Investigation of Robustness in Detecting and Localizing Sensor Malfunctions in Deteriorating Structural Health Monitoring Systems

MARK HOYER, NIKLAS R. WINNEWISSER,
JAN-HAUKE BARTELS, THOMAS POTTHAST
and MICHAEL BEER

Mark Hoyer, Niklas R. Winnewisser, Thomas Potthast: Leibniz University Hannover, Institute for Risk and Reliability, Appelstr. 9A. 30167 Hannover, Germany
Jan-Hauke Bartels: LPI Ingenieurgesellschaft mbH, Volgerstr. 9, 30519 Hannover, Germany
Michael Beer: Leibniz University Hannover, Institute for Risk and Reliability, Appelstr. 9A, 30167 Hannover, Germany; Department of Civil and Environmental Engineering, University of Liverpool, Liverpool, L69 3GH, UK; International Joint Research Center for Engineering Reliability and Stochastic Mechanics and International Joint Research Center for Resilient Infrastructure, Tongji University, Shanghai, China

ABSTRACT

Structural Health Monitoring (SHM) systems are essential for damage detection and maintenance planning in aging infrastructure. However, sensor degradation increases epistemic uncertainty and leads to incorrect SHM assessments, making it necessary to develop robust methods for handling sensor faults. This work investigates the robustness of a recently proposed framework for detecting, localizing, and compensating faulty sensors. The approach uses the Mahalanobis distance between sensor data frames from a current and a reference period, relying on cross-validation among sensor outputs. It includes a so-called α-level mapping, which links evaluated distances to corresponding levels of uncertainty relying on common statistical thresholds, as well as an algorithmic mechanism for excluding faulty sensors based on individual trust scores. The robustness study, using artificial FEM data from a steel lattice mast, identifies three key parameters that primarily govern the correct classification of faulty and healthy sensors over time. Results from the case study show that the method remains effective under multiple sensor faults, and identifies a favorable parameter domain, yielding the optimal performance. These findings support practical parameter recommendations to maintain confidence in SHM systems during long-term operation.

Keywords: Structural Health Monitoring, Sensor Fault Detection, Epistemic Uncertainty, Robustness

INTRODUCTION

Today's advances in sensing have enabled ever more precise condition monitoring of both aging and newly constructed structures, such as bridges and tunnels. The growing scale and complexity of these engineering systems demand a robust Structural Health Monitoring (SHM) for reliable damage detection and predictive maintenance. SHM is not just a complementary tool but essential for reducing costs, extending service life, and most importantly, protecting human lives [1]. An SHM system depends on a variety of sensor measurements that capture the structure's dynamic responses, such as displacement, acceleration, strain etc. However, these measurements are subject to uncertainties arising from the natural variability of structural properties, environmental influences, and sensor degradation over the decades-long lifespans for which such structures are de-

signed [2], [3], [4]. Neglecting these uncertainties can lead to misleading conclusions about structural integrity, compromising safety and increasing life-cycle costs [3].

Inherent uncertainties are commonly classified into aleatory uncertainty, coming from inherent randomness or variability, and epistemic uncertainty, which arises from incomplete knowledge. The latter is reducible through enhanced data quality or improved modeling [4], [5], [6]. Aleatory uncertainty includes acceptable sensor noise, while epistemic uncertainty grows over time as sensors degrade, leading to measurement drift, bias, or gain errors. To maintain reliable SHM, it is essential to quantify both uncertainty types. Addressing sensor faults requires a four-stage fault-diagnosis process [7]: detection, localization, classification, and compensation.

Winnewisser et al. [6] present a novel, self-adaptive SHM framework that (i) quantifies epistemic uncertainty via α-levels derived from probabilistic consistency measures, (ii) identifies and localizes sensor malfunction using cross-validation across combinatorial subsets of sensor measurements, and (iii) compensates for faulty sensors by excluding those with low trust, thus preserving high-confidence measurement data. The Mahalanobis distance serves as the core consistency metric, comparing real-time data frames against a reference frame. The resulting δ-evidence is mapped to α-level evidence, from which expected epistemic uncertainty (α-confidence) and sensor trust are derived based on the expected value of the α-level distribution.

To assess the robustness of the proposed framework, it is defined as the algorithm's ability to deliver correct sensor state assessments, even in the presence of multiple sensor faults or parameter variations. This paper focuses on a robustness analysis of the proposed method [6] and is structured as follows: First, the method is briefly described. This is followed by a detailed description of the algorithm's parameters. The case study analyzes stepwise parameter variations for a structure with multiple sensor faults. The results are discussed, followed by conclusions and an outlook.

SENSOR FAULT DETECTION AND LOCALIZATION

Based on prior work [8], [9] and [10], Winnewisser et al. [6] developed a methodology to detect, localize, and compensate faulty sensors in SHM systems affected by aging. The approach uses combinatorial consistency evaluation by forming sensor subgroups between outputs from a reference period (healthy sensors and structure) and the current data frame (potentially faulty sensors) to quantify sensor trust over time. Figure 1 illustrates the method.

A time point j is used to denote a specific data frame $\widetilde{X}_{K,j}$ given in terms of samples along the considered timeline. To convert samples to seconds, divide the number of samples by the sampling rate $f_s = 40$ Hz. Sensor groups for cross-validation are formed with $\binom{|Q|}{|k|}$, where $|Q|$ is the total sensor number and $|k|$ is the group size. For each group the Mahalanobis distance between the windowed current measurements and a undisturbed reference frame $\widetilde{X}_{K,0}$ is used to approximates the true distance as $\delta_{\text{true}} \approx \delta_M \left(E\left[\widetilde{X}_{K,c}\right], E\left[\widetilde{X}_{K,0}\right] \right) + \Delta_{\delta_M}$, where Δ_{δ_M} denotes the approximation error, caused by the Mahalanobis distance, as it does not fully capture differences in the shapes of multivariate distributions. For linear distribution shifts, as exclusively considered in this work, $\Delta_{\delta_M} \approx 0$ can be assumed. $\delta_{M,i}$ are then mapped onto a level of epistemic uncertainty (α-level) between 0 and 1, using a partly linear mapping between two statis-

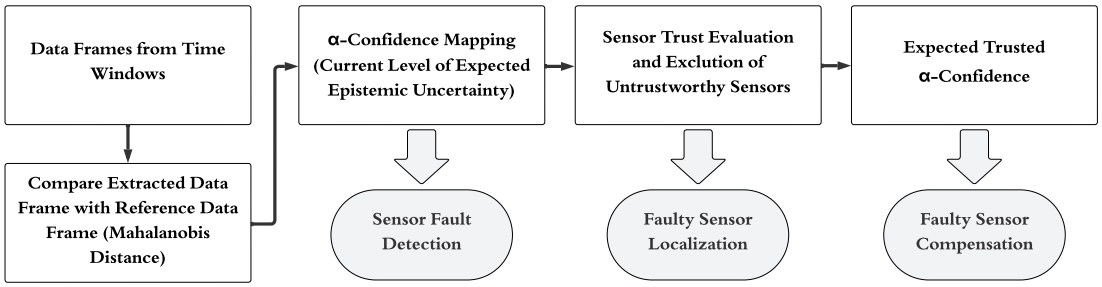

Figure 1. Simplified scheme of the proposed confidence and trust determination method.

tical thresholds, for characterizing affiliation of new sample points to known probability distributions δ_{min} and δ_{max}, representing minimal and maximal consistency. Those are derived using common statistical knowledge or can be determined through a χ^2-test. To model and quantify the current expected level of epistemic uncertainty, within in the overall SHM system, a Beta distribution is fitted to collect evidence for specific α-levels as $C_j^{d_M} \sim \text{Beta}\left(\alpha \,|\, \alpha_j^C, \beta_j^C\right)$, specific for the permutation set of possible sensor subgroups P and distance d_M with parameters α_j^C, β_j^C specific for $\widetilde{D}_j$. These parameters are derived from distribution fitting for α-evidence $\alpha_{e,j} = [\alpha_{l,j}]$ and characterize the Beta distribution, where $\alpha_{l,j} = u(\delta_{l,j} \,|\, \delta_0, \delta_{\text{max}})$ with $\delta_{l,j} = d_M\left(E\left[\widetilde{X}_{P(l),j}\right], \widetilde{X}_{P(l),0}\right)$ for $P(l) = \{K \subseteq Q\}$, $l = \{1, 2, \ldots, p\}$, and $p = \binom{|Q|}{|K|}$. $E\left[C_j^{d_M}\right]$ represents the expected α-confidence level, serving as an indicator of the current level of uncertainty in the system. The trustworthiness of each sensor is assessed based on positive and negative evidence counted as integers, if its associated combinations tend to increase or decrease the expected α-confidence. Following [11], the trust for a sensor s_i on the considered time frame T_c is modeled with a Beta distribution $Tr_j^{s_i} \sim \text{Beta}\left(\alpha_j^{s_i}, \beta_j^{s_i}\right)$. The integer $\alpha_j^{s_i}$ indicates that sensor s_i reduces or maintains epistemic uncertainty as $\alpha_{l,j} \geq E\left[C_j^{d_M}\right]$ increases the expected α-Level. β indicates negative evidence, where $\alpha_{l,j}$ is obtained from s_i and in the case $\alpha_{l,j} < E\left[C_j^{d_M}\right]$ s_i increases epistemic uncertainty while reduces the α-level. To identify the set of all potential corrupted sensors, a trust threshold is defined based on the $2/3$ tertile of all expected trusts of the sensors $E[Tr_j^{s_i}]$. A sensor s_i falling below the trust threshold $E[Tr_j^{s_i}] = (\alpha_j^{s_i})/(\alpha_j^{s_i} + \beta_j^{s_i} + 2) < 2/3$ is excluded from subsequent confidence assessments (expected trusted α-level), ensuring that the system confidence remains high. This approach enables the SHM system to maintain robustness against sensor degradation by isolating untrustworthy sensors and preserving global monitoring capabilities.

CRITICAL PARAMETERS AND KEY INDICATORS FOR ROBUSTNESS

Before conducting a robustness analysis in the context of a case study, a set of the most critical parameters is considered. Subsequently, optimal domains are determined regarding a specific structure.

Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ be a vector of parameters. The associated parameter space

Θ is defined as the Cartesian product of the individual parameter domains:

$$\Theta = \prod_{i=1}^{n} \Theta_i, \quad \text{with } \Theta_i = \begin{cases} \{\theta_i^*\} & \text{if } \theta_i \text{ is fixed,} \\ [a_i, b_i] & \text{if } \theta_i \text{ is variable.} \end{cases} \tag{1}$$

The investigated parameters comprise: The **cross-validation group size** (k) refers to the number of sensors per permutation, given by $\binom{|Q|}{|k|}$. The choice of k directly influences the number of α-evidences, which in turn affects the shape and variance of the α-confidence distribution. The **window size** (w) determines the length of the data segment over which sensor measurements are aggregated and evaluated. A smaller w enables quicker reactions to changes but can yield unstable estimates due to the limited sample size. Conversely, a larger w potentially improves the convergence of central moments but may lead to delayed fault detection and less responsiveness to abrupt changes. The **update interval** (u) specifies the time points j for which the proposed framework is evaluated, as the time points are spread equidistantly starting from $j = 0$. Consequently, this also determines how often α-confidence within the overall sensor system and the sensors' trust are evaluated. A smaller u means a higher temporal resolution and vice versa. Depending on the chosen window size, i.e., $w > u$, overlaps of the considered data frames occur. It should be noted that different u pose distinct classification problems to the proposed framework. The values for w and u are given in terms of considered sample points. For the reason of completeness, the **probabilistic limit values** δ_{min} and δ_{max} should be mentioned. However, these are not considered extensively as their influence is marginal.

To assess the detection and localization quality of the algorithm, the **confusion matrix** is used which categorizes the outcomes of the binary classification of sensor states (faulty or healthy) in a 2×2 matrix. A true positive (TP) occurs when a faulty sensor is correctly identified as defective. A true negative (TN) refers to a healthy sensor that is accurately classified as functional. In contrast, a false positive (FP) describes a healthy sensor that is incorrectly labeled as faulty. Finally, a false negative (FN) occurs when a defective sensor is not detected and is mistakenly classified as healthy.

Further, the $\boldsymbol{F_1}$**-score** is employed to evaluate the robustness of the algorihtmic framework based on the confusion matrix. It is defined as the harmonic mean of precision (the proportion of correctly identified positive cases among all predicted positives) and recall (the proportion of correctly identified positives among all actual positive cases). It ranges from 0 (poor classification) to 1 (accurate classification).

$$F_1\text{-score} = \frac{2\,\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}} \tag{2}$$

CASE STUDY AND ROBUSTNESS ANALYSES

To assess the robustness of the proposed framework, a comprehensive case study was conducted using a 9-meter steel lattice mast instrumented with nine acceleration sensors. The structure in Figure 3 (a), described in [12] and [6], was modeled using a Finite Element model, with parameters fitted based on long-term measurement data from the real structure. It was further used to generate artificial acceleration signals given a

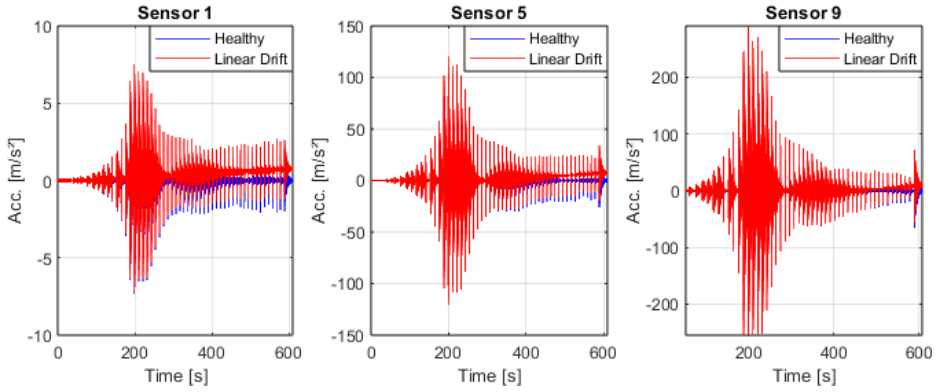

Figure 2. Healthy and drifted synthetic raw data output of the faulty sensors. A sensor drift starts for S_1 at 100 sec., S_5 at 300 sec. and S_9 at 450 sec.

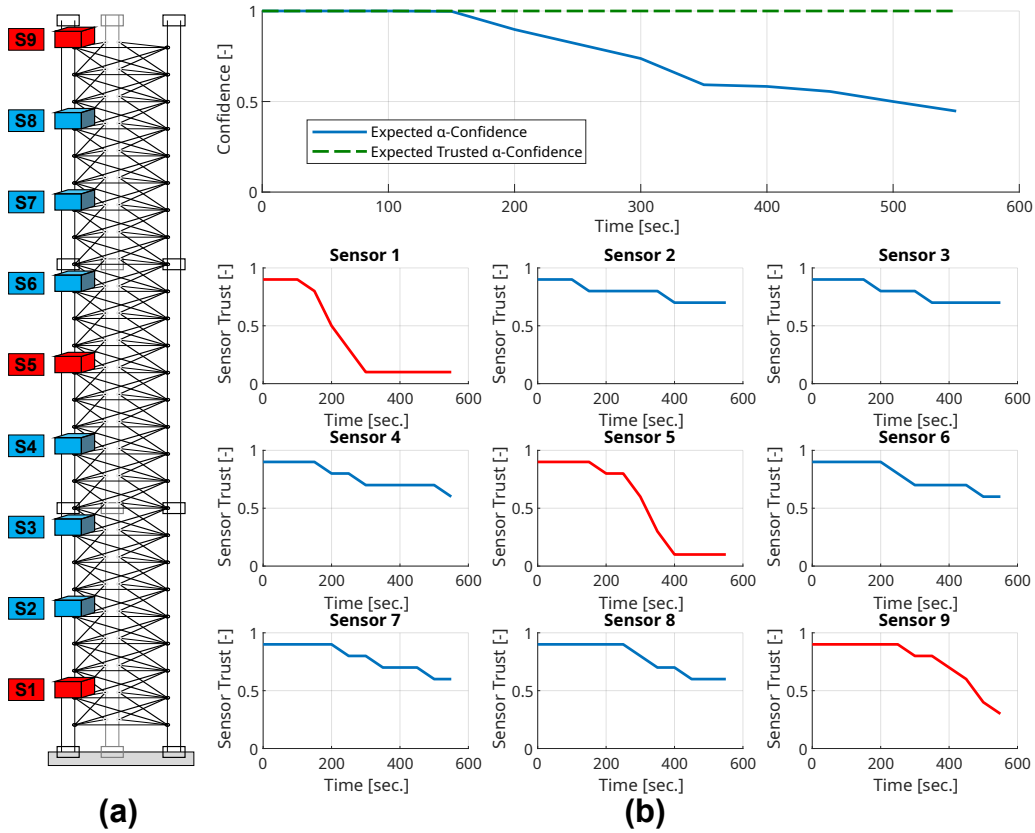

Figure 3. **(a)** Structure showing faulty sensors in red. **(b)** Expected α-confidence and expected trusted α-confidence (top); nine individual sensor trust scores (bottom).

single harmonic excitation as in [13]. Eventually, sensor drifts were induced at different time points in sensors S_1, S_5, and S_9 to imitate sensor aging (see Figure 2).

As baseline parameter configuration, this work follows the suggestions in [6] and [13]. This initial setting is given by $\Theta_{initial} = \{(k, w, u, \delta_{min}, \delta_{max}) \in \{2\} \times \{2000\} \times \{2000\} \times \{0.95\} \times \{0.997\}\}$. In [13] a w and u of equal size is recommended. Figure 3 (b) shows the results of the initial configuration with $F_{1,intial} = 0.878$. Faulty sensors decrease α-confidence and can be identfied by significantly reduced expected trust (red).

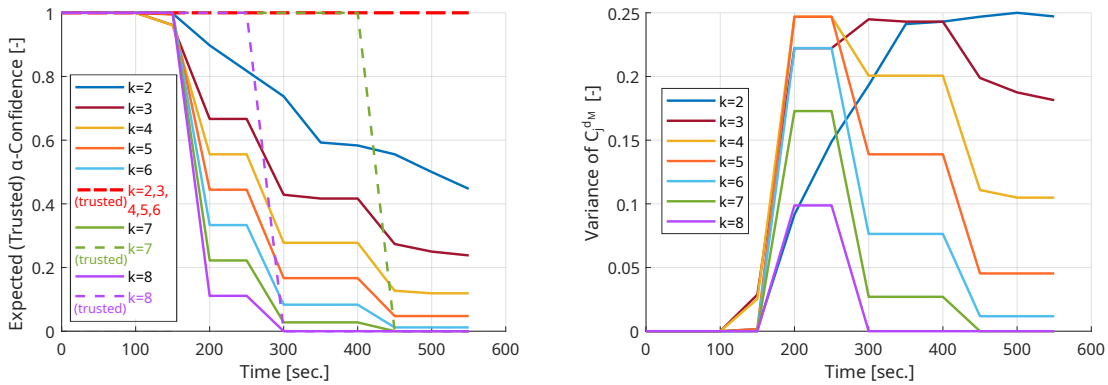

Figure 4. Expected (trusted) α-confidence (left) and corresponding variance (right) for the investigated structure with three asynchronous drifting sensors.

Note that, unlike for the computation of the expected trusted α-confidence, the evidence from faulty sensors was not omitted for the trust computations for illustrative purposes. The expected trusted α-confidence remains stable due to the exclusion of the faulty sensors.

Based on $\Theta_{initial}$ the influence of different group sizes is investigated. Figure 4 shows the results for various $k \in \{2, 3, 4, 5, 6, 7, 8\}$. The slope of the curve of expected α-confidence serves as an indicator for detection sensitivity, i.e., a steep decline reflects rapid identification of anomalies. Accordingly, a larger k appears preferable. However, $k = 7$ and $k = 8$ fail to compensate faulty sensors, clearly indicated by expected trusted α-confidence dropping to 0 instead of maintaining an α-level of 1. Achieving an expected (trusted) α-level of 1 corresponds to maintaining a consistent majority with no epistemic uncertainty present. The compensation after sensor malfunction is critical to preserving the capability of the algorithm to detect, localize and eventually compensate arising sensor faults. The variance of α-confidence $C_j^{d_M}$ reflects the arising inconsistency or discrepancy among the arising α-evidence. A smaller k exhibits higher variance in the long run ($k = 2, 3$). While for an intermediate k temporarily high variance is a good sign due to a larger amounts of α-evidence (compare the corresponding binomial coefficient), for small k this comes from indiscernibility. This emphasizes that small k is not preferable. The F_1-score is used to assess the robustness regarding the localization of faulty sensors. Intermediate group sizes, particularly $k = 4$ and $k = 5$, show the most promising results by achieving high F_1-scores ($F_{1,k=4} = 0.973$, $F_{1,k=5} = 0.900$). Consequently, it can be stated that $k = 4$ provides the most stable and robust overall performance in this case study.

In the next step, the window size and the update interval are vary for a fixed $k = 4$ in order to find robust regions in the parameter space $\Theta = \{(k, w, u, \delta_{min}, \delta_{max}) \in \{4\} \times [200, 4000] \times [200, 4000] \times \{0.95\} \times \{0.997\}\}$. According to [13], a minimum of 100 data points per window is recommended. Hence, both parameters start at a minimum of 200 data samples, with increments of 200 to define the sampling granularity. The heat map in Figure 5 shows the F_1 score of the detection accuracy for the underlying setting for w and u. Varying over an interval for u can be considered as posing distinct classification problems. A vertical belt along u of parameter configuration with $F_1 = 1$ lies between

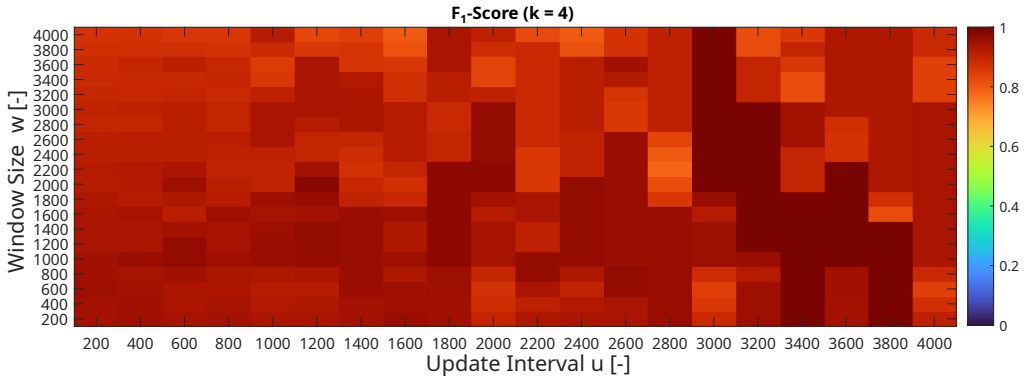

Figure 5. Heat map for $k = 4$ with variable window size w $[-]$ and update interval u $[-]$, showing a vertical belt of the robust configuration (dark red). Both parameters can be converted to time in seconds via T [sec] $= w$ $[-]/f_s$ [Hz] and T [sec] $= u$ $[-]/f_s$ [Hz].

$[3000, 3800]$ for w. A very similar pattern was observed for the cases $k = 3$ and $k = 5$. In addition, heat maps for smaller or larger k show an overall worsening of the F_1-score in the parameter space, but it is noticeable that the robust region also remains approximately in the same boundaries. This indicates that there are potentially simpler problems posed to the algorithmic framework with $u \in [3000, 3800]$. It can be concluded that there appear more or less difficult classification tasks in different phases of sensor anomaly. Considering $u = 3000[-] = 75[sec]$, evaluations of the framework appear at $j = 75, 150, 225, 300, 375, 450, 525$ and a large k performs better than a smaller k. This indicates that larger k are preferable in stable regions. In contrast, when considering $u = 3400[-] = 85[sec]$, evaluations appear at $j = 85, 170, 255, 340, 425, 510$ and a smaller k achieves higher F_1-scores. Consequently, a smaller k is favorable in regions where sensor anomalies are in an early phase.

CONCLUSIONS

This work investigated the robustness of a trust- and confidence-based framework for detecting, localizing, and compensating sensor malfunctions in degrading SHM systems. Through systematic variation of core parameters, including the sensor group size, the data frame size that determines measurement samples, and the update interval size, the methodology proved to be robust against multiple asynchronous sensor drifts. The robustness study revealed that group size has the most influence on the detection and localization quality and intermediate group sizes are preferable due to the highst number of possible permutations that can be used for cross-validation. The results further indicate that the data frame size depends on the posed classification problem arising from different phases regarding the occurence of sensor degradation.

The results demonstrate that the proposed method supports robust SHM by maintaining confidence in the level of uncertainty at the system-level despite localized sensor faults. Future work considers other engineering structures and corresponding sensor measurements. Further, it should extend the robustness analyses by considering different types of sensor anomalies, more sophisticated distances, examining trust thresholds,

and focusing on synchronous sensor failure leading to Byzantine fault problems.

ACKNOWLEDGMENT

The German Research Foundation (DFG) funded this research, as part of the Priority Program SPP 100+ (subproject D01, grant number 2388501624329) and the Collaborative Research Centre 1463 (SFB 1463) "Integrated Design and Operation Methodology for Offshore Megastructures" (C01, grant number 434502799). This article presents the opinions of the authors and does not represent opinions of the funding entities.

REFERENCES

1. Gatti, M. 2019. "Structural health monitoring of an operational bridge: A case study," *Engineering Structures*, 195:200–209.
2. Beer, M., S. Ferson, and V. Kreinovich. 2013. "Imprecise probabilities in engineering analyses," *Mechanical Systems and Signal Processing*, 37(1-2):4–29.
3. Salomon, J., N. Winnewisser, P. Wei, M. Broggi, and M. Beer. 2021. "Efficient reliability analysis of complex systems in consideration of imprecision," *Reliability Engineering & System Safety*, 216:107972.
4. Kiureghian, A. D. and O. Ditlevsen. 2009. "Aleatory or epistemic? Does it matter?" *Structural Safety*, 31(2):105–112.
5. Kamariotis, A., K. Vlachas, V. Ntertimanis, I. Koune, A. Cicirello, and E. Chatzi. 2025. "On the Consistent Classification and Treatment of Uncertainties in Structural Health Monitoring Applications," *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, 11(1).
6. Winnewisser, N. R., M. Hoyer, J.-H. Bartels, F. Mett, T. Potthast, S. Marx, and M. Beer. 2025. "How to Determine the Level of Epistemic Uncertainty and Exclude Faulty Sensors in Structural Health Monitoring Systems," in *IABSE Symposium Tokyo 2025*.
7. Patton, R. J. 1991. "Fault detection and diagnosis in aerospace systems using analytical redundancy," *Computing & Control Engineering Journal*, 2(3):127.
8. Bartels, J.-H., T. Potthast, S. Möller, T. Grießmann, R. Rolfes, M. Beer, and S. Marx. 2024. "Robust SHM: Redundancy approach with different sensor integration levels for long life monitoring systems," *e-Journal of Nondestructive Testing*, 29(7).
9. Fu, Y., C. Peng, F. Gomez, Y. Narazaki, and B. F. Spencer. 2019. "Sensor fault management techniques for wireless smart sensor networks in structural health monitoring," *Structural Control and Health Monitoring*, 26(7):e2362.
10. Lo, C., Y. Bai, M. Liu, and J. P. Lynch. 2015. "Efficient Sensor Fault Detection Using Group Testing," *arXiv preprint arXiv:1501.04152*.
11. Josang, A. and R. Ismail. 2002. "The Beta Reputation System," in *BLED 2002 Proceedings. 41*, pp. 324–337.
12. Wernitz, S., B. Hofmeister, C. Jonscher, T. Grießmann, and R. Rolfes. 2022. "A new open–database benchmark structure for vibration–based Structural Health Monitoring," *Structural Control and Health Monitoring*, 29(11).
13. Bartels, J.-H., F. Mett, N. Winnewisser, T. Potthast, M. Beer, and S. Marx. 2025. "Probabilistic Sensor Fault Detection in Structural Health Monitoring Systems Using Mahalanobis Distance," in *Proceedings of the 11th International Conference on Experimental Vibration Analysis of Civil Engineering Structures*, pp. 1–12.