COL/NW/05

Network-enabled data-based query tool

FY 2003 Proposal to the NOAA HPCC Program

August 19, 2003

| Title Page | Proposed Project | Budget Page |

Principal Investigator:	**James E. Overland**
Line Organization:	OAR
Routing Code:	R/PMEL
Address:	NOAA/PMEL 7600 Sand Point Way NE Seattle, WA 98115
Phone:	(206) 526-6795
Fax:	(206) 526-6845
E-mail Address:	James.E.Overland@noaa.gov

Willa H. Zhu
Willa.Zhu@noaa.gov

Donald W. Denbo
Donald.W.Denbo@noaa.gov

Proposal Theme:	**Technologies for Collaboration, Visualization, or Analysis – Enabling applications**
Funding Summary:	FY 2003 $ 89,800

James E. Overland Oceanographer PMEL	Steve Hammond OERD Division Leader PMEL	Cynthia L. Loitsch Program Support Officer PMEL	Eddie Bernard Director PMEL

Network-enabled data-based query tool

Proposal for FY 2003 HPCC Funding

Prepared by: James E. Overland

Executive Summary:

We propose the development of an NGI data-based network query tool that will assist the scientist who wishes to base a network query on characteristics of the data itself. Examples of such a query might be a request for temperatures exceeding a specific value, or values along a bathymetry contour. The ability of a scientist to productively work with very large datasets is presently limited because of the difficulties related to the network transfer of significant subsets of the data and the subsequent search through the dataset for specific observations or data characteristics of interest. This difficulty is compounded when the scientist needs to inter-relate these results with the extensive climatologic data sets that have been created and are now available directly from the network. We propose to develop a network-enabled data-based query tool that that load pertinent subsets from geographically separated and multi-disciplinary datasets into a temporary, on-the-fly relational database, optionally perform local calculations, and then allow a scientist to construct sophisticated SQL queries.

The proposed tool greatly expands the ability of a scientist to effectively work with the diverse datasets which are increasingly available on-line. Programs with which we are affiliated and which would benefit from this functionality include Fisheries Oceanography Combined Investigations (FOCI), NOAA's Study of Environmental Arctic Change (SEARCH) program, DODS/OPeNDAP and the NOAA Operational Model Archive and Distribution System (NOMADS). Broad based environmental retrospective studies anticipated for FOCI and SEARCH require such a tool. We will utilize our research funding ($100K) to support the development of the data-specific search algorithms, and are asking HPCC to support only the networking and infrastructure part of the application. We will make the tool freely available for use throughout the community and will present the tool in venues such as Technical Working Group meetings and professional society meetings.

Problem Statement:

Scientists routinely work with datasets that contain enormous numbers of individual observations. Although several projects are addressing the issues involved utilizing metadata for locating and navigating these datasets, little has been done to help the scientist who wishes to base a network query on characteristics of the data itself (for example, temperatures exceeding a specific value, or values along a bathymetry contour). The ability of a scientist to productively work with very large datasets is limited because of the difficulties related to the network transfer of significant subsets of the data and the subsequent search through the dataset for specific observations or data characteristics of interest. This difficulty is compounded when the scientist

need to inter-relate these results with the extensive climatologic data sets that have been created and are now available directly from the network. If, for example, the scientist is interested in those observations where the mixed layer depth is greater than 50 meters and warmer than 20 C, there are no data-based network query tools, and so the datasets must be local in order to compute the mixed layer depth and test the mixed layer temperature.

As cross-disciplinary datasets have become increasingly available on the network, it is now possible for a scientist ask data-based questions that require several different and geographically separated datasets be accessed and have subsets transferred to a local machine for processing, for example, looking at the relationship between atmospheric conditions and the ocean state. A network-enabled data-based query tool that could load pertinent subsets from geographically separated and multi-disciplinary datasets into a temporary, on-the-fly relational database, optionally perform local calculations, and then allow a scientist to construct sophisticated SQL queries, would greatly expand the ability of a scientist to effectively work with these datasets.

This proposal is more than a generic exploration of a new technology. The absence of such a tool limits our own research in FOCI and SEARCH, and we are well aware that this is a long-standing problem in the research community and that the scope of applications for the proposed tool is wide.

Relationship to NOAA HPCC objectives: The NGI data-based network query tool that we propose directly supports the HPCC objective in the Collaboration, Visualization and Analysis Theme to develop "*modern network-based applications that demonstrate new techniques for working with NOAA data and information*". This is a major, forward-looking effort that addresses a need that has been expressed in the science community for some time, but for which no solution has yet been developed. Therefore, our proposed implementation "*pushes the envelope of what is currently available*". It is also extensible and scalable for use across NOAA.

Proposed Solution:

Synopsis: We propose to build a Java application that will enable a scientist to issue data-based network queries to explore the relationships between data features in observations and climatology from geographically distributed datasets and multiple disciplines. This network query application will enable a scientist to specify subsets by variable, geographic region, and time, and load them into a temporary, on-the-fly relational database. Once the specific subsets are loaded then SQL queries can be made on both the metadata and data. Results would then be saved into a local file. Initially, the application will be able to access Climate Data Portal and OPeNDAP (formerly DODS) servers and use a temporary, on-the-fly, JDBC compliant embedded database. For example, a researcher would be able to select from a list of servers the datasets he/she wished to include in a query. Queries could be of the form "get me all the data where Temperature>12 for Salinity<24".

Implementation: The Network Query application will be developed in Java and draw heavily on the code and libraries produced for the Climate Data Portal (FY01 HPCC) and ncBrowse (FY02 HPCC) projects. The application will be able to connect to both in-situ datasets (via Climate Data Portal) and gridded archives (via OPeNDAP servers) using a SQL-like command

language or a flexible and powerful GUI interface. Requests will be created that will search these datasets for data that satisfy the constraints.

The capabilities of both the Climate Data Portal and OPeNDAP servers will be used to maximize the efficiency of creating and downloading subsets of remote datasets. Observational data downloads will be restricted to the space and time range of interest and use any other selection criteria available. Gridded climatologies will be subset by the server, significantly reducing the number of grid points downloaded. In both cases only the variables of interest will be transferred. Our testbed for this tool will utilize the Climate Data Portal, ncBrowse, and OPeNDAP datasets. We will focus on testbed applications involving Arctic, North Pacific, and El Nino datasets and incorporate the feedback we receive from scientists. The data archived at the National Snow and Ice Data Center and the Scott Polar Research Institute are also candidate data sources.

A freeware, for non-profit applications, embedded Java Database like HypersonicDB (hsqldb) will be used to implement the local database. HypersonicDB supports much of the SQL standard and uses JDBC database access. Support of JDBC and the SQL standard is critical to minimize the Network Query applications dependence on a specific database product.

Other features of the Network Query application include:

> Computation of additional variables from those downloaded, for example, density, potential temperature, and pseudo stress.
> Variable name translations via a dictionary. Since each database (or even individual files at a server location) may not have a common naming scheme it will be necessary to produce translations for the variables, for example, T -> Temperature, u_surf -> UVelocity, S -> Salinity.
> Unit conversion. Unit conversions may also be necessary, for example, 10 cm/sec -> 0.1 m/sec, to enable comparisons.
> The name translation dictionary for both variables and units will be managed by the Network Query application, allowing a user to add, modify, and delete entries.
> Queries can be made against the local temporary relational database with SQL statements. A GUI interface will be available in a future version.

While this could be a memory intensive application, memory and processor speed of desktop computers is getting greater all the time. 1Gb of memory for a 2 GHz computer is not out of the question for the average scientist who may want to use this tool.

Leveraging: This leverages off the recent development made in creating on-line network enabled datasets. It will use OpenSource software to keep acquisition costs low. The Climate Data Portal and ncBrowse (both are previous HPCC projects) would be used to access CDP servers and OPeNDAP servers, respectively.

Matching Funds: Our SEARCH and FOCI funding ($100K) will support the research side of this effort, which consists of developing and implementing the required data searching

algorithms. We are asking HPCC to support only the development of the required networking infrastructure application.

Cost/benefit for this effort is superb when one considers the magnitude and importance of the problem being addressed, the science support and matching funds devoted to the effort, and the technical excellence of the programming team at PMEL.

Scope: Programs with which we are affiliated and which would benefit from this functionality include Fisheries Oceanography Combined Investigations (FOCI), NOAA's Study of Environmental Arctic Change (SEARCH) program, DODS/OPeNDAP and the NOAA Operational Model Archive and Distribution System (NOMADS).

Analysis:

We chose this solution because it leverages from existing software and can access the majority of geophysical data that is available over the network. Commercial applications, at present, do not have the capability to connect to these geophysical data servers.

While commercial solutions do exist for the database components of the project, free alternatives offer sufficient performance for the initial implementation.

Performance Measures:

Milestones

Month 1 – Collect requirements for the Network Query application.
Month 3 – Design Network Query application
Month 5 – Initial development of application
Month 6 – User feedback
Month 7 – Modify design based on feedback.
Month 12 – Final release of the Network Query application.
Month 12 – Present results at NOAATech

Deliverables

Network Query Application. The application will subset remote datasets into a local temporary relational database and perform SQL based queries on the metadata and data.