Rethinking Sensitivity Analysis of Nuclear Simulations with Topology

Dan Maljovec* Bei Wang† Paul Rosen‡ Andrea Alfonsi§ Giovanni Pastore¶ Cristian Rabiti‖
Valerio Pascucci**

ABSTRACT

In nuclear engineering, understanding the safety margins of the nuclear reactor via simulations is arguably of paramount importance in predicting and preventing nuclear accidents. It is therefore crucial to perform sensitivity analysis to understand how changes in the model inputs affect the outputs. Modern nuclear simulation tools rely on numerical representations of the sensitivity information – inherently lacking in visual encodings – offering limited effectiveness in communicating and exploring the generated data. In this paper, we design a framework for sensitivity analysis and visualization of multidimensional nuclear simulation data using partition-based, topology-inspired regression models and report on its efficacy. We rely on the established Morse-Smale regression technique, which allows us to partition the domain into monotonic regions where easily interpretable linear models can be used to assess the influence of inputs on the output variability. The underlying computation is augmented with an intuitive and interactive visual design to effectively communicate sensitivity information to the users. Our framework is being deployed into the multi-purpose probabilistic risk assessment and uncertainty quantification framework RAVEN (Reactor Analysis and Virtual Control Environment). We evaluate our framework using an simulation dataset studying nuclear fuel performance.

Keywords: Sensitivity analysis, probabilistic risk assessment, uncertainty, nuclear simulation, computational topology.

1 INTRODUCTION

Nuclear fuel performance behavior is a very complex and multi-physical phenomenon that is, nonetheless, crucial in determining both the economical and safety performance of a nuclear power plant. Since the 2011 nuclear accident in Fukushima Japan, significant research has been focused on improving the simulation capability of fuel behavior to design safer (i.e. accident-tolerant) fuel. Qualification of a new fuel design is a long process lasting for years and costing billions of dollars. It is therefore natural to try to improve this process by increasing the understanding of the new fuel design behavior via sensitivity analysis (e.g. [32, 41]) before starting any physical experiment phase.

Given the large span of phenomena impacting the fuel performance, engineers are tasked with finding the optimal design in a multidimensional simulation space where the system behavior is potentially non-linear. Sensitivity analysis via regression has historically been used to guide engineers in the optimization process and in identifying the leading phenomena. However, using a simple linear regression model in *global* sensitivity analysis may fail to capture intrinsic local behaviors when the system is highly non-linear. For this reason, we perform *local* sensitivity analysis and visualization of multidimensional nuclear simulation data using partition-based regression models. We use the established Morse-Smale regression technique [25, 26] to identify regions of approximate monotonic behavior within the system response, and then perform regression and sensitivity analysis locally on these regions.

This process is paired with a user interface tailored specifically for the nuclear scientists. Such an interface guides the users to the discovery of the parameters driving the system behavior in the various regions of the input space. It also provides information on the non-linearity of the system response. Combining domain decomposition with local regression on the analysis side provides us the opportunity to create more targeted and detailed visual encodings than a global approach. Through close interactions with the nuclear scientists, the visual encodings are designed with specific requirements in mind and we aim to keep each view as simple as possible to enhance its usability. We follow the core stages of a design study [47]–discover, design, implement, and deploy–in this application-driven research and highlight our contributions:

- We use the established Morse-Smale regression [25, 26] in the context of sensitivity analysis, together with other common sensitivity metrics, to represent the main drivers within local regions of the model domain. Such a representation leads to a unique set of visualization design requirements.
- We report on the iterative process of refining and extending an existing visualization tool, HDViz, [24, 34, 36, 37] to match the needs of nuclear scientists. We report on the successful integration of our framework into the workflow of RAVEN [43], a multi-purpose risk assessment and uncertainty quantification software in nuclear engineering.

Our system targets nuclear simulation datasets used in sensitivity analysis, up to tens of dimensions of un/structured point cloud data. We validate and reflect on the efficacy of the new design using a 3D dataset from nuclear fuel analysis.

2 RELATED WORK

Our proposed technique focuses on performing sensitivity analysis and visualization for multidimensional nuclear simulation datasets modeled as scalar functions observed on sampled data. We employ a domain partitioning of the model, followed by local regression and analysis within each partition, and provide intuitive visualization for nuclear scientists. We review the most relevant related work on partition-based regression, sensitivity analysis methods and visualization systems designed for sensitivity analysis.

Partition-based regression. Partition-based regression techniques typically employ a systematic domain partitioning based on certain criteria, coupled with regression models fitted to each partition, to identify and understand local structures of the model. They can be classified based on their partitioning criteria: those seeking to minimize numerical errors [2, 7, 11, 13, 21], and those based on geometric analysis [25, 26, 33]. In particular, regression trees are constructed by recursively partitioning the domain into multiple partitions at *optimal* locations in the domain (where the optimal criteria can vary), and each partition is fitted with constant [7], linear [2], low-order splines [21] or polynomials [11]. On the other

*SCI Institute, University of Utah, e-mail: maljovec@cs.utah.edu
†SCI Institute, University of Utah, e-mail: beiwang@sci.utah.edu
‡University of South Florida, e-mail: prosen@usf.edu
§Idaho National Laboratory, e-mail: andrea.alfonsi@inl.gov
¶Idaho National Laboratory, e-mail: giovanni.pastore@inl.gov
‖Idaho National Laboratory, e-mail: cristian.rabiti@inl.gov
**SCI Institute, University of Utah, e-mail: pascucci@sci.utah.edu

hand, Principal Hessian direction (PHD) tree regression [33] splits the domain into areas of high curvature.

We employ the established Morse-Smale regression (MSR) [26], which, being geometrically motivated, is similar to the PHD approach. MSR utilizes a domain partitioning induced by the Morse-Smale complex (MSC) of the regression surface. The MSC partitions the domain into monotonic regions, and MSR takes advantage of the monotonicity implied by this partitioning and builds linear models to fit each partition to capture local interactions between input and output parameters. The MSC is at the core of the HDVis visualization system as seen in [24, 34, 36, 37]. However, the insights made by nuclear scientists are limited by the complexity and the lack of intuition provided in the visualization. In this paper, we follow an iterative process of designing and refining a new visualization system, using HDVis as a basis, that specifically targets nuclear scientists as end users and sensitivity analysis as the end task. We collaborate closely with nuclear scientists throughout the discover, design, implement and deploy stages; and we document such an effort in Section 4.

Sensitivity analysis methods. Sensitivity analysis (SA) studies how changes in the model inputs affect the outputs, see [46] for a survey. SA approaches can be categorized into local SA and global SA. *Local* SA addresses sensitivity relative to point estimates of the parameter values; while *global* SA focuses on information for the entire parameter distribution [29]. Local SA studies the change of model response by varying one parameter while keeping other parameters fixed. A common approach, differential SA, computes the location-dependent partial derivative of the output with respect to an input parameter. Global SA, on the other hand, explores the change of model response by varying all parameters simultaneously. Common global SA approaches include generalized SA [31] and those based on the design of experiments.

Contemporarily available SA methods include correlation analysis [50], regression analysis [22], Sobol sensitivity indices [49], Morris one-at-a-time screening (MOAT) [39], Gaussian process (GP) screening [44], multivariate adaptive regression splines (MARS) screening [21], etc.; see [23] for a comprehensive evaluation. For example, correlation analysis measures parameter sensitivity by correlation coefficients, such as the Pearson correlation coefficient (PEAR) and Spearman rank correlation coefficient (SPEAR), which measure the strength of a linear (or monotonic) relationship between model parameters and model responses [23]. Regression analysis involves fitting a linear regression to the model response and using standardized regression coefficients to evaluate parameter sensitivity. MARS and GP screenings are both examples of response surface methods used to derive relative scores of parameters' overall effects [23]. For a survey on SA methods in the field of nuclear engineering specifically, see [1].

Our analysis based on the MSR falls somewhere between a local SA and a global SA approach. It is a variation on traditional regression analysis, as it employs topology-inspired partition-based regression models. MSR allows us to partition the domain into regions of uniform gradient flow such that a linear model is fitted within each partition. We rely on first derivative information and the fitness of our local models, and also provide the option to compute PEAR, SPEAR and canonical correlation analysis (CCA) coefficients. The objective of this paper is not on the choice of superior statistic in SA, but rather on the use of topology to provide more meaningful domain partitioning.

Sensitivity analysis and visualization. We give a brief review of several visualization systems designed for SA, in particular, systems that enable visual exploration of local sensitivity information. HyperMoVal [42] uses support vector regression (SVR) [48] for high-dimensional data and visually validates a model against the ground truth. It highlights discrepancies between the data and the model, and computes sensitivity information on the model. We provide similar capabilities in our system to report on the fitness of our regression model per partition. Berger et al. [4] utilize both nearest neighbor regression and SVR to create a visual interface geared toward optimizing performance in car engine designs, where the software displays local sensitivity information on and near a user-selected focal point. Vismon [5] provides SA, comprehensive and global trade-offs analysis, and a staged approach to the visualization of the uncertainty of a simulation for fisheries scientists.

Canonical correlation analysis (CCA) has been used in the heliograph [17] and Slycat [16] system to determine correlations between multiple inputs and multiple outputs in a global setting for ensemble data. Other related works visualize local sensitivity information by encoding partial derivative, using glyphs [27], flow-based scatterplots [9], generalized sensitivity scatterplots [10], and histograms [3].

We differentiate ourselves from the prior methods by providing the capabilities to interactively refine the partitioning during the analysis to give a multi-scale view of the sensitivity information. As such, our method is, in spirit, similar to systems provided in May et al. [38] and Muhlbacher et al. [40]. However, data is partitioned along one or two dimensions (i.e. input parameters) in both cases while our partitioning spans all dimensions.

3 Technical Background

Morse-Smale regression and Morse-Smale complex. We apply Morse-Smale regression (MSR) [25, 26] in the context of SA. MSR builds upon a domain partitioning of a dataset induced by the Morse-Smale complex (MSC), and employs a linear regression fit within each partition thereby exploiting its monotonicity. The MSC itself has been successfully utilized in visual exploration of simulation data modeled as high-dimensional scalar functions [24, 34, 36, 37]. Fig. 1a shows the MSC of a 2D test function with four maxima and nine minima. The MSC decomposes the domain into sixteen partitions such that each partition can be well-approximated by a linear model as shown in Fig. 1b.

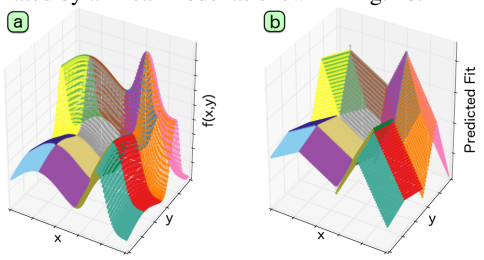

Figure 1: (a) MSC of a 2D height function that induces a partitioning of the domain. (b) Linear models are fit to each of the partitions.

The topological characteristics of the MSC is at the core of MSR. Here, we give a high-level description, see [19] for details. The MSC partitions the domain of a scalar function into monotonic regions, where points in each region have gradient flow that begins at the same local minimum and ends at the same local maximum of the function. Furthermore, the MSC can be simplified based on the notion of *topological persistence* [18, 20]. For a fixed scale, the main idea behind the simplification is to merge its corresponding partitions based on a measure of their significance (i.e. persistence), see Fig. 1d-e of [36] for an example applied to the MSC.

For point cloud data, the MSC can be approximated [12, 24], enabling MSR to be applied in high dimensions. Points are connected by neighborhood graphs such as the k-nearest neighbor (kNN) graph, and gradients are estimated along the edges of the graph. In our context, we utilize the same approximation schemes [24, 25, 26], where points are connected using the relaxed Gabriel graph, which was shown to give superior results in extracting topological features [15] compared to the kNN graph.

Linear regression and sensitivity analysis. For our analysis, we use a *least square linear regression* as the linear model for each

partition. To obtain the coefficient estimates, such a least-squares fitting minimizes the sum of squared residuals. For a given partition with n data points, let $y = [y_1, ..., y_n]^T$ be the n-by-1 vector of observed response values, X be the n-by-m design matrix of the model (that is, X_{ij} is the j-th dimension of the i-th data point), and β be the m-by-1 vector of coefficients, we minimize the error estimate: $s(\beta) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} X_{ij}\beta_j)^2$. In matrix form, we obtain the coefficient estimates $\hat{\beta}$ in the following way, $\hat{\beta} = \arg\min_\beta s(\beta) = (X^T X)^{-1} X^T y$. For SA, we use the regression coefficient $\hat{\beta}_i$ (for $1 \leq i \leq m$) to evaluate the sensitivity of the i-th parameter/dimension.

Coefficient of determination. For a given partition fitted with a linear regression model, it is important to evaluate how well the data points fit the model by computing the *coefficient of determination*, or the R^2 score. Given a partition with n data points, for the i-th data point, y_i is the observed response value and $\hat{y}_i = \sum_{j=1}^{m} X_{ij}\beta_j$ is the fitted response value. Let $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ be the mean of the observed response values. The *coefficient of determination* is computed as $R^2 = 1 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2$.

We extend such a notion by ranking and considering *how many* input dimensions are sufficient to provide an optimal fit. We select a subset of input dimensions for the n data points, apply least square linear regression on these points with reduced dimensions, and evaluate the R^2 score of the linear fit. The closer the value of R^2 is to 1, the better the linear regression fits the data with the selected subset of input dimensions.

4 Problem Characterization and Abstraction

We follow the core stages of a design study [47], namely, discover, design, implement, and deploy, in our application-driven research. During the *discover* stage, we focus on problem characterization and abstraction. We learn the targeted domain of study (in this case, SA for nuclear simulations) through close interactions with nuclear scientists. In this process, we study their *existing work flow* and *design requirements* to better enable knowledge discovery via analysis and visualization.

Existing workflow. Before designing our visualization solution, we actively engage nuclear scientists in understanding their domain problems and the existing workflow (i.e. their common practices) in SA. As a first approximation, the scientists would typically perform a global SA via linear regression or correlation analysis. Pointwise SA are used to estimate gradient information at specific locations via back-of-envelope computation, or a reduced order model (ROM) is constructed from experimental data whereupon statistical information can be collected. The scientists would also manually divide the data domain into subregions that exhibit changes in gradient behavior based on axis-aligned scatterplot projections; then global SA that combines resampling and ROM construction is applied to each manually extracted subregion.

During the discover stage, we help the nuclear scientists formulate and examine their data analysis and visualization needs through an iterative process, where we listen to their description of domain problems, characterize and abstract their problems into design requirements, and obtain feedback regarding abstractions for continuous refinement. We identify several challenges in the existing workflow that are summarized into the following design requirements.

A: Structure-based domain partitioning amenable to local SA. Existing local SA is restricted to be point-wise and domain partition is done manually via a time-consuming process; while our proposed SA, built upon MSR, applies automatically and efficiently to partitions at multiple scales. Our collaborating scientists have previous exposure to MSC and MSR from their use of HDViz [24] in probabilistic risk assessment of nuclear datasets [35, 36, 37]. We have agreed that the MSC provides a meaningful, structure-based domain partition suitable for local SA. MSR takes advantages of

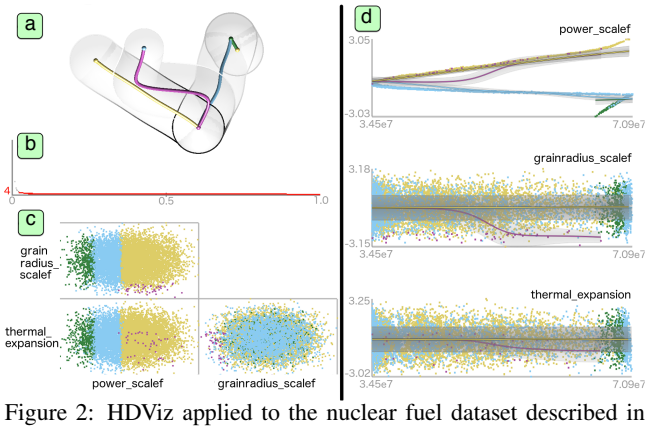

Figure 2: HDViz applied to the nuclear fuel dataset described in Section 7. (a) Topological skeleton: each summary curve in the visual space corresponds to a partition of the data; Transparent tubes capture the spread (width) and density (luminance) of the partition. (b) Persistence chart: number of partitions plotted as a function of scale. (c) Scatterplot matrix of partitioned data. (d) Inverse coordinate plots: summary curves and partitioned data are projected onto 2D plot where the x-axis represents the output dimension, and each y-axis represents an input dimension.

the underlying topology of the data and is optimized to give reasonable data fitting qualities across partitions. Using HDViz as a starting point (see Fig. 2), it is an ongoing process for us to define new visual design requirements as well as refine existing ones that satisfy the various SA tasks.

B: Intuitive visualization of data hierarchy. Existing visualization using HDViz has a very steep learning curve for data practitioners in general, is particularly non-intuitive for nuclear scientists, and has very limited support for the SA pipeline. For example, when presented with the topological skeleton seen in Fig. 2, nuclear scientists have questioned the meaning and utility of the oscillations produced by the inverse regression and the geometric interpretation of their 2D projection. During our contextual inquiries [30], we notice that the scientists typically disable the transparent tubes surrounding the inverse regression curves (Fig. 2a). Further interactions reveal that the density and geometry information provided is considered "distracting" and uninformative in their context. Instead, the scientists use the topological tube view to count the number of partitions and extrema at various scales.

Selecting the "appropriate" scale (i.e. persistence level) from the persistence chart remains an unintuitive process when the scientists are rarely aware of how many "noisy" features actually exist in the data. We have spent a lot of effort in designing various ways to convey such information to the users, as evidenced in Section 5. Understanding how the scientists interact with the topological view has greatly influenced its design in our iterative process. We understand that the scientists are primarily interested in a high-level picture of their data, therefore we simplify our visualization to an abstract 2D representation that enables easier selection and manipulation of partitions, and supports well-integrated, interactive selection of scales. Finally, the new visual representations should convey information about different scales within the data hierarchy, but also provide context about the partitions within a fixed scale. The topology map described in Section 5 is designed to fulfill such requirements.

C: Integration of common practices with new designs. We incorporate visual tools that are familiar to the nuclear scientists (e.g. scatterplots) into our new visual design to ease the knowledge discovery process. For example, scientists could select a MSC-based partition and observe its associated point cloud clustered in a geometrically coherent space in the corresponding scatterplot. Such an integration gives the users some intuition of how the topology-inspired partition is being performed. In addition, the scientists find

that the scatterplots also provide some sense of data density within each partition – such information is vital to the users as it is directly related to the confidence associated with the derived sensitivity information. Knowing this also allows us to encode density information into other representations not prone to the occlusion problem.

D: Presentation of comparative and quantitative information. Initial visualizations of the sensitivity information focus on a comparative analysis that uses shapes of different sizes to convey differences among separate partitions in the data. Though it is useful for quickly detecting major trends, the scientists have requested the numeric measurements of each partition to be displayed as this information is more familiar to them and gives them an increased amount of accuracy necessary for decision-making. Additional designs have been suggested (such as the fitness view detailed in Section 5) by the scientists which allow us to better understand their mental model of the data and design our visualization accordingly.

E: Scalable analysis and visualization. The existing analysis capabilities do not scale well with increasing number of parameters/dimensions; while our proposed SA using MSR scales well in high dimensions. Given point cloud data, the analysis in [26] has shown that the algorithm in [24] provides a good approximation of the true MSC of the underlying (unknown) function when the smallest feature (signal) of f has a persistence that is an order of magnitude larger than the standard deviation of the noise. The running time of the MSC does not depend on the ambient dimension of data points, but rather the topological complexity (e.g. number of local maxima and local minima) of the underlying function. In addition, we require that the new visual design should also remain intuitive and informative even with increasing data dimensions.

5 Visualization Design

In the *design* stage, we focus on data abstraction, visual encoding as well as interaction mechanisms [47]. We have proposed multiple visual encodings where the feedback from nuclear scientists help us narrow them down to a few usable solutions. These solutions are integrated into a linked view system with multiple visual components providing interactive functionalities as illustrated in Fig. 3.

A typical workflow begins with the *topology map*, where a user can navigate through partitionings at different scales, and at a chosen scale, explore the structure of the partitions. Within this view, one can understand at a glance the number of local extrema for a given partitioning, their relative importance encoded by persistence, and their connectivity. The appropriate choice of scale is then validated by the *persistence diagram* [14] and the *barcode* [8]. At a fixed scale, the user then selects a subset of partitions for further SA. 2D or 3D scatterplots can be constructed for selected partitions by choosing any two or three input/output dimensions, where the output dimensions include both observed and predicted values. Subsequently, the user can choose to build a *histogram* of any chosen input/output dimension. Such a histogram can also be used for selecting and filtering data for further analysis. Finally, sensitivity information is computed when requested and then visualized on a per dimension basis using the *sensitivity view*, and linear fitness information in terms of R^2 score is given in the *fitness view*. We give a detailed description of the *primary* visual encodings below: topology map, scatterplot projection, sensitivity view and fitness view; followed by a brief introduction to the *secondary* visual encodings: persistence diagram, barcode, and histograms.

A: Topology map. The topology map is a key data abstraction within our design study. It is an abstract, 2D representation of the data that highlights its topological structure. We generate and validate such an abstraction through an active and cyclic design process with the scientists to arrive at its current form. The topology map encodes the locations of local extrema defined by their persistence and function values, as well as their connectivities (i.e. curves describing flow from a minimum to a maximum). Such a visualization is, in spirit, similar to the topological skeleton proposed by Gerber et al. [24] and employed by Maljovec et al. [35, 36, 37]. However, based on the new requirements by the nuclear scientists outlined in Section 4, we have completely redesigned such a visual encoding to provide a topological summary that omits geometric information but preserves the underlying topology essential to understanding the data partitioning, therefore greatly improving its usability.

As illustrated in Fig. 3A, each local extrema is mapped onto a 2D plane, whose *x-axis* represents its *persistence* and *y-axis* corresponds to its *function value* (i.e. $f(x,y)$). Therefore, local maxima move toward the top of the display, while local minima gravitate towards the bottom; more robust/significant features appear to the right, while noisy ones tend to the left. Encoding function value to the y-axis is well-aligned with the scientists' understanding of typical 2D function plots where the dependent variables are mapped to the y-axis; in fact, the inverse coordinate plots used in HDViz (Fig. 2d) are often misinterpreted for violating this common notion. In addition, the local maxima (minima) are upward (downward) pointing red triangles for fast differentiation and counting.

Selecting the appropriate scale (i.e. persistence level) for data partitioning is essential during the exploratory process as it helps to understand the partition-based data hierarchy and sensitivity information. The above design also provides a natural separation between features and noise. The user is able to choose a scale by clicking anywhere in the plot which creates a vertical line passing through the cursor location, separating the local extrema into those that represent topological features (grey region on the right) and those that represent topological noise (white region on the left). Extrema in the grey region are visualized together with their connectivities to summarize robust topological structure of the data. Their sizes correspond to the point densities in their surrounding regions. Extrema in the white region are rendered with a default minimum size, therefore not drawing the user's attention away from more salient features, but still providing context.

The selection of the scale parameter is an interactive process where the user is free to explore various levels and construct ROMs at arbitrary levels. A typical rule of thumb is to choose a scale (a vertical line) that serves as a clear divider between well-separated clusters of extrema; in other words, to look for large separations along the horizontal direction among the extrema in the visualization. This strategy is inline with the idea of *persistence simplification* [14, 20] where the features and noise of the data are assumed to exist on two separable scales.

The scale selection is a complicated process that is hard to automate. Additional visual cues in our design also help guide such a selection process. Fitness metrics in the fitness view could be compared to aid in selecting the appropriate scale, where higher R^2 scores typically correspond to better local fits of the model at a chosen scale. On the other hand, the sizes of the extrema capture the point densities in their local neighborhood, therefore signifying a possible over-segmentation or under-sampling in their regions of interest (ROI). Such density information influences the interpretation of the R^2 score for a given partition, where high R^2 score for a partition with low point density may not be trustworthy. Under some circumstances, the user may purposefully over-segment their data in order to obtain better linear fits when enough data are available. Selecting the most informative scale for data partitioning can be ambiguous and depends on the user's intention.

Closed user feedback loops help us refine our visual encoding of the topological map. For example, the extrema in the grey region are connected via color-coded, user-adjustable cubic Bezier curves, each of which represents a Morse-Smale cell (i.e. a data partition). Such a representation affords a level of flexibility to counteract visual clutters. Each partition is identified by one of nine colorblind-safe colors [51] that is used throughout the visual interface. The topological map is used as a high-level atlas to orient users in the

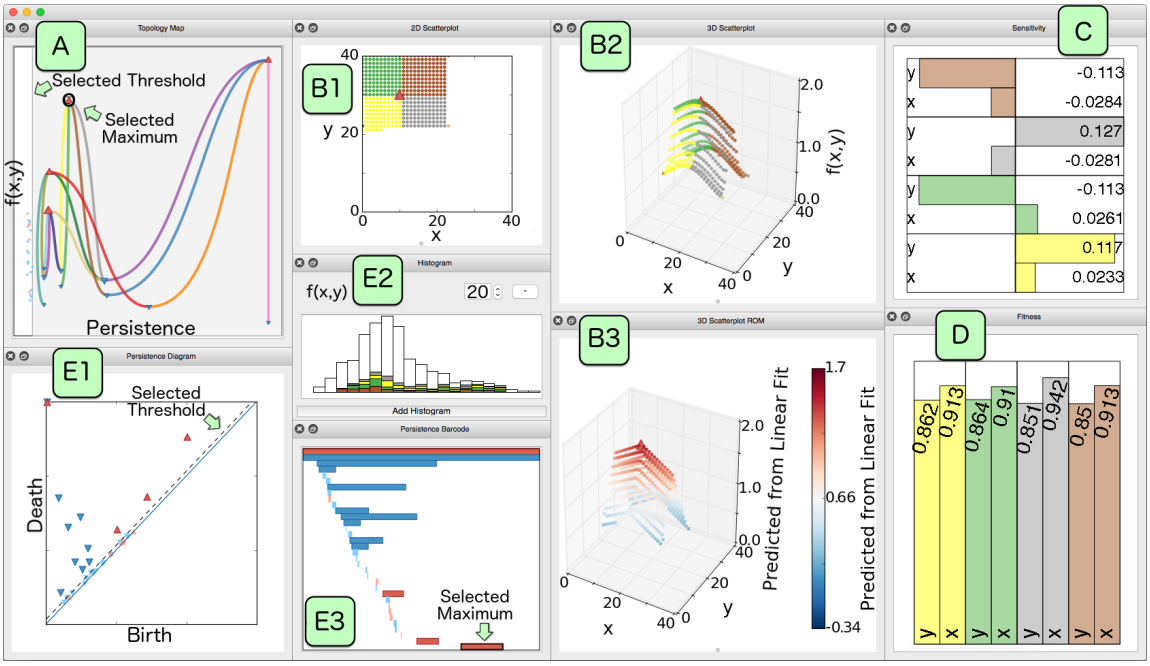

Figure 3: Our linked view visualization system using the same 2D test function of Fig. 1. It includes (A) topology map, (B1-B3) scatterplots, (C) sensitivity view, (D) fitness view, (E1) persistence diagram, (E2) histogram and (E3) persistence barcode.

data exploration process. It enables efficient comparisons among partition schemes across multiple scales. The user can select local extrema and partitions via its interactive interface for in-depth analysis via the remaining views. We also defer additional geometric information to other views to avoid sensory overload.

B: Scatterplot projection. The scatterplot projection is a common tool familiar to the nuclear scientists. It allows the user to select up to three dimensions to be mapped to spatial coordinates, and a potential fourth dimension to be mapped to color. The dimensions of choice include the input parameters, the observed and predicted output parameters, and the residuals of the ROM. The users are therefore provided with detailed spatial information on demand. The 2D height function example in Fig. 3 includes a 2D scatterplot (involving two inputs x and y in B1), a 3D scatterplot (involving x, y, and an observed output $f(x,y)$ in B2) and a 3D scatterplot for the local ROMs (involving x, y, and the predicted output).

C: Sensitivity view. We show per dimension coefficients associated with each partition of the data in signed, horizontal bar chart format. The coefficients used include the linear coefficients, the PEAR and the SPEAR. The user therefore gains first derivative behavior for each selected partition. There are two important pieces of sensitivity information to convey in our visual encoding: the sign and relative magnitude of the different coefficients. Using signed bar charts, this information becomes available at a glance, and the ubiquity of bar charts ensures their universal interpretation. These bar charts are clustered by partition for easy comparison.

D: Fitness view. For each partition, the fitness view reports the fit accuracy as well as its incremental improvement by increasing the dimensionality of the fit. Based on the linear coefficients described in the sensitivity view, the dimensions are ordered by their magnitude and visualized via vertical bar charts. Based on this sorted order of dimensions as a heuristic, we build lower dimensional linear fits beginning with only the dimension with the largest coefficient in magnitude. We then iteratively add one dimension at a time and recompute an updated linear model. For an m-dimensional dataset, we will end up with m different linear fits and for each fit, we compute the R^2 score given in Section 3. The use of vertical bar charts conveys the stepwise improvement of adding a single dimension to the regression model. Typically (but not always) the largest changes in R^2 score occur at the beginning, and we are interested to know at what point the value added by increasing the dimensionality becomes negligible, signifying potentially extraneous dimensions.

E: Secondary visual encodings. We include a few secondary visual encodings to enhance and validate insights obtained from the primary ones. The persistence diagram is the classic tool in the form of a 2D scatter plot used for separating signal from noise [18]. In a nutshell, a point in the persistence diagram corresponds to a topological feature represented by the pairing of a pair of critical points. The *persistence* of a point (and its corresponding feature) in the diagram is its distance to the diagonal. When selecting an appropriate scale for partition-based SA, the chosen vertical line in the topology map corresponds to a dotted line in the persistence diagram; a large separation between clusters of extrema in the topology map corresponds to a large separation between clusters of points in the persistence diagram. The persistence diagram provides a standard and less cluttered view of the distribution of topological features in the data, and offers a complementary and validating alternative when selecting the appropriate scale. In addition, we also include a persistence barcode (see [8] for details) that encodes similar information as the persistence diagram as another alternative for visual comparison. Histograms are provided on demand and allow the user to visualize the distribution of inputs, outputs, and various computed values on a per partition basis and also provides an interactive filtering system. These alternative visual encodings are included as part of our parallel prototyping stage and they are implemented in our final tool to offer diversity and increase efficacy.

6 Implementation

During the *implement* stage of our design study, we carefully choose algorithms, techniques and programming languages to meet the design requirements of the nuclear scientists, addressing issues such as scalability and usability. Our implementation is closely integrated into RAVEN, an uncertainty quantification and probabilistic risk assessment tool actively being developed by and for nuclear scientists. This creates a platform with good exposure for the new workflow being adopted by the scientists. Close user interactions help us refine our design and implementation in an iterative process. RAVEN is written in Python with a collection of C++ bindings for

back-end algorithms. The approximate MSC computation and our costumed version of MSR are implemented in C++, and the graphical interface is implemented using the PySide binding of Qt. Such a programming environment is chosen to support rapid software prototyping and to ensure tight integration with the existing RAVEN Python code. In addition, our modular implementation and linked view design [45] are also suitable for extending and expanding the visualization capabilities of RAVEN in the future.

7 Deployment & Evaluation: Nuclear Fuel Design

The final *deploy* stage of our design study involves software release and evaluation. To evaluate the efficacy of our design and to gather feedback from the users, our framework is applied to an actual simulation being studied by the scientists involving nuclear fuel performance. As described in Section 4, the deployment is tightly integrated into the design stage as a feedback loop, allowing both the user and the designer to exchange ideas on the data and the design. We recount such an iterative process below.

7.1 Data from Nuclear Fuel Analysis

We consider analysis of a nuclear fuel rodlet simulation performed using the BISON software [28], a modern finite-element based nuclear fuel performance code. The rodlet is axisymmetric and composed of ten stacked UO_2 pellets surrounded by a zirconium alloy shield known as the cladding. We track the midplane von Mises stress occurring on the 3D cladding. A high stress can cause the cladding to crack and allow radioactive gas to leak into the plant environment. The simulation looks at a ramping of the linear power in the reactor from time $t = 0$, up to $t = 10000$ (seconds), whereupon the power level maintains a constant value for the remainder of the simulation. The linear power begins at 0 W/m at $t = 0$ and linearly climbs to a value of 25000 W/m before leveling off.

The goal is to gain a better understanding of the physics happening at the *contact point* that occurs when the fuel rodlet expands to the point of touching the cladding. The stress on the cladding at $t = 0$ is due to a compressive force from the water pressure outside of the cladding. As fission occurs, the fuel rod expands as it is heated due to thermal expansion and swelling from the release of fission gas within the microstructure of the fuel. Once contact is made, thermal expansion (of the cladding) exerts a force that counteracts the compressive force from the water pressure, therefore the stress in the cladding decreases for a time until these forces reach an equilibrium. After the equilibrium point, the expansive forces on the cladding become dominant, causing more stress on the cladding as it expands.

In the above scenario, contact is not indicative of a failure state, and is actually expected in these types of environments. The described problems arise only when the stress in the cladding becomes too high. For this study, the scientists vary three input parameters during the simulation: (a) the linear power scaling factor (**power_scalef**), a multiplier for the previously described ramping linear power; (b) the grain radius scaling factor (**grainradius_scalef**), a multiplier for the size of microstructure elements known as the grains, and is related to the swelling caused by the fission gas; and (c) the thermal expansion coefficient for the fuel rodlet (**thermal_expansion**), which dictates how quickly the rod expands and thus how quickly it contacts the cladding. The output parameter of interest is the final midplane von Mises stress **midplane_stress** (a scalar quantity derived from the Cauchy stress tensor useful in determining at what point the cladding may yield/crack) recorded after a simulation time of $t = 10^6$ seconds. All parameters are scaled using z-score standardization to align with both domain and algorithmic practices.

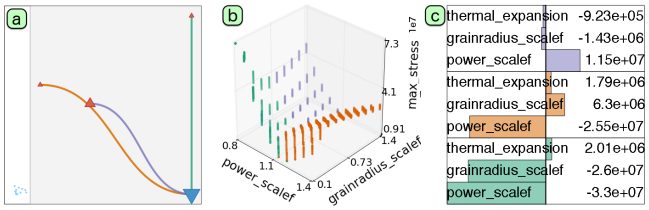

Figure 4: SA for the initial nuclear fuel dataset. (a) Topological map used to separate signals from noise. (b) Scatterplot projecting the most significant input **power_scalef**, **grainradius_scalef**, and the output **midplane_stress**. (c) Linear coefficients.

7.2 Anomaly Detection in an Initial Dataset

The nuclear scientists start their exploration with an initial dataset sampled from a uniform $10 \times 10 \times 10$ grid of the parameter space. During the exploration process illustrated in Fig. 4, they notice an unexpected behavior within the sensitivity view. The scientists expect that an increase in the **thermal_expansion** of the fuel would cause it to expand more rapidly and come into contact with the cladding sooner. The net result would cause everything to precipitate faster: for a simulation that reaches the equilibrium point, this leads to a higher stress in the cladding; for a simulation that ends before reaching the equilibrium point (exhibited via lower values of **power_scalef**), this results in a lower stress. The expectation is that the higher the **thermal_expansion** is, the higher the stress becomes. However, the linear coefficients for **thermal_expansion** across partitions in the sensitivity view contradict such an expectation: (a) **thermal_expansion** is positive for the green and orange partitions when **power_scalef** is low, and increases when approaching the equilibrium point (located at the intersection of these three partitions); (b) **thermal_expansion** is negative for the violet partition when **power_scalef** is high, and decreases past the equilibrium point. PEAR and SPEAR coefficients (not shown here) exhibit similar behaviors.

After further investigation with the fuel experts, simulation parameters aside from the three sampled ones were found to be improperly initialized, leading to this anomaly in the initial dataset. This is a very important observation as this particular dataset is being used by multiple ongoing projects where the validity of this dataset is crucial to their success. Such an anomaly in the dataset itself would not have been exposed via a global SA since the problem was detected by comparing the behavior of an input parameter from one partition to another while the partitions carry actual physical meaning. The fuel experts are able to re-run the simulation and generate a new, corrected dataset.

7.3 Workflow with a New Dataset

After fixing the bug associated with the simulation input, a new dataset is generated for SA from 9517 simulations using Monte Carlo sampling of the three input dimensions from independent distributions. Starting with the topology map in Fig. 5a, a large horizontal gap between two clusters of extrema signifies well separation between signals and noise in the data. The scientists therefore select a scale (a vertical line) that preserves the three extrema in the grey region as topological features, resulting in two partitions of the data domain. Such a selection is also supported by observing a large gap between two clusters of points within the persistence diagram (Fig. 5b), where one cluster contains points near the diagonal $y = x$ and the other contains points far away from the diagonal. As a first order approximation, the scientists build local linear regression models for these two partitions.

The resulting linear coefficients are shown in Fig. 5d where the **power_scalef** dominates the behavior of the **midplane_stress**. Furthermore, the fitness view (Fig. 5e) demonstrates that little to no information is gained by incorporating the remaining two dimen-

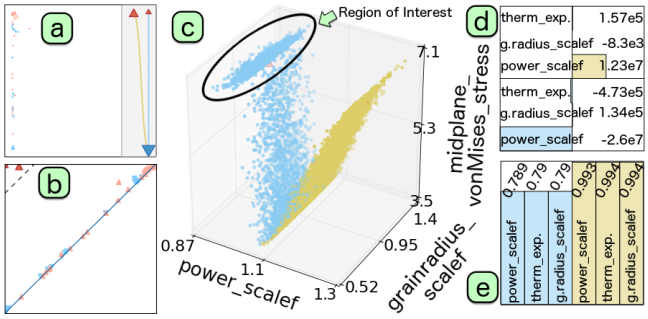

Figure 5: SA of the new nuclear fuel dataset: (a) topology map, (b) persistence diagram, (c) linked scatter plot projection, (d) linear coefficients and (e) fitness view with stepwise R^2 scores.

sions, in either partition of the data. The fitness view also shows that the blue partition is not well-described by a linear fit of any number of dimensions (i.e. with all three R^2 scores valued roughly at 0.79). The scientists investigate further by projecting the data onto a scatterplot (Fig. 5c) that includes the most sensitive input dimension (**power_scalef**) and the output dimension (**midplane_stress**). From this scatterplot, they detect the non-monotonic behavior of a region of interest (ROI) within the blue partition, which has low **power_scalef** and high **midplane_stress**. The scientists then focus on a more refined analysis at a finer scale to try to capture the behavior of the ROI within its own partition.

Fig. 6 shows the results after decomposing the domain into three and four partitions at finer scales. The scientists, guided by both the topology map and the scatterplot view, iteratively choose finer and finer scales until the ROI separates itself from the larger blue partition. The first level of refinement produces three partitions in the data (Fig. 6a). Compared to the approximation with two partitions (Fig 5c), the newly constructed magenta partition has relatively low point density (as shown in Fig. 6b) and does not correspond to the ROI. Under the second level of refinement (Fig. 6c), the ROI forms its own partition in green, as illustrated in Fig. 6d. In addition, there is significant improvement in terms of fitness for the blue region, that is, the three R^2 scores increase from roughly 0.790 to 0.898 (Fig. 7). Therefore, extracting the desired ROI requires two additional levels of refinements.

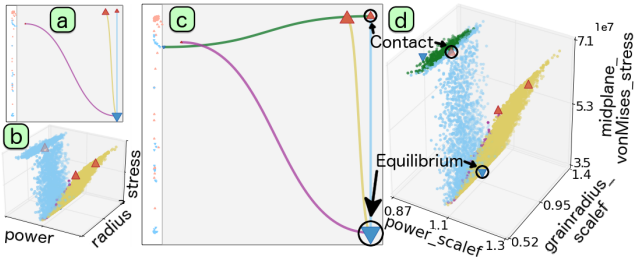

Figure 6: SA of the new nuclear fuel dataset under two refined settings: topology maps and scatterplots with three partitions (a)-(b) and four partitions (c)-(d), respectively.

Under the refined setting with four partitions, scientists are able to obtain insights that are aligned with their expectation and domain knowledge. Using topology-based domain partitioning actually allows them to decompose the domain in a physically meaningful way. In particular, the four partitions within the data domain are shown to correspond to various stages of the simulation. The scientists focus on examining the characteristics of each partition and how the partitions interface with one another. First, the interface between the green and blue partitions represents the *contact* point where the fuel touches the cladding (Fig. 6d). Within the green partition (which has very low **power_scalef** values), the net pressure acting on the cladding originates from the external water pressure. Second, the interface between the blue and the combined

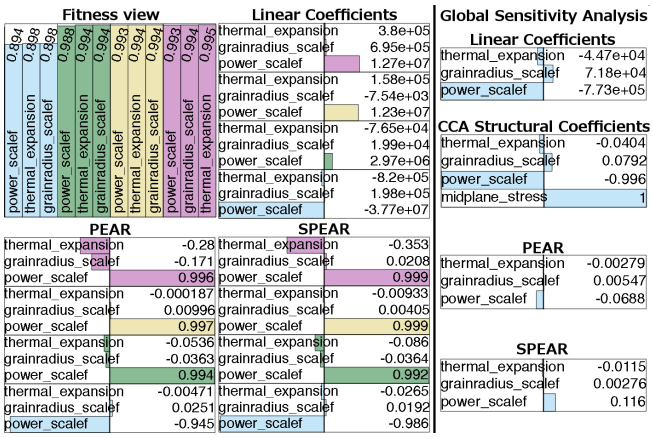

Figure 7: Left: sensitivity information of the new nuclear fuel dataset under the refined setting with four partitions. Right: global SA of the same dataset.

magenta and gold partitions corresponds to the *equilibrium* point. The blue partition represents the simulations where the expansive forces of the cladding begin to counteract the compressive water pressure force. Within the blue partition, as **midplane_stress** decreases and **power_scalef** increases, the simulation moves closer to the equilibrium. Finally, the remaining two partitions (gold and magenta) contain scenarios after the equilibrium point between compressive and expansive stresses, therefore within these partitions, an increase in **power_scalef** leads to an increase in the dominating expansive stress (i.e. **midplane_stress**). In terms of sensitivity information (Fig. 7), **power_scalef** has a strong positive effect on **midplane_stress** across the partitions with the exception of the blue region where it is strongly negative.

Finally, the scientists compare the results from the partition-based local SA (Fig. 7 left) with that of a global SA (Fig. 7 right). For the global SA, the linear coefficients and CCA structural coefficients [16] are used to identify **power_scalef** as the most sensitive parameter, yet its sign heavily depends on the amount of data on either side of the contact and equilibrium points. Similarly, the non-monotonic global behavior masks the sensitivity associated with **power_scalef** for both PEAR and SPEAR coefficients. On the other hand, the topology-inspired partition-based SA is able to capture three distinct behaviors in the data, and highlights the high sensitivity associated with **power_scalef** within each partition.

Using our framework, the scientists have validated the behaviors of a nuclear fuel simulation to be well-aligned with their expectations. They are actively investigating higher dimensional problems where this particular type of topological partition could offer them additional insights that are not readily available via traditional visualization such as scatterplot projections. Our objective in validating our deployed system is to figure out whether the scientists could be helped by our analytic and visual solutions. Our objective has been confirmed by the scientists where they could perform SA faster and with higher accuracy. The topology-inspired local SA also offers a new paradigm in rethinking about SA in nuclear engineering.

8 Conclusion

Since the recent deployment of our framework in RAVEN, our tool has generated much interest in the nuclear engineering community. According to our collaborating scientists, our software is the *first* screening tool to understand the dominant impacts of the uncertain parameters with respect to design figures of merit (e.g. internal pressure, stress, etc.). It is also useful for ranking the uncertain parameters and for the construction of effective and powerful ROM for an extensive UQ study. Recent studies [6, 32, 41] have demonstrated the application of SA to the modeling of nuclear fuel behavior, where the considered scenarios have covered both steady-state

and transient irradiation, and different burnup levels. However, existing results have been presented in terms of output range given the whole input domain (i.e. global SA). Instead, we have introduced a technique that allows fuel designers to decompose the analysis into partitions based on actual regimes of fuel behavior (e.g., open fuel-cladding gap or fuel-cladding contact). Fuel performance during these different regimes can be driven by different aspects (e.g., rod fill gas pressure or fuel-cladding contact pressure), hence, such physically based decomposition may offer a more meaningful insight into the specific situations targeted by the analysis.

REFERENCES

[1] H. S. Abdel-Khalik, Y. Bang, and C. Wang. Overview of hybrid subspace methods for uncertainty quantification, sensitivity analysis. *Ann. Nucl. Energy*, 52:28–46, 2013.

[2] W. P. Alexander and S. D. Grimshaw. Treed regression. *J. Comp. Graph. Stat.*, 5:156–175, 1996.

[3] S. Barlowe, T. Zhang, Y. Liu, J. Yang, and D. Jacobs. Multivariate visual explanation for high dimensional datasets. In *IEEE VAST*, 2008.

[4] W. Berger, H. Piringer, P. Filzmoser, and E. Gröller. Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. *Comp. Graph. Forum*, 30(3):911–920, 2011.

[5] M. Booshehrian, T. Möller, R. Peterman, and T. Munzner. Vismon: Facilitating analysis of trade-offs, uncertainty, and sensitivity in fisheries management decision making. *Comp. Graph. Forum*, 31, 2012.

[6] A. Boulor, C. Struzik, and F. Gaudier. Uncertainty and sensitivity analysis of the nuclear fuel thermal behavior. *Nucl. Eng. Des.*, 253:200 – 210, 2012. {SI} : CFD4NRS-3.

[7] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth & Brooks, Monterey, CA, 1984.

[8] G. Carlsson, A. Zomorodian, A. Collins, and L. Guibas. Persistence barcodes for shapes. In *Eurographics/ACM SIGGRAPH SGP*, 2004.

[9] Y.-H. Chan, C. Correa, and K.-L. Ma. Flow-based scatterplots for sensitivity analysis. In *IEEE VAST*, pages 43–50. IEEE, 2010.

[10] Y.-H. Chan, C. Correa, and K.-L. Ma. The generalized sensitivity scatterplot. *IEEE TVCG*, 19(10):1768–1781, Oct 2013.

[11] P. Chaudhuri, M. ching Huang, W. yin Loh, and R. Yao. Piecewise-polynomial regression trees. *Statistica Sinica*, 4:143–167, 1994.

[12] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Analysis of scalar fields over point cloud data. In *ACM-SIAM SODA*, 2009.

[13] H. A. Chipman, E. I. George, and R. E. McCulloch. Bart: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1):266–298, 03 2010.

[14] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete Comput. Geom.*, 37(1):103–120, 2007.

[15] C. Correa and P. Lindstrom. Towards robust topology of sparsely sampled data. *IEEE TVCG*, 17(12):1852–1861, Dec 2011.

[16] P. Crossno, T. Shead, M. Sielicki, W. Hunt, S. Martin, and M.-Y. Hsieh. Slycat ensemble analysis of electrical circuit simulations. In J. Bennett, F. Vivodtzev, and V. Pascucci, editors, *Topological & Statistical Methods for Complex Data*, Mathematics and Visualization, pages 279–294. Springer Berlin Heidelberg, 2015.

[17] A. Degani, M. Shafto, and L. Olson. Canonical correlation analysis: Use of composite heliographs for representing multiple patterns. In *Diagrammatic Representation & Inference*, volume 4045 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2006.

[18] H. Edelsbrunner and J. Harer. Persistent homology - a survey. *Contemp. Math.*, 453:257–282, 2008.

[19] H. Edelsbrunner, J. Harer, and A. J. Zomorodian. Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete Comput. Geom.*, 30(87-107), 2003.

[20] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. In *IEEE FOCS*, Washington, DC, 2000.

[21] J. H. Friedman. Multivariate adaptive regression splines. *Ann. Stat.*, 19(1):1–67, 03 1991.

[22] F. Galton. Regression towards mediocrity in hereditary stature. *J. Anthropol. Inst. G B Irel.*, 15:246–263, 1886.

[23] Y. Gan, Q. Duan, W. Gong, C. Tong, Y. Sun, W. Chu, A. Ye, C. Miao, and Z. Di. A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model. *Environ. Model. Softw.*, 51:269–285, 2014.

[24] S. Gerber, P. Bremer, V. Pascucci, and R. Whitaker. Visual exploration of high dimensional scalar functions. *IEEE TVCG*, 16(6), Nov 2010.

[25] S. Gerber and K. Potter. Data analysis with the morse-smale complex: The msr package for r. *J. Stat. Softw.*, 50(2):1–22, 7 2012.

[26] S. Gerber, O. Rübel, P.-T. Bremer, V. Pascucci, and R. T. Whitaker. Morse-smale regression. Manuscript, 2011.

[27] Z. Guo, M. Ward, E. Rundensteiner, and C. Ruiz. Pointwise local pattern exploration for sensitivity analysis. In *IEEE VAST*, 2011.

[28] J. Hales, S. Novascone, G. Pastore, D. Perez, B. Spencer, and R. Williamson. *BISON Theory Manual*, 2013.

[29] D. M. Hamby. A comparison of sensitivity analysis techniques. *Health Phys.*, 68(2):195–204, 1995.

[30] K. Holtzblatt and S. Jones. *Contextual inquiry: a participatory technique for system design*. Lawrence Erlbaum Assoc., Hillsdale, 1993.

[31] G. Hornberger and R. Spear. An approach to the preliminary analysis of environmental systems. *J. Environ. Manag.*, 12(1):7–18, 1981.

[32] T. Ikonen and V. Tulkki. The importance of input interactions in the uncertainty and sensitivity analysis of nuclear fuel behavior. *Nucl. Eng. Des.*, 275(0):229 – 241, 2014.

[33] K.-C. Li, H.-H. Lue, and C.-H. Chen. Interactive tree-structured regression via principal hessian directions. *JASA*, 95(450), 2000.

[34] D. Maljovec, S. Liu, B. Wang, D. Mandelli, P.-T. Bremer, V. Pascucci, and C. Smith. Analyzing simulation-based PRA data through traditional and topological clustering: A BWR station blackout case study. *RESS*, 2015.

[35] D. Maljovec, B. Wang, D. Mandelli, P.-T. Bremer, and V. Pascucci. Adaptive sampling algorithms for probabilistic risk assessment of nuclear simulations. *PSA*, 2013.

[36] D. Maljovec, B. Wang, D. Mandelli, P.-T. Bremer, and V. Pascucci. Analyzing dynamic probabilistic risk assessment data through clustering. *PSA*, 2013.

[37] 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. *M&C*, 2013.

[38] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer. Guiding feature subset selection with an interactive visualization. In *IEEE VAST*, pages 111–120, 2011.

[39] M. Morris. Factorial sampling plans for preliminary computational experments. *Technometrics*, 33(2):161–174, 1991.

[40] T. Muhlbacher and H. Piringer. A partition-based framework for building and validating regression models. *IEEE TVCG*, 19(12), 2013.

[41] G. Pastore, L. Swiler, J. Hales, S. Novascone, D. Perez, B. Spencer, L. Luzzi, P. V. Uffelen, and R. Williamson. Uncertainty & sensitivity analysis of fission gas behavior in engineering-scale fuel modeling. *J. Nucl. Mater.*, 456:398–408, 2015.

[42] H. Piringer, W. Berger, and J. Krasser. Hypermoval: Interactive visual validation of regression models for real-time simulation. volume 29, pages 983–992. Blackwell Publishing Ltd, 2010.

[43] C. Rabiti, A. Alfonsi, J. Cogliati, D. Mandelli, R. Kinoshita, and S. Sen. *RAVEN User Manual*, 2015.

[44] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[45] J. Roberts. State of the art: Coordinated multiple views in exploratory visualization. In *CMV 2007*, pages 61–71, July 2007.

[46] A. Saltelli. *Global Sensitivity Analysis: The Primer*. John Wiley, 2008.

[47] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE TVCG*, 18(12):2431–2440, 2012.

[48] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Stat. Comput.*, 14(3):199–222, Aug 2004.

[49] I. Sobol'. Sensitivity analysis for non-linear mathematical models. *Math. Model. & Comput. Exp.*, 1:407–414, 1993.

[50] C. Spearman. The proof and measurement of association between two things. *Am. J. Psychol.*, 15(1):72–101, 1904.

[51] P. Tol. Colour schemes. Technical Note SRON/EPS/TN/09-002, SRON Netherlands Institute for Space Research, Dec 2012.