# NSF/ONR Workshop on Data Assimilation in Ocean Research

# LOOPS/Poseidon: A Distributed System for Real-Time Interdisciplinary Ocean Forecasting with Adaptive Modeling and Sampling

P.F.J. Lermusiaux, C. Evangelinos, P.J. Haley Jr, W.G. Leslie, N.M. Patrikalakis, A.R. Robinson, R. Tian

PIs: N.M. Patrikalakis, J.J. McCarthy, A.R. Robinson, H. Schmidt
Scientists: C. Evangelinos, P.J. Haley Jr., P.F.J. Lermusiaux, R. Tian

http://czms.mit.edu/poseidon

# Ocean Science and Data Assimilation

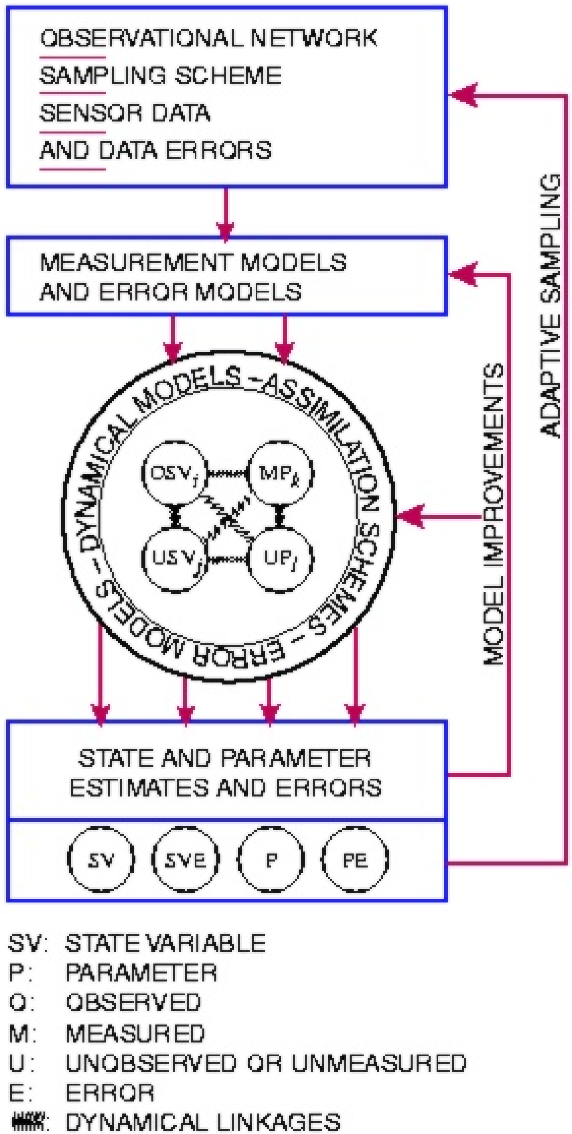



Fig. 1. Data assimilation system schematic

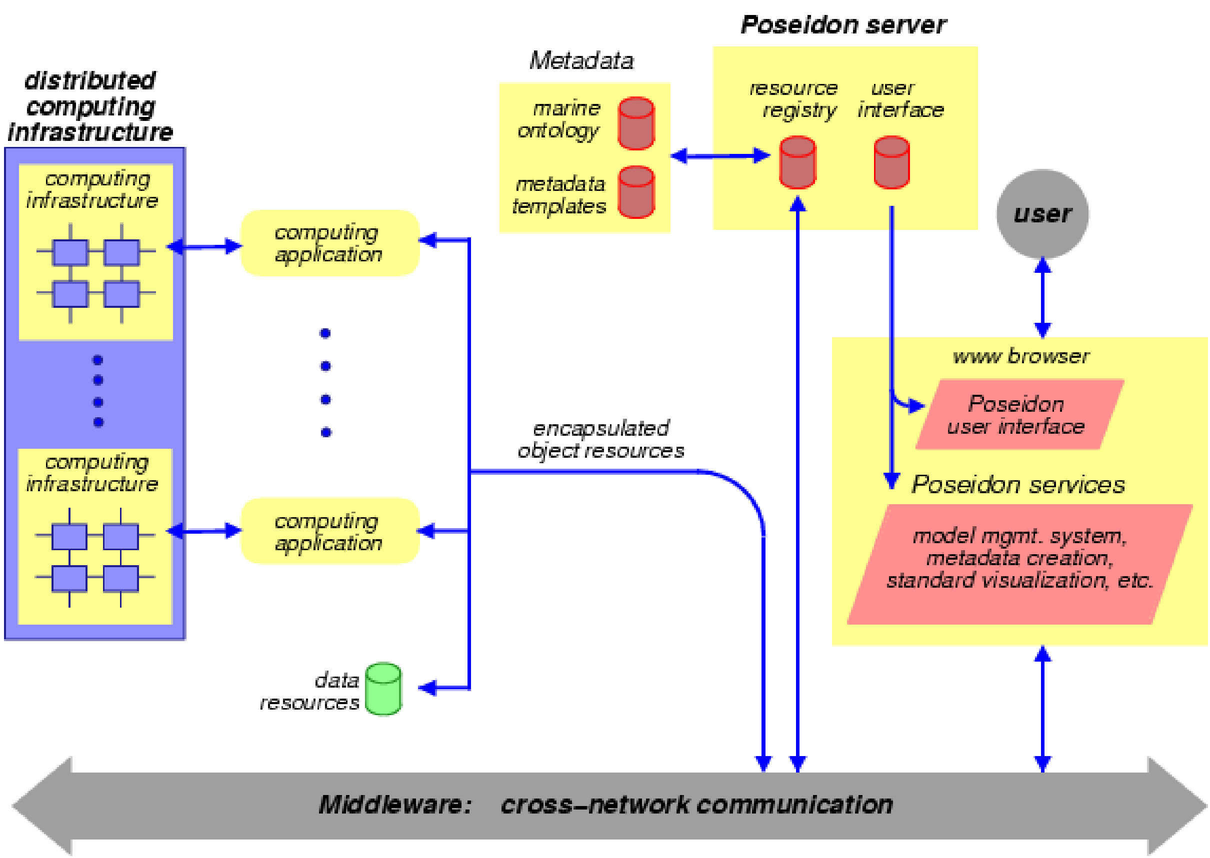
- Field and remote observations
- Models:
  - – Dynamical
  - – Measurement
  - – Error
- Assimilation schemes
- Sampling strategies
- State and parameter estimates
- Uncertainty estimates
- *A Dynamic Data-Driven Application System (DDDAS)*

# LOOPS/Poseidon
## Adaptive Interdisciplinary Ocean Forecasting in a Distributed Computing Environment

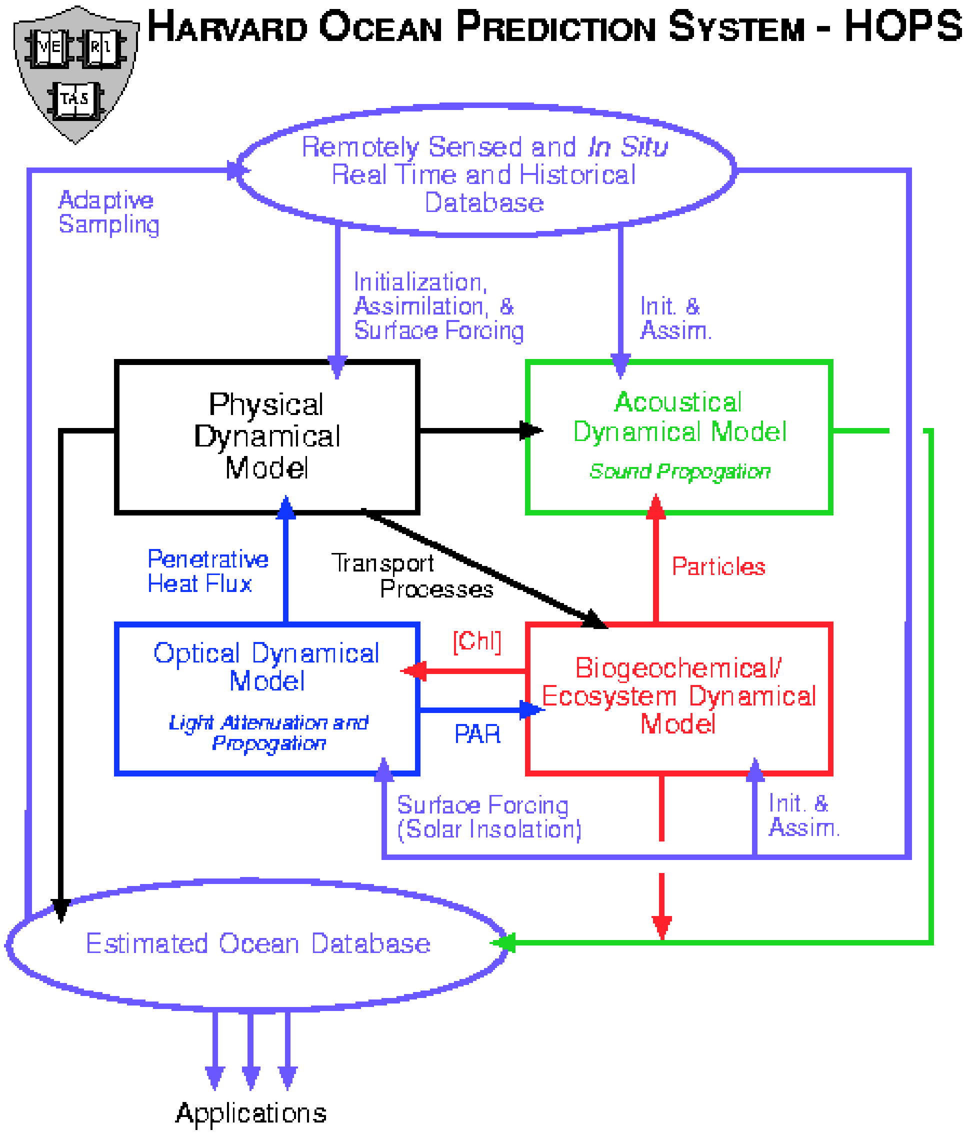
- Research coupling Physical and Biological Oceanography with Ocean Acoustics.
- More effective Real-Time Ocean Forecasting for Naval and Maritime Operations, Pollution Control, Fisheries Management, Scientific Data Acquisition, etc.
- MIT OE (IT, Acoustics) and Harvard DEAS (Ocean Physics-Biology-Acoustics).

**Key points**

- Web interface
- Remote visualization
- Metadata for code and data
- Metadata/Ontology editors
- Legacy application support
- Grid computing infrastructure
- Transparent data access
- Data assimilation (ESSE, OI)
- Interdisciplinary interactions
- Adaptive modeling
- Adaptive sampling
- Feature Extraction
- Prototype for community-use

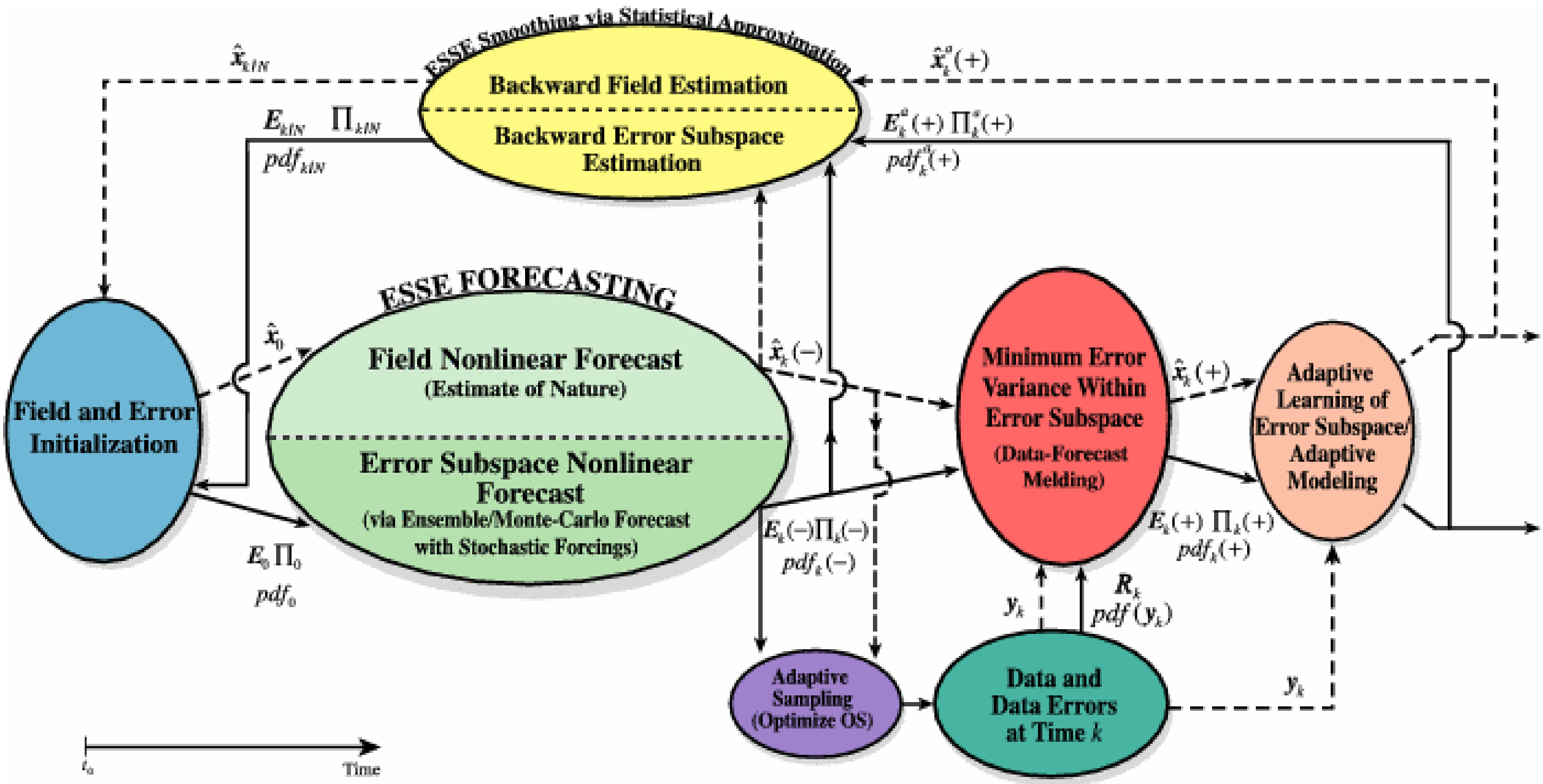


# Physical-Biological-Acoustical Oceanography with HOPS




- Primitive Equation (PE) physical dynamics model
- Multiple biological models
- Interfaces to acoustical models
- Adaptable to different domains
- Nested-domains parallelism
- Software: F77-matlab-C
- I/O: NetCDF, stdin

# Error Subspace Statistical Estimation (ESSE)



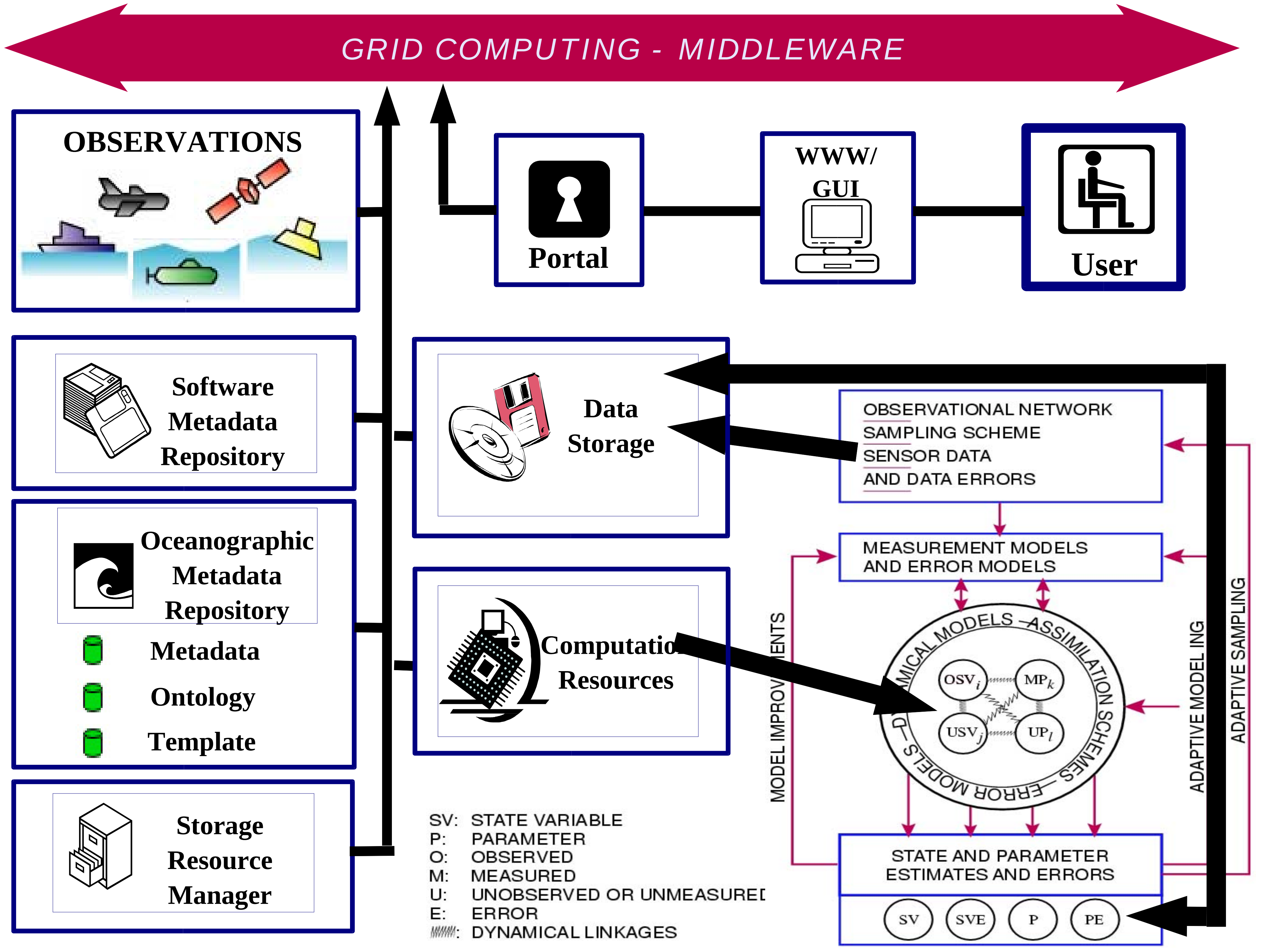

- Uncertainty forecasts (with dynamic error subspace, error learning)
- Ensemble-based (with nonlinear and stochastic model)
- Multivariate, non-homogeneous and non-isotropic DA
- Consistent DA and adaptive sampling schemes
- Software: not tied to any model, but specifics currently tailored to HOPS

# IT Design Motivations

- Real-time predictions of **interdisciplinary** ocean fields and *uncertainties*
  - Data Assimilation (DA) using ESSE is currently ensemble-based and thus ideal for high throughput distributed computing
  - Interdisciplinary interactions and multiscale/nested simulations ideal for parallel computing
- Develop autonomous adaptive models for physics & biology
  - Adaptive parameter values, model structures and state variables
  - Error metrics and criteria for adaptation
- Towards automated, distributed management of observed and modeled data
  - Consistent use of metadata helps provide transparent data management, including quality control
  - Forecasting workflow is being automated, including DA
- Web access from lightweight clients eases operational use and system control
- Interactive visualizations for better understanding and decision-making

# Software Strategies

- Exploit parallelism (especially throughput) opportunities
- Maximize performance, facilitate users, but limited changes
  - For new generalized adaptive biological model: MPI coding
  - For existing software: automate file I/O based workflows
  - Work to the maximum extent possible at the binary level
  - Metadata for software use (and installation) in XML
- Use Grid technologies
  - For user: compute and data access solutions
  - Drive forecasting, visualization workflows on the Grid
  - Present results to user's web browser


GRID COMPUTING - MIDDLEWARE
OBSERVATIONS
Portal
WWW/ GUI
User
Software Metadata Repository
Data Storage
Oceanographic Metadata Repository
Metadata
Ontology
Template
Computational Resources
Storage Resource Manager
OBSERVATIONAL NETWORK
SAMPLING SCHEME
SENSOR DATA
AND DATA ERRORS
MEASUREMENT MODELS
AND ERROR MODELS
DYNAMICAL MODELS – ASSIMILATION SCHEMES – ERROR MODELS
OSV
MP
USV
UP
MODEL IMPROVEMENTS
ADAPTIVE MODELING
ADAPTIVE SAMPLING
STATE AND PARAMETER ESTIMATES AND ERRORS
SV
SVE
P
PE
SV: STATE VARIABLE
P: PARAMETER
O: OBSERVED
M: MEASURED
U: UNOBSERVED OR UNMEASURED
E: ERROR
DYNAMICAL LINKAGES

# Interdisciplinary Data Assimilation (DA)

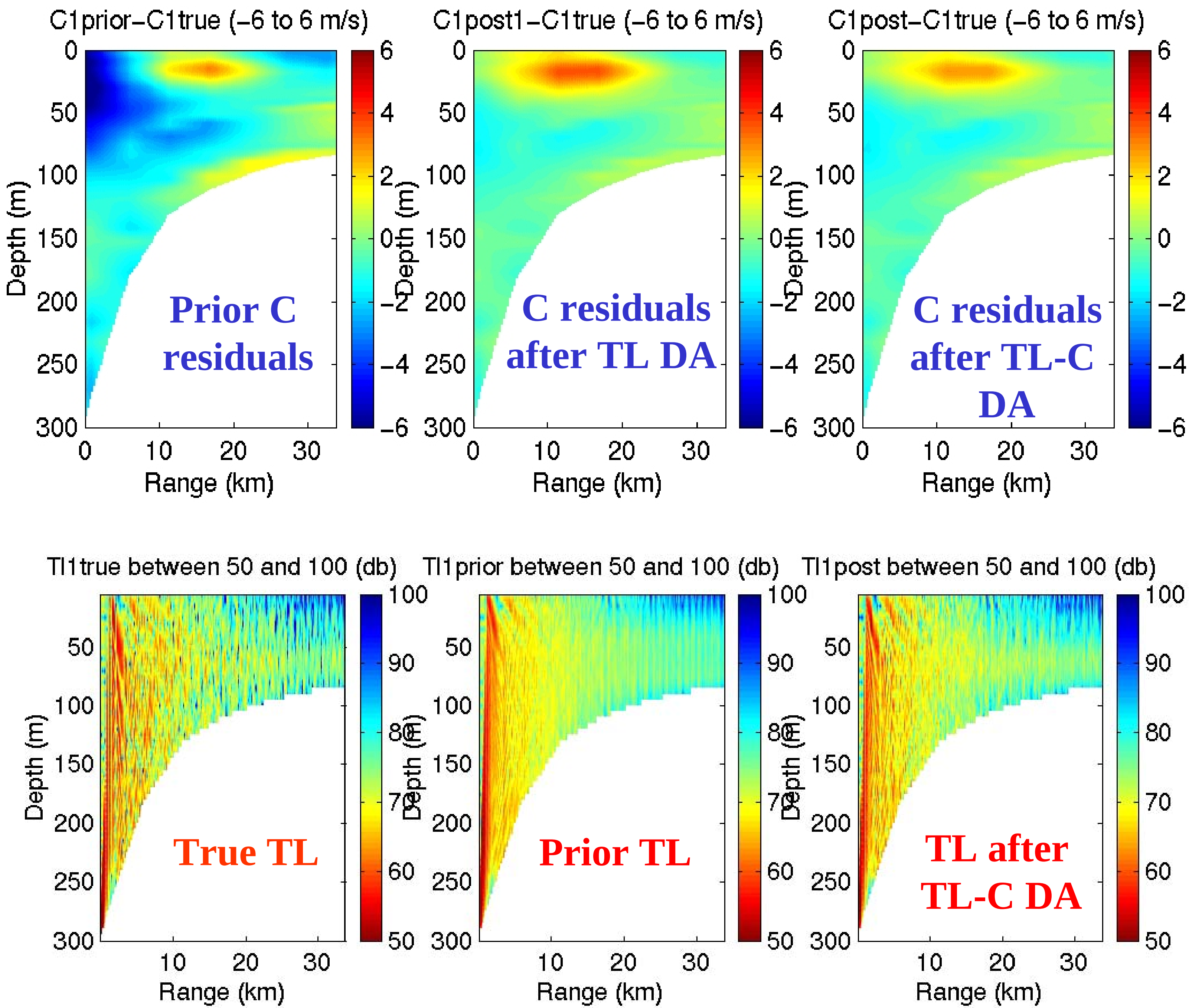
- **Is in its infancy, but can contribute significantly to understanding physical-acoustical-biogeochemical processes, including quantitative development of fundamental models**
- **Required for interdisciplinary ocean field prediction and parameter estimation**
- **Model-model, data-data and data-model compatibilities are essential**
- **Care must be exercised in understanding, modeling and controlling errors and in performing sensitivity analyses to establish robustness of results**
- **Dedicated interdisciplinary research needed**

# Coupled Physical-Acoustical **Filtering** via ESSE

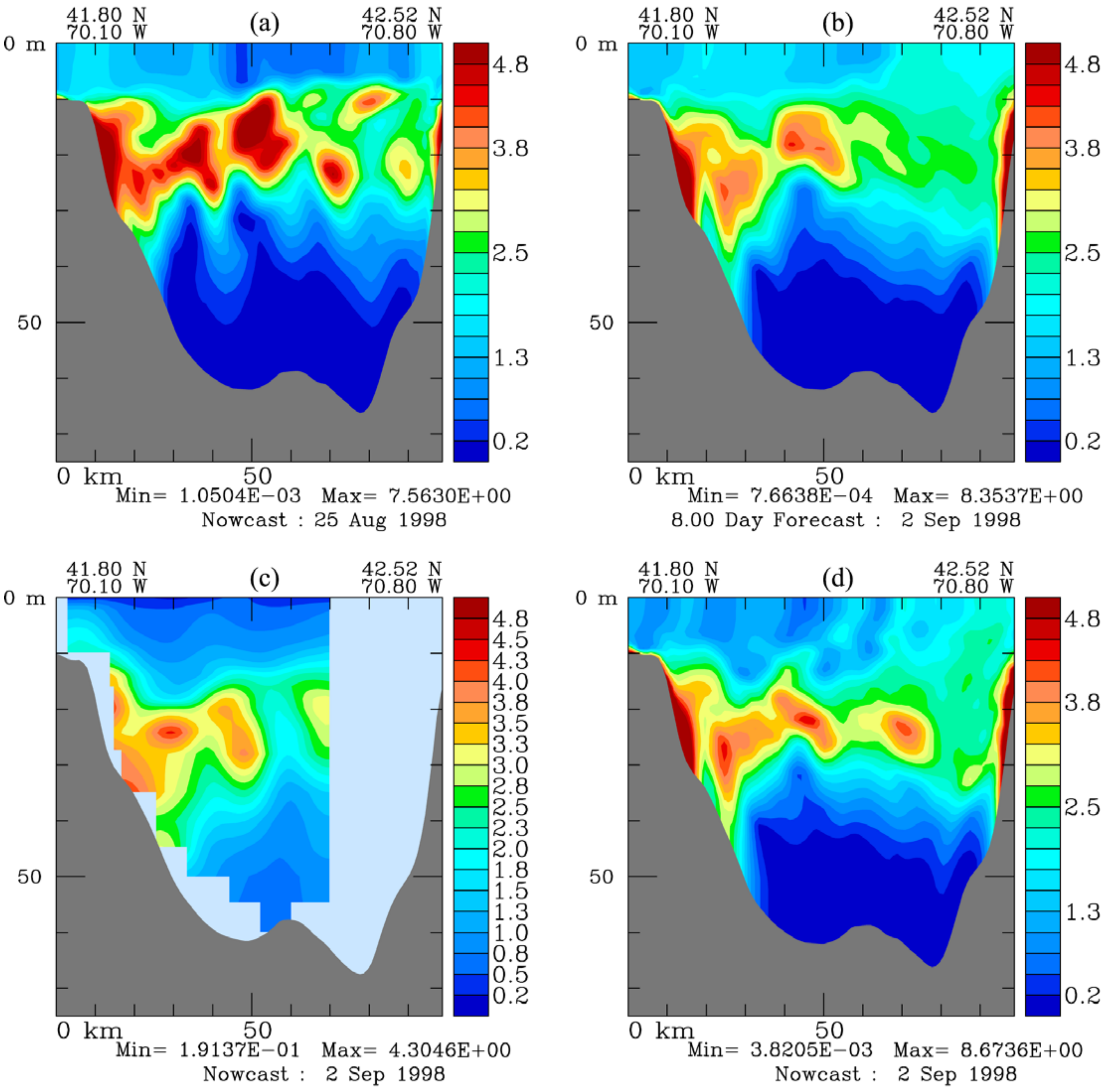
**Coupled assimilation of sound-speed and TL data for a joint estimate of sound-speed and TL fields**

**Twin-experiments:**

- **"Truth" ocean physics assimilates natural data**
- **Provides 3 CTDs**
- **Corresponding TL "truth" provides towed-receiver TL data, every 500m at 75m depth**



# Coupled Physical-Biogeochemical **Smoothing** via ESSE




**Cross-sections in Chl-a fields, from south to north along main axis of Massachusetts Bay, with:**

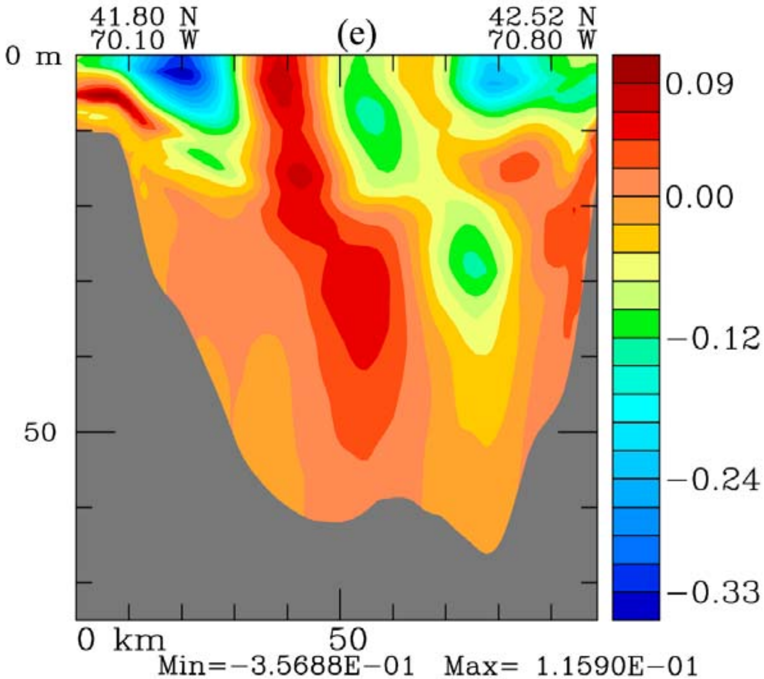
**a) Nowcast on Aug. 25**

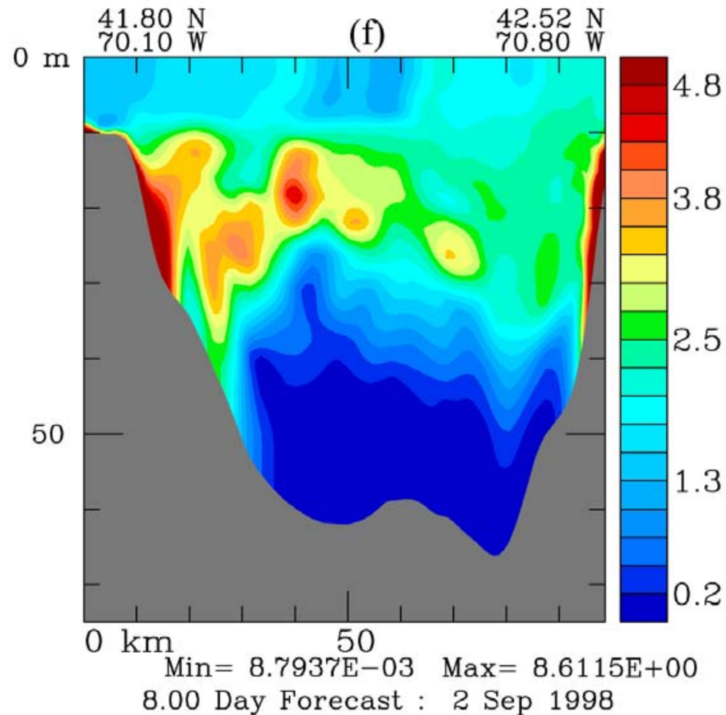
**b) Forecast for Sep. 2**

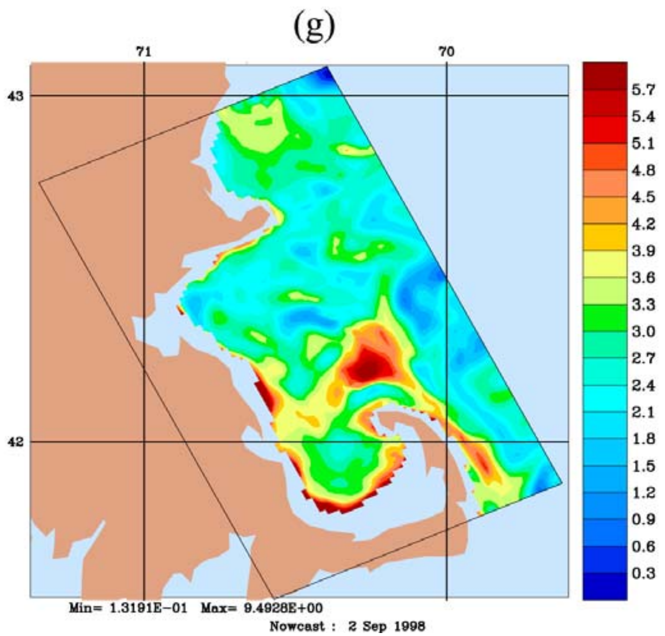
**c) 2D objective analysis for Sep. 2 of Chl-a data collected on Sep. 2–3**

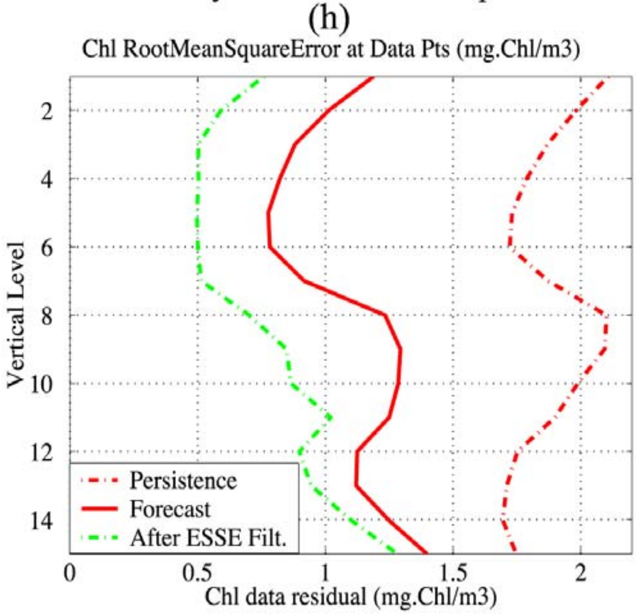
**d) ESSE filtering estimate on Sep. 2**

# Coupled Physical-Biogeochemical DA via ESSE (continued)



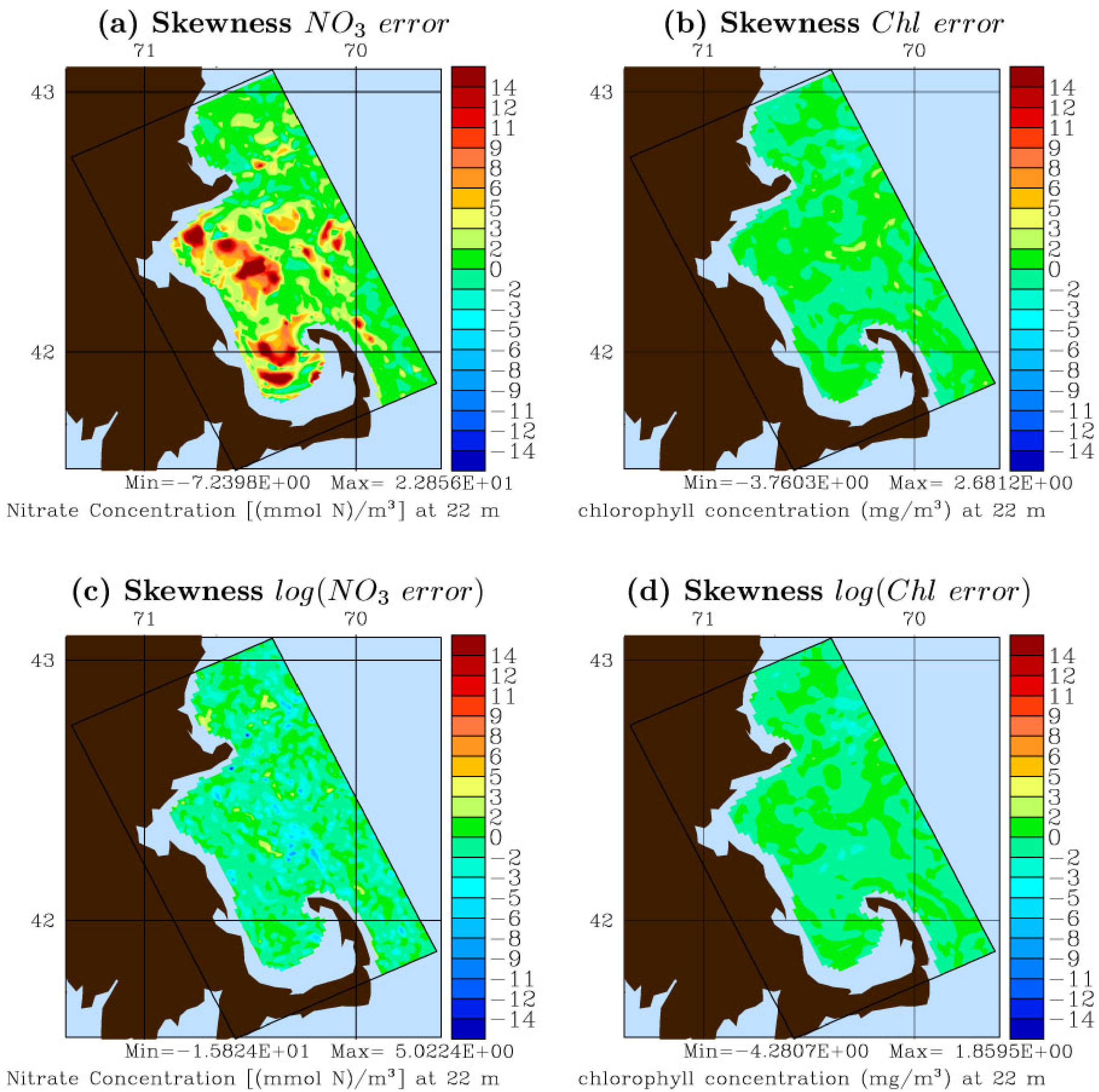

**e) Difference between ESSE smoothing estimate on Aug. 25 and nowcast on Aug. 25**

**f) Forecast for Sep. 2, starting from ESSE smoothing estimate on Aug. 25**

**(g): as d), but for Chl-a at 20 m depth**

**(h): RMS differences between Chl-a data on Sep. 2 and the field estimates at these data-points as a function of depth (specifically, "RMS-error" for persistence, dynamical forecast and ESSE filtering estimate)**

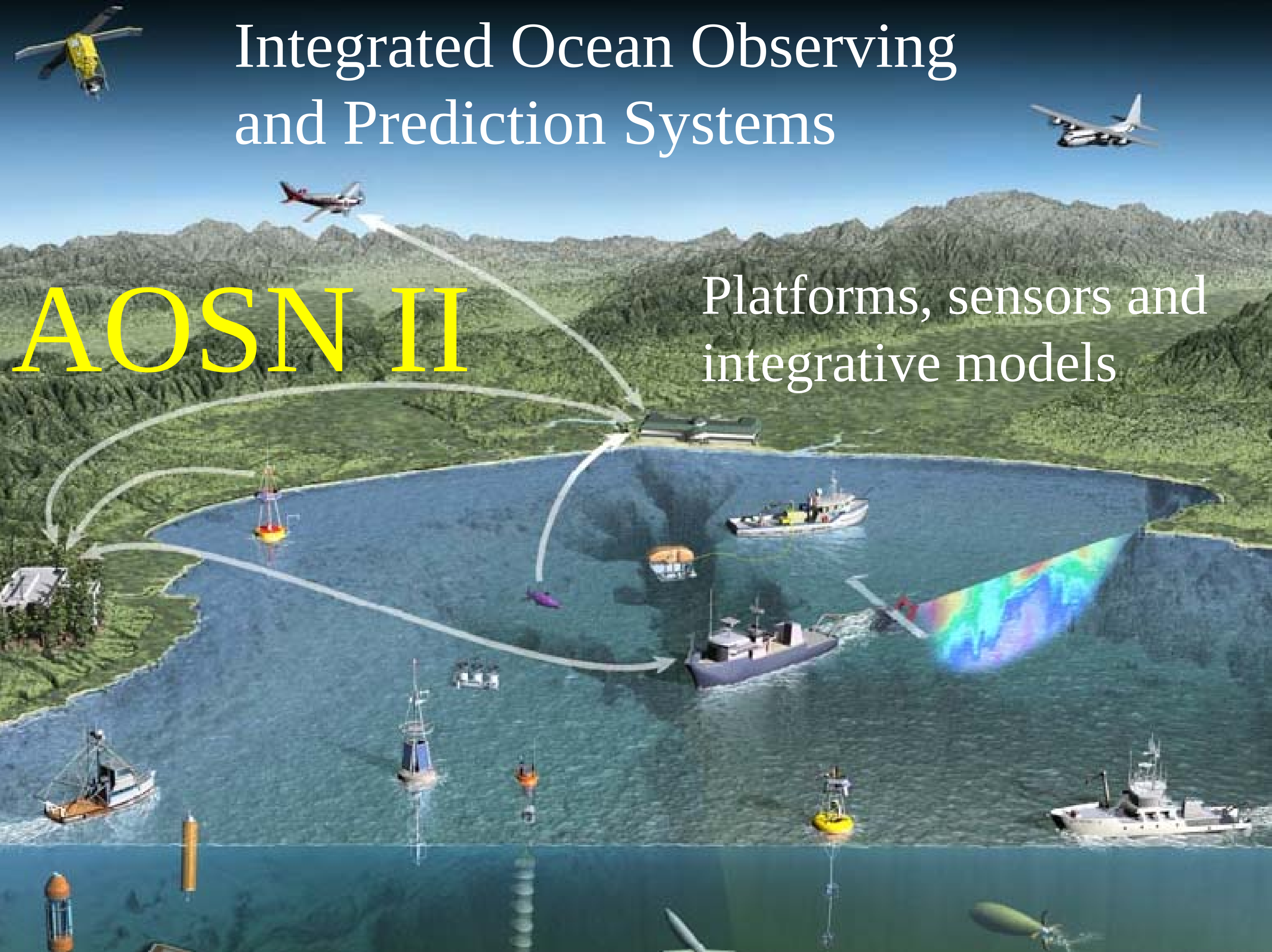
# How Gaussian are biogeochemical error forecast distributions?



# Interdisciplinary Adaptive Sampling

- Use forecasts and their uncertainties to alter the observational system in space (locations/paths) and time (frequencies) for physics, biology and acoustics.
- Locate regions of interest, based on:
  - Uncertainty values (error variance, higher moments, pdf's)
  - Interesting physical/biological/acoustical phenomena (feature extraction, Multi-Scale Energy and Vorticiy analysis)
  - Maintain synoptic accuracy
- Plan observations under operational, time and cost constraints to maximize information content (e.g. minimize uncertainty at final time or over the observation period).

# Integrated Ocean Observing and Prediction Systems



## AOSN II

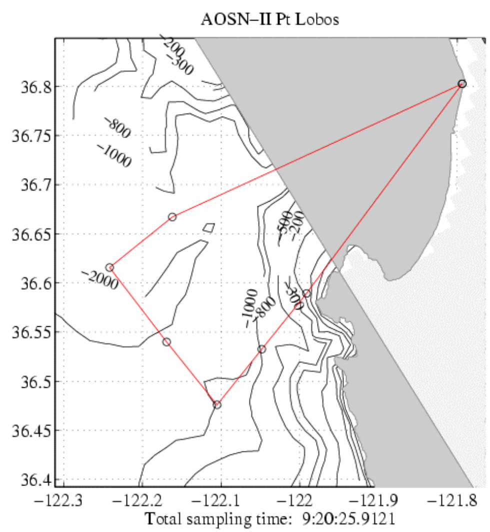
Platforms, sensors and integrative models

# HOPS/ESSE– AOSN-II Accomplishments

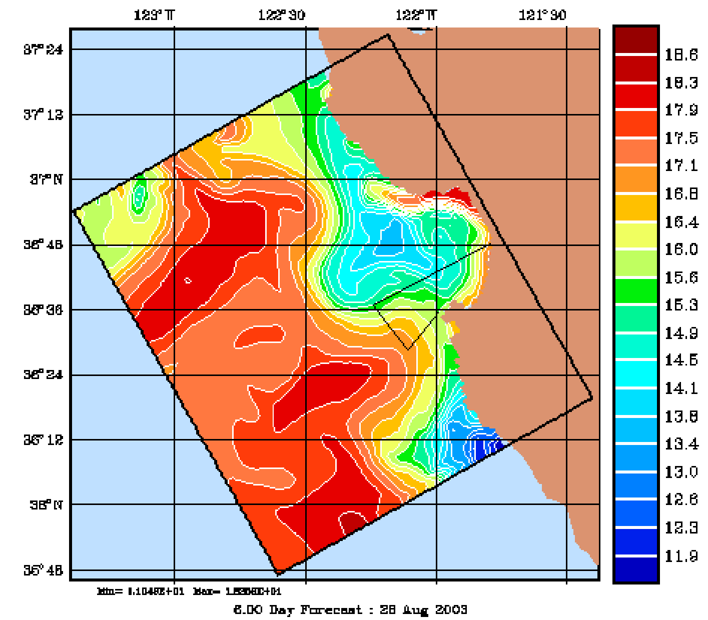
- 23 sets of real-time nowcasts and forecasts of temperature, salinity and velocity released from 4 August to 3 September
- 10 sets of real-time ESSE forecasts issued over same period: total of 4323 ensemble members (stochastic model, BCs and forcings)
- Adaptive sampling recommendations suggested on a routine basis
- Web: http://www.deas.harvard.edu/~leslie/AOSNII/index.html for daily distribution of forecasts, scientific analyses, data analyses, special products and control-room presentations
- Assimilated ship (Pt. Sur, Martin, Pt. Lobos), glider (WHOI and Scripps) and aircraft SST data, within 24 hours of appearance on data server (after quality control)
- Forecasts forced by 3km and hourly COAMPS flux predictions

# Real-time Adaptive Sampling – Pt. Lobos

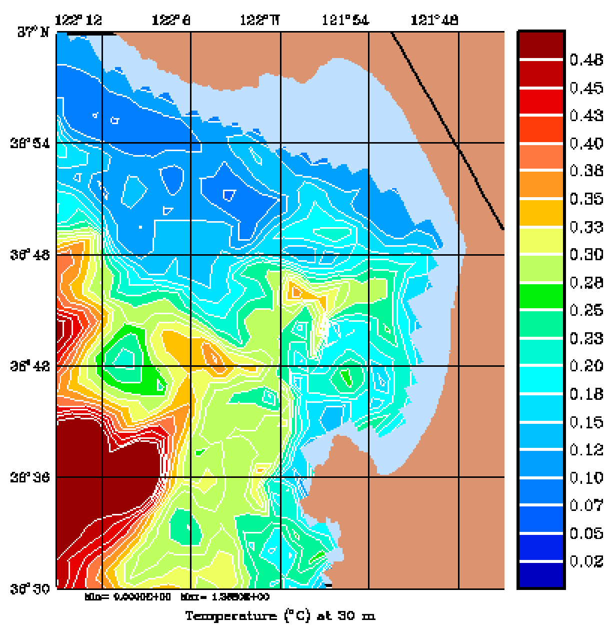
- Large uncertainty forecast on 26 Aug. related to predicted meander of the coastal current which advected warm and fresh waters towards Monterey Bay Peninsula.
- Position and strength of meander were very uncertain (e.g. T and S error St. Dev., based on 450 2-day fcsts).
- Different ensemble members showed that the meander could be very weak (almost not present) or further north than in the central forecast
- Sampling plan designed to investigate position and strength of meander and region of high forecast uncertainty.



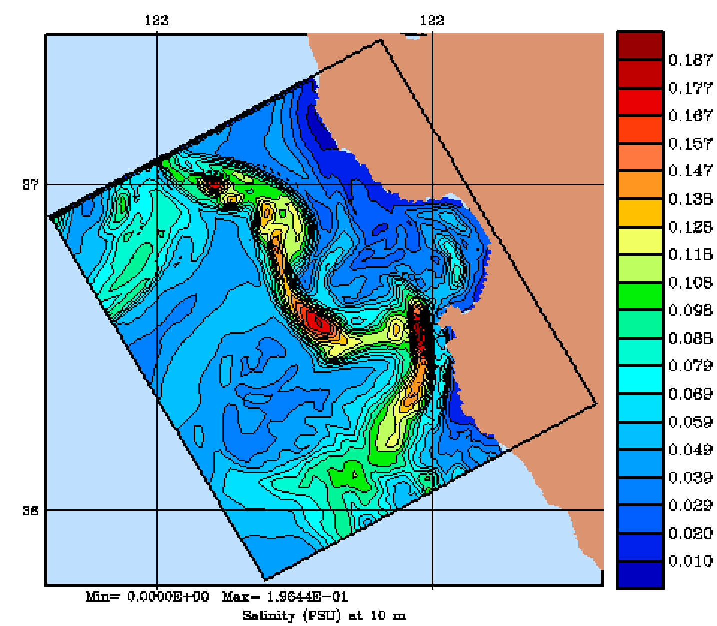

## Surf. Temperature Fct



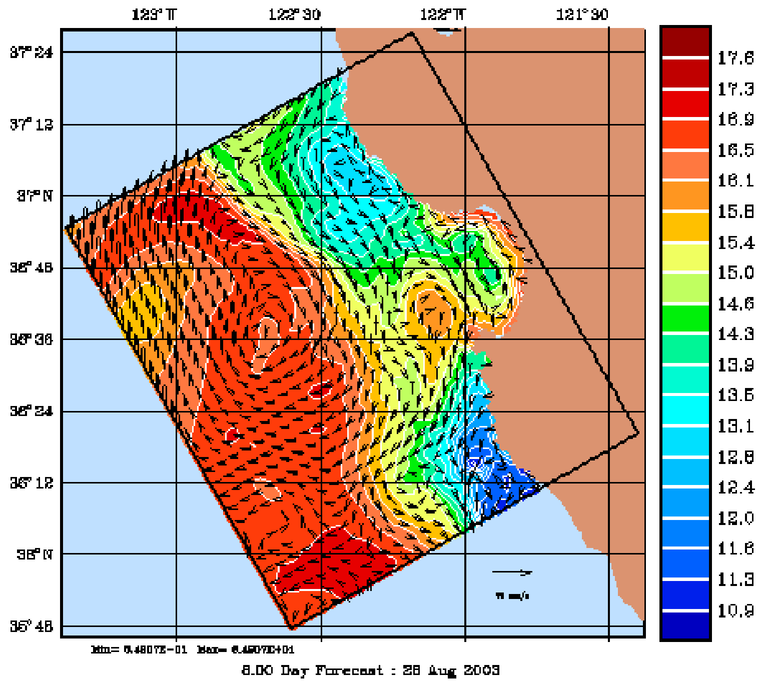

## Temperature Error Fct

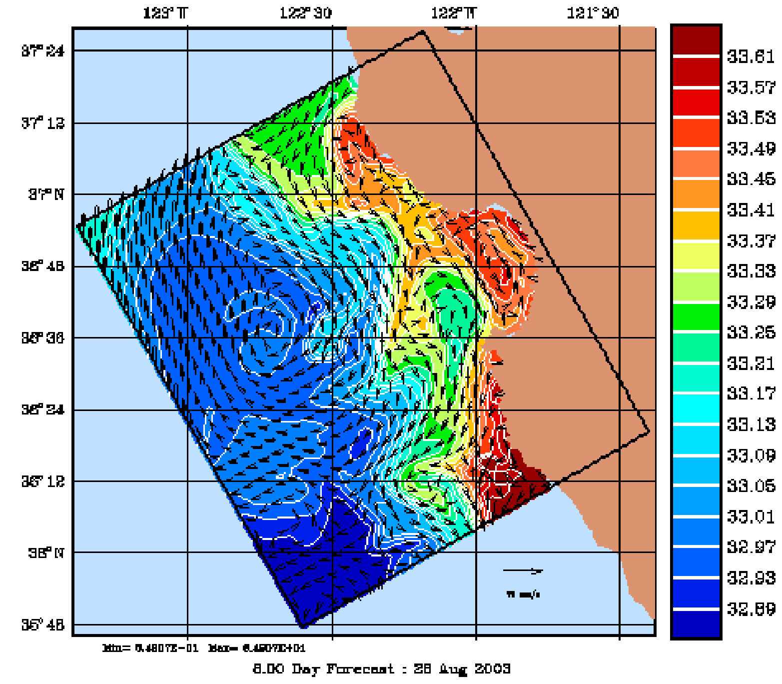



## Salinity Error Fct

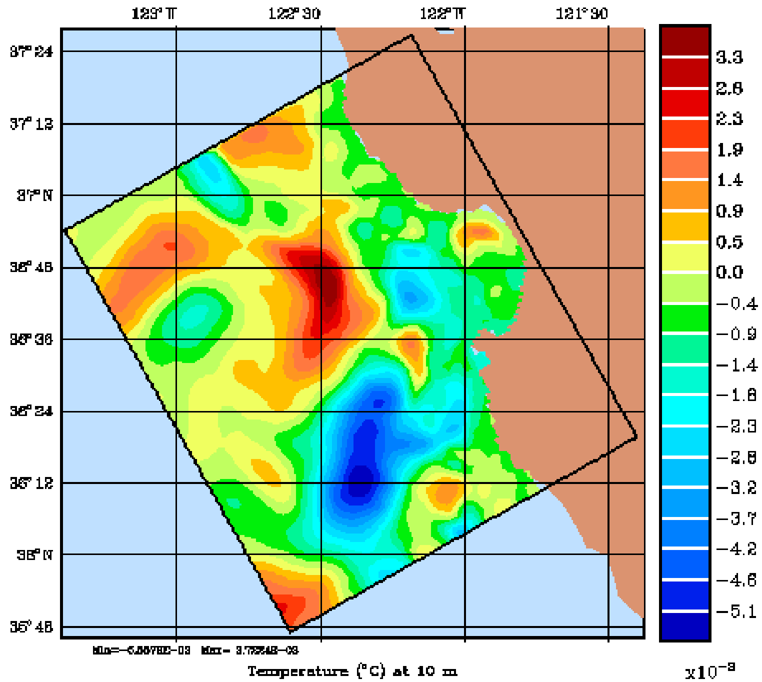


# ESSE field and error modes forecast for August 28 (all at 10m)

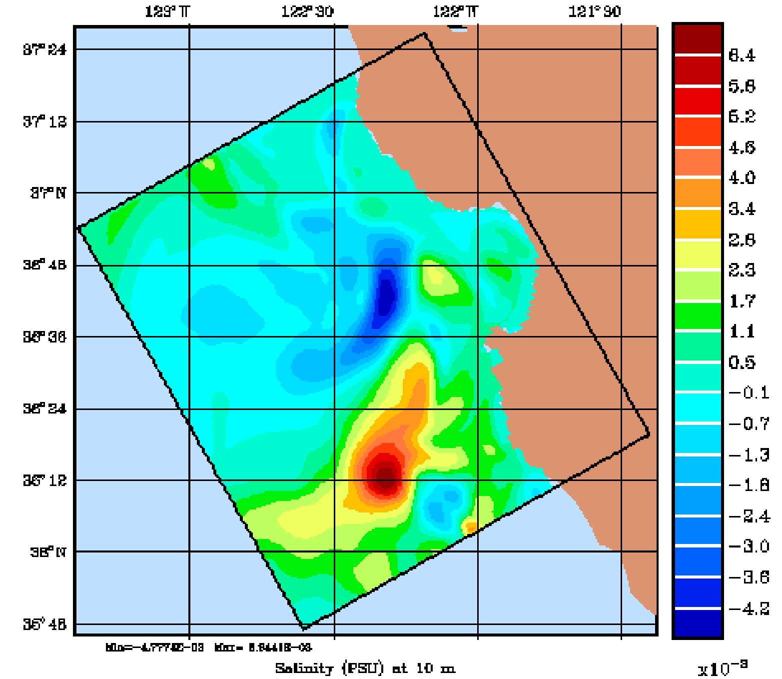


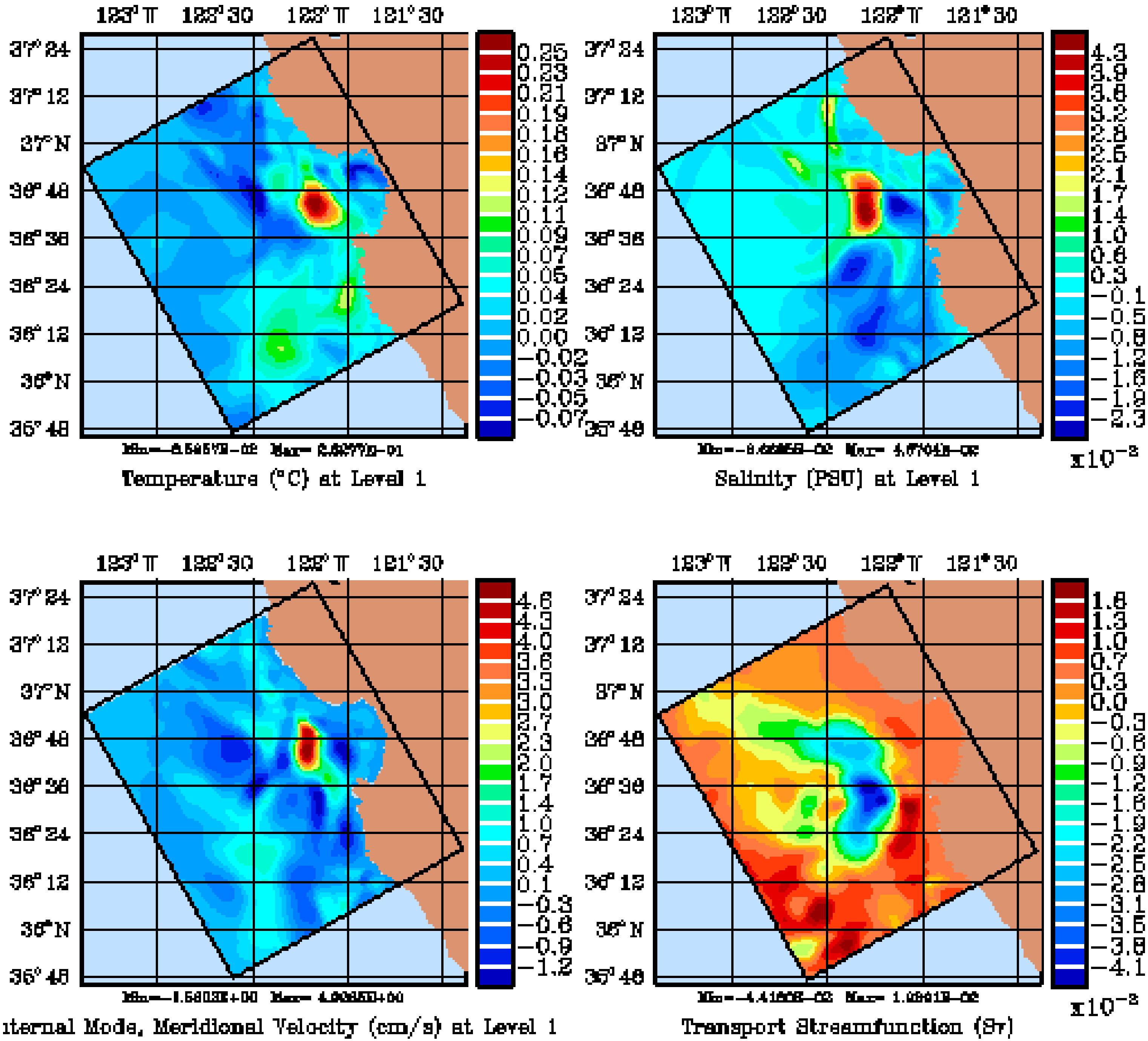

ESSE T error-Sv



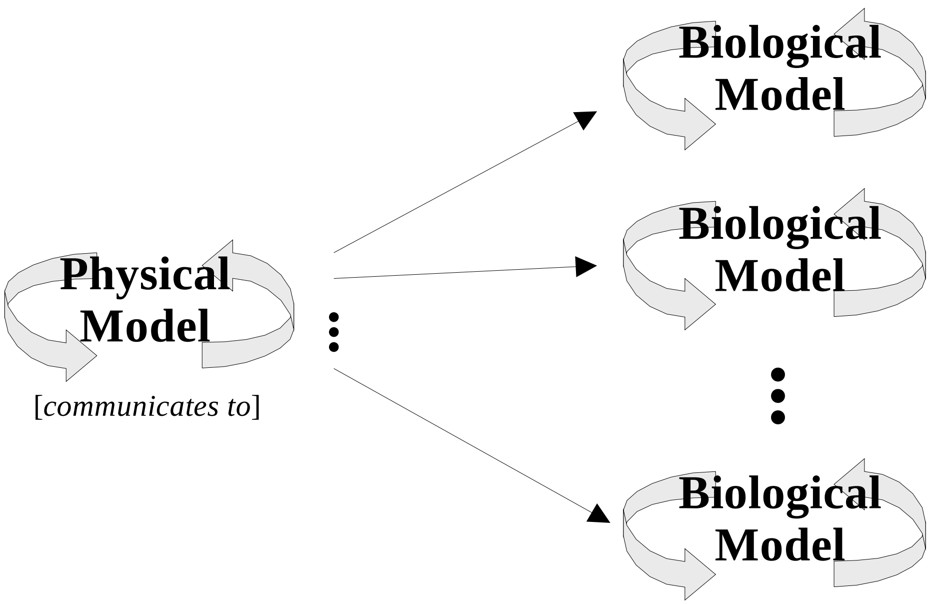

ESSE S error-Sv

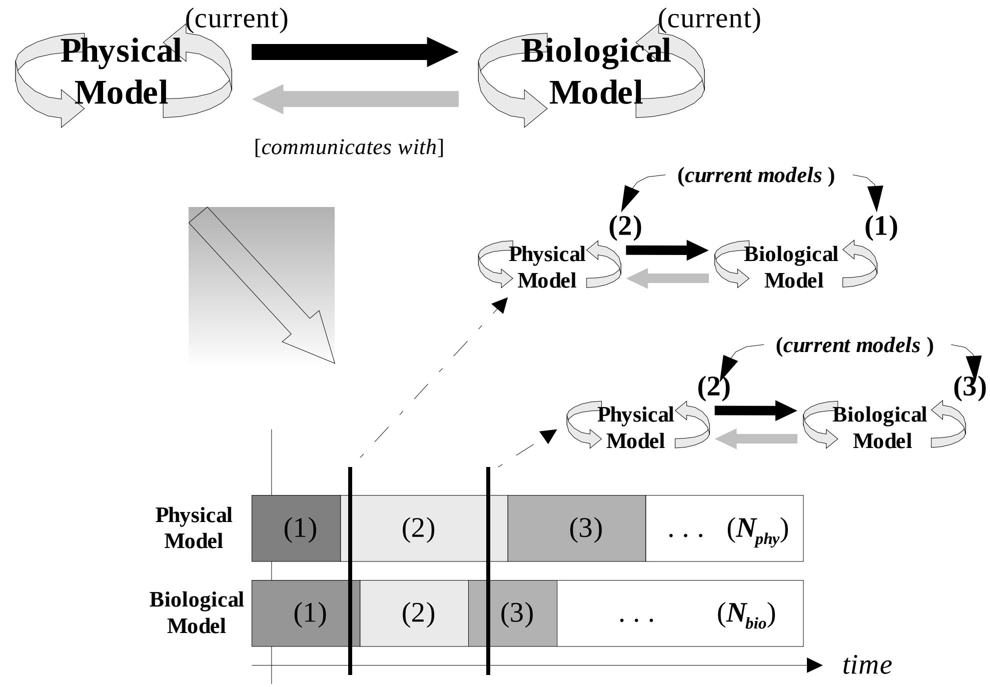


# Error Covariance Forecast for 28 August



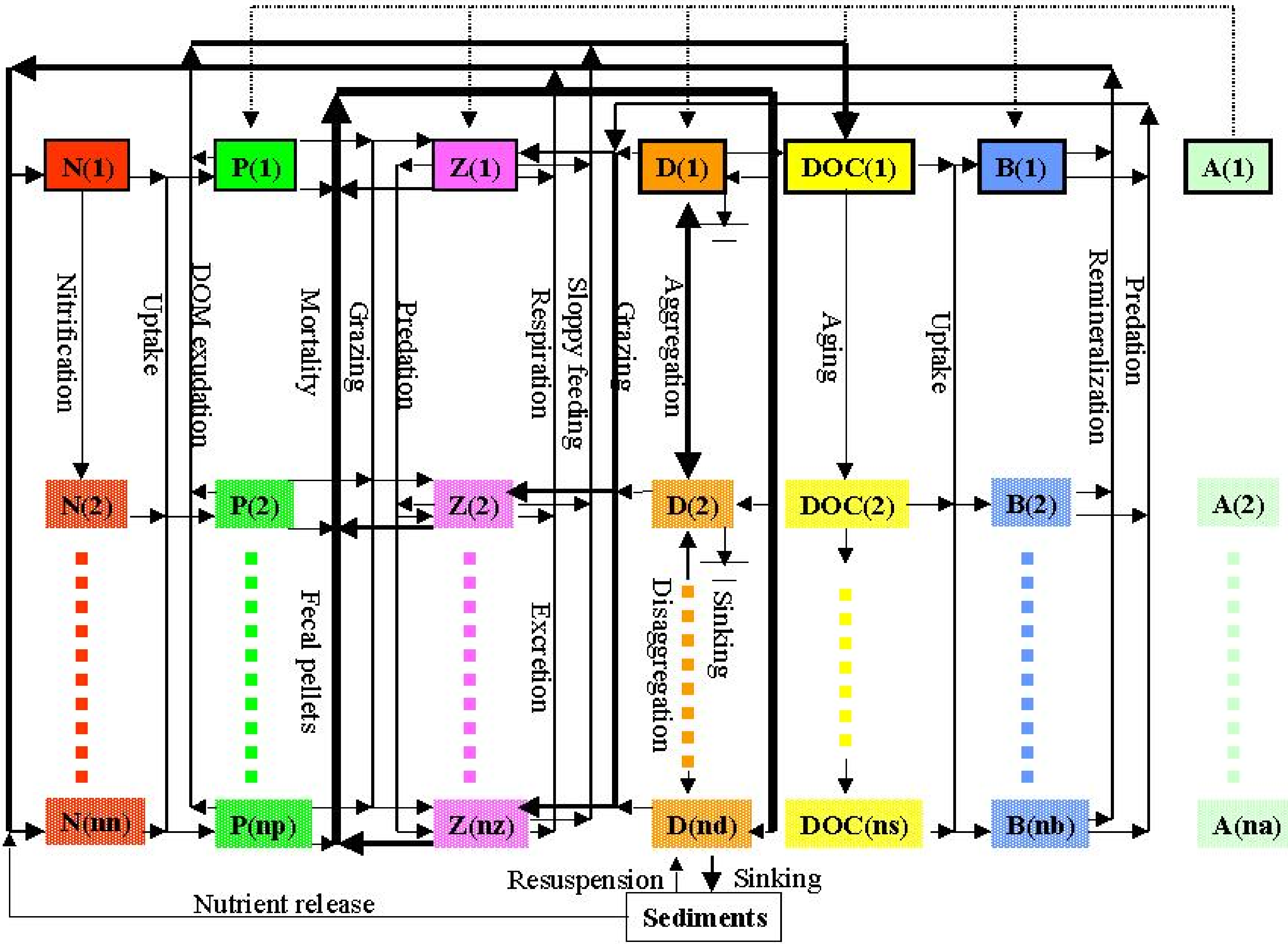
# Real-time Adaptive Coupled Models



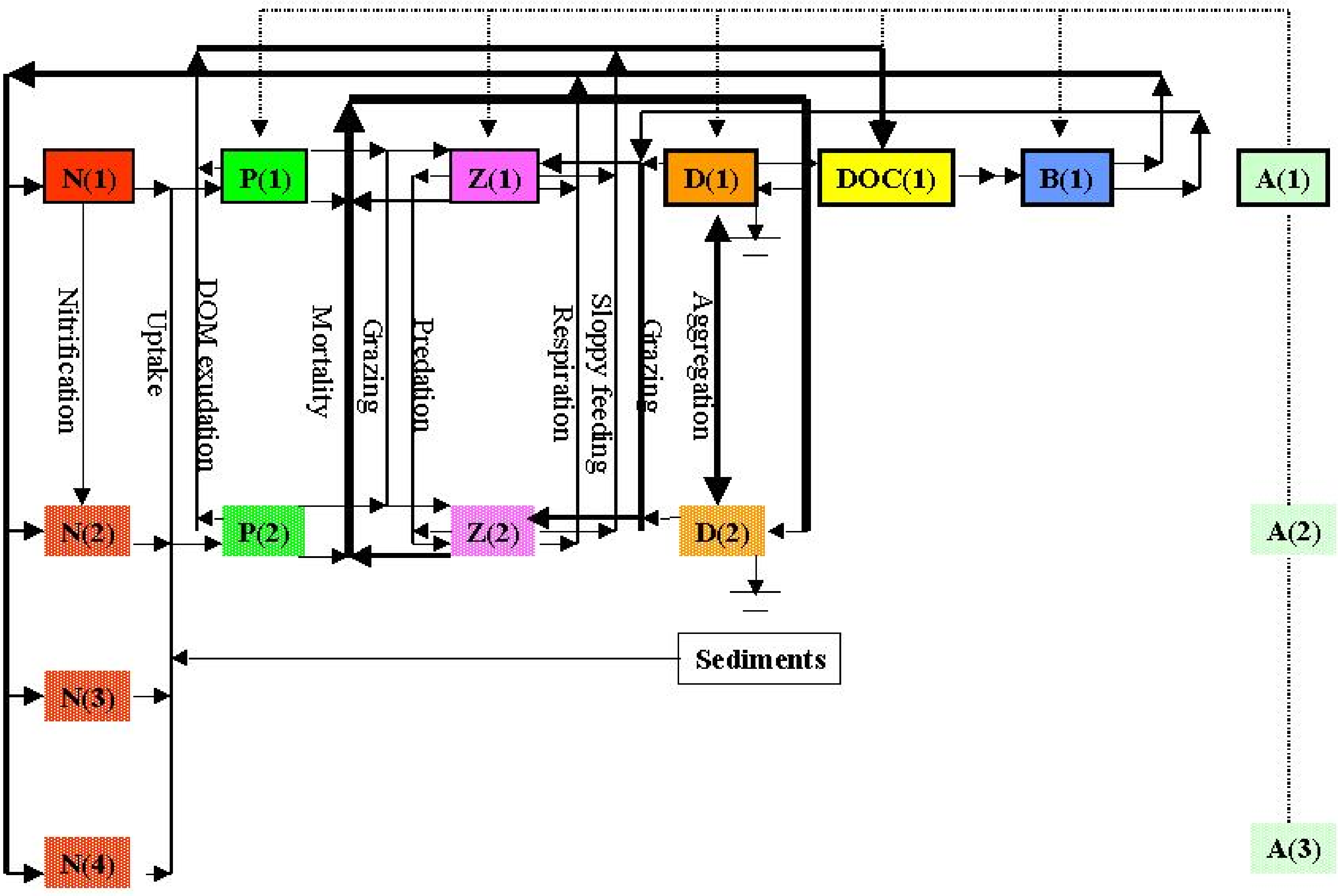




- Different Types of Adaptive Couplings:
  - Adaptive physical model drives multiple biological models (biology hypothesis testing)
  - Adaptive physical model and adaptive biological model proceed in parallel, with some independent adaptation
- Implementation
  - For performance and scientific reasons, both modes are being implemented using message passing for parallel execution
  - Mixed language programming (using C function pointers and wrappers for functional choices)

# Generalized Adaptable Biological Model



# *A Priori* Biological Model



# Example: Use P data to select parameterisations of Z grazing

Table 1. Parameterization of grazing on multiple types of prey with passive selection ($g_{max}$: maximum grazing rate; K: Half-saturation constant (but saturation constant in Eq. 1); $P_0$ threshold below which grazing is zero; $p_i$: preference coefficient; ? , *a*, ? : constant).

| Function | References |
|---|---|
| (1) Rectilinear $g_i = \begin{cases} g_{max}\frac{p_i P_i}{K}, & for \quad R \le K \\ g_{max}, & for \quad R > K \end{cases}, \; R = \sum_{i=1}^{n} p_i P_i$ | Armstrong, 1994 |
| (2) Ivlev function for each prey type: $g_i = g_{max}(1 - e^{-\alpha_i P_i})$ | Leonard et al., 1999 |
| (3) Ivlev function with interference between prey types: $g_i = g_{max}\left(1 - e^{-\alpha R}\right)\frac{p_i P_i}{R}$, with $R = \sum_{i=1}^{n} p_i P_i$ | Hofmann and Ambler, 1988 |
| (4) Mechanistic disc function: $g_i = g_{max}\frac{a_i N_i}{1 + \sum_{j=1}^{n} a_j \tau_j N_j}$ | Murdoch and Oaten, 1975; Holt, 1983; Gismervik and Anderson, 1997; Strom and Loukos, 1998 |
| (5) Michaelis Menten Function: $g_i = g_{max}\frac{p_i P_i}{K + \sum_{j=1}^{n} p_j P_j}$ | Murdoch, 1973; Real, 1977; Moloney and Field, 1991; Verity, 1991; Gismervik and Anderson, 1997; Strom and Loukos, 1998 |
| (6) Threshold MM function: $g_i = g_{max}\left(\frac{R - P_0}{K + R - P_0}\right)\frac{p_i P_i}{R}$, with $R = \sum_{i=1}^{n} p_i P_i$ | Evans, 1988; Lancelot et al., 2000 |
| (7) Modified MM function: $g_i = g_{max}\frac{p_i P_i}{1 + \sum_{i=1}^{n} p_j P_j}$ | Verity, 1991; Fasham et al. (1999) and Tian et al. (2001) |

Table 2. Parameterization of grazing on multiple types of prey with active switching selection ($g_{max}$: maximum grazing rate; K: Half-saturation constant; $P_0$ threshold below which grazing is zero; $p_i$: preference coefficient; $\alpha$, *a*, $\tau$: constant).

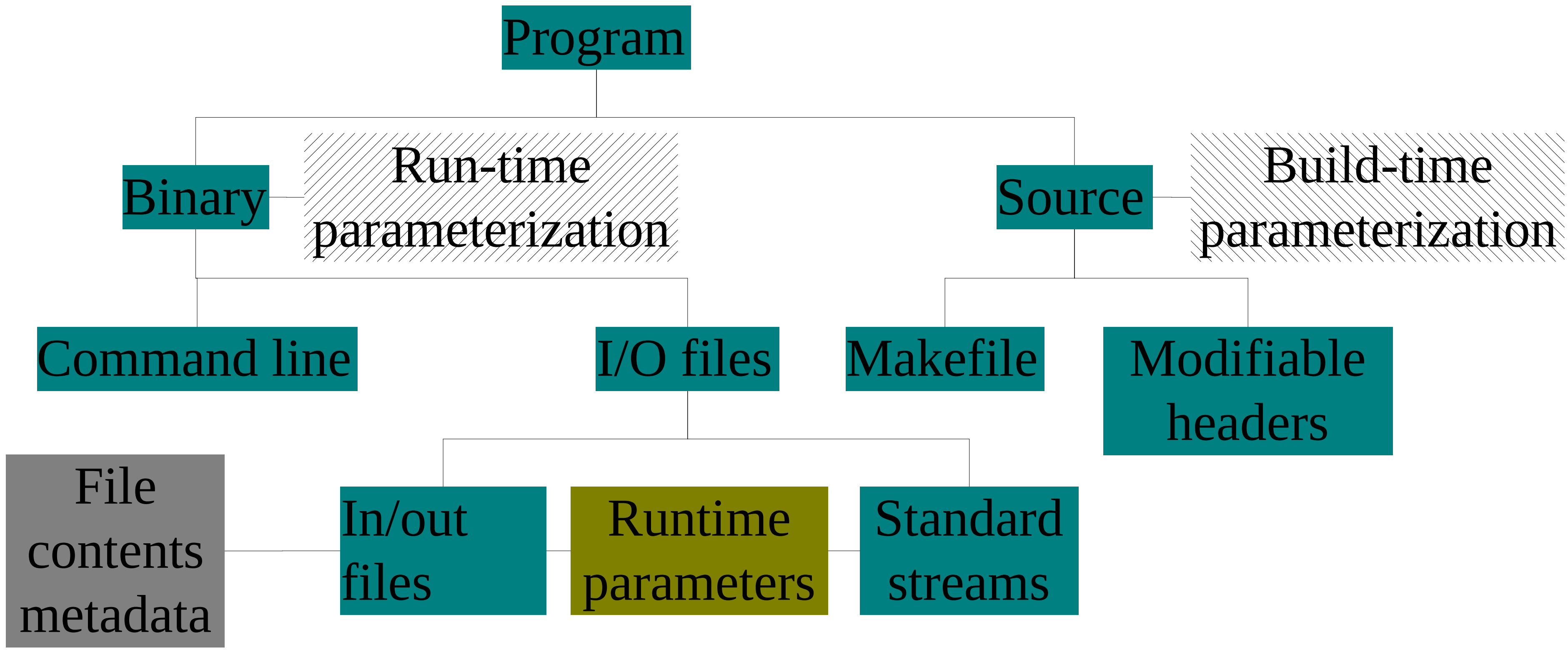
| Function | References |
|---|---|
| (1) Switching MM predation: $g_i = g_{max}\frac{p_i P_i^2}{K\sum_{j=1}^{n} p_j P_j + \sum_{j=1}^{n} p_j P_j^2}$ | Fasham et al., 1990; Strom and Loukos, 1998; Pitchford and Brindley, 1999; Spitz et al., 2001 |
| (2) Mechanistic disc switching predation: $g_i = g_{max}\frac{b_i N_i^2}{(1 + c_i N_i)(1 + \sum_{j=1}^{n}\frac{b_j h_j N_j^2}{1 + c_j N_j^2})}$ | Chesson, 1983 |
| (3) Generalized switching function: $g_i = g_{max} a_i \frac{(p_i P_i)^m}{\sum_{i=1}^{n}(p_i P_i)^m}$ | Tansky, 1978; Teramoto, 1979; Matsuda et al., 1986 |
| (4) Generalized switching function: $g_i = g_{max}\frac{(p_i P_i)^m}{\left(\sum_{i=1}^{n}(p_i P_i)\right)^m}$ | Vance, 1978 |
| (5) Generalized switching MM function: $g_i = g_{max}\frac{(p_i P_i)^m}{1 + \sum_{i=1}^{n}(p_i P_i)^m}$ | Gismervik and Andersen (1997) |
| (6) Generalized switching MM function: $g_i = g_{max}\frac{(p_i(P_i - P_{0i}))^m}{1 + \sum_{i=1}^{n}(p_i(P_i - P_{0i}))^m}$ | This work |

# Distributed/Grid Computing, Forecasting and Data assimilation with Legacy codes

- Distributed technologies (Sun Grid Engine) with web portal front-end ready to be tested with ESSE and HOPS
- Partial parallelism within ESSE easy because open-source routines (Sun Lapack) were used from the start
- HOPS, ESSE and acoustics codes: Fortran-matlab legacies
  - Relatively complex codes and makefile options
  - Hundreds of build and runtime parameters
- For other (future) codes, source code might not be available
- Classic encapsulation techniques that compartmentalize the code into subroutines, called from wrappers require constant reworking
- Thus: *we chose to encapsulate at the binary level, with generic approach, so as to handle new codes with limited/no rewriting*

# Metadata for handling legacy software

- Hierarchical structure for describing code (can also handle binary-only case)
- Basic assumptions about codes thus encapsulated:
  - No independent GUI, all runtime control from the command line and input/stdin files
  - All build-time parameterization done by altering the makefile and selecting values (parameters) in include-files
- Datatypes and relevant ranges for each parameter checked to ensure validity



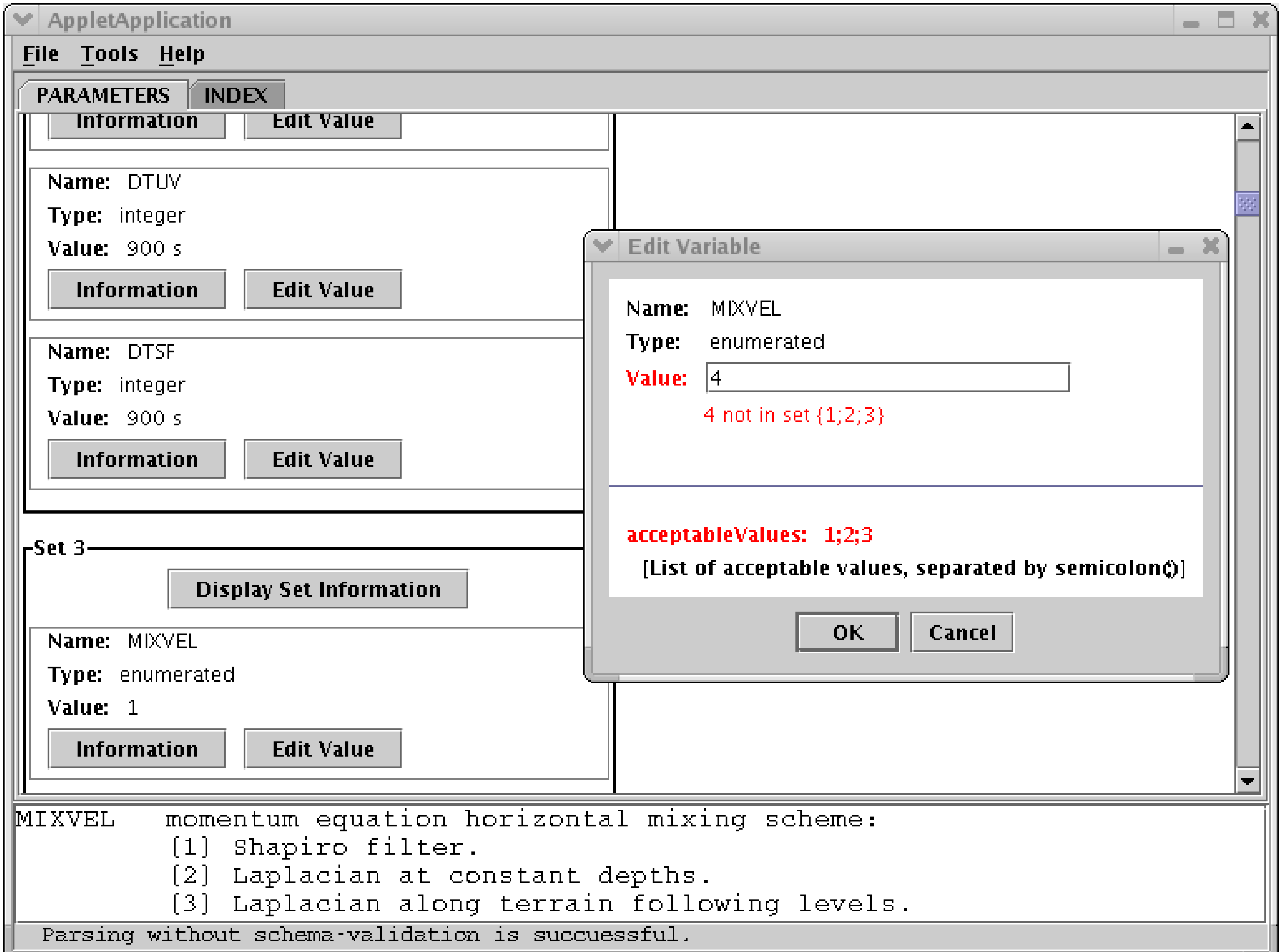
# *XML Encapsulation for Legacy Binaries*

- Descriptions of I/O files, runtime parameters, stdin and command line arguments, makefile parameters, requirements and conflicts for options, invocation mechanisms are needed:
  - Essentially a computer readable install and user guide
  - XML description provides software use and build metadata
  - Design of appropriate hierarchical XML Schemas (evolutionary)
  - Simulation datafile metadata are also usable (e.g. NcML for NetCDF)
  - Provides the constraints for generation of workflows (file I/O based)
- Binaries can be built on demand from generated makefiles
- Developers need to keep XML description up-to-date with their code (incremental effort) without switching to more elaborate approaches
- Concept is generally applicable, directly useful with other ocean models

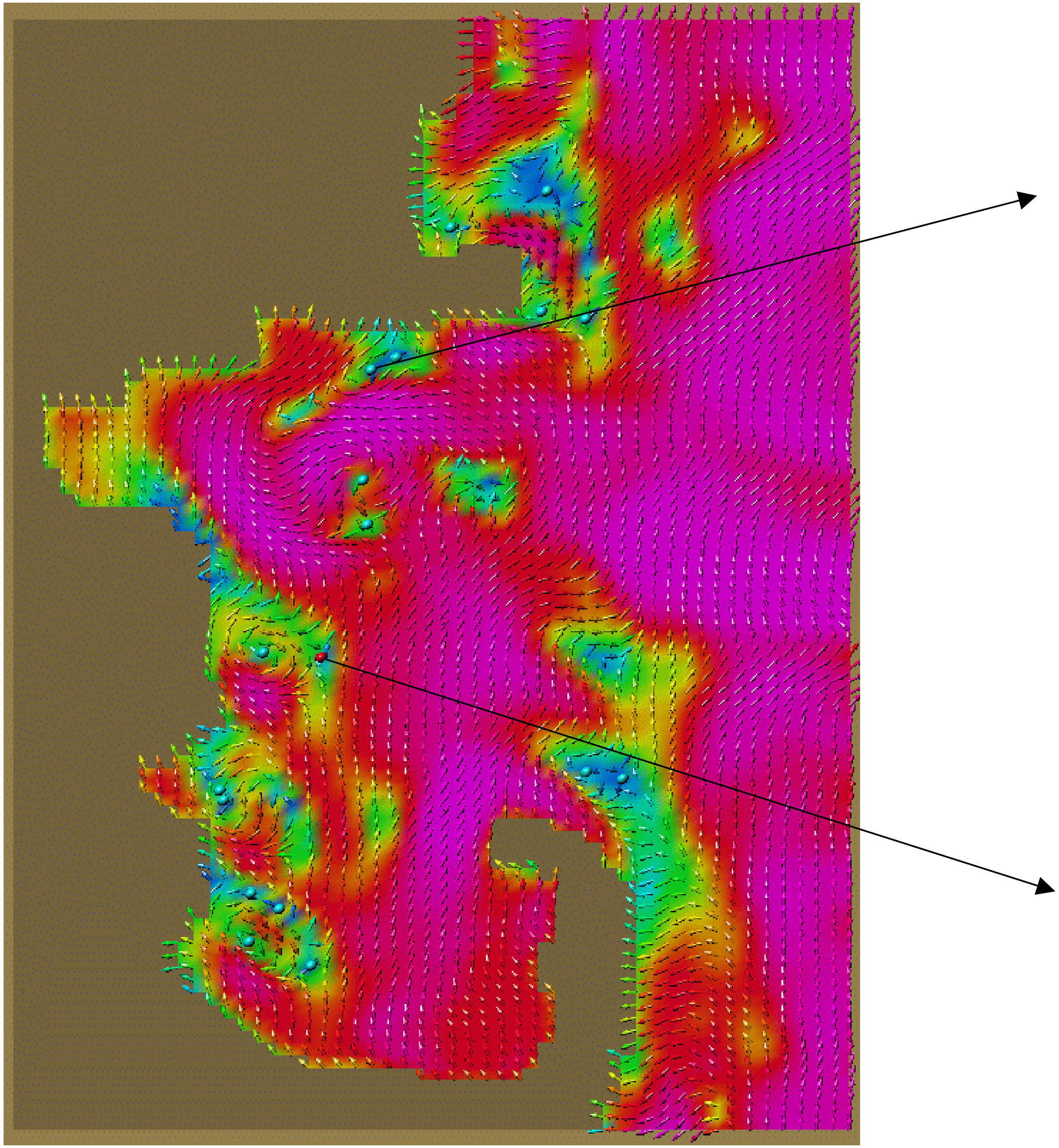
# Java-Based GUI for Legacy Binaries

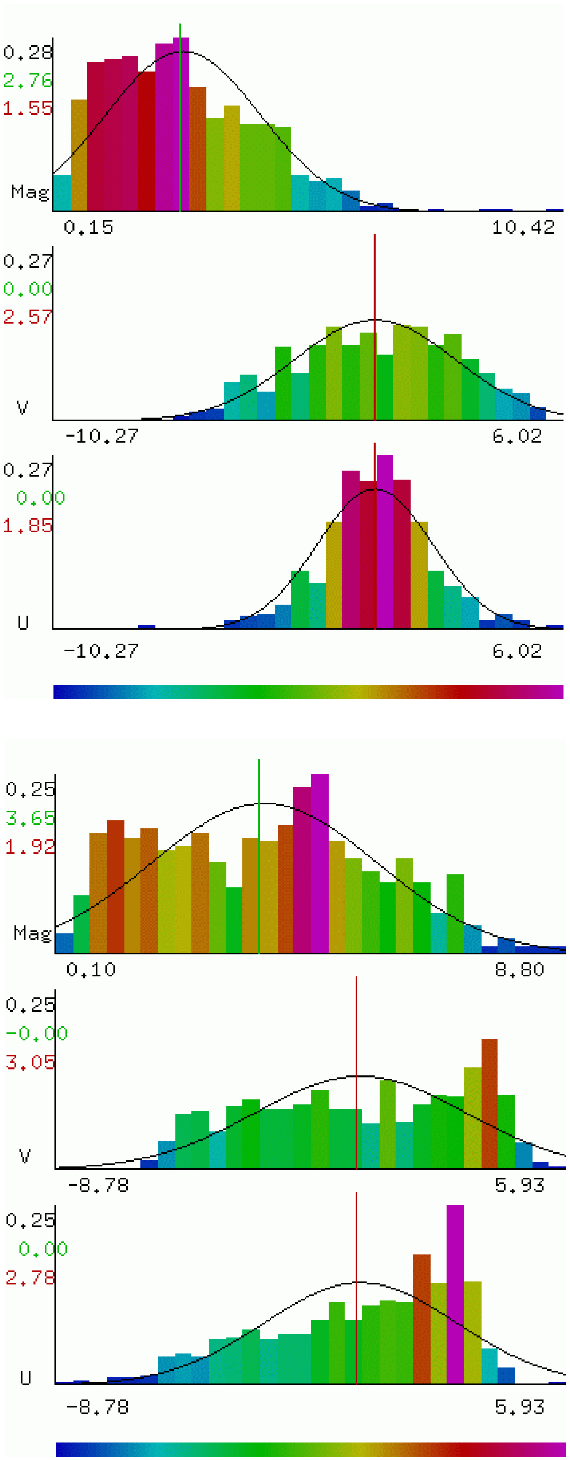
- Prototype GUI, accepts generic set of description files and generates user interface for building and running the binary. Implemented as an applet.
- Validates user choices, generates relevant scripts
- Integral part of the Grid-portal for LOOPS/Poseidon, it can be re-implemented in a more server-centric way (JSP etc.)
- Future directions for enhancement include:
  - Workflow composition: Employing the descriptions of the binaries and their input/output files as constraints. *We are currently using predefined workflows.*
  - Context mediation: When dataflow endpoints mismatch

# GUI: validity checking

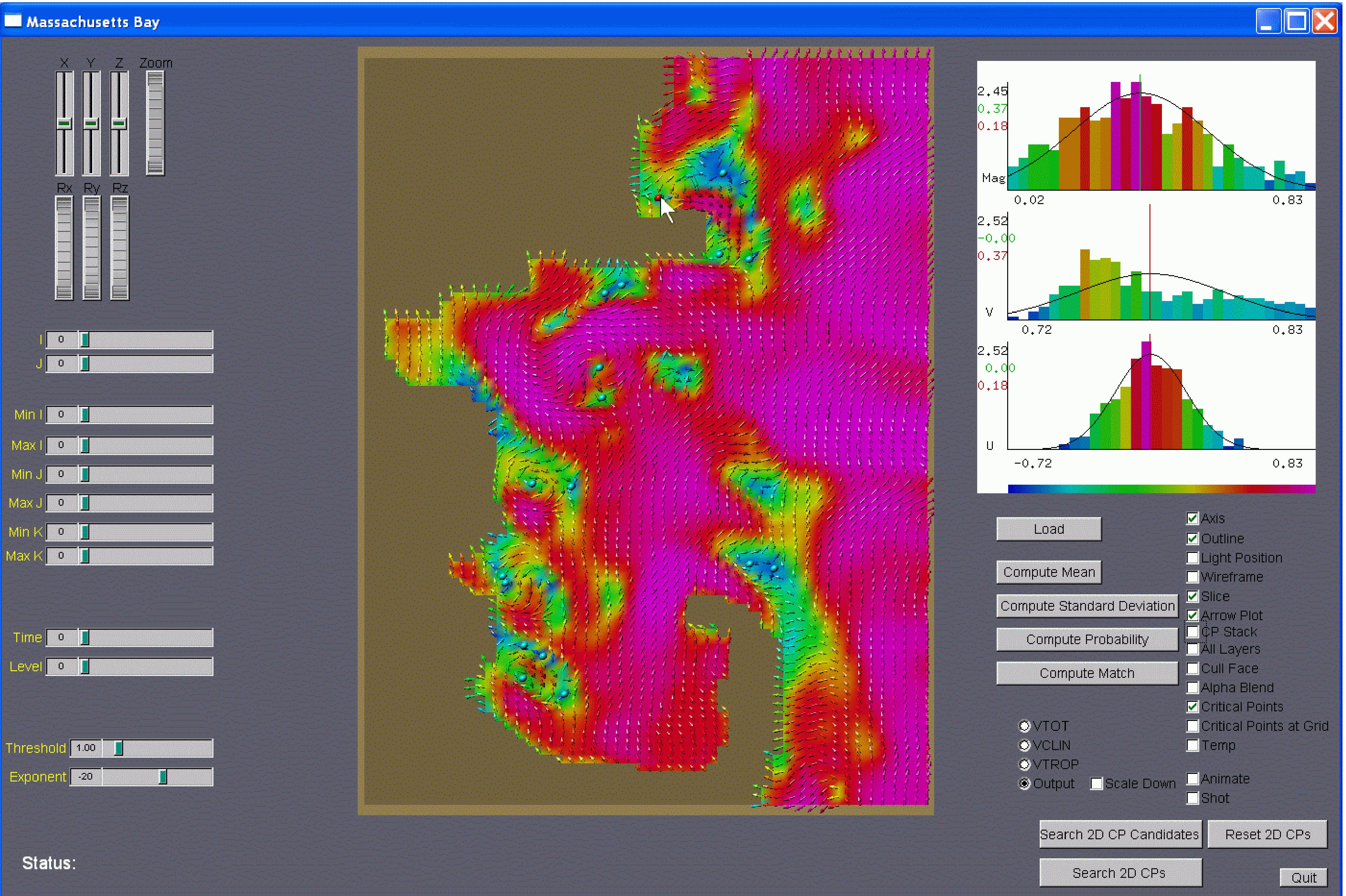


# Interactive Visualization and Targeting of pdf's





# Interactive Visualization and Targeting of pdfs (cont.)



# CONCLUSIONS: Present and Future

- Advanced systems for adaptive sampling and adaptive modeling in a distributed computing environment
- Web interface, Remote visualization, Metadata for code and data, XML-based encapsulation of software, Grid computing infrastructure (SunGridEngine)
- Interdisciplinary data assimilation should contribute significantly to understanding, especially to the quantitative development of fundamental/simplified coupled models
- More interdisciplinary research and education needed: mathematics, computer science, physical-biogeochemical-acoustical ocean science, atmospheric science, earth science and complex system science
- Short-term impacts likely overestimated, long-term effects likely under-estimated

# Feature Extraction for Adaptive Sampling

- Developing automated procedures to identify physical features of interest in the flow: upwelling, eddies & gyres, jets/fronts etc.
- Procedure can be based on a threshold for a derived quantity or a more complicated set of rules.
- Graphical output (in conjunction with uncertainty information) helps the user plan sampling patterns and vehicle paths.