A Flexible Approach for the Statistical Visualization of Ensemble Data

Category: Research

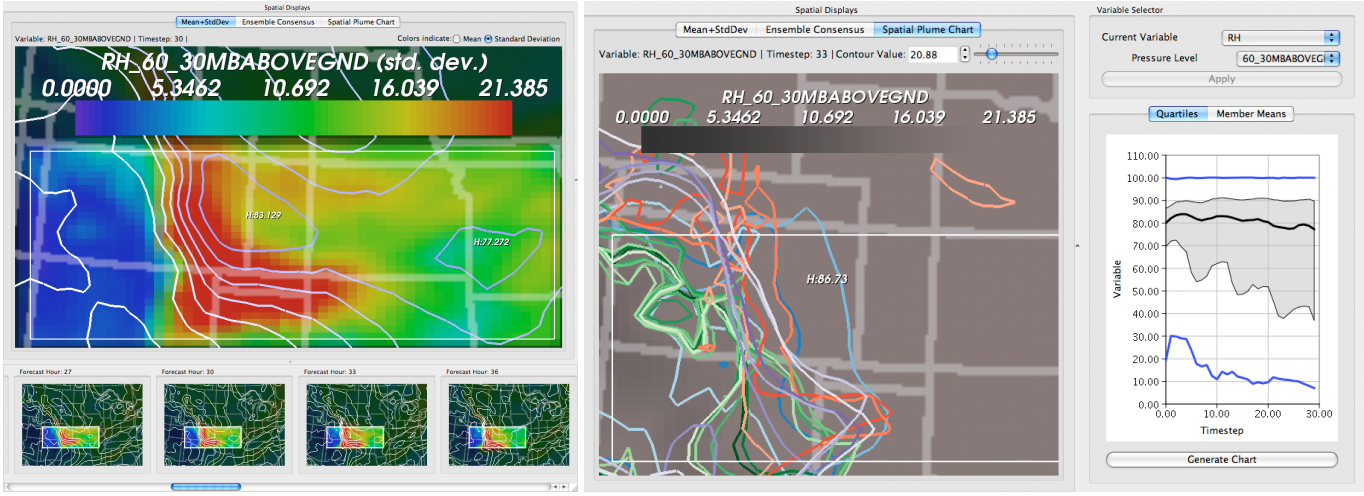

Fig. 1. This figure shows examples of ensemble data analysis at different levels of abstraction. On the left we show a high-resolution overview of a single simulation timestep where an area of high variance in one important variable (relative humidity) is shown in red. Below the overview is a filmstrip widget that shows several time steps at once as well as context for the highlighted region. On the right we show a lower-level, more quantitative examination where the user compares the locations of a single isocontour from each of the simulation results in the ensemble. The chart at far right shows a summary plot of the values of relative humidity in the selected region over time.

Abstract—
Scientists are increasingly moving towards *ensemble data sets* to explore the relationships between changing initial conditions or model assumptions and the probabilities of associated simulation outcomes. Ensemble data sets combine time- and spatially-varying 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 also pose significant visualization and analysis challenges due to their size (tens of gigabytes to petabytes), complexity (tens or hundreds of state variables) and organization (many values for each variable at each point in space and time). We present algorithms for gaining key scientific insight through a collection of overview and statistical displays linked through a high level of interactivity. In contrast to methods that present large amounts of diverse information in a single display, we argue that combining multiple linked displays that show the data from a variety of statistical viewpoints yields a clearer presentation of the data and facilitates a greater level of visual data analysis. We demonstrate these algorithms using driving problems from climate modeling and meteorology and discuss applications to other fields such as mechanical and biomedical engineering, economics, and the geosciences.

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

1 Introduction

Ensemble data sets are becoming increasingly common as tools to help scientists simulate complex systems, mitigate uncertainty and error in models and initial conditions, and investigate a model's sensitivity to its input parameters. These ensemble data sets are large, multidimensional, multivariate and multivalued over both space and time. Because of their complexity and size, ensembles provide both data management and visualization challenges.

In this paper we present a general approach to the visual analysis of ensemble data focused on the discovery and evaluation of simulation outcomes. We combine a variety of statistical visualization techniques to allow scientists to quickly identify areas of interest and ask quantitative questions about the ensemble behavior. We demonstrate our approach on driving problems from meteorology and climate modeling and discuss its applications to other fields. By linking visualization techniques from scientific and information visualization we are able to provide a cohesive view of the ensemble that permits analysis at multiple scales from high-level abstraction to the direct display of data values. Additionally, we provide high quality annotations and labels within the interactive system providing an important insight to the data values as well as the capability to produce publication-ready illustrations. 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. While many options exist to reduce an ensemble data set to manageable size, the specific set of data reduction algorithms applicable to any given scenario depend principally upon the particular application and the needs of the domain ex-

pert 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.

1.2 Driving Problems

In this paper 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.

Weather Forecasting Meteorologists increasingly turn to probabilistic data sets to forecast the weather rather than relying on singular, deterministic models [15]. Uncertainties and errors exist in every weather simulation due to the chaotic nature of weather itself as well as the impossibility of accurately measuring the state of the entire atmosphere 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 weather models with a variety of perturbed initial conditions and parameters. The resulting collection of simulation results yields a richer characterization of likely weather patterns than any single, deterministic simulation.

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 [3].

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 [12]. 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 [2]. This experiment produces an ensemble whose statistical trends are of utmost interest to climate researchers.

	Perturbations										
Model	ctl	ctl1	ctl2	n1	n2	n3	n4	p1	p2	p3	p4
ETA	•			•				•			
EM		•	•	•	•	•	•	•	•	•	•
NMM	•			•				•			
RSM		•		•	•			•	•		

Table 1. The 21 SREF members: four Numerical Weather Prediction (NWP) models, their input perturbations, and a color scheme for each model type used in Figure 2. Members generated with the ETA model are shown in yellow, EM in blue, NMM in green and RSM in red. The 21 model/perturbation configurations are run using four distinct initial conditions, leading to 84 total ensemble members at each point in time and space.

1.3 Ensemble Data Sets

In this paper 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, different simulation models, and different parameters to those simulations. We will use the term *ensemble* to refer interchangeably to the family of simulations that generated a particular collection of data sets or to the collection of data sets itself.

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.

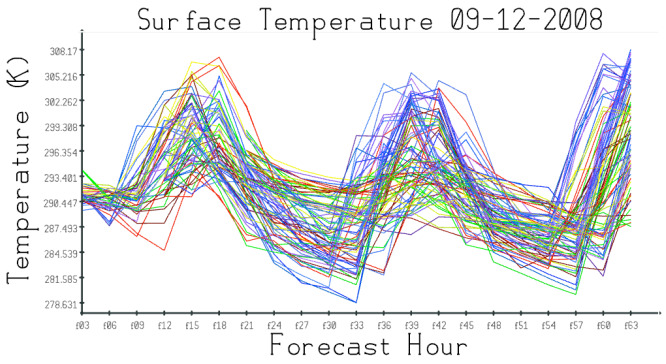

Fig. 2. An example of the complexity of an ensemble data set. Here, surface temperature data is shown at a single weather station all forecast hours. The model types are color-labeled using the color scheme in Table 1. While this plot reduces the overall data, it is still too visually cluttered to assist 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 (Table 1). 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 2, 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 techniques for displaying weather and climate datasets include software systems such as SimEnvVis [21] and Vis5D [16]. These systems include 2D geographical maps with data overlaid via colormaps and contours, as well as more sophisticated visualization techniques such as isosurfacing, volume rendering, and flow visualization. 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.

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 [10]. Multivariate correlation in the spatial domain is an often used approach for reducing the complexity of the task of data understanding [4], as is reducing the data to a hierarchical form which is conducive to 2D plots [20]. 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 [9].

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 [18]. Three approaches to visualizing this type of data are proposed; a parametric approach which summarizes the PDFs using statistical summaries and visualizes them using colormapping 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 [4, 25]. 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 [5], or querying [28]. One interesting technique uses smooth brushing to select data subsets and then visualize the statistical characteristics of that subset [29]. Many of these methods use graphical data analysis techniques in the individual windows, such as scatterplots, histograms, and boxplots to show statistical properties and uncertainty of the underlying PDFs [11, 24]. 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 [17]. 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, 23], 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 [14]; using flow visualization techniques to show the mean and standard deviation of wind and ocean currents [30]; uncertainty contours to show variations in models predicting ocean dynamic topography [22]; and expressing the quality of variables in multivariate tabulated data using information visualization techniques such as parallel coordinates and star glyphs [31].

3 Our Approach

In this section we discuss a framework for ensemble visualization and analysis through the use of multiple views, each of which condenses space, time, or the multiple values at each point in order to highlight some aspect of the data behavior. These views share selections, camera information, and contents wherever appropriate. We present our algorithms 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

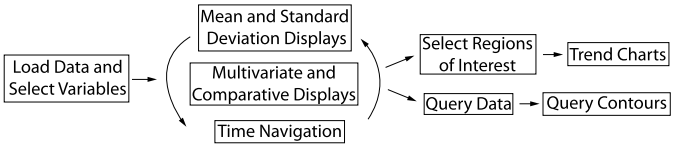

Fig. 3. An organization of the typical flow of data analysis through our framework. The user first chooses a data set and one or more variables to display. They are then provided with mean and standard deviation views, comparative, and multivariate visualizations, all of which can be explored in the 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.

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, they need to understand how likely each outcome is relative to the other possibilities. To this end, a typical session follows the structure shown in Figure 3. An analyst begins by connecting to a data source and choosing one or more variables to display. The selected variable is used to populate a *summary view* showing a statistical and spatial overview of data from one time step as well as a *filmstrip view* showing small multiples of the summary view 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.

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

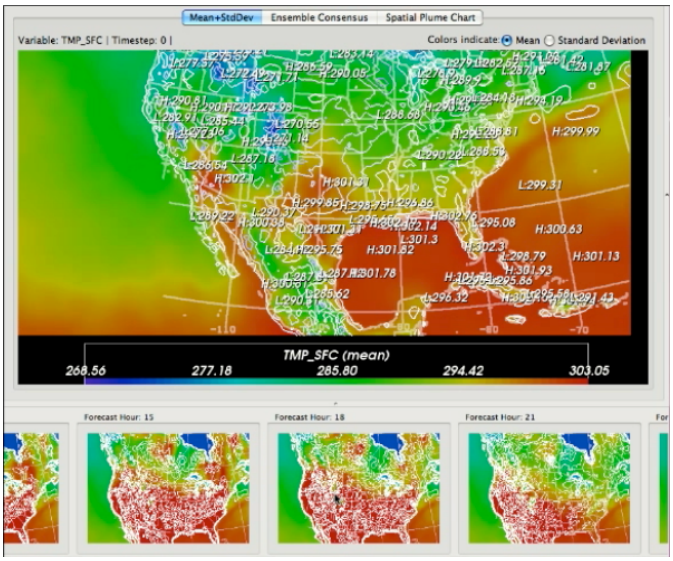

Fig. 4. We combine two representations to summarize each variable in the ensemble. A high-resolution spatial display (top) displays mean, standard deviation, and local minima and maxima for a given time step. An arrangement of lower-resolution multiples into a filmstrip (bottom) shows the same information over several time steps at once. The user can scroll through the filmstrip and transfer any time step to the high-resolution display.

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 summary view (Figure 4, top) shows the behavior of one variable over space at one time step. The filmstrip view (Figure 4, bottom) shows the same variable at lower spatial resolution over several time steps at once.

3.3.1 Spatial Summary Views

The purpose of the 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.

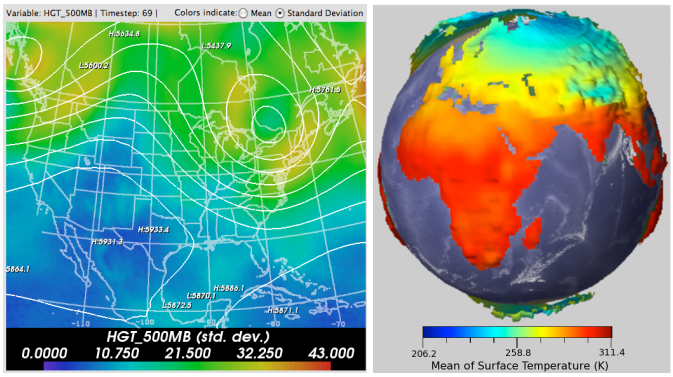

Fig. 5. We illustrate mean and standard deviation simultaneously using either color plus overlaid contours (left) or color plus height (right).

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 [27]. Second, many ensembles do not have enough members to support more sophisticated, precise characterizations.

By default, we display the variable mean using color and the standard deviation using overlaid contours (Figure 5, left). Although the rainbow colormap is generally a poor choice for scientific visualization [8], it is familiar and appropriate 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 7. 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

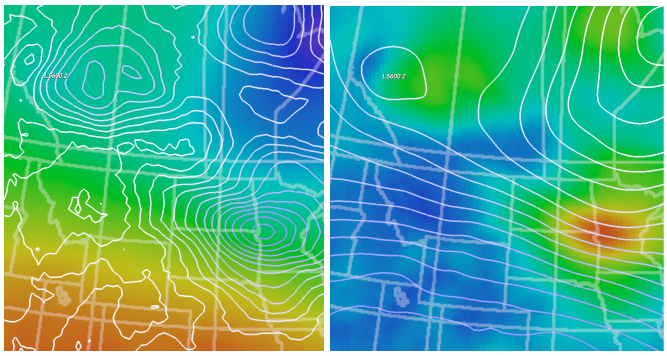

Fig. 6. The user can toggle the assignment of mean and standard deviation to colors and contours, respectively (left) or the reverse (right). Both images show the same region of the data.

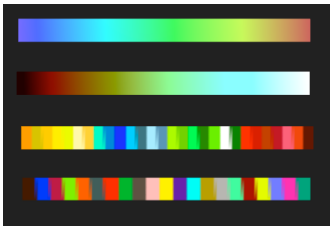

Fig. 7. Examples of our colormaps. We use a subdued rainbow colormap and a sequential low to high map for scalar variables and two categorical color maps for labeling.

deviation using a height field instead of contours. This is particularly effective when displaying 2D data projected onto the globe, as is common in climate simulations (Figure 5, left), since the height is easily visible along the silhouettes of the globe.

3.3.2 Time-Domain Summary Views

In addition to the spatial summary view, which shows a high-resolution overview of a single time step, we also provide time-domain summary views that sacrifice visible detail in order to allow quick navigation and inspection across time steps.

The *filmstrip view*, Figure 8, shows the current variable 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. Double-clicking a frame transfers it to the higher-resolution summary, query contour views, and spaghetti plot views. We also include the ability to animate climate data over time on the surface of a rotating globe. Each of these approaches has its own advantages. The filmstrip view allows the user to quickly identify exactly where in time something interesting is happening. The animated globe gives a clearer sense for the velocity of large-scale phenomena and 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. We provide two such views.

3.4.1 Quartile Charts

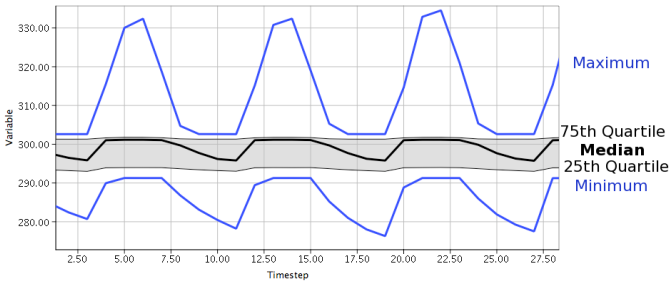

Fig. 9. Quartile trend charts. These charts show the quartile range of the ensemble within a user-selected region. Minimum and maximum are shown in blue, the gray band shows the 25th and 75th percentiles, and the median is indicated by the thick black line.

A quartile trend chart (Figure 9) displays the minimum, maximum, 25th and 75th percentiles and median of a selected 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 majority of the data values fall. As with the choice of mean and standard deviation in the summary view, this is most appropriate for unimodal distributions and can grow less informative as the data distribution grows 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.

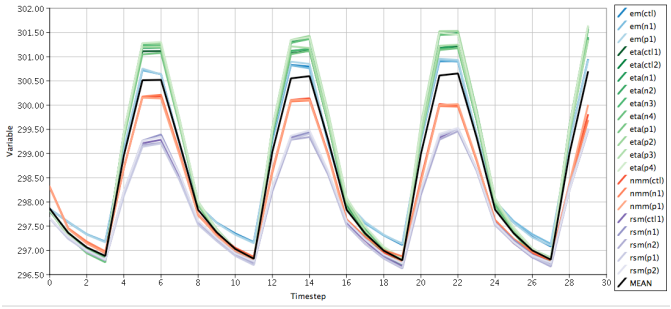

Fig. 10. Plume trend charts. These charts show the average of each ensemble model within a user-selected region of interest. Each model is type is color-coded. The thick black line shows the mean across the entire ensemble.

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. In addition, multimodal distributions can be detected in this display since multiple strong clusters of members should be readily apparent.

3.5 Condition Queries

The spatial and temporal 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 sets is for the analyst to specify a set of circumstances and ask for information about where they may occur. Such query-driven techniques [28] constrain the visualization to the subset of data deemed interesting by the analyst and discards the rest. We refer to these sets of circumstances as *conditions*.

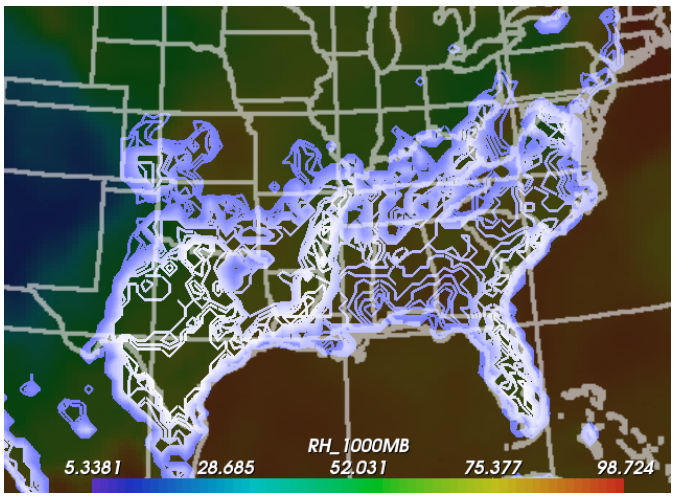

Fig. 11. The condition query view shows the probability that a given set of conditions will occur as a set of nested contours. Contour values are the fraction of the ensemble that predicts that the condition will be satisfied. In this figure we see a query for regions of dangerously high heat, defined as temperatures above 95° Fahrenheit and relative humidity above 50%.

Once an analyst specifies a condition, 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

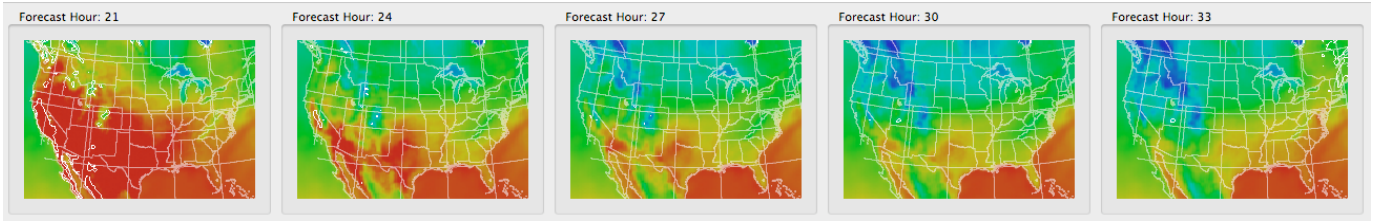

Fig. 8. The filmstrip summary view. Each frame in the filmstrip shows a single time step from the ensemble. The filmstrip also displays selection information from other views to help the user maintain a sense of context.

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 directory or (more usefully) drawn as a series of contours on a summary display as shown in Figure 11.

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.

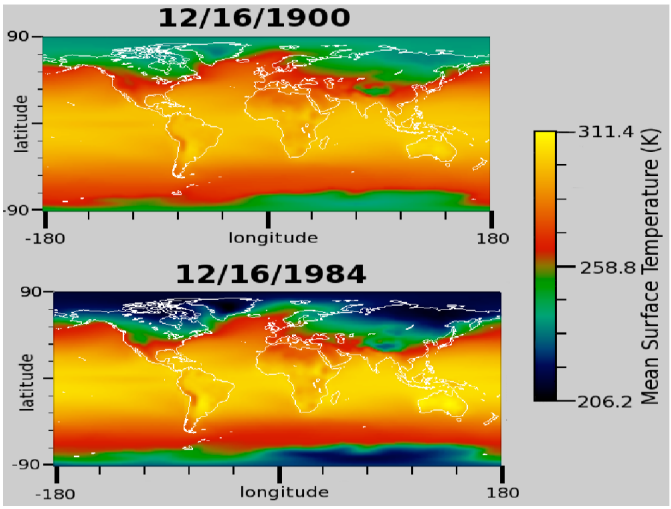

Fig. 12. Screenshot of the multivariate display. In this figure we display the average surface temperature on December 12th during both 1990 and 1994.

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 12 demonstrates the change in surface temperature for 1900 and 1984, while Figure 16 shows three elevations which add to the cloudiness across the globe.

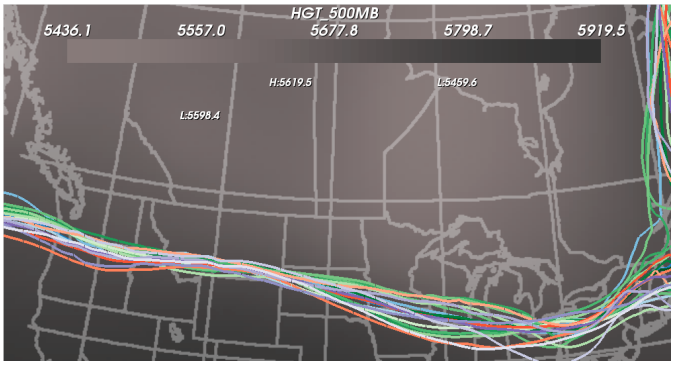

Fig. 13. A spaghetti plot displays a single isocontour from each ensemble member in order to allow examination of differences across space. When the members are in agreement the contours form coherent bundles as seen here.

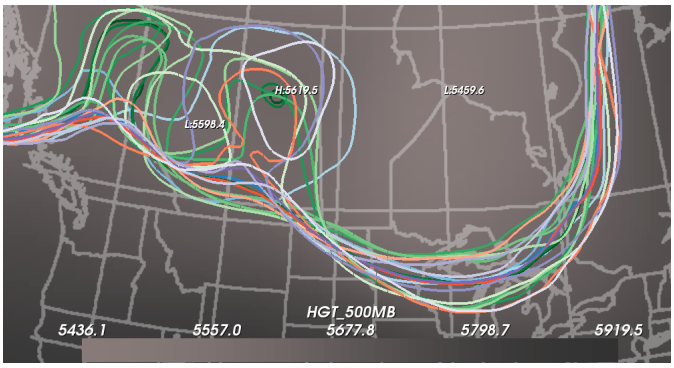

Fig. 14. Another example of spaghetti plots. In this case the ensemble members disagree in the upper left section of the image. This is visible where outliers diverge from the main bundle.

3.7 Spaghetti Plots

A *spaghetti plot* [13], 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 13, the contours will fall into a coherent bundle. When minor variation exists, a few outliers may diverge from the bundle (Figure 14). 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.

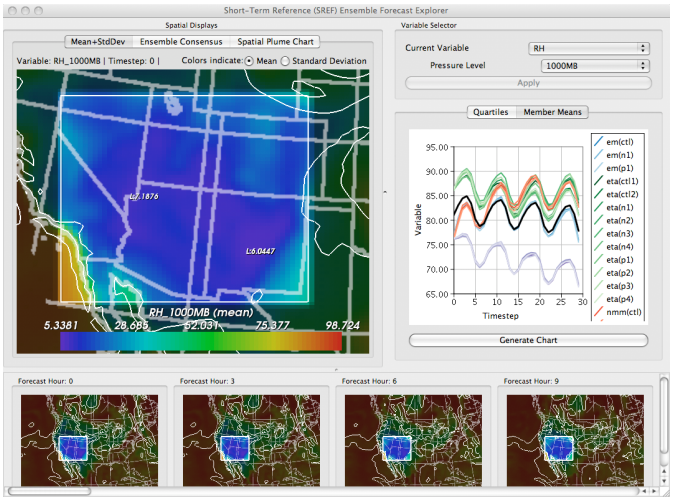

Fig. 15. Screenshot of the SREF Weather Explorer. This prototype is implemented as a set of VTK filters and can thus be easily integrated into tools deployed to domain scientists.

4.1 SREF Weather Explorer

The SREF Weather Explorer application permits ensemble analysis of a single instance of the NOAA Short-Term Reference Ensemble Forecast (SREF) data set [3]. 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 15.

The visualization algorithms in SREF Weather Explorer are implemented as filters in VTK [26], a well-known open-source toolkit for scientific visualization. The user interface components were implemented as Qt widgets [6]. 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.

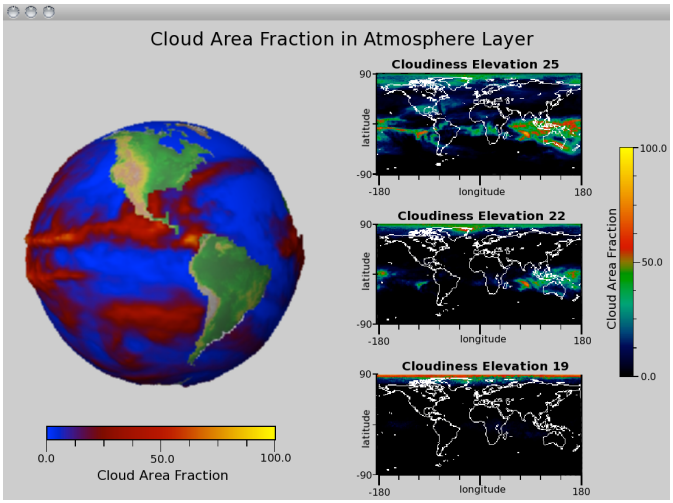

Fig. 16. Screenshot of the ViSUS prototype. This system is integrated into the CDAT framework used by climate scientists.

4.2 ViSUS/CDAT

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) [1] 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, and enhancing the 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 16.

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 and 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.

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 well 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, as we see from the distribution of members in Table 1, 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 [7]. 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 paper we have presented an approach to ensemble visualization using a federation of simple, familiar representations that are immediately familiar to domain scientists. Our main contribution is the flexible organization and linking of these representations to focus on the formulation and evaluation of hypotheses 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. We believe that our work constitutes early progress toward the many new challenges posed by these large, complex and rich data sets.

References

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

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

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

[4] 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.

[5] R. Becker and W. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, May 1987.

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

[7] J. S. D. Bonet and Q. Zaidi. Comparison between spatial interactions in perceived contrast and perceived brightness. *Vision Research*, 37(9):1141–1155, May 1997.

[8] D. Borland and R. T. II. Rainbow color map (still) considered harmful. *IEEE CG & A*, 27(2):14–17, Mar/April 2007.

[9] A. Buja, D. Cook, and D. F. Swayne. Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics*, 5(1):78–99, March 1996.

[10] R. Bürger and H. Hauser. Visualization of multi-variate scientific data. In *Eurographics 2007 STAR*, pages 117–134, 2007.

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

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

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

[14] S. Djurcilov, K. Kim, P. Lermusiaux, and A. Pang. Volume rendering data with uncertainty information. In *Data Visualization*, pages 243–52, 2001.

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

[16] B. Hubbard, J. Kellum, B. Pual, D. Santek, and A. Battaiola. Vis5d. http://vis5d.sourceforge.net.

[17] C. R. Johnson and A. R. Sanderson. A next step: Visualizing errors and uncertainty. *IEEE CG & A*, 23(5):6–10, 2003.

[18] A. Love, A. Pang, and D. Kao. Visualizing spatial multivalue data. *IEEE CG & A*, 25(3):69–79, May 2005.

[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*, 32(3):139–160, July 2005.

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

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

[22] R. S. A. Osorio and K. Brodlie. Contouring with uncertainty. In I. S. Lim and W. Tang, editors, *6th Theory and Practice of Computer Graphics Conference*, pages 59–66, 2008.

[23] A. Pang, C. Wittenbrink, and S. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, Nov 1997.

[24] K. Potter, J. Kniss, and R. Riesenfeld. Visual summary statistics. Technical Report UUCS-07-004, Univeristy of Utah, 2007.

[25] 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*, pages 61–71.

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

[27] J. Sivillo, J. Ahlquist, and Z. Toth. An ensemble forecasting primer. *Weather Forecasting*, 12:809–817, 1997.

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

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

[30] C. M. Wittenbrink, A. T. Pang, and S. K. Lodha. Glyphs for visualizing uncertainty in vector fields. *IEEE Transactions on Visualization and Computer Graphics*, 2(3):266–279, September 1996.

[31] 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*, pages 183–190, 2006.