TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 01/08 pp111-124
Volume 18, Number 2, April 2013

An Interactive Visual Analytics Framework for Multi-Field Data in a Geo-Spatial Context

Zhiyuan Zhang*, Xiaonan Tong, Kevin T. McDonnell, Alla Zelenyuk, Dan Imre, and Klaus Mueller

Abstract: Climate research produces a wealth of multivariate data. These data often have a geospatial reference and so it is of interest to show them within their geospatial context. One can consider this configuration as a multi-field visualization problem, where the geo-space provides the expanse of the field. However, there is a limit on the amount of multivariate information that can be fit within a certain spatial location, and the use of linked multivariate information displays has previously been devised to bridge this gap. In this paper we focus on the interactions in the geographical display, present an implementation that uses Google Earth, and demonstrate it within a tightly linked parallel coordinates display. Several other visual representations, such as pie and bar charts are integrated into the Google Earth display and can be interactively manipulated. Further, we also demonstrate new brushing and visualization techniques for parallel coordinates, such as fixed-window brushing and correlation-enhanced display. We conceived our system with a team of climate researchers, who already made a few important discoveries using it. This demonstrates our system's great potential to enable scientific discoveries, possibly also in other domains where data have a geospatial reference.

Key words: geospatial visualization; visual analytics; information visualization; multivariate visualization; parallel coordinates; coordinated displays; linking and brushing

1 Introduction

Our paper describes a comprehensive framework and system for interactive visual analytics with multi-field data. Here, a multi-field is considered a multivariate extension of a scalar field, that is, each point in Euclidian space offers values of multiple properties at that location and time. The definition of field as a physical quantity associated with each space-time point is one that has been adopted by all branches of physics: electricity, electro-magnetism, gravity, but also fluid dynamics, and so on. Likewise, geography has adopted a similar notion of field than physics, but with the important distinction that geographic fields are not necessarily produced by strict physical laws. Rather, they can be due to demographic sampled assessments, behavioral modeling, population and cultural effects, environmental measurements, and many others.

Geographic fields are often visualized using cartographic techniques, contour plots, and choropleth maps. They are in many cases multivariate – just consider a map of households with incomes, number of children, cars, and so on. However, multiple variables are difficult to plot with choropleth maps,

• Zhiyuan Zhang and Klaus Mueller are with the Visual Analytics and Imaging Lab at Computer Science Department, Stony Brook University, Stony Brook, NY 11790, USA. E-mail: {zyzhang, mueller}@cs.sunysb.edu.
• Klaus Mueller is also with SUNY, Korea.
• Xiaonan Tong is an undergraduate student at Stanford University, Stanford, CA 94305, USA. E-mail: tongxn2009@gmail.com.
• Kevin T. McDonnell is with the Department of Mathematics, Dowling College, Oakdale, NY 11769, USA. E-mail: mcdonnek@dowling.edu.
• Alla Zelenyuk and Dan Imre are with Pacific Northwest National Lab, Richland, WA 99354, USA. E-mail: alla.zelenyuk@pnnl.gov; dimre2b@gmail.com.
* To whom correspondence should be addressed.
Manuscript received: 2013-02-21; accepted: 2013-03-01

and so a number of researchers have linked them with multivariate visualization displays, most frequently parallel coordinates. In such a visual geo-analytical framework, the analyst would obtain insight about the spatial distribution of the color-coded populations in the geographical display, and then visualize the multivariate signatures/composition of these populations in the parallel coordinate display, using a color legend as a reference. Typically these geo-graphical displays are two-dimensional (2-D) maps.

Our system was developed in close collaboration with two groups of climate scientists. With the growing intensity of local climate fluctuations, the melting of polar ice caps, and the emergence of other related processes, researching the cause of these trends has gained tremendous importance in recent years. As opposed to demographic data, phenomena that change the earth's atmosphere bear important three-dimensional (3-D) relationships, and as such climate researchers typically acquire their data using probes in 3-D space, either aided by airplanes or other sensors situated at diverse altitudes. In their research, climate scientists often pursue efforts in which they seek to proof or disprove hypotheses involving different aspects of the data. This mandates an interactive system that not only supports both data types – spatial and the non-spatial fields – but also uses 3-D maps for the geo-spatial display.

An attractive platform for interactive geo-visualization is Google Earth. It provides easy access to 3-D geographical data and for that reason has often been lauded as a "democratization of GIS". In recent years, various efforts have also used Google Earth as a geographically-referenced canvas for the presentation of many types of field data, both of physics and geographic nature. However, the purpose of these Google Earth-based data displays is typically merely data visualization, with only some selection capabilities available. The opportunities for user interaction are mostly restricted to navigating the environment – the globe – using the standard Google Earth spatial navigation tools and possibly a slider for time-animation of the data. To the best of our knowledge no application so far has utilized Google Earth as a platform for full-fledged visual analytics in which users can interact with the data directly in the Google Earth display, by ways of standard information interrogation techniques such as brushing, filtering, and aggregating, and communicating these interactions back to a linked information display. Specifically, the major contributions of our paper are:

- We extend an interactive geo-browser – Google Earth – to support a set of common interactive information interrogation techniques such as brushing, filtering, and aggregating.
- We link this extended geo-browser to a popular multivariate information display – parallel coordinates – and establish bi-directional interaction propagation.
- We devise a new family of design primitives, conceived to show some multivariate data aspects directly in Google Earth.
- We demonstrate the viability of our framework in the context of climate research, enabling a collaborating team of climate scientists to make important discoveries in their research domain.

We consider the majority of our contributions an engineering effort, but one that is fairly sophisticated. A valuable element of our system is that its design is well-informed by the research workflow of climate scientists. This is confirmed by the fact that a team of such scientists was able to make a number of significant discoveries using our system. While they might have made these discoveries with conventional tools as well, our system enabled them much quicker and easier, and so accelerating progress in this important research domain.

2 Related Work

This paper addresses the visualization of multi-field data from a geographic perspective, and one in which the data originate from non-regular sampling of real physical phenomena in a 3-D domain. Conversely, the multi-field work targeted towards data visualization from a physics perspective has focused more on simulations on regular grids. Recent work in that area includes that of Nagaraj and Natarajan[1] who visualized the data by assessing the relationship among multiple scalar fields in terms of covariate critical points and topology. Gosink et al.[2] devised a query-driven framework that can reveal statistically important interactions between any three field variables. Most similar to our efforts, although in a different domain, is the system by Blaas et al.[3] who explored multi-field medical data by linking object views with information displays.

The association of multivariate information displays with geographical maps has a fairly long history. The first development in that direction appears to

be the system by MacEachren et al.[4] who joined a geographical display with scatterplots and parallel coordinates[5]. Andrienko and Andrienko[6] added to that work by linking the parallel coordinate display to dominant attribute classification maps. Jern, in collaboration with various colleagues, devised the GeoAnalytics Visualization (GAV)[7] framework and class library that adds to the standard choropleth maps a collection of standard information visualization representations including scatterplots, a parallel coordinate plot with percentile handles, table lens, and treemap. Finally, Guo et al.[8] added to this mix of tools a Self-Organizing Map (SOM) to perform multivariate clustering, sorting, and coloring. These tools are available in a geo-visual analytic software package, VIS-STAMP.

Google Earth, widely available since 2005, has attracted a great deal of attention as a platform for the visualization of scientific and geographic data. In a paper published relatively early, Wood et al.[9] described the use of Google Earth for visualizing a large mobile directory service log file with spatial, temporal, and attribute components as tags on top of the geography. Users can mouse-click on tags which triggers the system to zoom into the affiliated geographical area. The OpenEarth toolbox[10] has enabled an impressive set of scientific visualizations[11] on Google Earth, but so far mainly in the domain of marine and coastal science.

The system most similar to ours is GEO-SPADE, recently described by Kisilevich et al.[12] It also integrates Google Earth as a plug-in into a larger system, but it does not provide linked information displays. The authors demonstrate the use of their system by ways of the analysis of tourist travel sequences in an arbitrary region of the world. Users are able to specify geographic boundaries via a text interface which triggers the clustering of photo-collections in the selected region and a subsequent display of the cluster boundaries on Google Earth. Again using a text interface, users may then modify the clusters. Our framework differs from GEO-SPADE in that it allows all interactions to occur directly in the Google Earth interface using the mouse. This provides for a much more direct and intuitive user experience and also allows for more free-form shape specifications.

3 Domain Applications and Requirements

Our efforts were motivated by two specific applications in climate science. In the following we describe their data as well as the research workflow for one of them.

3.1 Domain data

3.1.1 ISDAC dataset

The ISDAC dataset was acquired by a single particle mass spectrometer (SPLAT II)[13,14] on Flight 26 (F26) which took place on April 19-20, 2008 as part of the Indirect and Semi-Direct Aerosol Campaign (ISDAC)[15], a month-long field campaign at the North Slope of Alaska (see Figs. 1a and 1b). F26 began in Barrow, Alaska and ended in Fairbanks, Alaska (see Figs. 1c and 1d). The flight began with a short transit over a Department of Environment (DOE) ground site, followed by about 90 minutes of sampling a cloud at low-altitude. The aircraft then performed a spiral, climbing to an altitude of about 7000 m proceeding for a landing in Fairbanks. The main scientific objective of ISDAC was to improve the understanding of how changes in the size, composition, and concentration of aerosols particles influence cloud properties and their associated radiative forcing. During the month-long campaign, SPLAT II measured the number

(a) (b) (c) (d)

Fig. 1 Capturing the ISDAC dataset. (a) The single particle mass spectrometer (SPLAT II) operated by the collaborating scientist in-flight in the Arctic aboard a Convair-580 research aircraft. (b) Various sensor probes mounted on the aircraft wing. The aircraft flew various missions over Alaska to measure concentrations, size distributions, shape, density, and compositions of millions of particles in clear atmosphere to establish a large and highly resolved data set of Arctic aerosol particles. Other environmental variables, such as cloud density, pressure, and density were also sampled. (c) Overview of the flight path. (d) Profile as seen from the side, both captured with Google Earth.

concentrations, size distributions, shapes, densities, and compositions of millions of particles in clear atmosphere to establish a large and highly resolved data set of Arctic aerosol particles. In the cloud, SPLAT II characterized the properties of Cloud Condensation Nuclei (CCN) particles, on which cloud droplets form, and those of interstitial particles to develop a highly detailed dataset.

The ISDAC dataset consists of more than 2 million data points, each a 33-dimensional vector: latitude, longitude, altitude, time stamp, temperature, and pressure. It also contains measurements on the cloud particles (cloud droplets presence, cloud particle concentration, etc.) and on the aerosol particles (size and composition: soot, sulfate levels, organics, dust, sea salt, etc.). The dataset was obtained by fusing measurement files of different instruments using the time stamp for alignment/binning.

3.1.2 Global seawater oxygen-18 database

This global seawater oxygen-18 database[16] is a collection of about 26 000 seawater measurements from all around the world, each an 8-dimensional vector: longitude, latitude, month, year, depth, temperature, salinity, and oxygen composition ratio δ^{18}O. The δ^{18}O value is a very good tracer of water origin and highly correlated with salinity, but it varies regionally and seasonally under some specific conditions. For example, when salinity is nearly 0 psu, then the δ^{18}O typically has a wide range of values. Possible reasons can be precipitation[17], river inflow[18], or glacier calving[19]. These various geographic and multivariate dependencies make this dataset an excellent test case for our system.

3.2 Domain requirements for the ISDAC dataset

The goal of the team of domain scientists we primarily collaborated with – they are also co-authors of this paper – was to gain a better understanding of particle composition and size at various geo-spatial locations, as well as the relations to other particle properties, atmospheric conditions, and particle activation probabilities. Since the domain data share the calamities of most such datasets – many outliers, unspecified values for some attributes, etc. – these relationships are difficult to discern via automatic analytical algorithms, and this has motivated the use of visual analytics techniques to overcome these shortcomings[20].

In the following we list the set of basic requirements our collaborators expressed in the onset of the project:

R1. Ability to visually interact with the multi-field data.

R2. Ability to summarize the data in terms of different variables.

R3. Ability to visualize the relations among variables as a multi-variate display.

R4. Support of geo-spatial references, whereby the geo-spatial display should fully support interactions such as selection, filtering, and brushing.

R5. Support of coordinated displays – all displays should be linked such that operations on one display are reflected on the other.

Although our collaborators had access to a variety of visualization frameworks, such as parallel coordinates, our ClusterSculptor framework[21], Microsoft Excel, Google Maps and Earth, etc., these systems were disjoint and could not provide the holistic dual-domain interaction that was needed to produce the desired insights. For quite some time, the scientists would create pie charts of particle distributions with MS Excel and then overlay them on a static Google Earth map to visualize the data – clearly a rather cumbersome workflow which greatly slowed the pace of research.

4 System Design

Our primary aim was to devise an integrated framework that would allow climate scientists to interactively visualize and analyze their multi-field data. Here we had a number of choices. In the following we address each in the context of the five domain requirements (marked as Rx below) listed in Section 3.2.

4.1 Visualizing multivariate data (R1)

To visualize multivariate data, among the most popular techniques are parallel coordinates[5] and scatterplot matrices[22]. However, due to the distributed 2-D tiled layout of the scatterplot matrix, it can be difficult to discern relationships that involve more than two variables. We chose parallel coordinates since they visualize high-dimensional data as flows across vertical axes and so yield a more connected representation. They also conveniently support brushing, selection, and filtering by simple axis interactions.

4.2 Summarizing different variables (R2)

Pie charts (for proportions) and histograms or bar charts (for distributions) are fairly low-tech but well understood visualizations, and have also been widely

used by our collaborators. Therefore, to reduce the learning curve we made use of these paradigms in our system, but merged them into a combined design for added expressiveness. We call this design the Pie Chart-histogram Design (PHD). In order to avoid potential overcrowding in the map, our primary goal was to make the PHD space-efficient. This ruled out designs that would place a pie chart next to a bar chart, as such configurations would waste much empty space. Conversely, as is well known[23], circles – when sized equally – achieve the tightest packing in 2-D space. Therefore, we strived to create designs with circular geometry. We derived two different designs mainly distinguished by their renditions of the histogram. Figure 2 compares the two designs side by side, along with a pie chart with no histogram. One design leaves the pie chart non-occluded and wraps the bar chart around its perimeter. The other places the bar chart into the center but allows the pie chart sector lines to shine through. Finally, our interface also provides a dedicated window that shows the histogram of a selected space-point.

The feedback from our domain collaborators was that although the circular bar chart looked artistic, they felt that it was too different from traditional representations, and also difficult to read. They preferred the second rendition.

4.3 Visualizing relationships (R3)

Here we chose an illustrative correlation rendering technique to display the correlation information in the parallel coordinate display. Since domain experts are not as fluent as visualization researchers in the visual language of parallel coordinate displays, we bridged this gap by adding illustrative hints to help the interpretation of trends[20] and correlations[24]. In parallel coordinates, negative correlations give rise to lines that aggregate into bow-tie shaped line bundles, however subtle. In Ref. [24] we proposed the following graphical-design inspired scheme that can make these relationships more obvious for less experienced users. First, for each adjacent dimension pair a bounding

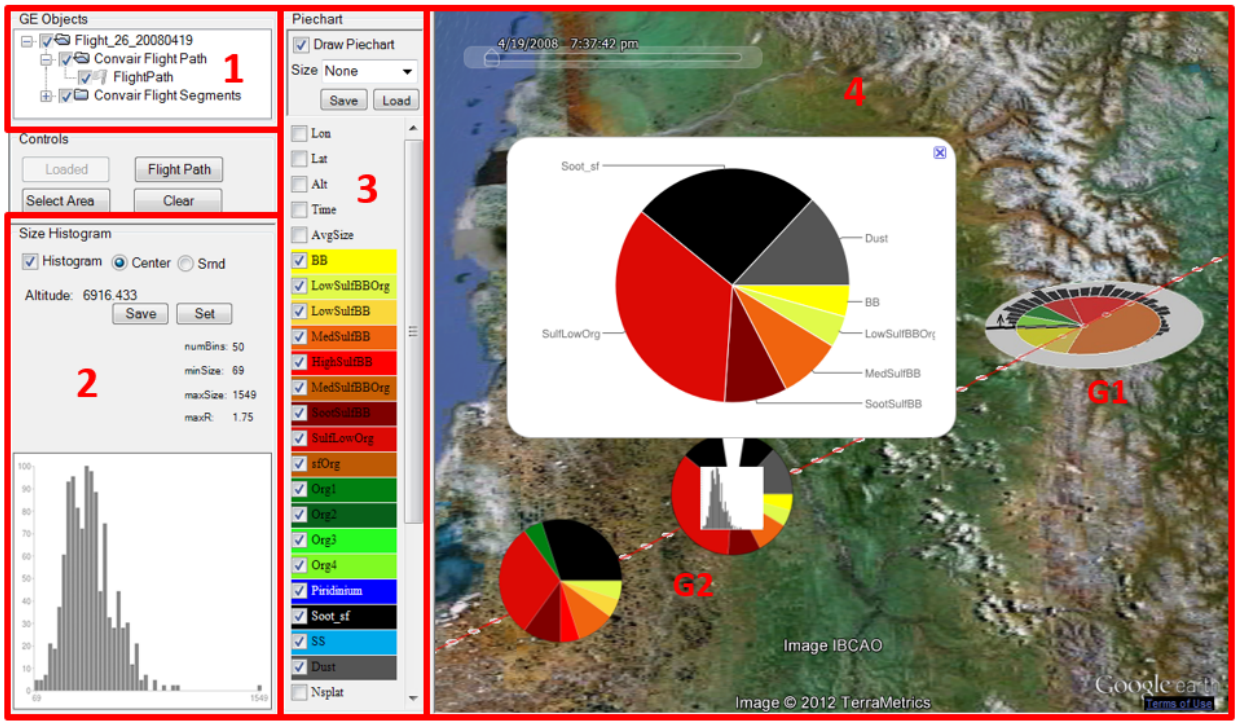


Fig. 2 Interface and pie chart-histogram design. Panel 1 is the GE object control panel which allows users to show/hide a GE object. Panel 2 is the bar-chart (or histogram) panel – here configured for ISDAC particle size. Its control panel allows the analyst to control various parameters, including the design of the PHD in the GE display. Panel 3 is the pie chart control panel. It contains the parameters for configuring the pie chart, such as what attributes will be displayed in the chart and which attribute is used to determine the size of the PHD. Users can also save/load the previous settings or export the current pie chart/histogram information into text files for further research. Panel 4 is the Google Earth display showing the three different PHD styles we provide. When the user clicks on the PHD, the pie chart detail is shown nearby, and the corresponding size histogram is shown in the histogram panel.

hull of the line bundles is computed based on the dimensional means and standard deviations. If the correlation is positive, then we can use this bounding hull as an abstracted band shape. Conversely, if the two dimensions are negatively correlated, the characteristic bow-tie shape is employed. Then the bounding hull is colored in terms of correlation strength where less saturation maps to lower correlation. Figure 3e provides an example for this scheme.

4.4 Supporting geo-spatial references (R4)

Map-based methods, such as Google Maps, Bing Maps, and Google Earth, have been widely used to provide geo-spatial or location references in many applications. In climate research, the altitude (elevation or pressure) plays an important role in the analytical process. Although 2-D map-based methods can show the entire world in one display using Mercator projection and the like, due to the fact that the altitude information will be inevitably lost after projecting the 3-D data onto the 2-D maps, they are not useful for our purposes. Hence, we chose Google Earth as the geo-spatial reference display.

4.5 Supporting coordinated displays (R5)

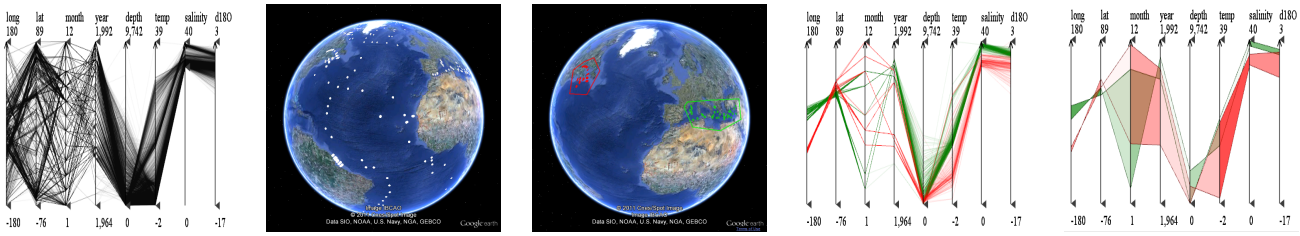
Our framework consists of two displays: a multivariate visualization display and a geographic display. These two displays are linked such that operations on either display will be reflected on the other. Upon reading a dataset, longitude, latitude, and altitude are used to populate Google Earth (GE) with simple icons (placemarks). Meanwhile, the Parallel Coordinate Plot (PCP) is populated with the data spectrum of the data points. Analysts can then use the mouse and brush in either display (GE or PCP) to select a subset of points, assign a color to these points, and see them reflected in the same color in the other display. To make up for the shortcomings of GE to display multivariate information, we have stricken a compromise and display a pie chart and a histogram at each selected measurement site (see Section 4.2). Both size and color of the GE site icons can be linked to any variable using the attribute mapping checkboxes in the pie chart control panel.

5 Methods

Koua et al.[25] proposed 10 general exploratory goals that geo-analytic systems should address: identify, locate, distinguish, categorize, cluster, distribute, rank, compare, associate, and correlate (see Section 5.4 for a more detailed description in the context of our system). Since geo-data have two types of attributes – geospatial and non-geospatial – only an interface that links (at a minimum) two dedicated displays – one geospatial and one multivariate – together can achieve a comprehensive visualization experience that meets these goals. This was recognized already early on (see Section 2). However, another issue is how users can express their goals, which they typically do via manual *selection* operations – one might call them gestures. A PCP information display typically facilitates selection interactions by allowing users to manipulate range handles on the individual axes. On the other hand, a geographical display such as GE would support selection operations by allowing users to click on points (or placemarks) or draw bounding contours around sets of points. To the best of our knowledge these direct selection interactions are thus far not supported in GE-based visual analytics systems.

Finally, once these data points have been selected in GE and color-tagged they would be represented in the PCP display as a group in the same color (and vice versa). This completes the full brushing operation.

Our framework is developed in C# using Direct3D



(a) (b) (c) (d) (e)

Fig. 3 Dual-domain analytics. (a) The analyst first uses the PCP brushing handles to select the normal ocean data points (salinity from 32 to 40). (b) The GE display responds by showing only these remaining data points. (c) Next the analyst uses mouse clicks to outline some interesting regions in the GE display (Mediterranean shown in green and Gulf of St Lawrence shown in red). (d) The points inside the selection polygon appear highlighted in the PCP display. (e) Correlation-enhanced PCP display.

for graphics. Our GE display uses the GE plug-in and also a C# custom-built API for the GE plug-in – the Winforms-Geplugin-Control-Library (WGCL)[26]. This library defines a list of methods that can be used to interact with the GE plug-in by dynamically injecting JavaScript code into a browser page during run-time to interact with the GE plug-in. WGCL works in the .NET Framework (C#) and provides a bridge between the client-based application and the web-based GE plug-in.

5.1 Embedding the PHDs into Google Earth

All embedded pie chart histogram designs are rendered on the fly whenever a data point is selected. Figure 4 describes this process as a flow chart. On the other hand, if a region is selected by brushing either the GE or PCP display, the corresponding PHD for all selected points will be shown and placed nearby the average geographical location.

There are in fact two approaches to render a PHD in GE. One is to render each PHD using GE's polygon rendering functionalities (see G1 in Fig. 2). A downside of this approach is that after rendering, both location and tilt angle of the PHD are fixed in GE. This means that if we wish to make the PHD always face the viewer, we must delete it from GE and re-render it every time we change the viewing direction, which can be rather time consuming. However, this method provides better depth and height information than the second method in which we use an icon (image) to represent the PHD. GE supports a feature in which an icon/image of a placemark consistently faces the viewer, no matter how we change the view direction. But the challenge here is how to render the PHD image on the fly. Fortunately, Google Chart Tools (https://developers.google.com/chart/) provide good support for rendering various charts into an image which can be later retrieved via a URL. Using Google Chart Tools, we render the image of a PHD by passing its parameters (such as pie chart compositions and histogram information) and then load the image into GE by passing it the image URL. The result is shown in G2 in Fig. 2. In our system, we allow the user to choose the rendering method, dependent on the underlying task.

5.2 Brushing in the Google Earth display

We support two types of brushing tasks in GE:

- **Single data point brushing**. The user can select any data point in GE by a single mouse click. Attributes associated with the data point will be shown in a popup window and at the same time the corresponding polyline will be highlighted in the PCP display.
- **Region-based brushing**. This addresses the need to visualize the behavior of an entire geographic region in the PCP display. However, since mouse dragging is reserved by GE for view rotation, we can not simply drag the mouse to outline the region of interest. Instead, we impose the moderate requirement that users can use a series of mouse clicks to specify the vertices of the selection polygon. The polygon can be either convex or concave. After polygon completion, we employ a quick points inside the polygon test to determine the data points inside the selection polygon. These interior points are then marked and the corresponding polylines highlighted in the PCP display. Likewise, selecting points in the PCP display will highlight them in GE. Figure 3 shows an example for both brushing directions.

5.3 Brushing in the PCP display

In our PCP display users are able to manually interchange axes, flip (invert) axis directions, and perform statistics-guided outlier filtering[7] and clustering in each dimension. The following is a set of further capabilities our domain scientists found useful.

- **Fixed-window brushing**. We extended the range handles typically used to bracket data intervals from the top and bottom to *fixed-window brushing* – essentially an interval slider. In this mode, the distance between the two handles remains fixed and as the user drags the handles they will move up and down simultaneously. This feature is very helpful to show how other attributes behave as one attribute changes, and the GE display will visualize the corresponding changes in geo-spatial

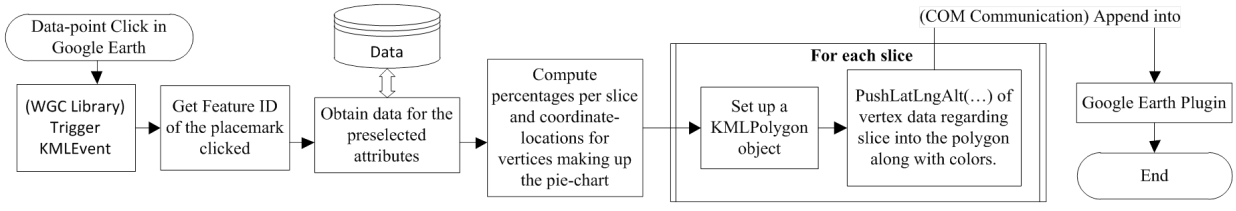


Fig. 4 Flowchart showing how to dynamically render an object into Google Earth. Here the pie chart is used as an example.

locations.

• **Adjust-window brushing**. Users can also drag the mouse up/down to make the width of the bracketed window bigger/smaller. This operation is useful to see the behavior of the entire dataset as the range of an attribute spreads out/shrinks.

5.4 Addressing exploratory tasks

To rate how well our system supports geo-analytics, we examine our framework via Koua's exploratory tasks, also in the context of our domain applications.

• **Identify**. The now fully bi-directionally linked displays make it easy to identify relationships.

• **Locate**. GE provides an effective way to directly see not only the longitude and latitude information, but also the altitude information, which is a significant factor in climate research.

• **Distinguish and distribute**. By using the PHDs in GE, users can easily judge the differences among sample points directly in the geographic domain.

• **Categorize and cluster**. The range-handle, cluster, and tag operators in PCP and the point/region selection operators in GE well support interactive classification directly in the most suitable domain. Any cluster can be subdivided further or re-assigned to another cluster, all of which is maintained via a cluster checkbox.

• **Rank and compare**. The PHDs in GE provide the user with an effective means to visually assess ranking and perform comparisons. Likewise, the brushing and clustering operations in PCP allow users to compare cluster behaviors in a multivariate context.

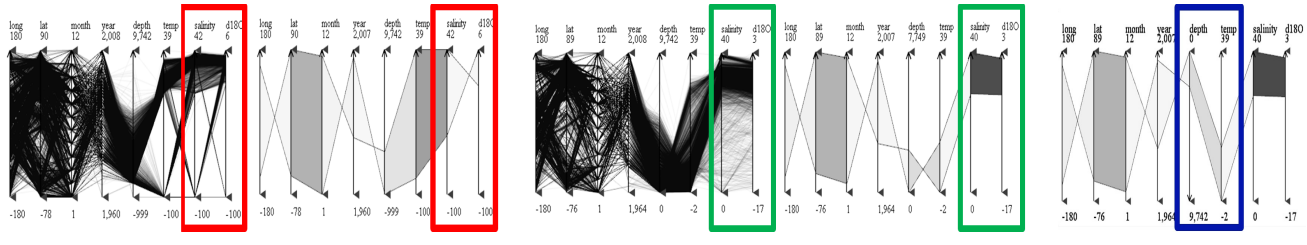
• **Associate and correlate**. Both PHDs and brushing operations in PCP and GE aid users in assessing relationship between attribute(s) and geographical information. In addition, the PCP correlation visualization allows users to easily recognize data relationships.

6 Use Cases and Results

6.1 Global seawater oxygen-18 database

First, since the data is an amalgamation from different sources and different tasks, several attributes have undefined values (-999 for depth and -100 for temperature, salinity, and δ^{18}O in Fig. 5a). These undefined values can significantly influence the analysis results. As shown in Fig. 5b very weak correlations among all dimensions are observed. Following, Fig. 5c shows the parallel coordinate display after brushing the four dimensions, filtering out the undefined values, and re-normalizing their bottom axis brackets. We can now readily see in the correlation display (Fig. 5d) that there is in fact a strong correlation between dimension salinity and δ^{18}O.

Though salinity and δ^{18}O are highly correlated with each other, they behave quite differently at some conditions. One condition is when salinity is nearly 0 psu, the δ^{18}O might have widely differing values. Factors that can influence the δ^{18}O values are precipitation[17], river inflow[18], or glacier calving[19]. To get insight on which factor influences the sample points most, we first use the PCP brush handles to select the regions that have nearly 0 salinity. We notice that the δ^{18}O varies greatly (see Fig. 6a). To determine the reason, we turn to the GE display (Figs. 6b and 6c). After zooming into all the regions that contain the filtered data points, we can clearly see that all the



(a) (b) (c) (d) (e)

Fig. 5 Interactions in the PCP display. (a) PCP and (b) correlation-enhanced PCP display of the original dataset with undefined values in dimensions depth, temp, salinity, and δ^{18}O. (c) and (d) The same set of displays after filtering out the undefined values and zooming into the dimensions. We now have a clear view of the remaining data. For example, we can see a strong positive correlation between salinity and δ^{18}O (green square in (c) and (d)) after the dimension zoom-in, while in the original dataset with undefined values, δ^{18}O is falsely negative correlated with salinity (red square in (b)). (e) Correlation display after axis depth inversion. Sea depth and temperature now have a positive relationship (blue square in (e)), which is consistent with our knowledge.

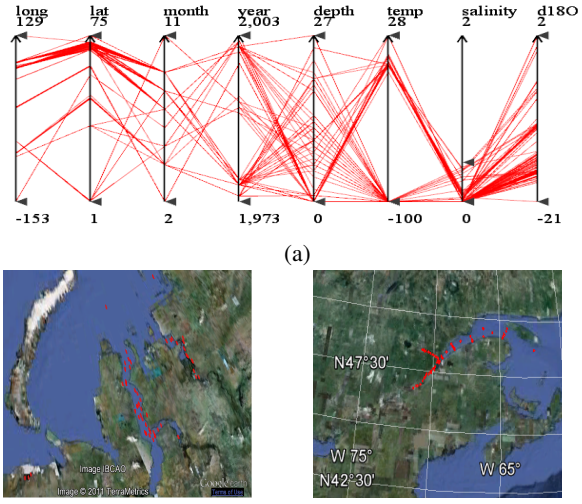


Fig. 6 Salinity is nearly 0 psu while the δ^{18}O has widely differing values due to river inflow. (a) Brushing in PCP to select data that have near-zero salinity. These areas are: (b) Obskaya Gulf (estuary of Ob River), Yenisey (estuary of Mal. Taz River), and White Sea in Russia. (c) St. Lawrence River area in Canada.

remaining points are at the mouths of rivers, where rivers meet the sea. Thus for this dataset, the most influencing factor appear to be junctures where fresh water meets salty water.

Our framework can also be used to confirm (or reject) hypotheses. For instance one hypothesis for sea water oxygen is that sensitivity of δ^{18}O can be greater than that of salinity in the deep ocean[27]. To test this hypothesis, we first use the brush handles to select the deep ocean dataset, here we choose depth>2000 m (Fig. 7a). From the poly lines, we observe that the δ^{18}O varies more than salinity does, but we can not see the difference clearly. To get a better view, we set the PCP rendering mode to correlation-enhanced (Fig. 7b). We observe that the width of the branches for δ^{18}O is larger than that for salinity. Since each dimension is normalized in the PCP display, the hypothesis is confirmed.

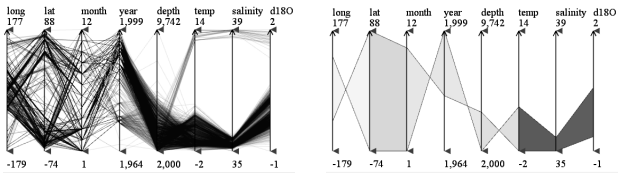


Fig. 7 Deep seawater (depth>2000 m) analysis. Both displays show that at deep sea the variation of δ^{18}O is larger than salinity. (a) Parallel coordinate plot for data with depth greater than 2000 m. (b) Correlation enhanced rendering of parallel coordinate plot.

The dual-domain analytics facilitated by the two linked displays also helps to make some interesting discoveries. For example in Fig. 3, the analyst first used the brush handles to select the normal ocean data points (salinity from 32 to 40), shown in Fig. 3a. The GE display then only displayed the data points that were sampled in normal ocean waters. The analyst then outlines some interesting regions in the GE display – the Mediterranean area (green cluster in Fig. 3c) and Gulf of St. Lawrence (red cluster). Following these interactions, the points inside these selection polygons appear highlighted in the PCP display (Fig. 3d). By comparing these two clusters in the PCP display we can make the following observations:

(1) The samples observed in the Gulf of St. Lawrence appear much earlier (year) than those in the Mediterranean.
(2) The depths of the sample points in the Mediterranean have higher variation than in the Gulf of St. Lawrence.
(3) But the temperature, salinity, and δ^{18}O in the Mediterranean have much lower-variant distributions than those in the Gulf of St. Lawrence, which means the water conditions in Mediterranean are more stable. The observations can be confirmed in the correlation display (Fig. 3e) and are actually easier to see.

6.2 ISDAC dataset

Much of the visualization work conceived for this paper was developed in tight collaboration with a team of climate researchers studying the effects of aerosols on global warming. One might consider the following section a result of a formative, user-in-the development-loop user study.

6.2.1 Background

Atmospheric aerosols play an important role, affecting both global and regional climate change during the last century. They do so by scattering and absorbing solar radiation and by determining cloud properties[28,29] in their role as CCN and Ice Nuclei (IN). Yet the relationship between the properties of aerosol particles and clouds, i.e., the aerosol indirect effect, remains the most uncertain aspect in our current understanding of climate change. Scattering and absorption probabilities depend on particle number concentrations, size distributions, individual particle compositions, and Relative Humidity (RH), which requires sophisticated instrumentation and data analysis

tools to mine the vast amount of detailed data that needs to be acquired.

Laboratory data suggest that cloud formation and cloud properties are tightly connected with the properties of the particles, on which they form. Because particle's hygroscopicity, which is determined by particle composition, is related to their CCN activity, we expect particle compositions to play an important role in determining cloud formation and properties. The fraction of aerosol particles that activate to form cloud droplets depends on particle composition, size distribution, and number concentration, presenting a complex dependence. Most importantly, because CCN activation is not linear, it is essential to know the spatial distributions of the aerosol fields with high resolution.

The Arctic region represents an important and interesting location to study the forces that affect the global climate. Arctic aerosols are advected into the region from Asia, Europe, and North America, their loadings, compositions, and other properties vary significantly with meteorology. Biomass Burning aerosol (BB) transported from Asia and North America is presently one of the most significant aerosol sources in the Arctic Spring. During Arctic spring, high BB concentrations produce the "Arctic haze". One of the interesting aspects of Arctic haze is that it is often found to be in distinct stratified layers[30]. Previous measurements of aerosol chemical composition point to sulfates as dominant constituents of arctic aerosol, with smaller contributions of soot, Sea Salt (SS), organics, and dust[31-34]. However, these composition measurements provide information on the bulk aerosol composition only with poor spatial and temporal resolution.

The effects of aerosol in climate models strongly relate to size and composition of individual aerosol particles[35], requiring high sensitivity and temporal resolution aircraft measurements. Because these types of measurements generate massive amounts of complex, multidimensional data there is a defined need for a specialized data analysis tool. In the following, we illustrate how our collaborating climate researchers, who are co-authors of this paper, used our framework to analyze their ISDAC dataset (see Section 3.2). Prior to this analysis, we first combined classes with particles of similar compositions to 17 distinct classes.

6.2.2 Visual analytics outcome

As the first step, to get an overview of the particle compositions, sizes, and compression information, we use the region brushing feature in GE display to outline some interesting regions during the flight, such as the low elevation sea area, cloud area, spiral area, and high elevation land area (see Fig. 8). Then for each region, the system computed the PHDs based on the summation of the particle compositions and particle sizes and plotted it nearby. The PHDs are sized by the NSplat variable to visualize compression information. The overview shows significant spatial variability in particle composition and size, which is consistent with previous reports showing a highly stratified atmosphere. Most are BB particles that were transported over long distances during which they adsorbed sulfates and additional organics. As the aircraft approached Fairbanks, at high altitude, the number of organic, dust, and sea salt particles increased.

Next they used the framework to perform a visual analysis for the cloud particles (Fig. 9). A simple comparison between the two pie charts indicates that the compositions of activated and un-activated interstitial particles are virtually the same. According to our collaborators, this is a significant discovery and changes much what has been assumed in the field so far.

The feedback from our collaborators was tremendously inspiring. They first checked that the analysis of clouds probed on all other flights yielded the same results. This provides, for the first time, direct experimental evidence that particle compositions play only a minor role in determining cloud activation probability. The finding is in contradiction with laboratory experiments that have shown a simple relationship between activation probability and particle hygroscopicity, which is directly related to particle composition. It suggests that laboratory-based cloud activation instruments operate on a different principle than that controlling cloud activation in the ambient atmosphere. Moreover, climate models use laboratory derived composition dependent activation probabilities to simulate cloud formation. The data presented here, if reproducible in other parts of the globe, indicate that cloud activation needs to be reformulated.

As noted above, stratification is one of the more interesting features of the Arctic atmosphere. To be properly characterized it requires instrumentation with high temporal resolution and data analysis tools that

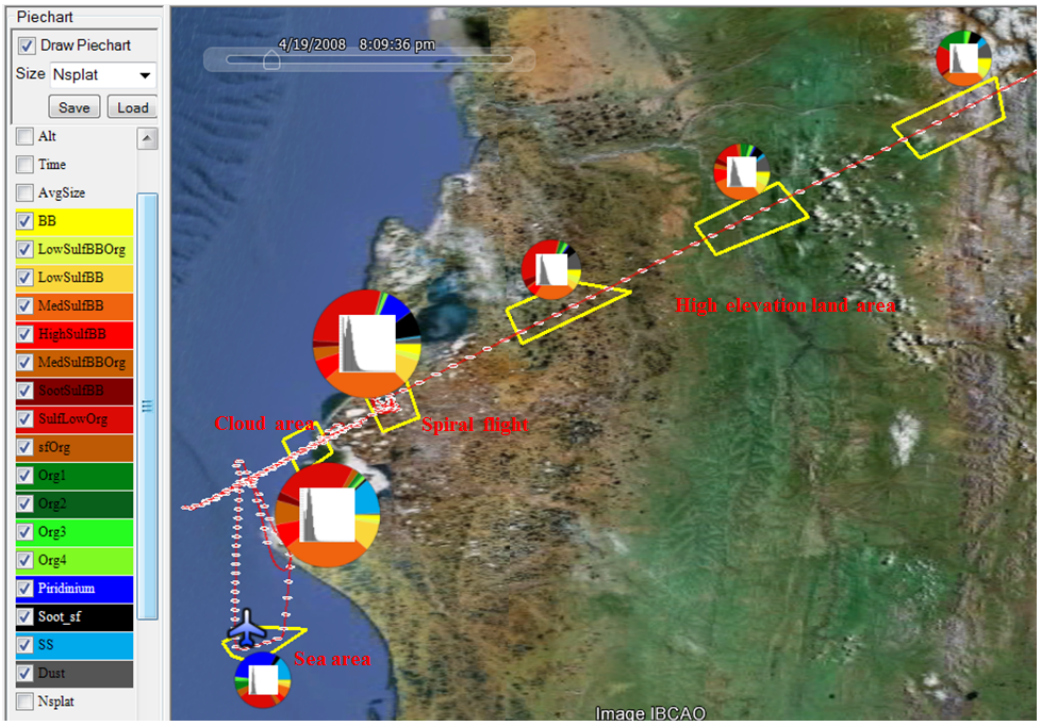


Fig. 8 **An overview of particle compositions and size changes along the flight. The flight track is marked as a red line and each of the one-minute-spaced data points is superimposed as a grey ellipse. The polygon selection tool is used to outline several interesting areas (indicated by yellow polygons and labeled in red) and the corresponding PHDs are drawn nearby. PHDs are sized by the variable NSplat to allow an assessment of compression. The colors for the pie charts are assigned by the scientists via a color map popup widget, applying their domain standards. This overview visualization shows significant spatial variability in particle composition and size, which is consistent with previous reports that show a highly stratified atmosphere.**

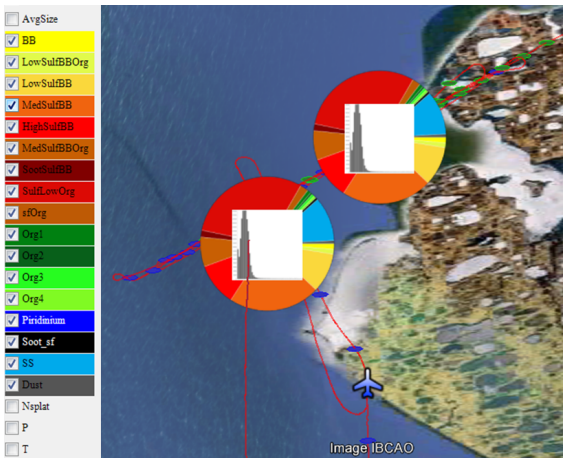


Fig. 9 **Cloud data analysis. Here the scientists first used the PCP display to filter data points with clouds (Nd>20). They then applied a second filter to select only data points in which cloud droplets are sampled (CVI =1, green points on the flight track) to make the upper PHD. They then changed the filter to select data points where only the particles between the cloud droplets are sampled (CVI=0, blue track points) and obtain the lower PHD. The fact that these two PHDs are virtually identical is a significant discovery in climate research. This has never been done before and changes much of what has been assumed in the field.**

make possible to mine the large and complex data these instruments produce, with the appropriate resolution. In Fig. 10 we present an example in which we analyze changes in particle composition as a function of altitude. We zoom in on the spiral ascent, where the aircraft climbed from a few hundred meters to an altitude of 7000 m. Our program makes it possible to view the particle composition and particle size distributions at each data point by simply clicking on its icon. The figure shows 13 pie charts, clearly illustrating that particle compositions change significantly with altitude and that the changes are not monotonic. Similar filamentous structures are observed on horizontal legs as well (not shown here) and exhibit large changes include large changes in particle number concentrations and size distributions as well. Our collaborators state that these complex structures have very important implications for climate modeling as well.

7 Conclusions and Future Work

We presented an interactive framework that allows users to visualize and analyze multi-field geospatial data in

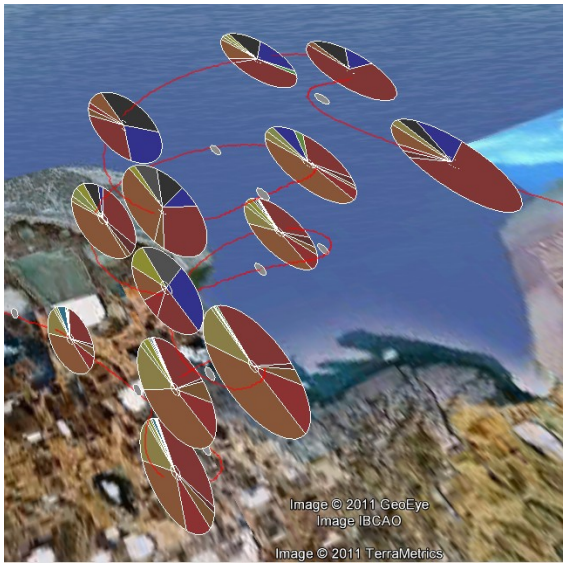


Fig. 10 Changes in particle composition as a function of altitude. We zoom into the flight's spiral ascent, where the aircraft climbed from a few hundred meters to an altitude of about 7000 m. The pie charts clearly illustrate that particle compositions change significantly with altitude and that the changes are not monotonic. Here we use the first PHD rendering method (G1) in Section 5.3 because it better shows the alleviation.

a dual-domain framework consisting of a geographic and a multivariate visualization interface. We used parallel coordinates for the latter and proposed a few enhancements both for interaction and for visualization. But our main focus was put on the interactions in and with the geo-display. Here we have taken advantage of Google Earth as a versatile 3-D geo-browser. We showed how this platform can be programmed to enable in-window selection and brushing operations and how these can be coupled with the parallel coordinate display. We also showed how some multivariate information can be directly embedded into Google Earth, in form of pie and bar chart design primitives.

Future work will focus on refining our system also with other domain scientists. Further, we also plan to make some advances on the topic of dimension reduction. In conversations with these and other scientists we found that one way to accomplish this might be to employ mathematical relationships and physics laws, which are often non-linear. These types of dimension reductions and aggregations will be difficult to detect with traditional dimension reduction which are purely algorithmic. However, we believe that our greatest contribution is the fact that our system was conceived in tight collaboration with a team of domain scientists and already enabled them to make a number of groundbreaking discoveries in their research field – the effect of aerosols on global warming. Our system is able to manage large heterogeneous data collections without problem and makes it easy to associate spatial effects in multi-field geo-referenced data.

Acknowledgements

Partial support for this research was provided by the US National Science Foundation (Nos. 1050477, 0959979, and 1117132), by a Brookhaven National Lab LDRD grant, by the US Department of Energy (DOE) Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences, and by the IT Consilience Creative Project through the Ministry of Knowledge Economy, Republic of Korea. Some of the research was performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the DOE's OBER at Pacific Northwest National Laboratory (PNNL). PNNL is operated by the US DOE by Battelle Memorial Institute under contract No. DE-AC06-76RL0 1830.

References

[1] S. Nagaraj and V. Natarajan, Relation-aware isosurface extraction in multifield data, *IEEE Trans. Vis. Comput. Graph.,* vol. 17, no. 2, pp. 182-191, 2011.

[2] L. Gosink, C. Garth, J. Anderson, W. Bethel, and K. Joy, An application of multivariate statistical analysis for query-driven visualization, *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 3, pp. 264-275, 2011.

[3] J. Blaas, C. Botha, and F. Post, Interactive visualization of multi-field medical data using linked physical and feature-space views, in *EuroVis*, 2007, pp. 123-130.

[4] A. MacEachren, M. Wachowicz, R. Edsall, D. Haug, and R. Masters, Constructing knowledge from multivariate spatiotemporal data: Integrating geographical visualization with knowledge discovery in database methods, *Int' J. of Geographical Information Science*, vol. 13, no. 4, pp. 311-334, 1999.

[5] A. Inselberg and B. Dimsdale, Parallel coordinates: A tool for visualizing multi-dimensional geometry, in *Proc. IEEE Visualization*, 1990, pp. 361-378.

[6] G. Andrienko and N. Andrienko, Exploring spatial data with dominant attribute map and parallel coordinates, *Env. and Urban Systems*, vol. 25, no. 1, pp. 5-15, 2001.

[7] M. Jern, T. Astrom and S. Johansson, GeoAnalytics tools applied to large geospatial datasets, in *Proc. IEEE Information Visualization*, 2008, pp. 362-372.

[8] D. Guo, J. Chen, A. MacEachren, and K. Liao, A visualization system for space-time and multivariate patterns, *IEEE Trans. Vis. Comput. Graph.*, vol. 12, no. 6, pp. 1461-1474, 2006.

[9] J. Wood, J. Dykes, A. Slingsby, and K. Clarke, Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup, *IEEE TVCG*, vol. 13, no. 6, pp. 1176-1183, 2007.

[10] *http://publicwiki.deltares.nl/display/OET/OpenEarth*, 2012.

[11] *http://publicwiki.deltares.nl/display/OET/KML+Screenshots*, 2012.

[12] S. Kisilevich, D. Keim, and L. Rokach, GEO-SPADE: A generic Google Earth-based framework for analyzing and exploring spatio-temporal data, in *Int' Conference on Enterprise Info. Systems*, 2010, pp. 13-20.

[13] A. Zelenyuk and D. Imre, Beyond single particle mass spectrometry: Multidimensional characterisation of individual aerosol particles, *Int. Rev. Phys. Chem.*, vol. 28, no. 2, pp. 309-358, 2009.

[14] A. Zelenyuk, J. Yang, D. Imre, and E. Choi, SPLAT II: An aircraft compatible, ultra-sensitive, high precision instrument for in-situ characterization of the size & composition of fine & ultrafine particles, *Aerosol Sci. Technol.*, vol. 43, no. 5, pp. 411-424, 2009.

[15] G. M. McFarquhar, S. J. Ghan, J. Verlinde, A. Korolev, J. W. Strapp, B. Schmid, J. M. Tomlinson, M. Wolde, S. D. Brooks, D. J. Cziczo, M. K. Dubey, J. Fan, C. J. Flynn, I. Gultepe, J. M. Hubbe, M. K. Gilles, A. Laskin, P. Lawson, W. R. Leaitch, P. S. Liu, X. Liu, D. Lubin, C. Mazzoleni, A. M. Macdonald, R. C. Moffet, H. Morrison, M. Ovchinnikov, M. D. Shupe, D. D. Turner, S. Xie, A. Zelenyuk, K. Bae, M. Freer, and A. Glen. Indirect and semi-direct aerosol campaign (ISDAC): The impact of arctic aerosols on clouds, *Bulletin of the American Meteorological Society*, vol. 92, no. 2, pp. 183-201, 2011.

[16] G. Schmidt, G. Bigg and E. Rohling, Global seawater oxygen-18 database — v1.21, http://data.giss.nasa.gov/o18data, 2011

[17] W. Dansgaard, Stable isotopes in precipitation, *Tellus*, vol. 16, no. 4, pp. 436-468, 1964.

[18] W. Mook, The oxygen-18 content of rivers, *SCOPE*, vol. 52, pp. 565-570, 1982.

[19] V. Ferronsky and V. Brezgunov, Stable isotopes and ccean dynamics, in *Environmental Isotopes in the Hydroshpere*, John Wiley, New York, 1982, pp. 1-27.

[20] K. McDonnell and K. Mueller, Illustrative parallel coordinates, *Computer Graphics Forum*, vol. 27, no. 3, pp. 1031-1027, 2008.

[21] E. Nam, Y. Han, K. Mueller, A. Zelenyuk, and D. Imre, ClusterSculptor: A visual analytics tool for high-dimensional data, in *IEEE Symposium on Visual Analytics Science and Technology*, 2007, pp. 75-82.

[22] J. Hartigan, Printer graphics for clustering, *J. of Statistical Computation and Simulation*, vol. 4, no. 3, pp. 187-213, 1997.

[23] J. Conway and N. Sloane, *Sphere Packings, Lattices and Groups, 3rd Edition*, Springer Verlag, 1998.

[24] Z. Zhang, K. T. McDonnell, K. Mueller, A network-based interface for the exploration of high-dimensional data spaces, in *IEEE PacificVis*, 2012, pp. 17-24.

[25] E. Koua, A. Maceachren, and M. Kraak, Evaluating the usability of visualization methods in an exploratory geo-visualization environment, *J. of Geographical Information Science*, vol. 20, no. 4, pp. 425-448, 2006.

[26] F. Chapman, Winforms-geplugin-controls-library, http://code.google.com/p/winforms-geplugin-control-library/, 2011.

[27] G. Bigg, and E. Rohling, An oxygen isotope data set for marine water, *J. Geoph. Res.*, vol. 105, no. C4, pp. 8527-8535, 2000.

[28] B. A. Albrecht, Aerosols, cloud microphysics, and fractional cloudiness, *Science*, vol. 245, no. 4923, pp. 227-1230, 1989.

[29] K. N. Liou and S. C. Ou, The role of cloud microphysical processes in climate — An assessment from a one-dimensional perspective, *J. Geophys. Res.-Atmos.*, vol. 94, pp. 8599-8607, 1989.

[30] L. A. Barrie, Arctic air pollution: An overview of current knowledge, *Atmospheric Environment — Part A General Topics*, vol. 20, no. 4, pp. 643-663, 1986.

[31] K. Hara, S. Yamugata, T. Yamanouchi, K, Sato, A. Herber, Y. Iwasaka, M. Nagatani, and H. Nakata, Mixing states of individual aerosol particles in spring Arctic troposphere during ASTAR 2000 campaign, *J. of Geophysical Research-Atmospheres*, 2003.

[32] P. K. Quinn, T. S. Bates, E. Baum, N. Doubleday, A. M. Fiore, M. Flanner, A. Fridlind, T. J. Garrett, D. Koch, S. Menon, D. Shindell, A. Stohl, and S. G. Warren, Short-lived pollutants in the Arctic: Their climate impact and possible mitigation strategies, *Atmospheric Chemistry and Physics*, vol. 8, pp. 1723-1735, 2008.

[33] L. F. Radke, J. H. Lyons, D. A. Hegg, P. V. Hobbs, and I. H. Bailey, Airborne observations of arctic aerosols, I: Characteristics of Arctic Haze, *Geophysical Research Letters*, vol. 11, no. 5, pp. 393-396, 1984.

[34] A. Sirois and L. A. Barrie, Arctic lower tropospheric aerosol trends and composition at Alert, Canada: 1980-1995, *Journal of Geophysical Research-Atmospheres*, vol. 104, pp. 11599-11618, 1999.

[35] G. McFiggans, P. Artaxo, U. Baltensperger, H. Coe, M. C. Facchini, G. Feingold, S. Fuzzi, M. Gysel, A. Laaksonen, U. Lohmann, T. F. Mentel, D. M. Murphy, C. D. O'Dowd, J. R. Snider, and E. Weingartner, The effect of physical and chemical aerosol properties on warm cloud droplet activation, *Atmospheric Chemistry and Physics*, vol. 6, no. 9, pp. 2593-2649, 2006.

Zhiyuan Zhang received his BEng from Shandong University, China and currently pursues a PhD degree in computer science at Stony Brook University. His research interests are information visualization and visual analytics, with a focus on healthcare informatics, multivariate data visualization, and correlation analysis. He was awarded the IBM PhD Fellowship for 2013. For more information, see http://www.cs.sunysb.edu/ zyzhang.

Xiaonan Tong is an undergraduate student at Stanford University. His academic interests are computer science, human computer interaction, visualization, and product design. He participated in this project during summer 2011.

Kevin T. McDonnell received his BS, MS, and PhD degrees in computer science from Stony Brook University in 1998, 2001, and 2003, respectively. Since 2004 he has been a member of the full-time faculty of Dowling College, where he is a tenured associate professor of computer science and mathematics and where he is co-PI of Dowling's NSF-funded Robert Noyce Teacher Scholarship Program. His research interests include scientific and information visualization, visual, analytics and human computer interaction.

Alla Zelenyuk received her PhD in chemical physics from Moscow Institute of Physics and Technology, Russia, and is currently senior research scientist at Pacific Northwest National Laboratory. Her research interests include real-time multidimensional characterization of physical and chemical properties of individual aerosol particles. She has authored more than 60 peer-reviewed papers and reports.

Dan Imre received his PhD in physical chemistry from Massachusetts Institute of Technology and is currently working as a consultant with Imre Consulting. His expertise includes single particle mass spectrometry, multidimensional single particle characterization, and data analysis. He has authored more than 110 peer-reviewed papers.

Klaus Mueller received his PhD degree in computer science from the Ohio State University. He is currently a professor of Computer Science at Stony Brook University, where he also holds co-appointments in the Biomedical Engineering and Radiology Departments, and he serves as the chair of the Computer Science department at SUNY Korea in Songdo, Korea. His present research interests are computer and volume graphics, visualization, visual analytics, human computation, medical imaging, and computer vision. He won the US National Science Foundation CAREER award in 2001 and the SUNY Chancellor Award for Excellence in Scholarship and Creative Activity in 2011. He served as a co-chair at various conferences, such as IEEE Visualization and he is the current chair of the IEEE Technical Committee on Visualization and Computer Graphics. He has authored more than 150 journal and conference papers which have been cited more than 4 200 times according to Google Scholar, and he has participated in 15 tutorials at international conferences on various topics in visualization and medical imaging. He is a senior member of the IEEE. For more information, see http://www.cs.sunysb.edu/ mueller.