14th World Congress on Computational Mechanics (WCCM)
ECCOMAS Congress 2020
11–15 January 2021, Paris, France

A PROBLEM OF PROBABILITY DENSITY FUNCTION ESTIMATION FOR LARGE DIMENSIONAL SPACES WITH MANY LOW-INFLUENTIAL DIMENSIONS

Jergus Suja[1], **Martin Kubicek**[2] **and Tomas Koutnik**[3]

[1] Brno University of Technology, Faculty of Mechanical Engineering, Technická 2896/2, 616 69 Brno, Czech Republic, 170200@vutbr.cz

[2] UptimAI, Vinohradská 2396/184, 130 00 Praha 3, Czech Republic, martin.kubicek@uptim.ai

[3] UptimAI, Vinohradská 2396/184, 130 00 Praha 3, Czech Republic, tomas.koutnik@uptim.ai

Key Words: HIGH DIMENSIONAL PROBABILISTIC MODELING, UNCERTAINTY PROPAGATION, UNCERTAINTY QUANTIFICATION.

All engineering problems consider uncertainties. These range from small production uncertainties to large-scale uncertainties coming from outside, such as variable wind speed or sunlight. Currently, modern methods for uncertainty propagation have large difficulties with estimation of statistics for large-scale problems which considers hundreds of these uncertain parameters. Due to the complexity of the problem and limitations of the modern methods, a common approach for modelling large scale problems is to select a few important parameters and model statistics for these parameters.

However, this can lead to an important problem. In this paper, an application of the UptimAI's UQ propagation algorithm is used to discuss a new problem arising from very high dimensional spaces where a large number of parameters have negligible impact on the final solution. In other words, when a problem consists of a great number of uncertain design parameters, common practice is to focus on the most important ones and neglect the non-influential ones. However, a combination of a great number of non-influential parameters can lead to completely different results. This is especially a problem for modelling large dimensional statistical models where a common approach is to perform sensitivity analysis and neglect the non-influential variables, i.e. set the non-influential variables to nominal value. Therefore, using a common approach of neglecting the non-influential variables could lead to a dramatic error and hence, we call this problem "many times nothing killed a horse". This problem cannot be observed for cases with a small number of design parameters, which are commonly solved in statistical modelling. The reason for this issue is that the combined influence of neglected variables is extremely small and such that has no influence on the final output.

Application of the UptimAI's UQ propagation algorithm to modern engineering problems and the possibilities of mitigation of the cumulative influence of non-influential parameters is discussed in detail.

The problem is shown on a case of economic load dispatch (ELD) problem which consist of 140 dimensions [1]. To this problem was applied UptimAI's UQ propagation algorithm to obtain accurate statistics for the problem and to get deeper insight into the statistics. Using the accurate model obtained by UptimAI's algorithm, we compare statistics of using only important variables and using all variables. This lead to a significant difference between results and such that put a large question mark on standard

approach. The obtained results are validated with the Monte Carlo simulation applied directly to ELD problem.

Application of UptimAl's UQ propagation algorithm to modern engineering problems and the possibilities of mitigation of the cumulative influence of non-influential parameters is discussed in detail.

1 Introduction

In order to be able to correctly determine the mathematical model of a large-scale problem, we need to overcome several limitations. The basic pitfalls of this issue includes computational expense, course of dimensionality [2], nonlinearity [3], correlated inputs [4], and others. The main issue in large-scale problems is the curse of dimensionality. With an increasing number of dimensions, the volume of the space rises so fast that the available data becomes sparse. To be able to estimate reliable statistics, the amount of data needed, often grows exponentially. It is natural that most research on large-scale problems has taken the path of reducing dimensions with the vision of improving the efficiency, and accuracy of statistical properties.

In [5], authors focused on the division and comparison of existing methods for dimension reduction. They contrasted the classical linear method (principal component analysis - PCA [6]) with twelve novel non-linear methods such as Kernel PCA [7], Isomap [8], [9], Sammon mapping [10], etc. From the obtained results, they concluded that nonlinear techniques for dimensionality reduction are, despite their large variance, not yet capable of outperforming the traditional PCA.

The topic of dimensionality reduction significantly interferes with the field of artificial intelligence, where the topology of feature-selection methods was proposed in [11]. PCA method was there classified as feature extraction method. They divided all dimension reduction methods into two classes. Feature extraction methods interprets the transformation of input high-dimensional space into a lower dimension, while feature selection methods choose specific, most influential input variables and neglecting others.

To identify the most influential input variables, there are plenty of methods. In case of mapping (how variations in all input variables vary with model output), one should focus on Sensitivity Analysis (SA) [12]. It is used for feature selection as a wrapper method for filtering input variables/features in AI [13].

Another field of use of SA is the mathematical modeling of real-world systems. For this purpose, it is an attempt to break out of the course of dimensionality. One such problem is the reservoir operating rule [14], [15], [16] where an adaptive surrogate model and SA are used to improve the performance of multi-objective optimization. For the adaptive surrogate model they chose the multivariate adaptive regression splines (MARS) and for SA they used Sobol's method. The work of [17] is more closely devoted to hydrological modeling.

Water models were the subject of the research in [18]. They paid attention to the total maximum daily load in the watershed model. Through Hydrologic Simulation Program-FORTRAN (HSPF) model and implementation of SA, it was determined, the water temperature is the most influential input variable which controls the overall dissolved oxygen concentration. The statistical properties of dissolved oxygen data were obtained using the HSPF model which was compared to Rosenblueth's method [18].

The aim of this article is to show that common methods for dimension reduction could have difficulties with solving very-high dimensional space problems. We propose an example of hyper-large problem (Economic Load Dispatch problem - ELD) [1] considering 140 input variables. In this case, we use

global variance-based SA method to find out the global sensitivity indices [19]. It reveals which input variables have the least effect on the resulting variance. On this problem we will show how these methods fail to select the proper amount of variables.

This work contributes to:

- drawing attention to the risks of dimensionality reduction
- opening research into modeling of very high-dimensional problems

The work is structured in the following way. In the first section, we present the ELD problem. In the second section, we perform the SA using the UptimAI software. According to the result of the analysis, we make feature selection on the ELD problem. Then we look closer at the feature-selected model, e.g. position and the shape of PDF and basic statistics, such as mean, median and mode. We perform the same process for the full-dimensional model and compare the results from the feature-selected model. In the third section, we verify the results using the Monte Carlo simulation. In the last section, we discuss the conclusion of the raised issue.

2 ELD problem

A large number of neglected input variables can cause the "many times nothing killed a horse" phenomenon. To show the influence of neglected variables, we choose the high dimensional static Economic Load Dispatch (ELD) problem.

The task is to optimize an amount of electricity produced by individual power generators which are all connected to one network [1]. The amount of energy produced must meet the demand requirements. Hence, we want to determine the optimal production capacity of individual units using the amount of energy required. The main goal is to minimize the production and operational costs while meeting the constraints of unit production capacity and energy demand requirements.

There is a quadratic dependence between the size of the generated electricity (P_i) by i^{th} generation unit and the price of production and operating costs ($f_i(P_i)$). Production and operating costs are defined as

$$f_i(P_i) = a_i P_i^2 + b_i P_i + c_i, \quad i = 1,2,3,...,N \tag{1}$$

where a_i, b_i, c_i are cost coefficients found in [1]. The network contains N generation units. Hence, the total operating costs are

$$F = \sum_{i=1}^{N} f_i(P_i), \tag{2}$$

and the demand requirement is an equality constrain noted as follows:

$$P_{load} = \sum_{i=1}^{N} P_i. \tag{3}$$

We limit the production capacities of an individual unit by using lower and higher bounds. The lower and upper bound of i^{th} power generation unit is P_i^{min}, P_i^{max}, i.e.

$$P_i^{min} \leq P_i \leq P_i^{max}. \tag{4}$$

However, the power units draw energy from renewable sources. Hence, they are dependent on random factors, such as wind, water or sunlight. The magnitude of the generated energy of the i^{th} generator is a random variable. In such a case, we transfer the task of minimizing operating costs to a uncertainty propagation problem. Hence, we estimate the probability distribution of the price of operating costs and obtain standard statistics, such as mean, median and mode. PDFs of the real generated volumes of an electric energy of individual power unit corresponding to time period t are described in Table 1. We assume that each dimension corresponds to an independent random input variable.

Variable name	Probabability distribution	Statistical parameters
x1	uniform	min: 71.0 max: 119.0
x2	uniform	min: 120.0 max: 189.0
x3	uniform	min: 125.0 max: 190.0
...	...	...
x81	**normal**	**mean: 200.0** **std: 542.0**
...	...	...
x138	uniform	min: 7.0 max: 19.0
x139	uniform	min: 7.0 max: 19.0
x140	uniform	min: 26.0 max: 40.0

Table 1: Input uncertain variables.

Note that variable **x81** was modified by replacing the uniform distribution with a normal one.

3 Uncertainty Quantification of ELD Problem

A standard way is to reduce the dimensionality of the problem. For this purpose we use the SA method such as Sobol indices [19]. Using UptimAI's SA, we filter input variables with minimal influence on the output. We plot the PDF of the feature-selected model. Also, we create model considering all the low-influential variables – the full-dimensional model. Comparing the shape of the PDFs and their basic statistics, we show a significant impact of the combination of neglected input variables on the final statistics.

3.1 Sensitivity Analysis

The field of sensitivity analysis is extensive and includes several different methods. Our goal is to quantify the influence of individual input variables on the variance of the output. For this purpose, we will use the variance-based method – Sobol indices [19]. Decomposition of variance expression of dimension d follows

$$\mathrm{Var}(Y) = \sum_{i=1}^{d} V_i + \sum_{i<j}^{d} V_{ij} + \cdots + V_{12\ldots d} \tag{5}$$

where the first order variances are defined as

$$V_i = \mathrm{Var}_{X_i}\left(E_{\mathbf{X}_{\sim i}}(Y \mid X_i)\right) \tag{6}$$

and the second order variances are defined as

$$V_{ij} = \mathrm{Var}_{X_{ij}}\left(E_{\mathbf{X}_{\sim ij}}(Y \mid X_i, X_j)\right) - \mathrm{V}_i - \mathrm{V}_j \tag{7}$$

and so on. X_i refers to uncertain input variable i and $X_{\sim i}$ indicates the set of all variables except X_i. The overall contribution of input variable i to the resulting variance is called global sensitivity index defined as

$$S_i = \frac{V_i}{\mathrm{Var}(Y)} \tag{8}$$

where $Var(Y)$ is defined in equation 5.

UptimAI's sensitivity analysis [20] is performed using the Monte Carlo simulation directly to the full-dimensional model. This process is done independently for each input variable to examine variance sensitivity. The SA results of the ELD example are listed in Table 2. The most influential variables are shown.

Variable name	Sensitivity variance	Variable name	Sensitivity variance
x72	0.45%	x73	0.83%
x74	0.84%	x75	0.87%
...	...	...	...
x80	0.84%	**x81**	**33.29%**
...	...	...	...
x102	2.57%	x103	2.54%
x104	3.37%	x105	3.47%

Table 2: Global sensitivity indices

The results of the most influential variables are shown in Figure 1. We choose the filter-based SA method with the threshold set to 5%. The threshold is highlighted in the figure as a red line. The global sensitivity index of the $x81$ variable is the only one over the threshold.

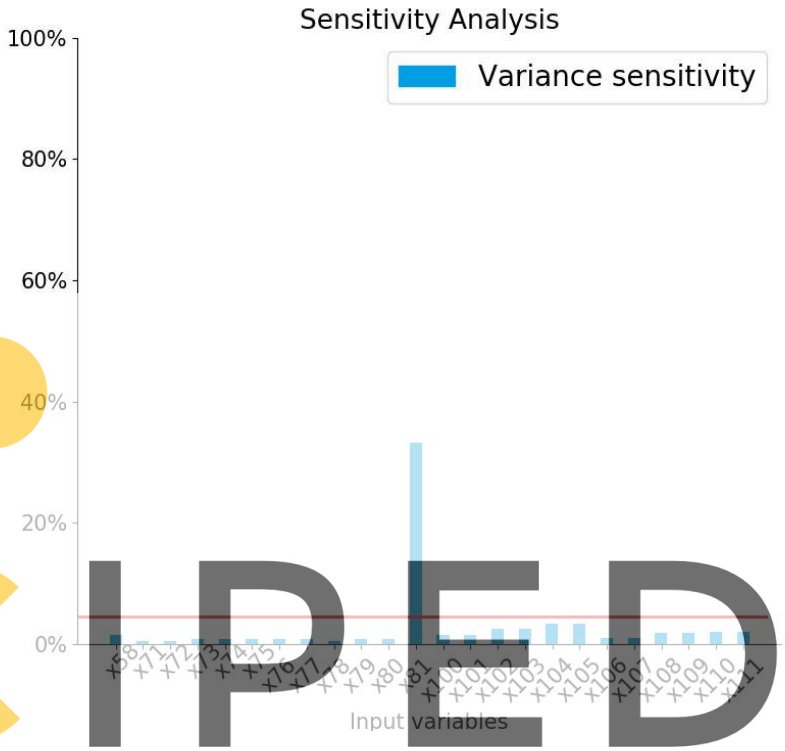

Figure 1: Global Sensitivity Indices

3.2 Feature-selected model

Using the results from SA, we select the most influential variables (over the threshold). In our case, only variable $x81$ is selected to represent the final statistics. All the other non-influential input variables are fixed to nominal values. Hence, model response $F(P_k, P_l)$ is dependent on the set of uncertain input variables $(P_k)_{k=1,\ldots,K}$ and a set of nominal inputs variables $(P_l)_{l=1,\ldots,L}$.

Note that we assign individual input variable to P_k if its global sensitivity index is over a 5% threshold. Otherwise, the variable is assigned to the P_l set. For the P_k set, we use probability distributions defined in Table 1. For the P_l set, the variable is fixed to its nominal value.

For the presented ELD problem we excluded all input variables, except for variable $x81$. The final PDF is shown in Figure 2. According to the shape, it is obviously not a normal distribution.

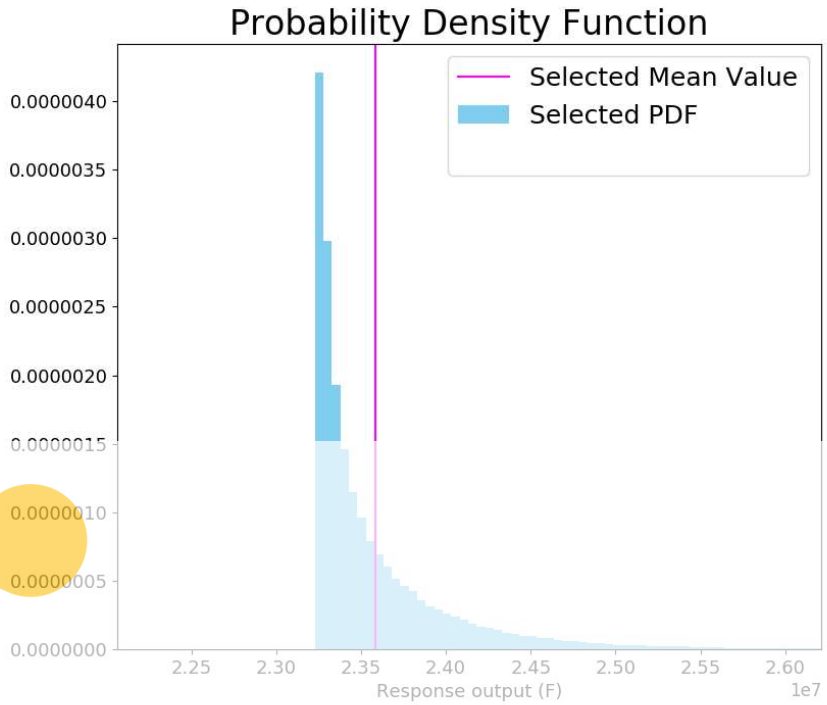

Figure 2: PDF of the feature-selected model

3.3 Full Dimensional Model

For the full-dimensional model, we assume the following model: $F(P_i)$ is considering all input variables, i.e. $i = 1, \ldots, N$. All variables acknowledge PDFs defined in Table 1. Then model response F has the following probability density function.

Based on the Figure 3, it is clearly visible that the response output of the ELD problem is normally distributed with the mean value around 24 millions.

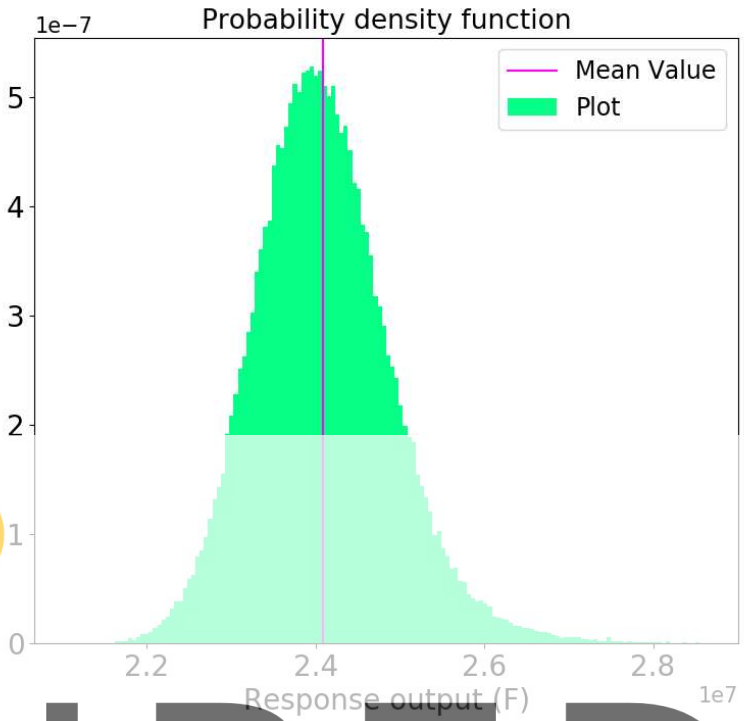

Figure 3: PDF of the full-dimensional model

3.4 Model comparison

Let us now consider a comparison of the previously shown figures. Obviously, the PDF of the feature-selected model (Figure 2) differs significantly from the PDF of the full-dimensional model (Figure 3). Two main differences can be seen, namely the shape and the position of the resulting PDFs (Figure 4).

To compare the position of PDFs, we use three statistics: mean, median and mode. Table 3 shows a comparison for both models.

Statistics	Full-dimensional model	Feature-selected model
mean	24 137 492.12	23 588 020.23
median	24 085 838.18	23 405 351.93
mode	24 017 408.25	23 262 748.57

Table 3: Compared statistics

We performed two normality tests (Shapiro-Wilkinson [21] and D'Agostino's K-squared [22]) for both models. The p-values for both tests are shown in Table 4.

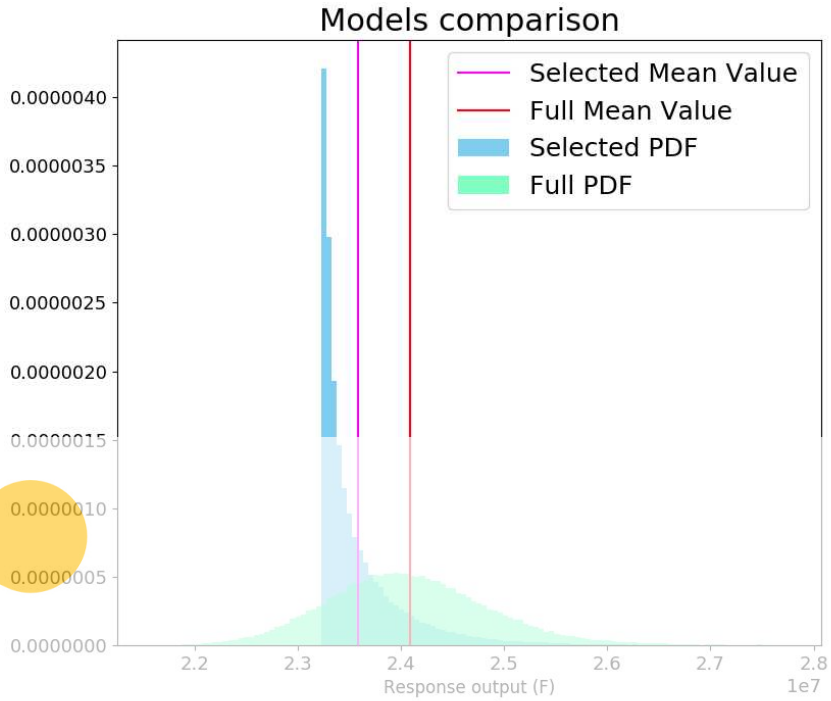

Figure 4: Comparison of PDFs

Statistical hypotesis test	p-value	
	Full-dimensional model	Feature-selected model
Shapiro-Wilkinson	0.40	0.00
D'Agostino's K-squared	0.41	0.00

Table 4: Comparison of p-values

4 Monte Carlo confirmation

To confirm the results obtained by the UptimAI's algorithm, we generated the PDFs of both models, i.e. the feature-selected model and the full-dimensional model. To these models the Monte Carlo simulation was applied using 100000 samples. The results are shown in Figure 5 where the phenomenon *"many times nothing killed a horse"* can be observed. The model with the neglected input variables is significantly different from the full-dimensional model. Hence, the commonly used method for dimension reduction in the field of SA, such as filtering on variance-based global sensitivity indices, has failed in this case. Not considering the low-influential input variables led to low accuracy of the final model. The accurate model was obtained with UptimAI's unique algorithm.

We confirmed both errors using statistical properties such as mean, median and mode. In the ELD

example, the dimension reduction method caused an error of the 10^5 order.

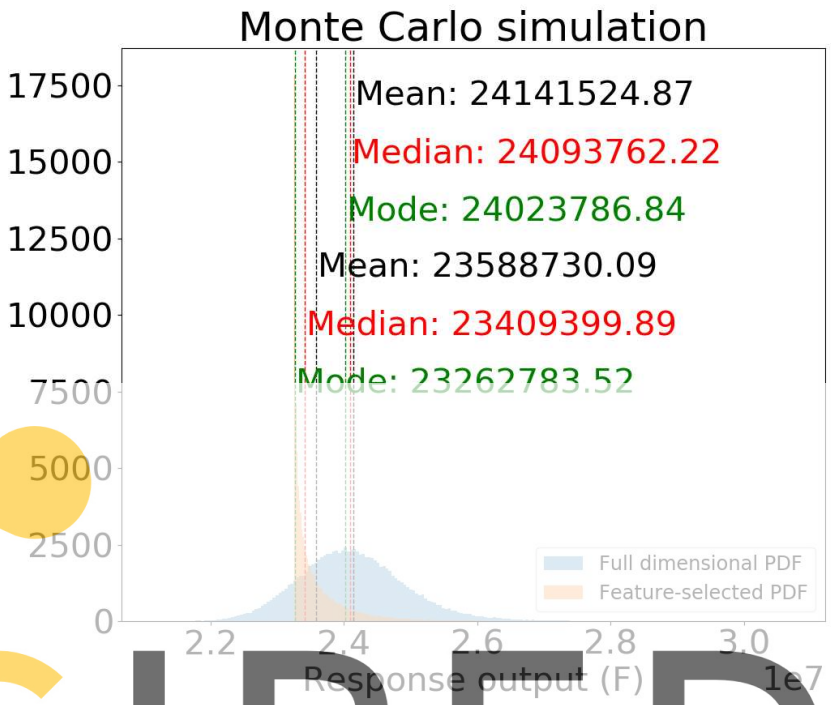

Figure 5: Monte Carlo simulations

Additionaly the shapes of the PDFs are clearly different, i.e. the full-dimensional model has a normal distribution, while feature-selected model has not.

5 Conclusion

Using a standard dimension reduction methods can be very risky. On the example of the ELD problem, we showed that uncertainty propagation considering dimensional reduction can lead to a significant error. This was shown on two models:

- feature-selected model (applied sensitivity analysis based reduction)
- full-dimensional model

where uncertain distributions were propagated and results were compared.

The statistical results of these two models proved that dimension reduction cannot be applied so simply. However, more research in this topic is needed to confirm our findings. Our future research will focus on other dimension reduction methods such as PCA or Isomap, to approve our results.

6 Acknowledgment

We would like to thank the UptimAI's team and the UptimAI UQ algorithm for showing this problem. Also, it was very pleasant to work with UptimAI software, which is focused on efficient propagation uncertainty and provides insight into this process. It helps to uncover unthinkable findings of the statistical nature of solved problems.

This paper has been prepared with the support of the UptimAI company and the department of applied mathematics at Brno University of Technology.

REFERENCES

[1] DAS, S. and Suganthan, P.N. *Problem Definitions and Evaluation Criteria for CEC 2011 Competition on Testing Evolutionary Algorithms on Real World Optimization Problems*. Technical report, Dec. 2010.

[2] BELLMAN, R., Rand Corporation and Karreman Mathematics Research Collection. 1957. *Dynamic Programming*. Princeton University Press.

[3] Explained: Linear and nonlinear systems. 2010. *MIT - Massachusetts Institute of Technology* [online]. MIT News Office. Available at: https://news.mit.edu/2010/explained-linear-0226

[4] CROXTON, F.E., D.J. COWDEN and S. KLEIN. 1968. *Applied General Statistics*. Pitman.

[5] VAN DER MAATEN, Laurens, Eric POSTMA and H. HERIK. 2007. Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research - JMLR*. **2007**(01), 10.

[6] PEARSON, Karl. 2010. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. **2**(11), 559-572.

[7] SCHÖLKOPF, Bernhard, Alexander SMOLA and Klaus-Robert MÜLLER. 1998. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*. **10**(5), 1299-1319.

[8] TENENBAUM, J B, V DE SILVA and J C LANGFORD. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)* [online]. **290**(5500), 2319.

[9] DE SILVA, V. and J.B. TENENBAUM. 2003. Global versus local methods in nonlinear dimensionality reduction. In: *Advances in Neural Information Processing Systems* [online]. Neural information processing systems foundation. Available at: books.nips.cc

[10] SAMMON, J.W., Alexander SMOLA and Klaus-Robert MÜLLER. 1969. A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*. **C-18**(5), 401-409.

[11] PUDIL, P., J. HOVOVICOVA and Klaus-Robert MÜLLER. 1998. Novel methods for subset selection with respect to problem knowledge. *IEEE Intelligent Systems*. **13**(2), 66-74.

[12] SALTELLI, Andrea, J. HOVOVICOVA and Klaus-Robert MÜLLER. 2002. Sensitivity Analysis for Importance Assessment. *Risk Analysis*. **22**(3), 579-590.

[13] SÁNCHEZ-MAROÑO, N. and A. ALONSO-BETANZOS. 2007. Feature selection based on sensitivity analysis. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* [online]. 4788, p. 239-248.

[14] CHU, J., C. ZHANG, G. FU, Y. LI and H. ZHOU. 2015. Improving multi-objective reservoir op-

eration optimization with sensitivity-informed dimension reduction. *Hydrology and Earth System Sciences*. **19**(8), 3557-3570.

[15] LI, J., Q. Y. DUAN, W. GONG, et al. 2013. Assessing parameter importance of the Common Land Model based on qualitative and quantitative sensitivity analysis. *Hydrology and Earth System Sciences*. **17**(8), 3279-3293.

[16] ZHANG, Jingwen, Xu WANG, Pan LIU, et al. 2017. Assessing the weighted multi-objective adaptive surrogate model optimization to derive large-scale reservoir operating rules with sensitivity analysis. *Journal of Hydrology*. **544**(8), 613-627.

[17] GAN, Yanjun, Qingyun DUAN, Wei GONG, et al. 2014. A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model. *Journal of Hydrology*. **51**(8), 269-285.

[18] PATIL, Abhijit, Zhi-Qiang DENG, Wei GONG, et al. 2010. Analysis of uncertainty propagation through model parameters and structure: A case study with a hydrological model. *Water Science and Technology*. **62**(6), 1230-1239.

[19] SOBOL, I.M, Zhi-Qiang DENG, Wei GONG, et al. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates: A case study with a hydrological model. *Mathematics and Computers in Simulation*. **55**(1-3), 271-280.

[20] KUBICEK, Martin. 2020. *UptimAI software - help*. Praha.

[21] SHAPIRO, S. S. and M. B. WILK. 1965. An analysis of variance test for normality (complete samples). *Biometrika*. **52**(3-4), 591-611.

[22] D'AGOSTINO, RALPH B. and M. B. WILK. 1970. Transformation to normality of the null distribution of g 1. *Biometrika*. **57**(3), 679-681.