Reliability Engineering and System Safety 145 (2016) 262–276

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Analyzing simulation-based PRA data through traditional and topological clustering: A BWR station blackout case study

D. Maljovec [a,*], S. Liu [a], B. Wang [a], D. Mandelli [b], P.-T. Bremer [c], V. Pascucci [a], C. Smith [b]

[a] Scientific Computing and Imaging Institute, University of Utah, 72 S Central Campus Drive, Salt Lake City, UT 84112, United States
[b] Idaho National Laboratory, 2525 Fremont Avenue, Idaho Falls, ID 83415, United States
[c] Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550, United States

ARTICLE INFO

Available online 14 July 2015

Keywords:
Probabilistic risk assessment
Computational topology
Clustering
High-dimensional data analysis

ABSTRACT

Dynamic probabilistic risk assessment (DPRA) methodologies couple system simulator codes (e.g., RELAP and MELCOR) with simulation controller codes (e.g., RAVEN and ADAPT). Whereas system simulator codes model system dynamics deterministically, simulation controller codes introduce both deterministic (e.g., system control logic and operating procedures) and stochastic (e.g., component failures and parameter uncertainties) elements into the simulation. Typically, a DPRA is performed by sampling values of a set of parameters and simulating the system behavior for that specific set of parameter values. For complex systems, a major challenge in using DPRA methodologies is to analyze the large number of scenarios generated, where clustering techniques are typically employed to better organize and interpret the data. In this paper, we focus on the analysis of two nuclear simulation datasets that are part of the risk-informed safety margin characterization (RISMC) boiling water reactor (BWR) station blackout (SBO) case study. We provide the domain experts a software tool that encodes traditional and topological clustering techniques within an interactive analysis and visualization environment, for understanding the structures of such high-dimensional nuclear simulation datasets. We demonstrate through our case study that both types of clustering techniques complement each other for enhanced structural understanding of the data.

Published by Elsevier Ltd.

1. Introduction

A recent trend in the nuclear engineering field is the implementation of computationally intensive codes to perform safety analysis of nuclear power plants. In particular, the new generation of system analysis codes aims to address thermohydraulic phenomena, structural behaviors, system dynamics, etc. Often these codes are coupled with stochastic analysis tools, such as dynamic probabilistic risk assessment (DPRA) methodologies, to perform probabilistic risk analysis, uncertainty quantification, and sensitivity analysis.

DPRA methodologies account for possible coupling between triggered or stochastic events through explicit consideration of the time element in system evolution, for example through the use of dynamic system simulators. Such methodologies are useful when the system has multiple failure modes, control loops, processes, software/hardware components, or human interactions. A DPRA is typically performed by sampling values of a set of parameters from the space of interest with uncertainty (using the simulation controller codes) and then simulating the system behavior for that specific set of parameter values (using the system simulator codes). Due to the intrinsically high level of details within such a process, large amounts of data are generated within the simulation [18]. The main challenge in employing DPRA methodologies is how to explore and understand such large amounts of data through effective analysis and visualization.

Related work. A first approach towards understanding such data follows fuzzy classification [11] and classic clustering algorithms [18]. In particular, a clustering algorithm such as the mean-shift [6] partitions the set of scenarios generated by DPRA based on their similarities and the observation density and enables the organization and interpretation of trends and risk contributors in scenario evolution [18,20].

On the other hand, for effective analysis and visualization of DPRA and nuclear datasets in general, we have been investigating the use of topology-based clustering techniques to obtain local, in-depth structural understanding of the data. The clustering technique we utilize focuses on a domain-partitioning algorithm based on a topological structure known as the Morse–Smale complex [3,4], which partitions the data points into clusters based on their uniform gradient flow behavior. In [14], we have built upon a well-established framework that visualizes high-dimensional

* Principal corresponding author: Tel.: +1 814 688 7884.
E-mail address: maljovec@cs.utah.edu (D. Maljovec).

http://dx.doi.org/10.1016/j.ress.2015.07.001
0951-8320/Published by Elsevier Ltd.

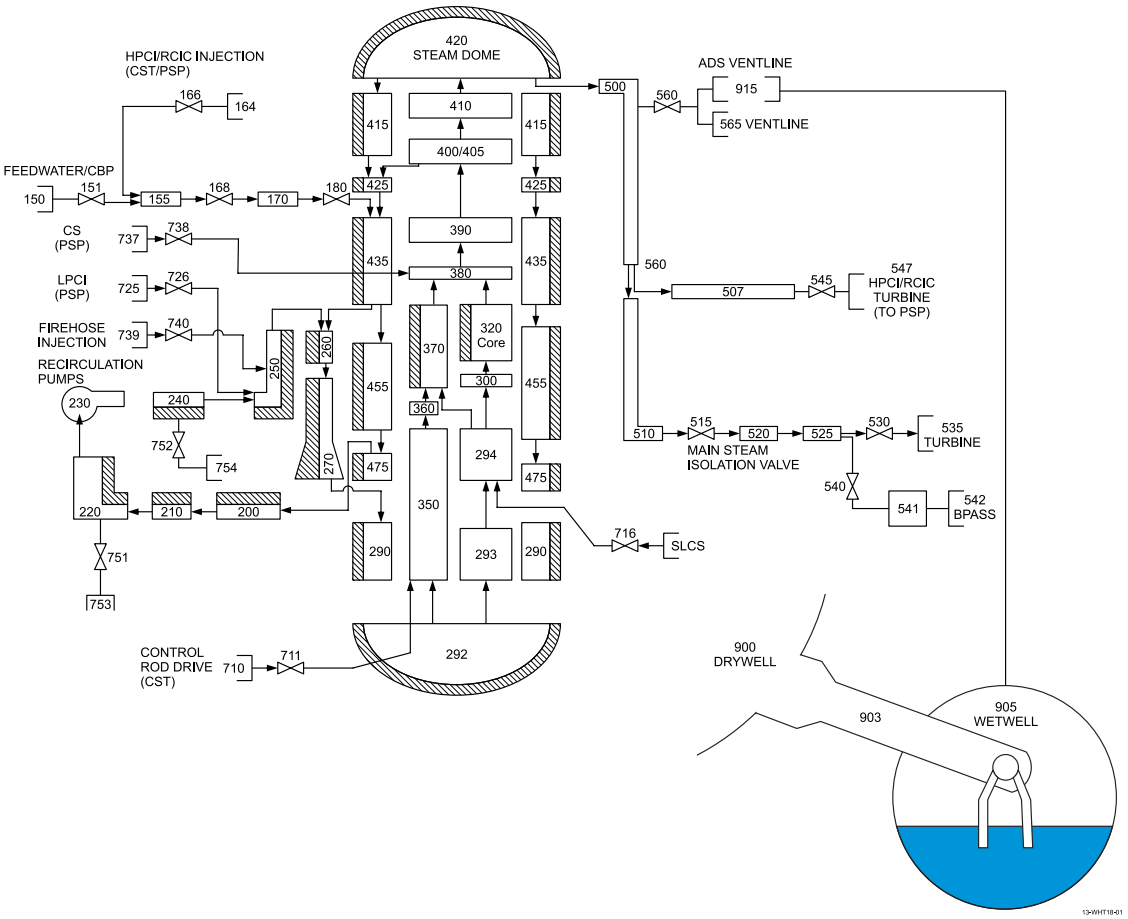

Fig. 1. BWR system considered.

scalar functions through a topological segmentation of its input domain [7,8]. The input of such a high-dimensional function arises from the set of n uncertain parameters $x_1, x_2, \ldots, x_n$, whereas the output originates from some safety-related outcomes, such as the maximum core temperature of each simulation. Our topological tools aim to reconstruct the topological structure of such a function, i.e., the response surface, in the high-dimensional space. We have demonstrated, for the first time, such a framework to nuclear engineers by applying it to data extracted from a VR_2^+ nuclear reactor simulator where a SCRAM event occurs due to system failure. We have further explored the topological clusterings that lie beneath such a framework for DPRA datasets [13] in terms of end-state analysis (which classifies the scenarios into clusters based on their end state, e.g., final outcome, [21]) and transient analysis (which considers the complete system dynamics, e.g., time evolution of scenarios, and identifies clusters having similar temporal behavior of the state variables [18]). The tools we develop have been briefly described in surveys and technical reports that summarize methodologies and algorithms that are implemented within the RISMC project and are under development for RAVEN [15,16].

Our contribution. This paper includes and extends our earlier work in [12]. Compared to prior work mentioned above, our main contributions are as follows : First, we present an in-depth application discussion that focuses on the analysis of two particular nuclear simulation datasets that are part of the risk-informed safety margin characterization (RISMC) boiling water reactor (BWR) station blackout (SBO) case study [17]. Second, we enrich our previously developed tool [13,14] by combining traditional hierarchical clustering and topological clustering, as well as dimensionality reduction (DR) techniques.

We demonstrate through our first data example that both types of clustering techniques complement each other in enhancing structural understanding of the data. In particular, the topological clustering helps highlight key features of the data that are otherwise hidden using traditional techniques. In the second example, we explore new ways of thinking about risk-informed data by incorporating probability information into the topological analysis in order to characterize the most probable area of the identified failure region, in addition to a well-established analysis of the data's observed output, namely, the maximum temperature reached by the cladding.

Compared to [12], this extended version includes one additional, more complex BWR SBO dataset in the analysis and visualization, as well as provides a more complete exposition of our enriched analysis and visualization toolset in understanding such nuclear simulation datasets. In a nutshell, the power of clustering comes from the notion that a large number of simulations spanning a large input and potentially large output space can be distilled into a few canonical or interesting cases. Agglomerative hierarchical clustering and topological clustering provide two very different views of the data, and in this work, we attempt to highlight the advantages of each and demonstrate, through the case study provided, when one method may be preferred over the other depending on the user's needs. We also demonstrate that, in some cases, the two methods can validate trends in the data in a complementary fashion.

BWR system. The system considered in both simulation datasets is a generic BWR power plant with a Mark I containment as shown in Fig. 1. The three main structures are the reactor pressure vessel (RPV), a pressurized vessel that contains the reactor core; the primary containment including the drywell (DW) that houses the

RPV and circulation pumps; and the pressure suppression pool (PSP), also known as the wetwell. The PSP is a large torus-shaped container that contains a large amount of water (almost 1 million gallons of fresh water) and is used in specific situations as an ultimate heat sink. The BWR system includes a large number of subsystems, but for the scope of this paper and for the case study considered, we use a smaller subset of systems that includes the RPV level control system, the RPV pressure control system, the cooling water inventory, and the AC power system. The AC power system consists of two power grids, emergency diesel generators (DGs), and battery systems for the instrumentation and control systems.

The RPV level control system provides manual and automatic control of the water level within the RPV and consists of two components, the reactor core isolation cooling (RCIC) and the high pressure core injection (HPCI). The RCIC provides high-pressure injection of water from the condensate storage tank (CST) to the RPV. Water flow is provided by a turbine-driven pump that takes steam from the main steam line and discharges it to the suppression pool. The HPCI functions similarly but allows a much greater water flow rate.

The RPV pressure control system provides manual and automatic control of the RPV internal pressure and consists of a set of safety relief valves (SRVs), safety valves, and the automatic depressurization system (ADS). The SRVs are DC-powered valves that control and limit the RPV pressure, and the ADS is a separate set of relief valves that are employed in order to depressurize the RPV.

The cooling water inventory includes the CST, the PSP, and the fire water system. The CST in the considered plant is a 375 kgal fresh water reservoir that can be used to cool the reactor. The PSP contains a large amount of fresh water that is relied upon as an ultimate heat sink when AC power is lost. Water from the fire water system can be injected into the RPV when other water injection systems are disabled and when the RPV is depressurized.

*SBO scenario.*The scenario considered in this paper is the loss of offsite power (LOOP) event followed by the loss of the diesel generators (DGs), i.e., the station blackout (SBO) initiating event. In particular, at time $t=0$, a LOOP condition occurs due to an external event. Therefore, the LOOP alarm triggers the following events:

1. A successful scram of the reactor is performed by the operators.
2. Main steam isolation valves are successfully closed, isolating the primary containment from the turbine building.
3. Emergency DGs start successfully to keep the AC power buses energized.

It is assumed that the DC systems (i.e., batteries) are functional, and the decay heat generated by the core is successfully removed from the RPV through the residual heat removal system.

At some point, an SBO condition may occur due to some internal failure, where the set of DGs fails, thus impeding the removal of decay heat. Reactor operators then start the SBO emergency procedures and perform RPV level control using RCIC or HPCI, RPV pressure control using SRVs, and containment monitoring (both drywell and PSP). At this point, plant staff members start to bring the DGs back online while recovering the off-site power grid. Due to heavy usage, battery power can be depleted. When this happens, all remaining control systems become off-line, causing the reactor core to heat until the maximum temperature limit for the clad is reached, where a core damage (CD) condition occurs.

If DC power is still available and one of three conditions is met (i.e., failure of both RCIC and HPCI, HCTL limits have been reached, or RPV water level becomes too low), then the reactor operators activate the ADS in order to depressurize the RPV and allow fire water injection when available. As an emergency action, when RPV pressure is below 100 psi, plant staff can connect the fire water system to the RPV in order to cool the core and maintain an adequate water level. Such a task is, however, hard to complete since physical connection between the fire water system and the RPV inlet has to be made manually. When AC power is recovered, through successful restart/repair of DGs or off-site power, the residual heat removal system can be employed to keep the reactor core cool.

Overview. In our case study, we investigate datasets that model the maximum temperature reached by the reactor cladding and the overall system success or failure in terms of recovering from an SBO event, while varying the timings of failure or recovery of the various subsystems described above. We therefore model the data as a high-dimensional function of these timing parameters, whose real-valued output corresponds to the maximum temperature of the reactor cladding, the time it takes for a failure to occur (i.e., when the cladding breaches a preset maximum temperature), or the overall simulation success or failure. Our objective is to summarize a large amount of scenarios into a manageable number of meaningful categories by performing traditional and topological clusterings. We describe these methods and the subsequent visualizations of their results in detail.

2. Technical background

Dimensionality reduction (DR) and traditional hierarchical clustering are widely used techniques for high-dimensional data analysis. To extend the existing framework we have developed in [13,14], we employ a visualization system that utilizes more standard clustering and DR techniques in addition to the topological methods. The topological methods require a slightly different treatment of the data, yet follow the same basic principle as using DR to construct a mapping of the clustering results for intuitive visual analysis. We begin with a brief description of DR and traditional hierarchical clustering techniques, and then focus on the topological clustering, which may be unfamiliar to non-specialists. We include some technical details in our system which are most relevant to the related work reviewed here.

Dimensionality reduction (DR) and traditional hierarchical clustering are widely used techniques for high-dimensional data analysis. To extend the existing framework we have developed in [13,14], we employ a visualization system that utilizes more standard clustering and DR techniques in addition to the topological methods. The topological methods require a slightly different treatment of the data, yet follow the same basic principle as using DR to construct a mapping of the clustering results for intuitive visual analysis. We begin with a brief description of DR and traditional hierarchical clustering techniques and then focus on the topological clustering, which may be unfamiliar to non-specialists. We include some technical details in our system that are most relevant to the related work reviewed here.

Dimensionality reduction. DR techniques [1], such as principal component analysis (PCA) [9], multi-dimensional scaling (MDS) [10], and Isomap [19], are common tools for analyzing high-dimensional data by constructing its low-dimensional representation. Since direct visualization of high-dimensional data is extremely challenging, we would like to obtain some intuition regarding the structure of the data through its low-dimensional embedding. Such embeddings are typically constructed in 2D or 3D spaces for visualization purposes. We have integrated a number of DR techniques into our system. For the purpose of our study, we use primarily PCA, a linear DR technique, due to its simplicity and computational efficiency. Using DR alone as a black

box solution in the analysis suffers a major limitation, that is, the results could be hard to interpret as a certain amount of structural information could be lost during the DR process. Therefore, we try to impose structural context for the embeddings by combining DR results with clusterings obtained from the original high-dimensional data.

The projection offered by PCA is not axis-aligned, which makes interpreting the horizontal and vertical axes somewhat difficult in the context of the original input parameters. We can retain some of this information by either showing the projected axes or by employing a colormap to show how well a particular parameter aligns with one or more of the principal components. The latter method is used during our initial analysis, as it helps highlight which parameters most strongly influence the global structure of the data.

Traditional Hierarchical Clustering. A clustering groups the data in such a way that points are more similar to those in the same cluster than to those outside the cluster. There are numerous criteria (based on density, distribution, distance, or connectivity, etc.) for defining what constitutes a cluster. In our current analysis, we choose average-linkage hierarchical clustering [2] (among others available in the system). Such a clustering technique is based on point-wise connectivity where points are considered more related to nearby points than points that are farther away. Starting from individual points as their own clusters, this technique builds a dendrogram from the bottom up, merging nearby clusters. In our system, the number of clusters does not need to be specified a priori; instead, the user interactively expands or collapses different levels of clustering in the hierarchy during the analysis.

Visualizing high-dimensional clusters. We visualize high-dimensional clusters obtained by hierarchical clustering using their PCA projections. An example is shown in Fig. 2 (left) where sampled points from a 3D paraboloid (defined by the function $f(x,y)=x^2+y^2$) are visualized by combining their hierarchical clustering results with PCA. The data is normalized using z-score standardization to ensure each dimension has a mean of zero and a standard deviation of one. After clustering, the points are projected onto their first two principal components and colored according to their cluster labels. We see that the five clusters correspond to the four corners as well as the area surrounding the global minimum of the paraboloid. In our visualization toolset, we utilize various visualization techniques to understand the relative size of the clusters, detect outliers, and identify key parameters that characterize each cluster. In particular, we enhance the comparison of clusters by providing the ability to interactively drag the clusters apart on the 2D canvas (Fig. 2, bottom right) to prevent occlusion and to help us understand the relative size of the clusters and the dispersions of points within. Furthermore, we include statistical summaries of the individual clusters (Fig. 2, top right), enabling us to characterize the key contributors of their distinct behaviors.

Approximated Morse–Smale complex and topological clustering. We consider an alternative method for clustering high-dimensional data based on the concept of the Morse–Smale complex (MSC). We give a brief overview of these concepts; see [13,14] for details. The MSC is a type of topological structure that serves as a structural summary of a given scalar function. We consider a scalar function $f:\mathbb{X}\rightarrow\mathbb{R}$ defined over a finite set of points $\mathbb{X}$ in $\mathbb{R}^n$. The approximated MSC, at its finest level, partitions the points in $\mathbb{X}$ based on their uniform gradient behavior. First, points in $\mathbb{X}$ are connected with a neighborhood graph (e.g., k-nearest-neighbor (KNN) graph). Second, the steepest ascending edge adjacent to a given point is used to estimate the gradient flow at the point. All points with no neighbors of higher/lower values are considered local maxima/minima. Finally, points are clustered based on the unique minimum-maximum pair from which their gradient flows start and end. A topological clustering at the finest level for a height function defined on a 2D domain is illustrated in Fig. 3(a) and (b). We can then merge clusters based on persistence simplification [5], where less (topologically) significant clusters are merged into more significant ones. We avoid the technical details here but simply illustrate such a process in Fig. 3(d) and (e).

Topological skeleton obtained through DR. Given a topological clustering at a fixed scale, we further our analysis by computing a collection of summary curves that serves as the topological skeleton of the data in the visual space. We follow a three-step process, as detailed in [7]: 1) perform inverse linear regression with data in each cluster and obtain a 1D curve embedded in $\mathbb{R}^n$, 2) project the curves in $\mathbb{R}^n$ to a curve in the visual space using PCA [9], and 3) align the curves in the visual space to meet at their shared extrema to maintain the coherency of the extracted structure. The resulting topological skeleton serves as a structural summary of the data, and it is visualized to encode structural information, as illustrated in Fig. 4. Finally, the topological skeleton can also be visualized based on the cluster labels. In addition, we distinguish the clusters based on configurations of their input dimensions through a collection of inverse coordinate plots.

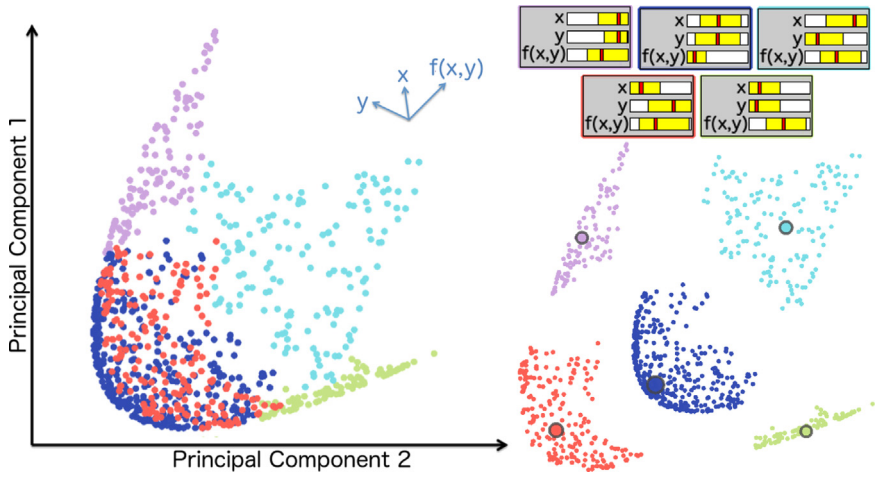

Fig. 2. Left: Points sampled from a 3D paraboloid dataset are projected by PCA and colored according to their cluster labels. Top right: Cluster summaries demonstrate the mean and range of each dimension within a cluster. Bottom right: The clusters are manually rearranged on the 2D canvas for comparison. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

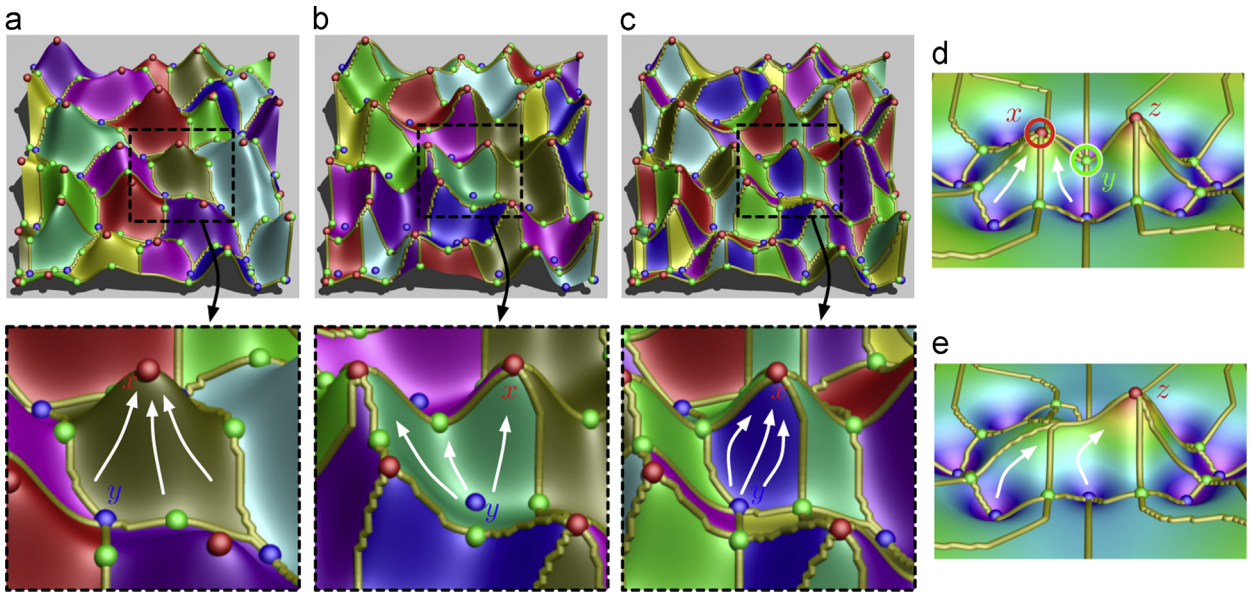

Fig. 3. For a height function defined on a 2D domain (where maxima, minima, and saddles are colored red, blue, and green, respectively): (a) For each point in the brown region, the gradient flow (white arrow) ends at the same maxima x. (b) For each point in the green region, the gradient flow starts at the same minimum y. (c) For each point in the blue region (i.e., a cluster based on the MSC), the gradient flow begins and ends at the same maximum–minimum (i.e., (x,y)) pair. To illustrate merging of clusters based on persistence simplification, in (d), the left peak at the local maximum x is considered less topologically important than its nearby peak at the local maximum z, since x is lower. Therefore, at a certain scale, we would like to represent this feature as a single peak instead of two separate peaks, as shown in (e), by redirecting gradient flow (white arrow) that originally terminates at x to terminate at z. In this way, we simplify the function by removing (canceling) the local maximum x with its nearby saddle y. On the cluster level, the clusters (i.e., decompositions of the domain separated by edges connecting the saddles and extrema) surrounding the left peak x are merged into clusters surrounding the right peak z. Figures are reproduced from [13]. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

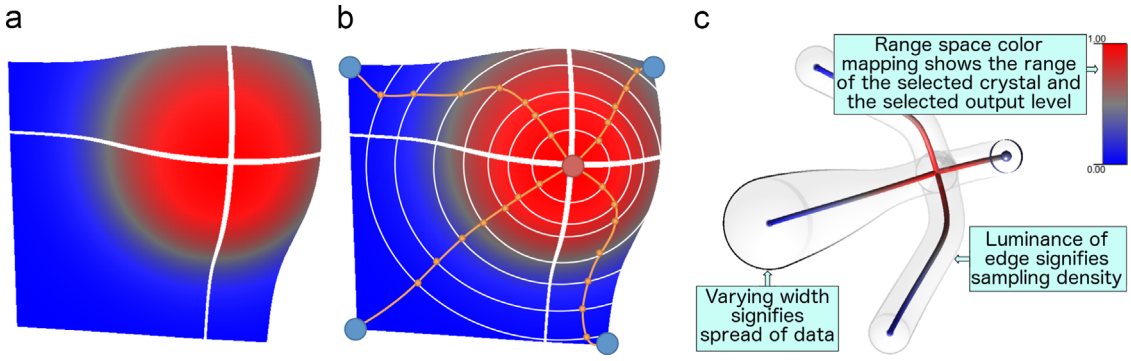

Fig. 4. An illustrative example of our visualization of a topological skeleton extracted from a 2D height function: (a) the surface is first segmented into clusters of uniform gradient flow; (b) then each level set (white line) is averaged to a single point, and consecutive level sets are connected to form a curve per cluster (orange curves); and (c) finally the resulting topological skeleton is visualized. Each summary curve in the visual space corresponds to a cluster of the original high-dimensional data. In the visualization, the color of each curve signifies the average value of each level set, and a transparent region encloses a given curve, where its width represents a direction-independent estimate of the spread of data and the luminance of its boundary edges signifies the sampling density. Figures are reproduced from [13]. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Suppose we employ a point sampling of the same 2D height function in Fig. 4. The above process is illustrated in Fig. 5. For more details of the visualization pipeline, see [7,13,14].

3. Case study dataset 1: 7d simulation ensemble

3.1. Data description

An ensemble of 4997 transient simulations has been generated using classical Monte-Carlo sampling of seven input parameters. Among these simulations, 833 scenarios resulted in system failure (where the core temperature reached the clad failure temperature threshold of $2200^{\circ}\mathrm{F} \approx 1477$ K), whereas the rest of the 4164 scenarios ended in system success (where AC power is recovered or the fire water becomes available when the RPV is depressurized early enough to prevent the cladding from reaching a dangerous temperature). Each simulation includes information regarding the timing of various recovery attempts (e.g., cooling recovery, fire water, etc.) and component failures (e.g., battery life is exhausted or a safety relief valve gets stuck open, etc.). The seven input parameters are listed below, as they are the only uncertain parameters under consideration.

1. *FailureTimeDG*: Failure time of the DGs corresponding to the time of the SBO event.
2. *ACPowerRecoveryTime*: The minimum between the recovery time of DGs and the off-site power recovery time. The minimum of these two will determine when the AC power is considered recovered.

3. *SRVStuckOpenTime*: The time when an SRV is stuck in the open position.
4. *CoolingFailtoRunTime*: The maximum between the HPCI failure time and the RCIC failure time. As long as one of the two high pressure cooling systems (i.e., HPCI and RCIC) is functioning, the reactor is being actively cooled, so it is important to understand when both systems have failed.
5. *ADSActivationTimeDelay*: The time when the operator manually depressurizes the RPV by activating the ADS system. This parameter measures the time delay from when the PSP heat capacity limits are reached.
6. *FirewaterTime*: As an emergency action, when RPV pressure is below 150 psi ($\approx 1.03 \times 10^6$ Pa)150 psi($\approx 1.03 \times 10^6$ Pa), plant staff can connect the fire water system to the RPV to cool the core and maintain an adequate water level. This parameter indicates the time needed to connect the fire water system for injection.
7. *ExtendedECCSOperation*: Battery life combined with extended ECCS operation. That is, operators may extend RCIC/HPCI and SRV control even after the batteries have been depleted. They manually control RCIC/HPCI by acting on the steam inlet valve of the turbine and/or supply DC power to the SRVs through spare batteries.

All the above time-related parameters are measured from the time of the SBO event (in seconds), which is the FailureTimeDG, with the exception of FailureTimeDG, which is measured from the LOOP event, and the ADSactivationTimeDelay, which is measured from the time the PSP reaches its heat capacity limits. The output parameters obtained from the simulations are:

1. *MaxCladTemp*, which is the maximum temperature attained anywhere on the cladding during the entire course of the simulation;
2. *SimulationEndTime*, which for failure cases represents the time to reach the failure temperature of 2200 °F($\approx$ 1477 K).

We study the topology of scalar functions with each of these outputs as the scalar value in isolation. The above data is pre-processed with a Z-score standardization, whereby values V of each dimension are recomputed as $(V - \text{mean}(V))/\text{std}(V)$; therefore all input parameters have the same mean (0) and standard deviation (1) but may vary in their ranges.

In this study, we are interested in what combination of conditions (in the form of input simulation parameters) can cause potential reactor failure.

3.2. Results

We provide analysis under both traditional (Section 3.2.1) and topological clustering (Section 3.2.2) using the 7D input data. For each subsection, we consider two separate cases. In the first case, referred to as the *All scenarios case*, we analyze all 4997 simulations, using maximum clad temperature (maxCladTemp) as the observed output parameter. Note that in this case, all failure cases have the same output parameter of 2200°F ($\approx$ 1477 K) 2200 °F($\approx$ 1477 K). In the second case, referred to as the *Failure scenarios case*, we focus on clustering of the 833 failure scenarios.

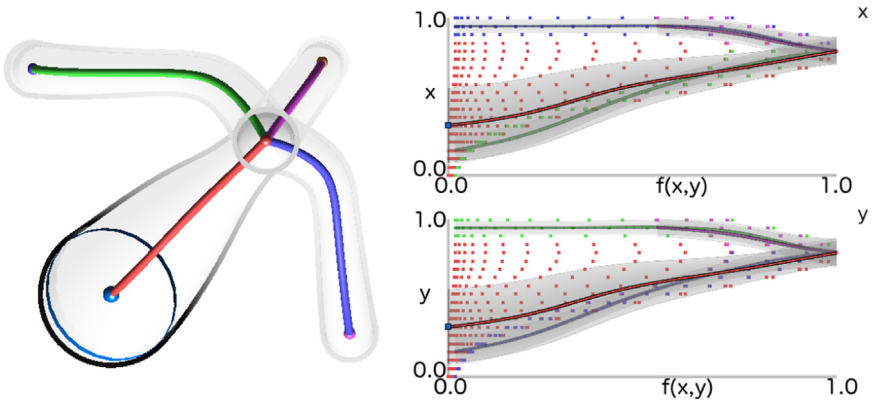

Fig. 5. Left: topological skeleton colored by cluster labels. Right: inverse coordinate plots. Data points are visualized by their cluster labels, and summary curves are projected. For the inverse coordinate plots, the horizontal axis represents the output dimension (e.g., height values), and each vertical axis represents an input dimension (e.g., x or y coordinates of the domain). The projected summary curve in each inverse coordinate plot gives the average value (of the input dimension of interest) at each level set and uses a dimension-specific standard deviation for the width of the transparent region. Visualizations of the data points, summary curves, and their associated standard deviations could be enabled/disabled based on user specifications. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

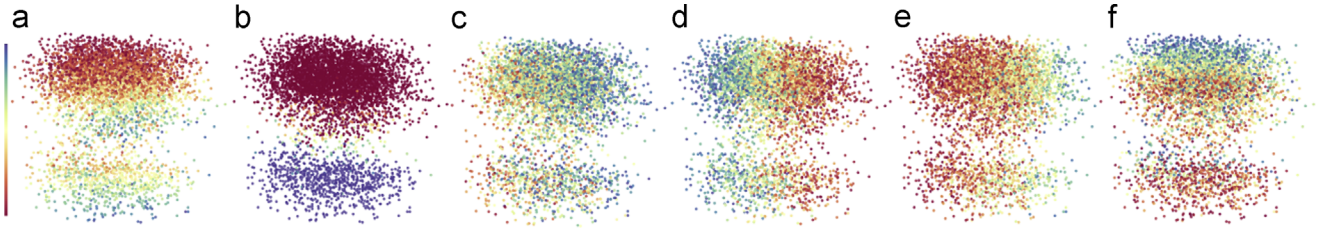

Fig. 6. PCA embedding for the 8D dataset under the *All scenarios case*. The dimensions shown exhibit relatively strong correlation patterns within the embedding. We use a spectral colormap (color bar on the left) where red/blue represents low/high value. (a) ACPowerRecoveryTime, (b) MaxCladTemp, (c) CoolingFailToRunTime, (d) FirewaterTime, (e) SRVStuckOpenTime, and (f) ExtendedECCSOperation. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Since the maximum clad temperature does not vary for these cases, we treat the time of the failure (SimulationEndTime) as the output parameter. We give a comprehensive picture by comparing between the two clustering techniques and discussing the benefits and limitations of each approach.

3.2.1. Traditional clustering

For traditional hierarchical clustering, we map the data into an 8D space by considering the seven input parameters and the output parameter, maximum clad temperature (MaxCladTemp). We start our analysis by applying PCA to reduce the 8D data to its 2D embedding for direct visual analysis.

All scenarios case. To study the distribution/variation of each dimension with respect to the embedding, we first color the points according to each dimension, as illustrated in Fig. 6. All the dimensions shown exhibit a certain amount of visual correlation within the embedding. The two omitted dimensions, ADSActivationTimeDelay and FailureTimeDG, on the other hand, show little to no visual correlation, indicating they account for the least amount of variability in the data.

It is important to note that a vertical or horizontal pattern of variation corresponds to the variance of the dimension. That is, a larger variance corresponds to a more noticeable pattern, which is likely because PCA is inherently optimized for capturing dominant directions of maximum variance.

In Fig. 6(b), there appear to be only a few data points with a moderate MaxCladTemp as the top portion of the embedding is dominated by success scenarios characterized by low MaxCladTemp values (in red), and the bottom portion of the data consists of mostly failure scenarios characterized by high (constant) MaxCladTemp (in blue). It is therefore obvious that MaxCladTemp separates the success from failure scenarios in the embedding. This claim can be further validated by coloring the points with known labels of success/failure.

In Fig. 6(a), ACPowerRecoveryTime varies smoothly within both the success and failure scenarios, but it does not serve as a differentiating factor between the successes and failures. Furthermore, in Fig. 6(f), relatively high ExtendedECCSOperation time can be observed among all the success scenarios, so we suspect that a long extended ECCS operation time is a main contributing factor for stable system recovery. However, ExtendedECCSOperation is likely not a sufficient condition to separate successes from failures as there are a few points with high ExtendedECCSOperation values within the lower half of the embedding (i.e., failures scenarios). In Fig. 6(c)–(e), the remaining three dimensions vary orthogonally with respect to maxCladTemp. This observation implies that these dimensions have less impact on the outcomes of the simulation, which are characterized by variations in maxCladTemp.

In addition, combined with traditional hierarchical clustering, our analysis framework enables us to color the points in the embedding based on cluster labels. Furthermore, the tool also visualizes the statistical summary of each dimension for points within each cluster (bottom of Fig. 7). In the statistical summary of a given cluster, each row represents a dimension of the data, where the yellow bar corresponds to its min–max range, and the red marker indicates its mean value across all points in the cluster. The span of the horizontal bars signifies the total range of values for each dimension. With these summaries across all clusters, we can quickly compare and investigate the defining characteristics of each cluster at a glance (see Fig. 7).

During the interactive exploration of the embedding, we apply cluster expansions recursively to study the data from coarse to fine resolutions. At the coarsest level, the data is split into two clusters,

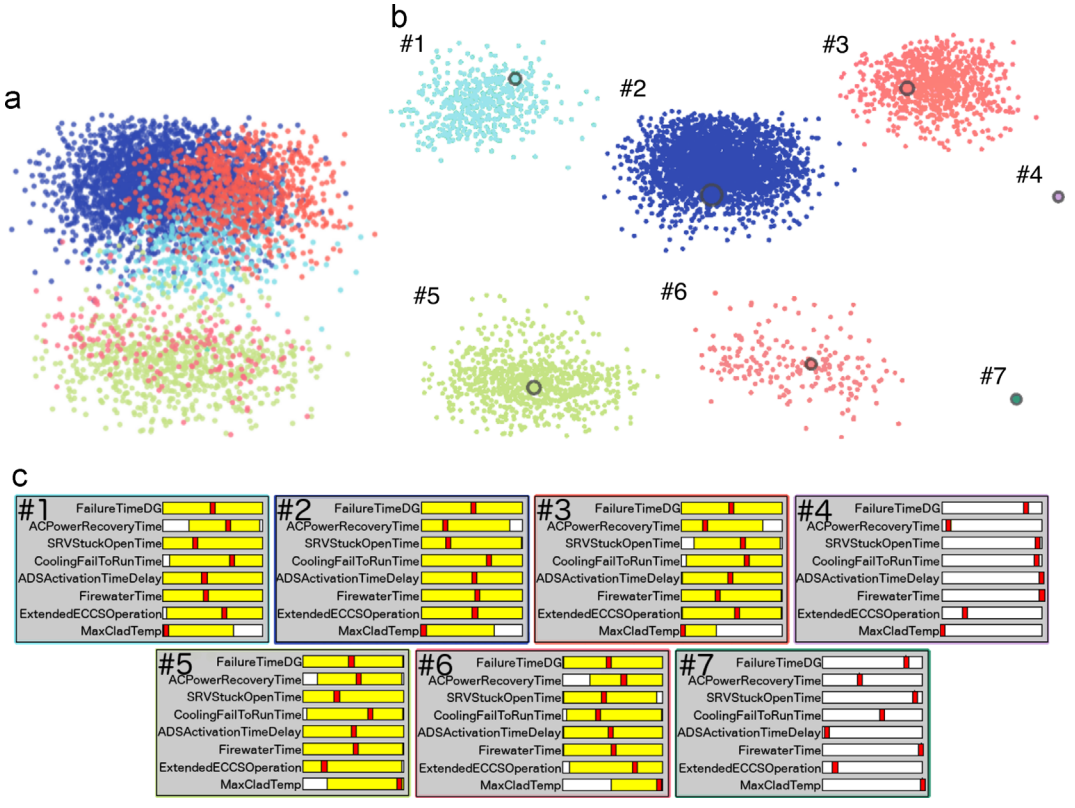

Fig. 7. (a) 2D embedding of the data colored by cluster labels. (b) In order to provide a more clear view for the clusters, we provide a separate illustration of each individual cluster and (c) its summary statistics. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

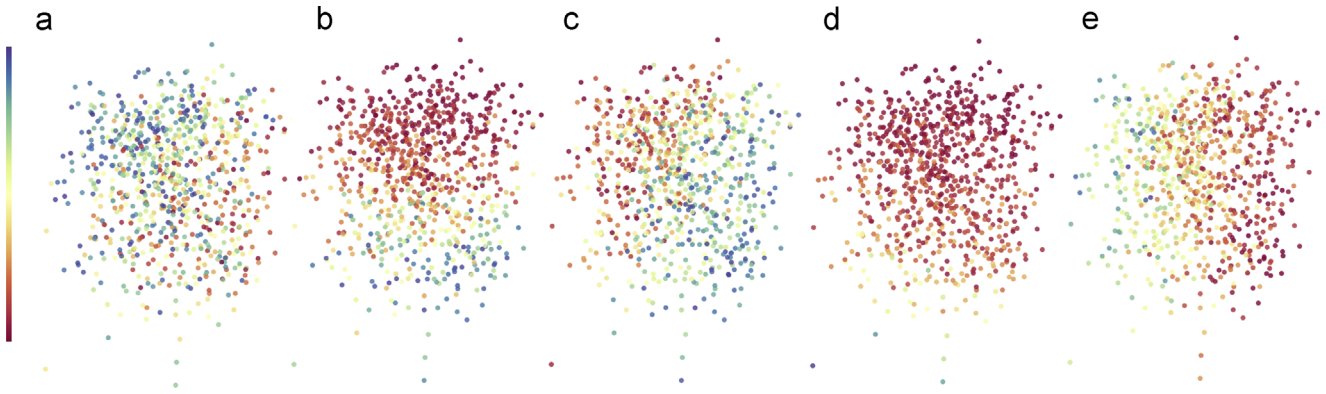

Fig. 8. PCA embedding for the 8D dataset under the *All scenarios case*. The dimensions shown exhibit relatively strong correlation patterns within the embedding. (a) CoolingFailToRunTime, (b) ExtendedECCSOperation, (c) FirewaterTime, (d) SimulationEndTime, and (e) SRVStuckOpenTime.

where the upper cluster contains exclusively success scenarios, and the lower cluster contains all failure scenarios and a small number of successes (verified via known labels of success/failure cases). We subdivide these clusters by applying a few steps of cluster expansion. We then arrive at a level in the clustering hierarchy that consists of seven clusters, as shown in Fig. 7.

The four top clusters decompose all of the success scenarios (top half of the embedding). The single point purple cluster (#4) likely consists of an outlier in the data, since it exhibits extremely low ACPowerRecoverTime and MaxCladTemp. This point corresponds to a success scenario where AC power is recovered very quickly and the clad temperature never increases drastically. Although the blue (#2) and cyan (#1) clusters share similar statistical summaries across most dimensions, ACPowerRecoveryTime seems to be the most likely factor that differentiates these two clusters. The fact that the cyan (#1) cluster has a late ACPowerRecoveryTime but still records success scenarios suggests that this factor is not important for successful system recovery for this cluster, but may be more involved in the blue (#2) cluster. The differentiating factor between the red (#3) cluster and the blue (#2) and cyan (#1) clusters is its late SRVStuckOpenTime.

The three bottom clusters partition primarily the failure cases. The dark green cluster (#7) again contains an outlier exhibiting extremely late SRVStuckOpenTime and FirewaterTime. These clusters correspond to the failure scenarios where all SRVs operate correctly for a long time, and the fire water is injected very late, not in time to avoid the core damage from overheating. The light green (#5) and pink (#6) clusters differ mostly in ExtendedECCSOperation and CoolingFailToRunTime. The light green (#5) cluster is concentrated with data points exhibiting lower ExtendedECCSOperation and higher CoolingFailToRunTime compared to the pink (#6) cluster. In this analysis, we demonstrate that differentiating clusters based on variations across different dimensions allows the user to organize and interpret the trends in scenario evolution and risk contributors for each scenario.

Failure scenarios case. Once again, we color the points in the PCA embedding for all failure scenarios, as illustrated in Fig. 8. There are clear variations among points in the embedding under ExtendedECCSOperation, FirewaterTime, and SRVStuckOpenTime. FirewaterTime and SRVstuckOpenTime vary along the horizontal direction, whereas ExtendedECCSOperation varies vertically. We also notice that very few points exist with a high SimulationEndTime among all the failure scenarios. Comparing this case with the *All scenarios case*, it is much more difficult to obtain insights from the original data based on this visualization alone.

Using clustering expansion, we arrive at a level of the hierarchy where five clusters are presented in the data (Fig. 9). In this focused analysis of all the failure scenarios (without the interference from the dominating dimension MaxCladTemp), we obtain various insights regarding the separation of clusters that can be used to identify the significant failure modes.

For example, the purple (#1) cluster contains an outlier with a late ACPowerRecoveryTime and CoolingFailToRunTime. Both the green (#2) and red (#3) clusters consist of early failure scenarios, but their reasons for failing early are evident in their corresponding parameter settings. In particular, the differentiating factors here are the CoolingFaillToRunTime and ExtendedECCSOperation. In the green cluster (#2), we see that the cooling system fails early and leads to an early set of failure cases; whereas in the red cluster (#3), the cooling system is available for longer, and instead the extended ECCS operation time is very short. Both conditions lead to similar rates of failure; thus loss of either system will yield similar performance.

3.2.2. *Topological clustering*

For topological clustering, we map the data into a 7D scalar function, where its input includes the seven input parameters of the simulation, and its output corresponds to MaxCladTemp for the *All scenarios case* and EndSimulationTime for the *Failure scenarios case*.

All scenarios case. We investigate several levels of the topological hierarchy before arriving at the clustering shown in Figs. 10 and 11. Beginning at the coarsest level, we continually refine the clustering looking for a stable persistence level, indicated by a wide red bar in the persistence chart (Fig. 10, bottom left), while avoiding over-segmentation involving small or uninformative clusters. A small cluster in the Morse–Smale complex approximation often indicates noise in the data and is typically considered unstable. If adding a new cluster does not significantly change the segmentation, such an addition is considered uninformative. In this example, we arrive at a level with four clusters.

In Fig. 10, three of the clusters share a common global maximum, whereas the remaining cyan cluster (#2) consists of points exhibiting low MaxCladTemp values, which correspond to success scenarios. Here we study the conditions that lead to distinct local minima, that is, the different parameter settings that yield stable success scenarios, by focusing on the behavior of the projected summary curves in the inverse coordinate plots of Fig. 10.

Recall the vertical axis of each inverse coordinate plot is labeled by one input parameter, and the horizontal axis corresponds to MaxCladTemp. Since we study conditions that lead to minimal values of MaxCladTemp, we focus on the left side of the horizontal axis of each plot, which corresponds to low values of MaxCladTemp.

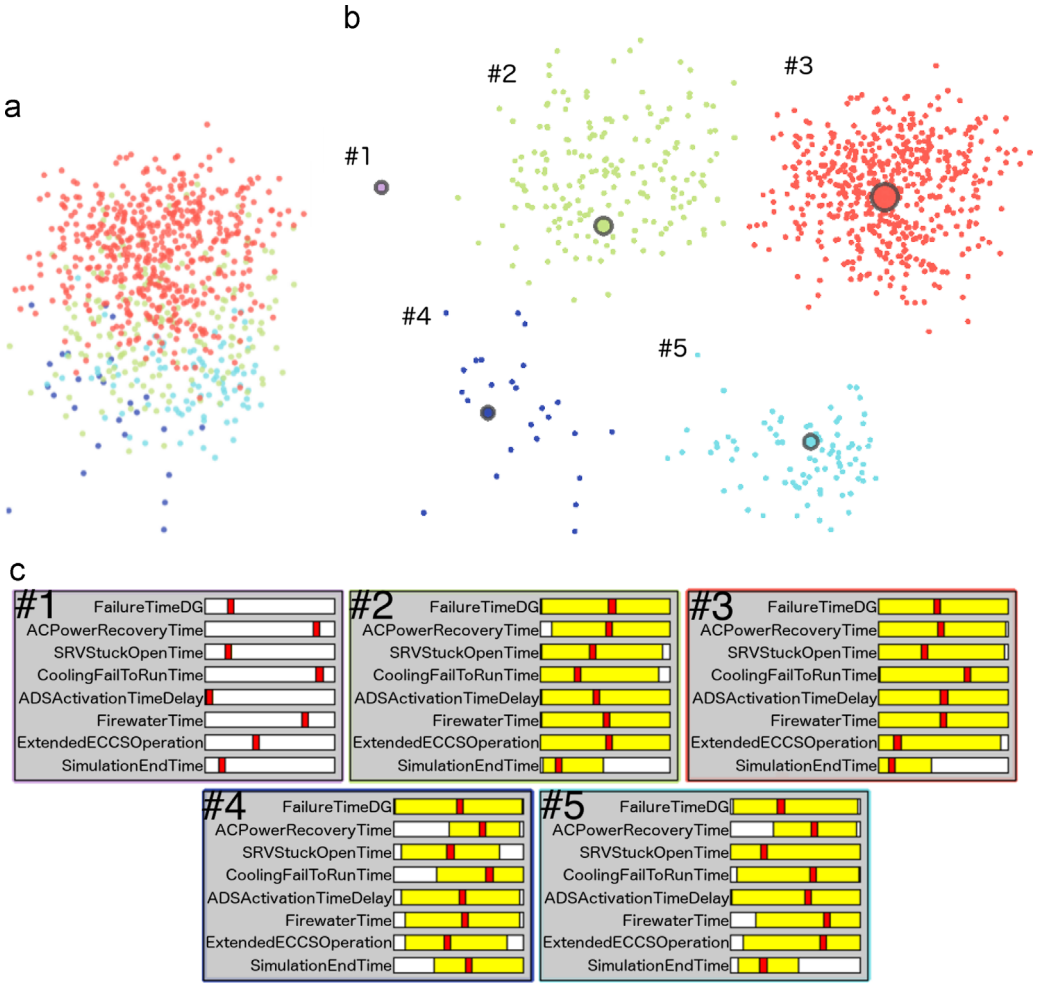

Fig. 9. (a) 2D embedding of the data colored by cluster labels. (b) A separate illustration of individual clusters and (c) their summary statistics. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

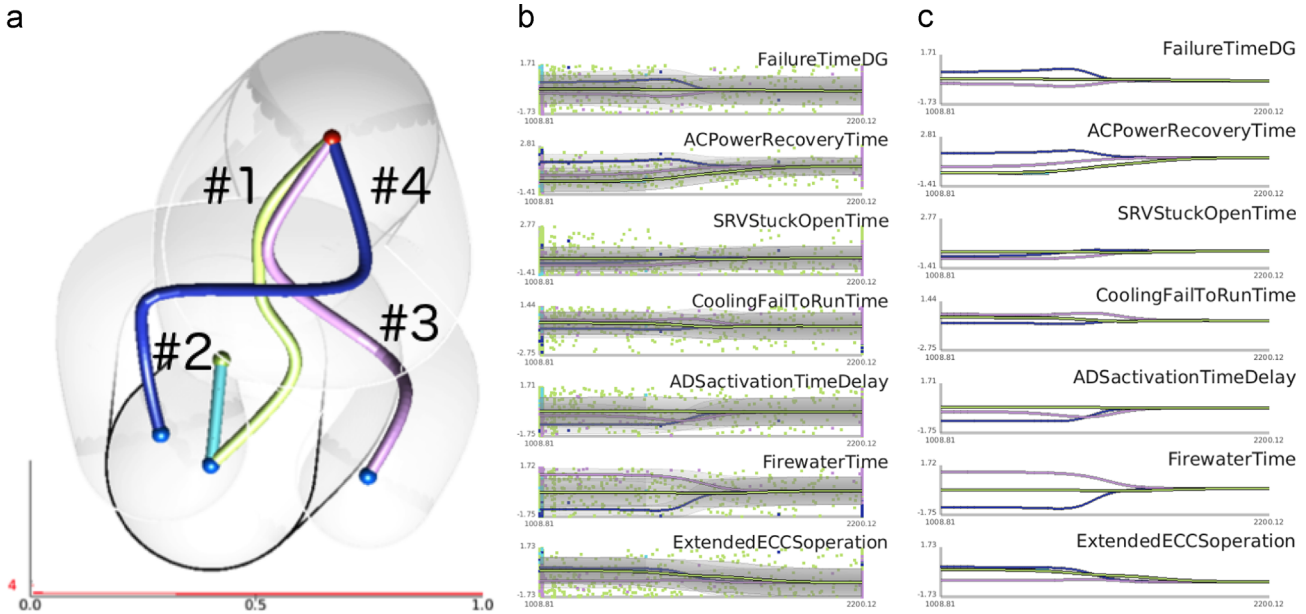

Fig. 10. (a) The topological skeleton of all 4997 scenarios. Inverse coordinate plots with (b) and without (c) points projected. Points and summary curves are colored by cluster labels. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

In Fig. 10 (right) the local minimum that belongs to the pink cluster (#3) exhibits a short ACPowerRecoveryTime, a long FirewaterTime, and a short ExtendedECCSOperation time. The local minimum of the blue cluster (#4), on the other hand, has a late ACPowerRecoveryTime, a very short FirewaterTime, a short ADSActivationTimeDelay, and a long ExtendedECCSOperation time. The third local minimum, shared by the green (#1) and cyan (#2) clusters, has a moderate FirewaterTime paired with a short ACPowerRecoveryTime and a long ExtendedECCSOperation time.

The input parameters that seem to be irrelevant in differentiating these clusters are the FailureTimeDG, the CoolingFailToRunTime, and the SRVStuckOpenTime. This last observation seems to be well aligned with the observations made in the beginning of Section 3.2.1, where we see no visual correlation between the MaxCladTemp and the FailureTimeDG (therefore we omitted the plot for FailureTimeDG in Fig. 6), and that the CoolingFailToRunTime and SRVStuckOpenTime are orthogonal in variation direction to the maxCladTemp in the PCA embeddings.

The new information we obtain from topological clustering is that the FirewaterTime does play a role in differentiating the pink (#3), green (#1), and blue (#4) clusters, as we see clear separation among the left end points of all three summary curves in the inverse coordinate plot (Fig. 10, right). Therefore, from a safety analysis perspective, we observe that, in order to assure a low value of maximum clad temperature, the high pressure injection system needs to be available for a long time for scenarios to remain system successes. On the other hand, the failure time of the diesel generators (FailureTimeDG, initial time of the SBO condition) does not play a relevant role in guaranteeing a low value of maximum cladding temperature.

For the pink cluster (#3) in (Fig. 10, right), an early AC recovery time guarantees system success even for early failures of two subsystems (low SRVstuckOpenTime and ExtendedECCSoperation) and late availability of the fire water (high FirewaterTime). This means, even in the case of an early RPV depressurization (i.e., SRV stuck open), the core heating rate is slow enough that an early AC recovery time guarantees low values of MaxCladTemp.

Failure scenarios case. In this case, we consider only failure scenarios and use SimulationEndTime, that is, the time to reach the failure temperature of 2200 °F(≈ 1477 K), as the output parameter. We obtain a topological clustering that consists of four clusters. Results are shown in Figs. 12 and 13.

In Fig. 12 (left) four clusters share a global minimum, characterized by a SimulationEndTime of 434.82 s. There are four distinct local maxima. One interpretation is to look at the local maxima as independent, near-success scenarios, as they represent, within their

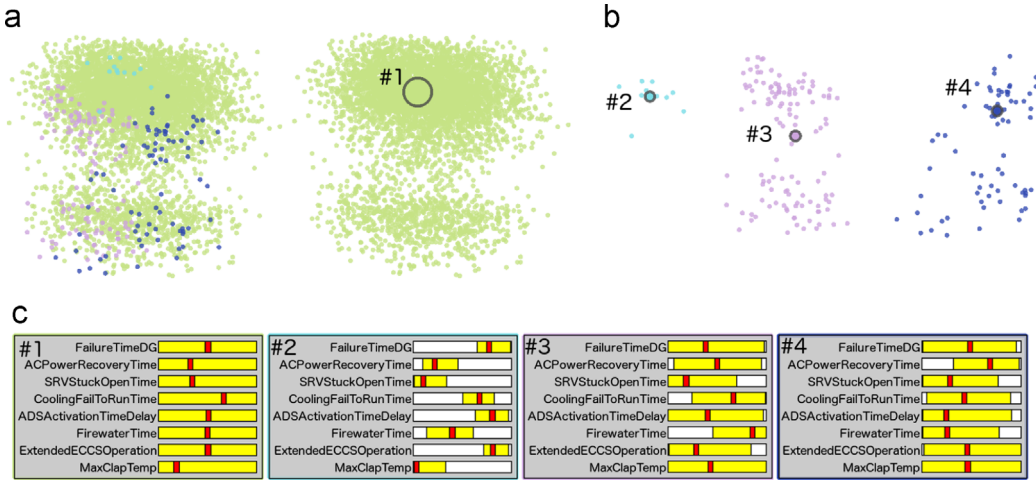

Fig. 11. (a) 2D embedding of the data colored by topological clustering labels. (b) A separate illustration of individual clusters and (c) their summary statistics with respect to the input dimensions.

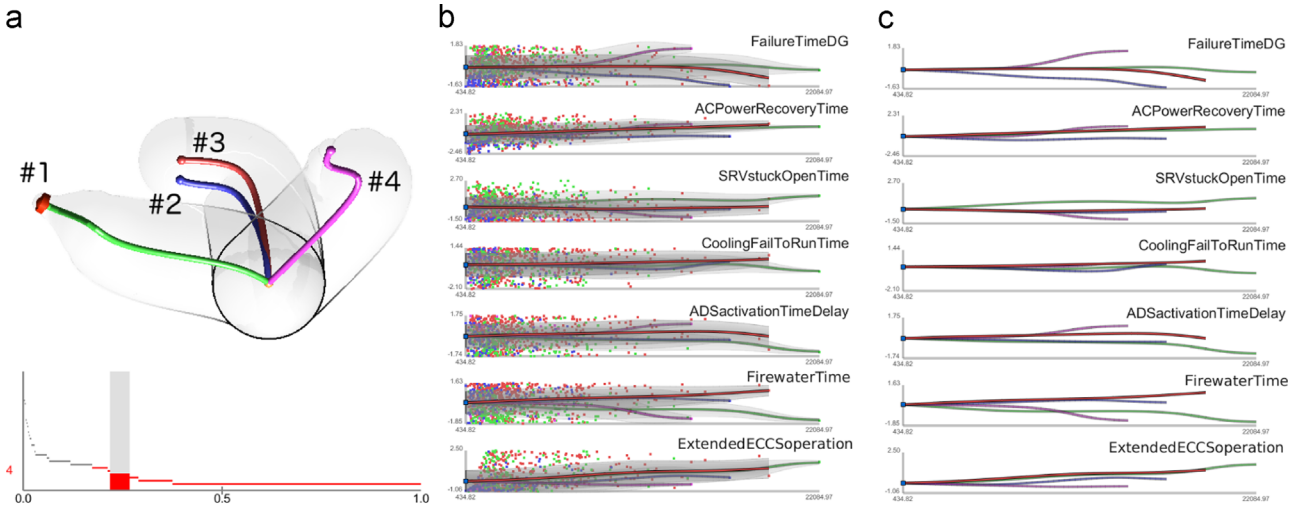

Fig. 12. (a) Topological skeleton of all failure scenarios. Inverse coordinate plots with (b) and without (c) points projected. Points and summary curves are colored by cluster labels.

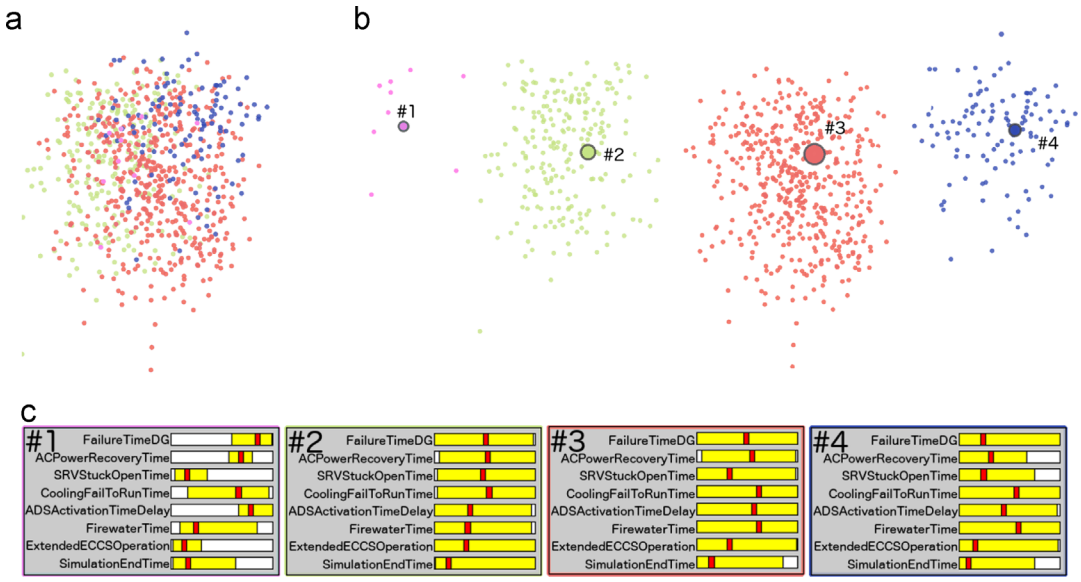

Fig. 13. (a) 2D embedding of the data colored by topological clustering labels. (b) A separate illustration of individual clusters and (c) their summary statistics.

own cluster, the latest time to reach the failure state (e.g., when the simulations terminate). In other words, the temperature for each of these local maxima scenarios grows slowly during the simulation, thereby allowing a longer simulation time.

From a safety analysis perspective, we are interested in understanding the conditions under which we have a late core damage event. Recall in the inverse coordinate plots of Fig. 12 (right) that the horizontal axis corresponds to the SimulationEndTime. Therefore, we focus our analysis on the right side of the horizontal axis, where a long simulation corresponds to a late core damage event.

For the green cluster (#1) in Fig. 12 (right), as expected, a driving factor to reach a late core damage is a high value of ECCS operation. This observation implies that it is preferable to keep the RPV pressurized as long as possible and maintain high pressure cooling, instead of activating the ADS system and obtaining cooling through the fire water system. Also note for this same cluster that a late core damage is also correlated with a late ACPowerRecoveryTime.

For all scenarios contained in the purple cluster (#4), we notice that the latest core damage within the cluster is reached for high values of FailureTimeDG, since a large quantity of heat has been discharged before reaching the SBO condition. On the contrary, for the red cluster (#3), the latest core damage within the cluster occurs when a small quantity of heat has been rejected from the core following reactor scram (i.e., low value of FailureTimeDG) and late failure of the high pressure core cooling system (i.e., high value of CoolingFailToRunTime).

In summary, for all clusters, a late failure of the high pressure core cooling system and a late ACPowerRecoveryTime are always needed in order to guarantee a late core damage condition. The latter should be an obvious observation as an early recovery of the AC power system will restore normal function before a large amount of heat is able to accumulate in the core. The more useful information is the importance of maintaining the high-pressure cooling system in order to delay core damage. In addition, FailureTimeDG when coupled with the FirewaterTime also plays a relevant role in understanding the conditions for reaching late core damage.

For comparison, as before, we color points in their 2D embedding based on the topological clustering results, as shown in Fig. 13. We are able to see how the clusters differ in terms of the statistical summaries of the input dimensions. However, the information regarding how the output parameter varies among the clusters remains hidden. For example in Fig. 13, ACPowerRecoveryTime varies in its range and mean value across the four clusters; however, the inverse coordinate plot in Fig. 12 reveals that such an input parameter is not a differentiating factor across the four clusters at the local maxima. As a matter of fact, the summary curves of this parameter overlap significantly in its inverse coordinate plot.

Table 1
The 10 input parameters from our simulation ensemble and their PDFs with associated parameters. For the SRV_soTime, the probability is p if SRV_soTime < ADS_actTime - DG_failTime; otherwise the probability is $1-p$.

Input name (units)	Range	Dist. type	Parameters
RCIC_failTime (h)	(0,8)	Exponential	$\lambda = 4.43{*}10^{-3}$
HPCI_failTime (h)	(0,8)	Exponential	$\lambda = 4.43{*}10^{-3}$
SRV_soTime (h)	(0,8)	Bernoulli	$p = 8.56{*}10^{-4}$
FW_availTime (m)	(0,480)	Lognormal	$\mu = 45, \sigma = 30$
DG_failTime (h)	(0,8)	Exponential	$\lambda = 1.09{*}10^{-3}$
DG_recTime (h)	(0,8)	Weibull	$\alpha = 0.745,\ \beta = 6.14$
PG_recTime (h)	(0,8)	Lognormal	$\mu = 0.793,\ \sigma = 1.982$
BATT_recTime (m)	(0,480)	Lognormal	$\mu = 45,\ \sigma = 15$
BATT_life (h)	(4,6)	Triangular	(4,5,6)
BATT_failTime (h)	(0,8)	Exponential	$\lambda = 3.5{*}10^{-6}$

4. Case study dataset 2: 10d simulation ensemble

4.1. Data description

A second dataset consisting of 10 000 station blackout simulation trials using a Monte Carlo sampling of ten input parameters has also been investigated. These input parameters are similar to the first dataset and are explained below:

1. *RCIC_failTime*: the time when the RCIC system fails to run.
2. *HPCI_failTime*: the time when the HPCI system fails to run.
3. *SRV_soTime*: the time when a Safety Relief Valve (SRV) gets stuck in the open position.

4. *FW_availTime*: the time when the fire water is available for injection into the RPV.
5. *DG_failTime*: the time when the diesel generators (DGs) stop providing power to the plant (i.e., the time when the SBO condition starts).
6. *DG_recTime*: the time when the power provided by the DGs is restored to the plant.
7. *PG_recTime*: the time when the AC power provided by the external power grid is restored to the plant.
8. *BATT_failTime*: the time when the battery system fails and must be repaired.
9. *BATT_recTime*: the time when the battery system is recovered.
10. *BATT_life*: the total uptime provided by the batteries before they become expended.

In addition, each parameter comes with a pre-defined probability density function (PDF), given in Table 1. This information can be used to compute the probability of occurrence for each simulation trial. We assume that all 10 parameters are independent of one another, and the probability associated with a given sample $\vec{x} = (x_1, \ldots, x_{10})$ is given by the equation below:

$$P(\vec{x}) = \prod_{i=1}^{10} p_i(x_i), \tag{1}$$

where p_i is the one-dimensional PDF associated with the i-th input parameter. Therefore, for this dataset, the output parameters of interest are:

1. *MaxCladTemp*: the maximum temperature attained anywhere on the cladding during the entire course of the simulation;
2. *OccurenceProb:* the probability of occurrence associated with each point in the domain, as computed from Eq. (1).

In this dataset, we set the clad failure temperature at 1800 °F(≈ 1255 K), and we have recorded 1243 (out of the 10 000) sampled trials correspond to failure scenarios. Unlike in the first dataset, the simulations are not terminated when they reach the threshold temperature; therefore, we see more variations in the range space for MaxCladTemp. The data is again pre-processed with a *Z*-score standardization.

4.2. Results

We now apply traditional clustering to the above dataset followed by topological clustering. Recall the failure region is defined as all parameter settings in the input domain whose corresponding clad temperature reached or exceeded 1800 °F(≈ 1255 K). We identify the failure region of the input domain and further analyze this region in detail. In particular, we study the topology of the probability landscape over the failure region (i.e., *Failure scenarios case*). That is, we construct a 10D scalar function based on the 10 simulation input parameters based on the failure scenarios and use occurenceProb as its scalar output. We aim to characterize the failure region according to areas of high probabilities, whereupon further efforts could be made to reduce the risks associated with these areas.

4.2.1. Traditional clustering

We map the data into a 11D space by considering the 10 input parameters and the output parameter MaxCladTemp. Similar to Section 3.2.1, we perform agglomerative hierarchical clustering using average linkage on this 11D data and then apply PCA to project the points into a 2D domain.

All scenarios case. We illustrate the results for the hierarchical clustering of the dataset into 14 clusters in Fig. 14. We show PCA projections of the data points colored by both cluster labels (Fig. 14, top left) as well as their success/failure conditions (Fig. 14, top right). From a safety perspective, the interesting cases occur near the failure region of the input domain, namely, the regions that contain failure or near-failure cases. Therefore, we remove the clusters that contain only success scenarios, and focus our statistical analysis on the remaining eight clusters (Fig. 14, bottom), for which we analyze the mean and range of each input parameter.

In Fig. 14 (bottom), two data points (#4 and #8, respectively), both corresponding to failure scenarios, exist as their own clusters. The purple outlier (#8) exhibits late failure times for the RCIC, HPCI, DG, and battery systems (i.e., high values of RCIC_failtime, HPCI_failTime, DG_failTime and BATT_failTime), as well as late recovery times for DG and PG systems (i.e., high DG_recTime and PG_recTime), leading to overheating of the cladding due to a prolonged exposure to the heat in the system as there is not sufficient time for recovery. The red outlier (#4) characterizes a scenario with early failures of the HPCI, SRV, DG, and battery systems (i.e., low values of HPCI_failTime, SRV_soTime, DG_failTime, and BATT_failTime) as well as a short battery life even with a fast recovery of the battery system (i.e., low BATT_life and BATT_recTime). Even though such a scenario has access to the fire water early, loss of the battery system impedes adequate cooling of the core. In addition, an early SRV failure allows an RPV depressurization but not fast enough to be able to use the fire water injection before the maximum temperature of the cladding reaches its threshold. Further analysis of these two scenarios could be conducted to verify these hypotheses.

The next smallest cluster in light green (#7) consists exclusively of failure scenarios. In these cases, the fire water is available early; however, this set of cases exhibit early failures of other subsystems: RCIC, HPCI, DG and battery systems, thus impeding an adequate rate of heat removal. Among the larger clusters, the blue cluster (#6) consists mainly of failure scenarios most likely due to the late recovery times of both the DG and PG systems. The cyan cluster (#5) also consists mainly of failure cases. Analysis of the mean values of all input parameters shows mostly moderate values except for a late recovery time of the PG system and an early failure time of the battery system. Meanwhile, the brown (#1), orange (#2), and dark green (#3) clusters contain mainly success scenarios, where the failure scenarios within these clusters typically have low MaxCladTemp, making them less interesting for further analysis, but they can be used to contrast the behaviors of the mostly or exclusively failure scenarios.

4.2.2. Topological clustering

We map the data into a 10D scalar function, where its input includes the 10 input parameters of the simulation, and its output corresponds to MaxCladTemp for the case where all scenarios are considered and OccurenceProb for the case where only the failure scenarios are considered.

All scenarios case. At an appropriately chosen scale, topological clustering of the data results in a clustering consisting of three clusters whose topology is characterized by a shared global minimum and three distinct local maxima within its topological skeleton (Fig. 15, left). As illustrated in Fig. 15 (middle), the data points are sampled at varying densities within the range space. That is, relatively dense samples are obtained within the range [750 °F, 1000 °F](≈ [672 K, 811 K]) of the MaxCladTemp (which corresponds to a large number of success scenarios that have been safely recovered) and within the failure region, that is, on or above 1800 °F(≈ 1255 K). Data points within the green cluster (#1) represent the smallest span of the range space, between 585 °F(≈ 580 K) and 2378 °F(≈ 1576 K).

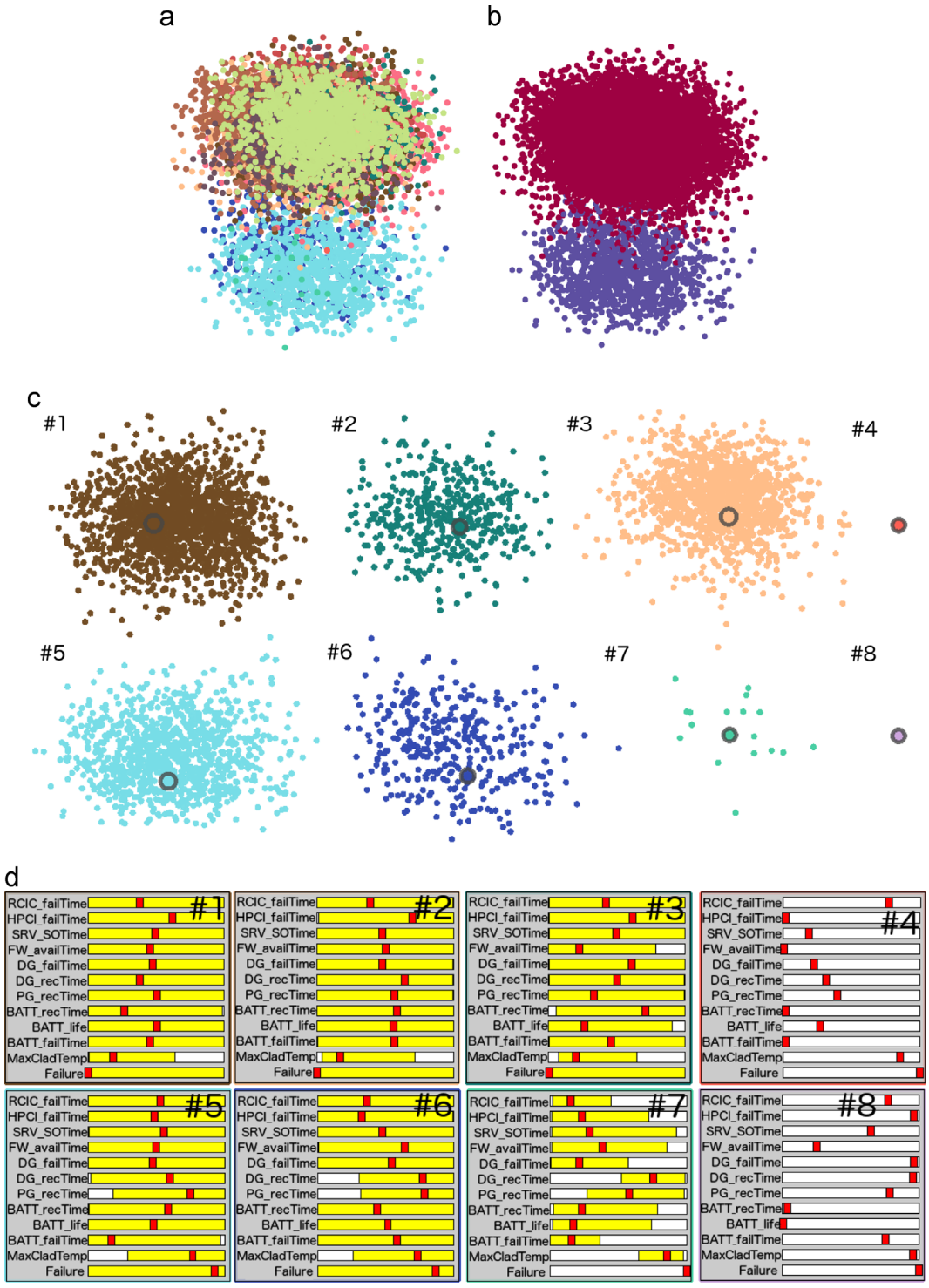

Fig. 14. Results of traditional clustering of the 11D SBO data for all scenarios. The top row shows the PCA projections of the 11D point cloud, colored by cluster labels (a) and success (red) or failure (blue) conditions (b), respectively. (c) The subset of clusters containing failure scenarios. (d) Detailed statistical analysis of the clusters containing failure scenarios. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Within the failure region, the blue (#2) and green (#1) clusters combined account for less than 1% of the observed failure scenarios, whereas the red cluster (#3) contains the majority of the failure scenarios. Two input parameters stand out in the inverse coordinate plot. As shown in Fig. 15 (middle), an early PG_recTime is the most likely parameter setting to avoid reaching failure conditions, as evidenced by an area with low sample density within the failure region. Meanwhile, a large number of failure cases share an early BATT_failTime, as witnessed by an area with high sample density within the failure region.

We focus our visual sensitivity analysis surrounding the failure region to understand how different input parameters influence the observed output parameter, MaxCladTemp, by further exploration of the inverse coordinate plots highlighting the summary curves in Fig. 15 (right). Within the failure region in Fig. 15 (right), the defining characteristics of the green cluster (#1) are its distinctly

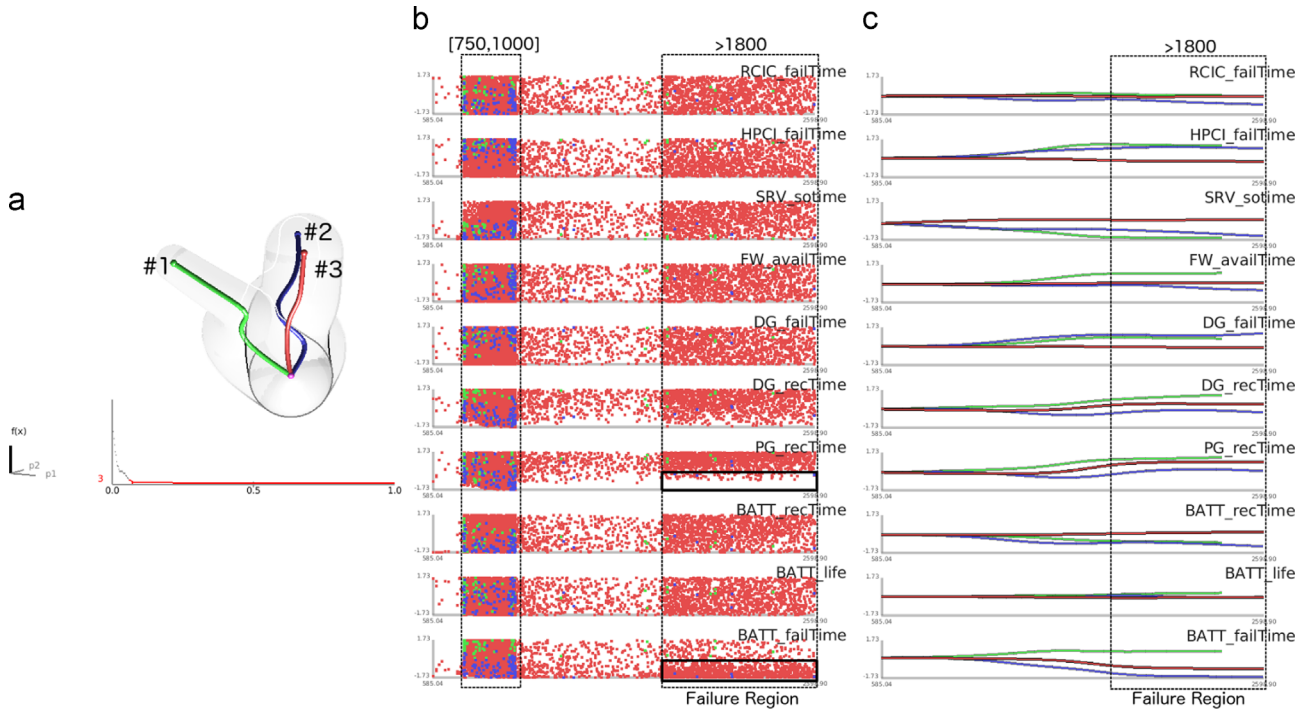

Fig. 15. (a) The topological skeleton of all scenarios. (b) Inverse coordinate plots highlighting the point samples colored by cluster labels. (c) Inverse coordinate plots showing only the summary curves associated with each cluster. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

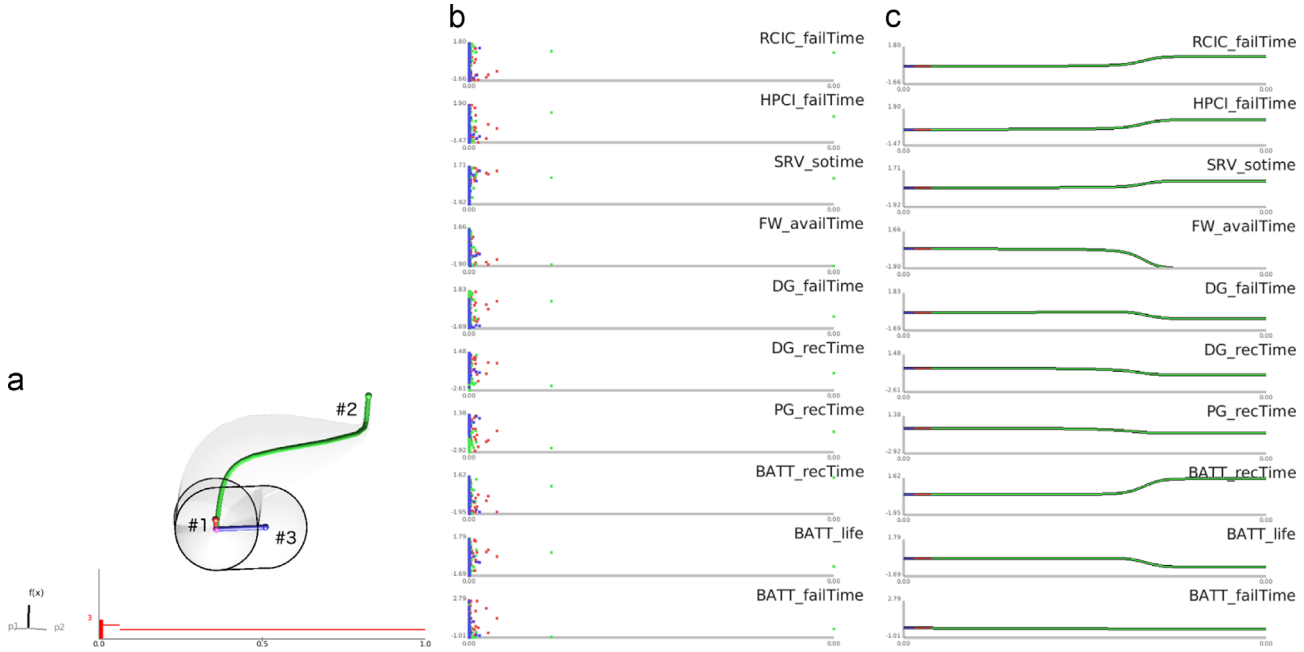

Fig. 16. (a) The topological skeleton of the failure scenarios. (b) Inverse coordinate plots that highlight the point samples colored by cluster labels. (c) Inverse coordinate plots that highlight the summary curves associated with each cluster. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

late FW_availTime, late DG_recTime, and late BATT_failTime. The blue cluster (#2) shares several similar behaviors with the green cluster (#1) within the failure region, namely, a late HPCI_failTime, an early SRV_soTime, a late DG_failTime, and an early BATT_recTime. However, it differentiates itself from the green cluster (#1) by having an early RCIC_failTime, an early FW_availTime, an early PG_recTime, an early DG_recTime, and an early BATT_failTime. The DG_recTime and BATT_failTime are the most relevant input parameters that distinguish all three clusters in the failure region.

In this example, we are able to gain insight by evaluating the sampling density within various projections, as well as by observing the summary trend information given by the inverse regression plots. Decomposing the data by topological clustering allows us to separate the different trends occurring locally within the high-dimensional space and to compare and contrast them with one another.

Failure scenarios case. We focus on studying areas within the failure region that have a high probability of occurrence (i.e., high values of

occurenceProb). Based on a topological clustering Fig. 16 (left), we obtain three clusters that have a shared global minimum and three distinct local maxima valued at $7.39 \times 10^{-5}, 2.79 \times 10^{-5}$, and 9.75×10^{-4} for the red (#1), blue (#3), and green (#2) clusters, respectively. Fig. 16 (middle) illustrates a very sparse sampling within the range space as most samples are concentrated towards the low probability regions. The green cluster (#2) contains the most interesting failure scenario, that is, the global maximum, which corresponds to the data point with the highest probability of occurrence. Such a global maximum corresponds to a FW_availTime valued at 22.9 s (near its lower bound of 0, see Table 1) and a BATT_recTime valued at 2.82×10^4 s($\approx$ 470 m, near its upper bound of 480 m; see Table 1). Further sampling of the input parameter space surrounding such a global maxima could reveal more structures associated with the failure region in highly probable areas. In addition, sampling could be extended towards regions surrounding minor local maxima to identify yet unwitnessed, distinct, high probable areas of the failure region.

5. Conclusion

We apply both traditional and topological clusterings in conjunction with dimensionality reduction techniques on DPRA datasets. We provide the domain scientist with an analysis and visualization tool for obtaining insights with respect to system responses under the simulated accident scenarios. We focus on two datasets simulating the response of a BWR system during an SBO accident scenario. We obtain such datasets by performing a series of simulations where, for each simulation run, we randomly change timing and sequencing of a specified set of events. We aim to identify how timing or sequencing of these events affects the maximum core temperature.

Clustering is a powerful tool that can be used to summarize large amounts of scenarios into digestible pieces for effective analysis and visualization. As the two clustering algorithms considered take very different approaches, they offer different insights regarding the data. We have observed that a traditional clustering combined with dimensionality reduction is adequate to distinguish failure scenarios and success scenarios and to group points with similar parameter settings. On the other hand, topological clustering captures information regarding how input parameters are correlated with the output and how input parameter settings help differentiate local extrema of the output. Topological clustering takes the dependencies among the input and output parameters into consideration and performs global analysis that highlights topological structures encoded within these dependencies. In addition, topological clustering leads to novel visualizations. We believe that pairwise comparisons and validations of both types of clustering techniques complement each other in enhancing structural understanding of the data.

Acknowledgments

This work was performed in part under the auspices of the US DOE by LLNL under Contract DE-AC52-07NA27344, LLNL-CONF-658933. This work is also supported in part by NSF 0904631, DE-EE0004449, DE-NA0002375, DE-SC0007446, DE-SC0010498, NSG IIS-1045032, NSF EFT ACI-0906379, DOE/NEUP 120341, and DOE/Codesign P01180734.

References

[1] Carreira-Perpinan Miguel A. A review of dimension reduction techniques. Department of Computer Science. University of Sheffield. Technical report CS-96-09; 1997. p. 1–69.
[2] Defays Daniel. An efficient algorithm for a complete link method. Comput J 1977;20(4):364–6.
[3] Edelsbrunner Herbert, Harer John, Natarajan Vijay, Pascucci Valerio. Morse–Smale complexes for piecewise linear 3-manifolds. In: Proceedings 19th ACM symposium on computational geometry; 2003. p. 361–70.
[4] Edelsbrunner Herbert, Harer John, Zomorodian Afra J. Hierarchical Morse–Smale complexes for piecewise linear 2-manifolds. Discret Comput Geom 2003;30:87–107.
[5] Edelsbrunner Herbert, Letscher David, Zomorodian Afra J. Topological persistence and simplification. Discret Comput Geom 2002;28:511–33.
[6] Fukunaga K, Hostetler L. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inf Theory 1975;21(1):32–40.
[7] Gerber Samuel, Bremer Peer-Timo, Pascucci Valerio, Whitaker Ross. Visual exploration of high dimensional scalar functions. IEEE Trans Visualization Comput Graph 2010;16:1271–80.
[8] Gerber Samuel, Rübel Oliver, Bremer Peer-Timo, Pascucci Valerio, Whitaker Ross T. Morse–Smale regression. J Comput Graph Stat 2013;22(1):193–214.
[9] Jolliffe IT. Principal component analysis. Springer-Verlag, New York; 1986.
[10] Kruskal Joseph B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964;29(1):1–27.
[11] Di Maio F, Stasi M, Zio E, Mandelli D, Aldemir T. Identification of faults in a level control dynamic system. In: Proceedings of NPIC-HMIT 2009, Knoxville, TN; 2009.
[12] Maljovec Dan, Liu Shusen, Wang Bei, Pascucci Valerio, Bremer Peer-Timo, Mandelli Diego, et al. Analyzing simulation-based PRA data through clustering: a BWR station blackout case study. In: Probabilistic safety assessment & management conference; 2014.
[13] Maljovec Dan, Wang Bei, Mandelli Diego, Bremer Peer-Timo, Pascucci Valerio. Analyze dynamic probabilistic risk assessment data through topology-based clustering. International topical meeting on probabilistic safety assessment and analysis (PSA); 2013.
[14] Maljovec Dan, Wang Bei, Pascucci Valerio, Bremer Peer-Timo, Pernice Michael, Mandelli Diego, et al. Exploration of high-dimensional scalar function for nuclear reactor safety analysis and visualization. In: International conference on mathematics and computational methods applied to nuclear science & engineering; 2013.
[15] Mandelli D, Smith C, Rabiti C, Alfonsi A, Youngblood R, Pascucci V, et al. Dynamic PRA an overview of new algorithms to generate, analyze and visualize data. ANS winter meeting; 2013.
[16] Mandelli D, Smith C, Riley T, Nielsen J, Schroeder J, Rabiti C, et al. Overview of new tools to perform safety analysis: BWR station black out test case. In: Probabilistic safety assessment & management conference; 2014.
[17] Mandelli Diego, Smith Curtis, Riley Thomas, Schroeder John, Rabiti Cristian, Alfonsi Aldrea, et al. Support and modeling for the boiling water reactor station black out case study using RELAP and RAVEN. Technical Report INL EXT-13-30203. Idaho National Laboratory (INL); 2013.
[18] Mandelli Diego, Yilmaz Alper, Aldemir Tunc, Metzroth Kyle, Denning Richard. Scenario clustering and dynamic probabilistic risk assessment. Reliab Eng Syst Saf 2013;115:146–60.
[19] Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290:2319–23.
[20] Zio E. Reliability engineering: old problems and new challenges. Reliab Eng Syst Saf 2009;94(2):125–41.
[21] Zio E, Maio FD. Processing dynamic scenarios from a reliability analysis of a nuclear power plant digital instrumentation and control system. Ann Nucl Energy 2009;36:1386–99.