THE VISUALIZATION OF UNCERTAINTY

by

Kristin Potter

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

School of Computing

The University of Utah

August 2010

THE UNIVERSITY OF UTAH GRADUATE SCHOOL

SUPERVISORY COMMITTEE APPROVAL

of a dissertation submitted by

Kristin Potter

This dissertation has been read by each member of the following supervisory committee and by majority vote has been found to be satisfactory.

	Chair: Richard F. Riesenfeld
	Chrisopher R. Johnson
	Elaine Cohen
	Robert M. Kirby
	Bruce Gooch

ABSTRACT

The graphical depiction of uncertainty information is emerging as a problem of great importance in the field of visualization. Scientific data sets are not considered complete without indications of error, accuracy, or levels of confidence, and this information is often presented as charts and tables alongside visual representations of the data. Uncertainty measures are often excluded from explicit representation within data visualizations because the increased visual complexity incurred can cause clutter, obscure the data display, and may lead to erroneous conclusions or false predictions. However, uncertainty is an essential component of the data, and its display must be integrated in order for a visualization to be considered a true representation of the data. The growing need for the addition of qualitative information into the visual representation of data, and the challenges associated with that need, command fundamental research on the visualization of uncertainty.

This dissertation seeks to advance approaches for uncertainty visualization by exploring techniques from scientific and information visualization, creating new visual devices to handle the complexities of uncertainty data, and combining the most effective display methods into the *Ensemble-Vis* framework for visual data analysis. Many techniques exist for graphical data display. However, their usage on data with uncertainty information is not straightforward. This work begins by first exploring existing methods for data visualization and assessing their application to uncertainty. New visual metaphors are then presented for the depiction of salient features of data distributions, including indications of uncertainty. These new methods are inspired by proven visual data analysis techniques, but account for the requirements of large, complex data sets. Finally, *Ensemble-Vis* is presented, which combines effective uncertainty visualization techniques with interactive selection, linking, and querying to provide a user-driven, component-based framework for data investigation, exploration, and analysis.

To Mom and Dad

CONTENTS

ABSTRACT . . . . . **iii**

LIST OF FIGURES . . . . . **viii**

ACKNOWLEDGEMENTS . . . . . **xi**

CHAPTERS

1. INTRODUCTION **1**

1.1 Types of Uncertainty 2
1.1.1 Experimental Uncertainty 3
1.1.2 Geometric Uncertainty 3
1.1.3 Simulation Uncertainty 4
1.1.4 Visualization Uncertainty 5
1.2 The Need for Uncertainty Visualization 5
1.3 Contributions 10
1.4 Overview 11

2. TECHNICAL BACKGROUND **13**

2.1 Ensemble Data 13
2.2 Probability Density Functions 14
2.3 Descriptive Statistics 15
2.4 Uncertainty 17

3. VISUALIZATION OF SUMMARY STATISTICS **19**

3.1 Introduction 19
3.2 Related Work 20
3.2.1 Statistical Plotting Techniques 21
3.2.1.1 The Boxplot 21
3.2.1.1.1 Origins. 22
3.2.1.2 Modifications to the Boxplot 23
3.2.1.2.2 Density information. 24
3.2.1.2.3 Additional descriptive statistics. 26
3.2.1.3 Bivariate Extensions 29
3.3 The Summary Plot 32
3.3.1 The Abbreviated Boxplot 32
3.3.2 Quartiles and the Histogram 34
3.3.3 Moments 35
3.3.3.1 Mean and Standard Deviation 36
3.3.3.2 Skew 37
3.3.3.3 Kurtosis 37

3.3.3.4 Tail . . . 38
3.3.3.5 Moments, Sample Size, and Outliers . . . 38
3.3.4 Distribution Fitting . . . 39
3.3.5 User Interface for Reduction of Visual Clutter . . . 40
3.3.6 Comparison of the Box and Summary Plots . . . 40
3.4 Joint 2D Summaries . . . 50
3.4.1 Joint Mean and Standard Deviation . . . 50
3.4.2 Joint Density . . . 50
3.4.3 Covariance and Skew Variance . . . 54
3.4.4 Correlation . . . 56
3.5 Discussion . . . 57
3.6 Conclusion . . . 59

4. THE VISUALIZATION OF MULTIDIMENSIONAL UNCERTAINTY DATA . . . 60

4.1 Introduction . . . 60
4.2 Application Data . . . 61
4.3 Related Work . . . 64
4.4 Two-Dimensional Techniques . . . 65
4.4.1 Colormapping . . . 65
4.4.2 Bivariate Colormaps . . . 69
4.4.3 Perceptual Considerations . . . 73
4.5 Three-Dimensional Techniques . . . 75
4.5.1 Displacement Mapping . . . 75
4.5.2 Volume Rendering and Isosurfacing . . . 78
4.5.3 Streamlines and Particle Tracing . . . 80
4.6 Conclusion . . . 82

5. ENSEMBLE-VIS . . . 84

5.1 Introduction . . . 84
5.1.1 Motivation . . . 85
5.1.2 Driving Problems . . . 85
5.1.2.1 Weather Forecasting . . . 85
5.1.2.2 Climate Modeling . . . 86
5.1.3 Ensemble Data Sets . . . 86
5.1.3.1 Ensembles and Uncertainty . . . 87
5.1.3.2 Challenges for Analysis . . . 87
5.2 Related Work . . . 88
5.2.1 Visualization of Climate and Weather Data . . . 88
5.2.1.1 Multidimensional Data Visualization . . . 89
5.2.2 Uncertainty Visualization . . . 91
5.2.2.1 Comparison Techniques . . . 91
5.2.2.2 Attribute Modification . . . 93
5.2.2.3 Glyphs . . . 94
5.2.2.4 Image Discontinuity . . . 95
5.3 The Ensemble-Vis Framework . . . 95
5.3.1 Work Flow . . . 96
5.3.2 Data Sources . . . 96

5.3.3 Ensemble Overviews . . . 97
5.3.3.1 Spatial-Domain Summary Views . . . 97
5.3.3.2 Time Navigation Summary Views . . . 99
5.3.4 Trend Charts . . . 103
5.3.4.1 Quartile Charts . . . 103
5.3.4.2 Plume Charts . . . 103
5.3.5 Condition Queries . . . 104
5.3.6 Multivariate Layer Views . . . 105
5.3.7 Spaghetti Plots . . . 107
5.3.8 Coordination Between Views . . . 107
5.3.9 Clustering . . . 108
5.4 Implementation Details . . . 110
5.4.1 SREF Weather Explorer . . . 110
5.4.2 ViSUS/CDAT . . . 113
5.5 Discussion . . . 113
5.5.1 Data Challenges . . . 113
5.5.2 Where Summary Statistics Break Down . . . 115
5.5.3 Glyphs for Standard Deviation . . . 116
5.6 Conclusion . . . 116

6. CONCLUSIONS AND FUTURE WORK . . . 117

6.1 Future Work . . . 117

REFERENCES . . . 119

LIST OF FIGURES

1.1 Geometric uncertainty. 4
1.2 Visualization uncertainty. 6
1.3 A computed tomography (CT) scan of a brain with a tumor. 7
1.4 Brain tumor volume renderings. 8
1.5 A photograph and a 3D reconstruction of a Mayan temple. 9
1.6 Weather variable outlooks for aviation. 10
2.1 The Gaussian, or normal, distribution. 16
3.1 The boxplot and visual modifications. 23
3.2 The density modifications to the boxplot 26
3.3 The sectioned density plot. 27
3.4 Data attribute modifications of the boxplot. 28
3.5 Descriptive statistical plots. 29
3.6 Bivariate extensions of the boxplot. 31
3.7 Anatomy of the summary plot. 33
3.8 Cumulant and regular histograms. 35
3.9 Moment arm abstraction. 37
3.10 Mean and median. 37
3.11 Glyphs for the higher order central moments. 39
3.12 Distribution fitting. 41
3.13 User interface for data information. 42
3.14 User interface for display options. 43
3.15 User interface for density information. 44
3.16 User interface for distribution fitting options. 45
3.17 A comparison of the box and summary plots. 46
3.18 Temperature and humidity data. 47
3.19 Wet bulb and pressure data. 48
3.20 Bash *du* command run in VTK directories. 49
3.21 Joint summary for two 1D categorical data sets. 51

3.22 Joint summary and histogram on a single data set. 52
3.23 Joint summary and histogram on multiple data sets. 53
3.24 Close-up of the joint summary plot. 55
3.25 Correlation between the data from Figures 3.18 and 3.19. 57
4.1 The classified human torso. 63
4.2 The torso mesh. 63
4.3 Mean and standard deviation of the torso data. 63
4.4 Colormaps of mean. 67
4.5 Colormaps of standard deviation. 68
4.6 Colorspaces. 71
4.7 Bivariate colormap of mean and standard deviation. 72
4.8 Uncertainty encoded in the triangular mesh. 74
4.9 Various visual phenomenon. 76
4.10 The (x, y, potential) space. 77
4.11 Direct volume rendering of the κ-volume. 79
4.12 Isosurfaces of voltages in the κ-volume. 80
4.13 Streamlines. 81
4.14 Particle Tracing. 83
5.1 An example of the complexity of an ensemble data set. 89
5.2 Volume rendering of ocean salinity. 92
5.3 Flow visualizations of currents. 93
5.4 Uncertainty contours. 94
5.5 Uncertainty encoded using parallel coordinates and star glyphs. 94
5.6 The Ensemble-Vis work flow. 97
5.7 Spatial and time domain overviews. 98
5.8 Mean and standard deviation display. 100
5.9 Colormaps used within Ensemble-Vis. 101
5.10 Toggle interface between colormaps and contours. 101
5.11 Height mapped onto the earth globe. 102
5.12 The filmstrip summary view. 102
5.13 Quartile trend charts. 104
5.14 Plume trend charts. 105
5.15 Condition queries. 106

5.16 Multivariate display. 108
5.17 Spaghetti plots. 109
5.18 Clustering based on spatial location and mean data value. 111
5.19 SREF weather explorer. 112
5.20 ViSUS prototype. 114

ACKNOWLEDGEMENTS

Many thanks to:

⋆ My committee.

⋆ Collaborators: Andrew Wilson, Peer-Timo Bremer, Valerio Pascucci, Jens Krueger, Roni Choudhury, Sarah Geneser, Joe Kniss, and Erik Anderson.

⋆ The Scientific Computing and Imaging Institute and the School of Computing.

⋆ Karen Feinauer, a.k.a. wonder woman.

⋆ Nathan Galli, Chems Touati, and Erik Jorgensen for making me look good.

⋆ The ladies: Liz, Miriah, and Betty. Ladies night is cheaper than therapy!

⋆ The old school: Amy Gooch, Aaron Lefohn, Shaun Ramsey, Chris Wyman, Erik Rheinhard, Gordon Kindlman, Xaiver Tricochoe, Mike Stark, Ramy Sadek, Taylor Erickson, Joel Daniels, and Dave Demarle.

⋆ The new school: Sam Gerber, Roni Choudhury, Sylvain Gouttard, Austin Robison, Jacob Hinkle, Sam Preston, Dave Nellans, Mike Steffen, Ryan Vance, and Chelsea Robertson.

⋆ Mom and Dad for their unwaivering support.

CHAPTER 1

INTRODUCTION

This dissertation is about the visualization of data that is accompanied by some measure of uncertainty. By including information describing qualitative aspects of a data set within visualization, scientists are able to improve their understanding of the data by becoming more aware of its intrinsic characteristics. Such information is difficult to express visually. However, its inclusion is necessary to further the field of visualization. The work presented in this dissertation investigates the challenges encountered when working with uncertainty data and presents novel visualization methods to address these challenges.

Computational advances are allowing large, complex data sets to gain in popularity as a way for mitigating errors in numerical models, reducing unknowns in initial conditions and evaluating the influence of parameter sensitivity. As computational power, memory limits, and bandwidth have inexorably increased, so has the corresponding size and complexity of the data sets generated by scientists. Because of the reduction of hardware limitations, scientists are able to run simulations at higher resolution, for longer amounts of time, and using more sophisticated numerical models. We can compute more exhaustively, store more data, and access data more rapidly, all of which leads scientists to create more complex systems to increase the accuracy and reduce the error in simulations of complex systems.

With increased size and complexity of data becoming more common, visualization and data analysis techniques are required that not only address issues of large scale data, but also allow scientists to understand better the processes that produce the data, and the nuances of the resulting data sets. Information about uncertainty, including confidence, variability, as well as model bias and trends are now available in these data sets, and methods are needed to address the increased requirements of the visualization of these data. Too often, these aspects remain overlooked in traditional visualization approaches; difficulties in applying preexisting methods, escalating visual complexity, and the lack of

obvious visualization techniques leave uncertainty visualization an unsolved problem.

Effective visualizations present information in a manner that encourages data understanding through the appropriate choice of visual metaphor. Data are used to answer questions, test hypotheses, or explore relationships and the visual presentation of data must facilitate these goals. Visualization is a powerful tool allowing great amounts of data to be presented in a small amount of space, however, different visualization techniques are better then others for particular types of data, or for answering specific questions. Using the most befitting visualization method based on the data type and motivated by the intended goals of the data results in a powerful tool for scientists and data analysts.

The effective visualization of uncertainty, however, is not always possible through the simple application of traditional visualization techniques. Often, the visualization of the data itself has a high visual complexity, and the addition of uncertainty, even as a scalar value, over-complicates the display. Issues of visual clutter, data concealment, conflicts in how the data and the uncertainty are represented, and unintentional biases are just some of the problems incurred when visualizing data accompanied by uncertainty. Also, the complexity of these data sets may not lend themselves to the straightforward application of existing visualization methods, and thus, the added burden of uncertainty can be overwhelming.

Uncertainty data are becoming more prevalent and can be found in fields such as medical imaging, geoscience, and mechanical engineering. The simulation of complex systems, compilation of sensor data, and classification of tissue type are but a few sources of uncertainty data and their expression, size, and complexity can drastically vary. Uncertainty can arise in all stages of the analysis pipeline, including data acquisition, transformation, sampling, quantization, interpolation, and visualization. It can be a single scalar value presented alongside the original data, or can be an integral aspect of the data, derived from the description of the data itself. In any case, uncertainty is an imperative component of scientific data sets and should not be disregarded in visualizations.

1.1 Types of Uncertainty

Uncertainty can be classified as particular types depending on how they arise in the data set. Experimental, geometric, simulation, and visualization uncertainties are some of the most often seen types of uncertainty. While each of these types of uncertainty

result in data sets with additional indications of quality, understanding the origin of the uncertainty can help in determining appropriate visualization paradigms.

1.1.1 Experimental Uncertainty

The National Institute for Standards and Technology (NIST) defines experimental uncertainty as the standard deviation of a collection of measured results [77]. This type of uncertainty comes from running an experiment numerous times or performing a non-deterministic simulation in which the outcome varies after each run. Incorporating this uncertainty into the resulting data set allows for scientists to reduce the error present in a data set by reproducing an experiment over and over, and using an average of the collection of results to estimate the true outcome. The assumption is that the majority of the experimental results will have values close to each other, and outlying results, stemming from errors in the experiment, will be identified as flawed results and thrown out, or their effect reduced through averaging.

1.1.2 Geometric Uncertainty

Another frequently occurring type of uncertainty is called *geometric uncertainty*, in which the spatial position of some or all of the data set is in question. This type of uncertainty arises, for example, in molecular simulations where the final location of a molecule may be within a region of space rather than at a specific location, or laser scanned data sets where the laser scanner may not have a high enough resolution to accurately sample a model. In these cases, uncertainty may describe the amount of possible variation a point may lie from a particular location, or may describe a boundary region within which the data point will be positioned.

One example of a visualization of geometric uncertainty can be seen in Figure 1.1. In this image, the yellow spheres represent the spatial uncertainty of the points on the surface. The larger the uncertainty in the position, the larger the sphere [55]. Thus, this figure shows that the largest uncertainties exist at the minimum and maximum peaks of this surface because this is where the largest spheres exist. Displaying uncertainty in this way helps the viewer understand not only where the location of the surface is in question, but also the relative quantity of uncertainty that exists at each point.

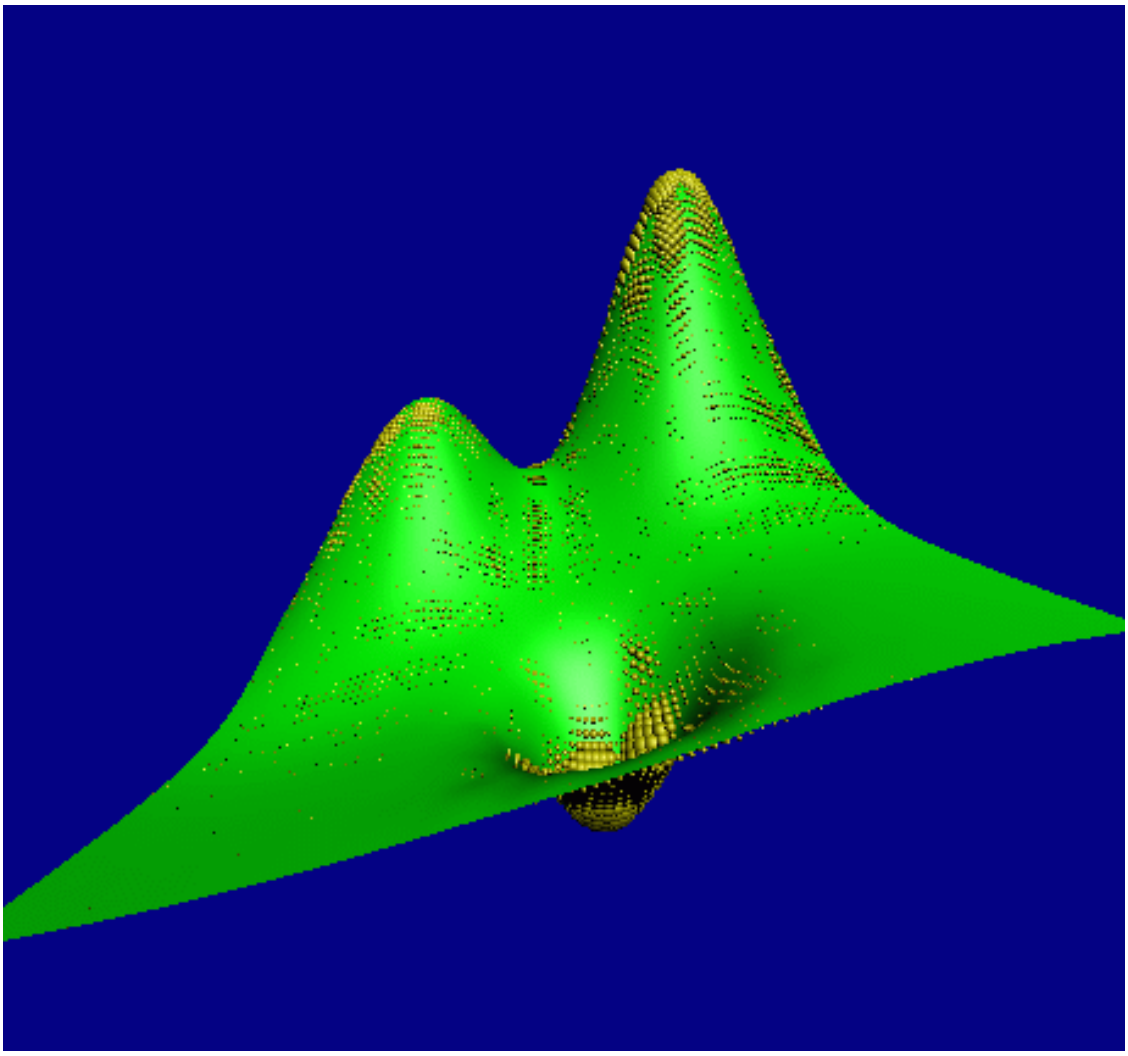


Figure 1.1. Geometric uncertainty. Unknowns exist in the spatial positions of the surface. The spheres represent the possible positions of the uncertain data points; the larger the sphere the more uncertainty in the position.

1.1.3 Simulation Uncertainty

Simulation uncertainty arises from multimodel simulations, such as the kind found in weather and climate simulations. These simulations combine multiple numerical models that each estimate scientific variables. This is similar to experimental uncertainty in that multiple values are used to estimate an outcome, however, in this instance, distinct models are combined rather than a single experiment. Each of the models may be run numerous times, often using a variety of initial conditions or perturbations on parameters. While experimental uncertainty relives errors from an erroneous model run or experiment, simu-

lation uncertainty mitigates errors that can stem from particular models. For example in weather simulation, a particular model may exhibit a biased prediction in a certain region of the country, while be a reliable estimator in another. Combining multiple models will reduce the overall error of the simulation by reducing the effect of the erroneous model in the results.

1.1.4 Visualization Uncertainty

Uncertainty can also arise from the visualization process itself. Visualization techniques must often infer aspects of a data set that may not be explicitly present in order to visually display the data. For example, different methods for calculating isosurfaces exist, and these methods may not generate exactly the same isosurface for the same data set [70]. Similarly, the choice of transfer function or colormap can be highly influential on understanding the data; a particular colormap may highlight aspects of the data that are hidden by the choice of another color scheme.

A concrete example of uncertainty within visualizations is shown in Figure 1.2. In this example, blood flow through an artery is shown [57]. On the left, the artery appears to shrink, signaling to the doctor that a stenosis exists. However, on the right, the same data are shown using a different transfer function and the artery appears healthy. A doctor using the visualization that diagnosed a stenosis would likely misprescibe some sort of procedure, such as bypass surgery, when simply changing the visualization parameters reveals that no such actions are required. This leads us to the question of which visualization is actually right, and how do we choose appropriate visualization techniques in order to accurately convey the data? Unfortunately, this question is extremely challenging and extends beyond the research in this dissertation. For now, we leave it to the visualization experts to understand the questions asked of their data, and conscientiously select display parameters in order to responsibly visualize the data.

1.2 The Need for Uncertainty Visualization

A specific example of a scientific data set with uncertainty comes from the field of medical imaging. The central image in Figure 1.3 shows a computed tomography (CT) scan of a human brain with a brain tumor. The tumor can be seen as the lighter grey, fuzzy mass and two close-ups of the tumor's boundary can be seen in the insets. On the left, the boundary is fairly clear. We see that the two solid lines encompass the region within which the boundary lies, and the central dotted line is an appropriate approximation of

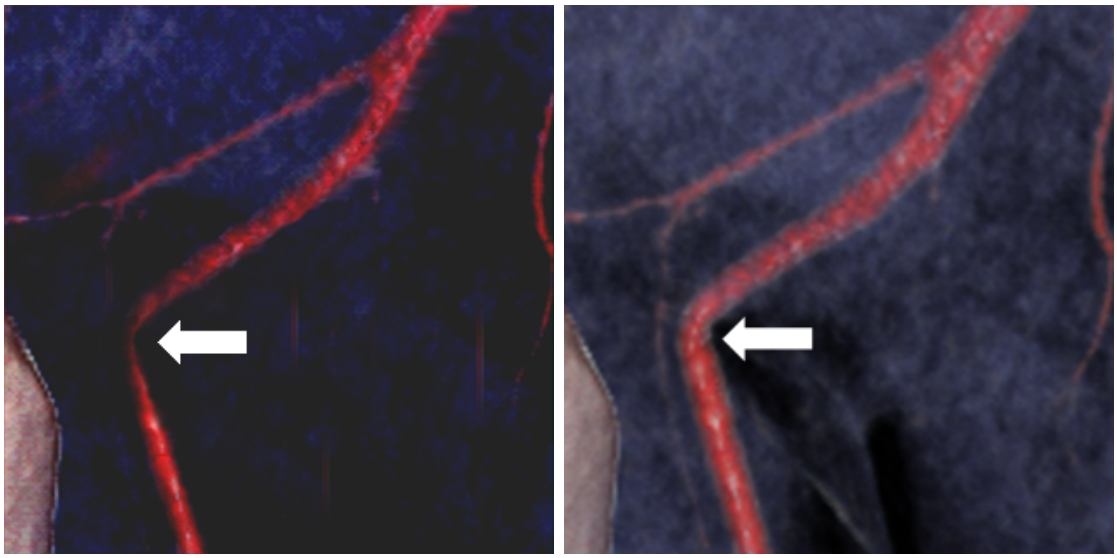


Figure 1.2. Visualization uncertainty. Uncertainty arising from the visualization itself. On the left, a stenosis, or blocked artery, is detected because in the visualization the artery appears to shrink in width. However, by simply changing the visualization parameters (right) the artery appears to be healthy.

the boundary. Because this boundary is visibly distinct from its surrounding tissue types, the confidence of the dotted line being the tissue boundary is high. However, on the right side, the position of the tumor boundary is not as clear. The difference in intensity of the image value between tumor and surrounding tissue is very small, leading to a large area in which the tumor boundary can fall, again denoted by the solid blue lines. Again, the dotted line indicates the possible position of the tumor boundary, but in this case, the confidence of this line is much lower.

Another visualization of these data can be seen in Figure 1.4. Here, the data are volume rendered and two different boundaries of the tumor are shown on the left and right. While the differences between the two representations of the tumor are not huge, the general location of the tumor is readily apparent, and the consequences of deciding upon the fine details of a particular boundary position may have a great impact. Because this tumor lies within the white and grey matter of the brain, destruction of the surrounding tissue may have serious negative repercussions. If the boundary of the tumor is chosen such that excess brain tissue is removed, the patient can be significantly injured. However, if the tumor is not completely removed, it can regrow, forcing surgeons to repeat costly and invasive surgeries. Additionally, understanding the possible position of boundaries

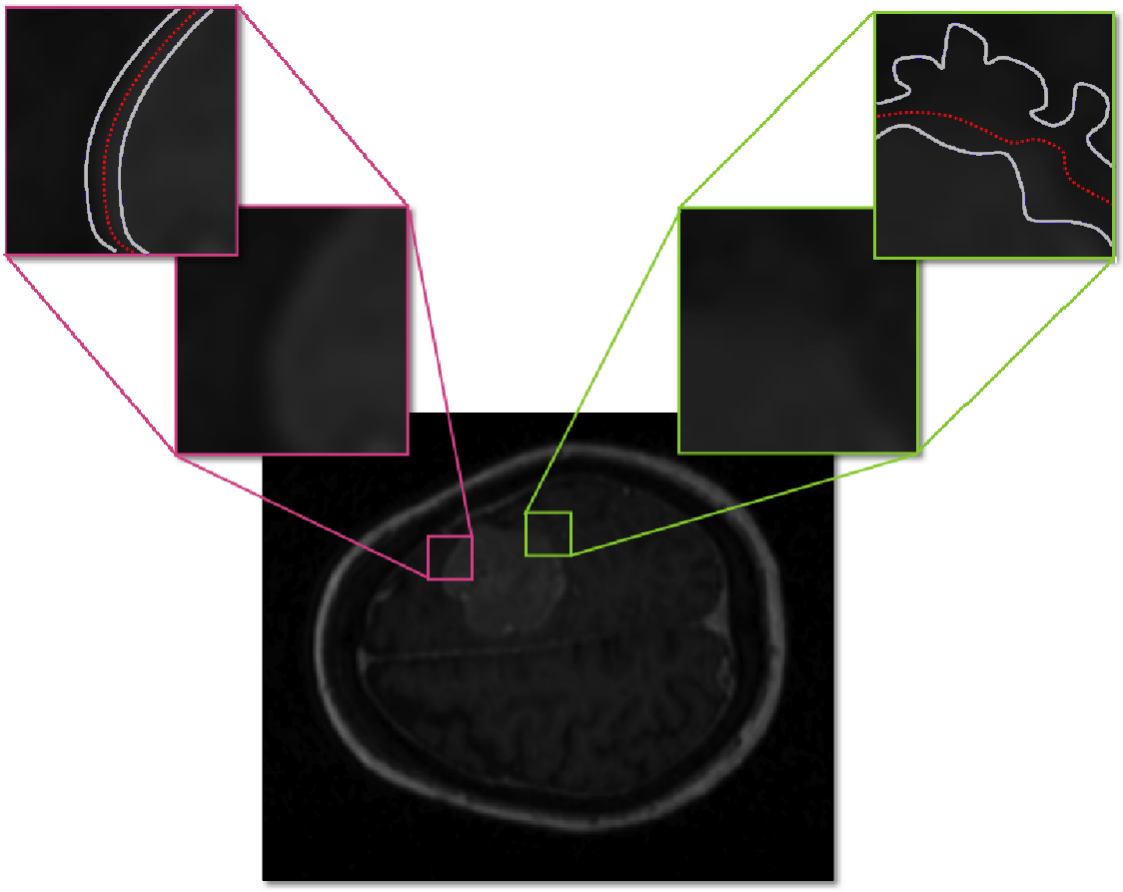


Figure 1.3. A computed tomography (CT) scan of a brain with a tumor. The first level of insets shows closeups of the boundary between tumor and brain tissue for two different locations. The second level shows the region within which the boundary lies (outer solid lines) and a possible position for the boundary (inner dotted lines). The closer together the solid lines are the higher the confidence is that the dotted line is the actual tumor boundary; thus, the left boundary position has a higher confidence level than the one on the right.

can help in surgery planning; incisions close to the actual surgery location minimize damage and hasten recovery time for the patient. Thus, informing the doctors and surgeons of the boundaries of tumor tissue, as well as the level of confidence in those boundaries, can improve surgery outcomes.

Another example of uncertainty data outside of the medical field is shown in Figure 1.5. In this example, a photograph of an incomplete Mayan temple is shown on the top, and on the bottom, a three-dimensional (3D) computer model of the temple with indications of the uncertainty [67]. The existent structure of the temple, as seen in the photograph,

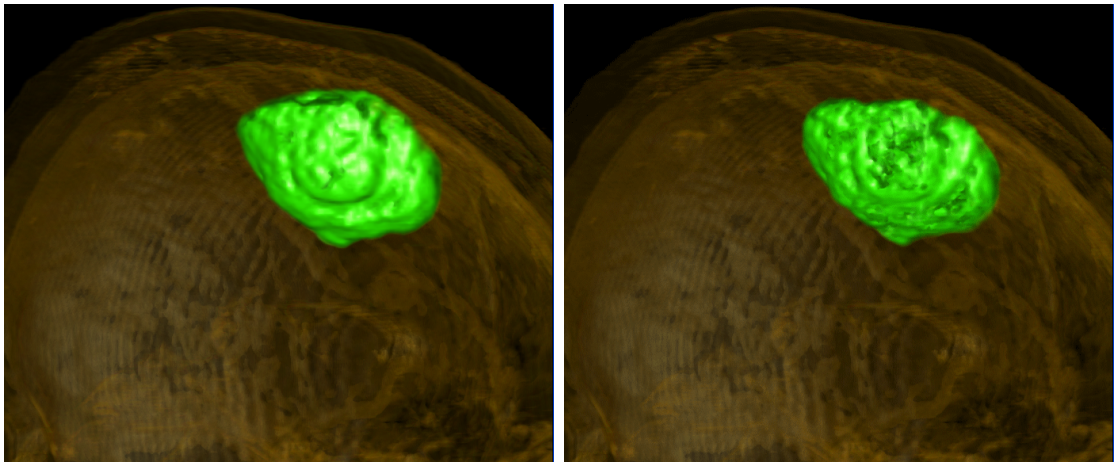


Figure 1.4. Brain tumor volume renderings. Two possible tumor tissue boundary positions are displayed on the left and right.

consists of a lower base and an upper pile of stone rubble. The reconstruction resurrects the rubble pile as another level of stonework and shows a wooden hut structure on top. The rendering expresses the amount of confidence in the reconstruction through the use of line textures. The structure of the base of the temple has a high confidence level since it is still in existence, and thus, it is rendered using solid, straight lines. The next level is rendered using a less connected and more random line texture, which expresses the known fact that stones were used in the construction of this level, as evidenced by the rubble pile, but the actual structure of the level is assumed from knowledge of nearby temples. The topmost hut is rendered in a very sketchy style, without any textural indications of texture. This reflects that the hut is a matter of speculation based on finding wood fragments scattered throughout the rubble, and its actual existence is arguable.

Both of the previous examples view uncertainty as a modifier expressing the amount of confidence across the data set. Uncertainty can also describe error present in various stages from acquisition to visualization, or illuminate the amount of variation within the data. An example of the latter comes from the field of meteorology in which numerous computational models are combined to form a short-term weather outlook for aviation. In the example shown in Figure 1.6, the results of numerous runs of simulations are combined, and the amount of variation between the runs is encoded using a blue to red rainbow colormap. This type of uncertainty is less of a modifier and more an integral part of the data set. Each simulation run produces a sample across the spatial domain,

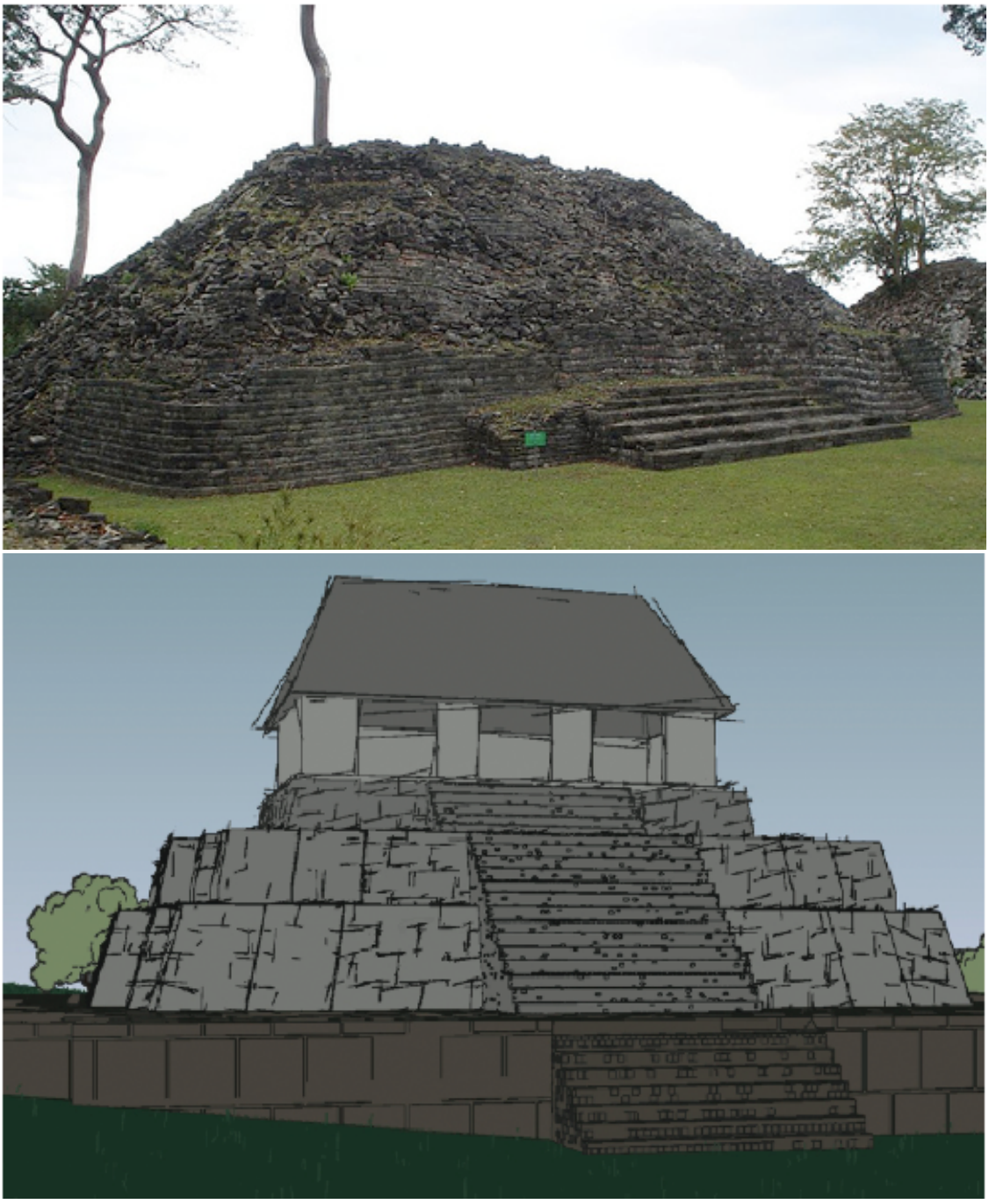


Figure 1.5. A photograph and a 3D reconstruction of a Mayan temple. The photograph shows the actual state of the temple, while the model expresses how the site may have once looked. Variations in the rendering style express levels of confidence in the reconstruction.

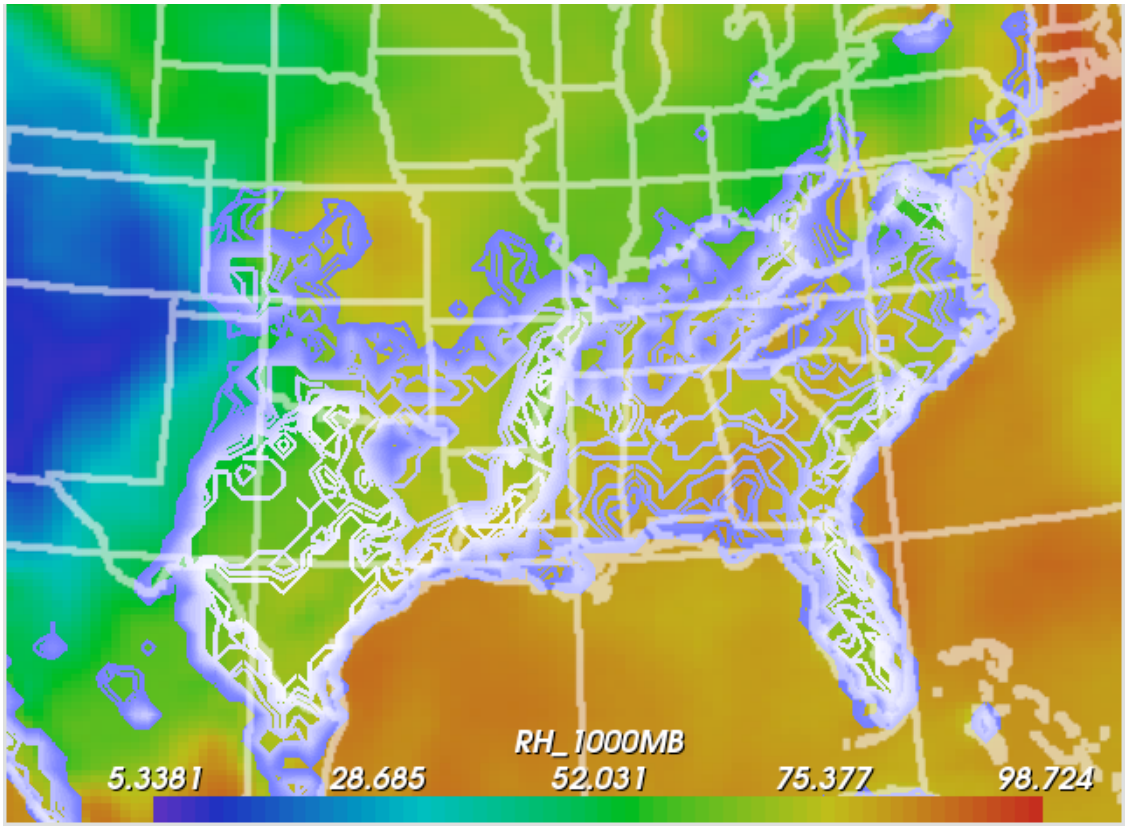


Figure 1.6. Weather variable outlooks for aviation. Uncertainty is expressed through standard deviation and shown in the white to blue contours.

and the collection of samples provides the scientist with a data set including quantitative information on the quality of the data.

The ultimate goal of uncertainty visualization is the seamless integration of uncertainty information within the data presentation. Keeping this possibly unattainable goal in mind, this dissertation seeks to further uncertainty visualization by investigating methods for the visual expression of uncertainty. To this end, quantifiable measures of uncertainty are focused on, existing visualization techniques are evaluated, novel visual metaphors are introduced, and finally a framework for the analysis and visualization of uncertainty data is presented.

1.3 Contributions

This dissertation investigates visualization techniques for data with associated uncertainty information. Pursuing the above stated goal, this dissertation provides the

following contributions:

- *An investigation of visualization techniques for uncertainty data.* The first step this work takes towards understanding uncertainty data is to investigate existing visualization techniques for this type of data. First, the data are reduced to mean and standard deviation, and methods for displaying these values in a single display are examined. Next, using the entire data set, 3D visualization techniques are explored and the issues of these types of approaches are analyzed.
- *New visualization techniques are proposed for displaying significant characteristics of a data distribution describing the data.* Because much of this type of data can be thought of as a probability distribution function, or PDF, a new approach for visualizing 1D PDFs is presented, as well as its extension to 2D. A use case is presented to illustrate both the application of this method as well as highlight the advantages and limitations.
- *The Ensemble-Vis framework for the visualization and graphical data analysis of PDF data is presented.* Because many of the approaches to visualization require data to be abstracted in some manner, we present a visualization and graphical data analysis framework whose goal is to provide a collection of summarization and statistical displays: techniques that afford the scientist the power to drive the visualization.

1.4 Overview

The remainder of this dissertation is organized as follows:

Chapter 2 presents a technical background on how uncertainty is derived from various types of data. This chapter presents an overview of how these data are created and the representation of uncertainty. Ensuing discussion is devoted to considerations of the specific data sets employed in this work.

Chapter 4 presents an investigation into existing visualization techniques for uncertainty data. Traditional visualization techniques are explored, including colormaps and heightfields. However, the simplicity of these methods require the data to be reduced to a single or bivalued datum at each grid point. Because this distills away large amounts of data, 3D visualization methods are examined, including isosurfacing and volume rendering. These approaches provide a greater insight into the processes used to create the data and the relationships present in this type of complex data.

Chapter 3 proposes new visualization techniques for expressing notable attributes of a 1D data set and an extension of this plot to 2D. A use case is presented that demonstrates the usability of this approach, as well as the challenges of expressing this type of data in dimensions greater than two.

Chapter 5 presents the Ensemble-Vis framework for visualization and graphical data analysis of *ensemble* data, which is data created by combining the results of multiple numerical simulation models. Here, a collection of visualization methods are presented that enable the scientist to explore various aspects of the data through summarization and reduction. This framework is demonstrated employing data from the fields of meteorology and climate modeling.

Finally, Chapter 6 discusses some of the issues confronted in working with data of this type and addresses some of the remaining unresolved challenges. Future work is discussed, including extensions of visualization techniques presented in this work to higher-dimensional data. Techniques for improved understanding of the shape of probability distribution functions in three or more dimensions are also visited.

CHAPTER 2

TECHNICAL BACKGROUND

The unifying theme of this dissertation is the visual representation of data that uses multiple values to estimate a variable and the uncertainty that arises from disagreements within this collection of values. This type of data can be generated, for example, from simulations using more than one numerical model to approximate results, nondeterministic processes that produce different outcomes when run multiple times, and experiments that explore the sensitivity of control conditions such as the initial state of a system or empirical parameters. These data sets are used to mitigate error stemming from a single, imperfect model and to account for inaccuracies in measurements of starting conditions or the influence of parameter sensitivities and biases. While each distinct data set may vary from others in ways such as the number of predicted variables, the spatial or temporal domains, or the size of the sample set, they all indicate measures of uncertainty. This chapter discusses the general aspects of this type of data, including how uncertainty is characterized.

2.1 Ensemble Data

An exemplary type of data expressing uncertainty is ensemble data. This class of data comprises a collection of models, called *members*, each predicting on the same variable. In the most broad sense, an ensemble can be thought of as a committee in which each member has a vote on the outcome. The strength of the ensemble relies on the fact that the members do not always agree, however, each member is competent in casting a vote; that is, the models predicting the outcome may not agree on the resulting outcome, however, each model predicts a valid value. Thus, all member votes are contributing; if predictions from the members always agreed, there would be no need for the ensemble. Disagreements within the ensemble can indicate erroneous member predictions. The effects of member error can typically be diminished by the accurate predictions of other members. Member variation can also indicate general dissent in the outcome, which may

be caused by accumulated errors or model biases, and corresponds to high uncertainty in the outcome. Such dissent may signal that an outcome should be less trusted or even disregarded, or reveal the need to investigate the individual votes of each member more closely as this may be a more interesting result than conformity.

Ensemble data sets typically simulate complex systems, express numerous interrelationships, and are generally quite large. The multiple models used for simulation can each be perturbed on initial input conditions as well as parameters. This accounts for inaccuracies in measurement of the initial state of the system, for example the impossibility to exactly measure the continuous and chaotic nature of the atmosphere, by running the models using various initial conditions that encompass the possible initial state. Parameter perturbations can indicate the sensitivity of a model to a particular parameter; small changes in parameters that result in large changes in the predictions reveal a high dependence of the model on that parameter. Combining parameter perturbations into an ensemble lessens the influence of that single parameter on the overall outcome.

While the number of individual simulation models may be small, incorporating the various perturbations into the ensemble can greatly increase the size of the data set. An ensemble using only a few base models can quickly grow as each model is run numerous times using a variety of perturbations, the number of which is dependent on aspects of the individual model and may not be consistent across all models. Often, each model predicts for numerous variables, adding to the size of every simulation run. In addition, models are run across various time steps and can even incorporate results from previous runs. Thus, the size and complexity of ensemble data sets can be great. A real-world example of this type of data complexity is discussed further in Chapter 5.

2.2 Probability Density Functions

Another way to think about the data discussed in this dissertation is to look at it simply as a set of samples that predicts a variable. That is, we have some number of values all of which represent a possible outcome of the variable. In this collection, the more a particular outcome occurs, the more likely that outcome is the actual value of the variable. Likewise, one can think of this set of samples as describing the probability that a variable has a particular value. For example, if 95 out of 100 samples have a particular value x, the probability of the variable to be x is 95%. More formally, these data can be thought of as characterising a probability distribution function, or PDF, which describes

the probability that a random variable takes on a particular value.

The most well-known probability distribution function is the Gaussian, or normal, distribution. This distribution describes data that clusters around the mean, or average, and resembles the bell curve. Figure 2.1 shows four variations on the normal distribution by changing the mean, μ, or standard deviation, σ^2. The peak of this distribution indicates that the most expected outcome will be near the mean. The width of the bell curve shows the spread of the data, and wider curves reveal a higher variation in data samples. The normal distribution is often used as a simple model to approximate complex phenomenon and thus, throughout this dissertation, predicted variables are assumed to be normally distributed, unless otherwise stated.

2.3 Descriptive Statistics

One typical method for discussing properties of a data set is through the use of descriptive statistics [7]. These are measures that are used to quantitatively describe features of a data set and include measures of central tendency and dispersion. Descriptive statistics are commonly used to summarize data sets through a small collection of numbers and this approach is used extensively throughout this dissertation.

The most commonly used descriptive statistic is central tendency, which describes the middle, or expected, value of a data set. There are a variety of measures for central tendency, including mean, median, and mode. Given a data set $\{x_i\}_{i=1}^{N}$, the mean is defined by:

$$\mu_1 \simeq \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.1}$$

where N is the number of samples in the data set. Median is the middlemost number in the data set, that is, the data value that separates the upper half of the data from the lower half. The mode is the most frequently occurring value of the data set, and for many distributions, there is assumed to be only one. Each of these measures describe a single value that is often used as an estimate of the value of the entire data set.

Variance and standard deviation are statistics used to describe how much the data deviates from the estimation of the mean. They are calculated by finding the difference between every point in the data set and the mean value. Thus, a data set with most of

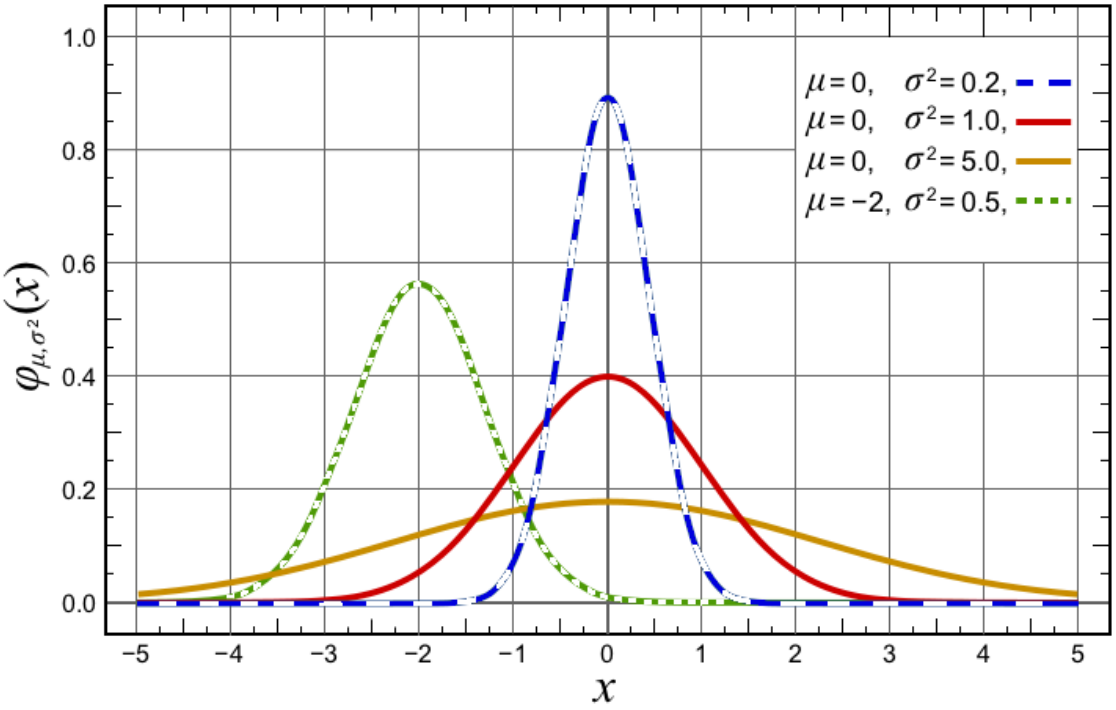


Figure 2.1. The Gaussian, or normal, distribution. Each curve shows the change in the distribution when mean or variance is modified.

its values located near the mean will have a smaller variance than one in which the data values are very spread out. Formally, variance is defined by:

$$\mu_2 \simeq \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_1)^2 \tag{2.2}$$

and standard deviation as:

$$\sigma = \sqrt{\mu_2} \tag{2.3}$$

While mean and standard deviation are the most common descriptive statistics, other descriptors exist. Many of these can be categorized as *central moments*, which are defined as:

$$\mu_k \simeq \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_1)^k \tag{2.4}$$

where k is the number of the desired moment. The first and second moments are mean and variance. The next moment, skew, is the amount of asymmetry in the data and can also be defined in terms of standard deviation:

$$\gamma = \frac{\mu_3}{\sigma^3} \tag{2.5}$$

Kurtosis is the fourth central moment and describes the "peakedness" of the distribution. Data sets with high kurtosis have sharp peaks and fatter tails while low kurtosis is associated with rounded peaks and thinner tails. Kurtosis is defined as:

$$\kappa = \frac{\mu_4}{\sigma^4} \tag{2.6}$$

Excess kurtosis is the standard kurtosis measure normalized by the kurtosis of a Gaussian:

$$\kappa_e = \kappa - 3 \tag{2.7}$$

Data sets with large, negative excess kurtosis (i.e., $\kappa_e < 0$) are called *platykurtic*. Kurtosis values very close to zero are associated with data sets resembling a Gaussian distribution and are called *mesokurtic*. Highly peaked, or *leptokurtic* data exhibit strongly positive kurtosis values (i.e., $\kappa_e > 0$).

2.4 Uncertainty

The National Institute of Standards and Technology (NIST) [77] defines the uncertainty of a measured result to be the standard deviation of the collection of data samples approximating the measurand. A measurand is a particular quantity to be measured, such as the temperature reading at a certain location and time of day. This measurand is estimated by a series of measurements; the true value of the measurand is expected

to lie within the range of these measurements. Uncertainty can then be expressed as the standard deviation of the set of measurements. This uncertainty value can then be used to describe a level of confidence that the true value of the measurand does lie, in fact, within an indicated range. Uncertainty information must accompany a set of data samples for the set to be considered a scientifically complete result. Thus, uncertainty is becoming a more important value as the quantity of generated data grows with the increase in the sophistication of scientific simulation and measurement devices.

CHAPTER 3

VISUALIZATION OF SUMMARY STATISTICS

An important visualization research problem is to effectively convey uncertainty information along with traditional visual data representations. This chapter investigates the problem from a graphical data analysis standpoint. By using descriptive statistics to summarize both characteristic features of a data distribution and measures of uncertainty, a cohesive understanding of the information can be achieved. In this chapter, the boxplot, a convenient and well-known method for graphically depicting summary statistics, is reexamined. A new hybrid summary plot that combines moment, cumulant, and density information along with higher order descriptors that rely on distribution fitting is developed. The summary plot is extended to a two-dimensional (2D) comparative plot for correlated data sets. In view of the important role summarizing plots have in decision making, this work focuses on using advanced visualization techniques to incorporate additional descriptive parameters, leading to a stronger understanding of the data.

3.1 Introduction

The goal of visualization is to effectively present large amounts of information in a comprehensible manner, however, most visualizations lack indications of *uncertainty* [46, 45]. Uncertainty is a term used to express descriptive characteristics of data such as variation, error, and level of confidence, and is often statistically described by the standard deviation. Because this additional information is crucial in understanding the reliability of information and thus in decision making, its absence can lead to misrepresentations and incorrect conclusions. However, as the importance of visualizing these data sets grows, the actual task of visualizing them becomes more complicated, and incorporating the additional data parameter of uncertainty into the visualizations becomes even less straightforward.

Tukey [80] proposed graphical techniques to summarize and convey interesting char-

acteristics of a data set not only to facilitate an understanding of the given data but also to further investigation and hypothesis testing. These tested graphical methods, such as the boxplot, histogram, and scatter plot, provide identifiable representations of a data distribution, and their simplicity allows for quick recognition of important features and comparison of data sets. In addition, they can be substituted for the actual display of data, specifically when data sets are too large to plot efficiently.

This work takes inspiration from the visual devices used in exploratory data analysis and extends their application to uncertainty visualization. The statistical measures often used to describe uncertainty are similar to measures conveyed in graphical devices such as the histogram and boxplot. This research investigates the creation of the *summary plot*, which combines the boxplot, histogram, a plot of the central moments (mean, standard deviation, etc.), and distribution fitting. The boxplot has a canonical feel; the "signature" of the plot is easily recognizable and does not need much explanation to allow for a full understanding. The goal of this work is to create a summary plot that similarly incorporates higher order information, allowing for the quick identification of characteristic features. This higher order signature provides at-a-glance recognition of variations from normal and allows easy comparison of data distributions in detail. In addition, a 2D extension of the summary plot is presented that provides for the comparison of correlated data and an exemplary application of the method is presented to demonstrate how these techniques can be applied on large scale data sets.

3.2 Related Work

Understanding data sets is an essential part of the scientific process. However, discerning the significance of data by looking only at numerical values is a formidable task. Descriptive statistics are a quick and concise way to extract the important characteristics of a data set by summarizing the distribution through a small set of parameters. Median, mode, mean, variance, and quantiles are typically used for this purpose. The main goal of descriptive statistics is to quickly describe the characteristics of the underlying distribution of a data set through a simplified set of values. Often, these parameters provide insights into the data that would otherwise be hidden. In addition, data summaries facilitate the presentation of large scale data and comparison of multiple data sets.

Creating graphics for data presentation is a difficult task involving decisions not only about data display but also about data interpretation. Often, the graphic is intended to

show specific characteristics of the data, and the presentation style should make these characteristics clear. Poor presentation style can distract the viewer or even lead to erroneous conclusions. To alleviate this problem, design practices for effective data visualization are outlined in numerous sources [20, 24, 79, 82]. These references not only guide the scientist towards the graphical technique of choice for specific data types, but also describe how a visualization may be interpreted by the viewer and suggest methodologies to influence this interpretation.

3.2.1 Statistical Plotting Techniques

The display of statistical information is ubiquitous in all fields of visualization. Whether aided by graphs, tables, plots, or integrated into the visualizations themselves, understanding the best way to convey statistical information is important. Highlighting the boxplot, the following section presents a survey of traditional methods for expressing specific statistical characteristics of data. Reviewing techniques for the expression of statistical measures will be increasingly important as data sets become larger and data quality, confidence, and uncertainty become influential characteristics to integrate into visualizations.

Methods for visually presenting summary statistics include tables, charts, and graphical plots. Graphical plots pictorially convey a large amount of information in a concise way that allows for quick interpretation and understanding of the data. This section will focus on one of the most well-known techniques for summarizing data, the boxplot. In addition to various ways to construct the standard boxplot, modifications that increase the amount of information presented in the plot will be discussed as well as extensions into higher dimensions.

3.2.1.1 The Boxplot

One of the most ubiquitous approaches to graphing summary statistics is the boxplot [41, 75, 80], which is the standard technique for presenting the *five-number summary*, consisting of the minimum and maximum range values, the upper and lower quartiles, and the median, as illustrated in Figure 3.1. This collection of values quickly summarizes the distribution of a data set, including range and expected value, and provides a straightforward way to compare data sets. In addition, the reduced representation afforded by the five-number summary provides a concise tool for data analysis, since only these characteristic values need to be analyzed. Surveys on the introduction and evolution

of the boxplot can be found in [22, 66].

The typical construction of the boxplot, which can be seen in Figure 3.1, left, partitions a data distribution into quartiles; that is, four subsets with equal size. A box is used to indicate the positions of the upper and lower quartiles; the interior of this box indicates the *innerquartile range*, which is the area between the upper and lower quartiles and consists of 50% of the distribution. The box is intersected by a crossbar drawn at the median of the data set. Lines (sometimes referred to as whiskers) are extended to the extrema of the distribution, either minimum and maximum values in the data set, or to a multiple, such as 1.5, of the innerquartile range [34] to remove extreme outliers. Often, outliers are represented individually by symbols; this type of plot is sometimes referred to as a schematic plot [80]. The width and fill of the box, the indication of outliers, and the extent of the range-line are all arbitrary choices depending on how the plot is to be used and the data it is representing. Figure 3.1 (a-d), shows various visual modifications on the boxplot.

3.2.1.1.1 Origins. The origins of the boxplot can be traced to the range-bar chart. Haemer [41] suggested the use of range-bar charts not only for the comparison of ranges of data, but also for expressing central measures such as median, mean, mode, standard deviation, and tolerance limits through annotations on the chart. This idea was extended to displaying the five-number summary on the range-bar chart [75], as seen in Figure 3.1(a), by shortening the bar to encompass only the central 50% of the data, using a thin line to indicate the entire range, and a perpendicular line to show the median. This is the first appearance of the form of the boxplot we know today. Introduced in 1977, the Tukey boxplots [80] became a popular representation, and is shown in Figure 3.1(b). This plot truncated the length of the range-line to 1.5 times the length of the innerquartile range. Outliers are indicated by independently marking them on the plot. The look of Tukeys boxplot is also refined from that of the range-bar chart. The box fill is removed and the end of the range-line is clearly marked.

The visual refinement of the boxplot continued with the introduction of the quartile plot [79], shown in Figure 3.1(c), which sought to reduce visual clutter and maximize the ink-to-paper ratio by removing the box completely, and indicating the innerquartile range by an offset line. The median is simply a break in the innerquartile line. Other versions of this plot indicate the median using a small square and remove the innerquartile line, letting the empty space between the two range-line segments represent the central

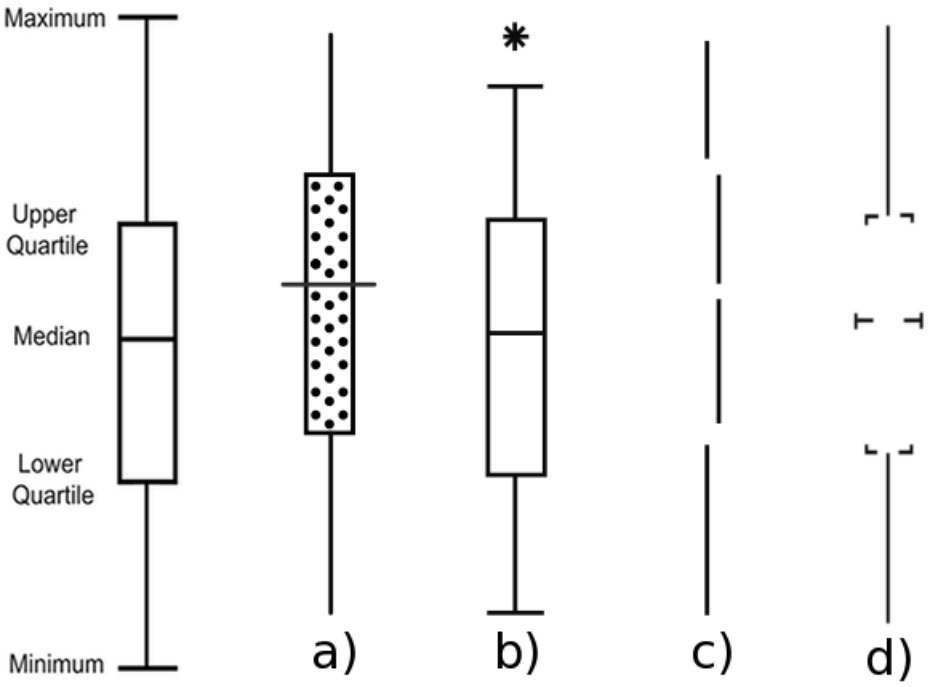


Figure 3.1. The boxplot and visual modifications. (Left) The construction of the boxplot. (Right) Various visual modifications: a) Range plot [75]. b) Boxplot [80]. c) Innerquartile plot [79]. d) Abbreviated boxplot [68].

quartiles. While these plots do reduce the amount of ink used to indicate the five-number summary, they may also reduce the ease in interpretation of the plot due to the subtle way that the median is indicated, and the similar technique used to show both the range of the data, and the innerquartile range. Furthermore, reducing the amount of area taken up by the innerquartile representation is counter-intuitive since this region contains the majority of the data, a fact that the plot should clearly express. The abbreviated boxplot [68], as seen in Figure 3.1(d) and more formally covered later in this chapter, is another approach that reduces the ink needed to convey the five-number summary, specifically for the purpose of superimposing further summary statistics on top of the plot. This method maintains the original form of the boxplot, but removes the sides of the box, leaving only the corners indicative of the box.

3.2.1.2 Modifications to the Boxplot

One of the major advantages of the boxplot is its simplicity of design. Critical information about a data set is quickly expressed, and the box itself is a signature of

the distribution. General characteristics such as the symmetry of the distribution, the location of the central value, and the spread of the observations are immediately apparent. This concise representation allows for the inclusion of additional information about the data set, and permits the user to customize the plot for specific purposes.

3.2.1.2.2 Density information. One of the most common types of information added to the boxplot is a description of the distribution of the data values. The boxplot summarizes the distribution using only five values, but this overview may hide important characteristics. For instance, the modality (or number of most often occurring data values) of a distribution is not expressed in the boxplot and similarly looking plots may encode distinctive distributions with varying modality. This is especially problematic when no prior information is known about the distributions, as comparing distributions with differing modalities or other statistical properties may not be appropriate. One solution to these types of problems is to add into the boxplot indications of the density of the underlying distribution.

The histplot [10], shown in Figure 3.2(a), is a simple approach for adding density information to a boxplot. In the histplot, the density of the distribution is estimated at the median and the two quartiles. The width of the boxplot at these locations is then modified to be proportional to the density estimation, and lines are drawn to connect these widths, essentially changing the box of the boxplot into a polygon. The histplot adds a quick summary of the density of the central area of the distribution, but it is still possible for important features to be missed. The vaseplot [10], shown in Figure 3.2(b), is a refined version of the histplot that adds in estimated densities for every point between the upper and lower quartiles. A line is drawn between each density estimation point (on both sides), and the polygon of the histplot is replaced with something that, depending on the distribution, resembles a vase. This modification explicitly shows the density of the central 50% of the data. In addition, confidence intervals can be added to both of these plots by superimposing a light gray shaded bar over the median with the height of the bar signifying confidence.

The box-percentile plot [33], shown in Figure 3.2(c), is another method for adding the empirical cumulative distribution of the data set into the boxplot. In this type of plot, both sides of the boxplot are used to plot the percentile of the distribution at each point. Thus, for each position in the plot, the width of the box is proportional to the percentile of that data value, up to the 50th percentile, at which point the width is switched to

being proportional to minus the percentile. The sides of the plot are symmetric and the 25th, median, and 75th percentiles are marked with a line. The advantages of this plot are that there is no question as to how it should be drawn, it covers the entire range of data, and it does not use any arbitrary choices for its creation. Additionally, the plot is straightforward enough to be understandable by untrained readers, but includes details for trained readers.

The violin plot [42], shown in Figure 3.2(d), combines the standard boxplot with a density trace [20] to exploit the information contained in both types of diagrams. The boxplot is used to show the innerquartile range, however, it is modified in two ways. The first modification changes the boxplot by making the box solid black and replacing the median line with a circle; this allows for quick identification of the median and easy comparisons. The second modification removes the individual symbols for outlying data values since the outliers are contained in the density trace and individual points would clutter the diagram. A density trace is added as an alternative density estimator to the histogram and gives a smoother indication of frequency by allowing the intervals in which density is calculated to overlap, in contrast to the histogram. The density trace $f(y)$, at value y, is defined by:

$$f(y) = \frac{1}{hn}\sum_{i=1}^{n} W(\frac{y - y_i}{h}), \qquad \text{where} \qquad W(u) = \begin{cases} 1 & \text{if } \mid u \mid \leq \frac{1}{2}, \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where n is the sample size, and h is the interval width. The trace is added to the violin plot as two symmetric curves on either side of the boxplot, making the density and magnitude easy to see. The main factor that controls the look of the density trace is the size of the interval width, h. There is no specific size that works best in every situation, but an h value around 15% of the data range often produces good results; and the h values should stay between 10 and 40% of the data range to maintain a pleasing smoothness of the density trace curve.

The sectioned density plot [25], Figure 3.3, exploits characteristics of the human visual system to present, in implied 3D, shape information of a data distribution and trends in variance and central tendency. The human visual system is capable of using occlusion and intensity variation as cues to spatial depth. The sectioned density plot uses these cues to display the distribution of a data set in order to create the illusion of 3D. To create a sectioned density plot, the data are partitioned into fixed-width intervals, the number

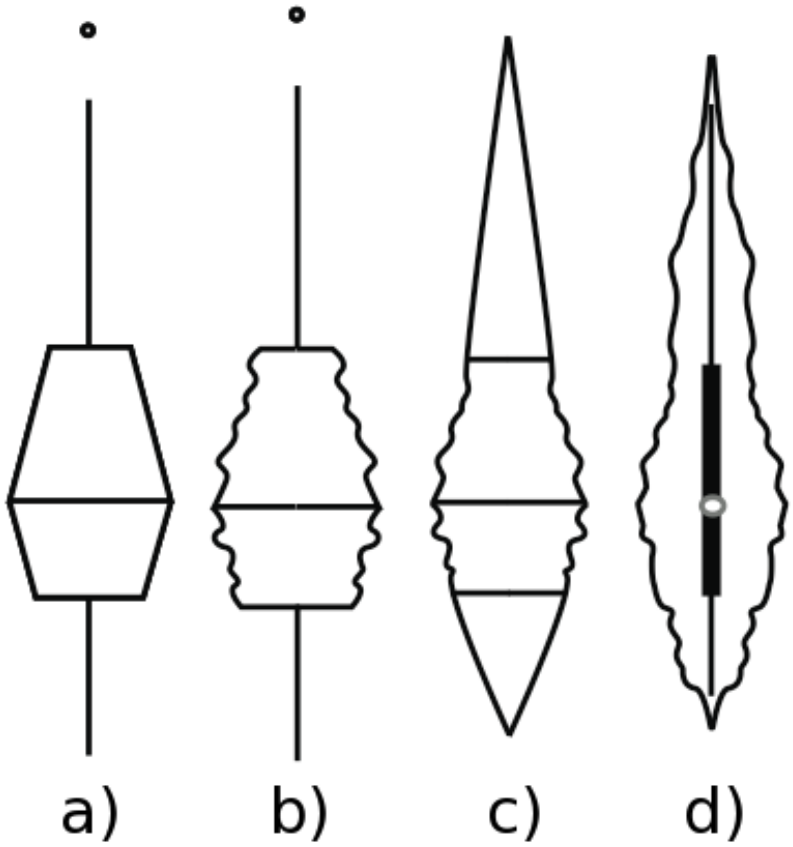


Figure 3.2. Density modifications to the boxplot. (a) Histplot [10] b) Vaseplot [10] c) Box-percentile plot [33]. d) Violin plot [42].

of which is variable. Each of these intervals is plotted onto a black background. From lowest density to highest, each interval is plotted using a rectangle shifted slightly to the left, occluding the previous interval, and filled with a monotonically increasing intensity. The five-number summary is incorporated into these plots by using the coordinate axis to show the range of the values, indicating the upper and lower quartiles with thin rectangles superimposed on the axis, and the median as a break in the range line. Each of these values is extended through the graph as a thin white line.

3.2.1.2.3 Additional descriptive statistics. Often there are instances in which the five-number summary is not enough information, however, adding a density plot is not feasible or necessary. For instance, when doing a comparison of multiple data sets, adding the density distribution of each data set may clutter the plot, however, it would be useful to have information such as the relative number of observations. Additional

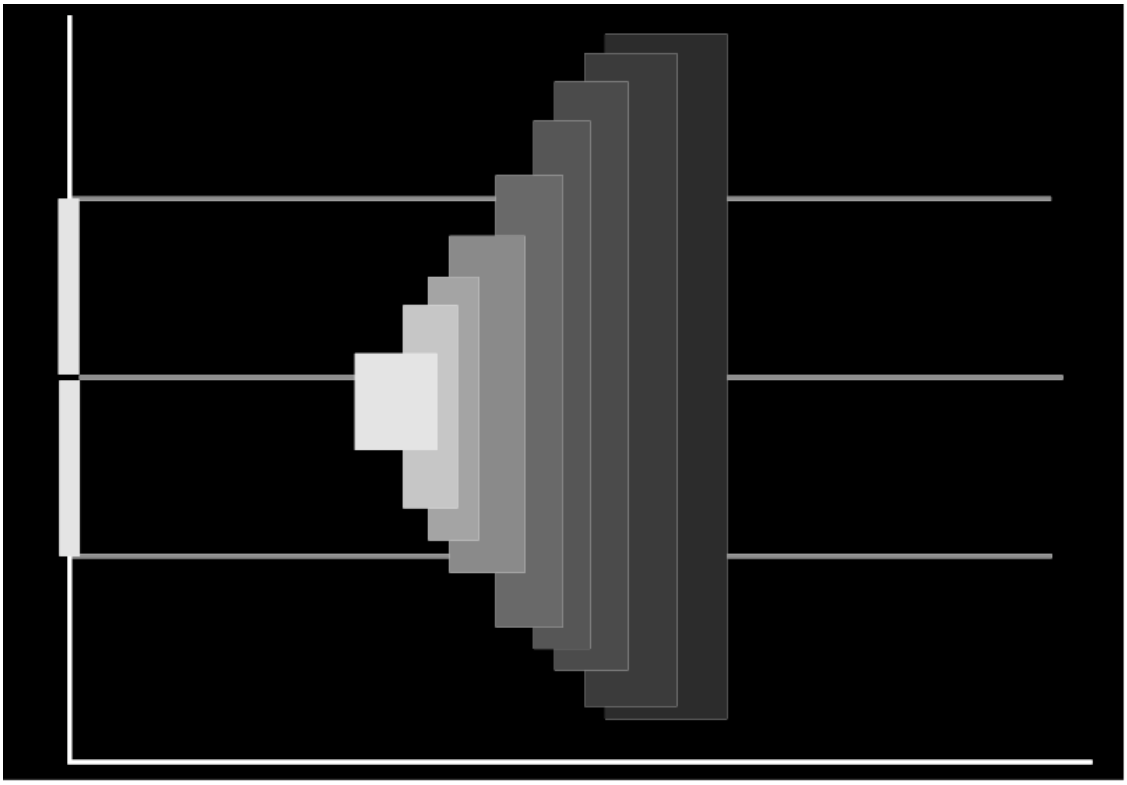
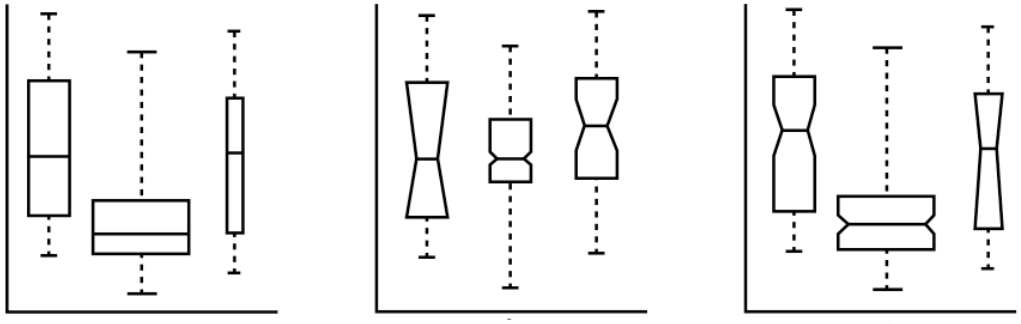


Figure 3.3. The sectioned density plot. This plot indicates density information through the size and darkness of the central rectangles. The five-number summary is shown on the coordinate axis.

information may also reduce the possibility of the user making false conclusions.

McGill et al. [60] suggested a few minor modifications of the original boxplot to address these issues. The first variant is the variable width boxplot, which can be seen in Figure 3.4(a). This plot uses the width of the box to proportionally encode the size of the data set. The addition of this size clue easily alerts the viewer to distinctions in the number of observations in each data set and can help the viewer avoid misinterpretation. The second variant proposed is the notched boxplot, as shown in Figure 3.4(b). In this plot, notches are added to the box to roughly indicate the significance of differences between values or the confidence level of the data. The last proposed plot is the variable width notched plot that combines the information contained in the previous two plots and can be seen in Figure 3.4(c).

One of the drawbacks of the simplicity of the boxplot is that the plot can hide distinguishing features of a distribution, and possibly encode very different distributions



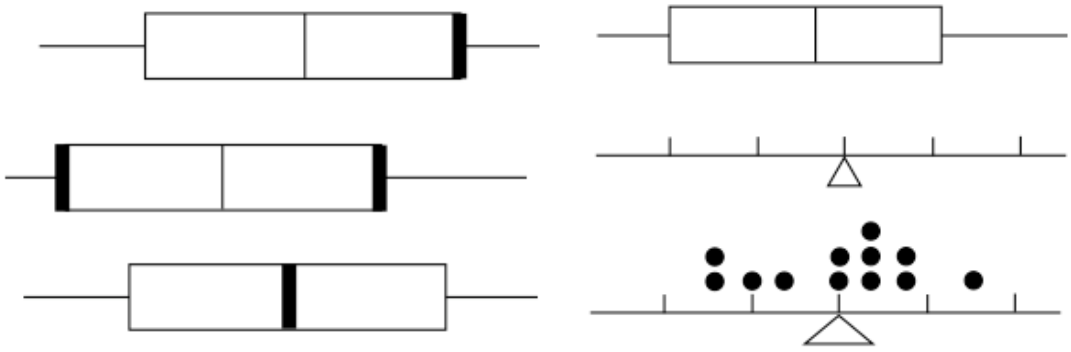
(a) Variable width boxplot. (b) Notched boxplot. (c) Variable width notched boxplot.

Figure 3.4. Data attribute modifications of the boxplot. (a) The variable width boxplot encodes differences in sample sizes by modifying the width of each individual plot. (b) The notched boxplot indicates confidence levels or other significant areas of the data. (c) The variable width notched boxplot combines plots (a) and (b).

using similar plots. The addition of density information attempts at solving this problem, but it is not always feasible to add in this (possibly) large amount of data. An alternative to using density information is to use statistics that describe specific characteristics of a distribution.

An example of adding descriptive statistics to the boxplot is the addition of skew and kurtosis measures. Skew and kurtosis are statistics that describe the symmetry and peakiness of the distribution and can indicate modality. One method for adding these measures into a boxplot thickens the sides of the box when these measures indicate skew or high kurtosis in a specific direction [22], Figure 3.5(a). The topmost plot indicates that the distribution is skewed toward the right. A bimodal distribution would be skewed in both directions, and this is shown in the center plot in which both ends of the boxplot are thickened. Finally, a distribution that is centrally peaked has the median line of the boxplot thickened, as shown in the bottom plot. This technique quickly conveys an indication of these statistics and can be used to distinguish between differing distributions.

The beam and fulcrum display [31] is a complementary diagram to the boxplot and this combination can be seen as the two diagrams at the top of Figure 3.5(b). In this type of display, the range is represented as a line (or beam) and the fulcrum, represented as a triangle, is placed at the mean. On each side of the fulcrum, tick marks are used to show standard deviation points. As seen in Figure 3.5(b), bottom, a dot plot can be added to



(a) Skew, modality, and kurtosis.

(b) A beam and fulcrum diagram.

Figure 3.5. Descriptive statistical plots. (a) Boxplots with varying skew and kurtosis. From top to bottom: right-skewed, bimodal, and centrally peaked distributions. (b) Beam and fulcrum diagram. A dot plot is added to the bottom gure to indicate frequency and the size of the fulcrum base shows the width of a condence interval.

the beam and fulcrum display to show the frequency of data values, and the size of the fulcrum base can be modified to express the width of a confidence interval. The benefits of such a diagram when presented alongside a boxplot are that the user is able to quickly pick out non-normal distributions (i.e., when the mean and median are not equal), see where the data are distributed with respect to the standard deviation scale ($\pm\sigma$, $\pm 2\sigma$, ...), and easily find outliers (i.e., data points outside three standard deviations). It is also a useful learning tool, in as much as students can easily understand that the mean balances the distribution.

3.2.1.3 Bivariate Extensions

Standard implementations of the boxplot focus on univariate data distributions. The five-number summary is a useful descriptor of not only univariate, but also bivariate data distributions. The main challenge in extending the boxplot for use with higher dimensional data is how to translate the five-number summary values, which are vector values in the bivariate case, into visual metaphors with meaningful spatial positions, while maintaining the simplicity of the original boxplot. A bivariate boxplot can show not only the location and a summary of the data distribution, but also skew, spread and correlation.

A rangefinder boxplot [9], as seen as the solid back lines in Figure 3.6(a), is a simple extension of the boxplot into 2D. To create a rangefinder boxplot, all data values are

plotted as points on a 2D graph (this is often called a scatter plot). For each variable, the five-number summary is calculated, a line segment is drawn along the innerquartile range and perpendicular lines are placed at the adjacent values of the variable, where the 1D boxplot would terminate. The intersection of the two central line segments is the cross-median value. This idea was further improved upon, as shown as the thick gray lines in Figure 3.6(a), to emphasize the quartiles rather than the range, by moving the perpendicular lines from the adjacent values to the upper and lower quartile positions and extending whisker lines to the extrema value of the variable [53]. These extensions of the boxplot into 2D are an unobtrusive expression of the summary of each variable, but the correlation between the two variables is not visible.

Other techniques for extending the boxplot into 2D all use the notion of a hinge that encompasses 50% of the data and a fence that separates the central data from potential outliers. The distinctions between each of these methods are the way the contour of the hinge and fence are represented, and the methods used to calculate the contours.

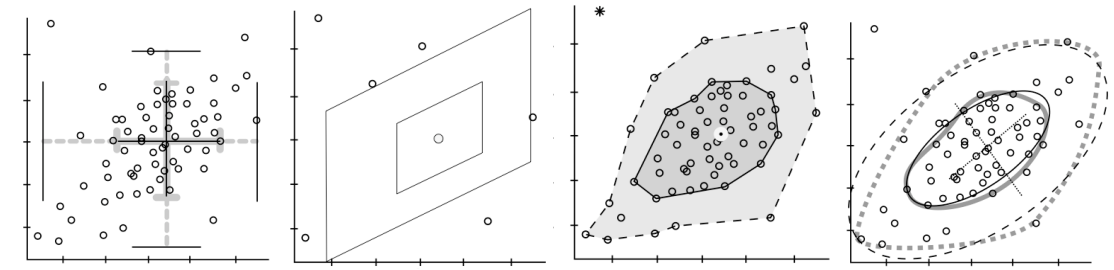
The 2D boxplot [78], as seen in Figure 3.6(b), computes a robust line through the data by dividing the data into three partitions, finding the median value of the two outer partitions, and using these points as the line. Depending on the relationship between the slope of the line and each variable, the quartile and fence lines are drawn either parallel to the robust line, or parallel to the variables coordinate axis. The lines not comprising the outer-fence and the inner-hinge boxes are removed.

The bagplot [72] uses the concept of halfspace depth to construct a bivariate version of the boxplot, as seen in Figure 3.6(c). The halfspace depth $ldepth(\theta|Z)$ of some point θ is the smallest number of data points $z_i \in Z = z_1, z_2, ..., z_n$ contained in any closed halfplane with a boundary line through θ. The depth region D_k, which is a convex polygon, is the set of all θ with $ldepth(\theta|Z) > k$ and $D_{k+1} \subset D_k$. To construct the bagplot, a scatter plot of the data is first created. The depth median is then found, which is the θ with the highest $ldepth(\theta|Z)$, if there is only one such θ, otherwise, it is the center of gravity of the deepest region. This point, which is at the center of the plot, is represented as a cross. The bag is a dark gray region in the plot encompassing 50% of the data. The fence separates the outliers of the data set, but is not drawn, and the loop is a light gray region of the plot that contains points outside of the bag, but inside of the fence. Outliers are highlighted as black stars. Options for reducing the visual clutter of the bagplot are to not plot data points contained in the bag, and to not fill the regions contained in the bag

and the loop, but instead surround the bag with a solid line and the loop with a dashed line. In addition, a confidence region can be added into the bagplots as a blotch drawn around the depth median.

The relplot and the quelplot [38] use concentric ellipses to delineate between the hinge and fence regions. Both the relplot and quelplot can be seen in Figure 3.6(d). The relplot uses full ellipses that assume symmetric data and are constructed using a robust estimator such as the minimum volume ellipsoid. In the figure, the relplot is shown as ellipses drawn in thin black lines. The quelplot divides the ellipses into four quarters aligned on the major and minor axes, and computed using an M-estimator. The quelplot is shown in the figure as thick, gray lines. The quelplot can show skewed data, since each quarter ellipse can be transformed individually.

The boxplot is a standard technique for presenting a summary of the distribution of a data set. Its use has become prevalent in all forms of scientific inquiry, and understanding its construction, origins, and modifications can help not only with interpretation of the information presented by the boxplot, but also in its creation and use. The concise representation provides not only insights to the important characteristics of a distribution, but permits the addition of information that enables the customization of the boxplot to specific scenarios. Overall, the simplicity of the boxplot makes it an elegant method for the presentation of scientific data.



(a) Rangefinder boxplot. (b) 2D Boxplot (c) Bagplot (d) Quel and Relplots.

Figure 3.6. Bivariate extensions of the boxplot. (a) The rangefinder boxplot [9]. (b) The 2D boxplot [78]. (c) The bagplot [72]. (d) The quel- and relplots [38]

3.3 The Summary Plot

The hybrid boxplot that this work introduces can be more formally titled the *summary plot*. This display includes not only the quartile information present in the form of a modified boxplot, but also a collection of descriptive statistics and density information. As shown in Figure 3.7, we use an abbreviated form of the traditional boxplot to convey the five-number summary and a symmetrically drawn histogram to convey density information. While this technique is similar to that of the violin plot [42], we have extended it to include minimum and maximum, rather than truncating extreme values, and incorporated a colormapped histogram to further aid the understanding. We express descriptive statistics in the form of mean, standard deviation, and higher-order moments as glyphs, with the design of each reflecting the semantic meaning of the statistic. Finally, we add distribution fitting capabilities to allow the user to find a best fit from a library of distributions and to compare against well-known distributions.

The main goal of the summary plot is to create a data distribution signature, providing for fast recognition of interesting properties. The higher order glyphs clearly display deviations from a normal distribution and are easily compared. Because the statistical meanings of the moments are more complicated than the meaning of the boxplot, there will be a learning curve associated with understanding the additional information. However, because of the simplicity of the technique, a user who has learned it will easily recognize desired characteristics. The summary plot also reduces the amount of information needed to convey the data distribution - a reduction that is often desirable when the amount of information is too large to easily display let alone understand.

3.3.1 The Abbreviated Boxplot

As discussed above, the traditional approach to presenting summary information is through the boxplot, which has been refined numerous times in efforts to maximize the ratio of information to ink consumption and improve aesthetics. We have chosen to refine the plot further, as shown in Figure 3.7 (leftmost). Our plot builds on Tukey's boxplot [80] (Figure 3.1(b)) with a few important distinctions. The first modification removes the sides of the innerquartile box. This not only reduces the visual clutter of the plot, but also removes possible assumptions about the density of the distribution. The prevalence of using the sides of the plot to indicate density is due to the visual metaphor created by the box itself. Since 50% of the data samples lie within the box, it is easy to assume a density

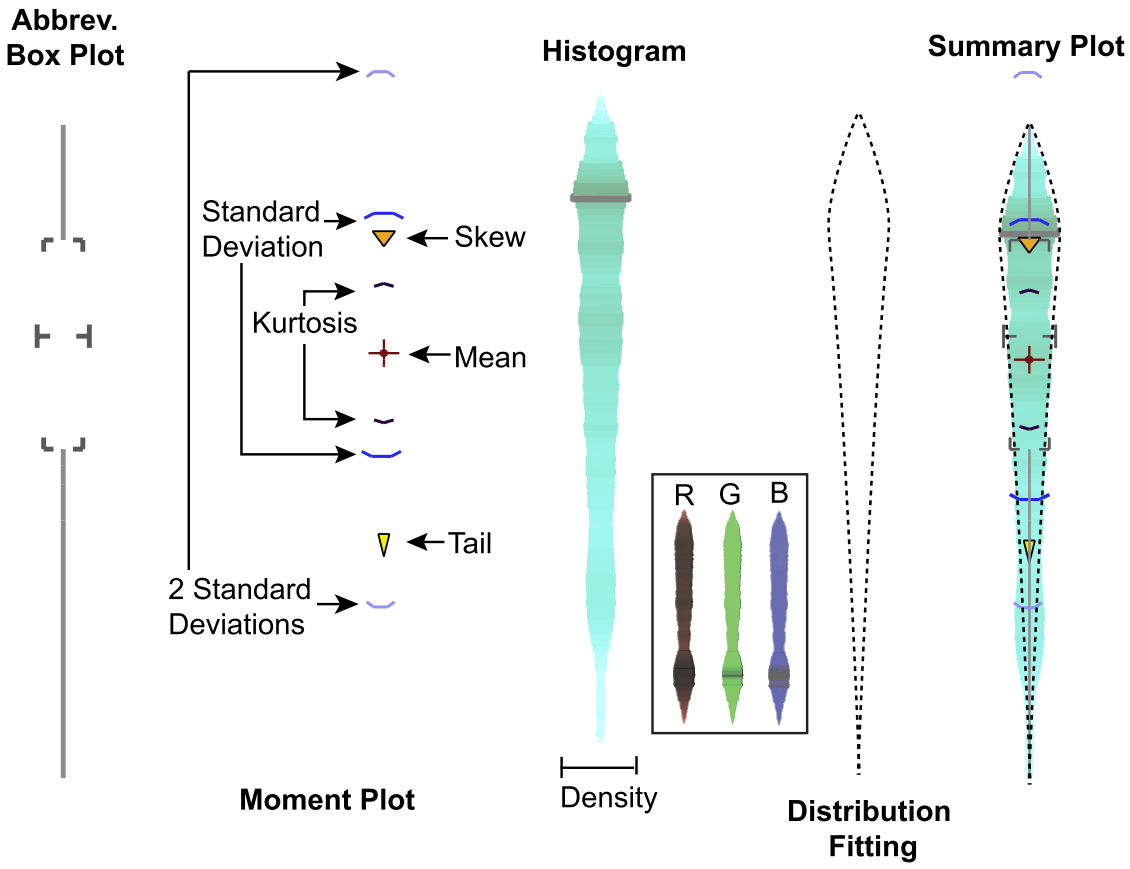


Figure 3.7. Anatomy of the summary plot. The abbreviated boxplot displays the range of the data distribution. The moment plot shows higher order moments that describe feature characteristics. The histogram estimates the density of the distribution and is displayed using a symmetric display and a redundant colormap. Distribution fitting allows the user to compare the data against well-known distributions.

distribution that resembles the plot itself, with the highest densities falling close to the median. However, this restricts the plot to normal or Gaussian-like distributions, which is not always the case. Often, the mode (or most frequently occurring sample value) lies outside of the innerquartile range, which is only evident when the boxplot is combined with a density display. The next modification extends the median lines outward in order to emphasize the position of the median and ensure this position does not get lost with the addition of more information. Finally, outliers are not removed from the plot. While this choice may at times stretch out the plot to extreme values, we prefer to express the entire range of the data set.

3.3.2 Quartiles and the Histogram

One of the simplest ways to describe a data distribution is to calculate the *quartiles* of the data set. Recall that a quartile partitions the ordered data into four equally sized subsets such that 25% of the data are less than the lowest quartile, 50% of the data are less than the next quartile (i.e. the median), and 75% of the data are less than the highest quartile. There are conflicting conventions concerning whether the term "quartile" refers to the specific data value that cuts off the partition or to the subset. In this work, the former definition is adopted in order to be able to place the quartile values spatially. Figure 3.8 demonstrates this distinction, plotting a single data distribution as density and cumulative histograms. At the top of the figure, a histogram is displayed; the height at each point reflects the density of the distribution at that data value. Below is a cumulant histogram in which density is successively added. Reference lines illustrate the quartile partitioning. At the bottom is a violin plot [42] displaying both the distribution density and cumulant information, via the quartiles, in a compact form.

The calculation of the cumulant quartile is based on the histogram. The histogram employs a user-specified number of bins to sort the data based on value, giving a rough estimate of the density of the data distribution. From the histogram, the quartile positions are found using a straightforward counting algorithm. The position of the quartile value is determined by dividing the number of data points by the desired quartile position and counting the sorted data in the histogram until the quartile position is reached.

Density information is added to the summary plot as a histogram, which is displayed using quadrilaterals whose widths are varied with the density at each bin location. The colormap used for the histogram was designed to be both redundant and nonintrusive.

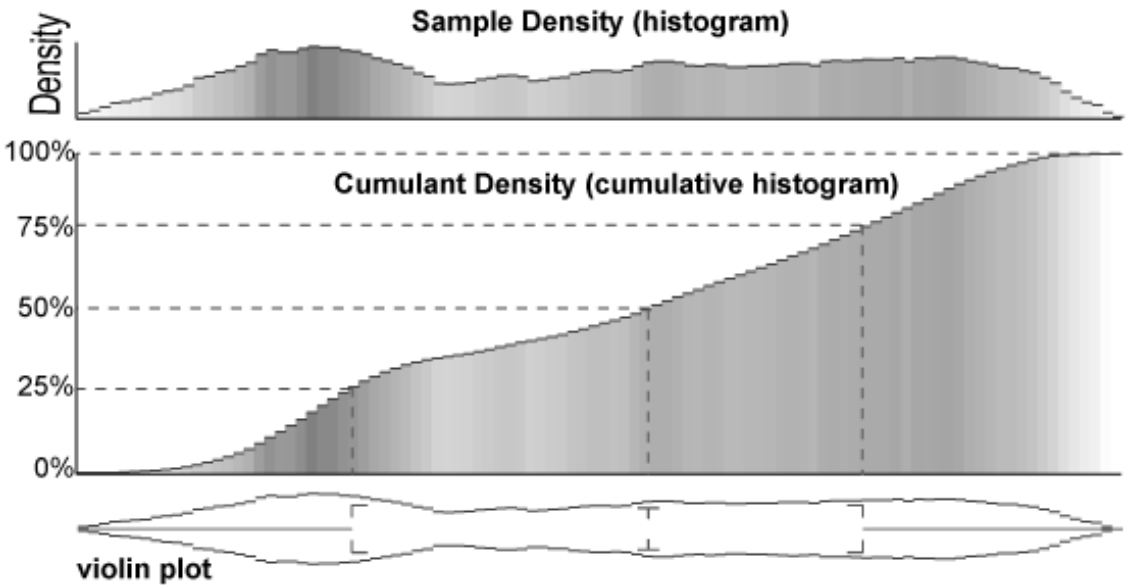


Figure 3.8. Cumulant and regular histograms. The regular histogram (top) shows the density of the distribution at each data point, and the cumulant histogram (middle) shows the additive density of the distribution up to that data point. The violin plot [42] (bottom) is a symmetric display of the sample density combined with a boxplot.

Each color channel uses a distinct mapping that when combined, clearly emphasizes areas of high density without overpowering the plot with color. The color is defined as follows: the red color channel is normalized log density, the green color channel is normalized density, and the blue color channel is normalized linear density. Each color channel can be seen in Figure 3.7 (inset). The distinction between the maps for the individual channels is subtle and intended to encode the density information in a manner that is reiterative, aesthetically pleasing, and subdued so as to act as a backdrop for the more saturated color scheme used for other glyphs.

3.3.3 Moments

The moments of a distribution are statistical measures of feature characteristics, the most well known being mean and standard deviation. The main distinction between the summaries presented by the boxplot and the moment plot is that the quartiles give information about the location and variation changes in the data, while moments express descriptive characteristics of the look of the distribution such as "peakiness." One of the drawbacks of using only a boxplot to summarize a distribution is that multiple, distinct

distributions can have the same boxplot signature. For example, one may come across two distributions, one unimodal (having one data value occurring most frequently) and the other multimodal (multiple most frequent values), having identical quartiles and thus whose boxplot signatures are the same. Adding moment information exposes differences between distributions and allows for the expression of non-Gaussian distributions, while maintaining the simplistic characteristic of the simpler boxplot.

The creation of the moment plot was inspired by the beam and fulcrum display in Figure 3.5(b) and the use of moments in physics. As seen in Figure 3.9, the mean can be thought of as a fulcrum under a beam, and the moments as weights used to balance the beam, each moment having a specific role in dynamically balancing the system. While this approach is not meant to be a physically-based explanation of moments, those unfamiliar with the role of moments in statistics may find this abstraction helpful.

3.3.3.1 Mean and Standard Deviation

The most familiar and frequently used moments are mean and variance (the first and second moments). The average of the data samples is an estimator of the mean of the underlying distribution, or the expected value of a random variable. Variance is a measure of the dispersion of the data indicating the distance a random variable is likely to fall from the expected value. Standard deviation is simply the square root of variance. For the summary plots, we use only mean and standard deviation, as standard deviation is derived from variance.

The addition of mean and standard deviation to the summary plot is very straightforward. The mean is rendered as a dark red cross. The width of the lines making up the cross are constructed so that when the mean and median are displayed at the same location, the glyphs coincide and form a straight line across the plot. This emphasizes symmetrical distributions and quickly reveals when a distribution varies from normal. A close up of this can be seen in Figure 3.10.

Standard deviation, like all even moments, is rendered as two glyphs on the plot. Blue curved lines are placed on either side of the mean to express the average variation from the mean. The glyphs are placed at mean $\pm$ standard deviation and mean $\pm$ 2$\times$standard deviation. This allows the user to easily see where the majority of the data lies, as well as to identify samples outside two standard deviations, typically referred to as extrema.

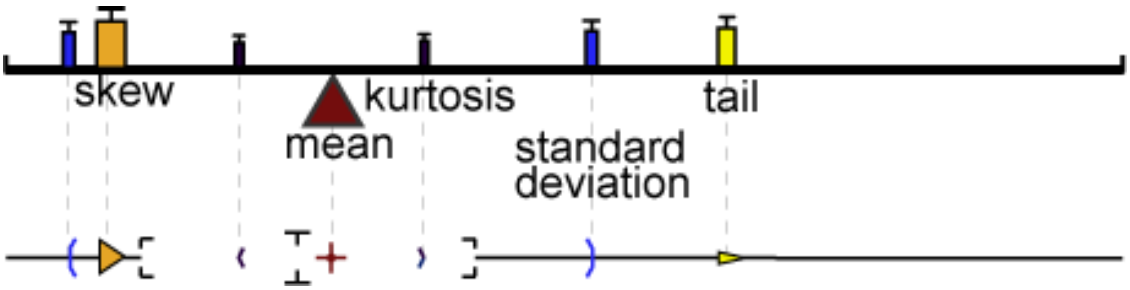


Figure 3.9. Moment arm abstraction from which we designed the moment glyphs.

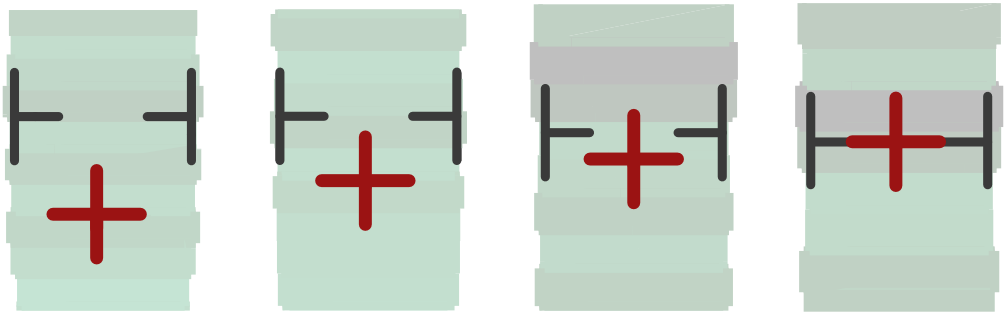


Figure 3.10. Mean and median glyphs align when their values are equal.

3.3.3.2 Skew

Skew is a measure of the asymmetry of a distribution, or the extent to which the data are pushed to one side or the other. Figure 3.11(a) shows three distributions in which skew varies from negative to positive. Based on the balance beam abstraction (see Figure 3.9), a large triangle is used to denote skew in the summary plot and is placed so that it rests on the side of the distribution with the highest density, pointing at the tail. Mathematically, the placement of the skew glyph is calculated by first finding skew (γ) as defined in Equation 2.5 and placing the glyph $-\gamma$ distance away from the mean, with the apex of the triangle pointing toward the tail of the distribution. Thus, the placement of the skew glyph indicates on which side of the mean the largest spatial grouping of samples lies.

3.3.3.3 Kurtosis

Kurtosis is a measure of how peaked or flat topped a distribution is compared to a normal distribution. Figure 3.11(b) shows three distributions with varying kurtosis,

where a flat, box-like distribution can be seen on the left. The glyphs chosen to represent kurtosis reflect the aforementioned categories of kurtosis. The glyphs are rendered using a deep purple color and are scaled so that their size reflects their magnitude away from zero (excess kurtosis). To distinguish between flat and peaked, the glyphs assume a flat or sharp shape depending on the sign of kurtosis. Thus, for a highly positive value, the glyph is very pointy; the more negative the kurtosis value, the flatter the glyph.

3.3.3.4 Tail

The final moment in the summary plot is what we refer to as tail, which is based on the fifth central moment, μ_5. The quantity is sensitive to distribution asymmetry farther away from the mean when compared to the skew. Tail will have a high magnitude when there are additional modes in the distribution or strong outliers. Like skew, tail is rendered as a triangle pointing in the direction of asymmetry. However, unlike skew, tail is rendered on the same side of the mean as its sign. The tail glyph is rendered as a sharp arrowhead, where both the size and sharpness is dependent on the tail quantity. The visual effect of this glyph should indicate that there is a significant number of samples biased far from the mean. Figure 3.11(c) shows a set of distributions with tail values varying from very negative to very positive. Upon close inspection, one can see a cluster of outliers in the rightmost distribution, which is indicated by the large tail glyph.

3.3.3.5 Moments, Sample Size, and Outliers

It should be noted that the higher order moments are very sensitive to noise, outliers, and variations in sample size. This can be problematic when the number of samples is not large enough to adequately characterize the underlying distribution. In such cases, the histogram visualization becomes extremely important, both in providing a redundant encoding of the characteristics expressed by the moments and also making it readily apparent to the user that the summary is based on a sparse number of samples. An alternative approach to the histogram is kernel density estimation, which approximates a distribution from a small number of samples. Using this technique, a smooth histogram can be generated, and the moments from this estimation can be calculated. However this does remove information about the number of samples. Work has been done to investigate methods for calculating higher order moments in the presence of noise, such as [40], however, these approaches increase the complexity of calculating the moments and are often application-dependent. We have chosen to use the more simplistic formulation

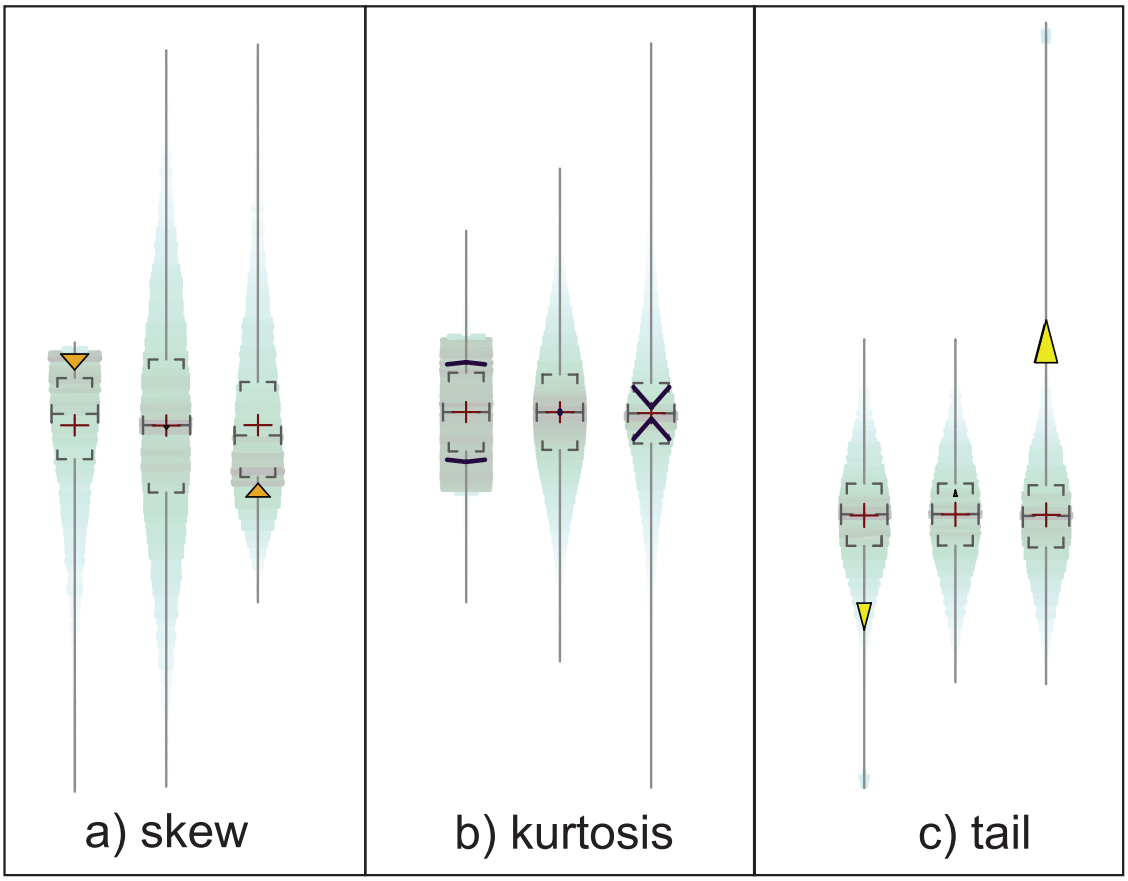


Figure 3.11. Glyphs for the higher order central moments. Each triplet of distributions shows negative, close to zero, and positive values for the respective moment. Each higher order moment is relative to the moments of a Gaussian distribution, which is the central distribution in each set.

of moments, and rely on the redundancy of the summary plot to highlight unreliabilities in the moment summary.

3.3.4 Distribution Fitting

Often, understanding the characteristics of a particular data set is less interesting than determining the canonical distribution that best fits the data because the feature characteristics of the canonical distributions (such as Gaussian, Uniform, etc.) are well known. The final element of the summary plot is a distribution fit plot, which represents either a best-fitting distribution or a distribution chosen by a user. The user is provided with a library of common distributions, including Gaussian, Uniform, Poisson, Rayleigh, Laplace, and others, as well as the fitting of multiple Gaussians and asymmetric

distributions to allow for the mean to be not centrally located. The fit distribution is displayed symmetrically as a dotted line showing the density of the distribution along the axis, as seen in Figure 3.12. Through the user interface, Figure 3.16, the user can also get information about the parameters of the fit distribution and closeness-of-fit statistics. The interface also allows the user to learn the summary plot on a Gaussian distribution, so variations from normal become easy to spot. In addition, any distribution can be used as a learning set, allowing the user to quickly identify data sets that resemble specific distributions, as well as to explore relationships between distributions.

3.3.5 User Interface for Reduction of Visual Clutter

While our design of the summary plot attempts to create glyphs that can be presented together harmoniously and minimize visual clutter, the presentation of all of this information at once may still at times be overwhelming. To ease this problem, we have designed a user interface that allows the user to interactively choose the desired information to investigate by turning the glyphs, boxplot, and histogram on and off at will, by modifying distribution fitting parameters, and by querying the plots for quantitative attributes such as number of samples. This user dialog comes to the forefront when the user clicks on the plot, and can be seen in Figures 3.13, 3.14, 3.15, 3.16. In this dialog, the user is presented with a variety of options for reducing the summary plot to only the statistics he/she is interested in, as well as providing options for some of the display modalities and pertinent information about the data.

3.3.6 Comparison of the Box and Summary Plots

The summary plot supports four different modalities for understanding a data distribution. Each modality expresses slightly different characteristics of a distribution, and the combination of the four reiterates and reinforces the specific characteristics separately expressed by each.

To demonstrate the effectiveness of the summary plot, Figure 3.17 compares different displays of rent data in five different US cities named Springfield, obtained from the 1990 census [4]. As shown by Figure 3.17(a), the boxplot alone gives a good comprehension of the extent of a distribution and indicates where the inner 50% of the data lies. This suggests the range in which a data sample is most likely to fall. For example, in Springfield, OH, it is reasonable to expect monthly rent to fall between $250 and $500. However, Figure 3.17(b) shows that the most frequently paid rent value is actually slightly less

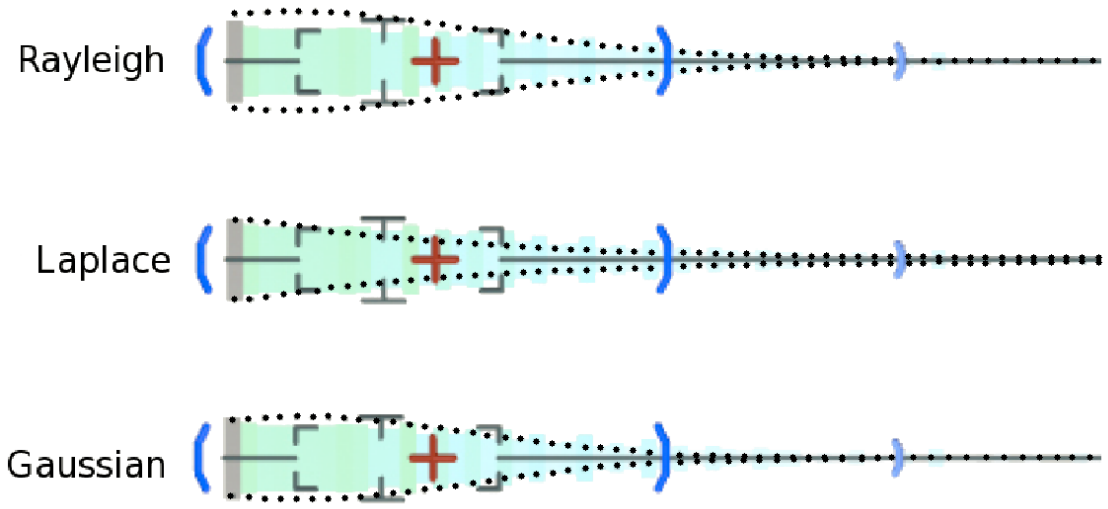


Figure 3.12. Fitting results for 3 distributions are shown as the dotted, curved lines on either side of the plot.

than $250. That is, while most renters pay a monthly rent between $250 and $500, the most frequently occurring amount of rent is slightly less than $250. Thus, the mode of the distribution lies outside of the inner-quartile range, and therefore the distribution of the rents paid in Springfield, OH is skewed slightly, which is indicated by the slightly enlarged skew glyph. This is not apparent in the boxplot alone, and the addition of both the density display and moment plot aide in the understanding of the data.

Another benefit of the summary plot is evident in the comparison of distributions. In Figure 3.17(a), the boxplots of Massachusetts (MA) and Oregon (OR) have fairly similar boxplots, which leads one to believe that their data distributions are similar as well. However, on inspection of the summary plot, we see that the underlying distributions are very different. The rent in OR has a distribution that is centered at the median with very little skew, while rent in MA is skewed and has a density bulge below the first standard deviation. Using the boxplot alone easily leads to erroneous assumptions about the similarity of distributions. The additional information in the summary plot still allows for easy yet accurate comparisons.

The comparison of summary plots is also useful when examining global trends. Figures 3.18 and 3.19 demonstrate the use of the summary plot on 4 variables from meteorological data: temperature, humidity, wet bulb, and pressure. At a glance, it is easy to

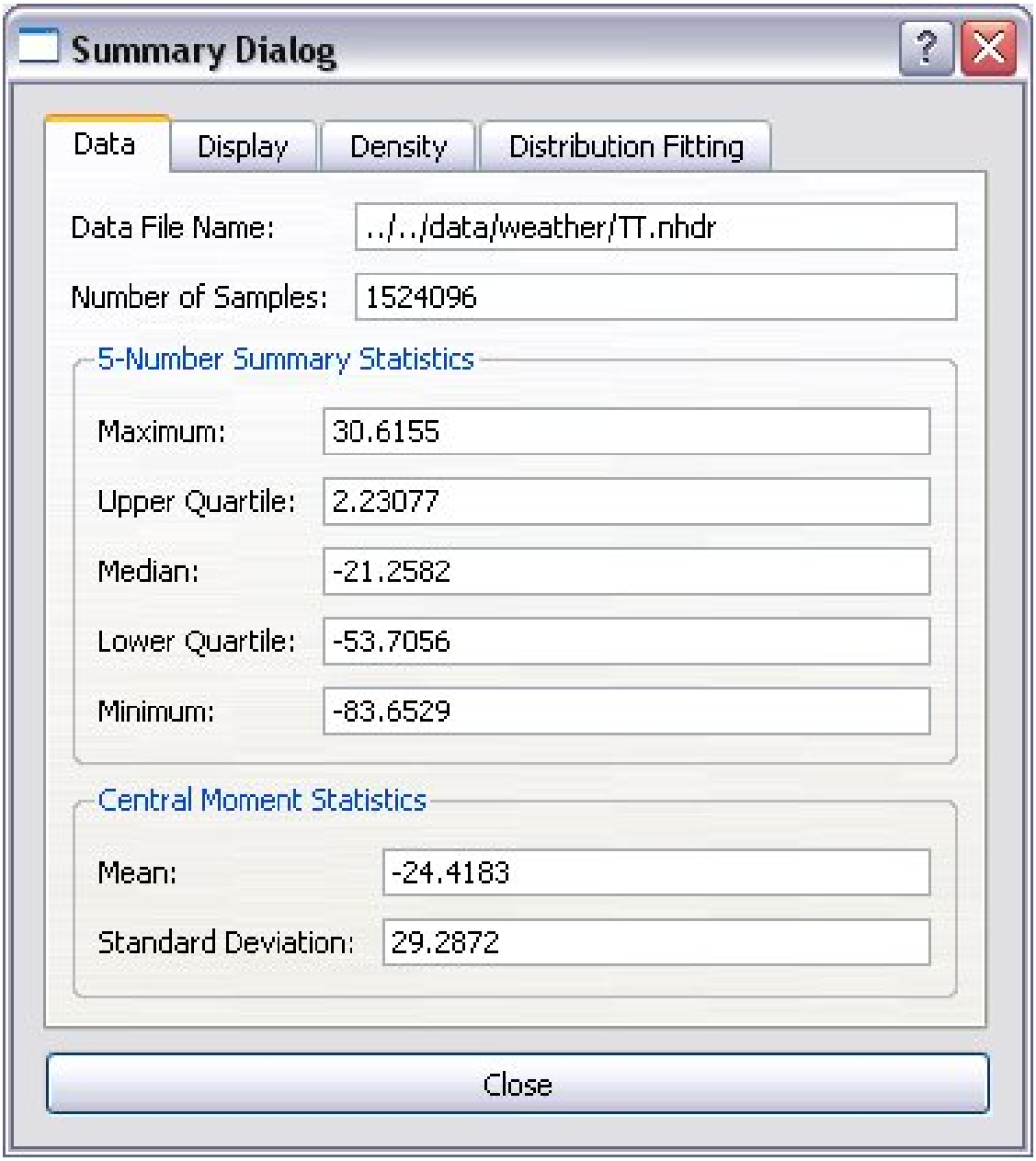


Figure 3.13. User interface for data information.

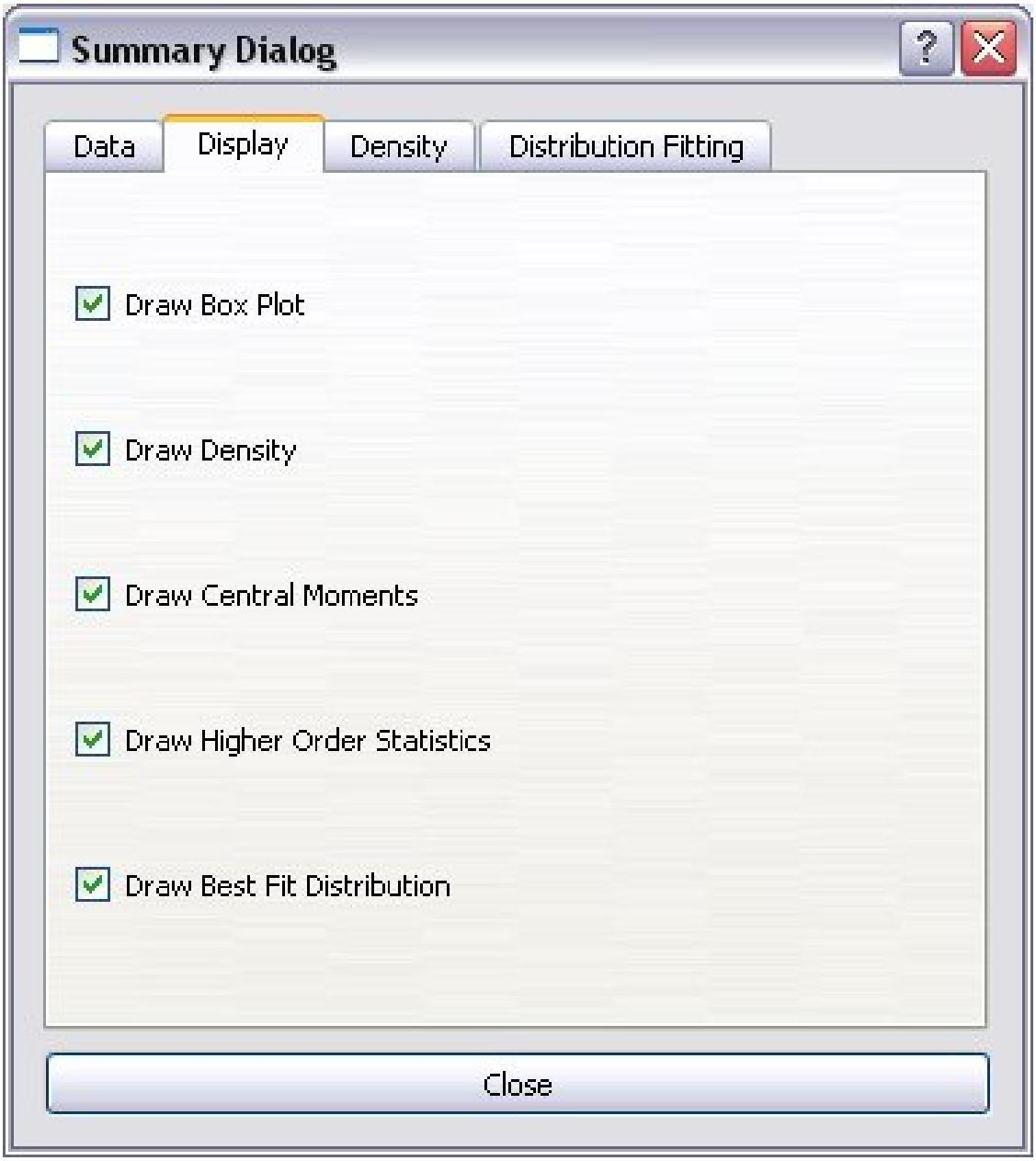


Figure 3.14. User interface for display options.

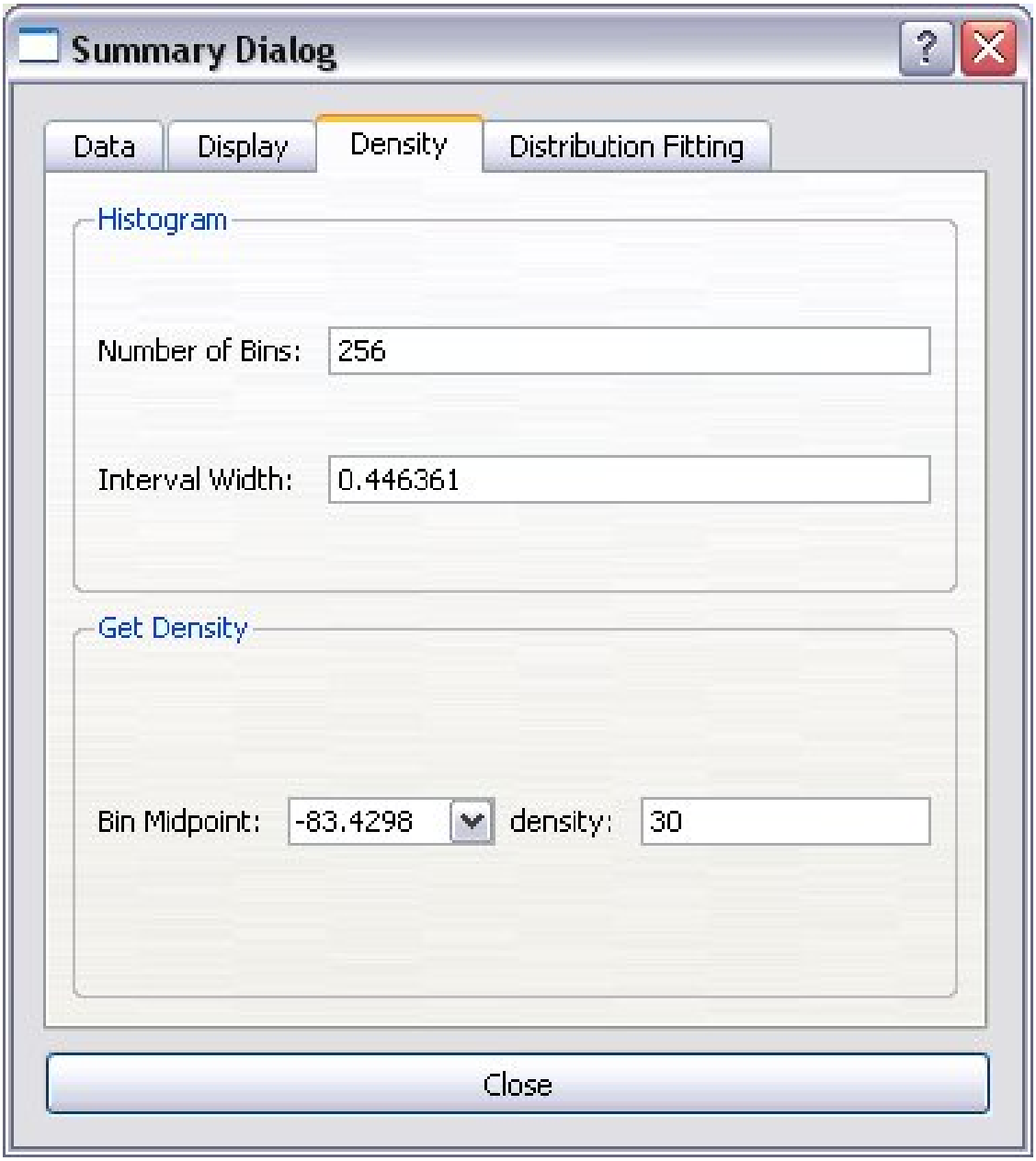


Figure 3.15. User interface for density information.

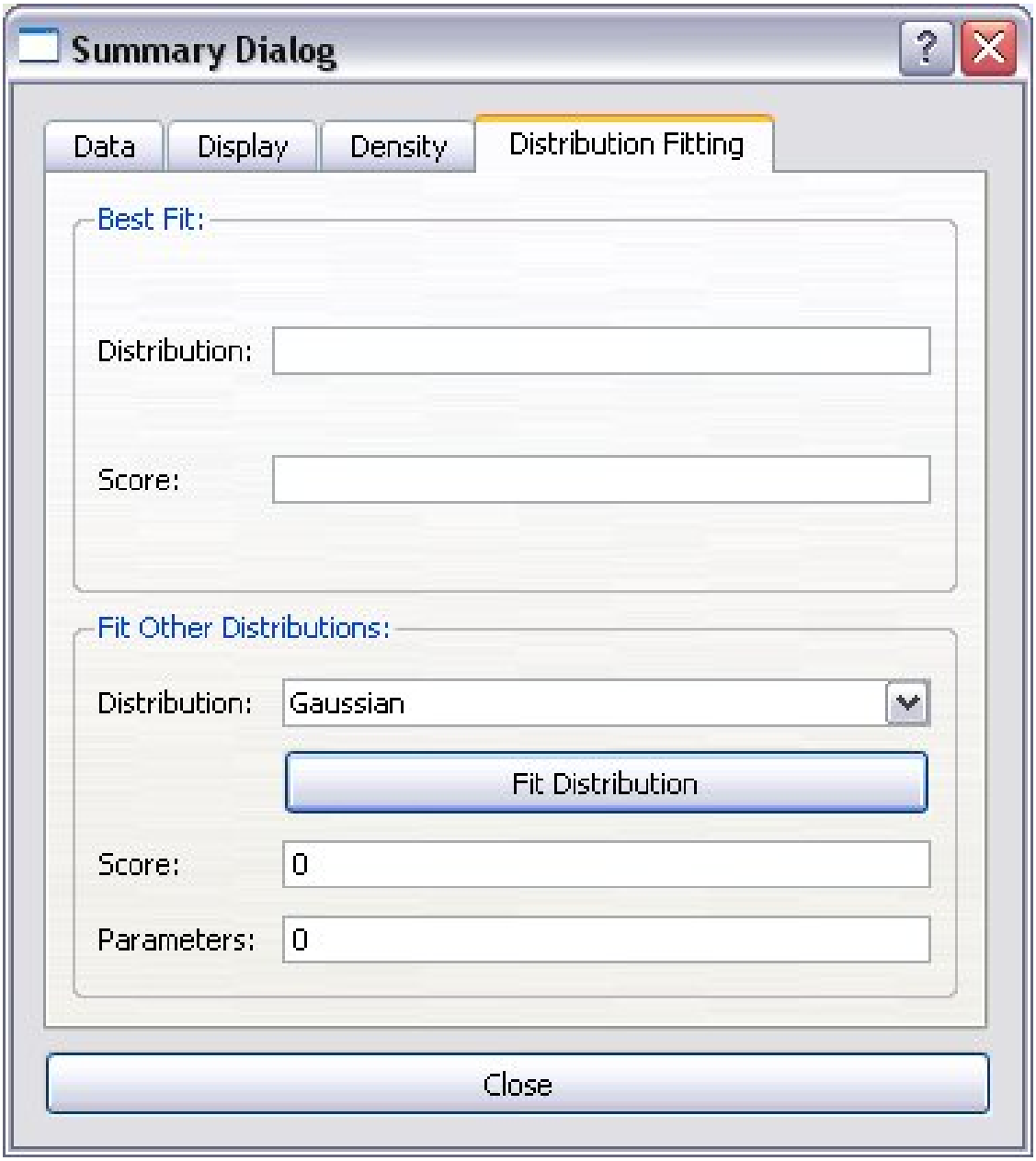
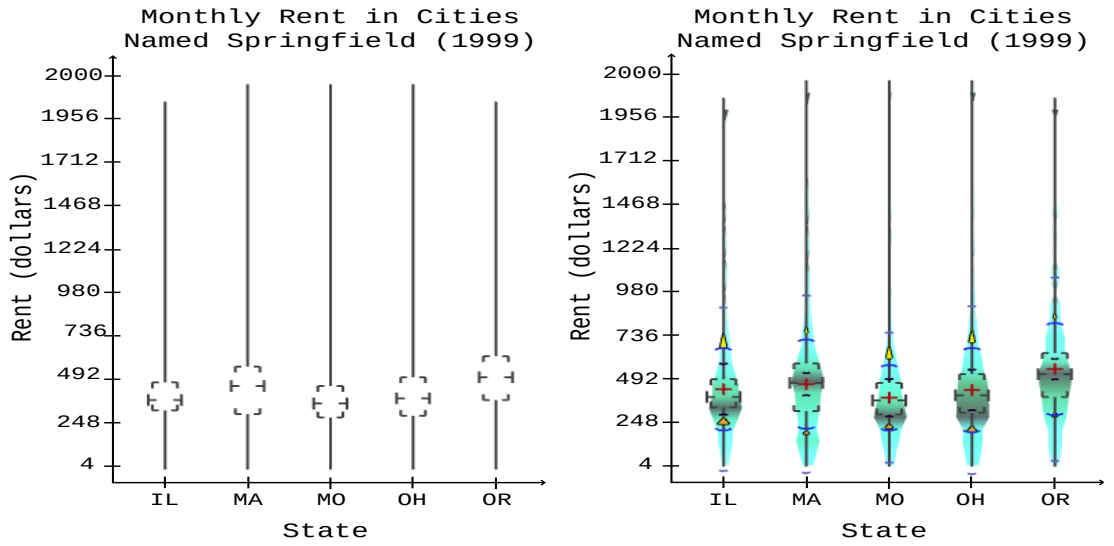


Figure 3.16. User interface for distribution fitting options.



(a) Boxplot

(b) Summary plot.

Figure 3.17. A comparison of the box and summary plots. While the boxplot effectively shows the range of the data, the summary plot shows that the mode is surprisingly outside of the inner quartile range for numerous states (MO and OH). The summary plot also clearly indicates distributions with high skew and large outliers (IL and OH) as well as emphasising that the OR distribution is close to a Gaussian.

see the trend of each variable as altitude increases. For each variable, the change between altitude slices is minor, however, comparing the plots for the first and last altitude shows a dramatic evolution. While a simple boxplot would show the trend of the ranges, the summary plot gives the indication of where the majority of the data lies as well as how the distributions evolve across the data. For instance, the temperature data has a modal value well above the mean across the entire data set. This is shown by the darkening of the density display and the placement of the skew glyph. Likewise, the pressure data starts out as a highly skewed, peaky distribution and evolves into a distribution that resembles a slightly skewed Gaussian. While the data range of pressure for each altitude slice is very limited, we can easily deduce the characteristics of the data from the moment glyphs. Using only the boxplot or the density display would not give any indication of the data distribution in this case.

The overall goal for the summary plot display is as an investigational tool to be used to both test and pose hypotheses. As such, summary plot displays must provide a global overview as well as drilling down into the local data. Figure 3.20 shows an example of this.

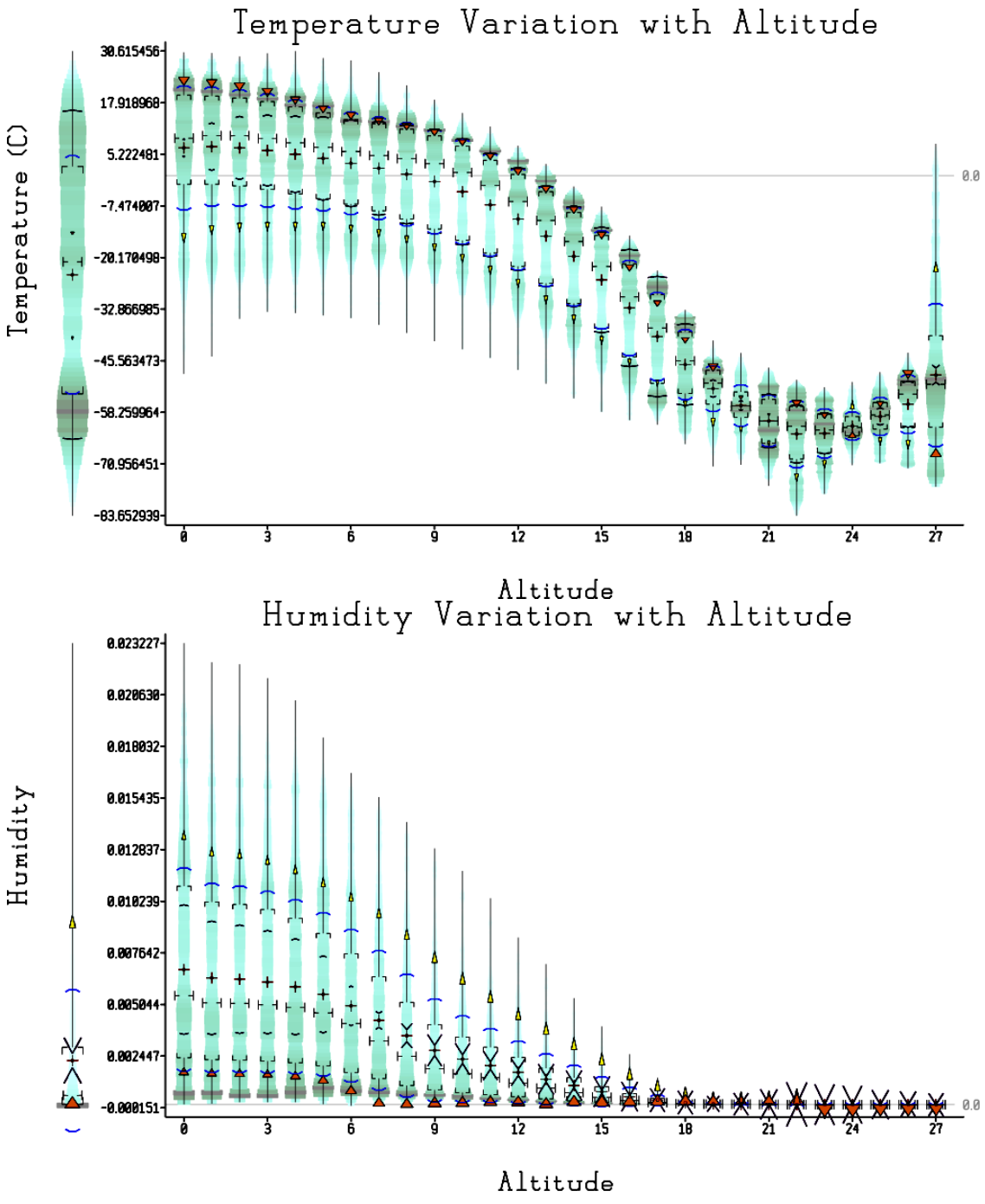


Figure 3.18. Temperature and humidity data. Taken at a single moment in time across the world. Data courtesy of the Canada Meteorological Centre.

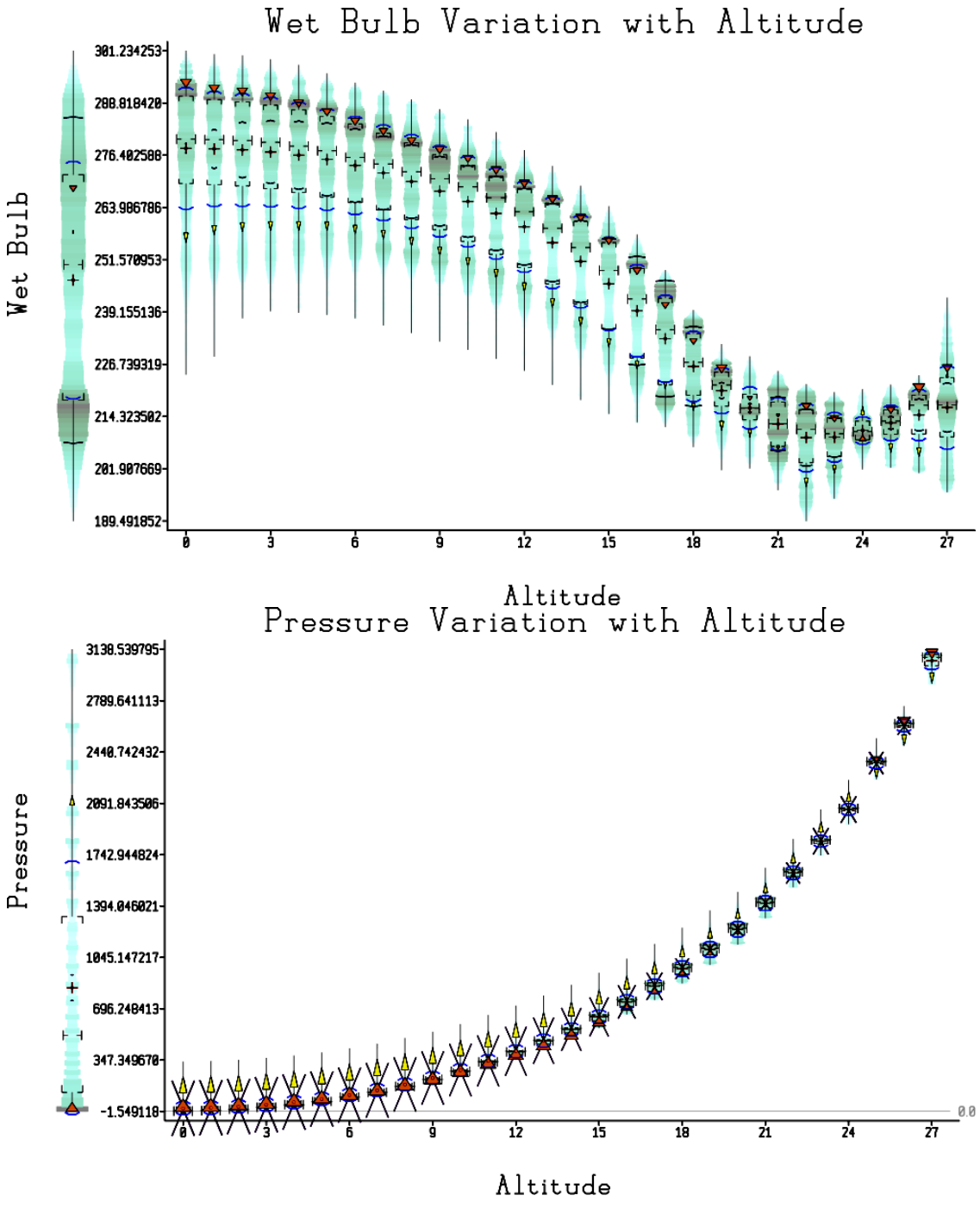


Figure 3.19. Wet bulb and pressure data. Taken at a single moment in time across the world. Data courtesy of the Canada Meteorological Centre.

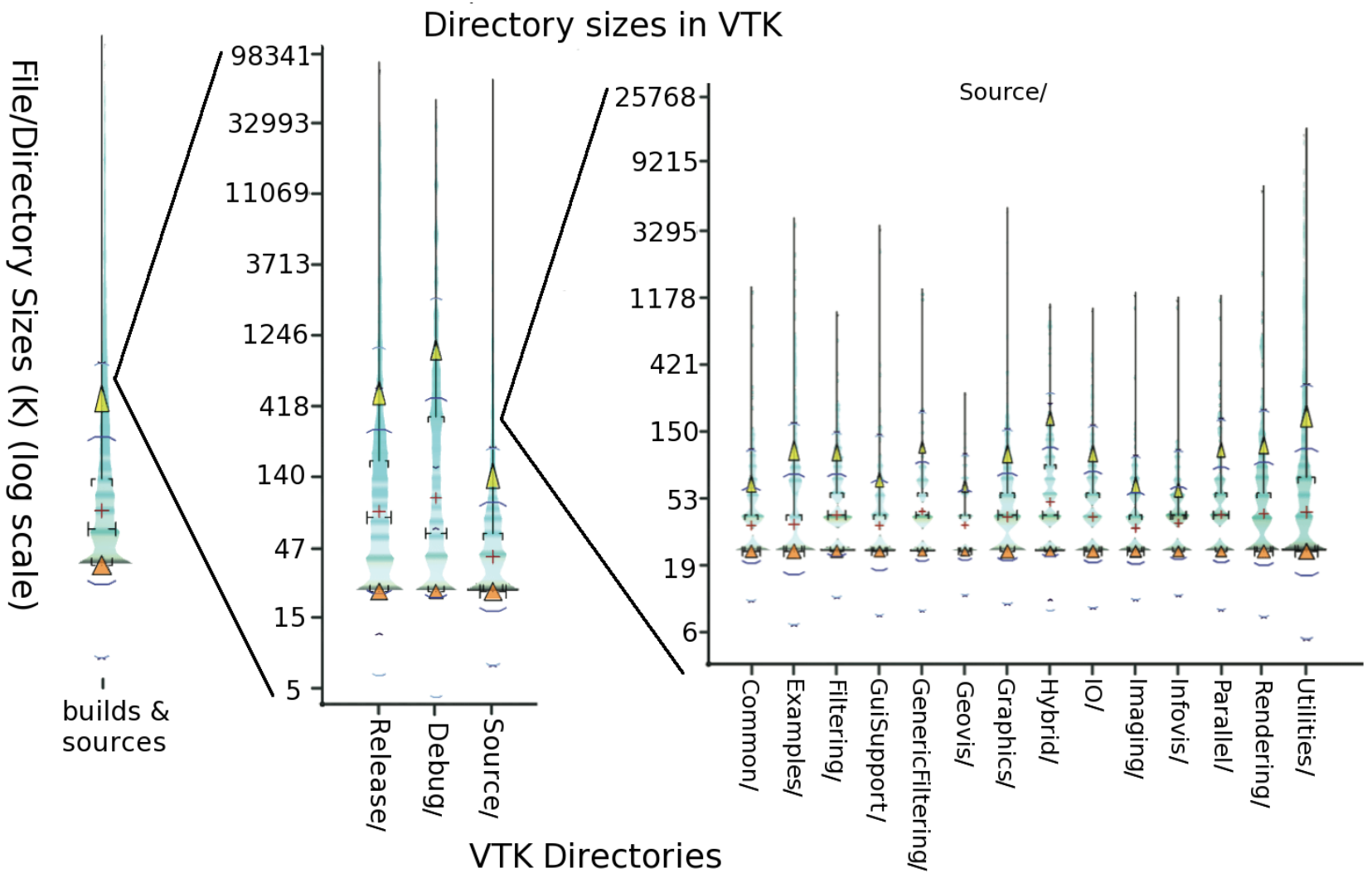


Figure 3.20. Bash *du* command run in VTK directories. On the left is the results of *du* on the release, debug, and source directories. On the right is a drill-down into the source directory.

On the left are the results of running *du* on the VTK [73] directory structure. A summary of the entire VTK directory is shown in the summary plot displayed in the left median, while the release, debug, and source directories are shown within the graph. Next, the source directory is expanded, showing each of the directories that make up the source of VTK. The most obvious characteristic of the source directories is that there is minimum size of source files, most likely due to the overhead of compulsory documentation at the top of each file. The move from global to local summaries exemplifies the interrogative nature of summary plots.

3.4 Joint 2D Summaries

While a statistical summary for a 1D categorical data set is highly useful, users require methods for comparing multiple, correlated data sets to understand how samples with multiple distinct data values are related. In this section, methods are explored for summarizing categorical data with pairs of values associated with each sample. Figure 3.21 shows the joint summary plot for two 1D data sets. The joint summary uses the 1D summary plots for each data set and places them perpendicularly to allow the summary plots to be used to orient the viewer. Joint mean and standard deviations, a joint histogram, and a reduced higher order moment display are added, providing a plot that shows how the correlated data sets relate.

3.4.1 Joint Mean and Standard Deviation

The first measures of correlation that are added to the display are joint mean and standard deviation. These measures are mainly used to orient, however, they do encompass the inner quartile of the correlated data set. The display uses lines that connect the mean and standard deviation of one distribution to the corresponding values in the other, using the colors of these measures from the summary plot. An image of these two lines drawn for a single comparison can be seen in Figure 3.21.

3.4.2 Joint Density

The density of a set of samples drawn from a 2D distribution can be directly visualized using a joint histogram as shown in Figure 3.22 for a single data set and in Figure 3.23 for multiple data sets. A joint histogram can be generated by subdividing the 2D domain into NxN bins, and, for each sample, incrementing the bin-count indexed by its pair of data values. Our system displays the joint histogram by rendering a quadrilateral at each

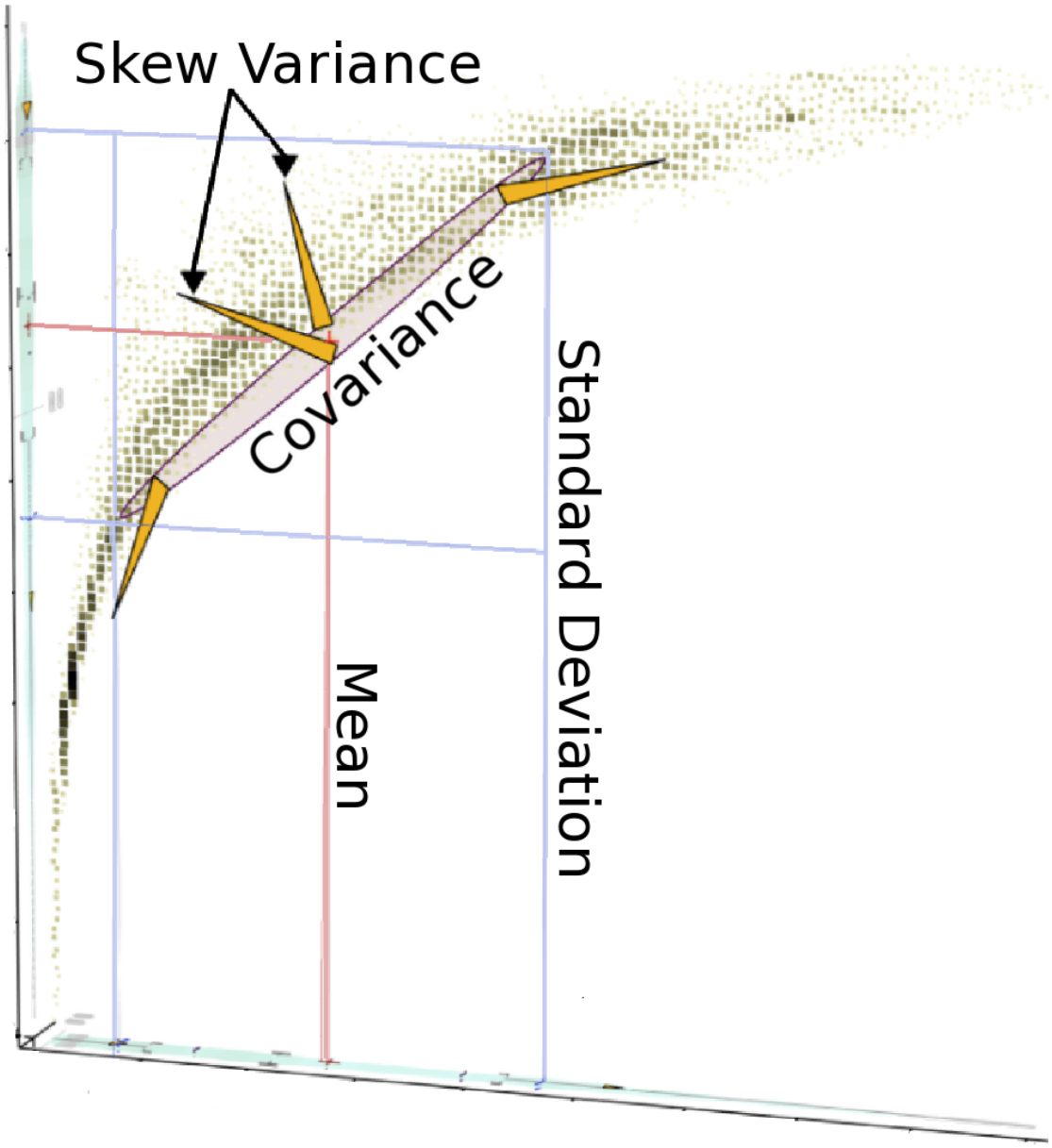


Figure 3.21. Joint summary for two 1D categorical data sets.

bin location scaled by the square-root of the normalized density for that bin. The joint histogram is colormapped such that darker quadrilaterals refer to larger joint densities.

Because of the regularity of the bin spacing, joint histograms used to simultaneously summarize multiple categories tend to produce aliasing artifacts. Jittering alleviates this problem by perturbing the position of the quadrilaterals for each bin, where the magnitude of the jitter is inversely proportional to the quadrilateral's scale. This is shown

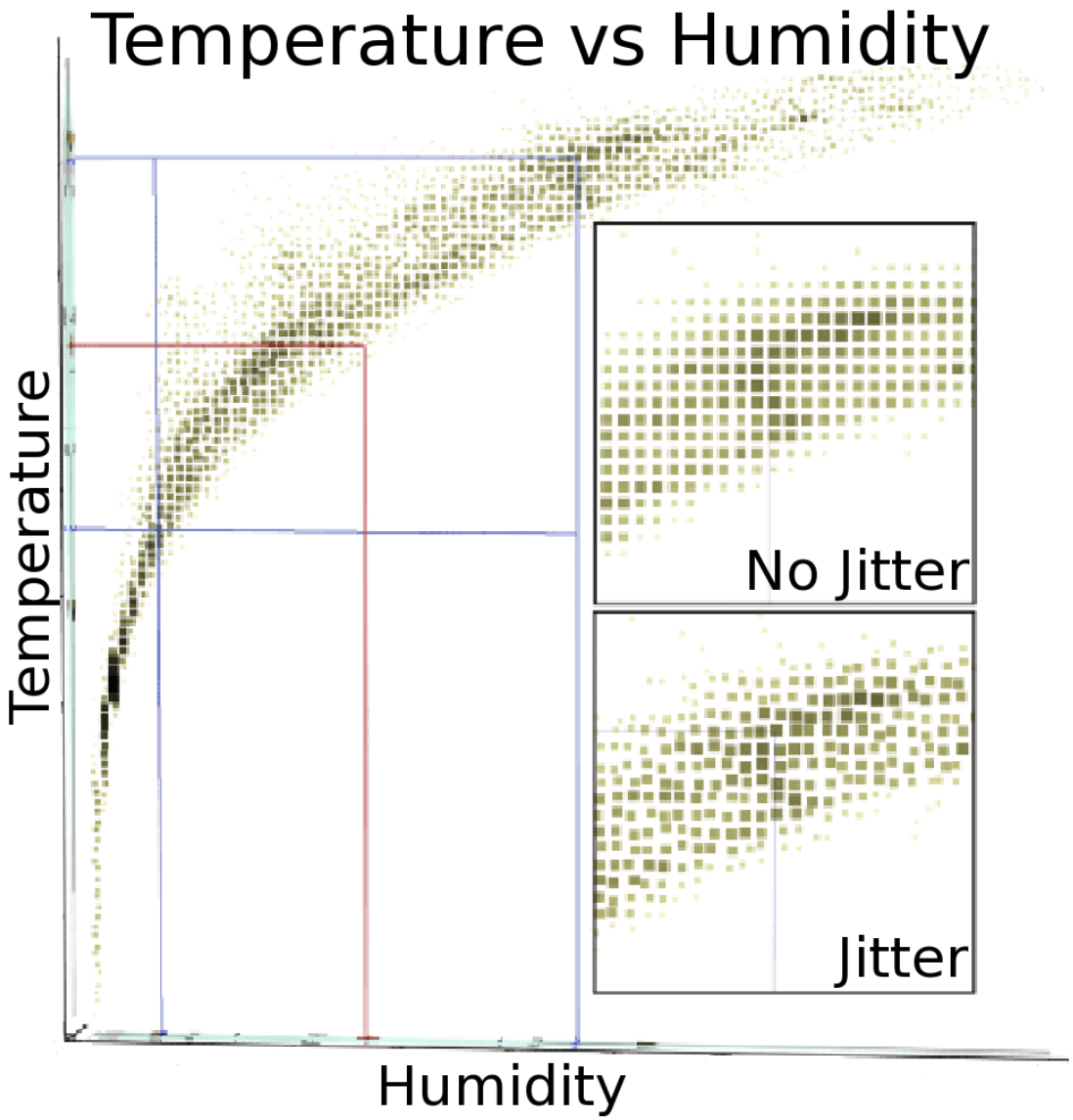


Figure 3.22. Joint summary and histogram for a single data set.

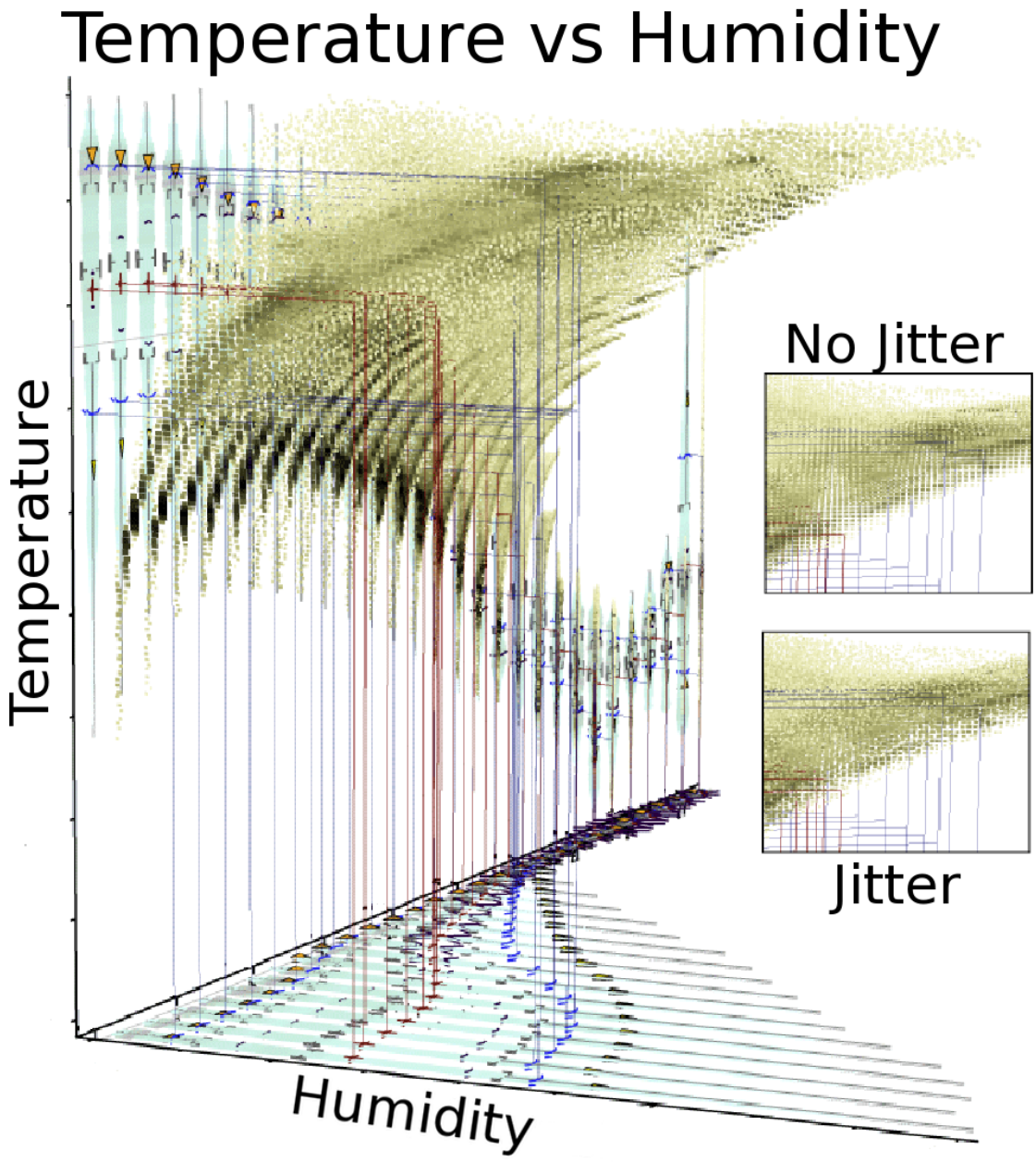


Figure 3.23. Joint summary and histogram on multiple data sets.

in the insets of Figure 3.22. This constraint ensures that the quadrilateral is drawn at a randomized location but is always inside the bin.

3.4.3 Covariance and Skew Variance

For multivariate distributions, the covariance matrix is the analogue of variance in 1D distributions. The covariance of two data sets, $\{x_i\}_{i=1}^{N}, \{x_j\}_{j=1}^{N}$ can be defined,

$$V_{ij} = \frac{1}{N}\sum_{k=1}^{N}(x_{i_k} - \mu_i)(x_{j_k} - \mu_j) \tag{3.2}$$

where μ_i and μ_j are the means for each data set. Covariance is a measure of how the two data sets vary in relation to each other. For these presentations, the covariance matrix is used to transform a unit disk so that the way the disk is stretched visually relates to covariance of the data sets. Since the interest is in a multivariate analogue of standard deviation, the covariance ellipse-disk glyph is scaled as follows:

$$\text{scale} = \frac{\sqrt{\text{ev}_{\max}}}{\text{ev}_{\max}}, \tag{3.3}$$

where $\text{ev}_{\max}$ is the maximum eigenvalue of the covariance matrix. The covariance glyph is laid on top of the joint histogram, and, for reference, the mean and standard deviation of both data sets are extended into the joint space using lines. Figure 3.21 demonstrates this approach on temperature and humidity data sets. The image shows the covariance ellipse between the first categories of the temperature and humidity data sets without the joint histogram. Figure 3.24 shows a close up of the joint summary plots with and without the joint histogram.

Just as covariance is the analogue of variance, higher order multivariate moments can also be described with matrices. The so called "skew variance" of two data sets, $\{x_i\}_{i=1}^{N}, \{x_j\}_{j=1}^{N}$ can be expressed by two matrices, $V_{i^2j^1}$ and $V_{i^1j^2}$ where

$$V_{i^m j^n} = \frac{1}{N}\sum_{k=1}^{N}(x_{i_k} - \mu_i)^m(x_{j_k} - \mu_j)^n \tag{3.4}$$

In general, these matrices are neither symmetric nor positive definite. Skew variance is visualized using four sharp arrows pointing in the direction of the skew located at the

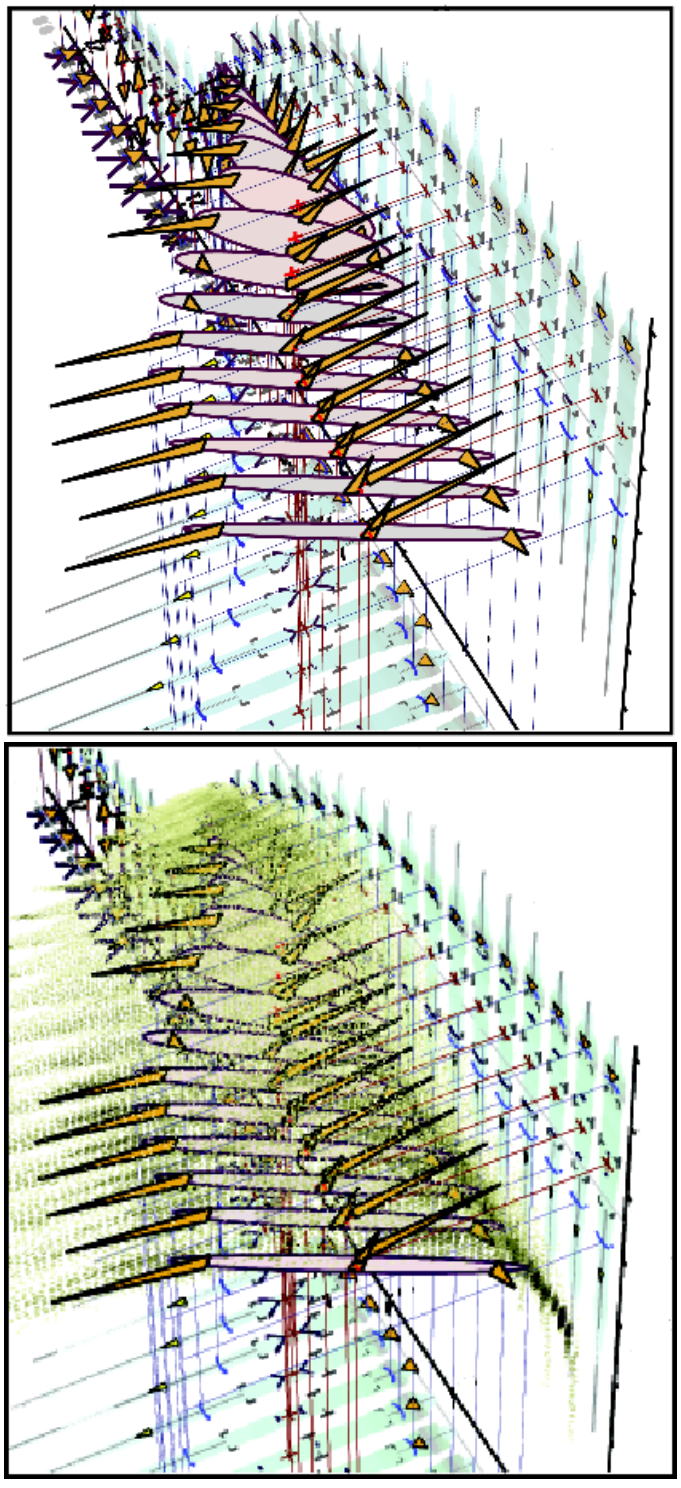


Figure 3.24. Close-up of the joint summary plot. The top image is without and the bottom image with the joint histogram.

endpoints of the covariance eigenvectors. The directions of the skew variance arrows are defined by the column vectors of $V_{i^2 j^1}$ and $V_{i^1 j^2}$. As with covariance, skew-variance visualizations are scaled:

$$\text{scale} = \frac{\sqrt[3]{\text{ev}_{\max}}}{\text{ev}_{\max}}, \tag{3.5}$$

where $\text{ev}_{\max}$ is the maximum eigenvalue of the skew-variance matrix.

The use of skew-variance glyphs in 2D (or higher-dimensional) distributions is important, since joint distributions can be very asymmetrical even when their 1D distributions are symmetrical. Figure 3.21 illustrates this. While the covariance ellipse indicates the overall trend of the joint distribution, it gives no indication that the majority of the distribution's density is outside the ellipse. The skew variance glyphs indicate the strong asymmetry of this distribution. When multiple 2D distributions are combined, as seen in Figure 3.24, the moment glyphs allow the user to visually identify each category. Without them, the individual joint histograms would be difficult to separate.

3.4.4 Correlation

The joint histograms in Figure 3.25 illustrate the use of higher order moment signatures in multidimensional distributions. The moment glyphs help guide the eye through the important characteristics of the distribution and disambiguate overlapping joint histograms. The figure shows a distribution that is essentially derived from two others (wet bulb-temperature can be derived from temperature and pressure). In this case, the expected behavior is a strong correlation between temperature and wet bulb, however, at high altitudes, the correlation trend breaks down and is clearly visible in the extreme skew-variance glyphs. In this example, the sensitivity of higher order moments to outliers can actually be viewed as an advantage. Because the joint histogram is readily available in the visualization, the skew-variance glyphs point us toward samples that may be missed due to their low frequency. In many cases, these outliers represent interesting phenomena or discrepancies in the data. These are situations that we feel are important to draw to the user's attention, even if they are to be disregarded as true outliers later in the analysis

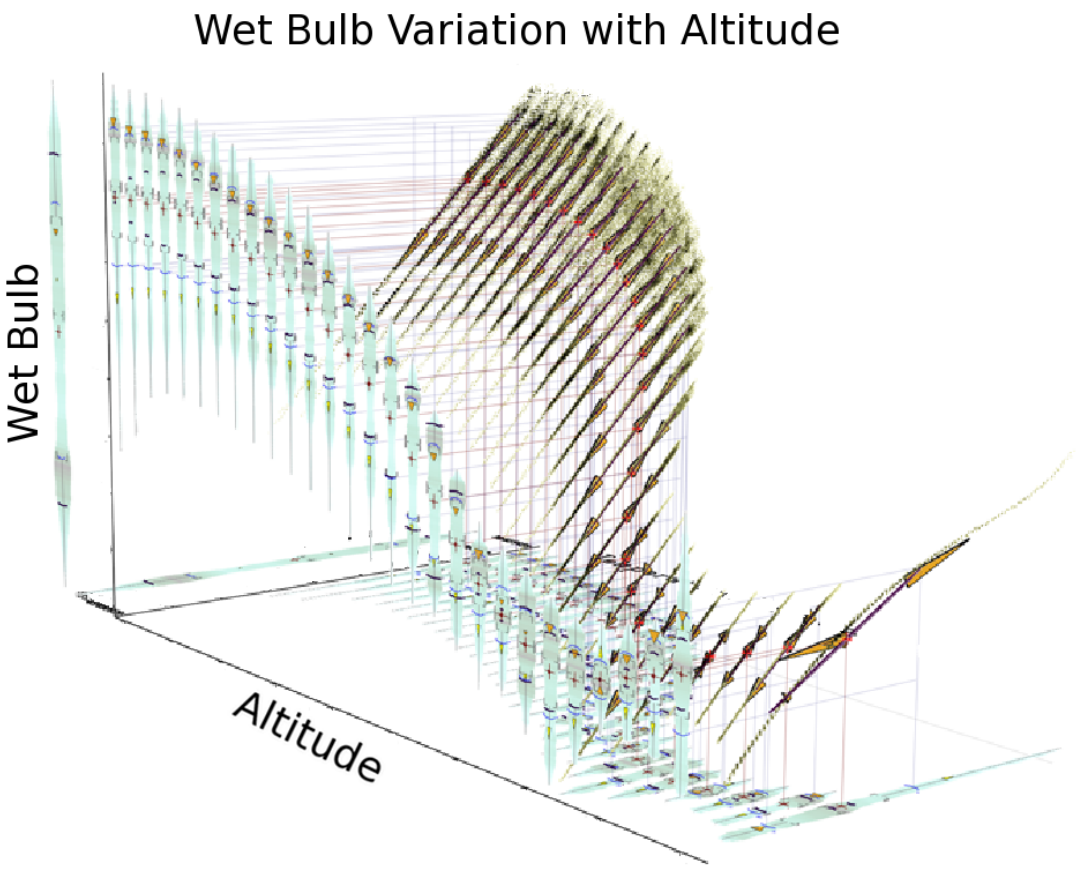


Figure 3.25. Correlation between the data from Figures 3.18 and 3.19.

3.5 Discussion

Uncertainty information has been inadequately addressed in the visualization community, largely because of the difficulties involved with visually expressing this additional data. However, if visualization is to become a robust decision-making tool, it must in some form represent uncertainty to the audience. This work provides a method for investigating visual characteristics of a data distribution, both for learning about the shape of the data set and for expressing the associated uncertainty.

The summary plot provides a simple way to annotate features of a distribution, enhance distinguishablity between data sets, and allow for the easy comparison of multiple distributions. It contains by nature uncertainty information in the form of standard deviation. In comparison to the boxplot, the summary plot quickly shows the location of outliers (i.e., on either side of the 2nd standard deviation), the position and size

of the skewness of the distribution, and the relative shape. When any of the glyphs is outstanding, the user can quickly understand that the corresponding moment has significance, and thus further examination may be in order. For instance, a user interested in the skew of a data set may simply look at Figure 3.17(b) and compare the size of the skew glyph to see the distributions with the most skew. In fact, this information is clear from the moment plot alone. Likewise, to compare the skew between distributions, the user need simply compare skew glyphs rather than making a close inspection of the histogram or inferencing from the boxplot.

The full summary plot, Figure 3.17(b), includes density information in the form of a symmetrically displayed histogram with the box and moment plots. The combination of the boxplot and histogram is conventional, since the boxplot can hide common aspects of a distribution that may be clear in the histogram. However, direct inspection of the histogram can be overwhelming, since the histogram conveys a large amount of information. Thus, the moment plot extracts much of this information into a subset, which is reinforced by the histogram.

The final modality of the summary plot is the use of distribution fitting. This encourages the use of the summary plot as a learning tool, since canonical distributions can illustrate the expected look of the summary plot. The fit distribution can also be used to generate robust higher order moments and glyphs when the number of data samples is too low to provide good direct estimates.

While the meaning of the statistics of the summary plot may not be initially intuitive, they do provide a standard set of features. This feature set has been translated into a learnable signature, which can serve both as a learning tool for the understanding of data distributions and also as a method for reducing a distribution to an easy to read summary, providing for quick identification of interesting characteristics.

While the boxplot has been used, almost universally, to summarize statistical data for nearly 60 years, there are many characteristics of a distribution that it cannot express. One such attribute is the mean or expected value of the distribution. Without this moment in the summary, a user may incorrectly assume that the median and mean are identical or closely correlated. Certainly, the same can be said about summaries based solely on the moments. This is especially true when the only moments considered are the first two, mean and variance. Such a summary would imply a symmetric uni-modal distribution, such as a Gaussian distribution. Together, boxplot, cumulant, and moment

summaries and distribution fitting express different yet complementary aspects of the data. However, they may still fail to expose important subtleties of the distribution. On the other hand, density plots or histograms simply summarize the data itself. While the histogram summary makes the modes of the data easily discernible, it does not allow the user to predict the median or mean values. By combining all summary methods, we can feel more confident in the analysis of the data and the questions that the summary is intended to help answer.

The combination of moment, cumulant, and density information into a single display can introduce visual clutter. Because each of the distribution components are overlapped in the display, characteristics of the summary may be occluded. These issues are significantly reduced using a dynamic display, where the user is allowed to zoom and rotate the view. In static displays, it is important that the choice of summary components is selected and arranged in a clear manner, such as side-by-side elements rather than overlapping.

3.6 Conclusion

The boxplot has been a highly effective means for conveying cumulant summary statistics. Using the boxplot as inspiration, a hybrid summary plot has been created that incorporates cumulant statistics, density, higher order moments, and distribution fitting. A generalized approach has been demonstrated for providing joint 1D comparisons as well as summaries of 2D categorical data. This approach targets reducing visual clutter while redundantly encoding information and simultaneously summarizing a large amount of data as a visual signature. The presentation of data in a summarized and easy to read form can quickly communicate information about both large amounts of data and the data's uncertainty, emphasizing meaningful characteristics and facilitating visual comparisons.

CHAPTER 4

THE VISUALIZATION OF MULTIDIMENSIONAL UNCERTAINTY DATA

Uncertainty information is an important characteristic associated with much of the data scientists encounter. While such information is often available, employing existing visualization techniques to uncover and explore uncertainty, as well as to integrate uncertainty information into data display, has proved challenging and remains an underdeveloped area of research. This chapter surveys the application of well-known visualization methods to uncertainty data. The main goal is to assess the ability of existing techniques to adequately display data with associated uncertainty as well as to uncover possible relationships present in the data.

4.1 Introduction

The estimation and visualization of uncertainty information is an important research problem in both simulation [27] and visualization [45]. Uncertainty concerns the error, confidence, and variation of a data set, information critical for allowing a scientist to understand the accuracy and assess validity not only of the data but also of the processes used to create the data. One such technique, sensitivity analysis, enhances the scientists' understanding of the effects of perturbing input parameters of a function. Small perturbations of the input parameters that create large perturbations in the output results can indicate areas of the function that are highly dependent on the input parameters. These regions of high variation might be interpreted as unstable or possibly incorrect. Sensitivity analysis techniques can be used not only to explore the mathematical models used to generate uncertainty data but also to understand better the effects of input parameters associated with visualization techniques. While this chapter focuses on uncertainty data generated from the sensitivity analysis of a mathematical model reconstructing a

biological experiment, the methods presented are applicable to many data sets of similar type.

Visualizing data sets that include uncertainty information can quickly become complicated. Incorporating the additional parameter of uncertainty into visualizations is challenging both because adding uncertainty increases the complexity and information content of the visualization, and because there is no generally adopted standard convention for the visual representation of uncertainty for three-dimensional (and higher) problems. Thus far, uncertainty visualization approaches represent uncertainty as a scalar, vector, or tensor value. This chapter explores the underlying probability distribution function used to create the data. Mean and standard deviation (or variance) are straightforward and robust ways to summarize quickly the data into an understandable, concise quantity. If the underlying distribution is Gaussian (normal), the mean and standard deviation completely characterize the underlying distribution. However, most distributions from scientific data sets are not entirely normal, and the deviation from normal is often the more interesting information. The sensitivity analysis data used in this chapter can be viewed in several ways; reducing it to mean and standard deviation is the simplest. However, since there is a functional relationship that such an approach ignores, a variety of visualization techniques are also presented that have the goal of learning about the relationship between input parameters and output results.

The remainder of this chapter is structured as follows. In Section 4.2, the data and its generating process is described more fully. Section 4.3 discusses previous work related to the visualization of distribution data sets. The next section, 4.4, presents visualization methods developed for distribution data, including methods that solely display mean and variance measures, as well as techniques to explore the full data space. Finally, Section 4.6 discusses the approach more thoroughly and some future directions.

4.2 Application Data

The data used in this chapter comes from the study of the impact of conductivity variation within computational models of bioelectric fields of the heart. In other words, scientists are working to understand how electrical signals, emanating from the heart and monitored on the body surface, propagate through the human torso. These signals, sensitive to variations in tissue conductivity and the geometry of the torso, attenuate as they propagate through the torso. Understanding how these signals change as they

move through the human body can help medical professionals discern normal changes in electrocardiogram readings from signals that can indicate abnormalities in heart function. This work looks at data generated by a sensitivity analysis of a mathematical model for reconstructing a biological experiment in which the voltages on the human torso are estimated based on the input electrical conductivities [36].

Such an approach quantifies the sensitivity of the cardiac electrophysiological models by stochastically varying the input conductivities of different tissues such as fat, lungs, or muscle and examining the resulting changes in potential across the torso. Figure 4.1 depicts the torso classified by tissue type, and Figure 4.2 shows the two-dimensional (2D) triangular domain on which the simulation is computed. The orientation of the torso is according to the standard radiological view, looking towards the head of a patient who is lying on their back. The simulation stochastically varies the lung conductivity uniformly $\pm 50\%$ from the reference lung conductivity. For each individual variation of input conductivity, κ, a set of heart voltages is generated at each mesh point of the torso. This pairing of input conductivity and the set of output voltages is called a realization; the data set consists of 10,000 such realizations. A straightforward way to consider these data is as a volume of slices (realizations) with axes (x, y, κ), called the κ-volume.

The data set resulting from numerous runs of the simulation describes the sensitivity of voltages across the torso to small changes in the input conductivity, thus describing the uncertainty associated with the mean voltage at each mesh point. Its large size is the main challenge in visualizing these data; not only is it hard to visually understand 10,000 2D slices, it is unclear how to present this information in a way that leads to a proper interpretation of the sensitivity. An alternative way to look at these data is from the point of view of the individual mesh points. Each mesh point has 10,000 samples associated with it. This is, in essence, 10,000 realizations from which one can attempt to form an approximation of the probability distribution function, or PDF, located at each mesh point.

A simple approach to visualizing these data is to summarize through statistical measures on the PDF, thus reducing the dimensionality of the data to a small number of parameters, such as mean and standard deviation. This can be seen in Figure 4.3, in which mean (a) and standard deviation (b) are displayed through a typical rainbow colormap. Thus, we can see that the mean, or expected value of the data set, ranges roughly from -13 to 13 and that the highest variation in the data lies in the lower right

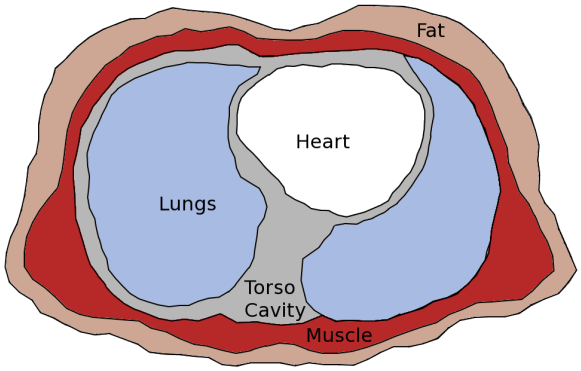


Figure 4.1. The classified human torso.

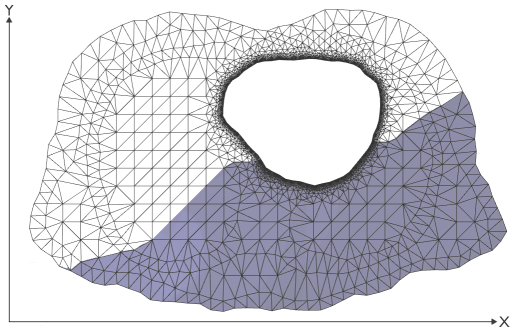
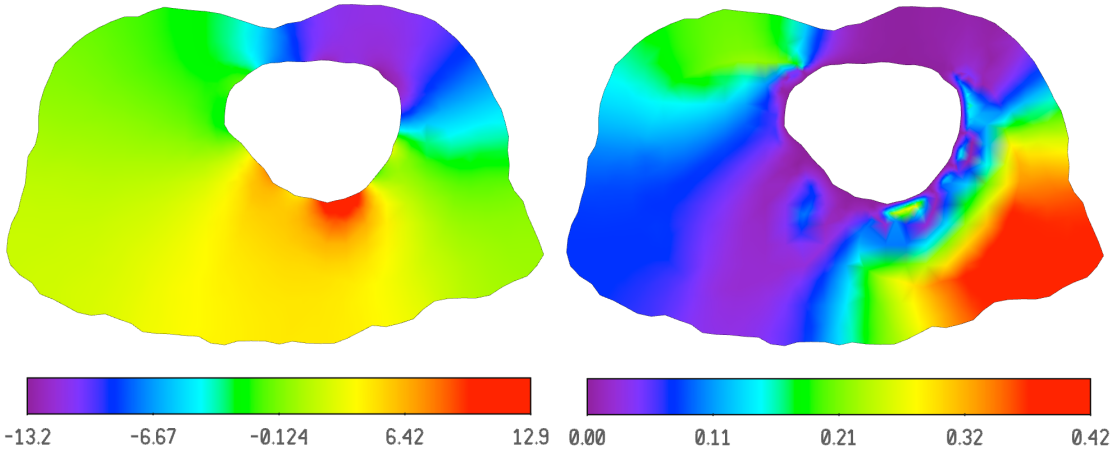


Figure 4.2. The torso mesh. The (x,y) triangular mesh is the domain space of the potentials data.



(a) Mean for ±50% lung conductivity.

(b) Standard deviation for ±50% lung conductivity.

Figure 4.3. Mean and standard deviation of the torso data.

corner of the torso domain. Such a presentation is straightforward in that it readily allows the scientist to see the data ranges and estimate data values in regions across the spatial domain. However, simple colormaps discard a large amount of data, provide presentations of only one aspect of the data at a time, and decouple the perturbations on input conductivity parameters from the resulting output voltages, a relationship that is important to preserve. The work in the rest of this chapter looks at visualization techniques that explore the data space to better understand the intrinsic relationships as

well as to effectively present the results of the sensitivity analysis.

4.3 Related Work

Work related to the visualization of the type of data presented in this chapter can be categorized in numerous ways, including visualization approaches that reduce the data through summarizing quantities, methods that strive to incorporate measures of quality or uncertainty into traditional scientific visualizations, and techniques that seek to visualize probability distribution functions.

Most of the visualization techniques for presenting uncertainty use mean and standard deviation as the usual measure of uncertainty. These measures are primarily used as statistical descriptors of data distributions and are often available alongside a data set as tables, graphs, and charts [75, 80], but they must be incorporated into existing visualization techniques if we are to portray the data comprehensively, completely, and accurately [83]. Pang et al. [65] have created a taxonomy of visualization approaches based on the type of uncertainty data. Common approaches include colormapping, glyphs [39, 84], isosurfacing [47, 70], volume rendering [30], and annotation [19]. A more thorough discussion of uncertainty visualization techniques is provided in Section 5.2.

There is a body of research investigating methods for displaying probability distribution functions with spatial positions. Each of these methods takes an exploratory approach to the presentation of the data by filtering down the amount of data, and then providing a user interface for the scientist to explore the data sets. Ehlschlaeger et al. [32] present a method to smoothly animate between realizations of surface elevation. Bordoloi et al. [14] use clustering techniques to reduce the amount of data, while providing ways to find features of the data sets such as outliers. Streamlines and volume rendering have been used by Luo et al. [58] to show distributions mapped over two or three dimensions.

Kao (2002) [50] uses a slicing approach to show spatially varying distribution data. This approach is interesting in that a colormapped plane shows the mean of the PDFs, and cutting planes along two edges allow for the interactive exploration of the distributions. Displaced surfaces as well as isosurfaces are used to enhance the understanding of the density of the PDFs.

Case studies of specific data have been performed by Kao [48, 49]. Their data sets come from NASA's Earth Observing System (EOS) Satellite images and Light Detection And Ranging (LIDAR) data. The methods used to show this data include encoding

the mean as a 2D color map, and using standard deviation as a displacement value. Histograms are also employed to understand better the density of the PDFs. To explore the mode of specific distributions, a small set of PDFs are plotted onto a color mapped spatial surface.

This small body of work on visualizing probability density functions is an interesting starting point for this research. However, the aim of this work is to avoid the data reduction employed in these previous techniques in order to preserve the correspondence between input and output parameters and to explore the full, nonreduced data set. This additional directive requires a visualization paradigm that can appropriately handle large data sets with high dimensionality.

4.4 Two-Dimensional Techniques

Central main challenges to visualizing these data stem from the large number of realizations with which we are confronted and the fact that it is unclear how to visualize the data in order to bring out the most salient aspects. The meaning of the collection of realizations is not intuitive, however, this work seeks to impart understanding not only of general results of the simulation, but also of the relationship between input and out parameters. Thus, this work begins by investigating methods for visualizing only mean and standard deviation, as these quantities quickly give insight into the data results. Then, more sophisticated approaches for exploring the realization space are presented.

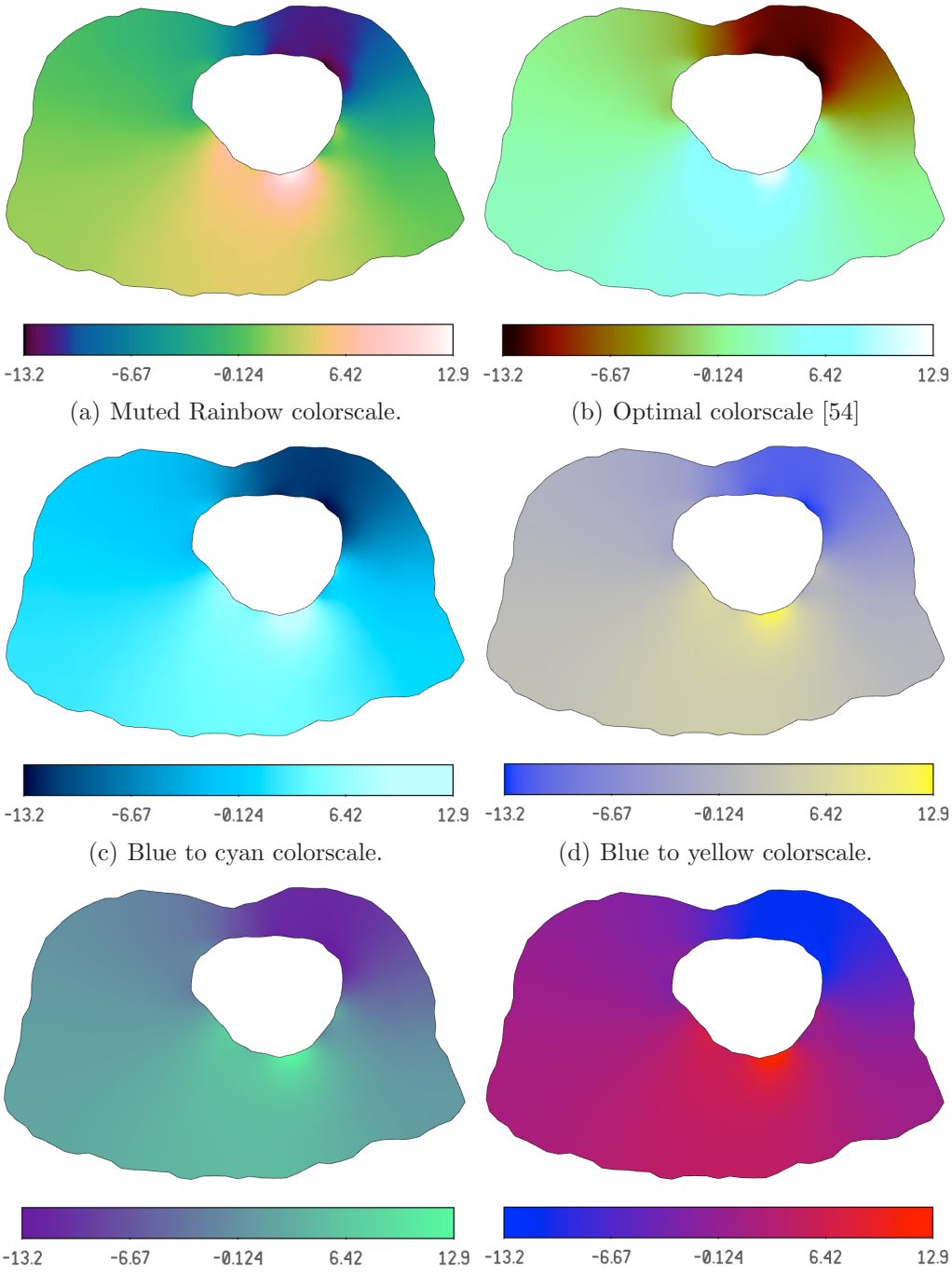
4.4.1 Colormapping

The first approach explored in this work looks at colormaps to display the mean and standard deviation of the data. The visualizations in Figure 4.3 present an overview of the results of the sensitivity analysis in that the expected value and variation are shown using a colormapped representation. These figures use a standard rainbow colormap. While the rainbow colormap is a standard device used to present data, and is used by default in many visualization systems, it often is not the best choice to present data [15]. Often the data values being presented are strictly increasing and the rainbow colormap does not universally convey an intuitive sense of low to high, and may even suggest some categorization. For example, how do we know if an area colormapped purple has a lower data value than one colormapped yellow? The typical answer is to reference the color bar legend (if provided) and match the color of interest to the scale. While the prevalence of the rainbow colormap affords recognition of the juxtaposition of the colors in the map, the

ordering of the colors is not a naturally intuitive ordering and previous knowledge (such as the ordering and color of the wavelengths of light) is assumed. There may also exist cultural biases, audience preferences, or viewer disabilities such as colorblindness that further strengthen the argument against naively using the rainbow colormap. Depending on the type of data and the questions being asked, as well as the visualization audience, number of salient features, and context of the display, a different colormap should be used. However, the choice of the correct colormap to use is difficult and too often subjective [69]. In the following, a collection of different colormaps is presented and their strengths and weaknesses discussed. While this is not an exhaustive list of colormaps, the intuition imparted from this discussion should aid in a more informed choice of colormap.

Figures 4.4 and 4.5 show a collection of different colormaps displaying both the mean and standard deviation of the torso data using the same colormaps. Although scaled so that lower data values correspond to darker colors in the colormap, the top row shows colorscales similar to the rainbow colormap. These colormaps move through a continuous range of colors while simultaneously increasing the brightness. These types of scales can be very effective in that high and low values are shown in bright and dark, respectively, and the color variations between the highest and lowest colors lets the viewer pick out particular data values. For example, it is straightforward to pick out the areas of the torso with the highest and lowest mean values, and we can find values between -1 and 1 because they are colormapped using a greenish hue in both depictions. These scaled colormaps help the viewer to easily differentiate between high and low values, rather than having to match colors as is the case in the traditional rainbow, however, mapping through numerous hues may not be appropriate for the data, or may be problematic when used in conjunction with other visualization schemes. Reducing the map to one or two colors may be a much more effective choice.

The bottom two rows of colormaps display variants of two-color gradients. Here, the lowest and highest values in the data set correspond to one of the two colors on the end of the scale. Data values then blend between the two ends of the scale. These colormaps are useful for data that have a low to high ordering, as these values are the easiest to pick out. Central values manifest as blends between the end values; if these central values are important, the blending between the two colors is then also important. For example, the blue to yellow 4.4(d) and blue to red 4.4(f) both blend to a distinctive color that both stands out as another color (grey in the blue to yellow map and purple in the blue to



(a) Muted Rainbow colorscale.

(b) Optimal colorscale [54]

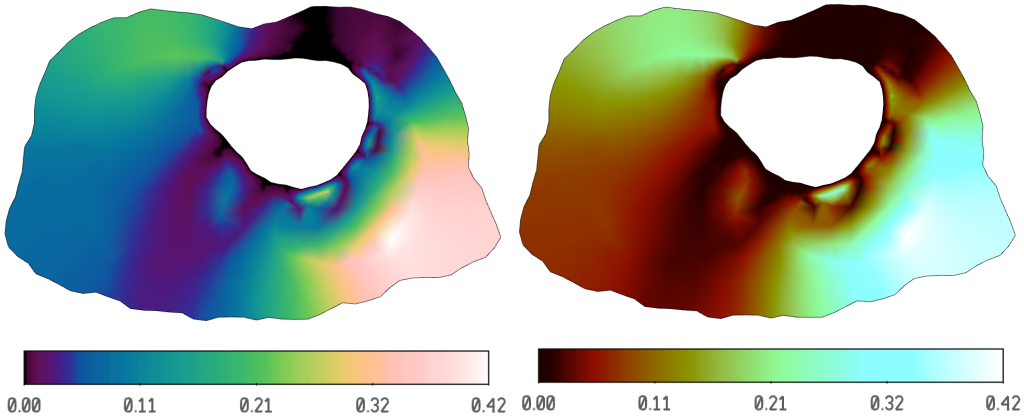
(c) Blue to cyan colorscale.

(d) Blue to yellow colorscale.

(e) Purple to seafoam green colorscale.

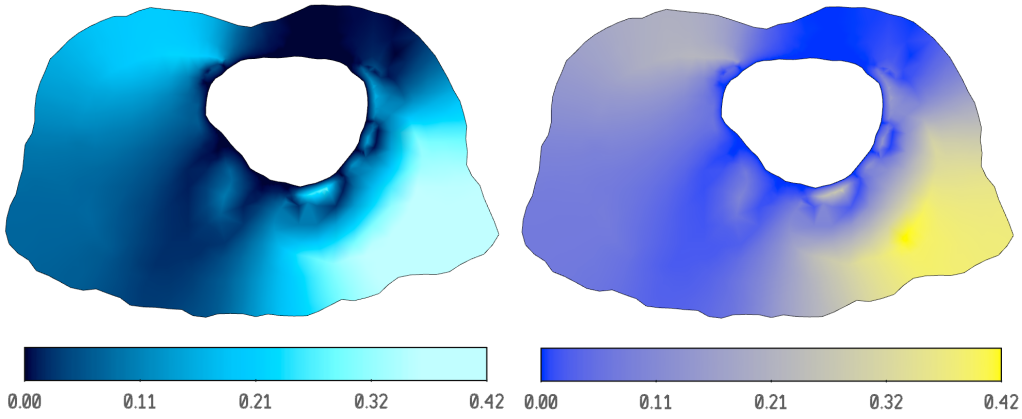
(f) Blue to red colorscale.

Figure 4.4. Colormaps of mean.



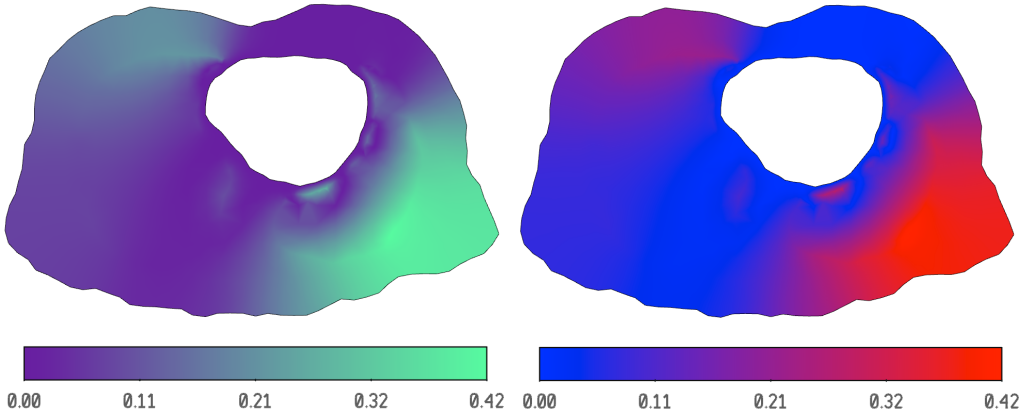
(a) Muted Rainbow colorscale.

(b) Optimal colorscale [54]



(c) Blue to cyan colorscale.

(d) Blue to yellow colorscale.



(e) Purple to seafoam green colorscale.

(f) Blue to red colorscale.

Figure 4.5. Colormaps of standard deviation.

red) but also is visually clear as a blend between the endpoints. In contrast the purple to seafoam green 4.4(e) does not compliment itself with a central color, and while the two colors blend, a central-most color does not stand out. Colormaps with this characteristic lend themselves to data that is highly two-valued, or where the point of the visualization is to show the data tending towards one value or the other.

The blue to cyan colormap 4.4(c) can be thought of either as a single or two-color gradient. This colormap is interesting in that is has a very apparent low to high ordering, however, its positioning in the blue spectrum allows for other colors to be employed in other areas of the visualization without interfering with the colormap. Also, the human visual system is particularly adept at distinguishing between variations in blue [16], and thus distinct data values can easily be picked out of this colorscale.

One interesting effect of colormapping is how distinct the effectiveness of one colorscale is compared to another, especially with respect to the display of different data. For example, Figure 4.5 shows the same colormaps as Figure 4.4, however, they appear rather different. For example, the mean data varies smoothly through the range, and thus a lot of the data is colormapped to the middle values of the color space, while the standard deviation has lots of disjoint values and thus values low on the colorscale often appear in juxtaposition to values encoded high on the colorscale. Thus, the goal of the visualization should drive the choice of colormap. In the standard deviation images, the blue to red colormap highlights areas of low and high, but the central values of std are not as easy to see, as they are in the blue to yellow colorscale. Likewise, very low standard deviation values are distinctive when using the Optimal colorscale [54]. Thus, having an a priori understanding of what the viewer wants to study in the data is a strong consideration for choosing an appropriate, effective colorscale.

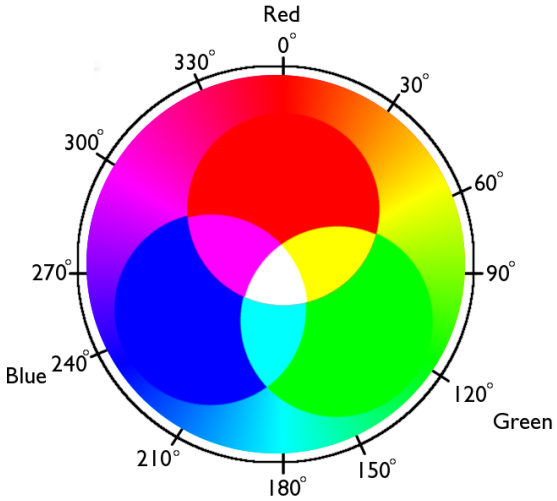
The simplicity of traditional colormapping is the strongest argument for displaying the data in this way, however, this simplicity may be an argument against its use. Only a single variable can be presented at a time, and the viewer must investigate multiple visualizations simultaneously to understand multiple variables. An improvement to the standard colormapping approach combines multiple variables into a bivariate colormap, allowing the viewer to gain simultaneous insight into the behavior of two variables at once.

4.4.2 Bivariate Colormaps

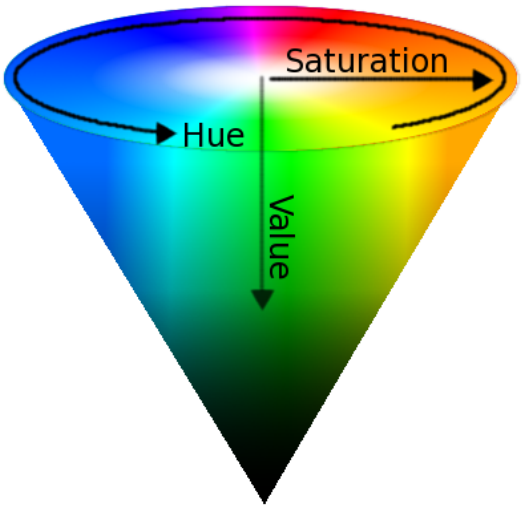
The design of bivariate colormaps is a difficult process and the choice of colorspace becomes very important. The typical Red, Blue, Green (RGB) or Cyan, Magenta, Yellow (CMY) colorspaces are used primarily for computer hardware such as monitors or printing, respectively. The RGB space is well suited to the human-visual system's strong perception to these three primary colors. However, both spaces are difficult to navigate and are not great at describing colors intuitively [5]. An easier colorspace to work in is Hue, Saturation, Value (HSV), which can be seen in Figure 4.6. This colorspace is also 3D, however, the dimensions do not refer to percentages of primary colors, as they do in RGB and CMY, but are a more precise description. The relationship between RGB, CMY, and the hue component of HSV can be seen in Figure 4.6(a).

The navigation of the HSV colorspace is a simpler task than the navigation of RGB or CMY because of the separation of the space into easy to understand components. The colorspace describe by HSV can be thought of as a cone, as shown in Figure 4.6(b). Hue can be thought of as degrees around a circle. As one travels around the outer edge of the circle, the hue changes in a rainbow-like ordering. Saturation is defined as the amount of white light that dilutes a pure color and can be thought of as the movement from the center of the circle that is pure white (corresponding to no saturation) along the radius to the edge of the circle that has high saturation. Likewise, value is the amount of grey present in the color and follows the tip of pure black cone (value of 0) to the outer edge of the hue circle. These three dimensions of the space are shown separately in Figure 4.6(c). In comparison to the RGB and CMY colorspaces, the HSV has a more natural interaction. First, the hue is chosen and then value and saturation attributes are applied to modify that hue. This is in contrast to RGB or HSV where the dimensions of the colorscale must be combined not only to create the desired color, but also to modify the light and darkness of that color. For example, when describing a light brown color in RGB, one must know that red and green have to be combined, however, it is not clear as to how much of each, and there is a not a single slider that can be adjusted to change the shade of that brown.

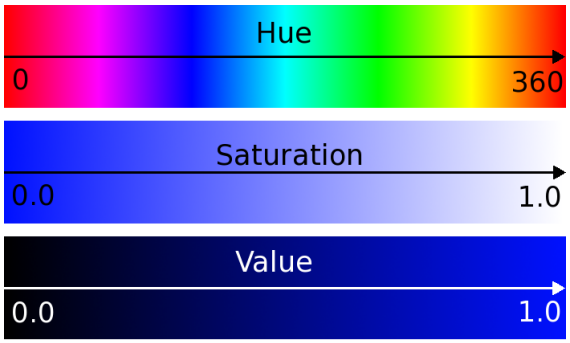
In the case of the mean and standard deviation of the torso data, Figure 4.7 shows the creation of a bivariate colormap. Here, mean is colormapped using a blue to green colorscale, and standard deviation is displayed by interpolating through the *value* component of the HSV color space. Because we consider areas of high standard deviation to



(a) RGB, CMY, and hue.

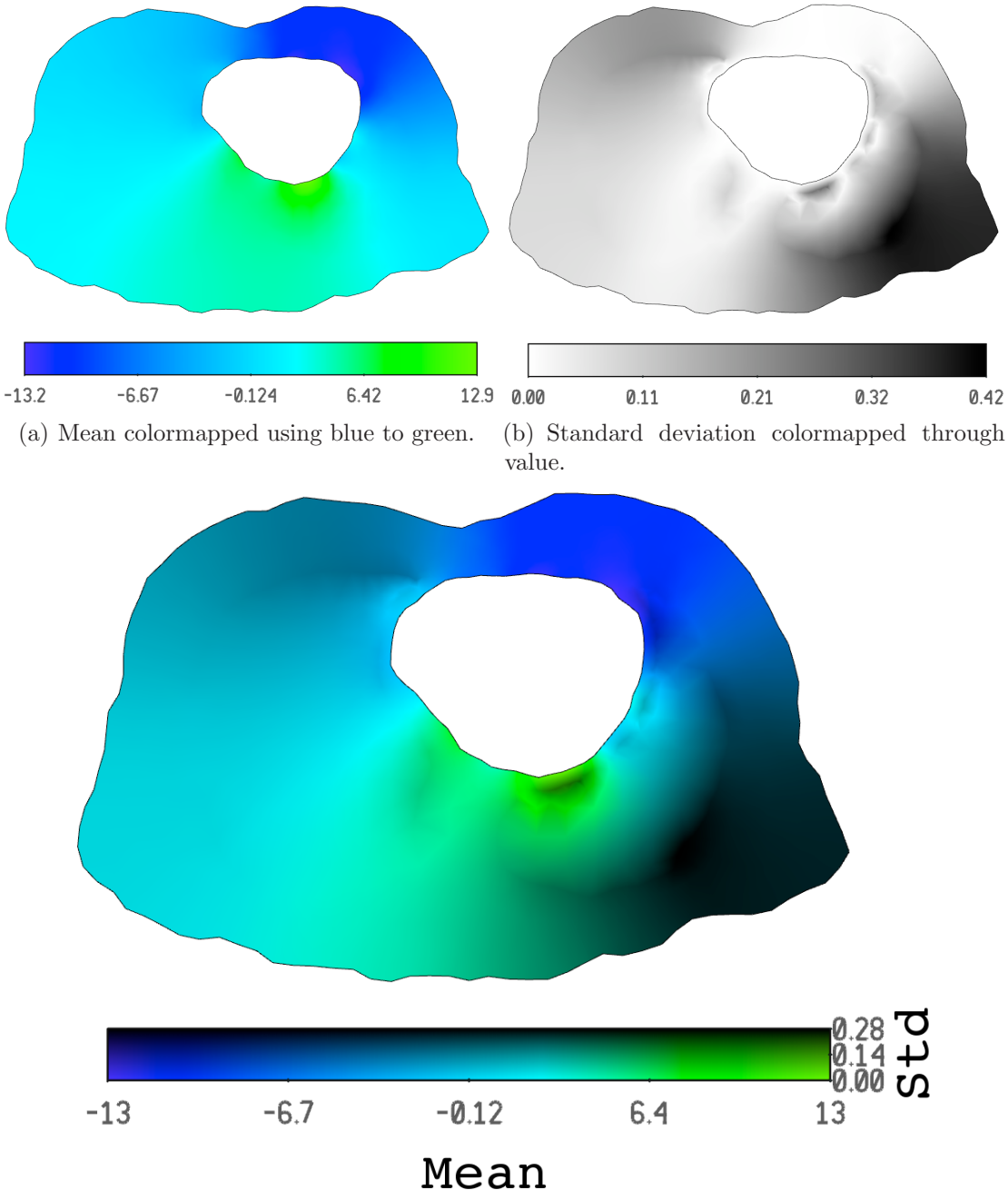


(b) The HSV cone.



(c) The 3 dimensions of HSV.

Figure 4.6. Colorspaces.



(a) Mean colormapped using blue to green.

(b) Standard deviation colormapped through value.

(c) Combined bivariate colormap displaying both mean (left to right on the colorbar) and standard deviation (bottom to top on the colorbar).

Figure 4.7. Bivariate colormap of mean and standard deviation.

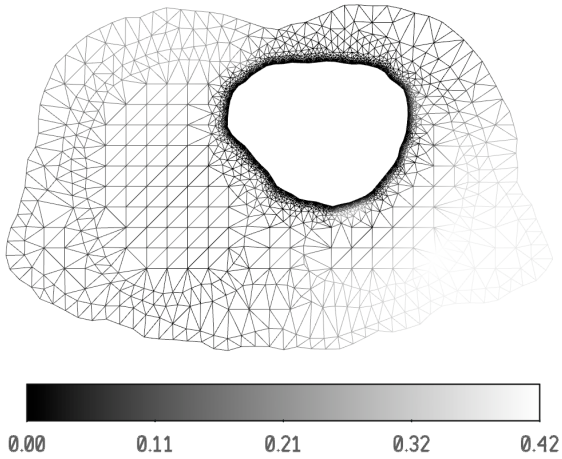
relate to areas of high uncertainty, the value component is mapped from light to dark, causing areas with high standard deviation to be dark and harder to see, and areas with low deviation to be bright and clear. This allows users to have an impression of the variance of the data in areas where they find the expected value interesting. While this approach does obscure some data by encoding areas of high standard deviation in a dark color that can leave the mean at those locations hard to interpret, this trade-off can be acceptable. This is especially the case when this type of approach is used in combination with other techniques such as texture or displacement mapping.

Similar to the approach presented in [19], another technique for adding to the readability of these types of displays is to use the wire frame of the spatial domain to encode uncertainty. Figure 4.8 shows the construction of this approach. The triangular mesh is used to encode the standard deviation of the data through a linear mapping of the opaque-to-transparent scale; areas of low variation are clear and readily perceived. As shown in Figure 4.8(a), the mesh becomes more transparent as the uncertainty grows. This mesh can then be added to the bivariate colormap (Figure 4.8(c)) to encode redundantly standard deviation, and possibly enhance the readability of this type of display. Or it can be used as the only means for encoding standard deviation, as seen in Figure 4.8(b).

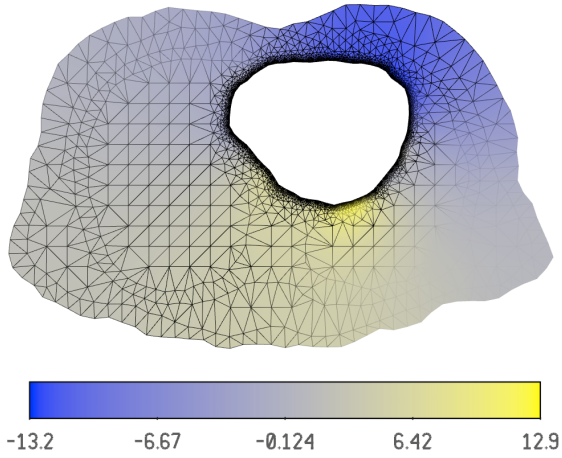
4.4.3 Perceptual Considerations

While the use of color to encode data is ubiquitous and seemingly straightforward, there exist many factors that may influence the interpretation of the colormap. Complexities inherent in the human visual system should be acknowledged in the creation and use of colormaps. Characteristics such as favoring particular wavelengths, quick detection of discontinuities, and artifacts deriving from the fusion of the images between the left and right eye lend themselves to quirks in visual perception that may help or hinder the interpretation of colormaps. Ongoing research is being conducted on harnessing attributes of the visual system to enhance visualizations, however, this section aims to simply create an awareness of the idiosyncrasies of the eye.

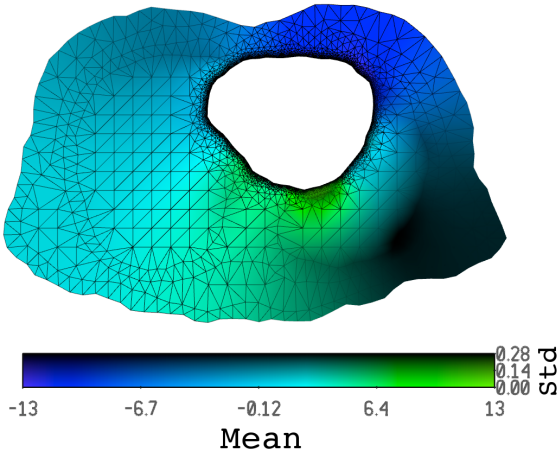
Simplistically, the eye consists of photoreceptor cells that detect changes in light. Two major groups of these cells are rods and cones. The cones of the eye detect bright light in three wavelengths, short (corresponding to blue), medium (green), and long (yellow to red). The rods are adapted to detect low light and are not sensitive to color. This view of receptor cells may misrepresent the action of rods and cones of just detecting a particular



(a) Encoding standard deviation in the triangular mesh.



(b) Blue to yellow colorscale with mesh.



(c) Bivariate colormap with mesh.

Figure 4.8. Uncertainty encoded in the triangular mesh.

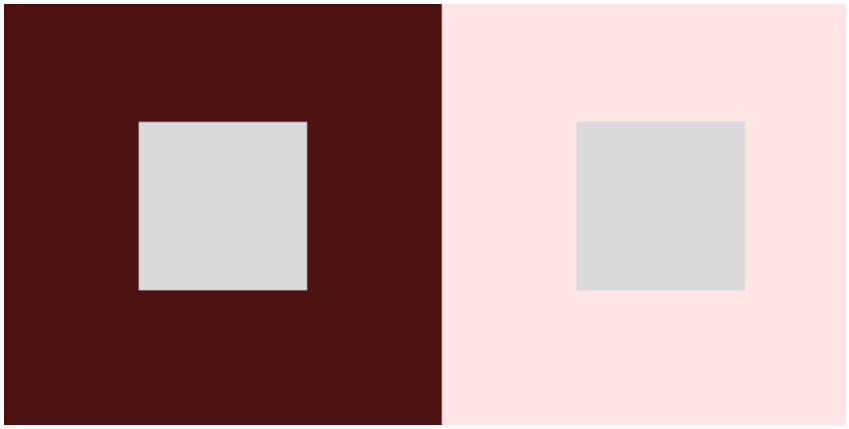
color, while in truth there exist complex relationships between cells. The wavelength sensitivity of the cells overlaps and the perception of color is really the interaction of signals between cells.

The interconnectedness of the cells in the visual pathway creates levels of processing that specialize in detecting particular aspects of the visual field. Signals from the photoreceptor cells pass through numerous intermediate cells that are adapted to detect things such as horizontal movement, line orientation, and contrasting colors or work to fuse the left and right visual planes into a 3D view of the world. Each of these levels of processing imparts artifacts that can affect the perception of visualizations. Examples of these visual phenomena are shown in Figure 4.9 and include mach banding, simultaneous contrast, and relative size [16]. Mach banding is the occurrence of amplified edges between regions with differing contrast. Figure 4.9(a) demonstrates this effect, where each band of color appears lighter on the left and darker on the right. Figure 4.9(b) shows the illusion of simultaneous contrast where the central grey squares are both the same color, however, the leftmost square appears lighter than the rightmost. This illusion exposes the effect of the juxtaposition of colors. Finally, the relative size example, Figure 4.9(c), exemplifies how context can impact size perception. Here, both yellow circles are the same size, however the size and placement of the purple circles creates the illusion that their sizes differ. In addition to optical illusions, other factors including color blindness, display medium (i.e., monitor or print), viewing conditions such as ambient lighting, and even personal preference should be taken into consideration when designing visualizations, particularly colormaps.

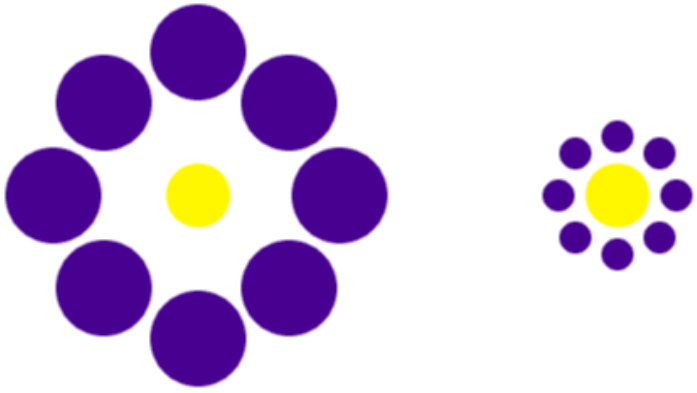
4.5 Three-Dimensional Techniques

Colormapping approaches can, if designed well, effectively portray one or two properties of a data set across the spatial domain, such as mean and variance, or a single instance of a data realization. However, colormaps suffer from quantization errors and it is difficult to precisely interpret data values at specific spatial locations. Large amounts of data are replaced, in favor of a few, summarizing parameters. The reduction of data allows for very little further exploration and limiting the visualization to 2D, while simplifying the display, ignores the third dimension that can be exploited for display. In addition, the complexities existent in the data are obfuscated by summarization. Thus, more sophisticated visualization techniques must be employed to bring out subtle relationships

(a) Mach banding.



(b) Simultaneous contrast.



(c) Relative size.

Figure 4.9. Various visual phenomenon.

and explore the entire data space.

4.5.1 Displacement Mapping

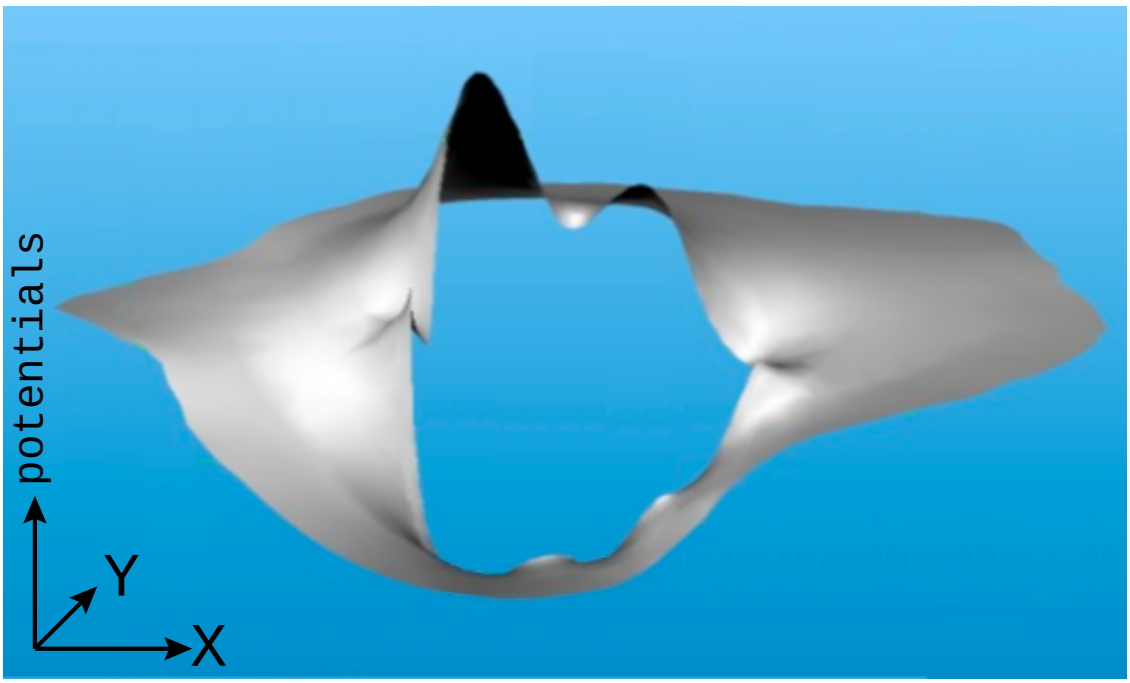
The first step toward improved visualizations is to define a space in which the visualization will exist. Because the third dimension is not used in the spatial domain of the data, it is free to be defined as we choose. To this end, as demonstrated in Figure 4.10, the (X,Y) plane is the coordinate system from the torso spatial mesh, and the Z axis will be defined as the potential, or resulting heart voltage. Mean is encoded as height along the Z-axis; positive Z values correspond to positive potentials and conversely, negative Z values show negative potentials. Standard deviation is colormapped and added to the displacement mapping of mean in order to summarize the results of the sensitivity analysis.

4.5.2 Volume Rendering and Isosurfacing

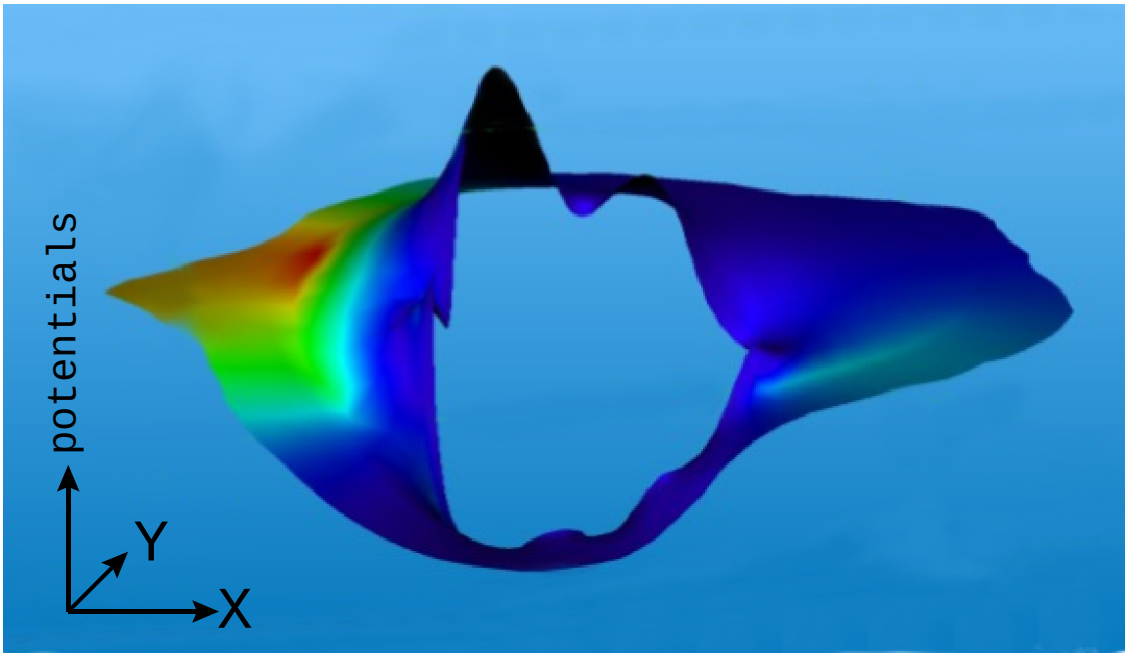
While taking advantage of the extra dimension available from the spatial domain, displacement mapping techniques simply summarize the output of the sensitivity analysis and do not highlight relationships between perturbations between input and output. Rather than defining the Z-axis as the space in which the outcome potentials are defined, the Z-axis can be used to represent the realizations of the simulation. Thus, a volume is created that stores all realizations of the voltages in 2D space for every input conductivity,κ. To create the κ-volume, the 2D realizations, or slices, of the data set are stacked in to a 3D volume. The Z-axis is associated with κ, in contrast to the (x, y, voltages) axes used for the displacement mapping of mean. For this particular data set, using the Z-axis to represent realizations results in a volume with 10,000 slices in the κ direction. Because of the large number of realizations, the κ-volume is very large. In order to facilitate the use of an existing graphics processing unit (GPU) visualization system [52], the volume has been subsampled to 512 slices. Thus, the κ-volume is really the mean of roughly 20 voltages at each voxel.

Because this would lead to a very large data set, the volume had been subsampled to 512 slices, computing the mean of roughly 20 voltages for every voxel (or volumetric pixel). This scalar volume can be visualized using direct volume rendering or isosurface raycasting. Figure 4.11 shows the κ-volume superimposed onto the mean/variance height field.

Once the κ-volume has been constructed, it can be visualized using direct volume



(a) Mean as height.



(b) Standard deviation is colormapped.

Figure 4.10. The (x, y, potential) space. Mean and standard deviation are color and displacement mapped into the (x, y, potentials) space.

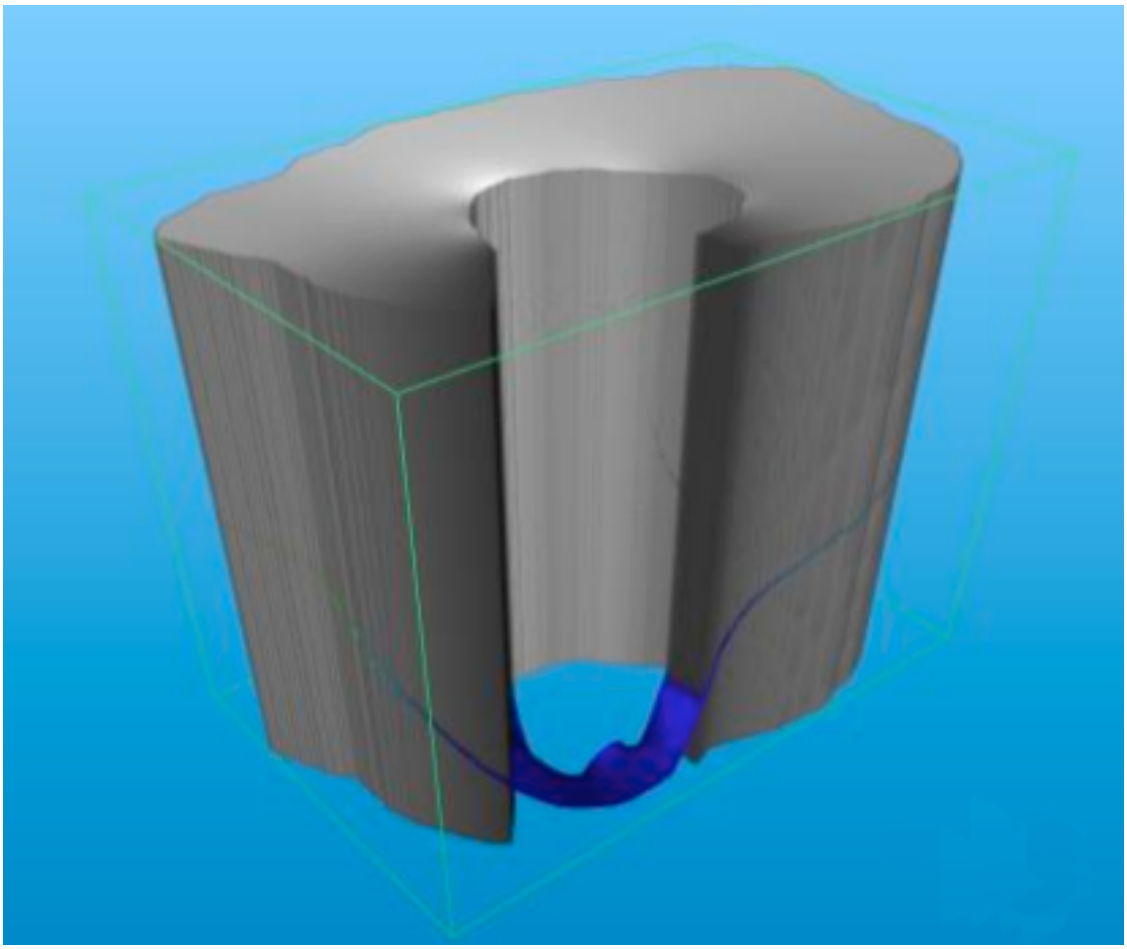
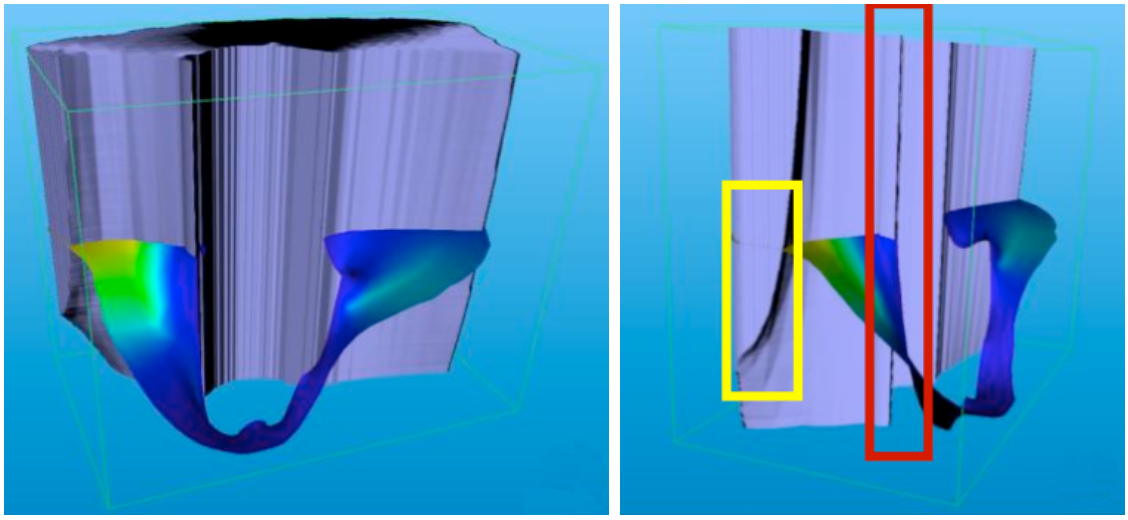


Figure 4.11. Direct volume rendering of the κ-volume.

rendering. In this technique, the color of every voxel is defined using a transfer function. Figure 4.11 shows the κ-volume rendering using a greyscale transfer function. Although direct volume rendering allows the user to see the entire volume, this technique is problematic for these data because some voltages occlude each other. As shown in the figure, it is difficult to distinguish between voltages at each slice Moreover, if the transparency is reduced, the voltages are indistinguishable. Volume rendering these data does not provide the hoped for insight and thus, alternative visualization techniques must be employed.

Isosurface raycasting is another such technique for exploring the κ-space. This method has the advantage in that the desired isovalue that defines the value of interest must be defined by the user, and thus the volume can be explored by changing this isovalue. Figure 4.12 shows two different isosurfaces of the potential data generated by changing



(a) (b)

Figure 4.12. Isosurfaces of the voltages in κ-volume. Bends in the isosurfaces indicate high dependence on the input conductivity (κ) (highlighted by the yellow bow), while straight isosurfaces (red box) indicate stochastic independence.

the isovalue. For this data set, the structure of the isosurface rather than the actual isovalue is of interest. An isosurface that falls straight down through κ-space (red box in Figure 4.12(b)) indicates that the potential at this point in 2D space does not vary when changing the input conductivity; thus, it is considered independent at that point. A bending isosurface (yellow box in Figure 4.12(b)) indicates areas of high dependence on the input conductivity. As can be seen in comparison with Figures 4.3 and 4.10, the bend in the isosurface corresponds to an area of high variance; however, in this visualization, the actual reason for this high variance becomes clear. Even in the region with the highest variance, only small κs (lowest values on the κ-axis) result in potential changes. This relationship becomes evident only in the isosurface visualization.

4.5.3 Streamlines and Particle Tracing

Even though the isosurface rendering qualitatively depicts stochastic dependency and presents a quick method to get a global overview of potential distribution, a more quantitative measurement is also desirable. For this pursuit, the gradient volume of the outcome potentials is considered. That is, the change in potentials, rather than the input connectivity or the potentials themselves, is used as the data space. This highlights changes in the data space, which in the case of this particular data, indicates locations

in which perturbations in input conductivities highly influence the outcome potential.

Viewing the data in this manner allows streamline tracing to be used to visualize potential gradients. Figure 4.13 shows streamline visualizations of the electrocardial potential data. The physical interpretation of these streamlines is as follows. If a streamline's course is primarily horizontal, meaning that its tangents (i.e., the gradient in the potential field) lie mainly in x and y planes, not into κ-direction, the points along the streamline have a potential that is locally independent of the realizations of the input conductivities. This, in turn, means that a streamline pointing upwards- or downwards indicates a high dependency on κ. In fact, the length of the streamline influenced by the gradient-magnitude is a quantitative measure of these dependencies.

Particle tracing can be used to further emphasize the gradient-magnitude; particles are seeded into the gradient volume, as shown in Figure 4.14. These particles not only convey the direction of the gradient but also allow gradient magnitude to be derived easily from their speed.

4.6 Conclusion

This chapter presents an exploration of data with uncertainty information using preexisting visualization techniques. Two- and three-dimensional methods were discussed, including how to apply these techniques to this specific type of data and the effectiveness of each technique to bring out particular characteristics of the data. While much of the complex structure of the data set can be highlighted by using particular visualization methods, no single approach presented all of the information afforded by the data. As the sources of uncertainty data become more diverse, techniques for understanding complicated data sets will be required. Many of the traditional visualization techniques will be applicable, however, understanding the meaning of the generated images becomes more difficult and in many cases more than one technique will be required.

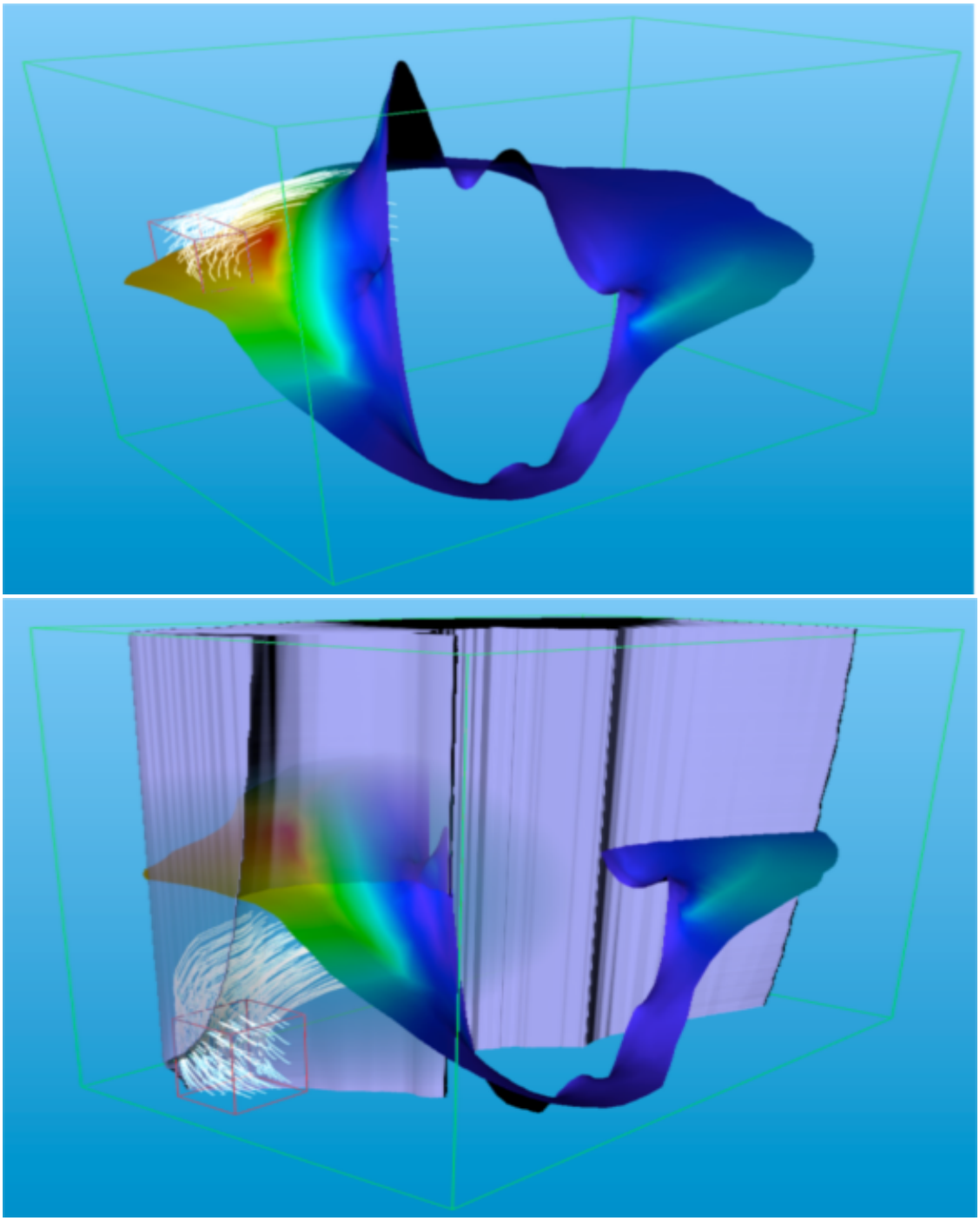
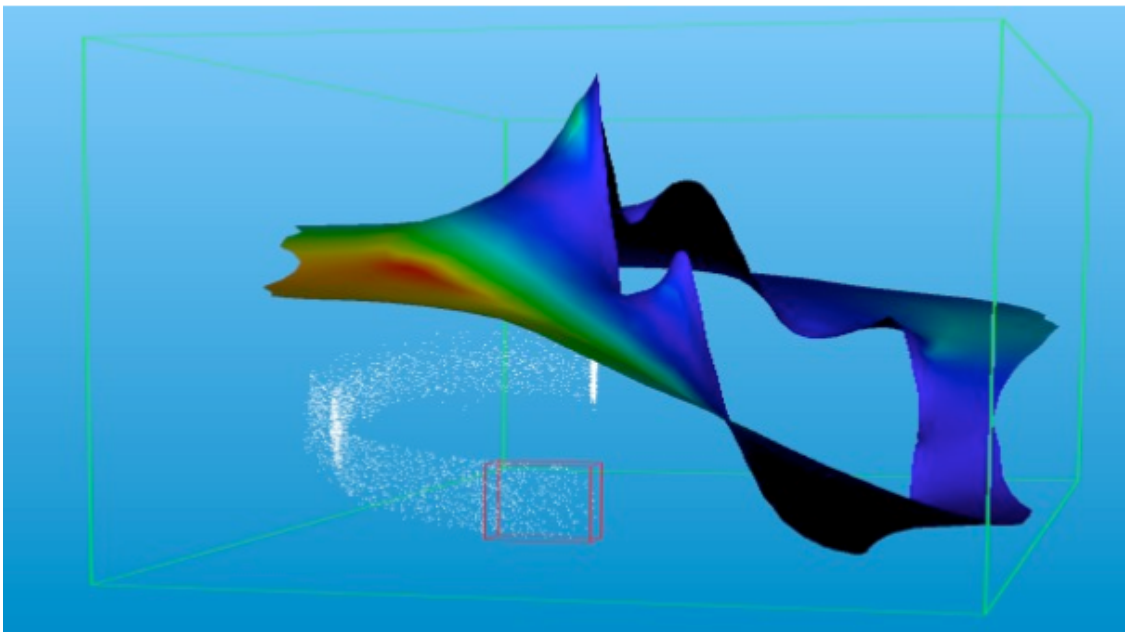
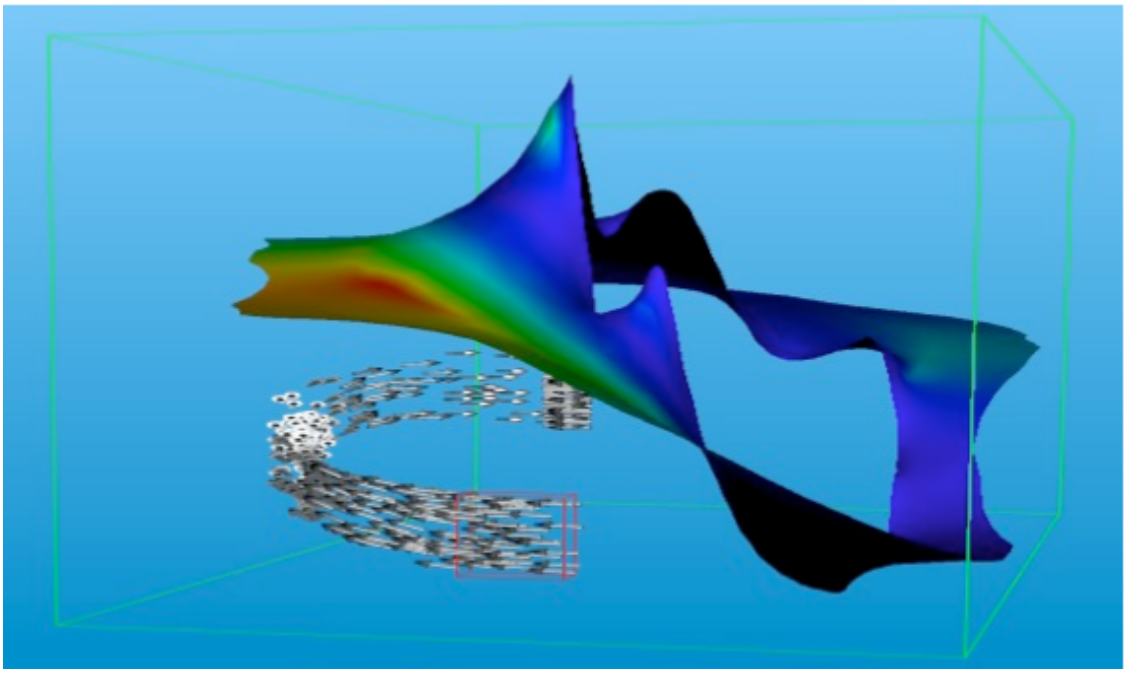


Figure 4.13. Streamlines. The direction and slope of the streamlines indicate dependency of the potential on the conductivity.



(a) Particle tracing, simple points as particles.



(b) Arrow glyphs are used as particles.

Figure 4.14. Particle tracing. Simple particles help show the gradient of the volume, however, arrow glyphs are easier to understand in still images in that they add more information.

CHAPTER 5

ENSEMBLE-VIS

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. In this chapter, *Ensemble-Vis* is presented that combines a collection of overview and statistical displays linked with 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, combining multiple linked statistical displays yields a clearer presentation of the data and facilitates a greater level of visual data analysis. The Ensemble-Vis framework is demonstrated using driving problems from climate modeling and meteorology and is easily generalized to other fields.

5.1 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 chapter, a general approach to the visual analysis of ensemble data is presented 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, this framework provides a cohesive view of the ensemble that

permits analysis at multiple scales from high-level abstraction to the direct display of data values. This work is developed in a component-based framework, allowing it to be easily adapted to new applications and domains.

5.1.1 Motivation

A 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 experiencing freezing rain in a given area over a given period of time, and qualitative aspects, like 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 element common among many applications using ensembles is the goal stated above: to predict and quantify the range of outcomes from a range of initial conditions. Here, a data analysis framework is presented that allows domain scientists to explore and interrogate an ensemble both visually and numerically in order to reason about those outcomes.

5.1.2 Driving Problems

This work focuses on two driving problems: short-term weather forecasting and long-term climate modeling. While this approach is informed by some of the specific needs of meteorology and climatology, in particular, the applications described in this section, the structure and algorithms presented are sufficiently general to apply to analysis problems involving ensemble data across a wide variety of fields.

5.1.2.1 Weather Forecasting

Rather than relying on singular, deterministic models [37], meteorologists increasingly turn to probabilistic data sets to forecast the weather. Uncertainties and errors exist in every weather simulation due to the chaotic nature of the atmosphere as well as our inability to accurately measure its exact state at any specific time. Moreover, biased or inaccurate numerical weather prediction models often lead 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 [37].

The weather forecasting data used is 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. This data set is obtained from the National Centers for Environmental Prediction's (NCEP) Environmental Modeling Center and Short-Range Ensemble Forecasting Project [3].

5.1.2.2 Climate Modeling

In contrast to the goal of meteorologists 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 (namely, the whole planet) instead of being restricted to a small geographical region of interest [26]. 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.

5.1.3 Ensemble Data Sets

An *ensemble data set* is defined 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 sample point.

5.1.3.1 Ensembles and Uncertainty

Ensemble data sets are frequently 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.

The multiple values for each variable at each point in an ensemble can be interpreted 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.

5.1.3.2 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, as shown in Table 5.1. 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 5.1, the result is visual chaos and confusion 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

Table 5.1. Four numerical weather prediction models, as well as the parametric perturbations run on each model. For every predicted variable in the simulation, there are 21 data values.

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

most 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 only four.

5.2 Related Work

Because of the complexity of the data, 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.

5.2.1 Visualization of Climate and Weather Data

Current state-of-the-art techniques for displaying weather and climate data sets include software systems such as Vis5D [43] and SimEnvVis [62]. 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 five 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 multirun experiment results.

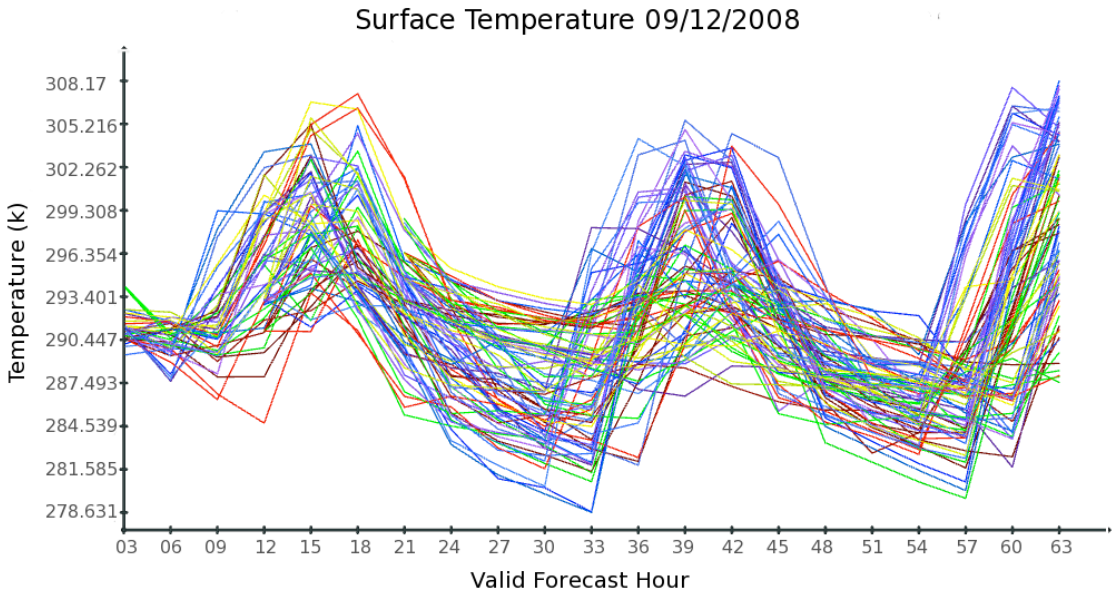


Figure 5.1. An example of the complexity of an ensemble data set. In the graph, each of the 21 models for a single weather station are charted across all valid forecast hours. The line colors relate the 4 colored models in Table 5.1. While this graph reduces the overall data by selecting only 1 out of 23685 possible weather stations, it is still too visually cluttered to assist in data analysis beyond giving a notion of the general outcome of the simulation.

The main distinction between these previous efforts and the approach presented here is the stress on understanding the uncertainty available from the data by providing visualization tools that emphasize the probabilistic characteristics of ensemble data. Overviews are provided to initially drive the analysis of ensemble data and highlight changes in uncertainty. This framework 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. The focus is 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.

5.2.1.1 Multidimensional Data Visualization

The type of data addressed by this work 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 [18]. Multivariate correlation in the spatial domain is an often used approach for reducing the complexity of the task of data understanding [6], as is reducing the data to a hierarchical form that is conducive to 2D plots [61]. Likewise, the visualization of multidimensional data is challenging and often involves dimension reduction and user interaction through focusing and linking. Buja et al. have provided a taxonomy of such techniques that can assist in determining an appropriate approach [17].

Among the most relevant work using ensemble type data treats observations in terms of PDFs describing the multiple values at each location and each point in time [56]. Three approaches to visualizing this type of data are proposed: a parametric approach that summarizes the PDFs using statistical summaries and visualizes them using color mapping and bar glyphs; a shape descriptor that 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 this approach also uses a variety of statistical measures to describe the underlying PDF, it provides statistical views from a number of summarization standpoints in a single framework, allowing the user to actively 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 [6, 71]. 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 over areas of interest [8], or querying [76]. One interesting technique uses smooth brushing to select data subsets and then visualize the statistical characteristics of that subset [81]. 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 [23, 68]. The resulting collection of views provides for complex investigation of the data by allowing the user to drive the data analysis.

5.2.2 Uncertainty Visualization

Most visualization techniques that incorporate uncertainty information treat uncertainty like an unknown or "fuzzy" quantity; [65] is a survey of such techniques. These methods employ the meaning of the word *uncertainty* to create the interpretation of *uncertainty* or *unknown* to indicate areas in a visualization with less confidence, greater error, or high variation. Ironically, while blurring or fuzzing a visualization accurately indicates the lowered confidence in that data, it does not lead to more informed decision making. On the contrary, it obfuscates the information that leads to the measure of uncertainty. Because it obscures rather than elucidates the quantitative measures leading to the uncertain classification, such a solution to the problem of adding qualitative information to visualization misses important information.

Much of this work is motivated by the growing need for uncertainty information in visualizations [46]. 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 [59, 65], 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 [29], shown in Figure 5.2; using flow visualization techniques to show the mean and standard deviation of wind and ocean currents [84], Figure 5.3; uncertainty contours to show variations in models predicting ocean dynamic topography [64], Figure 5.4; and expressing the quality of variables in multivariate tabulated data using information visualization techniques such as parallel coordinates and star glyphs [85], Figure 5.5.

5.2.2.1 Comparison Techniques

Often, uncertainty describes a comparison that can most clearly be understood visually, such as the difference between surfaces generated using different techniques, or a range of values that a surface might fall in. A simple approach to the visualization of this type of information is a side-by-side comparison of data sets. However, this

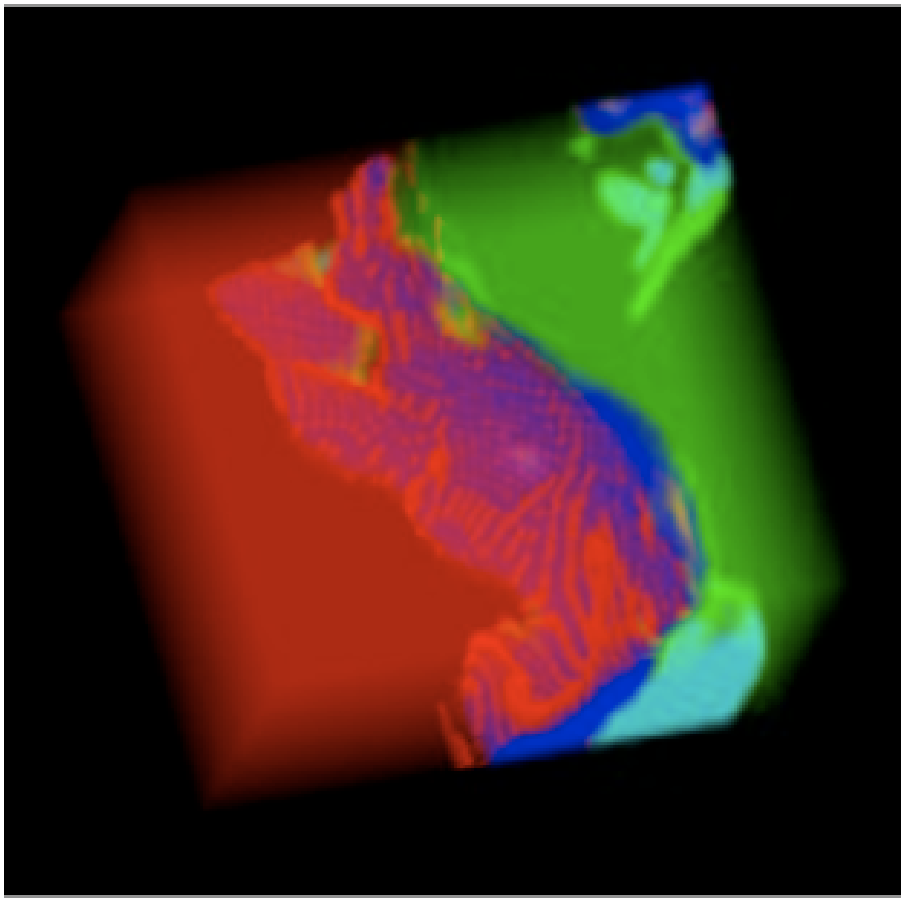


Figure 5.2. Volume rendering of ocean salinity [29].

approach may not clearly manifest subtle differences when the data are nearly the same, and it becomes harder to perform this comparison as the visualization becomes more complicated. Another simple approach is to overlay the data to be compared [47]. With this technique, the addition of transparency or wire frame can produce a concise, direct comparison of the data sets. A similar approach uses difference images to display areas of variation [83]. These approaches are less effective, however, when the uncertainty can be categorized as more of a range of values rather than just two distinct ones. In such cases, a surface sweep, known as a fat surface [65], can be used to indicate all possible values. Another approach is the integration of isosurface and volume rendering. Here, an opaque isosurface can be used to indicate the most likely value, and a transparent volume rendering surrounding the isosurface can indicate the range of possible values [46]. Uncertainty information for large collections of aggregated data can be presented using

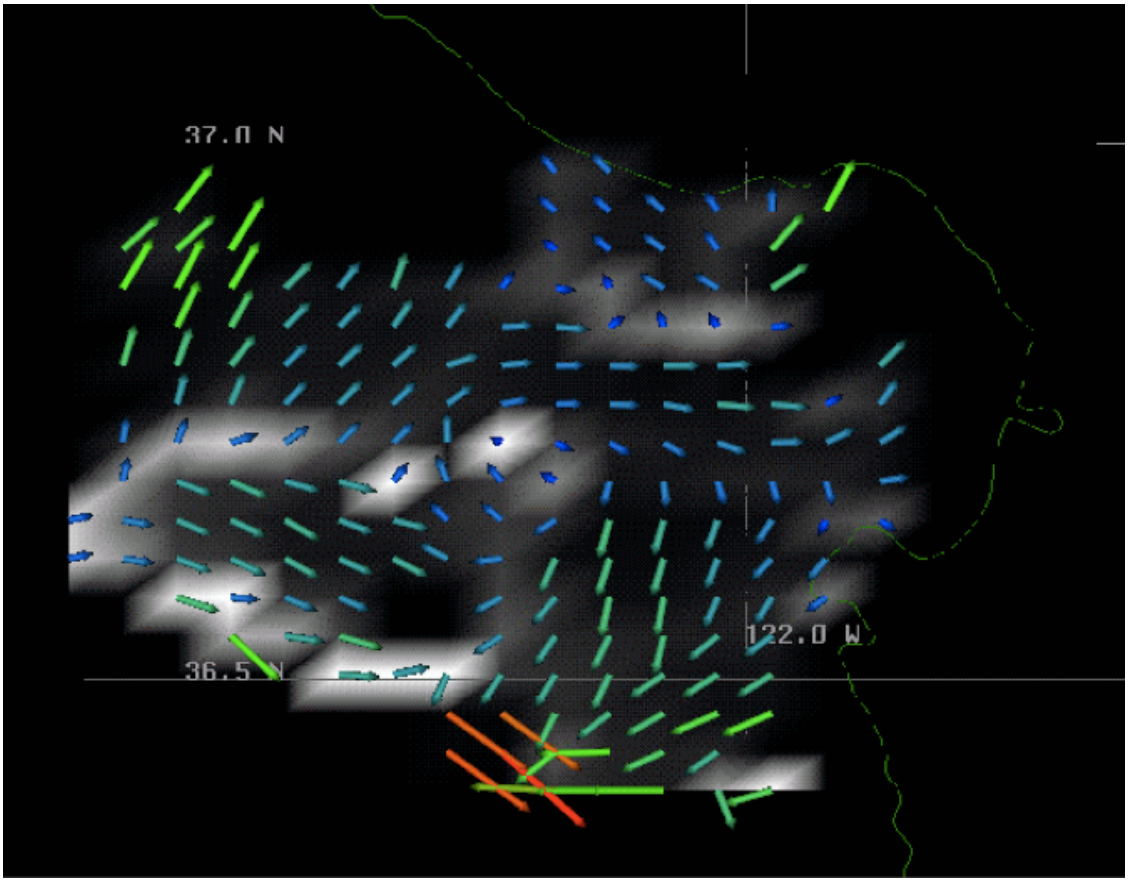


Figure 5.3. Flow visualizations of currents. The mean and standard deviations of wind and ocean currents are shown using glyphs [84].

hierarchical parallel coordinates [35]. Finally, bounded uncertainty, while not effectively visualized in 3D, can be expressed through the ambiguation of boundaries and edges of pie charts, error bars, and other 2D abstract graphs [63].

5.2.2.2 Attribute Modification

Another standard method to visualize uncertainty involves mapping it to free variables in the rendering equation or modifying the visual attributes of the data. Such methods include modifying the bidirectional reflectance function (BRDF) to change surface reflectance, mapping uncertainty to color through a 2D transfer function, or pseudo-coloring using a look-up table [65]. This technique has been used as a means for conveying uncertainty in the areas of volume rendering [30] and isosurfacing [47, 70], and is often combined with other uncertainty visualization methods. An example technique colormaps flowline

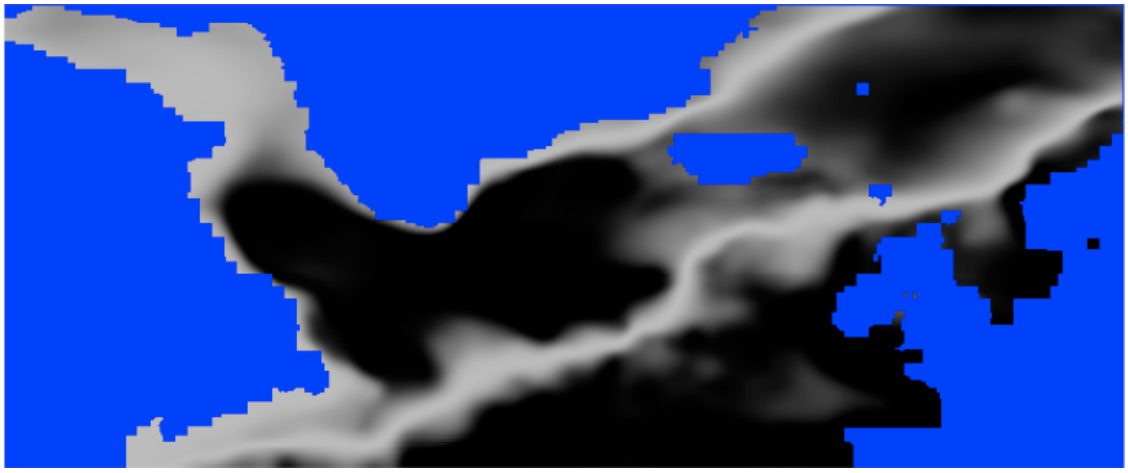


Figure 5.4. Uncertainty contours. Variations in models predicting ocean dynamic topography [64] are displayed using contours.

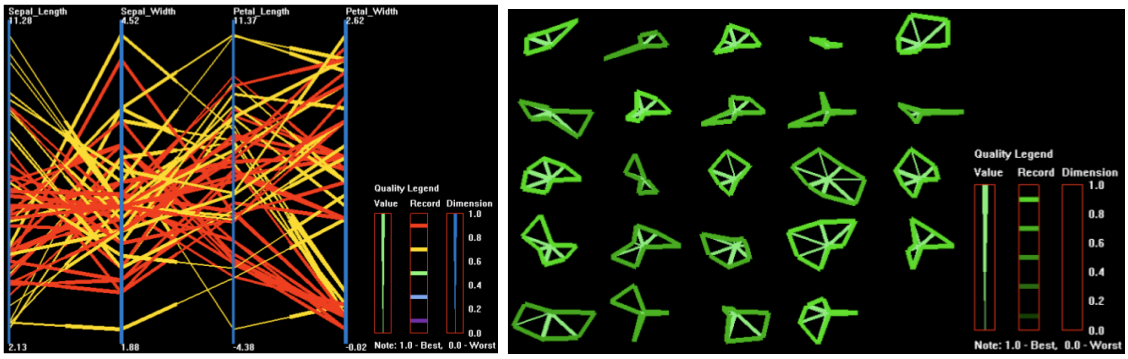


Figure 5.5. Uncertainty encoded using parallel coordinates and star glyphs [85].

curvature onto volume rendered surfaces, highlighting areas in which small changes in isovalue lead to large changes in isosurface orientation and thus indicating areas where the isosurface is a poor representation of material boundary [51]. Texture can be used similarly to convey uncertainty and is also often modified by opacity, hue, or texture irregularities [44].

5.2.2.3 Glyphs

Glyphs are symbols used in visualization to signify data through parameters such as location, size, shape, orientation, and color. Because of the multivariate nature of glyphs, they can be used in visualization to map uncertainty to a free parameter. One such

approach uses glyphs to present the distribution of multivariate aggregated data over a range of values [21]. These glyphs show the average, standard deviation, and distribution of three attributes of the data set.

An approach that modifies attributes of glyphs already present in the visualization is presented as a procedural generation algorithm [19]. In this work, the data is sampled on a regular grid and the size, color, and placement of glyphs are taken directly from the data samples. The uncertainty is then used to distort the glyphs so that glyphs with low uncertainty are very sharp, with the sharpness level decreasing as the uncertainty level increases. This distortion provides a clear indication of uncertainty and error while not placing heavy emphasis on areas of high uncertainty.

Because not all data is visualized effectively using glyphs, the addition of glyphs to convey only uncertainty information is often a preferable approach. A specific example is the UISURF system [47], which visually compares isosurfaces and the algorithms used to generate them. In this system, glyphs are used to express positional and volumetric differences between isosurfaces by encoding the magnitude of the differences in the size of the glyphs. Similarly, line, arrow, and spherical glyphs can be used to depict uncertainty in radiosity solutions, interpolation schemes, vector fields, flow solvers, and animations through variation of placement, magnitude, radii, and orientation [65, 84].

5.2.2.4 Image Discontinuity

Uncertainty visualization often relies on the human visual system's ability to quickly pick up an image's discontinuities and to interpret these discontinuities as areas with distinct data characteristics. Techniques that utilize discontinuities rely on surface roughness, blurring, oscillations [19, 39, 83], depth shaded holes, noise, and texture [30], as well as on the translation, scaling, rotation, warping, and distortion of geometry already used to visualize the data [65], to visualize uncertainty.

5.3 The Ensemble-Vis Framework

In this section, a framework for the visualization and analysis of ensemble data is discussed that emphasizes the probabilistic nature of the data. Changes in uncertainty are highlighted 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. The algorithms are presented in two prototypical systems, the SREF Weather Explorer, and the ViSUS Climate Data application.

5.3.1 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 5.6. 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 *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, this work incorporates 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 that are characteristically biased or unreliably predict specific regions across the spatial domain.

5.3.2 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, the visualization of these data does 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,

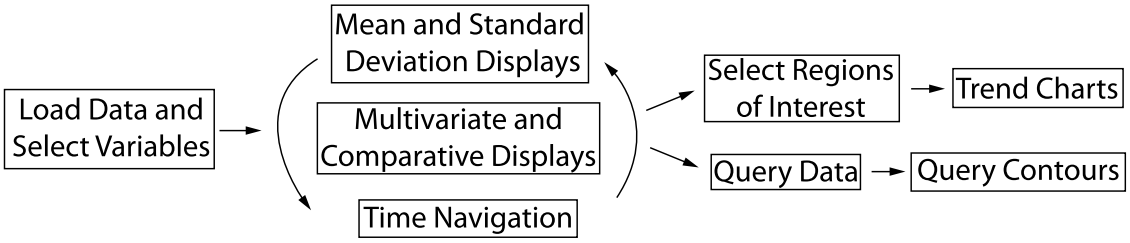


Figure 5.6. The Ensemble-Vis work flow. An organization of the typical flow of data analysis through the Ensemble-Vis framework. Users first choose a data set and one or more variables to display. They are then provided with mean and standard deviation views, comparative, and multivariate visualizations, 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.

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 presented here, the only requirement is that the data repository be 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.

5.3.3 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 5.7, top) show the behavior of one variable over space at one time step. The time navigation summary views (Figure 5.7, bottom) show the same variable at lower spatial resolution over several time steps at once or through an animation.

5.3.3.1 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

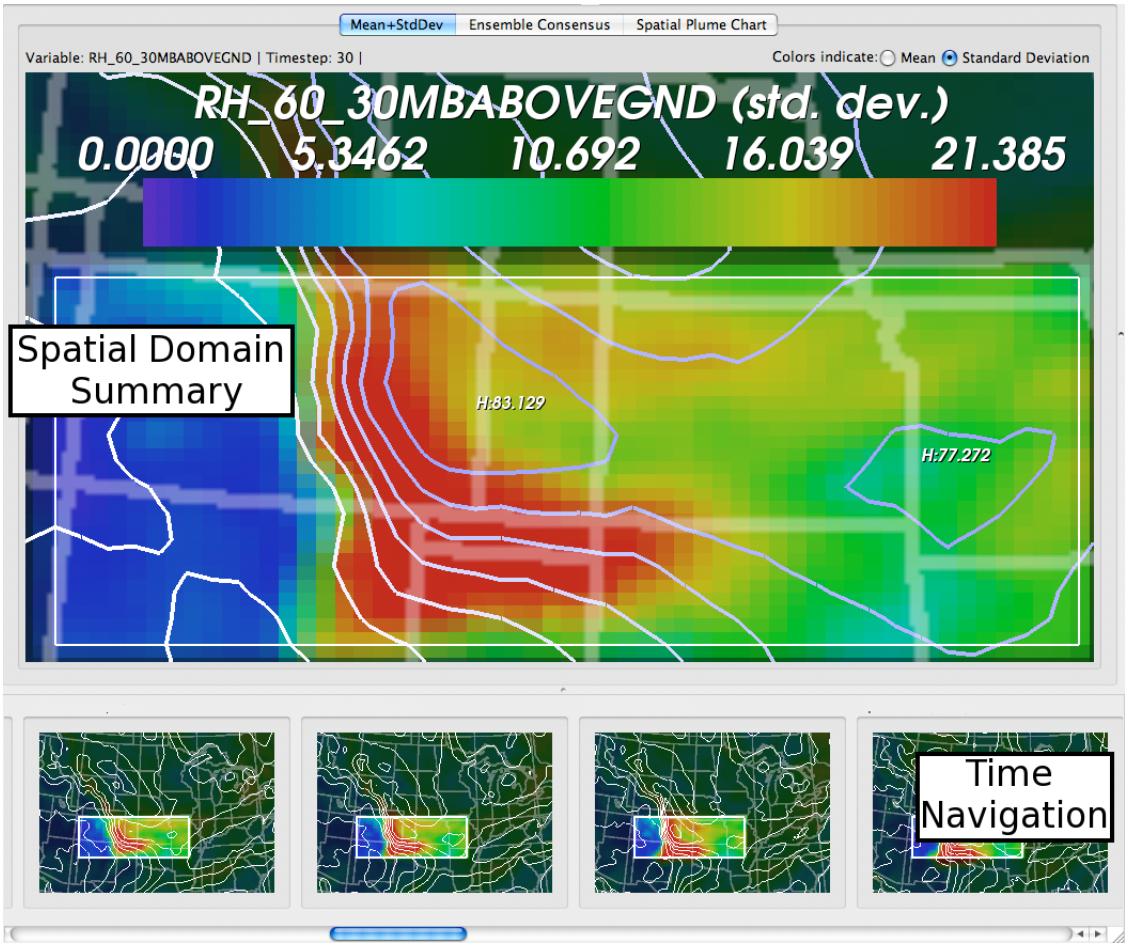


Figure 5.7. Spatial and time domain overviews. Two representations to summarize each variable in the ensemble are combined. A high-resolution spatial display (top) shows mean, standard deviation, and local minima and maxima for a given time step. An arrangement of lower-resolution multiples into a filmstrip (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.

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: precise scalar values for both mean and standard deviation need not be conveyed. An approximate sense of the value of the mean plus an indication of high or low standard deviation is all that is required.

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 data set. 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.

The default display of the variable mean uses color and the standard deviation uses overlaid contours (Figure 5.8). Although the rainbow color map is generally a poor choice for scientific visualization [15], 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 (the reflectivity of the sun's radiation) or probability of precipitation, the user can select to use a different sequential color map, examples of which can be seen in Figure 5.9. 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 5.10.

A heightfield can also display standard deviation instead of contours. This is particularly effective when displaying 2D data projected onto the globe, as is common in climate simulations (Figure 5.11), 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 [74]. Second, many ensembles do not have enough members to support more sophisticated, precise characterizations.

5.3.3.2 Time Navigation Summary Views

In addition to the spatial summary view, which shows a high-resolution overview of a single time step, time navigation summary views are also provided that 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 steps in time. As shown in Figure 5.12, 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

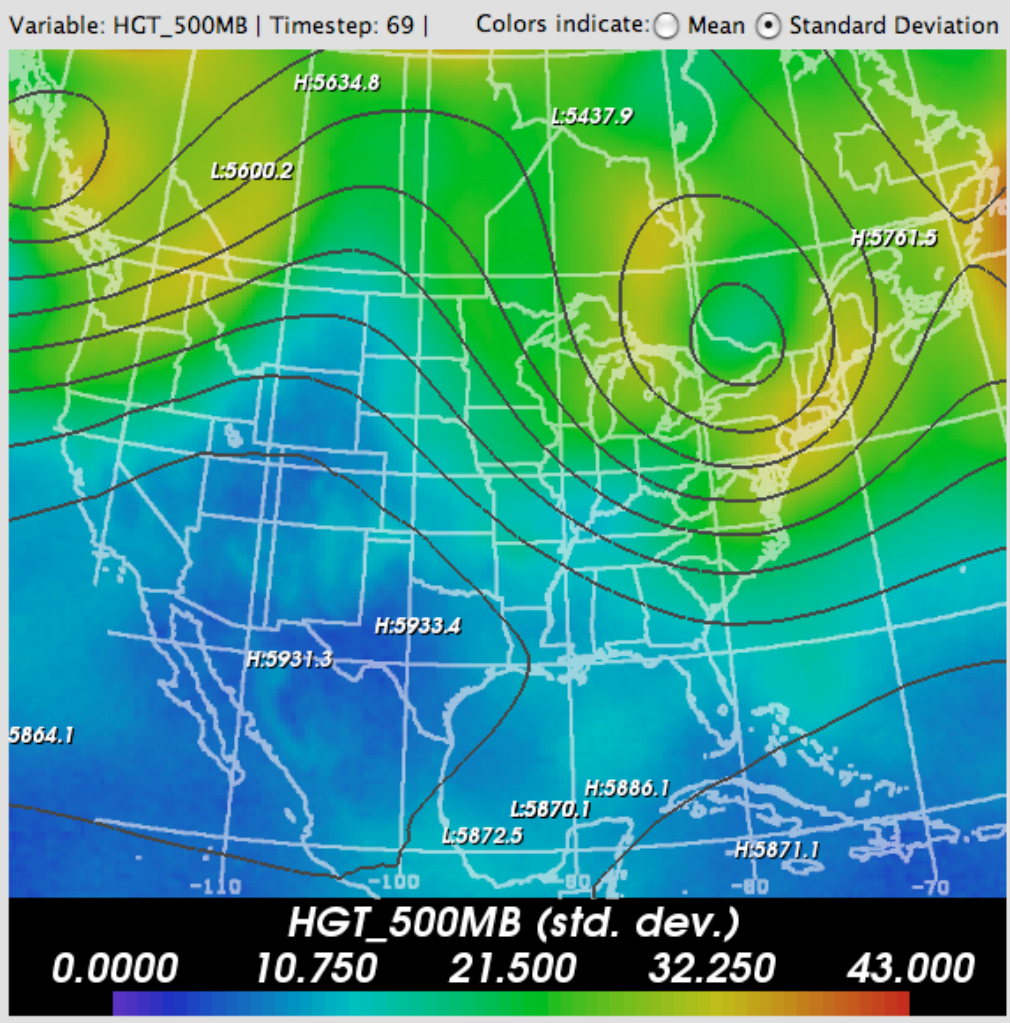


Figure 5.8. Mean and standard deviation display. We illustrate mean and standard deviation simultaneously using color plus overlaid contours

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, 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.

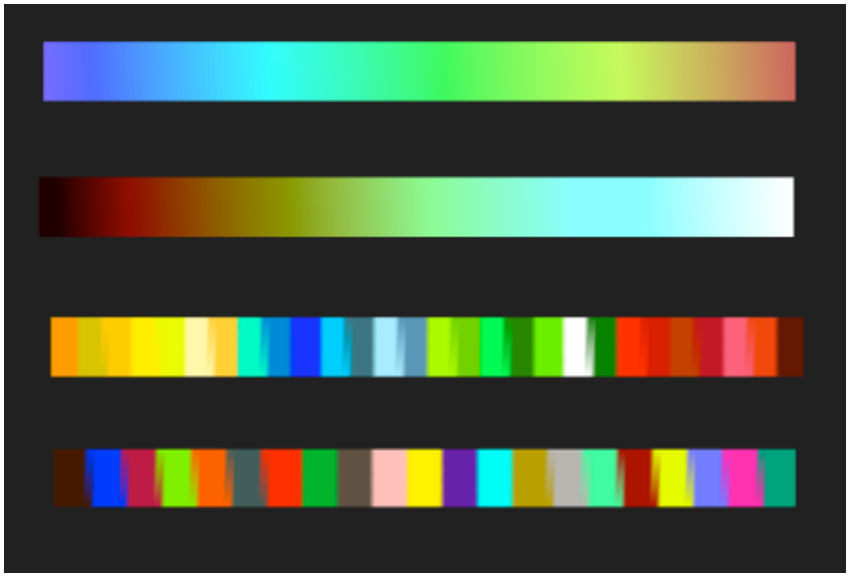


Figure 5.9. Colormaps used within Ensemble-Vis. A subdued rainbow color map is used and a sequential low to high map for scalar variables and two categorical color maps for labeling.

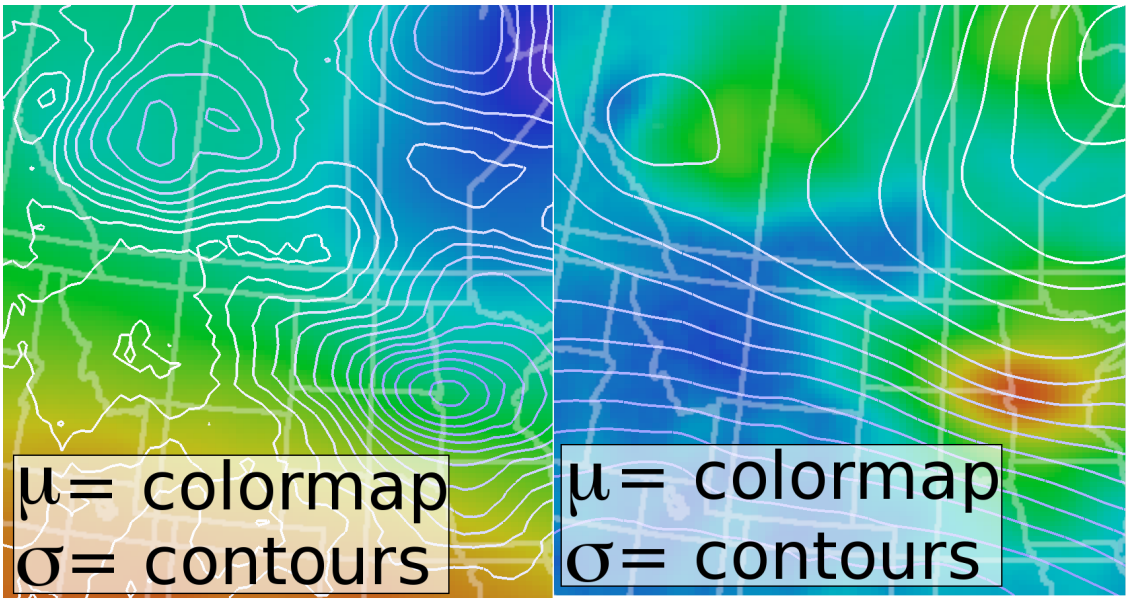


Figure 5.10. Toggle interface between colormaps and contours. The user can toggle the assignment of mean and standard deviation to colors and contours, respectively (left), or the reverse (right). Both images show the same region of the data.

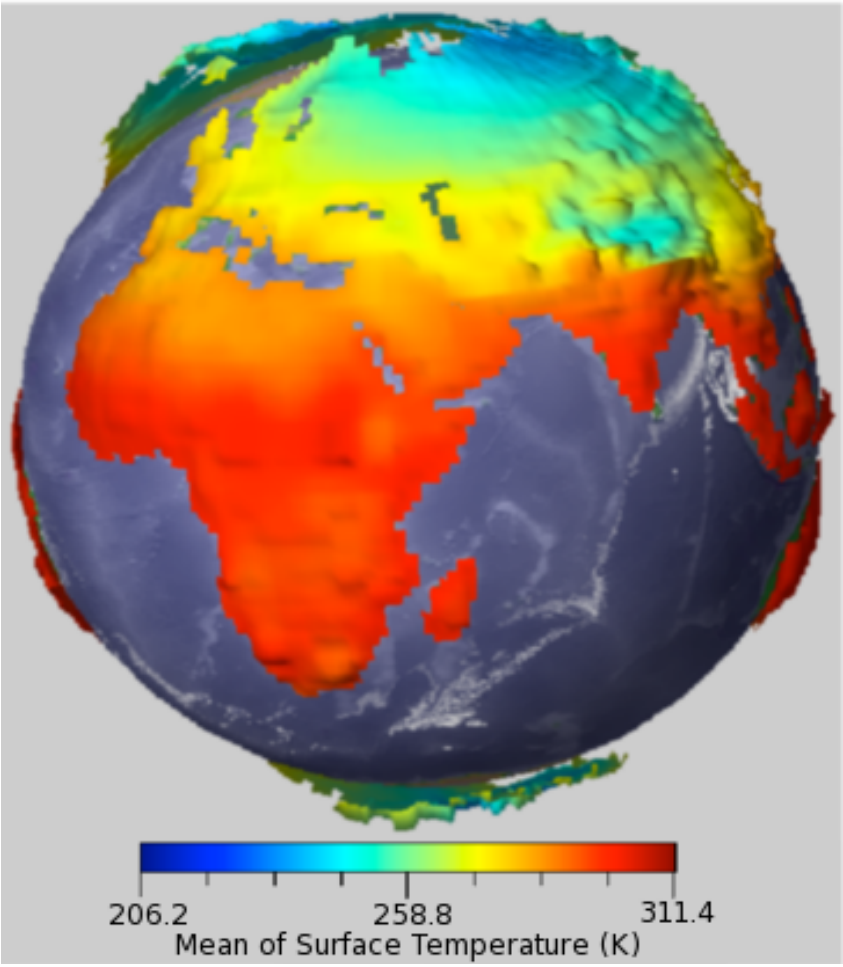


Figure 5.11. Height mapped onto the earth globe. Height is another channel available for data presentation. Here, standard deviation is displayed as a heightfield, and mean is shown through color. Especially visible on the silhouettes of the globe, highly displaced regions indicate high uncertainty.

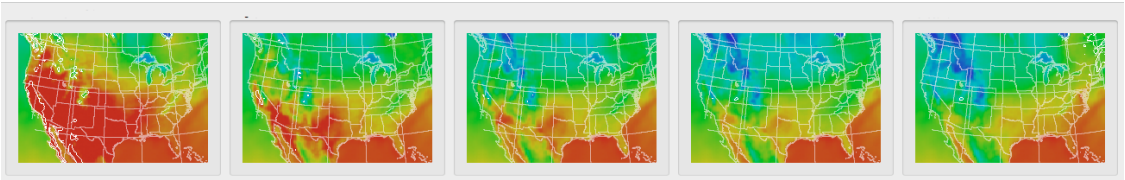


Figure 5.12. The filmstrip summary view. Each frame in the filmstrip shows a single time step from the ensemble. The filmstrip also displays selection information from other views to help the user maintain a sense of context.

5.3.4 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 – aggregation over space can instead be employed 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. Two such views are provided in the Ensemble-Vis framework.

5.3.4.1 Quartile Charts

A quartile trend chart (Figure 5.13) displays the minimum, maximum, 25th, 75th percentiles, and median of a particular variable in a selected region over time. These values are computed 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.

5.3.4.2 Plume Charts

A plume chart (Figure 5.14) 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), the mean of each ensemble member's values is computed 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 this 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

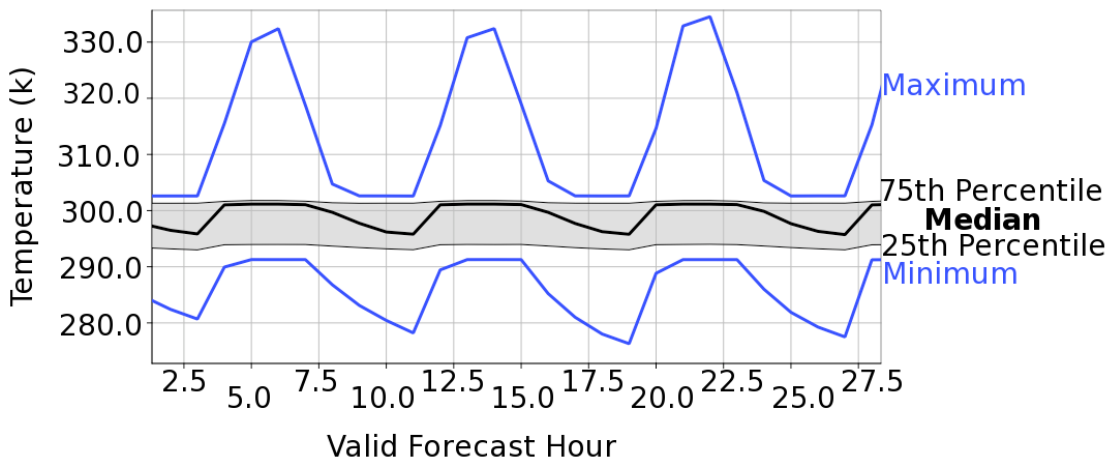


Figure 5.13. Quartile trend charts. These charts show the quartile range of the ensemble within a user-selected region. Minimum and maximum are shown in blue, the gray band shows the 25th and 75th percentiles, and the median is indicated by the thick black line.

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.

5.3.5 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 [76] constrain the visualization to the subset of data deemed interesting by the analyst and discards the rest. These sets of circumstances are referred to as *conditions*.

Once an analyst specifies a condition, as shown in the inset of Figure 5.15, 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.

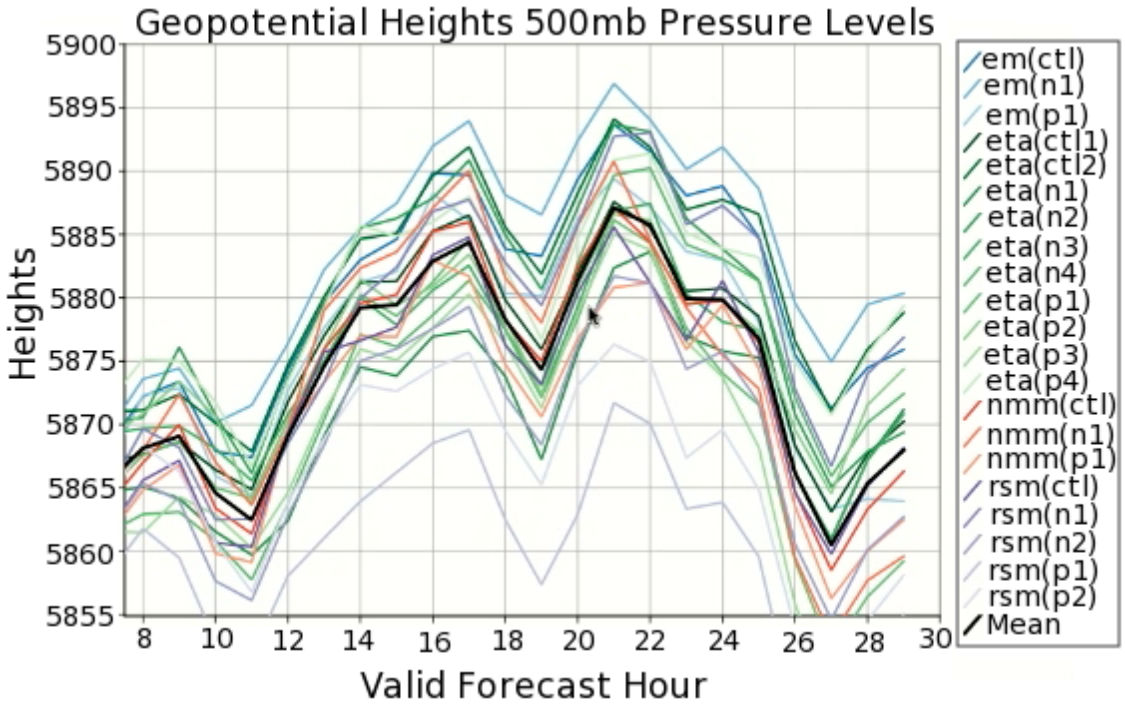


Figure 5.14. Plume trend charts. These charts show the average of each ensemble model within a user-selected region of interest. Each model type is color-coded. The thick black line shows the mean across the entire ensemble.

In this 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 a very simple dialog to specify a condition is used, there exist a wide variety of query languages and mechanisms for visual query specification. This 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.

5.3.6 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. These slices are displayed using multiple 2D views in the same window. The data are displayed using a common colormap in a single window.

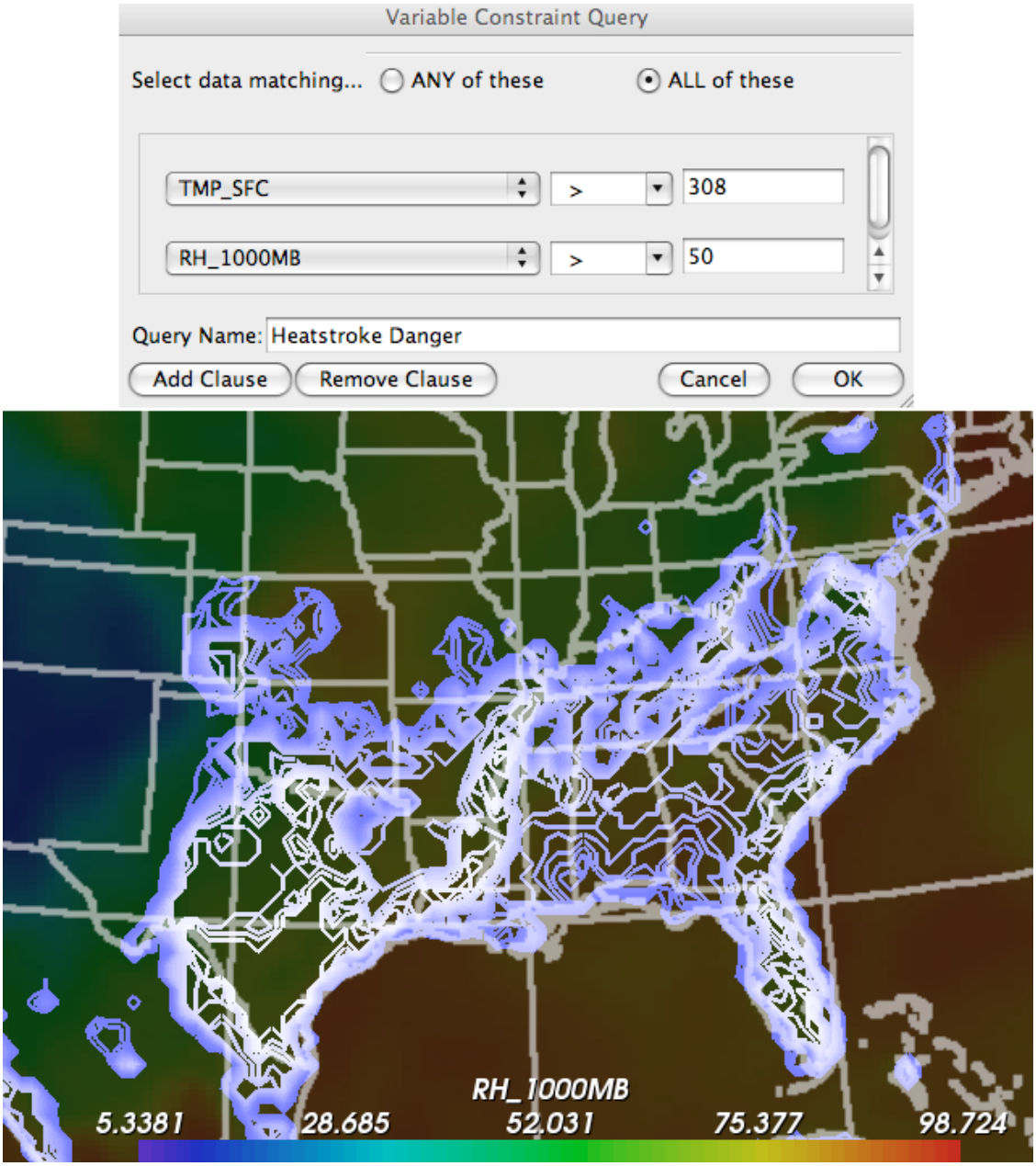


Figure 5.15. Condition queries. The condition query view shows the probability that a given set of conditions will occur as a set of nested contours. Contour values are the fraction of the ensemble that predicts that the condition will be satisfied. In this figure, we see a query for heatstroke danger (defined as relative humidity above 50% and temperatures above 95° F) and the resulting visualization.y

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 5.16 shows three elevations that add to the cloudiness across the globe.

5.3.7 Spaghetti Plots

A *spaghetti plot* [28], 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 all of the ensemble members are in agreement, that is they all predict relatively similarly, as shown in Figure 5.17, left, the contours will fall into a coherent bundle. When minor variation exists, a few outliers may diverge from the bundle (Figure 5.17, 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, colors are assigned to the contours in a spaghetti plot so that contours that arise from the same simulation model will have similar colors. The user is also allowed 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.

5.3.8 Coordination Between Views

The various views in this 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 the retrieval of new data from storage. Since these operations take several seconds to complete their execution is deferred until the user specifically requests them.

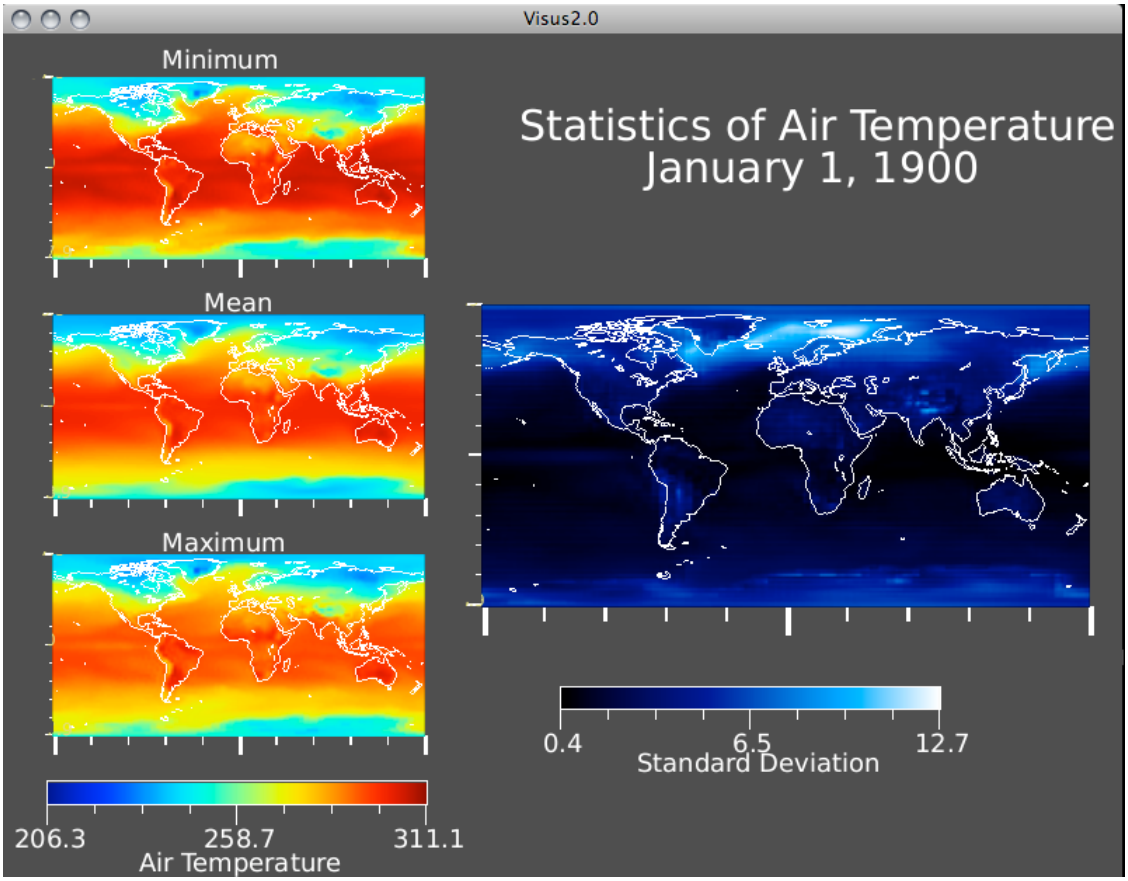


Figure 5.16. Multivariate display. In this figure, various statistics for surface temperature are displayed.

5.3.9 Clustering

Clustering techniques offer another method for reducing the size of data. Clustering groups the data into similar regions based on a particular measure of similarity. Such algorithms, common in the fields of machine learning and data mining, can readily be applied here. The K-Means/Mediods algorithm [11] is a technique for grouping data based on their distance to a middlemost representative of each cluster. First, the user specifies the number of clusters in which they are interested. The algorithm then iterates through each data point, finding the closest cluster, and correspondingly labeling each point. It then updates the central data point of that cluster. These steps are repeated until the cluster groupings are stable and do not change. The distinction between the K-Means and K-Mediods algorithm is simple: either the values representing the central location of each cluster are actual data values or the values are simply the mean of all the data grouped

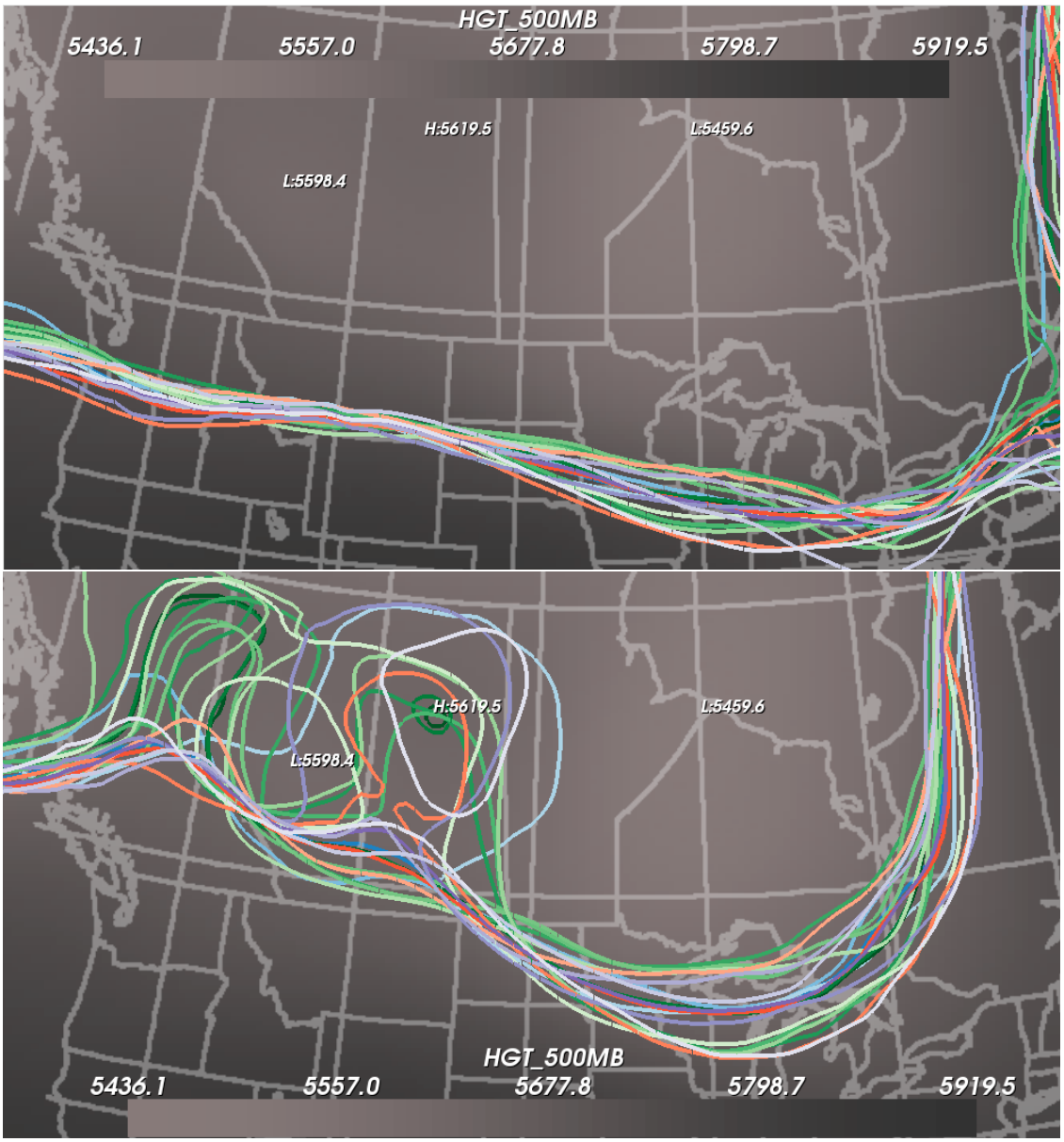


Figure 5.17. Spaghetti plots. A spaghetti plot displays a single isocontour from each ensemble member in order to allow examination of differences across space. (Left) When the members are in agreement, the contours form coherent bundles. (Right) When ensemble members disagree, as in the in upper left region of the image, outliers diverge from the main bundle.

into that cluster. For the climate and weather data, the main concern is not only with the value of the data we are clustering, but also the spatial locations of those data values. Thus, two measures of distance when labeling a data point are combined. The spatial measure of distance uses Euclidean distance based on the grid location of the data. Since the grids are regular, the resulting clusters reduce to the Voroni diagram describing the space, as seen in Figure 5.18(a). However, if the spatial grid is irregular, the calculation of the mean location of the cluster can be skewed and a more complex estimation of the central cluster location may be required. This calculation of spatial distance is then combined with a distance measure between the actual data values used. In this example, the mean of surface temperature is used, which can be seen in Figure 5.18b. Finally, these values are combined using a weighting function, which allows the user to choose which distance metric is more important, space or data value. The results of the clustering algorithm, which can be seen in Figure 5.18c, are then used to populate the 2D Trend Charts, providing the user with an insight into regions of similarity in the data.

5.4 Implementation Details

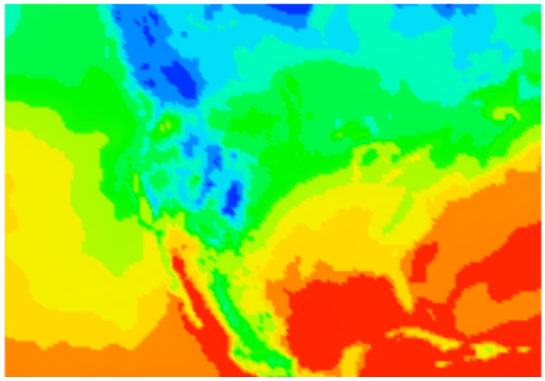
The algorithms described above have been implemented in two prototype systems for weather and climate simulation analysis. This demonstrates the flexibility of the component-based approach. In this section, the purpose and system architecture of each prototype is briefly described. Working memory is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

5.4.1 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 the system can be seen in Figure 5.19. The visualization algorithms in SREF Weather Explorer are implemented as filters in the Visualization Toolkit (VTK) [73], a well-known open-source toolkit for scientific visualization. The user interface components were implemented as Qt widgets [12].

Standard relational databases were used as the storage engine for the SREF ensemble data. This allowed the application to offload the task of storage management and thus run

(a) Results of space-based clustering.



(b) Mean of surface temperature.

(c) Resulting clusters based on spatial location and data value.

Figure 5.18. Clustering based on spatial location and mean data value. On the left is the results of clustering based only on spatial information. The middle image shows a colormap indicating the mean data values across North America for Surface Temperature, and the rightmost image shows the results of clustering based on both spatial distance measures and mean value information.

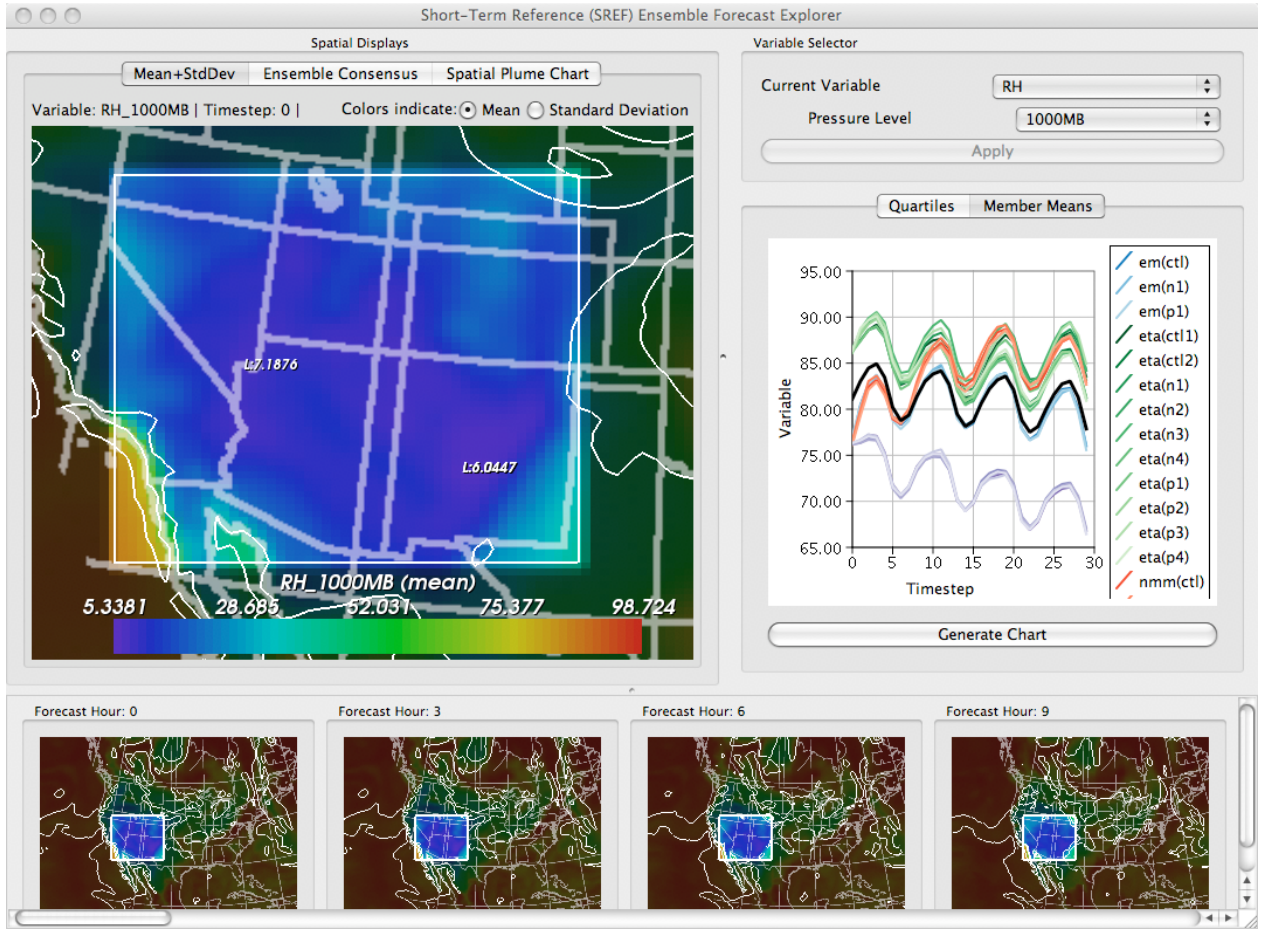


Figure 5.19. SREF weather explorer. This prototype is implemented as a set of VTK filters and can thus be easily integrated into tools deployed to domain scientists.

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, it was possible 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.

5.4.2 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, the 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 5.20.

5.5 Discussion

Visual analysis of ensemble data sets is challenging and complex on all levels. No single view or collection of views is ideal for all analyses. In this section, some of the trade-offs in this approach and the rationale behind the motivating decisions are discussed.

5.5.1 Data Challenges

The first major challenge encountered 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

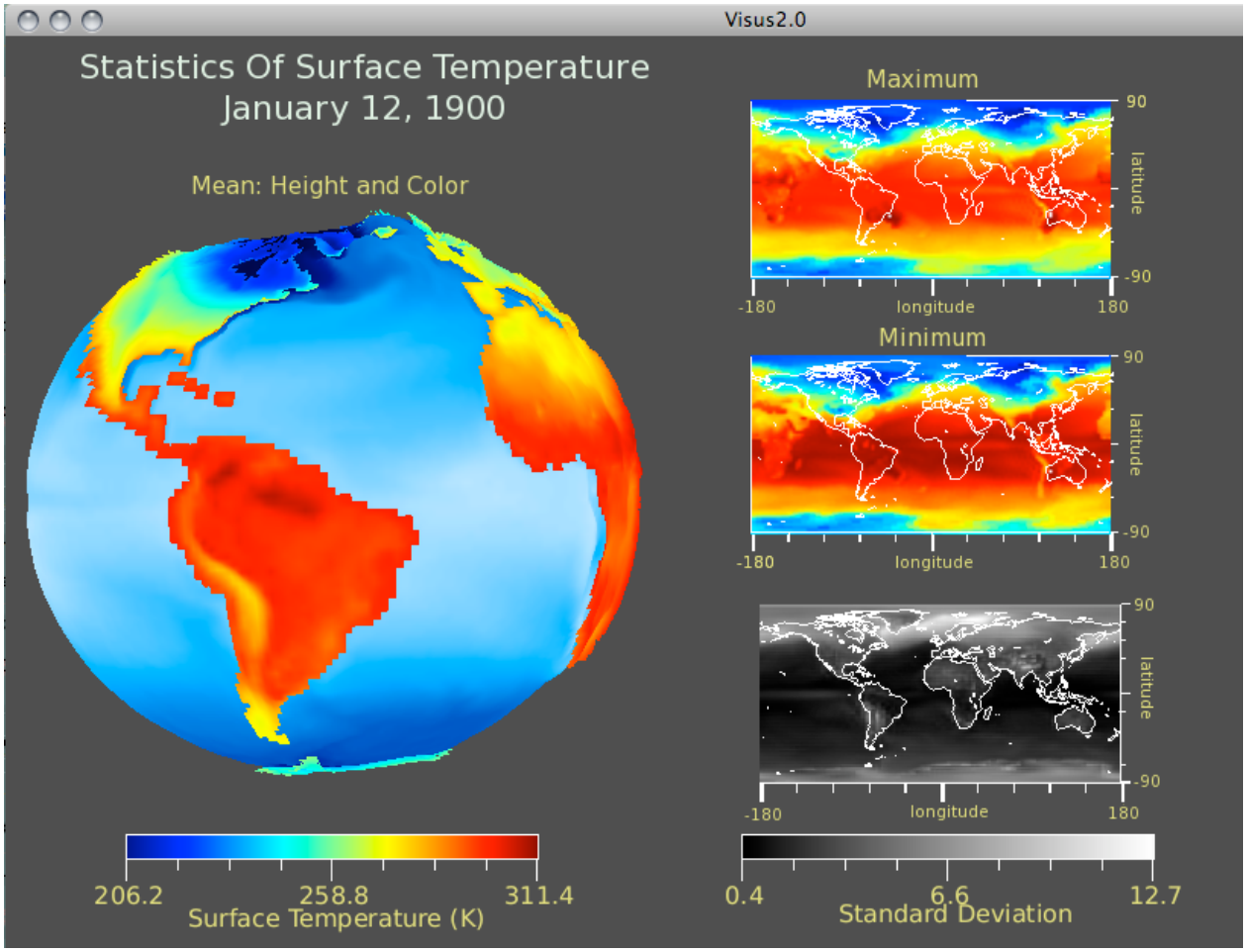


Figure 5.20. ViSUS prototype. This system is integrated into the CDAT framework used by climate scientists.

amount of multivariate data. Somehow one must 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, the 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.

5.5.2 Where Summary Statistics Break Down

Fortunately, most weather and climate data 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. One must also consider whether there exist enough data values to justify using any given measure. Moreover, the use of simple summary statistics in this 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.

5.5.3 Glyphs for Standard Deviation

Experiments were conducted 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. This approach was discarded 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 [13]. A dark glyph placed next to a bright glyph of the same size will appear smaller. Instead of glyphs at every point, the decisions was made to move toward the use of glyphs to highlight highs and lows in the data.

5.6 Conclusion

In this chapter, a framework to ensemble visualization has been presented 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 this approach and the coordination between views allows data analysts to focus on the formulation and evaluation of hypothesis in ensemble data. The strengths of this 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. The framework has been demonstrated in two different software prototypes that allow the analysis of large data sets with hardware requirements easily met by present-day laptops.

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. This work constitutes early progress toward the many new challenges posed by these large, complex, and rich data sets.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

This dissertation examined numerous methods for visualizing data with indications of uncertainty. Given the lack of preexisting visualization techniques for this type of data, this work naturally began by employing simple summarization techniques of the data through mean and standard deviation. These values were employed throughout the work as well-known measures of the expected value and variation in the data, and as useful overview values. Methods for combining the visualization of these values were discussed as was the need for more refined ways for exploring the data. To this end, visualization tools that explore the relationship between input and output parameters were discussed. These methods used volume rendering, isosurfacing and streamline tracing to better understand underlying relationships.

However, these types of relationships are not always present in the data, and it is often more useful to explore the density distribution of the data. For this, a new type of boxplot was introduced as well as visual metaphors for higher order summary statistics. These techniques were extended into two dimensions, and a use case of the approach was presented.

Finally, this dissertation culminates in the Ensemble-Vis framework for the graphical data analysis of ensemble data. Because of the complexity of ensemble data and its prevalence in science, the Ensemble-Vis framework strives to create a general approach to visualization that provides the scientist with a collection of display devices to allow for a thorough investigation of the data.

6.1 Future Work

Data sets of the type examined in this work continue to grow in size and complexity. Avenues for future work are extensive and include dealing with the increased complexity and size of the data. Any technique applied to this type of data will have to employ dimension reduction algorithms, parallel methods, or streaming approaches.

Most of the data used in this work maintained a two-dimensional spatial domain. This simplified many of the visualizations required to understand the data, as well as simplified the user interfaces. Increasing the spatial domain, even only to 3D, increases the complexity of the required visualizations and may render many of the techniques presented here useless. This work takes advantage of the extra spatial dimension through the use of displacement maps and stacking slices for volume rendering and isosurfacing. When the data takes up this third dimension, it is no longer free to be used in the visualization. In this case, the data will either have to be reduced in dimension, or visualization methods must encode the information outside of the spatial domain. This problem obviously becomes even harder when the dimension of the data domain exceeds 3, which is the case in 3D data augmented by a time component. In addition, data sets may have input parameters that have a higher order such as vector or tensor parameters, and understanding how these parameters interact with output parameters will be important. In these cases, visual clutter and occlusion will be increasingly problematic.

Another open problem is the visualization of probability distribution functions in higher dimensions. The summarization statistics used by the summary plots are helpful for 1D data sets, or correlating 2D data sets, however, it is unclear how well they will work with N-D data. The summary plots are but one way of depicting PDFs, and their use in higher dimensions may be through data reduction schemes or as sidebars within larger visualization frameworks such as Ensemble-Vis.

As data sets grow larger, the use of mean and standard deviation may also become troublesome in that the data may become too large to efficiently calculate these values and methods will be needed for insitu methods for calculating statistics, as well as visualization techniques that can handle large quantities of streaming data.

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] Us census bureau american community survey public use microdata. http://www.census.gov/acs/www/Products/PUMS/index.htm.

[5] *Digital Image Processing*. Prentice Hall, 2002.

[6] Anselin, L., Syabri, I., and Smirov, O. Visualizing multivariate spatial correlation with dynamically linked windows. In *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting. CD-ROM* (2002).

[7] Bain, L. J., and Engelhardt, M. *Introduction to Probability and Mathematical Statistics*. Duxbury Press, 1992.

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

[9] Becketti, S., and Gould, W. Rangefinder box plots. *The American Statistician 41*, 2 (May 1987), 149.

[10] Benjamini, Y. Opening the box of a boxplot. *The American Statistician 42*, 4 (November 1988), 257–262.

[11] Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

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

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

[14] Bordoloi, U., Kao, D., and Shen, H.-W. Visualization techniques for spatial probability density function data. *Data Science Journal 3* (2005), 153–162.

[15] Borland, D., and II, R. T. Rainbow color map (still) considered harmful. *IEEE Computer Graphics and Applications 27*, 2 (Mar/April 2007), 14–17.

[16] Bruce, V., Green, P. R., and Georgeson, M. A. *Visual Perception*. Psychology Press, 2003.

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

[18] Bürger, R., and Hauser, H. Visualization of multivariate scientific data. In *Eurographics 2007 STAR* (2007), pp. 117–134.

[19] Cedilnik, A., and Rheingans, P. Procedural annotation of uncertain information. In *IEEE Proceedings Visualization 2000* (2000), pp. 77–84.

[20] Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. *Graphical Methods for Data Analysis*. Wadsworth, 1983.

[21] Chlan, E. B., and Rheingans, P. Multivariate glyphs for multi-object clusters. In *Proceedings of InfoVis '05* (2005), pp. 141–148.

[22] Choonpradub, C., and McNeil, D. Can the box plot be improved? *Songklanakarin Journal of Science and Technology 27*, 3 (2005), 649–657.

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

[24] Cleveland, W. S. *The Elements of Graphing Data*. Hobart Press, 1994.

[25] Cohen, D. J., and Cohen, J. The sectioned density plot. *The American Statistician 60*, 2 (May 2006), 167–174.

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

[27] DeVolder, B., Glimm, J., Grove, J., Kang, Y., Lee, Y., and Ye, K. Uncertainty quantification for multiscale simulations. *Journal of Fluids Engineering 124*, 1 (March 2002), 29–41.

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

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

[30] Djurcilov, S., Kim, K., Lermusiaux, P., and Pang, A. Visualizing scalar volumetric data with uncertainty. *Computers and Graphics 26* (2002), 239–248.

[31] Doane, D. P., and Tracy, R. L. Using beam and fulcrum displays to explore data. *The American Statistician 54*, 4 (November 2000), 289–290.

[32] Ehlschlaeger, C. R., Shortridge, A. M., and Goodchild, M. F. Visualizing spatial data uncertainty using animation. *Computers in GeoSciences 23*, 4 (1997), 387–395.

[33] Esty, W. W., and Banfield, J. D. The box-percentile pot. *Journal of Statistical Software 8*, 17 (2003).

[34] Frigge, M., Hoaglin, D. C., and Iglewicz, B. Some implementations of the box plot. *The American Statistician 43*, 1 (February 1989), 50–54.

[35] Fua, Y.-H., Ward, M., and Rundensteiner, E. Hierarchical parallel coordinates for exploration of large datasets. In *Proceedings of Vis '99* (1999), pp. 43–50.

[36] Geneser, S. E., Kirby, R. M., and MacLeod, R. S. Application of stochastic finite element methods to study the sensitivity of ecg forward modeling to organ conductivity. *IEEE Transations on Biomedical Engineering 55*, 1 (January 2008), 31–40.

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

[38] Goldberg, K. M., and Iglewicz, B. Bivariate extensions of the boxplot. *Technometrics 34*, 3 (August 1992), 307–320.

[39] Grigoryan, G., and Rheingans, P. Point-based probabilistic surfaces to show surface uncertainty. In *IEEE Transactions on Visualization and Computer Graphics* (September/October 2004), pp. 546–573.

[40] Gruber, M., and Hsu, K.-Y. Moment-based image normalization with high noise-tolerance. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19*, 2 (1997), 136–139.

[41] Haemer, K. W. Range-bar charts. *The American Statistician 2*, 2 (April 1948), 23.

[42] Hintze, J. L., and Nelson, R. D. Violin plots: A box plot-density trace synergism. *The American Statistician 52*, 2 (May 1998), 181–184.

[43] Hubbard, B., and Santek, D. The vis-5d system for easy interactive visualization. In *IEEE Vis '90* (1990), pp. 28–35.

[44] Interrante, V. Harnessing natural textures for multivariate visualization. *IEEE Computer Graphics and Applications 20*, 6 (November/December 2000), 6–11.

[45] Johnson, C. R. Top scientific visualization research problems. *IEEE Computer Graphics and Applications 24*, 4 (July/August 2004), 13–17.

[46] Johnson, C. R., and Sanderson, A. R. A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications 23*, 5 (2003), 6–10.

[47] Jospeh, A. J., Lodha, S. K., Renteria, J. C., and Pang., A. Uisurf: Visualizing uncertainty in isosurfaces. In *Proceedings of the Computer Graphics and Imaging* (1999), pp. 184–191.

[48] Kao, D., Dungan, J. L., and Pang, A. Visualizing 2d probability distributions from eos satellite image-derived data sets: a case study. In *VIS '01: Proceedings of the conference on Visualization '01* (2001), pp. 457–460.

[49] Kao, D., Kramer, M., Love, A., Dungan, J., and Pang, A. Visualizing distributions from multi-return lidar data to understand forest structure. *The Cartographic Journal 42*, 1 (June 2005), 35–47.

[50] Kao, D., Luo, A., Dungan, J. L., and Pang, A. Visualizing spatially varying distribution data. In *Information Visualization '02* (219–225), p. 2002.

[51] Kindlmann, G., Whitaker, R., Tasdizen, T., and Moller, T. Curvature-based transfer functions for direct volume rendering: Methods and applications. In *Proceedings of the 14th IEEE Visualization 2003 (VIS'03)* (2004), pp. 67–74.

[52] Krueger, J., and Westermann, R. Acceleration techniques for gpu-based volume rendering. In *In Proceedings IEEE Visualization 2003* (2003), pp. 287–292.

[53] Lenth, R. V. Comment on rangefinder box plots. *The American Statistician 42*, 1 (February 1988), 87–88.

[54] Levkowitz, H., and Herman, G. Color scale for image data. *Computer Graphics and Applications 12*, 1 (January 1992), 72–80.

[55] Lodha, S., Sheehan, B., Pang, A., and Wittenbrink, C. Visualizing geometric uncertainty of surface interpolants. In *Proceedings of the conference on Graphics interface '96* (1996), pp. 238–245.

[56] Love, A., Pang, A., and Kao, D. Visualizing spatial multivalue data. *IEEE Computer Graphics and Applications 25*, 3 (May 2005), 69–79.

[57] Lundstrm, C., Ljung, P., Persson, A., and Ynnerman, A. Uncertainty visualization in medical volume rendering using probabilistic animation. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (Nov./Dec. 2007), 1648–1655.

[58] Luo, A., Kao, D., and Pang, A. Visualizing spatial distribution data sets. In *VISSYM '03: Proceedings of the symposium on Data visualisation 2003* (2003), pp. 29–38.

[59] MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., and Hetzler, E. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science 32*, 3 (July 2005), 139–160.

[60] McGill, R., Tukey, J. W., and Larsen, W. A. Variations of box plots. *The American Statistician 32*, 1 (February 1978), 12–16.

[61] Mihalisin, T., Timlin, J., and Schwegler, J. Visualization and analysis of multivariate data: a technique for all fields. In *IEEE Vis '91* (1991), pp. 171–178.

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

[63] Olston, C., and Mackinlay, J. D. Visualizing data with bounded uncertainty. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)* (2002), pp. 37–40.

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

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

[66] Potter, K. Methods for presenting statistical information: The box plot. *In Hans Hagen, Andreas Kerren, and Peter Dannenmann (Eds.), Visualization of Large and Unstructured Data Sets, GI-Edition Lecture Notes in Informatics (LNI) S-4* (2006), 97–106.

[67] Potter, K., Gooch, A., Gooch, B., Willemsen, P., Kniss, J., Riesenfeld, R., and Shirley, P. Resolution independent npr-style 3d line textures. *Computer Graphics Forum 28*, 1 (2009), 52–62.

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

[69] Rheingans, P. Task-based color scale design. In *Proceedings of Applied Image and Pattern Recognition '99, SPIE* (1999), pp. 35–43.

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

[71] Roberts, J. State of the art: Coordinated and multiple views in exploratory visualization. In *5th International Conference on Coordinated and Multiple Views in Exploratory Visualization* (2007), pp. 61–71.

[72] Rousseeuw, P. J., Ruts, I., and Tukey, J. W. The bagplot: A bivariate boxplot. *The American Statistician 53*, 4 (November 1999), 382–287.

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

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

[75] Spear, M. E. *Charting Statistics.* McGraw-Hill, 1952.

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

[77] Taylor, B. N., and Kuyatt, C. E. Guidelines for evaluating and expressing the uncertainty of nist measurement results. Tech. rep., NIST Tecnical Note 1297, 1994.

[78] Tongkumchum, P. Two-dimensional box plot. *Songklanakarin Journal of Science and Technology 27*, 4 (2005), 859–866.

[79] Tufte, E. R. *The Visual Display of Quantitative Information.* Graphics Press, 1983.

[80] Tukey, J. W. *Exploratory Data Analysis.* Addison-Wesley, 1977.

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

[82] Wilkinson, L. *The Grammar of Graphics.* Springer-Verlag New York, Inc., 1999.

[83] Wittenbrink, C., Pang, A., and Lodha, S. Verity visualization: Visual mappings. Tech. rep., University of California at Santa Cruz, 1995.

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

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