}$ must be well satisfied for the ensemble average defined in Eq. (13), and it

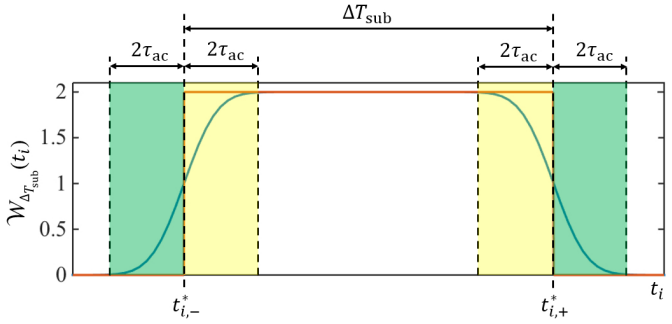

Figure 3: An illustration of exact (blue) and approximated (red) $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$. At $t_i = t^*_{i,\pm}$, we have $\mathcal{W}_{\Delta T_{\rm sub}}(t_i) \approx 1$. In yellow (green) shaded regions whose widths are $2\tau_{\rm ac}$, the approximated $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$ over(under)-estimates the exact value. We wish to minimize the fraction of the shaded regions in $\Delta T_{\rm sub}$.

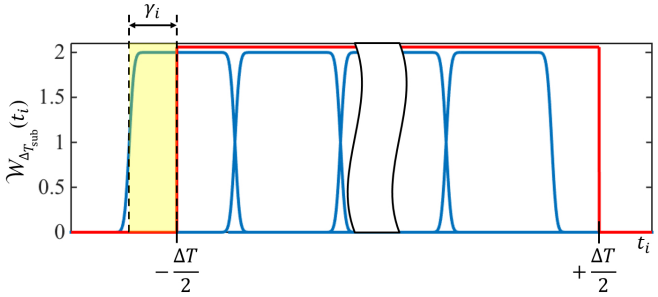

Figure 4: Blue lines show the multiple $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$'s with the widths of $\Delta T_{\rm sub}$ inside the total ΔT window set by the $P(t_i)$ depicted by the red line. All the $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$'s are shifted by an amount of γ_i. The correlation value is underestimated for the far left subtime window containing the yellow shaded region. We wish to have the size of γ_i negligible compared to ΔT.

is typically the case that the synthetic data are generated with a large enough ΔT. However, for the sake of completeness, we provide more quantitative criterion on ΔT such that the true correlation value can be estimated from the synthetic data generated with the 'finite' ΔT. We realize that $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$ is shifted by an amount of γ_i as depicted in Fig. 4 where the blue lines show the shifted $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$, and the red line shows the uniform probability density function of t_i, $P(t_i)$. The $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$ in the middle of $P(t_i)$ would not cause any problems on estimating correlation values, but the $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$ at the edge of $P(t_i)$ may cause the underestimation of the true correlation value. This is because no eddies are generated in the shaded region, i.e., $P(t_i)$=0, while the size of the averaging subtime window is kept to be $\Delta T_{\rm sub}$ as other subtime windows. This violates the condition of data being stationary. To minimize this effect of underestimation we, thus, need to have the maximum of $|\gamma_i|$ much smaller than ΔT.

From the definition of γ_i in Eq. (11), the maximum possible value of $|x_i|$ is the $\Delta L/2$ from the probability density function of x_i, i.e., $P(x_i)$, and that of $|x_a + x_b|$ is $\Delta L_{\rm view}$ since $x = x_a$ and x_b are the 'measurement' positions where we apply statistical analyses. Note that this is valid for the center of $\Delta L_{\rm view}$ coinciding with that of ΔL. Thus, we obtain the condition on ΔT, in addition to $\Delta T_{\rm sub} \ll \Delta T$, such that the synthetic data satisfy the stationary process:

$$\max(|\gamma_i|) = \frac{\tau_{\rm ac}^2}{\lambda_x^2} v \left[\frac{\Delta L}{2} + \frac{\Delta L_{\rm view}}{2}\right] \ll \Delta T \quad \text{and} \quad \Delta T_{\rm sub} \ll \Delta T, \tag{20}$$

which depends on the spatial domain size due to the finite velocity of eddies.

To determine the minimal size of spatial domain ΔL given the 'viewing' domain $\Delta L_{\rm view}$, we need to examine Eq. (13) where the approximation of $\mathcal{W}_{\Delta L} \approx 2$ is supported by assuming large ΔL. Again, this assumption is regarded to be well satisfied if the argument is larger than two:

$$\left|\frac{x_a + x_b \pm \Delta L}{2\sqrt{\lambda_x^2 + \tau_{\rm life}^2 v^2}}\right| \geq 2. \tag{21}$$

Since $\Delta L_{\rm view} \geq |x_a + x_b|$, we find a criterion on ΔL as

$$4\sqrt{\lambda_x^2 + \tau_{\rm life}^2 v^2} + \Delta L_{\rm view} \leq \Delta L, \tag{22}$$

and this condition guarantees that the ensemble averaged correlation values do not depend on the 'measurement' positions, hence the synthetic data satisfy the homogeneous state.

If we delve into the structure of Eq. (13) deeper, one may raise a question: what happens if $\mathcal{W}_{\Delta L}$ is almost constant at a value other than two within the 'viewing' domain $\Delta L_{\rm view}$? If this happens, then we realize that $\langle C_{a_i,b_i} \rangle$ does not depend on the spatial position as long as the distances between the two points, $|x_a - x_b|$, are the same. This consequence may be argued for the data being homogeneous without satisfying the condition on ΔL set by Eq. (22). We provide more detailed explanation on this in Appendix A.

3.2. Simulation results with parameter scans

We now have the conditions on the sizes of temporal (ΔT) and spatial (ΔL) simulation domains given the viewing domain size ($\Delta L_{\rm view}$), correlation length, lifetime (decorrelation rate) and velocity of eddies such that the synthetic data are stationary and homogeneous. If the proposed conditions, Eq. (20) and Eq. (22), are correct, then we can generate statistically valid synthetic data, i.e., stationary and homogeneous data, while keeping the computation resource minimal. Here, we examine the conditions on ΔL with synthetic data using the auto-correlation function because this is the easiest way to see the effect of the finite spatial domain, and whilst the correlation lengths and times will also be affected these require fitting functions that add unnecessary complexity to the problem. We do not examine the ΔT condition because synthetic data are usually generated with many time points such that ensemble average can be performed, in which case Eq. (20) is readily satisfied.

We generate synthetic data with $\lambda_x = 0.1$ m, $\tau_{\rm life} = 15\,\mu$s, $v = 5,000$ m/s, $\Delta T = 48,000\,\mu$s and $\Delta T_{\rm sub} = 480\,\mu$s. We set the variance of amplitudes, A^2 in $P(A_i)$ (see Eq. (3)), constant in space with the intention of generating homogeneous synthetic data. To cover a couple of correlation lengths, we set $\Delta L_{\rm view} = 0.2$ m. Based on Eq. (22), we find that $\Delta L \geq 0.7$ m for this case.

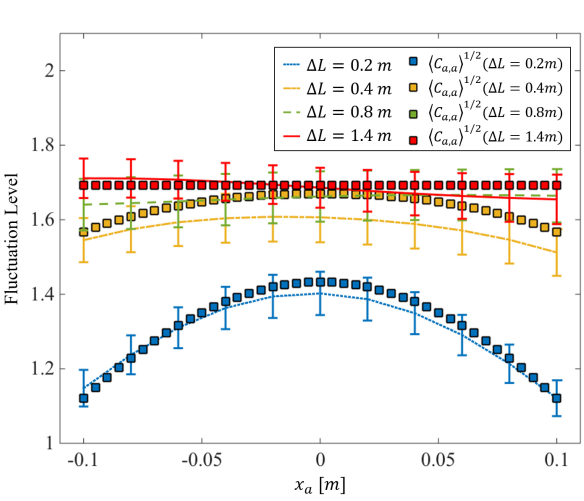

Figure 5: Analytically calculated fluctuation levels of the synthetic data $\langle C_{a,a}\rangle^{1/2}$ (squares) using Eq. (16) with the actual values of $\mathcal{W}_{\Delta L}$, i.e., no approximation of $\mathcal{W}_{\Delta L}$ to two, and numerically estimated fluctuation levels $\left\langle C^{\rm SYN}_{a,a}\right\rangle^{1/2}$ from the synthetic data as a function of spatial position x_a within $\Delta L_{\rm view}$ for $\Delta L = 0.2$ (blue dot), 0.4 (yellow dash dot), 0.8 (green dash) and 1.4 m (red line). Uncertainties represent the 95% confidence level of estimating fluctuation levels. Here, $\Delta L \geq 0.7$ m is the condition for the data to be homogeneous.

Fig. 5 shows, as functions of spatial position x_a, the $\langle C_{a,a}\rangle^{1/2}$ (squares), i.e., the calculated fluctuation levels using Eq. (16) with the actual values of $\mathcal{W}_{\Delta L}$ (no approximation of $\mathcal{W}_{\Delta L}$ to two) for $\Delta L = 0.2$ (blue), 0.4 (yellow), 0.8 (green) and 1.4 m (red). Note that the green squares are not visible as they are overlapped with the red squares. Numerically estimated fluctuation levels, $\left\langle C^{\rm SYN}_{a,a}\right\rangle^{1/2}$, based on four sets of synthetic data with $\Delta L = 0.2$ (blue dot), 0.4 (yellow dash dot), 0.8 (green dash) and 1.4 m (red line) are also shown. We see that the data are not homogeneous for the cases of $\Delta L = 0.2$ and 0.4 m which do not satisfy the ΔL condition set by Eq. (22) even if A^2 is set to be constant in space while generating the synthetic data. The underestimation of the fluctuation level towards the edge of $\Delta L_{\rm view}$ for these cases are caused by the 'edge effect,' i.e., $\mathcal{W}_{\Delta L} < 2$ towards the edge. Data are homogeneous for $\Delta L = 0.8$ and 1.4 m, i.e., $\Delta L \geq 0.7$ m is satisfied.

Fig. 6(a) shows the required minimum ΔL normalized to λ_x, calculated with Eq. (22), as a function of λ_x and v while keeping $\tau_{\rm life} = 15\ \mu$s with $\Delta L_{\rm view} = 2\lambda_x$. We, then, generate 342 sets of synthetic data with various values of λ_x and v with the domain size of $\Delta L/\lambda_x = 9$ shown as the red line in Fig. 6(a). Thus, we expect that the sets of synthetic data which fall into the region above the line of $\Delta L/\lambda_x = 9$ do not satisfy the homogeneous condition, i.e., these sets of data require a larger ΔL. With the line (red) of $\Delta L/\lambda_x = 9$, Fig. 6(b) shows the normalized average distance $\mathcal{D}(\lambda_x, v)$ of the numerically estimated fluctuation levels $\left\langle C^{\rm SYN}_{a,a}\right\rangle^{1/2}$ from the expected value $\langle C_{a,a}\rangle^{1/2}$ (Eq. (16))

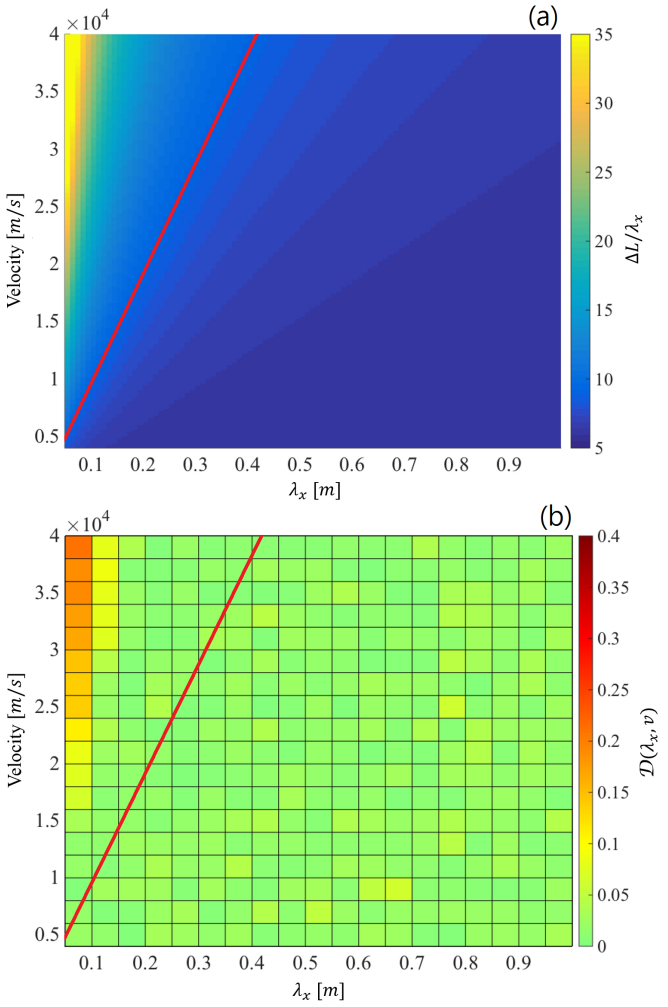

Figure 6: (a) Required minimal ΔL normalized to λ_x as a function of λ_x and v with fixed $\tau_{\rm life} = 15\ \mu$s and $\Delta L_{\rm view} = 2\lambda_x$. (b) The normalized average distance $\mathcal{D}(\lambda_x, v)$, defined in Eq. (23), for 342 (18×19) sets of synthetic data with various values of λ_x and v while keeping $\Delta L/\lambda_x = 9$ (red lines in both figures). If $\mathcal{D}(\lambda_x, v)$ is not close to zero, then the synthetic data may not be regarded as homogeneous, and the region violating the required ΔL condition, i.e., above the red line, has values of $\mathcal{D}(\lambda_x, v)$ greater than zero.

defined as

$$\mathcal{D}(\lambda_x, v) = \frac{\sqrt{\dfrac{1}{N_{x_a}}\displaystyle\sum_{x_a}\left[\left\langle C^{\rm SYN}_{a,a}\right\rangle^{1/2} - \langle C_{a,a}\rangle^{1/2}\right]^2}}{\langle C_{a,a}\rangle^{1/2}}, \tag{23}$$

where the sum is performed on the all spatial positions within the $\Delta L_{\rm view}$, and N_{x_a} is the number of spatial points. $\mathcal{D}(\lambda_x, v)$ is a zeroth order proxy for the homogeneity of synthetic data (see Fig. 5). If $\mathcal{D}(\lambda_x, v)$ is not close to zero, then we speculate that the data are not homogeneous. Fig. 6(b) clearly shows that $\mathcal{D}(\lambda_x, v)$ is conceivably larger than zero above the line, hence vindicating our proposed condition on ΔL.

4. Conclusion

Motivated by the recent trend of wide usage of synthetic data either in a direct comparison of data from a local turbulence

simulation to experimental data or in evaluating the reliability of a statistical algorithm, we have investigated how the synthetic data must be generated while minimizing the computation resource. The conditions on the total simulation domains, ΔT and ΔL, given the 'viewing' domain $\Delta L_{\rm view}$ (or the domain of interest) are summarized in Eq. (20) and Eq. (22) such that the data are stationary and homogeneous. We emphasize that many statistical analyses require the data to be homogeneous and stationary at least within the domain of interest. Furthermore, if one generates a synthetic data based on a local turbulence simulation such as a gyro-kinetic simulation, the conditions can be applied to $\Delta L_{\rm view}$ given ΔT and ΔL.

We found the conditions by realizing that two error functions $\mathcal{W}_{\Delta T_{\rm sub}}(t_i)$ in Eq. (9) and $\mathcal{W}_{\Delta L}$ in Eq. (13), which take roles of weighting factors on each eddy, must be approximately two throughout the domain of interest. As these error functions are manifestations of the coherent structure of Gaussian-shaped eddies, similar weighting factors would appear for different shapes of eddies as long as the eddies have non-zero coherent structures in temporal and/or spatial domains. Although we do not provide exact forms of weighting factors for different shapes of eddies as we have done for Gaussian-shaped eddies in this paper, our rationale can be applied to other shapes of eddies. We do not extend the conditions to include other shapes of eddies here because 1) the purpose of this paper is not to provide the conditions for all possible shapes of eddies (which is not possible), rather to provide the 'rationale' how one must choose a domain of interest such that the synthetic data can be used for statistical analyses and 2) we have chosen to apply our rationale on the Gaussian-shaped eddies as we believe that such a shape is a good approximation [22] to describe turbulent fluctuations.

Acknowledgement

This work is supported by National R&D Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (grant number 2014M1A7A1A01029835) and the KUSTAR-KAIST Institute, KAIST, Korea.

Appendix A. Constant $\mathcal{W}_{\Delta L}$ within the 'viewing' domain $\Delta L_{\rm view}$

$\mathcal{W}_{\Delta L}$ within the viewing window $\Delta L_{\rm view}$ can be stated as constant if the difference between the maximum and the minimum of $\mathcal{W}_{\Delta L}$ denoted as $\mathcal{W}^{\rm max}_{\Delta L}$ and $\mathcal{W}^{\rm min}_{\Delta L}$, respectively, are small (for instance, less than 5%):

$$\frac{\mathcal{W}^{\rm max}_{\Delta L} - \mathcal{W}^{\rm min}_{\Delta L}}{\mathcal{W}^{\rm max}_{\Delta L}} \leq 0.05. \tag{A.1}$$

Here, we provide the definition of $\mathcal{W}_{\Delta L}$ again as a matter of convenience:

$$\mathcal{W}_{\Delta L} = \mathrm{Erf}\left[\frac{x_a + x_b + \Delta L}{2\sqrt{\lambda_x^2 + \tau_{\rm life}^2 v^2}}\right] - \mathrm{Erf}\left[\frac{x_a + x_b - \Delta L}{2\sqrt{\lambda_x^2 + \tau_{\rm life}^2 v^2}}\right]. \tag{A.2}$$

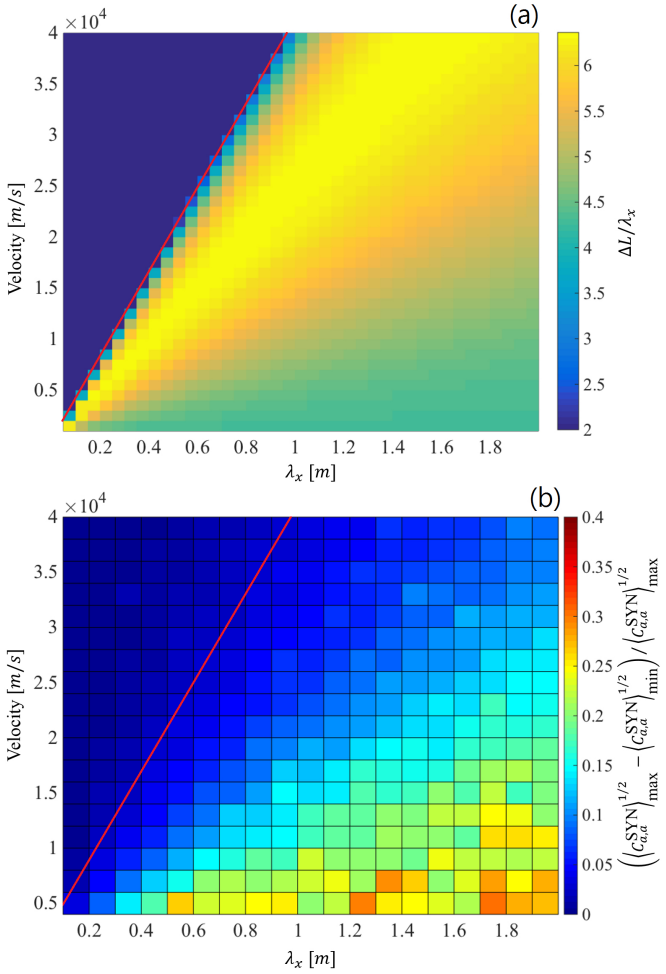

Figure A.7: (a) The minimum size of ΔL normalized to λ_x such that $\mathcal{W}_{\Delta L}$ is almost constant within $\Delta L_{\rm view}$, i.e., satisfying Eq. (A.1). Note that if this calculated ΔL is less than $\Delta L_{\rm view}$, then we force $\Delta L = \Delta L_{\rm view}$. (b) Normalized difference between the maximum and minimum fluctuation levels from the synthetic data with $\Delta L/\lambda_x = 2$ (red lines in both figures). The region above the line has constant $\mathcal{W}_{\Delta L}$ within $\Delta L_{\rm view}$, thus the fluctuation levels are almost constant.

To estimate the fluctuation level, we set $x_a = x_b$ as in $\langle C_{a,a}\rangle^{1/2}$. By taking the first and the second derivatives of $\mathcal{W}_{\Delta L}$ with respect to x_a, we find that the maximum point is at $x_a = 0$. Furthermore, because $x_a = 0$ is the only critical point, the minimum occurs at the boundary of the $\Delta L_{\rm view}$. Then, we have

$$\begin{aligned}
\mathcal{W}^{\rm max}_{\Delta L} &= \mathrm{Erf}\left[\frac{\Delta L}{2\sqrt{\lambda_x^2 + \tau_{\rm life}^2 v^2}}\right] - \mathrm{Erf}\left[\frac{-\Delta L}{2\sqrt{\lambda_x^2 + \tau_{\rm life}^2 v^2}}\right], \\
\mathcal{W}^{\rm min}_{\Delta L} &= \mathrm{Erf}\left[\frac{\Delta L_{\rm view} + \Delta L}{2\sqrt{\lambda_x^2 + \tau_{\rm life}^2 v^2}}\right] - \mathrm{Erf}\left[\frac{\Delta L_{\rm view} - \Delta L}{2\sqrt{\lambda_x^2 + \tau_{\rm life}^2 v^2}}\right].
\end{aligned} \tag{A.3}$$

Thus, the data may resemble the condition of homogeneous state if Eq. (A.1) is satisfied with Eq. (A.3).

Fig. A.7(a) shows, as a function of λ_x and v, the minimum size of ΔL normalized to λ_x satisfying Eq. (A.1) with fixed $\tau_{\rm life} = 100\ \mu$s and $\Delta L_{\rm view} = 2\lambda_x$. Note that if this calcu-

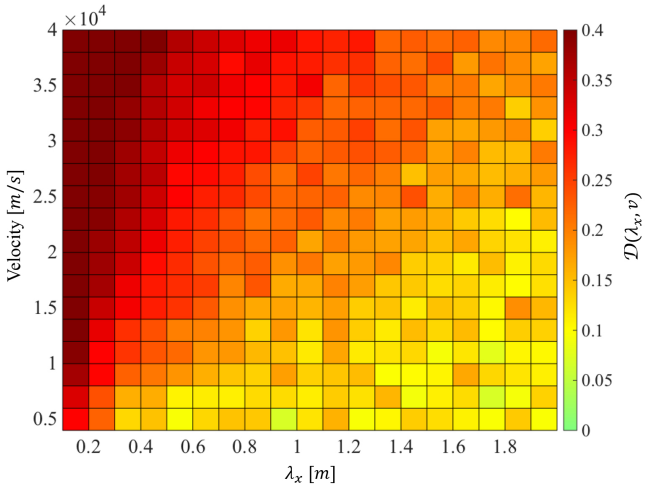

Figure A.8: Normalized average distance $\mathcal{D}(\lambda_x, v)$ defined in Eq. (23) for the 342 sets of synthetic data used to generate Fig. A.7. None of the data sets have the values of $\mathcal{D}(\lambda_x, v)$ close to zero as in Fig. 6.

lated minimum size of ΔL happens to be less than ΔL_{view}, we force $\Delta L = \Delta L_{\text{view}}$ since it is non-sense to consider the larger 'viewing' window than the total simulation window. Then, we generate 342 sets of synthetic data with $\Delta L/\lambda_x = 2$, i.e., $\Delta L = \Delta L_{\text{view}}$. The lines of $\Delta L/\lambda_x = 2$ are depicted as red lines in both Fig. A.7(a) and (b). Fig. A.7(b) shows that in the region above the line of $\Delta L/\lambda_x = 2$, where Eq. (A.1) is satisfied, the maximum and minimum fluctuation levels from the synthetic data, $\left\langle C^{\text{SYN}}_{a,a}\right\rangle^{1/2}_{\text{max}}$ and $\left\langle C^{\text{SYN}}_{a,a}\right\rangle^{1/2}_{\text{min}}$, respectively, within the ΔL_{view} are quite similar; while the region below the line shows non-negligible differences.

Although the fluctuation levels are constant over the viewing window for the data sets satisfying Eq. (A.1), we can easily find that none of the data sets in Fig. A.7 satisfy the 'true' homogeneous condition set by Eq. (22). If we plot the normalized average distance $\mathcal{D}(\lambda_x, v)$ defined in Eq. (23) for the same sets of the synthetic data, we see that all the synthetic data have values of $\mathcal{D}(\lambda_x, v)$ greater than zero as shown in Fig. A.8 (cf. Fig. 6(b) where we have deliberately used the same color scale).

As a summary, we state that constant $\mathcal{W}_{\Delta L}$ is not a sufficient condition for data being homogeneous, rather we require $\mathcal{W}_{\Delta L} \approx 2$.

References

[1] B. Carreras, IEEE Trans. Plasma Sci. 25 (1997) 1281.
[2] K. H. Burrell, Phys. Plasmas 4 (1997) 1499.
[3] P. H. Diamond, S.-I. Itoh, K. Itoh, and T. S. Hahm, Plasma Phys. Controlled Fusion 47 (2005) R35.
[4] T. S. Hahm and K. H. Burrell, Phys. Plasmas 2 (1995) 1648.
[5] T. Huld, A. H. Nielsen, H. L. Pecseli, and J. Juul Rasmussen, Phys. Fluids B 3 (1991) 1609.
[6] G. R. McKee, R. J. Fonck, D. K. Gupta, D. J. Schlossberg, M. W. Shafer, C. Holland, and G. Tynan, Rev. Sci. Instrum. 75 (2004) 3490.
[7] Y.-c. Ghim, A. A. Schekochihin, A. R. Field, I. G. Abel, M. Barnes, G. Colyer, S. C. Cowley, F. I. Parra, D. Dunai, S. Zoletnik, and the MAST Team, Phys. Rev. Lett. 110 (2012) 145002.
[8] Y.-c. Ghim, A. R. Field, A. A. Schekochihin, E. G. Highcock, C. Michael, and the MAST Team, Nucl. Fusion 54 (2014) 042003.
[9] S. Levinson, J. Beall, E. Powers, and R. Bengtson, Nucl. Fusion 24 (1984) 527.
[10] Y. c Ghim, A. R. Field, D. Dunai, S. Zoletnik, L. Bardo czi, A. A. Schekochihin, and the MAST Team, Plasma Phys. Controlled Fusion 54 (2012) 095012.
[11] R. D. Durst, R. J. Fonck, G. Cosby, H. Evensen, and S. F. Paul, Rev. Sci. Instrum. 63 (1992) 4907.
[12] R. J. Fonck, G. Cosby, R. D. Durst, S. F. Paul, N. Bretz, S. Scott, E. Synakowski, and G. Taylor, Phys. Rev. Lett. 70 (1993) 3736.
[13] A. E. White, L. Schmitz, G. R. McKee, C. Holland, W. A. Pee- bles, T. A. Carter, M. W. Shafer, M. E. Austin, K. H. Burrell, J. Candy, J. C. DeBoo, E. J. Doyle, M. A. Makowski, R. Prater, T. L. Rhodes, G. M. Staebler, G. R. Tynan, R. E. Waltz, and G. Wang, Phys. Plasmas 15 (2008) 056116.
[14] M. W. Shafer, R. J. Fonck, G. R. McKee, C. Holland, A. E. White, and D. J. Schlossberg, Phys. Plasmas 19 (2012) 032504.
[15] A. R. Field, D. Dunai, Y.-c. Ghim, P. Hill, B. McMillan, C. M. Roach, S. Saarelma, A. A. Schekochihin, S. Zoletnik, and the MAST Team, Plasma Phys. Controlled Fusion 56 (2014) 025012.
[16] A. Bencze and S. Zoletnik, Phys. Plasmas 12 (2005) 052323.
[17] B. Tal, A. Bencze, S. Zoletnik, G. Veres, and G. Por, Phys. Plasmas 18 (2011) 122304.
[18] D. Guszejnov, A. Bencze, S. Zoletnik, and A. Kra mer-Flecken, Phys. Plasmas 20 (2013) 062303.
[19] G. R. McKee, C. Fenzi, R. J. Fonck, and M. Jakubowski, Rev. Sci. Instrum. 74 (2003) 2014.
[20] J. S. Bendat and A. G. Piersol, Random Data - Analysis and Measurement Procedures (4th ed), Wiley, 2010.
[21] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.
[22] P. Davidson, Turbulence - An introduction for scientists and engineers, Oxford, 2004.