iomac
International Operational Modal Analysis Conference

11th International Operational Modal Analysis Conference

20 - 23 May 2025 | Rennes, France

Robust data-driven health monitoring of full-scale concrete dam structures via stochastic ML schemes

Simona Bogoevska[1], *and Eleni Chatzi*[2]

[1] Faculty of Civil Engineering, University of Ss. Cyril and Methodius, Skopje, North Macedonia, simona.bogoevska@gf.ukim.edu.mk
[2] Department of Civil, Environmental and Geomatic Engineering, ETH Zurich, Switzerland, chatzi@ibk.baug.ethz.ch

ABSTRACT

Dams form critical components of civil infrastructure, as they play an essential role in flood control, hydroelectricity generation and water supply. The reliable and uninterrupted operation of dams throughout their designed lifespan bears regional and national impact. However, as any other infrastructure, dams are similarly vulnerable to ageing-related material degradation, extreme climatic events and overloading scenarios. In this context, for dam owners, operators and policy architects adequate maintenance and prediction of their structural condition become vital.

Data-driven Structural Health Monitoring (SHM) schemes are becoming particularly valuable for infrastructures that bear critical importance for modern societies. Historically, dam maintenance strategies have been founded on regular scheduled visual inspections. As the sensor technology and software-hardware interface advance, the opportunities for minimizing the shortcomings of the traditional approaches (e.g. manpower demand, inaccessibility, subjectivity and difficulty to predict) are quickly expanding. By exploiting sensory data from an instrumented dam structure, in addition to real-time structural diagnosis, Dam Health Monitoring (DHM) allows for development of a reliable prediction model.

This work focuses on testing the performance of a data-driven prognostic model based on the Polynomial Chaos Expansion (PCE) uncertainty quantification method. The model is applied on an experimental benchmark from a double curvature arch dam, located in the south of France and data collected from an instrumented double curvature arch concrete dam located in North Macedonia. Data handling, preparation of inputs, as well as specifics and characteristics of the developed models will be discussed. The presented results reveal the potential for robust real-time DHM and further integration into a multi-targeted diagnostic tool for operational infrastructure.

Keywords: Dam Health Monitoring, Prognostics, Data driven, Real time, Full scale, Robust

1. INTRODUCTION

The advancing capabilities of new sensor technologies and software-hardware solutions have made it possible to gather large amounts of data, which has, in turn, facilitated the integration of soft computing methods (machine learning and deep learning tools) into data-driven structural health monitoring approaches [1]. Data-driven SHM systems can play a key role in improving the reliability and longevity of civil infrastructure, making them increasingly essential for structures of great societal importance. Civil engineering structures, particularly during operation, can exhibit complex and unpredictable behavior due to their constant exposure to the surrounding environment. In this context, data-driven SHM schemes must extend beyond structural damage detection, offering early warning alarms capable of identifying deviations from established behavioral patterns typical of a specific structure-environment cohabitation (interaction) [2].

The PCE method enables the development of data-driven models that approximate an input-output relationship, offering a relatively simple polynomial mathematical formulation and affordable computational time [3]. Another key advantage that makes the PCE model ideal for practical applications is its ability to provide an effective mathematical framework for analyzing the sensitivity of modeled system outputs (i.e. the measured response of a structure) to multiple random input variables (i.e. reflecting environmental and operational factors) affecting the system. Several studies have investigated the use of PCE for meta-modeling in civil engineering applications, addressing various numerically-based problems [4-7]. The resulting proxy models showed high accuracy, efficiency in predictions and reduced computational burden.

However, in the context of real-time data-driven SHM schemes, a major challenge in developing a robust PCE-based tool is meeting the theoretical requirement of the PCE model, which necessitates the use of uncorrelated input variables [8]. Within such a purely data-driven ML framework, the application of the PCE method was explored in [9,10]. Spiridonakos and Chatzi [9] introduced the PCE-Independent Component Analysis (PCE-ICA) tool, successfully validating it on damage detection for the benchmark SHM project of the Z24 bridge. In [10], the authors combined the PCE-ICA tool with the parametric smoothness priors time-varying autoregressive moving average (SP-TARMA) method to track the long-term behavior of operational wind turbine structures. Building on this previous research, the current paper explores the robustness and applicability of the PCE method as a data-driven ML prognostic tool for the purpose of DHM. Dams, as vital components of public infrastructure, are equally susceptible to aging-related material degradation, extreme climatic events, and overloading. Given their critical importance, they require reliable and continuous operation throughout their designed lifespan. Leveraging collected multivariate sensor data for real-time structural diagnostics can greatly reduce the established limitations of traditional periodic inspections [11]. In a preliminary investigation [12], Georgijev and Bogoevska explored straightforward employment of uncorrelated input set of measured environmental data for the first year of operation of a concrete arch dam, demonstrating the effect of incomplete training set on the accuracy of the PCE model prediction.

In the work presented herein, the PCE is successfully applied as a multi-output modelling approach on two real-world concrete dam structures in a long-term framework, on data collected throughout 10 and 13 years of monitoring. The case studies are an experimental benchmark from a double curvature arch dam, located in the south of France and an instrumented double curvature arch dam located near Skopje, North Macedonia. Various multivariate input configurations for model generation and their effects on the prediction accuracy are compared, showcasing the method's potential for efficient use in prognostic and diagnostic tasks within autonomous DHM frameworks.

2. APPLIED METHODS

2.1. Polynomial Chaos Expansion

PCE is an effective metamodeling approach that seeks to approximate an input-output relationship through a spectral representation on a suitably built basis of polynomial functions. For a given system S, the PCE uncertainty quantification method generates a mathematical expansion of the model's

random output variable Y using multivariate polynomial chaos basis functions, which are appropriately linked to the model's random input data vector X [3,8].

Specifically, if we consider the system $Y = S(X)$, where M random input parameters are represented by independent random variables (e.g., water levels and temperature values), collected in the random vector X with a prescribed joint Probability Density Function (PDF) f_X, and the output variable has finite variance, the PCE model takes the following form [10]:

$$Y = S(X) = \sum_{\alpha \in N^M} y_\alpha \psi_\alpha(X) \tag{1}$$

where $\psi_\alpha(X)$ are polynomials orthonormal with respect to f_X, $\boldsymbol{\alpha} \in N^M$ is a vector of multi-indices identifying the components of the polynomials ψ_α and $y_\alpha \in R$ are the corresponding unknown deterministic coefficients. The multivariable polynomials $\psi_\alpha(X)$ are obtained through the tensor product of corresponding one-dimensional orthonormal polynomials, selected based on the PDF of the random input variables and the known Askey scheme for orthonormal polynomials [10]. In practical applications, the sum in Eq. (1) is truncated to a finite sum, usually by restricting the total maximum degree p of the polynomials in the polynomial basis, which ensures that the total number of terms in the polynomial basis remains:

$$P = \frac{(M+p)!}{M!\,p!} \tag{2}$$

where *M* designates the number of random variables and *p* denotes maximum basis degree.

Finally, truncating the PCE model to the first P terms results in a finite parameter vector $\boldsymbol{y_\alpha}$ which can be estimated by solving Eq. (1) in a least squares sense. Furthermore, by exploiting the orthonormality of the PC basis and their consequent suitable properties, a global sensitivity analysis can be performed based on Sobol' decomposition [8]. Sobol' indices are calculated as the sum of squared PC coefficients and represent portion of the total variance *D* of the model output that can be attributed to each input variable or combinations of variables X_i and X_j (referred to as higher-order Sobol' indices $S_{ij}, i \neq j$). More precisely, the index for a single input variable is called the first-order Sobol' index and represents the effect of X_i alone:

$$S_i = \sum_{\alpha \in A_i} y_\alpha^2 / D, \quad A_i = \{\alpha \in A: \alpha_i > 0, \alpha_{j \neq i} = 0\} \tag{3}$$

3. APPLICATION CASE STUDIES

3.1. Case study I: Saint Petka concrete arch dam

The first presented case study herein is a concrete arch dam, located 30 km southwest of Skopje, North Macedonia. The structure represents a thin concrete shell with double curvature, with a structural height of 64 meters. The dam is mostly unreinforced, with the exception of the upper third, which should withstand increased vibrations during an earthquake. The thickness is 10 m at the bottom, gradually decreasing to 2.0 m at the crest. The structure was built in 2012, utilizing 27362 m^3 concrete for the dam body. Since its construction the structure is equipped with a comprehensive monitoring system measuring: reservoir water level, underground water levels, ambient temperature and temperatures of concrete and water, rainfall, displacements at the crest and dam body, strains, rotations, accelerations during triggered earthquake events, contact stresses, etc. The distribution of instruments is with a nearly symmetrical layout, at five different altitude levels. The system continuously collects data from total of 332 instruments, recording once in six hours. Clearly, extracting valuable information from such a large volume of data necessitates the use of robust multivariate data-driven models.

To test the applicability and reliability of the PCE method, two variants of multi-output PCE models were developed for this case study. The models were built using a total of 15 384 data samples, measured between years 2012 and 2021. The selected output variables for both models are crest displacements in two orthogonal directions, measured by hanging pendulum anchored at the dam crest. Displacements relative to the wire are collected via remote automatic x-y telecoordinometer. In order to robustly capture the effects of the Environmental and Operational Parameters (EOPs) influencing the dam behavior, the models were built on input sets consisting of 18 measured temperature variables (12 concrete, 3 water and 3 air) and measured reservoir water level. A schematic overview of the measurement location of employed input/output variables is presented in Figure 1. The first 3000 samples correspond to the first year, which includes the planned operational testing of the newly constructed dam in the following sequence: water reservoir filling-partial (half-height) emptying-filling-full emptying-filling. Following these operations, the water level remains relatively stable for the remainder of the analyzed period. The PCE models were generated with the application of the UQLab toolbox [13] by setting a training length of three consecutive years (35% of the data) and validation period of 65% of the full collected database. Given the annual cyclic behavior of the 18 EOPs, the stationarity of the reservoir water level and the temperature-correlated displacements over the 10 years of monitoring, the relatively short training length is shown to be sufficient.

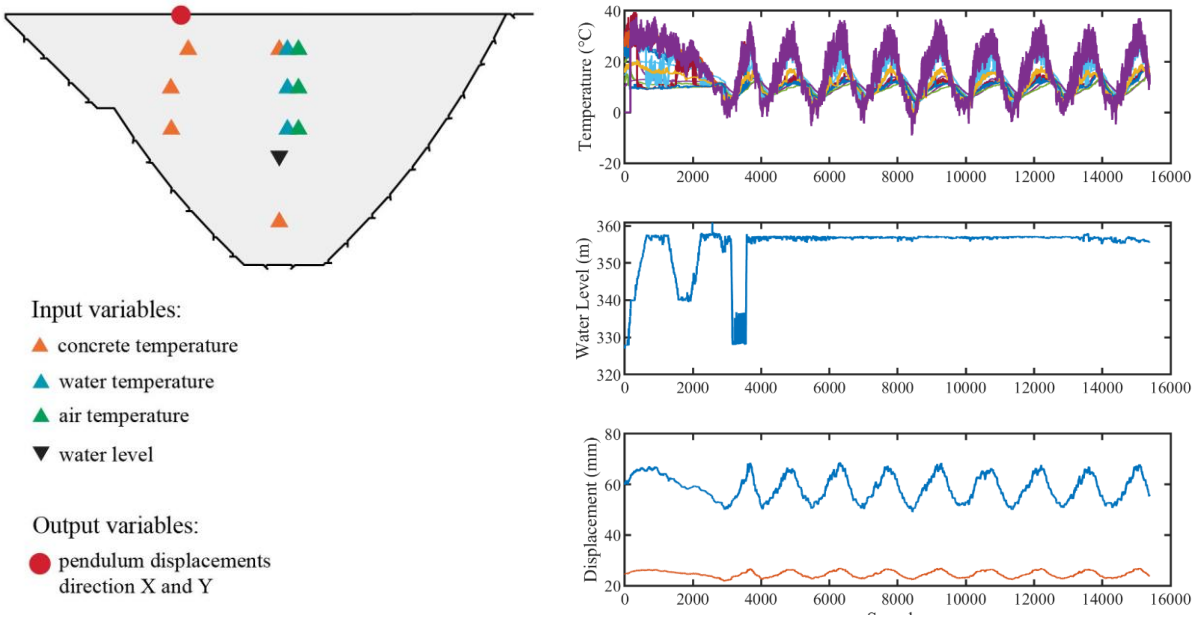

Figure 1. Schematic overview of employed sensors and time history plots of PCE inputs/outputs.

In order to setup a robust strategy for real-time treatment of highly correlated input data, for the 18 measured sets of correlated temperature data, the two estimated PCE model variants explore distinct modeling strategies for the input set, specifically: (a) Straightforward (unrefined) utilization of all 18 temperature sets; (b) Reduced input set based on Principal Component Analysis (PCA), where temperature data is represented by only 4 PCA latent variables. The first strategy builds the PCE polynomial basis orthonormal with respect to the original data-driven input marginals, thus ignoring the multivariate input correlation. Such a concept was demonstrated to produce more accurate pointwise predictions and proved to be more effective in [14]. The second variant deals with the correlation by first estimating PCA latent variables. In order not to lose the complete physicality of the problem, the PCA is calculated on the 18 temperature sets, keeping the water level variable intact, as fifth PCE input variable. The estimated Pearson correlation matrices for both variants are demonstrated in Figure 2. Initially, the quite good performance of the first PCE model variant, for both orthogonal crest displacements, is presented in Figure 3. The vertical magenta line separates the timeline into training (left) and validation period (right). A comparison of the two variants, shown here for brevity only for crest displacement along the X direction, is presented in Figure 4. For the estimation set both models produce the same LOO error of 0.022. However, for the prediction part the PCA-based model outperforms the full database variant.

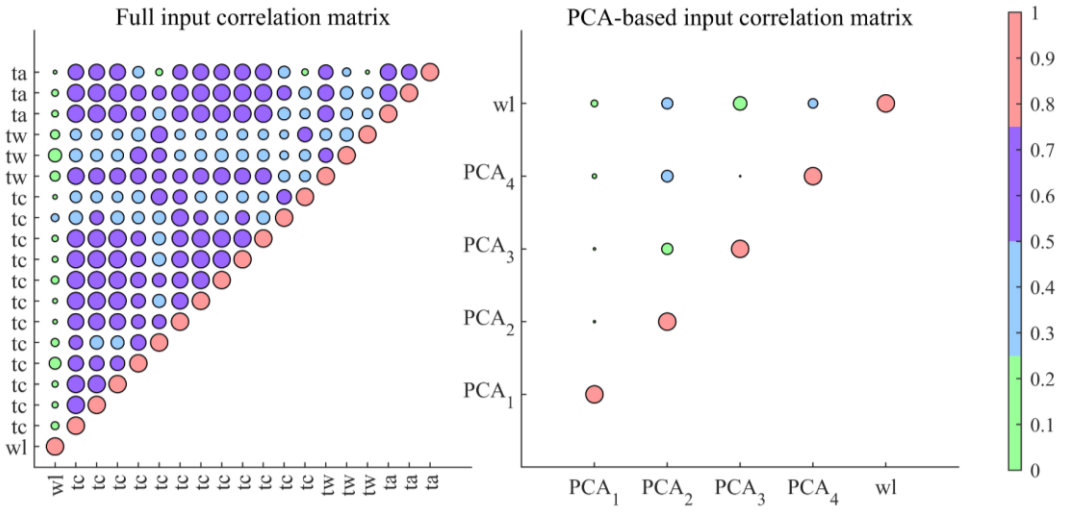

Figure 2. Comparison of Pearson correlation matrices for the two model variants: (Left) Full input set with 19 variables: water level (wl), concrete temperature (tc), water temperature (tw) and air temperature(ta); (Right) PCA-based input set with 5 variables: PCA-based latent variables and water level (wl).

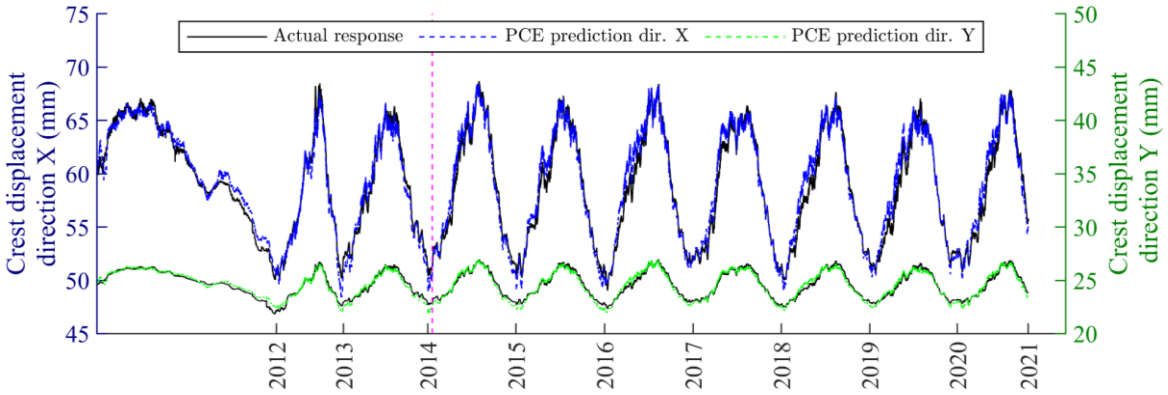

Figure 3. PCE model performance for full data set of 10 years of monitoring (training period equals 5538 model evaluations, 19 input variables, maximal degree 1, size of basis 20, method OLS).

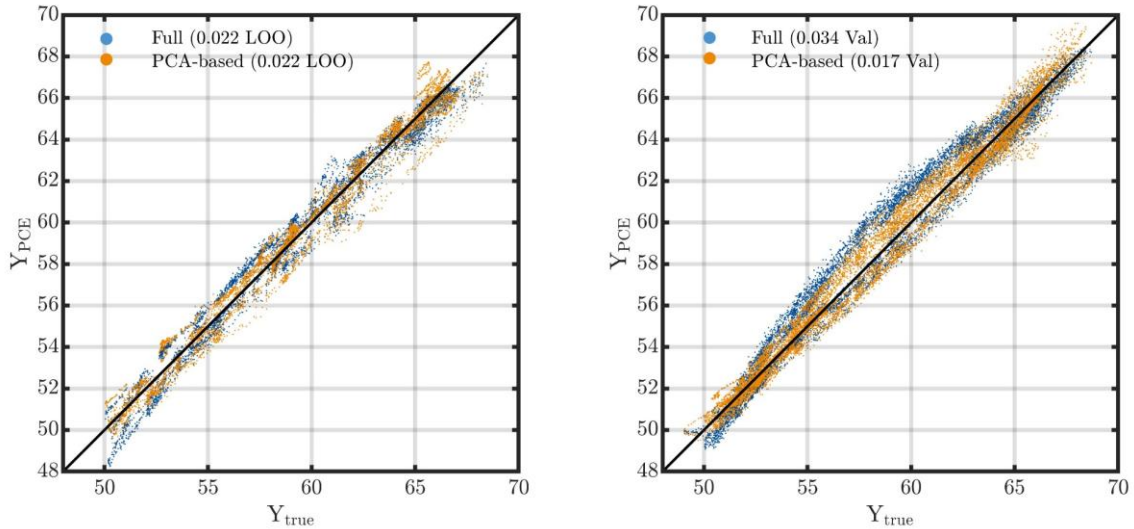

Figure 4. Comparison of PCE model performance for crest displacement in direction X for the two different variants of input sets: (Left) Estimation set; (Right) Validation set.

3.2. Case study II: Benchmark concrete arch dam

The second case study is a concrete dam located in the south of France, provided as a benchmark study in [15]. The dam was constructed between 1957 and 1960. It is a double curvature arch dam, which is asymmetric due to the terrain disposition. Dam height above foundation is 45m. The thickness of the crest is 2 m, gradually increasing to 6 m at the bottom. The crest length is 166 m. The dam is equipped with a comprehensive continuous monitoring system, including pendulums, crack opening displacement sensors, piezometers and seepage measurements. The air temperature is not measured at the location of the dam, instead it is calculated by interpolation from several air temperature measuring stations taking into account the altitude of the dam. It is calculated on a mesh of 1 square kilometer. Daily cumulative precipitation time series from a rain gauge located about 5 km from the dam is provided as well in mm. The measurements are automatically verified with a delay of 48 hours after acquisition, disregarding data stemming from sensor malfunctions.

A multi-output PCE model was built with total of 680 data samples, measured in the period 2000 – 2012. A schematic overview of the measurement location of employed input/output variables is presented in Figure 5. The selected output variables are radial crest and foundation displacements, measured by two hanging pendulums anchored at the crest and foundation height at the central block of the dam body, where an increasing radial displacement indicates a movement of the point in the downstream direction. The PCE model was built on weakly correlated input set consisting of the 3 measured variables (air temperature, water level and rainfall). The PCE model was generated with the application of the UQLab toolbox [13] by setting a training length of 75% of the data and validation period 25% of the full benchmark database. The longer training length is due to several key differences compared to case study I. Specifically, the measured reservoir water level exhibits greater variability, and the temperature data is not as consistently periodic as in the first test case.

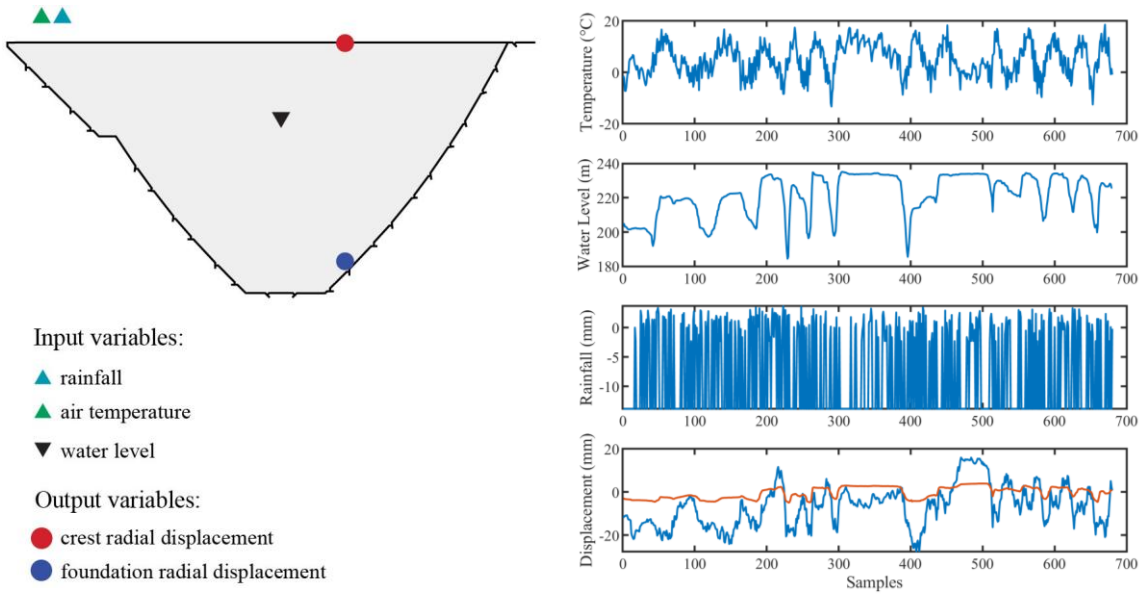

Figure 5. Schematic overview of employed sensors and time history plots of PCE inputs/outputs.

The performance of the PCE model, for both crest and foundation displacements, is presented in Figure 6. The vertical magenta line separates the timeline into training (left) and validation period (right). The LOO error is estimated as 0.095, and the validation error equals 0.28 for the crest displacement, and 0.042 and 0.35 for the foundation displacement, accordingly.

3.3. Dimensionality reduction

When it comes to utilization of the models within a real-time monitoring framework, the possibility for dimensionality reduction becomes vital. The models' performances with reduced input sets, based on the obtained sensitivity of outputs via Sobol' indices analysis (Figure 7), is presented with the control charts in Figure 8. Both models yield a residual within predefined ±3std threshold values.

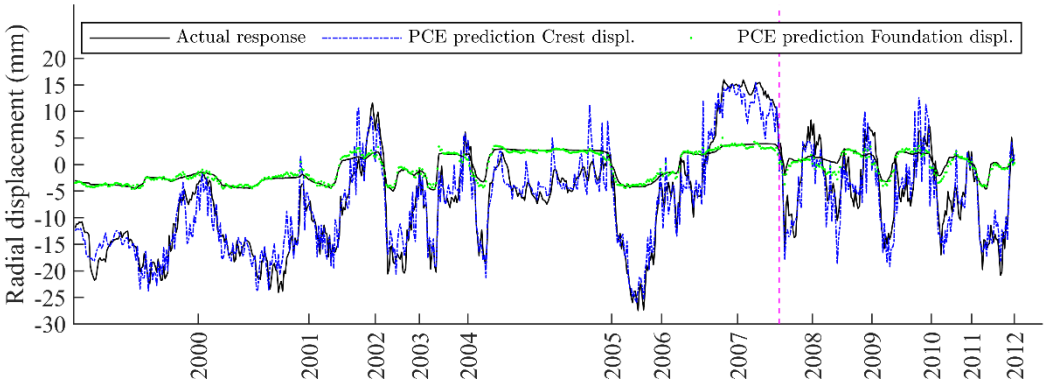

Figure 6. PCE model performance for full data set of 13 years of monitoring (training period equals 510 model evaluations, 3 input variables, method LARS), for crest displacement PCE maximal degree 6, size of basis 84; for foundation displacement PCE maximal degree 3, size of basis 20.

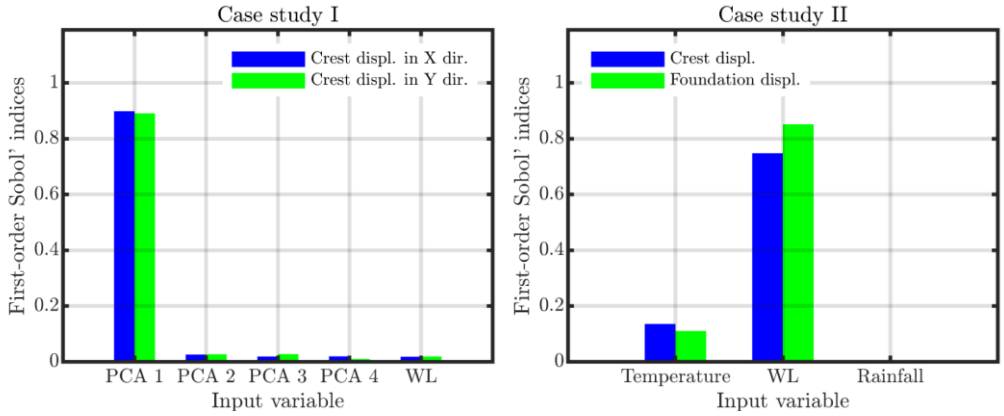

Figure 7. PCE-based Sobol' indices estimated for case study I and case study II.

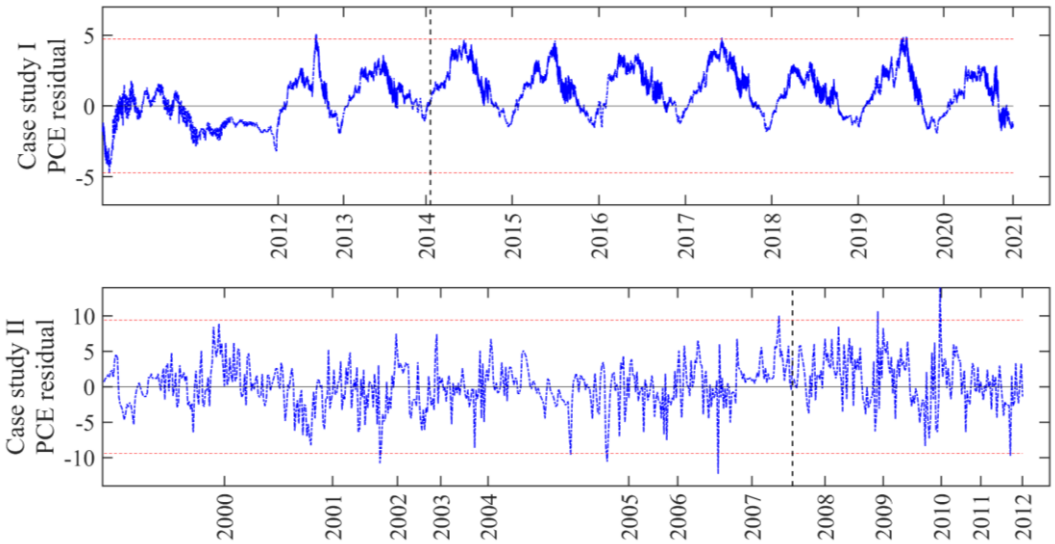

Figure 8. Control charts of obtained PCE residuals with Sobol-based reduced input sets (for case study I remaining input variable PCA1, for case study II remaining input variables temperature and WL), vertical dashed lines separate training (left) from validation (right) period.

4. CONCLUSIONS

The presented research investigates the robustness of the PCE model for application within a long-term DHM framework. The model was successfully implemented as a ML approach for two distinct case studies of full-scale concrete arch dams, monitored for periods of 10 and 13 years, respectively.

The results highlight the prediction efficiency of the PCE model, particularly when it comes to extensive monitoring systems, due to its ability to handle both multivariate inputs and outputs. For both case studies, the model performance was validated under various input configurations, demonstrating its simplicity in construction and its convenient mathematical formulation for dimensionality reduction. Future research efforts are targeted towards comparing the performance of the PCE tool with ML schemes based on Convolutional Variational Autoencoders, as a state-of-the-art unsupervised generative machine learning tool, which further bears potential for uncertainty quantification.

ACKNOWLEDGEMENTS

The authors would like to gratefully acknowledge JSC ESM for generously providing access to the monitoring data of the St. Petka dam, used in this study. This data played a crucial role in the successful completion of the presented research.

REFERENCES

[1] M. Mishra, P.B. Lourenço and G.V. Ramana. Structural health monitoring of civil engineering structures by using the internet of things: A review. *Journal of Building Engineering*, 48, p.103954, 2022.

[2] A. Malekloo, E. Ozer, M. AlHamaydeh and M. Girolami. Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights. *Structural Health Monitoring*, 21(4), pp.1906-1955, 2022.

[3] O. Le Maître and O.M. Knio. *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics*. Springer Science & Business Media, 2010.

[4] M.D. Spiridonakos and E.N. Chatzi. Metamodeling of dynamic nonlinear structural systems through polynomial chaos NARX models. *Computers & Structures*, 157, pp.99-113, 2015.

[5] P. Ni, Y. Xia, J. Li and H. Hao. Using polynomial chaos expansion for uncertainty and sensitivity analysis of bridge structures. *Mechanical Systems and Signal Processing*, 119, pp.293-311, 2019.

[6] M.A. Hariri-Ardebili and B. Sudret. Polynomial chaos expansion for uncertainty quantification of dam engineering problems. *Engineering Structures*, 203, p.109631, 2020.

[7] L. Novák, H. Sharma and M.D. Shields. Physics-informed polynomial chaos expansions. *Journal of Computational Physics*, 506, p.112926, 2024.

[8] B. Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability engineering & system safety*, 93(7), pp.964-979, 2008.

[9] M. Spiridonakos and E. Chatzi. Polynomial chaos expansion models for SHM under environmental variability. *In Proceedings of the 9th International Conference on Structural Dynamics (EURODYN)*, Porto, Portugal, 2014.

[10] S. Bogoevska, M. Spiridonakos, E. Chatzi, E. D. Jovanoska and R. Höffer. A data-driven diagnostic framework for wind turbine structures: A holistic approach. *Sensors*, 17(4), p.720, 2017.

[11] F. Salazar, R. Morán, M.Á. Toledo and E. Oñate. Data-based models for the prediction of dam behaviour: a review and some methodological considerations. *Archives of computational methods in engineering*, 24, pp.1-21, 2017.

[12] V. Georgijev and S. Bogoevska. Application of the polynomial chaos expansion method for forecasting structural response of two full-scale case studies. *Journal of Building Materials and Structures*, 2024.

[13] UQLab: A Framework for Uncertainty Quantification in MATLAB, S. Marelli and B. Sudret, *In The 2nd International Conference on Vulnerability and Risk Analysis and Management (ICVRAM 2014)*, University of Liverpool, United Kingdom, July 13-16, 2014.

[14] E. Torre, S. Marelli, P. Embrechts and B. Sudret. Data-driven polynomial chaos expansion for machine learning regression. *Journal of Computational Physics*, 388, pp.601-623, 2019.

[15] https://www.unesco-floods.eu/education/