14th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP14
Dublin, Ireland, July 9-13, 2023

Finite dimensional surrogates for extreme events

Hui Xu
PhD Student, Center for Applied Mathematics, Cornell University, Ithaca, USA

Mircea D. Grigoriu
Professor, Dept. of Civil Engineering, Center for Applied Mathematics, Cornell University, Ithaca, USA

Kurtis R. Gurley
Professor, Dept. of Civil Engineering, University of Florida, Gainesville, USA

ABSTRACT: Numerical solutions of stochastic problems require the representation of random functions in their definitions by finite dimensional (FD) models, i.e., deterministic functions of time and finite sets of random variables. It is common to represent the coefficients of these FD surrogates by polynomial chaos (PC) models. We propose a novel model, referred to as the polynomial chaos translation (PCT) model, which matches exactly the marginal distributions of the FD coefficients and approximately their dependence. PC- and PCT- based FD models are constructed for a set of test cases and a wind pressure time series recorded at the boundary layer wind tunnel facility at the University of Florida. The PCT-based models capture the joint distributions of the FD coefficients and the extremes of target times series accurately while PC-based FD models do not have this capability.

1. Introduction

The solution of a broad range of problems in science and engineering involves extremes of random processes $X(t)$ over finite times intervals, e.g., extreme climate events and design responses of dynamical systems subjected to random loads Grigoriu. (2020); Easterling et al. (2000); Grigoriu and Samorodnitsky. (2015). Yet, most practical methods for calculating the distribution of the extreme random variable $X_t = \sup_{0tt} jX(t)j$ are based on the mean rate at which the real-valued process $X(t)$ or its absolute value crosses with positive slope specified levels Leadbetter et al. (1983) (Chap. 7), which is available analytically for mean square differentiable Gaussian processes $X(t)$ and memoryless transformation of these processes, referred to as translation processes Gioffré et al. (2000). If $X(t)$ does not have these properties, the distribution of the extreme random variable X_t can be approximated from crossing of time series $X(t_0); X(t_1); : : : ; X(t_n)$ defined by the values of $X(t)$ at a finite set $0 = t_0 < t_1 < < t_n = t$ of times in $[0; t]$ Naess and Gaidai. (2008). The accuracy of this approximation depends on the time step and the properties of the samples of $X(t)$. For example, the approximation fails if the samples of $X(t)$ are not differentiable, e.g., the Brownian motion process.

It is proposed to approximate the distribution of $X_t = \sup_{0tt} jX(t)j$ by that of $X_{d;t} = \sup_{0tt} jX_d(t)j$, where $X_d(t)$ is a finite dimensional (FD) model of $X(t)$, i.e., a deterministic function of time and $d < ¥$ random variables which has the following two properties. First, the distributions of X_t and $X_{d;t}$ are similar for a sufficiently large stochastic dimension d. Accordingly, the distribution of X_t can be estimated from samples of $X_d(t)$. Second, samples of $X_d(t)$ can be generated by standard Monte Carlo algorithms. In contrast, samples of $X(t)$ cannot be generated since, generally, stochastic processes have infinite stochastic dimensions as uncountable families of random variables indexed

by time.

2. FINITE DIMENSIONAL MODELS

Let $X(t)$ be a real-valued, zero-mean stochastic process on a bounded time interval $[0;t]$ with correlation function $c(s;t) = E[X(s)X(t)]$. Denote by $\{l_k\}$ and $\{j_k(t)\}$, $k = 1;2;:::$, the eigenvalues and the eigenfunctions of the correlation function of $X(t)$. It is assumed that $c(s;t)$ is continuous so that its eigenfunctions are real-valued continuous function on $[0;t]$ Mercer. (1909).

The family of FD models of $X(t)$ has the form

$$X_d(t) = \sum_{k=1}^{d} Z_k j_k(t); \quad 0 \le t \le t; \qquad (1)$$

where the random coefficients $\{Z_k = \int_0^t X(t) j_k(t)\, dt\}$ are the projections of $X(t)$ on the basis functions $\{j_k(t)\}$. Simple calculations show that $E[Z_k] = 0$ and $E[Z_k Z_l] = l_k d_{kl}$, so that the zero-mean random variables $\{Z_k\}$ are uncorrelated. They are independent if $X(t)$ is Gaussian.

The FD models $X_d(t)$ have two notable properties. First, they are defined on the same probability space as $X(t)$ so their samples are paired with those of $X(t)$. Second, for given time t, the random variables $X_d(t)$ converge in mean square to $X(t)$ as $d \to \infty$ since

$$E\left[X_d(t) - X(t)\right]^2 = \sum_{k=d+1}^{\infty} l_k j_k(t)^2 \to 0;$$

as $d \to \infty$ by Mercer's theorem Mercer. (1909). This converges implies the converges in probability of $X_d(t)$ to $X(t)$ and, therefore, in distribution. This observation and Theorem 18.10 of van der Vaart. (1998) imply that the finite dimensional distributions of $X_d(t)$ converge to those of $X(t)$ as $d \to \infty$.

3. FD-BASED ESTIMATES OF EXTREMES

Denote by F_t and $F_{d;t}$ the distributions of the extremes $X_t = \sup_{0 \le t \le t} |X(t)|$ and $X_{d;t} = \sup_{0 \le t \le t} |X_d(t)|$ of $X(t)$ and $X_d(t)$ in the bounded time interval $[0;t]$. We give conditions under which $F_{d;t}$ converges to F_t as $d \to \infty$. Under these conditions, F_t can be estimated from samples of $X_d(t)$ provided that d is sufficiently large. This is essential in applications since samples of $X_d(t)$ can be generated by standard Monte Carlo algorithms while samples of $X(t)$ are not available. For simplicity, we assume as in the previous section that $X(t)$ is real-valued. Extension to vector-valued processes is straightforward.

Property 1. If $X(t)$ has continuous samples and it-s correlation function $c(s;t) = E[X(s)X(t)]$ is continuous and the finite dimensional distributions of $X_d(t)$ converge to those of $X(t)$ as $d \to \infty$, then the distribution $F_{d;t}$ of $\sup_{0 \le t \le t} |X_d(t)|$ converges to the distribution F_t of $\sup_{0 \le t \le t} |X(t)|$ as $d \to \infty$.

This property results by showing that the condition-s of the Theorem 8.2 of Billingsley. (1968) are satisfied. Details can be found in Xu and Grigoriu. (2022). The practical implication of this property is that the distribution of the extreme random variable $\sup_{0 \le t \le t} |X(t)|$ can be estimated from samples of FD models $X_d(t)$ of $X(t)$ provided that d is sufficiently large. This is essential in applications since the distribution of $\sup_{0 \le t \le t} |X(t)|$ is available analytically in special cases of limited practical interest and samples of $X(t)$ cannot be generated.

Property 2. If $X(t)$ is a Gaussian process with continuous samples and its correlation function $c(s;t) = E[X(s)X(t)]$ is continuous and if the finite dimensional distributions of $X_d(t)$ converge to those of $X(t)$ as $d \to \infty$, then the sequence of random variables $\sup_{0 \le t \le t} |X_d(t) - X(t)|$ converges to zero in probability as $d \to \infty$.

The proof can be found in Xu and Grigoriu. (2022). This means that the "bad" subset

$$W_d(e) = \{w \in W : \sup_{0 \le t \le t} |X_d(t) - X(t)| > e\}$$

of the sample space W on which the samples of $X(t)$ and $X_d(t)$ differs by more than any $e > 0$ can be made as small as desired by increasing d since $P[W_d(e)] \to 0$ as $d \to \infty$. Accordingly, most of the samples of $X(t)$ can be represented by the samples of $X_d(t)$ for a sufficiently large stochastic dimension d.

4. PC AND PCT MODELS

Our objective is to construct models of $Z = (Z_1; \ldots; Z_d)^T$ from its samples which are accurate in the sense that their joint distributions match the joint distribution of Z, and efficient, i.e., standard Monte Carlo algorithms can be used to generate samples of these models.

The Rosenblatt transformation Rosenblatt. (1952) provides a model with these features. It shows that the components of $Z = (Z_1; \ldots; Z_d)$ can be related to the components of, e.g., a vector $G = (G_1; \ldots; G_d)$ with independent standard Gaussian variables, by the mapping

$$\begin{aligned} Z_1 &= F_1^{-1}(G_1) \\ Z_k | Z_{k-1}; \ldots; Z_1 &= F^{-1}_{k|k-1;\ldots;1}(G_k); \end{aligned} \tag{2}$$

where F_k is the distribution of Z_k, $F_{k|k-1;\ldots;1}$ is the distribution of $Z_k | Z_{k-1}; \ldots; Z_1$. If the mapping in (2) is available, samples of Z can be obtained from samples of G, which can be generated by standard algorithms. Since the conditional distributions in mapping $G \to Z$ are available analytically only in special cases, they have to be constructed numerically in most applications. Their construction from the joint distribution of Z is computationally demanding and the resulting conditional distributions are likely to be unsatisfactory, particularly when dealing with heavy tail distributions. The construction of the conditional distributions $F_{k|k-1;\ldots;1}$ from data is not feasible when dealing with high dimensional vectors and relatively small data sets.

This section develops approximations of the Rosenblatt transformation for the random coefficients $(Z_1; \ldots; Z_d)$ of the FD models in (1) based on polynomial chaos (PC) and an extension of this representation, referred to as PCT models. These models of $(Z_1; \ldots; Z_d)$ are denoted by $Z^{PC} = (Z_1^{PC}; \ldots; Z_d^{PC})$ and $Z^{PCT} = (Z_1^{PCT}; \ldots; Z_d^{PCT})$.

The PC models considered here are quadratic forms of independent standard Gaussian variables $G_1; \ldots; G_d$ defined by

$$\begin{aligned} Z_k^{PC} &= E[Z_k] + \sum_{j=1}^{d} a_{k;j} G_j \\ &\quad + \sum_{1 \le j < l \le d} a_{k;j;l} (G_j G_l - E[G_j G_l]). \end{aligned} \tag{3}$$

The coefficients $\{a_{k;j}; a_{k;j;l}\}$ in (3) are determined by minimizing the objective function

$$\begin{aligned} & e_1(a_{k;j}; a_{k;j;l}) \\ &= g_1 E[\|Z - Z^{PC}\|_2^2] \\ &\quad + g_2 \max_{i_1 < i_2} \|h_{i_1;i_2}() - h^{PC}_{i_1;i_2}(a_{k;j}; a_{k;j;l})\|_2 \\ &\quad + g_3 (\|E[ZZ^T] - E[Z^{PC}(Z^{PC})^T]\|) \end{aligned} \tag{4}$$

where $h_{i_1;i_2}()$ is the histogram of $(Z_{i_1}; Z_{i_2})$ and $h^{PC}_{i_1;i_2}(a_{k;j}; a_{k;j;l})$ is the histogram of $(Z^{PC}_{i_1}; Z^{PC}_{i_2})$ for given coefficients $\{a_{k;j}; a_{k;j;l}\}$. The Matlab function histcounts2 is used to construct the two dimensional histograms of $(Z_{i_1}; Z_{i_2})$ and $(Z^{PC}_{i_1}; Z^{PC}_{i_2})$. The error between the two matri-ces is described by the norm $\|\cdot\|_2$, i.e., the absolute largest eigenvalue of the error matrix. We consider the set of all pairs of components rather than all components to minimize calcu-lations. The weighting coefficients $g_1; g_2; g_3$ are such that the components $E[\|Z - Z^{PC}\|_2^2]$, $\max_{1 \le i_1 < i_2 \le d} \|h_{i_1;i_2}() - h^{PC}_{i_1;i_2}(a_{k;j}; a_{k;j;l})\|_2$ and $\|E[ZZ^T] - E[Z^{PC}(Z^{PC})^T]\|$ contribute equally to the objective function (4). We set $g_1 = 0$ if Z and Z^{PC} are not defined on the same probability space since the mean error $E[\|Z - Z^{PC}\|_2^2]$ cannot be calculated.

The components of the PCT models are defined by

$$Z_k^{PCT} = F_k^{-1} F^{PC}_k(Z^{PC}_k); \quad k = 1; \ldots; d; \tag{5}$$

where F_k^{PC} is the distribution of Z_k^{PC} for given coefficients $\{a_{k;j}; a_{k;j;l}\}$. The coefficients $\{a_{k;j}; a_{k;j;l}\}$ in (5) are determined by minimizing the objective function

$$\begin{aligned} & e_2(a_{k;j}; a_{k;j;l}) \\ &= w_1 E[\|Z - Z^{PCT}\|_2^2] \\ &\quad + w_2 \max_{1 \le i_1 < i_2 \le d} \|s_{i_1;i_2}() - s^{PCT}_{i_1;i_2}(a_{k;j}; a_{k;j;l})\|_2 \\ &\quad + w_3 \max_{1 \le i_1 < i_2 \le d} \|h_{i_1;i_2}() - h^{PCT}_{i_1;i_2}(a_{k;j}; a_{k;j;l})\|_2; \end{aligned} \tag{6}$$

where $h_{i_1;i_2}()$ is as in (4), $s_{i_1;i_2}()$ is the spectral measure of $(Z_{i_1}; Z_{i_2})$, $s^{PCT}_{i_1;i_2}(a_{k;j}; a_{k;j;l})$ and

$h^{PCT}_{i_1;i_2}(|a_{k;j};a_{k;j;l}|)$ are the spectral measure and the histogram of $(Z^{PCT}_{i_1};Z^{PCT}_{i_2})$ for given coefficients $\{a_{k;j};a_{k;j;l}\}$. Spectral measures of $(Z_{i_1};Z_{i_2})$ are metrics which quantify the likelihood that $(Z_{i_1};Z_{i_2})$ are simultaneously large, see (5.3) and (5.4) in Grigoriu. (2019) for definitions and Resnick. (2007), Chap.6 for technical details. We sort the samples of the two-dimensional vectors $(Z_{i_1};Z_{i_2})$ and $(Z^{PCT}_{i_1};Z^{PCT}_{i_2})$ according to their lengths such that the first sample is the furthest to the ori-gin and construct the spectral measures from the top 10% of these samples. The Matlab function histcounts2 is used to construct the two dimensional histograms and spectral measures of $(Z_{i_1};Z_{i_2})$ and $(Z^{PCT}_{i_1};Z^{PCT}_{i_2})$. We consider the set of all pairs of components rather than all components to minimize calculations. The weighting coefficients $w_1;w_2;w_3$ are such that the components $E[\|Z - Z^{PCT}\|_2^2]$, $\max_{1i<_1i\ d_2}\|s_{i\ ;i\ (1)}\|_2$ $s^{PCT}_{i_1;i_2}(|a_{k;j};a_{k;j;l}|)\|_2$ and $\max_{1i_1<i_2 d}\|h_{i_1;i_2}()$ $h^{PCT}_{i_1;i_2}(|a_{k;j};a_{k;j;l}|)\|_2$ contribute equally to the objective function (6). We set $w_1 = 0$ if Z and Z^{PCT} are not defined on the same probability space since the mean error $E[\|Z - Z^{PCT}\|_2^2]$ cannot be calculated. The second and third terms of $e_2(a_{k;j};a_{k;j;l})$ quantify differences between the dependence structure of Z and Z^{PCT}. The third term is an approximate metric for the differences between the joint distributions of Z and Z^{PCT} while the second term measures the differences between the tail dependence of these random vectors.

Example 1. Let $X_1(t);X_2(t)$, $0 \ t \ t$, be real-valued processes defined by the differential equations

$$\begin{aligned} \ddot{X}_1(t) + a_1\dot{X}_1(t) + b_1X_1(t) &= k_1V(t); \\ \ddot{X}_2(t) + a_2\dot{X}_2(t) + b_2X_2(t) &= k_2V(t) \end{aligned} \quad (7)$$

with the initial conditions $X_i(0) = 0$ and $\dot{X}_i(0) = 0$, $i = 1;2$, where $a_i;b_i;k_i > 0$, $i = 1;2$ are constants. The input is the translation process $V(t) = F^{-1}(W(t))$, where F is the Gamma distribution function with the shape parameter n and scale parameter 1, $W(t)$ is the stationary solution of $dW(t) = -JW(t)\,dt + \sqrt{2J}\,dB(t)$, $J > 0$, and B denotes the standard Brownian motion.

From Grigoriu. (2021) (Chap.2), the solution of (7) is

$$X_i(t) = \int_0^t \frac{k_i}{y_i} e^{-a_i(t-u)/2} \sin(y_i(t-u))V(u)du; \quad (8)$$

where $y_i = (b_i - a_i^2/4)^{1/2}$, $i = 1;2$.

Our objective is to construct FD models for the vector-valued process $X_1(t);X_2(t)$. Since (7) has to be solved numerically, $V(t)$ and $X_1(t);X_2(t)$ are defined and calculated at a finite set of times, e.g., the equally spaced times $t_i = iDt$, $i = 1;:::;n$, where $Dt = t/n$ denotes the integration time step. Denote by $h = (V(t_1);:::;V(t_n))$ and $z_i = (X_i(t_1);:::;X_i(t_n))$, $i = 1;2$, the discrete versions of the input $V(t)$ and of the processes $X_i(t)$, $i = 1;2$. The random vector h admits the representation $h = \sum_{k=1}^{n} Z_k v_k$, where $\{v_k\}$ are the eigenvectors of the covariance matrix $E[h\,h^T]$ and the random coefficients $\{Z_k\}$ are defined sample by sample by projection, i.e., $Z_k(w) = h^T(w)v_k$, $w \in W$. The corresponding FD model is $h_d = \sum_{k=1}^{d} Z_k v_k$. Since the differential equations (7) are linear, their solutions to h and h_d are linear forms of $\{Z_k\}$ denoted by $z_i = \{z_{i;j}\}$ and $z_{d:i} = \{z_{d:i;j}\}$, $i = 1;2$, $j = 1;:::;n$.

The thin solid lines of the left and right panels of Fig. 1 are estimates of $P(|z_i| > x)$ for $i = 1$ and $i = 2$ which are obtained directly from data, where $|z_i| = \max_{1jn}|z_{i;j}|$. These probabilities are viewed as truth. The other lines of the figure are calculated from samples of $z_{d:i}$ (heavy solid lines), $z^{PC}_{d:i}$ (dotted lines) and $z^{PCT}_{d:i}$ (dashed lines) for the first and second components (left and right panels). The heavy solid lines are the closest to the truth. The next best model is $z^{PCT}_{d:i}$ while $z^{PC}_{d:i}$ differs significantly from the truth. We prefer $z^{PCT}_{d:i}$ to $z_{d:i}$ since the set of samples of $z_{d:i}$ is defined by the available data so that it cannot be extended. In contrast, samples of any size can be generated from $z^{PCT}_{d:i}$ since its probability law is known.

5. FD MODEL FOR WIND FORCES

Construct FD models for the vector-valued wind pressure time series $X(t) = X_1(t);:::;X_m(t)$ recorded in the wind tunnel of the University of Florida and estimate the distributions of extremes of the time series.

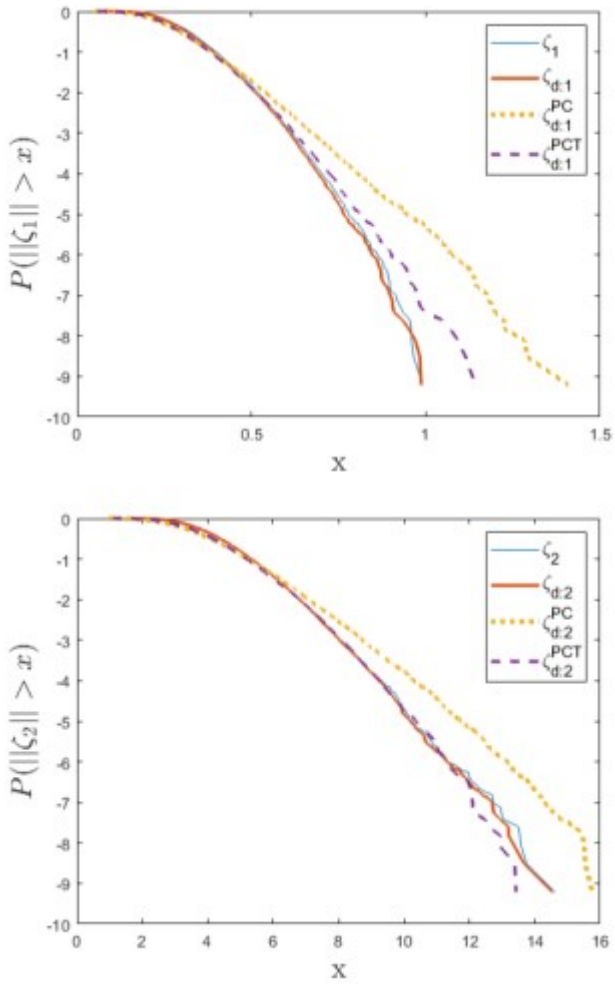

Figure 1: Estimates of the target probability $P(|z_i| > x)$ (thin solid line), estimates based on FD model (heavy solid line), estimates based on finite dimensional PC model (dotted line) and estimates based on finite dimensional PCT model in logarithmic scale (dashed line) for z_1 and z_2 (left and right panels).

Data is used to estimate the correlation functions of the vector-valued process $X(t)$, find the eigenfunctions of these functions and calculate the samples of the random coefficients $\{Z_{i;k}\}$ of the FD models of the components of $X(t)$ by projection as discussed in Sect. 2.

The joint distribution of the random vector whose components are the random variables $\{Z_{i;k}\}$ is obtained by translating polynomial chaos representation such that they match exactly the target marginal distribution, see Grigoriu. (2019). These models can be used to generate samples of the random coefficients $\{Z_{i;k}\}$ which are used to find the corresponding samples of FD models of $X(t)$.

The plots of Fig. 2 explore the effects of the dependence between the random coefficients of FD models on extremes. The estimates are unsatisfactory if the components of Z^{PCT} are assumed independent, an expected result since the resulting FD wind model is approximately Gaussian. They approach the target probability as the degree of the polynomial chaos is increased from two to four since this increases results in a superior representation of the dependence between the random variables $Z_{i;k}$. However, increasing the degree of the polynomial chaos does not improve the estimates based on PC models, since Z and Z^{PC} have different marginal distributions.

6. CONCLUSIONS

Finite dimensional (FD) models, i.e., deterministic functions of time and finite sets of random variables, have been constructed for a set of test cases and a wind pressure time series recorded at the UFBLWT facility in Gainesville by using polynomial chaos (PC) and polynomial chaos translation (PCT) models to represent their random coefficients. The components of PCT models are obtained from those of PC models by translation, so that they match exactly the target marginal distributions irrespective of the coefficients in their definition. The optimal values of the PCT coefficients minimize the discrepancy between the PCT and target joint properties, which are quantified by joint distributions and spectral measures. In summary, the PCT models match exactly the marginal distributions of the random coefficients of FD models by construction and capture their dependence with an accuracy that increases with the truncation level of the underlining PC models.

FD models with random coefficients represented by PC and PCT models have been constructed for a set of test cases and a 6-dimensional wind pressure time series recorded in the UFBLWT facility. The FD models with PCT random coefficients are superior to those with PC coefficients in the following sense. First, the PCT models provide a more accurate representation of the joint distributions of the random coefficients of FD models

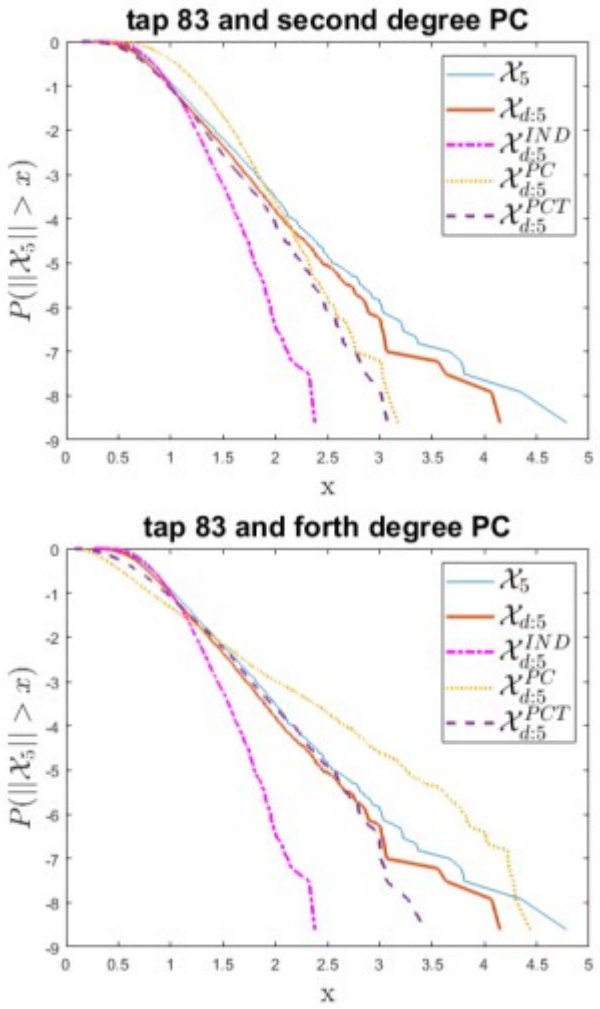

Figure 2: Estimates of the target probability P(jjX$_5$jj > x) (thin solid line), estimates based on FD model (heavy solid line), estimates based on finite dimen-sional independent model (dash-dotted line), estimates based on finite dimensional PC models (dotted line) and estimates based on finite dimensional PCT model in logarithmic scale (dashed line) based on second and forth degree polynomial chaos (left and right panels).

than the PC models. Second, the distributions of extremes of PCT-based FD models are similar to those of target time series while PC-based FD models do not have this capability. It is also shown that the performance of PCT-based FD models can be further improved by increasing their stochastic dimension and/or the order of their underlining PC models.

7. REFERENCES

Billingsley., P. (1968). Convergence of probability measure 1st Edition. Wiley, New York.

Easterling, D. R., Evans, J., Groisman, P. Y., Karl, T. R., Kunkel, K. E., and Ambenje., P. (2000). "Observed variability and trends in extreme climate events: a brief review." Bulletin of the American Meteorological Society, 81(4), 417–425.

Gioffré, M., Gusella, V., and Grigoriu., M. (2000). "Simulation of non-gaussian field applied to wind pressure fluctuations." Probabilistic Engineering Mechanics, 5(4), 339–346.

Grigoriu., M. (2019). "Pc translation models for random vectors and multivariate extremes." SIAM J. Sci. Comput., 41(2), A1228–A1251.

Grigoriu., M. (2020). "Data-based importance sampling estimates for extreme events." Journal of Computational Physics, 412, 1–21.

Grigoriu., M. (2021). Linear Dynamical Systems 1st Edition. Springer.

Grigoriu, M. and Samorodnitsky., G. (2015). "Reliability of dynamic systems in random environment by extreme value theory." Probabilistic Engineering Mechanics, 38, 54–69.

Leadbetter, M. R., Lindgren, G., and Rootzén., H. (1983). Extremes and Related Properties of Random Sequences and Processes. Springer-Verlag, New York.

Mercer., J. (1909). "Functions of positive and negative type and their connection with the theory of integral equations." Philosophical Transactions of the Royal Society A, 209, 415–446.

Naess, A. and Gaidai., O. (2008). "Monte carlo methods for estimating the extreme response of dynamic systems." Journal of Engineering Mechanics, 134(8), 628–636.

Resnick., S. (2007). Heavy-Tail Phenomena: Probabilistic and Statistical Modeling. Springer, New York.

Rosenblatt., M. (1952). "Remarks on a multivariate transformation." Ann. Math. Statistics, 23, 470–472.

van der Vaart., A. W. (1998). Asymptotic statistics.

Xu, H. and Grigoriu., M. (2022). "Finite dimension-al models for extremes of gaussian and non-gaussian processes." Probabilistic Engineering Mechanics, 68.