Navigating Oceans of Data

David Maier[1], V.M. Megler[1], António M. Baptista[2], Alex Jaramillo[2], Charles Seaton[2], and Paul J. Turner[2]

[1] Computer Science Department, Portland State University
[2] Center for Coastal Margin Observation & Predication, Oregon Health & Science University
{maier,vmegler}@cs.pdx.edu,
{baptista,jaramilloa,cseaton,pturner}@stccmop.org

Abstract. Some science domains have the advantage that the bulk of the data comes from a single source instrument, such as a telescope or particle collider. More commonly, big data implies a big variety of data sources. For example, the Center for Coastal Margin Observation and Prediction (CMOP) has multiple kinds of sensors (salinity, temperature, pH, dissolved oxygen, chlorophyll A & B) on diverse platforms (fixed station, buoy, ship, underwater robot) coming in at different rates over various spatial scales and provided at several quality levels (raw, preliminary, curated). In addition, there are physical samples analyzed in the lab for biochemical and genetic properties, and simulation models for estuaries and near-ocean fluid dynamics and biogeochemical processes. Few people know the entire range of data holdings, much less their structures and how to access them. We present a variety of approaches CMOP has followed to help operational, science and resource managers locate, view and analyze data, including the Data Explorer, Data Near Here, and topical "watch pages." From these examples, and user experiences with them, we draw lessons about supporting users of collaborative "science observatories" and remaining challenges.

Keywords: environmental data, spatial-temporal data management, ocean observatories.

1 Introduction

The growth in the variety and numbers of sensors and instrument platforms for environmental observation shows no signs of abating. In the past, measuring an environmental variable (such as the chlorophyll level in water) might have required collection of a physical sample, followed by laboratory analysis (say on a monthly basis). Now an in-situ sensor can monitor the variable continuously. Laboratory analysis of samples still occurs, but some tests now generate gigabytes of data, such as high-throughput DNA sequencing. Observational and analytic data is itself dwarfed by outputs of simulation models. Oceans of data are upon us.

An equally important trend is a change in how science is performed. Traditionally, for much of ocean science, data collection was on a per-investigation basis, with the same researcher or group analyzing the data as gathered it. That data might be shared

A. Ailamaki and S. Bowers (Eds.): SSDBM 2012, LNCS 7338, pp. 1–19, 2012.

with other scientists, but typically months or years after initial collection. Answering questions about human effects on the environment, or influences of climate change, require data collection on spatial and temporal scales beyond the abilities of any individual or small group. Thus we are seeing shared environmental observatories (much like in astronomy [15]), such as that operated by the NSF-sponsored Center for Coastal Observation and Prediction (CMOP, www.stccmop.org). In such observatories, those planning and carrying out the data collection are often different from those using the data, and the common pool of data supports a "collaboratory" in which scientists from disparate disciplines work together on complex environmental questions. This shift in the nature of the scientific enterprise presents challenges for data dissemination and analysis. It is no longer reasonable to expect an individual scientist to have comprehensive knowledge of the complete type and extent of data holdings in an observatory such as CMOP's. Moreover, with an expanded base of users, providing display and analysis tools for each type of user separately would be challenging. Thus it is important to have a common base of capabilities that can help investigators locate and judge datasets relevant to their work, as well as carry out initial graphing and analysis tasks on line, without having to download and work locally with that data (though that mode of interaction must also be supported).

The cyber-infrastructure team of CMOP is charged with managing the storage and dissemination of data assets associated with observation and modeling activities, as well as producing web-based interfaces for navigating, accessing and analyzing those assets. We begin by surveying the main user groups that CMOP supports (Section 2), then briefly describe the data-collection and management process (Section 3), along with several of the tools that support these groups (Section 4). We touch on some of the techniques we are investigating to meet the performance demands of one of these tools, Data Near Here (Section 5) and recount some of our lessons learned (Section 6). Section 7 concludes by laying out current issues and challenges that could be the basis for future research on scientific data management.

2 The User Base

There are broadly three classes of users of CMOP systems.

Operations: This class consists of internal users with responsibility for the day-to-day operation of the CMOP observatory, from sensors and telemetry to data ingest to quality assurance to data download and display services. The needs of this class concern detecting problems in the data chain, such as fouled or failed sensors, records corrupted in transmission, failed loads and inoperative interfaces. In many cases, such problems are currently exposed via the tools used by "regular" users, such as the pages that display recent observations from sensor stations. However, this time-intensive approach can lead to delays in problem detection and in data quality assurance. Sometimes specialized interfaces are needed for status reporting, station viewing and quality-assessment tasks. In addition, recording information on collection of water and other samples, and the results of laboratory analyses, is also needed to support observatory operations.

Science: This class consists of researchers internal and external to CMOP. The data holdings of CMOP are becoming extensive enough that few scientists are aware of their totality, in terms of time, location and type of observation. Even when someone might know where a sensor station is located, and when it first became operational, he or she might be unaware for exactly what times data is available – some instruments are deployed only seasonally, some may be removed temporarily for repairs, some segments of data might be dropped during quality assurance. Thus, there need to be tools to help a scientist find data that is potentially relevant to his or her research question, and also to get a quick view of temporal coverage of a specific observation station. Once a dataset of possible interest has been identified, a scientist often wants a simple plot of it, to assess its suitability. She might be checking if there are dropouts during the period of particular interest, or if it contains some event she is seeking, say, unusually low dissolved-oxygen levels. Once a dataset is deemed useful, she might want to download all or part of it in a form suitable for use in a desktop tool, such as Excel or Matlab. However, there should be some capability to analyze the data online, such as charting several variables on the same graph, or plotting one variable against another. Finally, scientists want to comment on or annotate data or products of analysis, to point out suspected problems or to highlight interesting subsets.

While this paper focuses on observed data, there are places where observed data and simulated data intersect. One is in comparing observed and modeled behavior of an environmental variable, such as salinity. To judge model *skill* (a model's ability to reproduce real-world physical phenomena), it is useful to plot observed and model data together. Since the model data has much denser coverage than observed data, it must be sub-sampled to a dataset that matches the location and times of a corresponding sensor dataset. Such sub-sampling is essentially a "virtual sensor" operating in the simulated environment at a place and time that matches the corresponding physical sensor in the real environment. A second interaction between observed and modeled data is when the latter provides a context for sensed and sampled observations. To this end, *climatologies* are useful for comparing current conditions to historical trends. A climatology is an aggregation of a particular variable, generally over both time and space. Examples are the monthly average of maximum daily *plume* volume (the portion of the ocean at a river mouth with reduced salinity), and the average weekly temperature of the estuary. A scientist can then see, for example, if water samples were taken when temperatures were relatively high for the time of year.

Education: An important subset of science use is educational use. For students pursuing undergraduate or graduate research, the needs for data access and analysis largely match those of scientific staff. For classrooms and science camps, the user base is quite different in motivation and sophistication. Currently, we do not have interactive tools specifically for K-12 use. However, this class of users is considered in the design of interactive tools, particularly in choosing default settings that are likely to yield viewable results on initial encounter.

Resource Management: There is a growing class of users who use observatory data in reaching decisions, both in day-to-day resource management as well as for longer-range policy making. For example, the Quinault Indian Nation is highly interested in the timing and spatial extent of hypoxic (low-oxygen) conditions near their tribal

lands, to understand the possible effects on the shellfish harvest. A second example is a manager at a fish hatchery deciding when to release juvenile fish to the estuary. Research points to a correlation between estuary conditions (properties of the freshwater plume extending into the ocean from the river's mouth) with survivability of hatchlings [3]. Comparing predicted conditions for the coming week against the typical range of conditions at the same time in past years might help optimize the release time. In general, for such users, it helps to organize data thematically, bringing together data from a range of sources related to a theme (such as hypoxia) on a single web page, preferably with accompanying commentary that highlights important current trends or conditions. Such thematic pages are also useful to scientists studying a particular phenomenon or condition. For example, the Columbia River often exhibits red-water blooms in the late summer. It is useful to collect information that indicates the onset of such events so that, for example, additional sampling can take place.

Our Goal: The data and environment we support are complex (and our resources are bounded); the users have a wide spectrum of skill sets (K-12 students, resource managers, ocean scientists); and we have a huge range of scales of analysis and processing that we must deal with (models of the entire coastal shelf versus RNA in one water sample; decades of data versus phenomena that manifest in a few seconds). Further, a given line of research can require different levels of detail at different stages. We do not want to require people to learn (nor expend the resources to build) different, specialized tools for each of these combinations (which has often been the norm in the past). We also do not want to enforce simplicity by dictating a single workflow or by limiting the user to only one set of data or analysis. Thus our goal is to find simple, consistent abstractions that expose the complexity in the data (which is relevant to scientists) while hiding the complexity in the infrastructure (generally not of concern).

3 The CMOP Observatory

CMOP is funded by the National Science Foundation's Science and Technology Centers program, along with matching contributions from center participants. It studies conditions and processes in the estuary, plume (the jet of fresher water that protrudes from a river's mouth into the ocean) and near-ocean systems, trying, in particular, to anticipate and detect the influences of human activity and climate change. A major component of CMOP's common infrastructure is an environmental observatory focused on the Columbia River Estuary, but also extending up river as well as to the near ocean off the Oregon and Washington coasts. (See Figure 1.) The observatory collects measurements of environmental variables (henceforth just *variables*) via sensors for physical (temperature, salinity), geochemical (turbidity, nitrate) and biological (chlorophyll, phycoerythrin) quantities [12]. These sensors are mounted on fixed (pier, buoy), profiling (moving through the water column) and mobile platforms. The mobile platforms include staffed research vessels as well as autonomous vehicles. The readings from many of the sensors are immediately relayed back, via wired and wireless links, to CMOP servers. However, some information, particularly from mobile

platforms, is downloaded in bulk, for example, at the end of a cruise or mission. Sensor readings are supplemented with laboratory tests of water and other samples for chemical and biological properties, including RNA and DNA assays. (However, technology is developing to allow in-situ performance of some of these tests [6].) While most sensors deliver a few floating-point numbers per reading, others can produce vectors of values (e.g., density profiles) or 2-D images (for example, of surface waves or micro-organisms [8]). Observation frequencies can be as often as every few microseconds, or as few as tens per year for DNA assays.

In terms of volume and growth, CMOP collected about 75K observations of physical variables in 1996 from fixed stations. A decade later, the rate was about 10M observations per year, and rising to 42M observations in 2011. Collection of biogeochemical variables began in 2008, with 38M observations collected in 2011. In 2002, total observations from mobile platforms was just 5K. Since then, it has been as high as 17M observations (2008), though it dropped off last year because of fewer cruises.

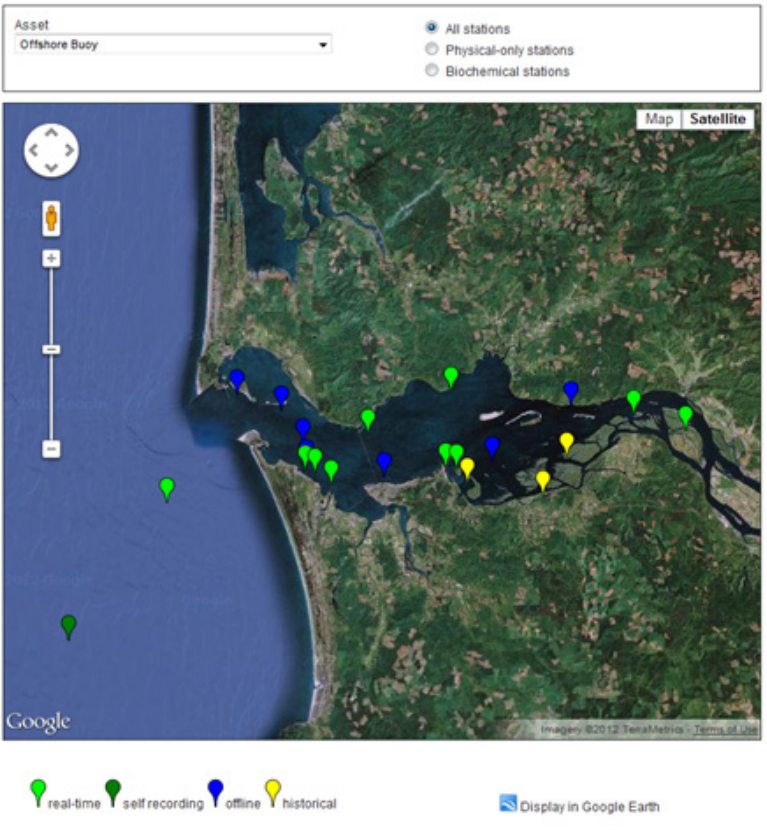


Fig. 1. An overview of the CMOP observation network, including both current and past positions of sensor stations. This map also serves as an interface for navigating to the information pages for specific stations.

While observational data is our main focus here, another major component of the shared CMOP infrastructure is a modeling capability for the Columbia estuary and other coastal systems. These simulation codes are used both to prepare near-term (days) forecasts of future conditions as well as long-term (decades) retrospective runs, called *hindcasts* [2]. Historically, these models have addressed the 4-D hydrodynamics of the river-ocean system, including velocity, temperature, salinity and elevation over time. More recently, the models are being extended to include geochemical and biological aspects. Currently, each forecast run produces almost 20GB of data, mostly as time series of values on a 3-D irregular mesh, but also including pre-generated images and animations. The hindcast databases are approaching 20TB of data.

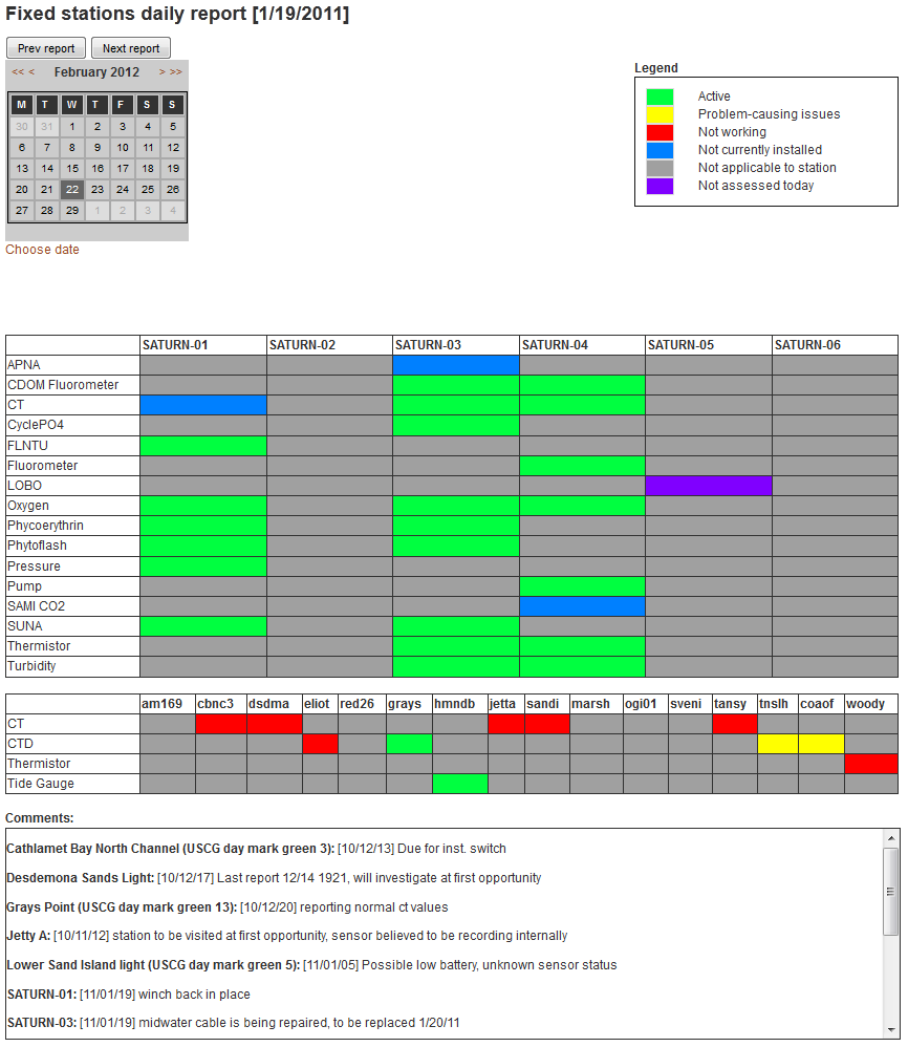


Fig. 2. The daily status page for fixed observatory stations. It indicates deployment and operational status for various instrument types at various CMOP stations.

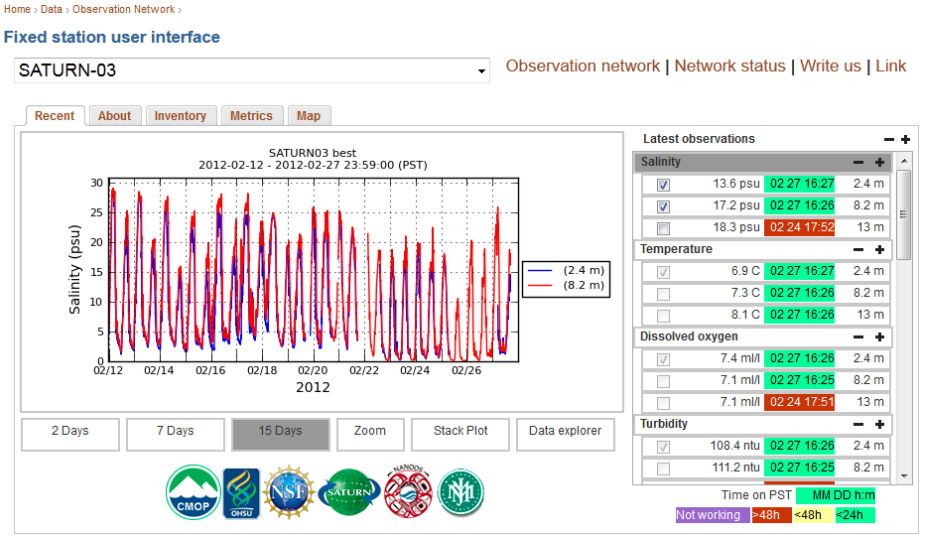


Fig. 3. The station page for SATURN03, plotting salinity offerings for two different depths

CMOP makes as much data as it can available online as soon as possible. The preponderance of the sensor data is routed into a relational DBMS. Ingest processes work directly with network feeds or through frequently polled remote files to get sensor records, which are parsed and inserted into database tables. The unit of designation for time-series observational data is the *offering*, which generally refers to the data for a particular environmental variable coming from a specific instrument at a particular position (often given as a station name, e.g., SATURN04, and a depth, e.g., 8.2 meters). There are also offerings from mobile platforms, where position is itself captured as a time series. A physical dataset can give rise to multiple offerings: a raw stream, as well as one or more corresponding streams that are the output of quality assurance and calibration procedures. High-frequency (multiple readings per second) data can be "binned" down to a coarser time step (1 minute, 5 minutes), and registered to a common time scale. Offerings also exist for derived variables (for example, conductivity and temperature used to compute salinity) and "virtual" observations from the simulation models. A few offerings provide monitoring information about the observatory infrastructure, such as the status of pumps, for use by operations staff.

4 CMOP Interfaces and Tools

While observed data are often available on CMOP database servers within minutes (if not seconds), they have little value if they are not easily accessible to CMOP scientists and other users. CMOP endorses the vision of a "collaboratory" where there is open sharing of data, and scientists of multiple disciplines can easily interact with each other and CMOP information resources. Often the easiest way for a scientist to get an initial impression of data is through a plot or graph. Thus, a key strategy is

making plot production a basic service in the cyber-infrastructure. The CMOP *offeringPlot* service is available via a RESTful API, where a URL details both the offerings of interest and the plot parameters (kind of plot, extent of axes, aspect ratio, etc.). For example, the URL

```
http://amb6400a.stccmop.org/ws/product/offeringplot.py
?handlegaps=true&series=time,saturn03.240.A.CT.salt.PD0
&series=time,saturn03.820.A.CT.salt.PD0&series=time,
saturn03.1300.R.CT.salt.PD0&width=8.54&height=2.92
&days_back=2&endtime=2012-03-09
```

produces a scatter plot of two salinity offerings at the SATURN03 station versus time. The plot will be generated at a particular width and height, and will cover data going back two days from 9 March 2012.

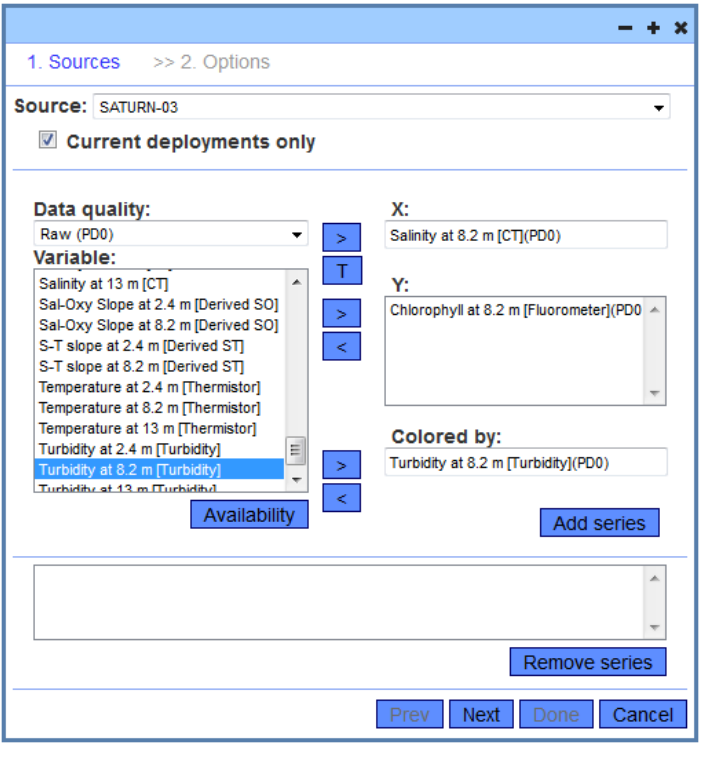


Fig. 4. The first configuration screen for Data Explorer, where the offerings to be plotted are selected

Plotting as a service is used heavily by CMOP interfaces, but supporting interactive response times was a bit of a challenge. In theory, the plotting service could get its data directly from the RDBMS. However, our experience was that direct access often resulted in significant latency, likely influenced by the fairly constant load of ingest tasks. Also, users are often interested in the most recent data from a station, so

as use increases, redundant access to the same data is likely. Thus, we moved to an information architecture where we maintain a cache of extracts from the database, and, in fact, pre-populate the cache. The cache consists of about 36GB of files in netCDF format [13], arranged in a directory structure with a file for each offering for each month. (We may switch to individual days in the future, to avoid regenerating files for the current month and to support extra detail in some of our tools.) While an interface can access the database directly—and some do—the netCDF caches satisfies much of the read load. As a side benefit, the cache also supports programmatic data download, via a THREDDS [5] server using the OpenDAP protocol [4].

We now turn to some of our existing interfaces, plus one under development. Figure 2 shows the observatory *status page*, as used for fixed and profiling stations. It reflects daily reports by field staff on the dispositions of various instruments installed at CMOP observation stations. Operations users record and report overall status, which CMOP management monitors via this page.

Operations staff and resource managers coordinate many of their day-to-day activities using the *station pages*, which are also a starting point for researchers with specific instruments supporting their studies. These pages provide immediate display of data transmitted from instruments. Figure 3 shows an example of a station page, for the station named SATURN03. Shown is a 15-day plot of the salinity offerings at the station for 2.4 meters and 8.2 meters. Such pages are designed to be easy to interact with, so present a limited set of choices for configuration. Along the right side are different offerings associated with the station, such as temperature at 2.4 meters and turbidity (cloudiness) at 8.2 meters, grouped by variable type; a user can quickly display other instruments' data using the checkboxes. For each offering, the time and value of the latest available measurement is shown. The colors over the time indicate if new data has appeared in the past day (green), last two days (yellow) or longer ago (red). Along the bottom are time periods. It is also possible to obtain all offerings in a single page of plots (a *stack plot*). The station pages are intentionally limited in their capabilities, to keep them simple to work with and give fast response times.

The simple plots of the station pages are limited in many ways: no arbitrary time periods, no combination of offerings for different variables or different stations in one plot, only plots of variables against time (as opposed to plotting one against another). The Data Explorer, accessible directly from this page, is a more sophisticated tool, that allows control of these aspects (and more), but with a more complicated interface. It supports both variable-time and variable-variable plots, optionally coloring the plot by an additional variable. The configuration process in Data Explorer involves sequencing through several set-up screens. The first screen (shown in Figure 4) is used to select the offerings to include in the plot. Here Chlorophyll and Salinity from station Saturn03 at 8.2m are selected for a scatter plot, to be colored by Turbidity at the same depth, perhaps to contrast the influences of the river and the ocean on the estuary.

Additional screens allow selection of a time period, axes limits, and aspect ratio of the plot. Figure 5 shows the requested plot. The Data Explorer supports saving and annotating plots, as well as downloading the underlying data. This powerful tool is used by researchers for everything from exploratory research to producing diagrams for publication. Operations staff use it as well, to identify the onset of instrument malfunctions, for further annotation and analysis during quality-control processing.

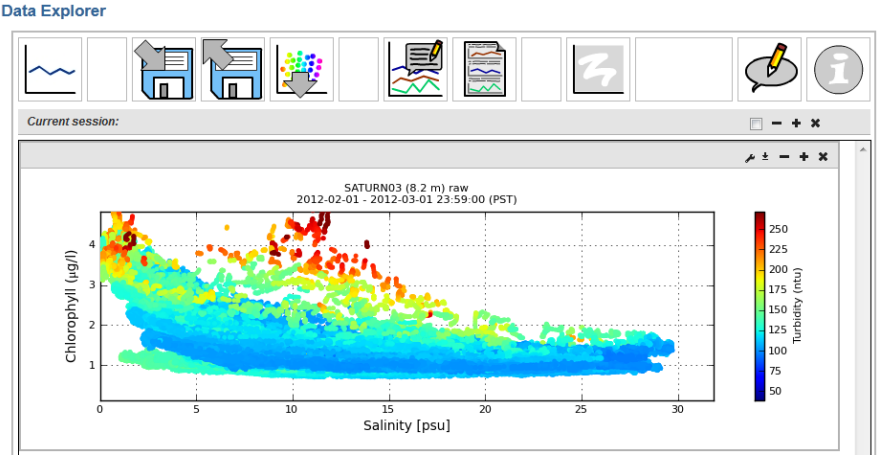


Fig. 5. The resulting plot from Data Explorer

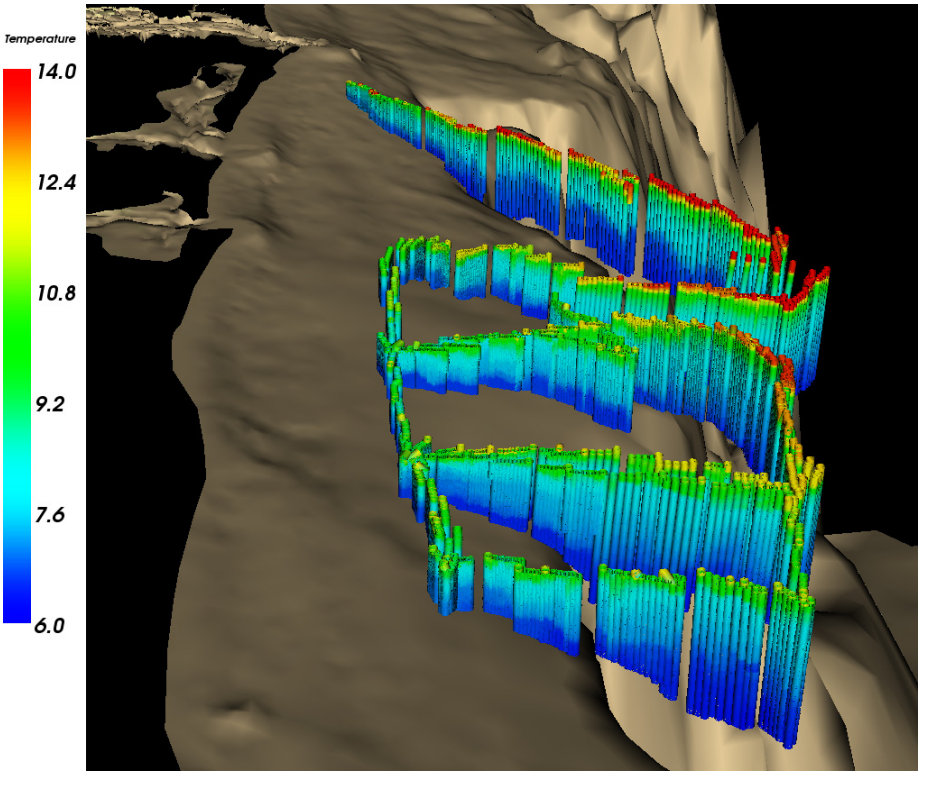


Fig. 6. Specialized plots for a glider mission, showing the trajectory for the glider, colored by temperature in this case, superimposed on the sea-bottom topography

The design of the plot-specification interface for Data Explorer has been challenging. One issue is avoiding creating plots where there is no data, usually due to selecting a time period before a sensor was deployed or during an outage, or choosing a data-quality level that has not yet been produced. Once an offering is selected, one can see an inventory of data for it (the "Availability" button). However, it might be helpful to default the time selection on the next screen to the most recent period with data.

A related issue is the order in which plot aspects are specified. Currently, the configuration interface aims at a work pattern where a user is first interested in one or more stations, then selects offerings from those stations, followed by choosing a time period. But there are certainly other patterns of work. A scientist might be interested in a particular variable, say dissolved oxygen, at a particular time (say corresponding to field work), and want any stations that have an offering for that variable at the time. We have not yet devised a means to simultaneously support a variety of work patterns with the Data Explorer interface.

Some of the data-collection platforms have specialized displays related to particular properties of the platform. For example, CMOP's underwater glider, called Phoebe, runs multi-day missions over a pre-programmed trajectory. The gathered data are time series, hence can be used with time-series oriented interfaces such as the Data Explorer However, because of the nature of the glider path (repeated dives from the sea surface to near the bottom), it often makes more sense to plot depth versus time, with the plot colored by the variable of interest, such as salinity. Thus we provide a special interface for specifying such plots (which are generated by the plot service). As shown in Figure 6, there are also renderings that depict the 3-dimensional trajectory of the glider. These plots are pre-computed for each glider mission and variable.

For resource managers (and scientists), CMOP provides *watch* pages for particular interest areas. A watch page has a selection of plots connected to the interest area, along with commentary. Figure 7 shows the Oxygen Watch page, which targets hypoxic (low-oxygen) conditions [14]. It contains a plot of dissolved oxygen from multiple observation stations, along with reference lines that reflect different definitions of hypoxic conditions from the literature. Additional plots show environment conditions (river discharge, north-south wind speed) that are often correlated with oxygen levels in the estuary. Commentary in the "Blog" tab interprets current conditions.

Other watch pages under development include one for Myrionecta rubra (a microorganism) causing red-water bloom in the river [9], and one directed at steelhead survivability. The latter features displays that compare predicted plume area, volume and distance off shore to historical conditions for the same day of year, which could provide hatchery managers with guidance on the best time to release young steelhead (a fish related to salmon) [3].

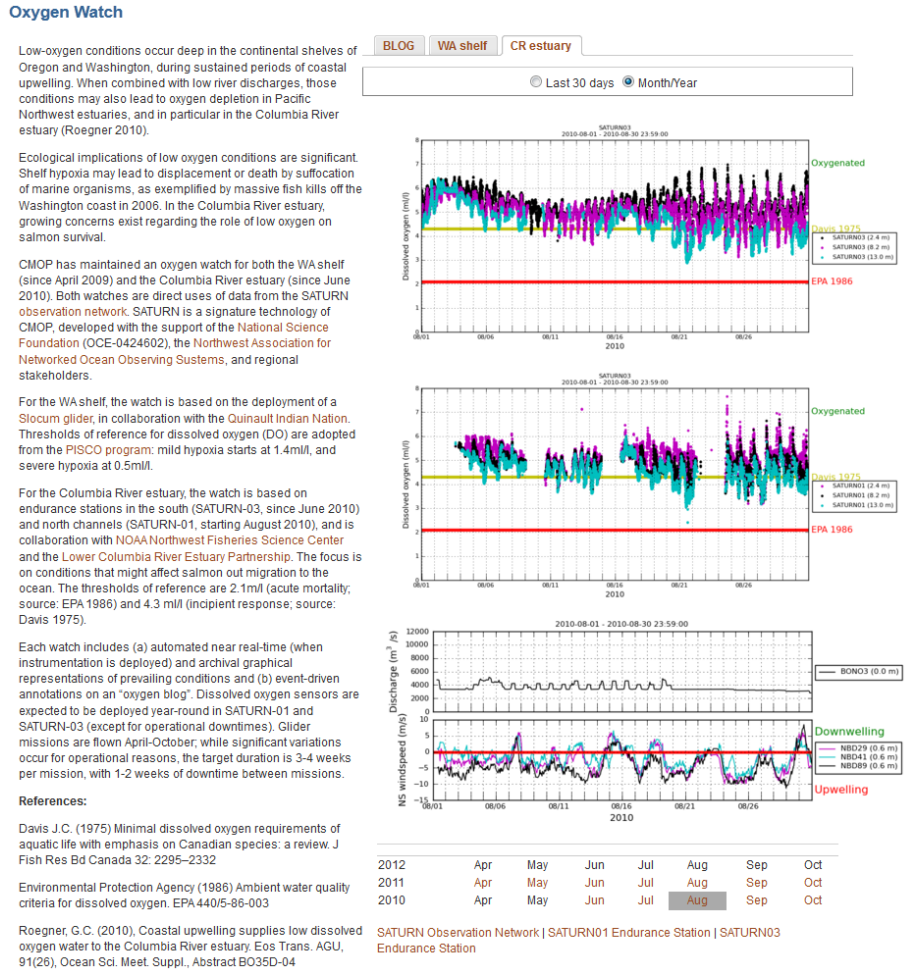


Fig. 7. The Oxygen Watch page, showing conditions during August 2010

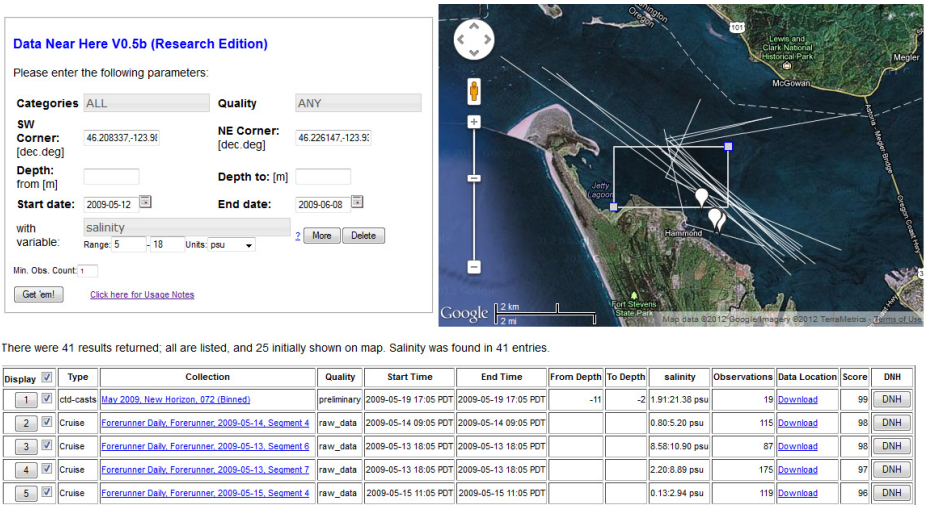
5 Supporting Ranked Search for Datasets

One challenge for CMOP scientists is knowing what datasets might be relevant to their current work. Database and basic spatial search techniques (contains, overlaps) often prove unsatisfactory, in that it is easy to get answers that return no datasets or thousands of them, requiring iterative tweaking of search conditions to get a candidate set of answers. As an alternative, we are developing an interface that applies Information Retrieval approaches to give ranked search of datasets. Data Near Here (inspired by the "search nearby" in map services) makes use of similarity search over spatial-temporal "footprints" that are computed from the datasets. Our initial work [10] focused on identifying a similarity measure that would balance geospatial and temporal search conditions in a way that resonated with our user community. At scientists' request, we have since added "dimensions" of depth, variable existence and variable

values to our search capabilities. Figure 8 shows the results of a Data Near Here query, with the top few matching datasets shown.

Most data in the archive treats depth as a separate field; also, the currently used version of spatial tools (PostGIS 1.5) does not fully support three-dimensional spatial functions. As a result, depth is currently treated as a separate search condition, and the search condition is given the same weight as geospatial location. An alternate approach is to treat the geospatial locations, including depth, as true three-dimensional locations. The current spatial distance metric does not change if given fully three-dimensional data, although some implementation details will need to change.

Scientists may also wish to search for data based on variable values; for example, all places and times where low oxygen conditions occurred. A scientist may even be searching for places and times where a variable was collected, irrespective of the variable's values. We added the capability to search over variables and their values into the same metadata extraction and search framework. The metadata extraction tools were extended to identify and store the variable names for each dataset. The variables are generally represented by column names, and so we assume that each column represents a variable. For netCDF files, this information is available in the header; for comma-separated value files it is often in the first row, and for data served from CMOP's relational database, it is in the database catalog. If available, we also capture the data type and units for each variable. If the units for a variable cannot be inferred, they are shown in the catalog as "unknown". Data types are treated the same way; alternately, techniques exist (such as those used in Google Fusion Tables [7]) to infer likely data types from the data itself. We also read the data and store the maximum and minimum values found for each variable, handling character and numeric data similarly. We intend to provide search capabilities over the modeled data.



There were 41 results returned; all are listed, and 25 initially shown on map. Salinity was found in 41 entries.

Display	Type	Collection	Quality	Start Time	End Time	From Depth	To Depth	salinity	Observations	Data Location	Score	DNH
1	ctd-casts	May 2009, New Horizon, 072 (Binned)	preliminary	2009-05-19 17:05 PDT	2009-05-19 17:05 PDT	-11	-2	1.91:21.38 psu	19	Download	99	DNH
2	Cruise	Forerunner Daily, Forerunner, 2009-05-14, Segment 4	raw_data	2009-05-14 09:05 PDT	2009-05-14 09:05 PDT			0.80:5.20 psu	115	Download	98	DNH
3	Cruise	Forerunner Daily, Forerunner, 2009-05-13, Segment 6	raw_data	2009-05-13 18:05 PDT	2009-05-13 18:05 PDT			8.58:10.90 psu	87	Download	98	DNH
4	Cruise	Forerunner Daily, Forerunner, 2009-05-13, Segment 7	raw_data	2009-05-13 18:05 PDT	2009-05-13 18:05 PDT			2.20:8.89 psu	175	Download	97	DNH
5	Cruise	Forerunner Daily, Forerunner, 2009-05-15, Segment 4	raw_data	2009-05-15 11:05 PDT	2009-05-15 11:05 PDT			0.13:2.94 psu	119	Download	96	DNH

Fig. 8. The Data Near Here prototype, showing a search based on a particular X-Y region, with no constraint on depth, seeking datasets that contain salinity in a certain range. Results are ranked on a weighted combination of similarity to the search conditions, rather than on exact match. Data can be directly downloaded, or plotted in the Data Explorer.

Once we have extracted metadata for each dataset to identify contained variables and their values, we are able to search over it using extensions of the techniques and formulae we use for geospatial-temporal search. We provide two types of search conditions for variables. The first specifies a variable name and a desired range of values, in some specified units. For each dataset that contains the desired variable in the specified units, the range of values is compared to the desired range and a similarity score computed; the computed score contributes to the overall dataset similarity score. A dataset that does not contain the desired variable can still be returned in the query results if it has high scores on the other query conditions. However, a dataset that contains the desired variable with values similar to the desired data range is likely to receive a higher overall score, even if its scores on the geospatial and temporal query conditions are lower. Unit translation is possible in many cases, and we are experimenting with approaches to this problem. If the units for a dataset are unknown, we assume the values are in the desired units but substantially discount the score.

The second type looks for datasets that contain a certain variable but does not specify a range. In concept, this condition specifies a variable with an infinite range of values; thus, any dataset that contains a column of that name, with any values at all, is considered "closer" to that query condition than a dataset that lacks that variable. In effect, the resulting score is binary: a dataset is a perfect match to the query condition if the desired variable is found in that dataset, or a complete non-match if it does not.

At present, we only match on exact variable names; a search for "temperature" will not match "air temperature" or "airtemp". In a large archive built over more than a decade, inconsistencies and changes in variable names are common. We are considering methods to match on "close" variable names, as these inconsistencies frustrate our scientists. One possibility is to extend our approach to variable existence, so that the existence of a variable with a similar name is given a score reflecting the higher similarity, converting variable existence from a binary to a continuous similarity score.

We are finding that as Data Near Here queries become more sophisticated, it becomes expensive to apply the similarity function to the footprints of all the data sets. Figure 9 illustrates the problem for queries with increasing numbers of search conditions on variables. The "cast variables" are those typically measured by lowering an instrument package from a cruise vessel, whereas "station variables" are those typically seen at fixed stations. The alternating line is for queries that include queries asking for datasets containing both kinds of variables, which none of the existing datasets will match very closely. As can be seen, response times start to grow out of the interactive range rather quickly for this last category of query.

One technique we are investigating starts by selecting a cut-off on minimal similarity score, and incorporating a pre-filter into the query that can quickly rule out certain datasets being over that score without applying the full (and more expensive) similarity calculation. As can be seen in Figure 9, incorporating the cut-off does improve response times on the more expensive searches. An area for further work is determining how to initially set the cut-off threshold for a given query and limiting the number of expensive geospatial comparisons by using cheaper pre-filters.

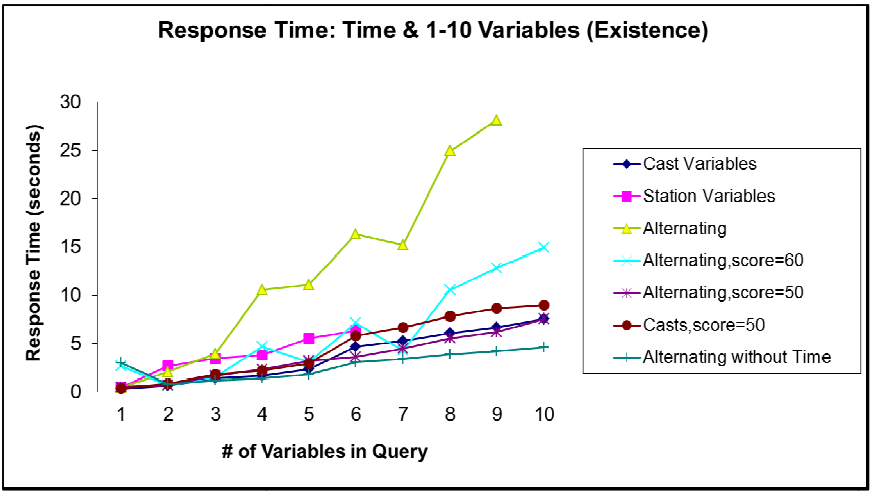


Fig. 9. The effect of incorporating an initial cut-off threshold on similarity score in Data Near Here Queries

Data Near Here currently provides access to more than 750B observations from fixed stations, glider, cruises, casts, and water samples, at three quality levels (raw, preliminary and verified). The largest dataset indexed has 11.5M values, the smallest, one. The mean dataset size is about 33K entries.

6 General Lessons

While the development of CMOP information access and analysis capabilities is an on-going process, we can identify some important guidance for similar endeavors.

1. *Don't make users repeat work.* For example, if a user has gone through a data-selection process in order to specify a plot or chart, do not make him or her repeat the specification to download the underlying data. Similarly, if a user has invested time in configuring a graph of an interesting segment of data using an online tool, there should be a way to share the result. At a minimum, the resulting image should be savable, but much better is providing a URL that can put the tool back into the same state. We are not perfect on avoiding repeated work–in some cases a user must re-specify some aspect of a plot to change it–but we are working to reduce such cases.
2. *Default to where the data is.* Upon initially coming to an interface it is useful to have default settings selected. It is tempting to choose these settings in a uniform way. For example, every station page could be set to display salinity at that station for the past two days. However, such settings can result in no data being displayed, because there is a problem with the salinity sensor or transmission of its data. In such cases we find it preferable to adjust the settings so data is available for display–for example, expanding the time period or selecting a different

variable. More generally, we try to not offer the user selections in the interface where no data is available. For instance, in Data Explorer, if the user has selected an observation station in Figure 4, only variables for that station are then listed to select from. (We could go further in this direction. For example, if a station and an offering are selected, then offer only choices of date range where data is present.)

3. *Give access to underlying data.* Any display of data should provide a ready means to download that data. While we hope in most cases a user can meet his or her data-location and analysis needs through our interfaces, in many cases a user will want to view the data using a plot style our tools do not provide, or carry out more advanced computation, say in Matlab. Thus, whenever a tool displays a data set it should be possible to download the underlying data, preferably in a choice of formats. Currently, data can be downloaded from any station page, such as shown in Figure 3 (via the "Inventory" tab), or from the other tools we discuss.
4. *Integrate the tools*. Each of the tools provided has a place in the scientists' workflows. A scientist can quickly search for or browse to a likely source of data using Data Near Here, use Data Explorer to plot some variables to confirm its relevance, and download data in a variety of formats directly from these tools. Such workflows are inherently iterative. By allowing multiple tools to operate over the same data and, where possible, pass settings and selections from tool to tool, we allow the scientists to focus on the research and not on the complexities of the tooling and infrastructure.
5. *Balance pre-computation with production on demand.* Ideally, we could provide any possible data display with zero delay. The realities are that there is a bounded amount of processing that has to support data ingest, quality assurance, model evaluation and servicing of analysis and retrieval requests. If the last grows to consume too great a share of resources, the observation system cannot keep up with the other functions. Even if we could upgrade to meet all these demands today, the continuing increasing volume and density of the data being collected would make this goal unattainable tomorrow. Obviously we can control the cost of analysis and display requests by how complex of processing we support in interactive mode. To do more resource-intensive operations, the user needs to download data and compute locally. We also pre-compute and cache display plots that are likely to be requested by multiple users, such as the plots displayed upon entering the station pages, as that shown in Figure 3. We also pre-compute and cache plots that are hard to produce at interactive speeds, for example the track plots for glider missions shown in Figure 6. The output of the simulation model is handled similarly. For forecast simulations we pre-compute various data products at the time of model generation. Many of these products are animations of a particular variable along a 2-D horizontal or vertical slice. (In fact, these animations can be computed incrementally as the model runs, and provide a means for diagnosing computations going awry.) However, it is also possible to produce map layers from model data (via a WMS [11]) on demand.

7 Issues and Challenges

While the various CMOP information interfaces described here have gone a long way towards meeting the needs of the various user groups, there are still areas that could be expanded and enhanced. Here are list of areas of work, ranging from ones where we are fairly certain how to handle to ones that will require extensive research.

1. With the wide range of interfaces, there can of course be inconsistencies. We have discussed how we try to use common components, such as the plot service, across interfaces for uniformity. We also try to drive menus and choices (such as available offerings from a station) out of a common database of metadata. However, there can still be variations in grouping or ordering of options, which could possibly become more table-driven.
2. While we have various means of showing the inventory (for different time periods) of holdings for a given offering, we lack means to depict "joint availability". For example, a scientist might want to know for what time periods is temperature available at both SATURN03 and SATURN04, in order to cross-compare them.
3. Our current plotting facility can deal with datasets spanning many months. However, we are only beginning to develop representations for multiple years of data that allow short-term trends and events to be discerned. Simple plots and aggregates can lose the fine detail.
4. As mentioned in Section 2, fault diagnosis and quality assurance are often handled with general purpose interfaces, requiring a fair amount of manual effort. We need more automated methods to allow limited staff to support continued growth in the sensor arrays. We have had some success in the past applying machine-learning techniques to detecting biofouling of sensors [1], but there remains a wide range of approaches to explore in specifying or learning normal reporting patterns and detecting divergence from them.
5. Another open area is the display and indication of uncertainty. While we are currently expanding our capabilities for flagging and suppressing problem data, we do not know of good methods to portray the inherent systemic uncertainty of our various datasets, nor can we propagate such knowledge through analysis and charting tools. We welcome the suggestions of other researchers here.
6. We have over a decade of historical simulated data, and one chief use for these "hindcasts" is *climatology queries*. Such a query aggregates possibly the whole hindcast database over time and space, for example, daily maximum temperatures over the estuary averaged by month, or fresh-water plume volume on a daily basis. A variety of these queries are pre-computed and constitute the CMOP Climatological Atlas [16], but given the size of the hindcast database (tens of terabytes), we do not support climatological queries on demand. The size of the hindcasts similarly makes download of the database for local use generally infeasible. This problem will become more challenging as we include chemical and biological quantities in our models. We also contemplate producing hindcast databases for "what-if" scenarios, such as different river-discharge levels and

changes in bathymetry (bottom topography) of the estuary. While reduced-resolution databases might address on-demand climatologies for quick comparisons, detailed analysis of differences will require computation at full resolution. Putting the hindcast databases in the cloud, and having users pay for their processing is an intriguing possibility; especially as most climatology queries are easily parallelizable. However, current cost schedules for cloud storage are prohibitive for the amount of data contemplated. One issue is that even the "cheap" option at such services has availability guarantees (99.9%) beyond what we really require. (Even 90% availability would probably satisfy most of our demands.)

Going forward, the ocean of data will continue to swell and present greater challenges for navigation. On one hand, we want to minimize both the complexity of interfaces and their need for manual support. On the other, the questions scientists are trying to answer, and their processes for investigating them, are becoming more complicated. It will be a balancing act not to constrain them by making interfaces too limited to handle their needs or too difficult to work with efficiently.

Acknowledgments. This work is supported by NSF award OCE-0424602. We would like to thank the staff of CMOP for their support.

References

1. Archer, C., et al.: Fault detection for salinity sensors in the Columbia estuary. Water Resources Research 39(3), 1060 (2003)
2. Burla, M., et al.: Seasonal and Interannual Variability of the Columbia River Plume: A Perspective Enabled by Multiyear Simulation Databases. Journal of Geophysical Research 115(C2), C00B16 (2010)
3. Burla, M.: The Columbia River Estuary and Plume: Natural Variability, Anthropogenic Change and Physical Habitat for Salmon. Ph.D. Dissertation. Beaverton, OR: Division of Environmental and Biomolecular Systems, Oregon Health & Science University (2009)
4. Cornillon, P., et al.: OPeNDAP: Accessing Data in a Distributed, Heterogeneous Environment. Data Science Journal 2, 164–174 (2003)
5. Domenico, B., et al.: Thematic Real-time Environmental Distributed Data Services (THREDDS): Incorporating Interactive Analysis Tools into NSDL. Journal of Digital Information 2(4) (2006)
6. Ghindilis, A.L., et al.: Real-Time Biosensor Platform: Fully Integrated Device for Impedimetric Assays. ECS Transactions 33(8), 59–68 (2010)
7. Gonzalez, H., et al.: Google Fusion Tables: Data Management, Integration and Collaboration in the Cloud. In: Proceedings of the 1st ACM Symposium on Cloud Computing, pp. 175–180. ACM, New York (2010)
8. Haddock, T.: Submersible Microflow Cytometer for Quantitative Detection of Phytoplankton (2009), https://ehb8.gsfc.nasa.gov/sbir/docs/public/recent_selections/SBIR_09_P2/SBIR_09_P2_094226/briefchart.pdf
9. Herfort, L., et al.: Myrionecta rubra (Mesodinium rubrum) bloom initiation in the Columbia River Estuary. Estuarine, Coastal and Shelf Science (2011)

10. Megler, V.M., Maier, D.: Finding Haystacks with Needles: Ranked Search for Data Using Geospatial and Temporal Characteristics. In: Bayard Cushing, J., French, J., Bowers, S. (eds.) SSDBM 2011. LNCS, vol. 6809, pp. 55–72. Springer, Heidelberg (2011)
11. Open Geospatial Consortium, Inc.: OpenGIS® Web Map Server Implementation Specification Version: 1.3.0 (2006)
12. Plant, J., et al.: NH 4-Digiscan: an in situ and laboratory ammonium analyzer for estuarine, coastal and shelf waters. Limnology and Oceanography: Methods 7, 144–156 (2009)
13. Rew, R., Davis, G.: NetCDF: an interface for scientific data access. IEEE Computer Graphics and Applications 10(4), 76–82 (1990)
14. Roegner, G.C., et al.: Coastal Upwelling Supplies Oxygen-Depleted Water to the Columbia River Estuary. PLoS One 6(4), e18672 (2011)
15. Szalay, A.S., et al.: Designing and mining multi-terabyte astronomy archives: the Sloan Digital Sky Survey. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, vol. 29(2), pp. 451–462 (2000)
16. Climatological Atlas, Center for Coastal Margin Observation & Prediction, `http://www.stccmop.org/datamart/virtualcolumbiariver/simulationdatabases/climatologicalatlas`