TTR/NE/05

Using Network-enabled Query Tool at NODC

FY 2005 Proposal to the NOAA HPCC Program

September 13, 2004

| Title Page | Proposed Project | Budget Page |

Principal Investigator: **L. Charles Sun**

Line Organization: NESDIS
Routing Code: E/OC1
Address:

NODC
Bldg: SSMC3 Rm: 4752
1315 East West Hwy
Silver Spring, MD 20910-33282

Phone: (301) 713-3272 ext. 111
Fax: (301) 713-3302
E-mail Address: Charles.Sun@noaa.gov

James E. Overland James.E.Overland@noaa.gov
John Osborne John.Osborne@noaa.gov
Norman Hall Norman.Hall@noaa.gov

Proposal Theme: **Technology Transfer**

Funding Summary:

L. Charles Sun	Kurt J. Schnebele
Oceanographer	Acting Director
NOAA/NESDIS/NODC	NOAA/NESDIS/NODC

Using Network-enabled Query Tool at NODC

Proposal for FY 2005 HPCC Funding

Prepared by: L. Charles Sun and John Osborne
In consultation with James E. Overland

Executive Summary:

The goal of this Technology Transfer proposal is to package and enhance the HPCC-developed NQuery[1] for implementation at the NODC. NQuery is a program to allow queries based upon the actual measured parameters in a set of data files. At NODC, NQuery will enhance and expand the existing data selection and sub-setting capabilities for the Argo and Global Temperature-Salinity Profile Program data by scientists, ocean applications and operational programs. This proposal is independent of any other proposed FY05 effort.

Given the popularity of these important datasets to users of the NODC data distribution (over 98,000 requests from about 2,000 distinct hosts in August 2004), the scope of this proposal will be very far-reaching.

Problem Statement:

Background: The National Oceanographic Data Center (NODC) operates the Global Argo Data Repository for Argo and assembles both real-time and delayed-mode GTSPP data into a continuously managed database. The GTSPP database contains 2,300,000 individual station files. Additional files are generated and loaded into the database weekly. The GTSPP individual files are sorted and distributed by calendar quarters and ocean areas. The NODC receives 11,526 requests/month for GTSPP data from 1457 unique hosts (monthly averages over the past sixteen months). The Argo database consists of about 100,000 individual station files. Additional files are generated and loaded into the database daily. Monthly Argo data requests increased from 287 in March of 2003 to 97,560 in July 2004 and the number of distinct hosts served increased dramatically from 35 to 620 during the same period.

Since many large data collections are becoming available on the network through OPeNDAP (*in-situ* data collections and gridded datasets), it is increasingly difficult for scientists to determine whether each data file will suit their needs, they must sort through each file, wasting time and energy. This difficulty is compounded when the scientist needs to inter-relate these results with the extensive climatological data sets that have been created and are now available directly from the network.

[1] Overland, FY03 HPCC Project titled with *"Network-enabled data-based query tool"* (COL/NW/05)

The Problem: Once a scientist has subsetted a large data collection based on time/space criteria, the remaining data subset may still be extremely large (tens of thousands of station files). There remains the need to further refine the data selection and subsequent subsetting in order to winnow out unsuitable datasets. Scientists have long wished for the capability to select based on the values of the data within the files.

Relationship to NOAA HPCC objectives: This proposal supports the HPCC objective "*to improve technology for access to critical data, information and unique resources*", and the HPCC technology transfer objective "*to extend the utilization of previously funded and successful HPCC projects" by helping "others utilize and enhance existing prototypes*".

Proposed Solution:

The need: This proposal addresses the need of a scientist who has located data based on time and space, and now wishes to refine that data selection based on the value of the data within the files. As an example, a large data collection may contain millions of profiles. The subset falling within the scientist's desired time-space range may contain only tens of thousands of profiles. From those, the scientist may wish, for example, to obtain only those profiles for which temperatures at the surface exceed a certain value

Solution: The time/space subsetting issue has been addressed by the HPCC-supported Argo NdEdit[2], a spatial/temporal filtering tool for pre-selecting data sets of interest from much larger online archives. The final data selection/subsetting step can be accomplished by enhancing the HPCC-supported NQuery[1], implementing it at the NODC, and integrating it with the NODC Argo NdEdit. The modified version of NQuery (called Argo-GTSPP NQuery) will allow a scientist who has selected Argo and GTSPP datasets that meet the desired time-space criteria, to then determine whether each data file will suit their needs, based on the characteristics of the data within each of the data files. The scientist specifies the desired data characteristics, and the NQuery tool returns references to specific profiles that meet those criteria. This proposal integrates HPCC supported projects (see footnotes 1 and 2) by enhancing and expanding other existing and proposed Argo/GTSPP tools. It complements but is completely independent of any other proposed FY05 effort.

Argo-GTSPP integrates the existing HPCC-supported Argo NdEdit, a spatial/temporal filtering tool for pre-selecting data sets of interest from much larger online archives (Sun, Denbo, and Osborne, FY04 HPCC Funded Project titled with *"In-Situ Data Selection/Subsetting Tool for Argo CD-ROM"*). This is how NdEdit and NQuery work together:

- **NdEdit** allows the user to make an initial data selection based on detailed time/space criteria
- **NQuery** allows the user to refine that data selection based on the characteristics of the data.

[2] Sun, Denbo, and Osborne, FY04 HPCC Project titled with *"In-Situ Data Selection/Subsetting Tool for Argo CD-ROM"*

- Together, **NdEdit** and **NQuery** provide a data selection or subsetting capability wished for by scientists for decades.
- This proposal integrates NQuery and the NODC Argo NdEdit data selector that was developed under FY04 HPCC funding.

Specifics of Solution: The mechanism used within the NQuery tool is to load pertinent subsets of user-selected multi-disciplinary datasets into a temporary, on-the-fly relational database, perform local calculations, and then uses the scientists' specifications to construct sophisticated SQL queries to locate subsets of interest. A critical design consideration, with implications for the implementation methodology, was the speed of this application. Selecting calculations only as necessary makes the query process as efficient as possible.

A simple two-step process takes the user from pre-selected data (via Argo NdEdit and/or the scientist's own data collections) to a subset of data files selected from their observed values. The user is unaware of the complexity of the underlying process. First, the user determines what observed and computed variables will be included in the database, and therefore available for subsequent queries. This selection is accomplished via a simple graphical interface that lists all available observed variables in the selected dataset. Additional interface dialogs allow the user to select from a set of built-in summary statistics variables (e.g., average value and depth of maximum value) and "user-defined" calculated variables, such as mixed-layer depth, apparent oxygen utilization, and potential density. A major advantage of NQuery is that many of the user-defined calculations can be customized by the scientist for their specific needs. A single click starts the process of reading the individual files in the pre-selected dataset, extracting the observed values and computing the requested statistical or user-defined variables. An on-the-fly MySQL database is created and populated as each data file is processed.

Second, after the database has been populated with the requested calculated values, the user can create a SQL query using a second graphical user interface. This interface allows a user to build simple to fairly complex queries by using either a point-and-click query builder interface, or entering an SQL query statement by hand. Once a satisfactory query is constructed, it is executed by the database, resulting in a list of files that satisfies the search criteria. The scientist can then use these files in other research tools (e.g., ncBrowse and Java OceanAtlas).

Implementation: The primary task is to integrate Argo NdEdit into NQuery to allow spatial/temporal pre-selection of the Argo and GTSPP archives in NQuery. Argo NdEdit will ingest a special form of the Argo inventory file that contains URL's for individual profiles. After a user has selected a subset of the Argo profiles, NdEdit will then produce an intermediate data object that will be used by NQuery to locate the actual profiles, ingest the data (via the Internet and HTTP) and construct the on-the-fly database. For the much larger GTSPP archives, user will use the data pre-selection pages available from NODC at http://www.nodc.noaa.gov/GTSPP/access_data/inv.htm to download an extended inventory file that can be ingested into Argo NdEdit. Note that NQuery is a desktop application that does not rely on any host resources at NODC except for making the NQuery installer available on a web page for download. What NODC does to make NQuery an operational system is to provide the extended Argo and GTSPP inventory files available for download and to assure that the individual are "reachable" via an HTTP URL.

Specific Implementation Tasks (and who performs them):

1) Provide extended Argo and GTSPP inventory files that include a URL for individual netCDF profiles (NODC)
2) Integrate Argo NdEdit into NQuery user interface to provide spatial/temporal pre-selection of data subsets from the inventory files (PMEL)
3) Enhance Argo NdEdit to provide intermediate data products recognizable by NQuery to present the NQuery variable-selection interface (PMEL)
4) Enhance NQuery to ingest netCDF files via the HTTP protocol (PMEL: based upon existing code developed at PMEL by Donald Denbo)
5) Develop a double-clickable installer for Argo-GTSPP (NODC, PMEL)
6) Provide user documentation for installing Argo-GTSPP and setting up a database server (e.g., mySQL) on their desktop machine. (PMEL, NODC)

Analysis:

Rationale: NQuery greatly expands the ability of a scientist to effectively work with the diverse datasets which are increasingly available on-line. Examples of programs which can 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. It will be most cost-effective for NODC to transfer and implement NQuery without "re-inventing the wheel".

Comparison: We know of no other applications that that provide NQuery's ability to ingest local and distributed data, construct an on-the-fly database of computed values, and allow users to select profiles based upon the measured characteristics of a profile.

Benefits: The successes of this proposal will benefit the HPCC program office, the government and academic marine science communities, the general public community, and NODC in the following ways. The impact can be seen from the large and growing number of Argo and GTSPP users (see Executive Summary and Problem Statement sections).

- Provide an opportunity for HPCC to leverage previously funded projects.
- Provide a seamless network-enabled in-situ data selection /sub-setting tool for the Argo and Global Temperature-Salinity Profile Program data by scientists, ocean applications and operational programs.

This proposal leverages from previous work. It complements, but is completely independent of other proposed FY05 efforts.

Performance Measures:

Milestones

Month 1 - Collect requirements
Month 2 - Develop and test Argo-GTSPP *in-situ* selection tool
Month 3 - Deploy Argo-GTSPP *in-situ* selection tool

Deliverables

- Deliverable 1 Draft system design document
- Deliverable 2 Proof of concept and prototype
- Deliverable 3 Argo-GTSPP NQuery tool
- Deliverable 4 Final project report