Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

EFFICIENT RISK ESTIMATION USING EXTREME VALUE THEORY AND SIMULATION METAMODELING

Joseph J. Kennedy
Armin Khayyer
Alexander Vinel
Alice E. Smith

Industrial and Systems Engineering
Auburn University
Shelby Center
Auburn, Al 36849, USA

ABSTRACT

This paper considers a new approach for constructing metamodels for capturing tail behavior in stochastic systems, e.g., simulation outputs. Specifically, we are concerned with the problem of global estimation of the conditional value-at-risk (CVaR) surface, given (stochastic) responses from a collection of design points. The approach combines stochastic kriging, which has previously been shown to work well for metamodeling of discrete-event simulation output, with extreme value theory, which is a powerful statistical tool for estimating tail behavior. We present the general methodology and promising results of preliminary computational experiments.

1 INTRODUCTION

Executing and analyzing high-fidelity simulation models covering a wide range of parameters are often quite computationally expensive, leading to significant attention paid in the literature to the area of metamodeling, or surrogate modeling, for simulation. Applications that are concerned with characterizing risk are particularly in need of such efforts. In these applications, the decision-makers are interested in accurately describing tail behavior for the underlying distributions, which usually requires proportionally more computational effort to observe sufficient rare events.

This paper proposes a metamodeling approach for globally characterizing a particular tail measure, Conditional Value-at-Risk (CVaR), combining two relevant methodologies: stochastic kriging (SK) and the peaks-over-threshold (POT) method for tail estimation. Stochastic kriging has been shown to be a powerful modeling tool for predicting discrete-event simulation outputs, including forecasting tail measures (Chen et al. 2012). While Chen et al. (2012) demonstrated promising results, their methodology is based on empirical (nonparametric) estimators of risk measures. An alternative approach to characterizing tail behavior has been studied under the umbrella of extreme value theory (EVT). Widely used in many computational areas, it so far has received rather limited attention in the operations research community in general and simulation applications in particular. EVT is known to improve tail estimates for various stochastic systems (McNeil and Saladin 1997). Consequently, in this paper, we are interested in evaluating whether this improvement persists if used within a metamodeling framework, such as stochastic kriging. To this end, we propose a two-stage metamodeling framework, which employs the peaks-over-threshold (POT) method from EVT to estimate CVaR for a set of given design points on the response surface and then applies stochastic kriging to the estimated values to enable global predictions.

Note that in this research, we explicitly concentrate on CVaR as the primary measure of risk and tail behavior. While other approaches exist, CVaR is widely accepted as a de facto standard in many stochastic modeling and optimization applications (Krokhmal et al. 2011). At the same time, it is worth emphasizing that our methodology is also applicable for predicting other measures of tail behavior, such as Value-at-Risk (VaR), quantiles or higher moment measures, as long as a corresponding expression can be derived for the peaks-over-threshold estimator.

The remainder of this paper is organized as follows. In Section 2, we briefly discuss relevant literature, focusing on recent developments in stochastic kriging and EVT applications. Section 3 presents an overview of the relevant aspects of EVT, risk measures, and stochastic kriging. In Section 4, we present the proposed method and then describe its performance in computational experiments in Section 5. Section 6 gives concluding remarks.

2 LITERATURE REVIEW

The problem of interpolating a response surface over a design space given stochastic observations has received significant attention from numerous perspectives, see Barton (1998) for a seminal paper. In this section, we concentrate only on the literature directly related to the two components of our proposed approach: stochastic kriging for estimating tail measures and EVT approaches to interpolation. The former has been recently emerging in the simulation community, and the latter have found success in various applied fields, such as hydrology. Note that we also provide a general introduction and overview (with the corresponding references) of EVT, stochastic kriging, and measures of risk in the next section.

Use of simulation experiments paired with surrogate models with the goal of predicting tail measures has been proposed by Chen and Kim (2016) and Chen et al. (2012) where they achieved global estimators of CVaR and its derivative by using stochastic kriging and explored VaR and CVaR estimation methods by fitting nonparametric estimators to a SK model. These methods showed promising results for creating global estimators for VaR and CVaR but only employed empirical nonparametric estimators. In 2009 (Liu and Staum 2009) used an adaptive allocation technique for allocating computational effort to approximate CVaR using stochastic kriging in a nested Monte Carlo environment.

Beguería and Vicente-Serrano (2006) employed regression techniques to the parameters of a generalized Pareto distribution (GPD) to model the extremes via a Peaks-Over-Threshold model. They were then able to achieve a spatial model for rainfall based on the parameters for the extreme distribution. Reza Najafi and Moradkhani (2013) modeled the GPD scale parameter using a hierarchical Bayesian model with a homogeneous shape parameter. Their work was based on real rain data sampled from gauges across the Columbia river basin. Their work focused on fitting the GPD parameters, which vary based on the spatial dimensions of their study; this work is similar to what we perform here; however, we will not assume a prior distribution for the parameters of the extreme distribution. Wu et al. (2019) conducted a similar analysis for China, and Bracken et al. (2018) proposed a multivariate nonstationary model where the parameters that are being estimated by the Bayesian process vary in time. Lima and Lall (2010) performed similar work by designing a Bayesian hierarchical model for seasonal maxima and showed effective results. Their results rely on the seasonal extreme (i.e., block-maxima methods, not peaks over threshold), so the ability to analyze the lower level extreme limits the analysis relative to the work done by Reza Najafi and Moradkhani (2013).

3 BACKGROUND OF TECHNICAL COMPONENTS

3.1 Extreme Value Theory

Extreme Value Theory (EVT) commonly describes one of two classes of theoretical results describing the asymptotic behavior of extreme events. The first branch (Block Maxima method) results from the limiting distribution of max/min of a sample, which is characterized by the Fisher-Tippet-Gnedenko Theorem. The

second branch (Peaks over Threshold, POT, method) characterizes the limiting behavior of exceedances beyond a threshold via the Generalized Pareto Distribution (GPD).

The Generalized Pareto Distribution is a family of continuous probability functions defined by the following cumulative distribution function:

$$G_{\xi,\beta}(x) = \begin{cases} 1-(1+\frac{\xi x}{\beta})^{\frac{-1}{\xi}} & \text{if } \xi \neq 0 \\ 1-e^{\frac{-x}{\beta}} & \text{if } \xi = 0 \end{cases} \tag{1}$$

where ξ is the shape parameter and $\beta > 0$ is the scale parameter. This family of distributions describes the behavior of observations exceeding some (high) threshold, as follows from the following result.

Theorem 1(Balkema and de Haan 1974; Pickands III 1975) Let x_0 be the right endpoint (finite or infinite) of a distribution $F(x)$ and $0 \leq u < x_0$, then we can find a positive measurable function $\beta(u)$ such that,

$$\lim_{u \to x_0} \sup_{0 \leq x \leq x_0 - u} |F(x) - G_{\xi,\beta(u)}(x)| = 0 \tag{2}$$

if and only if $F \in MDA(H_\xi)$ (the maximum domain of attraction of distributions with tail parameter ξ) i.e., $F \in MDA(H_\xi)$ if there exist sequences a_n and b_n such that $F^n(a_n x + b_n) \to H_\xi(x)$ as $n \to \infty$, for all $x \in \mathbb{R}$.

The power of this result comes from being able to describe the distribution of large values of a random variable by fitting the corresponding generalized distribution. There exists a rich variety of literature dedicated to best strategies to organize this calculation, most importantly, rules for selection of the threshold u. See for example Scarrott and MacDonald (2012), Ho and Wan (2002), Ferreira* et al. (2003). In this effort we consider a simple approach to allow for easy implementation, that is, setting a u threshold equal to some pre-set quantile (90% in our experiments).

3.2 Conditional Value at Risk and Its POT Estimates

A rigorous methodology for evaluating risk in stochastic models was proposed in Artzner et al. (1999). It describes risk measures in terms of coherence, which relies on several axioms such as translation invariance, subadditivity, positive homogeneity, and monotonicity. The most popular coherent measure is Conditional Value-at-Risk, which is a widely used way of estimating risk, especially in finance. For a random variable X, the quantity Value-at-Risk at level α is defined as the corresponding quantile: $VaR_\alpha(X) = F_X^{-1}(1-\alpha)$, where F_X^{-1} is the right continuous inverse of the cumulative distribution of X. In the case of continuously distributed X, CVaR is then defined as the expectation of values above the value-at-risk, i.e.,

$$CVaR_\alpha(X) = E\Big[X | X > VaR_\alpha(X)\Big].$$

Note that in the case of a discrete distribution, the formal definition is somewhat less intuitive, but allows for a similar interpretation, see Rockafellar and Uryasev (2002) for details. Parameter α is usually taken to be small, less than 5%, thus an accurate empirical approximation of CVaR requires at least $n \gg 1/\alpha$ observations. However, EVT provides a method of approximating it by fitting a model to only the tail of the distribution and then extrapolating the conditional expectation directly from it. The following description is discussed in more detail in McNeil (1999). This is achieved by considering the conditional probability of X above a threshold value, say u, with the corresponding exceedance distribution given by

$$F_u(x) = \frac{F(x+u) - F(u)}{1 - F(u)} \to F(x) = F_u(x-u)[1-F(u)] + F(u). \tag{3}$$

Since the distribution function is often unknown, we can approximate the exceedance distribution with a GPD with estimated values for $\xi < 1$ and β. Given N_u exceedances and N total observations above the

threshold we can approximate $1-F(u)$ by N_u/N. This gives a tail estimator of F for a value larger than u as

$$\hat{F}(x) = G_{\xi,\beta}(x-u)[1-F(u)] + F(u) \approx 1 - \frac{N_u}{N}\left(1 + \frac{\hat{\xi}(x-u)}{\hat{\beta}}\right)^{-\frac{1}{\hat{\xi}}} \tag{4}$$

where ξ and β are estimates of the GDP parameters. Thus, we can estimate the value at risk by inverting equation (4). To calculate the conditional value-at-risk the parameter β is shifted by $\xi(VaR_\alpha - u)$, giving the following estimate of $CVaR_\alpha(X)$

$$CVaR_\alpha(X) \approx \frac{VaR_\alpha(X)}{1-\hat{\xi}} + \frac{\hat{\beta}-\hat{\xi}u}{1-\hat{\xi}}. \tag{5}$$

This estimate is known as the peaks over threshold model for conditional value at risk (POT CVaR). The estimation of $\hat{\xi}$ also informs the modeler of the tail behavior of the distribution and its finite moments. See McNeil (1997) for more details on this approach. It should be noted that the methods for estimating the GPD parameters range widely and, depending on the data available and underlying distribution, can achieve different levels of bias and variance, see de Zea Bermudez and Kotz (2010), and Mackay et al. (2011). For the purposes of this paper we do not discuss these in any more detail or consider other approaches.

The other, more readily intuitive, method for estimating CVaR, that is not based on EVT, is the empirical average exceedance. Empirical CVaR is calculated by taking the expectation of the samples that exceed the value at risk threshold, say $q(\alpha)$. More specifically, for a sample $\{x_i\}_{i=1}^N$ the empirical $CVaR_\alpha(X)$ is defined as

$$\widehat{CVaR_\alpha}(X) = \frac{\sum_{i=1}^N x_i \mathbf{1}_{\{x_i \geq q(\alpha)\}}}{\sum_{i=1}^N \mathbf{1}_{\{x_i \geq q(\alpha)\}}}. \tag{6}$$

3.3 Stochastic Kriging

Ankenman et al. (2010) extended the theory of kriging to the stochastic simulation setting by considering both intrinsic and extrinsic uncertainties, where the former is inherent to the stochasticity of the simulation itself, and the latter is due to the unknown response surface. Given a training set $X = [x^{(1)}, x^{(2)}, ..., x^{(k)}]^T$, where $x^{(i)} = [x_1^{(i)}, x_2^{(i)}, ..., x_d^{(i)}]$, $i = 1, 2, ..., k$ are d-dimensional decision variables, the corresponding observed response vector is $\bar{Y} = [\bar{Y}^{(1)}, \bar{Y}^{(2)}, ..., \bar{Y}^{(k)}]^T$ and $\bar{Y}^{(i)}$ is the average of observations over the entire number of replications at design point i. A linear model, a realization of a mean 0 random field that represents fluctuations around the trend and exhibits spatial correlation, and a noise term are the three components representing the output of a stochastic simulation model of replication j at design point x:

$$y_j(x) = f(x)^T\beta + M(x) + \varepsilon(x) \tag{7}$$

where $M(x)$ is referred to as extrinsic uncertainty, $\varepsilon(x)$ is intrinsic noise, and $f(x)^t$ is a $(m \times 1)$ vector of polynomial basis functions.

Let $\Sigma_M(x,x') = Cov[M(x), M(x')]$ be a $k \times k$ extrinsic covariance matrix achieved using the spatial correlation function $r_M(x,x' : \theta)$ assuming that M is second-order stationary, then $\Sigma_M(x,x')$ can be rewritten as $\tau^2 r_M(x,x' : \theta)$. And let$\sum_M(x_0,.)$ be a $k \times 1$ vector containing the covariance between x_0 and all of the k design points. Also, let Σ_ε be the $k \times k$ intrinsic covariance matrix. Given the mean and covariance function of the GRF, the prediction at a new point x is:

$$\hat{y}(x) = f(x)^T\beta + \Sigma_M(x_0,.)^T[\Sigma_M + \Sigma_\varepsilon]^{-1}(\bar{y} - F\beta) = f(x)^T\beta + r(x)\left[R + \frac{C}{\tau^2}\right]^{-1}(\bar{y} - F\beta), \tag{8}$$

where R is $\Sigma_M/(\tau^2)$, C is the intrinsic noise, and $F_i = f(x_i)$. Note that in extreme cases where C (the intrinsic covariance matrix) is negligible compared to the extrinsic variance, the above equation becomes the same

as that of kriging. Additionally, if τ^2 (otherwise known as extrinsic variance) is negligible and close to zero, then the above equation approaches $\hat{y}(x) = f(x)^T\beta$ which is the same as a regression equation. (β, θ, τ^2) can be estimated using maximum likelihood, assuming Σ_ε is known and equal to $Diag\{V_1, V_2, ..., V_k\}$ where $V_i = V(x_i)/n_i$. n_i shows the number of replications at the design point x_i and the log-likelihood function of (β, θ, τ^2) is:

$$\ell(\beta, \theta, \tau^2) = -\ln[(2\pi)^{\frac{k}{2}}] - \frac{1}{2}\ln[|\tau^2 R_M(\theta) + \Sigma_\varepsilon|] - 1/2(\bar{y} - \beta B)^T[\tau^2 R_M(\theta) + \Sigma_\varepsilon]^{-1}(\bar{y} - \beta B) \quad (9)$$

The estimated parameters then can be used to predict $\hat{y}(x)$.

4 METHODOLOGY

The overall metamodeling framework is given in Figure 1. First, in the Initialization step, we select metamodel parameter values, which affect the model accuracy and computational effort. The two main phases are: Estimation and Metamodeling. In the estimation phase, after obtaining the simulation response for each design point, we apply the POT CVaR estimator. Then, the intrinsic variance of this estimator is evaluated by replicating this step. Given these observations, we can then construct the SK metamodel. Note that instead of the POT estimator in Step 4 we can also use a more naive empirical estimator (and will do so in our computational comparisons in Section 5), which results in the same model as the one proposed in Chen and Kim (2016). Finally, the whole process can be repeated if the desired accuracy is not achieved.

To estimate the GPD parameters, we use the probability-weighted moments method, which has been shown to be more reliable over the MLE estimates for sample sizes less than 500 (Hosking and Wallis 1987). For the experiments reported below, the threshold for exceedances is selected to be constant across the entire design space, set as the 90th percentile of the samples taken at each design point.

In the next section, we use a test case function with different kinds of random noise to represent a complex simulation, and sample from the design space using different computational budgets. The intent is to demonstrate the capability of POT-CVaR estimates compared to empirical CVaR estimates. The test case consists of different scenarios to capture potentially different behaviors of the randomness of a simulation.

5 EXPERIMENTATION

We demonstrate the capability of SK surrogate models for the proposed methodology with a test case of three different scenarios. We use a test case function with three kinds of random noise to represent a complex simulation, and sample from the design space using different computational budgets. The test case then represents a complex surface, and the scenarios represent various forms of randomness that will affect the estimation of CVaR, namely a finite tailed, a heavy-tailed, and a normal (thin-tailed) distribution of noise. The experiment consists of a fixed design strategy for selecting test points in the design space and is repeated for different computational budgets, employing both empirical and POT-CVaR estimation methods, and then fitting the surrogate models. For each computational budget we develop two stochastic kriging metamodels based on the same framework: one as described on Figure 1, and one based on empirical CVaR estimate, as proposed in Chen and Kim (2016). We then compare the two models in terms of the accuracy for a set of common test points (separate from the design points used in model construction), as measured against the known true values of CVaR.

5.1 Test Case Design and Experimental Parameters

The test case for the experiment is a well-known two-dimensional function with an additive noise term:

$$f_i(x_1, x_2) = x_1 \sin(\pi x_2) + x_2 \sin(\pi x_1) + \varepsilon_i(x_1, x_2), \quad (10)$$

where $\varepsilon_i(x_1, x_2)$ is random noise and $x_1, x_2 \in [-\pi, \pi]$. The scenario description is as detailed in Table 1 below. The function, as well as true CVaR values, are as described in Norton et al. (2019).

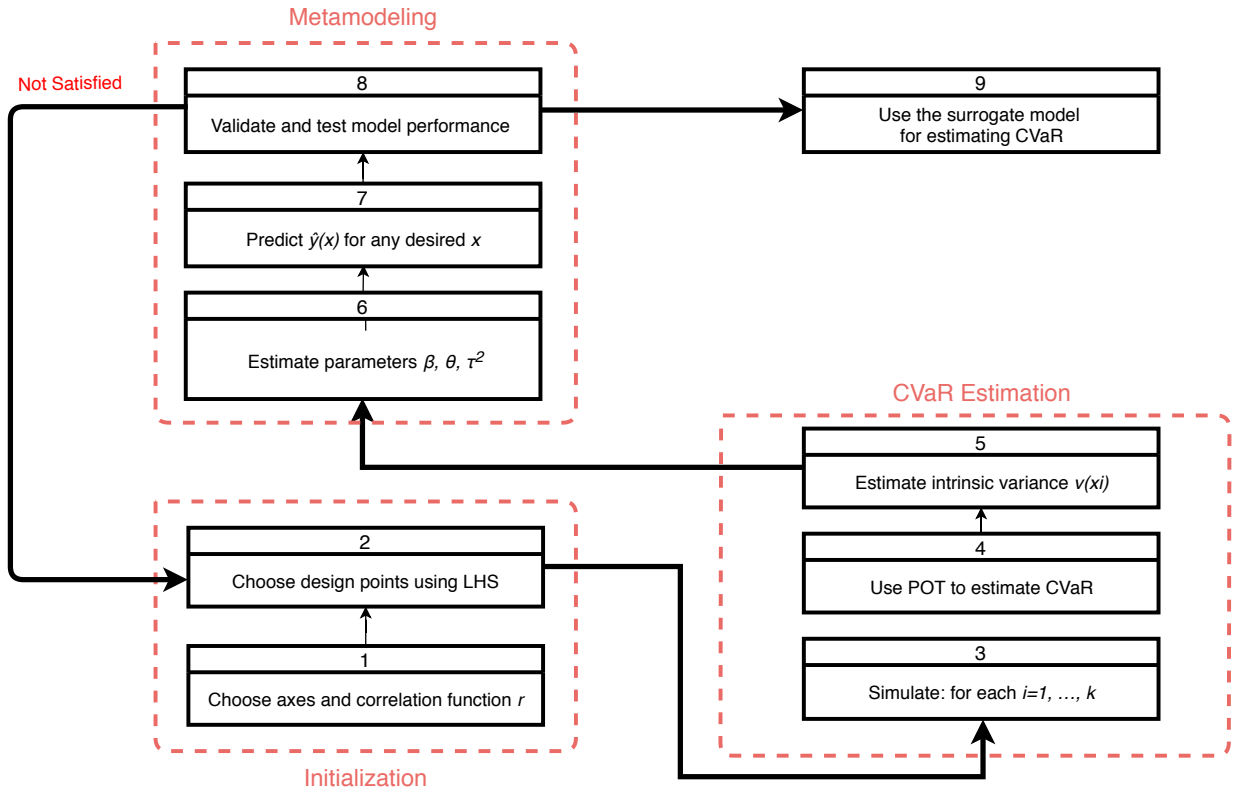

Figure 1: Proposed metamodeling framework

Table 1: Description of noise terms and expressions for the true CVaR values

Scenario Description	$\varepsilon_i(x_1,x_2)$	True CVaR $[\varepsilon_i(x_1,x_2)]$
normal (thin) tail	$N(\mu=0,\sigma=\sqrt{(x_1^2+x_2^2)})$	$\frac{\sqrt{(x_1^2+x_2^2)}\phi(\Phi^{-1}(\alpha))}{\alpha}$
finite tail	$Tri(\min=0,\max=\sqrt{(x_1^2+x_2^2)})$	$\sqrt{(x_1^2+x_2^2)}-\frac{2\alpha^{1/3}}{3}\sqrt{\frac{(x_1^2+x_2^2)}{2}}$
heavy tail	$Pareto(a=2,x_m=2+\sqrt{(x_1^2+x_2^2)})$	$\frac{2(2+\sqrt{(x_1^2+x_2^2)})}{(1-\alpha)^{1/2}}$

We sample the design space using a Latin hypercube sampling (LHS) design, which is a statistical sampling method to generate a near-random sample of parameter values. A one-dimensional LHS with n samples partitions the CDF evenly into n regions and then randomly picks one sample in each region. The d-dimensional LSH, on the other hand, generates one-dimensional samples for each input and then randomly combines the samples. This design has shown favorable performance in the literature, for example, (Helton and Davis, 2003). The computational budget for creating the SK model is divided over the number of runs, replications, and design points. The SK model is based on observations from k design points. In each design point, N runs (samples of the test function) are observed to obtain a POT (or empirical) estimate as described earlier. To evaluate the noise for constructing an SK model, POT (or empirical) estimation is repeated over m replications. Thus, the computational budget (total number of simulated values) is given by $N \times m \times k$. For our experiment we select three budget values and consider a variety of corresponding values for the number of runs, replications, and design points for a total of 16 trials, see Table 2.

For all trials we use $\alpha = 0.01$. Note that this implies that the lowest number of runs to empirically estimate CVaR is 100, in which case both empirical VaR and CVaR estimates are simply the worst-case values. We then expect that introduction of POT estimator can lead to improved accuracy at each design point, and consequently of the metamodel itself for the cases with $N = 100$. For each trial, we compare the proposed CVaR predictor (referred to as POT) with a straightforward approach based on the empirical estimate at each design point (referred to as EMP), constructed analogously to (Chen and Kim 2016). Our implementation relies on the Python SciPy package for minimizing the likelihood function and estimating the SK parameters. As different components may have very different variabilities, it is helpful to normalize the design points first. We use the Sklearn MinMaxScaler function to do so.

Table 2: Experimental trials summary

Total Budget	Number of Design Points (k)	Number of Replications of each design Point (m)	Number of Runs for Each Replication (N)	Budget Identifier
10000	10	10	100	1
	20	5	100	2
100000	10	100	100	3
	20	50	100	4
	10	10	1000	5
	20	5	1000	6
	50	20	100	7
	100	10	100	8
1000000	10	10	10000	9
	20	5	10000	10
	10	1000	100	11
	20	500	100	12
	100	10	1000	13
	200	5	1000	14
	100	100	100	15
	200	50	100	16

A major issue in getting good predictions from kriging is choosing an appropriate spatial correlation function and value of its parameter θ. In this study a Gaussian correlation function is used, given by $r(x-x',\theta_i) = exp\Big(-\sum_i \theta_i (x_i - x_i')^2\Big)$. θ_i determines the correlation decay with distance measured along dimension i. We choose the initial value of τ^2 as the variance of the residuals. Since our test function includes many oscillations, finding a well-specified polynomial or linear trend can be burdensome and might result in misspecification of the model. Therefore, we perform kriging with $b = 1$ and this results in estimating a constant mean over the domain, as opposed to a trend that varies over X. However, it should be noted that SK without the trend term can still lead to good predictions if the parameters are well-chosen.

Overfitting, misspecification, and too few design points are three of the many things that can result in an invalid and inaccurate metamodel. Posterior variance is one of the measures that can be used to validate a metamodel along with standard error measures, such as Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). Since posterior variance can be misleading due to misspecification of GRF, we utilize MAPE and MAE to report the performance of the surrogate models.

The same set of validation points are used to validate the metamodel corresponding to each budget combination, which enables us to draw comparison between the different noise distributions, and also between the different simulation budgets. The validation set contains 100 points, which are sampled from the problem domain using LHS.

5.2 Experimental Results

Table 3 reports the error measures (MAPE and MAE) between the estimated CVaR using the surrogate stochastic kriging model (POT or EMP) and the true CVaR value, averaged over the 100 test points. Additionally, Table 4 reports the outcome of the nonparametric Wilcox Signed Rank Test over the absolute error of the empirical and POT CVaR SK models. Note that the bolded values are significant at 95% confidence. Figure 2 give representative predicted (from trial 15) and true response surfaces for the three noise types.

Table 3: Measures of error of the surrogate stochastic kriging models averaged over 100 random test points. "Average improvement" row refers to the improvement achieved by the POT model over empirical, averaged over all trials.

Noise	Normal				Triangular				Pareto			
Measures	MAPE (%)		MAE		MAPE (%)		MAE		MAPE (%)		MAE	
Budget	POT	EMP	POT	EMP	POT	EMP	POT	EMP	POT	EMP	POT	EMP
1	**21.88**	27.23	**1.93**	2.33	610.71	**593.88**	1.71	**1.68**	**32.45**	49.05	**30.64**	44.30
2	25.99	**25.95**	**2.20**	2.43	440.27	**423.61**	**1.42**	1.43	**27.71**	47.01	**24.89**	42.37
3	**26.11**	32.77	**2.72**	3.40	362.63	**347.51**	**1.54**	1.56	**10.69**	35.17	**10.24**	31.53
4	**19.34**	23.09	**1.70**	2.06	575.86	**559.36**	1.54	**1.53**	**17.84**	39.78	**16.34**	35.44
5	**20.17**	28.22	**1.82**	2.29	493.72	**491.91**	1.53	1.53	**17.90**	19.19	**15.90**	17.63
6	**19.74**	25.34	**1.80**	2.05	412.42	**410.59**	1.63	1.63	**12.07**	15.54	**11.43**	14.71
7	**8.83**	14.96	**0.83**	1.43	235.43	**223.35**	**0.98**	1.00	**17.26**	38.95	**15.41**	34.33
8	**4.99**	12.76	**0.47**	1.12	**19.08**	29.18	**0.12**	0.13	**28.31**	45.54	**25.30**	40.39
9	**18.48**	26.65	**1.73**	2.23	450.81	**450.33**	1.55	1.55	**3.45**	3.58	**3.14**	3.27
10	**21.23**	25.11	**1.81**	1.93	442.17	**441.85**	1.46	1.46	**4.87**	5.08	**4.44**	4.65
11	**21.59**	29.46	**1.80**	2.28	551.64	**534.68**	1.54	**1.52**	**9.38**	34.76	**8.84**	31.03
12	**20.60**	24.94	**2.03**	2.38	528.12	**511.90**	1.46	1.46	**9.65**	34.97	**8.87**	31.05
13	**2.04**	2.38	**0.20**	0.24	**15.83**	16.52	0.13	0.13	**6.29**	10.59	**5.29**	9.03
14	**2.47**	3.09	**0.22**	0.27	**7.21**	22.22	0.05	0.05	**10.08**	13.89	**8.91**	12.24
15	**3.21**	13.01	**0.30**	1.13	**20.72**	26.82	**0.12**	0.14	**12.51**	36.52	**11.29**	32.31
16	**2.33**	13.15	**0.21**	1.15	**5.24**	17.19	**0.04**	0.08	**13.41**	35.47	**11.72**	31.19
Average Improvement	5.57		0.43		-4.44		0.00		14.45		12.68	

As it can be expected, in general a higher number of design points results in a better surrogate model across the different noise distributions. It can also be seen from the results that more design points are better than very low variance at a few points. For example, budget identifiers 7 and 8 have 5 times as many design points as budget identifiers 3 and 4 but have fewer replications at each design point, however the surrogate models achieved using budget identifiers 7 and 8 outperform those achieved using budget identifiers 3 and 4.

Most importantly, we can observe that for most trials and both measures, use of the POT estimator improves the accuracy of the stochastic kriging model for both normal and Pareto noise experiments. This improvement is statistically significant for all trials with Pareto noise, and most trials with normal noise. As expected, the difference is especially large for trials with N=100 (trials 1-4, 7, 8, 11, 12, 15, 16), for which the empirical estimator is not very accurate. This implies that the POT based SK model can more effectively manage the tradeoff between the three budget elements (k,N,m). For example, it may be possible to consider more design points (resulting in a more accurate SK model), while still maintaining adequate accuracy of the CVaR estimate. More thorough experiments are needed to carefully examine this tradeoff in practical applications. For the triangular noise, the difference between the two methods is smaller and not statistically significant in most trials (note that the relatively large MAPE measure in this case can be explained by relatively low values for the predicted CVaR, since the error term is bounded). This is not surprising, since for a triangular (finite) tailed distribution, with sufficiently many samples and

small α, the CVaR value quickly converges to the maximum of the domain, meaning that both empirical and POT estimates are very accurate.

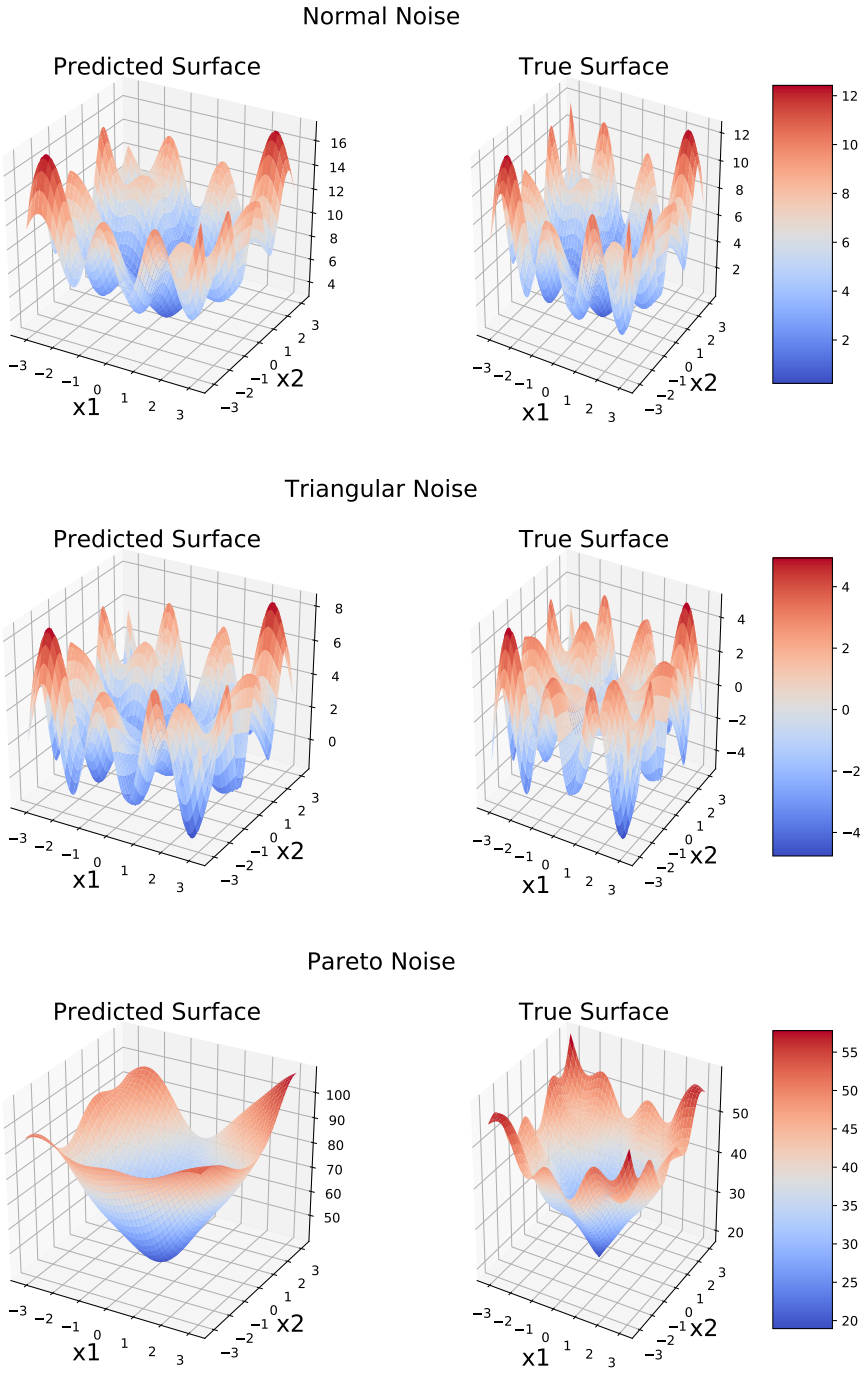

Figure 2: CVaR response surface, true and predicted by the POT SK model for the three noise types.

Table 4: P-value results for the Wilcox Signed Rank Test over the absolute error of the empirical and POT CVaR SK models (bolded value indicate significance at 95% confidence)

	Wilcox Signed Rank Test: P-Value Absolute Error		
Budget	Normal	Triangular	Pareto
1	0.1591	**0.0008**	$<$ **0.0001**
2	0.9465	0.2015	$<$ **0.0001**
3	0.7819	**0.0045**	$<$ **0.0001**
4	0.0003	0.3383	$<$ **0.0001**
5	**0.0002**	0.7168	**0.0108**
6	$<$ **0.0001**	**0.0284**	$<$ **0.0001**
7	$<$ **0.0001**	**0.0267**	$<$ **0.0001**
8	$<$ **0.0001**	**0.0050**	$<$ **0.0001**
9	0.7323	0.4402	**0.0315**
10	$<$ **0.0001**	0.4906	**0.0002**
11	**0.0017**	0.2190	$<$ **0.0001**
12	0.1794	0.4161	**0.0001**
13	0.6315	0.2660	$<$ **0.0001**
14	$<$ **0.0001**	0.6711	$<$ **0.0001**
15	$<$ **0.0001**	**0.0142**	$<$ **0.0001**
16	**0.0284**	$<$ **0.0001**	$<$ **0.0001**

6 CONCLUDING REMARKS

In this paper we proposed combining a stochastic kriging metamodeling approach to globally characterizing tail behavior of stochastic systems with a more accurate local estimator based on extreme value theory. Kriging metamodels depend on a wise choice of how to spend a computational budget, and hence a more accurate estimator at each design point has potential to translate into better global accuracy of the metamodel. Our experiments generally confirm this phenomenon. Specifically, we observed that for stochastic functions that exhibit a spatially dependent fat tailed noise the use of POT estimation for CVaR at high α levels outperforms the common empirical approach. We have shown (experimentally) that for a triangular (finite) tailed distribution of noise the POT does not exhibit any gained accuracy. The observed improvement varies with the computational budget and type of noise. Clearly, more work is needed to confirm these observations and learn more about the value of EVT for kriging metamodels in stochastic simulation.

ACKNOWLEDGMENTS

Part of this work was funded by the project "SBIR Phase III – Analytical Framework and Modeling to Support Wargaming Logistics" by Frontier Technology, Inc.

REFERENCES

Ankenman, B., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling". *Operations Research* 58(2):371–382.

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath. 1999. "Coherent Measures of Risk". *Mathematical finance* 9(3):203–228.

Balkema, A. A., and L. de Haan. 1974. "Residual Life Time at Great Age". *The Annals of Probability* 2(5):792–804.

Barton, R. R. 1998. "Simulation Metamodels". In *1998 Winter Simulation Conference. Proceedings (Cat. No. 98CH36274)*, Volume 1, 13–16 December, Washington, DC, USA, 167–174. Institute of Electrical and Electronics Engineers, Inc.

Beguería, S., and S. M. Vicente-Serrano. 2006. "Mapping the Hazard of Extreme Rainfall by Peaks over Threshold Extreme Value Analysis and Spatial Regression Techniques". *Journal of applied meteorology and climatology* 45(1):108–124.

Bracken, C., K. Holman, B. Rajagopalan, and H. Moradkhani. 2018. "A Bayesian Hierarchical Approach to Multivariate Nonstationary Hydrologic Frequency Analysis". *Water Resources Research* 54(1):243–255.

Chen, X., and K.-K. Kim. 2016. "Efficient VaR and CVaR Measurement via Stochastic Kriging". *INFORMS Journal on Computing* 28(4):629–644.

Chen, X., B. L. Nelson, and K.-K. Kim. 2012. "Stochastic Kriging for Conditional Value-at-Risk and Its Sensitivities". In *Proceedings of the Winter Simulation Conference*, WSC '12. 9-12 December, Berlin, Germany: Winter Simulation Conference.

de Zea Bermudez, P., and S. Kotz. 2010. "Parameter Estimation of the Generalized Pareto Distribution—Part I". *Journal of Statistical Planning and Inference* 140(6):1353 – 1373.

Ferreira*, A., L. de Haan*, and L. Peng. 2003. "On Optimising the Estimation of High Quantiles of A Probability Distribution". *Statistics* 37(5):401–434.

Ho, A. K., and A. T. Wan. 2002. "Testing for Covariance Stationarity of Stock Returns in the Presence of Structural Breaks: An Intervention Analysis". *Applied Economics Letters* 9(7):441–447.

Hosking, J. R., and J. R. Wallis. 1987. "Parameter and Quantile Estimation for the Generalized Pareto Distribution". *Technometrics* 29(3):339–349.

Krokhmal, P., M. Zabarankin, and S. Uryasev. 2011. "Modeling and Optimization of Risk". *Surveys in Operations Research and Management Science* 16(2):49 – 66.

Lima, C. H., and U. Lall. 2010. "Spatial Scaling in a Changing Climate: A Hierarchical Bayesian Model for Non-Stationary Multi-Site Annual Maximum and Monthly Streamflow". *Journal of Hydrology* 383(3-4):307–318.

Liu, M., and J. Staum. 2009. "Estimating Expected Shortfall with Stochastic Kriging". In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, November 29 –December 2, Austin, Texas, 1249–1260.

Mackay, E. B., P. G. Challenor, and A. S. Bahaj. 2011. "A Comparison of Estimators for the Generalised Pareto Distribution". *Ocean Engineering* 38(11-12):1338–1346.

McNeil, A. J. 1997. "Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory". *ASTIN Bulletin: The Journal of the IAA* 27(1):117–137.

McNeil, A. J. 1999. "Extreme Value Theory for Risk Managers". *Departement Mathematik ETH Zentrum* 12(5):121–237.

McNeil, A. J., and T. Saladin. 1997. "The Peaks Over Thresholds Method for Estimating High Quantiles of Loss Distributions". In *Proceedings of 28th International ASTIN Colloquium*, August 10 – 13, Cairns, Australia, 23–43.

Norton, M., V. Khokhlov, and S. Uryasev. 2019, October. "Calculating CVaR and bPOE for Common Probability Distributions with Application to Portfolio Optimization and Density Estimation". *Annals of Operations Research*.

Pickands III, J. 1975. "Statistical Inference Using Extreme Order Statistics". *The Annals of Statistics* 3(1):119–131.

Reza Najafi, M., and H. Moradkhani. 2013. "Analysis of Runoff Extremes Using Spatial Hierarchical Bayesian Modeling". *Water Resources Research* 49(10):6656–6670.

Rockafellar, R. T., and S. Uryasev. 2002. "Conditional Value-At-Risk for General Loss Distributions". *Journal of Banking & Finance* 26(7):1443–1471.

Scarrott, C., and A. MacDonald. 2012. "A Review of Extreme Value Threshold Estimation and Uncertainty Quantification". *REVSTAT–Statistical Journal* 10(1):33–60.

Wu, Y.-b., L.-q. Xue, and Y.-h. Liu. 2019. "Local and Regional Flood Frequency Analysis Based on Hierarchical Bayesian Model in Dongting Lake Basin, China". *Water Science and Engineering* 12(4):253–262.

AUTHOR BIOGRAPHIES

JOSEPH J. KENNEDY is a doctoral student in the Industrial and Systems Engineering Department of Auburn University. He holds a MS in Mathematical Sciences from University of West Florida. His email is jjk0023@auburn.edu.

ARMIN KHAYYER is a third-year Ph.D. student in the Industrial and Systems Engineering and a second-year MSc student in Computer Science and Software Engineering Department of Auburn University. His email is azk0100@auburn.edu.

ALEXANDER VINEL is an Assistant Professor in the Industrial and Systems Engineering Department of Auburn University. He holds a doctorate degree in industrial engineering from the University of Iowa. His research interests are in the areas of stochastic optimization and risk-averse decision making, with applications in transportation systems and data analytics. His email address is alexander.vinel@auburn.edu.

ALICE E. SMITH is the Joe W. Forehand/Accenture Distinguished Professor of the Industrial and Systems Engineering Department at Auburn University with a joint appointment with the Department of Computer Science and Software Engineering. Her research focus is analysis, modeling, and optimization of complex systems with emphasis on computational intelligence. She holds a U.S. patent and has authored more than 200 refereed publications. These have accumulated over 12,500 citations

with an H Index of 50 (Google Scholar). Dr. Smith has been a principal investigator on over U.S. $10 million of sponsored research. She is a four-time Fulbright Scholar with appointments in Turkey, Colombia, and Chile. She is a Fellow of IEEE and of ISE and a registered Professional Engineer. Dr. Smith is Editor in Chief of *INFORMS Journal on Computing* and Area Editor of *Computers & Operations Research*.Her email is smithae@auburn.edu.