Visualization of the CReSIS Greenland Data Sets for the Polar Grid High Performance Computing System

Michael Austin (ECSU), Shaketia McCoy (MVSU), Frensha Slaughter (ASU)
Undergraduate Research Experience in Ocean, Marine, and Polar Science 2009
Campus Box 672, Elizabeth City State University, Elizabeth City, NC 27909

Abstract— **The Center for Remote Sensing of Ice Sheets (CReSIS) has been compiling Greenland ice sheet thickness data since 1993. The airborne program utilizes a 150 MHz radar echo sounder to measure the ice thickness. The data is currently available on the CReSIS web site in various formats including PDF, Matlab, and plain text files. These formats are not readily usable in the classroom environment as a visual representation of the ice depth is not available for each expedition.**

During the Undergraduate Research Experience in Ocean, Marine and Polar Science 2009 program, the Greenland Data Visualization Team took the CReSIS data and created a 4-D visualization consisting of depth, time, latitude, and longitude. This visualization was created utilizing HTML, JavaScript, and PHP. Microsoft Excel was used to filter the raw data downloaded from the CReSIS site. The team then statistically analyzed the Greenland ice sheet thickness data for calculated, missing, and actual depth readings. The goal of this project was to present the CReSIS data via the web in a visual format to elementary, undergraduate, and graduate students for research and education. This visualization package and corresponding data will eventually be migrated to the Elizabeth City State University Polar Grid High Performance Computing System.

I. INTRODUCTION

A. Overview

The Center of Remote Sensing of Ice Sheets (CReSIS) was established in 2005 by the National Science Foundation (NSF). The CReSIS mission is to develop new technologies and computer models to measure and predict the response of sea level change to the mass balance of ice sheets in Greenland and Antarctica. In addition to the CReSIS headquarters being located at University of Kansas, CReSIS also works in partnership with Elizabeth City State University, Haskell Indian University, Ohio State University, Pennsylvania State University, and University of Maine. Research, education, and knowledge transfer are the CReSIS's three main programs.

The Polar Grid goal is to obtain and organize the computing communications needed to investigate the critical problems in glacial melting. The Polar Grid project is collaboration with Elizabeth City State University and Indiana University funded through the National Science Foundation (NSF) Major Research Instrumentation (MRI). Polar Grid is in partnership with CReSIS, CERSER, and the Byrd Polar Research Center.

B. Data Visualization

Data Visualization is making data in raw format available in a visual representation. Data Visualization can be visual representations of information, data, or knowledge. Data Visualization is used often in the area of teaching, research, and development.

The Greenland Data Visualization team goal was to effectively and clearly communicate data collected during the Greenland expeditions from 1993 to 2007. The team made scatter plots and line graphs on Excel to convey the flight paths and depth lines. The team then produced 4-D images consisting of time, depth, latitude, and longitude using the Google Map application, HTML, JavaScript, and PHP.

C. Terminology

During the process of the research, many terms needed to be defined. It was evident that the team needed to learn the terminology and get familiarized with the terms. In addition, the team needed to create an online glossary for future reference. The glossary would include unfamiliar terms that the group encountered with during the research process. Also incorporated were the formulas used in Excel in the glossary.
http://nia.ecsu.edu/ureomps2009/teamgreenland/glossary.html

D. Data Collection

The purpose of data collection is to acquire information to keep on file, to reach choices about significant issues, and to pass information on to other people. The data that our team used was collected by the CReSIS project. A Twin-Otter fixed wing airplane was utilized to chart the ice sheet layers. The data was obtained with antennas attached to the wings of the plane that produced radio waves. The data was received every 100 microseconds over the course of multiple flights. After the data is collected, the raw data received must be filtered and placed in context through visual means.
https://www.cresis.ku.edu/research/data/greenland_data.html

E. Data Format

The data found on CReSIS came in many formats such as PDF, Matlab, and plain text files. This project entailed the use of the plain text files. During the process of collecting files from the web, several file extensions were encountered that included text (.txt), height (.ht), and (.th). The plain text files were converted to Excel workbooks (.xlsx).

Microsoft Office Excel 2007 was used to format and analyze plain text files. After the team data formatting was complete, the files were converted to comma separated value (.csv) to be read by PHP code. This also removed the calculations and formulas from the spreadsheet.

II. EXCEL

A. Overview

Microsoft Office Excel 2007 is a software package that can be used to create and format spreadsheets. Excel is also used to analyze and share data in an understandable form. This project used Excel to filter and manage collected data as well as create line graphs and scatter plots. In addition, the team utilized Excel to statistically analyze the Greenland ice sheet thickness data for calculated, missing, and actual depth readings.

B. Capabilities

Microsoft Office Excel 2007 has increased it row amount to one million, and the column amount has expanded to sixteen thousand. Excel quickly formats cells and tables. Excel can create many different charts including radar, scatter, and line. Sorting and filtering are two important basic analyses that organize and locate data. Sorting allows the arranging items in sequence. Filtering allows the reduction of information by different sets. In addition, Excel permits the input formulas. Formulas are equations that perform calculations on values in a workbook.

C. Final Output Formats

The final output formats using Excel were Excel workbook (.xlsx) and comma separated value (.csv) files. By saving the data files as spreadsheet and plaintext files, it gives the researcher a choice file format to work with.

III. GOOGLE MAPS

A. Overview

Google Map is a free web mapping service application. Google Maps offer aerial photos as well as plain maps and satellite views. It allows a user to embed maps on third-party websites.

B. Inputs

The Greenland Data Visualization team used Google Maps to input flight paths and the depth of the ice. Flight paths included the latitude and longitude for the plane's position. The depth of the ice included the time in addition to the depth. In order to insert the depth, a line graph was created using Excel and converted to JPG format.

http://mmt.cs.ecsu.edu/pgdataviz/pg-data-set-request.php

C. PHP

PHP is used widely for general purpose especially suited for Web development and can be embedded into HTML. Originally PHP was designed for producing web pages but scripting language can do far more. PHP has evolved the capability to use command line interface and can also create standalone graphical applications. Excel offers the opportunity to save files as .csv. The PHP script in this project parsed the CSV files into separate variables and created JavaScript code to be used by Google Maps.

D. JavaScript

JavaScript is a scripting language that allows you access to objects within other applications. JavaScript code was used to display the information in the Google Maps application. In addition, JavaScript allowed the team to show the line graphs created in Excel as representation of depth via the web. JavaScript was used in combination with PHP in order to present the CReSIS data via the web in a visual format.

IV. PROJECT PHASES

A. Phase I: Download

The first phase of our project was to download the needed files from the CReSIS archive of data files located at: http://www.cresis.ku.edu/research/data/Greenland_data.html. These files were the datasets that contained the information recorded while using radar to determine the depth of ice from 1993 to 2007 in Greenland. The files initially contained: latitude, longitude, time, depth, and a height column. The data consisted of 103 days spread throughout hundreds of text files. After downloads were complete, the files were then saved to portable storage devices and personal computers for backup purposes.

B. Phase II: Formatting

In order to be able to use the files for our project, some formatting needed to be completed. The files initially were text files and in order to edit them efficiently and use calculations, the files were opened and saved with Microsoft Excel. Upon opening them, certain columns needed to be added to the ones that already existed. The columns labeled *Index*, *Date* and *Comment* were added. The Index column was added for the sole purpose of reducing the number of lines later on. This was placed in column "A". The Date column was added to better organize the data, set to column "B". If one of the files were to be misnamed, the date column would quickly identify the file. Finally the Comment column was added to determine what data was missing, actual, or calculated data. With this information percentages can be taken for the above information. Figure 1 shows the format of the column in Excel.

C. Phase III: Converting and Replacing

When starting with this phase it was brought to our attention that the longitude was calculated by 0-180 but needed to be 0-360. To correct the longitude for our data we needed to create an extra column that would take place in column D (New Longitude). The equation that took place here to correct our longitude was "=E1-360". Now that the Longitude was corrected there were certain data in each year that was not consistent. To have a finalized product the files

Index	Date	Latitude	New Long	Longitude	Time	New Depth	Depth	Comment	Depth Type
A	B	C	D	E	F	G	H	I	J
36488	20060603	66.363073		-17.985462	56042.027	0	0		M
36487	20060603	66.383171		-17.985225	56042.239	0	0		M
36488	20060603	66.363268		-17.984989	56042.451	0	0		M
36489	20060603	66.383365		-17.96475	56042.663	0	0		M

Fig. 1. Image shows the format that the columns were set to.

needed to all be consistent. The problem here was when some data could not be recorded it was represented as NaN, on others it was represented as -999 or 0.

To correct this, all the above representations were located with the find and replace function that is located within Excel. Once the instances of any of these names were found, they were then replaced with "0". The problem that occurred with longitude was now happening with Depth. In order to represent our data this needed to be altered as well. A new column was added to the left of the G column (New Depth) with the following equation "=IF((D1="0"),"0",-(D1))".This would give the negative number for the depth and if the depth in a particular row was 0 then it would continue to stay 0. The last procedure for this phase was to ensure that there was a Data type column. This was placed in the J column in Excel. This column would then be used to Determine whether or not the data is "M" (missing), "A" (actual), or "C" (calculated). The equation that was made to fit this need was "=IF((D1="0"),"M",IF((I1="CALCD"), "C","A"))".

D. Phase IV: Combine Data

Initially the data sets that were downloaded consisted of many files for each day. To look at this file as a whole and to have it work for the final product these files needed to be joined. To join these files together within one spread sheet, it would consist of many hours of copying and pasting data into one big file. To save time on this Phase, a macro was created to combine the data more efficiently and quicker. These files were then saved into a new Excel file.

E. Phase V: Sorting and Re-Indexing

After combining the data by their days the data needed to be ordered by the time is was recorded at. This can be done by choosing the time column and going to the sort & filter option. Once here the data must be sorted by A to Z. After sorting, the Index column needed to be re-calculated. This was due to the fact that it is now repeating for every time a copy and paste procedure took place. To perform the sort all data must be highlighted in that column. Once that is done series was clicked and the column was numbered with a step value of 1.

F. Phase VI: Reducing and Graphing

When it came time to graph the data, so that there was visual representation, certain changes needed to be made to the files. In order to graph the files there needed to be a maximum of 32 thousand lines in each file. To reduce the amount of lines in each data set, an equation was created. In order to start reducing a new work sheet had to be created within Excel. In the A1 column the equation was entered. The equation for reducing is as follows

"IF("sheetname"!$A1/ROUNDUP(COUNT("sheetname"!$A:$A)/32000,0)=INT("sheetname"!$A1/ROUNDUP(COUNT ("sheetname"!$A:$A)/32000,0)),"sheetname"!A1,"")", where "sheetname" is the name of the first work sheet in Excel. The roundup portion of the equation is what rounds a number to a specified number of digits. The IF is a function that tests the equation and gives output that is determined on if the statement if true or false. The INT is simply an integer.

After entering the equation in the second worksheet, it needs to be filled horizontally to the J column and dragged down to the location where the first work sheet ends. Upon doing so your computer may seem to be non responsive for a few minutes to a couple of hours depending on the file size of which you are dealing with. After the calculations were complete, there were gaps in between the data. These gaps can be easily fixed by highlighting the first column, in this case the A column and choosing sort A to Z. When sorting, the option to expand the selection should be chosen. After the data is sorted graphs have to be made.

Graphs started out with line graphs. Figure 2 shows the line graph. The data that need to go into the graph was column G (New Depth). This involved highlighting the data in column G then clicking the insert tab at the top of the program and selecting line graph. Once the graph showed on the screen the title, legend, and the key needed to be deleted if it was there. After this right click the top axis of data and select format. In the window that pops up, Axis Labels must be set to LOW. On the graph the actual data needed to be changed. This was done by right clicking the data and selecting format data. On the new window, edit must be selected under the horizontal axis. A range will be asked to be given. For the range the column F was selected. After that the color of the line on the line graph needs to be changed to 25, 0,100 in the RGB scale. This is done by right clicking the line and selecting the format data. Here the line color was changed to meet the RGB numbers. Finally the graph had to be resized by choosing the format tab, locking the aspect ratio and setting the height to 7". This would then be taken into Photoshop by means of copy and pasting. The image is then saved for the web as a jpeg with a width of 900 pixels.

When creating the second graph the data that needed to go into the graph was column C and E (Latitude and Longitude). Figure 3 shows the scatter plot. This involved highlighting both the data in column C and E, then clicking the insert tab at the top of the program and selecting scatter plot. Once the graph showed on the screen the title, legend, and the key needed to be deleted if it was there. After this right click the top axis of data and select format. In the window that pops up, Axis Labels must be set to LOW. On the graph the actual data needed to be changed. This was done by right clicking the data and selecting format data. On the new window, edit must be selected under the vertical axis. A range was given. In the ranges given C and E needed to be swapped. Where ever there is a C replace with E and wherever there is a E replace with a C. After that change has taken place, the color of the line on the line graph needs to be changed to 25, 0,100 in the RGB scale. This is done by right clicking the line and selecting the format data. Here marker fill color was changed to no fill. Under the marker line option a solid line with the RGB requirements was selected. Finally the graph had to be resized by choosing the format tab, locking the aspect ratio and setting the height to 7". This would then be taken into Photoshop by means of copy and pasting. The image is then saved for the web as a jpeg with a width of 900 pixels.
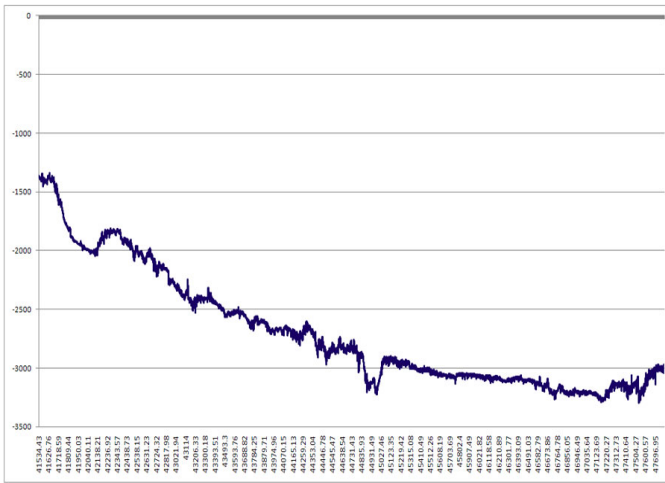
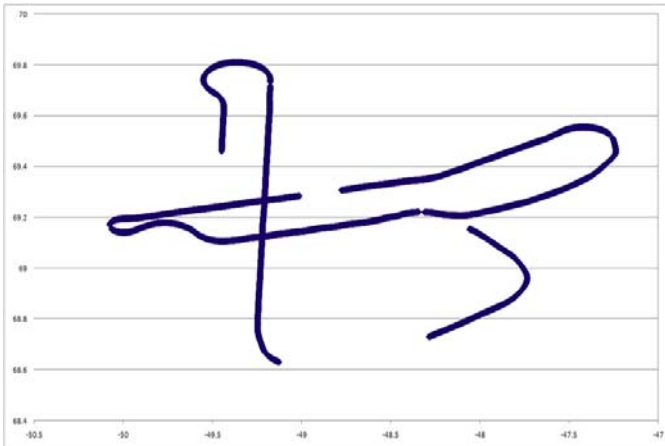Fig. 2. The line graph is used to visualize the depth of the ice over the entire flight.

Fig. 3. Scatter plot is used to visualize the flight path taken when the data was recorded.

G. Phase VII: Determine the Type of Data

This phase involved making a copy of the phase VI data because there was no need for the graphs and the worksheets that accompanied the data after that phase. Now with the correct data it was time to determine a few things. The occurrences for the following needed to be recorded to better understand what kind of data was being worked with, total lines, actual lines (A), missing lines (M), and calculated lines (C). To successfully perform this task several equations needed to be inserted within the Excel file. To calculate for the number of total lines the equation "=COUNTA(A:A)" was used. For the number of A, M, and C lines the equation "=COUNTIF(J:J, "M")" was used by replacing the quotation letter with what you were looking for. The reason these two equations are different is because the A, M, and C data was all in one column and had mixed data. During this phase a macro was recorded using the macro option in Excel and applied to all spreadsheets to decrease the amount of time spent on it. Figure 4 shows the format of the determined data.

Fig. 4. Format that phase VII was set in to find the occurrences of each type of data.

H. Phase VIII: Reducing for CSV

To begin this phase the phase V files needed to be copied in order to just use the pure data. In order to save the Excel files there needed to be a maximum of 200 lines. To achieve this requirement an equation was used to reduce the data. This equation was "=IF(S1!$A1/ROUNDUP(COUNT(S1!$A:$A)/200,0)=INT(S1!$A1/ROUNDUP(COUNT(S1!$A:$A)/200,0))1!A1,"")". Once the equation was processed through and the data was reduced the file was saved as Comma Separated Value files (.CSV) through the Microsoft Excel "save as" feature.

I. Phase IX: Formatting for PHP

To format for PHP the recently saved .CSV file needed to be opened in Excel to make a few adjustments. While in Excel the columns A, E, H, and I (Index, Longitude, Depth, and Depth Type) needed to be deleted. After this alteration the file was then saved. The same file was then opened with Microsoft Word. The find and replace function within Microsoft Word was then used to find "^p" (paragraph mark) and then replace it with a ",^p" (comma paragraph mark). Without replacing the marks, the PHP code would not be able to operate. The last step in assuring the PHP code would operate as needed was to open the file in Notepad (a basic text editor that accompanies Microsoft Windows). While in Notepad there were two blank lines at the end of the file. These lines were added by Microsoft Word and needed to be deleted in order for the PHP code to operate.

J. Phase X: Downloadable files

This phase involved removing the calculations and formulas from the Excel sheet and giving future researchers the choice of both a .CSV and an .XLSX. To start this procedure Phase VIII files needed to be copied and used. The files were then opened in Excel and the second worksheet was then deleted. After the deletion the file was saved as a .CSV and the Excel file was closed. The .CSV files were reopened in Excel and the columns A, E, H, and I were deleted. These files were then saved as an Excel file and the previous file was overwritten.

V. OUTCOME

A. Data Visualization

Upon completing the project phases the data could now be visualized. The data was visualized not only by the graphs that were created within Excel but also within Google Map. The

team utilized the capabilities of Google Map, JavaScript, PHP and Hyper Text Markup Language (HTML) to create a user interactive representation of the data. With the Google Map representation it first asks the user to select the day and year of which data is needed, shown by figure 5. After the user chooses their preference in data, a new window loads up the flight line along with the depth that was recorded for that day, shown by figure 6.

B. URLs

The finish product consisted of a page that contained the flight lines and their location over Greenland (http://mmt.cs.ec su.edu/pgdataviz/pg-data-set-flightlines.php). Also located on this site was a 4-D visualization that consisted of the latitude, longitude, depth and time.

http://mmt.cs.ecsu.edu/pgdataviz/pg-data-set-request.php

For future work and for those who would like to download the files that were used in the research, a page was added so this is possible. The files are available in Excel format and in .CSV.
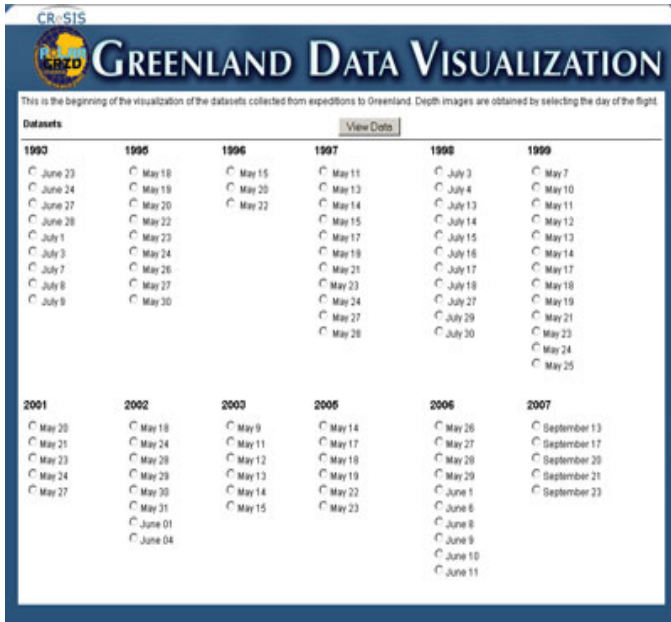
http://mmt.cs.ecsu.edu/pgdataviz/dd.php

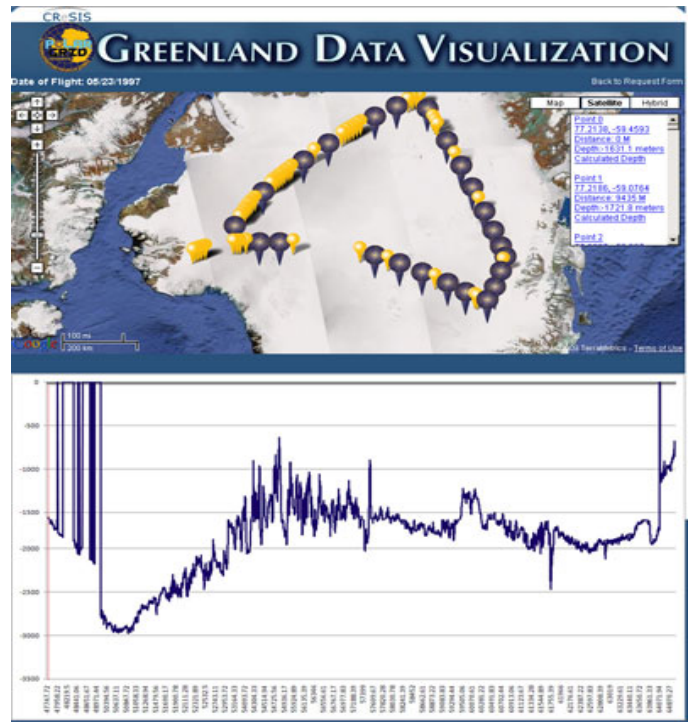Fig. 5. Format that phase VII was set in to find the occurrences of each type of data.

Fig. 6. Format that phase VII was set in to find the occurrences of each type of data.

C. Statistics

While working with the data sets, certain characteristics were recorded from them. These characteristics were "M" (missing), "A" (actual), and "C" (calculated). These were calculated during phase VII. The statistics for M, A, and C are shown by figure 7. Statistics were also taken on the total lines for each day (figure 8) and also the amount of day's data was recorded each year (figure 9). Graphs were then made to reflect the statistics.
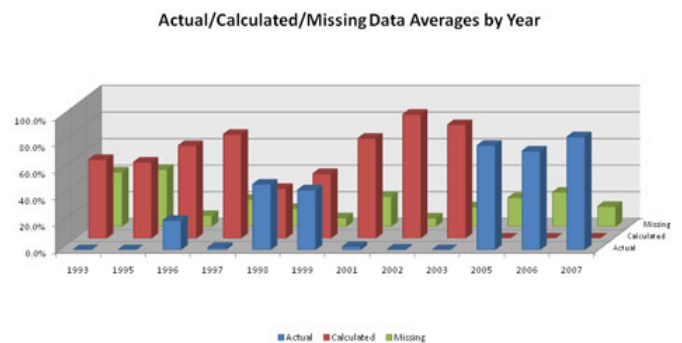

Fig. 7. This graph shows the comparison for each year of how much of the data was Actual, Calculated, and Missing.
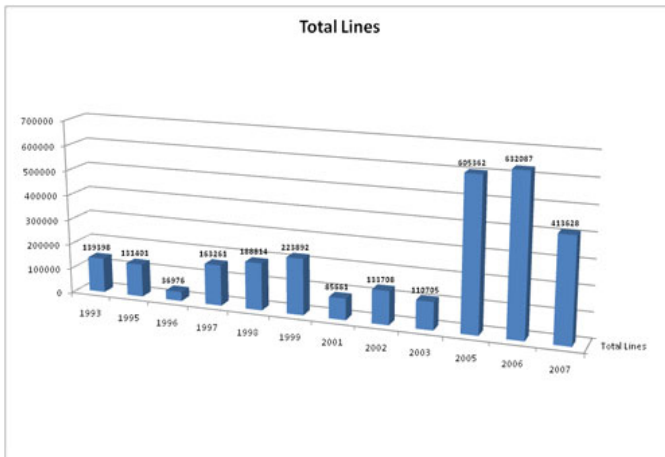
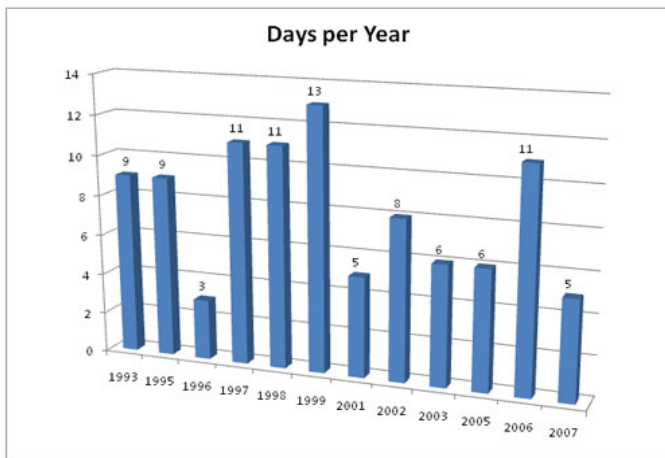Fig. 8. This graph shows the total lines of data for each year

Fig. 9. This graph shows how many days data was recorded in each year.

VI. FUTURE WORK

A. 2008 - Present

In this research the data that was utilized was spanned from 1993 to 2007. At the current time the years from 2008 and 2009 have not been released for research. In the future this data could also be utilized to provide a more accurate visual of what is happening to the ice now.

B. Antarctic Data

On the CReSIS website where the Greenland ice sheet thickness data was downloaded from, there are also Antarctic data as well. This data could also be of use to make comparison with the Greenland data. The calculation of lost ice could be compared to see if there is the same amount of loss or if the climate is changing differently in the two locations. The Antarctic data is in .txt format. Each year also has the days combined into one file. This would eliminate some of the phases that were needed in the Greenland data set.

C. Excel Macros

In phases IV and VII macros were used to perform some of the tedious task. In phase IV the task was to combine the data by days. Since there were multiple files per day, copying and pasting would be an all day task. To save time a macro was created in Excels macro program, using Visual Basic (VB).

D. Standardized Data Collection Format

On the CReSIS site the data was recorded in different formats. Making the data consistent added to the work. If CReSIS could come up with a more standardized collection format it would make everything more consistent and those using the data for research would not have to alter the sets to meet their needs.

E. Relocation

At some point the data will be moved to a high computing cluster at ECSU. This cluster will be able to handle the data faster. The high computer cluster will be located in the E.V. Wilkins Building on the ECSU campus.

VII. ACKNOWLEDGEMENTS

During the duration of our project there were many people we would like to thank. The Greenland Data Visualization Team would like to thank Mr. Jeff Wood, not only for being the teams mentor but for being dedicated to the project. His hard work and persistence is what made the project possible. Also to be acknowledged is the scientists that work with CReSIS. Without them the data sets we worked with would not have been available. Thank you, Dr. Linda Hayden, for providing us the opportunity to attend such a great internship. A thank you also goes to the Center of Excellence in Remote Sensing Education and Research (CERSER).

REFERENCES

[1] Turner and Barksdale. HTML & JavaScript: Programming Concepts; South Western Educational Publishing, 2000.
[2] Greenland Data Sets,CReSIS,2009 https://www.cresis.ku.edu/research/data/greenland_data.html
[3] Climate Change,CReSIS,2008 https://www.cresis.ku.edu/news/documents/ClimateChangeFAQ.pdf
[4] Microsoft Excel, 2007 http://office.microsoft.com/en-us/Excel/default.aspx
[5] Roman, Steven. Writing Excel Macros with VBA. O'Reilly Media, Inc, 2002.
[6] Microsoft, "VBA for Excel". Microsoft. 2008 http://www.Excel-vba.com/Excel-vba-contents.html.