Tropical Cyclone Trend Analysis using Enhanced Parallel Coordinates and Statistical Analytics

Chad A. Steed[a], Patrick J. Fitzpatrick[b], J. Edward Swan II[d], and T.J. Jankun-Kelly[c]

[a]Naval Research Laboratory, Mapping, Charting, and Geodesy Branch,
Stennis Space Center, MS, USA, 39529, chad.steed@nrlssc.navy.mil,
(voice) 228-688-4558, (fax) 228-688-4558;

[b]Northern Gulf Institute, Mississippi State University,
Stennis Space Center, MS, 39529, fitz@ngi.msstate.edu;

[c]Department of Computer Science and Engineering, Mississippi State University,
Mississippi State, MS, 39762, tjk@acm.org;

[d]Department of Computer Science and Engineering,
Mississippi State University, Mississippi State, MS, 39762, swan@acm.org.

July 23, 2009

Abstract

This work presents, via an in-depth case study, how parallel coordinates coupled with statistical analysis can be used for more effective knowledge discovery and confirmation in complex, environmental data sets. Advanced visual interaction techniques such as dynamic axis scaling, conjunctive parallel coordinates, statistical indicators, and aerial perspective shading are combined into an interactive geovisual analytics system. Moreover, the system facilitates statistical processes such as stepwise regression and correlation analysis to assist in the identification and quantification of the most significant predictors for a particular dependent variable. Using a systematic workflow, this approach is demonstrated via a North Atlantic hurricane climate study in close collaboration with a domain expert. By revealing several important physical associations, the case study reveals that the visual analytics approach facilitates a deeper understanding of multidimensional climate data sets when compared to traditional techniques.

1 Introduction

One of the most challenging tasks in multivariate data analysis is to identify and quantify the associations between a set of interrelated variables. In climate studies, this task is even more daunting due to the uncertainty and complexity of dynamic, environmental data sets. Notwithstanding the difficulty, the variability and destructiveness of recent hurricane seasons has invigorated efforts by weather scientists to identify environmental variables that have the greatest impact on the intensity and frequency of seasonal hurricane activity. In general, the goal of such efforts is to improve the accuracy of seasonal forecasts which should, in turn, improve preparedness and reduce the impact of these devastating natural disasters.

Traditional visual data analysis tools used in climate studies do not support the increasing quantity and number of different parameters in the data. Consequently, scientists are forced to reduce the problem in order to fit the tools, thereby limiting the insight that can be obtained from today's wealth of data. To overcome these limitations, this research brings to bear a new interactive geovisual analytics approach that is based on automated statistical analysis to enhance knowledge

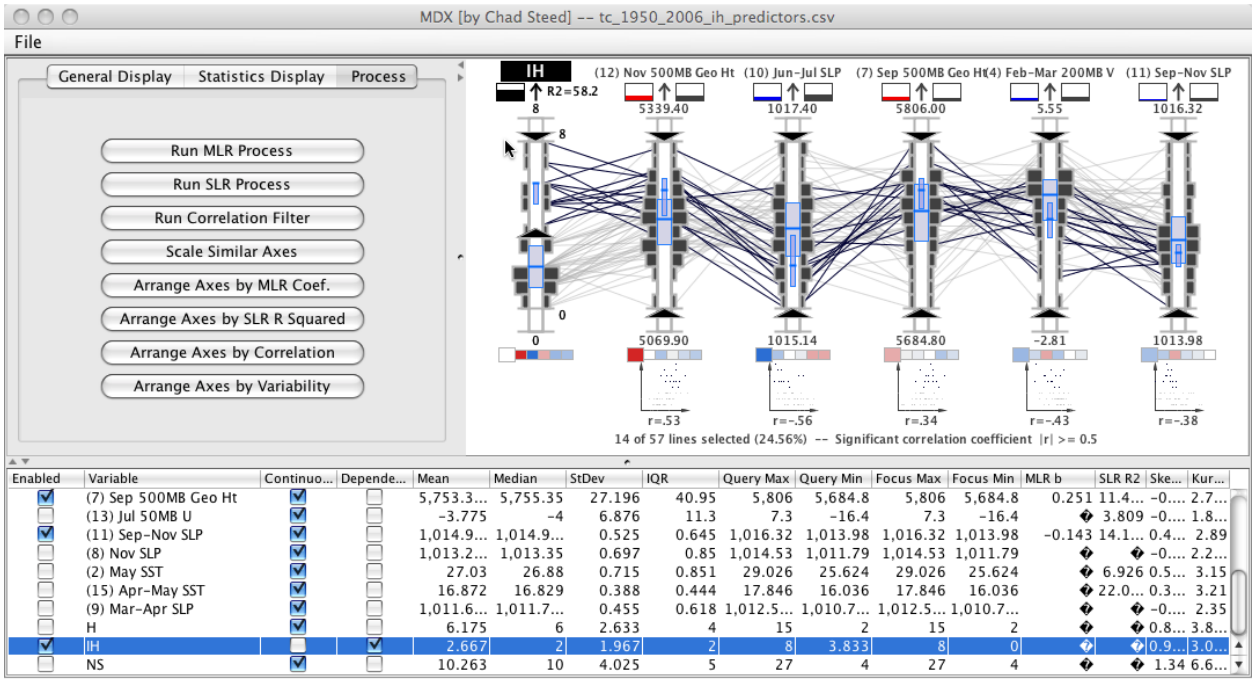


Figure 1: The Multidimensional Data eXplorer (MDX) is composed of a settings panel (upper left), enhanced parallel coordinates plot (upper right), and a table view panel (lower).

discovery. As shown in Fig. 1, the resulting system, which is called the Multidimensional Data eXplorer (MDX), combines several fundamental parallel coordinates capabilities and variants of more advanced techniques from prior works with new interactive capabilities and statistical analysis processes. In this paper, it will be shown that statistical analytics and interactive parallel coordinates can compliment each other, with the statistical processes highlighting the relevant associations and the parallel coordinates providing a deeper level of understanding about the relationships.

In the current work, the promise of this approach is demonstrated via detailed analysis of potential predictors for topical cyclone activity. By significantly expanding earlier work [Steed et al., 2008], the new MDX system provides a comprehensive environment for climate analysis. Therefore, the main contributions of this work to geographic visualization are as follow:

- A unique geovisual analytics system has been formulated that combines several interactive parallel coordinates capabilities with automated statistical processes to meet the needs of complex climate studies.

- The results of a comprehensive tropical cyclone trend analysis case study reveal several im-

portant physical associations in the environmental predictor data set, thereby highlighting the promise of a geovisual analytics approach for facilitating a deeper understanding of the data.

Furthermore, this research fulfills the NIH/NSF Visualization Challenges Report recommendation 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 [Johnson et al., 2006]" through the inclusion of a hurricane expert, Dr. Patrick Fitzpatrick, who is also the second author of this paper. The research team also includes an earth scientist who specializes in geovisual analytics and two computer graphics and data visualization professors.

2 Background

One particularly useful method for predicting seasonal hurricane variability is based on the idea that there are predictors of the main dynamic parameters that affect storm activity which can be observed up to a year in advance. Using historical data, their importance is estimated using statistical regression techniques similar to those described by Vitart [2004]. Klotzbach et al. [2006] used this technique to determine the most important variables for predicting the frequency of North Atlantic tropical cyclone activity. Similarly, Fitzpatrick [1996] developed a multiple regression scheme call the Typhoon Intensity Prediction Scheme (TIPS) to understand and forecast tropical cyclone intensity for the western North Pacific Ocean. TIPS represented the first tropical cyclone multiple regression scheme that combined satellite information with other environmental predictors; and it revealed the vital information contained in the satellite data that distinguishes between fast and slow developing tropical cyclones. Although sometimes complicated to establish, regression analysis techniques provide an ordered list of the most important predictors for the dynamic parameters. Scientists gain additional insight and identify the more informative variables in these studies by evaluating descriptive statistics and performing correlation analysis.

In conjunction with statistical analysis, researchers have relied on simple scatter plots and histograms which require several separate plots or layered plots to analyze multiple variables. Using

separate plots, however, is not an optimal approach in this type of analysis due to perceptual issues described by Healey et al. [2004] such as the extremely limited memory for information that can be gained from one glance to the next. These issues are illustrated through the so-called change blindness (a phenomenon described by Rensink [2002]) and they are especially detrimental when searching for combinations of conditions. A potential solution often employed by statisticians is the scatterplot matrix (SPLOM), which presents multiple adjacent scatterplots for all the variable comparisons in a single display with a matrix configuration [Wong and Bergeron, 1997]; but it requires a large amount of screen space and forming multivariate associations is still mentally challenging. Wilkinson et al. [2006] used statistical measures to organize the SPLOM and guide the viewer through exploratory analysis of high-dimensional data sets. The effectiveness of the SPLOM technique is improved but the perceptual issues mentioned above remain to some degree. The application of these sorting methods to the parallel coordinates plot is also briefly demonstrated using weather data. Another alternative is to use layered plots, which condense the information into a single display; but there are significant issues due to layer occlusion and interference as demonstrated by Healey et al. [2004].

Furthermore, the geographically-encoded data used in climate studies are usually displayed in the context of a geographical map; although certain important patterns (those directly related to geographic position) may be recognized in this context, additional information may be discovered more rapidly using non-geographical information visualization techniques. Few multivariate visualization techniques provide access to integrated, automatic statistical analysis techniques that are commonly utilized in climate studies to identify significant associations. To compensate for these deficiencies, new visualization methods are needed that intelligently integrate statistical processes and accommodate the simultaneous display of real-world, multivariate data.

In the current work, a popular multivariate visualization technique, called parallel coordinates, forms the basis of an approach designed to address these needs. The parallel coordinates concept was first introduced by Inselberg [1985] to represent hyper-dimensional geometries. Later, Wegman [1990] applied the technique to the analysis of multivariate relationships in data. In general, the technique yields a compact two-dimensional representation of even large multidimensional data sets.

Since the introduction of parallel coordinates, several innovative extensions have been described in the visualization research literature. For example, Hauser et al. [2002] described a histogram display, dynamic axis re-ordering, axis inversion, and some details-on-demand capabilities for parallel coordinates. In addition, Siirtola [2000] presented a rich set of dynamic interaction techniques (e.g., conjunctive queries) and T.J. Jankun-Kelly and Waters [2006] and Johansson et al. [2005] described new line shading schemes for parallel coordinates. Furthermore, several focus+context implementations for parallel coordinates have been introduced by Fua et al. [1999], Artero et al. [2004], Johansson et al. [2005], and Novotńy and Hauser [2006]. More recently, Qu et al. [2007] introduced a method for integrating correlation computations into a parallel coordinates display. The MDX system described in the following section utilizes variants of these extensions to the classical parallel coordinates plot.

In Seo and Shneiderman [2005], a framework is used to explore and comprehend multidimensional data using a powerful rank-by-feature system, that guides the user and supports confirmation of discoveries. Recently, Piringer et al. [2008] expanded this rank-by-feature approach with a specific focus on comparing subsets in high-dimensional data sets. The MDX system is designed to support a similar rank-by-feature framework with subset selection capabilities using stepwise regression and interactive visual analysis.

3 Climate Analysis Workflow

To facilitate the development of a geovisual analytics approach for climate analysis, a systematic workflow has been formulated to capture the main tasks. This workflow is depicted in the system context diagram shown in Fig. 2. Although this workflow is described in a sequential order, typical analysis involves several iterations and moving between the various processes

After preparing and loading the data set into the system, the scientist will manually filter the data to remove unnecessary variables. Then, descriptive statistical indicators will be employed to acquire a preliminary overview of the entire data set. In this initial exploratory analysis phase, graphical statistical indicators and variable comparison capabilities provide vital assistance to the scientist.

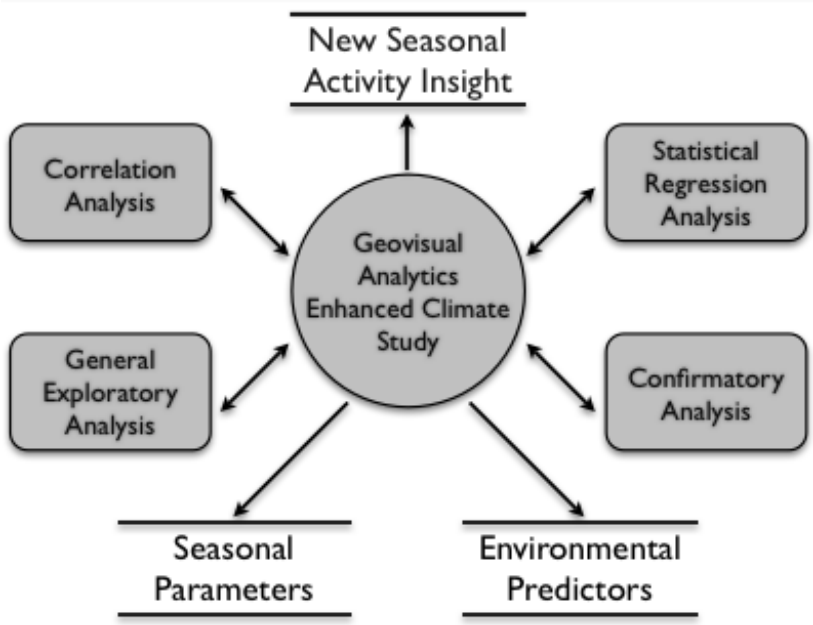


Figure 2: Climate study system context diagram depicting the climate analysis workflow.

Next, the scientist will observe the statistical correlations in the data using the correlation indicators and axis arrangement methods. In this stage, the scientist will prepare for regression analysis by identifying the most informative variables and taking steps to reduce multicollinearity among the independent variables. Automated filters for ensuring the independence between predictors and identifying significant associations are key capabilities in this stage. The scientist will gain additional insight in this phase by observing correlations between the predictors as well as correlations between each predictor and the dependent variable.

After the correlation analysis, the scientist will utilize regression analysis capabilities to continue investigation of the data set. Simple Linear Regression (SLR) processes quantitatively indicate the individual associations between predictors and the dependent variable. Multiple Linear Regression (MLR) is also used to quantify the significance of several predictors for a dependent variable. The result of the MLR process is a ranked list of the most important variables. Unlike the SLR and correlation analysis, the MLR analysis considers the contribution in relation to the other predictors.

By following this workflow, the scientist will develop new ideas about how the specific variables can be used to predict the dependent variable. That is, the scientist will have formed hypotheses about the associations between the variables. Then, the scientist can continue to explore the data

to attempt to prove or disprove the new hypotheses; a process that Tukey [1977] calls confirmatory data analysis. For example, the scientist can analyze the data from 1950 to 2006 to discover patterns that can be used to predict above normal hurricane activity. These patterns would be identified by selecting specific ranges for a set of variables that coincide with above normal hurricane seasons. The scientist can then verify the accuracy of the patterns by examining its performance in particular years of interest, such as the active 2005 season or the below normal 2006 season. If the patterns perform well, the scientist may use the new insight in future forecasts.

4 Multidimensional Data Explorer

The MDX system described in this paper provides an innovative geovisual analytics interface by combining variants of previously introduced interactive parallel coordinates techniques with statistical indicators and automated analysis processes. MDX extends an earlier version of the software described by Steed et al. [2008], which focused on the visual analysis of regression models, to include more statistical indicators, correlation analysis capabilities, and better integration with regression processes. The implementation of the new MDX system facilitates subsequent effectiveness evaluations of the geovisual analytics approach to the investigation of climate trends.

4.1 Interactive Visualization Capabilities

MDX provides a powerful parallel coordinates interface by fusing variations of several interactive visualization techniques. Fundamental parallel coordinates extensions such as frequency information display, dynamic axis re-ordering, axis inversion, and details-on-demand are included. In addition, MDX provides advanced capabilities derived from recent research publications. For example, the double-ended sliders on each axis (see Fig. 3) facilitate the investigation of subsets of data in a manner similar to the approach presented by Siirtola and Räihä [2006]. The user can drag these sliders to dynamically adjust which lines are highlighted. Lines that fall within the sliders for all the axes are rendered in a more prominent manner giving the user the ability to perform rapid Boolean AND selections.

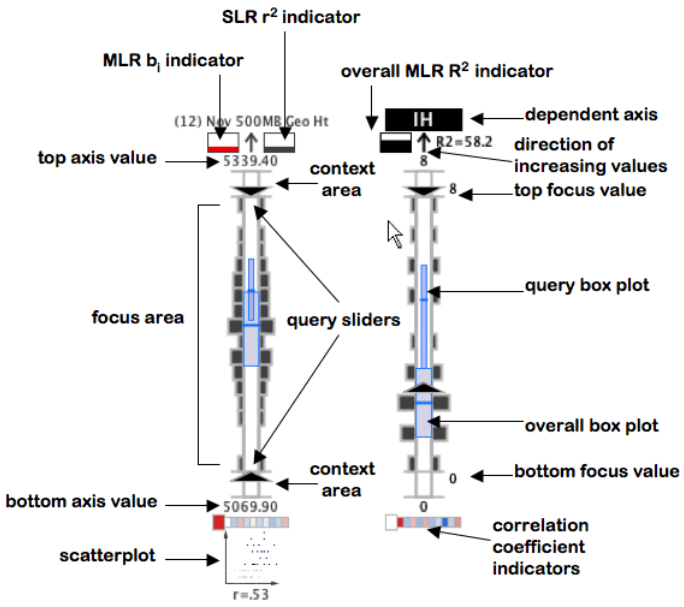
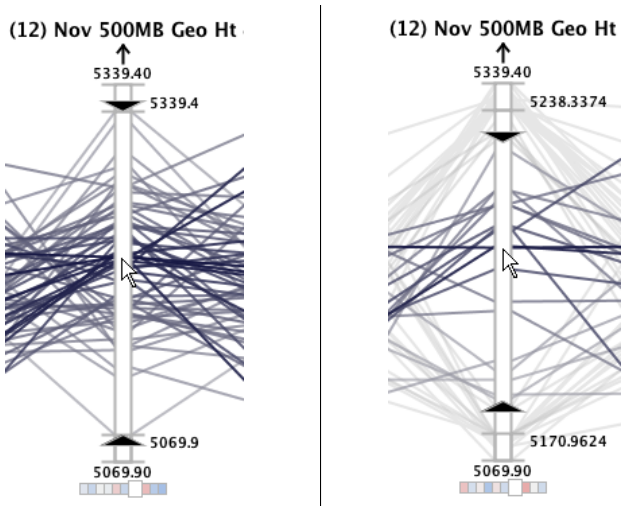


Figure 3: An annotated view of the enhanced parallel coordinate axis component in MDX highlighting the visual interaction components and the statistical indicators. The axis on the left shows the normal axis shading while the axis on the right illustrates the shading for a highlighted, dependent axis.

4.1.1 Dynamic Axis Scaling

MDX's dynamic axis scaling capability provides a method to interactively tunnel through the data until a smaller subset of the original data is in focus. This capability allows the user to modify the scale (minimum and maximum values) for a focus area of a selected axis using mouse wheel movement—a unique approach to dynamic axis scaling. As shown in Fig. 3, 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 is hovering over the central area, an upward mouse wheel motion causes a smaller range of values to be displayed in the focus area. This action expands the display of the focus area outward and pushes extreme values into the two context areas. A downward mouse wheel motion causes the inverse effect: focus region compression. The user may also use the mouse wheel over either of the two context areas to alter the minimum or maximum values separately. Furthermore, the user may also manually enter the minimum and maximum values by typing them in the appropriate fields of the table view panel (see Fig. 1). This intuitive axis scaling capability helps to free space and reduce line clutter, thereby making it easier to analyze relation lines of interest.



(a) Before scaling (b) After scaling

Figure 4: 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, effectively zooming into the central axis region.

4.1.2 Aerial Perspective Shading

Expanding prior proximity shading techniques for parallel coordinates [Johansson et al., 2005, T.J. Jankun-Kelly and Waters, 2006], MDX also provides an enhanced line shading scheme that enables rapid monitoring of trends due to the similarity of data values over multiple dimensions. This shading scheme seeks to simulate the human perception of aerial perspective whereby the contrast between objects and the background decreases as the distance between the viewer and the object increases. In this implementation, aerial perspective shading can be used in either a discrete (see Fig. 1) or a continuous mode (see Fig. 4). In the discrete mode, the lines are colored according to the axis region that they intersect. If any point of a relation line is in the context (non-focus) area of at least one axis, the line is shaded with a light gray color and drawn beneath the non-context lines. If all the points on a relation line fall within the query area of each axis (the area between the two query sliders), the line is colored using a dark gray value that attracts the viewer's attention and the remaining lines (non-query and non-context) are colored a shade of gray that is slightly darker than the context lines but lighter than the query lines. The resulting discrete shading effect is illustrated in Fig. 1 and Fig. 6 where the upper range of the *IH* axis is queried.

In the continuous mode, non-context lines go through an additional step to encode the distance of the line from the mouse cursor (see Fig. 4). 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. The proximity shading scheme in MDX supports rapid investigation of multidimensional relationships, thereby enhancing the scientists ability to achieve a deeper understanding of the environmental data.

4.2 Statistical Analysis Capabilities

The interactive visualization capabilities in MDX are complemented with statistical indicators and automated statistical analysis capabilities that provide guided exploratory analysis capabilities. These statistical analysis capabilities help spotlight the significant associations in the data thereby enhancing knowledge discovery.

4.2.1 Descriptive Statistical Indicators

To support the interactive analysis capabilities of MDX, each axis graphically represents key descriptive statistics (see Fig. 3). In addition to representing frequency information, the axes represent central tendency (median or mean) and variability (standard deviation or interquartile range) statistics in the form of box plots for the data in the central focus area of each axis and a histogram display. As shown in Fig. 1, the wide box plots represent the descriptive statistics for all the axis samples while the more narrow query box plots represent the samples that are currently selected with the axis query sliders.

4.2.2 Correlation Analysis

MDX also provides graphical indicators that facilitate correlation analysis. Correlation analysis helps one to measure the strength of the relationships between two variables by using a single

number, which is called the correlation coefficient. Specifically, MDX uses the Pearson product-moment correlation coefficient, r, to estimate the correlation between two variables. Graphical indicators that encode r between axis pairs are represented beneath each axis in a set of color-coded blocks (see Fig. 3). Each block uses color to encode the correlation coefficient between the axis directly above it and the axis that corresponds to its position in the set of blocks. For example, the first block in the indicators shown in Fig. 1 represents the correlation between the axis above it and the first axis, *IH*. When the mouse hovers over an axis, the axis is highlighted and its corresponding correlation indicator blocks are enlarged. Furthermore, the blocks are shaded blue for negative correlations and red for positive correlations. The stronger the correlation, the more saturated the color. When the absolute value of a correlation coefficient is greater than or equal to the significant correlation threshold, the block is shaded with the fully saturated color. The significant correlation threshold is a user-defined value that is displayed at the bottom of the parallel coordinates panel (see Fig. 1).

In addition to the correlation indicators, MDX displays small scatterplots below the axes when an axis is highlighted. These 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. The scatterplots provide a visual means to quickly confirm the type and strength of the correlation as well as a straightforward representation to investigate non-linear or multi-modal distributions.

The correlation analysis features facilitate complex tasks such as manual multicollinearity filtering. Furthermore, an automatic multicollinearity filter is available; if any axes are correlated with each other by more than the significant correlation threshold, one axis (the axis with the weaker correlation coefficient with the dependent axis) is automatically removed from the display. After the application of this filter, the remaining axes are truly independent of one another.

4.2.3 Regression Analysis

In addition to SLR, MDX offers stepwise MLR capabilities with a backwards glance which selects the optimum number of the most important variables using a predefined significance level. MDX executes the MATLAB *regress* and *stepwisefit* utilities that perform simple and stepwise regression,

respectively. These regression capabilities complement the visual analysis capabilities of MDX by isolating the significant variables in a quantitative fashion. In the MLR analysis, a normalization procedure is used so that the y-intercept becomes zero and the importance of a predictor can be assessed by comparing regression coefficients, b_i. Denoting σ as the standard deviation of a variable, y as the dependent variable, $\overline{x}$ as the predictor mean, and $\overline{y}$ as the dependent variable mean, a number k of statistically significant predictors are normalized by the following equation:

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

As shown in Fig. 3, MDX visually encodes the MLR coefficients in the parallel coordinates plot using the boxes below the axis label and to the left of the arrow. Like a thermometer, the box is filled from the bottom to the top based on the magnitude of the coefficient. The box is colored red if the coefficient is positive and blue if it is negative. The box to the right of the arrow encodes the r^2 output from the SLR process. In addition to the regression coefficients, the MLR analysis returns the coefficient of multiple determination, R^2, which quantitatively captures the variance explained by the predictors for the dependent variable. The box beneath the dependent variable axis name encodes the R^2 value from the MLR analysis (see Fig. 3).

4.2.4 Automatic Axis Arrangement

MDX can also automatically arrange the axes in the parallel coordinates plot using one of the dynamically computed statistical measures previously mentioned. That is, the user can choose to sort the axes by one of the following statistical measures: correlation coefficient, interquartile range / standard deviation range, MLR coefficient, or SLR r^2 value. When sorting by the correlation coefficients, the axes with negative correlations are arranged to the left of the highlighted axis in ascending order and axes with positive correlations are arranged to the right in descending order (see Fig. 7). When sorting by the other statistical measures, the dependent axis is placed at the leftmost position and the other axes are arranged in descending order based on the selected measures.

Table 1: Tropical cyclone climate variables evaluated as predictors in the climate study.

	Variable Name	Geographical Region
(1)	June–July Niño 3	5S-5N, 90-150W
(2)	May SST	5S-5N, 90-150W
(3)	February 200-mb U	5S-10N, 35-55W
(4)	February–March 200-mb V	35-62.5S, 70-95E
(5)	February SLP	0-45S, 90-180W
(6)	October–November SLP	45-60N, 120-160W
(7)	Sept. 500-mb Geopotential Height	35-55N, 100-120W
(8)	November SLP	7.5-22.5N, 125-175W
(9)	March–April SLP	0-20N, 0-40W
(10)	June–July SLP	10-25N, 10-60W
(11)	September–November SLP	15-35N, 75-97W
(12)	Nov. 500-mb Geopotential Height	67.5-85N, 50W-10E
(13)	July 50-mb U	5S-5N, 0-360
(14)	February SST	35-50N, 10-30W
(15)	April–May SST	30-45N, 10-30W
(16)	June–July SST	20-40N, 15-35W

SST – sea surface temperature
SLP – sea level pressure
U – wind zonal component (or x-coordinate)
V – wind meridional component (or y-coordinate)

5 A Tropical Cyclone Trend Analysis Case Study

This research is motivated by the notion that a geovisual analytics approach will reveal a deeper level of understanding about the data than traditional techniques. The effectiveness of this approach has been studied by utilizing MDX in the context of a practical tropical cyclone climate study, conducted in close collaboration with a hurricane expert, Dr. Patrick Fitzpatrick, who, in addition to being the second author of this paper, is the author of two books and several other publications on tropical cyclones. The discovery of new associations and the confirmation of known patterns in this data set reveals the promise of this approach for improving knowledge discovery in environmental data analysis.

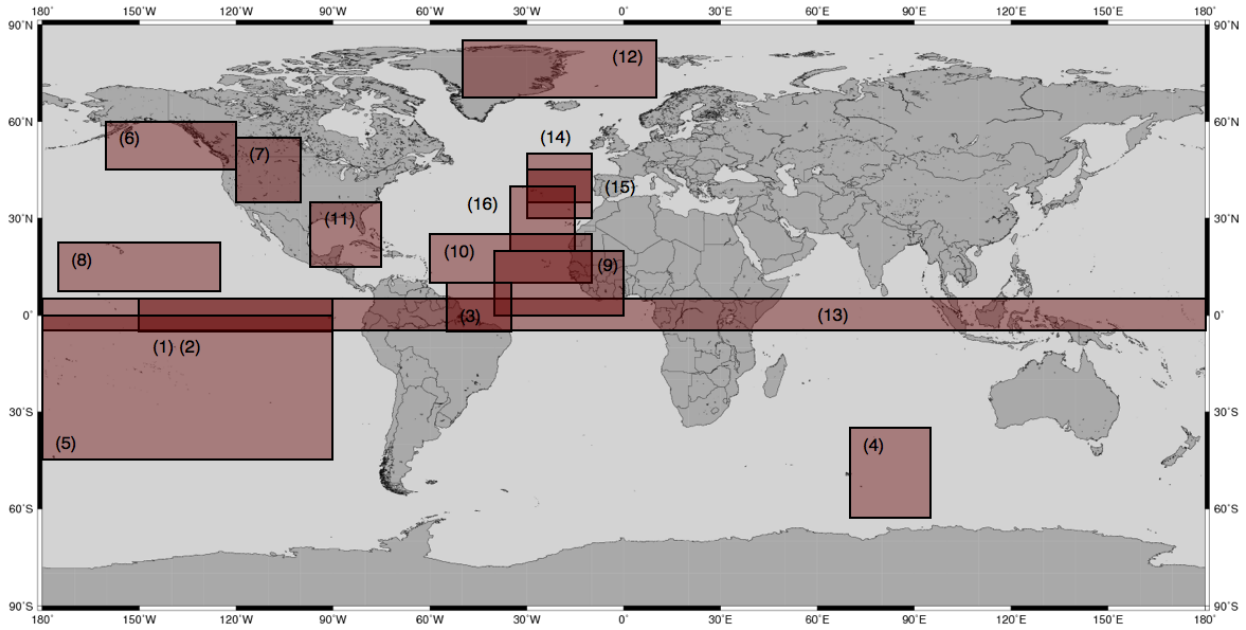


Figure 5: Geographical regions for the tropical cyclone climate predictors.

5.1 Climate Study Data

In this climate study, a data set that contains potential environmental predictors observed annually from 1950 to 2006 (57 records) has been analyzed. Table 1 lists the sixteen potential environmental predictors from this data set along with their geographical region. In Fig. 5, the geographical regions of the predictors are plotted in a world map. This data set has been provided by Dr. Phil Klotzbach [Klotzbach, 2007] of the Tropical Meteorology Project at Colorado State University, where it has been used to predict 2007 North Atlantic tropical cyclone activity by categories. Although many categories are considered in practice, the focus of this study is on the number of intense hurricanes (IH) in a tropical cyclone season. A hurricane is classified as intense when its sustained 10-meter winds are at least 49 ms^{-1} [Fitzpatrick, 1999]. Although intense hurricanes account for just over 20% of the tropical storms and hurricanes that strike the United States, these storms warrant special attention because they are responsible for over 80% of the damage [Goldenberg et al., 2001].

These variables have known relationships to Atlantic tropical cyclone activity. For example, Chu [2004] describes how the North Atlantic basin has fewer tropical cyclones during El Niño Southern

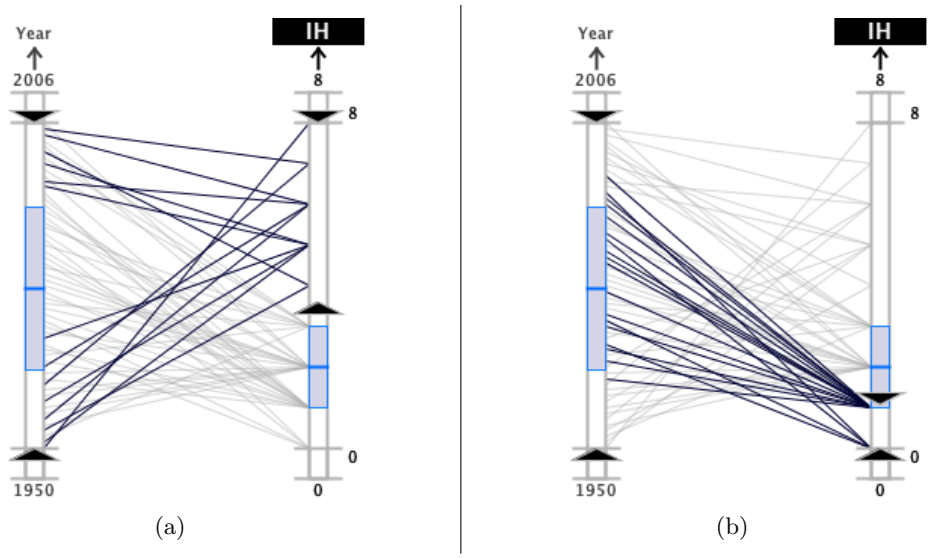


Figure 6: A temporal trend is discovered by visually comparing the years with above average *IH* activity (a) to years with below average activity (b). The gap shown in (a) is filled by below normal seasons in (b) revealing the multidecadal hurricane activity cyclone.

Oscillation (ENSO) years, and active seasons in La Niña years. Because of this relationship, scientists use ENSO signals as one of the main predictors for seasonal storm activity. In Table 1, variables 1 through 8 are believed to characterize ENSO events. An expanded discussion of the parameters in this data set is provided in the previous work [Steed et al., 2008], that describes the initial results of this research.

5.2 Initial Analysis (Overview)

After loading the predictors and seasonal storm statistics, the MDX visual analysis capabilities are used to explore the data set and rearrange the axes. Since the objective of this study is to use the climate variables to predict seasonal activity, the overall axis box plot on the *IH* axis is used to identify the seasons with normal activity. That is, the seasons that intersect the overall box plot of this axis are classified as normal. Then, the query sliders are used to investigate the behavior of each axis in above normal and below normal seasons.

By interactive visual queries on the *Year* axis, a temporal trend is discovered in the initial exploration. When the above average seasons are queried, a gap is visible on the *Year* axis (see Fig. 6(a)). This gap reveals that from 1960 to 1994, a period of normal to below normal activity is

evident since there are no seasons with an above normal number of intense hurricanes. What's more, a query of the below normal seasons reinforces this discovery since these seasons are clustered into this same time period (see Fig. 6(b)). This visual finding is validated by recent findings presented in weather research literature [Goldenberg et al., 2001, Klotzbach and Gray, 2006, Klotzbach et al., 2006] that suggest a strong multidecadal variability in the number of intense hurricanes per year in the North Atlantic.

5.2.1 Correlation Analysis

Correlation analysis of the data set reveals several important physical associations. After arranging the predictor axes by the correlation coefficient with the *IH* axis, the correlation indicators reveal that two variables have correlations above the significant correlation threshold (0.5 for this study): *June–July SLP in the tropical Atlantic* (10) and *November 500–mb Geopotential Height in the far North Atlantic* (12). Fig. 7 shows a portion of the resulting plot after the axis arrangement showing the eight strongest correlations. In this figure, the enlarged color-coded correlation indicator box, polyline 'X'-shaped polyline crossings, downward slope in the scatterplot, and numerical display of the correlation coefficient reveal that axis (10) has the strongest negative correlation. Likewise, the strongest positive correlation with axis (12) is evident by the correlation indicator, fewer polyline crossings, the upward slope of the scatterplot, and the numerical display of the correlation coefficient.

The image sequence shown in Fig. 8 illustrates the use of the continuous aerial perspective shading capability to investigate a strong negative correlation between *October–November SLP in the Gulf of Alaska* (6) and *November SLP in the Subtropical NE Pacific* (8) axes. This intuitive visual query technique, which shades the polylines according to their proximity to the mouse cursor, highlights the 'X'-shaped polyline crossings between the axes, which is indicative of a negative correlation in parallel coordinates.

In Fig. 9(a), the correlations between three SST variables are shown. This plot reveals that a relatively strong positive correlation exists between *April–May SST off the northwestern European coast* (15) and both the *February SST off the northwestern European Coast* (14) and the *June–July*

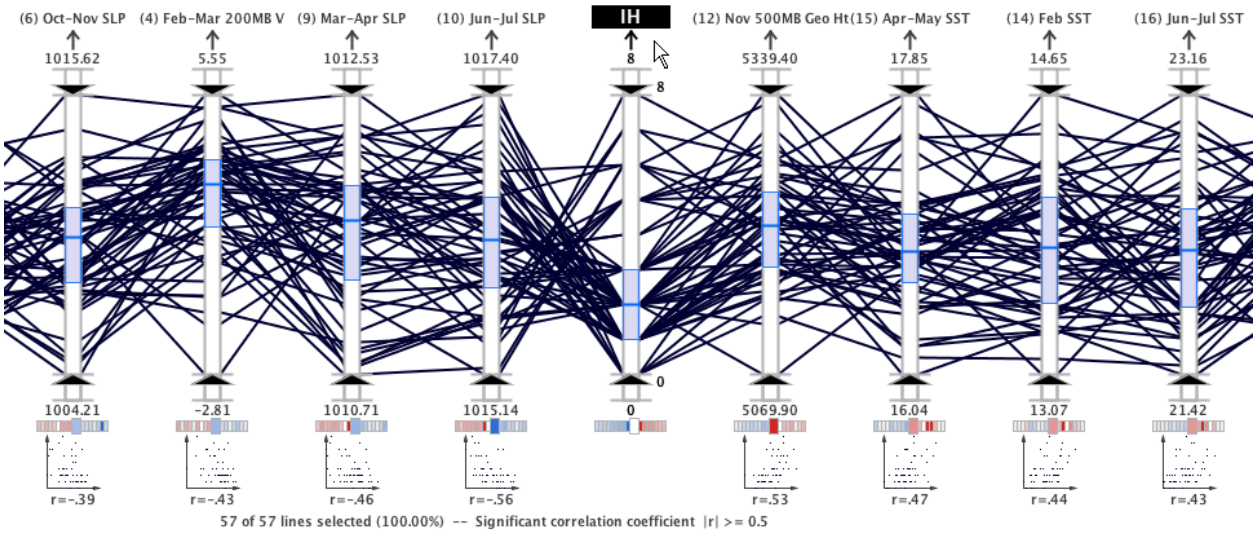


Figure 7: Portion of plot used in *IH* axis correlation analysis. The axes have been automatically arranged according to the magnitude of the correlation coefficients with the dependent axis. Negatively correlated axes are placed to the right of the dependent axis, *IH*, while positively correlated axes are placed to the left.

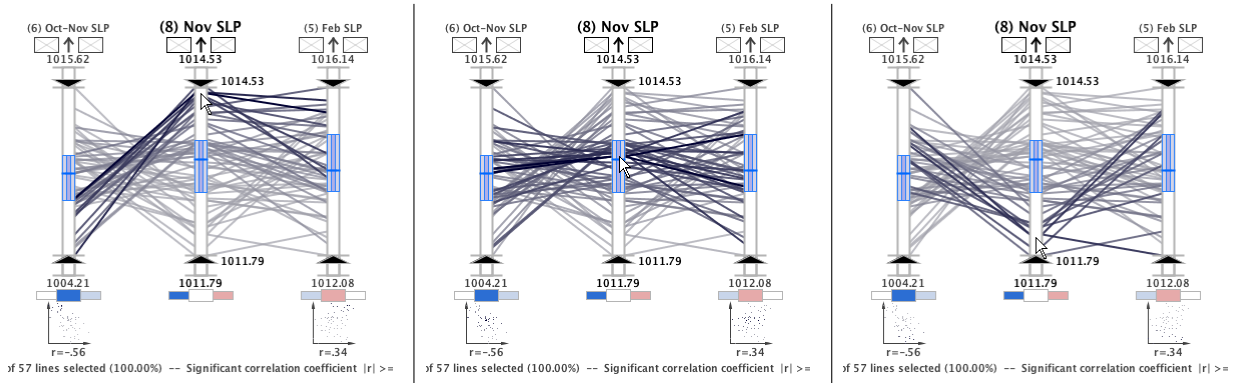


Figure 8: The sequence of images demonstrates how the aerial perspective shading can be used to analyze the SLP variable correlations by moving the move from the top (a) to the bottom (c) of axis (8) through the middle (b).

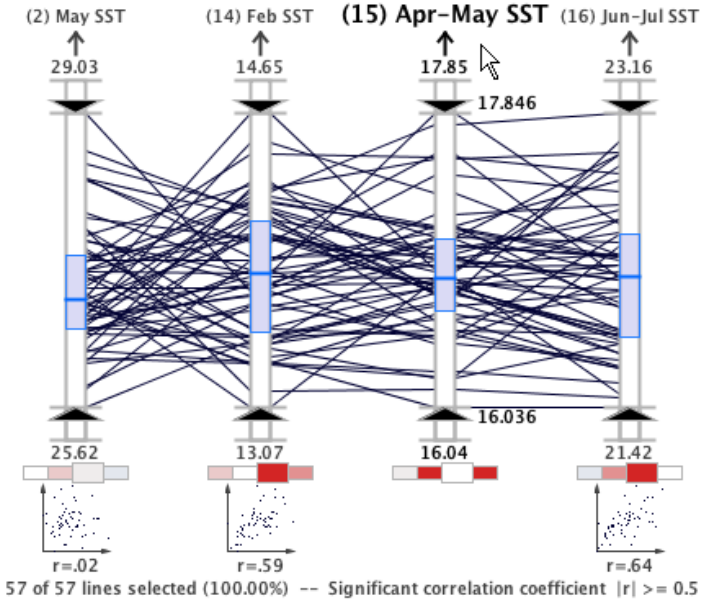
SST in the northeastern subtropical Atlantic (16) axes. Meanwhile, the *May SST in the eastern equatorial Pacific* (2) axis exhibits almost no correlation ($r = .02$). To reduce the multicollinearity between the SST predictors, axis (14) and (16) must be removed since they have a strong correlation with axis (15) and axis (15) has a stronger correlation with the *IH* axis. Removing these and any other variables with strong correlations between predictors will ensure the independence of the predictors and thus improve subsequent regression analysis.

Before removing axis (14) and (16), the physical relationships between these variables can be considered in order to investigate the cause of the strong correlation. From the geographic extents of these variables (see Fig. 9(b)) one can observe that the three SST predictors with strong correlations are all sampled in the North Atlantic Ocean. However, axis (2), which exhibits a very weak correlation, is measured in the Pacific Ocean. Therefore, the strong correlations among axis (14), (15) and (16) can be mostly attributed to the close geographical proximity of the measurements whereas the low correlation of axis (2) can be attributed to the fact that it is measured in the Pacific Ocean. The scientist can continue to manually find and eliminate the highly correlated predictors, or use MDX's automatic multicollinearity filter. The automatic filter removes six predictor axes from the display leaving a total of ten independent axes: (9), (14), (16), (8), (1), and (3).

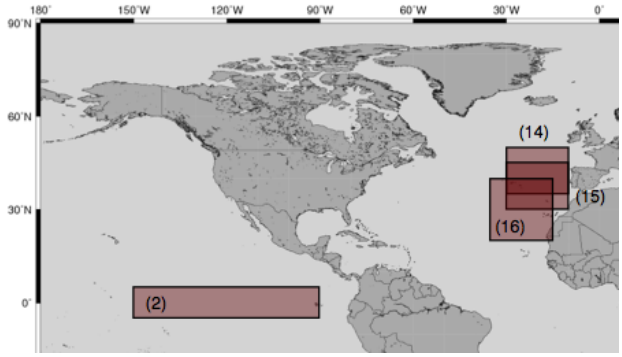
5.2.2 Identifying Most Important Predictors

Using MDX's automatic SLR and stepwise MLR processes, the predictors are statistically analyzed to determine the most important predictors with respect to the number of intense hurricanes in a season. The significance level in the stepwise regression analysis is 80%. In Fig. 10, the results of the MLR and SLR analysis are shown with the axes arranged according to the magnitude of the normalized MLR coefficient. In this figure, only the axes selected in the regression model are displayed. Furthermore, the above normal seasons are queries, which reveals how each specific variable behaves in the most destructive years.

We can use the above normal season query to develop a prediction model that is based on the ranges of these significant environmental measurements. First, the query sliders are used to interactively formulate the model which consists of normal to above normal values on axes (12) and



(a)



(b)

Figure 9: SST correlation analysis using the statistical indicators (a) reveals strong positive correlation of variable (15) with variables (14) and (16). Analysis of the geographic extents for these variables (b) reveal the primary cause of these associations is tied to the location where variables are measured.

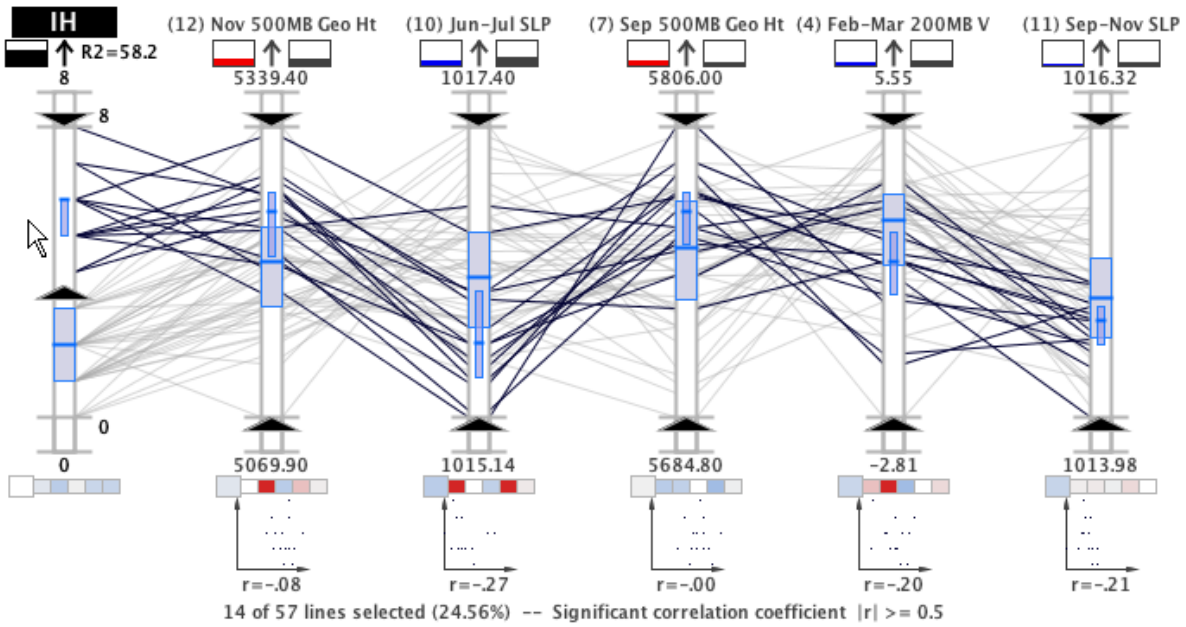


Figure 10: *IH* regression model for active seasons (1950 to 2006) with the axes arranged in descending order based on regression coefficients. The ordering reflects the significance of the variables for predicting seasonal *IH* activity.

(7), as well as normal to below normal ranges on axes (10), (4), and (11). Using these predictor ranges would have resulted in successfully identifying eleven of the fourteen seasons (74%) that had an above normal number of intense hurricanes between 1950 and 2006. On the other hand, the ranges omit three above normal activity seasons (with 7, 6, and 5 intense hurricanes). In particular, one of the storm seasons that is not selected is the infamous 2005 hurricane season which had seven intense hurricanes, including the devastating Hurricane Katrina. Using the visual query capabilities, minor adjustments can be applied to the query sliders of the significant predictors to ensure that all fourteen seasons with active intense hurricane activity are captured (see Fig. 11). The predictor ranges are listed numerically in Table 2. Then, these predictor ranges can be used to predict the activity of future tropical cyclone seasons with respect to the number of intense hurricanes.

In Fig. 12, the query sliders are reset on the *IH* axis to include all storm years that fall within the refined predictor ranges. From this figure, it is evident that an additional eleven seasons are falsely identified as seasons with above normal *IH* activity when, in fact, the seasons show normal activity levels. These eleven seasons represent type I errors (also called false positives) where the test returns a positive result when the actual condition is absent. From the perspective of human

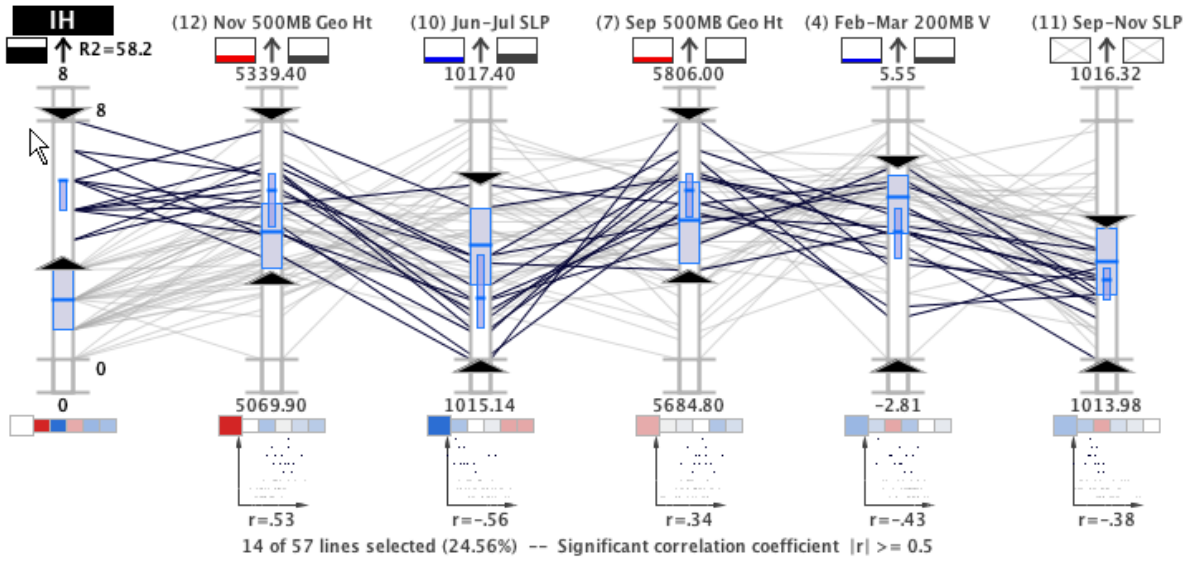


Figure 11: Visually tweaked ranges for forecasting intense hurricane activity using the predictors that the regression model identified as the most important parameters. These variables ranges can be evaluated as a prototype forecast model.

Table 2: Above normal *IH* activity predictor ranges. The bold ranges are those that have been visually adjusted.

Predictor Name	Initial Range	Tweaked Range
Nov. 500-mb Geopot. Ht. (12)	$>$ 5170.00	$>$ 5170.00
June–July SLP (10)	$<$ 1016.60	$<$ **1016.78**
Sep. 500–mb Geopot. Ht. (7)	$>$ 5732.00	$>$ **5730.50**
Feb.–Mar. 200-mb V (4)	$<$ 3.62	$<$ **3.87**
Sep.–Nov. SLP (11)	$<$ 1015.26	$<$ 1015.26

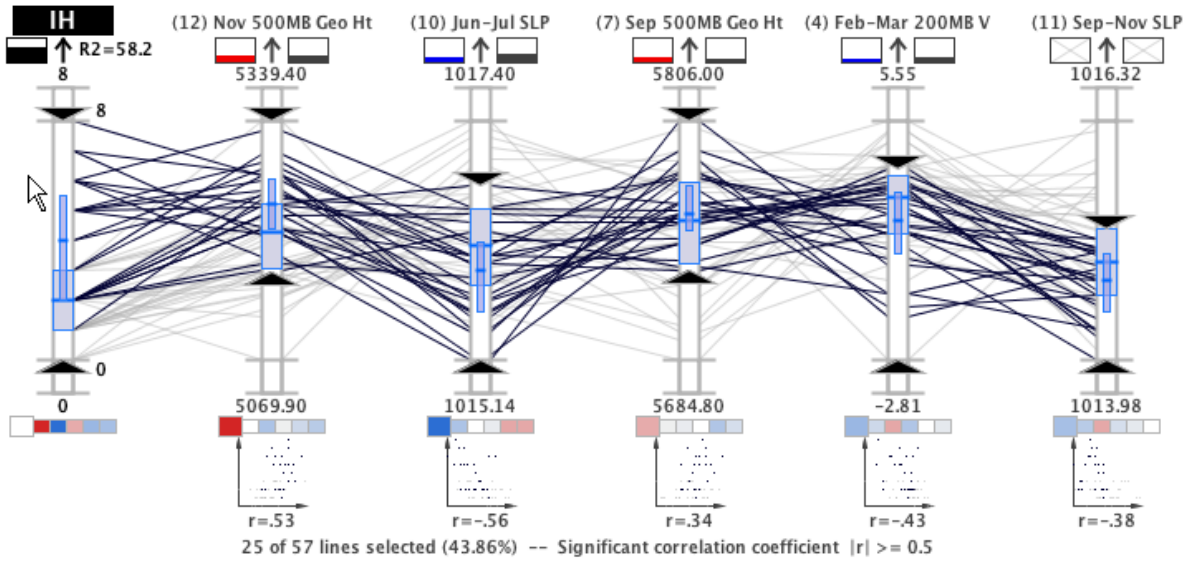


Figure 12: All seasons that fall within ranges for forecasting intense hurricane activity. This query reveals eleven false positives that meet variable range constraints developed to forecast seasonal *IH* activity.

safety, type I errors are preferred instead of type II errors (also called false negatives) where there is a failure to forecast the above normal activity season when, in truth, the activity is above normal. After the minor range adjustments, the three type II errors are eliminated, but the eleven type I errors remain.

5.2.3 Confirmatory Analysis

To be complete, the physical relationships of the selected predictors can be evaluated to ascertain the validity of the selections from a weather science perspective. Although a detailed physical evaluation is beyond the scope of this article, the selections of these five predictors can be validated by briefly describing how each variable influences the development of tropical cyclones.

The most significant predictor, axis (12), measures the the long-term oscillations which impact global wind patterns, known as teleconnections. When these oscillations are in one phase, they cause more ridges in the Atlantic, which corresponds to less wind shear. Also, weaker zonal winds in the subpolar areas are indicative of a relatively strong thermohaline circulation and therefore a warmer Atlantic Ocean. The MLR results indicate that when predictor (12) is normal or above normal, the environment is more favorable for the development of intense hurricane systems.

Pressure in the Atlantic Ocean is inversely related to tropical cyclone activity; low sea-level pressure in the tropical Atlantic implies increased atmospheric instability, moisture, and ascent (more favorable for the genesis of tropical cyclones), and weaker trade winds (which correspond to less wind shear that can tear up the thunderstorms in tropical cyclones). This relationship explains the selection of axis (11) and axis (10), which are normal or below normal in the active intense hurricane seasons.

The MLR analysis also identified two variables that characterize El Niño events which inhibit tropical cyclone formation and intensification in the Atlantic. The first clues of an impending El Niño can be detected in February by observing three variables. The MLR analysis selected one of these variables, axis (4), which measures the anomalous late winter north-south winds at 200 mb in the southern Indian Ocean (a condition associated with El Niño). As shown in Fig. 11, normal to below normal values of (4) correspond to more favorable conditions for intense hurricane development. The MLR model includes one Fall variable that is correlated to El Niño conditions for the following year, axis (7), which is more favorable for hurricane intensification in normal to above normal measurements.

5.3 Discussion

In a prior publication [Steed et al., 2008], early results from this research demonstrated the promise of parallel coordinates for confirming and clarifying the results of stepwise regression analysis in a tropical cyclone climate study. In this paper, an expanded version of the parallel coordinates system is used in conjunction with automated decision support algorithms to discover and confirm tropical cyclone trends. The expanded MDX system effectively blends the analytic spotlight of statistical processes with the inferential floodlight of visual exploration to facilitate guided multivariate data analysis. Using traditional climate study techniques alone would have required close examination of hundreds of plots to observe the same associations that are captured in a single frame of MDX. Perhaps the greatest evidence of the promise of this approach comes from the hurricane expert in the study, Dr. Fitzpatrick, who indicated that MDX made it possible to explore the associations in the climate data set quicker and more comprehensively than conventional climate study approaches.

MDX automated and streamlined many of the manual processes typically employed in his studies and it fused the capabilities into an integrated interface that is built upon a new interactive visual analytics approach for multivariate climate analysis.

In the process of developing MDX, a significant amount of time has been devoted to achieving an optimal design for the user interface. The color scheme and layout of elements have been formulated by drawing on color design principles from graphic design [Itten, 1970], empirical perceptual studies in information visualization [Ware, 2004], and hands-on evaluations during development. For example, muted colors are used in most graphical elements reserving the most saturated colors for small portions of the display. This creates a visual balance that is aesthetically pleasing to the viewer. Furthermore, the most vivid colors are placed on the peripheral of the display to further balance the view. A well-planned design for the visual analytics interface will greatly improve the user experience by reducing fatigue and making important relationships stand out to the viewer. What's more, the design can also improve the viewer's confidence in the software capabilities, which is crucial to effective communication of results.

6 Conclusion

This research corroborates the notion that interactive parallel coordinates can be used in conjunction with statistical analytics to discover and confirm hypotheses in climate data. While the statistical analysis processes guide the viewer to significant associations among the set of inter-related parameters, the dynamic visual interactions facilitate a deeper understanding of the relationships. These techniques have been fused together into a powerful interactive analytics system and evaluated in a real-world hurricane climate study. This research offers the following contributions to geographical visualization:

- A unique geovisual analytics system has been formulated that combines several interactive parallel coordinates capabilities with automated statistical processes to meet the needs of complex climate studies.

- The results of a comprehensive tropical cyclone trend analysis case study reveal several important physical associations in the environmental predictor data set, thereby highlighting the promise of a geovisual analytics approach for facilitating a deeper understanding of the data.

In the future, this research will be expanded to include additional seasonal statistics and climate study data sets. In addition, new multivariate visualization capabilities will be developed that enhance environmental data analysis, thus providing scientists with more effective visual alternatives for exploring the climate.

7 Acknowledgments

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 Dr. Phil Klotzbach of Colorado State University's Tropical Meteorology Project for providing the Atlantic tropical cyclone data set as well as the Geospatial Visual Analytics Workshop organizers and reviewers.

References

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, Austin, Texas, Oct. 2004. IEEE Computer Society.

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.

P. J. Fitzpatrick. *Understanding and Forecasting Tropical Cyclone Intensity Change.* PhD thesis, Department of Atmospheric Sciences, Colorado State University, Fort Collins, CO, 1996.

P. J. Fitzpatrick. *Natural Disasters, Hurricanes: A Reference Handbook.* ABC–CLIO, Santa Barbara, California, 1999.

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, San Francisco, California, Oct. 1999. IEEE Computer Society.

S. B. Goldenberg, C. W. Landsea, A. M. Mestas-Nuñez, and W. M. Gray. The recent increase in atlantic hurricane activity: Causes and implications. *Science*, 293:474–479, July 2001.

H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *Proceedings of IEEE Symposium on Information Visualization 2002*, pages 127–130, Boston, MA, 2002. IEEE Computer Society.

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.

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

J. Itten. *The Elements of Color*. Van Nostrand Reinhold Publishing, Ravensburg, Germany, 1970.

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, Minneapolis, Minnesota, Oct. 2005. IEEE Computer Society.

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 31 Mar. 2008).

P. J. Klotzbach. personal communication, Jan. 2007.

P. J. Klotzbach and W. M. Gray. Summary of 2006 atlantic tropical cyclone activity and verification of author's seasonal and monthly forecasts. Technical report, Nov. 2006. http://hurricane.atmos.colostate.edu/Forecasts/2006/nov2006/ (current 31 Mar. 2008).

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 31 Mar. 2008).

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.

H. Piringer, W. Berger, and H. Hauser. Quantifying and comparing features in high-dimensional datasets. In *International Conference on Information Visualization*, pages 240–245, London, UK, Jul. 2008. IEEE Computer Society.

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, November/December 2007.

R. A. Rensink. Change detection. *Annual Review of Psychology*, 53:245–277, 2002.

J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, 2005.

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.

H. Siirtola and K.-J. Räihä. Interacting with parallel coordinates. *Interacting with Computers*, 18 (6):1278–1309, Dec. 2006.

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*, 2008.

T.J. Jankun-Kelly and C. Waters. Illustrative rendering for information visualization. In *Posters Compendium: IEEE Visualization 2006*, pages 42–43, Baltimore, MD, 2006. IEEE Computer Society.

J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts, 1977.

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.

C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, 2nd edition, 2004.

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

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.

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, Los Alamitos, California, 1997.