OOI CYBERINFRASTRUCTURE
NSF OCEAN OBSERVATORIES INITIATIVE

OCEAN OBSERVATORIES INITIATIVE

OOI – CyberInfrastructure

Requirements Workshop
University of California, San Diego, CA
January 23-24, 2008

Workshop Report

FINAL

May 2008

NSF

CONSORTIUM FOR Ocean Leadership
DISCOVERY • UNDERSTANDING • ACTION

UCSD

Document information

Project	Network for Ocean Research, Interaction and Application (NORIA)
Document Custodian	Oscar Schofield/Scott Glenn (Rutgers University, NJ) OOI CIIO Project Scientists
Approval	Frank Vernon
Created on	January 24, 2008
Last Changed on	May 9, 2008
Document Status	FINAL

Document History

Date	Version	By	Description of Changes
2008-03-10	0.1	M. Meisinger	Initial workshop report from notes and previous workshop report structure; requirements elaborated and refined
2008-03-14	0.2	M. Meisinger	Presentation summaries, discussion session one summary, domain models, feedback session summary, conclusion, questionnaire and agenda added
2008-03-15	0.3	C. Farcas, E. Farcas	Group two discussion section added
2008-03-19	0.4	M. Meisinger	Requirements walk-through and validation sections added; list of requirements enhanced; individual questionnaire input added (in presentations, discussions, requirements sections); RWS1 requirements recategorized together with RWS2 requirements
2008-03-25	0.5	A. Chave	Review, modifications and comments across text
2008-04-14	0.6	O. Schofield	Review, modification and comments across text
2008-04-26	0.7	O. Schofield	Shortened science section
2008-04-29	0.8	M. Meisinger	Moved science section to App. A; minor corrections across the text; requirements explanations corrected where missing; candidate version
2008-05-09	1.0	M. Meisinger	Final corrections, changed logo, added science figures, set to FINAL

Table of Contents

OOI - CyberInfrastructure
Requirements Workshop at UCSD, January 2008
Outcome and Summary

1 Executive Summary

In an effort to elicit and document community user requirements and constraints for the planned Ocean Observatories Initiative (OOI) CyberInfrastructure (CI), the OOI CyberInfrastructure implementing organization (IO) is holding a series of workshops with scientists and other future users of the CI. One of these was a user requirements workshop January 23-24, 2008 at the California Institute for Telecommunications and Information Technology (Calit2) of the University of California, San Diego (UCSD). This workshop was the second in the series and succeeded the first requirements workshop held July 23-24, 2007 at Rutgers University (NJ). It was based on the results of the first requirements workshop, documented in the outcome report [CI-RWS1].

Oceanographic scientists from the US West Coast numerical ocean modeling and oceanography communities were invited to the second workshop. The workshop goals were CyberInfrastructure science user requirements identification and elicitation, validation of existing requirements, as well as a further early outreach measure to immediate CI user communities – in this case the numerical ocean modelers from the West Coast. UCSD's Calit2 provided the scientific environment for a 2 day workshop that covered introductions to the planned CI and the OOI program, oceanographic science presentations, CI requirements elicitation and validation sessions, domain modeling and usage scenario development sessions as well as feedback opportunities.

The workshop outcome and results include

- Additional CI user requirements provided by the numerical modeling community
- Refined and validated previously existing user requirements
- Prioritization of existing user requirements during the workshop
- Domain models elaborated during the workshop
- CI usage scenarios for modeling, analysis, mission planning and control elaborated during the workshop
- Collected workshop presentation materials including introductory presentations (OOI, CI, science) on the OOI CI Confluence web site [RWS2-WEB]
- Science user questionnaires for requirements elicitation (extended and short versions)
- Filled out participant questionnaires
- Existing user requirements prioritization and validation by the participants

2 Introduction

2.1 Goals and Background

In order to provide the U.S. ocean sciences research community with access to the basic infrastructure required to make sustained, long-term and adaptive measurements in the oceans, the National Science Foundation (NSF) Ocean Sciences Division has initiated the Ocean Observatories Initiative (OOI). The OOI is the outgrowth of over a decade of national and international scientific planning efforts. As these efforts mature, the research-focused observatories enabled by the OOI will be networked, becoming an integral partner to the proposed Integrated and Sustained Ocean Observing System (IOOS;

www.ocean.us). IOOS is an operationally-focused national system, and in turn will be the enabling U.S. contribution to the international Global Ocean Observing System (GOOS; http://www.ioc-goos.org) and the Global Earth Observing System of Systems (GEOSS; www.earthobservations.org). Additionally, the OOI will provide an ocean technology development pathway for other proposed net-centric ocean observing networks such as the Navy's proposed Littoral Battlespace and Fusion Integration program (LBSFI). Additionally the global community spanning Canada, Asia, and Europe are also developing new ocean networks which all contribute to the GEOSS. Developing a robust capability to aggregate these distributed but highly linked efforts is absolutely key for them to achieve success.

The OOI comprises three distributed yet interconnected observatories spanning global, regional and coastal scales that, when their data are combined, will allow scientists to study a range of high priority processes. The OOI CyberInfrastructure (CI) constitutes the integrating element that links and binds the three types of marine observatories and associated sensors into a coherent system-of-systems. The objective of the OOI CI is provision of a comprehensive federated system of observatories, laboratories, classrooms, and facilities that realizes the OOI mission. The infrastructure provided to research scientists through the OOI will include everything from seafloor cables to water column fixed and mobile systems. Junction boxes that provide power and two-way data communication to a wide variety of sensors at the sea surface, in the water column, and at or beneath the seafloor are central to these observational platforms. The initiative also includes components such as unified project management, data dissemination and archiving, and education and outreach activities essential to the long-term success of ocean observatory science. The vision of the OOI CI is to provide the OOI user, beginning at the science community, with a system that enables simple and direct use of OOI resources to accomplish their scientific objectives. This vision includes direct access to instrument data, control, and operational activities described above, and the opportunity to seamlessly collaborate with other scientists, institutions, projects, and disciplines.

A conceptual architecture for the OOI CyberInfrastructure was developed and published by a committee established by JOI in 2006 (see http://www.orionprogram.org/organization/committees/ciarch) [CI-CARCH]. It describes the core capabilities of such a system. Initial requirements were derived from similar CyberInfrastructure projects.

In May 2007, the consortium led by SIO/UCSD, including JPL/NASA, MIT, MBARI, NCSA, NCSU, Rutgers, Univ Chicago, USC/ISI and WHOI, was awarded the development of the CI as an Implementing Organization (IO). The first six months of the design phase focused on architecture and design refinement and consolidation, and an initial science user requirements analysis and community involvement effort. In December 2007, the preliminary CI design [CI-PAD] was successfully reviewed in a PDR (Preliminary Design Review) by a panel of independent experts appointed by NSF who provided very positive review comments.

Ongoing and future efforts focus on advancing the CI design and that of its subsystems to the next level to be ready for the start of OOI construction. At the same time, the validation of any previously elicited and documented CI science user and system requirements through the community remains a main concern. Direct involvement of prospective CI user communities is of paramount importance to the success of the program. The requirements elicitation and management process is planned to be an ongoing activity in close collaboration with the user communities involved throughout the design and construction phases.

The initial direct science user involvement occurred during the first CI requirements workshop (RWS1), July 23-24, 2007 at Rutgers University. A summary of the outcome of this workshop was documented in the form of a publicly available report of similar format to this one [CI-RWS1]. In addition to involving a cross-section of the numerical modeling community, this meeting explicitly requirements given the parallel development of the OOI, IOOS and LBSFI. This report covers the outcome of the recent second re-

quirements workshop (RWS2); this time with the numerical ocean modeling community from the West Coast. It took place January 23-24, 2008 at UC San Diego.

This workshop is the second in a series of CI architecture and design team organized workshops to identify and elicit requirements from domain users. The first two workshops were targeted mainly at the numerical ocean modeling communities. Subsequent workshops will complement this input

- Workshop one covered the Mid-Atlantic community with a focus on the range of data assimilative numerical continental shelf models [CI-RWS1]
- Workshop two covered Global-Climate modeling communities and included West Coast institutions [CI-RWS2]
- Further planned workshops will cover ocean observing and instrument management, data product generation and integrated observatory management.
- The next workshops will also consider the linkages to the autonomous system and the education/outreach communities.

Goals of the second workshop described in this report were:

- Establish further direct contact to the West Coast ocean modeling community
- Provide the CI engineering team with further detailed insight into the current situation and present issues of the coastal ocean modeling community, and provide insight into current research projects.
- Identify and elicit new user requirements for the CI from the view of this specific community
- Validate, refine and prioritize existing user requirements from the first requirements workshop.
- Validate existing CI system requirements
- Develop a thorough domain understanding through direct collaboration with domain scientists in order to increase language tangibility, and document this understanding in the form of domain models.
- Refine and consolidate the basis for further requirements elicitation and domain modeling in subsequent instances of this workshop and in ongoing requirements and architecture design work

2.2 Outline

The remaining parts of this report are structured as follows: Section 3 summarizes the presentations given at the workshop and places them into the context of the scientific background. 4 documents the direct workshop outcomes, such as discussion summaries, domain models, elaborated scenarios and prioritized requirements. Section 5 lists all current science user requirements for the OOI CI, which are a result of the refinement of existing requirements and the addition of new requirements identified in this workshop. Section 6 documents participant feedback and provided conclusions from the organizers. The appendices contain further details about the workshop organization and background materials.

2.3 Preparation

The CI ADT has developed an extensive questionnaire with relevant questions for user requirements elicitation that was structured into selected categories. A shortened and tailored version of the questionnaire was sent to the workshop participants. The scientists were asked to provide answers to the questions prior to the workshop.

Each scientist was asked to prepare an overview presentation covering projects, research interests and further relevant background information related to the OOI CI. The presentations were supposed to address the main topics covered by the questionnaire. The presentations covered approximately 15-20 minutes each, including questions.

During the workshop, the extended version of the questionnaire was used to structure the general requirements discussion session. Appendix B of this report documents the extended questionnaire.

2.4 Acknowledgements

This report was developed by the OOI CI architecture and design team; it contains input from many sources, such as the workshop presentations by the organizers and invited science users, the filled out participant questionnaires, the CI preliminary architecture and design, OOI science background information by the project scientists, and notes taken by Michael Meisinger and Emilia Farcas. Furthermore, this report contains summarizing and general statements by the organizers.

We thank the participating scientists profoundly for their time and efforts during the workshop and their valuable contributions to the OOI CI requirements elicitation process. Furthermore, we would like to thank them for their efforts in filling out the participant questionnaire and providing further materials after the workshop, and for reviewing and validating this report.

2.5 Disclaimer

The contents of this report reflect the understanding and analyses of the CI architecture and design team, based on written workshop notes and general background materials – no guarantee for the correct representation of any of the participant contributions can be given. No statements in this report are verbatim quotations of the participants; there were no audio recordings of the discussions taken during the workshop.

3 Presentations

3.1 OOI Project Overview

Oscar Schofield (Rutgers University), OOI CI Project Scientist, welcomed the workshop participants and described the current status of the OOI developments. The project went successfully through the preliminary design review (PDR) and is now refining designs and planning to enter the construction phase. Next step will be a review by the SRB. This workshop's goal is the collection of new requirements and validation of existing user requirements directly by science users out of the numerical modeling community. There will be other requirements workshops focusing on different topics.

The science motivating the OOI network is based on the research community input. The numerous community reports emphasized the need for simultaneous, interdisciplinary measurements to investigate a spectrum of phenomena, from episodic, short-lived events (tectonic, volcanic, biological, severe storms), to more subtle, longer-term changes in ocean systems (circulation patterns, climate change, ecosystem trends). The introduction of high power and bandwidth will allow the transition from ship-based data collection to the management of interactive, adaptive sampling in response to remote recognition of an "event" taking place. Sophisticated CI tools will enable individual and communities of researchers to tackle their specific research questions. The following are integrative examples of some of the broad science questions that the OOI network will be able to address.

- What is the ocean's role in the global cycle?
- How important are extremes of surface forcing in the exchange of momentum, heat, water and gases between the ocean and atmosphere?
- How important are severe storms and other episodic mixing processes affect the physical, chemical, and biological water column processes?
- How does plate scale deformation mediate fluid flow, chemical and heat fluxes, and microbial productivity?

- What are the forces acting on plates and plate boundaries that give rise to local and regional deformation and what is the relation between the localization of deformation and the physical structure of the coupled astenosphere-lithosphere system?
- How do tectonic, oceanographic and biologic processes modulate the flux of carbon into and out of the submarine gas hydrate "capacitor," and are there dynamic feedbacks between the gas hydrate methane reservoir and other benthic, oceanic and atmospheric processes?
- How do cyclical climate signals at the ENSO, NAO and PDO timescales structure the water column and what the corresponding impacts on the chemistry and biology in the ocean?
- What are the dynamics of hypoxia on continental shelves?

Figure 1 shows visualization of numerical model output showing the global CO_2 flux as one important process necessary to understand the ocean's role in the global cycle. Appendix A provides detailed explanations for all these questions and shows how the OOI will be able to address them.

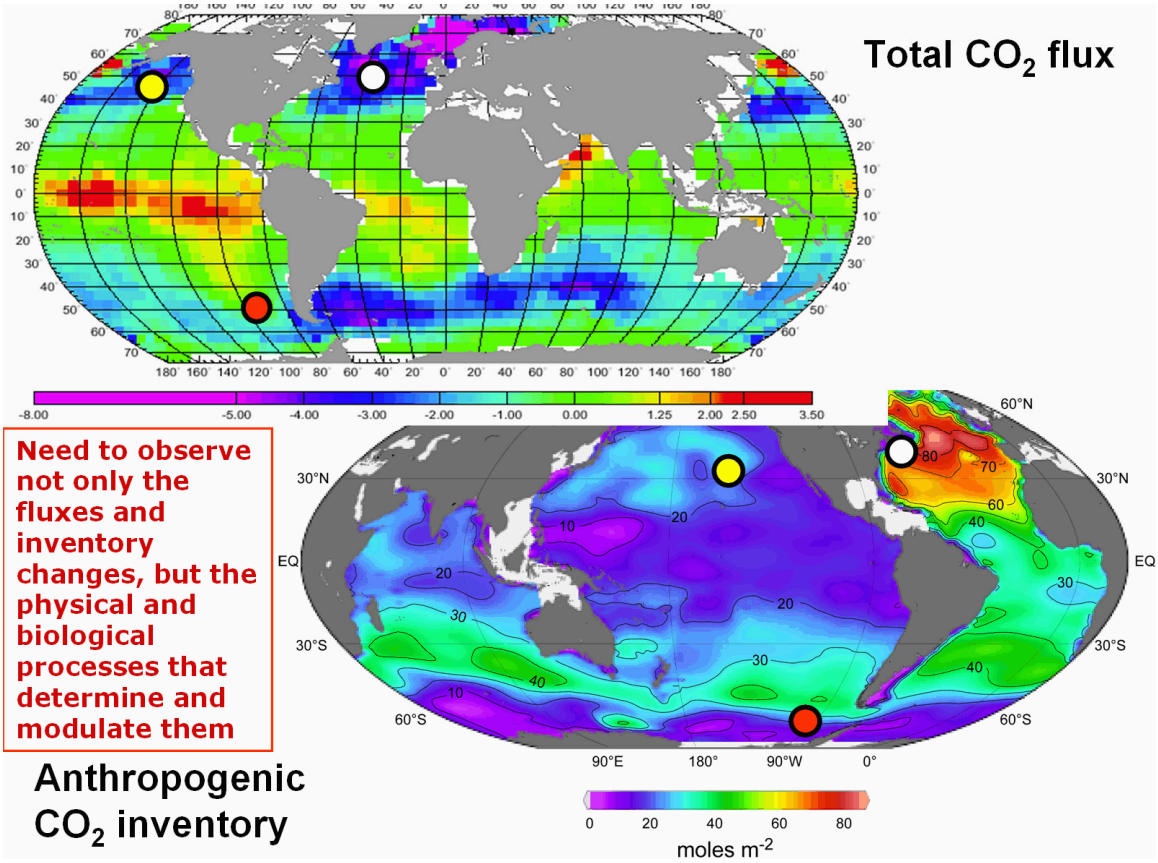

Figure 1: Visualization of total CO_2 flux and anthropogenic CO_2 inventory model output

3.2 CI Overview, Requirements, Architecture

Matthew Arrott (UCSD/Calit2), OOI CI Project Manager, provided an overview of the OOI CyberInfrastructure project and the CI project organization. The main goal of the CI is to support the three main research activities of observing, modeling and exploiting knowledge through a set of well-rounded resources and services. The CI infrastructure will be distributed across the country and will have points of presence at the sites of the main OOI observatory components on the east and west coasts. Numerical modelers are among the first recipients of data streams that are processed by the OOI CI, and therefore are a very important source for direct user requirements.

The CI will make use of the Internet2 and National LambdaRail network backbone infrastructure that enable high volume data transfer at reasonable cost for academic and research institutions. The CI will enable transport and storage of all data collected on the OOI observatories; actual decisions need to be made about which data are to be handled how based on available funding and secondary infrastructure cost. The OOI project including the CI continues to analyze similar and related efforts, as well as candidate standards and technologies, in order to make the OOI and CI design as effective and interoperable to the community as possible. Examples mentioned include SCCOOS (http://www.sccoos.org) with its data acquisition, data processing & transport and modeling workflow, the ESMF (Earth System Modeling Framework) framework for hierarchical numerical model composition and commercial elastic computing and storage service providers such as Google and Amazon. In particular the virtualization of computing, storage and instrumentation provides very powerful means for flexible deploying resources at any point of the network, on demand.

Ingolf Krueger (UCSD/Calit2), OOI CI System Architect, presented the current status of the CI preliminary architecture and design. He briefly introduced domain modeling notation and techniques that are required as a structured way to capture domain knowledge for requirements purposes and as foundation for system architecture and design. He also presented a requirements elicitation and management methodology and described the purpose of systematic and iterative requirements elicitation efforts involving multiple user communities over the course of the OOI/CI project. The CI infrastructure will provide a data distribution network with general public availability, subject to access policy. It will be possible to interface with IOOS, Argo, NASA, Codar, etc. for data exchange.

3.3 Project and Research Overview: Andy Moore

Andy Moore (Ocean Sciences Department, UC Santa Cruz) described some representative research projects and their scientific background. He has a strong interest in El Nino Southern Oscillation (ENSO) dynamics, and works on seasonal forecast models using the ROMS model platform. He is working on developing a suite of modeling tools for ROMS.

In one of his projects, a large Caribbean cruise ship was equipped with sensors and computing hardware. The ship ran every 2 weeks on the same Caribbean route and measured specific ocean parameters. Additional input came from satellite data. A computer system was installed on board the ship; it performed fully automated ensemble prediction and displayed selected results on a screen on board.

Specific statements:

- Models in the community: ROMS (MPI), OPA (MPI, OpenMP)
- There are at least 4 relevant versions of ROMS to consider: Nonlinear ROMS, Tangent Linear (TL) ROMS, Finite Amplitude TL ROMS, Adjoint ROMS
- Besides the ROMS model, there are also applications (drivers), for instance for sensitivity analysis, GST, 4DVAR. Some applications enable to compute model output for different parameter values in a linear effort instead of multiple runs.
- ROMS can be big and its application complex. Students often avoid working directly with ROMS because of these characteristics.
- Ensemble model prediction relates to computing models with slightly different realizations of forcing, boundary conditions, etc.
- Mathematical tools from the dynamical stochastic domain can be applied to model ensemble run results, if available.
- Adaptive observations, for instance mobile sensor platforms, help improve forecasting quality
- Observation sensitivity can determine the quality of a model run
- The European mid-range weather forecast project ECMWF uses these kinds of models. The modeling results are of high quality, but are not available free of charge to US researchers.

- Data and meta-data formats used include: NetCDF, NetCDF header, Grib, ASCII, Matlab
- Subversion (svn) management of model source codes – this is extremely important since code changes, corrections and additions must be made simultaneously to 4 different sets of ROMS codes

3.4 Project and Research Overview: Bruce Cornuelle

Bruce Cornuelle (Scripps Institution of Oceanography, UC San Diego) presented on the circulation in the Southwest Tropical Pacific as one example of his research projects and scientific interest. He has many similar research interests to Andy and collaborates with him on some projects.

The primary goal of ocean modeling is providing realistic models for certain ocean regions that are close to the actual observations. To run a numerical model, initial conditions, boundary conditions and forcing are required as input, influencing the model outcome. The mathematical foundation of model computation is differential equation solving. All that is needed for ocean models are initial and boundary conditions. The ocean physics, i.e. the forcing, are in this case less of a problem to create a model of the ocean. Practical problems in numerical modeling come from the availability and quality of the input data. Data are typically incomplete; uncertainties and many further kinds of problems exist. Furthermore, there exist many background constraints, such as the accuracy of the estimated wind, topography, and temperature. The entire set of challenges goes well beyond the simulation. The numerical modeler's challenge is to bring together numerous distinct data streams, make them consistent with physical observations and then create a data product. The quality of the data is a significant concern. In many model applications, disparate data that was never compared before are brought together.

A typical modeling scenario is to run the biggest model grid possible and compute the covariance of the state variables at every grid point with those at every other grid point. Such model runs may be performed many times (in the 100s) with parameters adjusted to fit to actual observations. Each run provides some new insight. Ideally, the later model runs fit closer to the observations.

Specific statements:

- External data sources are provided by Q(uick)Scat and NCEP satellites.
- Part of the numerical model is not only the model code but also items derived from it: compiler, parser.
- Numerical modeling requires a lot of technology and mathematics such as adjoint and iterative descent
- Models in the community: RSM, GSM, WRF, Delft-3d, MOM, POP, HYCOM, ROMS, POM, FVCOM, MITGCM
- Models used in particular include MITgcm, ROMS, RSM
- Metadata are present in form of standard NetCDF headers

3.5 Project and Research Overview: Bill O'Reilly

Bill O'Reilly (Scripps Institution of Oceanography, UC San Diego) presented on wave prediction and modeling. In particular, he works in a coastal data information program, measuring waves along the coastline. The data produced serve as input for wave models that are used for instance by surfers to make wave and swell maps. Wave prediction and modeling is a fairly mature research topic, where several lessons have been learned and applied. Much current work in wave modeling targets model applications, and is done often by engineers rather than scientists.

The capability limits of current wave modeling and prediction are reached because of the number of available observations. Often, models become a quality control metric for the observations. Wave model

predictions can be better than some observations from low-quality sensors. Measuring suitable wave data is very difficult. It is hard to measure waves with moored buoys, in particular with multi-purpose buoys (e.g. that also measure wind). Sensors on moored buoys only measure low order moments of relevant parameters. They are often not sensitive to the full wavelength spectrum because of their shared sensor nature, which interferes with and influences other sensors and the platform. It is physically complicated and generally not well understood how the cross-influences work. It general, it is not possible to simply add another wave measuring sensor to a given buoy. This is the reason for a substantial data QC problem. Scripps for instance has waves-only buoys, which are small, behaving like a particle, and work very well. But these buoys are not extensible with other sensors which could potentially impact the quality of the measured data; therefore they are not suitable for joint observations. Other observation stations exist as well. It would be possible different buoys in relatively close proximity. In general, it is very desirable to have different measurements at the same location. Currently, the existence of cheap sensors on cheap buoys even hurts the data quality and model outcome because they introduce uncertainties. For instance, certain instruments do not produce reliable data under certain conditions ("bad dates"), which is a hard problem for scientists.
Some community wave models exist, such as the original global wave model (WAM), developed by the Klaus Hasselmann at the MPI and Walter Munk at Scripps, a proprietary community standard model for a long time. Community model development is somewhat episodic, where individuals rebuild existing community models over time. New community models have been developed in this way, such as WaveWatch and later WaveWatch-II and WaveWatch-III, driven by certain individuals and institutions. A problem with community models is reaching the consensus to extend them or modify them, as many researchers rely on these models and have different opinions. Because of plural opinions, and no clear consensus, many things got added to WAM, but no changes to existing features were made. Eventually the model code became very hard to change and cumbersome. The idea of sharing models and model communities is good in general but scientists should not be forced to agree on specific solutions – this would stifle the general research productivity. The suggestion to the OOI is not to get too hung up on implementing "the" community wave models. It is an important research process that creates new results and insight in an unrestricted research environment.

Specific statements:

- The bathymetry (i.e. the underwater topography) is well known at large scales, but is not so well known in many shallow areas
- Many models are regional ones. Regional scale models need input from global scale wave models on their boundaries. On the global scale, however, not so much science exists.
- Normally, models are nested to satisfy boundary conditions.
- SWAN from Delft University is a model for shallow waters.
- Typical wave modeling outputs are 2D wave spectra as functions of frequency and time.
- "Bin" is a measure of frequency directional space. Frequency is split up in a way to get consistent wavelength bands.
- Wave modeling topics of concern are data sources, data sinks and non-linear interactions
- Model size and execution limits: Global models (~50km grid) e.g. have 130000 ocean grid points; there are also regional models (~10km grid) and shallow water models (~100m grid).
- All models consume about the same amount of memory and CPU time; they may differ in the scale of the model grid. Resource limitations prescribe how models run operationally.
- Available computational resources limit the model resolution. Current goals are to increase the resolution of direction from the present 15 to 10 or even 5 degrees as more computational power becomes available.
- Models are for instance run by NCEP on grids. Individual grid points get assigned to separate CPUs.

- Typical model resolution: Models predicting 24/48/72 hour time interval in 3 hour time steps (1 hour steps for shallow water models). Models are re-run every 6 hours
- Today, groups often don't distribute model output for all computed grid points. For instance, on FTP servers, only subsets of model output are provided for download for reasons of data bandwidth conservation. The complete data sets are too large. However, different scientists are interested in different subsets of the data. It could be very valuable for some to get data out of the published data points.

3.6 Project and Research Overview: Libe Washburn

Libe Washburn (Department of Geography and ICESS, UC Santa Barbara) presented on coastal physical oceanography. His research interests center on the processes forcing coastal currents, low frequency wave phenomena, and sub-mesoscale & coastal flow structures. As an oceanographer, he works closely with numerical modelers and on many interdisciplinary studies. For instance, in the Moorea (MCR-LTER) project, an understanding of waves is very important, because waves break at reefs and create strong currents that transport nutrients. Also important are salinity levels.

Specific statements:

- Real-time continuous monitoring is very important to sample episodic events; currently such events cannot really be observed except serendipitously.
- The project public websites are used in E&O and get many hits from schools and from the community.
- It is difficult to find and provide good metadata. Challenges are to find ways to define "controlled vocabularies" in order to describe the data so that someone can reproduce them.
- For instance, some sensors gather raw data (archived forever, some offline), processed in a non-linear way to provide a "data product". It would be good to have archived raw data available online.
- Observations of currents with CODAR requires waves that are big enough to reflect radio waves. In the absence of such waves, some currents cannot be measured, which leads to data gaps or wrong readings.
- Models used include Regional Ocean Modeling System (ROMS) for ocean dynamics, HydroLight for radiative transfer, SB DART for radiative transfer, MM5 for mesoscale atmospheric dynamics
- Data sources used currently include: numerous oceanographic archives including: NOAA National Data Buoy Center (NDBC), Coastal Data Information Program (CDIP)
- Data formats: ASCII, binary, OPeNDAP; meta-data formats: XML, EML

3.7 Project and Tool Presentation: Yi Chao

Yi Chao (Jet Propulsion Laboratory, NASA) presented about projects and technologies regarding the JPL OurOcean portal (http://ourocean.jpl.nasa.gov/ourocean.html). One of its goals is to do adaptive sampling. For instance, glider owners can use the portal to run several model configurations, and then identify where to deploy the gliders next. It is also possible to create model prototypes on demand. A web interface exists to create a model assimilating different data. The portal interfaces to the LAS server provide a standard user interface to NetCDF files. Ferret is used for visualization.

Modeling areas include real-time modeling for forecasting. On a dedicated computer every six hours, data are obtained (in situ, land-based, and satellite) from a variety of data servers, assimilated into a numerical model, and a nowcast (also known as analysis) is produced. In addition, batch-job modeling for research is done on the supercomputers at JPL or NASA Ames Research Center with a variety of ocean models

ranging from Pacific basin-scale to the regional and coastal scales. The goal is to test the model parameterizations and various boundary and forcing conditions with the goal of yielding the best agreement between model simulation and data.

Specific statements:

- A model can be empirical, statistical, or numerical.
- Numerical models start with the continuous equations, digitize in finite space and time, and integrate them with time.
- Data are the information coming from observing platforms including in situ and remotely sensed (land-based or spacecraft). Metadata describe details needed to fully interpret the data.
- A workflow represents a sequence of commands and instructions that process data, run models and manage the output.
- Models within the community include: MOM, POP, HYCOM, ROMS, POM, FVCOM, MITGCM.

4 Workshop Outcome

4.1 Questionnaire Response Analysis

The CI ADT received substantial input from the participating scientists through the questionnaires that were handed to them prior to the workshop with the request to provide answers to as many questions as possible. The input from the questionnaires went directly into refining and validating the science user requirements. Selected statements are listed in various sections throughout this report, in particular the individual participant presentation sections and the general requirements discussion section.

4.2 Present Day Numerical Ocean Modeling Scenario

The goal of the first breakout session was discussion and analysis of present day numerical modeling and related workflows, processes, responsibilities, technologies, etc. The workshop split into two groups with the same task for each group. After the analysis and discussion, each group was tasked to work on a depiction of the numerical modeling domain as a domain model. Background on the notation used for domain models can be found in product AV2 of [CI-PAD]. The following sections contain summaries of the two discussion sessions and the two subsequent domain modeling sessions.

4.2.1 Group 1 Discussion Summary

Andy Moore and Bruce Cornuelle provided input in group one. The discussion started with the activities leading to a usable forecast model. As mentioned above, the typical scenario in numerical modeling includes filling in past initial and boundary conditions, for instance taken from global scale models and observations. Then models are executed for a hind-cast simulation window, e.g. 2 weeks in the past up to the present. Subsequently, simulation outcome data are compared with observations from the model covered area over this hindcast time; model parameters can then be optimized for a best fit with the observations, which potentially leads to a rerun of the models. After this initial model development and tuning phase, forecast models can be calculated, using current initial and boundary condition values and external weather forecast data. While the model is running, no further current observations are considered. Models are typically computed as ensembles: based on a selection of slightly varied starting conditions and parameters, several model runs are computed and their relationships are evaluated (by means of averaging, variation, etc.). An ensemble member is an individual forecast with slightly different initial and forcing conditions, which represent specific values in a probability space of initial conditions. A typical number of models in an ensemble is 50.

A numerical model is based on a set of differential equations describing ocean physics. A number of settings go into a given model run that are needed to reproduce it. The number of possible settings can go into the millions. It makes sense to distinguish control parameters (which can be tuned for more accurate models) and fixed parameters (which modify the model characteristics and are better left alone). Model tuning is performed manually. Parameter tuning is based on extensive research and on answering research questions. Data assimilation is also a form of model adaptation that can be performed automatically. Model tuning ends after the initial development phase; subsequently, the model can be used for forecasts.

For each model run, there is a need to retrieve various data products from the internet. This is mostly done automatically by a script or program. Data are processed in proximity to the model. Irrelevant data can be stripped out and data can be reformatted. A current major impediment is getting all the data in real-time to the model platform on a regular basis. Picking the best and criteria to select data sources cannot be achieved automatically. Often, knowledge is exchanged between members of the community, for instance in the special model collaboration group. For most of the developed models, there is only one possible source of data that are provided in a format chosen by the particular data provider. If there is a choice of several data input products, data sources can be changed throughout a project with some manual adaptation effort for supporting tools. A change of data products often affects hind-casting outcome, however. For the presented projects, data are transformed after download into the NetCDF format (from ASCII, GRIB, etc.) and then filtered if needed. Most of the required data transformations involve reformatting. The ROMS model, for instance, requires one to many NetCDF files as input for the forcing and several more for the open boundary conditions.

Development and tuning of models is typically done on local development machines, while production runs execute in grid environments with reserved resources. Currently, access to grids is cumbersome because of the job queuing systems that are used. Dealing with shared resources is a big issue that grids should ideally fix, but this is not always achieved. Further problems occur when submitting jobs to grids. One problem is the time delay in getting models executed. This delay can be unpredictable for shared use grid environments. In the presented projects, most models are computed on local research institution clusters, but infrastructure such as NCAR / NCSA / SDSC clusters are used is well. The community is currently not willing to accept any impact on runtime and network latency for more flexible models. Currently the demand is to optimize the number of model runs and output grid resolution, not to increase portability, etc.

Numerical models can produce different kinds of output products: near real-time (“quick and dirty”), and QC’ed products with a latency. Model output data are 3D/4D fields, often in one big block of data (e.g. 1.8GB bulk of data). A forecast model outputs 3D fields with many variables (temperature, salinity, current) for several times (e.g. every 24h for 14 days) for all (e.g. 50) runs in an ensemble. A practical concern is the huge size of data sets. Choosing the model products to archive presents difficult choices. It is possible to compute an almost infinite amount of output data. The decision to publish and/or archive data needs to be made based on informed decisions

Currently, the development of numerical models is complex and requires substantial technological and scientific knowledge. It is not realistic for interested third-parties to get and run a numerical model e if they don’t know anything about the model and its environment. It is relatively straightforward to provide the source code and everything necessary to reproduce the model results. ROMS is already set up like this. For instance, a model that is under configuration control can be reused. However, one problem is to reproduce the execution stack of compilers and libraries in the appropriate versions so that it is possible to run the makefile. The size of the model core application is relatively small (on the order of several 100MB). If it would be possible to reliably recompute model output when needed based on archived data and model version, this would be preferable to storing the model output. This would exchange computation resources for storage resources.

Visualization can be decoupled from model computation based on availability of resources. ROMS models define the type and extent of output they produce, for instance only a snapshot of circulation at specific times or averages of values. ROMS works together with the NCAR plotting package for ROMS. Custom Matlab processing and visualization is often done as required. There is no consistent manner of visualization in the community. There are many different technologies existing with different capabilities and each with a learning curve. There is no standard tool to plug in and fly through data. Plotting tools instead develop into computational tools assuming more and more functionality, and get more and more complicated to use separately from the actual models. Currently used technologies include Ferret for oceanographic visualization, which is not easy to use, GMT with a steep learning curve, Matlab, and ROMS with its NCAR graphics package.

4.2.2 Group 1 Domain Model

Group 1 designed the domain model in Figure 2 describing the numerical ocean modeling process. Appendix C contains a larger scale version of this domain model.

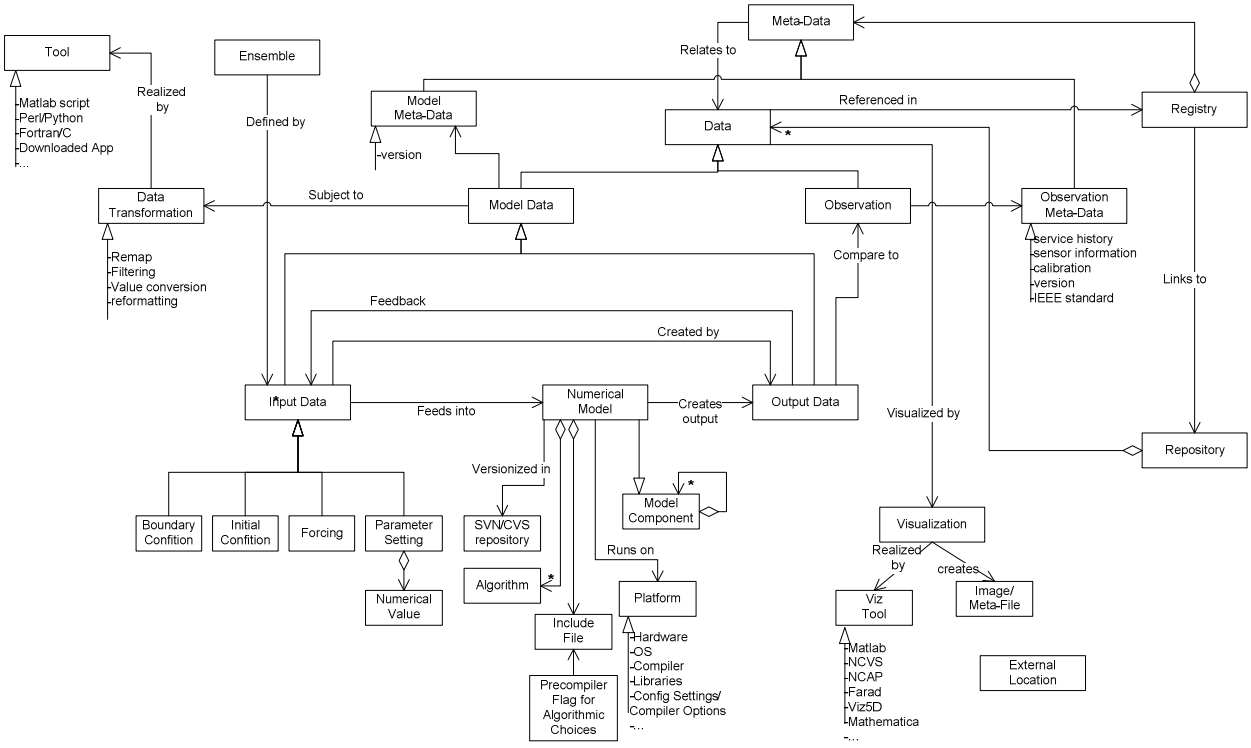

Figure 2: Domain Model Group 1

4.2.3 Group 2 Discussion Summary

Bill O'Reilly and Libe Washburn provided input in group two. The discussion started with data collection challenges. A primary problem is the collection and archival of data. Sometimes, the researchers who collect the data are not the data users – this means that the level of quality may not be the one needed by the users. For remote-sensing data processing, quality control is not a straightforward process. A substantial improvement would be the consistent and pervasive introduction of metadata associated with data products. This is a basic feature that has not been done before.

For instance, NDCS uses a network of buoys that provide their data in real-time. Nevertheless, it takes one year until the researchers process the data, quality control them, and then make the processed data

public. Other scientists display generated images of all data on the web and provide the data themselves only by request. Until the QC process is performed, such data come with the disclaimer that there could be some issues that are not yet fully understood. While a lot of QC can be performed automatically, there are still some inconsistencies or faults that can only be detected by users. A standard way of including the feedback from data users into the data products is not yet available.

Data discovery is an important issue - to be able to locate where the repositories with interesting data are located. Nowadays, there is a lot of effort invested in identifying relevant data, especially for interdisciplinary data sets which generally come from national repositories of oceanographic and meteorological, or ecological data) Hence, data availability is a real bottleneck.

Aside from the discovery problem, a remaining challenge is that data are often not well described. Generally, the data come in plain ASCII flat files accompanied by a "Readme" file about how the data have been collected and processed. In the worst case, the data comes in some form that cannot be reused such as bitmap images and plots.

Just a few data providers support subsetting of data; most of them do not provide this capability, resulting in unnecessary bandwidth and storage usage as subsetting is carried out locally using tools such as Matlab or Excel. For instance, a scientist would benefit from specifying a region and a time period and obtaining all the data from that area and interval. A useful visualization option for this feature would be to get a grid and have the different data sources plotted on the grid. Each data source on the grid would have descriptive text identifying the method of obtaining the data at that point. The reason is that many measurements are obtained for the same point using different methods and instruments. Furthermore, a data stream has to have some reliability information such as experimental state or quality control state associated with it .

Data distribution comes at a high cost for data producers. For instance, at the Department of Civil and Environmental Engineering at UC Berkeley, a lot of wave data come from their own instruments, although historically the NDBC data was also used. The delay between the measuring time and when data are presented on their website is about 45min. The buoy saves the data for half an hour and sends it to the shore seven times for redundancy purposes. After half an hour, the data represents the spectra for the previous half an hour. Some systems transmit processed data via satellite; some save the data on board and send also some basic QC data about the time series. After QC (sometimes in both places, on the buoy and at the shore), the processed data are presented on their website and sent to NDBC for archival. The results of wave modeling – get all the buoy data on a region and predict values of all other grid points – are also available on their website as wave maps/plots. This information is very valuable to the general public (e.g., surfers); thus, thousands of people come to their website every day, generating a high bandwidth and computation demand on their web-servers.

Another issue is the availability of computational power, as the models may require huge amounts of computation, which may not be readily available to the scientists. For instance, at UC Berkeley researchers perform real-time analysis and a little nowcast. The next step is to process five years of data (which is now possible for Southern California) and do a forecast. For better prediction, they run simulations and compare the buoy data with the Wavewatch III forecasts. Trend detection requires good historical data. For hindcasting, it is very important to obtain solid numbers that can be used for a long time, instead of real-time data with lower quality. For example, some scientist might think that he discovered a big shift and, in fact, a buoy was moved, but he could not find that information from NDBC. Platforms are changed all the time for navigational purposes or to detect new things, but such changes affect historical values, relevant for example in climate change. To avoid such mistakes, data sources that change their location or purpose should make clear that they provide a different data stream by having their name changed, adjusting the accompanying metadata, or notifying subscribers of the original data stream. These

changes also have to be reflected in the discovery process, where in a history search some parts would show the data, while others would not.

Metadata requirements apply also to archives of model output. An example is the ROMS model (a collection of algorithms) that can be easily downloaded as a software package. As each run of the model requires many configuration parameters that are stored into settings files full of coefficients, the researchers share their setting files to replicate experiments. Most adjustments to the model come from adjusting its parameters, whereas the algorithms are changed just from time to time. The boundary conditions limit the range over which the model operates. Through transformations, the QC data are used to obtain the boundary conditions, which may also come from other models. The initial conditions – the state the model begins with – come from observations. The models focus on a particular spatial-temporal domain bordered by the boundary conditions, but they rarely offer predictions on the boundaries themselves. However, instead of downloading and executing the model, most people are more interested in working directly with the output of the model (about 1GB every six hours as a compressed ASCII file with the spectra). For comparison, the instrument data stream are much smaller than the model output. The parameters of the model can be attached to the output as metadata. This is important as the models could feed into other models. Metadata should describe all aspects of the workflow, e.g., sampling rate, position of the instrument, meaning of the data output.

4.2.4 Group 2 Domain Model

The second group developed the domain model in Figure 3 describing the numerical ocean modeling process. Appendix C contains a larger scale version of this domain model.

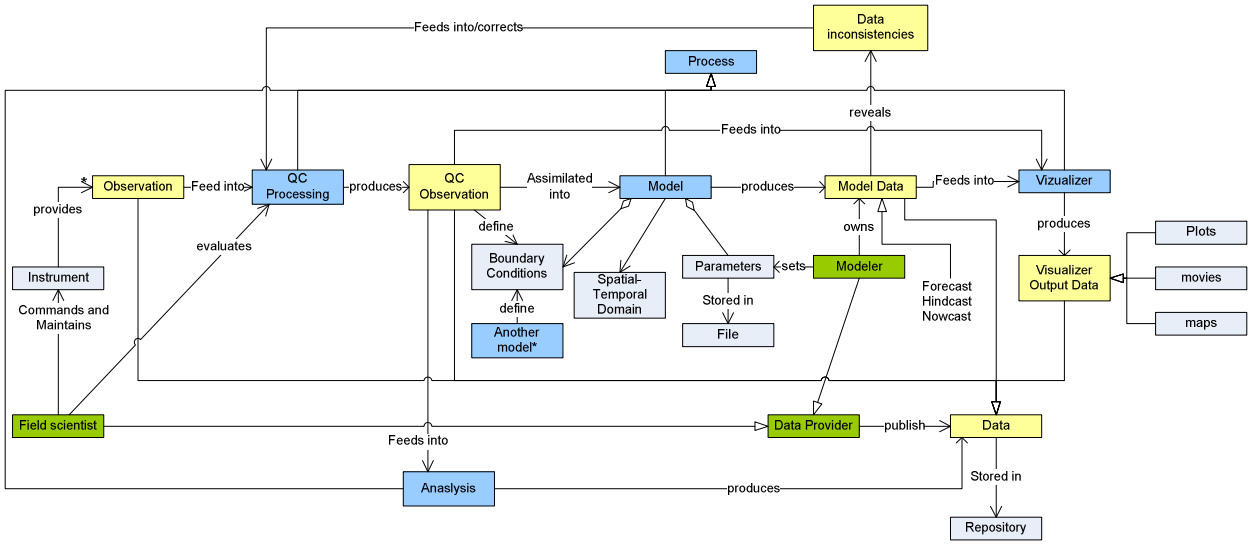

Figure 3: Domain Model Group 2

4.3 Existing User Requirements Discussion

4.3.1 Requirements Walk-Through

The workshop participants discussed the list of existing CI science user requirements, as documented in the first requirements workshop report [CI-RWS1]. The goals of the walk-through were a validation and refinement of these requirements.

Comments made in this session lead to updated RWS1 science user requirements as documented below in Section 5.2. Changes include refinements to the requirements and their explanations, as well as subsumed and dropped requirements to documented reasons.

4.3.2 Requirements Prioritization

The participants discussed and rated the science user requirements, in their form as documented in [CI-RWS1] using the following attributes:

- Essential (product unacceptable unless existent),
- Conditional (would enhance the product),
- Optional (may or may not be worthwhile),
- Reject (should not be considered as requirement)
- Rephrase (in this form not ratable)

R-ID	RWS1 Requirement	Importance
RWS1-R1	The CI shall support distributed resources and actors	Essential
RWS1-R2	The CI shall facilitate user offline operation	Essential
RWS1-R3	The CI shall facilitate adding new resources and applications	Essential
RWS1-R4	The CI shall facilitate the translation between specified data and message formats	Essential
RWS1-R5	The CI shall facilitate the translation between user-specified message formats	Essential
RWS1-R6	The CI shall provide application program interfaces (APIs) to CI services	Essential
RWS1-R7	The CI shall provide synoptic time throughout the OOI observatories	Essential
RWS1-R8	The CI shall utilize open standards and software to the maximum possible extent	Essential
RWS1-R9	The CI shall provide a catalog for all resources under CI governance	Essential
RWS1-R10	The CI shall provide the capability to discover all resources based on provided selection criteria	Essential
RWS1-R11	The resource catalog shall include information about physical samples	Essential
RWS1-R12	The CI shall support links to non-OOI resource catalogs and metadata	Essential
RWS1-R13	The CI shall provide unique identification for resources, including data streams and data sets	Essential
RWS1-R14	The CI shall provide pointers from entries in the resource catalog to the resource subject	Essential
RWS1-R15	The CI shall provide pointers from entries in the resource catalog to their associated metadata	Essential
RWS1-R16	The CI shall bind metadata to all resources connected to an OOI observatory from inception to removal	Essential
RWS1-R17	The CI shall incorporate information on citation and correspondence of resources into the bound resource metadata	Essential
RWS1-R18	OOI-standard metadata shall include, but not be limited to, a complete description of behaviors, content, syntax, semantics, provenance, quality, context and lineage	Essential
RWS1-R19	The CI shall relate different data streams that are based on the same source data	Essential
RWS1-R20	The CI shall offer data stream subscribers fallback options with similar data in case of original data stream unavailability	Optional
RWS1-R21	All data or data products associated with an OOI observatory shall be archivable	Essential
RWS1-R22	The CI shall facilitate the archival of versioned data	Essential
RWS1-R23	The CI shall verify the accuracy of archived data throughout the OOI life cycle	Essential
RWS1-R24	The CI shall ensure that archived data are up to date	Essential
RWS1-R25	The CI shall facilitate the integration of multiple data streams or data sets into a single stream or set, including elimination of redundant entries	Conditional

RWS1-R26	The CI shall support notification of changes in resource state	Essential
RWS1-R27	The CI shall provide a standard set of tools to compose and execute processes	rephrase
RWS1-R28	The CI shall facilitate data manipulation such as re-projection, re-gridding, sub-setting, averaging, filtering and scaling	Reject
RWS1-R29	The CI shall facilitate alignment of data gridlines based on resource meta-data when combining multiple models	Essential
RWS1-R30	The CI shall facilitate publication of processed data streams as new data streams	Essential
RWS1-R31	The CI shall provide subscription facilities to data streams	Essential
RWS1-R32	The CI shall provide time zone conversion capabilities for subscribed data re-sources	Reject
RWS1-R33	The CI shall provide resource access statistics	Conditional
RWS1-R34	The CI shall provide web-based user interfaces	Essential
RWS1-R35	The CI shall provide the capability to make OOI-standard metadata human read-able	Essential
RWS1-R36	The CI shall facilitate the integration of user-friendly 4D data visualization tools	Essential
RWS1-R37	The CI shall facilitate resource listing based on user selected sort criteria	Essential
RWS1-R38	The CI shall provide real time tailorable data plotting capabilities	Essential
RWS1-R39	Web-based documentation for all components of the CI shall be available	Essential
RWS1-R40	A mechanism to incorporate user-suggested modifications to the CI shall be pro-vided	Essential
RWS1-R41	CI source code developed by the CIIO shall be publicly available	Essential
RWS1-R42	The CI shall provide documentation for any application program interfaces (APIs) to CI services	Essential
RWS1-R43	The CI shall provide mechanisms to enforce user privacy policies	rephrase
RWS1-R44	The CI shall provide for the sharing of resources subject to specified policies	Essential
RWS1-R45	The CI shall provide access to resources subject to use policy	Essential
RWS1-R46	The CI shall deliver messages with accuracy comparable to that of the Internet	Essential
RWS1-R47	The CI shall support real time, guaranteed delivery, pull mode, streaming and register to receive communication capabilities	Rephrase

4.4 Requirements Discussion Summary

The requirements discussion was structured along the extended questionnaire as documented in Appendix B. This section documents facts and statements made during this session. Stated requirements were added to the list of requirements in Section 5.2 in an abstracted form.

Expected changes and transformative vision of a community infrastructure

Expected changes in the next years within the community:

- Existence of an infrastructure and tools that enable non-specialists to run and perform minor model manipulations
- Ability of individual users to influence how any model output is stored and used.
- Data management. One of the biggest challenges is managing the data and model output. The data are archived in many distributed locations. The model output is huge (on the order of many GB per day); therefore extracting a subset of a few months or even a few years duration will require going through TB of data and extracting KB of data for analysis. It would be a major breakthrough to develop a data extracting/query tool for distributed data and large-size model output.

- Web-based analysis and visualization tool. Currently, we are using off-the-shelf desktop visualization tools (e.g., Matlab, IDL, GMT, ncview, vis5D etc) that are locked to a particular CPU. It is desirable to be able to analyze data through a web-based interface accessing local and remote data sets.
- 3D visualization tool. Ocean information is 3D in nature and has to be visualized using the right tools. Most 3D visualization tools are on the high end, and usually are very difficult to employ for general users. It would be highly desirable to have access to some simple-to-use 3D visualization tools.
- A single portal with all the available modeling codes, documentations, and users' experiences (shared via wiki for example) would be useful; it is also desirable to have experts available for questions and hands-on help.
- A single portal with all of the available data assimilation schemes and the associated documentations would be useful.
- Perhaps a transition to more real-time products, meaning more need for real-time data streams.
- Improved assimilation methods enabled by expanded computing resources
- More cooperative quality control, so that datasets used by many are QC'd by the most appropriate group and the changes are supplied to all. The assimilation is the ideal QC tool, but it is sometimes more cost-effective to find outliers other ways.

Envisioned transformative changes over 5-10 years, for instance through the OOI CI:

- Numerical Modeling
    - Enable users to determine different configurations for existing models
    - Provide the ability to combine very disparate data sets with suitable user interfaces and visualization
    - Provide the ability to run large scale models across different locations on the network
    - Users should be able to interact with models directly via a simple web interface (on-demand modeling).
    - Modeling and data assimilation codes should be as portable and easy to use as Matlab subroutines.
- Data retrieval, research
    - Availability of "one place to go shopping for everything", "a sophisticated Google"
    - Everything hooked into everything else (external data sources, computation, storage resources)
    - An infrastructure with a uniform and streamlined user interface with homogeneous presentation
    - Enabling the ability to find all kinds of data in one place (e.g. temperature, salinity, wind, multiple satellite data for one region)
    - Provision of the ability to ask generic, multi-disciplinary research questions with answers supported by identification (suggestions) of suitable input data, transformations, resources, tools, visualizations
    - Transition to a service oriented way of providing information
    - Provide students with flexible research capabilities across disciplines and data sources
    - Users should have ways to find out what model outputs are available in what regions, and download them if needed.
    - Datasets should be available with wiki-style QC and CVS-style revision logs.
- Advocate CI design and requirements such that other data centers adopt them to achieve similar capabilities for effective linkage of data in a transparent way
- Advance of social networks, ad-hoc communities
- Technology advances
    - Transmitting data without wires underwater to observing stations

  - Virtualization of computing, storage, processing ("magic" resources)
  - Support execution of large scale models with varying resource needs
  - Provide flexible access to remote resources (c.f., Condor tool)
  - Decouple observation request definition from resource matching to execute a job
  - Availability of ubiquitous wireless networking to connect sensors, applications, users
  - Access to unlimited distributed data storage on the network
  - Ability to use the unlimited computing capability on the Grid/virtual-computer
  - Open-source tangent linear and adjoint compilers, such as Open AD

The most beneficial advances for the community would be:

- Making more subsurface measurements available (e.g., from vertical profiles)
- Providing the ability to access data in real time
- Universal data delivery

Specific statements:

- The Condor tool (Condor High Throughput Computing) makes resources available in disparate networks
- An IT infrastructure can only complement solutions for underwater wireless data transport, such as by reliable communication, retransmission, buffering, time-stamping, bandwidth scheduling etc. Other issues, such as power management need to be addressed elsewhere.
- Virtualization of computation could lead to the distribution of individual processes to available resources across the network and thus to increased communication overhead and bottlenecks. The infrastructure's responsibility will be to avoid such bottlenecks by assigning processes to large enough resources (e.g., a massive model better be deployed on a mainframe) and considering network proximity. The main goal of computation virtualization is not to enable distribution but to support all diverse computation needs in the system. For instance, different instruments may require the execution of processes.
- Setting priorities for resources, such as determining the precedence of certain computations over others is one form of policy.

Shared resources and community infrastructure

Resources that the researchers envision sharing with the community through a community infrastructure include:

- Resources as part of the Condor resource pool
- Desktop, laptop, running tools
- Modeling knowledge, experience and expert advice through CI community mechanism
- Numerical models
- Data assimilation algorithms

The community infrastructure should act as a single point of access for any kind of registered resource in the form of a clearinghouse. It should enable the search of resources and point to the places where the resources are available. The infrastructure could for instance provide various tools in a similar way to the Matlab toolbox where tools can be selected based on various criteria.

Comparable and exemplary efforts

Processes and platforms that function well exist for

- Numerical weather prediction
  - Diverse ways of accessing the model data (e.g. through mobile devices)

- Weather and atmospheric re-analysis (NCAR)
  - Geosciences "Collaboratory"
  - Can integrate several data sources (weather stations) into a coherent framework
  - The OOI can learn from data storage, data processing, filtering ("what to keep?") policies.
    - Meteorologists often have to exclude substantial parts of satellite data. Could learn from the applied tools and strategies
    - The challenge is in incorporating data. Having more data will not necessarily make the estimate better.
    - In some ocean models, about 30% of the available input data that could go into the models has the potential for degrades the forecasts, because of lower quality sensor characteristics and measurement interferences.
  - Have good visual products (e.g. "cloud picture in newspaper") rather than plots
- Medical research
  - Data published are available to the public with metadata in a documented format
  - Analyses can be reproduced in exact form by anyone

Numerical Models and Modeling

Numerical model development and use:

- It is not realistic to expect to have ROMS or MIT GPL available on the CI, because these models are developed and maintained elsewhere. It might not be desirable for the OOI to have models resident on the CI. The CI could run community models though.
- There exists a fundamental difference between model use and development. Simplifying use would be a great advance, but development remains where it is because it is a research issue.
- The CI could make the use of existing models easier through user interfaces to define basic model parameters, data archival, model output etc. Currently, the learning curve for numerical models is substantial (3-6 months) – you have to know the right people, and the right setups.
- The CI could increase model configuration efficiency for instance through intuitive GUIs for non-experts that enable the choice of all the options for a model (see ROMS). In such a case, the community will be very receptive and there will be experts who will provide comments and contributions.
- These user interfaces could capture expert knowledge and make it available to non-experts, thus improving their learning curve and efficiency. This also reduces the communication load on the experts, because the CI acts as a broker for the knowledge.
- The CI should provide more intuitive choices for parameters and facilities for easy model diagnosis, debugging and tuning. This will especially help non-experts and students.
- Enforcing a certain logical framework for model configuration will probably have an impact on the models and constrain them somehow.
- The CI should facilitate model documentation and self-documenting models.
- There are on the order of 10 models in the community. Many models are similar enough, so that it is possible to pick one representative one for developing interfaces and parameterizations.
- It should be possible to upload the entire model configuration so that model runs are fully repeatable by third parties, including the visualization of the model output.

Numerical model execution:

- In real-time applications, models are run almost constantly.
- The biggest enhancement to numerical modeling was the availability of model ensembles.
- Computing ensembles is a benign problem, because each job (ensemble member run) is completely independent from the others and can be sent to a different cluster.

- Supercomputing is under current conditions infeasible for real-time models because of the unavailability of supercomputers on a regular basis with predictable latency
- Model algorithm changes can occur frequently. For instance the ROMS repository typically changes twice a day
- In general, the goal of developing a model is to keep a running model as long as possible, so that model results are comparable over longer periods of time
- Virtualization of resources for running models can provide many advantages. If the CI can pioneer a solution to the problem of elastic computing and a related execution environment, this would be a great improvement for everyone, not only inside the community.
- The CI can provide solutions to the community that are solved only once and thus avoid most redundant work on the side of the modelers.
- Currently, many nested models require a shared memory machine, because ocean model boundary conditions update over large volumes of data. There is often strong coupling between the tiles of the models.
- When downloading current input data for models, the available data bulk files are typically fine grained enough so that there is not too much overlap and required retransmission. Scripts can detect which data were already downloaded.
- For education purposes, on demand modeling could be very useful. This can be simple models.
- Currently, all model input and output is performed using exchanged data files. There is no concept of data streams used when executing models

Numerical model output and visualization:

- Model output visualization is currently quite cumbersome, for instance using the Life Access Server (LAS).
- Existing 2D plotting tools include Matlab, Ferret and GMT. Interactive 3D/4D plotting tools don't exist.
- The community needs a new way of analyzing the 3D ocean. Any existing visualization tool only provides batch mode and no interactive visualization similar to GoogleEarth

Data sources, data transformation

Selecting data sources:

- The question of how to satisfy all the boundary conditions for numerical models is currently answered manually. GODAE servers provide some data to facilitate decision. It needs to be determined what the decision criteria are.
- The CI could help to share the metrics that contain the assumptions and the knowledge about the data and quality of the models and input in several years
- The CI should provide decision support by providing information co-located in one place such that data possibilities can be compared side by side.

Data source format and quality:

- It would be beneficial to have an independent moderator for contributed data, who for instance decides what data products are archived. This also facilitates auditing processes. Archiving model output should be subject to policy and review. Some models produce 5-10GB of output data every six hours.
- Acknowledgements of data provenance ("who produced data, provided the data product") help to determine data ownership questions throughout the different uses of the same source data.
- Satellite data for instance are typically provided in HDF. The CI needs to provide conversion facilities, for instance through a Matlab layer.

- Variable naming in NetCDF is not common. However, there are standards, and compliance to these standards is required for instance when using the Lab Access Server. The "holy grail" of oceanography is defining standards fast enough to enable reliable automation.
- Tools used include TTide (a Matlab tool), water toolbox, EOF analysis

User interfaces

Data access and visualization:

- The JPL OceanPortal is an example of a good web portal
- Other examples are TEO-PME and CCAR UC Boulder, which is very simple
- It should be possible to deliver field science information to classrooms vs. lab science operations
- The current way of data representation is not too user-friendly. There should be better ways to present it to different communities.
- This observing system will compete with other projects, so you want many people to use it and therefore need good visualization and user interfaces.

Further concerns

Privacy concerns, security:

- Projects have designated data managers that would interact with the CI. The CI needs to provide interfaces for data managers to define privacy and access policies.
- Projects have made extensive investments in database formats, for instance as XML standards, MetaCAD. The CI needs facilities to provide adapters to such data formats.
- Institutional system administrators will be able to provide more information about security and policy. Use standard best practices for the development of the CI.
- It is often not a problem to share computation power with the community; however the policy of the sharing institution needs to be fulfilled.

Operations and maintenance:

- The SDSC ROCKS group provides good examples for operating and maintaining grid clusters.
- The distinction between recomputable data and home data is crucial. Data trees can be regenerated. Home trees needs to be backed up.
- In existing projects, there is mostly only backup of the models. The input data are backed up by the data providers; any downloaded subsets can be used and deleted afterwards.

4.5 CI Use Scenario Definition

In this session, the charge for the workshop participants was to brainstorm and discuss a hypothetical use scenario for a transformative community cyberinfrastructure. The following list documents this use scenario:

Project Preparation:

- A PI writes a proposal
  - Use Endurance Array data
  - Study the research question "what is the reason for a low oxygen region"
  - He identifies three hypotheses that need to be tested
- The PI gets funding and the project starts.

Project Definition:

- The PI is a registered user of an assumed hypothetical CI

- He opens his web browser and accesses a CI portal for PIs.
    - Creates a new project
    - Invites collaborators
    - Joins existing social networks for the community and related topics
- He defines a new CI project workspace for numerical model analysis, etc.

Research Phase:

- The PI performs a comprehensive review with resources provided by the CI
    - Finds similar research questions and answers
    - Surveys the historical database and existing model simulations to see if the scenario of interest has been investigated
        - For instance there is a browser available with access to all NASA databases
    - Finds relevant publications
    - These activities can be done during the proposal writing process
- Use collaboration tools
    - Within the social networks, asks for ideas, available data sources, models, etc.
    - Asks others to share data and models with the group
    - In his own workspace, a link is added to other projects
- The search results enable the PI to eliminate one hypothesis; two remain to be tested.
    - Two subsequent alternative approaches for testing hypotheses are available. One is to study model output to investigate the historical data. The other approach is to look at the observations and detect something.

Scenario I – "Test the shelf productivity hypothesis" (Numerical model analysis)

- Use the CI to find the relevant variables in the domain and what data products exist
    - Within the OOI, find data from the cabled observatories, gliders…, and other data that are available, and create a time series for the spatial & time domain he is interested in.
    - Use the CI to find links to external resources (NOAA, NASA satellites)
    - The CI shows available resources that potentially require authorization or impose certain use policies
    - Import and group data into the previously defined model workspace
        - For instance, after initial data retrieval, 10GB of data are now in the model workspace
- Search for models that provide the desired result. What model outputs are already available in the CI?
    - The OOI repository contains several data sets with models run for the selected input data.
    - For instance, there is a global model output data set available and a regional-coastal model output with a finer resolution but not exactly with the desired variables.
    - Model output size can be on the order of 1 TB
- Analyze the found model output
    - Use the OOI community toolbox to find suitable data analysis tools
    - Finding these tools requires the PI to enter a number of characteristic keywords
    - First, the PI performs an interactive analysis process with random ("poke around") plots, visualizations, tool applications etc. This is not really a defined workflow. The CI supports this, for instance by keeping a history and a stack of the commands used and keeping results and tools in the workspace.
    - This interactive analysis occurs iteratively, with refined analysis and searches. It realizes a spiral development cycle. The CI supports iterative analysis.
    - The CI also provides data services for statistical analysis and comparison of results and data series. The CI enables the scientist to investigate a region much faster.

- Publish the analysis to the CI community
- Further subsequent alternative research paths exist:
    - Run existing models with different data and parameterizations
    - Design observations and use the resulting data with the models

Option 1: Rerun models; create different model output using available models and data

- Situation: "I don't like any model output so far and want to combine a model with existing data "
- The PI goes to the data assimilation portal.
    - Looks at what models and data sets are available.
    - Retrieves desired model code and original configuration into the modeling workspace, or gets access to the model front-end server.
    - Selects archived data set as input to the model
    - Configures and parameterizes the model
    - Selects from available model extensions and applies them to the given model (similar to an ECMP model component tree)
- Model execution in workspace or externally
    - Use the selected model, model extensions, parameterization and input data sets
    - Schedule computation for execution on CI infrastructure resources
    - The resulting output data could be in the order of 1TB , now in the local workspace
- Model output analysis
    - As sketched before, using CI analysis and visualization toolbox
    - Share selected results by email
    - The CI can provide links to results in the workspace
    - Create an output report from available resources in the workspace
- Cleanup workspace (archive, prune).

Option 2: Schedule observations

- Situation: "I decide I will do my own experiment for a month."
- The PI uses the OOI command and control portal to:
    - Evaluate the available instruments in the OOI network
    - Define a "virtual model" with its inputs and outputs, properties and constraints as well as potential instrumentation.
    - A simulator such as the OSSE (Observing System Simulation Experiment) uses this virtual model and helps to answer the questions:
        - "How many gliders are needed, how many resources need to be deployed and where?"
        - "Supposed I have data on these points, how would the error be? What is an optimal tradeoff between resources and resulting model output quality"
    - Create an observation request
        - Depending on the cost, the PI decides on the list of instruments requested
        - Define the observations in terms of resource needs, application area, etc.
        - Examples of observation requests:
            - For gliders, AUVs and mobile sensors: define sampling grids for ("when and where you want the gliders?")
            - For moored sensors: "When cold water coming on the shelf is observed, provide vertical profile 10 times a day instead of once per day"
- Negotiation of the observation request
    - The PI files the request through the OOI to the marine operator who oversees instrument operations

        - The request could be "I want the glider to spend more time in this area. Please schedule this experiment in the next month."
        - The CI cannot execute this schedule, but facilitates communication with the marine operator
            - For instance, in this experiment, the scientist requests the use of many sensors from the OOI, but does not directly control the instruments. There are 30 requests to change where the gliders are flying, so the marine operators must solve the problem.
        - Scientist and marine operator negotiate an agreement facilitated by the CI
- Execution of the observation request
    - The experiment gets executed and new observations are made
    - Data from the observations is made available on the CI
- Perform data analysis as before
- Perform publication of results as before

Education and Outreach:

- Use the model output and analysis results to create a simple simulation to present the results.
- Create a report from the available documentation, resources and results

5 Science User Requirements

5.1 Requirements Elicitation Process

The requirements listed in the next section represent the current collection of science user requirements for the OOI CI. Some of the requirements were identified during the first requirements workshop in July 2007 (RWS1), and were validated and refined by the participants of the second requirements workshop (RWS2). The remaining requirements were identified through a thorough post-workshop analysis process. Requirements were either directly stated by the participants during the workshop discussions, called out in the participant questionnaires or inferred through a requirements analysis process by the CI architecture and design team. Requirements are grouped into categories and formatted according to a template as described below.

In order to uniquely identify the elicited requirements, each requirement in this report follows a standard template. Each requirement contains a unique identifier: [RWS2-Rn] for new requirements of the second workshop, [RWS1-Rn] for refined and validated first workshop requirements. Furthermore, each requirement contains a label and an explanation section. Requirement labels are constructed in a schematic way.

The listed requirements strive to be atomic (i.e., they cover one requirement statement only and do not contain sub-requirements). However, requirements might be related and one requirement might be influenced by another requirement. Also, the explanations might contain further details on the requirements.

The workshop participants validated the requirements documented in the first workshop report [CI-RWS1]. Each requirements category in the next section contains first a list of RWS2 requirements and subsequently a list of the refined RWS1 requirements. While the requirement labels might have changed, the requirement identifiers remain the same. For easier distinction, RWS1 requirements are formatted in italics in this report only. They are of the same quality as RWS2 requirements and together form the list of CI science user requirements. RWS1 and RWS2 requirements are intended to be non-redundant and non-overlapping to the highest degree possible.

All listed requirements have been vetted through the participating user community in a post-workshop agreement process.

5.2 CI Science User Requirements

This section contains all science user requirements form the first and second requirements workshops. Requirements are grouped into the following requirements categories:

5.2.1 Resource Management

This category contains requirements related to the management of CI governed resources. This covers in particular the resource life-cycle, resource registration, resource catalog etc.

[RWS2-R1] The CI shall notify registered users and applications when new resources are added to the system.

> Explanation: New resources, such as physical resources, sensors, instruments, but also computational and storage resources as well as data products and processing tools can be made available to the OOI user community at any time. CI authorized users shall be able to register to receive notifications of new resources that match their requested characteristics. The CI shall provide notifications to these users when such resources become available.

[RWS1-R3] The CI shall be extensible to allow the addition of new resources and applications to the OOI infrastructure.

> *Explanation: New proposals and grants lead to new and updated hardware in existing observatories as well as to new observatories. The CI shall be flexibly extensible to accomodate such resources.*

[RWS1-R9] The CI shall provide a catalog listing all resources under CI governance.

> *Explanation: A catalog provides references to the cataloged resources and further descriptive information and metadata. The catalog shall not be restricted to resources of certain types or characteristics. The CI shall provide unique identification for all resources under CI governance, including physical and information resources, in their different variants and versions. The CI shall provide pointers from entries in the resource catalog to the resource subject and all associated descriptions and metadata.*

[RWS1-R9A] The CI shall enable users to discover observatory resources together with their metadata based on resource characteristics and user-defined search criteria.

> *Explanation: A catalog enables the discovery of previously unknown resources using standard criteria, such as name, type, make, quality-of-service, version etc. The catalog provides references to the cataloged resources and further descriptive information and metadata. Selection criteria apply to resource descriptions, metadata, parameters, locations, observatories, etc. Discovery covers resources connected to the OOI observatories, as well as user-provided electronic and data resources.*

[RWS1-R11] The CI shall catalog physical samples in the CI resource catalog.

> *Explanation: Physical samples refer to biological, chemical or geological samples retrieved from the seafloor or water column, for example during an expedition. Some physical samples are collected within OOI observatories but not analyzed in it; other samples are collected outside of OOI observatories but should be available to OOI users. Cataloging physical*

samples in the CI requires that metadata be associated with these samples. This capability facilitates reaching out to many communities.

[RWS1-R12] The CI shall support cross-referencing from CI governed resources to external resource catalogs and metadata.

Explanation: Resources that are under CI governance, whether part of OOI or not, can also be listed and registered in external resource catalogs. Metadata about resources can be available at external locations. The CI shall facilitate cross-referencing such external catalogs and metadata locations from CI resources and CI catalog entries. This enables full resource information and cross-referencing availability from within CI interfaces.

[RWS1-R16] The CI shall bind metadata to all resources under CI governance throughout the resource life cycle.

Explanation: CI governed resources shall have metadata descriptions from inception to removal. This requirement does not specify the metadata format or content.

[RWS1-R18] The CI shall provide standard OOI metadata descriptions that include, but are not limited to, a complete description of resource behavior, content, syntax, semantics, provenance, quality, context and lineage.

Explanation: Metadata provide descriptive information about any kind of OOI resource. Metadata standards will be externally imposed since the OOI is federally funded, but the OOI standard will probably need to go beyond them. The term behavior refers to the inherent characteristics of a resource (such as the range of sample rates that an instrument is capable of). The term content refers to the characteristics of any externally presented information provided by a resource (for instance what an instrument measures, including calibration information). The term syntax refers to a model for the resource content based on structure. The term semantics refers to a model for the resource content based on meaning. The term provenance refers to the resource origin, e.g., how and by whom data were collected. For data products, this identifies the sensor and instrument platform where the data originated. The term quality refers to information on the QA/QC status of a resource. The term context refers to information about resource usage (such as the geographic location of an instrument). The term lineage refers to information about the evolution of a resource (such as versioning of data due to QA/QC.

[RWS1-R19] The CI shall allow the discovery of all information resources that are based on a given original information resource.

Explanation: For instance, it shall be possible to discover all distinct data streams that are based on the same instrument source with possible differences in sampling rate, quality of service parameters, metadata annotations or applied post-processing algorithms. This is useful when alternatives to a given data product need to be found, for instance because one becomes unavailable. This requirement also applies to finding all models and their output that are based on a given input data source.

[RWS1-R20] The CI shall provide information resource subscribers automatic and manual fallback options with similar characteristics in case the original resource becomes unavailable.

Explanation: In case of temporary or permanent unavailability of a subscribed data stream, the CI shall offer alternatives that are comparable to the original resource, for instance because they are based on the same source data or other characteristics. If desired by the user, this fallback shall be automatic for uninterrupted operation. The term similar relates to

OOI-standard metadata characteristics that pertain to both original and fallback resources. The exact choice of fallback resource selection criteria shall be left to the user.

[RWS1-R26] The CI shall provide notification of resource state change to all resource subscribers.

Explanation: The term state refers to behaviors or characteristics that persist (for instance whether an instrument is on- or off-line, or changes in QA/QC state for archived data, availability of new versions of data). Notification applies to all subscribers of data products.

[RWS1-R33] The CI shall collect and provide resource access statistics.

Explanation: The CI shall record and provide information about access of data products and general resources. These access statistics measure impact in the field and are very valuable for researchers when publishing about such a data product. The CI shall keep track of resource access and usage and provide statistics based on these collected data to interested parties as a data stream or on request.

5.2.2 Data Management

This category contains requirements related to the integrity, distribution, streaming, storage and archival of data resources as a special kind of CI resources. It also covers the manipulation and dissemination of data resources, data streams etc.

[RWS1-R21] The CI shall be capable of archiving all data and data products associated with an OOI observatory or other CI-governed information resource.

Explanation: The decision about which data product and information resourcse should be archived made by an authorized OOI operator and is subject to policy and resource availability. It might be driven by economics. All data products must be archived together with their metadata.

[RWS1-R22] The CI shall support the publication, distribution and archiving of different versions of the same data product.

Explanation: Different versions of a data product may occur due to changes in the QA/QC state of the data product, sensor calibration compensation, filtering, necessary post-processing etc. The owner of a data product may decide to publish an updated version of the data product. The CI shall offer all data product subscribers the new version of the product. Each data product shall be uniquely identified with its version. The CI shall be capable of archiving all versions of the same data product.

[RWS1-R23] The CI shall ensure the integrity and completeness of all data products throughout the OOI life cycle.

Explanation: The CI shall ensure the integrity and completeness of all data products throughout the entire OOI life cycle, no matter which transformations and archival processes the data products undergo. This encompasses the requirements to verify that archived data accurately reflect the original, and that archived data are protected from loss due to media degradation or technology change.

[RWS1-R24] The CI shall ensure that all archived data products can be restored in their complete and most recent state.

Explanation: The archiving process must ensure that all to-be-archived data get archived completely and immediately so that the most recent data are always archived. The restore

process must be capable of returning a data product from any persistent distributed CI storage, for instance in case of failure, in a complete and most recent state.

[RWS1-R30] The CI shall publish new data products resulting from processing of existing data products.

Explanation: New data products shall be publishable by users without functional restrictions, subject to policy. This includes data streams containing filtered, processed, aggregated data as well as model computation and simulation output. Such computed data streams shall be treated in the same way as their input data products and should have similar properties, including unique identification, catalog entry, meta-data etc.

[RWS1-R31] The CI shall enable users and applications to subscribe to information resources in the form of data streams.

Explanation: Users and applications can subscribe to any kind of information resource subject to the relevant policies. The CI shall be responsible for keeping track of the state of the data delivery and provide buffering and retransmission capabilities if needed. Data delivery shall be immediate when new information becomes available, as desired by the user. Provided information includes scientific data but also resource state changes and notifications of the availability of new data product versions, etc. Data streams are similar for unprocessed raw data and for processed and aggregated data.

[RWS1-R47] The CI shall provide a topic-based (publish-subscribe) data distribution infrastructure that supports real-time and near real-time delivery, guaranteed delivery, buffering and data streaming subject to resource availability.

Explanation: The CI data distribution infrastructure shall provide communication capabilities with different qualities of service to authenticated users and applications based on requests and available resources. In general, interfacing with the CI shall occur in publish-subscribe or register to receive styles. Quality of service of communication resources can be requested based on available resources and policy. (Near) real-time delivery, in this context, refers to minimum delay commensurate with the latency on the channel. Guaranteed delivery refers to storage of a message until an acknowledgement of receipt is received. Buffering refers to storage of a message pending receipt of an explicit request for it (pull mode). The term streaming refers to asynchronous, continuous transmission.

5.2.3 Science Data Management

This category contains requirements related to the ingestion, transformation, annotation and metadata description of science data resources. This category covers the domain specific aspects of these data resources.

[RWS2-R2] The CI shall interface with, ingest and distribute data from external data sources, databases, and data distribution networks of related scientific domains.

Explanation: Oceanographic data analysis and modeling often relies on multiple data products and interdisciplinary oceanographic data sets, for instance as boundary conditions in numerical models. Some of these data products originate from non-OOI data sources, such as atmospheric data, IOOS data products, Argo, NASA, Codar etc. Other data products are provided by local servers within the institutions of OOI users. Ingestion of external data products requires interfaces and potentially agreements with data sources and data distribution networks where the data products are published. All post-processing tools available for use with CI data products shall be applicable to such external data products. The CI shall

provide flexible capabilities for any CI user to add new external data sources to the OOI network. For instance, projects have made extensive investments in database formats, such as XML standards or MetaCAD. The CI needs to provide adapters for such data formats.

[RWS2-R3] The CI shall provide interactive and automated data quality control (QC) tools.

Explanation: Science data QC refers to the process of analyzing and post-processing raw data streams in order to assure accuracy and quality of the resulting published data product. The QC process involves manual steps that require the judgment of scientists or engineers and the execution of corrective actions. The CI shall support this process of interactive analysis and processing of data products and data streams. The CI shall also support an automated QC process, for instance by providing means, strategy and policy to filter data and to define which data should be filtered, and workflows that apply QC automatically to data streams.

[RWS2-R4] The CI shall provide standard and user-defined methods to assess the quality of data.

Explanation: Researchers might be interested in the maturity of data. Maturity refers to the quality of a data product, its accuracy and integrity, and the number of QC steps that have been applied. The CI shall define a standard that enables an assessment and ranking of data sources according to their maturity, for instance by analyzing provenance and lineage information. For user-defined quality assessment, the CI could help to share metrics that contain the assumptions and the knowledge about the data and quality of the models and input data over several years.

[RWS2-R5] The CI shall facilitate the moderation and auditing of published data.

Explanation: In an automated system such as the CI, any authorized user can publish data products independent of their quality and suitability. In order to maintain high quality levels for data products, it is important to facilitate oversight and manual QC of data subject to system-defined policies. The CI shall facilitate such policy compliant oversight processes by independent authorized parties that assess and rate new and existing data products and resources according to defined quality standards.

[RWS2-R6] The CI shall act as a broker for CI-managed data products.

Explanation: The CI shall ingest data products and provide access to these products on demand. The CI shall provide universal data delivery.

[RWS2-R7] The CI shall provide access to CI-manage data products in standard formats and subsets.

Explanation: The CI shall ingest data products and provide access to these products in various formats for any requested subset of the data. For instance, numerical models produce large data sets. In case the output of a numerical model is registered for publication as a CI data product, the CI shall handle (local) storage of it and make relevant subsets available to interested parties.

[RWS2-R8] The CI shall act as a broker between information and processing resources.

Explanation: The CI shall establish the facilitating element that connects and binds different resources, for instance models and models, observations and models, observations and observations etc. The CI shall facilitate binding models and required input data products,

whether observational data products or model computed nowcast and forecast products. This binding might require reformatting or partitioning of data products. One particular application is nested numerical models.

[RWS2-R9] The CI shall make unprocessed raw sensor data available on request.

Explanation: Typically, processed and QC'd versions of data products are made available to the public. The filtering and corrective actions applied to raw data are at the discretion of some scientists. Other researchers might be interested in the original raw data streams as produced by the sensor, for instance in order to apply different corrective actions. The CI shall keep raw sensor data and make them available on request, in addition to any version of processed data products based on the same raw data.

[RWS2-R10] The CI shall track data provenance and correspondence.

Explanation: Data provenance refers to the origin and history of data products, i.e. it identifies the creator and the related sensors, instruments, the filtering and processing that has been applied, model identification and parameterization etc. Correspondence refers to a statement of association between two or more resources.

[RWS2-R11] The CI shall credit data publishers when data products are accessed.

Explanation: The CI shall make provenance information available, for instance in resource descriptions and metadata and credit data publishers each time the data product is used, e.g. in the way of a citation index. The term citation refers to statements about the use (including its outcome) of a given resource by another resource or actor. This creates incentives for publishing data to the CI instead of keeping data for private use only.

[RWS2-R12] The CI shall create and distribute related data products from a given source data product that have different characteristics, such as resolution, level of detail, real-timeform and quality,.

Explanation: For instance one model run can result in output data that can be published as different data products which differ in resolution, manual quality control etc. A real-time, low resolution, unprocessed data product can be published immediately and without manual interaction, while a high resolution, quality controlled, post-processed data product based on the same input data source may be published with latency after manual interaction.

[RWS2-R13] The CI shall flag data stream state change.

Explanation: For instance, a sensor may provide a data stream for a set of measured variables. Remote sensing can undergo periodic revision due to improvements in the processing algorithm. There may also be ongoing calibration and validation efforts. Assimilating consistent data with non-consistent data produces inconsistent data as output. Therefore, subscribers to data products need a mechanism to be informed about changes to resources and to distinguish subsets of data with different characteristics.

[RWS2-R14] The CI shall support the provision of complete metadata by users.

Explanation: Metadata are essential to describe and interpret the meaning of data and resources. Metadata standards can be very complex to understand and providing metadata can be a daunting task for scientists and resource providers. The CI shall support the process of metadata definition as completely as possible through easy-to-use interfaces, on-demand assistance, detailed explanations, guidelines, examples, consistency checks etc.

[RWS1-R4] The CI shall support a standard set of data exchange formats.

Explanation: Interfacing with the CI requires compliance with CI interfaces. The CI shall provide a number of pre-defined data formats that will be compatible with the CI, either as a data producer or as a data receiver. Currently predominant formats in the community include NetCDF for data and OPeNDAP for data exchange, and hence these technologies shall be supported by the CI.

[RWS1-R4a] The CI shall translate between the standard data exchange formats without loss of information.

Explanation: The CI shall provide translators between the standard data formats, for instance based on shared ontologies.

[RWS1-R5] The CI shall allow the addition of user-defined data exchange formats and translators.

Explanation: The CI shall provide interfaces that allow the definition of user-defined data formats and execution or connector facilities for user-provided translators that can convert user-defined data formats into CI standard format.. This also applies to sensors and data sources that provide proprietary raw data formats that need to be connected to the CI network.

5.2.4 Research and Analysis

This category contains requirements related to research and analysis of science data through the CI.

[RWS2-R15] The CI shall provide capabilities and user/application interfaces for researching scientific materials and OOI-governed resources across disciplines.

Explanation: The study of a scientific problem involves the use of numerous resources external and internal to the OOI, such as CI-governed data sources, data products, publications, numerical models, model configurations, model output data, instruments, communities and so on. The CI shall support this research process by flexible search, cross-reference, collocation and comparison services and interfaces. For instance, the CI could provide a browser with access to all NASA databases. Flexible research capabilities across disciplines would benefit students and research grant proposal writers in finding related work and existing material about a subject.

[RWS2-R16] The CI shall suggest suitable data products, data transformations, observation resources, analysis tools, visualization tools and other OOI resources based on user-specified research questions in domain language.

Explanation: Getting from research questions to possible solutions in terms of available OOI and CI resources can be a complex and tedious task, in particular for generic, multi-disciplinary research questions. The CI shall support this task to the greatest degree possible by enabling users to state research questions in domain language, and then suggesting available resources that can be part of the solution. This decouples observation request definition from resource matching to execute a job. For instance, a scientist studies a phenomenon in the Monterey area. Relevant questions to answer include “are moorings there?”, “are any ships passing by?”, “are any gliders in the area?”, “are any satellites measuring something?”, or “are any models that some scientists run in the area?”. This capability will realize a transformative change for the community and beyond. It realizes a transition to a service way of providing information. For instance, ecology researchers will be able to get answers to questions about marine life that currently require searching massive amounts of data for very lit-

tle information. For numerical modeling, the CI shall provide decision support for selecting boundary condition input data sources for numerical models. The CI shall also enable ranking of observation sensitivity in respect to a given model.

[RWS2-R17] The CI shall support interactive and iterative analysis and visualization through infrastructure, tools and user interfaces.

Explanation: Analyzing observation and model data is a complex and highly scientific process. It requires specialists with substantial knowledge about the domain, literature, and technology. The CI shall facilitate the analysis of data by providing the means to perform user-driven interactive analysis. All applicable analysis and visualization tools shall be available to the analyst, with efficient ways to configure and run them on the available input data sets. The turn-around time to change and rerun analyses with different parameterization shall be as low as possible. The CI shall keep track of the sequence and configuration of analyses, and of the resulting outcomes, and provide these to the user. Iterative analysis relates to the concept of re-running similar analysis steps with refined data sets and parameterization to optimize the resulting output.

[RWS2-R18] The CI shall provide tools, user interfaces and visualization for the analysis, combination and comparison of disparate, heterogeneous data sets..

Explanation: Answering multi-disciplinary research questions often requires the combination, assimilation and comparative analysis of data set with diverse provenances using different formats and characteristics. Data sets are may be located on different networks with different ownership. The CI shall enable bringing these data sets together and transforming them into comparable representations, for instance by co-location, re-gridding, semantic transformation, etc. This requirement implies the existence of effective infrastructure, tools and user interfaces. For instance, the CI shall provide data services for statistical analysis and comparison of results and data series.

[RWS1-R25] The CI shall provide a standard, extensible set of data product processing elements that provide data assimilation, alignment, consolidation, aggregation, transformation, filtering and quality control tasks.

Explanation: The standard processes will be defined through a community decision process. Such data processing elements can include assimilation of various data products, elimination of redundant entries, spatial interpolation, collocation of data sets, merging multiple compatible data products into one data stream etc. Further standard data manipulation tasks for model integration and combination can include re-projection, re-gridding, subsetting, averaging, filtering and scaling. The set of processing elements shall be extensible; as standard OOI processes, as well as by users with their own processing elements. Such processes shall be able to use data product meta-data for automated processing, such as alignment of data geographical grid points based on resource meta-data when combining multiple models outputs.

5.2.5 Ocean Modeling

This category contains requirements related to modeling of the ocean, for instance as numerical modeling, as one form of processing observational data products.

[RWS2-R19] The CI shall enable the efficient configuration, execution, debugging and tuning of numerical ocean models.

Explanation: Currently, the learning curve for numerical models is substantial and requires technical knowledge, expert users and extensive specific model knowledge. Making numerical models accessible to non-experts is an important step in reaching a broader audience. Models, and their parameterizations, input data assimilations and output visualizations are typically developed over extended periods of time by groups of experts and require substantial technological and scientific knowledge. Ways to improve the handling of numerical models include improved user interfaces that enable the definition of basic model parameters, data archival settings, model output use etc. The CI could increase model configuration efficiency through intuitive GUIs for non-experts that enable them to choose all of the necessary options for a model (see ROMS). These user interfaces could capture expert knowledge and make it available to non-experts, thus flattening the learning curve and improving efficiency. The CI should provide facilities for easy model diagnosis, debugging and tuning. This will especially help non-experts and students. The CI shall also facilitate the provision of pre-developed model environments and documentation to the user community for execution in adaptable contexts. Non-specialists shall be enabled to run models and perform minor model manipulations. Non-specialists include grad students, interested community researchers, and people in education institutions within the community.

[RWS2-R20] The CI shall support the interaction of model developers and non-expert model users.

Explanation: The transfer of knowledge from model developers to model users is very important because model users are typically non-experts for the specific model. The CI shall enable the documentation and self-documentation of models and facilitate the interaction of expert model developers and non-expert model users through sharing parameter settings, easy to use configuration interfaces and pre-configured analysis and visualization tools. The CI shall also provide the communication tools enabling knowledge transfer, such as bulletin boards, discussion forums, wikis, public commenting features etc. With the existence of a capable infrastructure, the community will be very receptive and there will be experts who will provide comments and contributions.

[RWS2-R21] The CI shall provide facilities to develop and tune numerical models and their parameters.

Explanation: Besides the numerical model algorithm, an important step in developing an effective numerical model is finding an optimal parameterization. This requires many model runs with different parameter values and comparison of the results with actual observations, for instance by comparing hindcast simulations with real world observations. The CI shall support this process by enabling parameter optimization of CI managed models and provisioning of suitable analysis tools.

[RWS2-R22] The CI shall provide a virtual model environment and simulator to determine optimal model inputs, parameterizations and outcome qualities.

Explanation: The outcome of a numerical model is constrained by the quality and availability of the input data. Determining what the outcome quality will be with given input data and which changes of input data will lead to more optimal model output are important steps in the overall observation and analysis process. In an adaptive observation environment, this can lead to optimized measurement schedules for deployed sensors, such as gliders, AUVs etc. A simulator such as the OSSE (Observing System Simulation Experiment) uses such virtual models and helps to answer the questions.

[RWS2-R23] The CI shall enable the sharing of ocean modeling, data assimilation and visualization components, including the extension of models with new model components.

Explanation: Numerical models and related data transformation processes can be constructed either sequentially or hierarchically out of individual model components. The extension of an existing model with an extension component is a special case. The user shall be able to select from all available model extensions and apply them to a given model. Modeling and data assimilation process definitions should be as portable and easy to use as Matlab subroutines.

[RWS2-R24] The CI shall provide a repository and sharing capabilities for numerical model algorithms, model configurations, data processing tools and documentation.

Explanation: The availability of a central repository and archive for instances of all resources related to numerical model development and execution with a uniform access interface and search capabilities is of high value to the community. The CI shall provide such a repository in the form of a community-managed clearinghouse, available to the user community for any kind of modeling resource. The CI shall not restrict this repository to models managed by the CI or executable by the CI. For instance, a user can use the forcing fields that an expert adjusted and shared, and use them in his/her own model, which may lead to improved accuracy. Other shared artifacts include error covariances for forcing, background, boundary conditions, and observations.

[RWS2-R25] The CI shall archive numerical models under configuration control.

Explanation: The CI shall provide the means to archive the model execution workflow with a level of completeness necessary to reconstruct and rerun the model in the future. This requires packaging, configuration control and documentation effort on the side of the initial model developer, which shall be supported by the CI. The CI shall archive the binding of data sources, model parameterizations, model compile and execution environments, model execution workflows.

[RWS2-R26] The CI shall recompute model data products using archived models and workflows.

Explanation: The CI shall provide the means to rerun archived model workflows for any suitable past input data set. This enables recomputation of model output for past archived data sets. This makes later exact model reruns possible and avoids storing computed data.

[RWS2-R27] The CI shall enable the modification of archived numerical models and and workflows.

Explanation: The CI shall provide the means to reuse archived models and workflows and modify them by any interested, sufficiently knowledgeable party, at any time and any location. This includes execution of such models but also adaptation and modification of parameters, input data sources, pre- and post-processing and model algorithms based on the published model, subject to policy.

[RWS2-R28] The CI shall provide an environment for the development of community numerical models under community process support.

Explanation: Within the research community, there are often numerical models and standard data analyses that become widely used and accepted community standards. The CI shall facilitate the establishment of such standard models and data processing with support for

community processes such as communication, commenting, voting, moderation, submission of change requests, documentation etc.

[RWS2-R29] The CI shall provide a non-restricted environment for the development of independent numerical models.

Explanation: In addition to community-accepted standard models, the research community also requires support for emerging independent research models. The CI shall facilitate such research and development by providing an environment that enables these efforts; it should be non-restrictive by not requiring use of prescribed standard data sources, processing tools, output formats, model algorithms etc. The CI shall facilitate making the output of such independent models available to the user community in a similar manner to CI standard and community models.

[RWS2-R30] The CI shall support nesting of ocean models at different geographical scales.

Explanation: Ocean models typically exist at different spatial resolutions, covering different parts of the ocean surface. The CI shall support the combination of global, regional, and coastal models. The CI shall support nesting of numerical models such that smaller scale models can automatically use larger scale model output as boundary conditions within a selected geographical range or model domain. This requirement has implications for user interfaces, model configuration and model execution scheduling.

[RWS2-R31] The CI shall provide a framework for the adaptation of model resolution to the available resources.

Explanation: The number of computed model grid points – the resolution of a model – is typically limited by the available computational and storage resources, for instance through the time it takes to compute a model ensemble on the available hardware. For a given CI managed model, the CI shall facilitate an increase in the resolution and the number of grid points when further computational and storage resources become available, configurable by the user that executes the model.

[RWS2-R32] The CI shall support model ensemble definition, execution and analysis.

Explanation: A developed and parameterized numerical model provides output that has a comparable quality to the available input data. All input data have uncertainty associated with them. An effective way of improving model reliability is to run ensembles of models with slightly modified initial and boundary conditions. Analysis can then take the model ensemble outputs into account, for instance by averaging or by computing standard deviation values. The CI shall support model ensemble definition, ensemble execution and ensemble analysis. The CI shall facilitate model ensemble execution optimization, for instance one pass computation of entire model ensembles.

[RWS2-R33] The CI shall publish both elements and aggregated ensemble data products.

Explanation: The CI should make the entire ensemble and individual model runs available to the user community with clear descriptions of the parameters and conditions.

[RWS2-R34] The CI shall support flexible high performance model execution.

Explanation: Performance and flexibility are a classic trade-off in numerical model development. Economies of scale suggest that model execution flexibility requires higher initial development and runtime performance costs. Currently, the community is not willing to ac-

cept any impact on runtime and network latency for more flexible models; the requirement is optimization of the number of model runs and output grid resolution, but not portability, etc. The CI shall address this concern while at the same time providing means and mechanisms for the development of more flexible, portable models. This could include standardized virtual execution environments, more user-friendly parameterization interfaces, harmonized data input and output formats, and better documentation facilities.

5.2.6 Visualization

This category contains requirements related to the visualization of data analysis products and ocean modeling output.

[RWS2-R35] The CI shall provide a uniform and consistent for numerical model output visualization and analysis in 2D, 3D and 4D.

Explanation: Currently, visualization of numerical modeling output is very heterogeneous across the ocean modeling community. Commonly used visualization tools include Ferret, GMT, IDL, Matlab and the NCAR package for ROMS. The CI shall provide a basic set of visualization and analysis tools for CI-managed numerical model output, with uniform and consistent input and output interfaces. The current lack of user-friendly 4D visualization tools is a concern to the community – this is an area where a consistent CI interface can provide a substantial benefit. Examples of basic analysis tools are: time series, special maps, cross-sections etc.

[RWS2-R36] The CI shall provide interactive visualization of the 3D and 4D ocean.

Explanation: The community currently has no standard way to visualizae 3D ocean data, and especially interactively. Interactive refers to the concept of the user navigating through the available data in real-time, for instance in the way GoogleEarth provides this functionality for geographical information. The CI shall provide mechanisms to integrate 3D visualization tools and make them available to all users. The tools shall be applicable to any 3D or 4D ocean data set.

[RWS2-R37] The CI shall support the integration of external visualization and analysis tools.

Explanation: New analysis and visualization tools and technologies become available all the time. The CI shall provide interfaces to extend its set of available tools with new ones that are subsequently available to the OOI users. External visualization tools include GoogleEarth. Analysis tools can easily be added if the CI provides a Matlab interface.

5.2.7 Computation and Process Execution

This category contains requirements related to the use of computational resources and the execution of data manipulating processes within the CI.

[RWS2-R38] The CI shall support the execution of large scale numerical ocean models across different locations on the network.

Explanation: Some numerical models require extensive computational resources to run. The CI shall facilitate running such models with CI-provided computational and storage resources, and coordinate resource assignment and communication subject to policy and available resources.

[RWS2-R39] The CI shall support workflows for automated numerical model execution, including just-in-time input data preparation, model computation, output post-processing, and publication of results.

Explanation: Running a numerical model repeatedly requires a number of steps to be carried out just-in-time within a workflow. This includes retrieval of data from sources, data assimilation, resource allocation, model or model ensemble computation, data post-processing, quality control, visualization and publication of results as data products. The CI shall facilitate such just-in-time executions by providing the means to define such workflows with suitable robustness, failure-tolerance and resource awareness.

[RWS2-R40] The CI shall enable the one-time and recurring execution of numerical models on any networked computational resource with quality-of-service guarantees based on contracts and policy.

Explanation: The CI shall provide the means to define, develop, package and schedule executable numerical models with their associated workflows such that they can be run flexibly on any networked computational resource with sufficient capacity. This execution shall be possible during the development of the models with high flexibility and variable resource and quality-of-service requirements, as well as during the production phase with negotiated resource reservations and quality-of-service contracts. Quality-of-service parameters relevant for model execution are for instance in-time resource availability, maximum execution time and maximum delay until model run results are available. All CI guarantees shall be subject to policy and resource availability, but shall respect advance long-term resource allocation contracts. For instance, the CI could provide a priority scheduling system for computational resources. For the community, unpredictable latencies in job processing are prohibitive for the use of these resources in production mode. The user shall be able to schedule model runs for automatic, recurring executions based on long-term resource and quality-of-service agreements.

[RWS1-R27] The CI shall provide uniform and easy-to-use interfaces to computational resources with varying characteristics to define executable processes.

Explanation: Currently, there is no common interface to define processes (jobs) that should be executed on remote computational infrastructure, such as clusters and grids. In particular, larger scale computational infrastructure requires specific job definitions and management of the job execution. This requirement implies the existence of documented tools and interfaces to define, develop, configure, schedule and execute user-defined CI executable processes and general taskable resources. The CI shall provide uniform interfaces to all forms of computation with varying resource requirements, from small scale computations on embedded devices to large scale persistent Grid computations. A process is a sequence of human or machine executed steps that are applied to a data stream or data set. This includes data cleaning, filtering and aggregation as well as numerical modeling algorithms. The CI shall provide and document all that is necessary to define and run such processes in a controlled and repeatable way and as an extensible framework.

5.2.8 Sensors and Instrument Interfaces

This category contains requirements for sensors, instruments and interfaces to such resources as one responsibility of the CI infrastructure.

[RWS2-R41] The CI shall provide flexible and reliable access to remote resources.

Explanation: The availability of remote controllable resources is becoming more and more common. The CI shall support the management and control of remote resources, such as instruments, as well as computational and storage resources with effective capabilities. This includes providing flexible access to resources as well as reliable communication. The CI shall support the operation and maintenance of remote resources; this helps to significantly reduce maintenance costs and improve overall system reliability. The CI shall diagnose problems and automatically log events (e.g. outages, equipment failures, etc.) at remote sites.

[RWS2-R42] The CI shall provide real-time monitoring of remote sensors.

Explanation: For instance, episodic ocean events can occur at unforeseen times and locations. In order to capture such highly scientifically relevant events in areas with deployed sensors, it is necessary to sample and analyze remote sensor data on a continuous basis in (near) real-time, so that events can be detected immediately and adaptive observation actions can be scheduled, such as AUV and glider deployments and adapted sensor calibrations and measuring frequencies. The CI and the OOI infrastructure shall install the mechanisms to bring remote sensor data through the network to the scientist in near real-time. This includes communication support for acoustic modem and wireless data transmission from sensor to infrastructure.

[RWS2-R43] The CI shall provide continuous collection of scientific data during extreme weather events.

Explanation: Extreme weather conditions often interrupt or prevent data collection, for instance due to ship scheduling constraints. The CI shall provide the means on various levels to prevent data loss during extreme weather events and to enable uninterrupted collection of data from deployed instruments. For instance, the CI must be capable of handling temporarily disconnected instruments, and include buffering and reliable communications facilities on the wet side of the data link.

[RWS2-R44] The CI shall provide discovery for the number and characteristics of sensors deployed on an instrument platform.

Explanation: Instrument platforms typically host many different sensors. In general, communication, computation and power resources are limited. Additionally, sensors might interfere with each other. The CI shall make such deployment information available to scientists, enable them to track data provenance and determine sensor characteristics and status.

[RWS2-R45] The CI shall support adaptive observation.

Explanation: Adaptive observation refers to the ability to modify sensor characteristics, such as sampling rate, resolution, sensitivity, calibration and position in the case of mobile sensors such as gliders and AUVs. The detection of an episodic event might lead to adaptive observation through the deployment of a fleet of gliders and adjustment of mooring sensor parameters. The CI shall facilitate adaptive observations by providing the necessary capabilities and user interfaces.

5.2.9 Mission Planning and Control

This category contains requirements related to the planning and prosecution of observational missions.

[RWS2-R46] The CI shall provide capabilities and user/application interfaces for mission planning and control.

Explanation: Mission planning refers to definition of observation requests in a domain language to satisfy a scientific purpose. With the support of the CI, observation requests shall be broken down into specific missions plans, including resources and scheduling. Mission control refers to the execution of mission plans, adaptive observations if necessary and further adaptive actions in case of failures or necessary optimizations. The CI shall facilitate and support this process. For instance, the CI cannot execute a detailed observation request, but can facilitate communication with the marine operator that performs the actual scheduling and execution. In this case, scientist and marine operator can negotiate an agreement brokered by the CI.

5.2.10 Application Integration and External Interfaces

This category contains requirements related to the external data and application interfaces of the CI and to application integration into the CI network of resources and services.

[RWS1-R1] The CI shall provision an integrated network comprised of distributed resources, applications and users.

Explanation: A resource is any entity associated with an observatory that provides capability, and includes instruments, data, workflows, networks and more. Applications and users are entities that interact with observatories. All entities together form the CI.

[RWS1-R2] The CI shall enable non-persistent connection of resources, users and applications.

Explanation: The CI provides a network of distributed services and resources that can be temporarily unavailable and impermanently connected. Users and applications should be able to interact with the CI on a regular basis without the obligation to be connected and online. For instance, it shall be possible to perform automated, bulk data stream updates and downloads of subscribed data products with temporary connections to a CI point of presence without loss of the session, configuration and state. Resources are typically only temporarily availably. One consequence of this requirement is the need for data caching and buffering while either a resource or the connected user/application is offline.

[RWS1-R6] The CI shall provide application program interfaces (APIs) to all CI services.

Explanation: APIs are required to integrate external user applications that interact with CI services. This enables full automation of interactions with the CI and goes beyond the availability of user interfaces. The existence of CI APIs enables the development of user- and organization-specific extensions to common CI services and functionality that are seamlessly integrated with the CI infrastructure.

[RWS1-R7] The CI shall provide a synoptic time service with an accuracy of TBD to all resources connected to the OOI observatories.

Scoping: Synoptic time means uniform, global time corrected for network latency and jitter It is made available to all resources and applications. Presuming that they are appropriately time-stamped, data products can be aligned based on consistent time information.

5.2.11 Presentation and User Interfaces

This category contains requirements related to the public appearance of the CI, for instance in the form of user interfaces.

[RWS2-R47] The CI shall provide "one stop shopping" interfaces that provide and collocate relevant information regarding scientific research using OOI resources.

Explanation: The user interface shall leverage the services and resources of the CI infrastructure and provide the user with the collocation of all relevant information. Typical scenarios include the search for related work and for applicable resources. The CI shall present all information that matches a given search query in one spot and thus enable rapid comparison and decision support. For instance, the CI shall enable the discovery of all kinds of data products in one place, e.g. temperature, salinity, wind, or multiple satellite data sources for one region.

[RWS2-R48] The CI shall provide annotation, commenting, ranking and rating services for resources.

Explanation: This applies to all CI resources, from sensor data sources to computed data products, computational and storage resources etc. These services shall facilitate community communication, knowledge management, search-and-retrieval, selection and decision processes. For instance, third party users shall be able to rate the quality of CI data streams and sources. Other users can make use of the aggregated rating information when selecting data products.

[RWS2-R49] The CI shall provide project and user workspace capabilities and user interfaces.

Explanation: Workspaces refer to storage, presentation, archiving and cross-referencing capabilities that apply to project or individual users. Project workspace management refers to the definition of projects and project workspaces as well as project collaboration links and inter-project collaboration. Collaboration refers to mechanisms that enable the communication between members of different projects, for instance notification mechanisms in case new data products become available. A project workspace can contain links to other projects. Data can then be imported into the project workspace and structured as required.

[RWS2-R50] The CI shall provide long-term and ad hoc social networking and collaboration capabilities.

Explanation: Social networking refers to the establishment of contacts between individual users, communities, projects, or interest groups based on personal contacts, and the facilitation of communication within that network. Communication can occur through e-mail messages, instant messages, bulletin boards, discussion forums, commenting of resources etc. Typically, a social network is established for long-term use. The CI shall also support ad hoc communities that form because they utilize specific resources, projects, episodic events etc. The CI shall support the sharing of data and resources between users and projects. For instance, a group may wish to publish their analysis results to the wider CI community and define links to results in their own workspace. User shall be able to define links between data sets.

[RWS1-R34] The CI shall provide homogeneous, intuitive, easy-to-use web-based interfaces to all CI services and resources.

Explanation: Interactive user access to the CI services and resources shall be provided through comprehensive web-based user interfaces. Such interfaces can be provided by a portal site to the CI and the resources of the OOI observatories. The web-based user interfaces shall support all interactive CI functionality, from resource management to mission planning and control to data analysis and visualization, and provide a typical web-based browsing experience, with user sessions, profiles, customization etc. The number of clicks to get to the desired information shall be as few as possible (3 clicks or less). Uniform and homogeneous user interfaces provide efficient interaction and usability

[RWS1-R35] The CI shall provide the capability to make OOI-standard metadata human readable.

Explanation: Scientific metadata for resources and data products have a sophisticated structure and encoding. Science data product metadata, for instance, include information about the sensors, sampling rate, provenance, applied transformations, content and context of the data and much more. Community standards exist to capture and transport such metadata. It is important for scientists to understand and if necessary provide such metadata in the context of the OOI CI resources. The CI shall provide the means to display all OOI-relevant metadata in human understandable and human processable forms.

[RWS1-R38] The CI shall provide extensible configurable visualization capabilities for selected types of data streams.

Explanation: The CI shall provide standard visualization capabilities for selected classes of CI information resources. The visualization shall be flexibly adaptable to the users' needs, for instance by determining the variables of interest, output refresh rate, resolution, output schemes, area of interest etc. All visualization is subject to resource availability and policy. The list of applicable data stream resources needs to be determined through a community decision process. The CI shall facilitate the integration of additional tailorable visualization capabilities.

[RWS1-R49] The CI shall provide real-time analysis and visualization for data resources.

Explanation: The CI shall enable the monitoring of streamed information resources in real-time with basic visualization capabilities. Real-time data plotting refers to the visualization of all kinds of data products, including model output. All visualization is subject to resource availability and policy.

5.2.12 Security, Safety and Privacy Properties

This category contains requirements related to security, safety and privacy aspects of the CI.

[RWS2-R51] The CI shall provide interfaces to define security and policy for information managers at participating institutions.

Explanation: Authentication, authorization and policy mechanisms and levels are typically defined at the level of institutions that participate in the OOI network. The CI shall make user and application interfaces available to define these settings and to manage policies.

[RWS2-R52] The CI shall provide secure operations.

Explanation: Secure operation refers to OOI and CI capabilities and resources that are only available to authorized users, subject to all applicable policies, while guaranteeing privacy, integrity and authenticity. For instance, this is required to protect resources from damage by excessive use when power is limited. The CI shall also provide mechanisms to ensure availability of resources, prevent any abuse and denial of service, and other intentional or coincidental negative impacts on resources.

[RWS2-R53] The CI shall only permit authenticated and authorized users to access OOI resources.

Explanation: Authentication refers to establishing the identity of a user or application, for instance by exchanging secret credentials such as passwords. Authorization refers to permitting access to resources based on a user's attributes and permissions.

[RWS1-R43] The CI shall provide mechanisms to enforce user privacy policies.

Explanation: Privacy policies will be defined by the OOI contractors and NSF in consultation with representatives of the user community. The CI needs to provide the means to define, update and propagate these policies across the distributed OOI network when required and to enforce and guarantee privacy throughout the infrastructure subject to these policies.

[RWS1-R44] The CI shall enable any authenticated party to share their resources.

Explanation: CI collaborators and participants shall be able to share any of their resources with all authenticated OOI users.

[RWS1-R44A] The CI shall grant or restrict resource access subject to use policy.

Explanation: Sharing of resources with all authenticated OOI users shall respect the defined use policies of the sharing party as well as the applicable overall OOI policies. The CI needs to guarantee resource usage according to these policies, for instance to protect resources from damage, overload and abuse. The CI shall provide individual users with the ability to influence how model output is stored and used. Typically, the basic policies will be set by the OOI operators, and are constrained by resource providers and external entities such as the US Navy. The extent of access depends on explicit resource policies set by the OOI operators and resource providers. This particularly applies to accessing resources discovered in the resource catalog and affects the extent of the linkage provided from the catalog to the resource. All policy definitions are subject to review.

5.2.13 Quality Properties

This category contains non-functional constraints on the CI including quality properties, scalability and maintainability requirements.

[RWS1-R46] The CI infrastructure shall provide services and deliver messages with reliability and accuracy that is comparable to that of distributed Internet applications.

Explanation: The "accuracy of the Internet" is given by the availability, reliability and accuracy of open Internet protocols and RFCs as provided on public and commercial infrastructure. The standard reliability number tremor the Internet is 0.99999 [REF]. The CI shall provide comparable quality of service.

5.2.14 Education and Outreach

This category contains requirements related to education and outreach concerns that the CI needs to support and part of the overall OOI.

[RWS2-R54] The CI shall facilitate the creation of publicly available idealized numerical ocean models with a limited choice of modifiable parameters for educational purposes.

Explanation: The CI shall support the development of educational numerical ocean models with idealized environment assumptions and limited, easy-to-understand configuration parameterization possibilities. For instance, a global climate model could take CO2 level, ice melt, etc. into account and produce idealized climate predictions in pre-defined output formats. The development of such idealized models should be possible on demand.

5.2.15 Documentation

This category contains requirements related to documentation of CI services, capabilities and interfaces.

[RWS1-R41] The CI IO shall make all source code for the OOI CyberInfrastructure implementation and drivers publicly available, subject to applicable licenses.

Explanation: The availability of source code for CI software, interfaces and drivers together with all design documents, API documentations and interface descriptions creates a very transparent infrastructure implementation environment that is open to change, community contributions, third-party assessment, and reuse. All this will benefit the CI adoption process, the availability of third-party CI extensions and drivers and the overall CI robustness and reliability.

[RWS1-R42] The CI shall provide documentation for all components of the CI, including all application program interfaces (APIs) to CI services.

Explanation: The CI is a powerful application integration and data distribution environment with powerful automated services. The availability of extensive and up-to-date documentation of CI services and applications interfaces is the prerequisite for successful third-party extensions and tools and for successful community usage

[RWS1-R39] The CI IO shall provide all documentation in web-based formats.

Explanation: Documentation for all CI components and interfaces shall be available in hypertext format such as HTML, either online on the web or for download. Hypertext formats are one the most effective and intuitive way to date for documenting APIs and user interfaces. They enable online indexing and search if made available online. They enable quick cross-referencing and can be made available offline or in printable page-size formats as well, if needed.

5.2.16 Development Process

This category contains requirements related to the development process of the CI, user involvement, availability of CI development materials and artifacts, documentation etc.

[RWS2-R55] The CI IO shall circulate CI requirements and designs within and outside the OOI community so that comparable infrastructures can adopt them.

Explanation: This will lead to similar capabilities across infrastructures and to effective linkage of data in a transparent way.

[RWS1-R8] The CI shall utilize open standards and open source software to the maximum possible extent.

Explanation: Open standards and software facilitate easy integration of heterogeneous resources and applications with the CI, which increases CI maintainability and extensibility. Open source software also permits user specific extensions and modifications. Often, source code and documentation are publicly available, which facilitates user understanding and proposal of changes to the CI. In some instances, proprietary software packages (such as Matlab) may be used where no open source substitute exists. Open standards for information exchange enable interoperability with large classes of existing data distribution networks, information resources and applications.

[RWS1-R40] The CI IO shall provide a process for submitting and incorporating user-suggested changes to the CI.

Explanation: Users shall be able to submit feature requests, defect reports and change requests to the CI IO. They shall even be able to submit executable extensions and source code

fixes for consideration of inclusion into the core CI services and interfaces. All submissions and extensions and subject to a review process and must comply with CI and OOI policies.

[RWS1-R48] The CI shall provide for the flexible and transparent extension of CI services and interfaces to incorporate user-provided processes, user and application interfaces, applications and resources.

Explanation: The CI provides a core set of capabilities as services through user and application interfaces. Through application interfaces, the users shall be enabled to plug in their custom defined applications, tools, user interfaces and further extensions to the CI. The extension mechanism shall be transparently available to authorized users, such as the owner of these extensions. The mechanism must include checks to ensure compatibility and consistency with observatory policies (such as security).

5.3 Removed and Obsolete CI User Requirements

This section contains recent user requirements that were identified as obsolete or redundant. The requirements listed in this section are thus removed from the list of standing CI science user requirements and are present for tracing purposes only.

Requirement	Action	Justification
[RWS1-R10]	removed - subsumed	merged into [RWS1-R9]
[RWS1-R13]	removed - subsumed	merged into [RWS1-R9]
[RWS1-R14]	removed - subsumed	merged into [RWS1-R9]
[RWS1-R15]	removed - subsumed	merged into [RWS1-R9]
[RWS1-R17]	removed - subsumed	merged into [RWS2]
[RWS1-R28]	removed - subsumed	merged into [RWS1-R25]
[RWS1-R29]	removed - subsumed	merged into [RWS1-R25]
[RWS1-R32]	removed	Time zone conversion capabilities are a special case of data transformation according to user and application needs.
[RWS1-R36]	removed - subsumed	merged into [RWS2-R25]
[RWS1-R37]	removed - subsumed	merged into [RWS1-R9]
[RWS1-R45]	removed - subsumed	merged into [RWS1-R45]

6 Workshop Conclusions

6.1 Feedback from the Participants

The following list contains feedback statements from the workshop participants that were provided during and at the end of the workshop in specific feedback sessions. The statements are listed anonymously and in no given order. Statements from different persons are grouped together in order to ease understandability. Statements might be redundant, overlapping and contradictory due to the fact that they originate from different individuals.

- The existing participant questionnaire is intimidating when sent to the scientists without much prior knowledge of the planned CI and the project. Improve the questionnaire: extend, shorten, crispen up.
- In order to get participating scientists to donate their time answering the questionnaire, the CI team should point them to the benefits. Sell the questionnaire to people to make them feel good about making it happen. Engage people but don't tax them.
- The process could be improved by introducing the CI concepts first and presenting the OOI to the broader audience.
- It is better to go through the questionnaire as a group exercise as on day 2 than to fill it out beforehand.
- Different surveys should be used for different target audiences
- The time required to understand and fill out the questionnaire should be as small as possible. Anything that can be done to make it easier for people should be done.
- If the questionnaire were made available to the broader community, it needs some adjustment. The language was intimidating; it needs to be clarified.
- Make a 5 minute survey for the public version.
- Provide incentives for taking part in the survey or the workshop, e.g. give away CI mugs or CI T-Shirts.
- If a web-based interface will be used for the public questionnaire, it could be helpful to present explanations, further information and figures related to certain questions on request for clarification.
- Questionnaires can be handed out to participants at the upcoming ocean science conference.
- There is a choice for the next workshop between refining materials (which narrows the search space but makes it easier) or to start over. The material presented during this workshop can be the basis for further exploration and refinement. For the next workshop, the CI team can present existing materials to other participants and see how they react.
- The present audience was probably one of the easiest to approach with CI related topics. Instrument owners and oceanographers might require more effort.
- Other possible audiences include NCAR people, GFDL people, the geosciences collaboratory group, CI operators, etc.
- Too many people at the workshop do not work. The format does not scale. 3-5 scientist participants are a good number. Small meetings are good for developing requirements.
- Two days for the workshop was a good time. It should not exceed 2 days. After 2 days, saturation is reached
- "I came a skeptic but then I enjoyed the workshop"
- Further requirements elicitation, refinement and validation based on existing results can be done in smaller and shorter meetings and even remotely.
- For remote conferences, reliable tools and infrastructure should be used to avoid the common problem that everyone has with the connection, etc.

- Prospective workshop participants should be invited to a conference call prior to the workshop, where OOI and CI background can be presented and the questionnaire introduced. This eliminates the time required for context setting
- This meeting went smoother than the last workshop because of the availability of results such as the first report and the experience of the CI team.
- The CI background presentations and introductions should be dumbed down and adapted to the invited audience. The language and many technical concepts can be intimidating. Make it easier to digest the concepts
- In the future, atmospheric models need to be coupled to oceanographic models for a credible outcome.
- "It was a nice workshop; I learned a lot"

6.2 Next Steps and Action Items

Next steps include:

- Analyze the workshop results and compile the workshop report
- Refine the questionnaire and provide a web-based version for a broader audience
- Perform technology investigations (e.g. Condor)
- The modelers provide some example user questions (emails) for help on model-related topics. This will enable the CI ADT to estimate the degree of assistance required by the CI to non-expert users

6.3 Conclusions from the Organizers

The second OOI CyberInfrastructure Requirements Workshop hosted by the University of California, San Diego, was very successful in providing valuable outcomes for CI requirements definition and validation efforts, for refining and complementing the CI architecture and design, and in further fostering the mutual understanding of prospective CI user communities and the CI design team. Direct outcomes, such as the list of identified and validated requirements, the jointly developed domain models and the use scenarios, will be valuable assets in the subsequent CI design efforts. Further results include a validation of the requirements previously collected for the Conceptual Architecture, and initial outreach measures to future CI user communities.

The subsequent planned requirement workshop is expected to deepen the mutual understanding, and further refine, complement and validate OOI CI requirements and design, based on the previously elaborated results through a complementation with input from a different user community. All presentation materials can be found on the workshop website [RWS2-WEB].

Appendices

A OOI Supported Science Questions

The following are integrative examples of some of the broad science questions that the OOI network will be able to address (cf. [SCIPROSP]).

What is the ocean's role in the global cycle? *What are the dominant physical, chemical, and biological processes that control the exchange of carbon and other dissolved and particulate material (e.g., gases, nutrients, organic matter) across the air-sea interface, through the water column, and to the seafloor? What is the spatial (coastal versus open ocean) and temporal variability of the ocean as a source or sink for atmospheric CO_2? What is the seasonal to inter-annual variability in particulate flux? What is the impact of increasing pH to ocean chemistry and biology?*

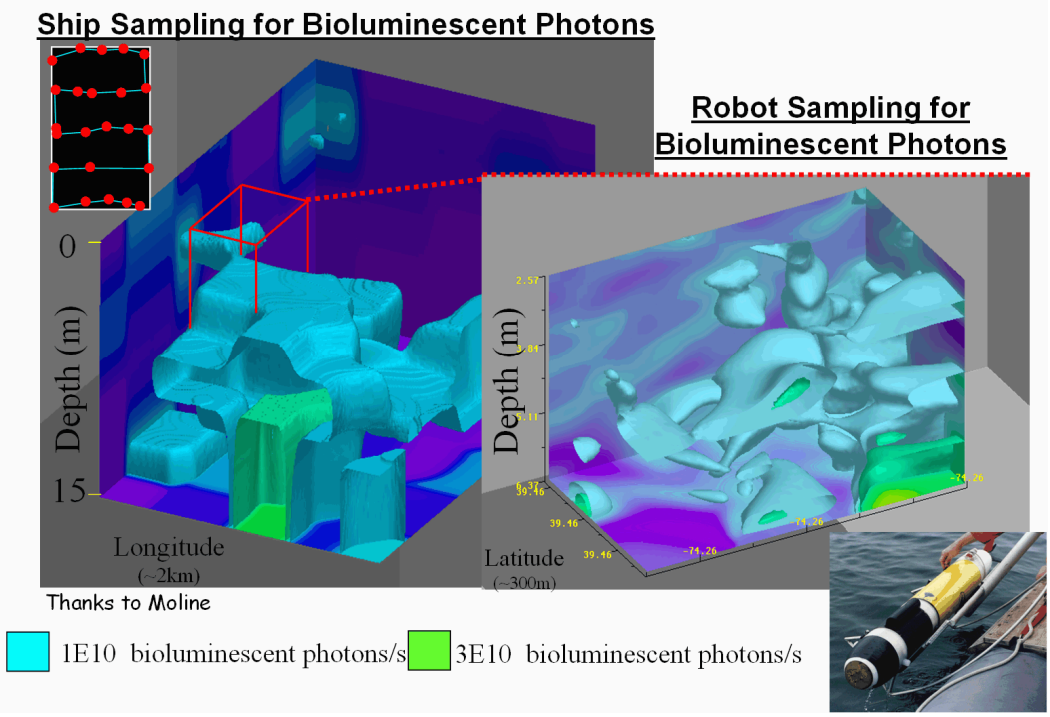

Figure 4: Small scale variability in sampling complex ocean processes

One of the most striking geochemical patterns observed in the twentieth century is the rising concentration of atmospheric CO_2. This discovery, only possible with sustained decadal observations, has few analogues in the ocean, despite that the ocean plays a dominant role in the global carbon cycle and represents the largest reservoir of carbon on Earth. Observations suggest complex processes that make the ocean a carbon sink are being modified as a result of increasing atmospheric CO_2 loading and climate change (25, 26). The exchange of CO_2 between the atmosphere and ocean is mediated by air-sea mixing and ocean ventilation, carbonate equilibrium (the solubility pump), and the conversion of dissolved CO_2 into particulate and dissolved organic carbon by marine phytoplankton and respiratory pathways (the biological pump). The fraction of the biologically fixed carbon that becomes sequestered in marine sediments is mediated by the structure of pelagic ecosystem. These ecosystem processes are predicted to change as increasing ocean CO_2 concentrations decrease ocean pH. Changes to high-latitude food webs, especially in the North Pacific and Southern Ocean, are disproportionately important regions of marine CO_2 biogeochemistry appear to be particularly sensitive.

The air-sea flux of CO_2, biological carbon fixation and sequestration rates are highly variable and understanding interactions between the biology and geochemistry is a major challenge for the research community. This research problem will require data collected at high sampling rates (hours to days) to quantify the importance of episodic events to air-sea fluxes and carbon fixation rates which are disproportionately important and yet are chronically under-sampled. These OOI infrastructure and real-time data will enable individual investigators by not only providing the pelagic network for deploying open/closing sediment traps but also providing the means to trigger adaptive sampling of episodic processes. Spatially the combination of highly instrument water column sites combined with a broad network of gliders will provide the mesoscale context required for interpretation. Figure 1 above shows visualizations of numerical model output covering CO_2 flux.

How important are extremes of surface forcing in the exchange of momentum, heat, water and gases between the ocean and atmosphere? *How important are extremes of surface forcing (such as severe storms) in the exchange of momentum, heat, water, and gases between the ocean and atmosphere? What is the effect of extreme wind on structure on the upper mixed layer depth? What are the air-sea fluxes of aerosols and particulates?*

Improving the knowledge of the mechanisms underlying air-sea exchange is crucial to the interpretation of larger scale physical and biogeochemical processes. The lack of observations at the air-sea boundary during high wind and sea states is a serious impediment to our understanding of air-sea exchange during extreme atmospheric forcing. This is problematic as for many air-sea interactions are disproportionately important. Measurements of the exchange of mass (including gases, aerosols, sea spray, and water vapor), momentum, and energy (including heat) across the air-sea interface during high wind conditions (> 20 ms^{-1}) are rare. Ships are not generally effective sampling platforms in severe storm conditions. The availability of these data have been identified as critical to improving the predictive capabilities of storm forecasting and climate change models, and for estimates of energy and material (e.g., carbon, nitrogen) exchange between the upper and deep ocean.

Continuous and simultaneous measurements above and below the air-sea boundary for periods of years to decades would provide the needed measurements of extreme surface forcing over time frames sufficient to observe episodic, seasonal, annual, and decadal processes. The difficult aspect in these processes is that measurements are required just above and below the sea surface. This has been difficult to accomplish with standard moored sensors especially in regions of high wind and thus OOI must provide the sufficient stability and power to support a suite of rugged meteorological and in-water sensors to enable studies of the dynamics marine storms, upper ocean circulation, primary productivity, ocean carbon fluxes, and climate. The real time communications is critical to enable adaptive sampling of subsurface measurements.

How important are severe storms and other episodic mixing processes affect the physical, chemical, and biological water column processes? *What are the effects of variable strength storms on surface boundary layer structure, nutrient injection in the photic zone? How does storm induced nutrient injections influence the primary productivity, and vertical distribution and size structure of particulate material?*

Water column mixing is central to driving ecosystem productivity by replenishing nutrients to the euphotic zone; however if mixing is too vigorous overall productivity is suppressed by the limitation of light. The nonlinear interaction between the mixing and light availability and the corresponding ecosystem response remains a central question to biological and chemical oceanography. These nonlinear processes impact the overall community composition of the phytoplankton which has cascading impacts on the entire foodwebs. Measuring the mixing of heat, energy, particulate and dissolved material and the

corresponding impact on ecosystem dynamics has been difficult problem and traditional sampling approaches have not allowed scientists to maintain persistent presence in the ocean to quantify the role high and low frequency mixing events. As the relative importance of episodic and seasonal mixing events on the overall productivity marine ecosystems remains an open question and placing in context the importance of large cyclical phenomena (ENSO, PDO, NAO) remains difficult.

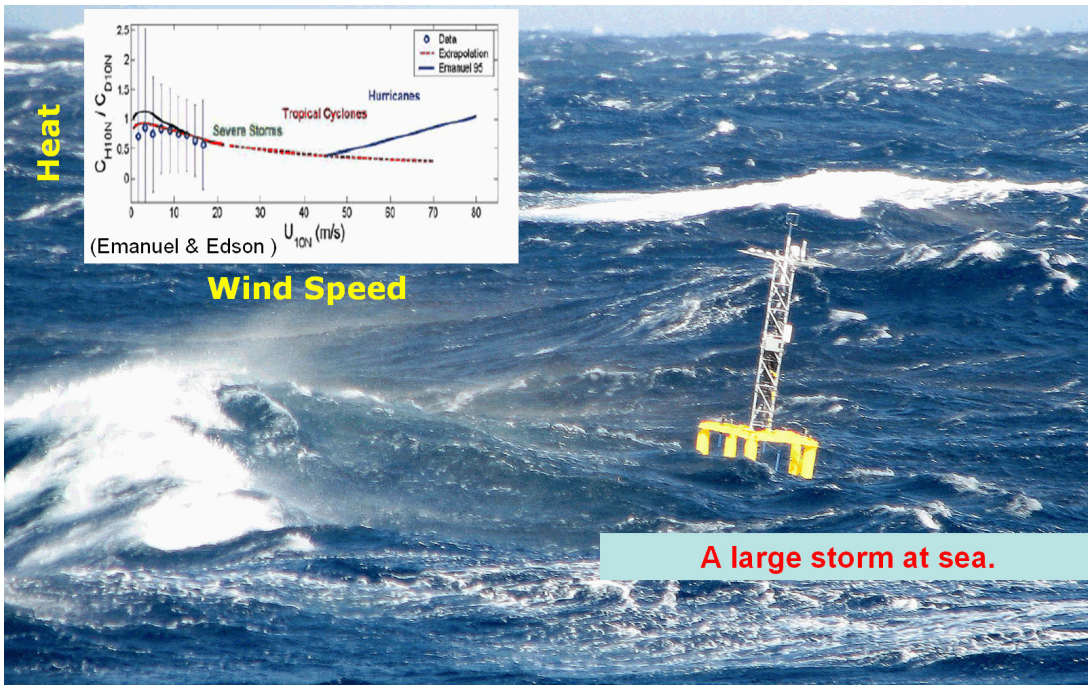

Figure 5: Severe storms and other episodic mixing processes affect the physical, chemical, and biological water column processes

The OOI will enable research by providing the infrastructure to persistently observe mixing processes in the ocean and assess the corresponding impact on the marine ecosystems. The distributed OOI assets will provide measurements for studying air-sea exchange processes, the dynamics in mixed layer depth, measurements of material exchange across the base of the mixed layer, internal wave dynamics, the evolution of benthic boundary layers, corresponding changes in the composition and size distribution of the phytoplankton. The measurements will be made on horizontal scales of meters to kilometers and vertical scales of millimeters to meters. Data will be collected on the time scales of minutes to hours which will be sustained for years providing for the first time a large time series that has high frequency data from a range of ocean ecosystems over all weather conditions. Data collected by profiling moorings will be critically important especially in the upper 200 meters of the water column. The high frequency sampling from the profiling moorings will be spatially extended by the coordinated transects collected from fleets of AUVs and gliders. Resuspension and benthic boundary layer dynamics will be enabled with sensors mounted at several depths.

How does plate scale deformation mediate fluid flow, chemical and heat fluxes, and microbial productivity? *What are the temporal and spatial scales over which seismic activity impacts crustal hydrology? How does the temperature, chemistry and velocity of hydrothermal flow change temporally and spatially in subsurface, black smoker, diffuse, cold seep and plume environments? How are these systems impacted by tectonic and magmatic events?*

The oceanic crust is the largest fractured aquifer on the planet. Thermally driven fluid circulation through the oceanic lithosphere profoundly influences the physical, chemical, and biological evolution of the crust and oceans. Fluid circulation within this aquifer provides heat and nutrients that sustains a vast microbial biosphere below the seafloor that is just beginning to be explored and may rival that on the continents. Despite some progress many of the most important fundamental questions remain such as the depth and extent to which life may occur within the subseafloor and overlying sediments, and the linkages between submarine plate tectonic and sedimentary process and this sub-seafloor biotope. Transient events such as magmatic eruptions at mid-ocean ridges increase nutrient (e.g. carbon dioxide) output and venting volume by as much as a factor of 100, resulting in extensive microbial blooms. Organisms sampled from high-temperature ecosystems at deep-sea hydrothermal vents have challenged our understanding of the physical and biochemical conditions under which life not only exists, but thrives. Studies of this vast biosphere in an such extreme environment are leading to biotechnical research for development of new pharmaceuticals important in fighting disease and infections and biocatalysts (enzymes) that are more efficient, thermally stable and cost-effective than synthetic catalysts important in material processing for industries.

A network on the Juan de Fuca Plate offers an unparalleled opportunity to examine the hydrological connectivity of the oceanic crust and crustal strain at a plate scale. This plate already hosts the highest density of instrumented Ocean Drilling Program and Integrated Ocean Drilling Program sites of any place within the worlds oceans, and Site 1027) will be connected to NEPTUNE Canada. Real-time access to the subseafloor via sealed boreholes coupled with suites of sensors will offer an unprecedented opportunity to study fundamental questions about the dynamics of the lithosphere and linkages with the subseafloor hydrosphere. The major volcanic and tectonic events that create the oceanic crust and that modulate the fluxes across the seafloor and that impact biological communities are inherently episodic on decadal time scales and are also short-lived. The only way to capture these events is to maintain a long-term monitoring capability at a number of sites with high probability for tectonic or magmatic activity. Because Axial Seamount is the most magmatically robust volcano on the Juan de Fuca Ridge and because it hosts several vent fields, it is an optimal site to study linkages among seafloor spreading, volcanic activity, and hydrothermal flow. Axial Seamount also hosts a robust subseafloor microbial community and is one of the few sites in the worlds oceans where several year time-series studies have documented temporal changes in microbial communities following an underwater eruption that are linked to changes in fluid chemistry-temperature.

What are the forces acting on plates and plate boundaries that give rise to local and regional deformation and what is the relation between the localization of deformation and the physical structure of the coupled astenosphere-lithosphere system? *What is the style of deformation along plate boundaries? What are the boundary forces on the Juan de Fuca plate and how do the plate boundaries interact? What are the causes and styles of intraplate deformation? What is the return flow from the ridge to the trench? How much oceanic mantle moves with and is coupled to the surface plate? How and why do stresses vary with time across a plate system?*

Tectonic plates are the fundamental building blocks of our planet, with boundaries defined by subduction zones, mid-ocean ridges, and transform faults. The Juan de Fuca plate incorporates a remarkable array of plate tectonic features within a relatively small area, including all major types of oceanic plate boundary and a continental ocean convergent margin capable of destructive earthquakes. As a consequence, whole-plate seismological and geodynamic observations provide a unique opportunity investigating the processes that control the formation, evolution, and destruction of an oceanic plate and of the interactions of that oceanic plate with the leading edge of a continental margin.

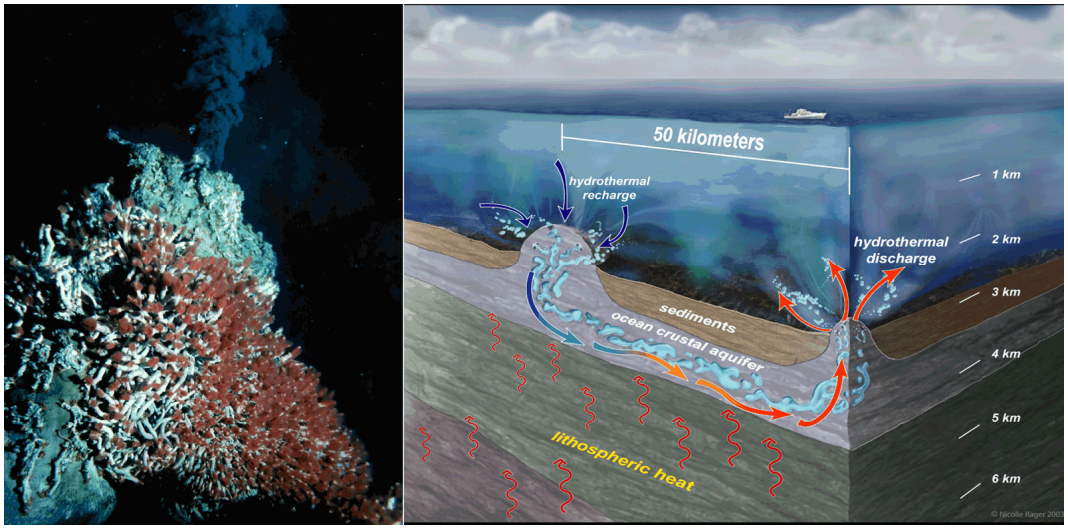

Figure 6: Major ecosystems in the ocean crust are modulated by Earthquakes – the OOI will provide the scientists the first sustained presence in the deep sea

A plate-scale seismic array will also facilitate studies of the structure and evolution of the lithosphere-asthenosphere system. In combination with data from land-based studies, a plate-scale seismic array will allow unprecedented imaging of the deep and shallow structure that accompanies plate formation, evolution and subduction. Such work would contribute to our understanding of mantle melting, the mechanical coupling of the asthenospheric mantle to the lithosphere, the pattern of return flow from trench to ridge, the nature of mantle flow near contrasting plate boundaries, the rheology of the mantle, and the importance of three dimensional plate-scale structure for localizing and influencing seismogenic deformation.

Because many of the problems of interest require observations at the plate scale, the first priority for a seismic network is to deploy a system capable of studying the entire Juan de Fuca plate with a network of broadband seismometers. Such a network could provide a regional context for other, more local experiments thus forming an integrated system of multi-scale observatories. Understanding the life cycle of an oceanic plate and interactions across the entire plate will require a series of experiments that address processes at a variety of scales, ranging from plate-scale monitoring and imaging (using arrays with apertures of about 1000 km) to more local experiments with apertures on the order of kilometers. Because seismic events cause significant perturbations to the hydrology of the oceanic crust, and are indicative of magmatic intrusion and eruptions, real-time data transmission would be the key to realization of these events and optimization of event response capabilities. At the plate-scale, observations of seismicity and deformation will constrain many important processes including the nature and causes of variations of stress with time across the entire plate, the styles and causes of intra-plate earthquakes, and the coupling of forces across plate boundaries.

How do tectonic, oceanographic and biologic processes modulate the flux of carbon into and out of the submarine gas hydrate "capacitor," and are there dynamic feedbacks between the gas hydrate methane reservoir and other benthic, oceanic and atmospheric processes? *What is the role of tectonic, tidal and other forces in driving the flux of carbon into and out of the gas hydrate stability zone? Can natural temperature fluctuations perturb the effects of long-term temperature change on hydrate stability, or are perturbation experiments required to artificially raise the temperature? What is the fate of hydrate/seep methane in the ocean and atmosphere?*

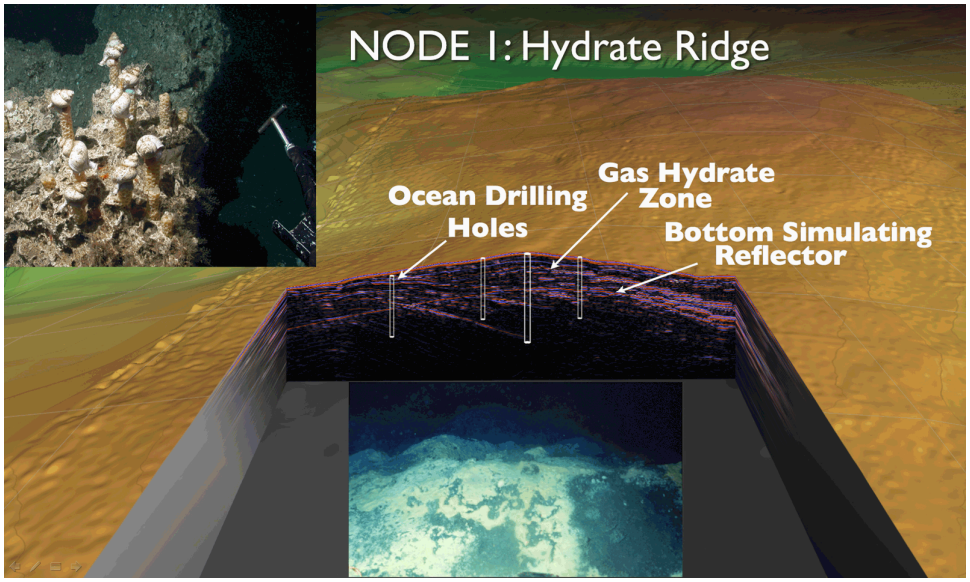

Figure 7: Tectonic, oceanographic processes modulating the flux of carbon – the OOI will provide the scientists the first sustained presence in hydrate fields

A significant amount of the methane near the surface of the Earth is locked into gas hydrates in shallow sediments on continental margins. The hydrates may act as a capacitor in the carbon cycle by slowly storing methane that can be suddenly released into the ocean and atmosphere. An over arching goal is understand how important gas hydrates are to seafloor and sub-seafloor environments and to understand the role of gas hydrate in modulating the flux of carbon between the solid earth, hydrosphere, atmosphere and biosphere. Given current uncertainties, understanding the possible biogeochemical feedbacks in this system is critically important. Long-term observations are required to constrain hypotheses about system evolution and response to transient internal and external forcing events.

Hydrate Ridge (Node 1) in the central Cascadia accretionary complex is one of the best-studied gas hydrate deposits. Seafloor venting and formation of gas-rich hydrate deposits near the seafloor have been documented at Hydrate Ridge through ODP drilling during Legs 146 and 204 and by a series of seafloor studies using submersibles and ROV's. These studies have provided a basis for understanding how gas hydrate is distributed in marine sediments and the processes that lead to heterogeneity in this distribution. In this area, the subsurface has been imaged with 3D seismic data, which define a focused plumbing system that provides a clear target for observatory instruments to define the temporal evolution of this system, determine material fluxes from the earth into the ocean and understand biogeochemical coupling associated with gas hydrate formation and destruction.

How do cyclical climate signals at the ENSO, NAO and PDO timescales structure the water column and what the corresponding impacts on the chemistry and biology in the ocean? *What are the effects of climate signals on variability in water column structure, nutrient injection in the photic zone, primary productivity, and vertical distribution and size structure of particulate material? Are secular climate change trends detectable in the oceans?*

Atmospheric forcing at seasonal to inter-decadal scales is a globally important factor that structures marine food webs. Understanding when and how the marine ecosystems shift between equilibrium states is a widely debated, and central, issue for both the research and marine resource management communities. This is especially important as many ocean food webs appear to be undergoing major shifts. For example limited time series in the sub-tropical north pacific show substantial changes in the phytoplankton, zoo-

plankton, and pelagic fish biomass during the mid-1970's and 1980's. Early evidence indicates another shift may have occurred in the late 1990's. The climate of the north Atlantic has also shifted over the last forty years with changes in the ocean circulation and declines in copepod and cod stocks.

Understanding the causes, processes, and consequences of inter-annual variability and inter-decadal scales requires high frequency (minutes), sustained (decades) time-series data across a range of ecologically relevant spatial scales spanning both global and deep sea ocean systems. High frequency data is required to resolve the physical structuring of marine food webs which often are disproportionately fueled by short-lived episodic events. The network of distributed assets will be required to provide measurements of both atmospheric and in situ physical, chemical, and biological properties. Vertical resolution of the system will need to range from less the one meter to tens of meters and horizontal scales ranging from meters to hundreds of kilometers. For many of the ecosystem questions resolving the chemistry and the particulate matter in the upper 200 meters of the water column will be particularly important. Given that many of the signatures of these large scale processes are resolved at local and regional scales it will be important that the network span a range of coastal and global sites. Locations in the North Pacific will be impacted by the ENSO and PDO cycles while sites in the subpolar and sub-tropical Atlantic will resolve the impact of the NOA. Chronic under-sampling in Southern ocean sub polar water would fill critical gaps in current data sets.

What are the dynamics of hypoxia on continental shelves? *What are the biological and chemical consequences of low DO on the continental shelves? How frequent are low DO intrusions and to what degree are they driven by atmospheric forcing and deep sea circulation? How does biological activity on the continental shelves modulate the intensity of the low DO events?*

Low dissolved oxygen concentrations have been documented in the coastal waters off Oregon during late spring to summer of 2002 to 2007. Large regional scale oxygen depletions have also been documented on the MAB in the recent past. Unlike hypoxic events fueled by anthropogenic nutrients and/or limited circulation of semi-enclosed estuaries or embayments, hypoxia on the PNW continental shelf is driven by atmospheric forcing, upwelling/downwelling, and variability in ocean circulation. Upwelling brings nutrient-rich, oxygen poor deep waters from polar and sub-polar seas onto the shelf fueling phytoplankton blooms which, in turn, reduce oxygen levels in the near-seafloor water column through decomposition. The alternating periods of upwelling and downwelling generally sets the stage for intensity and duration of shelf hypoxic events. Surveys have shown these are large-scale events (on the order of 3000 km2) and have serious impacts to the coastal ecosystem, including mass die-offs of commercially important shellfish and finfish. In contrast, the 2002 event was triggered by an invasion of low oxygen, subarctic water from the Gulf of Alaska. This "anomalous" source water was advected onto the Pacific Northwest shelf depressing dissolved oxygen levels in offshore waters from Vancouver Island to southern Oregon. The formation and duration of hypoxic areas are subject to climate variability and variations in oceanic flow on seasonal, inter-annual, ENSO, and inter-decadal scales. Understanding hypoxic events and impacts to PNW marine ecosystems requires the ability to observe physical, chemical, and biological conditions across the continental shelf to slope waters, for periods spanning years (seasonal to inter-annual change) to decades (ENSO and PDO shifts). This is an especially pressing problem as the impact of the low oxygen water can trigger mass mortalities in high tropic levels.

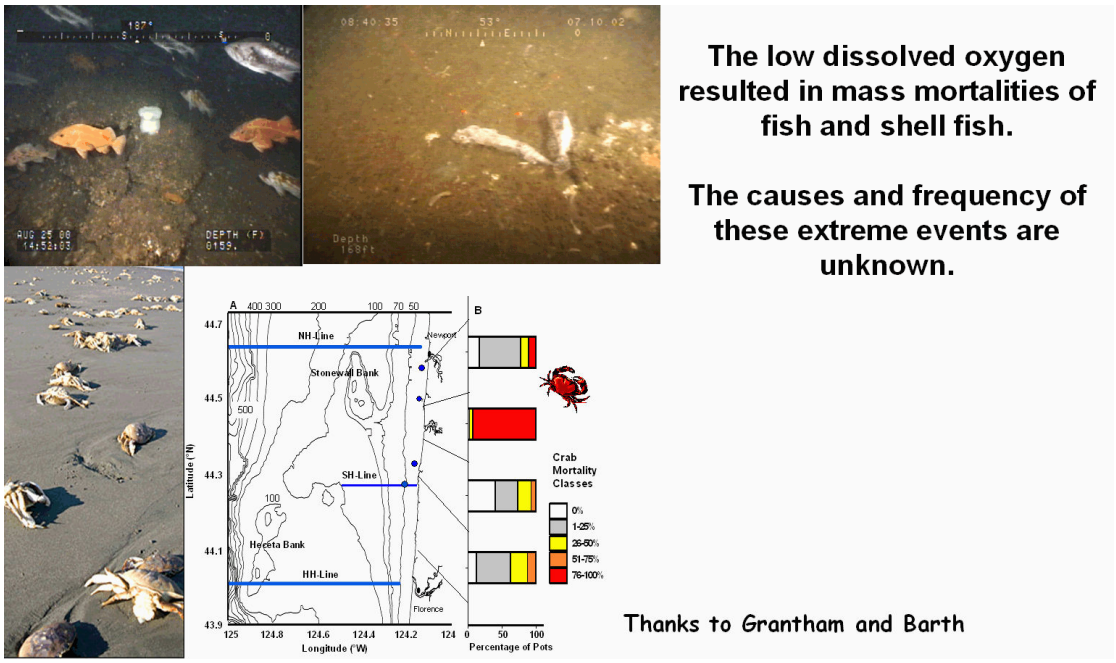

Figure 8: Dynamics of hypoxia on continental shelves resulting in mass mortalities of fish and shell fish – the OOI will provide sustained spatial time-series observations from the local to the mesoscale to understand interannual variability of hypoxia on the east and west coasts

Studying the frequency, intensity and mechanisms driving the invasion of low DO water on continental shelves will require a distributed network of fixed and mobile platforms. Large 3-D volumes of data collected by Gliders will provide maps of the low DO waters and they can adaptively map the spatial extent and morphology of the low DO intrusion. This volumetric data is then complemented with continuous, long-term operation of an instrumented array with sufficient power and bandwidth to support multi-disciplinary sensors providing the observations to study coastal ocean processes from event-scale to inter-annual variability to inter-decadal trends. Time series of the cross-shelf gradients in physical and biogeochemical properties across the continental shelf and slope combined with simultaneous observations of the meteorological forcing, oceanic flows, and a range of physical and biogeochemical properties measured with high vertical resolution will allow scientists to study the corresponding chemical and biological response to the low DO water with an unprecedented detail.

B Workshop Participant Questionnaire

The CI ADT identified several relevant categories for the CI science user requirements; for each of the categories, a number of questions were identified, which when answered could lead to new and refined CI science user requirements. All questions in the respective categories together with an introduction and context setting were compiled as a slide set presentation. The workshop participants received a significantly shortened version of this questionnaire prior to the workshop.

Intent of this template

- This slide set is a template for participants of the OOI CI requirements workshop in San Diego, January 2008
  - For presentations during the workshop
  - To capture relevant information in a structured way
- Goals of this exercise are
  - To capture as many CI relevant details as possible before the workshop
  - To capture structured, relevant information for use during and after the workshop
  - To enable quick information access for domain modeling during the workshop
  - To provide you some ideas about the expected outcome and materials covered during the workshop from a perspective of the CI design team
- We ask you to please fill it out to the degree possible/applicable. Please try to provide answers to as many (relevant) questions as you can
- You can use this template as you like. You can modify it, take only parts of it, add own slides, copy/paste out of it, use it to structure own text/spreadsheed/slideset documents …

Goals for the Requirements Analysis

- Analyze the Current Situation
  - Definition of basic terms: model, data, etc.
  - Tools, technologies, processes, data used and/or available
  - Organizational details (e.g. responsibilities, roles in team, workflows, policies)
  - Current shortcomings for whatever reason
- Determine Short-Term Improvements
  - What would make your every-day modeling tasks easier and more effective? List and rank, if possible.
  - Which shortcomings should be eliminated most urgently?
- Identify CI Transformative Vision and Requirements
  - Assumed there is a transformative community CI in place, what are your expectations to an “ideal CI”?
  - Capabilities, interfaces, made guarantees, resources provided, etc.
- Scope
  - As relevant to the OOI CyberInfrastructure
  - From a viewpoint of your community primarily, numerical modelers

Question Categories

- Basics
  - Current situation and expected changes
  - Definition of terms
- Technology

  - Models
  - Model Processing
  - Model Output, Visualization
  - Data, Data Sources
  - Technology, Infrastructure, Tools, Resources
  - Interfaces
- Organization
  - Workflow, Responsibilities
  - Privacy, Security, Policy
  - Operations and Maintenance
- Misc
  - Education and Outreach
  - Summary requirements
  - Comments, expectations, suggestions
  - Additional reading materials, concepts, sources, references

Current situation and Expected changes

- Please briefly describe your current situation, e.g. every-day tasks in numerical modeling and related activities (overview)
- What changes do you expect for the next 3-5 years?
- What transformative changes do you envision and/or anticipate for a 5-10 year time frame?
- What capabilities do you expect from a transformative cyber-infrastructure in the oceanographic domain?
- How would you use these capabilities if they were in place?
- What could and/or would you provide to the community as part if the infrastructure (e.g. data, tools, algorithms)?
- Are there any similar projects/communities that you like and/or that are technology-wise exemplary?
- What general developments would advance you/the community most?

Definition of Terms

- How do you define “model” or “numerical model”?
  - E.g. is it the algorithm, the output data, etc.
- How do you define “data”?
- How do you define “meta-data”?
- How do you define “workflow” resp. “process”?

Models

- What kind of models exist in your community? Or should be there?
- Which models do you use and/or develop?
- Please explain the specifics of (some of) these models
  - Size of the model algorithm
  - Parameterization possibilities
  - Number and type of input variables
  - Output variables or grid points, per time
  - Output data volume
  - Complexity of the model execution workflow

- Do you build models based on external models, tools, applications?
- Do you have an description of a typical every-day scenario using your models?
- What would make your modeling/analysis work more effective?
- Are your models open for change of formats, standards, platforms, technologies? Do you anticipate changes?
- To which degree would you accept change if it brings the community forward?

Model Processing

- Please detail some model execution characteristics
    - How often do you run the model?
    - How long does it take to run the model?
    - How often does the model change? Are changes parametric or algorithmic?
    - What are the execution platforms? Do the models have specific technology dependencies (e.g. compilers, platforms, libraries, computation resources)
    - Would your models benefit from parallelization and/or super-computing?
    - Could you run your models on a remote common infrastructure? Would you?
- Do you use external on-line resources (e.g. computation grids, data archive)?
- Do you support on-line processing?
    - If so what is your concept of real time?
    - What kind of connectors are you able to work with? Can you handle streams?
    - Is there a need for an infrastructure accessing data in both ways?
    - Are you able to cache incoming/outgoing data in files or databases?

Model Output, Visualization

- How do you store, publish, announce, and describe your model results?
- Do you provide different versions of the same data (e.g. lo-res, high-res, or filtered)?
- How often do you envision to update outputs?
- Do you envision revisions to data? How often does this occur in practice?
- Which meta-data do you associate with output data and how?
- What visualizations do and/or the community apply?
- How could a common infrastructure support (interactive) visualization?

Data, Data Sources

- What are the stages data undergoes from raw data to output data? E.g. filtering, processing, down-sampling, aligning steps
- What data should be stored and backed up by a common infrastructure and when?
- Who has “ownership” of data in different steps?
- What are typical data exchange formats?
- What meta-data is relevant to find the right data source? Are there specific meta-data standards used?
- What quality/reliability/accuracy/certification levels for data exist and how do you select if you have the choice?
- Which specific data sources do you use? How did you find them? How did you get access?
- Do you have backup sources for the same data in case of unavailability?
- Which manual interaction is required to check/validate/modify the data?
- What data volumes do you handle and/or anticipate? Any high-bandwidth data streams?

- Do you use streamed data or bulk data files or databases?
- What data filters and/or transformations do you apply?
- Which (external) tools do you use for data processing, transformation, etc.
- What's the frequency of data update? How often do you expect new data for new model runs? Can the models/applications handle continuously steamed data?
- Do you have example data files? Meta-data files?

Technology, Infrastructure, Tools, Resources

- If not done so with the data and models questions, please list the technologies, data standards and formats, tools, applications, computation platforms that are most prominent in your work and/or the community in general
- Interfaces
- What interactive user interfaces of the OOI CI do you envision and/or require?
- Should there be any other interactive interfaces besides web interfaces?
- Which user interface technologies are particularly efficient for your daily work?
- How much flexibility and/or expressive power should the user interfaces offer (vs. intuitive use)?
- Do you particularly like any current scientific web portals and their user interfaces?
- What are the biggest "must-haves" and "no-nos" with user interfaces that you plan to use regularly?
- What programmatic and application interfaces of the OOI CI do you envision and/or require?
- Do you need off line access capability?
- Do you require specific standards and/or technologies?

Privacy, Security, Policy

- What are the relevant roles and responsibilities in your organization? (E.g. data manager, operator, modeler, user)
- Are there any privacy concerns with your algorithms or used/produced data? Which?
- Is there a need for data embargoing for certain time frames?
- Are there intellectual property issues associated with data, algorithms, etc? Which?
- What security infrastructure do you or your organization use?
- What are the authentication mechanisms and policies?
- What are the authorization levels, granularity, privileges and mechanisms?
- Would you entrust currently self-operated models/computation/data/resources to a community infrastructure? Under which conditions?
- If other researchers had access to your models/data/resources, how would you like to see these protected? Roles, security, quota etc?
- Which governance and policy concerns (for resources, data, model use etc.) apply to you? Which strategies do you see as effective in applying them?

Operations and Maintenance

- How do you store (archive) your models?
- How do you store your data results?
- How do you manage different versions of data/models?
- Is there a responsible contact person available for operation/maintenance?
- How much of the total effort is required for operation and maintenance of hardware, models, network etc.?

- How do operation and maintenance requirements affect the design of you models and your daily work?

Education and Outreach

- How do you make modeling results available for education and outreach purposes?
- How do education and outreach concerns affect your models and the presentation of the results?
- How do you support publication of results? E.g. by making data available in special formats, for journals
- How do you integrate system with education environments?
- Do you consider releasing models, algorithms, tools as open source for the public? How does this affect your work?

Requirements Summary

- Do you have other specific requirements?
- Any specific standards to definitely incorporate
- What are current missing capabilities in general?
    - Such as higher sampling rates, better accuracy, more instruments, merging data, correlate data
    - Any data formats needed for processing, transfer?
- List the 3 short-term advances that would benefit you most
- List the 3 mid-term advances that would benefit you most
- List the 3 impediments for you/the community currently
- Can you provide a ranking for the requirements?

Comments, Expectations, Suggestions

- What do you expect from the upcoming workshop?
- Anything you think is relevant that you want to add?

Additional reading materials, References

- Reading materials
- References

C Workshop Developed Domain Models

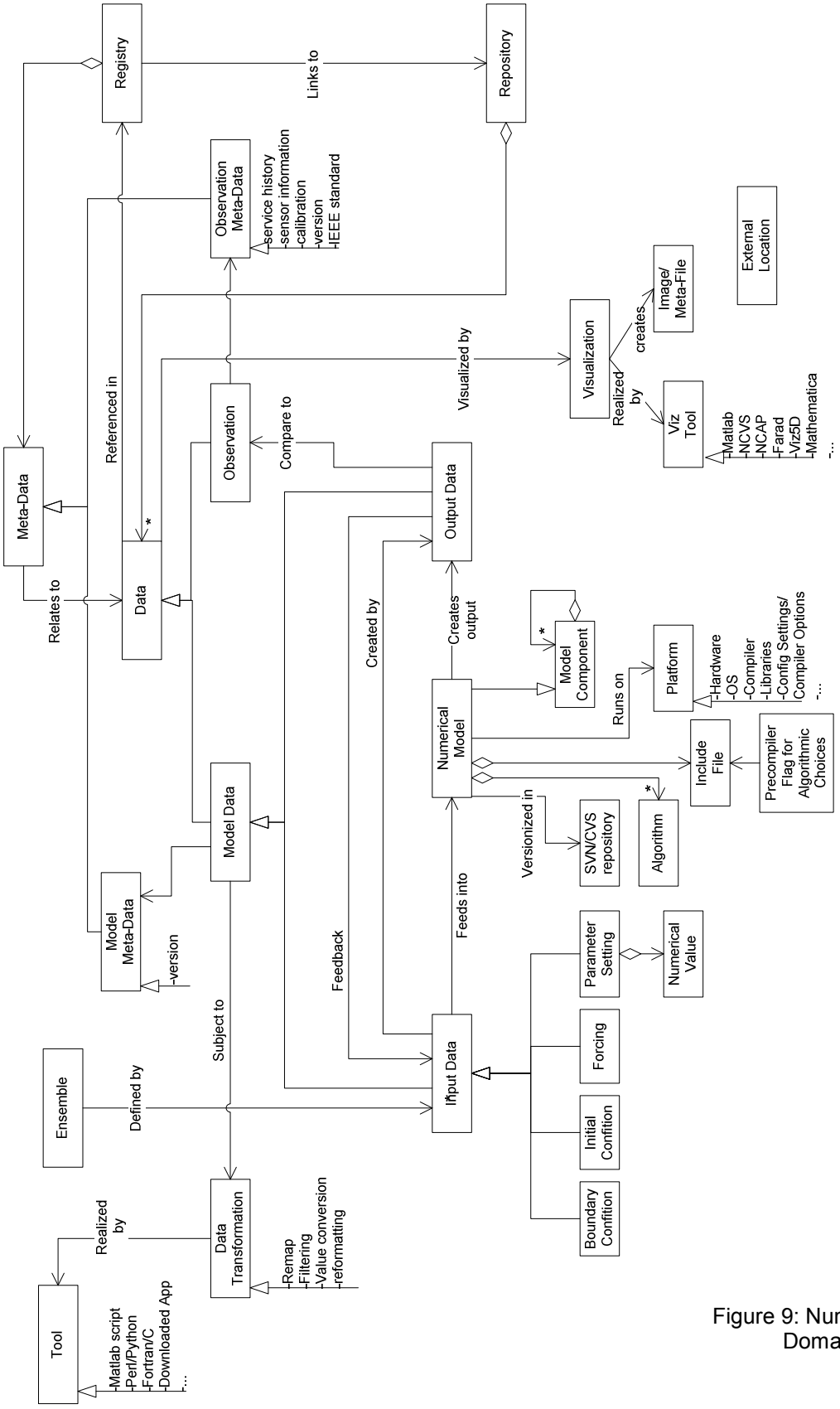

Figure 9: Numerical Modeling Process Domain Model, Group 1

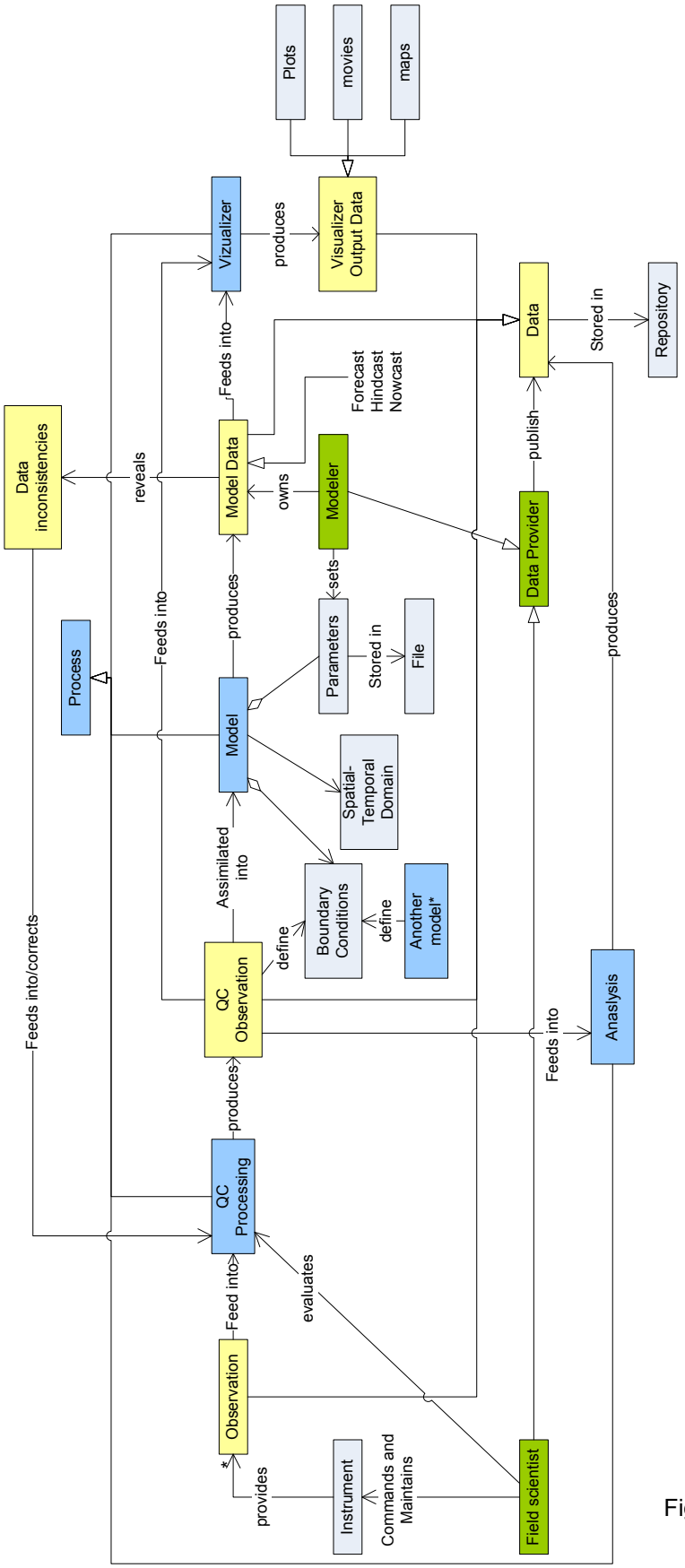

Figure 10: Numerical Modeling Process Domain Model, Group 2

D Workshop Agenda

Day 1, January 23, 2008 (Wednesday)

Time	Presenter(s)	Topics
08:30 AM	Oscar Schofield	Welcome & Introductions
08:40 AM	Oscar Schofield	Meeting Purpose, Goals and Agenda Overview of current OOI plans after PDR
09:15 AM	Matthew Arrott	CI Infrastructure for the OOI
09:50 AM	Ingolf Krueger	Developing science user requirements CI System Engineering
10:45 AM	Andy Moore	Project and research overview
11:00 AM	Bruce Cornuelle	Project and research overview
11:15 AM	Bill O'Reilly	Project and research overview
11:40 AM	Libe Washburn	Project and research overview
01:00 PM	CI ADT	Validation and review of the RWS1 user requirements
02:00 PM	CI ADT	Break-out sessions 1: Present day numerical modeling use cases & workflows
03:45 PM	CI ADT	Break-out sessions 2: Domain modeling of present day numerical modeling
05:00 PM	CI ADT	Day 1 wrap-up and charge for day 2

Day 2, January 24, 2008 (Thursday)

Time	Presenter(s)	Topics
08:30 AM	CI ADT	RWS1 requirements prioritization
09:30 AM	CI ADT	Requirements elicitation session 1 Participant questionnaire walkthrough
01:00 PM	Yi Chao	Project and tool overview
01:30 PM	CI ADT	Requirements elicitation session 2 Participant questionnaire walkthrough, continued
03:00 PM	CI ADT	Elaboration of a transformative CI usage workflow
04:30 PM	Oscar Schofield	Final feedback session Workshop Adjourns

E List of Participants

Name	Organization	Project Role
Arrott, Matthew	UCSD/Calit2	Project Manager
Chao, Yi	NASA JPL, Pasadena, CA	Domain Scientist, Subsystem Lead
Chave, Alan	WHOI	System Engineer
Cornuelle, Bruce	Scripps Institution of Oceanography	Domain Scientist
Farcas, Claudiu	UCSD/Calit2	System Modeler
Farcas, Emilia	UCSD/Calit2	System Modeler
Klacansky, Igor	UCSD/Calit2	System Modeler
Kleinert, Jack	Raytheon	System Engineer Support
Krueger, Ingolf	UCSD/Calit2	System Architect
Meisinger, Michael	UCSD/Calit2	System Modeler
Moore, Andrew	University of California, Santa Cruz	Domain Scientist
Schofield, Oscar	Rutgers	Project Scientist
O'Reilly, Bill	University of California, Berkeley	Domain Scientist
Washburn, Libe	University of California, Santa Barbara	Domain Scientist

Figure 11: Participant Group Picture at UCSD's Calit2

F Abbreviations

Abbreviation	Meaning
CI	OOI CyberInfrastructure
CI ADT	OOI CyberInfrastructure Architecture and Design Team
CI IO	OOI CyberInfrastructure Implementing Organization
ESMF	Earth System Modeling Framework
IOOS	Integrated Ocean Observing System
LAS	Life Access Server
LBSFI	Littoral Battlespace Sensing, Fusion and Integration
MPI	Max-Planck Institut
NCEP	National Centers for Environmental Prediction
NetCDF	Network Common Data Form
OOI	Ocean Observatories Initiative
OSSE	Observing System Simulation Experiment
PDR	Preliminary Design Review
ROMS	Regional Ocean Modeling System
SCCOOS	Southern California Coastal Ocean Observing System
WAM	WAve prediction Model

G References

Reference	Citation
[CI-CARCH]	CI conceptual architecture and initial requirements, available at http://www.orionprogram.org/organization/committees/ciarch
[CI-PAD]	OOI CI Architecture Document, PDR Final version, 16-Nov-2007
[CI-RWS1]	OOI CI First Science User Requirements Elicitation Workshop Report, OOI CI, Final version 1.0, 08-Nov-2007, available at: http://www.ooici.ucsd.edu/spaces/download/attachments/10453181/OOI-CI-ReqWS1-Report-FINAL.pdf?version=1
[CI-RWS2]	OOI CI Second Science User Requirements Elicitation Workshop Report (this document)
[CI-WEBSITE]	OOI CI Website, available at http://www.ooici.uscd.edu
[NORIA]	Network for Ocean Research, Interaction and Application (NORIA) Proposal, 22-Dec-2006
[RWS2-WEB]	Second OOI CI User Requirements Workshop, website, available at: http://www.ooici.ucsd.edu/spaces/display/MCW20080123/Home
[SCIPROSP]	OOI Science Prospectus, Dec 2007, available at: http://www.oceanleadership.org/files/Science_Prospectus_2007-10-10_lowres_0.pdf