Analysis of Large-Scale Scalar Data Using Hixels

David Thompson*
Sandia National Laboratories

Joshua A. Levine†
SCI Institute
University of Utah

Janine C. Bennett*
Sandia National Laboratories

Peer-Timo Bremer‡
Lawrence Livermore National Laboratory

Attila Gyulassy†
SCI Institute
University of Utah

Valerio Pascucci†
SCI Institute
University of Utah

Philippe P. Pébay*
Sandia National Laboratories

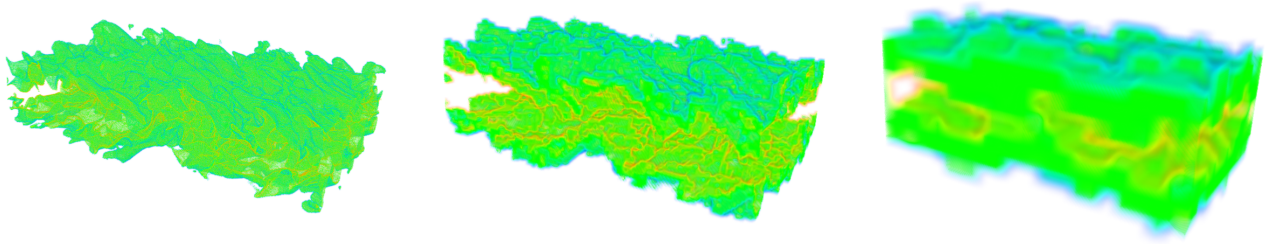


Figure 1: Volume rendering of the jet dataset, down-sampled using hixels. We visualize the scalar field g defined in Section 6, indicating the likelihood that the isosurface at isovalue $\kappa = 0.506$ passes through a voxel. From left-to-right, we show hixels that block 2^3, 8^3, and 32^3 data values. Opacity is a triangle function centered at $g = 0$ and color is a rainbow map, red for high values, green for middle, and blue for low.

Abstract

One of the greatest challenges for today's visualization and analysis communities is the massive amounts of data generated from state of the art simulations. Traditionally, the increase in spatial resolution has driven most of the data explosion, but more recently ensembles of simulations with multiple results per data point and stochastic simulations storing individual probability distributions are increasingly common. This paper introduces a new data representation for scalar data, called hixels, that stores a histogram of values for each sample point of a domain. The histograms may be created by spatial down-sampling, binning ensemble values, or polling values from a given distribution. In this manner, hixels form a compact yet information rich approximation of large scale data. In essence, hixels trade off data size and complexity for scalar-value "uncertainty".

Based on this new representation we propose new feature detection algorithms using a combination of topological and statistical methods. In particular, we show how to approximate topological structures from hixel data, extract structures from multi-modal distributions, and render uncertain isosurfaces. In all three cases we demonstrate how using hixels compares to traditional techniques and provide new capabilities to recover prominent features that would otherwise be either infeasible to compute or ambiguous to infer. We use a collection of computer tomography data and large scale combustion simulations to illustrate our techniques.

Keywords: Petascale Visualization, Scalar Field Data, Topology-Based Techniques, Ensemble Visualization

Index Terms: I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling—Boundary representations; G.1.1 [Mathematics of Computing]: Interpolation—Interpolation formulas;

*{dcthomp, jcbenne, pppebay}@sandia.gov
†{jlevine, jediati, pascucci}@sci.utah.edu
‡bremer5@llnl.gov

IEEE Symposium on Large-Scale Data Analysis and Visualization
October 23 - 24, Providence, Rhode Island, USA

1 Motivation

Prompted by the ever increasing performance of modern supercomputers, state of the art simulations continue to increase in size and complexity. Furthermore, not only are ever more complex phenomena simulated at higher resolutions but new forms of data are emerging. Traditionally growth in complexity has been driven by an increase in spatial resolution, enabling the study of more intricate situations. More recently, ensembles of simulations are becoming more common to study different outcomes and determine the sensitivity and uncertainty inherent in predictive simulations [22]. In such systems the spatial resolution is multiplied by the number of ensemble members (potentially hundreds or thousands of individual simulations) and each data point now contains many possible outcomes. Finally, new paradigms are being developed in which simulations natively act on random variables or probability distribution functions instead of scalar fields to characterize errors [9].

In the first two cases (high spatial resolution and ensembles of runs) the data in its native form is too large to be handled with traditional techniques while in the third case (intrusive uncertainty quantification or stochastic simulations) few techniques exist that are able to process such a representation. Therefore, a common approach is to pull samples from the continuous distributions thus creating an ensemble. Given the size of such data it is natural to search for reduced representations. Subsampling or computing means and standard deviations of spatial regions are both examples of this. However, such techniques have significant drawbacks, especially when data is not well-characterized by any single value which is the case for multi-modal data. As an alternative, we propose a new unified data representation called *hixels*, which at each sample point stores a histogram of values. Hixels naturally approximate all three original data representations, can be significantly smaller than the source data, and provide a more graceful degradation of the original information than subsampling or averaging since they preserve information of the scalar function range.

Consider the synthetic example shown in Figure 2. This "ensemble" is created by drawing samples from two different distributions, a Poisson distribution (left top) and a Gaussian distribution (left bottom). Simple statistical analysis of this ensemble is unable to identify basic characteristics. For example, when computing a

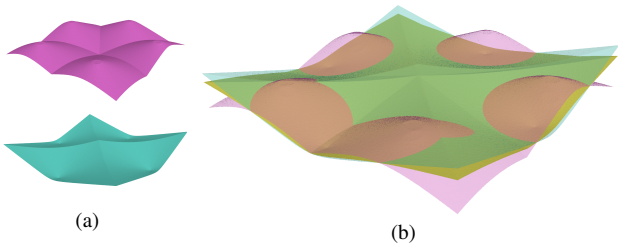


Figure 2: A 2D hixel data set generated by sampling images drawn from a Poisson and a Gaussian distribution. (a) Mean surfaces of each distribution, Poisson on top (magenta), Gaussian on the bottom (cyan). (b) Mean surface for all sampled values shown in yellow, overlaid with the Poisson and Gaussian. This surface loses characteristics of both.

single representative scalar field for the data such as the mean value (right) the resulting surface no longer shows characteristics of either distribution. Furthermore, computing the probability of traditional features occuring is inconclusive. For example, we have computed the probability that each hixel corresponds to an actual minimum, maximum, saddle, or regular point. The results of this test were entirely inconclusive; each sample had approximately 53% chance of being regular, 20% chance of being a minimum, 20% chance of being a maximum, and 7% chance of being a saddle point. An effective solution must in fact combine both traditional feature extraction and statistical methods. To address this need, we introduce three new techniques to visualize and analyze hixel data and demonstrate our results using large scale simulations and computer tomography data. Our contributions in detail are:

- We introduce hixels: per sample histograms as a new data representation for large scale and/or probabalistic data;
- We describe algorithms to extract approximations of common topological structures from hixel data;
- We develop techniques to segment multi-modal data by splitting individual histograms into their modes and correlate neighboring modes;
- We present a new method to define and render uncertain isosurfaces; and
- We demonstrate results using large scale simulation and CT data.

2 Related Work

Topology. The topological tools presented in this paper are based on Morse theory, a mathematical tool to study how the "shape" of a function is related to the shape of its domain [20, 21]. Morse theory is a well understood concept in the context of smooth scalar fields, and has been effectively extended to piecewise-linear [8] and discrete [10] domains.

Using this underlying mathematical framework, approaches for computing the Morse-Smale (MS) complex have been used for the analysis of scalar valued data. For example, Laney et al. [17] used the descending 2-manifolds of a two-dimensional MS complex to segment an interface surface and count bubbles in a simulated Rayleigh-Taylor instability. Bremer et al. [3] used a similar technique to count the number of burning regions in a lean premixed hydrogen flame simulation. Gyulassy et al. [14] used carefully selected arcs from the 1-skeleton of the three-dimensional MS complex to analyze the core structure of a porous solid.

Algorithms for computing the MS complex have been presented in the PL context [2, 8] as well as the discrete context [13, 16, 18]. One of the goals of this article is to extend the use of these approaches to include functions f that are probability functions instead of deterministic scalars. Thus, in addition to reviewing topological methods, we review literature characterizing uncertain or stochastic fields defined over a geometric space.

Data Representations. In this paper we model variable values at each node of a mesh with a distribution. Alternatively, we could represent the quantity or quantities of interest across the entire domain as a random variable or field indexed on spatial coordinates. For instance, under specific conditions [31], a stochastic process can be expressed as a spectral expansion in terms of suitable orthogonal eigenfunctions with weights associated with a particular density. A well-studied example is the Wiener process which can be formulated as an expansion in terms of multi-variate Hermite polynomials of normal distributions. Other formulations are available and have been examined for various classes of modeling such as flow-structure interaction or diffusion problems [32]. However, unless a simulation directly produces such a spectral expansion, it is difficult to obtain one by fitting postulated forms, as decisions regarding the family of basis functions, the number of terms in the series, and the order of the polynomials used amount to modeling choices [7] which are difficult to justify *a priori*. Another approach, taken by Adler and Taylor [1], is to compute the expectation of geometric properties, such as volume and the Euler characteristic, on excursion sets of spatial Gaussian processes. Again, these techniques require a specific model in order to yield results – one which may not always be simple or possible to obtain.

Computing the Modes of a Distribution. As part of our computational pipeline for hixelated data, we use topological persistence measures to decompose the distribution into a set of buckets in which the distribution behaves unimodally. Mode identification is a problem that has been explored in the statistics community. Existing mode finding algorithms identify a set of of potential modes and propose a variety of statistical tests [5, 6, 15, 30] or use function optimization [11] to identify modes of interest.

Uncertain Contours. Contouring in both two and three dimensions is one of the standard tools of scientific visualization and several methods exist to indicate uncertainty in the computation. In two dimensions the most common form of uncertainty visualization is to display the mean values enhanced by the standard error [19, 27]. Alternatively, Osorio and Broodlie [23] use a fuzzy colormap to indicate an uncertain contour. In three dimensions, Rhodes et al. [29] also use the color on a standard isosurface to indicate uncertainty but their approach does not represent the standard error or variance along the surface. Brown [4] uses a vibrating surface to indicate error on a surface and Grigoryan and Rheingans [12] use a displaced point cloud. Most closely related to the uncertain isosurfaces of Section 6 are the volume renderings of likelihood functions for atom positions proposed by Rheingans and Joshi [28]. However, their system is designed for a small, fixed number of uncertain objects rather than arbitrary isosurfaces. More recently, Pöthkow and Hege propose techniques for visualizing positional uncertainties in isosurfaces [25], where uncertainties are modeled as Gaussian distribution, later extended to include contingency information [26] and tag ray casts with probabilities of crossing the contour [24] similar to our own.

3 Foundations

In order to develop this new interpretation of scalar field data, we review mathematical foundations that we rely on from both the topology and statistics community. Finally, we extend the notation from the literature review with some new terms that will be useful in describing the algorithm.

3.1 Topological Concepts

We briefly recall that a d-manifold is a topological space that has a neighborhood homeomorphic to an open subset of the d-

dimensional space $\mathbb{R}^d$. For example, a line and a circle are 1-manifolds, and a plane and sphere are 2-manifolds. Let $\mathbb{M}$ be a smooth, compact d-manifold and let $f\colon \mathbb{M} \to \mathbb{R}$ denote a smooth real-valued function on $\mathbb{M}$. Assuming a local coordinate system at a point $x \in \mathbb{M}$, the *gradient* $\nabla f(x)$ consists of all first-order partial derivatives of f. A point $x \in \mathbb{M}$ is said to be *critical* when $\nabla f(x) = 0$. The *Morse Lemma* states that in the neighborhood of a non-degenerate critical point x, a local coordinate system can be constructed such that

$$f(x_0,\ldots,x_n) = f(x) \pm x_0^2 \pm \ldots \pm x_n^2.$$

Critical points are categorized by their *index* which is equal to the number of minus signs in the above summation. A function f is said to be *Morse* if it satisfies the two genericity conditions: (1) all critical points are non-degenerate (have invertible Hessians); and (2) $f(x) \neq f(y)$ whenever $x \neq y$ are critical. The critical points of a Morse function and their indices capture topological properties of the manifold domain of f.

An integral line in f is a path in $\mathbb{M}$ whose tangent vectors agree with the gradient of f at every point along the path. The upper and lower limits of an integral line are its *destination* and *origin*, respectively. *Unstable* and *stable* manifolds are obtained as clusters of integral lines having common origin and destination, respectively. *Basins* are traditionally defined as the unstable d-manifolds of a d-dimensional scalar function. A Morse function f is a *Morse-Smale function* if its unstable manifolds intersect stable manifolds only transversally. The intersection of the unstable and stable manifolds of a Morse-Smale function forms the *Morse-Smale (MS) complex*.

A function can be simplified by repeated cancellation of pairs of critical points that are connected by 1-dimensional cells of the MS complex. This process simulates a smoothing of the function and represents the function at multiple scales. We measure the weight of a cancellation by *persistence*, the absolute difference in value of the cancelled pair of critical points. A function is simplified to a persistence threshold by cancelling all critical points with lower persistence.

3.2 Terminology

As mentioned in the motivation, we are interested in characterizing an uncertain scalar field defined at many points in a metric space, $\mathbb{M}$. A *hixel* is a point $x_i \in \mathbb{M}$ with which we associate a histogram of scalar values, $h(x_i)$. In our setting, the $h(x_i)$ could either represent a collection of values in a block of data, collections of values at a location over a set of runs in an ensemble, or uncertainty about the potential values a location may represent. Figure 3 shows several empirical distributions with maxima identified.

3.2.1 Bucketing hixels

When a hixel is defined empirically as a number of values n_{f_i} on a finite support $\{f_j \mid j \in \{1,2,\ldots,N_f\}\}$, we call each entry of the support a *bin*. The probability distribution (specifically here a probability mass function) is thus given by:

$$h\colon f_i \longmapsto \frac{n_{f_i}}{\sum_{k=1}^{N_f} n_{f_k}}$$

for each possible value f_j. Whether this distribution is defined empirically or analytically, for instance as a weighted sum of Gaussians, we are interested in identifying regions of high probability associated with peaks in the probability density. For that we will perform topological segmentation of the histogram to identify peaks as well as range of function values associated with each peak. This range of function values is called a *bucket*. Thus, a bucket will contain one or more bins and will have a probability given by the cumulative distribution function over that range of function values. Figure 3 illustrates how the distributions have been bucketed by merging maxima of h with their lowest vertical-persistence neighbors in order of their associated probability.

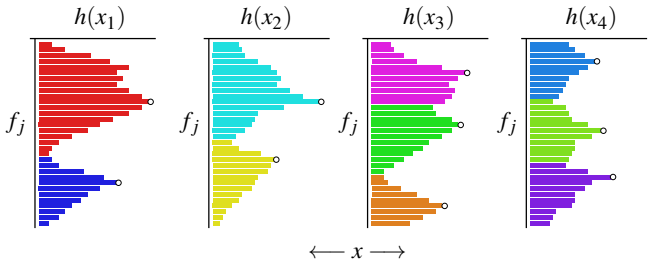


Figure 3: Four probability distributions represented as histograms $h(x_i)$ with 32 bins f_j (rotated 90 degrees). Maxima (identified with black circles) indicate function values with high probability. Colors indicate bucketing, the aggregation of bins of the histograms into modes based on the stable manifolds of persistent maxima.

Because our scalar function f is represented by probability distribution h and we are interested in identifying regions of high probability, we extend the notion of persistence to *areal persistence*. By ordering intervals between maxima and minima according to the area underneath them, peaks in probability density may be eliminated according to the probability associated with them. The decision of which of the two possible minima (assuming the maximum is interior) should be merged with the peak is made using vertical persistence: the smaller vertical persistence indicates the region to which corresponds the peak to be eliminated. Buckets can be merged in this fashion until the probability of the smallest bucket is above some threshold. When the number of sampled values is small, this threshold must be close to 1 since our confidence will be low. Assuming that f has a finite variance (so that the central limit theorem holds), the threshold may be lowered as the number of values increases. Eventually, each hixel will have one or more buckets corresponding to probable function values associated with a peak in the distribution function; each bucket thus corresponds to an estimated *mode* of the distribution.

Figure 4 shows the bucket counts for the jet dataset as areal persistence thresholds are varied. At low thresholds, hixels that encompass areas of turbulent behavior have high bucket counts. As persistence simplification is applied, but increasing the threshold of areal persistence, buckets are merging indicating the most probable modes of the dataset.

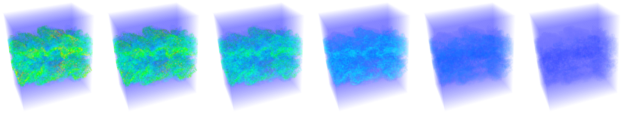


Figure 4: Varying areal persistence and its effect on bucketing for the Jet dataset. Using hixels with size 16^3 and 256 bins/histogram, we vary the areal persistence for all powers of two between 16 and 512, inclusive. Color indicates how many buckets at that hixel's position in (x,y,z) space, red indicating more, blue indicating fewer. At low levels of persistence, as many as 76 buckets can be selected in the hixel, but as persistence increases, most hixels have only 1 or 2 buckets.

4 Sampled Topology

Hixels encode the potential values along with their distributions at sample locations, a fact that can be exploited in visualizing the uncertainty in topological segmentations of down-sampled data. We use a three step process where we (1) sample of the hixels to generate individual instances of the coarser representation, (2) compute

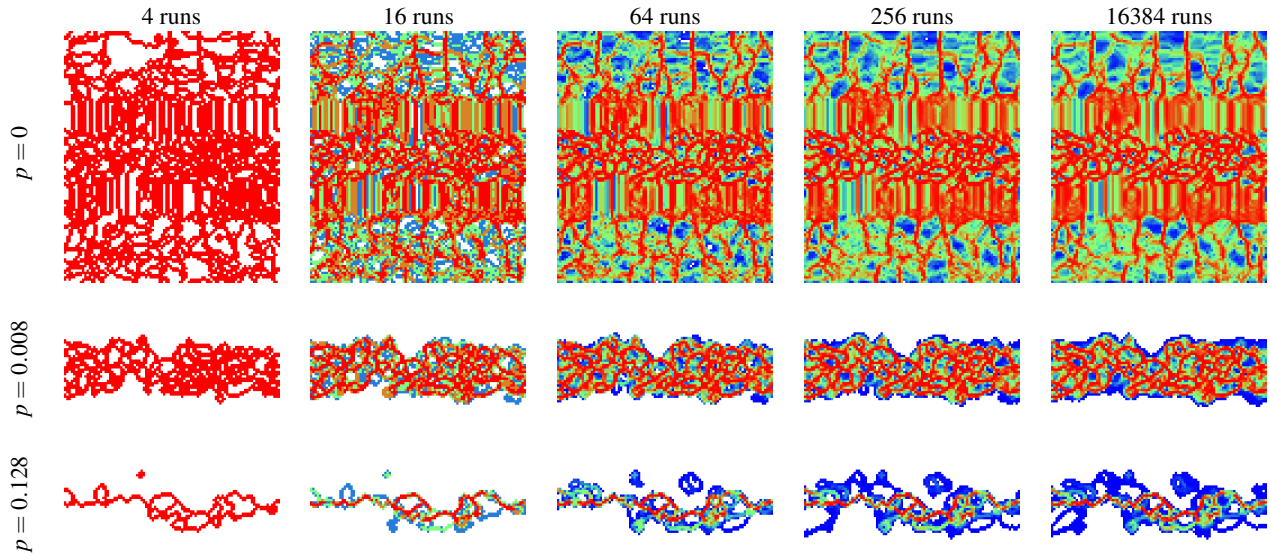


Figure 5: We sample the hixel data for an 8x8 blocking of combustion data, and compute the aggregate segmentation for a number of iterations, also varying the level of persistence simplification. Adjacent white pixels are identified in the interior of the same basin in every single run. The images converge as the number of iterations increases left to right.

the Morse complex on the instance, and (3) aggregate multiple instances of the segmentation to visualize its variability. We generate an instance V_i of the down-sampled data by picking values at each sample location from the co-located hixel. The value is picked at random as governed by the distribution encoded by the hixel. By picking values independently from neighboring values, we can simulate any possible down-sampling of the data, assuming each hixel's distribution is independent.

For each sampled field V_i we compute the Morse complex of the instance using a discrete Morse theory based algorithm [13], and identify basins around minima for varying persistence simplification thresholds. We next create a binary field C_i that encodes the geometric information of the arcs of the complex. Each sample location in V_i contributes a value of 1 to C_i if the sample is on the boundary of two or more basins, otherwise it contributes 0 if the sample is in the interior of a basin.

To visualize the variability of the topological segmentation of subsampled data, we repeatedly sample the hixels producing V_i's, and compute their basin boundary representations C_i. After n iterations, an aggregate function is computed over the boundary representations, recording the fractional identification of a sample location as a basin boundary. Formally, at each sampled location we compute the aggregate function $a_j = \frac{1}{n}\sum_{C_i} c_j$. Note that a_j can take values between one and zero, where one indicates it was identified as the boundary of basins in every instance, and zero meaning it was identified as interior in every instance. In this manner, we visualize rasterizations of the geometry of the Morse complex.

One point of interest is the amount of sampling required to capture a reasonable aggregate field. We perform convergence tests for a two-dimensional slice through a jet combustion simulation. In this experiment, we computed a hixel representation for the slice with blocks of size 8x8 and 16x16. We performed a number of iterations, recording the difference in the aggregates between n and $2n$ iterations, as shown in Figure 6. Furthermore, we visualize in Figure 5 each aggregate slice for the 8x8 block size, as number of iterations and topological persistence are varied. The convergence of these sequences indicates that the distribution represented by the hixels produces implies stable modes of segmentation.

We further visualize the aggregate segmentation as block size and level of persistence simplification are varied in Figure 7. In

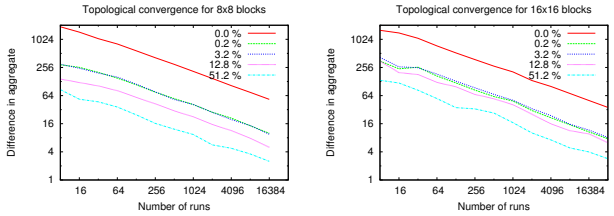


Figure 6: We compute the difference in the aggregate segmentations between n and $2n$ iterations. We use 8x8 (left) and 16x16 (right) sized blocks and varying thresholds of persistence simplification.

each case, we performed sufficient iterations for the visualization to converge to a stable image. In this figure, as the block size is increased, more variability is present in the analysis of an instance, as expected. Furthermore, this variability is reduced with increasing persistence simplification, since the segmentation itself in this case becomes more stable.

5 Topological Analysis of Statistically Associated Buckets

As HPC resources increase, ensembles of runs are being computed more frequently to explore the state space of phenomena of interest. The resulting ensemble data comprises a collection of simulation results, each of which represents a state in the system defined by different input parameters and/or models. While ensemble data sets are hailed as a useful mechanism for characterizing the uncertainty in a system, their large size and variability pose significant challenges for existing analysis and visualization techniques.

In this section we describe a novel statistical technique for recovering prominent topological features from ensemble data stored in hixel format. This computation is aided by the fact that ensemble data has a statistical dependence between runs that allows us to build a structure representing a predictive link between neighboring hixels. Our algorithm identifies subregions of space and scalar values that are consistent with positive association and we perform topological segmentation on only those regions.

After bucketing all hixels as described in Section 3 we identify

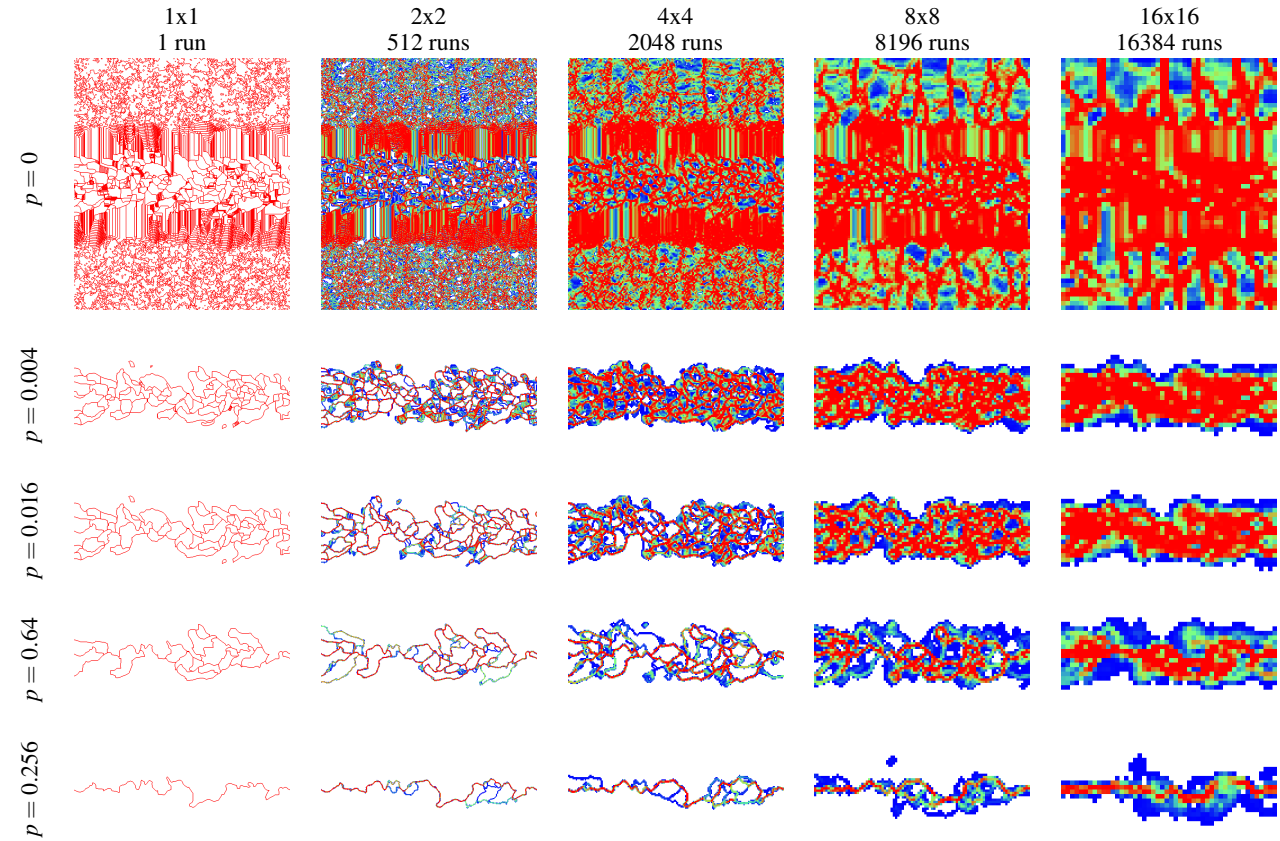
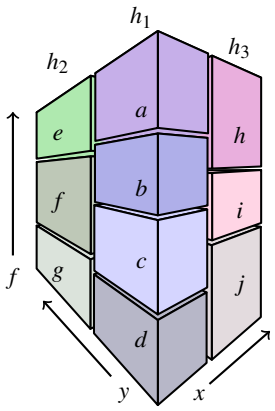


Figure 7: We sample the hixel data to construct one instance of the subsampled data, and find the pixels on the boundary of basins. We repeat this until the difference between aggregate segmentations is less than a user-specified threshold. We study the effects of varying block size and persistence simplification. The block sizes increase left to right, and the corresponding boundaries become less certain.



h_1–h_2	e	f	g
a	4	2	0
b	2	3	1
c	0	5	1
d	6	0	0

h_1–h_3	h	i	j
a	5	1	0
b	1	4	1
c	2	4	0
d	0	1	5

Figure 8: The spatial domain in this example is 2-dimensional (x and y) and the function, f, on which the data is being hixelated is the vertical axis. Hixel h_1 has buckets a, b, c, and d, hixel h_2 has buckets e, f, and g, and hixel h_3 has buckets h, i, and j. On the right two contingency tables are shown which tabulate the counts of simultaneously observed values for $h_1 - h_2$ and $h_1 - h_3$.

subregions with positive association. We begin by computing a *contingency table* or tabular representation between each pair of adjacent hixels, h_i and h_j, of the counts of all observed combinations of values as shown in Figure 8. By considering simultaneously observed values of h_i and h_j, it is possible to identify pairs of buckets that co-occur more frequently than if they were statistically independent by identifying those whose *pointwise mutual information* (pmi) is greater than zero. Pointwise mutual information is a statistical measure of association between realizations of discrete random variables. The pmi of a realization (x,y) of a pair of discrete random variables (X,Y) is defined as:

$$\operatorname{pmi}(x,y) := \log \frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)},$$

where p_X, p_Y, and $p_{(X,Y)}$ respectively denote the probability density functions of X, Y, and the joint probability (X,Y), for all possible outcomes of X and Y. When the joint probability vanishes the pmi is set to $-\infty$. Note that if X and Y are independent, then the pointwise mutual information vanishes everywhere the joint probability does not. Naturally, as this is a pointwise quantity, a zero value of the pmi does not indicate mutual independence of the random variables.

Taking some $\varepsilon \geq 0$ threshold on all contingency tables between neighboring hixels, we obtain a graph of connected buckets. We call these connected components *sheets*, illustrated in Figure 9. Sheets are geometrically like lower-dimensional surfaces in the product space of the spatial variables and the scalar data. Once we have selected sheets, we compute topological basins of minima and maxima on each sheet individually.

We demonstrate results on a mixture of 2 stochastic processes shown in Figure 10. This data highlights the fact that individual hixels can be multi-modal and can behave as both a minimum and maximum. A naive analysis that computes the mean or median of the hixels, followed by standard topological segmentation would fail to incorporate the multi-modal nature of the data. Our method

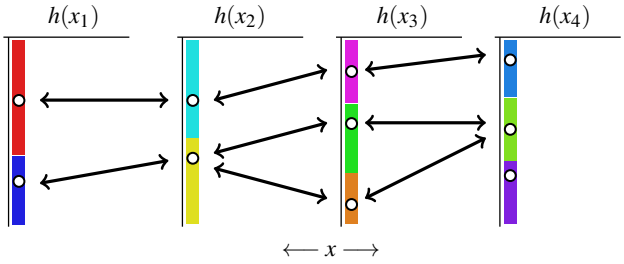
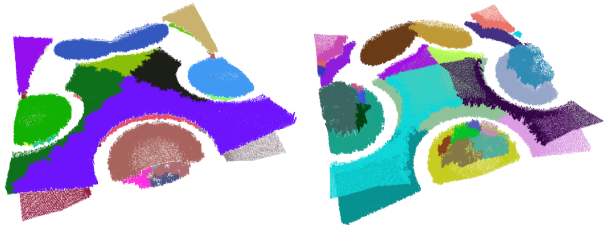


Figure 9: Once the hixels have been bucketed and modes have been identified, we compute the pointwise mutual informations between buckets that are spatially adjacent and connect those with positive associations to form sheets.

addresses this issue by performing topological analysis directly on sheets of the domain that have likely simultaneously observable sets of behavior.

Figure 10: Volume rendering of a hixel data set generated by sampling images 3200 values sampled from a Poisson distribution and 9600 values sampled from a normal distribution.There are 512x512 hixels in this data set, each with 128 bins. The shortest axis in the images corresponds to histogram bins, thus a spatially higher location along that axis indicates a higher function value. Color and opacity are used to illustrate the density of the samples. Thus the lower, right corner of shows a hixel with 2 distinct probable function values; the smaller function value is less probable than the larger.



(a) (b)

Figure 11: Basins of minima (a) and of maxima (b) are shown for the mixture model data set. By computing basins on sheets we are able to identify prominent features associated with each process in the mixture model.

There are 512x512 hixels in the mixture model data set, each with 128 equally-sized bins. The shortest axis in the images corresponds to histogram bins, thus a spatially higher location along that axis indicates a higher function value. The data is a mixture of two distributions at each hixel with 3200 values sampled from a Poisson distribution and 9600 values sampled from a Gaussian distribution. Hue and opacity are used to illustrate the density of the samples. When the number of values in a hixel bin is zero, the bin is rendered as a transparent blue. When the number of values in a bin is large, the bin is rendered as an opaque red.

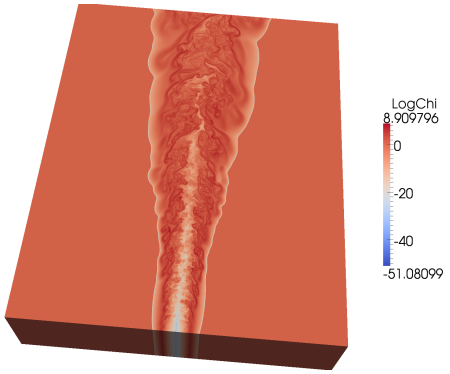
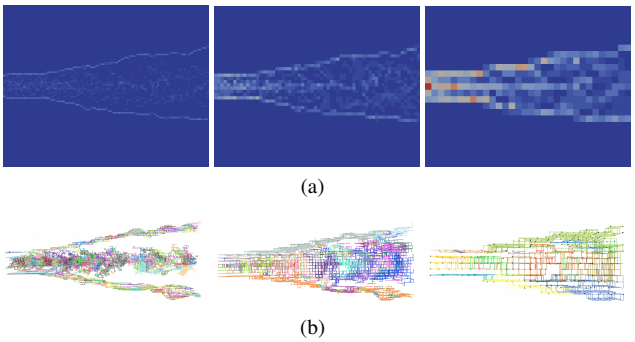


Figure 12: The logarithm of the χ field in the lifted ethylene jet data set with 1.3 billion grid points.

The Poisson and Gaussian distributions for each hixel have different parameter values that vary smoothly over the image. The Poisson lambda parameter is a maximum of 100 at five source points arranged in a circle and decreases to a minimum value of approximately 12 proportional to the distance to the nearest of these points. The Gaussian mean (standard deviation) is a minimum (maximum) at 4 points arranged in the same circle as the 5 Poisson source points. The mean varies from 32 to 108 while the standard deviation varies from 16 to 3.8. Topological basins of minima and maxima are shown in Figure 11 for all sheets with pmi greater than zero. Our approach clearly extracts separate sheets belonging to the two processes which then allows for topological analysis to identify the prominent features of each distribution.

To compare against down-sampling a large-scale dataset, we also demonstrate results of this method on a hixelated data set generated from the log of the χ field of a lifted ethylene jet combustion data set with 1.3 billion grid points, see Figure 12. The contingency tables between each pair of hixels are computed using observations between neighboring vertices along shared hixel faces. Fig. 13 shows the number of buckets per hixel with block sizes of 16 (top-left), 32 (top-middle), and 64 (top-right). The color map ranges from blue at 1 bucket per hixel to red with 27 buckets per hixel and, as is to be expected, the number of buckets per hixel increases significantly as the block size increases. On the bottom the basins of maxima are shown for corresponding block sizes.



(a)

(b)

Figure 13: (a) The number of buckets per hixel is displayed for block sizes of 16 (left), 32 (middle) and 64 (right). Blue regions have 1 bucket per hixel while the maximum number of buckets per hixel is 27 and is shown in red. (b) The basins of maxima are shown for corresponding block sizes.

6 Fuzzy Isosurfacing

When down-sampling larger datasets, hixels enable preserving the presence of an isosurface within the data. In particular, when hixels store the counts of all function values present within a block, we can use that to compute the likelihood of the presence of an isosurface within that block. Given a hixel h_i and an isovalue κ, we *slice* the histogram at κ and compute the number of voxels above and below κ. These two counts, a and b for the count above and below, respectively, give an indication to how frequently that isosurface may exist within the block. An alternate interpretation is these values can measure an approximation for the surface area of the isosurface within the block, Figure 14 visualizes this slicing process.

Using the values a and b, we can then compute a likelihood field. We let $g = \frac{a}{b} - \frac{b}{a}$. For hixels that have $a = b$, g takes on the value 0, while $g > 0$ for hixels that are strongly above κ and $g < 0$ for hixels that are strongly below. If $a = 0$, we set $g = b$, and when $b = 0$ we set $g = a$. By volume rendering the g field we can get a "fuzzy" depiction of where the isosurface exists in a hixelated field. By comparison, naive down-sampling of the scalar field could either move or destroy isovalues. By visualizing the field g we get a more honest depiction about where that isovalue was originally in the dataset, and can thus preserve that information.

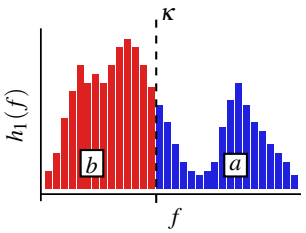
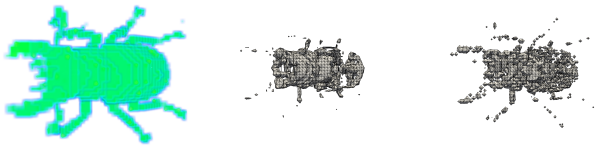


Figure 14: Slicing a histogram. For an isovalue κ, we can use the histogram to count the number of voxels with function values above (a) and below (b) i.



(a) (b) (c)

Figure 15: For hixel size 16^3, we compare the (a) likelihood field g volume rendered to the isosurfaces computed for $\kappa = 580$ for the (b) mean and (c) lower-left down-sampling. Because of the loss of information, the isosurfaces disconnect and lose features.

Figure 16 shows visualizations of the stag dataset for $\kappa = 580$, down-sampled from its original size of $832 \times 832 \times 494$ to $208 \times 208 \times 123$, $104 \times 104 \times 61$, $52 \times 52 \times 30$, $26 \times 26 \times 15$, and $13 \times 13 \times 7$. Hixels of block size b^3 used $2b^2$ bins. By tracking a histogram of values, at lower resolutions we can preserve the fidelity of the isosurface and display a more expressive view of the data. Using only a single value, it is challenging to preserve the thin features of the isosurface, as the legs, antenna, and mandibles are hard to preserve. Figure 15 shows a side-by-side comparison of the isosurfaces produced at $\kappa = 580$ for the mean and lower-left fields as compared to the volume rendering of the g field when the hixel block size is 16^3.

7 Discussion

By unifying the representations of large scalar fields from various modalities, hixels enable the analysis and visualization of data that would otherwise be challenging to process. We focus on exploring three proof of concept uses of hixels: generating topological views of the data by sampling hixels; computing topological basins on sheets of statistically associated hixels, and visualizing fuzzy isosurfaces, indicated by the likelihood that a hixel contains the isosurface. Moreover, we believe hixels provide a more honest view of data, that would otherwise be lost or discarded by simple down-sampling techniques.

While hixels have utility, they present a number of challenges and open questions to explore. One important question regards information preserved by the hixels vs. resolution loss. A study is required to explore the appropriate number of bins per hixel as well as persistence thresholds for bucketing and mode seeking algorithms. The performance of hixels was not currently emphasized in our work, but the complexity of many techniques used here should allow for scaling to larger data. Additional research is required to find a balance between data storage allotted for the histograms vs. feature preservation. Finally, further studies to distinguish which topological features can be easily preserved by hixelation from those which cannot is required.

Acknowledgements

This work was supported by the Department of Energy Office of Advanced Scientific Computing Research, award number DE-SC0001922. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. This work was also performed under the auspices of the US Department of Energy by the Lawrence Livermore National Laboratory under contract nos. DE-AC52-07NA27344 and LLNL-JRNL-412904L and by the University of Utah under contract DE-FC02-06ER25781. Additionally, this work was supported by National Science Foundation awards IIS-0904631, IIS-0906379, and CCF-0702817. We are grateful to Dr. Jacqueline Chen for the combustion datasets and M. Eduard Göller, Georg Glaeser, and Johannes Kastner for the stag beetle dataset.

References

[1] R. J. Adler and J. E. Taylor. *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer, New York, Feb. 2010.

[2] P.-T. Bremer, H. Edelsbrunner, B. Hamann, and V. Pascucci. A topological hierarchy for functions on triangulated surfaces. *IEEE Trans. Vis. Comp. Graph.*, 10(4):385–396, 2004.

[3] P.-T. Bremer, G. Weber, V. Pascucci, M. Day, and J. Bell. Analyzing and tracking burning structures in lean premixed hydrogen flames. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):248–260, 2010.

[4] R. Brown. Animated visual vibrations as an uncertainty visualisation technique. In *GRAPHITE '04: Proceedings of the 2nd international conference on Computer graphics and interactive techniques*, pages 84–89, 2004.

[5] P. Burman and W. Polonik. Multivariate mode hunting: Data analytic tools with measures of significance. *J. Multivariate Analysis*, pages 1198–1218, 2009.

[6] M.-Y. Cheng and P. Hall. Mode testing in difficult cases. *The Annals of Statistics*, 27(4):pp. 1294–1315, 1999.

[7] B. J. Debusschere, H. Najm, P. P. Pébay, O. Knio, R. Ghanem, and O. P. L. Maître. Numerical challenges in using polynomial chaos. *SIAM J. Scientific Computing, Special Issue on Uncertainty Quantification*, 26(2):698–719, 2004.

[8] H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete and Computational Geometry*, 30(1):87–107, 2003.

[9] D. Estep, M. Larson, and R. Williams. Estimating the error of numerical solutions of systems of reaction-diffusion equations. *Mem. Amer. Math Soc.*, 146, 2000.

[10] R. Forman. A user's guide to discrete Morse theory. In *Proc. of the 2001 Internat. Conf. on Formal Power Series and Algebraic Com-*

(a) Hixels (g field)

(b) Mean of each block

(c) Lower-left corner of each block

Figure 16: Volume rendering of g field for Stag, as compared to down-sampling with the mean and lower left corner of each block. The dataset is originally $832 \times 832 \times 494$, and from left-to-right we shown hixel sizes of 4^3 to 64^3, with all powers of 2 in between. To volume render the g field, a triangle function centered at 0 is used for opacity, for the mean and lower-left fields, a triangle function centered at 580 is used. The color map is rainbow, red are high values, blue are low, and green is in the middle of the range.

binatorics, A special volume of Advances in Applied Mathematics, page 48, 2001.

[11] J. H. Friedman and N. I. Fisher. Bump hunting in high dimensional data. *Statistics and Computing*, 9:123–143, 1999.

[12] G. Grigoryan and P. Rheingans. Point-based probabilistic surfaces to show surface uncertainty. *IEEE Trans. Vis. Comp. Graph.*, 10(5):546–573, 2004.

[13] A. Gyulassy, P.-T. Bremer, B. Hamann, and V. Pascucci. A practical approach to Morse-Smale complex computation: Scalability and generality. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1619–1626, 2008.

[14] A. Gyulassy, M. Duchaineau, V. Natarajan, V. Pascucci, E. Bringa, A. Higginbotham, and B. Hamann. Topologically clean distance fields. *IEEE Trans. Vis. Comp. Graph.*, 13(6):1432–1439, 2007.

[15] J. A. Hartigan and P. M. Hartigan. The dip test of unimodality. *Annals of Statistics*, 13:70–84, 1985.

[16] H. King, K. Knudson, and N. Mramor. Generating discrete morse functions from point data. *Experimental Mathematics*, 14(4):435–444, 2005.

[17] D. E. Laney, P.-T. Bremer, A. Mascarenhas, P. Miller, and V. Pascucci. Understanding the structure of the turbulent mixing layer in hydrodynamic instabilities. *IEEE Trans. Vis. Comput. Graph*, 12(5):1053–1060, 2006.

[18] T. Lewiner, H. Lopes, and G. Tavares. Applications of forman's discrete morse theory to topology visualization and mesh compression. *IEEE Trans. Vis. Comp. Graph.*, 10(5):499–508, 2004.

[19] A. Luo, D. Kao, and A. Pang. Visualizing spatial distribution data sets. In *Proceedings of the symposium on Data visualisation 2003*, pages 29–38, 2003.

[20] Y. Matsumoto. *An Introduction to Morse Theory*, volume 208 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI, 2002.

[21] J. Milnor. *Morse Theory*. Princeton University Press, New Jersey, 1963.

[22] J. Murphy, D. Sexton, D. Barnett, G. Jones, M. Webb, M. Collins, and D. Stainforth. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, 430:768–772, 2004.

[23] R. A. Osorio and K. Brodlie. Contouring with uncertainty. In *6th Theory and Practice of Computer Graphics Conference*, pages 59–66, 2008.

[24] T. Pfaffelmoser, M. Reitinger, and R. Westermann. Visualizing the positional and geometrical variability of isosurfaces in uncertain scalar fields. *Comput. Graph. Forum*, 30(3):951–960, 2011.

[25] K. Pöthkow and H.-C. Hege. Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *IEEE Trans. Vis. Comput. Graph.*, 17(10):1393–1406, Oct. 2011.

[26] K. Pöthkow, B. Weber, and H.-C. Hege. Probabilistic marching cubes. *Comput. Graph. Forum*, 30(3):931–940, 2011.

[27] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, V. Pascucci, and C. Johnson. Ensemble-vis: A framework for the statistical visualization of ensemble data. In *Proc. IEEE ICDM Workshop on Knowledge Discovery from Climate Data: Prediction, Extremes, and Impact*, pages 233–240, 2009.

[28] P. Rheingans and S. Joshi. Visualization of molecules with positional uncertainty. In *Data Visualization*, pages 299–306, 1999.

[29] P. J. Rhodes, R. S. Laramee, R. D. Bergeron, and T. M. Sparr. Uncertainty visualization methods in isosurface rendering. In *EUROGRAPHICS 2003 Short Papers*, pages 83–88, 2003.

[30] G. P. M. Rozál and J. A. Hartigan. The map test for multimodality. *Journal of Classification*, 11:5–36, 1994. 10.1007/BF01201021.

[31] W. Schoutens. *Stochastic Processes and Orthogonal Polynomials*. Springer-Verlag, New York, 2000.

[32] D. Xiu and G. E. Karniadakis. The Wiener-Askey polynomial chaose for stochastic differential equations. *SIAM J. Sc. Comput.*, 24:619–644, 2002.