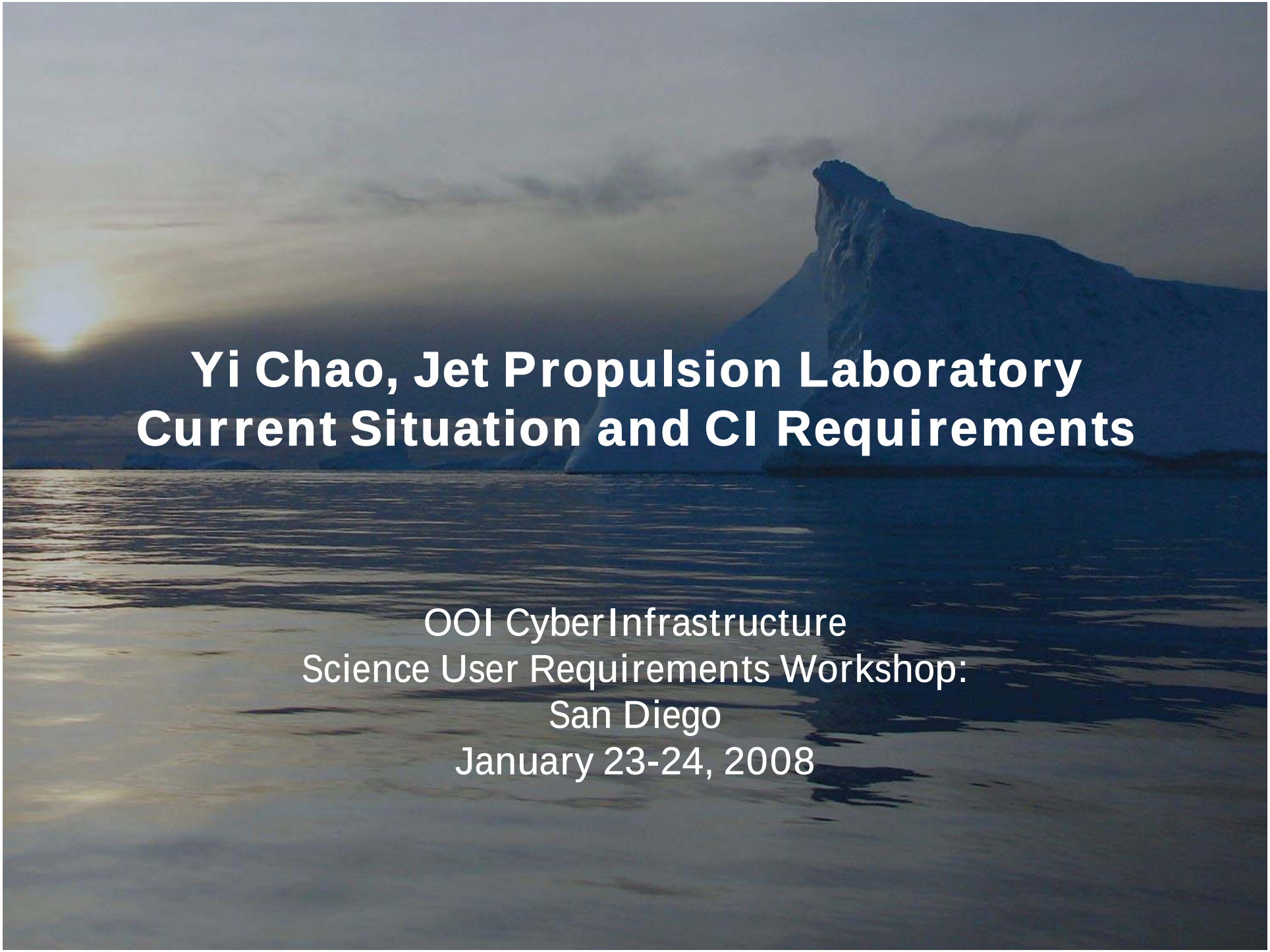
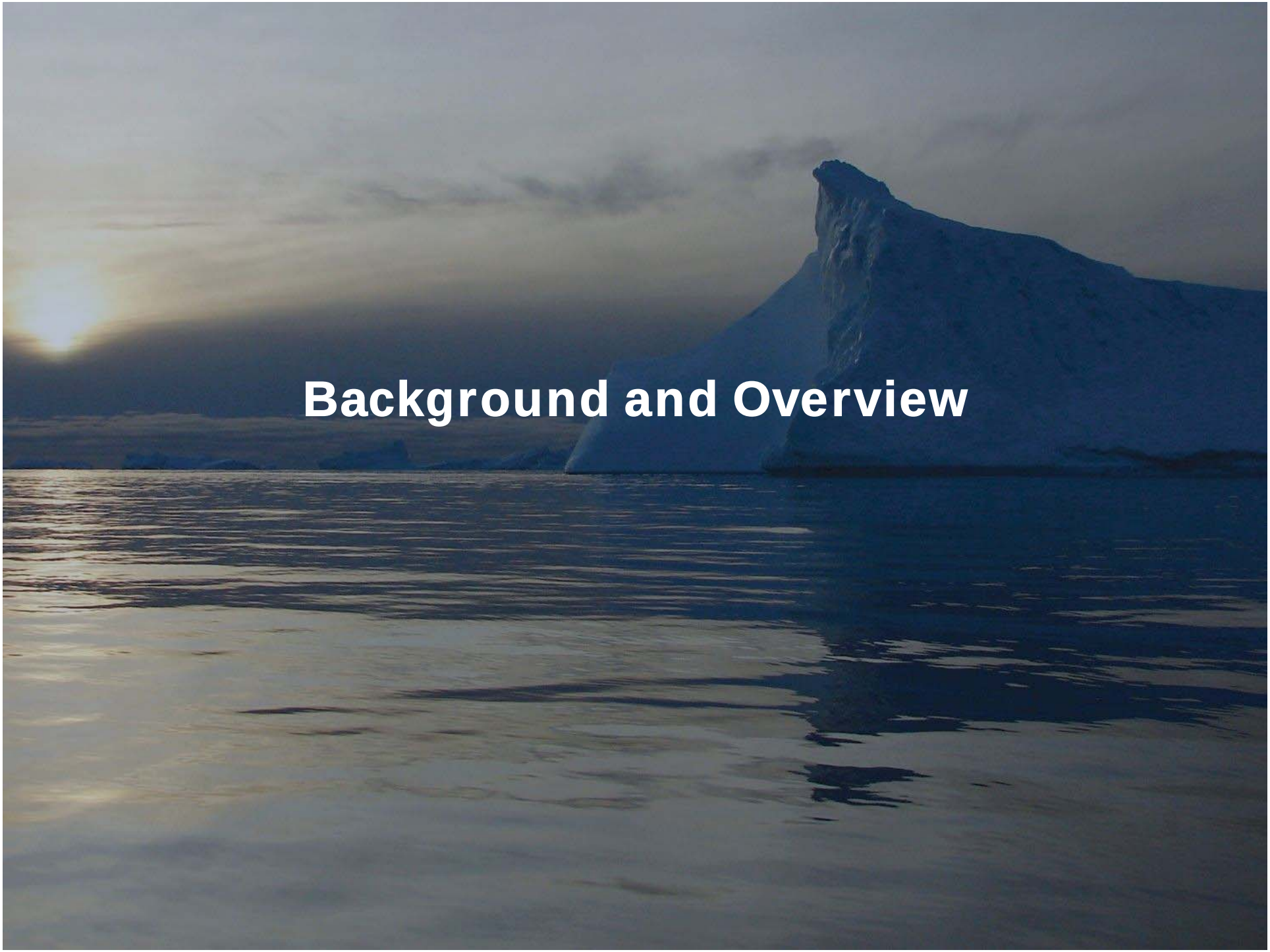


Yi Chao, Jet Propulsion Laboratory
Current Situation and CI Requirements

OOI CyberInfrastructure
Science User Requirements Workshop:
San Diego
January 23-24, 2008


Background and Overview

Intent of this template

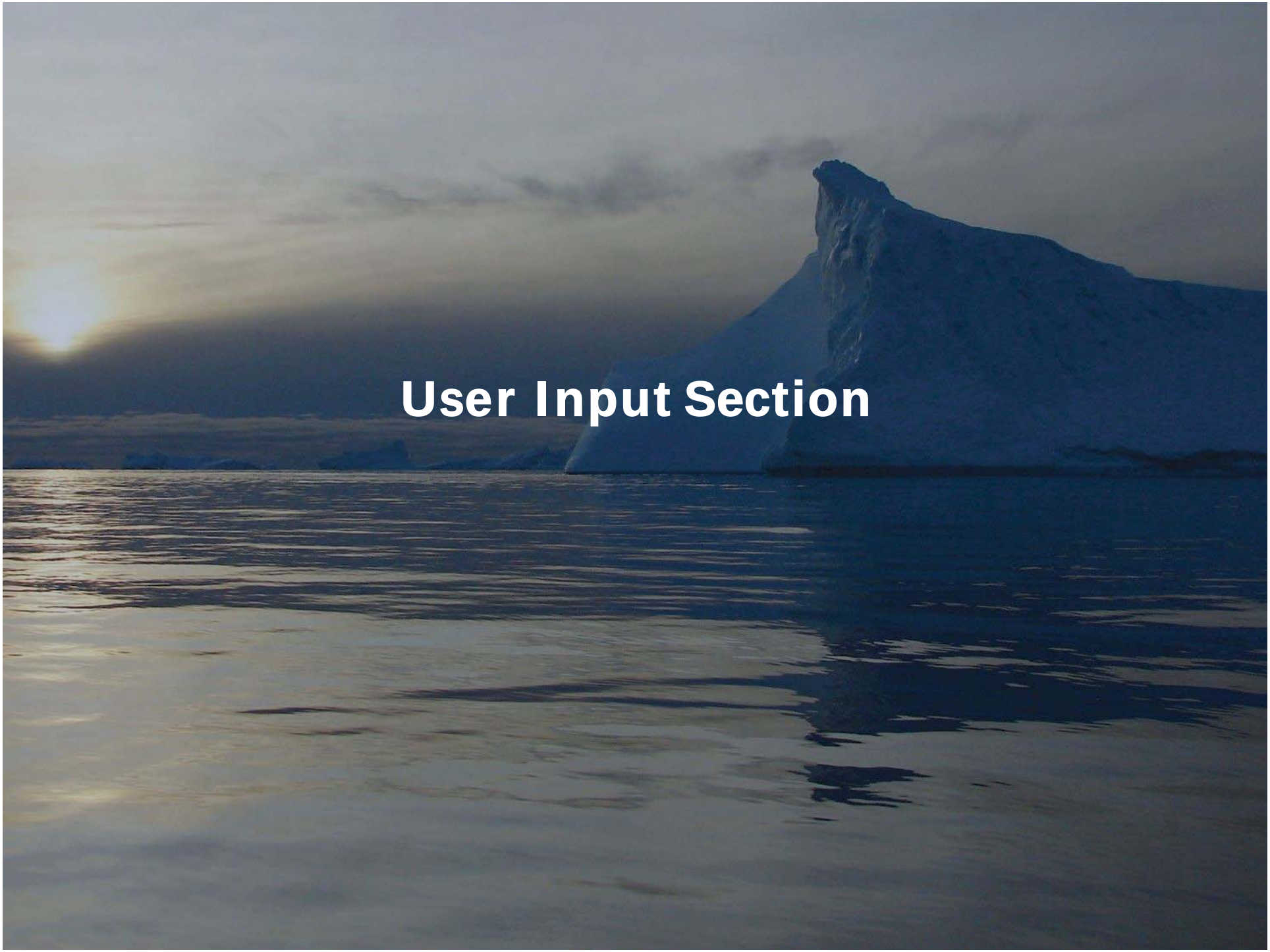
- This slide set is a template for participants of the OOI CI requirements workshop in San Diego, January 2008
  - For presentations during the workshop
  - To capture relevant information in a structured way
- Goals of this exercise are
  - To capture as many CI relevant details as possible before the 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


User Input Section

Current situation

- Please briefly describe your current situation, e.g. every-day tasks in numerical modeling and related activities (overview)

- Real-time modeling for forecasting
  - On the dedicated computer every six hours, we obtain all the data (in situ, land-based, and satellite) obtained from a variety of data servers, assimilate them into a numerical model, and produce a nowcast (also known as analysis).
- Batch-job modeling for research
  - On the supercomputers at JPL or NASA Ames Research Center, we run 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 an aim to yield the best Agreement between model simulation and data.

Expected changes

- What changes do you expect for the next 3-5 years?

• Data management. One of the biggest challenges is to manage 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 data of a few months or even a few years will require going through TB of data and extract the KB 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 on a particular CPU. It is desirable to analyze the 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 by the right tools. Most 3D visualization tools are on the high end, and usually very difficult to use by general users. It would be highly desirable to have access some simple-to-use, user-friendly 3D visualization tools.

Expected changes

- What changes do you expect for the next 3-5 years?

• A single portal with all the available modeling codes, documentations, and users' experiences (shared via wiki for example); it is also desirable to have experts available for questions and hand-on help.
• A single portal with all the available data assimilation schemes and the associated documentations.

Expected changes

- What transformative changes do you envision and/or anticipate for a 5-10 year time frame?

• Users should have ways to know what model output are available in what regions, and download them if needed.
• Users should be able to interact with models directly via a simple web interface (on-demand modeling).

Expected changes

- What capabilities do you expect from a transformative cyber-infrastructure in the oceanographic domain?

• Access to unlimited distributed data storage on the network
• Ability to use the unlimited computing capability on the GRID/virtual-computer

Expected changes

- How would you use these capabilities if they were in place?
  - Don't know.
- What could and/or would you provide to the community as part if the infrastructure (e.g. data, tools, algorithms)?
  - Modeling experiences in the areas of
    - Real-time modeling
    - On-demand modeling
- Are there any similar projects/communities that you like and/or that are technology-wise exemplary?
  - No.
- What general developments would advance you/the community most?
  - Sharing component models, data assimilation modules, and visualization tools

Definition of Terms

- How do you define “model” or “numerical model”?
  - 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.
- How do you define “data”?
  - Data are the information coming from observing platforms including in situ and remotely sensed (land-based or spacecraft)
- How do you define “meta-data”?
  - Meta-data describes the data.
- How do you define “workflow” resp. “process”?
  - Workflow represents a sequence of commands and instructions that process data, run models and manage the output.

Models

- What kind of models exist in your community? Or should be there?
  - MOM, POP, HYCOM, ROMS, POM, FVCOM, MITGCM
- Which models do you use and/or develop?
  - I used MOM, POP; now I use ROMS
- Please explain the specifics of (some of) these models
  - ROMS
    - 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

Models

- What kind of models exist in your community? Or should be there?
  - MOM, POP, HYCOM, ROMS, POM, FVCOM, MITGCM
- Which models do you use and/or develop?
  - I used MOM, POP; now I use ROMS

Models

- 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?

Models

- Please explain the specifics of (some of) these models
  - ROMS-coastal
    - Size of the model
      - Typical 300-km by 300-km square with 1-km resolution
    - Parameterization possibilities
      - Vertical mixing schemes
      - Horizontal mixing
    - Number and type of input variables
      - 5: temperature, salinity, east-west current, north-south current, sea level
    - Output variables or grid points, per time
      - all
    - Output data volume
      - 2 GB/month
    - Complexity of the model execution workflow
      - Many steps involved to setup the model configuration (3D grid, bottom bathymetry, open boundaries), prepare the initial condition and boundary conditions including surface and side, test and run the model, and analyze the output

Model Processing

- Please detail some model execution characteristics
  - How often do you run the model?
    - Every day
  - How long does it take to run the model?
    - 6-8 hours on a 16 processors SGI Altix
  - How often does the model change? Are changes parametric or algorithmic?
    - We try to minimize the change for operational models, while the research models are constantly changing. Most parametric, although occasionally algorithmic
  - What are the execution platforms? Do the models have specific technology dependencies (e.g. compilers, platforms, libraries, computation resources)
    - The nested model required a shared-memory cluster using OpenMP (e.g., SGI), while the single domain model can be run on most cluster computers using MPI
    - netCDF or HDF library for I/O
  - Would your models benefit from parallelization and/or super-computing?
    - yes
  - Could you run your models on a remote common infrastructure? Would you?
    - yes
- Do you use external on-line resources (e.g. computation grids, data archive)?
  - yes
- Do you support on-line processing?
  - If so what is your concept of real time?
    - Entire processing time is less than 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?
    - both

Model Output, Visualization

- How do you store, publish, announce, and describe your model results?
  - In files, FTP or LAS/OpenDAP, email notification
- Do you provide different versions of the same data (e.g. lo-res, high-res, or filtered)?
  - Yes, depending upon the users' requests
- How often do you envision to update outputs?
  - regularly
- Do you envision revisions to data? How often does this occur in practice?
  - Delayed data with better QCs; sometimes
- Which meta-data do you associate with output data and how?
  - Standard netCDF header
- What visualizations do and/or the community apply?
  - Matlab, IDL, ncview, GMT
- How could a common infrastructure support (interactive) visualization?
  - Visualize the remote data

Data, Data Sources

- What are the stages data undergoes from raw data to output data? E.g. filtering, processing, down-sampling, aligning steps
  - Spatial interpolation
- What data should be stored and backed up by a common infrastructure and when?
  - Only the final version to be distributed broadly
- Who has “ownership” of data in different steps?
  - Data collectors
  - Modeler and assimilators
  - Operational centers
- What are typical data exchange formats?
  - netCDF
- What meta-data is relevant to find the right data source? Are there specific meta-data standards used?
  - netCDF header
  - Read program

Data, Data Sources

- What quality/reliability/accuracy/certification levels for data exist and how do you select if you have the choice?
  - Critical needs, how difficult to reproduce
- Which specific data sources do you use? How did you find them? How did you get access?
  - Public available data sets
  - Via friends, community connections, various data centers, Google
  - FTP
- Do you have backup sources for the same data in case of unavailability?
  - No backup for the public data, which can be always downloaded again if needed
- Which manual interaction is required to check/validate/modify the data?
  - Size, content, numerical numbers
- What data volumes do you handle and/or anticipate? Any high-bandwidth data streams?
  - GB to TB via Gbits network

Data, Data Sources

- Do you use streamed data or bulk data files or databases?
  - Data files, database
- What data filters and/or transformations do you apply?
- Which (external) tools do you use for data processing, transformation, etc.
  - Most programs are developed in house
- 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?
  - Six hours
  - Not yet. As we upgrade our data assimilation scheme, reducing the assimilation window for 3DVAR or moving to 4DVAR, we will need the streamed data.
- Do you have example data files? Meta-data files?
  - Yes, can be provided upon request

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?
  - Web, cell-phone/PDA, phone, fax
- Should there be any other interactive interfaces besides web interfaces?
  - Some users don't have internet access
- Which user interface technologies are particularly efficient for your daily work?
  - web
- 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?
  - Interactive
  - Too many clicks to get the information you are looking for (3 clicks might be the limit)
- What programmatic and application interfaces of the OOI CI do you envision and/or require?
  - Data subsetting, data mining, data analysis
  - Basic visualization
- Do you need off line access capability?
  - yes
- 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?
  - No
- Is there a need for data embargoing for certain time frames?
  - Some new data and model products might need some times to validate before widely distributing
- 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?
  - Properly acknowledged when giving talks and publishing papers
- 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?
  - RAID disks in my lab; archive disk/tape in computer centers
- How do you store your data results?
- How do you manage different versions of data/models?
  - directories
- Is there a responsible contact person available for operation/maintenance?
  - yes
- How much of the total effort is required for operation and maintenance of hardware, models, network etc.?
  - Web programmer
  - Database/IT programmer
  - Model developer
  - Data assimilation developer
  - Workflow and process programmer
  - Researchers for data and model analysis
- 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?
  - Work closely with the educators to design the data product required
- How do education and outreach concerns affect your models and the presentation of the results?
  - Scientists always keep the data too complex for educational users, therefore input from educational expert throughout the process is the key
- 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?
  - Develop a separate educational data portal, different from the science data portal
- Do you consider releasing models, algorithms, tools as open source for the public? How does this affect your work?
  - Yes, more feedback from users

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
    - Data interpolation capability from one grid to the next for model-data and model-model comparisons
  - Any data formats needed for processing, transfer?
    - Programs to convert data from one format to the next, e.g., netCDF to ascii

Requirements Summary

- List the 3 short-term advances that would benefit you most
  - Web-based matlab tools
  - Basic graphic tools
  - Data and model subsetting tools
- List the 3 mid-term advances that would benefit you most
  - Easy access data and models
  - Interactive, on-demand modeling capability
  - Model components sharing so as to combine them into an optimal model for a particular application
- List the 3 impediments for you/the community currently
  - Data and model products are not transparent to users
  - Too many models to select from, no best model
  - Model output too large to analyze and visualize
- 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

Thanks!

- Thanks a lot for your important contributions!