A Flexible Approach for the Statistical Visualization of Ensemble Data

Kristin Potter, Andrew Wilson, Peer-Timo Bremer, Dean Williams,
Valerio Pascucci, and Christopher Johnson

Abstract—Scientists are increasingly moving towards *ensemble data sets* to explore relationships present in dynamic systems. Ensemble data sets combine spatio-temporal simulation results generated using multiple numerical models, sampled input conditions and perturbed parameters. While ensemble data sets are a powerful tool for mitigating uncertainty, they pose significant visualization and analysis challenges due to their complexity. We present a collection of overview and statistical displays linked through a high level of interactivity to provide a framework for gaining key scientific insight into the distribution of the simulation results as well as the uncertainty associated with the data. In contrast to methods that present large amounts of diverse information in a single display, we argue that combining multiple linked statistical displays yields a clearer presentation of the data and facilitates a greater level of visual data analysis. We demonstrate this approach using driving problems from climate modeling and meteorology and discuss generalizations to other fields.

Index Terms—Ensemble data, uncertainty, statistical graphics, coordinated and linked views.

1 Introduction

Ensemble data sets are becoming an increasingly common tool to help scientists simulate complex systems, mitigate uncertainty and error, and investigate sensitivity to parameters and initial conditions. These data sets are large, multidimensional, multivariate and multivalued over both space and time. Because of their complexity and size, ensembles provide challenges in data management, analysis, and visualization.

In this article we present a general approach to the visual analysis of ensemble data with a focus on the discovery and evaluation of simulation outcomes. The approach combines a variety of statistical visualization techniques to allow scientists to quickly identify areas of interest, ask quantitative questions about the ensemble behavior, and explore the uncertainty associated with the data. By linking scientific and information visualization techniques we provide a cohesive view of the ensemble that permits analysis at multiple scales from high-level abstraction to the direct display of data values. Our work is developed in a component-based framework allowing it to be easily adapted to new applications and domains.

1.1 Motivation

The goal of an ensemble of simulation runs is to predict and quantify the range of outcomes that follow from a range of initial conditions. These outcomes have both quantitative aspects, such as the probability of freezing rain in a given area over a given time, and qualitative aspects, such as the shape of a severe weather system. While ensemble data sets have enormous power to express and measure such conditions, they also present formidable challenges for both visualization and data management due to their multidimensional, multivariate, multivalued nature and their sheer size. Many options exist to reduce an ensemble data set to a manageable size, however the specific set of data reduction algorithms applicable to any given scenario depend principally upon the particular application and the needs of the domain expert performing the analysis. One important common element among most applications using ensembles is the goal stated above: to predict and quantify the range of outcomes from a range of initial conditions. We provide a data analysis framework that allows domain scientists to explore and interrogate an ensemble both visually and numerically in order to reason about those outcomes.

- *K. Potter, V. Pascucci, and C. Johnson are with the SCI Institute, University of Utah*
- *A. Wilson is with Sandia National Laboratories*
- *P.T. Bremer and D. Williams are with Lawrence Livermore National Laboratory*

1.2 Driving Problems

We focus on two driving problems: short-term weather forecasting and long-term climate modeling. While our approach is informed by some of the specific needs of meteorology and climatology and in particular the applications described in this section, the structure and algorithms presented here are general enough to be applied to analysis problems using ensemble data across a wide variety of fields.

1.2.1 Weather Forecasting

Meteorologists increasingly turn to probabilistic data sets to forecast the weather rather than relying on singular, deterministic models [1]. Uncertainties and errors exist

in every weather simulation due to the chaotic nature of the atmosphere as well as the impossibility of accurately measuring its exact state at a specific time. Moreover, the numerical weather prediction models are often biased or inaccurate, leading to further error in the results. Ensembles are used to mitigate these problems by combining a variety of models using perturbed initial conditions and parameters. The resulting collection of simulations yields a richer characterization of likely weather patterns than any single, deterministic model.

We use data from NOAA's Short-Range Ensemble Forecast (SREF), a publicly available ensemble data set regenerated each day that predicts atmospheric variables over the whole of North America for 87 forecast hours (roughly 3.5 days) from the time the simulation is run. We obtained this data set from the National Centers for Environmental Protection's Environmental Modeling Center and Short-Range Ensemble Forecasting Project [2].

1.2.2 Climate Modeling

In contrast to meteorologists' goal of predicting the weather over a span of days or weeks, climate scientists are interested in global changes in climate over hundreds of years. Moreover, the phenomena they study spans the entire simulation domain (i.e. the whole planet) instead of being restricted to a small region of interest [3]. Climatologists integrate models and data from multiple international climate agencies that predict (among other things) the state of the atmosphere, oceans, vegetation and land use. The goal of these ensemble simulations is to understand phenomena such as the impact of human activity on global climate or trends in natural disasters. Because these results are used for decision making and public policy formation, the reliability and credibility of the predicted data is of paramount importance. The models are currently being verified by recreating conditions over the past century by the Intergovernmental Panel on Climate Change's experiment on the Climate of the 20th century with data available from the Earth System Grid data holdings [4]. This experiment produces an ensemble whose statistical trends are of utmost interest to climate researchers.

1.3 Ensemble Data Sets

We define an *ensemble data set* as a collection of multiple time-varying data sets (called *ensemble members*) that are generated by computational simulations of one or more state variables across space. The variation among the ensemble members arises from the use of different input conditions, simulation models, and parameters to those simulations.

Ensembles are:

- *Multidimensional* in space (2, 2.5 or 3 dimensions) and time;
- *Multivariate*, often comprising tens to hundreds of variables; and
- *Multivalued* in collecting several values for each variable at each point.

1.3.1 Ensembles and Uncertainty

Ensemble data sets are chiefly useful as a tool to quantify and mitigate uncertainty and error in simulation results. These errors can arise through faulty estimations or measurements of the initial conditions, from the finite resolution and precision of the numerical model, and from the nature of a numerical simulation as an approximate model of an incompletely understood real-world phenomenon.

Ensembles mitigate uncertainty in the input conditions by sampling a parameter space that is presumed to cover all possible starting conditions of interest. They alleviate uncertainty and error due to a finite simulation domain by operating on finer and finer domain decompositions until convergence is demonstrated. Additionally, they dissipate the imperfect nature of any numerical model by allowing the use of multiple models that each provide greater or lesser fidelity in some aspect of the process of interest in order to deemphasize bias.

We can interpret the multiple values for each variable at each point in an ensemble as specifying a probability distribution function (PDF) at each of those points. This interpretation allows us to describe the uncertainty of the data as the variation between samples. High variation in the samples indicates higher uncertainty. Statistical properties of the PDFs can be used to predict the most likely simulation outcomes along with an indicator of the reliability of each prediction.

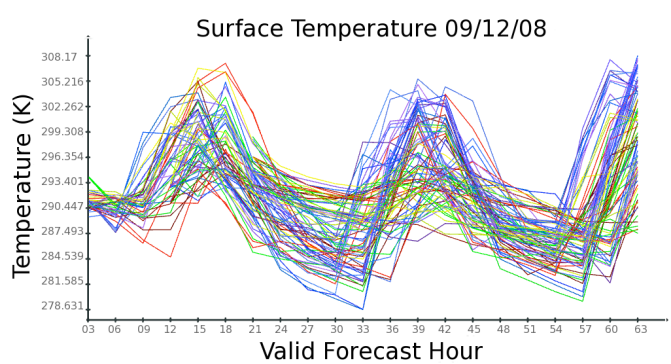

Fig. 1. An example of the complexity of an ensemble data set. Here, surface temperature data is shown at a single weather station across all valid forecast hours. While this plot reduces the overall data, it is still too visually cluttered to assist in data analysis beyond giving a notion of the general outcome.

1.3.2 Challenges for Analysis

The main challenges in using ensembles stem from the size and complexity of the data. For example, each of the

four daily runs of the SREF ensemble contains 21 members comprising four models and eleven sets of input conditions. Each member contains 624 state variables at each of 24,000 grid points and includes 30 time steps. A single day's output thus contains 84 members, each of which is a complex data set that poses visualization challenges in its own right. When information from all members is displayed together, as in the plume chart in Figure 1, the result is visual chaos that conveys only a general notion of the behavior of the predicted variable. Although the overall envelope defined by the minima and maxima can be discerned, the mostly likely outcome, the average across members, or even the course of any one member is difficult to extract. These challenges are exacerbated in more complex data sets such as climate simulations that incorporate 24 different models instead of four.

2 Related Work

Because of the complexity of the data we are working with, this research must draw from numerous fields within scientific and information visualization. Important topics include multidimensional, multivariate and multivalued data visualization, uncertainty visualization, statistical data display, and user interactivity.

Current state-of-the-art techniques for displaying weather and climate datasets include software systems such as Vis5D [5] and SimEnvVis [6]. These systems include 2D geographical maps with data overlaid via color maps, contours, and glyphs, as well as more sophisticated visualization techniques such as isosurfacing, volume rendering, and flow visualization. Vis5D focuses on displaying the 5 dimensional results of earth system simulations (3D space, time, and multiple physical variables) by combining the visualizations of multiple variables into a single image, and presenting a spreadsheet of these visualizations to show the various members of the simulation ensemble. SimEnvVis specializes in providing a library of comparative techniques to investigate and analyze multidimensional and multivariate climate-related simulation data. The system includes methods to compare and track features from a single simulation run, clustering to compare simulation and measured data, and information visualization approaches such as parallel coordinates to compare multi-run experiment results.

The main distinction between these previous efforts and the approach presented here is our stress on understanding the uncertainty available from the data by providing visualization tools that emphasize the probabilistic characteristics of ensemble data. We provide overviews to initially drive the analysis of ensemble data and highlight changes in uncertainty. Our system then provides a suite of statistical visualization tools to allow the analyst to understand where variations in the data arise, explore the relationships between ensemble members and directly present unaltered data values. We focus on providing qualitative information when appropriate, such as in the summary views, and quantitative statistics when necessary, for example when investigating the results of specific contributing members.

The data we are working with is multidimensional, multivariate, and multivalued. Previous work in visualizing these complex data types is extensive and can be investigated in a number of surveys and general techniques. Visualization of multivalued, multivariate data sets is a difficult task in that different techniques for dealing with the complexity of the data take effect through various stages of the visualization pipeline and are highly application specific. Knowing when to take advantage of these techniques through a categorization of methods is of great importance [7]. Multivariate correlation in the spatial domain is an often used approach for reducing the complexity of the task of data understanding [8], as is reducing the data to a hierarchical form which is conducive to 2D plots [9]. Likewise, the visualization of multidimensional data is challenging and often involves dimension reduction and user interaction through focusing and linking. A taxonomy of such techniques is very useful in determining an appropriate approach [10].

The most relevant work using ensemble type data views things in terms of probability distribution functions (PDFs) describing the multiple values at each location and each point in time [11]. Three approaches to visualizing this type of data are proposed; a parametric approach which summarizes the PDFs using statistical summaries and visualizes them using color mapping and bar glyphs, a shape descriptor which strives to show the peaks of the underlying distribution on 2D orthogonal slices, and an approach that defines operators for the comparison, combination, and interpolation of multivalued data using proven visualization techniques such as pseudocoloring, contour lines, isosurfaces, streamlines and pathlines. While our approach also uses a variety of statistical measures to describe the underlying PDF, we provide statistical views from a number of summarization standpoints in a single framework allowing the user to direct the data analysis, rather than automatically defining features of interest.

A major challenge for ensembles is in the wealth of information available. Depending on the application and the needs of the user, a single representation does not suffice. For example, a meteorologist may be interested in regional changes in temperature, as well as local variations at a specific weather station. The solution to this problem is to provide the user with multiple, linked views of the data [8], [12]. Such approaches let the user interactively select regions of interest, and reflect those selections in all related windows. The selection process can be through techniques such as brushing [13], or querying [14]. One interesting technique uses smooth brushing to select data subsets and then visualize the statistical characteristics of that subset [15]. Many of these methods use graphical data analysis techniques in the in-

dividual windows, such as scatterplots, histograms, and boxplots to show statistical properties and uncertainty of the underlying PDFs [16], [17]. The resulting collection of views provides for complex investigation of the data by allowing the user to drive the data analysis.

Much of this work is motivated by the growing need for uncertainty information in visualizations [18]. Understanding the error or confidence level associated with the data is an important aspect in data analysis and is too often left out of visualizations. There is a steadily growing body of work pertaining to the incorporation of this information into visualizations [19], [20], using uncertainty not only derived from data, but also present throughout the entire visualization pipeline. Specific techniques of interest to this work include using volume rendering to show the uncertainty predicted by an ensemble of Monte-Carlo forecasts of ocean salinity [21]; using flow visualization techniques to show the mean and standard deviation of wind and ocean currents [22]; uncertainty contours to show variations in models predicting ocean dynamic topography [23]; and expressing the quality of variables in multivariate tabulated data using information visualization techniques such as parallel coordinates and star glyphs [24].

3 Our Approach

In this section we discuss a framework for the visualization and analysis of ensemble data that emphasizes the probablisitic nature of the data. We highlight changes in uncertainty across the ensemble members and provide mechanisms for the investigation of areas deemed interesting by the analyst. Multiple windows are used which share selection, camera information and contents when appropriate. Each window presents the data condensed in space, time, or the multiple values at each point in order to highlight some aspect of the data behavior. Combining these windows into a single framework provides a unified platform for exploring the high complexity present in ensemble data sets. Our algorithms are presented in two prototypical systems, the SREF Weather Explorer, and the ViSUS Climate Data application.

We begin with an overview of the analysis work flow and then discuss each major component of our algorithm in detail, arranged from the most abstract view of the data to the most concrete and quantitative.

3.1 Work flow

A typical ensemble analysis is performed with two goals in mind. First, the analyst wishes to enumerate the possible outcomes expressed by the ensemble. Second, the user needs to understand how likely each outcome is relative to the other possibilities, and investigate how each member adds to the ensemble. To this end, a typical session follows the structure shown in Figure 2. An analyst begins by connecting to a data source and choosing one or more variables to display. The selected

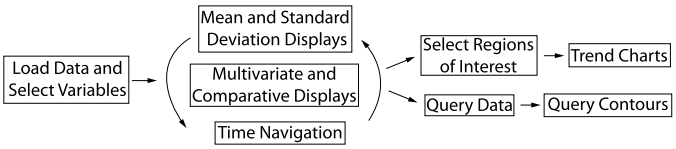

Fig. 2. An organization of the typical flow of data analysis through our framework. Users first choose a data set and one or more variables to display. They are then provided with mean and standard deviation views, comparative, and multivariate visualizations, all of which can be explored in the time domain via filmstrip views and animation. Next, the user selects a region of interest or queries the data. These selections drive the final stage of analysis by specifying interesting regions or data ranges, which are then displayed using more concrete representations such as trend charts and query contours.

variable is used to populate a *spatial-domain summary view* showing a statistical and spatial overview of data from one time step as well as a *time navigation summary view* showing a summary of the data over time.

From here the analyst can proceed in two directions. The *trend analysis* path reveals answers to questions of the form "What conditions will arise over time in a certain region of interest?" The *condition query* path addresses questions of the form "Where are the following conditions likely to arise and how probable are they?"

Since any investigation of average behavior is vulnerable to the influence of outliers, we incorporate methods to view ensemble members directly and include or exclude their effects from the various views. This is useful not only in understanding how simulation models influence the ensemble, but can also be used to eliminate members which are characteristically biased or unreliably predict specific regions across the spatial domain.

3.2 Data Sources

Ensemble data sets are usually too large for in-core processing on a single desktop computer. Each run of the SREF ensemble contains 36GB of data from each run; 106GB from each day. The climate data runs numerous models using fairly short time steps (15 minutes to 6 hours), over hundreds of years, resulting in hundreds of terabytes of data. However, unlike the simulations that generate the ensembles, we do not need fast access to all the data at all times. An analyst's investigation of the ensemble typically reduces the data by summarizing one or more of the spatial, temporal or probabilistic dimensions. These sorts of summaries are well suited to out-of-core methods. The ViSUS system traverses the ensemble using a streaming architecture. The SREF Weather Explorer stores the ensemble in a relational database and translates numeric queries into SQL.

The design of repositories for large amounts of scientific simulation data is itself an area of active research

with plenty of open challenges. For the purposes of the algorithms in this paper, we only require that the data repository is able to extract arbitrary subsets of an ensemble and, optionally, to compute summary information over those subsets. The underlying implementation details of the storage and retrieval system are orthogonal to requirements for visualization.

3.3 Ensemble Overviews

Immediately after connecting to a data source and selecting a variable of interest, the analyst is presented with a set of overview displays of the ensemble. The spatial-domain summary views (Figure 3, top) show the behavior of one variable over space at one time step. The time navigation summary views (Figure 3, bottom) show the same variable at lower spatial resolution over several time steps at once or through an animation.

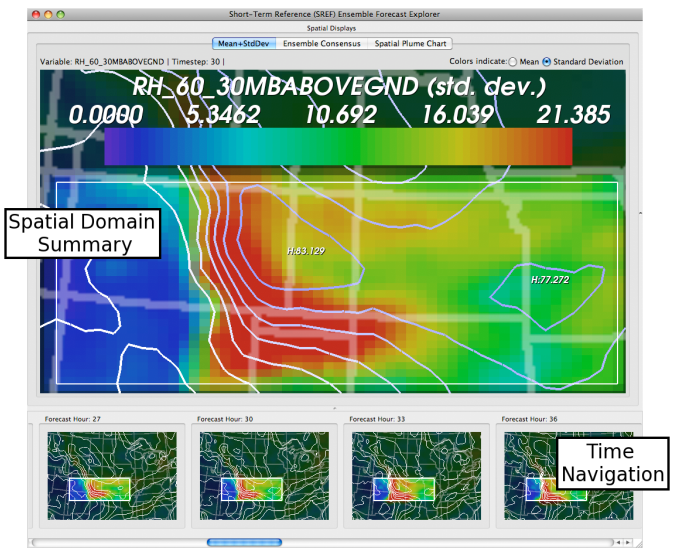

Fig. 3. We combine two representations to summarize each variable in the ensemble. A high-resolution spatial display (top) displays mean, standard deviation, and local minima and maxima for a given time step. An arrangement of lower-resolution multiples into a filmstrip (bottom) shows the same information over several time steps at once. The user can scroll through the filmstrip and transfer any time step to the high-resolution display.

3.3.1 *Spatial-Domain Summary Views*

The purpose of the spatial-domain summary view is to present a picture of the mean ensemble behavior at one point in time. Simple summary statistics such as mean and standard deviation work well as an approximate description of the range of values at each point. Since this is an overview, this approximation is sufficient: we need not convey precise scalar values for both mean and standard deviation. An approximate sense of the value of the mean plus an indication of high or low standard deviation is all that is required.

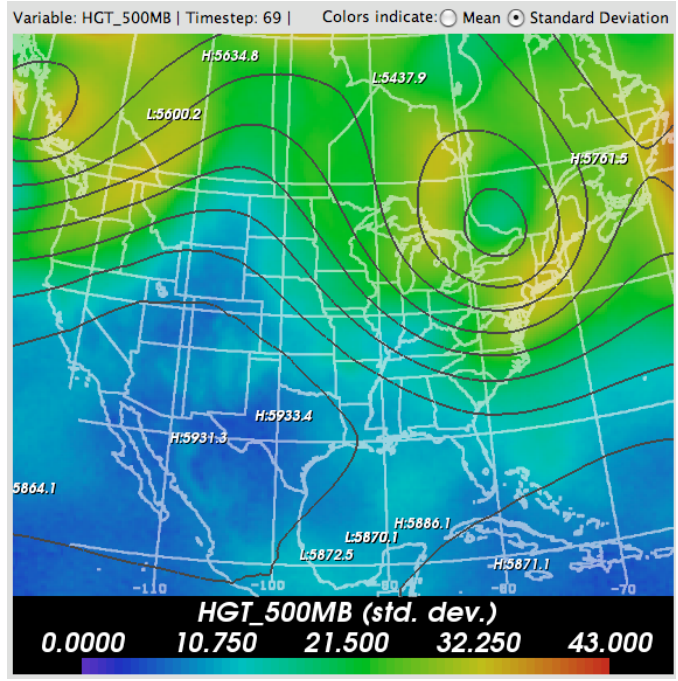

Fig. 4. We illustrate mean and standard deviation simultaneously using color plus overlaid contours

The spatial summary view also provides an indication of the uncertainty present across the spatial domain. Standard deviation, which characterizes the variation present in the data, is a measure typically used to describe the uncertainty of a dataset. In these summary views uncertainty is expressed either through color, height, or contours depending on the needs of the user. From this presentation, the analyst can quickly identify regions where the ensemble members converge indicating that the mean value at that location is a strong indicator of the predicted value, and where the members diverge indicating more exploration in that location is required to fully understand the ensemble's behavior.

By default, we display the variable mean using color and the standard deviation using overlaid contours (Figure 4). Although the rainbow color map is generally a poor choice for scientific visualization [25], it is familiar for variables such as temperature and relative humidity through its widespread use in print, television and online weather forecasts. For other variables such as surface albedo or probability of precipitation we allow the user to use a different sequential color map, examples of which can be seen in Figure 5. Still other scalar variables such as height and pressure are most easily interpreted using contour maps instead of colors. For these, the analyst can reverse the variable display so that the mean is shown as evenly spaced contours and the standard deviation is assigned to the color channel, as shown in Figure 6.

We can also display standard deviation using a height field instead of contours. This is particularly effective

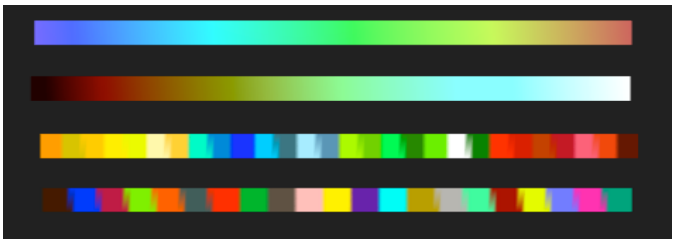

Fig. 5. Examples of our color maps. We use a subdued rainbow color map and a sequential low to high map for scalar variables and two categorical color maps for labeling.

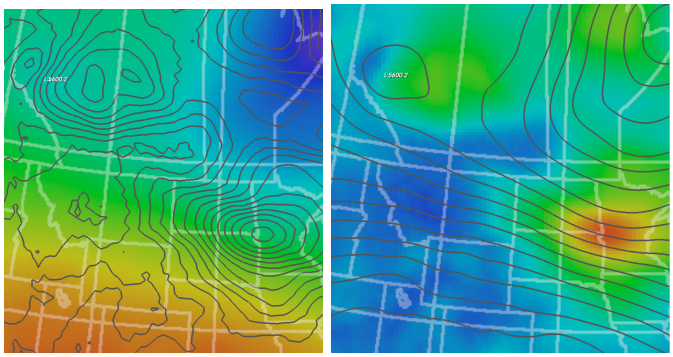

Fig. 6. The user can toggle the assignment of mean and standard deviation to colors and contours, respectively (left) or the reverse (right). Both images show the same region of the data.

when displaying 2D data projected onto the globe, as is common in climate simulations (Figure 7), since the height is easily visible along the silhouettes of the globe.

Although the mean and standard deviation cannot capture nuances of the underlying distribution, they are nonetheless appropriate here for two reasons. First, many observed quantities and phenomena in meteorology are well modeled by a normal distribution [26]. Second, many ensembles do not have enough members to support more sophisticated, precise characterizations.

3.3.2 Time Navigation Summary Views

In addition to the spatial summary view, which shows a high-resolution overview of a single time step, we also provide time navigation summary views which give an understanding of the evolution of the data through time.

The *filmstrip view*, sacrifices visible detail in order to allow quick traversal and inspection across time steps. As shown in Figure 8, the current variable is shown across all time steps using small multiples of the summary view. All of the frames in the filmstrip view share a single camera to allow the analyst to zoom in on a region of interest and observe its behavior over time. The user can scroll through the time steps and select the hour of interest. Double-clicking a frame transfers it to the higher-resolution summary, query contour views,

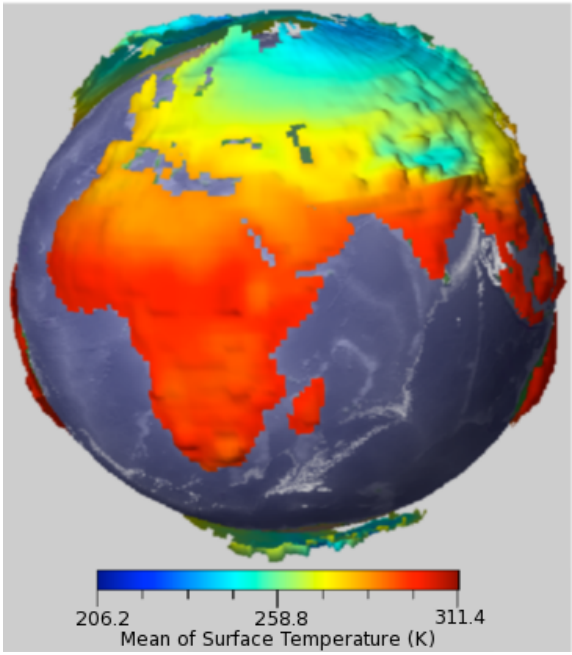

Fig. 7. Height is another channel available for data presentation. Here, standard deviation is displayed as a heightfield, and mean is shown through color. Highly displaced regions indicate high uncertainty, and this is especially visible on the silhouettes of the globe.

and trend chart views. This view allows the user to quickly select specific forecast hours, for example surface temperature 24 hours after initialization, or to quickly scroll through time and look for interesting events.

Animation is also used as a means to display time information. Here, the change of data with time is reflected in the summary view. This view emphasizes the evolution of the data and is best demonstrated through the climate data in which the animated globe gives a clear sense for the velocity of large-scale phenomena and global trends. This view is demonstrated in the accompanying video.

3.4 Trend Charts

The spatial and temporal summary displays discussed above summarize the distribution of values at each point into two numbers in order to preserve spatial information. In situations where the analyst specifies a region of interest – for example, when forecasting the weather for a particular region – we can instead aggregate over space and display detailed information about the distribution of values at each time step. This not only provides better detail at specific spatial locations, but also gives information about the behavior of the members making up the ensemble. We provide two such views.

3.4.1 Quartile Charts

A quartile trend chart (Figure 9) displays the minimum, maximum, 25th, 75th percentiles and median of

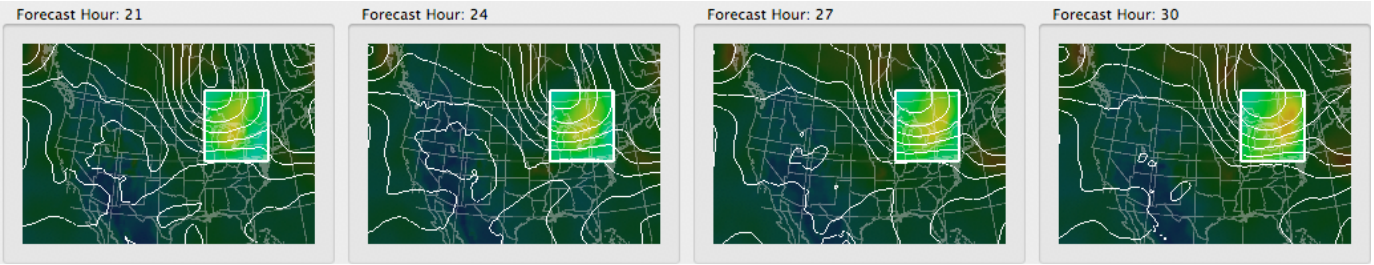

Fig. 8. The filmstrip summary view. Each frame in the filmstrip shows a single time step from the ensemble. The filmstrip also displays selection information from other views to help the user maintain a sense of context.

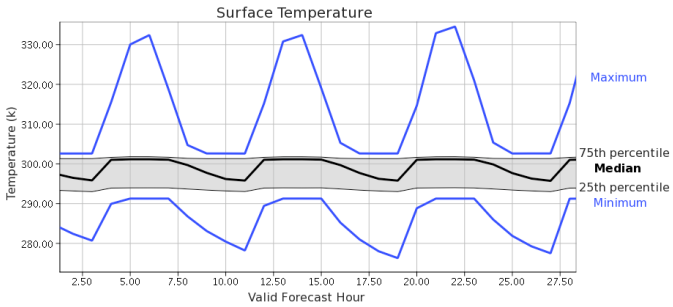

Fig. 9. Quartile trend charts. These charts show the quartile range of the ensemble within a user-selected region. Minimum and maximum are shown in blue, the gray band shows the 25th and 75th percentiles, and the median is indicated by the thick black line.

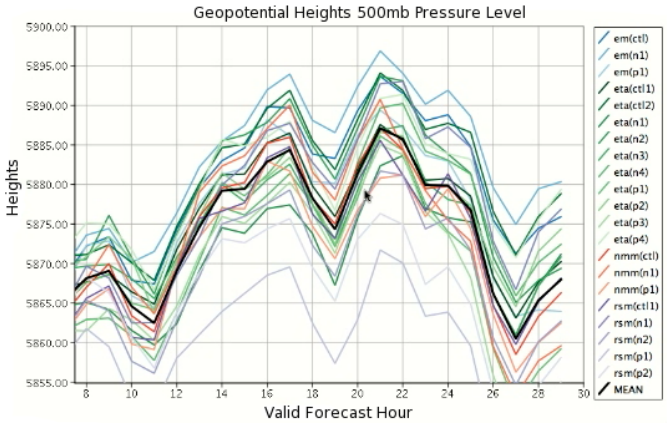

Fig. 10. Plume trend charts. These charts show the average of each ensemble model within a user-selected region of interest. Each model type is color-coded. The thick black line shows the mean across the entire ensemble.

a particular variable in a selected region over time. We compute these values over all the data for all ensemble members at each point in time. Order statistics give the analyst a view of the range of the possible outcomes as well as a notion of where the central 50% of the data values fall. This can be useful in quickly identifying minimal and maximal bounds at each forecast hour, as well as highlighting the range in which the majority of the members fall. As with the choice of mean and standard deviation in the summary view, this is most appropriate for unimodal distributions and can become less informative when the data distribution is more complex.

3.4.2 Plume Charts

A plume chart (Figure 10) shows the behavior of each ensemble member over time. Instead of aggregating all ensemble members into a single bucket (as is the case with quartile charts) we compute the mean of each ensemble member's values over the region of interest separately. Data series in the plume chart are colored so that all series that correspond to a single simulation model will have similar colors. The mean across all ensemble members is shown in black.

The plume chart is the most direct access to the data offered by our approach. Although it averages over the selected region, the analyst can obtain a view of raw values by selecting a region containing only a single data point. Since it displays data directly the plume chart also helps distinguish outliers and non-normal distributions. If the distribution is approximately normal, the mean represents the most likely outcome and should fall near the center of the members. If the distribution is non-normal, the mean is a poor estimation of the outcome, and the members will have high variation away from the mean line. Analysts can also track individual models of interest, or can discount heavily biased models. In addition, multimodal distributions can be detected since multiple strong clusters of members are readily apparent.

3.5 Condition Queries

The summary views and trend charts described above are *exploratory* views that illustrate behavior and possible outcomes over a region of interest. Another approach to ensemble data is for the analyst to specify a set of circumstances and ask for information about where they may occur. Such query-driven techniques [14] constrain the visualization to the subset of data deemed interesting by the analyst and discards the rest. We refer to these sets

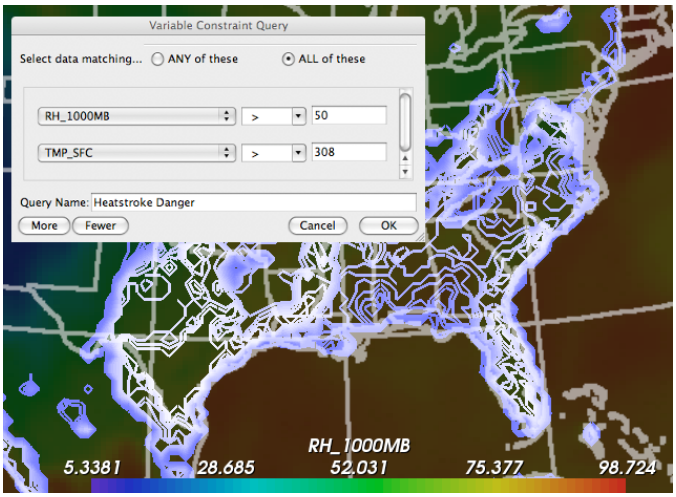

Fig. 11. The condition query view shows the probability that a given set of conditions will occur as a set of nested contours. Contour values are the fraction of the ensemble that predicts that the condition will be satisfied. In this figure we see a query for heatstroke danger (defined as relative humidity above 50% and temperatures above $95°$ Fahrenheit) and the resulting visualization.

of circumstances as *conditions*.

Once an analyst specifies a condition, as shown in the inset of Figure 11, the application translates it into a form understood by the data repository and retrieves a list of points where one or more ensemble members satisfies the condition. This list of points is transformed into an image where the scalar value at each point indicates the number of ensemble members (or, alternately, the *percentage* of the ensemble members) that meet the condition criteria. That image can in turn be displayed directly or (more usefully) drawn as a series of contours on a summary display.

In our example implementation using the SREF weather ensemble, conditions are translated into SQL and use the GROUP BY and COUNT constructs to aggregate individual data points into the image that represents the query contour. Although we used a very simple dialog to specify a condition, there exist a wide variety of query languages and mechanisms for visual query specification. Our component-based approach makes it straightforward to integrate any of these so long as an appropriate translation to the data source's native language exists.

3.6 Multivariate Layer Views

Although most ensemble analyses are performed using a single variable at a time, there are instances where an analyst wishes to compare multiple variables (especially multiple horizontal slices of a single 3D variable) across space at a single time step. This arises often when dealing with variables such as cloud structure that exhibit complex behavior across different altitudes. We display such slices using multiple 2D views in the same window. The data are displayed using a common color map in a single window. The analyst specifies the number of slices to be displayed and can also include a spatial summary (mean and standard deviation) along with the slice images. This type of display is assistive in comparing, for example, distinct time steps in the simulation, or the changes in a variable across the spatial domain. Figure 14 shows three elevations which add to the cloudiness across the globe.

3.7 Spaghetti Plots

A *spaghetti plot* [27], so named because of its resemblance to a pile of spaghetti noodles, is a tool frequently used in meteorology to examine variations across the members of an ensemble over space. An analyst first chooses a time step, a variable and a contour value for that variable. The spaghetti plot then consists of the isocontour for the chosen value for each different member of the ensemble. When the ensemble is in agreement, as shown in Figure 12, left, the contours will fall into a coherent bundle. When minor variation exists, a few outliers may diverge from the bundle (Figure 12, right). As the level of disagreement increases the contours become disordered and tangled and the spaghetti plot comes to resemble its namesake.

As with the plume charts, we assign colors to the contours in a spaghetti plot so that contours that arise from the same simulation model will have similar colors. We also allow the user to enable and disable different ensemble members in order to inspect and compare the behavior of different models or the effects of different perturbations of initial conditions.

3.8 Coordination Between Views

The various views in our system coordinate their displayed variables, time steps, camera parameters and selections to the greatest degree that is appropriate. Lightweight operations such as changes to the camera, selection, image/contour assignment and contour level (for the spaghetti plot) take effect immediately. More expensive operations such as changing the current variable, executing a condition query or generating trend charts from a selection require that we retrieve new data from storage. Since these operations take several seconds to complete we defer execution until the user specifically requests them.

4 Implementation Details

We have implemented the algorithms described in Section 3 in two prototype systems for weather and climate simulation analysis. This demonstrates the flexibility of our component-based approach. In this section we describe briefly the purpose and system architecture of each prototype. Working memory is not a major concern for either system: including OS overhead, our prototypes ran in under 300MB of RAM.

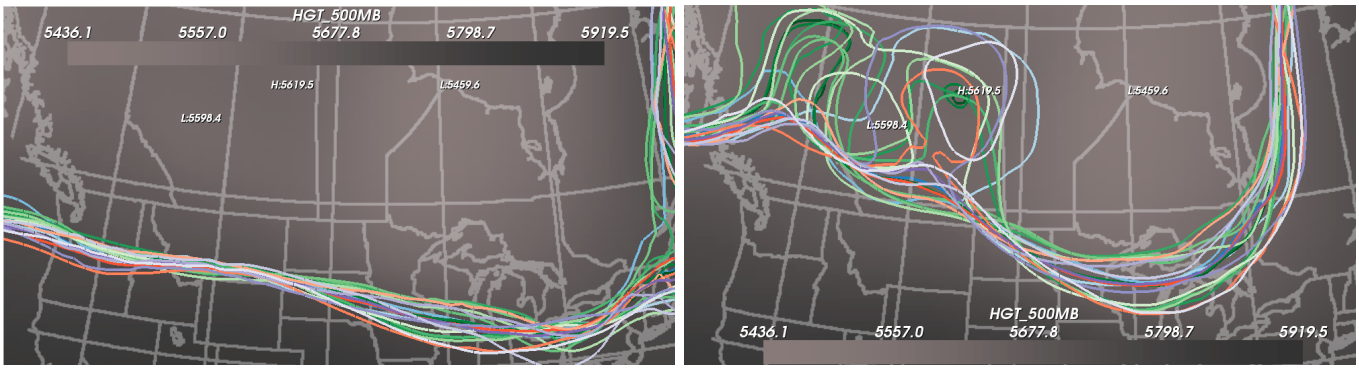

Fig. 12. A spaghetti plot displays a single isocontour from each ensemble member in order to allow examination of differences across space. (Left) When the members are in agreement the contours form coherent bundles. (Right) When ensemble members disagree, as in the in upper left region of the image, outliers diverge from the main bundle.

4.1 SREF Weather Explorer

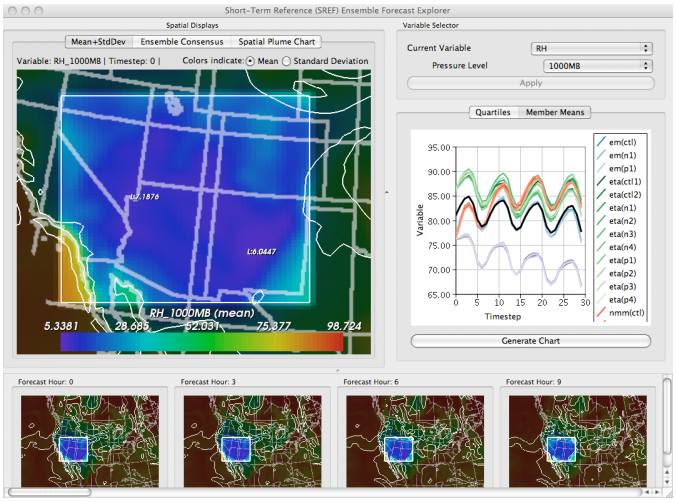

Fig. 13. Screenshot of the SREF Weather Explorer. This prototype is implemented as a set of VTK filters and can thus be easily integrated into tools deployed to domain scientists.

The SREF Weather Explorer application permits ensemble analysis of a single instance of the NOAA Short-Term Reference Ensemble Forecast (SREF) data set [2]. Since the SREF simulates weather conditions only in a region surrounding North America it lends itself to 2D display. This prototype incorporates 2D summary views, a filmstrip view, an ensemble consensus view using condition queries, spaghetti plots and trend charts, a screenshot of which can be seen in Figure 13. The visualization algorithms in SREF Weather Explorer are implemented as filters in VTK [28], a well-known open-source toolkit for scientific visualization. The user interface components were implemented as Qt widgets [29]. We plan to release these components as open source late in 2009.

We used standard relational databases as the storage engine for the SREF ensemble data. This allowed our application to offload the task of storage management and thus run identically on machines ranging from a five-year-old dual-processor Linux workstation to a Mac Pro with two 4-core processors and 16GB of local memory. By using VTK's modules for database connectivity we were able to switch between different database instances with no additional effort. These included one full 36GB run of the SREF ensemble stored on a 56-node Netezza parallel database appliance as well as a 5.5GB subset of the ensemble stored in a MySQL instance running on a single-processor laptop. From the user's perspective, the only difference was the hostname entered during application startup.

4.2 ViSUS/CDAT

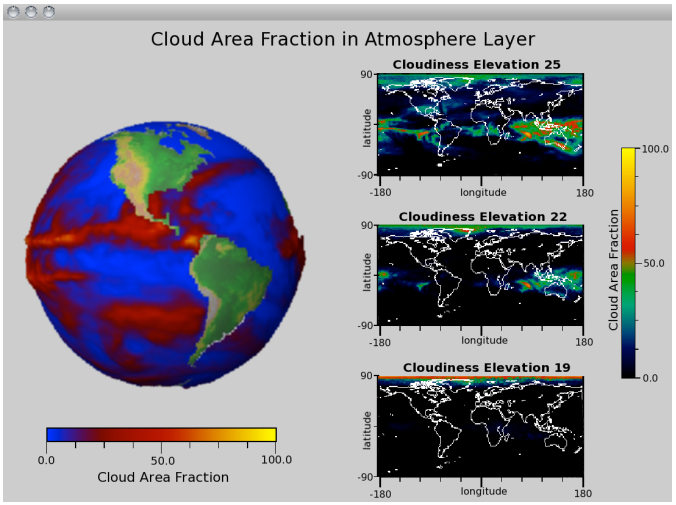

Fig. 14. Screenshot of the ViSUS prototype. This system is integrated into the CDAT framework used by climate scientists.

Climate scientists use a variety of special data formats and have domain specific requirements not common in general scientific visualization tools. The Program for Climate Model Diagnosis and Intercomparison (PCMDI) has developed a suite of Climate Data Analysis Tools (CDAT) [30] specifically tailored for this community.

ViSUS, our prototype, integrates into the CDAT infrastructure by providing a lightweight and portable, advanced visualization library based on an out of core streaming data model. ViSUS is developed to address the specific needs of climate researchers, and as such has specialized features such as projecting the data onto a model of the Earth, masking out land and ocean, and enhancing the visualizations with geospatial information such as satellite images and geographic boundaries. The algorithms contained in ViSUS are implemented in C++, OpenGL and python, and the system uses FLTK for user interaction. A screenshot of the ViSUS system can be seen in Figure 14.

5 Discussion

Visual analysis of ensemble data sets is challenging and complex on all levels. No one view or collection of views will be ideal for all analyses. In this section we discuss some of the trade-offs in our approach and the rationale behind our decisions.

5.1 Data Challenges

The first major challenge we encounter in ensemble visualization is to decide exactly what to display. Because an ensemble of simulations is expensive and difficult to compute, most ensemble data sets are written out with as much information as can be stored at the highest feasible resolution in both space and time. This quickly leads to an overwhelming amount of multivariate data. We must somehow determine which parts of the ensemble are important enough to keep and display.

However, guidelines for what data matters and what can be discarded are necessarily specific to each application domain, to each simulation, and even to each analysis session. Under these circumstances it seems most appropriate to preserve all the data and allow the analyst to specify exactly which data they want to see and the manner in which to display it.

To this end, our focus in this work is give the data analyst tools to understand the outcome of an ensemble simulation by providing insight into the statistics and uncertainty associated with the data. Each type of visualization was chosen to assist the in the discovery and evaluation of the ensemble from high to low level, and when combined provides a comprehensive tool for data analysis.

5.2 Where Statistics Break Down

We have been fortunate in working with weather and climate data because many of the variables of interest are well described by the normal distribution and thus sufficiently characterized by the mean and standard distribution alone. Simulations from other domains such as mechanical engineering and thermal analysis exhibit more complicated behavior where the mean and standard deviation are no longer appropriate. Such behavior can also arise in simulations of extreme conditions using an ordinarily well-behaved model.

The choice of summary statistics for any given distribution is dependent on the characteristics of the distribution itself. We must also consider whether we have enough data values to justify using any given measure. Moreover, the use of simple summary statistics in our work presumes relatively complete, unbiased, registered data as input. This is not always the case. Even under the assumption of a common simulation grid, some data may be missing; that is, some ensemble members may not compute all values for all time steps. Also some models may be better represented in an ensemble than others. These problems share a common theme of data bias. Once again, the solution is specific to each analysis. Perhaps an apparently over-represented model is actually desirable due to its superior predictive power. Perhaps missing data values were omitted deliberately where a model strays into a region of inapplicability. A robust solution would address these scenarios by allowing the analyst to assign relative importance to different ensemble members.

5.3 Glyphs for Standard Deviation

We experimented with a summary display comprising a glyph at each data point. The glyph's color indicated the mean at that point. Its size reflected the standard deviation. We discarded this approach in favor of the one presented above for two reasons. First, glyphs lead to unacceptable visual clutter. They occlude one another in areas of high standard deviation in 2D data sets and are even more troublesome when moving to 3D. A second, deeper problem is that humans do not perceive size and color separately [31]. A dark glyph placed next to a bright glyph of the same size will appear smaller. Instead of glyphs at every point, we chose to move toward the use of glyphs to highlight highs and lows in the data.

6 Conclusion and Future Work

In this article we have presented an approach to ensemble visualization using a federation of statistical representations that when used in combination provide an adaptable tool for the discovery and evaluation of simulation outcomes. The complexity of the ensemble data is mitigated by the flexible organization offered by our approach and the coordination between views allows data analysts to focus on the formulation and evaluation of hypothesis in ensemble data. The strengths of our approach include little or no preprocessing cost, low memory overhead through reliance on queryable out-of-core storage and easy extension and adaptability to new domains and new techniques. We have demonstrated our approach in two different software prototypes that allow the analysis of large data sets with hardware requirements easily met by present-day laptops.

We see three principal directions for future research. First, our methods are specialized for two- and 2.5-dimensional data. An approach to 3D data sets must address the classic problems of clutter and occlusion. We might be able to exploit the observed tendency of the amount of ensemble variation to change relatively slowly in space and time. Second, we need better methods for the display of mean and standard deviation. Here we will exploit the use of standard deviation as an approximate indicator of ensemble disagreement instead of a precise scalar variable. Finally, we will expand our methods to gracefully handle non-normal, multimodal and higher dimensional probability distributions. This will require runtime characterization of the shape of a distribution, perhaps including automatic model fitting and trend charts that show histograms as well as summary statistics and ensemble members.

The rapid increase in computational capacity over the past decade has rendered ensemble data sets a viable tool for mitigating uncertainty and exploring parameter and input sensitivity. Visualization and data analysis tools are needed to help domain scientists understand not only the general outcome of the data, but also the underlying distribution of members contributing to that outcome. We believe that our work constitutes early progress toward the many new challenges posed by these large, complex and rich data sets.

Acknowledgments

This work was funded in part by the DOE SciDAC Visualization and Analytics Center for Enabling Technologies (www.vacet.org) and the NIH NCRR Center for Integrative Biomedical Computing (www.sci.utah.edu/cibc), NIH NCRR Grant No. 5P41RR012553-02.

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.

References

[1] T. Gneiting and A. Raftery, "Atmospheric science: Weather forecasting with ensemble methods," *Science*, vol. 310, pp. 248–249, October 2005.

[2] "Short-range ensemble forecasting," http://wwwt.emc.ncep.noaa.gov/mmb/SREF/SREF.html.

[3] G. Compo, J. Whitaker, and P. Sardeshmukh, "Bridging the gap between climate and weather," http://www.scidacreview.org/0801/html/climate.html, 2008.

[4] "Climate of the 20th century experiment (20c3m)," https://esg.llnl.gov:8443/index.jsp.

[5] B. hubbard and D. Santek, "The vis-5d system for easy interactive visualization," in *IEEE Vis '90*, 1990, pp. 28–35.

[6] T. Nocke, M. Fleshig, and U. Böhm, "Visual exploration and evaluation of climate-related simulation data," in *IEEE 2007 Water Simulation Conference*, 2007, pp. 703–711.

[7] R. Bürger and H. Hauser, "Visualization of multi-variate scientific data," in *Eurographics 2007 STAR*, 2007, pp. 117–134.

[8] L. Anselin, I. Syabri, and O. Smirov, "Visualizing multivariate spatial correlation with dynamically linked windows," in *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting*, 2002.

[9] T. Mihalisin, J. Timlin, and J. Schwegler, "Visualization and analysis of multi-variate data: a technique for all fields," in *IEEE Vis '91*, 1991, pp. 171–178.

[10] A. Buja, D. Cook, and D. F. Swayne, "Interactive high-dimensional data visualization," *Journal of Computational and Graphical Statistics*, vol. 5, no. 1, pp. 78–99, March 1996.

[11] A. Love, A. Pang, and D. Kao, "Visualizing spatial multivalue data," *IEEE CG & A*, vol. 25, no. 3, pp. 69–79, May 2005.

[12] J. Roberts, "State of the art: Coordinated and multiple views in exploratory visualization," in *5th International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pp. 61–71.

[13] R. Becker and W. Cleveland, "Brushing scatterplots," *Technometrics*, vol. 29, no. 2, pp. 127–142, May 1987.

[14] K. Stockinger, J. Shalf, K. Wu, and E. W. Bethel, "Query-driven visualization of large data sets," in *IEEE Vis '05*, 2005, pp. 167–174.

[15] A. Unger, P. Muigg, H. Doleisch, and H. Schumann, "Visualizing statistical properties of smoothly brushed data subsets," in *12th International Conference on Information Visualization*, 2008, pp. 233–239.

[16] W. Cleveland, *Visualizing Data*. Hobart Press, 1993.

[17] K. Potter, J. Kniss, and R. Riesenfeld, "Visual summary statistics," Univeristy of Utah, Tech. Rep. UUCS-07-004, 2007.

[18] C. R. Johnson and A. R. Sanderson, "A next step: Visualizing errors and uncertainty," *IEEE CG & A*, vol. 23, no. 5, pp. 6–10, 2003.

[19] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler, "Visualizing geospatial information uncertainty: What we know and what we need to know," *Cartography and Geographic Information Science*, vol. 32, no. 3, pp. 139–160, July 2005.

[20] A. Pang, C. Wittenbrink, and S. Lodha., "Approaches to uncertainty visualization," *The Visual Computer*, vol. 13, no. 8, pp. 370–390, Nov 1997.

[21] S. Djurcilov, K. Kim, P. Lermusiaux, and A. Pang, "Volume rendering data with uncertainty information," in *Data Visualization*, 2001, pp. 243–52.

[22] C. M. Wittenbrink, A. T. Pang, and S. K. Lodha, "Glyphs for visualizing uncertainty in vector fields," *IEEE Transactions on Visualization and Computer Graphics*, vol. 2, no. 3, pp. 266–279, September 1996.

[23] R. S. A. Osorio and K. Brodlie, "Contouring with uncertainty," in *6th Theory and Practice of Computer Graphics Conference*, I. S. Lim and W. Tang, Eds., 2008, pp. 59–66.

[24] Z. Xie, S. Huang, M. Ward, and E. Rundensteiner, "Exploratory visualization of multivariate data with variable quality," in *IEEE Symposium on Visual Analytics Science and Technology*, 2006, pp. 183–190.

[25] D. Borland and R. T. II, "Rainbow color map (still) considered harmful," *IEEE CG & A*, vol. 27, no. 2, pp. 14–17, Mar/April 2007.

[26] J. Sivillo, J. Ahlquist, and Z. Toth, "An ensemble forecasting primer," *Weather Forecasting*, vol. 12, pp. 809–817, 1997.

[27] P. Diggle, P. Heagerty, K.-Y. Liang, and S. Zeger, *Analysis of longitudinal data*. Oxford University Press, 2002.

[28] W. Schroeder, K. Martin, and B. Lorensen, *The Visualization Toolkit*. Kitware, 2006.

[29] J. Blanchette and M. Summerfield, *C++ GUI Programming with Qt 4*. Prentice Hall, 2006.

[30] "Climate data analysis tools," http://www2-pcmdi.llnl.gov/cdat.

[31] J. S. D. Bonet and Q. Zaidi, "Comparison between spatial interactions in perceived contrast and perceived brightness," *Vision Research*, vol. 37, no. 9, pp. 1141–1155, May 1997.