A fast algorithm for the estimation of statistical error in DNS (or experimental) time averages☆

Serena Russo[a,*], Paolo Luchini[b]

[a] *DICMaPI, Universitá di Napoli Federico II*
Piazzale Tecchio, 80, 80125, Napoli, Italy
[b] *DII, Universitá di Salerno*
Via Giovanni Paolo II, 132, 84084, Fisciano (SA), Italy

Abstract

Time- and space-averaging of the instantaneous results of DNS (or experimental measurements) represent a standard final step, necessary for the estimation of their means or correlations or other statistical properties. These averages are necessarily performed over a finite time and space window, and are therefore more correctly just estimates of the 'true' statistical averages. The choice of the appropriate window size is most often subjectively based on individual experience, but as subtler statistics enter the focus of investigation, an objective criterion becomes desirable. Here a modification of the classical estimator of averaging error of finite time series, i.e. 'batch means' algorithm, will be presented, which retains its speed while removing its biasing error. As a side benefit, an automatic determination of batch size is also included. Examples will be given involving both an artificial time series of known statistics and an actual DNS of turbulence.

Keywords: sample average, variance, DNS, uncertainty quantification
2010 MSC: 00-01, 99-00

☆A fast algorithm for the estimation of statistical error in DNS (or experimental) time averages

*Corresponding author
Email address: aneres85@gmail.com (Serena Russo)

1. Introduction

Simulation is one of the most important tools used for characterizing the behavior of complex systems in a wide variety of engineering fields. Direct numerical simulation (DNS) of turbulence represents a useful computational tool for the study of turbulent flows. Statistical quantities computed from DNS results are commonly used both to understand flow physics and to test hypotheses regarding turbulence [1] [2] [3] as well as to calibrate and validate engineering turbulence models [4] [5]. DNS data are thus commonly used like experimental data. As with experimental data, it is important to understand the uncertainty in these data to have confidence in the interpretation of a DNS. However, uncertainties in the DNS literature are not usually reported because they are generally not systematically evaluated. Instead, it is common practice in DNS the choice of grid spacing requirements, required simulation time, etc. to a combination of knowledge gained from previous experience and observations of simulation outputs. The goal of this work is to improve upon this practice by providing a systematic method for estimating uncertainty in the statistics computed from DNS data.

Uncertainty estimation for DNS is an example of solution verification for the computed statistical quantities, whose goal is to ensure that numerical solutions of a mathematical model are sufficiently accurate approximations of the exact solution of the model [6]. A large quantity of research papers is devoted to the solution verification techniques for computational fluid dynamics (CFD). The simplest methods for estimating discretization error, that are common for CFD calculations, are based on standard Richardson extrapolation from a sequence of simulations on successively finer meshes. It is commonly used to estimate the leading order error in numerical results for both deterministic and stochastic problems. Since DNS data are generally statistical quantities, the results are contaminated not only by discretization error but also by sampling error. In the Richardson extrapolation, the sampling error is not taken into account, so by neglecting this, one implicitly assumes that it is small relative to the discretization error. However, since the goal of DNS is to resolve all relevant physical scales, output could be affected from errors due to finite sampling relative to discretization errors.

If DNS data used to compute the statistics were samples from independent, identically distributed random variables, the central limit theorem would allow easy estimation of the sampling error. However, the samples used to generate DNS statistics are drawn from a time history and/or spatial field and are generally not independent. To avoid the complications of correlated samples, many authors downsample instantaneous measurements until the retained samples are arguably uncorrelated [7]. However, optimally downsampling autocorrelated samples requires coarsening the data just enough to decorrelate the signal but not too much to avoid discarding useful data [8]. As increasing the number of independent samples is computationally expensive in DNS, it is imperative to extract all possible information from the data. Hoyas and Jiménez ([9]) propose an approach to accounting for the correlations in DNS statistics based on a sequence of coarse grainings of the data. The main difficulty of this process is the automation of the procedure.

Several research papers are dedicated to the problem of the estimation of the uncertainties in the statistics of steady-state simulation, especially in the fields of statistics and signal processing. Two of the classical categories of estimators are those based on the batch means (BM) algorithms and auto-regressive-moving-average (ARMA) methods. The first approach is based on the subdivision on the data into a certain number of groups having uncorrelated statistics. Conway, Johnson and Maxwell ([10]) and Conway ([11]) are the first papers to address the problem via the method of nonoverlapping batch mean (NOBM), Schmeiser ([12]) makes a major contribution to the theory of NOBM. Another important output analysis method is the overlapping batch means (OBM) method ([13] and [14]). Batch means methods are widely used because they are very fast, even if they are not very accurate in all situations.

ARMA models provide a detailed description of a (weakly) stationary stochastic process in terms of two polynomials, one for the auto-regression (AR) and the second for the moving average (MA) ([15]). This model allows you to understand and, perhaps, predict future values in time series, so the sample mean variance is obtained from the estimation of the complete correlation function to start with, resulting in a slower approach with respect to the NOBM. Following the work of Broersen [16], Oliver et al. [17] propose a Bayesian extension of the standard Richardson extrapolation that accounts for both discretization and sampling uncertainty. The last one is based on direct estimation of the correlations in the DNS data reconstructed with an Auto Regressive model.

In this paper we propose a Batch Means and Batch Correlation (BMBC) algorithm, that is inspired by the Batch Means algorithm. With respect to this one, it retains its speed while removing its biasing error. The paper is organized as follows. Sections 2 and 3 deal with one feature of problem of uncertainty quantification and the estimation of the variance of the sample average. Section 4 is devoted to a description of the methods presented in literature: this overview could be useful to the reader for the understanding of the novel approach. Section 5 introduces the method of BMBC, by comparing it with existing approaches. Two applicative examples are reported in Section 6 and Section 7 deals with the Conclusions.

2. The problem of uncertainty quantification

This paper deals with one aspect of the problem of uncertainty quantification in the variance of the sample average computed from correlated data. It draws its inspiration from another paper of the same authors ([3]), in which they calculated the linear response of turbulent flow to an external volume force. To do this, time-averaged physical quantities (such as flowrate, velocity and so on) were evaluated in the statistically stationary state, that is the state in which all multi-time statistics are invariant under a shift in time while the one-point statistics are independent of time. The main requirement for a linear response was the choice of the amplitude of the volume force to be selected small enough to have a linear response and sufficiently large so that the response exceeds

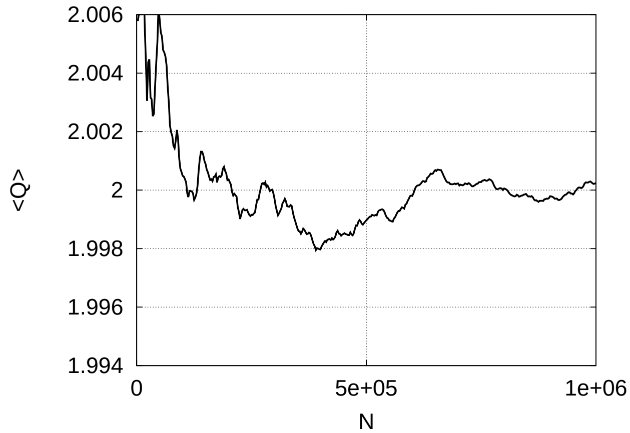

Figure 1: The convergence history of Q versus averaging time T

statistical fluctuations of turbulence. One of the greatest difficulties they encountered was represented by the computational time needed for convergence to a statistically stationary state, strongly influenced by the amplitude of the volume force. As the amplitude of the volume force decreases, larger averaging times become necessary to obtain a statistically stationary state.

Figure 1 (i.e. Figure 6 of paper [3]) shows the DNS profile of the time-averaged values of flowrate as a function of N, the length of the simulation. It only provides a qualitative indication of the time required for the statistically stationary state to be attained. In order to have a more reliable result, it is important to have a quantitative estimation of the variance of all the averaged values. It is possible to verify that, for time series with integrable correlation function, such variances are inversely proportional to $\sqrt{N}$, so they represent an important parameter that can provide a confidence interval for a given N (see [18]).

3. Estimation of the variance of the sample average

Let us start by examining a mathematical formulation of the problem of the estimation of the variance of the sample average. If X denotes a scalar flow quantity (for example a velocity component or flowrate), $\{X_n\}$ for $n = 1, 2, \ldots$ is a statistically stationary stochastic process representing the sequence of outputs generated by a single run of a DNS simulation, of which we are interested in steady-state average values. In particular, we want to estimate a confidence interval for some parameters, such as the steady-state mean $\mu = E[X]$ of the process, where $E[\cdot]$ denotes the expected value.

The simulation can only run for finite time and N denotes the length of the time series $\{X_n\}$ of outputs generated by a single run of the simulation. The classical estimator of μ is the sample mean

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} X_n, \tag{1}$$

where symbol '$\hat{}$' is used to indicate estimated quantities. In the hypothesis of ergodicity ([19]), the sample mean $\hat{\mu}$ converges to the steady-state mean μ as $N \to \infty$. Then the sampling error e_N is the difference between the sample average and the true mean

$$e_N = \hat{\mu} - \mu. \tag{2}$$

For sequences in which independence is approached for large time separations, as is expected to occur for turbulence time series, e_N converges to a normal distribution with zero mean as N becomes large ([19, 20]). So the variance of e_N is enough for the complete characterization of the sampling error.

According to a textbook definition ([23]), the variance of the sample average can be written as

$$\sigma^2(\hat{\mu}) = E\Big[(\hat{\mu} - \mu)^2\Big] = \frac{1}{N}\Big[\gamma_0 + 2\sum_{k=1}^{N-1}\Big(1 - \frac{k}{N}\Big)\gamma_k\Big], \tag{3}$$

where γ_k is the autocorrelation function.

Considering that these processes are weakly dependent, the lag-k correlation $\gamma_k \equiv E[(X_n - \mu)(X_{n+k} - \mu)]$ for $k = 0, \pm 1, \pm 2, \ldots$, satifies $\gamma_k \to 0$ as $k \to \infty$.

4. Classical estimators for $\hat{\sigma}^2(\hat{\mu})$

This section is devoted to a description of some of the existing methods found in literature for the estimation of the variance of the sample average for uncorrelated and correlated processes. Batch Means and Auto Regressive approaches are detailed for correlated processes.

4.1. The case of an uncorrelated process

The case of an uncorrelated process represents a classical example, already treated in textbook (see [14]) in which the samples $\{X_n\}$ are independent and identically distributed (i.i.d). In this situation, $\gamma_k = 0$ for $k \neq 0$, equation 3 simplifies to

$$\sigma^2(\hat{\mu}) = \frac{\gamma_0}{N} = \frac{1}{N^2}\sum_{n=1}^{N}(X_n - \mu)^2. \tag{4}$$

If μ is replaced by $\hat{\mu}$, the classical unbiased estimator is

$$\hat{\sigma}^2(\hat{\mu}) = \frac{\sigma^2(X_n)}{N} = \frac{1}{N(N-1)}\sum_{n=1}^{N}\Big(X_n - \hat{\mu}\Big)^2. \tag{5}$$

$\hat{\sigma}^2(\hat{\mu})$ is an unbiased estimator of $\sigma^2(\hat{\mu})$ and the classical demonstration (as explained in [14]) is reported in Appendix A.

4.2. The methods of Batch Means

Batch means are sample means of subsets of consecutive subsamples from a simulation output sequence. Methods based on Batch Means are very successfull because they are very fast and easy to apply. There are several methods based on Batch Means, but in this paper we will consider only the NonOverlapping Batch Means (NOBM) and the Overlapping Batch Means (OBM) approaches.

According to [14] and [21], in the approach of NOBM, the process sequence $\{X_n\}$ is divided into K adjacent non–overlapping batches, each of size M. If N

is a multiple of M (so that $N = MK$), the sample mean $\overline{X}_k$ for the k-th batch is

$$\overline{X}_k = \frac{1}{M} \sum_{i=kM+1}^{(k+1)M} X_i, \quad k = 1, 2, \ldots, K. \tag{6}$$

Assuming that the process $X_1, X_2, \ldots$ is stationary, the joint distribution of the X_is is insensitive to time shifts. If the process is weakly dependent and the batch size M is sufficiently large so that the batch means $\overline{X}_k$ are approximately independent and identically distributed normal random variables with mean μ, the NOBM estimator for $\sigma^2(\hat{\mu})$ can be derived from the same formula used for uncorrelated samples (see equation 5), i.e.

$$\begin{aligned}
\hat{\sigma}^2_{nobm}(\hat{\mu}) &= \frac{1}{(K-1)} \Big(\frac{1}{K} \sum_{k=1}^{K} \overline{X}_k^2 - \hat{\mu}^2 \Big) = \\
&= \frac{M}{N(K-1)} \sum_{k=1}^{K} \Big(\overline{X}_k - \hat{\mu} \Big)^2 = \\
&= \frac{M^2}{N(N-M)} \sum_{k=1}^{K} \Big(\frac{\sum_{i=kM+1}^{(k+1)M} (X_i - \hat{\mu})}{M} \Big)^2 = \\
&= \frac{1}{N(N-M)} \sum_{k=1}^{K} \Big(\sum_{i=kM+1}^{(k+1)M} (X_i - \hat{\mu}) \Big)^2 .
\end{aligned} \tag{7}$$

This method is very fast and simple, however one of the practical disadvantages is that M is an unknown parameter and needs to be investigated for each case (for details [14] and [21]).

The Overlapping Batch Means (OBM) estimator for $\sigma^2(\hat{\mu})$ is obtained by grouping data in $N - M + 1$ batches of length M

$$\hat{\sigma}^2_{obm}(\hat{\mu}) = \frac{M}{(N-M)(N-M+1)} \sum_{j=1}^{N-M+1} \Big(\overline{X}_j - \hat{\mu} \Big)^2 . \tag{8}$$

where the batch mean and the sample mean are defined respectively as

$$\overline{X}_j = \frac{1}{M} \sum_{n=j}^{j+M-1} X_n; \tag{9}$$

$$\hat{\mu} = \frac{1}{N-M+1} \sum_{j=1}^{N-M+1} \overline{X}_j. \tag{10}$$

Both $\hat{\sigma}^2_{nobm}$ and $\hat{\sigma}^2_{obm}$ estimators are statistically biased. According to [13], the bias of $\hat{\sigma}^2_{nobm}$ and $\hat{\sigma}^2_{obm}$ can be obtained by passing the expected value operator through the summations in the definitions of $\hat{\sigma}^2$, i.e.

$$E(\hat{\sigma}^2_{nobm}(\hat{\mu})) = \sigma^2(\hat{\mu}) - \frac{2}{K-1} \sum_{k=1}^{K-1} \Big(1 - \frac{k}{K}\Big)\hat{\gamma_k}, \tag{11}$$

$$E(\hat{\sigma}^2_{obm}(\hat{\mu})) = \sigma^2(\hat{\mu}) - \frac{2M}{N-2M+1} \sum_{k=1}^{K-1} \Big(1 - \frac{k}{K}\Big)\hat{\gamma_k}; \tag{12}$$

where $\hat{\gamma_k} = Cov(\overline{X}_h, \overline{X}_{h+kM})$ for all h. In the limit as batch size grows, $\hat{\gamma_k}$ decreases and the estimators become unbiased. It is also possible to verify that the OBM estimator has a smaller variance with respect to the the corresponding NOBM estimator by a factor $2/3$; however the computational cost of OBM is larger by a factor M.

4.3. Autoregressive Model

In this section we consider an alternative way to estimate the variance of the sample average. This is obtained by fitting an autoregressive (AR) model to the observed time series (see for instance [22], [23]). On defining $x_n = X_n - \hat{\mu}$, an AR process of order p is represented by a linear system excited by white noise, so it has the following form (see [8] and [15])

$$x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p} = \epsilon_n. \tag{13}$$

The first parameter a_0 is generally taken as one and ϵ_n denotes a white noise process with zero mean and variance σ^2_ϵ. The process parameters $a_1, \ldots, a_p$ and noise variance σ_ϵ^2 completely define the process and, thus, given these parameters, the exact autocorrelation function of the AR process and, so, the variance of the sample average may be computed [24]. More concisely, equation 13 can also be rewritten as

$$A(z)x_n = \epsilon_n \tag{14}$$

with

$$A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p}. \tag{15}$$

The roots of A(z) are called the poles of the AR(p) processes. As reported in [24], multiplying both sides of the AR(p) equation (13) by x_{n-k} with $k \geq 1$, and taking expectations gives

$$\gamma_n + a_1\gamma_{n-1} + \cdots + a_p\gamma_{n-p} = 0, \tag{16}$$

whose solution is

$$\gamma_n = C_1 p_1^n + C_2 p_2^n + \cdots + C_p p_p^n, \tag{17}$$

where p_i are the poles and C_i are constants whose vales are determined by $\gamma(0), \gamma(1), \ldots, \gamma(p-1)$. By substituting correlations 16 into equation 3, we can obtain the estimation of the variance of the sample average.

In section 6 we will present some results derived from the application of this AR model. They are obtained from an open source C++ implementation, developed in [17].

5. The new method of Batch Means and Batch Correlations (BMBC)

This section is devoted to the presentation of the new method of Batch Means and Batch Correlations, that takes its inspiration from the BM ones. The basic idea of the new algorithm is to collect together the best features of the methods described in the previous section. With respect to the BM methods, it preserves its speed and easy applicability, while it retains the accuracy typical of AR approaches.

The starting point is represented by the expression of the variance of the sample average (i.e. eq. 3), in which the autocorrelation function $\gamma_k \equiv E[(X_n - \mu)(X_{n+k} - \mu)]$ for $k = 0, \pm 1, \pm 2, \ldots$ is unknown. If μ is replaced by $\hat{\mu}$ and the expected value is substituted with the sample average, estimated expressions of γ_0 and γ_k can be derived

$$\hat{\gamma}_0 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \hat{\mu})^2, \tag{18}$$

$$\hat{\gamma}_k = \frac{1}{N-k} \sum_{i=1}^{N-k} (X_i - \hat{\mu})(X_{i+k} - \hat{\mu}), \quad for \quad k = 1, 2, \ldots, N-1. \tag{19}$$

By substituting γ_0 and γ_k, respectively, with $\hat{\gamma}_0$ and $\hat{\gamma}_k$ into eq. 3, an expression of the estimation of the variance of the sample average is derived

$$\hat{\sigma}^2(\hat{\mu}) = \frac{1}{N} \sum_{k=-(N-1)}^{N-i} \left(1 - \frac{|k|}{N}\right) \hat{\gamma}_k = \frac{1}{N^2} \sum_{k=-(N-1)}^{N-i} (N - |k|) \hat{\gamma}_k. \tag{20}$$

Since the process is weakly dependent, the estimated correlation function $\hat{\gamma}_k$ has its maximum at $k = 0$ and tends to zero when k increases. For this reason, eq. 20 cannot be used for the estimation of the variance of the sample average because the sum of all $\hat{\gamma}_k$ is equal to zero and it would lead to $\hat{\sigma}^2(\hat{\mu}) = 0$.

Now we decide to introduce a graphical illustration of all $\hat{\gamma}_k$'s to provide a different interpretation of the expressions used for the variance estimation according to BM methods (eq. 7 and 8) as introduced in the previous section. On defining $x_n = X_n - \hat{\mu}$, all the terms of equations 18 and 19 can be arranged in the following square matrix (named 'correlation matrix')

$$\begin{bmatrix} x_1x_1 & x_1x_2 & x_1x_3 & x_1x_4 & x_1x_5 & x_1x_6 & x_1x_7 & x_1x_8 & \dots & x_1x_N \\ x_2x_1 & x_2x_2 & x_2x_3 & x_2x_4 & x_2x_5 & x_2x_6 & x_2x_7 & x_2x_8 & \dots & x_2x_N \\ x_3x_1 & x_3x_2 & x_3x_3 & x_3x_4 & x_3x_5 & x_3x_6 & x_3x_7 & x_3x_8 & \dots & x_3x_N \\ x_4x_1 & x_4x_2 & x_4x_3 & x_4x_4 & x_4x_5 & x_4x_6 & x_4x_7 & x_4x_8 & \dots & x_4x_N \\ x_5x_1 & x_5x_2 & x_5x_3 & x_5x_4 & x_5x_5 & x_5x_6 & x_5x_7 & x_5x_8 & \dots & x_5x_N \\ x_6x_1 & x_6x_2 & x_6x_3 & x_6x_4 & x_6x_5 & x_6x_6 & x_6x_7 & x_6x_8 & \dots & x_6x_N \\ x_7x_1 & x_7x_2 & x_7x_3 & x_7x_4 & x_7x_5 & x_7x_6 & x_7x_7 & x_7x_8 & \dots & x_7x_N \\ x_8x_1 & x_8x_2 & x_8x_3 & x_8x_4 & x_8x_5 & x_8x_6 & x_8x_7 & x_8x_8 & \dots & x_8x_N \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_Nx_1 & x_Nx_2 & x_Nx_3 & x_Nx_4 & x_Nx_5 & x_Nx_6 & x_Nx_7 & x_Nx_8 & \dots & x_Nx_N \end{bmatrix}. \tag{21}$$

The sum of the terms belonging to the $k-$th diagonal denotes the estimated correlation $\hat{\gamma_k}$. Let's consider an example with $M = 3$. According to the NOBM approach, the only terms of the correlation matrix that are considered different from zero are those in black (framed in the black rectangles), while the others (grey) are considered equal to zero. If we start from equation 7, the variance estimator can be rewritten in the following form

$$N(N-M)\hat{\sigma}^2_{nobm}(\hat{\mu}) = \tag{22}$$

$$\sum\nolimits_{black} \begin{bmatrix} x_1x_1 & x_1x_2 & x_1x_3 & x_1x_4 & x_1x_5 & x_1x_6 & x_1x_7 & x_1x_8 & \dots & x_1x_N \\ x_2x_1 & x_2x_2 & x_2x_3 & x_2x_4 & x_2x_5 & x_2x_6 & x_2x_7 & x_2x_8 & \dots & x_2x_N \\ x_3x_1 & x_3x_2 & x_3x_3 & x_3x_4 & x_3x_5 & x_3x_6 & x_3x_7 & x_3x_8 & \dots & x_3x_N \\ x_4x_1 & x_4x_2 & x_4x_3 & x_4x_4 & x_4x_5 & x_4x_6 & x_4x_7 & x_4x_8 & \dots & x_4x_N \\ x_5x_1 & x_5x_2 & x_5x_3 & x_5x_4 & x_5x_5 & x_5x_6 & x_5x_7 & x_5x_8 & \dots & x_5x_N \\ x_6x_1 & x_6x_2 & x_6x_3 & x_6x_4 & x_6x_5 & x_6x_6 & x_6x_7 & x_6x_8 & \dots & x_6x_N \\ x_7x_1 & x_7x_2 & x_7x_3 & x_7x_4 & x_7x_5 & x_7x_6 & x_7x_7 & x_7x_8 & \dots & x_7x_N \\ x_8x_1 & x_8x_2 & x_8x_3 & x_8x_4 & x_8x_5 & x_8x_6 & x_8x_7 & x_8x_8 & \dots & x_8x_N \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_Nx_1 & x_Nx_2 & x_Nx_3 & x_Nx_4 & x_Nx_5 & x_Nx_6 & x_Nx_7 & x_Nx_8 & \dots & x_Nx_N \end{bmatrix}$$

where symbol $\sum_{black}$ means that the sum operator is only applied to the black elements of the matrix. The denominator of equation 7, $N(N-M)$, represents the number of elements of the matrix that are considered equal to zero (i.e. grey elements). Since $\hat{\gamma_k}$ decreases as k increases, grey terms belonging to the dotted panes, closer to the main diagonal, are important for the estimation of correlation (for example x_3x_4, x_4x_3 for $\hat{\gamma_1}$, x_2x_4 for $\hat{\gamma_2}$ and so on). This is the reason why the NOBM approach is biased. The same terms are taken into account in the OBM method, but with a wrong weight. For this reason, both of them are biased (see for details [13] and 7).

Since the variance decreases as N increases, a first solution to the problem of a biased variance estimation should be the inclusion of all the elements of the band around the main diagonal. In this approach of Central Band (CB), the expression for $\hat{\sigma}^2(\hat{\mu})$ should be

$$(N-M+1)(N-M)\hat{\sigma}^2_{cb}(\hat{\mu}) = \qquad (23)$$

$$\sum_{black} \begin{bmatrix} x_1x_1 & x_1x_2 & x_1x_3 & x_1x_4 & x_1x_5 & x_1x_6 & x_1x_7 & x_1x_8 & \dots & x_1x_N \\ x_2x_1 & x_2x_2 & x_2x_3 & x_2x_4 & x_2x_5 & x_2x_6 & x_2x_7 & x_2x_8 & \dots & x_2x_N \\ x_3x_1 & x_3x_2 & x_3x_3 & x_3x_4 & x_3x_5 & x_3x_6 & x_3x_7 & x_3x_8 & \dots & x_3x_N \\ x_4x_1 & x_4x_2 & x_4x_3 & x_4x_4 & x_4x_5 & x_4x_6 & x_4x_7 & x_4x_8 & \dots & x_4x_N \\ x_5x_1 & x_5x_2 & x_5x_3 & x_5x_4 & x_5x_5 & x_5x_6 & x_5x_7 & x_5x_8 & \dots & x_5x_N \\ x_6x_1 & x_6x_2 & x_6x_3 & x_6x_4 & x_6x_5 & x_6x_6 & x_6x_7 & x_6x_8 & \dots & x_6x_N \\ x_7x_1 & x_7x_2 & x_7x_3 & x_7x_4 & x_7x_5 & x_7x_6 & x_7x_7 & x_7x_8 & \dots & x_7x_N \\ x_8x_1 & x_8x_2 & x_8x_3 & x_8x_4 & x_8x_5 & x_8x_6 & x_8x_7 & x_8x_8 & \dots & x_8x_N \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_Nx_1 & x_Nx_2 & x_Nx_3 & x_Nx_4 & x_Nx_5 & x_Nx_6 & x_Nx_7 & x_Nx_8 & \dots & x_Nx_N \end{bmatrix}$$

Again, coefficient $(N-M+1)(N-M)$ represents the number of elements outside the band (i.e. grey terms). Compared with BM approaches, the computational cost of this method is greater because the inclusion of all terms of the main band of the matrix could be more expensive with respect to the product operation of sub-squares sums.

For this reason we propose the new method of Batch Means and Batch Correlations that is based on the inclusion of both diagonal and near–diagonal square blocks. In this approach, the expression for $\hat{\sigma}^2(\hat{\mu})$ is the following

$$(N-M)(N-2M)\hat{\sigma}^2_{bmbc}(\hat{\mu}) = \tag{24}$$

$$\sum\nolimits_{black} \begin{bmatrix} x_1x_1 & x_1x_2 & x_1x_3 & x_1x_4 & x_1x_5 & x_1x_6 & x_1x_7 & x_1x_8 & \dots & x_1x_N \\ x_2x_1 & x_2x_2 & x_2x_3 & x_2x_4 & x_2x_5 & x_2x_6 & x_2x_7 & x_2x_8 & \dots & x_2x_N \\ x_3x_1 & x_3x_2 & x_3x_3 & x_3x_4 & x_3x_5 & x_3x_6 & x_3x_7 & x_3x_8 & \dots & x_3x_N \\ x_4x_1 & x_4x_2 & x_4x_3 & x_4x_4 & x_4x_5 & x_4x_6 & x_4x_7 & x_4x_8 & \dots & x_4x_N \\ x_5x_1 & x_5x_2 & x_5x_3 & x_5x_4 & x_5x_5 & x_5x_6 & x_5x_7 & x_5x_8 & \dots & x_5x_N \\ x_6x_1 & x_6x_2 & x_6x_3 & x_6x_4 & x_6x_5 & x_6x_6 & x_6x_7 & x_6x_8 & \dots & x_6x_N \\ x_7x_1 & x_7x_2 & x_7x_3 & x_7x_4 & x_7x_5 & x_7x_6 & x_7x_7 & x_7x_8 & \dots & x_7x_N \\ x_8x_1 & x_8x_2 & x_8x_3 & x_8x_4 & x_8x_5 & x_8x_6 & x_8x_7 & x_8x_8 & \dots & x_8x_N \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_Nx_1 & x_Nx_2 & x_Nx_3 & x_Nx_4 & x_Nx_5 & x_Nx_6 & x_Nx_7 & x_Nx_8 & \dots & x_Nx_N \end{bmatrix}$$

This approach is unbiased and very fast. Compared with the classical BM approaches, in the method of BMBC, the near-diagonal terms (belonging to the dashed panes) are also included in the variance estimation. These terms are the same that are considered equal to zero in the NOBM, so responsible for the bias error (see Appendix B for further details about the bias of these methods). From the computational point of view, its effort is very similar to that of the method based on BM because it is based on the sum operation of sub-squares.

Let's assume that

$$\overline{X}_k = \frac{1}{M}\sum_{n=M(k-1)+1}^{kM} X_n; \quad \overline{x}_k = \frac{1}{M}\sum_{n=M(k-1)+1}^{kM}\Big(X_n - \hat{\mu}\Big); \tag{25}$$

The formulation based on the use of the correlation matrix (i.e. expression 24) can be rewritten in a following compact form

$$\hat{\sigma}^2_{bmbc}(\hat{\mu}) = \frac{1}{(K-1)(K-2)}(S_0 + 2S_1), \tag{26}$$

where

$$S_0 = \sum_{k=1}^{K}\Big(\overline{X}_k^2 - \hat{\mu}^2\Big) = \sum_{k=1}^{K}\overline{X}_k^2 - K\hat{\mu}^2 = \sum_{k=1}^{K}\overline{x}_k^2, \tag{27}$$

$$S_1 = \sum_{k=1}^{K-1} \left(\overline{X}_k \overline{X}_{k+1} - \hat{\mu}^2 \right) = \sum_{k=1}^{K-1} \overline{X}_k \overline{X}_{k+1} - (K-1)\hat{\mu}^2 = \sum_{k=1}^{K-1} \left(\overline{x}_k \overline{x}_{k+1} \right). \quad (28)$$

The ratio S_1/S_0 represents an important parameter to test and choose the proper batch size M.

6. Results

To illustrate the application of the BMBC approach discussed here, it is applied to the mean estimation computed for a synthetic process and data obtained from DNS. The results will be compared with BM methods (by using algorithms developed in [3]) and AR methods (through an open source C++ implementation developed in [17]).

6.1. The application of the method of BMBC to a test function

In this section the comparison between NOBM, BMBC and AR approaches is presented for a synthetic stochastic process, having a known variance. The artificial time series has the following expression

$$X(i+1) = 0.9X(i) + 0.1R(i), \quad (29)$$

where the function R gives uniformly distributed random numbers between 0 and 1. In Figure 2, we plot $N\hat{\sigma}^2(\hat{\mu})$ against N because of the inverse proportionality between $\hat{\sigma}^2(\hat{\mu})$ and N. Figure 2 shows the results for four different approaches: dotted line refers to the BMBC approach, while dashed and dash-dot lines refer to the NOBM, in which 1 and 2 represent, respectively, the cases in which the batch size is equal to M and $2M$. All of them are also compared with the exact variance which can be calculated analytically and the results obtained from the AR model (with an order $p = 1$). It is evident that the variance estimation obtained with the methods of BMBC and AR converges to a value that is closer to the exact one, but BMBC is faster. This represents

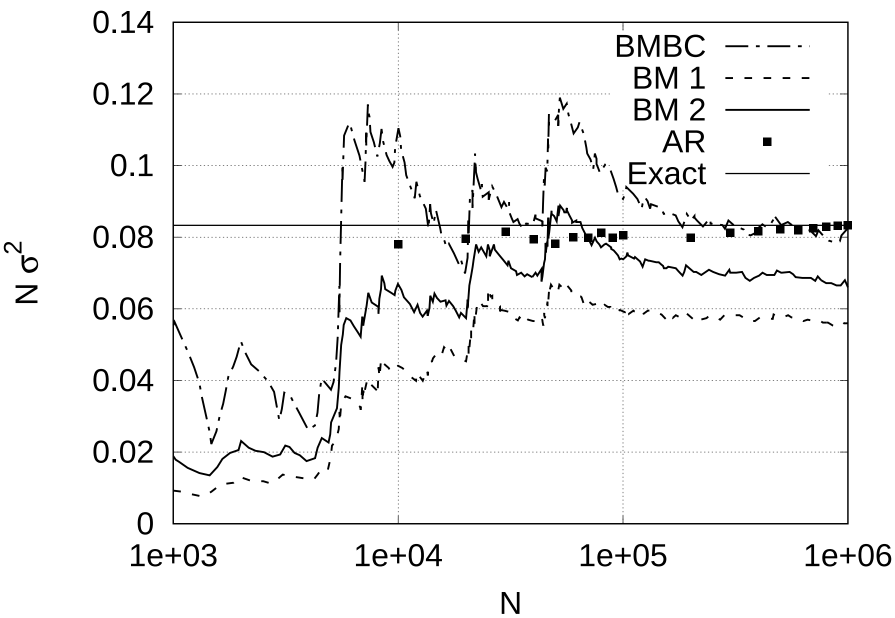

Figure 2: Comparison between NOBM and BMBC variance estimation with the exact variance.

METHOD	$N\hat{\sigma}^2(\hat{\mu})$
EXACT	0.0834
BMBC	0.0815
NOBM 1	0.0559
NOBM 2	0.0680
AR ([17])	0.0833

Table 1: Variance estimation for the test function ($N = 1 \times 10^6$)

METHOD	TIME [sec]
BMBC	0.52
AR ([17])	12.2

Table 2: Computational times

an important result because BMBC, with respect to NOBM 1 and NOBM 2, is unbiased.

The values of the estimated variances calculated at $N = 1 \times 10^6$ are reported in table 1. It is not strange that the AR approach converges better to the exact value: this is due to the fact that the test function defined by 29 is a particular case of 13 where $p = 1$, so the autoregressive reconstruction of the time series is exact.

Moreover it is interesting to compare the performances of both algorithms in terms of computational time. Table 2 shows that the corresponding times (excluding reading/writing phases) are very different with a net advantage for the BMBC approach.

6.2. The application to direct numerical simulation of turbulence

In this section we present the application of NOBM, BMBC and AR approaches to the estimation of the variance of the time-averaged flowrate (Q) as derived from a Direct Numerical Simulation (DNS) of turbulent flow in a plane channel. DNS is an important tool used for the study of such flows because it resolves all relevant physical scales.

In order to present you how uncertainty could affect the results in DNS, let's consider the linear response of turbulent flow to an undulated wall: this topic is addressed in [3] and [25]. In particular, in [3], the authors identified a benchmark problem simple enough that it can be solved both by an eddy-viscosity model and by direct numerical simulation: this is the linear response of a turbulent flow's mean-velocity profile to an external volume force. In [3], a channel flow subject to an external volume force, at moderate but turbulent Reynolds number, has been considered. The DNS solution of this problem has been compared to the corresponding RANS solution closed with an algebraic turbulence model. In particular we studied the linear response of turbulent flow, the one that can be observed when the amplitude of the applied perturbation is small enough, even though a turbulent flow is a nonlinear system. The mean linear response has been measured by imposing a volume force weighted by a small coefficient ϵ, performig a finite difference over ϵ and averaging over a long enough period of time. One of the requirement for a linear response has been the choice of the amplitude parameter ϵ to be selected small enough to have a linear response and sufficiently large so that the response exceeds statistical fluctuations of turbulence. It was observed that, by decreasing the value of ϵ, larger averaging times become necessary to obtain a statistically stationary state. As an example, for a bulk Reynolds number equal to 1450 and friction-velocity Reynolds number $Re_\tau = 100$, the case at $\epsilon = 0.00125$ has required a very long calculation time necessary to attain a statistically stationary state (about equals to 1.2×10^6 time steps and a time step size of 0.02) and it is twice that with $\epsilon = 0.0025$. According to these requirements, it is important to quantify the uncertainty of DNS statistics, in particular of time-averaged values. Let's consider the case of DNS in a plane channel flow as discussed above.

Figure 1 provides the time history of the flowrate time-averaged values: as already said in Section 2, that figure gives only a qualitative indication about the achievement of a statistic stationary state.

As for the test function, we plot $N\hat{\sigma}^2(\hat{\mu})$ against N because of the inverse proportionality between $\hat{\sigma}^2(\hat{\mu})$ and N. Figure 4 shows the results for four different approaches: dotted line refers to the BMBC approach, while dashed and dash-dot lines refer to the NOBM, in which 1 and 2 represent, respectively, the cases in which the batch size is equal to M and $2M$. All of them are, also, compared with the results obtained from the application of the AR model. Here the order of the AR process (p) is equal to 244: this fact means these DNS data are highly correlated signals.

The ratio S_1/S_0 represents an empirical parameter that can be used for the choice of M. As already said in section 5 (see equations 28), S_0 and S_1

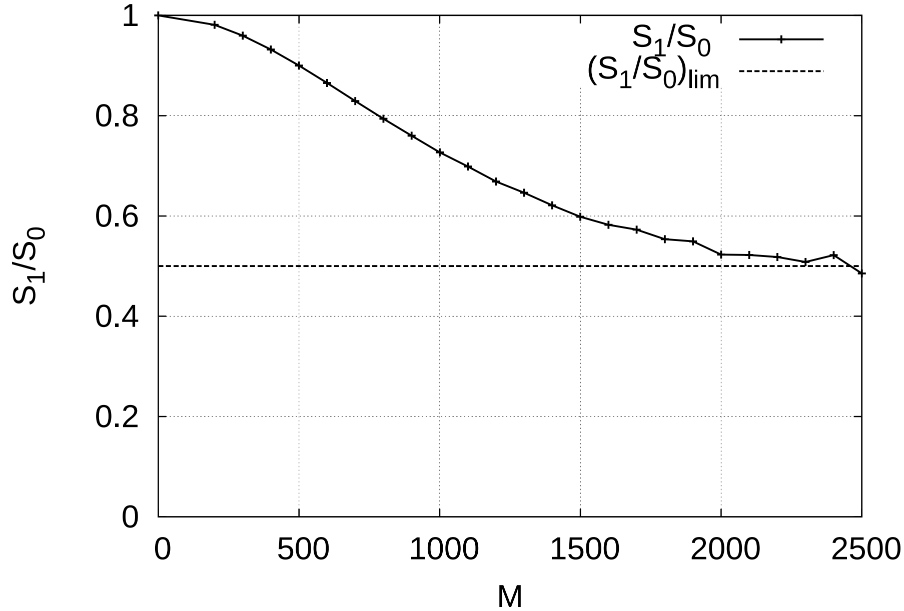

Figure 3: S_1/S_0

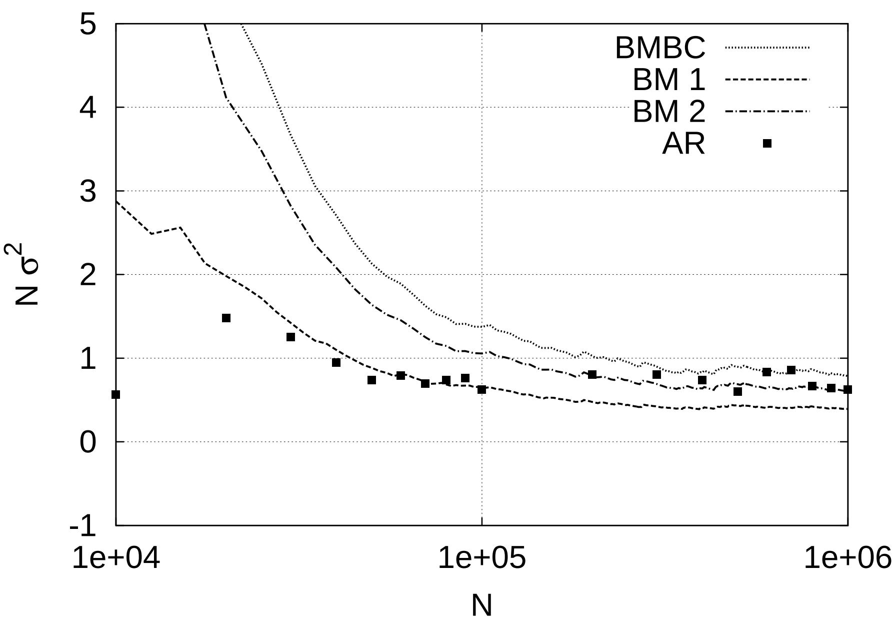

Figure 4: Comparison between NOBM 1, NOBM 2, AR and BMBC.

METHOD	N $\hat{\sigma}^2(\hat{\mu})$
BMBC	0.786
NOBM 1	0.393
NOBM 2	0.605
AR	0.623

Table 3: Variance estimation for DNS flow rate ($N = 1 \times 10^6$).

METHOD	TIME [sec]
BMBC	?
AR ([17])	?

Table 4: Computational times

represent the terms of the correlation matrix, respectively, closer and near the main diagonal. If M is choosen too small, many near-diagonal terms are not considered so they are responsible of the bias error in $\hat{\sigma}^2(\hat{\mu})$ because the $\overline{X}_k$'s are not uncorrelated. On the other hand, if M is too high, this could produce a loss of information. In order to control these two potential sources of error, the best compromise is to set S_1/S_0 (i.e. $(S_1/S_0)_{lim}$) equals to 0.5, so that the $\overline{X}_k$'s are sufficiently independent and uncorrelated. Figure 3 depicts the trend of the ratio S_1/S_0 as function of M: it is observable that it is necessary to impose at least $M = 2500$ to have the desired uncorrelation. The values of the estimated variances calculated at $N = 1 \times 10^6$ are reported in table 3. As for the test function, here the final values for BMBC, NOBM 1 and NOBM 2 are different, in particular the best value of $\hat{\sigma}^2(\hat{\mu})$ is obtained with BMBC. Conversely, AR expectation is not satisfactory as BMBC: this is due to the fact that the time history of flowrate is reconstructed with an AR process, while the test function represents a particular case of the AR model with $p = 1$.

Moreover, as for the test function, it is interesting to compare the performances of both algorithms in terms of computational time. Table 4 shows that the corresponding times (excluding reading/writing phases) are very different with a net advantage for the BMBC approach.

7. Conclusions

This paper focuses on one aspect of the uncertainty quantification in the variance of the sample average computed from correlated data. The idea originates from one paper of the same authors, in which time-averaged physical quantities need to be calculated in the statistically stationary state. In this work, they observed that the computation time needed for convergence could be very prohibitive and that a qualitative indication of the time required for a statistically stationary state can be obtained from the graphical representation of the time-averaged values against the length of the simulation. To have a more reliable result, it is important to consider a more precise quantification of the variance of all the averaged values.

Here we propose a Batch Means and Batch Correlation (BMBC) algorithm, that is inspired by the Batch Means (BM) algorithm. With respect to this one, it retains its speed and easy applicability and the accuracy typical of Auto Regressive (AR) approaches. The results obtained from the application of the BMBC are compared with the classical categories of estimators, i.e. those based on the BM algorithms and AR methods.

The results are reported for two types of functions: the first one is represented by a synthetic stochastic process having a known variance and the second one is the time history of flowrate as derived from a DNS of turbulent flow in a plane channel. From the comparison between BMBC, NOBM 1, NOBM 2 and AR, we observe that BMBC converges better to the expected value for both cases. Moreover, it is interesting to compare the performances of both algorithms in terms of computational time and find that BMBC is much more advantageous as compared with AR.

Appendix A. The demonstration of the unbiased estimator in an uncorrelated process

As reported in [23], it is possible to demonstrate that the expression for $\hat{\sigma}^2(\hat{\mu})$

given by eq. 5 represents the unbiased estimator of $\sigma^2(\hat{\mu})$ in case of uncorrelated process. Let's start from the quadratic part of eq. 5, i.e.

$$\begin{aligned} E(\hat{\sigma}^2(\hat{\mu})) &= \frac{1}{N(N-1)} E \sum_{n=1}^{N} \Big[X_n - \hat{\mu} \Big]^2 = \\ &= \frac{1}{N(N-1)} E \sum_{n=1}^{N} \Big[X_n^2 - 2X_n\hat{\mu} + \hat{\mu}^2 \Big] = \\ &= \frac{1}{N(N-1)} E \Big[\sum_{n=1}^{N} X_n^2 - 2\hat{\mu} \sum_{n=1}^{N} X_n + N\hat{\mu}^2 \Big] = \\ &= \frac{1}{N(N-1)} E \Big[\sum_{n=1}^{N} X_n^2 - 2N\hat{\mu}^2 + N\hat{\mu}^2 \Big] = \\ &= \frac{1}{N(N-1)} E \Big[\sum_{n=1}^{N} X_n^2 - N\hat{\mu}^2 \Big]. \end{aligned} \tag{A.1}$$

Now it is possible to commutate the E and $\sum$ operators, having

$$\begin{aligned} E(\hat{\sigma}^2(\hat{\mu})) &= \frac{1}{N(N-1)} \Big[\sum_{n=1}^{N} E[X_n^2] - NE(\hat{\mu}^2) \Big] = \\ &= \frac{1}{N(N-1)} \Big[NE[X_n^2] - NE[\hat{\mu}^2] \Big]. \end{aligned} \tag{A.2}$$

According to classical variance definition, i.e. $\sigma^2(X_n) = E(X_n^2) - E(X_n)^2$, and eq. 5, eq. A.2 can be rewritten in the following form

$$\begin{aligned} E(\hat{\sigma}^2(\hat{\mu})) &= \frac{N}{N(N-1)} \Big[\sigma^2(X_n) + E(X_n)^2 - \sigma^2(\hat{\mu}) - E(\hat{\mu})^2 \Big] = \\ &= \frac{N}{N(N-1)} \Big[\sigma^2(X_n) + \mu^2 - \frac{\sigma^2(X_n)}{N} - \mu^2 \Big]. \end{aligned} \tag{A.3}$$

Eq. A.3 simplifies to

$$E(\hat{\sigma}^2(\hat{\mu})) = \frac{N}{N(N-1)} \Big[\frac{(N-1)}{N} \sigma^2(X_n) \Big] = \frac{\sigma^2(X_n)}{N}. \tag{A.4}$$

Appendix A. The bias of BMBC estimator

This section is devoted to the characterization of the bias of the method of BMBC: it is interesting to compare it with the case of uncorrelated time series and the BM methods. The expression of the variance of the sample average (i.e. eq. 3) can be rewritten in the following form

$$E(\hat{\mu}^2) - \mu^2 = \frac{1}{N^2}\Big[N\gamma_0 + 2\sum_{k=1}^{N-1}(N-k)\gamma_k\Big]. \tag{A.1}$$

If we substitute eq. 1 for $\hat{\mu}$, eq. A.1 becomes

$$E\left(\frac{\sum_{k,l} X_k X_l}{N^2}\right) = \mu^2 + \frac{1}{N^2}\Big[N\gamma_0 + 2\sum_{k=1}^{N-1}(N-k)\gamma_k\Big], \tag{A.2}$$

i.e

$$E\Big(\sum_{k,l} X_k X_l\Big) = N^2\mu^2 + \Big[N\gamma_0 + 2\sum_{k=1}^{N-1}(N-k)\gamma_k\Big]. \tag{A.3}$$

In eq. A.3, the terms in square brackets are equal to $N^2\sigma^2(\hat{\mu})$ (according to eq. 3) and the l.h.s. is equal to $E(N^2\hat{\mu}^2)$. So eq. A.3 can be rewritten in the following form

$$E(N^2\hat{\mu}^2) = N^2(\mu^2 + \sigma^2(\hat{\mu})). \tag{A.4}$$

Let's suppose to divide the elements of the correlation matrix (i.e. eq. 21) into two groups, those belonging to the diagonal band (comprising the main diagonal) having a certain bandwidth and all the other elements. The bandwidth of the diagonal band is such that all the elements inside contribute to the estimation of the variance of the sample average and the off-diagonal elements are not correlated. Then, if N_{od} and $\sum_{od}$ are respectively the number of the elements that are off the diagonal band and the sum operator made on these elements, equation A.3 simplifies to

$$E\Big(\sum_{od} X_k X_l\Big) = N_{od}\mu^2. \tag{A.5}$$

and eq. A.4 becomes

$$E(N_{od}\hat{\mu}^2) = N_{od}(\mu^2 + \hat{\sigma}^2(\hat{\mu})). \tag{A.6}$$

where subscript index *od* stands for *off diagonal*. Let's subtract eq. A.6 from eq. A.5, to have the following expression

$$E\Big(\sum_{od} X_k X_l - \hat{\mu}^2\Big) = N_{od}\mu^2 - N_{od}(\mu^2 + \hat{\sigma}^2) = -N_{od}\hat{\sigma}^2; \tag{A.7}$$

now it is possible to derive a new relation for $\hat{\sigma}^2$ from equation A.7, i.e.

$$\hat{\sigma}^2 = \frac{1}{N_{od}} \sum_{od} \Big(\hat{\mu}^2 - X_k X_l\Big). \tag{A.8}$$

Since the sum of all the elements of the correlation matrix is equal to zero, i.e.

$$\sum_{k,l} \Big(\hat{\mu}^2 - X_k X_l\Big) = 0, \tag{A.9}$$

from the comparison between eq. A.8 and eq. A.9, by introducing the sum operator on the elements of the diagonal band ($\sum_b$) the following expression for $\hat{\sigma}^2$ can be derived

$$\hat{\sigma}^2 = \frac{1}{N_{od}} \sum_{b} \Big(X_k X_l - \hat{\mu}^2\Big). \tag{A.10}$$

In the hypothesis that the off-diagonal terms are not correlated, $\hat{\sigma}^2$, defined by equation A.10, is an unbiased estimator for σ^2, as long as N_{od} is correctly defined. It is possible to verify that $N_{o.d.}$ is equal to $N^2 - N$ in the uncorrelated case, $(N-2M)(N-M)$ in the method BMBC and $(N-M)(N-M+1)$ in the case of CB. These are, respectively, the denominators of equations 5, 23 and 24.

References

[1] M. P., M. K., Direct numerical simulation: A tool in turbulence research, Annual Review of Fluid Mechanics 30 (1998) 539–578. doi: 10.1146/annurev.fluid.30.1.539.

[2] J. J., M. R. D., What are we learning from simulating wall turbulence?, Philosophical Transactions of The Royal Society A Mathematical Physical

and Engineering Sciences 365 (1852) (2007) 715–732. doi:10.1098/rsta.2006.1943.

[3] R. S., L. P., The linear response of turbulent flow to a volume force: comparison between eddy-viscosity model and dns, Journal of Fluid Mechanics 790 (2016) 104–127. doi:10.1017/jfm.2016.4.

[4] P. E. E. P. S. Cheung S. H., Oliver T. A., M. R. D., Bayesian uncertainty analysis with applications to turbulence modeling, Reliability Engineering System Safety 96 (9) (2011) 1137–1149. doi:10.1016/j.ress.2010.09.013.

[5] O. T. A., M. R. D., Bayesian uncertainty quantification applied to rans turbulence models, Journal of Physics: Conference Series 318 (4) (2011) 042032. doi:10.1088/1742-6596/318/4/042032.

[6] A. C. on Standards, Aiaa guide for verification and validation of computational fluid dynamics simulations, AIAA Paper 98-G-077. doi:10.2514/4.472855.001.

[7] Y. P. K. Donzis D. A., S. K. R.

[8] von Storch H., Z. F. W., Statistical Analysis in Climate Research, 1st Edition, Cambridge University Press, 2001. doi:10.1017/CBO9780511612336.

[9] H. S., J. J.

[10] J. M. B. Conway R. W., M. L. W., Some problems of digital systems simulation, Management Science 6 (1) (1959) 92–110. doi:10.1287/mnsc.6.1.92.

[11] C. R. W., Some tactical problems in digital simulation, Management Science 10 (1) (1963) 47–61. doi:10.1287/mnsc.10.1.47.

[12] S. B. W., Batch size effects in the analysis of simulation output, Operations Research 30 (3) (1982) 556–568. doi:10.1287/opre.30.3.556.

[13] M. M. S., S. B., Overlapping batch means: something for nothing?, WSC '84 Proceedings of the 16th Conference on Winter simulation 27 (1984) 226–230. doi:10.1109/WSC.2011.6147767.

[14] L. A. M., W. D. Kelton, Simulation modeling and analysis, iii Edition, Boston: McGraw-Hill, 2000. doi:10.1080/01966324.1983.10737115.

[15] R. G. C. Box G., Jenkins G. M., L. G. M., Time Series Analysis: Forecasting and Control, 5th Edition, Wiley Prentice-Hall, 2015. doi:10.1111/jtsa.12194.

[16] B. P. M. T.

[17] U. R. Oliver T. A., Malaya N., M. R. D., Estimating uncertainties in statistics computed from direct numerical simulation, Physics of fluids 26 (3). doi:10.1063/1.4866813.

[18] H. J., Time Series Analysis, 1st Edition, Princeton University Press, 1994. doi:10.1017/S0266466600009440.

[19] B. P., Probability and Measure, 4th Edition, Wiley, 2012.

[20] Z. L., C. L., Limit Theory for Mixing Dependent Random Variables, Mathematics and Its Applications, Vol. 378, 1996.

[21] F. G. S., Grouping observations in digital simulation, Management Science 24 (5) (1978) 510–521. doi:10.1287/mnsc.24.5.510.

[22] P. D. B., Three curious properties of the sample variance and autocovariance for stationary processes with unknown mean, Journal of the American Statistician 47 (4) (1993) 274276. doi:10.1080/00031305.1993.10475997.

[23] W. W. W. S., Time Series Analysis: Univariate and Multivariate Methods, 2nd Edition, Pearson, 2005. doi:10.2307/2289741.

[24] B. P. M. T., Automatic autocorrelation and spectral analysis, xii Edition, Springer-Verlag, 2006. doi:10.1007/1-84628-329-9.

[25] L. P., Immersed-boundary simulations of turbulent flow past a sinusoidally undulated river bottom, European Journal of Mechanics - B/Fluids 55 (2015) 340–347. doi:10.1016/j.euromechflu.2015.08.007.