Analyzing Simulation-Based PRA Data Through Clustering: a BWR Station Blackout Case Study

Dan Maljovec[a*], Shusen Liu[a], Bei Wang[a], Valerio Pascucci[a], Peer-Timo Bremer[b], Diego Mandelli[c], and Curtis Smith[c]

[a]SCI Institute, University of Utah, Salt Lake City, USA
[b]Lawrence Livermore National Laboratory, Livermore, USA
[c]Idaho National Laboratory, Idaho Falls, USA

Abstract: Dynamic probabilistic risk assessment (DPRA) methodologies couple system simulator codes (e.g., RELAP, MELCOR) with simulation controller codes (e.g., RAVEN, ADAPT). Whereas system simulator codes accurately model system dynamics deterministically, simulation controller codes introduce both deterministic (e.g., system control logic, operating procedures) and stochastic (e.g., component failures, parameter uncertainties) elements into the simulation. Typically, a DPRA is performed by 1) sampling values of a set of parameters from the uncertainty space of interest (using the simulation controller codes), and 2) simulating the system behavior for that specific set of parameter values (using the system simulator codes). For complex systems, one of the major challenges in using DPRA methodologies is to analyze the large amount of information (i.e., large number of scenarios) generated, where clustering techniques are typically employed to allow users to better organize and interpret the data. In this paper, we focus on the analysis of a nuclear simulation dataset that is part of the risk-informed safety margin characterization (RISMC) boiling water reactor (BWR) station blackout (SBO) case study. We apply a software tool that provides the domain experts with an interactive analysis and visualization environment for understanding the structures of such high-dimensional nuclear simulation datasets. Our tool encodes traditional and topology-based clustering techniques, where the latter partitions the data points into clusters based on their uniform gradient flow behavior. We demonstrate through our case study that both types of clustering techniques complement each other in bringing enhanced structural understanding of the data.

Keywords: PRA, computational topology, clustering, high-dimensional analysis

1. INTRODUCTION

A recent trend in the nuclear engineering field is the implementation of computationally-intensive codes for the design and safety analysis of nuclear power plants. In particular, the new generation of system analysis codes aims to embrace phenomena such as thermo-hydraulic, structural behavior, system dynamics, human behavior, as well as uncertainty quantification and sensitivity analysis associated with these phenomena. The use of dynamic probabilistic risk assessment (DPRA) methodologies allows a systematic approach to uncertainty quantification.

DPRA methodologies account for possible coupling between triggered or stochastic events through explicit consideration of the time element in system evolution, often through the use of dynamic system simulators. Such methodologies are commonly needed when the system has multiple failure modes, control loops, processes, software/hardware components, or human interactions. A DPRA is typically performed by 1) sampling values of a set of parameters from the uncertainty space of interest (using the simulation controller codes), and 2) simulating the system behavior for that specific set of parameter values (using the system simulator codes).

[*] Corresponding author, maljovec@cs.utah.edu

Due to the intrinsically high level of detail within such a process, one would need to handle large amounts of data generated within the simulation [9]. In [7] we have presented a framework that visualizes high-dimensional scalar functions through a topological segmentation of its input surfaces. The input of such a high-dimensional scalar function arises from the set of n uncertain parameters $x_1, x_2, ..., x_n$, whereas the output originates from some safety-related outcomes, such as 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 further explore the topology-based clusterings that lie beneath such a framework for DPRA datasets [6].

In this paper, we focus on the analysis of a particular nuclear simulation dataset based upon our previously developed analysis and visualization framework [6, 7]. The dataset is part of the risk-informed safety margin characterization (RISMC) boiling water reactor (BWR) station blackout (SBO) case study [8]. We enrich our tool by combining traditional and topology-based (hierarchical) clustering, as well as dimensionality reduction (DR) techniques. We demonstrate through our case study that both types of clustering techniques complement each other in bringing enhanced structural understanding of the data. In particular, the topology-based clustering helps highlight key features of the data that are otherwise hidden using the traditional techniques.

BWR system. The system considered in our case study is a generic BWR power plant with Mark I containment. The three main structures are: 1) the reactor pressure vessel (RPV), a pressurized vessel that contains the reactor core; 2) the primary containment including the dry well (DW) that houses the RPV and circulation pumps; and 3) the pressure suppression pool (PSP), also known as the wet well. The PSP is a large torus-shaped container that contains a large amount of water (almost 1 M gallons of fresh water) and is used in specific situations as an ultimate heat sink. The original BWR Mark I includes a large number of systems, but for the scope of this report and for the case study considered, we use a smaller subset of systems that includes the RPV level control systems, the RPV pressure control systems, the cooling water inventory, and the AC power system, which consists of two power grids, emergency diesel generators (DGs), and battery systems for the instrumentation and control systems.

The RPV level control systems provide manual and automatic control of the water level within the RPV and consist 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 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 is similar, but allows much greater water flow rates. The RPV pressure control systems provide manual and automatic control of the RPV internal pressure and consist of a set of safety relief valves (SRV), 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 condensate storage tank (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 analysis considered is a **BWR Mk. I** system during a loss of offsite power (LOOP) event followed by loss of the diesel generators (DGs), i.e., station blackout (SBO). In more detail, at time $t = 0$, LOOP condition occurs due to an external event (i.e., the offsite power lines are damaged). The LOOP alarm triggers the following events: a successful scram of the reactor is performed by the operators; MSIVs are successfully closed, isolating the primary containment from the turbine building; emergency DGs successfully start keeping the AC power busses energized; DC systems (i.e., batteries) are functional; and the decay heat generated by the core is removed from the pressure vessel through the residual heat removal system.

Next, the SBO condition occurs due to internal failure, which results in the failure of the DGs. As a result of the loss of external power, removal of decay heat is impeded. Reactor operators start the SBO emergency procedures and perform the following: RPV level control using RCIC or HPCI, RPV pressure control using SRVs, and containment monitoring (both dry well and PSP). At this point, plant staff start recovery operations to bring back on-line the DGs while the recovery of the off-site power grid is underway as well. Due to heavy usage, battery power can be depleted. When this happens, all remaining control systems are off-line, causing the reactor core to heat until the maximum temperature limit for the clad is reached: 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; RPV water level becomes too low), then the reactor operators activate the ADS in order to depressurize the RPV and allow fire water injection, if available. When AC power is recovered, through successful restart/repair of DGs or off-site power, RHR can be employed to keep the reactor core cool.

2. BACKGROUND

Dimensionality reduction (DR) and traditional hierarchical clustering are widely used techniques for analyzing structures of high-dimensional data. To extend the existing framework we have developed in [6, 7], we employ a visualization system that utilizes both dimensionality reduction and clustering, where dimensionality reduction constructs a mapping for the clustering results for intuitive visual analysis. We begin with a brief description of DR and traditional hierarchical clustering and then focus on the topology-based clustering, which may be unfamiliar to nonspecialists.

Dimensionality reduction. DR techniques [1], such as principal component analysis (PCA), multidimensional scaling (MDS), and Isomap, 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, in the analysis due to its simplicity and computational efficiency. However, using DR alone as a black box solution in the analysis suffers a major limitation. That is, the results of DR could be hard to interpret as a certain amount of structural information has been lost during the DR process. Therefore, we try to impose structural context onto the embeddings by combining DR results with a clustering obtained from the original high-dimensional data.

Traditional hierarchical clustering. 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 for defining what constitutes a cluster, which are based on density, distribution, distance, connectivity, etc. 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 connectivities 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 clusters with nearby clusters. In our system, we do not need to specify the number of clusters we are looking for; instead we interactively expand or collapse different levels of clustering in the hierarchy during the analysis.

Approximated Morse-Smale complex and topology-based hierarchical clustering. We consider an alternative method for clustering high-dimensional data based on the concept of the Morse-Smale complex. We give a brief overview of these concepts. See [6, 7] for details. The Morse-Smale complex is a type of topological structure that serves as a structural summary of a given high-dimensional scalar function. We consider a scalar function $f : \mathbb{X} \to \mathbb{R}$ defined over a finite set of points $\mathbb{X}$ in $\mathbb{R}^n$. The approximated Morse-Smale complex, at its finest level, partitions the points in $\mathbb{X}$ based on their uniform gradient

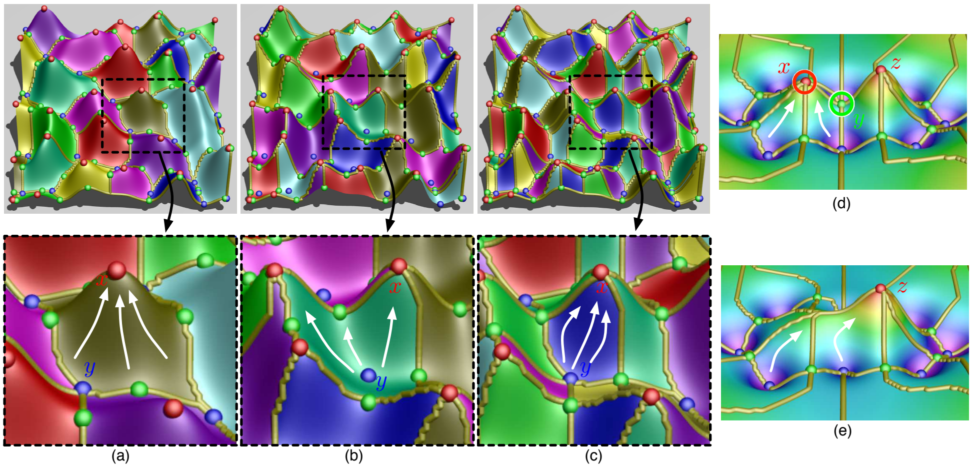


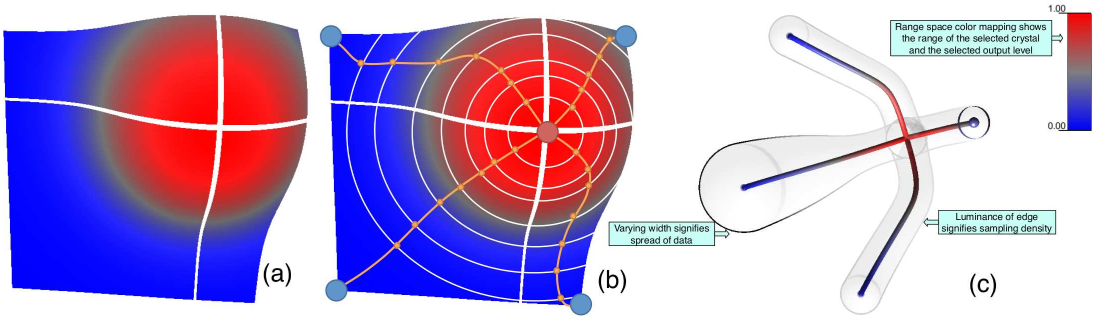
Figure 1: For a height function defined on a 2D domain (where maxima, minima, and saddles are colored red, blue, and green, respectively): (a) For each point in the brown region, the gradient flow (white arrow) ends at the same maxima *x*; (b) For each point in the green region, the gradient flow starts at the same minimum *y*; (c) For each point in the blue (Morse-Smale) cluster, 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 maximum *x* is considered less important topologically than its nearby peak at 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*. Figure reproduced from [6].

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 topology-based clustering at the finest level for a height function defined on a 2D domain is illustrated in Figure 1(a)-(b). We can then merge clusters based on persistence simplification [3], where less (topologically) significant clusters are merged into more significant ones. We avoid the technical details here but simply illustrate such a process in Figure 1(d)-(e).

Topological skeleton obtained through DR. Given a topology-based 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 [4]: 1) perform inverse 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 [5], 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 information, as illustrated in Figure 2. 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. Suppose we use a point sampling of the same 2D height function in Figure 2. The above process is illustrated in Figure 3. For more details of the visualization pipeline as well as additional views, see [4, 6, 7].

3. CASE STUDY DATASET

An ensemble of 4997 transient simulations has been generated using classical Monte-Carlo sampling of seven input parameters, among which 833 scenarios resulted in system failure (the core temperature



Range space color mapping shows the range of the selected crystal and the selected output level

Varying width signifies spread of data

Luminance of edge signifies sampling density

Figure 2: 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.

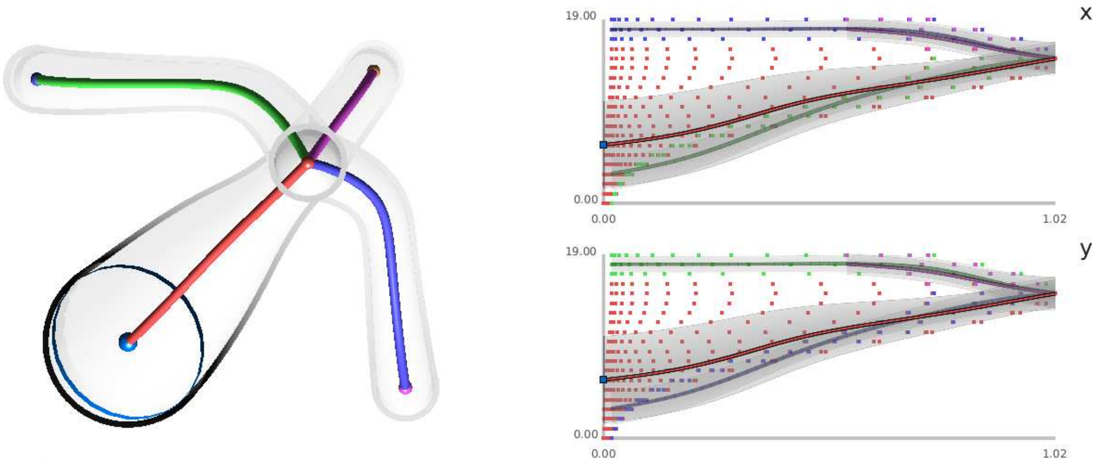


Figure 3: Left: topological skeleton colored by cluster labels. Right: inverse coordinate plots. Data points are visualized by their cluster labels, and summary curves are projected. For the inverse coordinate plots, the horizontal axis represents the output dimension (e.g., height values), and each 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.

reached the clad failure temperature threshold of 2200 F), whereas the rest of the 4164 scenarios ended up in system success (AC power is recovered or the firewater becomes available if the RPV is depressurized). 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 exhausted, a safety relief valve gets stuck open, etc.). The seven input parameters are listed below, as they are the only uncertain parameters under consideration.

- **FailureTimeDG**: Failure time of the DGs corresponding to the time of the SBO event.
- **ACPowerRecoveryTime**: $\min\{\text{Recovery time of DGs}, \text{off-site power recovery time}\}$. The minimum of these two times will determine when the simulation is considered recovered.
- **SRVStuckOpenTime**: The time when an SRV is stuck in the open position.
- **CoolingFailtoRunTime**: $\max\{\text{HPCI failure time}, \text{RCIC failure time}\}$. As long as one of these systems is functioning, the reactor is being actively cooled, so it is important to understand when

both systems have failed. Thus, we take the maximum of these two times.

- **ADSactivationTimeDelay**: The time when the operator manually depressurizes the RPV by activating the ADS system. This parameter measures the time delay from the HCTL event, not the time from 0 to when ADS is activated.
- **firewaterTime**: As an emergency action, when RPV pressure is below 150 psi, plant staff can connect the fire water system to the RPV to cool the core and maintain an adequate water level.
- **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 of 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 of the HCTL event. The output variables obtained from the simulations are: 1) **maxCladTemp**, which is the maximum clad temperature reached during the entire course of the simulation; and 2) **simulationEndTime**, which for failure cases represents the time to reach the failure temperature of 2200 F. 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, the domain scientists are interested in what combination of conditions (in the form of input simulation parameters) can cause potential reactor failure (i.e., nuclear meltdown witnessed by maximum core temperature exceeding a threshold value).

4. RESULTS

We provide analysis under both traditional (Section 4.1) and topological hierarchical clustering (Section 4.2) 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 as the observed output variable. Note that in this case, all failure cases have the same output variable of 2200 F. In the second case, referred to as the **Failure Scenarios Case**, we focus on clustering of the 833 failure scenarios. Since the maximum clad temperature does not vary for these cases, we treat the end simulation time as the output variable. We give a comprehensive picture by providing comparisons among the two clustering techniques and discuss the benefits and limitations inherent in each approach.

4.1. Traditional Clustering

For traditional clustering, we map the data into eight dimensional space by considering the seven input variables and the output variable, maximum clad temperature. We start our analysis by applying PCA to reduce the eight dimensional data to its two dimensional embedding for direct visual analysis.

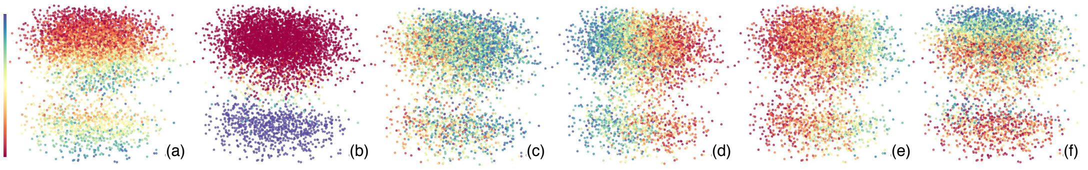


Figure 4: PCA embedding for the 8D dataset under the **All Scenarios Case**. The dimensions shown exhibit relatively strong correlation patterns within the embedding. We use a spectral colormap where red/blue represents low/high value. (a) ACPowerRecoveryTime; (b) maxCladTemp; (c) CoolingFailToRunTime; (d) firewaterTime; (e) SRVStuckOpenTime; (f) ExtendedECCSOperation.

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 Figure 4. All the dimensions shown exhibit a certain amount of visual correlation within the embedding, except for the omitted dimensions, ADSActivationTimeDelay and FailureTimeDG. 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 due to the fact PCA is inherently optimized for capturing dominant directions of maximum variance. In Figure 4(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 Figure 4(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 Figure 4(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 Figure 4(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.

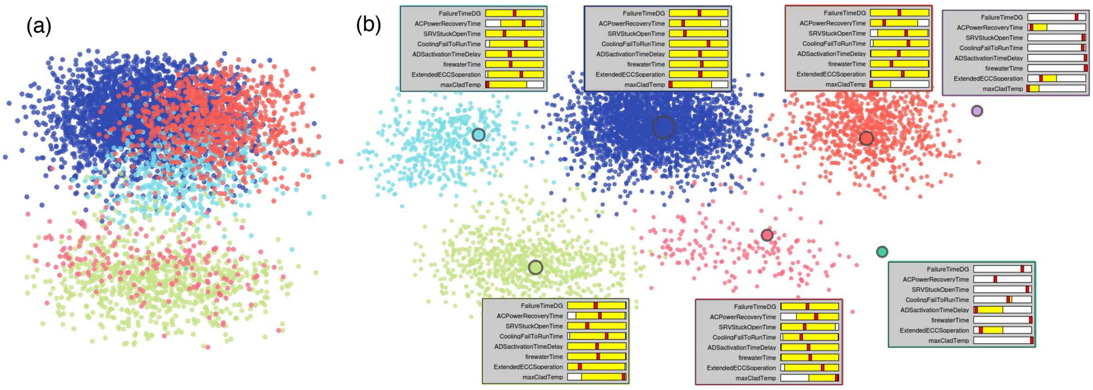


Figure 5: (a) 2D embedding of the data colored by cluster labels. (b) In order to provide a more clear view for the clusters,we provide a separate illustration of each individual cluster and its summary statistics.

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 (enclosed in a box next to the clustered points). 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. With these summaries across all clusters, we can quickly compare and investigate the defining characteristics of each cluster at a glance.

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, where the upper cluster contains exclusively success scenarios, and the lower cluster contains all failure scenarios and a few number of successes (via validations by known labels of success/failure). 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 Figure 5. Four of the top clusters decompose all of the success scenarios (top half of the embedding). The extremely small purple cluster likely consists of outliers in the data, since its points share extremely low ACPowerRecoverTime and maxCladTemp. These points correspond to the success scenarios where AC power is recovered very quickly and clad temperature never increases drastically. Although the blue and cyan clusters share similar statistical summaries across most dimensions, ACPowerRecoveryTime seems to be the most likely factor that differentiate these two clusters. The differentiating factor between the orange cluster and the other three clusters is its late SRVStuckOpenTime. On the other hand, three of the bottom clusters partition primarily the failure cases. The dark green cluster again contains the outliers and its points share extremely late SRVStuckOpenTime and firewaterTime. These correspond to the failure scenarios where all SRVs operate correctly for a long time and the firewater is injected very late, not in time to avoid the core damage from overheating. The light green and pink clusters differ mostly in ExtendedECCSOperation and CoolingFaillToRunTime. The green cluster is concentrated with data points exhibiting lower ExtendedECCSOperation and higher CoolingFailToRunTime compared to the pink cluster. Therefore, 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.

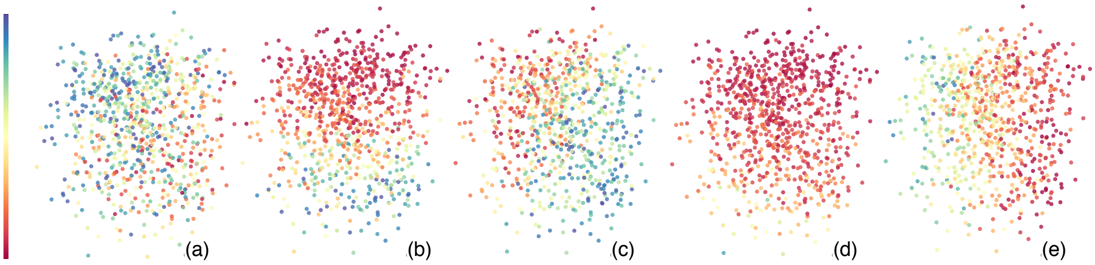


Figure 6: PCA embedding for the 8D dataset under the **Failure Scenarios Case**. The dimensions shown exhibit relatively strong correlation patterns within the embedding. (a) CoolingFailToRunTime; (b) ExtendedECCSOperation; (c) FirewaterTime; (d) simulationEndTime; (e) SRVStuckOpenTime.

Failure Scenarios Case. Once again, we color the points in the embedding for all failure scenarios, as illustrated in Figure 6. 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.

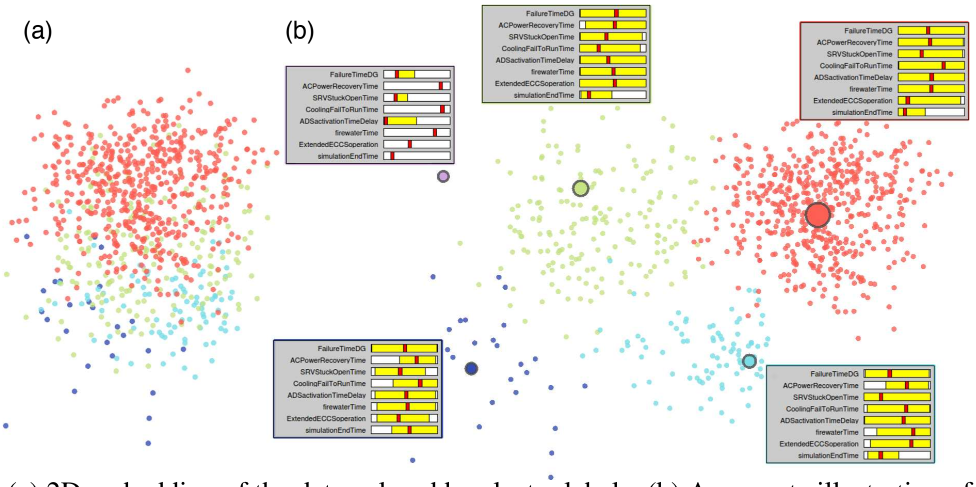


Figure 7: (a) 2D embedding of the data colored by cluster labels. (b) A separate illustration of individual clusters and their summary statistics.

Using clustering expansion, we arrive at a level of the hierarchy where five clusters are presented in the data (Figure 7). The purple cluster contains outliers that share late ACPowerRecoveryTime and

CoolingFailToRunTime. In this focused analysis of all the failure scenarios (without the interference from the dominating dimension MaxCladTemp), we obtain various insights regarding separations of clusters. For example, both green and red clusters consist of points with low simulationEndTime. The differentiating factors here are the CoolingFaillToRunTime and ExtendedECCSOperation.

4.2. Topology-Based Clustering

For topology-based clustering, we map the data into a seven dimensional 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. After careful analysis of the clustering hierarchy, we focus on a level consisting of four clusters. Figures 8 and 9 summarize our results. In Figure 8, three of the clusters share a common global maximum, whereas the fourth cluster (cyan) 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 Figure 8(right).

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. The local minimum that belongs to the pink cluster exhibits an early ACPowerRecoveryTime, a late firewaterTime, and an early ExtendedECCSOperation time. The local minimum of the blue cluster, on the other hand, has a late ACPowerRecoveryTime, a very early firewaterTime, an early ADSActivationTimeDelay, and a late ExtendedECCSOperation time. The third local minimum, shared by the green and cyan clusters, has a moderate firewaterTime paired with an early ACPowerRecoveryTime and a late 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 well aligned with the observations we have made in Figure 4, where we see that there is no visual correlation between the maxCladTemp and the FailureTimeDG (therefore we omit the plot for FailureTimeDG), and that the CoolingFailToRunTime and SRVStuckOpenTime are orthogonal in variations to the maxCladTemp in the PCA embeddings. The new information we obtain from topology-based clustering is that the firewaterTime does play a role in differentiating the pink, green, and blue clusters, as we see clear separation among the left end points of all three summary curves in its inverse coordinate plot.

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 DGs (FailureTimeDG, initial time of the SBO condition) does not play a relevant role in guaranteeing a low value of max clad temperature. For the pink cluster in (Figure 8), an early AC recovery time guarantees system success even for early values of SRVstuckOpenTime, ExtendedECCSoperation time, and late 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 max clad temperature.

For comparison, we could use the same 2D embedding from Section 4.1 and color the points based on the topological clustering results. This analysis is illustrated in Figure 9. Here we see four clusters colored in green, cyan, purple, and blue, respectively. This type of visualization gives less information regarding the correlations of input variables with respect to the output. On the other hand, for each cluster, it provides a more compact summary statistics of each input dimension.

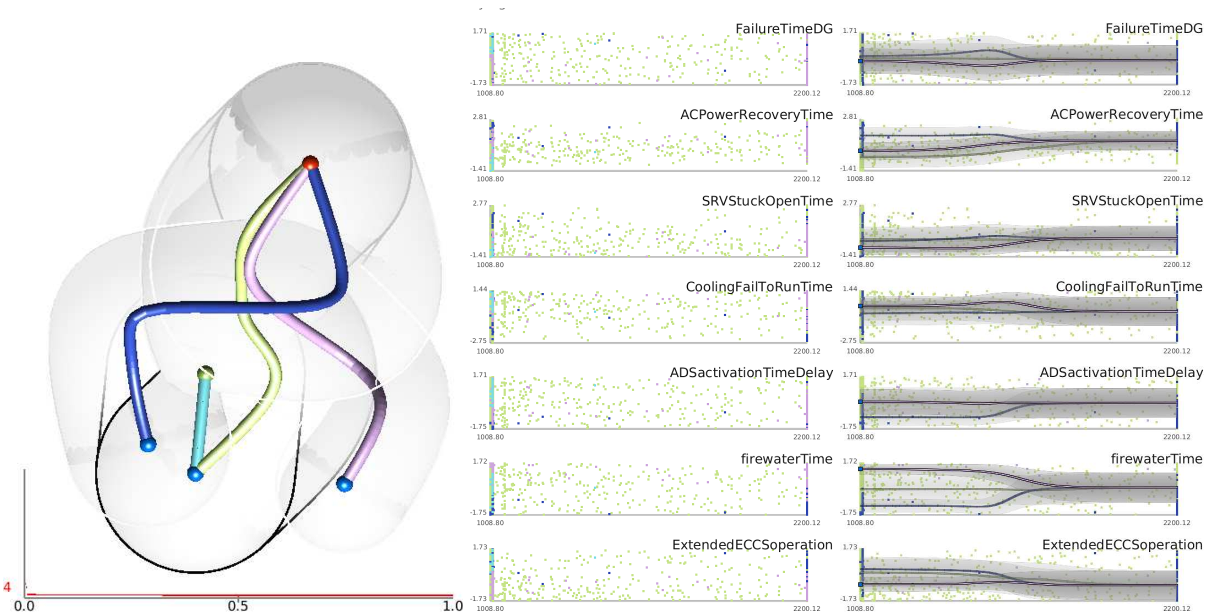


Figure 8: Left: the topological skeleton of all 4997 scenarios. Middle and right: inverse coordinate plots, where points (middle) along with summary curves (right) are colored by cluster labels.

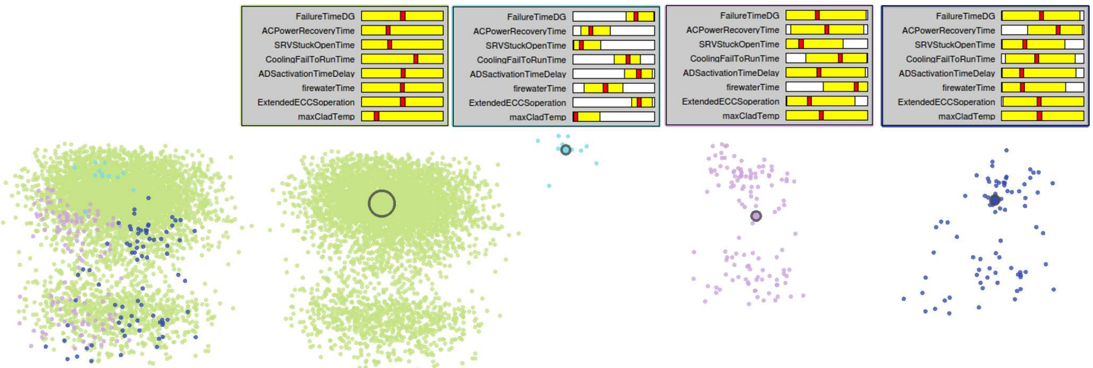


Figure 9: Left: 2D embedding of the data colored by topology-based clustering labels. Right: a separate illustration of individual clusters and their summary statistics with respect to the input dimensions.

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, as the output variable. We obtain a topology-based clustering that consists of four clusters. Results are shown in Figure 10 and Figure 11. In Figure 10(left), four clusters share a global minimum, characterized by a simulationEndTime of 434.82 seconds. 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 own cluster, the latest time to reach the failure states (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 Figure 10(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 in Figure 10(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 FW system. Also note for the green cluster that a late core damage is also correlated with a late ACPowerRecoveryTime. For all scenarios contained in the purple cluster, 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, 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. In addition, FailureTimeDG when coupled with the firewaterTime also plays a relevant role in understanding the conditions for reaching late core damage.

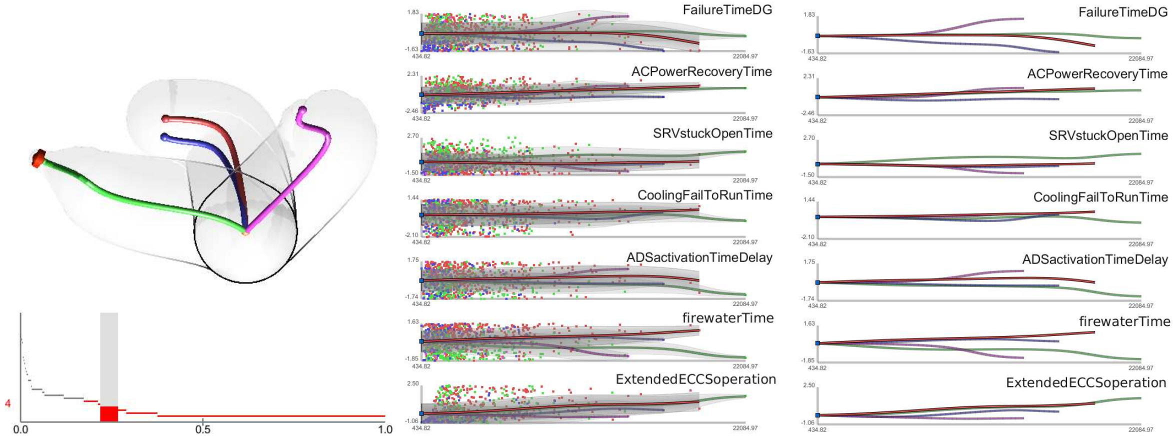


Figure 10: Left: topological skeleton of all failure scenarios. Inverse coordinate plots with (middle) and without (right) points projection. Points and summary curves are colored by cluster labels.

For comparison, as before, we color points in their 2D embedding based on the topological clustering results, as shown in Figure 11. 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 variable varies among the clusters remains hidden. For example in Figure 11, ACPowerRecoveryTime varies in its range and mean value across the four clusters; however, the inverse coordinate plot in Figure 10 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.

5. CONCLUSION

We investigate the use of both traditional and topology-based hierarchical clustering in conjunction with dimensionality reduction techniques on DPRA datasets. We provide the domain scientist with an analysis and visualization tool for obtaining insights about system responses under the simulated accident scenarios. We focus on a dataset that simulates the response of a BWR system during an SBO accident scenario. We obtain such a dataset by performing a series of simulations where, for each simulation run, we randomly change timing and sequencing of events. We would like to identify how timing or sequencing of these events affects the maximum core temperature.

We have observed that a traditional clustering combined with DR is adequate to distinguish failure scenarios with success scenarios, and to group points with similar parameter settings. On the other hand, topology-based clustering captures information regarding how input parameters are correlated with the output, and how input parameters settings help differentiate local extrema (i.e., local maxima or minima)

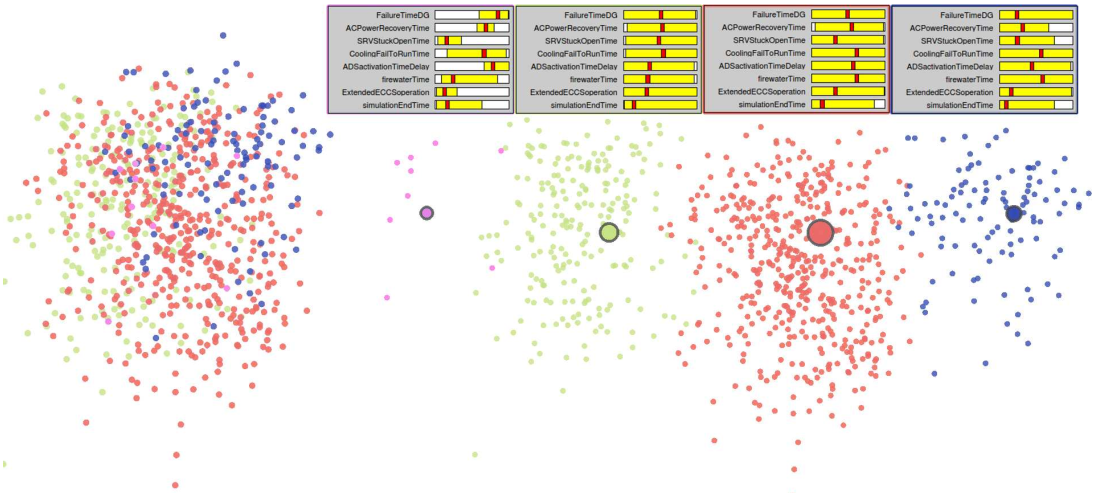


Figure 11: Left: 2D embedding of the data colored by topology-based clustering labels. Right: a separate illustration of individual clusters and their summary statistics.

of the output. We believe that pairwise comparisons and validations of both types of clustering techniques complement each other in bringing enhanced structural understanding of the data.

REFERENCES

[1] M. A. Carreira-Perpinan, “A review of dimension reduction techniques”, *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, pp. 1–69, 1997.

[2] D. Defays, “An efficient algorithm for a complete link method”, *The Computer Journal*, vol. 20, no. 4, pp. 364–366, 1977.

[3] H. Edelsbrunner, D. Letscher, and A. J. Zomorodian, “Topological persistence and simplification”, *Discrete and Computational Geometry*, vol. 28, pp. 511–533, 2002.

[4] S. Gerber, P.-T. Bremer, V. Pascucci, and R. Whitaker, “Visual exploration of high dimensional scalar functions”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 1271–1280, 2010.

[5] I. T. Jolliffe, *Principal Component Analysis*. Springer-Verlag, 1986.

[6] D. Maljovec, B. Wang, D. Mandelli, P.-T. Bremer, and V. Pascucci, “Analyze dynamic probabilistic risk assessment data through topology-based clustering”, *International Topical Meeting on Probabilistic Safety Assessment and Analysis (PSA)*, 2013.

[7] D. Maljovec, B. Wang, V. Pascucci, P.-T. Bremer, M. Pernice, D. Mandelli, and R. Nourgaliev, “Exploration of high-dimensional scalar function for nuclear reactor safety analysis and visualization.”, *International Conference on Mathematics and Computational Methods Applied to Nuclear Science & Engineering*, 2013.

[8] D. Mandelli, C. Smith, T. Riley, J. Schroeder, C. Rabiti, A. Alfonsi, J. Nielsen, D. Maljovec, B. Wang, and V. Pascucci, “Support and modeling for the boiling water reactor station black out case study using relap and raven”, Idaho National Laboratory (INL), Tech. Rep. INL EXT-13-30203, 2013.

[9] D. Mandelli, A. Yilmaz, T. Aldemir, K. Metzroth, and R. Denning, “Scenario clustering and dynamic probabilistic risk assessment”, *Reliability Engineering & System Safety*, vol. 115, pp. 146 –160, 2013.