Guided Analysis of Hurricane Trends Using Statistical Processes Integrated with Interactive Parallel Coordinates

Chad A. Steed* Naval Research Laboratory
J. Edward Swan II† Mississippi State University
T.J. Jankun-Kelly‡ Mississippi State University
Patrick J. Fitzpatrick§ Mississippi State University

ABSTRACT

This paper demonstrates the promise of augmenting interactive multivariate representations with information from statistical processes in the domain of weather data analysis. Statistical regression, correlation analysis, and descriptive statistical calculations are integrated via graphical indicators into an enhanced parallel coordinates system, called the Multidimensional Data eXplorer (MDX). These statistical indicators, which highlight significant associations in the data, are complemented with interactive visual analysis capabilities. The resulting system allows a smooth, interactive, and highly visual workflow.

The system's utility is demonstrated with an extensive hurricane climate study that was conducted by a hurricane expert. In the study, the expert used a new data set of environmental weather data, composed of 28 independent variables, to predict annual hurricane activity. MDX shows the Atlantic Meridional Mode increases the explained variance of hurricane seasonal activity by 7-15% and removes less significant variables used in earlier studies. The findings and feedback from the expert (1) validate the utility of the data set for hurricane prediction, and (2) indicate that the integration of statistical processes with interactive parallel coordinates, as implemented in MDX, addresses both deficiencies in traditional weather data analysis and exhibits some of the expected benefits of visual data analysis.

Keywords: Climate study, multivariate data, correlation, regression, interaction, statistical analysis, visual analytics.

Index Terms: I.3.6 [Computing Methodologies]: Computer Graphics—Methodologies and Techniques; H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces; J.2 [Computer Applications]: Physical Sciences and Engineering—Earth and Atmospheric Sciences;

1 INTRODUCTION

For years, weather scientists have used statistical analysis to investigate associations between weather phenomena and environmental observations. In particular, they have often used statistical regression, descriptive statistics, and correlation measures to estimate the importance of a set of predictors for seasonal hurricane activity. The motivation is to be able to predict, as far in advance as possible, the number and intensity of hurricanes that are likely to occur during a given hurricane season.

Weather scientists have also often used basic data graphics such as scatterplots and histograms. However, the inability of basic data graphics to convey complicated data associations is well understood by the visualization community [9]. Another limitation, which is directly addressed in the work reported here, is that visualizations have typically lacked a direct connection to statistical processes such as regression analysis. And yet, the authors have noted that for the environmental scientists that they work with—weather scientists, oceanographers, and physicists—a typical workflow involves performing statistical and graphical analysis simultaneously, where the statistical correlations suggest causal relationships, and the graphical analysis helps explain why the relationships exist.

These facts have motivated the work reported here. We have developed a system, called the Multidimensional Data eXplorer (MDX), that augments a parallel coordinates visualization with automated statistical processes. As shown in Fig. 1, MDX combines parallel coordinates capabilities with advanced interaction techniques integrated with MATLAB-based regression analysis. While most of the visualization and interaction techniques in MDX have been previously reported, we do not believe that any other parallel coordinates system has provided the range of capabilities as MDX.

In addition, this paper reports an extensive case study, in which a hurricane expert studied the ability of a new environmental weather data set, the 28-variable Colorado State University (CSU) and Atlantic Meridional Mode (AMM) data set, to predict annual hurricane activity. The hurricane expert is Dr. Patrick Fitzpatrick; he has been a long-term collaborator in this work and is a co-author of this paper. Approximately half of this paper describes the capabilities of MDX, while the other half describes the case study. Therefore, the research reported here is structured to directly address one of the main calls of the NIH/NSF Visualization Challenges Report [13], which recommends that visualization researchers "collaborate closely with domain experts who have driving tasks in data-rich fields to produce tools and techniques that solve clear real-world needs." In this paper, the tool is MDX, the techniques are parallel coordinates and interaction methods integrated with statistical regression routines, and the real-world need is the prediction of annual hurricane activity.

2 RELATED WORK

Using schemes similar to those described by Vitart [28], scientists have successfully employed statistical regression techniques using historical data to identify significant predictors for hurricane activity. Recently, Klotzbach et al. [15] used regression techniques to identify the most important variables for predicting the frequency of North Atlantic hurricane activity. Fitzpatrick [5] also created a multiple regression scheme called the Typhoon Intensity Prediction Scheme (TIPS) to understand and forecast storm intensity in the western North Pacific Ocean. Although sometimes complicated to establish, these and other statistical analysis methods (e.g. correlation analysis, descriptive statistics) provide valuable insight regarding key associations in the climate data.

In conjunction with statistical processes, climate researchers use scatterplots and histograms that utilize either layered plots or multiple separate plots. However, the use of separate plots causes perceptual issues due to the limited memory for information that can be gained from one glance to the next as discussed by Healey et al. [9]. Statisticians often use multiple adjacent scatterplots (a scatterplot matrix, or SPLOM), as described by Wong and Bergeron [32]; but it requires a large amount of space and forming multivariate asso-

*e-mail: csteed@acm.org
†e-mail: swan@acm.org
‡e-mail: tjk@acm.org
§e-mail: fitz@ngi.msstate.edu

IEEE Symposium on Visual Analytics Science and Technology
October 12 - 13, Atlantic City, New Jersey, USA
978-1-4244-5283-5/09/$25.00 ©2009 IEEE

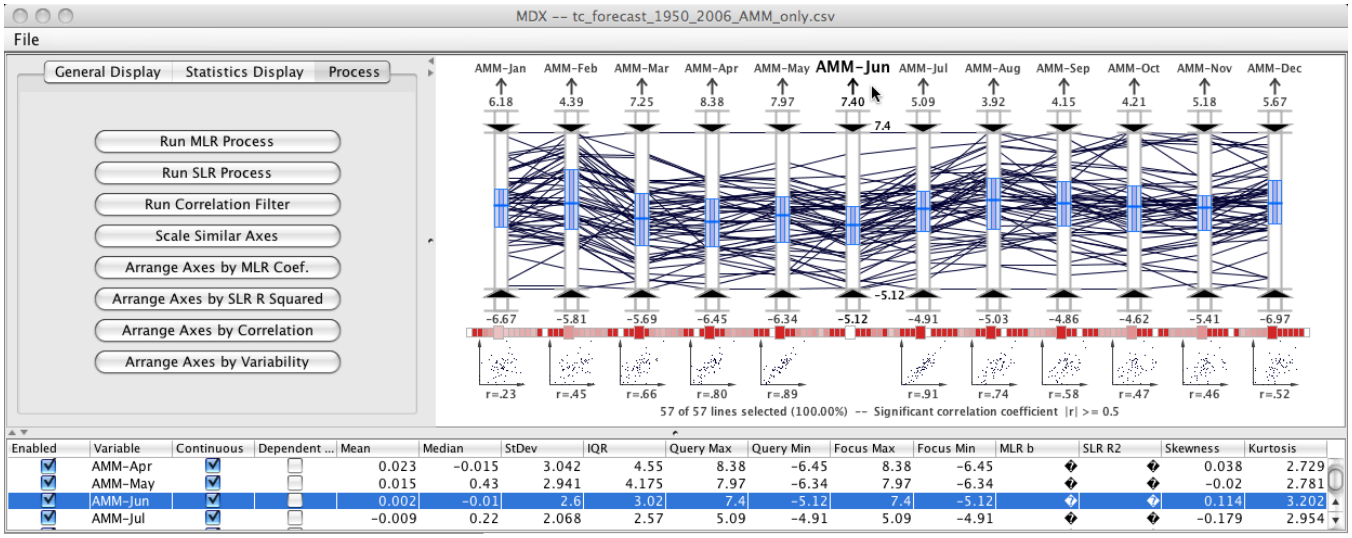


Figure 1: The MDX system developed in this research is composed of a settings panel (upper left), parallel coordinates panel (upper right), and a table view panel (lower). In this example, the mouse is used to highlight the *AMM-June* axis to examine its correlation with the other AMM variables. The correlation appears to be stronger for the months that are closer in time. This pattern is observed for all the AMM variables.

ciations is still difficult. Wilkinson et al. [31] employed statistical measures to organize the SPLOM and guide the viewer through exploratory analysis of high-dimensional data sets, but the perceptual issues mentioned above remain to some degree. Instead of separate plots, layer plots can be used, which condense the visualization into a single display; but there are significant issues related to layer occlusion and interference as described by Healey et al. [9].

In the current work, the parallel coordinates visualization technique forms the canvas for an approach designed to address the needs of climate analysis. The parallel coordinates concept, which was first popularized by Inselberg [11] to represent hyper-dimensional geometries and also used for direct analysis of multivariate relationships by Wegman [30], yields a compact, two-dimensional representation of even large multidimensional data sets. Several innovative parallel coordinates extensions have been described in the visualization research literature. For example, Martin and Ward [18] described data driven brushing techniques for parallel coordinates and Hauser et al. [8] described a histogram display, dynamic axis re-ordering, axis inversion, and details-on-demand capabilities. In addition, Siirtola [21] presented a rich set of dynamic interaction techniques (e.g., conjunctive queries) and parallel coordinate axes enhanced with the box plots developed by Tukey [26]. Johansson et al. [12] described new line shading schemes for parallel coordinates and Dykes and Mountain [4] introduce a shading approach similar to the aerial perspective shading used in the current work. Furthermore, several focus+context implementations for parallel coordinates have been introduced by Fua et al. [7], Artero et al. [2], Johansson et al. [12], and Novotńy and Hauser [19]. Theus [25] introduced a system that ties parallel coordinates to statistical processing functions and Qu et al. [20] introduced a method for integrating correlation computations into a parallel coordinates display. MDX utilizes variants of these extensions to enhance the classical parallel coordinates plot.

Parallel coordinates displays have also been used for environmental data analysis in earlier works. MacEachren et al. [17] used interactive parallel coordinates to investigate atmospheric data sets. Also, Theron [24] used parallel coordinates in a visual analytics system to analyze paleoceanographic conditions. In the current work, a case study is described on the use of parallel coordinates for conducting a tropical cyclone climate study.

3 Information Assisted Visual Analysis

MDX offers a comprehensive environment for visual multivariate data analysis by fusing automated statistical processes with enhanced parallel coordinates. The information derived from these statistical processes is used to augment the parallel coordinates axes in the form of graphical indicators that guide the user to important relationships, thereby reducing knowledge discovery timelines. The resulting visualization provides a rich blend of multivariate visualization and statistical information assistance that moves beyond conventional environmental data analysis. In the remainder of this section, we will describe the key information assistance features and the interactive visual analysis capabilities.

3.1 Statistical Information Assistance

The statistical processes in MDX yield information that guides the scientist via graphical indicators. In real-time, the descriptive statistics and correlation measures are computed to provide key characterizations of the variables. Furthermore, regression analyses are executed on demand to quickly indicate associative connections between multiple independent and dependent variables, thus amplifying cognition and reducing analysis time frames.

3.1.1 Regression Processes

Regression analysis techniques are effective for screening data and providing quantitative associations. In addition to simple linear regression (SLR), MDX offers stepwise multiple linear regression (MLR) with a backwards glance, which selects the optimum number of most important variables using a predefined significance level [29]. The stepwise MLR helps find a model that does a good job of predicting the dependent variable with as few independent variables as possible, which simplifies interpretation and usually means cheaper data collection and analysis. In general, the idea of stepwise regression is to start with an initial model and add or delete variables step by step, one at a time, to make the model better. The procedure stops when no appreciable improvement is gained by making another step. Although the resulting model may not be the best of all possible models, it is generally one of the best. In MDX, we use the MATLAB "regress" and "stepwisefit" utilities to perform simple and stepwise regression, respectfully. The information from these processes is parsed and graphically represented in the parallel coordinates panel.

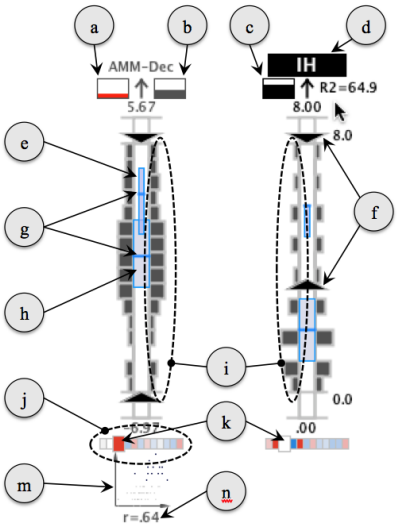


Figure 2: The parallel coordinates axes in MDX have been augmented with key statistical indicators. The call-outs in this annotated figure highlight the regression (see a, b, c, and d), descriptive statistics (e, g, h, and i), and correlation (j, k, m, and n) indicators. Query sliders (f) are also indicated, which facilitate dynamic visual queries.

In our regression analysis, a normalization procedure is executed so that the y-intercept becomes zero and the importance of a predictor may be assessed by comparing regression coefficients, b_i, between different predictors. Denoting σ as the standard deviation of a variable, y as the dependent variable, $\bar{x}$ as the predictor mean, and $\bar{y}$ as the dependent variable mean, a number k of statistically significant predictors are normalized by the following equation:

$$(y-\bar{y})/\sigma_y = \sum_{i=1}^{k} b_i(x_i-\bar{x}_i)/\sigma_i. \quad (1)$$

The b_i values are graphically encoded in the parallel coordinates plot using the box below the axis label and to the left of the arrow (Fig. 2, a). Like a thermometer, the box is filled from the bottom to the top based on the magnitude of b. The box is colored red if the coefficient is positive and blue if it is negative. The box to the right of the arrow (Fig. 2, b) encodes the r^2 output from the SLR process. In addition to the coefficients, the MLR analysis returns an overall R^2 value which provides an indication of how well the model captures the variance between the predictors and the dependent variable. The box beneath the dependent variable axis name (Fig. 2, c) encodes the overall R^2 value from the MLR analyses. It is important to note that the axis corresponding to the current dependent variable is indicated by light gray text on a dark gray box for its title (Fig. 2, d), the reverse shading of the other axes.

3.1.2 Descriptive Statistical Indicators

The axis display also provides visual representations of key descriptive statistical measures that aid the user in ascertaining the characteristics of variable distributions. The median, interquartile range (IQR), and frequency information are calculated for the data within the focus area of each axis. Alternatively, the user can configure MDX to display the mean and standard deviation range. These central tendency and variability measures provide quantitative measures for the typical value and how "spread out" the samples are in the distribution, respectively. This information can assist the user in deriving confidence metrics (e.g. tight line clusters may indicate better stability) as well as segmenting the data into above and below normal ranges.

The variability information is encoded in the boxes that are drawn on each axis interior. The wide boxes (Fig. 2, h) represent the descriptive statistics for all the axis samples, while the more narrow boxes (Fig. 2, e), which are drawn over the wide boxes, capture the descriptive statistics for the samples that are currently selected with the axis query sliders (Fig. 2, f). The thick horizontal lines that divide the variability boxes vertically (Fig. 2, g) represent either the median or mean value.

The frequency information can also be displayed on each axis bar as vertical histogram bins (Fig. 2, i) with widths that are indicative of the number of lines that pass through the bin's region. That is, the widest bins have the most lines passing through, while the more narrow bins have less lines. The user can also enable or disable the histogram display and fine tune the bin size parameter via the MDX settings.

3.1.3 Correlation Indicators

MDX also facilitates correlation analysis whereby the strength of relationships between pairs of variables are measured. The correlation coefficient, r, is used to quantify the relationship between two variables. Specifically, the system uses the Pearson product-moment correlation coefficient to measure the correlation for a series of n measurements of X and Y written as x_i and y_i where $i = 1,2,\ldots,n$ [29]. The value of r is given by:

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n\sum {x_i}^2 - (\sum x_i)^2][n\sum {y_i}^2 - (\sum y_i)^2]}} \quad (2)$$

As the user interacts with the system, the correlation information is updated in real-time and used to augment the parallel coordinates display. For each possible pairing of axes, the system automatically calculates r, which yields a correlation matrix. The correlation matrix is a square matrix where each i, j element is equal to the r value between the i and j variables. The rows from this matrix are displayed graphically beneath each axis as a series of color-coded blocks (Fig. 2, j). In each block, the color encodes r between the axis directly above it and the axis that corresponds to its position in the set of blocks. When the mouse hovers over an axis, the axis is highlighted and the correlation coefficient blocks corresponding to it below the other axes are enlarged (Fig. 2, k).

Similar to the regression indicators, the correlation indicator blocks are shaded blue for negative correlations and red for positive. The stronger the correlation, the more saturated the color so that the stronger correlations are more prominent. The r value of an axis with itself is always equal to one and the corresponding indicator block is shaded white. Furthermore, when the absolute value for r is greater than or equal to the significant correlation threshold, the block is colored with the fully saturated color. Displayed at the bottom of the parallel coordinates display (see Fig. 1), the significant correlation threshold is a user-defined value that controls the correlation block shading and the multicollinearity filter.

In addition to the indicator blocks, MDX displays small scatterplots below the correlation indicators for each axis when an axis is highlighted (Fig. 2, m). The scatterplots are created by plotting the points with the highlighted variable as the y axis and the variable directly above the scatterplot as the x axis. Each scatterplot also shows the numerical r value associated with the pair of axes below the scatterplot (Fig. 2, n). The scatterplots provide a visual means to confirm the type of correlation (positive or negative) and the strength. The type of correlation is also visually detectable in the line configuration of the parallel coordinates plot. Polylines that cross in an 'X' pattern are characteristic of a negative correlation while lines that appear to be more parallel indicate a positive correlation. Unlike the other correlation indicators, the scatterplot is useful for discovering nonlinear relationships between variables. For example, a nonlinear relationship can be observed in a scatterplot even if r is zero.

3.1.4 Multicollinearity Filter

Multicollinearity is a condition in which an independent variable is highly correlated with several other independent variables. This variable has much in common with several other variables and may have little information unique to itself [10]. The presence of this condition results in loss of power and makes interpretation more difficult in the results of regression models.

MDX provides an automated multicollinearity filter to ensure the proper selection of axes in subsequent regression analyses. The filter examines the visible axes in the parallel coordinates display for multicollinearity; if any axes are correlated with each other by more than the significant correlation threshold, one is removed from the display. The filter removes the axis that has a lower r with the dependent axis and the remaining axis are truly independent of each other. Although the user can use the correlation indicators to manually reduce the multicollinearity, the automatic filter can be used to ensure independence among the variables by simply clicking a user interface button.

3.1.5 Axis Ordering

MDX expedites statistical comparisons through its automatic axis arrangement capabilities, which use one of the previously mentioned statistical measures. When the axes are arranged by the correlation coefficient, one axis is selected initially as the target axis. The axes are then sorted according to the r value of the target axis and the other visible axes. Axes with negative correlations are arranged to the left of the target axis in ascending order. Axes with positive correlations are arranged to the right of the target axis in descending order. The strongest correlations are placed nearest to the target axis while the weakest correlations are placed farthest. When the axes are sorted in this manner, the user can quickly identify the strongest correlations with the target axis.

The IQR / standard deviation range, MLR b, and SLR r^2 options all sort the axes in descending order based on the statistical values. The dependent axis is placed at the leftmost position and the other axes are arranged accordingly. The IQR / standard deviation range arrangements are useful for examining the dispersion characteristics of each axis. The SLR r^2 and MLR b arrangements are useful for investigating individual associations of the axes with the dependent axis and identifying the most significant axes for the dependent axis, respectively.

3.2 Visual Analysis Techniques

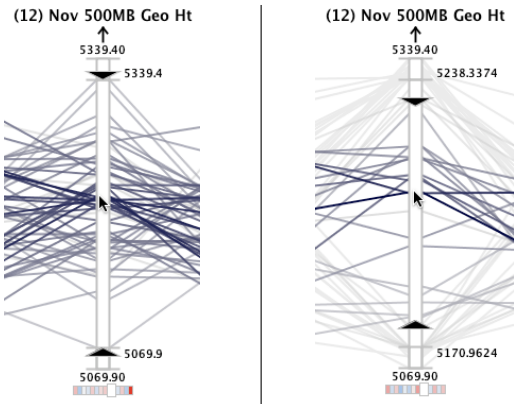
In addition to several fundamental parallel coordinates capabilities such as relocatable axes, axis inversion, and details-on-demand, MDX includes additional extensions such as rapid axis scaling, focus+context, aerial perspective shading, and dynamic visual queries. A detailed description of these capabilities are provided in earlier publications of this research [22, 23]. In the remainder of this section, we give an overview of the more significant extensions to set the stage for the subsequent climate study.

3.2.1 Dynamic Queries

MDX enables the rapid selection of subsets of polylines in the parallel coordinates display using double-ended sliders. Each axis has a pair of sliders (Fig. 2, f) for constraining the upper and lower limits of the highlighted lines. The user can drag these sliders using the mouse to perform rapid Boolean AND queries. Lines within the sliders of each axis are rendered with a darker shade of gray while the remaining lines are rendered with a less prominent shade of gray. An example of this capability is shown in Fig. 5 to highlight above normal *IH* activity.

3.2.2 Axis Scaling (Focus+Context)

The dynamic axis scaling capabilities in MDX enable interactive exploration of data subsets while retaining the context of the full data set. This capability is implemented by allowing the user to modify the upper and lower focus area values for a selected axis using the movement of the mouse wheel.



(a) Before scaling (b) After scaling

Figure 3: Image sequence illustrating dynamic axis scaling capability before (a) and after (b) axis scaling has been performed. In this example, scaling is performed by an upward mouse wheel movement in the focus area of the axis which moves the values of the upper and lower limits closer together, zooming into the central axis region.

Each axis is partitioned into three sections delineated by horizontal tick marks: the central focus area and the top and bottom context areas. When the mouse hovers over the focus area (see Fig. 3), an upward mouse wheel motion expands the display of the focus area outward and pushes lines outside the new upper and lower limits into the context areas. A downward wheel motion causes the inverse effect: focus area compression. Alternatively, the user may use the mouse wheel over either of the two context areas to alter the minimum or maximum limits, independently. This intuitive capability supports zooming into an area of interest for an axis to reduce clutter and facilitate focused analysis on specific polylines.

3.2.3 Aerial Perspective Shading

MDX also provides a proximity-based line shading scheme, which supports rapid investigation of multidimensional trends. This shading scheme simulates the human perception of aerial perspective whereby objects in the distance appear faded while objects nearer to the eye seem more vivid. The technique is similar to a fundamental painting technique often employed in landscapes. As the distance between the viewer and an object is increased, the contrast between the object and the background decreases.

The shading scheme is implemented in MDX in both a discrete and a continuous mode. In the discrete mode, the lines are colored according to the axis region that they intersect. As shown in Fig. 5, lines that are within the query slider limits for each axis are shaded with a dark gray value that draws the user's attention. Lines that fall within the context of any axis are shaded a light gray color and drawn beneath the other lines. The remaining lines (non-query and non-context) are colored a shade of gray that is darker than the context lines but lighter than the query lines.

In the continuous mode, non-context lines go through an additional step to encode the distance of the line from the mouse cursor. As shown in Fig. 3, query lines that are nearest to the mouse cursor receive the darkest value while lines farthest from the mouse cursor are shaded with a lighter gray. The other query lines are shaded according to a non-linear fall-off function that yields a gradient of colors between said extremes. Consequently, the lines that are nearest to the mouse cursor are more prominent to the viewer due to the color and depth ordering treatments and the viewer can effectively use the mouse to quickly interrogate the data set. In addition to the

Table 1: CSU+AMM Hurricane climate variables evaluated as predictors in the climate study.

	CSU Variables		AMM Variables
(1)	June–July Niño 3	(17)	AMM-January
(2)	May SST	(18)	AMM-February
(3)	February 200-mb U	(19)	AMM-March
(4)	February–March 200-mb V	(20)	AMM-April
(5)	February SLP	(21)	AMM-May
(6)	October–November SLP	(22)	AMM-June
(7)	Sept. 500-mb Geopotential Height		
(8)	November SLP	(23)	AMM-July
(9)	March–April SLP	(24)	AMM-August
(10)	June–July SLP	(25)	AMM-September
(11)	September–November SLP	(26)	AMM-October
(12)	Nov. 500-mb Geopotential Height		
(13)	July 50-mb U	(27)	AMM-November
(14)	February SST	(28)	AMM-December
(15)	April–May SST		
(16)	June–July SST		

SST – Sea Surface Temperature
SLP – Sea Level Pressure
U – zonal wind (or x-coordinate) component
V – meridional wind (or y-coordinate) component

query lines, the shading scheme is also applied to the small scatterplots that are displayed beneath each axis to link the elements over multiple views.

4 Case Study: North Atlantic Hurricane Trend Analysis

The objective of this research is to show the promise of an approach to interactive visual analytics that relies on guidance from statistical information. Building on our earlier case studies [22, 23], we have employed this approach in the current work by investigating trends in a new environmental data set and ascertaining the significance of the data set in comparison to a related data set of hurricane predictors. The discovery of hidden trends, confirmation of known patterns, and feedback from the domain expert reveal the potential for enhanced knowledge discovery in climate studies. In the remainder of this section, we provide an overview of the climate data along with details of the analyses.

4.1 Climate Data

This climate study involves two data sets that contain environmental predictors for hurricane seasons. The first data set contains 16 climate predictors observed annually from 1950 to 2006 (57 records) and they are listed along with the associated geographical region in Table 1. This data set was provided by Dr. Phil Klotzbach [14] of the Tropical Meteorology Project at Colorado State University (CSU), where it has been used in forecasts of the 2007 North Atlantic hurricane season activity by categories. These categories include the number of named storms (NS)[1], the number of hurricanes (H)[2], and the number of intense hurricanes (IH)[3] [6].

The CSU variables have known association with North Atlantic hurricane activity. For example, Chu [3] describes how this basin has less storms during the El Niño Southern Oscillation (ENSO) years, and more in La Niña years. Because of this relationship, scientists use the ENSO signals as one the primary predictors for seasonal activity. In Table 1, variables 1 through 8 are understood to characterize ENSO events. This data set's description is beyond the scope of this paper, and the reader is directed to Steed et al. [23] for those details.

In this climate study, we also include the Atlantic Meridional Mode (AMM) data set, which has been obtained from the Earth Systems Research Laboratory of the National Oceanic & Atmospheric Administration (NOAA). Recent research by Vimont and Kossin [16, 27] has shown strong relationships between the AMM and North Atlantic storm activity on decadal and interannual time series. The AMM is a dynamic mode of variability that is integral to the tropical coupled ocean-atmosphere system. These connections are due to the AMM's association with several local climate conditions, which all influence storm activity collectively as described by Vimont and Kossin [27]. AMM is highly correlated with a number of local climatic factors such as SST, shear, low-level vorticity and convergence, static stability, and SLP. The local factors cooperate to increase or decrease North Atlantic hurricane activity [16].

To align with the years covered by the CSU data set, we use a subset of the AMM data set in this study to cover the years 1950 to 2006. For each year, the AMM data set has 12 values, one for each month of the year. The AMM spatial pattern is determined by applying a Maximum Covariance Analysis to sea surface temperature and the zonal and meridional components of the 10-meter wind field. The data are defined over the region (21S–32N, 74W–15E), and smoothed spatially using 3 longitude and 2 latitude points [1].

4.2 AMM Exploratory Analyses

We began the study by investigating the trends and associations within the AMM data set. One of the first features we observed is that all of the correlation indicators are red (see Fig. 1), which indicates positive correlations. The correlation indicators also highlight that the AMM variables have strong positive correlations with months that are near in time. For example, the *AMM-June* variable is highlighted in Fig. 1, which reveals exceptionally strong correlations with the four nearest months (*AMM-April*, *AMM-May*, *AMM-July*, and *AMM-August*) and considerably weaker correlation with the other months. A glance at the correlation indicator patterns shows that each AMM variable exhibits this pattern—the correlation strengths falls off as the temporal separation increases.

When we evaluate the correlations of the AMM variables with the *IH* axis, we observe that the highest correlations occur in the months that historically have the highest intense hurricane activity. In Fig. 4, the *AMM-August* and *AMM-September* months have the second and third strongest correlation with the *IH* axis. It is not surprising that the weakest correlations with the *IH* activity are observed in the winter and spring months, which are outside the North Atlantic hurricane season (Jun. 1 to Nov. 30); but it is remarkable that the strongest correlation is with the *AMM-December* axis. The red correlation indicators and scatterplots beneath each axis also show that each AMM variable is positively correlated with the *IH* activity. We observed similar correlations trends with the *NS* and *H* categories; but the correlations with the other seasonal statistics were slightly weaker than with the *IH* statistic.

Using the axis sliders, we compare the AMM variable values for seasons with above normal and below normal *IH* activity. The above normal seasons are highlighted in Fig. 5 to show that as AMM values increase, the *IH* activity also increases. This observation reinforces the positive correlation between the AMM variables and the *IH* variable. The AMM variables were also examined for seasons with above and below normal *NS* and *H* activity and the same positive correlation trends were observed.

After exploring the AMM variables in isolation, the next phase of our analysis included the CSU data set. The 12 AMM were plotted with the 16 CSU parameters and the *IH* dependent variable for 29 total variables—the first simultaneous visualization of these

[1] A tropical cyclone is given a name when its winds reach 17 ms^{-1}.
[2] Hurricanes are tropical cyclones with winds that exceed 32 ms^{-1}.
[3] Intense hurricanes are hurricanes with winds that are at least 49 ms^{-1}.

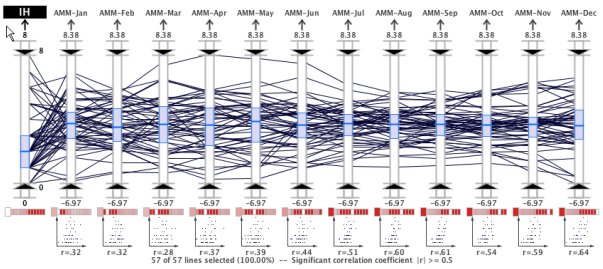


Figure 4: Correlation analysis of the AMM variables with the *IH* category shows that all correlations are positive and *AMM-December* has the strongest correlation, but December is known to be a quiet month for hurricane activity. The other strong correlations (*AMM-August* and *AMM-September*) are within the active portion of the hurricane season.

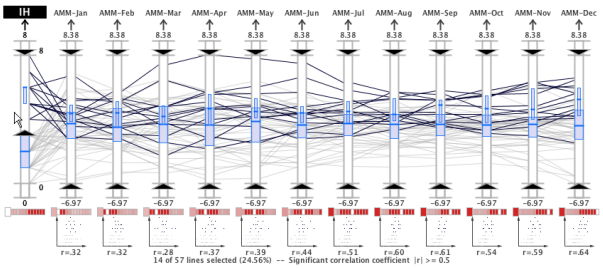


Figure 5: The query sliders are used to examine the seasons with above normal *IH* activity. The resulting line configuration shows that higher activity levels coincide with higher AMM values.

particular parameters. This figure is not shown due to space limitations, but it demonstrates that the visualization technique is only restricted by the horizontal resolution of the display. In this case, the overview of all the parameters revealed that the AMM have much more gradually changing line configurations between axes.

We investigated the correlations between the *IH* axis and the 28 independent variables. We used the MDX axis arrangement capabilities to order the axes according to the correlation coefficient with the *IH* axis. This analysis reveals that the 4 strongest correlations are with AMM variables (*AMM-December*, *AMM-September*, *AMM-August*, and *AMM-November*). Similar correlation analyses were executed for the *H* and *NS* categories. For the *H* statistic, the 6 strongest correlations are with AMM variables: *AMM-December*, *AMM-November*, *AMM-September*, *AMM-August*, *AMM-October*, and *AMM-July*. For the *NS* statistic, the 5 strongest correlations are with AMM variables: *AMM-November*, *AMM-December*, *AMM-October*, *AMM-September*, and *AMM-August*.

4.3 Identifying Significant Predictors

To determine the most significant predictors among the CSU and AMM variables for predicting seasonal hurricane activity, we used the MDX stepwise regression analysis capabilities. We generated 3 regression models for each of the seasonal categories: *NS*, *H*, and *IH*. These are the same dependent variables used in our earlier regression analysis using only the CSU predictors [23]. The significance level is 80%.

4.3.1 Multicollinearity Filter

Prior to running regression processes, the multicollinearity filter was executed in 3 separate cases (one for each for the *IH*, *H*, and *NS* dependent variables) for the full data set. In all three cases, the filter removed 17 of the independent variables from the axis configuration, leaving 11 variables in the display. For the *IH* and *H*

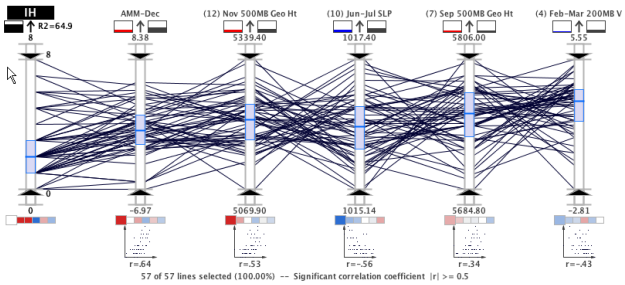


Figure 6: Stepwise regression model using AMM+CSU variables with the dependent variable set to *IH* (1950 to 2006) arranged by the b values. This model excludes *September–November SLP* (11), which was included in our prior regression that used only CSU data.

cases, only 2 AMM variables are included in the new configuration: *AMM-December* and *AMM-May*. With the *IH* case, the unfiltered axes were arranged according by the correlation coefficients with the dependent axis to reveal that the *AMM-December* axis has the highest correlation ($r = .64$). Similar visual analysis after multicollinearity filtering for the *H* category also reveals that *AMM-December* has the strongest correlation ($r = .62$)—a significantly stronger correlation than the second highest correlation ($r = .47$) with *November 500-mb Geopotential Height* (12).

For the *NS* case, 3 AMM variables remain after the filter execution: *AMM-November*, *AMM-January*, and *AMM-June*. Analysis of the remaining variables reveals that the highest correlation coefficient is with the *AMM-November* axis ($r = .62$), which is substantially stronger than the second strongest correlation ($r = .41$) with the *November 500-mb Geopotential Height* (12) axis.

4.3.2 Regression Analyses

After executing the multicollinearity filter, the stepwise regression capabilities are used to identify the most significant predictors among the remaining 11 variables. With the *IH* variable as the dependent variable, the regression model includes the 5 variables shown in Fig. 6. The R^2 value for this model is 65%, which is better than the 58% obtained in our prior case studies that used only the CSU predictors [22, 23]. In this model, the *AMM-December* variable has the highest regression coefficient ($b = 0.3421$). The other 4 variables included in the model are from the CSU data set, which were also included in the prior CSU case studies. The only variable not selected in this model that was selected in the previous studies is *September–November SLP in the southeast Gulf of Mexico* (11). This variable is omitted because the AMM explained more variance and variable (11) could no longer explain any new variance in this particular model.

When the *H* variable is set to the dependent variable, the regression model included the 4 variables shown in Fig. 7. The R^2 value for this model is 56%, which is significantly better than the 42% from the prior CSU data set analysis [23]. Similar to the *IH* regression model, the *AMM-December* variable had the strongest regression coefficient ($b = 0.5027$)—significantly higher than the next strongest coefficient ($b = -0.2404$). Like the previous CSU analyses, this model includes both the *October–November SLP* (6) and *November 500-mb Geopotential Height* (12) variables; but, in the current model, variable (6) has the lowest coefficient instead of the highest. Furthermore, the AMM+CSU model drops the *June–July SLP* (10) and *September–November SLP* (11) variables and includes the *February 200-mb zonal wind in equatorial East Brazil* (3) variable. Variables (10) and (11) are dropped for the same reason that the *IH* model drops variable (11). However, the model keeps variable (6), which also measures SLP. The reason variable (6) is retained and variables (10) and (11) are dropped is possibly due to geographic locations—variable (6) is measured in the Gulf

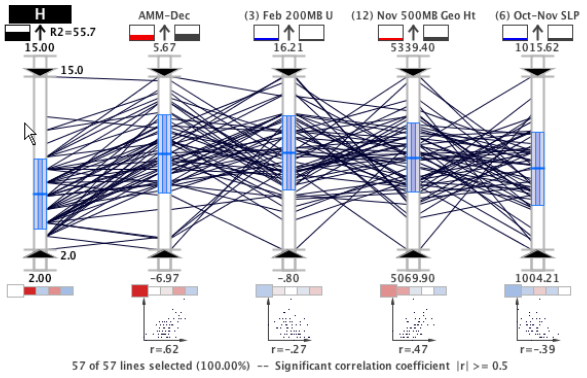


Figure 7: *H* stepwise regression model with AMM and CSU variables (1950 to 2006) arranged by the b values. The *AMM-December* b value is significantly greater than the other variables.

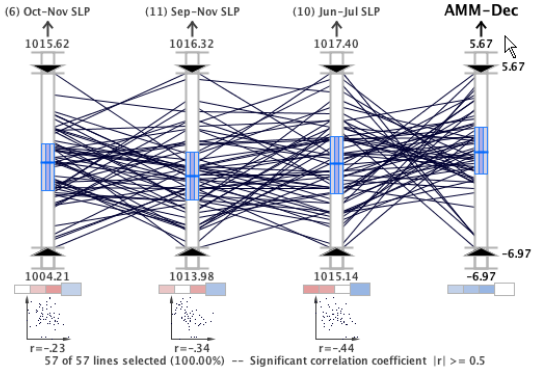


Figure 8: Correlation analysis between *AMM-December* and the 3 SLP variables shows that variable (6) has the weakest correlation.

of Alaska and variables (10) and (11) are measured in areas that overlap the AMM variables. Since variable (6) has the weakest correlation with *AMM-December* of these 3 SLP variables (see Fig. 8), it was not included in the model for the *H* category since there was little new variance to explain.

The *NS* regression model includes only 3 variables and yields an R^2 value of 45%, which is better than the 34% from the prior CSU analyses [23]. The only variable in the current model that was also included in the CSU analyses is *November 500-mb Geopotential Height* (12). Like the *H* model, the AMM+CSU *NS* model includes the *October–November SLP* (6) variable due to the weak correlation with *AMM-November* since it is measured in the Gulf of Alaska. The *February SST* (14) is removed by the multicollinearity filter process because it is strongly correlated with the *AMM-November* variable ($r = .56$) and it has a weaker correlation with the dependent variable ($r = .42$) than *AMM-November*. Similarly, the *September–November SLP* (11) variable is removed by the filter because it is strongly correlated with the *AMM-January* variable ($r = -.58$) and it has a weaker correlation with the dependent variable ($r = -.32$) than *AMM-January* ($r = .33$). Unlike the other two AMM regression models, this model results in the exclusion of the *AMM-December* variable, which is also removed during the multicollinearity filter process due to strong correlation with the *AMM-November* variable ($r = .92$).

In the *H* and *NS* models, the *February 200-mb U* (3) and *October–November SLP* (6) variables are CSU predictors that are not included in the prior CSU regression models [23]. These variables are correlated to El Niño conditions. They are likely included in the model because, as Vimont and Kossin [16] state, the AMM is largely independent of ENSO. In Figure 9, the independence of these variables can be observed in the lack of correlation between the 12 AMM variables and the *June–July Niño 3* (1) variable.

It is also remarkable that the *November 500-mb Geopotential Height* (12) variable was included in the 3 AMM+CSU regression models, as well as the 3 prior CSU regression models [23]. This variable measures the long-term oscillations that impact global wind patterns. Specifically, it represents a different climate signal from the AMM that is called the North Atlantic Oscillation (NAO). Vimont and Kossin [27] state that the AMM and the NAO are independent during the hurricane season which may explain its inclusion in all 3 AMM models. Furthermore, the fact that the geographic regions for variable (12) and the AMM do not overlap contributes to the independence of the variables. Since variable (12) is chosen in all regression case studies, it should be regarded as a significant signal for forecasting and studying hurricane seasons.

4.4 Discussion

As mentioned earlier, conventional visual analytics tools used in climate studies are usually restricted to static plots that lack linkages to statistical processes and are not well-suited to today's complex data sets. MDX overcomes these deficiencies by integrating the statistical and visualization processes for more rapid and creative knowledge discovery. This integration significantly reduced timelines for collaborative analyses with our domain expert, Dr. Fitzpatrick, from days to hours. In addition, the ability to directly interact with the data behind the visualization via dynamic visual queries is a vast improvement over conventional static plots. Dr. Fitzpatrick indicated that MDX made it possible to explore the associations in the climate data sets quicker and more comprehensively than traditional approaches. He also noted that MDX automated and streamlined the main statistical processes that are typically employed in his studies, such as manual multicollinearity analysis.

In addition to spotlighting significant associations in the data set, the graphical statistical indicators provide a visual uncertainty metric, which informs the user. For example, a tight clustering of lines in parallel coordinates might indicate a relatively stable predictor that may yield better results than another predictor with more dispersed lines. These statistical indicators can also directly guide the user, as illustrated in our use of the central tendency and variability measures to partition the distributions into active, normal, and inactive seasons.

5 Conclusion

This research has shown, via a practical climate study, the promise of an enhanced parallel coordinates visual analytics approach with statistical information assistance. The new approach overcomes several limitations of traditional climate analysis techniques and enhances knowledge discovery. In our climate study of hurricane trends, we demonstrated several remarkable associations in the analysis of the AMM and CSU data sets regarding seasonal measures of activity. In the future, we will explore additional statistical processes as well as new techniques to encode the resulting information. These enhanced analysis capabilities demonstrate tremendous potential to improve our understanding of complex environmental phenomenon.

Acknowledgements

This research is sponsored by the Naval Research Laboratory's Select Graduate Training Program, by the National Oceanographic and Atmospheric Administration (NOAA) with grants NA060AR4600181 and NA050AR4601145, and through the Northern Gulf Institute funded by grant NA06OAR4320264. The authors wish to thank the VAST reviewers of this paper for their constructive feedback.

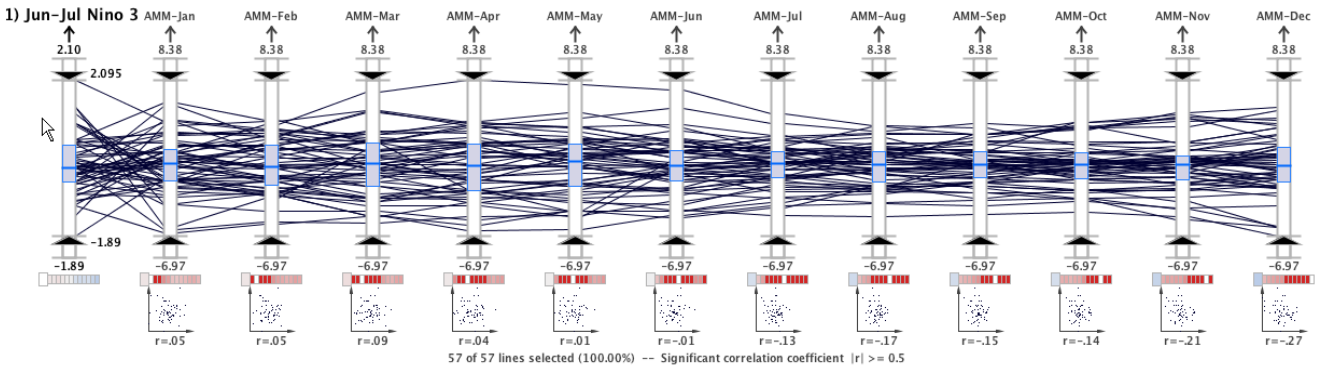


Figure 9: Correlation analysis between the AMM variables and *June–July Niño 3* (1) reveals weak correlations—very lightly saturated correlation blocks, dispersed scatterplots, and low r values—that reinforces that theory that AMM is largely independent of the ENSO.

REFERENCES

[1] Monthly climate timeseries: Atlantic Meridional Mode SST Index. National Oceanic & Atmospheric Administration, 2009. http://www.cdc.noaa.gov/data/timeseries/monthly/AMM/ (current 1 Jul. 2009).

[2] A. O. Artero, M. C. F. de Oliveira, and H. Levkowitz. Uncovering clusters in crowded parallel coordinates visualization. In *IEEE Symposium on Information Visualization*, pages 81–88, Oct. 2004.

[3] P.-S. Chu. ENSO and tropical cyclone activity. In R. J. Murnane and K.-B. Liu, editors, *Hurricanes and Typhoons: Past, Present, and Future*, pages 297–332. Columbia University Press, 2004.

[4] J. A. Dykes and D. M. Mountain. Seeking structure in records of spatio-temporal behavior: Visualization issues, efforts and applications. *Computational Statistics and Data Analysis*, 43(4):581–603, Aug. 2003.

[5] P. J. Fitzpatrick. Understanding and forecasting tropical cyclone intensity change with the Typhoon Intensity Prediction Scheme (TIPS). *Weather and Forecasting*, 12(4):826–846, 1997.

[6] P. J. Fitzpatrick. *Natural Disasters, Hurricanes: A Reference Handbook*. ABC–CLIO, Santa Barbara, CA, 1999.

[7] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *Proceedings of IEEE Visualization*, pages 43–50, Oct. 1999.

[8] H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *Proceedings of IEEE Symposium on Information Visualization*, pages 127–130, Oct. 2002.

[9] C. G. Healey, L. Tateosian, J. T. Enns, and M. Remple. Perceptually-based brush strokes for nonphotorealistic visualization. *ACM Transactions on Graphics*, 23(1):64–96, 2004.

[10] D. C. Howell. *Statistical Methods for Psychology*. Duxbury Press, Boston, MA, 2nd edition, 1987.

[11] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(4):69–91, 1985.

[12] J. Johansson, P. Ljung, M. Jern, and M. Cooper. Revealing structure within clustered parallel coordinates displays. In *IEEE Symposium on Information Visualization*, pages 125–132, Oct. 2005.

[13] C. Johnson, R. Moorhead, T. Munzner, H. Pfister, P. Rheingans, and T. S. Yoo, editors. *NIH/NSF Visualization Reserach Challenges*. IEEE Press, 2006. http://tab.computer.org/vgtc/vrc/index.html (current 1 Jul. 2009).

[14] P. J. Klotzbach. Personal communication, Jan. 2007.

[15] P. J. Klotzbach, W. M. Gray, and W. Thorson. Extended range forecast of Atlantic seasonal hurricane activity and U.S. landfall strike probability for 2007. Technical report, 2006. http://tropical.atmos.colostate.edu/Forecasts/2006/dec2006/ (current 1 Jul. 2009).

[16] J. P. Kossin and D. J. Vimont. A more general framework for understanding Atlantic hurricane variability and trends. *Bulletin of the American Meteorological Society*, 88(11):1767–1781, Nov. 2007.

[17] A. M. MacEachren, M. Wachowicz, R. Edsall, D. Haug, and R. Masters. Constructing knowledge from multivariate spatiotemporal data: Integrating geographic visualization (GVis) with knowledge discovery in database (KDD) methods. *International Journal of Geographical Information Science*, 13(4):311–334, 1999.

[18] A. R. Martin and M. O. Ward. High dimensional brushing for interactive exploration of multivariate data. In *Proceedings of the IEEE Visualization Conference*, pages 271–278, Oct. 1995.

[19] M. Novotńy and H. Hauser. Outlier-preserving focus+context visualization in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):893–900, 2006.

[20] H. Qu, W. Chan, A. Xu, K. Chung, K. Lau, and P. Guo. Visual analysis of the air pollution problem in Hong Kong. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1408–1415, 2007.

[21] H. Siirtola. Direct manipulation of parallel coordinates. In *Proceedings of the International Conference on Information Visualisation*, pages 373–378, London, England, 2000. IEEE Computer Society.

[22] C. A. Steed, P. J. Fitzpatrick, J. Edward Swan II, and T.J. Jankun-Kelly. Tropical cyclone trend analysis using parallel coordinates geovisual analytics. *Cartography and Geographic Information Science*, 36(3), Jul. 2009. (In Press).

[23] C. A. Steed, P. J. Fitzpatrick, T.J. Jankun-Kelly, A. N. Yancey, and J. Edward Swan II. An interactive parallel coordinates technique applied to a tropical cyclone climate analysis. *Computers & Geosciences*, 35(7):1529–1539, Jul. 2009.

[24] R. Theron. Visual analytics of paleoceanographic conditions. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 19–26, 2006.

[25] M. Theus. Interative data visualization using mondrian. *Journal of Statistical Software*, 7(11):1–9, Nov. 2002.

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

[27] D. J. Vimont and J. P. Kossin. The Atlantic Meridional Mode and hurricane activity. *Geophysical Research Letters*, 34:1–5, 2007.

[28] F. Vitart. Dynamical seasonal forecasts of tropical storm statistics. In R. J. Murnane and K.-B. Liu, editors, *Hurricanes and Typhoons: Past, Present, and Future*, pages 354–392. Columbia University Press, Dec. 2004.

[29] R. E. Walpole and R. H. Myers. *Probability and Statistics for Engineers and Scientists*. Prentice Hall, Englewood Cliffs, New Jersey, 5th edition, 1993.

[30] E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85(411):664–675, 1990.

[31] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1366–1372, Nov. 2006.

[32] P. C. Wong and R. D. Bergeron. 30 years of multidimensional multivariate visualization. In G. M. Nielson, H. Hagan, and H. Muller, editors, *Scientific Visualization - Overviews, Methodologies, and Techniques*, pages 3–33. IEEE Computer Society Press, 1997.