Comparison of strategies for the optimal sampling of random functions

Sena Mursel
Graduate Student, Dept. of Civil Engineering, Lehigh University, Bethlehem, PA, USA

Daniel Conus
Associate Professor, Dept. of Mathematics, Lehigh University, Bethlehem, PA, USA

Wei-Min Huang
Professor, Dept. of Mathematics, Lehigh University, Bethlehem, PA, USA

Manuel Miranda
Associate Professor, Dept. of Engineering, Hofstra University, Hempstead, NY, USA

Paolo Bocchini
Professor, Dept. of Civil Engineering, Lehigh University, Bethlehem, PA, USA

ABSTRACT: Engineering problems and applications are usually characterized by uncertainty. The accurate quantification of possible effects of those uncertainties is of great importance. The most common approach to do so is probabilistic simulation, but this strategy may become impractical if the associated deterministic problem is very complex and only a small number of its evaluations is possible. The selection of a small-to-moderate number of samples able to capture the probabilistic characteristics of the whole set of samples is an important open problem. One approach to perform this selection is a technique called "Functional Quantization by Infinite Dimensional Centroidal Voronoi Tessellation (FQ-IDCVT)". However, this method of selection focuses on mean-square optimality, and usually leads to reduction in the variability of the set of random sample functions. Hence, to capture the extreme values as well as the other probabilistic characteristics of the set of random samples, different methods for the selection of the set of samples are presented and new approaches are proposed in this study. To show differences among different methods, a numerical application is presented, consisting in the selection of ground motion time histories, and the results are compared using several performance measures.

1. INTRODUCTION

In engineering practice, a certain degree of uncertainty is frequently encountered. Specifically in structural engineering, the uncertainty can be seen in several areas including materials, loads, and geometry. This uncertainty needs to be characterized and studied in order to obtain an accurate and effective representation of the structural behavior. The traditional way of quantifying the uncertainty is probabilistic approaches. Monte Carlo Simulation (MCS) is commonly considered to be the only universal method that leads to the accurate representation of random processes, fields, or waves, in general random functions (Bergman et al., 1997). Although MCS is still one of the most powerful and versatile techniques, the major drawback is the computational time that restricts the areas of applicability, particularly when the problem is very complex, or only a small number of evaluations is possible. In these cases, a probabilistic technique that can represent the random function in an effective way is required. Various techniques, including Karhunen–Loève series expansion (Stefanou et al., 2007), truncated polynomial chaos series expansion (Blatman et al., 2010), and "Stochastic Reduced-Order Models" (Grigoriu, 2009) deal with this problem.

These techniques have computational and theoretical limitations, especially when non-Gaussian functions are involved (Field and Grigoriu, 2004).

As an alternative to the techniques aforementioned, "Functional Quantization by Infinite Dimensional Centroidal Voronoi Tessellation (FQ-IDCVT)" (Miranda and Bocchini, 2015) achieves an optimal selection of random samples by focusing on mean-square optimality. FQ-IDCVT provides "quanta", a predefined number *N* of samples of a random function and their corresponding probability masses (i.e., weights) to best approximate the given random function. It has been successfully applied to multi-dimensional, non-Gaussian and non-stationary problems (Christou et al., 2018). Although this technique offers a simple implementation and arguably it is the most versatile method among similar approaches, it intuitively reduces the variance as the optimality is in mean-square sense, and hence quanta end up focusing on the bulk of the distribution and being smoothened. In addition, FQ-IDCVT uses a methodology based on Lloyd's method (Lloyd, 1982) which leads to significant computational cost and it requires compromises in problems with high dimensionality and resolution.

Hence, the available methods have certain limitations and to overcome these limitations, this study proposes several alternatives. Different methods for the selection of the set of samples are presented and new combinations of approaches are proposed in this study to capture the tails as well as the other probabilistic characteristics of the whole set of random samples. In particular, strategies are applied to a stationary random process and to ground motion time histories. Since the sample size of ground motion records used is usually governed by computational costs and the risk assessment for earthquakes is highly sensitive to the uncertainty in the ground motion records, this is a proper example application.

This paper is organized in the following order. Section 2 introduces FQ-IDCVT and other sampling strategies. In Section 3, all strategies are applied to a non-Gaussian stationary one-dimensional random function and to ground motion time histories and the results are compared using several performance measures. The last section summarizes the conclusions and findings.

2. METHODOLOGY

2.1. Strategies in the original domain

2.1.1. FQ-IDCVT

Functional Quantization (FQ) aims to approximate the probabilistic description of the given random function with a small to moderate number of selected samples. The random function is considered to be given by a large number of realizations. Carefully constructed quanta and their associated weights $\{f_i, p_i\}_{i=1}^{N}$ will represent an approximation of all the realizations. Miranda and Bocchini proved the theoretical base of the algorithm. The implementation is simple and can be summarized with the following steps.

1. Define the inputs: the set of N_{sim} realizations of the random function, number of quanta N, the number of iterations n_{iter} or convergence criterion, and the probabilistic description of the field such as the marginal distribution and spectrum.
2. Randomly select N out of N_{sim} realizations as the initial set of quanta.
3. Compute the distances between each realization and each quantum based on the L^2 norm.
4. Define a set of tassels V_i, $i = 1,2,\dots,N$ by assigning each realization to one tassel so that the realization is closer to the ith quantum than to any other quantum.
5. Update each quantum by averaging all realizations belonging to the same tassel $(\widehat{f_k})$ $k = 1,2,\dots,N_i$ where $N_i$is the number of realizations in each tassel.
6. Repeat steps 3-5 until either convergence is met on the distortion metric (Δ) in Eq. (1) or n_{iter} iterations are completed.

$$\Delta\left(\{\mathrm{V_i}, \mathrm{f_i}\}_{\mathrm{i=1}}^{N}\right) = \sum_{i=1}^{N} \sum_{k=1}^{N_i} \frac{1}{N_i} \left\| \widehat{f_k} - f_i \right\|^2_{L^2(\Xi)}$$
$$with\ \widehat{f_k} \in V_i \quad (1)$$

7. Compute weights, p_i, as the fraction of realizations in each tassel.

2.1.2. *FQ-IDCVT combined with the Nearest Neighbor Search (NNS)*

In the traditional FQ-IDCVT, quanta are estimated through averaging all realizations in each tassel. Hence, the quanta end up being a smoothened form of the realizations. Although the information coming from realizations is transmitted to the quanta, the connection between the samples and quanta is not direct. Here, instead of using the quanta obtained through averaging, some of the original realizations are used, selected by proximity to the associated quantum. After the quanta are obtained through traditional FQ-IDCVT, an additional step is followed with nearest neighbor search (NNS). The realization which is the closest to its respective quantum (f_i) is used instead of the quantum as representative sample, with the weight computed as for FQ-IDCVT. The representative samples in this technique are thereby a subset of the initial set of realizations. In this way, the smoothing effect is reduced.

2.1.3. *Hybrid Method: Combination of the previous two approaches*

In the hybrid method, the specific characteristics of the previous two algorithms are combined. Some of the smoothing resulting from FQ-IDCVT is preserved but mitigated by the use of NNS. The strategy is straightforward: the quanta obtained by FQ-IDCVT and used for the nearest neighbor search. However, in this case instead of searching only the closest realization to each quantum, the *k* closest realizations to each quantum are found and averaged. These averages are used as representative samples of the random function, with weights computed as in FQ-IDCVT.

2.2. *Strategies in the frequency domain*

In previous strategies, all computations solely focused on the original domain of the random functions (i.e., time or space). Here, the spectral density function (SDF) of the quanta are calculated and used to select representative functions.

2.2.1. *FQ-IDCVT – SRM*

First the quanta are computed by FQ-IDCVT as explained in Sec. 2.1.1, then their SDF's are calculated. The Spectral Representation Method (SRM) is employed to generate one Gaussian sample with the SDF of each quantum (Shinozuka and Deodatis, 1991). Then, the samples are mapped to the probability distribution of the initial set of N_{sim} realizations using the Nataf transformation (Grigoriu, 1984). These samples are used as quanta, with the weights computed by FQ-IDCVT. In this way, the quanta match perfectly the probability distribution of the initial realizations, and this mitigates the loss of variance generated by FQ-IDCVT.

2.2.2. *FQ-IDCVT combined with the NNS and SRM*

In this case, the quanta are computed through FQ-IDCVT combined with the NNS, as explained in Sec. 2.1.2. Then, procedure continues as the one in Sec. 2.2.1.

All methods described in this section are summarized in Table 1.

Table 1. Steps in methods

Section No.				
2.1.1	FQ-IDCVT			
2.1.2	FQ-IDCVT	NNS		
2.1.3	FQ-IDCVT	k-NNS		
2.2.1	FQ-IDCVT	SDF	SRM	
2.2.2	FQ-IDCVT	NNS	SDF	SRM

In the table, blue background corresponds to time/space domain, whereas orange background represents calculations in the frequency domain.

3. APPLICATIONS

The methodology described in the previous section is applied to two different examples. The first application consists in studying a one-dimensional lognormal process and the second one involves the selection of ground motion time histories. Both examples address non-Gaussian processes, however, the first one is a stationary process and the second is nonstationary.

3.1. Lognormal Process

The first example focuses on a stationary one-dimensional process with a lognormal marginal distribution. Samples of the process are generated through SRM using the predefined SDF and a non-Gaussian marginal distribution (PDF). Three different SDF are considered:

$$S_{1_{FF}}(\omega) = \sigma^2 \cdot b^3 \cdot \frac{\omega^2}{4} \cdot exp(-b \cdot \omega^2) \quad (2)$$

$$S_{2_{FF}}(\omega) = \sigma^2 \cdot b^3 \cdot \frac{\omega^2}{64} \cdot exp(-b \cdot \omega^4) \quad (3)$$

$$S_{3_{FF}}(\omega) = \sigma^2 \cdot b^3 \cdot \frac{\omega^2}{16} \cdot exp(-b \cdot |\omega|) \quad (4)$$

where b is the correlation length, selected as $2\ sec$, while the upper cutoff frequency is set to be $3.14\ rad/sec$. The time domain of the random process in this example is $[0,6]\ sec$. For each SDF, a thousand realizations are generated and in total 3,000 realizations with a discretization size of 2,048 points are acquired. The quantizer size, *N*, and the number of iterations, n_{iter}, are set to 15 and 20, respectively. The mean and the standard deviation (σ) of the investigated lognormal distribution are 2 and 1, respectively.

A performance assessment of each technique discussed in Sec. 2 has been carried out to quantify the accuracy. The PDF and the autocorrelation function (ACF) of the quantizers are used to compare the accuracy of each technique. The statistical properties of the set of 3,000 samples were assessed and used as reference to determine the accuracy of the considered techniques. Four different metrics are utilized for the comparison: total normalized error in the empirical PDF (Eq. 5), maximum discrepancy in the empirical PDF (Eq. 6), total error in ACF (Eq. 7), and maximum discrepancy in ACF (Eq. 8).

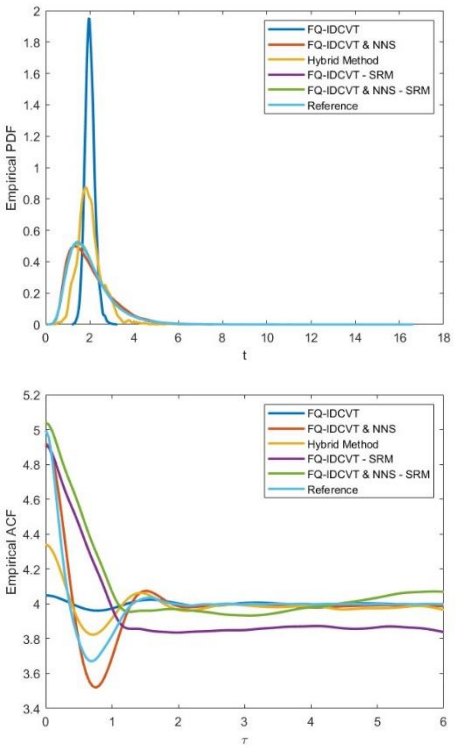

Figure 1. Empirical marginal distribution of the quanta and target (above) and ACF of quanta and target (below)

$$M_1 = \frac{1}{N_x}\sum_{j=1}^{N_x}\left|PDF\left(\widehat{f(x_j)}\right) - PDF\left(f(x_j)\right)\right| \quad (5)$$

$$M_2 = max\left(\left|PDF\left(\widehat{f(x_j)}\right) - PDF\left(f(x_j)\right)\right|\right) \quad (6)$$

$$M_3 = \frac{1}{N_x}\sum_{j=1}^{N_x}\left|R_{\widehat{ff}}(\tau_j) - R_{ff}(\tau_j)\right| \quad (7)$$

$$M_4 = max\left(\left|R_{\widehat{ff}}(\tau_j) - R_{ff}(\tau_j)\right|\right) \quad (8)$$

where $PDF(f(x_j))$ is the value of the empirical probability density of the quanta at time x_j, $PDF(\widehat{f(x_j)})$ is the probability density of the realizations at time x_j, $R_{ff}(\tau_j)$is the value of ACF of the quanta at the time lag τ_j, and $R_{\widehat{ff}}(\tau_j)$ is the value of ACF at the time lag τ_j of the realizations and N_x is the number of discretization points, set to be 2,048 as the resolution of the realizations.

The empirical marginal distribution of the quanta and the stationary ACF evaluated through these mechanisms for a single seed are shown in

Figure 1. Figure 2 shows Box and Whisker plots of these metrics using 50 different seeds.

3.2. Ground Motion Time Histories

The second application considers the selection of ground motion records. The Pacific Earthquake Engineering Research (PEER) Center, NGA-East and NGA-West2 strong motion database were used for this study (Peer Ground Motion Database, 2013).

In total, 2000 ground motion time history records were incorporated in this study and the emphasis is made on acceleration time histories. The magnitude of records used has a range between 4 and 8. There is no restriction on the number of recordings corresponding to the same earthquake. A set of 10 ground motion records used in this example is depicted in Figure 3.

Each record has its own length and time step. The maximum time step is initially determined to be $0.05\ sec$, and all realizations are then down sampled to such frequency, to achieve a more coherent collection of realizations. The ground motion time histories are shifted in such a way that the peak ground accelerations (PGA) for each record occur at the same instant in time. Zeros are then added to time histories with shorter lengths of time to make them equal in size. The quantizer size, N, and the number of iterations, n_{iter}, for this application are chosen as 50 and 20, respectively.

Due to the nonstationarity of the process, N needs to be larger compared to the first application to represent the samples more precisely.

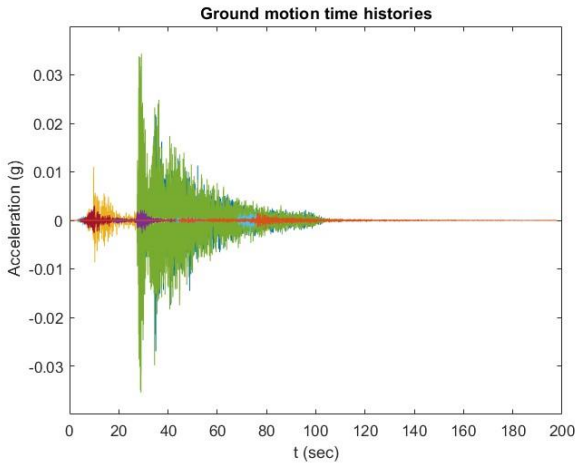

Figure 3. A set of 10 random records

The same assessment procedure is applied to this example, using the same metrics along with several additional performance measurements. Because of the nonstationary nature of the random function, additional metrics for the error in PDF and ACF at particular time instants ($t = 0.00\ sec$, $t = 20.00\ sec$, $t = 40.00\ sec$, $t = 60.00\ sec$, $t = 80.00\ sec$, $t = 100.00\ sec$) are introduced (Eqs. 9, 10 and 11). Hence, a total of 22 metrics (4 for the overall time history and 18 for the specified time instants) are utilized to assess and contrast the effectiveness of each strategy.

$$MP1_{t_i} = \frac{1}{\mathrm{N_x}} \sum_{j=1}^{N_x} \left| PDF\left(\widehat{f(t_i)}\right) - PDF\left(f(t_i)\right) \right| \quad (9)$$

$$MP2_{t_i} = \max\left(\left| PDF\left(\widehat{f(t_i)}\right) - PDF\left(f(t_i)\right) \right| \right) \quad (10)$$

$$MA_{t_i t_j} = \left| R_{\widehat{ff}}(t_i, t_j) - R_{ff}(t_i, t_j) \right| \quad (11)$$

where $PDF(f(t_i))$ is the value of the empirical probability density of the quanta at time t_i, $PDF\left(\widehat{f(t_i)}\right)$ is the value of the empirical probability density of the collection of realizations at time t_i, $R_{ff}(t_i, t_j)$ is the value of ACF of the quanta at time t_i, and $R_{\widehat{ff}}(t_i, t_j)$ is the value of ACF of the samples at time t_i and $t_i = 0, 10, 20, \cdots, 100\ sec$.

The Box and Whisker plots of whole set of time histories and at one specific time instant using 50 distinct seeds are presented in Figure 4. Similar results were obtained for the other time instants. The patterns of these plots in Figure 2 and Figure 4 are similar to each other. This demonstrates that the methodologies perform similarly, regardless of the process stationarity.

4. CONCLUSIONS

This study presents the comparison of effectiveness and accuracy of five different sampling strategies of a stationary and a nonstationary random processes using various measures. Each of these strategies has advantages of its own. Since FQ-IDCVT is based on a mean-square optimality, it underestimates the variability of the random function, thus the level of accuracy in the tails is only moderate. Moreover, in metrics

incorporating the PDF of quanta, FQ-IDCVT has large errors. However, FQ-IDCVT & NNS and

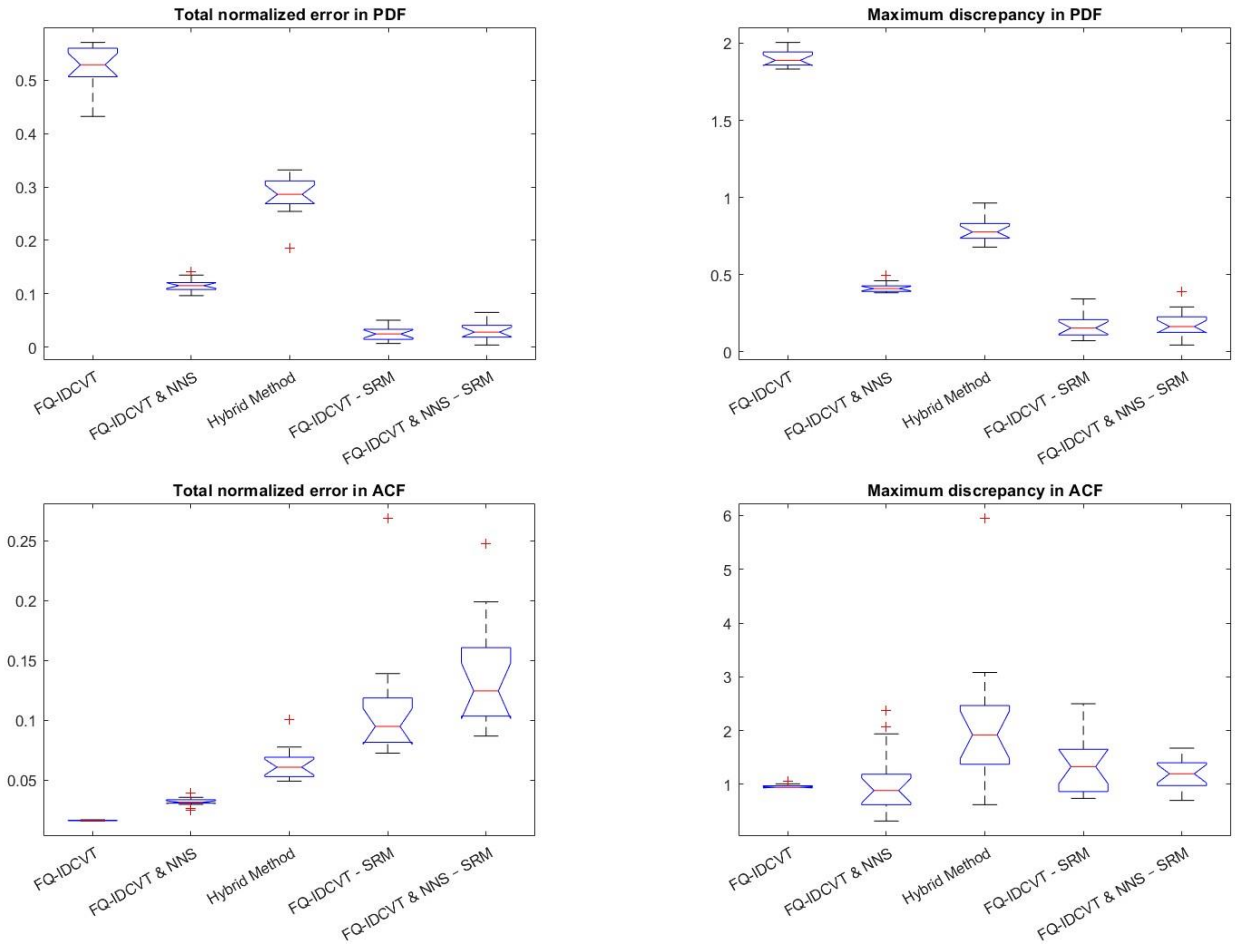

Figure 2. Box and Whisker plots of the metrics M_1, M_2, M_3, and M_4 for 5 different methods described in Sec. 2. The number of experiments is set to 50 for each metric. The metrics including PDF of the quanta (upper row) are higher in FQ-IDCVT, whereas metrics including ACF of the quanta (lower row) are higher in FQ-IDCVT & NNS and FQ-IDCVT & NNS - SRM.

FQ-IDCVT & NNS – SRM can represent the variability in the sample set more accurately. As a result, the marginal distribution of the sets resulting from these approaches is more similar to that of the entire sample collection. On the other hand, FQ-IDCVT and FQ-IDCVT – SRM predict autocorrelation far more accurately than FQ-IDCVT & NNS and FQ-IDCVT & NNS – SRM both across all time histories and at particular time instants. Despite the fact that FQ-IDCVT and FQ-IDCVT – SRM do a poor job of capturing variability (and therefore the first point of the ACF), the rest of the autocorrelation functions are more accurately estimated. The hybrid method on the other hand, can be a good compromise as it has smaller errors in the ACF but higher errors in PDF. It integrates the benefits and drawbacks of different approaches and might be viewed as desirable because it performs moderate to good on each criterion.

5. ACKNOWLEDGMENTS

This work is part of the activities of the "Catastrophe Modeling Center" at Lehigh

University (www.catmodeling.org). The financial support of Lehigh University through the "Accelerator" program and the "Research Futures: Major Program Development" grant is gratefully acknowledged. The opinions and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the sponsoring organizations.

6. REFERENCES

Blatman, G., & Sudret, B. (2010). An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis. *Probabilistic Engineering Mechanics*, *25*(2), 183-197.

Bergman, L. A., Shinozuka, M., Bucher, C. G., Sobczyk, K., Dasgupta, G., Spanos, P. D., & Zhang, R. (1997). A state-of-the-art report on computational stochastic mechanics. *Probabilistic Engineering Mechanics*, *12*(4), 197-321.

Christou, V., Bocchini, P., Miranda, M. J., & Karamlou, A. (2018). Effective sampling of spatially correlated intensity maps using hazard quantization: Application to seismic events. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, *4*(1), 04017035.

Field Jr, R. V., & Grigoriu, M. (2004). On the accuracy of the polynomial chaos approximation. *Probabilistic Engineering Mechanics*, *19*(1-2), 65-80.

Grigoriu, M. (1984). Crossings of non-Gaussian translation processes. Journal of Engineering Mechanics, 110(4), 610-620.

Grigoriu, M. (2009). Reduced order models for random functions. Application to stochastic problems. *Applied Mathematical Modelling*, *33*(1), 161-175.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE transactions on information theory*, *28*(2), 129-137.

Miranda, M. J., & Bocchini, P. (2015). A versatile technique for the optimal approximation of random processes by functional quantization. *Applied Mathematics and Computation*, *271*, 935-958.

PEER Ground Motion Database. (2013). Retrieved January 22, 2023, from https://ngawest2.berkeley.edu/site

Shinozuka, M., & Deodatis, G. (1991). Simulation of stochastic processes by spectral representation.

Stefanou, G., & Papadrakakis, M. (2007). Assessment of spectral representation and Karhunen–Loève expansion methods for the simulation of Gaussian stochastic fields. *Computer methods in applied mechanics and engineering*, *196*(21-24), 2465-2477.

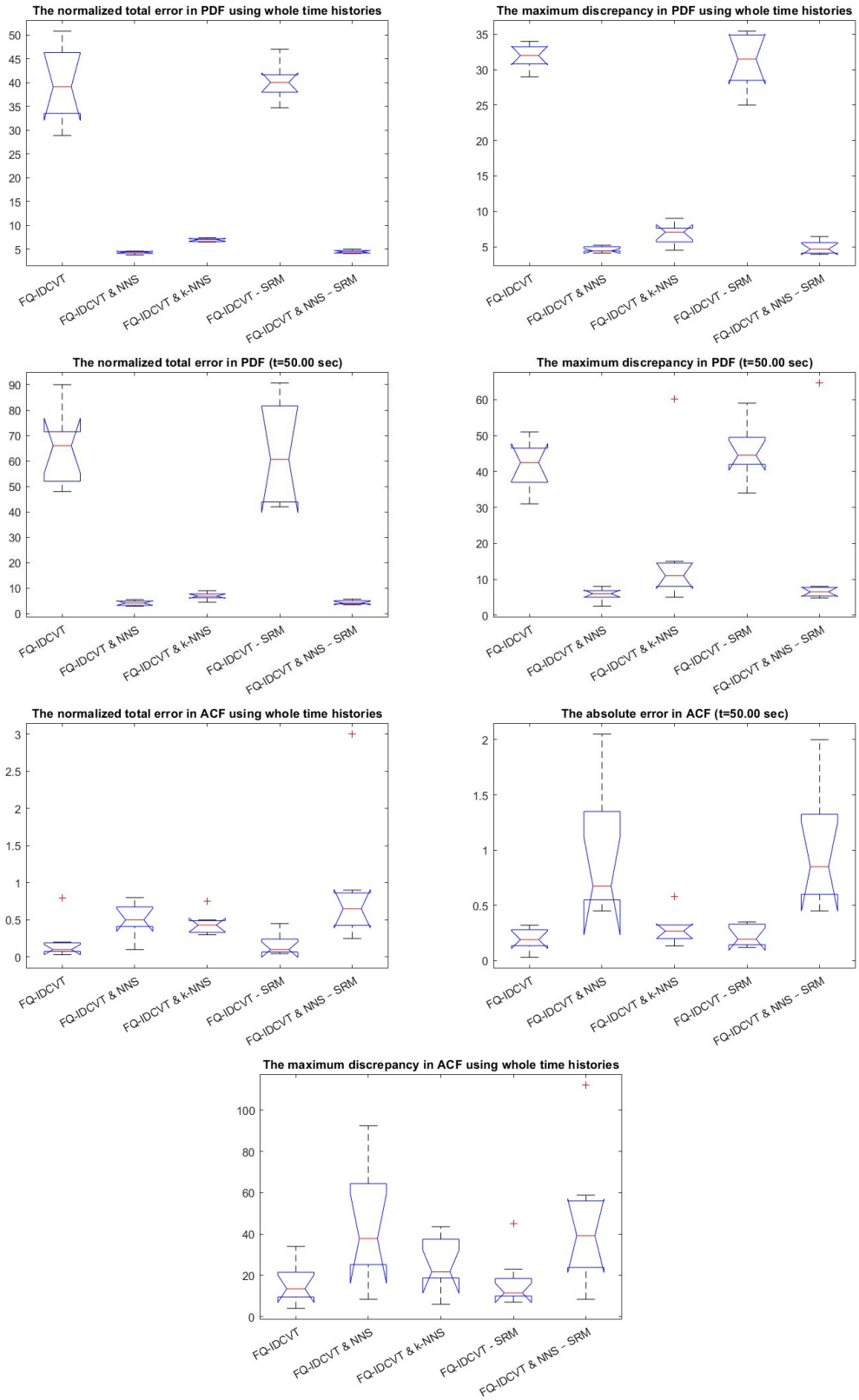

Figure 4. Box and Whisker plots of the metrics M_1, $M_2, M_3, M_4, MP1_{t=50}, MP2_{t=50}$ and $MA_{t=50}$ for 5 different methods described in Sec. 2. The number of experiments is set to 50 for each metric