# Analyzing Information Transfer in Time-Varying Multivariate Data

Chaoli Wang* Michigan Tech
Hongfeng Yu† SNL
Ray W. Grout‡ NREL
Kwan-Liu Ma§ UC Davis
Jacqueline H. Chen¶ SNL

**ABSTRACT**

Effective analysis and visualization of time-varying multivariate data is crucial for understanding complex and dynamic variable interaction and temporal evolution. Advances made in this area are mainly on query-driven visualization and correlation exploration. Solutions and techniques that investigate the important aspect of causal relationships among variables have not been sought. In this paper, we present a new approach to analyzing and visualizing time-varying multivariate volumetric and particle data sets through the study of information flow using the information-theoretic concept of transfer entropy. We employ time plot and circular graph to show information transfer for an overview of relations among all pairs of variables. To intuitively illustrate the influence relation between a pair of variables in the visualization, we modulate the color saturation and opacity for volumetric data sets and present three different visual representations, namely, ellipse, smoke, and metaball, for particle data sets. We demonstrate this information-theoretic approach and present our findings with three time-varying multivariate data sets produced from scientific simulations.

## 1 INTRODUCTION

Detecting interdependencies and causal relationships among multiple variables is one of the most important issues in multivariate time series data analysis. This issue is receiving increasing attention as our scientists' ability to generate data multiples every year. Applications of studying correlation and causation can be found in many fields of science, such as physics, economics, and physiology, to name a few. For example in brain studies, information about the interaction among recorded channels of an electroencephalogram (EEG) can aid clinical practice by identifying the region of the brain that is acting as a recruiting focus in epilepsy. In atmospheric prediction, improving the initial conditions through the investigation of information flow can reduce uncertainty in predictions at other locations and with respect to other dynamical variables.

Commonly-used techniques for the estimation of dependencies are linear cross-correlation and mutual information. However, these measures share the property of being *symmetric* and therefore are not suited for assessing causality within relationships. To study the *directional* aspect of interactions, Schreiber took a general nonparametric test of causality based on information theory and introduced the concept of *transfer entropy* [14] for quantifying the flow of information between time series. With minimal assumptions about the dynamics of the systems and the nature of their coupling, this information-theoretic measure can quantify the exchange of information between two systems, separately for each direction.

In visualization, existing work on time-varying multivariate data centers on query-driven visualization and correlation structure exploration. The challenging problem of identifying causal relationships among different variables in the time series has not been paid due attention. To respond, we focus on analyzing and visualizing causal relationships using transfer entropy. Besides being a general and easily applicable measure, two additional advantages regarding the nature of transfer entropy make it very attractive for our usage: First, this model-free approach does not assume anything about the coupling between the variables. Second, it allows us to detect statistical dependencies not limited to linear statistics and to reveal all types of temporal correlations.

The contributions of our work are: First, we have utilized the transfer entropy to study causal connections among time-varying multivariate data, which is seldom investigated in visualization due to the lack of appropriate techniques. Second, we have developed multiple views using information and scientific visualization techniques for effective display of information transfer. Third, we applied this new approach to visualizing volumetric and particle data sets, and verified the results with combustion scientists. Fourth, we expand the concept of transfer entropy by defining *relative transfer entropy* and discussing how to generalize the original pair-wise transfer entropy to simultaneously handle multiple variables. We also point out the challenges associated with this generalization.

## 2 RELATED WORK

**Information Theory and Visualization** Concepts of information theory have been applied to many areas of visualization including view selection for polygon and volume rendering [19, 2], camera path planning for focus of attention and visualizing time-varying data [20, 7], and analysis of the importance of multifield and time-varying data [6, 21]. These solutions are useful for coping with the massive growth of data in both scale and complexity. This paper leverages the concept of transfer entropy to study the influences among multiple variables.

**Multivariate Data Analysis and Visualization** Multivariate data analysis and visualization has gained considerable attention in recent years. Among them, one stream of research focused on query-driven visualization such as compound range queries [17] and fuzzy queries using textual pattern matching [4]. Another research stream focused on correlation study such as point-wise correlation coefficients [13, 10, 12, 4, 18] and gradient similarity measure [13]. Although correlation is often used to study the relationships between variables, it does not *imply* (or *suggest*) causation. In visualization, little work has been done to identify the causal relationships among multivariate data.

Multivariate data pose a unique challenge for visualization due to the large number of variables considered and the amount of data presented to the user simultaneously. Woodring and Shen [23] used boolean set operations to select voxels of interest and combine different variables together into a single volume for visualization. Sisneros et al. [16] presented a point classification algorithm that fuses key aspects of multiple data attributes into a single image for concurrent viewing. Other researchers explored the use of information visualization techniques to show variable relations. For example, Sauber et al. [13] developed *multifield-graph* for a complete visualization of scalar fields and their correlations. Qu et al. [12] created a weighted complete graph to reveal the overall correlation of all data attributes. Blaas et al. [1] used scatterplots in the high-dimensional multifield feature space and allowed arbitrary projection. Jänicke

*e-mail: chaoliw@mtu.edu
†e-mail: hyu@sandia.gov
‡e-mail: ray.grout@nrel.gov
§e-mail: ma@cs.ucdavis.edu
¶e-mail: jhchen@sandia.gov

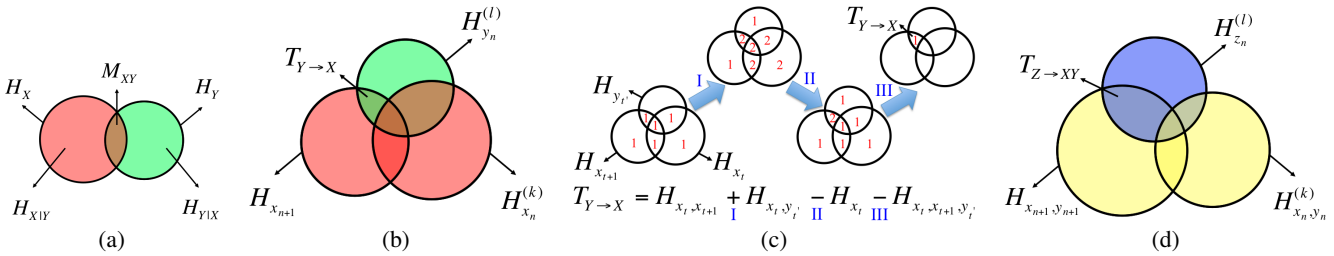



Figure 1: (a) and (b) illustrate the relations of entropies $H_X$, $H_Y$, mutual information $M_{XY}$, conditional entropies $H_{X|Y}$, $H_{Y|X}$, and transfer entropy $T_{Y\to X}$ using the Venn diagram. (c) illustrates the simplified calculation of $T_{Y\to X}$ where each step considers one more term in Equation 3 and the red number indicates how many times each portion gets counted. (d) shows a generalization of transfer entropy to three variables.

et al. [5] transformed multivariate data from their high-dimensional attribute space to a 2D *attribute cloud* for brushing and linking.

**Visualization of Causal Relations** There are only a few research efforts in causality visualization. Simple visualization includes the use of a node-link diagram such as Hasse diagrams to offer intuitive viewing of causal relations. Ware et al. [22] introduced the *visual causal vector*, an animation-enhanced metaphor that represents the perceptual impression of a causal relation between two graphical entities. Elmqvist and Tsigas [3] presented *growing polygons* for visualizing causal relations and information flow in a complex system with many nodes and relations. In scientific visualization, Silver et al. [15] juxtaposed 4D space-time vector fields where one contains a source variable and the other the response field to highlight the topological relationship between the two fields. In this work, we utilize time plot and circular graph to show pair-wise information flows and present different techniques to visualize causal relationships for volumetric and particle data sets.

## 3 Measuring Information Transfer

### 3.1 Transfer Entropy

To identify causal dependency or information transfer, we need to incorporate dynamical structure by investigating *transition* probabilities instead of *static* probabilities. Let us consider a system that can be approximated by a stationary Markov process of order $k$, that is, the conditional probability $X$ in state $x_{n+1}$ is independent of the state $x_{n-k}$: $p(x_{n+1}|x_n,\ldots,x_{n-k}) = p(x_{n+1}|x_n,\ldots,x_{n-k+1})$. Let us denote $x_n^{(k)} = (x_n,\ldots,x_{n-k+1})$ for words of length $k$, where the subscript denotes the state (or time step) and the superscript denotes the length of states (or time steps) considered. According to Schreiber [14], the *transfer entropy* between two variables $X$ and $Y$ is defined as follows

$$T_{Y\to X} = \sum p(x_{n+1}, x_n^{(k)}, y_n^{(l)}) \log \frac{p(x_{n+1}|x_n^{(k)}, y_n^{(l)})}{p(x_{n+1}|x_n^{(k)})}, \quad (1)$$

where $T_{Y\to X}$ denotes the influence of $Y$ on $X$. The most natural choices for $l$ are $l = k$ (the same number of time steps is considered for both $X$ and $Y$) or $l = 1$ (only one time step for $Y$ is considered at a time). Usually, the latter is preferable due to a lower computational cost. An illustration of transfer entropy and related terms is shown in Figure 1 (a) and (b).

Transfer entropy can be treated as a version of mutual information operating on conditional probabilities. It shares some of the desired properties of mutual information but takes the dynamics of information transport into account. As shown in Figure 1 (b), $T_{Y\to X}$ can be regarded as the information about future observations $x_{n+1}$ gained from past observations of $x_n^{(k)}$ and $y_n^{(l)}$ minus the information about future observations $x_{n+1}$ gained from past observations of $x_n^{(k)}$ only. Thus, it is the *information flow* from $Y$ to $X$. $T_{Y\to X}$ is explicitly nonsymmetric under the exchange of $X$ and $Y$ (a similar expression exists for $T_{X\to Y}$) and can thus be used to detect the directed exchange of information between the two time series. Transfer entropy has been used to study information flow among time series data in areas such as spatiotemporal systems [14], physiological studies [14, 11], financial markets [9], and sensorimotor networks [8]. We utilize transfer entropy to analyze and visualize information flow in scientific data sets.

### 3.2 Relative Transfer Entropy

As we can see from Figure 1 (b), transfer entropy $T_{Y\to X}$ shows the *amount* of influence of $Y$ on $X$. We point out that in other cases, it is also meaningful to consider the *rate* of influence by taking into account the amount of information in $X$ and $Y$, i.e., $H_{x_{n+1}}$ and $H_{y_n}^{(l)}$. As such, we define the *relative transfer entropy* as a normalized version of transfer entropy, i.e.,

$$RT_{Y\to X} = \frac{T_{Y\to X}}{\sqrt{H_{x_{n+1}} H_{y_n}^{(l)}}}. \quad (2)$$

If $H_{x_{n+1}} H_{y_n}^{(l)} = 0$, then $T_{Y\to X} = 0$ and we define $RT_{Y\to X} = 0$. $RT_{X\to Y}$ can be defined similarly.

### 3.3 Multiple Variables and Multiple Time Steps

The original definition for transfer entropy expressed in Equation 1 only operates on two variables $X$ and $Y$. We can generalize this to simultaneously handle multiple variables by replacing a single variable with a set of variables. In Figure 1 (d), we show an example where we consider the transfer entropy from variable $Z$ to the set of two variables $X$ and $Y$. To compute this, we use joint entropies $H_{x_{n+1},y_{n+1}}$ and $H_{x_n,y_n}^{(k)}$ to replace entropies $H_{x_{n+1}}$ and $H_{x_n}^{(k)}$, respectively, shown in Figure 1 (b). This treatment can be extended to study the general case of information transfer between a set of $s$ variables and another set of $r$ variables.

Note that Equation 1 already gives a general form in terms of time steps included. As such, the most general form of transfer entropy can deal with a set of $s$ variables over $k$ time steps and another set of $r$ variables over $l$ time steps. The central issue that prevents us from computing this generalized transfer entropy is the huge cost involved. As we shall see in Section 3.4, adding one more variable or one more time step would increase the dimension of joint histograms computed. For instance in Figure 1, if $k = 1$ and $l = 1$, we need 3D joint histograms for (b) but 5D histograms for (d). If we use the same number of bins for each dimension, the computation and storage costs for joint histograms will increase dramatically. Solutions that can efficiently evaluate the histogram for a high-dimensional data is the key to break the "curse of dimensionality", which we leave as our future work.

### 3.4 Simplified Calculation

For fast calculation of transfer entropy, we simplify Equation 1 by letting $k = 1$ and $l = 1$ and rewrite transfer entropy in the more convenient, albeit less intuitive form (refer to Figure 1 (c))

$$T_{Y\to X} = H_{x_t,x_{t+1}} + H_{x_t,y_{t'}} - H_{x_t} - H_{x_t,x_{t+1},y_{t'}}, \quad (3)$$

where $t = t' + \Delta t$ and $\Delta t$ ($\geq 0$) is some lag time. If $\Delta t > 0$, we imply that *past* time steps of $Y$ influence *current* or *future* time steps of $X$. $H_{x_t,x_{t+1}}$, $H_{x_t,y_{t'}}$, and $H_{x_t,x_{t+1},y_{t'}}$ are the joint entropies. Similar to Equation 3, we have

$$T_{X\to Y} = H_{y_{t'},y_{t'+1}} + H_{y_{t'},x_t} - H_{y_{t'}} - H_{y_{t'},y_{t'+1},x_t}. \quad (4)$$

For discrete data, entropy computation usually takes the histogram of data and uses the normalized heights as the probabilities. Note that in Equations 3 and 4, we only compute entropies and joint entropies. As such, we only need to compute joint histograms for $(x_t, x_{t+1}, y_{t'})$ and $(y_{t'}, y_{t'+1}, x_t)$. These 3D histograms are used to compute joint entropies $H_{x_t,x_{t+1},y_{t'}}$ and $H_{y_{t'},y_{t'+1},x_t}$. Other joint entropies (entropies) in Equations 3 and 4 can be computed by projecting the 3D histograms to 2D (1D) accordingly.

### 3.5 Calculation for Volumetric and Particle Data

For volumetric data sets, we take the Eulerian view and observe the information flow at fixed regions in the space through which the time-dependent phenomena evolve. We take a block-wise approach and partition the data into spatial blocks. Then, we evaluate the influence relation between any pair of variables within each individual data block. The computation of transfer entropy takes a data block as input and computes its joint histograms for $(x_t, x_{t+1}, y_{t'})$ and $(y_{t'}, y_{t'+1}, x_t)$ in Equations 3 and 4. Finally, we compute the transfer entropy of a particular time step as the summation of the transfer entropy values of all data blocks in the volume.

For particle data sets, we take the Lagrangian view and follow particle motions through space and time. We partition all particles at a time step into different groups according to certain criteria (a domain-specific partitioning example is presented in Section 5.2). Since particles can be traced forward and backward through time via their unique IDs, it is straightforward to build connections between a region of interest (particles within that region or group are selected) and its correspondence over time (identify where those particles have been drifted to). We thus compute the transfer entropy for each particle group where the particle correspondence in the joint histogram computation follows its ID.

When we consider information transfer in a block-wise or group-wise manner, we implicitly assume that data outside the block or group do not have influence on the block or group being investigated. This assumption is not entirely correct. For example, some of the data quantities could be derivatives, which are dependent on the values of neighboring blocks. Nevertheless, we make this assumption to simplify our transfer entropy calculation.

Two parameters affect the efficiency and effectiveness of joint histogram calculation. The first parameter is the block size (for volumetric data) or the number of particle in each group (for particle data). For a meaningful evaluation of histogram distribution, the number of voxels or particles considered must be at least an order of magnitude higher than the number histogram bins chosen. The block size also affects the efficiency of computation. Using smaller blocks will take more time to compute the transfer entropy for the whole volume. We should in general choose a block size that is in proportion to the volume size. As such, a good practice is to determine the desired number of blocks we want (typically in the order of hundreds or thousands), and then determine the proper block size. The second parameter is the number of histogram bins. Normally, using 256 bins that uniformly sample the 1D histogram is a good choice. In Section 6.1, we will provide the results with different parameter values chosen and discuss their impacts on computation efficiency and accuracy.

## 4 Visualizing Information Transfer

We present two different ways to visualize information transfer. One way is to utilize information visualization techniques to display information transfer in a separate view. Another way is to directly visualize information transfer in the spatial view. We use time plot and circular graph for overview of influence for all pairs of variables. For visualization with the spatial view, we adjust the color saturation and opacity of different variables to indicate influence relations in volumetric data blocks. For particle data, we present three different visual representations, namely, metaball, ellipse, and smoke, to simultaneously visualize the influences among different groups of particles.

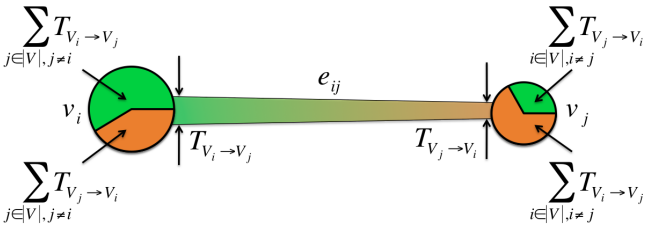



Figure 2: The mapping of transfer entropies to two vertices (corresponding to variables $V_i$ and $V_j$) and the edge in our circular graph.

### 4.1 Time Plot and Circular Graph

As shown in Figure 7, a straightforward way to show information transfer among variables is to draw time plots with the two axes for time step and transfer entropy respectively. Although time plots clearly display variable influence over time, it is difficult to observe the relation of one variable with all other variables in this view. Alternatively, we also display information transfer using the circular graph. As illustrated in Figure 2, we draw a circle at vertex $v_i$ to denote the total outgoing and incoming influence corresponding to variable $V_i$ at a certain time step. The size of circle shows the total amount of influence with green/orange for outgoing/incoming influence. An edge $e_{ij}$ indicates the influence between variables $V_i$ and $V_j$. We map transfer entropy $T_{V_i\to V_j}$ to edge width at vertex $v_i$ and $T_{V_j\to V_i}$ to $v_j$. Green/orange indicates more outgoing/incoming influence. The color saturation is adjusted according to the absolute difference between the two transfer entropies so that pairs of variables with larger difference would stand out. Edge width and color in between are linearly interpolated. To reduce the occlusion among edges, we first sort all edges in the decreasing order of their average thickness and then draw the edges accordingly. As shown in Figure 3, such circular graphs are intuitive for inferring the relation among all pair of variables, as a group and as an individual. Since every time step corresponds to such a graph, we can produce a time-varying graph showing information transfer.

### 4.2 Visualization for Volumetric Data

To visualize information transfer for volume data sets, we modulate the color saturation and opacity of variables for each data block with its normalized transfer entropy value. Given two variables $V_i$ and $V_j$, we use $T_{V_i\to V_j}$ to modulate $V_i$ and $T_{V_j\to V_i}$ to modulate $V_j$. As shown in Figure 4, data blocks of higher amount of influence would be highlighted as data blocks of lower amount of influence are attenuated. In this way, the user is able to intuitively infer information transfer between variables at different spatial regions over the time series when all time steps are animated.

### 4.3 Visualization for Particle Data

**Ellipse and Smoke Representations** For particle data sets, information transfer can be mapped to visual properties, such as color, opacity, size, or shape, of each individual particle in a group. In this case, each particle in the group undergoes the same change. We perturb the conventional disk representation (i.e., the view-aligned sphere) of particle to an ellipse representation. The ellipse's color maps to influence relation while the lengths of its two axes indicate their respective transfer entropy values. As shown in Figure 6, given

| Hurricane Isabel, 40GB in total $(x, y, z, v, t)$ = (500, 500, 100, 9, 48) | | Ionization Front Instability, 220GB in total $(x, y, z, v, t)$ = (600, 248, 248, 8, 200) | | Lifted-Flame Combustion, 201GB in total $(n, v, t)$ = (30M, 6, 300) | |
|---|---|---|---|---|---|
| variable | description | variable | description | variable | description |
| QC | cloud moisture mixing ratio | H | H mass abundance | YOH | hydroxyl radical mass fraction |
| QI | cloud ice mixing ratio | H+ | H+ mass abundance | YCH2O | formaldehyde mass fraction |
| QG | graupel mixing ratio | He | He mass abundance | YHO2 | hydroperoxyl mass fraction |
| QR | rain mixing ratio | He+ | He+ mass abundance | Z | nitrogen-based mixture fraction |
| QS | snow mixing ratio | He++ | He++ mass abundance | T | temperature |
| QV | water vapor mixing ratio | H2 | H2 mass abundance | G | mixture fraction gradient |
| PR | pressure | PD | total particle density | | |
| TC | temperature | GT | gas temperature | | |
| WS | wind speed (magnitude) | | | | |

Table 1: The three time-varying multivariate data sets and their variables tested. $x$, $y$, $z$, and $t$ are for the three spatial and the temporal dimensions, respectively. $v$ is the number of variables. $n$ is the number of particles. For the combustion data set, we investigated a subset of particles (around 109K, total 747MB) that is believed to participate in flame stabilization.

a pair of variables $(V_i, V_j)$, we use orange for $T_{V_i \to V_j} > T_{V_j \to V_i}$, white for $T_{V_i \to V_j} = T_{V_j \to V_i}$, and blue for $T_{V_i \to V_j} < T_{V_j \to V_i}$. The particles are rendered according to the visibility order. This ellipse representation, however, would lead to visual clutter if the number of particles drawn is fairly large. Therefore, as an option, we also use the smoke representation to reduce the clutter by decreasing the radii and opacities of particles. The color mapping stays the same. The influences among variables in different spatial regions will be more easily discernible.

**Metaball Representation** Another way of visualizing information transfer is to treat the group of particles as an entirety. Instead of adjusting the properties of individual particles, we construct two transparent layers which enclose all particles in the group. The two layers have distinct colors indicating directional transfer entropies and their sizes and enclosure relation show the corresponding influences between the two variables. To obtain this double-layer effect, one solution is to define the isosurfaces using two different thresholds in the same density field. However, this solution does not guarantee that the two surfaces agree with each other when the difference between the isovalues of two surfaces is large. Alternatively, we construct two density fields using two different radii for the same group of particles, and define the isosurfaces in the two fields using the same threshold. As shown in Figure 8, this treatment gives us desirable results.

## 5 Results

To demonstrate our approach, we investigated three time-varying multivariate simulation data sets and studied variable causal relationships. The first two data sets are volumetric: Hurricane Isabel data set from climate research and ionization front instability data set from astronomy research. They were made available through IEEE Visualization 2004 and 2008 Contests, respectively. The third one is a lifted-flame particle data set from combustion research. The variables used for the three data sets are listed in Table 1.

### 5.1 Hurricane Isabel and Ionization Front Instability

The Hurricane Isabel simulation data set was courtesy of NCAR and NSF. We set the block size to $20 \times 20 \times 20$ and the histogram size for each variable to 256. We used Equations 3 and 4 with $\Delta t = 0$ and calculated transfer entropy for every pair of variables and every pair of neighboring time steps.

In Figure 4, we show the visualization of information transfer among a pair of variables: QI and WS. The rendering of four selected time steps is displayed where we modulated color saturation and opacity for the variables according to their transfer entropy values. We can observe strong couplings of influence for the pair of variables in space and time. The regions that preserve original colors and opacities highlight data blocks with strong influence, which are around the hurricane's eye. The regions with attenuated colors and opacities indicate less inter-influence. Our visualization thus displays strong influence regions as the focus while keeping weak influence regions as the context. Regions with little influence are not displayed. The spatio-temporal coherence of the data leads to meaningful visualization as we observe information transfer across multiple continuous spatial data blocks over time. If we examine closely, we can also observe the change of influence over time. In early time steps, QI influences WS more in general as we can see more saturated red regions in the leftmost image. In later time steps, the influence of QI over WS drops as we can see less saturated red regions in the rightmost image. Sine the transfer entropy is calculated in the block-wise manner, we can see the variation of inter-influence among individual blocks as well.

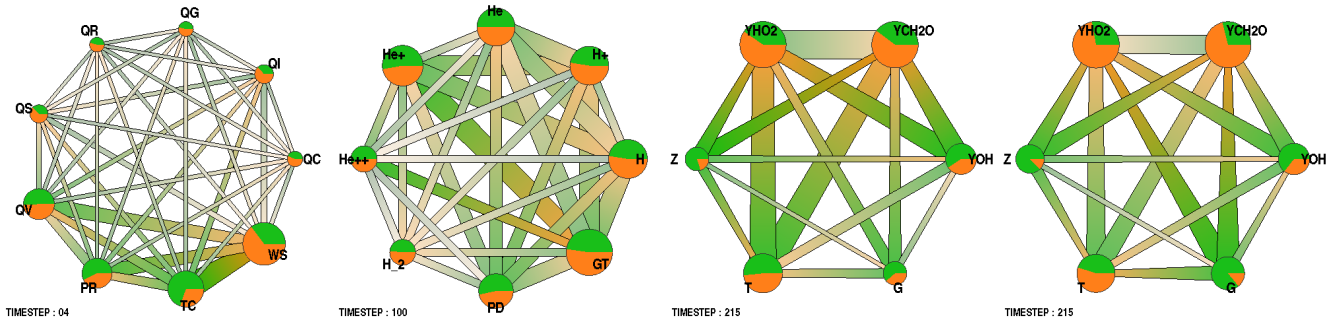
Scientists at LANL and SDSC performed three-dimensional radiation hydrodynamical calculations of ionization front instabilities to study a variety of phenomena in interstellar medium such as the formation of stars. To compute the transfer entropy, we set the block size to $30 \times 31 \times 31$ and the histogram size for each variable to 256. Figure 5 shows the visualization of information transfer for the He+ and H2 pair. We can observe that strong influence regions are around the plane on the front.

### 5.2 Lifted-Flame Combustion

The lifted-flame combustion particle data set was provided by SNL scientists. The combustion scientists employed a high-order finite difference algorithm to solve the fully compressible Navier-Stokes and chemically reacting species equations in their simulations and produced large-scale time-varying multivariate volumetric and particle data sets. One of the issues in their investigation is to identify the causal relationships among dozens of variables. In this experiment, we specifically focused on the particle data set. The scientists provided us with guidance for particle selection, partition, and tracing. The causal relationships discovered in this study have been confirmed by the scientists.

The spatial extent of the combustion data set is (2025, 1600, 400). We first performed a range query to retrieve particles of interest with $x \in (720, 740)$ and $y \in (533, 693) \cup (906, 1066)$ at time step 155. The particles were selected as those passing through a slab in the axial direction $x$, in the transverse direction $y$, and $y$ spanning the transition region which includes particles from both the hot air coflow and the interior of the fast moving fuel jet. The scientists conjectured that these particles were likely to participate in flame stabilization. A total of 109,483 particles were selected. We traced selected particles forward and backward over the 300 time steps. After that, the selected particles were partitioned based on the spatial range along the $y$ axis at time step 150. The resulting 30 groups have a nearly even number of particles in each group.

In Figure 6, we show the visualization of information transfer among all groups of particles for a pair of variables: YOH and T.




(a) (b) (c) (d)

Figure 3: Using circular graph to show information transfer. (a) hurricane, (b) ionization, (c) and (d) combustion. (a)-(c) show transfer entropy while (d) shows relative transfer entropy. Each graph shows the influence among all pairs of variables at the selected time step. Our graphs can answer questions such as which variable has the most total outgoing and incoming influence (find the largest circle), which variable's total outgoing influence is more than its total incoming influence and by how much (read green and orange portions of the circles), and for a pair of variables, which variable influences the other more and by how much (compare edge color and width difference at both ends).

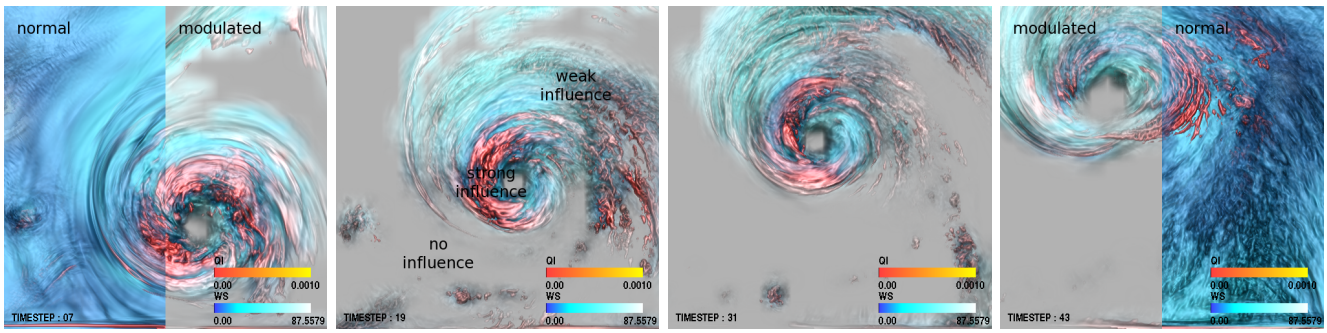



Figure 4: Visualization of information transfer on selected time steps of the hurricane data set. The QI and WS pair is shown. Regions that preserve original colors and opacities indicate a higher amount of influence. Comparisons of normal rendering and modulated rendering are given in the first and last images. Strong influence regions are around the hurricane's eye.

The rendering of four selected time steps is displayed with the ellipse representation. The rightmost image shows the rendering with the smoke representation. For ellipse rendering, the influence relation can be read from the colors of the particles. The lengths of the two axes of a particle indicate respective transfer entropies. That is, the size of a particle shows the magnitude of influence. Ellipse rendering is good at revealing influences around local regions although the inner particles could not be observed due to occlusion. Smoke rendering yields a less cluttered view and may serve better as an overview. However, the rendering could appear blurry.

Figure 7 shows transfer entropies between five pairs of variables calculated on the particles falling into the interior group with $y \in [690.0, 910.0)$. We can see strong couplings for pairs (T, G) and (YOH, T) and read the inter-influence among these five pairs of variables over time. The combustion scientists have confirmed that some of the major causal relationships are consistent with the established consensus in the combustion community, while the study also sheds new light on active areas of research. The relationship between mixing and combustion has long been a subject of inquiry: the scalar dissipation rate, $\chi \equiv 2D(\nabla Z)^2$, where Z is the mixture fraction and D is the mixture fraction diffusivity, is a key parameter in many models for turbulent combustion. This scalar dissipation rate is closely related to the quantity G considered here ($\chi = 2D(G)$). For a non-premixed flame, as in the latter portion of the domain ($t > 150$), the qualitative relationship is well known: the rate of reaction adjusts to balance the rate of mixing, and the dependent scalars (T, YOH, etc.), will vary as the quasi-steady balance adjusts to changes in $\chi$. In the early portion of the domain, where an ignition/stabilization process is occurring, the understanding is less clear. Some qualitative observations have been made that very high mixing rates hinder ignition, while the thermal expansion from fine flame fronts can dissipate the gradients and reduce the mixing rate. The findings of this analysis in the pre-stabilization region are of use to scientists seeking better understand of the coupling between mixing and chemistry in this regime.

Snapshots of three selected time steps with the double-layer metaball rendering are shown in Figure 8. The influences between the pair of variables YOH and T are illustrated while the color mapping of particles is based on a third variable Z. Note that when $T_{YOH \to T} \approx T_{T \to YOH}$, the two metaballs have very similar radii and they are perceived as a grey metaball when blended together.

## 6 Discussion

### 6.1 Parameter Choices

The top row of Figure 9 shows the side-by-side comparison of information transfer for the pair of variables QI and WS with three different block sizes used. We used the same transfer function and the same set of parameters to adjust the color and opacity for the three block sizes. As we can see, using a smaller block size leads to finer results. Note that the right image with the smallest block size shows the center of hurricane better where the inter-influence of the pair of variables is fairly weak, while the left image with the largest block size essentially averages the inter-influence within a very large region.

We point out that information-theoretic measures including transfer entropy require a large number of sample points as input. When much fewer samples are used, the estimation of the probabilities would be dominated by fluctuations. This is not a problem for analyzing volumetric and particle data sets as long as the region of

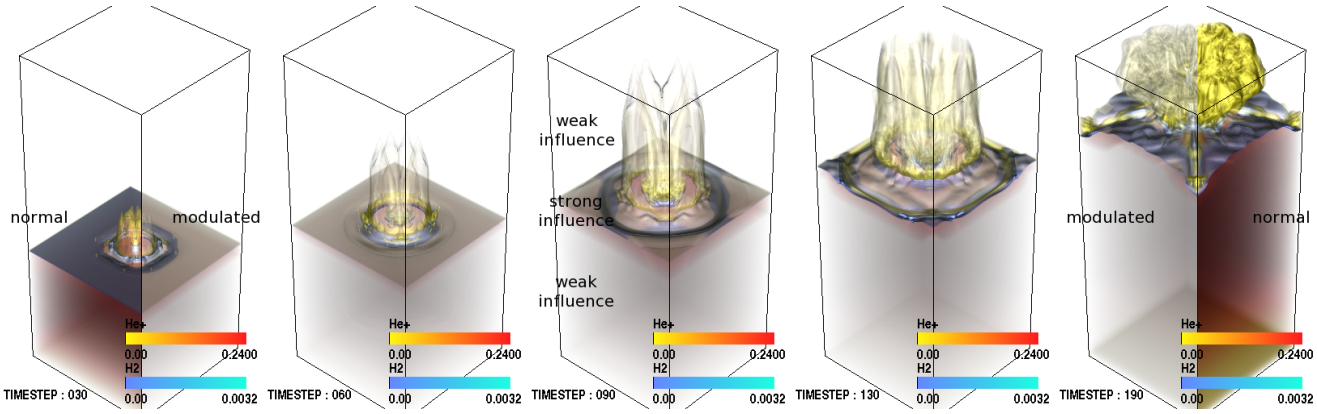



Figure 5: Visualization of information transfer on selected time steps of the ionization front instability data set. The He+ and H2 pair is shown. Strong influence regions are around the plane on the front moving upward.

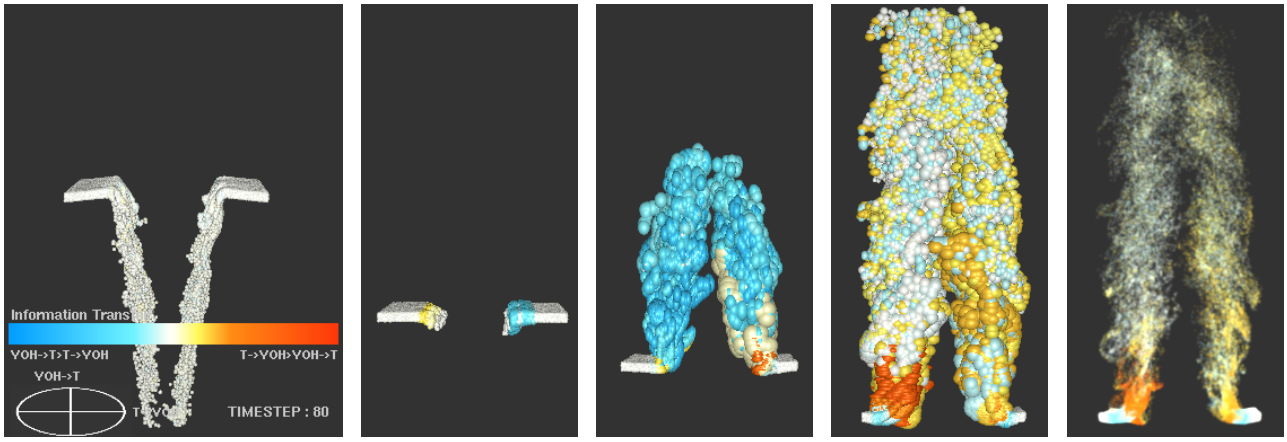



Figure 6: Visualization of information transfer on selected time steps for all 30 groups of the combustion particle data set. The YOH and T pair is shown and time steps are 80, 150, 220, 290, and 290 from left to right. The rightmost image shows the smoke representation while the rest show the ellipse representation. The influences between YOH and T are nearly equal in early time steps. After that, different groups of particles exhibit different inter-influence relations.

interest or the group of particles we select contains enough samples (e.g., at least in the order of thousands) for stable calculation. The bottom row of Figure 9 shows the side-by-side comparison of information transfer for the pair of variables YCH2O and YOH with three different histogram bins used. As we can see, using a larger bin size leads to more accurate results. The overall influence pattern, however, is very similar. Normally, using 256 bins for histogram suffices. If the value distribution of the data set is highly skewed, i.e., a large portion of the data values falls into a narrow range of the bins, nonuniform histograms should be used instead.

### 6.2 Timing

Table 2 lists the timing breakdown of transfer entropy calculation for a pair of variables per time step. The calculation was performed on an Intel Xeon 2.0GHz CPU. The numbers of groups (i.e., blocks) for the hurricane and ionization data sets are derived from the volume size and the block size used. For the hurricane data set, the block sizes are $50 \times 50 \times 20$, $20 \times 20 \times 20$, and $10 \times 10 \times 20$ respectively from top to bottom. For the ionization data set, the block sizes are $30 \times 31 \times 31$, $15 \times 31 \times 31$, and $10 \times 31 \times 31$ respectively from top to bottom. For a pair of variables, we read the time steps sequentially and only kept neighboring two time steps required for the computation in the memory. Every time step was read only once and the average data read per time step was calculated accordingly. The number of groups and the size of joint histograms largely determine the time required for joint histogram and transfer entropy computation. The computation time increases as we decrease the block size or increase the number of histogram bins used. On average, it took several minutes to process hundreds of megabytes (for a pair of variables per time step) and tens of hours to process hundreds of gigabytes (for all pairs of variables and all time steps).

The dominating time for transfer entropy calculation is due to the large numbers of log operations involved. Clearly, the computation is CPU bound. Since the calculation of transfer entropy is performed independently for every time step and for every data group, parallel preprocessing on a PC cluster or GPU implementation can speed up the computation. Another way of improving the performance is to replace the log function in the standard C/C++ library with direct table lookup or some fast approximation function. Moreover, as we discuss in Section 3.3, the increase of number of variables or time steps considered in the transfer entropy calculation has a significant impact on the timing performance. We will investigate efficient solutions for estimating joint histograms of high-dimensional data in the future.

### 6.3 Visualization Techniques

We use a separate view to visualize information transfer for all pairs of variables using time plot and circular graph. Time plots provide a good overview of influence changes over time in a single view, which is a familiar visualization to most users. Circular graphs allow us to easily capture variable relations, which is difficult to track with time plots. The time-varying graph shows influence over

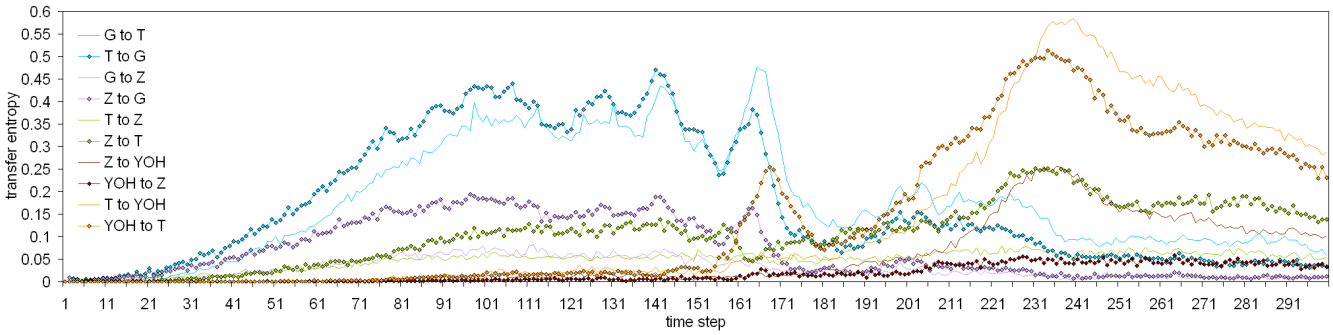



Figure 7: Using time plot to show transfer entropies between five pairs of variables calculated on the particles falling into the interior group.

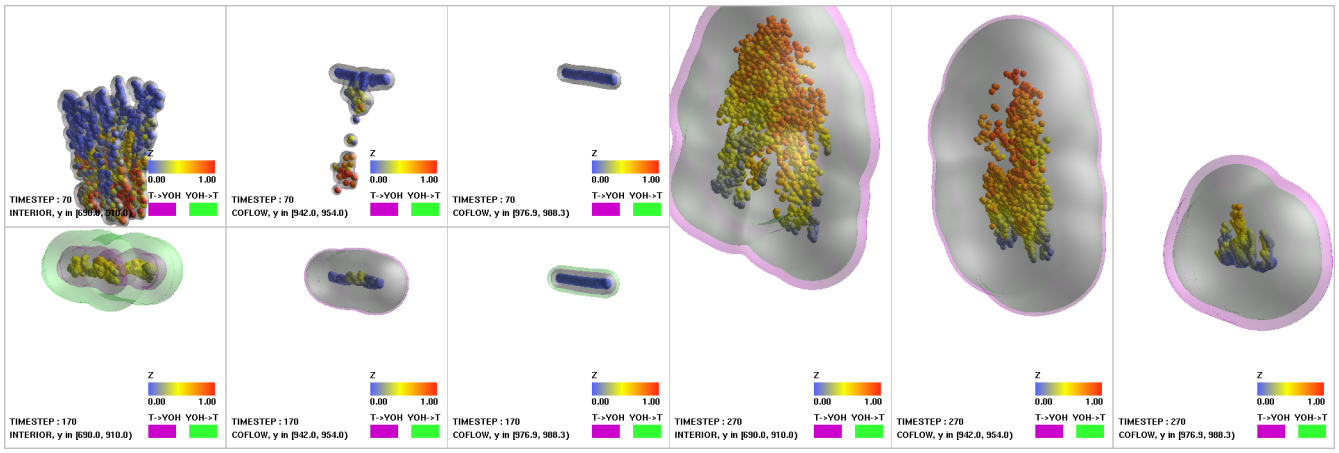


Figure 8: Side-by-side comparison of the three particle groups at three selected time steps: 70 (top-left), 170 (bottom-left), and 270 (right). The magnitudes of influence and their relations can be read from the sizes of the two metaballs.

| # groups | # bins | Read | Write | I/O | JH | TE |
|---|---|---|---|---|---|---|
| hurricane | | | | | | |
| 500 | 256 | 194MB | 5.5MB | 1.2s | 0.4s | 30s |
| 3125 | 256 | 194MB | 10.2MB | 1.4s | 0.3s | 190s |
| 12500 | 256 | 194MB | 17.9MB | 1.6s | 0.3s | 852s |
| ionization | | | | | | |
| 1280 | 256 | 282MB | 4.3MB | 3.1s | 0.7s | 76s |
| 2560 | 256 | 282MB | 5.2MB | 3.4s | 0.8s | 158s |
| 3840 | 256 | 282MB | 5.8MB | 3.7s | 0.6s | 218s |
| combustion | | | | | | |
| 30 | 128 | 858KB | 117KB | 4ms | 0.3s | 0.2s |
| 30 | 256 | 858KB | 206KB | 5ms | 0.3s | 1.8s |
| 30 | 512 | 858KB | 327KB | 8ms | 0.3s | 13s |

Table 2: The average timing for calculating transfer entropies for a pair of variables per time step. The output is joint histograms and transfer entropy values. The timing consists of reading and writing data (I/O), computing joint histograms (JH), and calculating transfer entropies (TE).

time as an animation. Both of these information visualization techniques, however, do not give the spatial context. This is complemented by integrating information flow directly into data rendering. The obvious advantage is that inter-influences can now be observed with respect to different spatial regions via visual properties such as color, opacity, or size. The downside is that exact transfer entropy values are not readable and we have to only show one pair of variables at a time. For particle rendering using metaball, showing multiple data groups would easily lead to difficulty in visual interpretation. This is generally not a problem with volume data block rendering and particle rendering using ellipse or smoke.

Accurately communicating the spatial nature of the relationships is a significant challenge both in terms of mental challenges (grasping the nature of the relationship and the implications) and the mechanics (occlusion, high-dimensional information). In both of these respects, the domain scientists found it useful to have a variety of representations available insofar as when their perceived understanding from different representations was inconsistent, they could actively correct their understanding. The scientists who are co-authors of our work found in particular that the ellipse and smoke representation were complimentary—taken together they alleviated the trade off between the occlusion level and the detail of information shown. The metaball visualization provided an overview of the relationships and the nature of the parcel of particle grouping to put the time series in context.

### 6.4 Limitations

Although we outline the generalization of transfer entropy to more than two variables, we need further research to address the fundamental challenge of dimension increase on both computation (time and storage) and visualization (information mapping and interpretation). In reality, studying more than two variables simultaneously is needed. For instance, in the combustion data set, it is known that there is a strong correlation between T, G, and Z, and it would be illuminating to explore the influence of T on G while conditioning both T and G on Z.

We use transfer entropy to measure information flow because it makes minimal assumptions about the dynamics of the time series and their coupling, captures both linear and nonlinear effects, and is numerically stable even for a reasonably small sample size (e.g., 1,000 samples). Transfer entropy is able to distinguish between

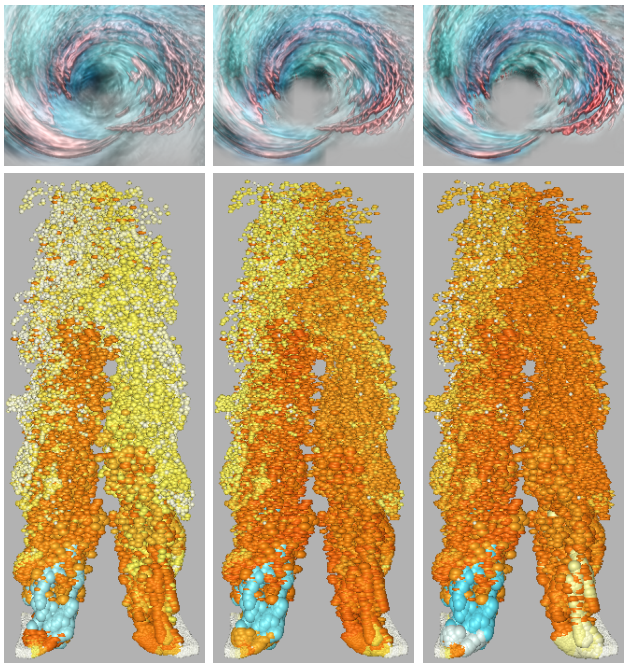


Figure 9: Parameter changes on information transfer results. Top row, left to right: QI and WS with block sizes of $50 \times 50 \times 20$, $20 \times 20 \times 20$, and $10 \times 10 \times 20$, respectively. Bottom row, left to right: YOH and YCH2O with histogram bin sizes of 128, 256, and 512, respectively. Although more refined/accurate results can be seen with a smaller block size/a larger bin size, the overall influence pattern is similar.

different stochastic time series where a pure visual investigation is difficult. However, care should be taken when interpreting causal relationships using transfer entropy. Without considering the nonlinear, transient, and noisy quality of the data, inferring "causal dependency" from mere time series data could be problematic. Moreover, the uncertainty introduced by the potential existence of unobserved variables or hidden common sources may be overlooked. In those cases, we tend to simply assume that no other factors have influence on the two systems or variables under investigation. As such, transfer entropy provides only one way of *suggesting* causation, not an *evidence* of causation. It is important to keep domain scientists in the loop so that they can make the final judgment.

## 7 Conclusions

Techniques for studying the causal relationships among multiple variables in time-varying data are in growing demand. Our technique is unique because it is based on measuring and visualizing information flow in the data. In a quantitative manner, we derive information transfer using the concept of transfer entropy from information theory. We show different ways to intuitively visualize information transfer for volumetric and particle data sets. Our causal analysis and visualization results provide valuable cues for scientists to understand complex time-varying multivariate data. Feedback from the scientists on this new approach is positive, suggesting it as a promising direction for studying important aspects of causal connections in the data.

## Acknowledgements

This work was supported by Michigan Technological University startup fund, the U.S. National Science Foundation through grants IIS-1017935, OCI-0749227, OCI-0905008, and OCI-0850566, and the U.S. Department of Energy through the SciDAC program with Award No. DE-FC02-06ER25777. The work at the Sandia National Laboratories (SNL) was supported by the DOE SciDAC Program. SNL is a multi-programme laboratory operated by the Sandia Corporation, a Lockheed Martin Company, for the DOE under contract DE-AC04-94AL85000.

## References

[1] J. Blaas, C. P. Botha, and F. H. Post. Interactive visualization of multifield medical data using linked physical and feature-space views. In *Proc. EuroVis*, pages 123–130, 2007.

[2] U. D. Bordoloi and H.-W. Shen. View selection for volume rendering. In *Proc. IEEE Visualization*, pages 487–494, 2005.

[3] N. Elmqvist and P. Tsigas. Causality visualization using animated growing polygons. In *Proc. IEEE Information Visualization*, pages 189–196, 2003.

[4] M. Glatter, J. Huang, S. Ahern, J. Daniel, and A. Lu. Visualizing temporal patterns in large multivariate data using textual pattern matching. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1467–1474, 2008.

[5] H. Jänicke, M. Böttinger, and G. Scheuermann. Brushing of attribute clouds for the visualization of multivariate data. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1459–1466, 2008.

[6] H. Jänicke, A. Wiebel, G. Scheuermann, and W. Kollmann. Multifield visualization using local statistical complexity. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1384–1391, 2007.

[7] G. Ji and H.-W. Shen. Dynamic view selection for time-varying volumes. *IEEE Trans. Vis. Comput. Graph.*, 12(5):1109–1116, 2006.

[8] M. Lungarella and O. Sporns. Mapping information flow in sensorimotor networks. *PLoS Comput. Biol.*, 2(10):1301, 2006.

[9] R. Marschinski and H. Kantz. Analysing the information flow between financial time series: An improved estimator for transfer entropy. *Eur. Phys. J. B*, 30(2):275–281, 2002.

[10] S. Oeltze, H. Doleisch, H. Hauser, P. Muigg, and B. Preim. Interactive visual analysis of perfusion data. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1392–1399, 2007.

[11] M. Paluš, V. Komárek, Z. Hrnčíř, and K. Štěrbová. Synchronization as adjustment of information rates: Detection from bivariate time series. *Phys. Rev. E*, 63, 2001.

[12] H. Qu, W.-Y. Chan, A. Xu, K.-L. Chung, K.-H. Lau, and P. Guo. Visual analysis of the air pollution problem in Hong Kong. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1408–1415, 2007.

[13] N. Sauber, H. Theisel, and H.-P. Seidel. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE Trans. Vis. Comput. Graph.*, 12(5):917–924, 2006.

[14] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85(2):461–464, 2000.

[15] D. Silver, M. Gao, and N. Zabusky. Visualizing causal effects in 4D space-time vector fields. In *Proc. IEEE Visualization*, pages 12–16, 1991.

[16] R. Sisneros, C. R. Johnson, and J. Huang. Concurrent viewing of multiple attribute-specific subspaces. *Comput. Graph. Forum*, 27(3):783–790, 2008.

[17] K. Stockinger, J. Shalf, K. Wu, and E. Wes Bethel. Query-driven visualization of large data sets. In *Proc. IEEE Visualization*, pages 167–174, 2005.

[18] J. Sukharev, C. Wang, K.-L. Ma, and A. T. Wittenberg. Correlation study of time-varying multivariate climate data sets. In *Proc. IEEE Pacific Visualization*, pages 161–168, 2009.

[19] P.-P. Vázquez, M. Feixas, M. Sbert, and W. Heidrich. Viewpoint selection using viewpoint entropy. In *Proc. Vision, Modeling, and Visualization*, pages 273–280, 2001.

[20] I. Viola, M. Feixas, M. Sbert, and M. E. Gröller. Importance-driven focus of attention. *IEEE Trans. Vis. Comput. Graph.*, 12(5):933–940, 2006.

[21] C. Wang, H. Yu, and K.-L. Ma. Importance-driven time-varying data visualization. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1547–1554, 2008.

[22] C. Ware, E. Neufeld, and L. Bartram. Visualizing causal relations. In *Proc. IEEE Information Visualization*, pages 39–42, 1999.

[23] J. Woodring and H.-W. Shen. Multi-variate, time-varying, and comparative visualization with contextual cues. *IEEE Trans. Vis. Comput. Graph.*, 12(5):909–916, 2006.