Globals for the Environment

Gaining Insights into CO2 Levels on the Global Oceans over the last 50+ years

by William Cheung
Toronto, February 17, 2012

The Opportunity

The Globals Database (*"Globals"*) from InterSystems with its unique flexibility can serve as the computational engine for performing big data analytics for environmental research. There is concern worldwide that global warming and climate change are among the greatest threats to our planet.[1] Addressing environmental issues of this scale requires analysis of large volumes of data collected over time to identify trends and raise alerts. Globals is ideal to fulfill this analytic requirement in opportunities with researchers in government, academia, and environmental groups. The use case presented in this paper is the application of Globals to gain insights into long-term trends in carbon dioxide levels on the global oceans.

Why should we care?

Carbon dioxide (*"CO2"*) absorbed by the oceans from the Earth's atmosphere makes seawater warmer and more acidic, creating low-oxygen ocean "*dead zone*s" where it's harder for marine life to survive.[2] Over the last 50 years especially, CO2 levels in the atmosphere have increased significantly due to human activities.[3] In prehistoric times high CO2 levels in the oceans coincided with mass extinctions of marine life.[4] It is important for us to monitor ocean CO2 levels with a goal of preserving the diversity of life in the seas.

The following sections will describe the process, data, program, and visualization of the *Globals for the Environment* project.

[1] http://www.globalissues.org/issue/178/climate-change-and-global-warming
[2] http://www.sciencedaily.com/releases/2009/04/090417161506.htm
[3] http://www.epa.gov/climatechange/science/recentac_majorghg.html#fig1
http://www.epa.gov/climatechange/science/recentac.html
[4] http://www.enn.com/enn_original_news/article/42717

The process

Globals for the Environment is a project which uses Globals adapted as a MapReduce process to perform analytics on CO2 measurements on the surfaces of the global oceans over the last 50+ years. The measurements are grouped by region of the Earth and by time period, with the measurement values used to compute minimum, maximum, and average values for each time / region combination. Results are visualized in an open web application where you are free to select a region of the Earth to see the trend in ocean CO2 levels in the region over time. For the purpose of this analysis, the granularity of the time periods is by year and that of the regions is by either 30x30 or 90x90 degree divisions of the Earth's surface (in degree units of latitude and longitude).

The data

The dataset used is the *LDEO Database Version 2010* provided freely by the CDIAC (Carbon Dioxide Information Analysis Center of the U.S. Department of Energy), Lamont-Doherty Earth Observatory department.[5] This dataset comprises more than 5.2 million measurements of surface water partial pressure of CO2 obtained over the global oceans from 1957-2011.

A small subset of this data (the first 25 rows of data) is available at:
http://globals-for-the-environment.herokuapp.com/LDEO_Database_V2010_subset.txt

The format of the data is a table of space-delimited columns:

FILENAME STN **LAT LON MONTH/DAY/YEAR** JDATE VCO2_SW TEMP_PCO2 TEMP SAL PCO2_SST PCO2_SSTPA **PCO2_TEQ** EQ_PBARO SHIPPBARO

0001 1 **-77.659 178.022 2/16/2000** 47.52030 208.24 -1.15 -1.680 34.040 198.28 20.091 **202.90** 992.70 990.89

0001 2 -77.660 178.048 2/16/2000 47.52159 207.95 -1.16 -1.690 34.050 197.99 20.061 202.60 992.70 990.79

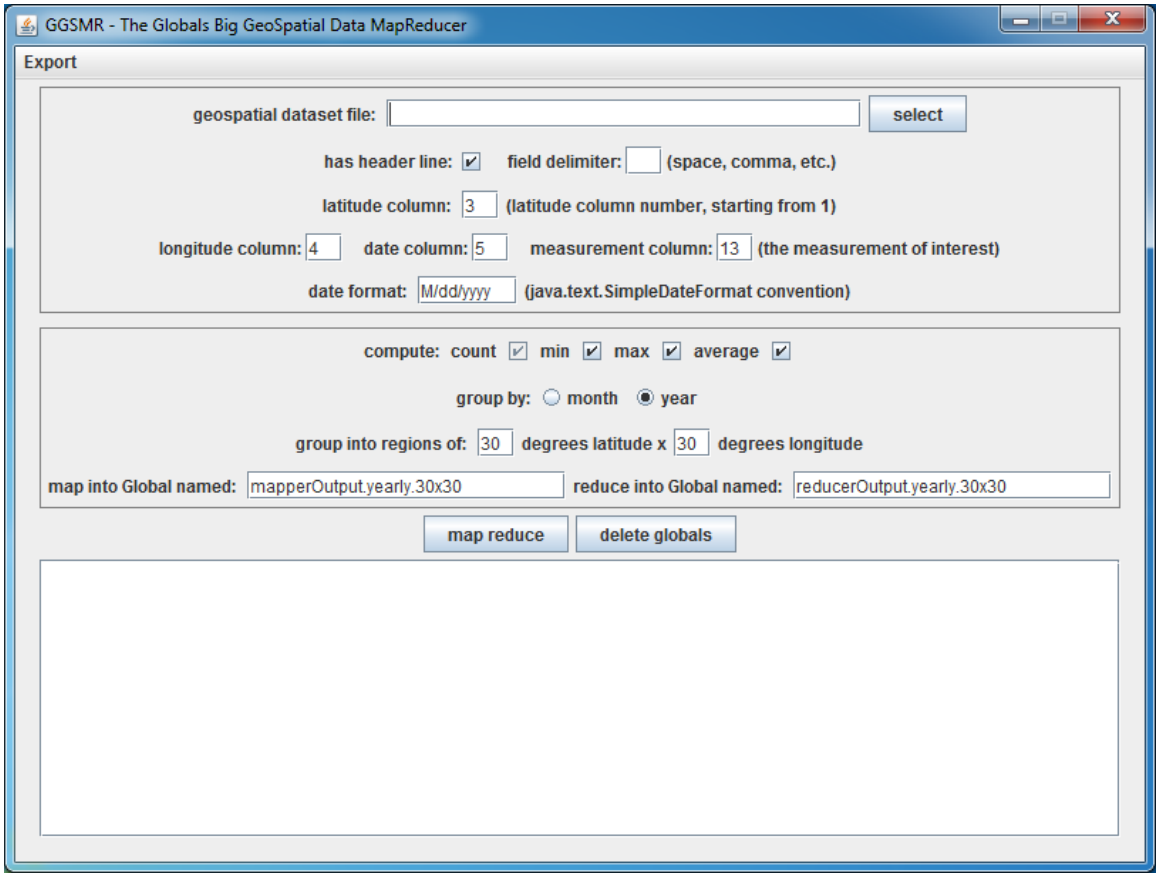
0001 3 -77.661 178.062 2/16/2000 47.52289 207.95 -1.15 -1.680 34.040 197.99 20.061 202.60 992.70 990.76

Each row in the table records a measurement taken at location LAT / LON (e.g., **-77.659 178.022**) on date MONTH/DAY/YEAR (e.g., **2/16/2000**). For our analysis we use the column PCO2_TEQ (e.g., **202.90**) as the measure of CO2 at the location and date, as this was the value actually measured (in units of microatmospheres).

The program

The program developed for this type of geospatial time-series analysis was *GGSMR - the Globals Big Geospatial Data MapReducer*, whose user interface on opening is shown on the next page.

[5] http://cdiac.ornl.gov/ftp/oceans/LDEO_Database/Version_2010/



The UI is initialized with default values corresponding to the specific LDEO dataset we want to analyze:

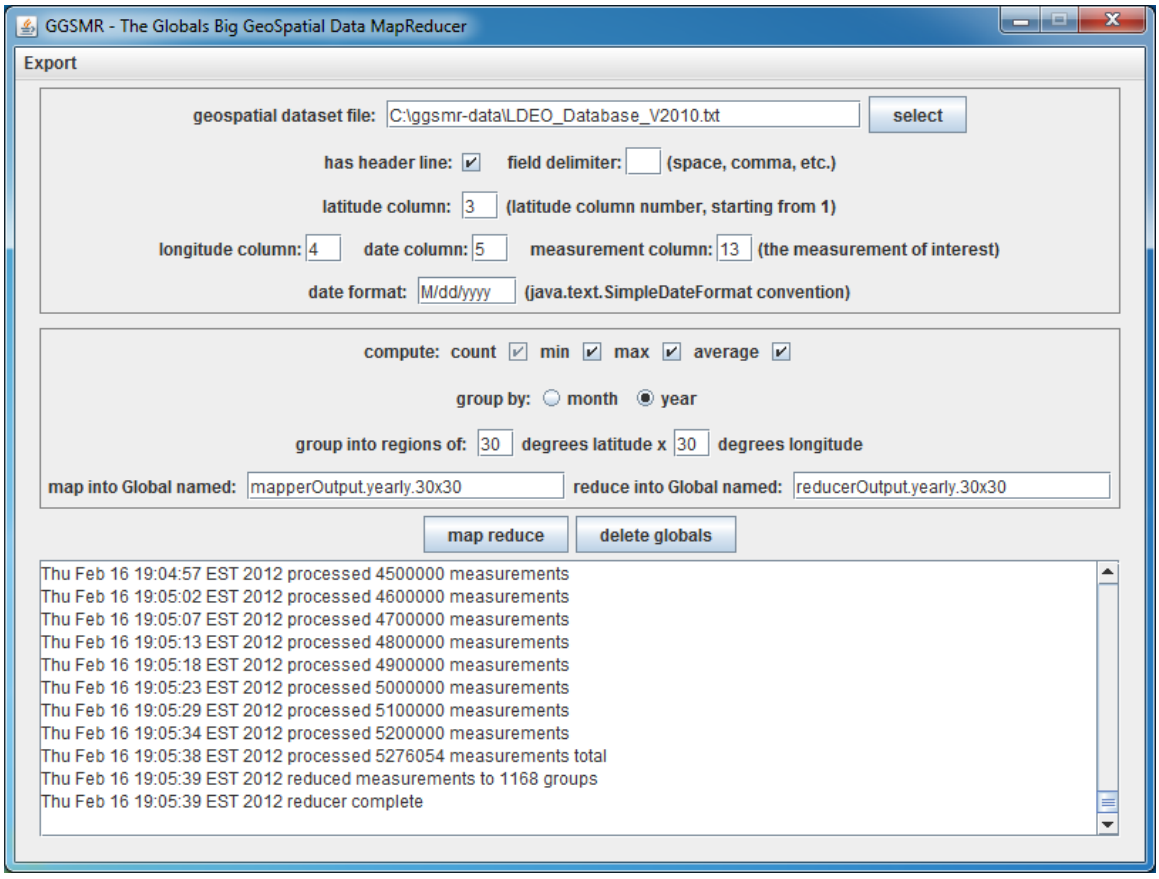
- has header line: *true*
- field delimiter: *space*
- latitude column: *3*
- longitude column: *4*
- date column: *5*
- measurement column: *13* (column PCO2_TEQ in the dataset is the 13th)
- date format: *M/dd/yyyy*

For an arbitrary geospatial dataset in another format, simply enter different values for these inputs.

You can see that by default the computations selected to be performed are: *count*, *min*, *max*, and *average*. And by default measurements will be grouped by *year* and into regions of *30* degrees latitude x *30* degrees longitude.

To perform the big data analysis, first download the dataset from http://cdiac.ornl.gov/ftp/oceans/LDEO_Database/Version_2010/LDEO_Database_V2010.txt.tar.gz and unzip it. This gives you a file named LDEO_Database_V2010.txt with size 543 megabytes.

Then in the GGSMR program, select the dataset file from where you unzipped it, and click the *map reduce* button. GGSMR will run a MapReduce job on the dataset, consisting of a Mapper run followed by a Reducer run. The Mapper maps each line to a *"key"* (a series of global node subscripts identifying a 30x30 degree region of the Earth and a year) and the CO2 measurement as a value. Note that there will be many values associated with each key (i.e., duplicate keys are allowed in this map). The Reducer groups those multiple values for each key (region + year combination) and performs the computations requested. The outputs of both Mapper and Reducer are stored in the globals you specified in the UI (*mapperOutput.yearly.30x30* and *reducerOutput.yearly.30x30* by default). During the MapReduce process GGSMR outputs time-stamped status to the text area, indicating lines processed (every 100,000 lines) until all 5,276,054 lines have been mapped and reduced (see screenshot below).



To access the results of the MapReduce job, select Export from the GGSMR menu and specify a file to save either Mapper or Reducer Output in CSV format. This is the *"gold"* which has been extracted from your dataset using Globals. A copy of the Reducer output is available at:

http://globals-for-the-environment.herokuapp.com/OceanCO2RegionYearlyStats-30.csv
Be aware that the Mapper output will be large and unlike the Reducer output, you would typically not export it except for verification of the MapReduce algorithm.

The next step in our big data analysis of global ocean CO2 trends is to group by 90x90 degree regions, so simply enter 90 in the UI everywhere 30 occurs. Make sure to change the names of the output globals, for example *mapperOutput.yearly.90x90* and *reducerOutput.yearly.90x90*). Otherwise the existing globals will be overwritten (GGSMR kills the globals before running the MapReduce job, although you can do this on demand using the *delete globals* button).

Run the MapReduce for 90x90 and export the Reducer output, a copy of which is available at:
http://globals-for-the-environment.herokuapp.com/OceanCO2RegionYearlyStats-90.csv

Opening the 30x30 CSV file, you see the dataset has been reduced to 1,168 yearly 30x30 degree region groupings as follows (heading column and spacing added here for clarity)

LatUp	LatDown	LonLeft	LonRight	Year	Samples	Min	Max	Average
-60,	-90,	-180,	-150,	1961,	11,	285.1,	306.1,	297.23636363636365

i.e., In the region bounded by latitudes -60° and -90° and longitudes -180° and -150°, in 1961, there were 11 samples taken, minimum value 285.1, maximum value 306.1, average value 297.23636363636365.

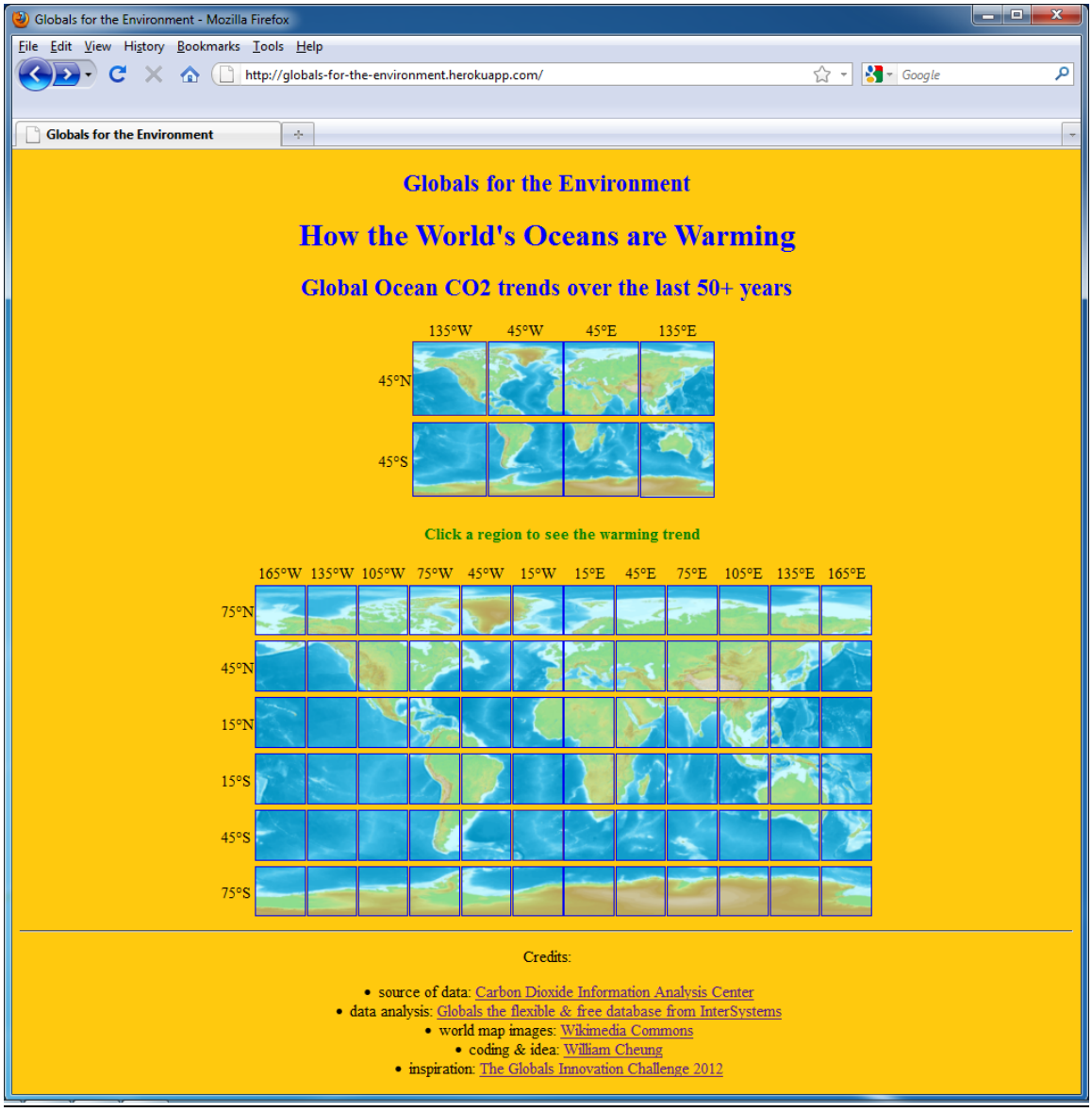
The representation of this reduced group in Globals is a leaf node with subscripts of [-60,-90,-180,-150,1961] and value of list: 11,285.1,306.1,297.23636363636365

The order of the computed values in the list corresponds to the order shown in the UI, from left to right. Also note that in the UI you can choose which values you want computed, although for our analysis we have chosen all.

The visualization

To visualize the output of GGSMR the Globals Big Geospatial Data MapReducer, the *Globals for the Environment* web app was developed, accessible at http://globals-for-the-environment.herokuapp.com and with UI shown below.

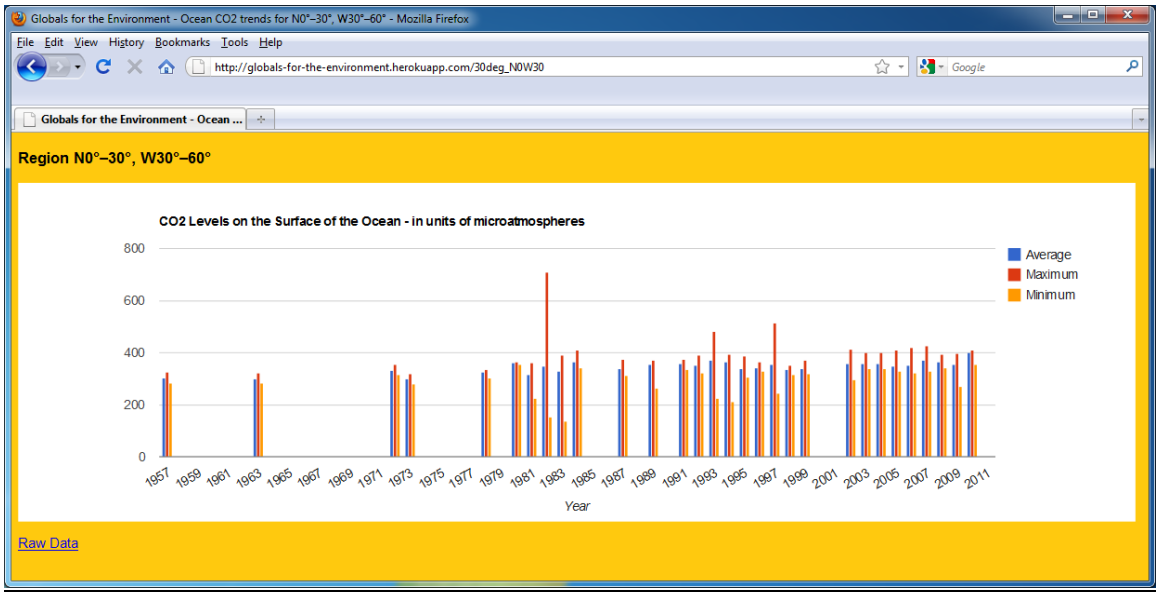
Home page:



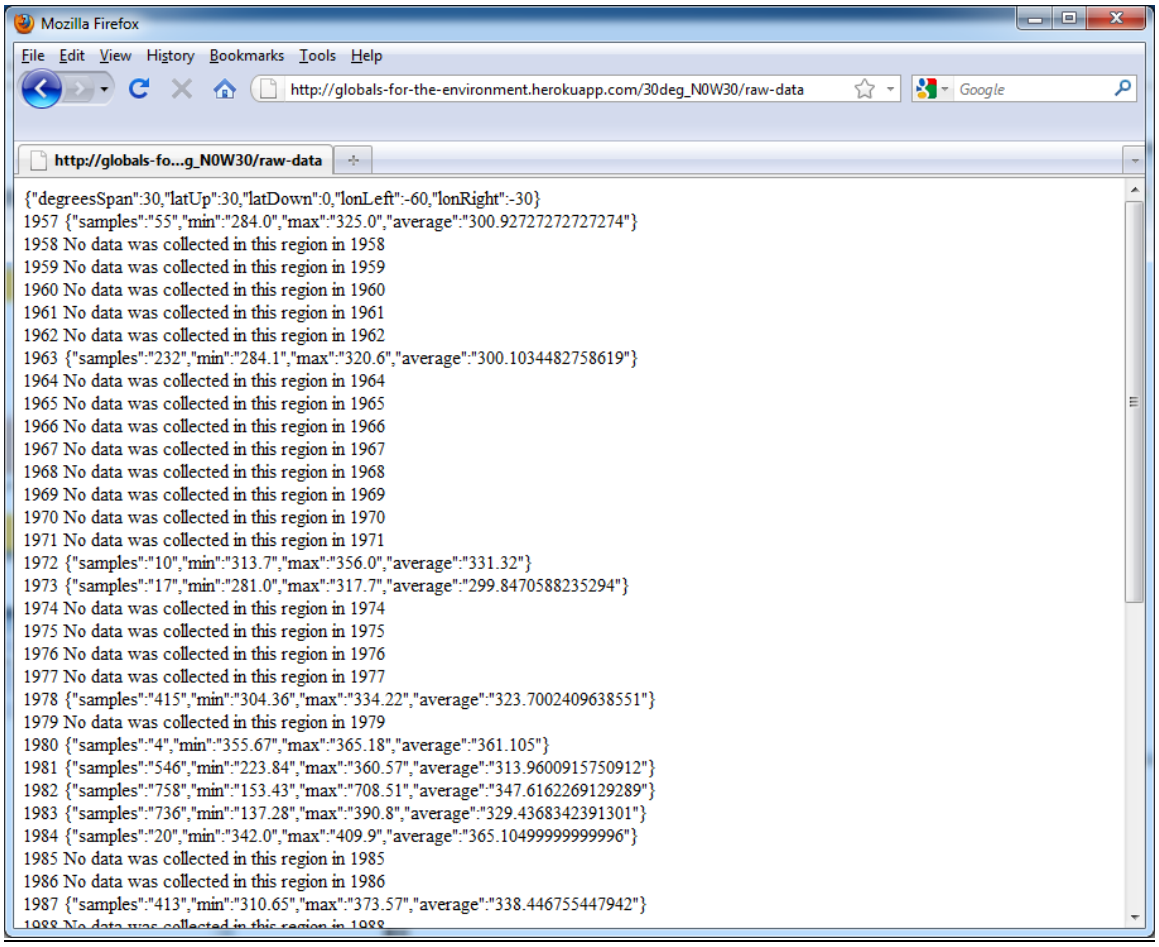
Note the upper map divides the Earth into 90x90 degree regions, while the lower map uses more granular 30x30 degree regions. You can select any region to drilldown on.

You will find the two output files exported from GGSMR (*OceanCO2RegionYearlyStats-30.csv* and *OceanCO2RegionYearlyStats-90.csv*) stored in the app's **public** directory. The app reads the raw data from these two files to generate the UI below on demand.

CO2 graph on map drilldown:



Raw data:



Conclusion

This *Globals for the Environment* project demonstrates how the Globals Database with its flexibility can be easily used to build a Big Geospatial Data MapReducer such as GGSMR. Tools such as this based on InterSystems free technology can be readily applied to big data analytic problems such as identifying long-term trends in carbon dioxide levels on the global oceans. The results of such analysis is the gold which governments, academia, and environmental groups seek to perform meaningful research to benefit the planet.