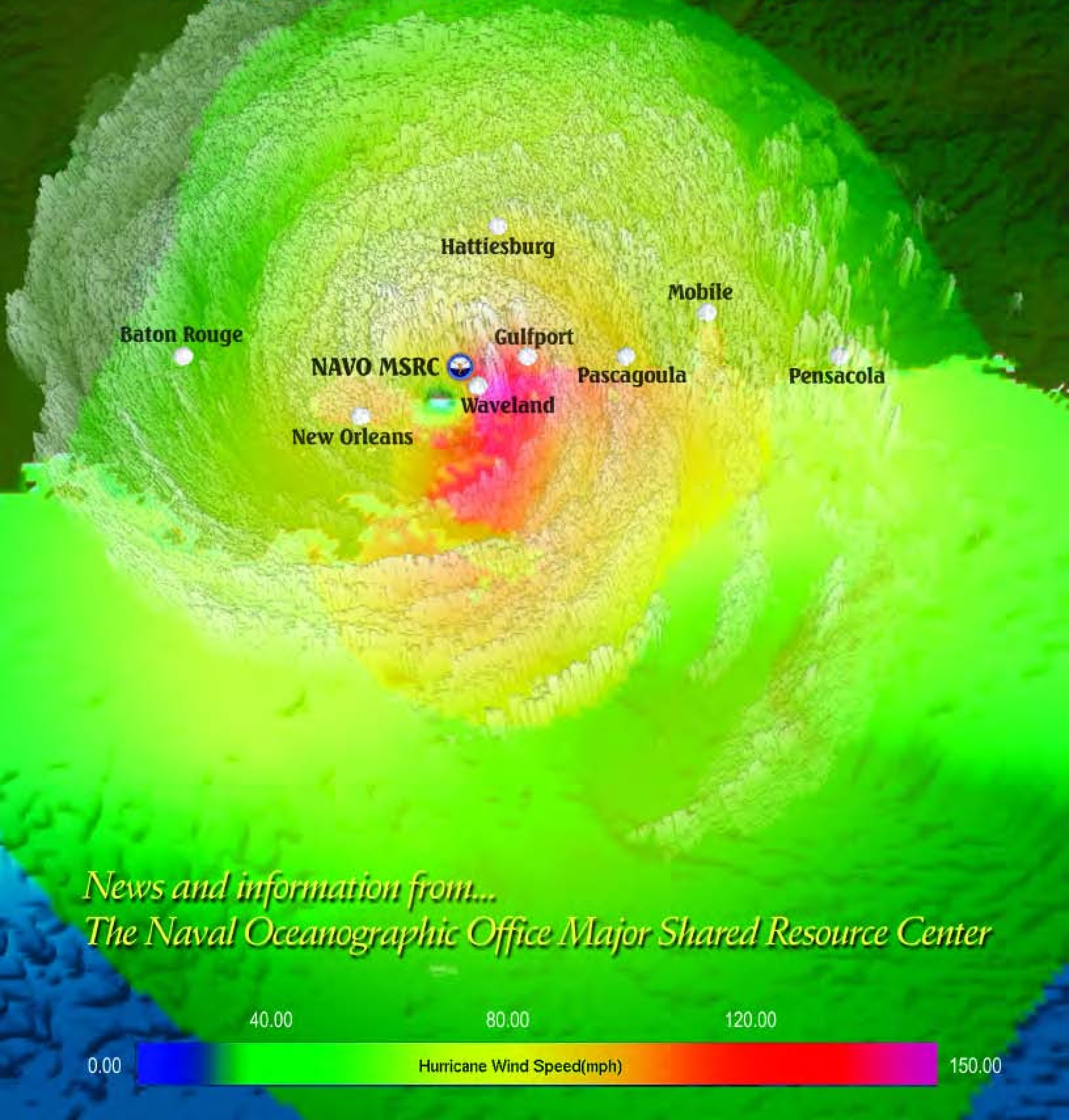
Navigator

NAVO MSRC FALL 2005



News and information from...
The Naval Oceanographic Office Major Shared Resource Center

The Director's Corner...

Steve Adamec, NAVO MSRC Director

NAVO MSRC and Katrina

Hurricane Katrina made landfall on the Mississippi Gulf Coast on August 29th—directly over Stennis Space Center (SSC) and the Naval Oceanographic Office. In a brief span of hours on that Monday morning, many lives were lost along the coast and the homes of hundreds of SSC employees were destroyed or severely damaged. The breadth and power of this storm were simply unimaginable, even for those of us who have weathered previous hurricanes and developed recovery and contingency plans for such storms.

In the days before and after the storm, it was my privilege to witness the absolute dedication of our employee community to their jobs throughout this crisis. A full day before the storm struck the coast, essential personnel from the extended NAVO MSRC staff had already moved their families to safety, reported for extended duty, and completed preparations to protect and continue MSRC operations during and after the storm.

I am very pleased to report that the MSRC systems, computing facilities, and data storage facilities suffered no significant storm-related damage, outages, or data loss—a direct result of superb preparation and teamwork by MSRC, Navy, and SSC staff. However, our Defense Research and Engineering Network (DREN) and commercial telephone connectivity were lost when the city of New Orleans lost power and flooded. Rodger Johnson, of the DREN support team, and the General Services Administration worked with us to quickly and innovatively establish satellite-based DREN and phone connectivity for the MSRC and SSC—weeks before reestablishment of reliable commercial communications in this area. When MSRC employees weren't actively engaged in MSRC work, they were helping with delivery of food, water, and support services to the many hundreds of storm refugees being housed in multiple buildings at SSC. Words cannot adequately express my gratitude for their performance—and that of the entire Navy and SSC community.

The enormous outpouring of concern, assistance, and prayers from all of you across the nation has made a clear difference for those of us here on the coast as we begin what will be a long period of recovery. Please be assured that as the gulf coast region heals and rebuilds, we will continue to provide and improve the premier HPC environment that you have come to expect from the NAVO MSRC.

The Naval Oceanographic Office (NAVO) Major Shared Resource Center (MSRC): Delivering Science to the Warfighter

The NAVO MSRC provides Department of Defense (DoD) scientists and engineers with high performance computing (HPC) resources, including leading edge computational systems, large-scale data storage and archiving, scientific visualization resources and training, and expertise in specific computational technology areas (CTAs). These CTAs include Computational Fluid Dynamics (CFD), Climate/Weather/Ocean Modeling and Simulation (CWO), Environmental Quality Modeling and Simulation (EQM), Computational Electromagnetics and Acoustics (CEA), and Signal/Image Processing (SIP).

NAVO MSRC
Code N7
1002 Balch Boulevard
Stennis Space Center, MS 39522
1-800-993-7677 or
msrchelp@navo.hpc.mil

NAVO MSRC Navigator
www.navo.hpc.mil/Navigator

NAVO MSRC Navigator is a biannual technical publication designed to inform users of the news, events, people, accomplishments, and activities of the Center. For a free subscription or to make address changes, contact NAVO MSRC at the above address.

EDITOR:
Gioia Furness Petro, petrogio@navo.hpc.mil

DESIGNERS:
Cynthia Millaudon, cynmill@navo.hpc.mil
Kerry Townson, ktownson@navo.hpc.mil
Lynn Yott, lynn@navo.hpc.mil

Any opinions, conclusions, or recommendations in this publication are those of the author(s) and do not necessarily reflect those of the Navy or NAVO MSRC. All brand names and product names are trademarks or registered trademarks of their respective holders. These names are for information purposes only and do not imply endorsement by the Navy or NAVO MSRC.

Approved for Public Release
Distribution Unlimited

Contents

The Director's Corner

Feature Articles

Scientific Visualization

The Porthole

Productivity Enhancement and Technology Transfer

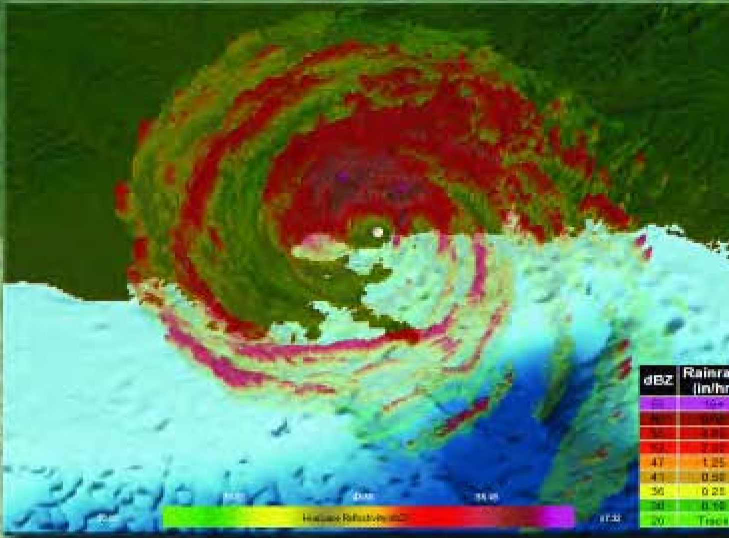
Navigator Tools and Tips

Upcoming Events

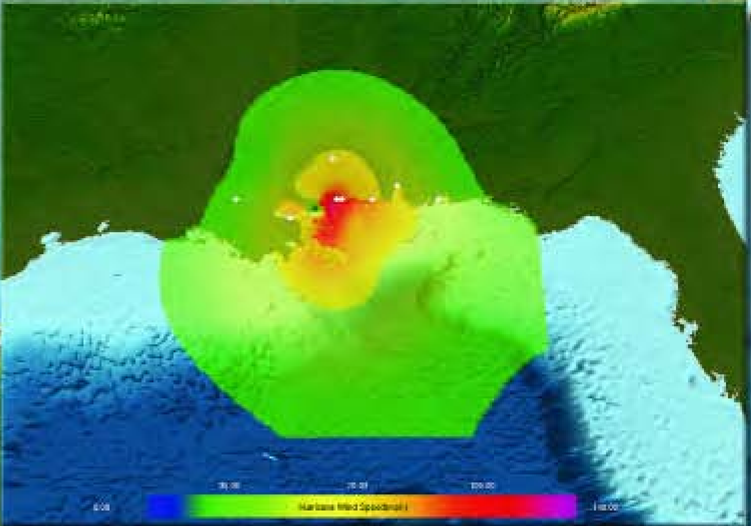
Mapping Katrina Data

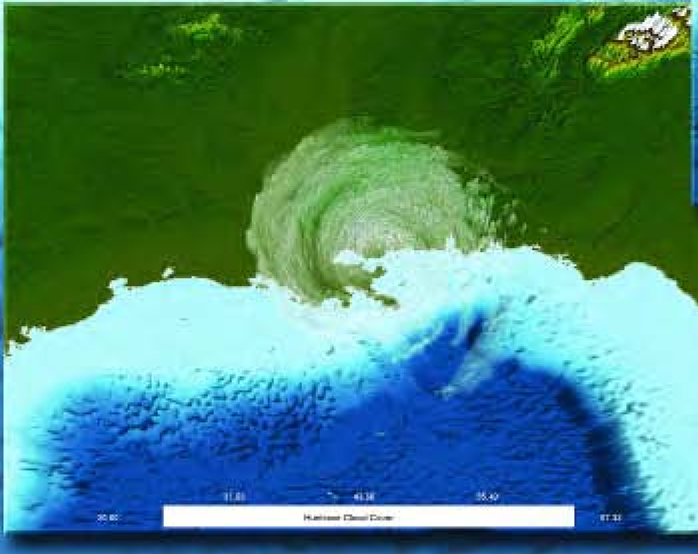
The bathemetry and topography data in these images and the cover were drawn by NAVO MSRC Visual Analysis and Data Interpretation Center (VADIC) Visualization Software Engineer Chad Saxon using OpenGL triangle strips with a water-to-land texture map and colored based on bathemetry and topography height. The data used to generate these images are courtesy of National Oceanic and Atmospheric Administration (NOAA).

On the cover: Hurricane Katrina's wind speed (in miles per hour) and reflectivity (in dBZ (radar reflectivity)), at 13:30 hours on 29 August 2005 as well as the bathemetry and topography data of the area affected by the storm.



Rain-dBZ levels are represented through energy reflected back to radar, corresponding directly to levels of precipitation intensity in terms of rainfall per hour. The higher the dBZ level, the more intense the precipitation. dBZ stands for decibels of Z, with Z being the actual reflectivity factor. The higher the reflectivity factor, the heavier the rainfall. This scale is logarithmic, so there is no factor one can use to multiply the dBZ level and arrive at the corresponding rainfall rate. *Data courtesy of NEXRAD (Next Generation Weather Radar).*





Wind and Cloud Cover (cloud height)—In order to see all the data at once, OpenGL blending, along with various 1-dimensional textures (color maps), had to be applied to both the wind speed and the hurricane reflectivity. For the wind data, the slower wind speeds were drawn as completely transparent so the location of the eye is more clearly visible, and data in the grid with irrelevant or no values at all were filtered out as well. As for the hurricane cloud cover, each grid cell was rendered between being opaque and transparent based on reflectivity and colored with a white to light gray color map. *The wind data are courtesy of NOAA, and the reflectivity data are courtesy of NEXRAD.*

A Regional Model of the 3-D Circulation of the Indonesian Seas

Kieran O'Driscoll, Naval Oceanographic Office, Department of Marine Science, The University of Southern Mississippi
Vladimir M. Kamenkovich, Department of Marine Science, The University of Southern Mississippi
Dmitri A. Nechaev, Department of Marine Science, The University of Southern Mississippi

BACKGROUND

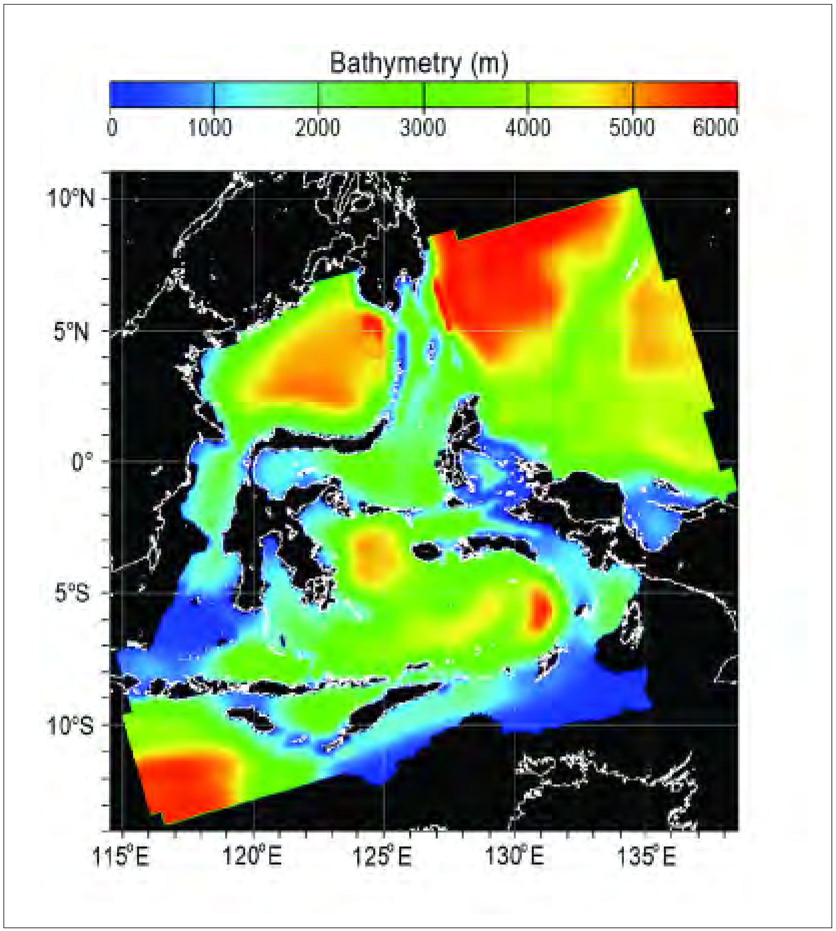
This regional numerical model with high spatial resolution is based on the Princeton Ocean Model. The horizontal resolution of approximately 10 Kilometers (km) allows for proper resolution of flows within straits such as Lombok, Ombai, the narrowest part of the Makassar Strait, and others. The vertical resolution has been chosen to properly resolve the surface and bottom Ekman boundary layers and the salinity maximum that is usually located at 150-200 Meters (m).

The bottom topography, based on Earth Topography - 5 Minute (ETOPO5) data, has been smoothed so as to eliminate overly steep slopes, on the one hand, and to retain all important sills and passages, on the other. Note that depths greater than 100m only are considered to be part of the ocean, and there are 29 sigma levels.

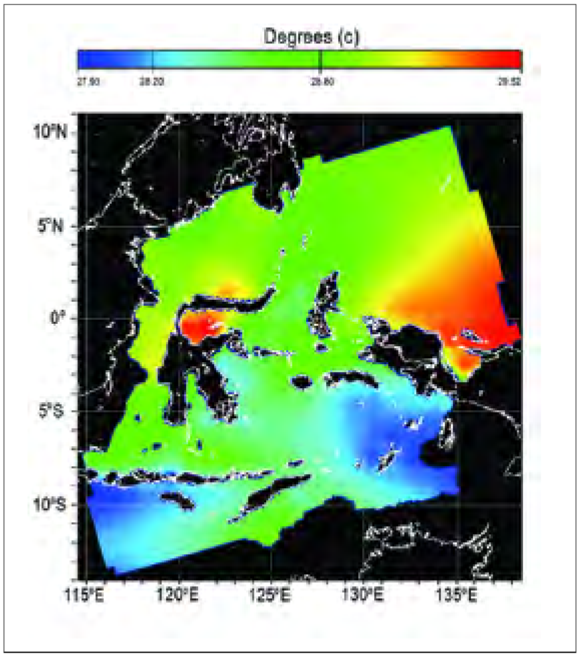
The motion in the whole Indonesian Seas area was assumed to be forced by the inflow and outflow of water due to well-pronounced currents such as the Mindanao Current, New Guinea Coastal Surface Current and New Guinea Coastal Undercurrent, North Equatorial Countercurrent, and the major outflow through an appropriately chosen section in the Indian Ocean. So the model has four ports simulating these inflows and outflows. The total transports through these ports have been taken from observations. Hence the total transport of the Indonesian Throughflow was specified, but the transports through various passages were determined by the internal dynamics. Simple distributions of the transport velocities across the ports have been assumed to provide the open boundary conditions for the barotropic velocities.

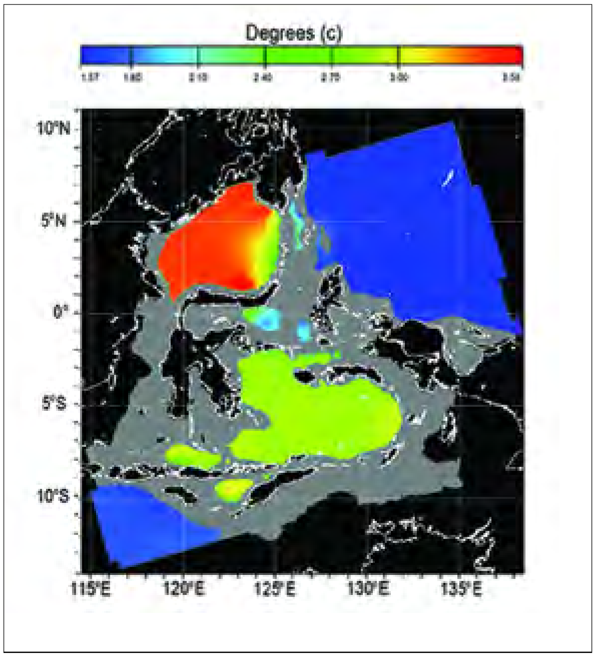
Typical vertical distributions of velocities within the ports, known from observations, have been incorporated into the model. At the entrance to the ports linearized momentum equations with nudging to observed velocities and modified friction were used. Such equations provide values for the baroclinic velocities at the open boundaries. The

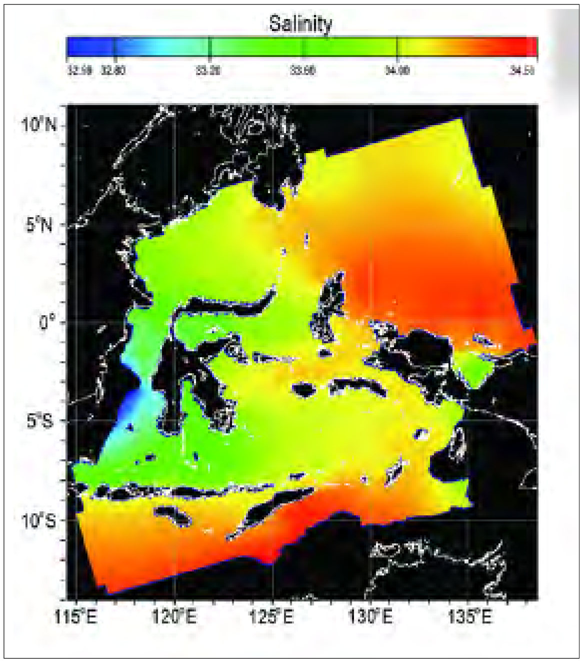
Continued Next Page...

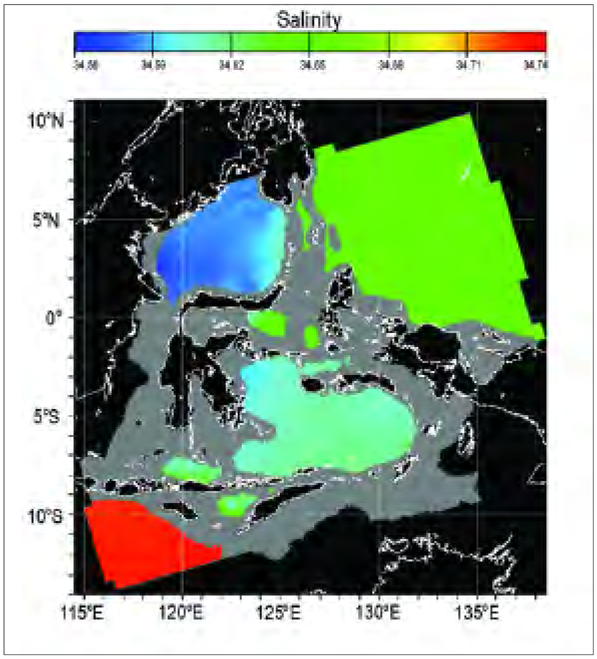


The domain of the model and bottom topography (smoothed). The four ports and the corresponding port channels are shown.









Top Left. Potential temperature distribution at the sea surface (specified). Note that the waters entering through the Mindanao and New Guinea ports (North Pacific and South Pacific waters) have substantially different temperatures.

Top Right. Potential temperature distribution at 2500m. Note that the sills break down the whole region into separate sub-basins with substantially different temperatures.

Bottom Left. Salinity distribution at the sea surface (specified). Note that the waters entering through the Mindanao and New Guinea ports (North Pacific and South Pacific waters) have substantially different salinities.

Bottom Right. Salinity distribution at 2500m. Note that the sills break down the whole region into separate sub-basins with substantially different salinities.

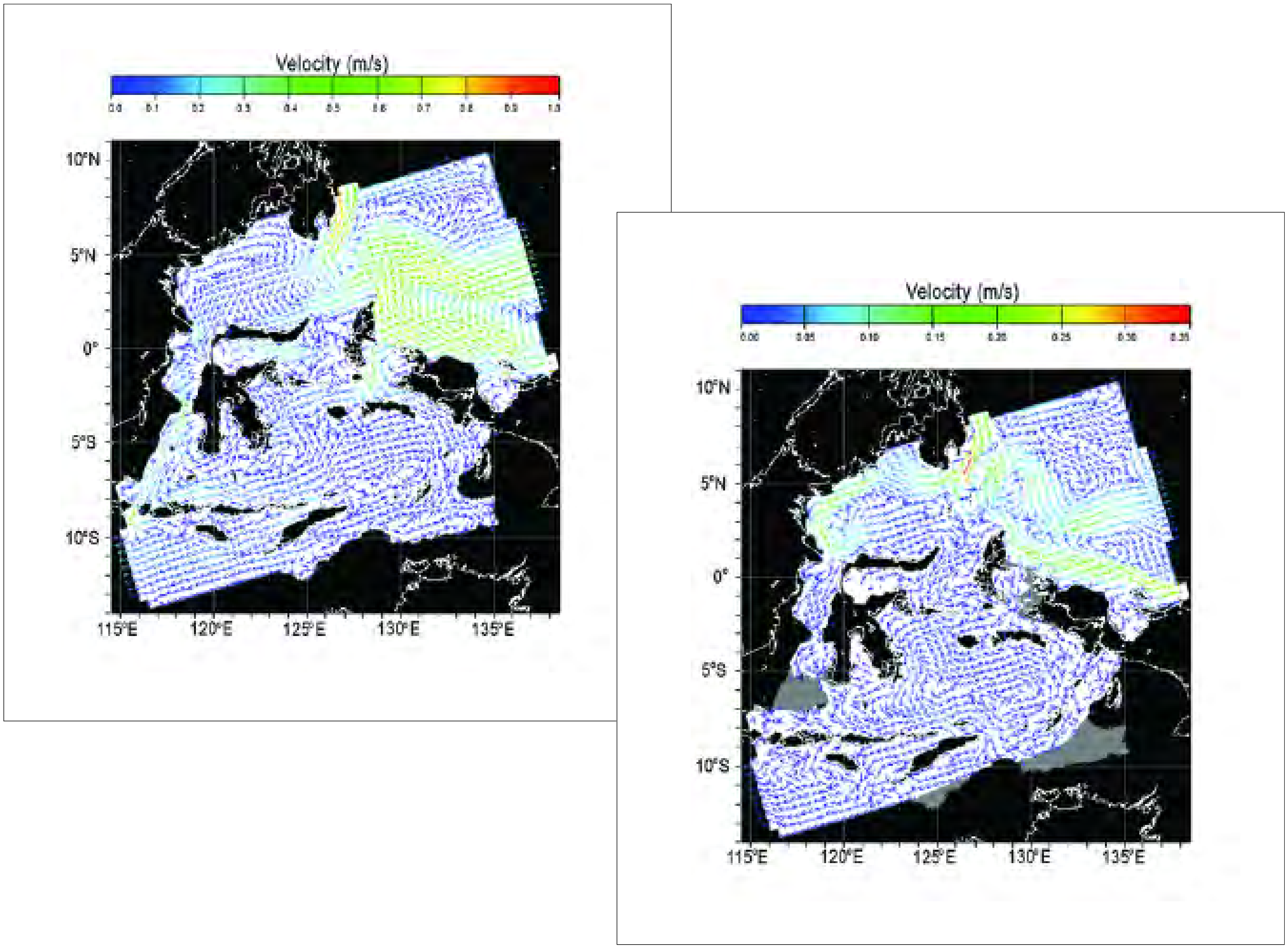
standard boundary conditions for potential temperature and salinity have been applied. The corresponding data have been taken from the Levitus database with some corrections in the Halmahera region provided by the Arlindo database.

Technically, it appeared convenient to introduce the so-called port channels for tapering off the nudging and additional friction. Such a technique made it possible to do all needed adaptation outside the main region of interest, thus not modifying any of the basic equations within the area. Some weak filtering has been applied to eliminate grid-scale oscillations caused by very complicated bottom topography. The results of the simulations with zero local wind stress are presented.

Conclusions

- ✍ A regional model of the Indonesian Seas circulation has been developed, which is capable of using very complicated bottom topography with numerous narrow passages and sills. In its present form, the model requires the specification of the total transports through the main ports, temperature and salinity distributions at the surface and lateral boundaries, and typical vertical profiles of the velocities at the entrance of the ports.

Continued Next Page...



Left. Horizontal velocity pattern at the sea surface. The arrows show the direction, while the colors show the magnitude of the velocity.

Right. The same as the figure on the left, but at a depth of 500m. The grey areas show the ocean sub-regions that are shallower than 500m.

- The partition of the specified water inflow between various passages is controlled mainly by the peculiarities of the bottom topography in the region. By and large, the simulated pattern of the circulation agrees satisfactorily with the pattern based on the available observations and some qualitative considerations.
- A two-layer type flow structure over distinct sills is obtained; the currents at the surface and at the bottom are directed oppositely. In contrast, the flow through passages is typically in one direction.
- It is shown that sills break the area under consideration into separate sub-basins with substantially different values of temperature and salinity.

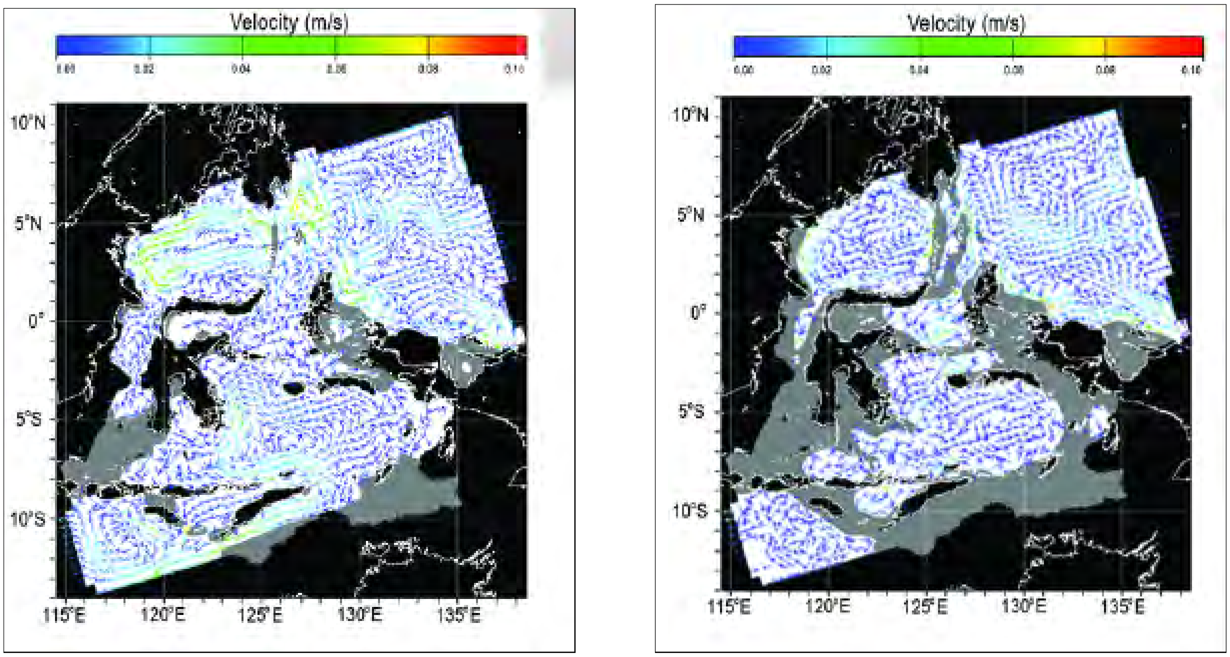
Future plans

Future plans include the incorporation of specified local winds, heat and fresh water fluxes at the surface, tidal friction, and seasonal variations into the model.

Scientific objectives of primary interest are:

- The propagation and interaction of the South and North Pacific waters.
- The role of the bottom form stress in the overall momentum balance in the area.

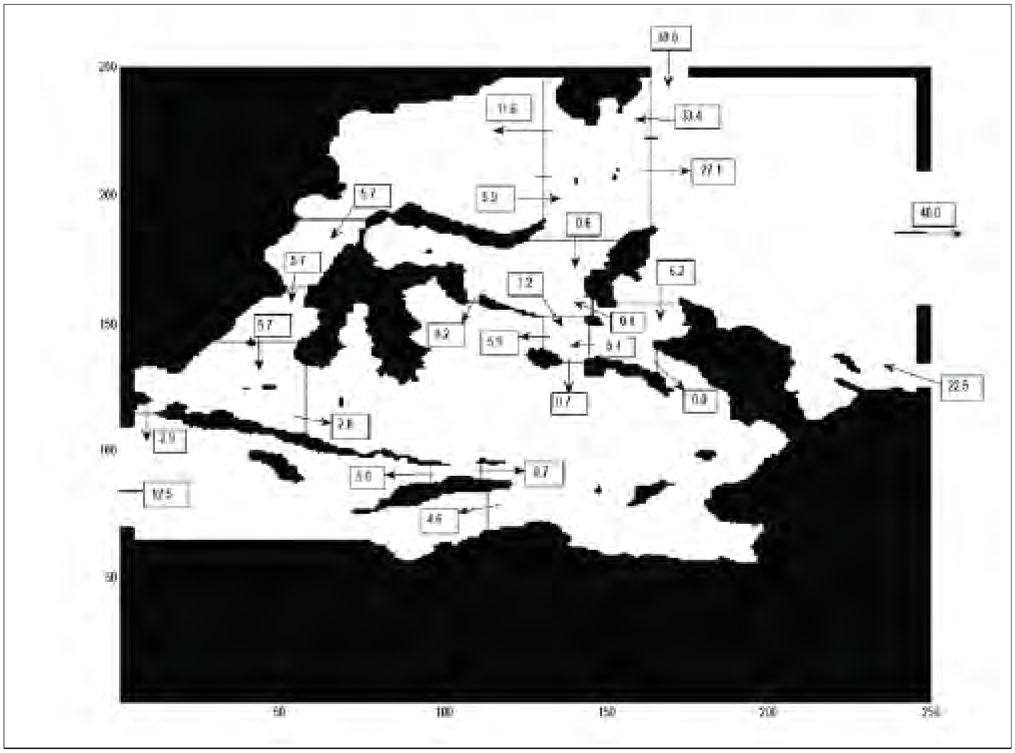
Images Continue Next Pages...



Left. The same as in the left-hand figure on Page 7, but at a depth of 1000m. The grey areas show the ocean sub-regions that are shallower than 1000m.
Right. The same as in the left-hand figure on Page 7, but at a depth of 2000m. The grey areas show the ocean sub-regions that are shallower than 2000m.

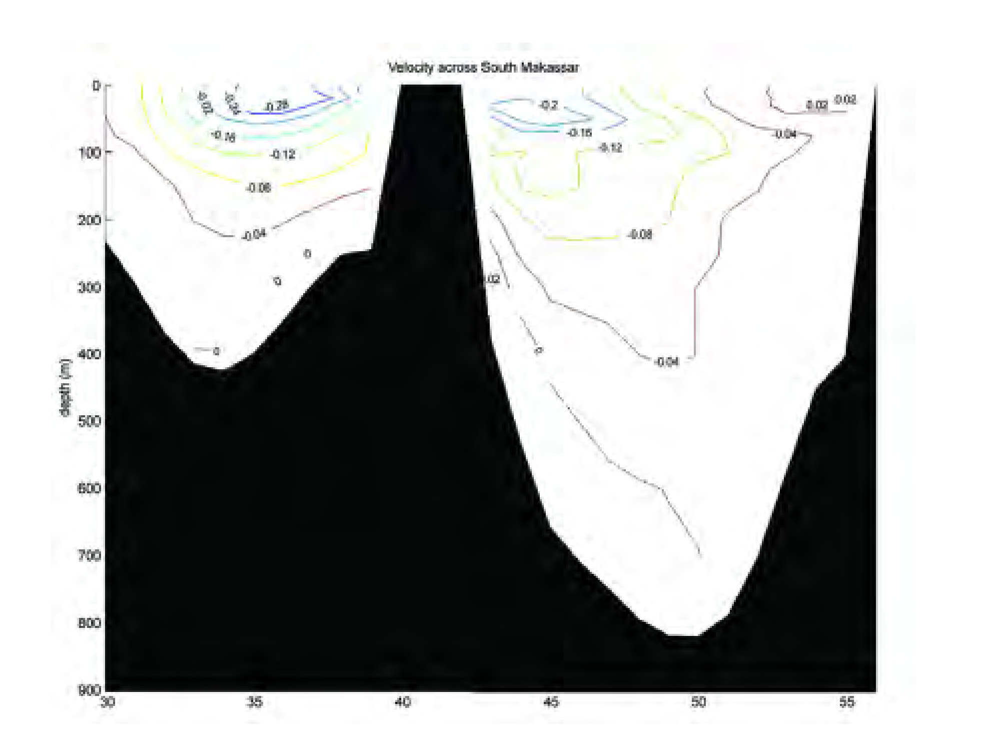
Acknowledgements

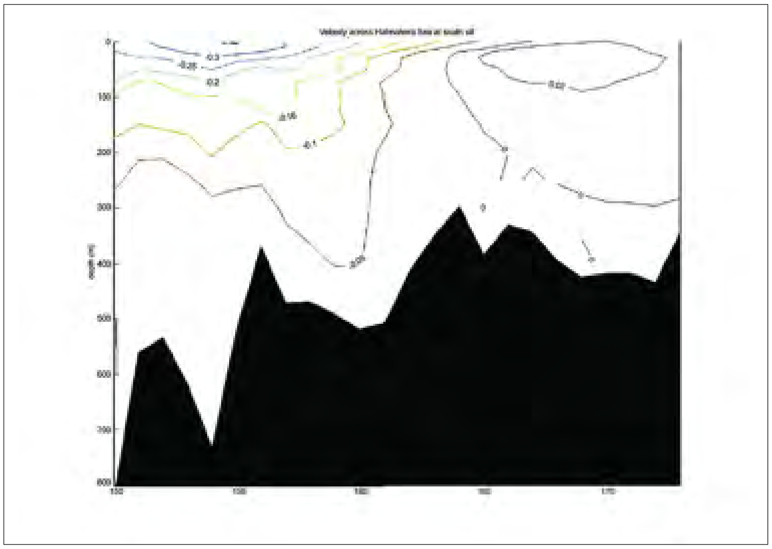
The authors would like to thank the NAVO MSRC for their assistance in preparing the graphics. Kieran O'Driscoll is grateful to the Naval Oceanographic Office (NAVOCEANO) for supporting his education. V. Kamenkovich and D. Nechaev were supported by National Science Foundation (NSF) grant OCE 01-18200.



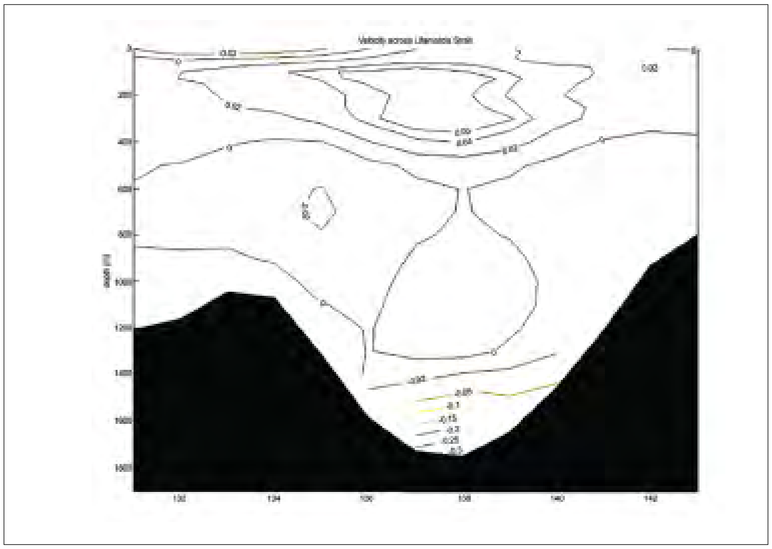
Top. Total transports (in Sv) through various sections, including ports, are shown. The position of grid cells is given by the corresponding i and j numbers ranging from 1 to 250.

Bottom. Isolines of the values of velocity normal to the section through the south of the Makassar Strait (section at j=194, see above). The values are in m/sec.

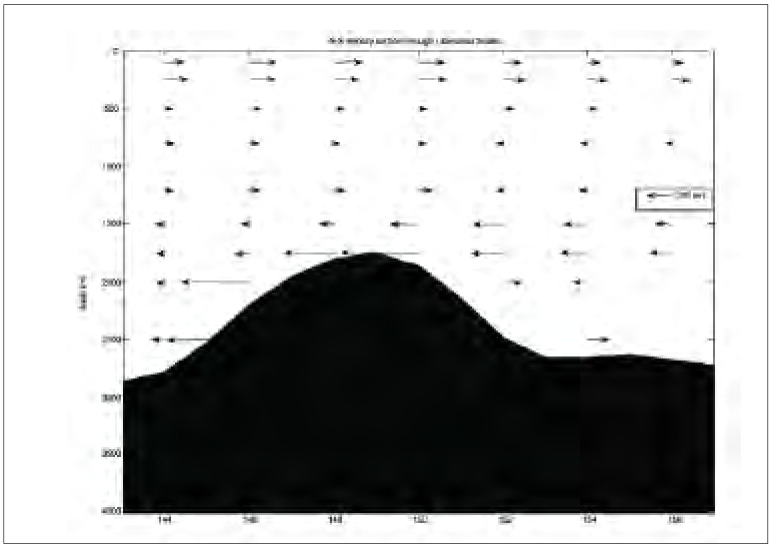




Right. Isolines of the values of velocity normal to the section through the southern Halmahera Sea sill (section at j=149, top image, Page 9). The values are in m/sec.



Left. Isolines of the values of velocity normal to the section through the southern Lifamatola Strait (section at j=150, see top figure, Page 9). The values are in m/sec.



Right. The j-component of the horizontal velocity at the section through the Lifamatola Strait (i=138). The scale of the arrows is shown. Note that the current near the bottom is opposite to the current near the sea surface.

The Baseline Configuration Initiative: An Overview

John Skinner, NAVO MSRC Support and Outreach

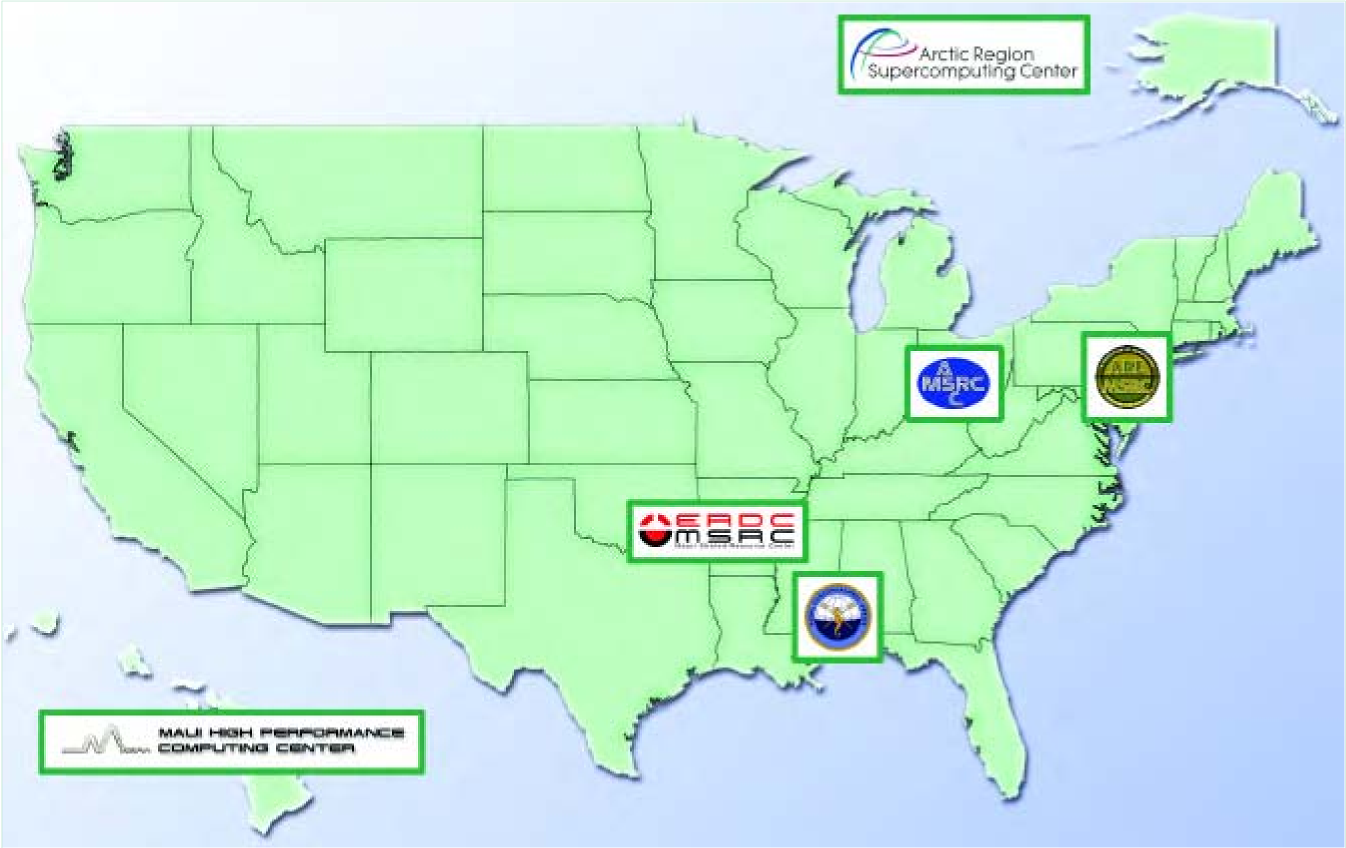
As part of an ongoing effort to increase commonality across Department of Defense (DoD) High Performance Computing Modernization Office (HPCMP) resources and simplify the global working environment for DoD users, the HPCMP has initiated the Baseline Configuration Initiative (BCI). The BCI will develop a "baseline configuration" of tools, software, configurations, and policies that should be common across the four Major Shared Resource Center (MSRC) centers, the Maui High Performance Computing Center (MHPCC), and the Arctic Region Supercomputing Center (ARSC).

These baselines will include relatively simple items (such as common location and definition of "scratch" workspace, a set of global environment variables, open source libraries/software packages), as well as more challenging items (such as commonality for site queue configurations, mass storage access and functionality, and default user environments). The ultimate goal is to create a cross-Center environment that makes it as easy as possible for users to work at any center or to move work among centers during a fiscal year without losing, and hopefully enhancing, productivity.

A planning team named the Baseline Configuration Team (BCT) has been formed to work on the BCI. This team is comprised of a Team Lead and Deputy Team Lead, and includes representatives from the six Centers: the User Advisory Group (UAG), the Programming Environment and Training (PET) program, the Space and Naval Warfare Systems Command (SPAWAR), and Instrumental, Inc.

The BCT will define requirements for common sets of capabilities and functions (the baseline configuration) on allocated High Performance Computing (HPC) systems, Mass

Continued Next Page...


Arctic Region Supercomputing Center
MSRC
ARL MSRC
ERDC MSRC
MAUI HIGH PERFORMANCE COMPUTING CENTER

Storage systems, and Scientific Visualization systems. The BCT is researching the development of documents and tool sets that will be used to verify initial site compliance with approved baselines and to verify that baselines stay in place at each Center.

As part of the compliance process, there is a need to create and maintain a Compliance Matrix Website accessible to the DoD community. This site will list the status of each baseline configuration item at each Center and will be automatically updated to reflect additions and descriptions of approved items.

The goal of the BCT is to enable users to move easily between Centers without having to learn and adapt to unique configurations, reduce the overall user learning curve when using multiple systems at different Centers, increase user efficiency on multiple systems, and provide new functionality without breaking existing conventions at the centers. A brief description of the first six Baseline Configuration projects that have been approved for implementation this year follows.

Common Environment Variables Project

The Common Set of Environment Variables documentation states:

> This project will develop a minimum set of environment variables that represent the same thing to users at each site. The baseline set of variables will be predefined via system login scripts, making them automatically available to users computing at various centers. Additional center-specific environment variables will still be appropriate in order to minimize impact on users. A common set of global environment variables will enhance the portability of user scripts and makefiles across the centers. This will improve user productivity by eliminating tedious, time-consuming and unnecessary efforts put forth by users to determine center-specific names for common variables. This should also help to alleviate the need for a common directory structure across centers. As long as common environment variables are defined and initialized on each system, the actual location of underlying software or file systems is not nearly as important.[1]

Minimum Scratch Space Retention Policy Project

The Minimum Scratch Space Retention Time Policy documentation states:

> This project will create a common baseline policy for minimum lifespan of files residing in volatile “scratch,” “work,” or “temp” file systems on computational servers. DoD users with allocations at multiple sites are currently faced with learning differing scratch purge policies.
>
> Once a standard baseline retention period for user files created within such "work" file systems is established and implemented, users can use the common cross-center minimum retention period as a guide when automating data archival to more permanent locations after job completion. Establishing a baseline policy across all six centers will provide users a consistent expectation from system to system and center to center.
>
> The $WORKDIR environment variable, which is a part of the Common Environment Variables project, will be used in this project. WORKDIR is an environment variable that points to each individual user's work directory on the appropriate large volatile temporary file system (local high speed disk) available on HPCMP high performance computing (HPC) systems. In order to provide sufficient free WORKDIR disk space to DoD users, the following maintenance policy will be implemented on all HPC systems.
>
> All user files in $WORKDIR that have not been accessed or modified within 5 days are subject to deletion. This minimum scratch space retention period was selected because it is deemed long enough to allow users to retain temporary work files under each individual $WORKDIR sub-directory, but is also short enough to prevent the need to delete user files and directory structures less than 5 days old, except under periods of unusually high usage on a computational server. Because workload can vary during the year, system administrators may occasionally need to delete WORKDIR files less than 5 days old to free up sufficient disk space to prevent a system failure. To minimize early deletion of WORKDIR files, users are strongly encouraged to use WORKDIR efficiently and economically and to always archive important data from computational runs in a timely manner after processing is completed.[2]

Common Queue Names Project

The Common Queue Names documentation states:

> This project will institute a baseline set of queues with the same names at each center. This common queue set will be designed to provide support for existing HPCMP-designated work classifications. Additional

Continued Page 31...

CNMOC Enterprise Teams Transform the Operational Environmental Forecasting Use of High Performance Computing

Charles Kleinschmidt, Director, Enterprise High Performance Computing

Good operations, as is good science, are determined by what the data tell you. A Fleet Numerical Meteorology and Oceanography Center (FNMOC) team is engineering the foundations for a new enterprise approach to operational environmental modeling using the High Performance Computing (HPC) resources at the Naval Oceanographic Office Major Shared Resource Center (NAVO MSRC).

Their goal: to reduce the total cost of operations and increase capability by augmenting the production systems at FNMOC in Monterey, CA, with access to the HPC systems over two thousand miles away at the NAVO MSRC, Stennis Space Center, MS.

This initiative was driven by the clear and compelling vision of the office of the Commander, Naval Meteorology and Oceanography Command (CNMOC): a vision of a single enterprise working as one "to provide accurate, timely characterization of the battlespace environment."

Under the guidance of Tom Dunn, the CNMOC Chief Information Officer (CIO), the FNMOC Enterprise High Performance Computing Team (led by Chuck Kleinschmidt and Jay Morford) worked in close collaboration with Dave Cole and Jeff Gosciniak of the NAVO MSRC User Services group to develop and implement an enterprise solution.

The NAVO MSRC is in the unique role of providing operational cycles for the Navy. Naval Oceanographic Office resources, working with the NAVO MSRC, provide critical capabilities to naval oceanographers and meteorologists through the creation of timely climatological and oceanographic forecasts.

These forecasts are distributed to the entire fleet and are essential for supporting naval operations. FNMOC, as part of the Naval Meteorology and Oceanography Command (CNMOC), is a key producer of these products.

The mission of FNMOC is "[to] prepare the marine and joint battlespace to enable successful combat operations from the sea. Exploit the meteorological and oceanographic opportunities and mitigate the challenges for Naval operations, plans, and strategy at all levels of warfare."

At the core of this mission is the timely, reliable execution of computationally intensive forecast models such as those generated by Operational Numerical Weather Prediction (NWP) activities.

NWP is one of the most challenging problems around. Not only are NWP jobs computationally intensive, but the requirement that these jobs be executed within a strict schedule compounds the challenge.

In recent years, the U.S. Navy has placed increasing emphasis on the littoral zone rather than the open ocean. This has increased the need for tactical Meteorological and Oceanographic (METOC) analysis and forecast products in coastal regions.

Phenomena of concern include coastal winds, squalls, and other organized convection, near-shore tides, currents, and surf. They typically occur on small (mesoscale: 1-50 Kilometers (km)) spatial scales and may go through their entire life cycle in a matter of hours.

The Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) was developed by the Naval Research Laboratory, Monterey (NRL-MRY), and implemented at the FNMOC to meet these emerging requirements. An example of COAMPS output is shown in Figure 1.

However, in a COAMPS output the atmosphere and the ocean are coupled, not separate; Surface winds help drive wave heights. Therefore, FNMOC uses coupled models to accurately predict both the weather and the ocean state.

For example, the forecasted wave heights are forced by surface winds derived from the FNMOC weather models. An illustration of the wave height products is shown in Figure 2.

Forecasting the weather and ocean conditions has another layer of complexity: the interactions between the large scale or Global forecasts with the small scale or Regional forecasts. Regional forecasts depend on a timely, global forecast. Global forecasts depend on the timely assimilation of satellite and conventional observations and the analysis of current conditions.

Essentially, forecasting depends on intensive jobs, complex dependencies, and tight schedules: because of these complexities, the resourcing and scheduling of operational weather

Continues Next Page...

forecasts have been historically performed at a single center where process controls are tight and the preeminent focus is getting the forecasts quickly to the warfighting customer. Late forecasts are not very useful.

What has changed to drive a transformation of this process? And, just as importantly, what has changed to make this transformation possible?

The business drivers for transformation were simple: Do more with less.

Warfighter requirements for improved environmental characterization of the battlespace continue to drive computational requirements up. Funding for capitalization programs and operations in support of environmental characterization continues to go down.

A prime technological enabler is the Defense Research and Engineering Network (DREN). Its high bandwidth, low latency connections provide the speed and reliability in the movement of data between geographically distant centers that approaches the transfers between systems on a single computer room floor.

Preliminary measures have shown that remote, distributed operations are feasible—with just marginal increases in turnaround times, schedule variance, and process reliability. Without this network, or one with similar capabilities, remote NWP operations could not work.

The second critical enabler is the implementation of a common “industrial-strength command and control” infrastructure, which will help to ensure the timely execution of critical jobs.

This means a robust, reliable job scheduler, metascheduler, and resource manager (like Platform Computing's Load Sharing Facility (LSF) and MultiCluster) that support real-time job monitoring and the enforcement of federated scheduling policies across collaborative sites. The NAVO MSRC technical team has been hard at work to finalize this installation.

With these components in place, serious parallel operations testing will begin. Then the data—the operational performance metrics–will drive the final engineered solutions, and will ultimately and objectively, determine success.

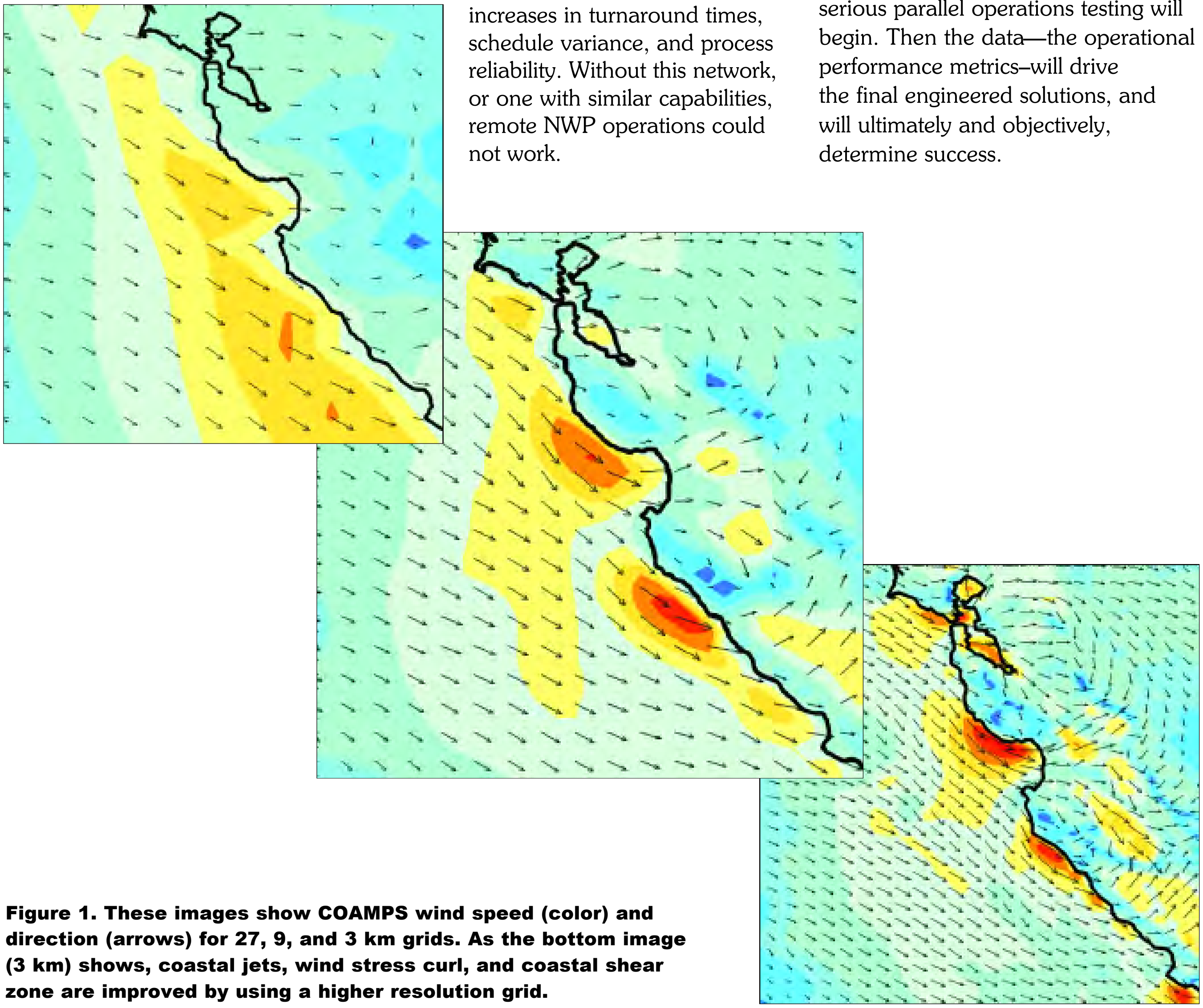


Figure 1. These images show COAMPS wind speed (color) and direction (arrows) for 27, 9, and 3 km grids. As the bottom image (3 km) shows, coastal jets, wind stress curl, and coastal shear zone are improved by using a higher resolution grid.

A footnote (prompted by the potential impact of Hurricane Katrina):

As recent events have unfortunately affirmed, while the quickest and easiest solution may be the single, point to point, stovepipe process, a more resilient and robust solution (with some additional design work) may be to build a grid or networked process.

As happens with the Internet, the failure of one node would not cause the failure of the entire process. This realization is the basis for the FNMOC promotion of a longer-term initiative to research the benefits of building a Globus Grid like the National Science Foundation's TeraGrid.

Significant advantages can be gained by building an HPC grid in accordance with the open standards like Globus, as outlined by the Global Grid Forum (GGF).

The GGF relates to global grids in a manner similar to the Internet Engineering Task Force's relationship to the Internet. These advantages would include a Grid Security Infrastructure (GSI), future interoperability with other similarly engineered grids, and additional methods for high-speed, high-volume data sharing.

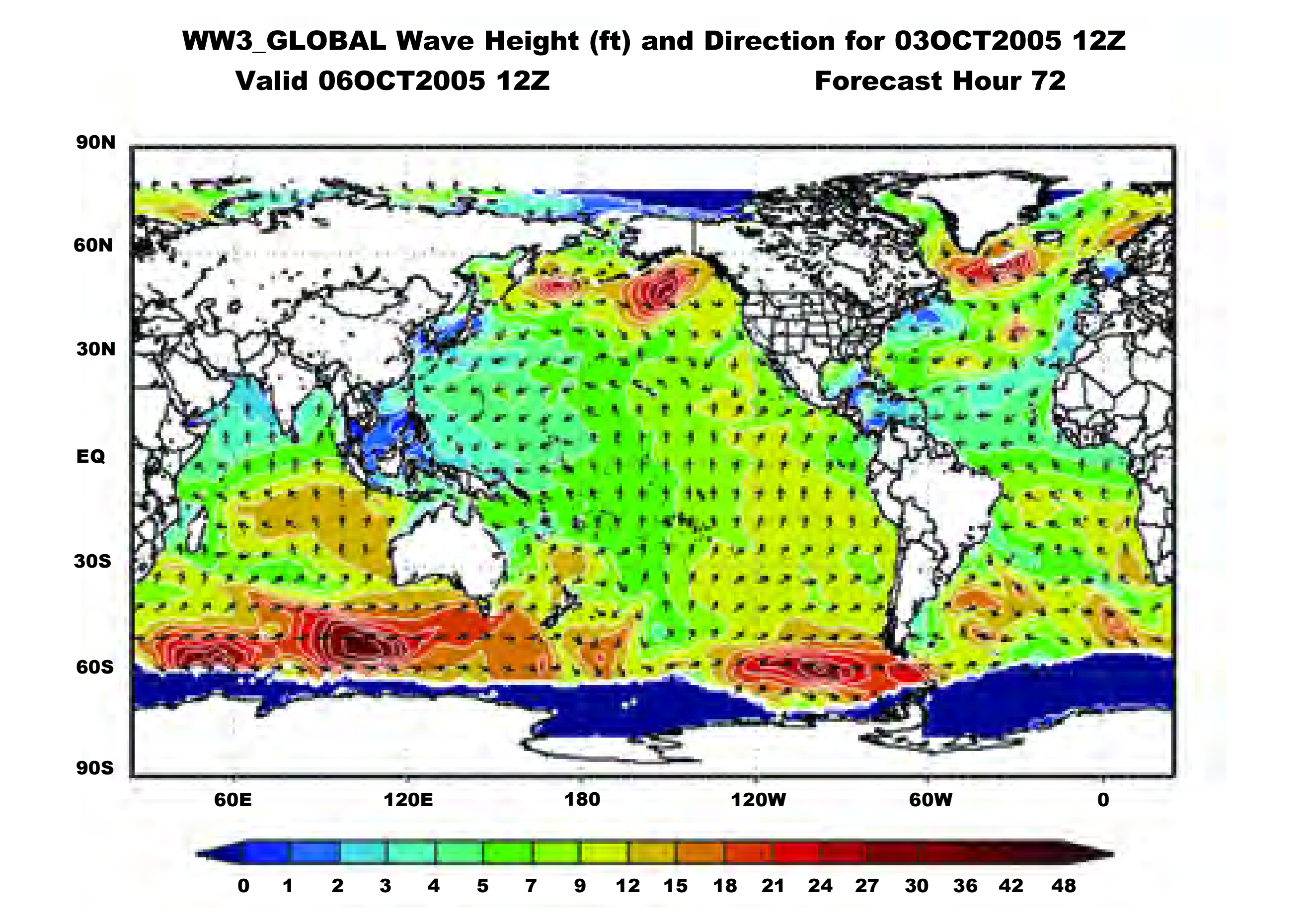


Figure 2. COAMPS output showing global wave height in feet and direction.

The Effects of Hurricane Katrina on NAVO MSRC, Stennis Space Center, MS

Hurricane Katrina and the NAVO MSRC

In the early hours of Monday, 29 August 2005, Hurricane Katrina made landfall, impacting the Louisiana and Mississippi Gulf coasts. This triggered the largest single natural disaster in the United States and provided a rigorous test of the Naval Oceanographic Major Shared Resource Center (NAVO MSRC) Disaster Recovery Plans (DRPs). However, the preparations for the impending hurricane started well before the actual landfall.

On Thursday and Friday, 25-26 August respectively, the NAVO MSRC team checked and verified the data storage resources and ensured that a second copy of the NAVO MSRC data was up-to-date and the second copy process was functioning as normal. In addition, each member of the team completed an employee evacuation plan. These plans were invaluable in allowing our DR Team (DRT) to locate employees after Katrina, verify that they were safe, and determine each employee's needs after the disaster.

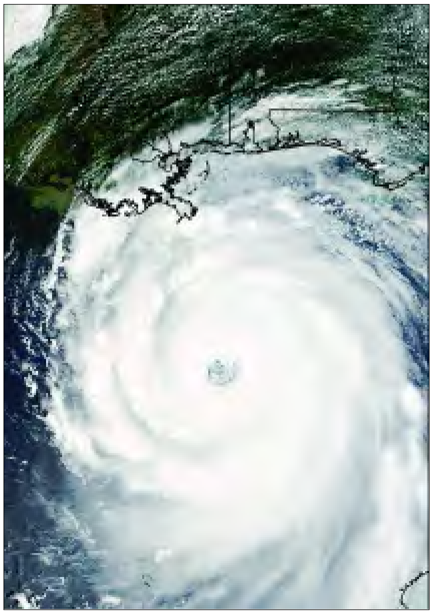
The Inclement Weather Crew (IWC) was notified to assemble by 0700 hours on Sunday, 28 August and prepare to ride out the storm; all other personnel were allowed to vacate the Mississippi and Louisiana coastal area and take their family members to safety. The IWC initially provided 24-hour, seven days a week coverage for both the Central Site Facility (CSF) and MSRC areas during the early phases of the storm.

Additional staff was added to the IWC after Katrina to provide adequate support once the extent and magnitude of the destruction was realized. Although all MSRC support personnel were personally impacted by this catastrophe, they displayed their utmost dedication and professionalism by returning to the Center to assist their colleagues to ensure the Center's continuity of operations.

The IWC performed a number of important tasks before, during, and after the storm. They monitored the 750S and 1750S generators, which were the sole source of electrical power to the MSRC. IWC members also ensured that fuel was delivered to the generator through the storm and its aftermath from existing stockpiles. Once Katrina had passed, the IWC began making inquiries about the status of employees and their well being.

After the storm, the DRT initiated a number of efforts to ensure consistent and reliable operations. They evaluated the state of each resource and performed emergency preventive maintenance on several servers to correct error conditions. While NAVO MSRC services were not significantly disrupted by the storm, the flooding in New Orleans, LA, impacted the MSRC operations. The flooding disrupted the OC-48 connection from the NAVO MSRC to the Defense Research and Engineering Network (DREN).

The High Performance Computing Modernization Program (HPCMP) was instrumental in resolving this problem. They furnished all the resources required to establish a satellite link to the DREN. Although this link was only at 20 megabits (Mbits) per second, it enabled users to continue utilizing NAVO MSRC resources. This satellite link was the Center's only lifeline to the DREN for two weeks while the Wide Area Network (WAN) provider established an alternate path to full connectivity.



Executing the DRP during and after such a catastrophic event provided a number of lessons learned. First, communications in the wake of a major hurricane are essential and almost impossible to establish. Both land lines and cell towers were rendered inoperable by Katrina—satellite telephones were the only means of communication immediately after the storm.

As cell towers were brought back on line, cell phone connectivity and reliability were still very unstable. The DRT and IWC employed text messaging as another means of communication. "Texting" provided an efficient means of communication once rudimentary cell connectivity was established.

Second, this storm demonstrated the risk of a single connection to DREN. While a second path to DREN from the NAVO MSRC was in the planning stages, Katrina underlined the critical need for this capability.

Although the NAVO MSRC successfully weathered the storm, Katrina severely impacted the lives of the MSRC team. Thirty percent of the MSRC team suffered catastrophic damage to their homes, rendering the home uninhabitable or totally devastated. The availability of the MSRC resources during and after Katrina is a great testament to the dedication, effort, and team work of the MSRC support staff.

Enhanced Graphics Capability: SEAHORSE-II

Sean Ziegeler, Visualization Software Engineer, NAVO MSRC VADIC

The Naval Oceanographic Office Major Shared Resource Center (NAVO MSRC) Visual Analysis and Data Interpretation Center (VADIC) has acquired a new system for graphics and visualization to serve as a replacement for the faithful, but dated, SEAHORSE. The new system, SEAHORSE-II, is a powerful graphics supercomputer provided by GraphStream[1] capable of much more than SEAHORSE or any modern graphics workstation alone.

System Differences

The original SEAHORSE was an SGI Onyx2. Its design was unique in that it truly behaved as a single computer. All eight Central Processing Units (CPUs) are located together, a feature known as Symmetric Multi-Processing (SMP). In addition, all of the CPUs could access the 8 Gigabytes (GB) of memory as a single entity, a feature known as a shared-memory model. This made it very easy to program the SGI to use multiple CPUs, a process not really much different than programming a single-CPU desktop system. Unfortunately, such a design is expensive, and the expense grows significantly faster for each CPU added.

Figure 1. SEAHORSE (left) and SEAHORSE-II (right).

SEAHORSE-II's design is much different. Commonly known as a "cluster," the system is really just a collection of ten workstation-class computers mounted together in a metal cabinet. Each computer is referred to as a "node," and the nodes are interconnected via high-speed networks to allow them to work together to form a supercomputer. Table 1 compares the salient characteristics of the two systems; the quantities for SEAHORSE-II result from adding together the components of all ten nodes.

This arrangement of multiple nodes connected together by a network requires a different approach when designing software. The total aggregate memory of the system (32 GB in this case) is no longer located in a single place but is divided among

	SEAHORSE	SEAHORSE-II
CPU	8 x MIPS R 1000: 250 MHz	20 x AMD Opteron: 2.0 GHz
GPU	2 x SGI InfiniteReality-2	2 x NVidia Quadro FX 3400 8 x NVidia 6800 GT
Memory	8 GB	32 GB
Disk	2 TB	6.5 TB
Operating System	SGI IRIX 6.5	RedHat Enterprise Linux 3

Table 1. Comprehensive point-by-point hardware comparison.

	Master Node (2)	Compute Node (8)
CPU	Dual Opterons: 2.0 GHz	Dual Opterons: 2.0 GHz
GPU	NVidia Quadro FX 3400	NVidia 6800 GT
Memory	8 GB	2 GB
Scratch Storage	250 GB	250 GB
Permanent Storage	One master node has 4 TB of storage	None

Table 2. Master and compute node configurations.

each of the nodes. (Table 2 shows how much memory is located in each node.) Effectively, no node can directly see the memory of another node. Known as a distributed memory model, this arrangement means two different nodes working with the same information must explicitly send messages to each other over the network rather than simply access the same location in memory. This is the primary disadvantage of a cluster because it requires the programmer to take extra steps to program for parallel processing.

So, why switch to a cluster when using it is more work? Since each node of the cluster is a standard workstation computer with commodity parts, it is inexpensive to assemble many workstations together to build a powerful computer. Because of this, many of the most powerful supercomputers in the world are clusters of some type. Moreover, toolkits exist to make interprocess communication across multiple nodes easier; for example, the Message Passing Interface (MPI).[2]

The SEAHORSE-II Cluster: Nuts and Bolts

Not all of SEAHORSE-II's nodes are the same. Two out of the ten are dubbed "master nodes." The purpose of the master nodes is to allow multiple external users to login to the system as well as handle large, single-CPU (serial) tasks. For this reason the master nodes contain more memory and a more capable Graphics Processing Unit (GPU). The remaining "compute nodes" can only be accessed from a master node and are used for multiple-CPU (parallel) tasks. Table 2 illustrates the hardware differences between master nodes.

Note that in Table 2, only one of the master nodes has 4 Terabytes (TB) of permanent storage. To make matters easier, that permanent storage is shared with all of the other nodes via the network using the Parallel Virtual File System 2 (PVFS2)[3]. PVFS2 allows each node to act as if the storage disks were physically attached to it. Even more convenient—all of the 250-GB "scratch" (temporary) storage disks on each node are tied together to create one large, virtual 2.5-TB scratch storage space. This simplifies software development since any program on a given node can assume that it will have transparent access to any data stored on disk.

One of the daunting tasks usually associated with a cluster is the management and administration of so many nodes. Traditionally, every node of the cluster must be taken care of individually for software installations, operating system updates, etc. SEAHORSE-II is provided with Warewulf[4] cluster management software, which allows a single configuration to be managed on one node, and then propagated to all of the other nodes automatically. Essentially, the cluster can be administered as a single system.

All nodes of the cluster are interconnected by two networks. The first, Gigabit Ethernet, running at 1 Gigabit per second (Gb/s), is for light traffic such as system-level communication. The second network, Infiniband[5], running at 8 Gb/s, is for the heavy traffic associated with parallel graphics applications and the parallel file system, PVFS2.

The Infiniband software bundle provided by Mellanox[6] includes an Infiniband-capable version of MPI known as MVAPICH[7] that has recently added support for multiplexing network traffic over two Infiniband channels. This is possible because the Infiniband adapters for each node are plugged into PCI-Express x16, which can support twice the Infiniband bandwidth. Since each node has two Infiniband connectors and the Infiniband switch that connects to all the nodes has enough ports for all 20 channels (two from each node), this configuration was adopted. Known as "Dual-Rail" Infiniband, this doubles the theoretical bandwidth to 16 Gb/s. The Infiniband connections are shown in Figure 2.

One additional feature of SEAHORSE-II is the Keyboard-Video-Mouse (KVM) switch. The KVM switch allows a single monitor, keyboard, and mouse to switch to viewing and controlling any node of the cluster. While most

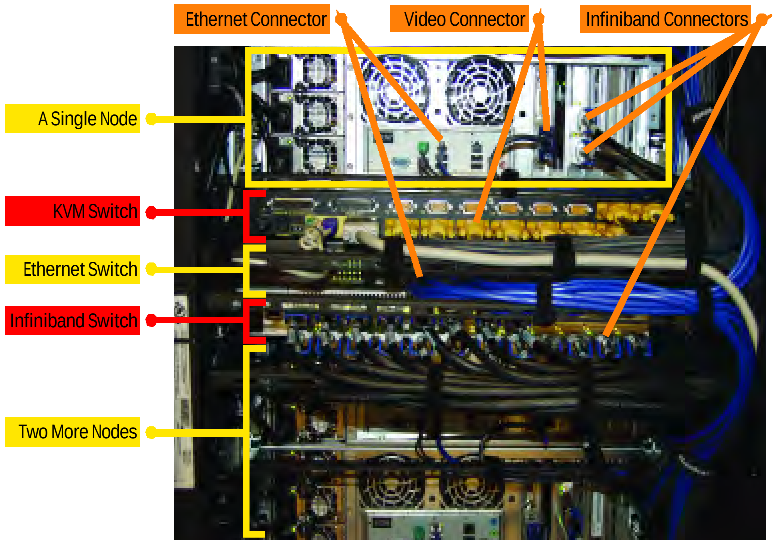


Figure 2. SEAHORSE-II internal components.

operations only require working at a master node, this allows the visualization staff to examine compute nodes to diagnose problems and optimize parallel graphics applications.

It's All in the Protocol

Besides being very fast, Infiniband is a flexible network fabric. It supports a number of types of communication, known as protocols. The most common of these is Internet Protocol (IP). Obviously, IP is the underlying protocol of the Internet and is the most popular because of its widespread use. Unfortunately, IP is well-known to be inefficient on Infiniband networks, so its use is discouraged, and it is typically only used for programs that cannot be modified to run on anything but IP. Despite that inefficiency, IP is still much faster over Infiniband than Gigabit Ethernet.

For applications that can be slightly modified, Infiniband supports Sockets Direct Protocol (SDP). SDP is significantly more efficient over Infiniband, yet its programming interface is so similar to IP that programs can often be converted in as little as one line of code.

The IP and SDP protocols for Infiniband are actually created on top of another protocol known as Verbs Application Programming Interface (VAPI). This is a more direct path to the Infiniband channel, but it requires a complete rewrite of the communications sections of any program. The reward for that effort, however, is use of nearly the full bandwidth (8 Gb/s) of an Infiniband channel.

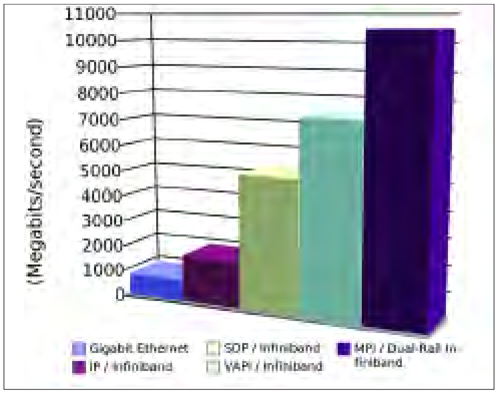
One other option is to use MPI. While MPI is much more than just a communications protocol, it does function as such. MPI for Infiniband is written on top of VAPI, making its communication efficiency similar to that of SDP. However, SEAHORSE-II's MPI implementation, MVAPICH, will automatically load-balance across two Infiniband channels, resulting in a total bandwidth greater than that of VAPI on a single channel. This makes MPI an attractive approach for parallel graphics applications. Figure 3 compares all of the above protocols to each other and to Gigabit Ethernet.

Modern Graphics Applications

In addition to MPI, the VADIC staff plans to make available several other parallel toolkits and programs that are specifically for graphics. The first of these is Chromium[8], which can automatically parallelize some graphics programs. Chromium is easier to use than MPI and in many cases is very efficient. Unfortunately, Chromium supports only the IP and SDP protocols, so the dual-rail Infiniband would not be utilized.

ParaView[9] is a single application designed to easily visualize large data sets. It is intended for a user to import, view, and manipulate the data all with an easy-to-use interface and no programming required. Another advantage of ParaView is that it is

Continued Next Page...



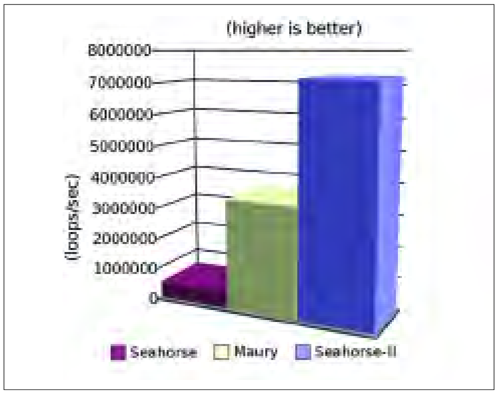
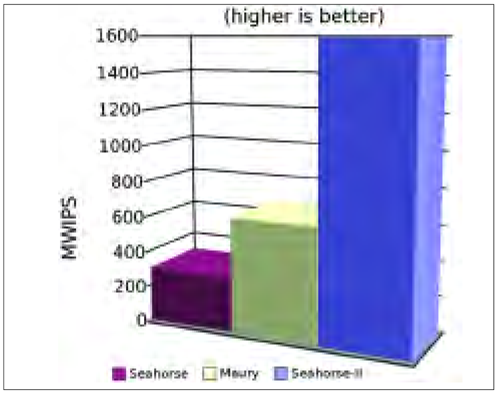


Figure 3. (above) Benchmarked bandwidths of Ethernet and Infiniband using several protocols.

Figure 4. (left) Integer CPU benchmarks.

Figure 5. (right) Floating-point CPU benchmarks.



parallelized using MPI, so it has the potential to take full advantage of dual-rail Infiniband.

Performance Benchmarks and Discussion

To illustrate the improvement of SEAHORSE-II over SEAHORSE, the VADIC staff performed a battery of benchmarks on both systems, targeting CPU, memory, disk, overall performance, and graphics throughput. However, such a comparison isn't really fair given the age of SEAHORSE, so a more modern system was included in the various benchmarks.

The first of these benchmarks (for the CPU) show the performance for both integer-based and floating-point math. The extra comparison system, MAURY, has dual 32-bit Intel Xeon 2.4 Gigahertz (GHz) processors. The reason for SEAHORSE-II's stellar performance with respect to MAURY, despite a slower clock speed (2.0 GHz), is that it is a 64-bit processor, which can handle more information per instruction. See Figures 4 and 5 for the results.

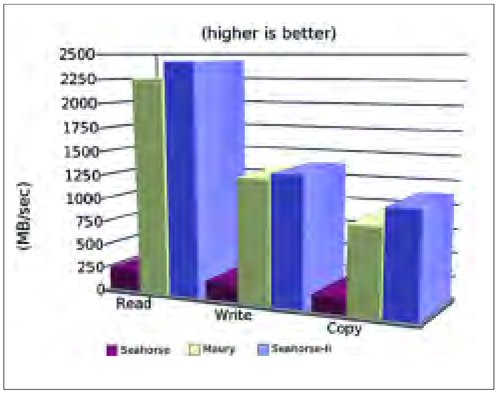
The second set of benchmarks show the memory performance, specifically the bandwidth and latency of memory accesses. Three types of access patterns were tested for bandwidth: (1) a read from memory; (2) a write to memory; and (3) a copy from one part of memory to another.

For latency, the three types of memory were tested for delays in access: (1) L1 cache; (2) L2 cache; and (3) main memory. Note that in this case, smaller values are better. The results are shown in Figures 6 and 7.

The storage disk benchmarks test the throughput of reading and writing data to and from the disks. Three file sizes were used: (1) small files testing short, bursty accesses; (2) large files testing long, streaming accesses; and (3) medium files testing moderate accesses.

SEAHORSE-II's significantly better performance here was surprising because MAURY's disks use Small Computer Systems Interface (SCSI) versus SEAHORSE-II's Serial Advanced Technology Attachment (SATA) interface. SCSI is supposedly better for bursty accesses (small files) and about comparable to SATA for streaming accesses (large files). However, it is likely that the SCSI on MAURY is a generation behind. The results are shown in Figure 8.

All previous benchmarks were concerned with a single node's performance. The following benchmarks examine the performance of the overall system. The first of these is an aggregate performance of all compute nodes or all processors in the case of SEAHORSE. The benchmark is known as Linpack, which is used to determine the Top-500[10] computing systems in the world. In this set of benchmarks, the VADIC staff used a different system for the additional comparison, PEPE, which is a 10-node IBM 1350 compute cluster. As expected, the performance of SEAHORSE-II was about twice that of PEPE, given that



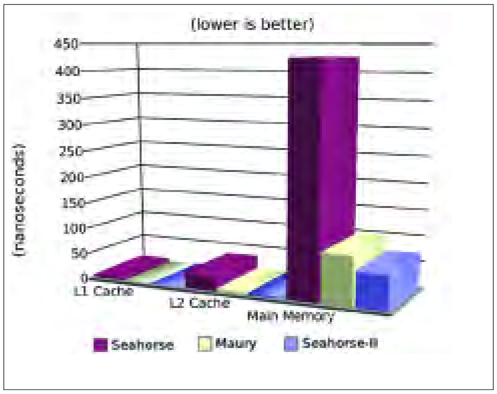
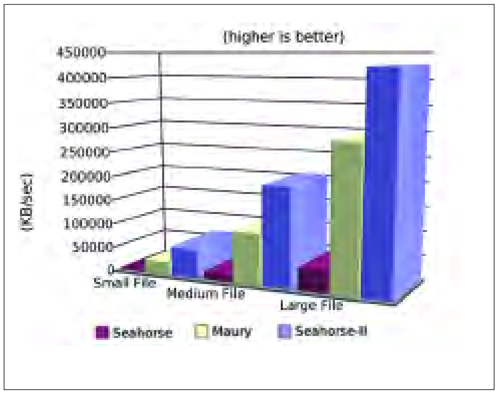


Figure 6. (above) Memory bandwidth benchmarks.

Figure 7. (left) Memory latency benchmarks.

Figure 8. (right) Storage disk benchmarks.



(1) PEPE has Xeon processors similar to that of MAURY, and (2) PEPE has a Myrinet interconnect which is about half of the speed of Infiniband. See Figure 9 for the comparison graph.

The final benchmark compares the graphics throughput (frames per second versus number of polygons) of SEAHORSE, MAURY, and SEAHORSE-II in various configurations. SEAHORSE and MAURY's benchmark ran serially on one GPU. This was also done for SEAHORSE-II on a master node to compare the individual GPUs directly.

As expected, SEAHORSE-II's GPU only slightly outperformed MAURY's, whose GPU is also NVidia but one generation behind. The true gain in performance was seen when using multiple GPU's on SEAHORSE-II.

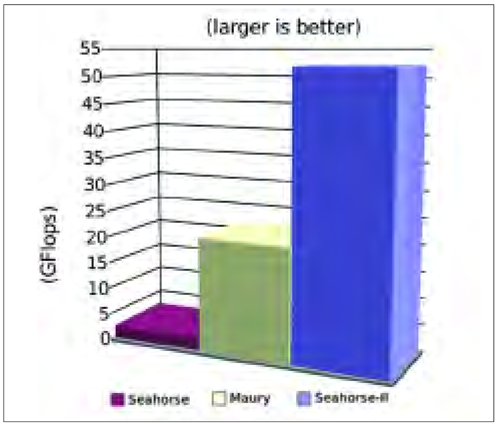
For low numbers of polygons, 5 Nodes (four to render parts of the image, one to composite them together) is faster because it reduces the number of required compositing operations. However, as the number of polygons begins to overwhelm any single GPU, the 9 Node configuration becomes the fastest approach. As expected, for large datasets, more nodes are required to maintain interactive rendering rates. Figure 10 illustrates these benchmark results.

The Final Transition

Not only is SEAHORSE-II a more than suitable replacement for SEAHORSE, it provides unparalleled graphics and visualization capability for the NAVO MSRC. When SEAHORSE is formally retired, SEAHORSE-II will be renamed to "SEAHORSE."

The VADIC staff plans to use it for visualizing large datasets and as a remote visualization server. Remote users of MSRC resources with a need to examine their potentially large computational output will be able to log in from their site and explore their data without having to transfer that data back to their site. Local users can work directly at the system and even route the display to a projection screen.

SEAHORSE-II is the first step in creating a truly interactive and accessible computing center.



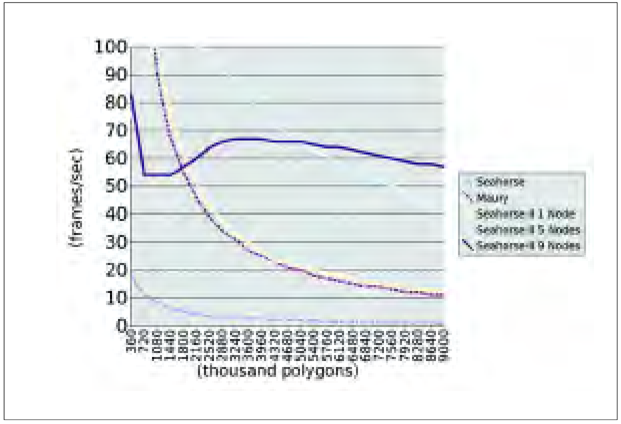


Figure 9. (left) Aggregate parallel performance (Linpack) benchmark.

Figure 10. (right) Serial and parallel graphics throughput benchmarks.

Resources

1. GraphStream
   http://www.graphstream.com
2. Message Passing Interface (MPI)
   http://www-unix.mcs.anl.gov/mpi/
3. Parallel Virtual File System 2 (PVFS2)
   http://www.pvfs.org/pvfs2/
4. Warewulf Cluster Management Software
   http://warewulf.lbl.gov/pmwiki/
5. Infiniband Trade Association
   http://www.infinibandta.org/home
6. Mellanox Technologies
   http://www.mellanox.com/
7. MVAPICH MPI-1 Implementation
   http://nowlab.cse.ohio-state.edu/projects/mpi-iba/
8. Chromium
   http://chromium.sourceforge.net/
9. ParaView
   http://www.paraview.org/
10. Top-500 Supercomputer Sites
    http://www.top500.org/

THE PORTHOLE

Representatives of the National Polar-orbiting Operational Environmental Satellite System (NPOESS) tour the MSRC Operations Center.

Naval Meteorology and Oceanography (METOC) students visit the Visual Analysis and Data Interpretation Center.

Naval Meteorology and Oceanography (METOC) students visit the MSRC Operations Center.

CAPT Andy Brown awards Steve Adamec and the MSRC team a special recognition award for service before and after Hurricane Katrina.

Mississippi State Representatives Dirk Dedeaux, J. P. Competta, and Billy McCoy, are accompanied by RDML Thomas Q. Donaldson, V., USN (Ret) on a tour of the MSRC Operations Center.

Personnel from the Republic of Korea (ROK) visit the NAVO MSRC Operations Center.

Dave Cole leads Susan MacLaughlin and LCDR Timothy Smith from the Defense Intelligence Agency,Technical Analysis and Development Office (DIA/DTT-1) on tour of the MSRC Operations Center.

NASA and NAVOCEANO personnel visit the Visual Analysis and Data Interpretation Center.

Representatives of the National Oceanic and Atmospheric Administration (NOAA) visit the MSRC Operations Center.

NAVO MSRC PET Update

Eleanor Schroeder, NAVO MSRC Productivity Enhancement and Technology Transfer (PET) Government Lead

My original plan was to write about some of the successes that Programming, Environment and Training (PET) Component 1 achieved in Contract Year 4, to bid a fond farewell to Dr. Leslie Perkins, the PET Program Manager for these past four years, to welcome Myles Hurwitz, the new PET Program Manager, and to say my own farewell from this program.

But then Hurricane Katrina visited us and left us with such a feeling of confusion and turmoil that I find it so difficult now to try and focus on the "normal" aspect of this article.

I and my family were very fortunate that the damages we suffered were minor in comparison to so many others. But others in our PET family here were not so fortunate.

Three of our staff have lost their homes and most, if not all, of their possessions. The devastation that has been wrought on this area will take many years to heal—and some scars will always remain. But the people in this area are for the most part resilient, and I have confidence that this area will rebuild and will be stronger and better for it.

As for PET Component 1, our remote site partners at the University of Texas and the University of Tennessee, as well as our on-site representatives at the U.S. Army Engineer Research and Development Center (ERDC), Air Force Weather Agency (AFWA), and in Monterey are continuing to help our users in every way they can.

Our Climate/Weather/Ocean Modeling and Simulation (CWO) On-Site here at Stennis, Dr. John Cazes, has been active in not only helping our CWO users continue in their activities, but has also pitched in to help the NAVO MSRC staff as well. And Dr. Tom Cortese, our Computational Environments (CE) On-Site, has been temporarily housed at the National Center for Supercomputing Applications (NCSA) in Illinois. We are very grateful to them for their assistance. So, our work continues.

That said, there are success stories for Contract Year 4, and they should be highlighted.

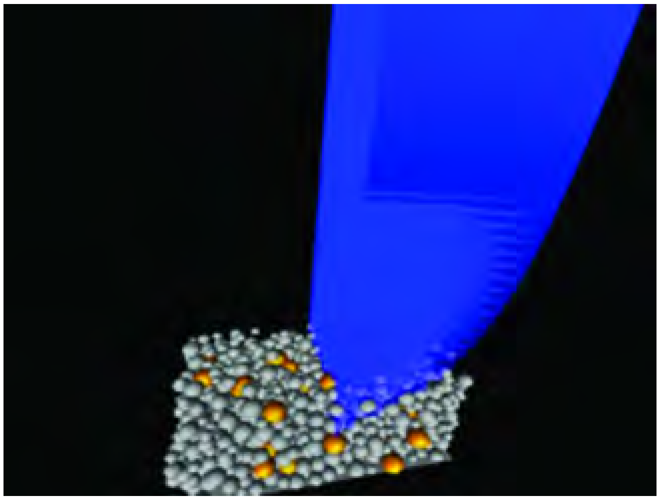
MPI-BASED ELCIRC

In CWO, Dr. Tim Campbell (former CWO On-Site at the Stennis Space Center (SSC)) developed a Message Passing Interface (MPI)-based, domain decomposition version of ELCIRC, with communication for backtracking across subdomain boundaries. The parallel model ensured that no pre- or post-processing was required and also ensured bit-for-bit matching.

Because coastal ocean modeling techniques are still evolving (especially unstructured models), there is no single model that satisfies the physical demands of modeling regions of interest for the Navy. ELCIRC is state-of-the-art, providing Naval Research Laboratory (NRL) researchers with new modeling abilities that are not available through other unstructured grid models such as ADCIRC or QUODDY.

Unfortunately, but as is typical in modeling, in order to gain better physical representation one must "pay" a higher computation price. The parallel version of ELCIRC is now available to NRL (DoD) researchers, enabling them to apply the model to regions of interest that were previously unfeasible with the serial code. This will result in improved coastal ocean predictions for the Navy.

Littoral Dynamics addresses basic research problems utilizing high resolution numerical models to understand and predict physical processes at fine scales. Hydrodynamic modeling utilizes a Navier-Stokes solver to determine depth-dependent velocity profiles, vorticity, bed shear stress, cross-shore pressure gradients, and water depths-parameters that are important for understanding sediment transport. Sediment transport modeling at the grain scale is achieved through discrete particle simulations based on the equations of motion for individual grains and conservation of momentum principles for interactions.



MPI-BASED ELCIRC—Schematic of discrete particle simulation; sand grains are shaded and colored spheres; mean sand grain size is about 1.1 mm; fluid motion (blue slabs) is parallel to the bed.

Modeling of sediment transport usually requires parameterized continuum descriptions. A Discrete Particle Model (DPM) can be used to directly model the motions of individual sediment grains (particles) on a small differential element of the seafloor.

A parallel version of a simplified DPM was developed to solve Newton's equations of motion, which are solved for each particle (translational and rotational motion): inter-particle forces (combination of elastic theory and empirical) and fluid-particle forces (buoyancy, drag, fluid acceleration). This work is supported by Joseph Calantoni and Todd Holland, both of NRL-SSC. These simulations have already proved useful by showing the importance of horizontal pressure gradients to net sediment transport: http://postoffice.nrlssc.navy.mil/littoral%20dynamics/nummodel.html

JOINT ENSEMBLE FORECASTING SYSTEM

Dr. John Romo, our CWO On-Site in Monterey, worked with Chuck Kleinschmidt, Mike Clancy, Doug Wegner, and Mike Sestak at the Fleet Numerical Meteorology and Oceanography Center (FLENUMMETOCCEN) to prepare for the Joint Ensemble Forecasting System (JEFS). FLENUMMETOCCEN and AFWA were recently awarded a joint Distribution Center (DC) project to create an ensemble forecasting "grid" with new IBM computing systems installed at each location.

The plan is to integrate these two systems with grid computing technologies in order to facilitate high-throughput forecasting ensembles based on the Weather Research and Forecast (WRF) framework, using grid scheduling, automated data transfers, etc., and then extend this grid to systems at other locations.

Their work has accelerated the rate at which users can begin using the new HPC system deployed at Monterey, and thus the pace of the JEFS project. This work will allow accurate assessment of resource requirements (both time and space) for batch scheduling of CWO jobs on the JEFS grid and will help FLENUMMETOCCEN gauge its effectiveness in using HPC resources and meeting efficiency and utilization targets.

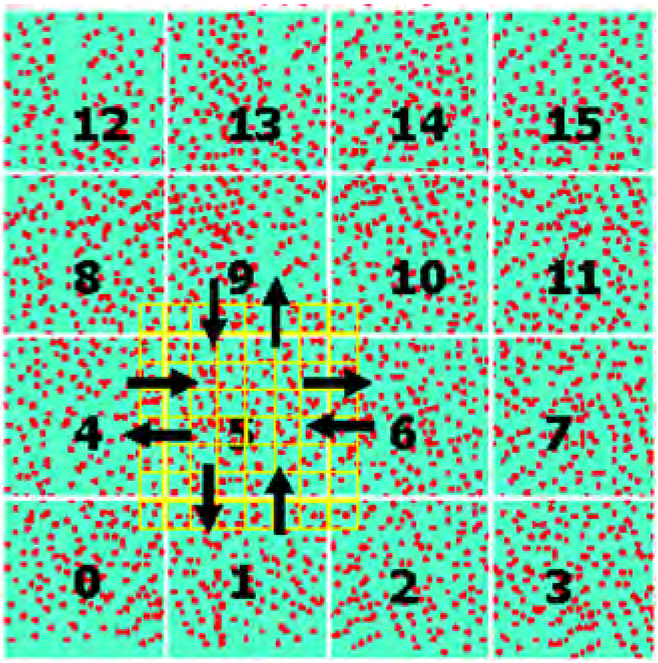
CHSSI EQM04

Dr. Jeff Hensley, our Environmental Quality Modeling and Simulation (EQM) On-Site at ERDC, assisted Jim Westervelt (ERDC/CERL), Tom Cole (ERDC/EL), and Jerry Lin and Aaron Byrd (ERDC/CHL) to beta test CHSSI EQM04.

This portfolio project involved parallelization and enhancement of four codes: (1) WASH123-D - a Finite Element (FE) surface water/groundwater code; (2) GSSHA - a two-dimensional rainfall/runoff watershed model; (3) CE-QUAL-W2—a hydro/water quality code; and (4) mLEAM—the Military Land Evolution and Assessment Model used to model the effect of urban growth around military bases.

In this project the four codes were coupled together to demonstrate that they could interoperate to model the effect of urbanization on a large watershed. Dr. Hensley's expertise in modeling in this field and his familiarity with the HPC systems made him an excellent resource to serve as a beta tester.

The beta test of CHSSI EQM04 highlights a strong feature of the PET program: the ability to effectively interact with and assist other programmatic efforts in the HPCMP arena. In an e-mail R. F. Athow (Deputy to the EQM Computational Technology Area (CTA) Lead) stated, "I am reporting to you that Dr. Hensley did an excellent job as the Beta Tester for the CHSSI EQM-4 test series. His efforts (and of course those of the developers) resulted in a successful test conclusion for EQM-4. Dr. Hensley received collective praise from all the developers as well as thanks from the EQM CTA Lead. PET support to this CHSSI project was vital to its successful conclusion."



MPI-BASED ELCIRC—Parallel domain decomposition of a particle system; white boxes with numbers indicate subdomains with processor assignment; yellow boxes indicate individual subcells used for particle linked-lists and interprocessor communication; arrows indicate nearest neighbor communication of subdomain boundary particles.

APPLYING DG METHODS TO MODELING INFILTRATION IN A TWO-PHASE AIR-WATER MODEL

The EQM team at the University of Texas—Dr. Mary Wheeler, Dr. Clint Dawson, and Dr. Owen Eslinger—successfully applied Discontinuous Galerkin (DG) methods to modeling infiltration in a two-phase air-water model. This approach involved a primal DG/LDG IMPES formulation. Recent results have shown that DG can treat the wetting and drying problem.

CONSISTENT COMPUTATIONAL ENVIRONMENT

In CE, the Consistent Computational Environment effort continues. A suite of software has been identified that should be consistently maintained across the MSRCs and Allocated Distributed Centers (ADCs). This suite includes performance analysis tools, numerical libraries, and data management systems.

These software packages are currently maintained across resources at the ERDC MSRC, U.S. Army Research Laboratory (ARL) MSRC, NAVO MSRC, Aeronautical Systems Center (ASC) MSRC, Maui High Performance Computing Center (MHPCC), and the U.S. Army Space and Missile Defense Command (SMDC). Furthermore, efforts have been initiated to add the Arctic Region Supercomputing Center (ARSC) to the supported centers.

Currently, the software packages are installed and maintained under a publicly available directory accessible through the predefined environment variable $PET_HOME. This allows users to use the $PET_HOME environment variable to locate the software regardless of machine or Center. Further deployment information is being maintained in the Computational Environments Software Repository, available through the Online Knowledge Center (OKC) or http://rib.cs.utk.edu/rib3app/catalog?rh=35. This acts as a centralized location where users can get information about what software is being maintained by CE on what resources.

Finally, efforts have begun to formalize the mechanism used to provide users with support regarding the software being maintained by the CE team and to inform them when versions change. Since the CE team is installing and maintaining the software, a mechanism is needed for user support questions to be forwarded to the CE team. It is often important for users to continue to use the same version of some software. By formalizing this notification mechanism, we ensure users are not surprised by a change in the software.

DYNAMIC PROCESS MANAGEMENT

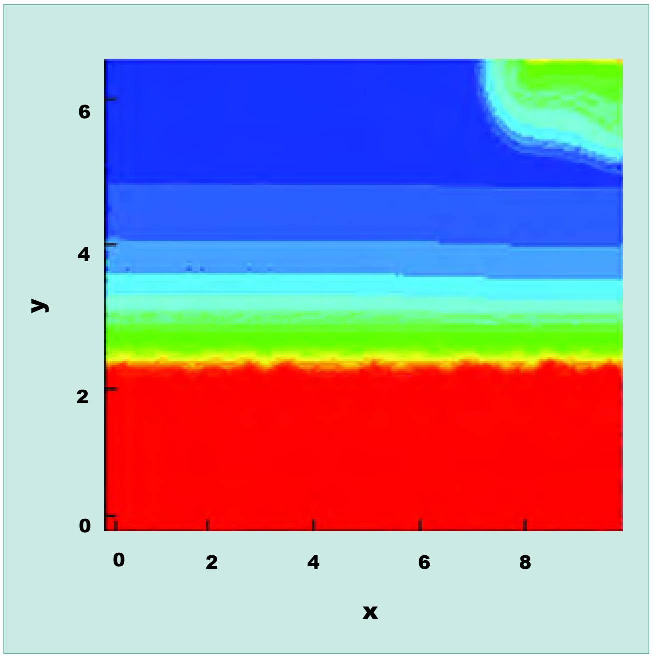
A project dealing with Dynamic Process Management (DPM), led by Dr. Shirley Moore at the University of Tennessee, included Dr. Richard Linderman of AFRL/IF and the Joint Battlespace Infosphere (JBI) system. This team developed the MPI-2 standard, which introduced the ability to dynamically add processes during program execution.

The original MPI standard specification provided for the ability to define the number of processes needed at the time of program initialization. While this approach works well for many regular applications, it is insufficient for certain classes of applications.

Furthermore, under the original MPI standard, increasing performance reduces efficiency, and vice versa. An ideal situation would allow resources to be added during times of heavy workload and allow resources to be returned to the system during times of reduced load.

The MPI-2 standard introduces the ability to dynamically add processes during program execution. This new functionality was studied, and an additional method for dynamically terminating processes at runtime was discovered. Though this functionality is not explicitly part of the standard, the method for terminating processes was tested extensively and found to be robust.

This allowed the DPM team to explore using this dynamic process management functionality as a solution to the problem of varying load applications. The code for the JBI application was studied to determine the communication



APPLYING DG METHODS TO MODELING INFILTRATION IN A TWO-PHASE AIR-WATER MODEL—Two-phase air-water model benchmark problem.

patterns in an effort to find the best method for adding DPM. Once the application characteristics were understood, a prototype was developed, mimicking the interactions of the real application.

This prototype was designed so processes can be added during periods of heavy load and processes can be terminated during periods of reduced workload. While this work was done on a particular application, the methodologies used are valid for many applications that experience varying loads. This is particularly true for pipelined applications that receive workload from a sensor or other collection device.

In an e-mail dated 16 February 2005, Dr. Linderman stated: "...what he has done may be very valuable in the short term. The parallel-pipeline he converted from MPI to MPI2 is the core of our pub-sub info management system we are currently deploying on the Distributed Interactive HPC test bed. It could be enhanced by some auto load balancing capabilities that this code may bring..."

Farewell

These are just a few of the successes that PET Component 1 has achieved this past year and are indicative of the continued support that the user community will see in future years.

To close, I would like to bid farewell to Dr. Leslie Perkins. She was the driving force behind the successes that PET has achieved in its new iteration. Her leadership and her determination to ensure that the user community's needs were being met by our team have helped to make this a stronger and much more user-focused program. We wish her well in her new career with the Air Force.

And we do welcome Mr. Myles Hurwitz as the new PET Program Manager. Myles worked closely with Dr. Perkins over the past three years and is aware of what the PET program is capable of achieving. I am confident that he will continue to ensure that PET continues on in its mission and become and even stronger program.

Lastly, I'd like to bid my own farewell. After almost eight years as the PET government lead here at NAVOCEANO, I've decided to move to different area of NAVOCEANO and explore new career opportunities. I want to thank all of you who I've worked with over the past eight years for making this job become a much easier one with time. I will truly miss working with many of you and hope that our paths will again cross one day.

PET Summer Interns Update

Tom Cortese, NAVO MSRC Productivity Enhancement and Technology Transfer (PET) Computational Environments (CE) Onsite

Fall is here, and once again it is time for our PET summer intern report. There were two PET interns at the Stennis Space Center (SSC) this summer, both working in the Climate, Weather, and Ocean Modeling (CWO) functional area with researchers at the Naval Research Laboratory-Stennis Space Center (NRL-SSC). The PET summer interns spent ten weeks learning about High-Performance Computing, working on a project, and making a final presentation. What follows is a summary of their experiences based on their final presentations.

Gerald Franklin, a senior majoring in Computer Science at Jackson State University, worked with Drs. Cheryl Ann Blain and Chris Massey at NRL-SSC. His project involved using a finite-element code called ADCIRC to model a section of the Mississippi River near Audubon Park in New Orleans, LA, and then comparing the results with multi-line bathymetry data from Tulane University.

ADCIRC is a numerical model that describes hydrodynamics in rivers and coastal waters. More specifically, it is a system of computer programs that use finite-element discretization on unstructured meshes for solving time-dependent, free-surface circulation and transport problems in two and three dimensions. It was developed by Dr. Rick Luettich at the University of North Carolina at Chapel Hill and Dr. Joannes Westerink at the University of Notre Dame; development is now conducted by a team of researchers including several at NRL.

Gerald's first task was to create a computational mesh, which was achieved using MATLAB functions written by NRL researchers. One function created a grid outline describing a rectangle and an annular region, while another function created a mesh within the outline with specified height, width, and evenly-distributed nodal spacing.

The bathymetry data obtained from Tulane University consisted of coordinates in degrees, minutes, and seconds, bathymetry, and mean water level. However, high-resolution computations with ADCIRC are carried out in Cartesian x-y coordinates. To overcome this, Gerald wrote a MATLAB function that first converted the coordinates from degrees, minutes, and seconds to degrees of latitude and longitude, and then finally into Cartesian x-y coordinates. The mean water level was also subtracted from the bathymetry.

After generating the computational mesh, Gerald was ready to run ADCIRC, but first he had to familiarize himself with the format of ADCIRC's input (which determine what computations will be performed by ADCIRC during a particular run) and output (which store the numerical results) files.

Since Gerald was modeling a section of a river, water flows into one boundary of the grid and out through a different boundary; the banks of the river are considered impermeable. The inlet and outlet boundary conditions can be either specified-elevation or specified-flux, and were computed using a program called XGRIDIT.

Several different ADCIRC simulations were performed using different input parameters: coarse-grid and fine-grid, as well as elevation or flux boundary conditions. Boundary conditions for the high-resolution runs were extrapolated from the results of lower-resolution runs.

Gerald used the KRAKEN NAVO MSRC supercomputer–an IBM SP-4 with 368 eight-processor nodes–to execute his ADCIRC runs. One of his models required 1,728,000 time steps in order to simulate three days of river flow. He tried running this model on a single processor, but only about 150,000 time steps had completed after three days of wallclock time. This same model run was able to complete in 90 minutes when using 128 processors on KRAKEN.

After the numerical simulations were complete, Gerald was able to analyze the predicted elevation and velocity results, and noted several differences and similarities due to the defined bathymetry shown in Figure 1.

Gerald would like to thank PET Component One for the opportunity to participate in the PET summer intern program, Dr. Blain for her patience and instruction of elevation and velocity, and Dr. Chris Massey for his help in MATLAB and other problems.

The other PET intern this summer was Allison Scogin, a senior majoring in Computer Engineering at Mississippi State University. Allison spent her summer working for Jim Dykes at NRL-SSC, running the Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS) and WaveWatch III ocean model codes.

COAMPS was developed by the Marine Meteorology division of NRL-Monterey in California. The atmospheric components of COAMPS, a complete three-dimensional data assimilation system comprised of data quality control, analysis, initialization, and forecast model components, are used operationally by the U.S. Navy for short-term numerical weather predictions for various regions around the world.

COAMPS features include a globally-relocatable grid, user-defined grid resolutions and dimensions, nested grids, an option for idealized or real-time simulations, and code that allows for portability between mainframes and workstations. The analysis component uses OpenMP, which means that it can run on multiple threads on a shared-memory machine. The forecast component runs with Message Passing Interface (MPI), allowing for distributed-memory parallelism.

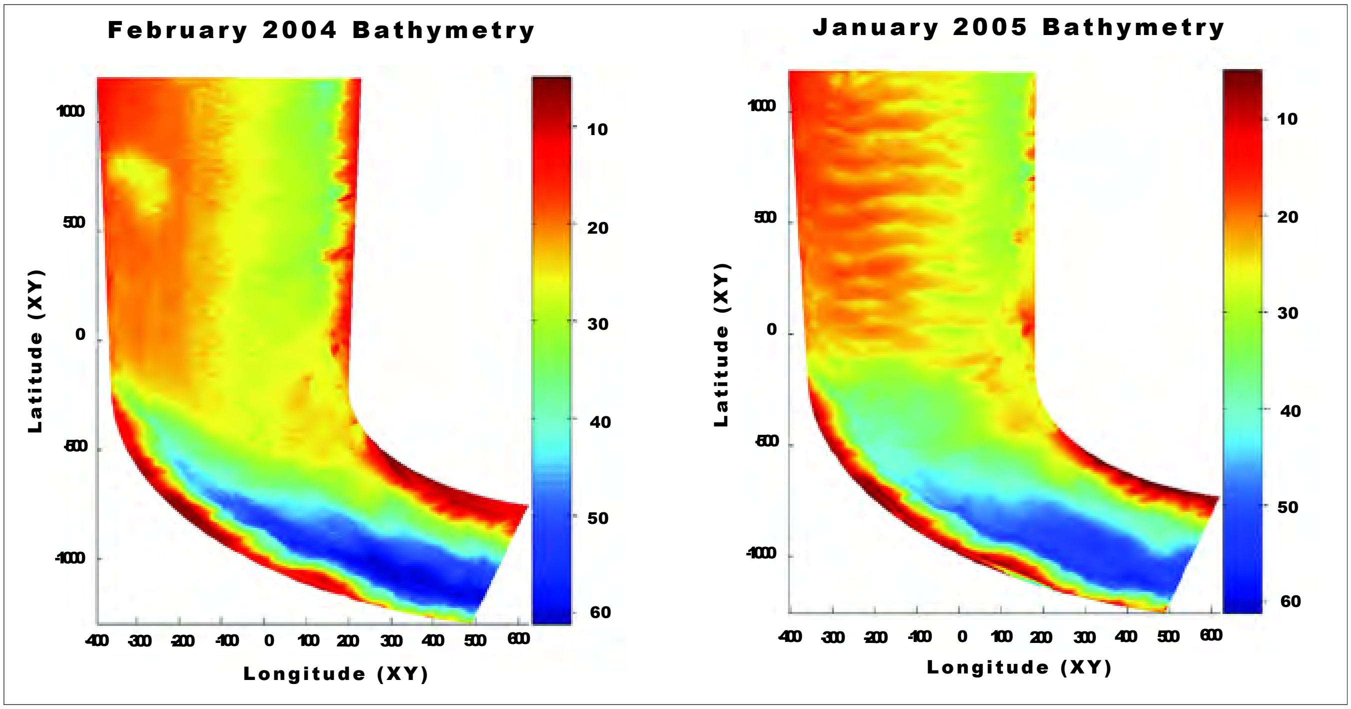


Figure 1. Bathymetry of the Mississippi River near Audubon Park in New Orleans, LA, showing differences and similarities between February 2004 and January 2005. ADCIRC simulations were run for two different time periods usung this bathymetry.

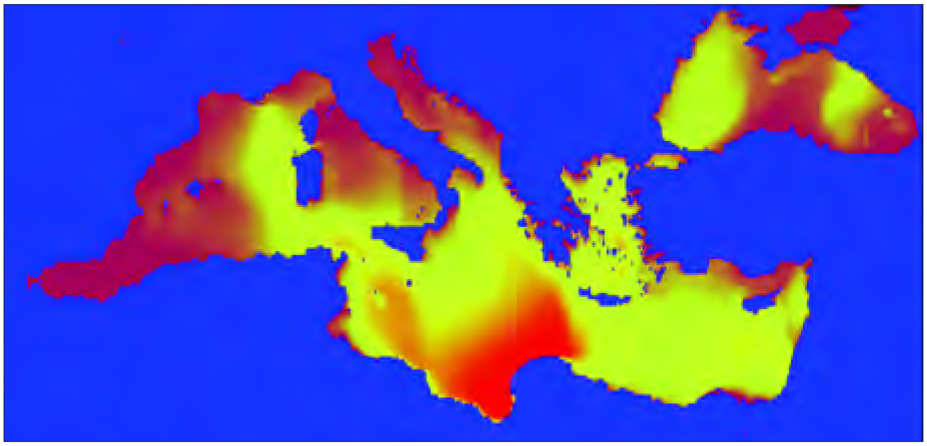


Figure 2. Wave heights over the Mediterranean and Black Seas during November 2002 as predicted by WaveWatch III.

Allison had also participated in the 2004 PET Summer Intern Program and had completed a ten-year hindcast, or re-analysis, with WaveWatch III, the third-generation spectral method wave model developed at the National Oceanic and Atmospheric Administration (NOAA), during the previous summer. This summer, she went one step further and set up WaveWatch III to run a hindcast for the Mediterranean Sea and Black Sea. In addition, she ran a COAMPS hindcast and monitored operational runs of COAMPS, which needed to be completed within rather strict time constraints. Figure 2 shows wave heights over the Mediterranean and Black Seas during November 2002 as predicted by WaveWatch III. She was hoping to complete a month of simulation of the Mediterranean Sea, but there was insufficient time at the end of the summer to re-attempt this analysis.

One purpose of running WaveWatch III was to contribute wave information to the climatological data server at the Fleet Numerical Meteorology and Oceanography Detachment in Asheville, NC. They are an office within the Department of Defense (DoD) that handles Navy, Marine Corps, and other DoD agency climatological requirements. This work is also related to the Slope to Shelf Energetics and Exchange Dynamics (SEED) Project and Dynamics of the Adriatic in Real Time (DART), a North Atlantic Treaty Organization (NATO) Undersea Research Center, Italy, project.

ROMULUS and KRAKEN were used to run the models, while VINCENT was used to store both input data and output from the models. A daily run used about 2 Gigabytes (GB) of space on VINCENT to store the input data required to run the models. Afterwards, the runs for the Adriatic Sea and Gulf of Mexico produced about 1 GB and 840 Megabytes (MB) of data, respectively.

Allison had to spend some time preparing scripts, which helped to streamline the process of submitting model runs to the batch queueing system. Besides presenting numerical results from running ocean models (a nice example is given in Figure 3, from a COAMPS visualization of the Gulf of Mexico during 18-20 July 2005–Hurricane Emily is clearly visible), she was also able to perform some performance analysis, showing good scalability of the COAMPS code between 8 and 32 Central Processing Units (CPUs) on both KRAKEN and ROMULUS.

Allison would like to thank her mentor Jim Dykes, the PET Component One staff, and the NAVO MSRC.

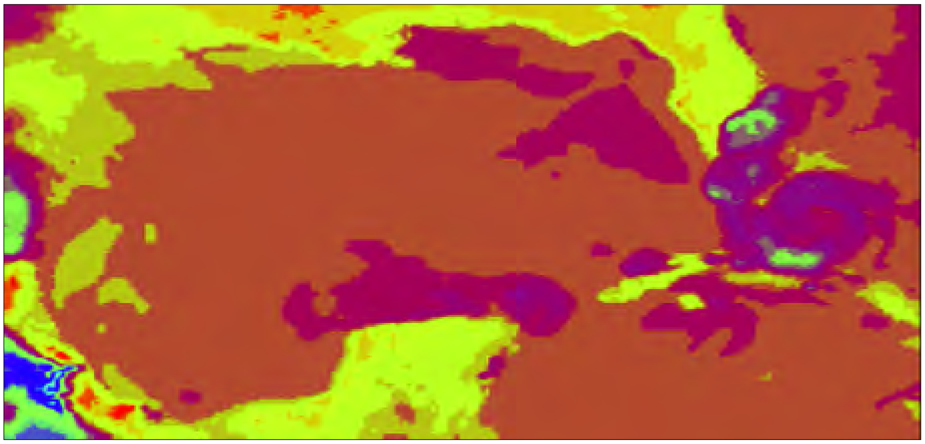


Figure 3. Numerical results from running ocean models illustrated in a COAMPS visualization of the Gulf of Mexico during 18-20 July 2005.

> center-specific queues are acceptable to allow sites an opportunity to offer customized support for specialized work loads and access to unique hardware resources. The project will seek to retain existing queue names/structures for a period of time in order to minimize impact to user productivity.
>
> A minimum set of common queue names will allow users who spend their processing time primarily within the standard HPCMP workload classifications to more easily move jobs and scripts among centers. This will improve user productivity by requiring less time and effort to support many different scripts and lessen the need to learn different queue environments at every center.
>
> At NAVO MSRC, the common queue names baseline is implemented for the FY06 computing year. The common queue names on each is now complete and includes the following queue names on each allocated computational server for FY2006 are listed below:
>
> - urgent
> - debug
> - challenge
> - high
> - Standard
> - background[3]

Common Login Shells Project

The common login shells documentation states:

> Lack of login shell consistency across Centers limits DoD user interactive/batch work environments, causes researcher frustration during interactive sessions, forces users accessing multiple sites to develop redundant shell-based programs and tool versions, and leads to reduced productivity for the DoD user community. Therefore, users need a common set of login shells at all Centers and a common policy to follow for login shell selection and subsequent changes.
>
> The Common Login Shells Project will provide DoD users a common set of supported login shells available to users on all allocated systems at the six Centers. A consistent, common policy for handling login shell requests on new user accounts and login shell change requests for existing accounts will also be provided.[4]

Kerberos Ticket Life Documentation Policy Project

The ticket life consistency documentation states:

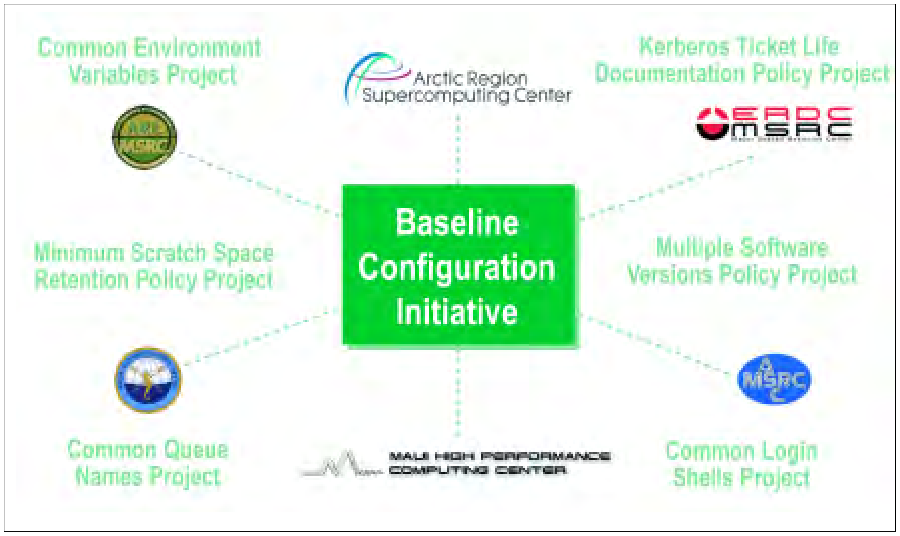
> DoD users have indicated some confusion about ticket lifetimes for Kerberos commands and authentication. This may be a side affect of the rapid changes that were deployed following increased internet security activities. Each Center should review all user documentation, including references in web pages, to ensure consistency with the policy as described at the Web-based Kirby, the HPCMP Kerberos and SecurID Information Center (https://kirby.hpcmp.hpc.mil).
>
> Each of the six allocated Centers that make up the BCI will incorporate a link from their respective Web home pages to the Kirby site in order to provide users with a consistent information source for DoD Kerberos, and in particular, the DoD Kerberos Ticket Life matrix.[5]

Multiple Software Versions Policy Project

The Multiple Software Version Policy documentation states:

> The Multiple Software Versions Policy Project will develop a common software work

Continued Next Page...


Common Environment Variables Project
Arctic Region Supercomputing Center
Kerberos Ticket Life Documentation Policy Project
ARL MSRC
ERDC MSRC
Baseline Configuration Initiative
Minimum Scratch Space Retention Policy Project
Multiple Software Versions Policy Project
ASC MSRC
Common Queue Names Project
MAUI HIGH PERFORMANCE COMPUTING CENTER
Common Login Shells Project

environment for HPCMO Center users. The common software work environment will enable users to more easily move between Centers, and use the resources of all Centers more easily. This should enhance overall user productivity and limit user frustration at having to find the current software version at a particular center or translate their application to another version of software. Reduction of site-specific limitations will allow for users to match their applications to HPCMP site resources without artificial obstacles.

The Centers currently maintain multiple versions of software stems because it is sometimes difficult to transition to newer versions. Part of the difficulty is that users often have to re-validate existing codes when system software is updated. This can be a problem, especially if there are differences in the results generated by the old and new software versions. The ability to maintain an archive of outdated software versions at multiple Centers will allow for quick evaluation and problem identification when newer versions of software are installed.

The Project will also develop a policy for the maintenance of and archive of outdated versions of software across all Centers. The final policy will include: a justification for a program-wide policy; scoping of the policy (software/number of versions/centers); an understanding of current practices; a discussion of issues concerning current practices and any program-wide implementation; a feasibility assessment of implementing a new policy (including impact on site infrastructures); a schedule for implementation; and an analysis of costs that may be incurred.[6]

Conclusion

The six initial BCI projects highlighted in this article are only the beginning.

Once these initial projects are implemented, the BCT will identify additional opportunities for the improvement of computing environment commonality at the Centers participating in the initiative in FY06. These new projects will continue to improve DoD user productivity and the overall user experience.

The NAVO MSRC, as part of its ongoing efforts to provide the best user support possible, has participated in this initial effort and will continue to participate in future projects. As additional baseline items are established, the NAVO MSRC will implement these items at the earliest possible opportunity.

References

1. Baseline Configuration Team, “Common Set of Environment Variables,” Baseline Configuration Project FY05-04, 2005
2. Baseline Configuration Team, “Minimum Scratch Space Retention Time Policy,” Baseline Configuration Project FY05-04, 2005
3. Baseline Configuration Team, “Common Queue Names,” Baseline Configuration Project FY05-04, 2005
4. Baseline Configuration Team, “Baseline Set of Login Shells and User Selection Policy,” Baseline Configuration Project FY05-04, 2005
5. Baseline Configuration Team, “Ticket Life Consistency,” Baseline Configuration Project FY05-04, 2005
6. Baseline Configuration Team, “Multiple Software Version Policy,” Baseline Configuration Project FY05-04, 2005

Navigator Tools and Tips

Looking for the best way to archive large amounts of data to the Archive Servers

Sheila Carbonette, NAVO MSRC User Support

So, you have a large Message Passing Interface (MPI) job running on the IBM P4+ system, KRAKEN, that is going to generate over 1000 output files in the /scr filesystem.

What are you going to do with the files? You can't leave the files in the /scr filesystem because the scrubber will eventually delete them. And, this may occur a lot quicker than you might expect.

Hmmm, you wonder, what is the most efficient way to transfer all of these files?

Should you create a script and transfer each file one at a time? This will work, but it is not the most efficient way.

An alternate approach is to create archive files using the tar command and then transferring the archive files. Depending on the size of the original output files, you may want to create several archive files to transfer. This will not only be quicker in archiving the files, but also in retrieving the files for a future job run. An example script follows:

```
#!/bin/ksh

cd /scr/shecar/run1

# CREATE THE ARCHIVE FILE USING TAR

/usr/bin/tar cvof archive_file.tar files_from_job1*

let RC=0

# STAGE A FILE TO JULES USING RCP

/usr/bin/rcp archive_file.tar jules:/u/home/
  shecar/data/archive/

let RC=$?

if (($RC != 0))

then

  /bin/echo "$0-ERROR: RCPing files TO Jules;
    RC=$RC"

else

  /bin/echo "File Staging TO Jules Completed
    SUCCESSFULLY!"

fi

exit $RC
```

Another scenario: You have a job running on KRAKEN that produces several large data files. Each of these files may be over 100 GB in size. What is the best way to transfer these files? The first step should be to compress the files using the "gzip" command. This may add CPU time to the job, but it will save on file transfer time. The second step should be to copy the files using the unkerberized "rcp" command. An example script follows:

```
#!/bin/ksh

cd /scr/shecar/run2

# COMPRESS the data file

/usr/bin/gzip data_file

let RC=0

# STAGE A FILE TO JULES USING RCP

/usr/bin/rcp data_file.zip jules:/u/home/
  shecar/data/archive/

let RC=$?

if (($RC != 0))

then

  /bin/echo "$0-ERROR: RCPing file TO Jules;
    RC=$RC"

else

  /bin/echo "File Staging TO Jules Completed
    SUCCESSFULLY!"

fi

exit $RC
```

The above scripts can either run interactively or under the LoadLeveler queueing system. Depending on the size or the number of files, the script may not finish within the 30-minute interactive time limit. If this is the case, the following script can be used to submit to the "transfer" queue on any of the IBM systems:

Continued Next Page...

```
#@ shell = /bin/ksh

#@ job_type = serial

#@ node_usage = shared

#@ output = transfer.$(jobid)

#@ error = transfer.$(jobid)

#@ job_name = transfer

#@ wall_clock_limit = 4:00:00

#@ notification = never

#@ account_no = NAVOSLMA

#@ class = transfer

#@ resources = ConsumableCpus(1)
  ConsumableMemory(512)

#@ queue

#

cd /scr/shecar/run2
```

```
# COMPRESS the data file

/usr/bin/gzip data_file

let RC=0

# STAGE A FILE TO JULES USING RCP

/usr/bin/rcp data_file.zip jules:/u/home/
  shecar/data/archive/

let RC=$?

if (($RC != 0))

then

  /bin/echo "$0-ERROR: RCPing file TO Jules;
    RC=$RC"

else

  /bin/echo "File Staging TO Jules Completed
    SUCCESSFULLY!"

fi

exit $RC
```

Coming Events

HotNets-IV
Fourth Workshop on
Hot Topics in Networks
November 14 - 15
College Park, MD
www.acm.org/sigs/sigcomm/HotNets-IV/

PDCS 2005
Parallel and Distributed
Computing and Systems
November 14 - 16
Phoenix, AZ
www.iasted.org/conferences/2005/phoenix/c466.htm

CNIS 2005
Communication, Network
and Information Security
November 14 - 16
Phoenix, AZ
www.iasted.org/conferences/2005/phoenix/cnis.htm

WISE 2005
Web Information
Systems Engineering
November 20 - 22
New York, NY
www.cse.unsw.edu.au/~jas/wise05/

ICDM '05
International Conference
on Data Mining
November 27 - 30
Houston, TX
www.cacs.louisiana.edu/~icdm05/

I-SPAN 2005
Int'l Symposium on
Parallel Architectures,
Algorithms, and Networks
December 7 - 9
Las Vegas, NV
http://sigact.acm.org/ispan05

ACSAC 2005
Computer Security
Applications Conference
December 5 - 9
Tuscon, AZ
www.acsac.org

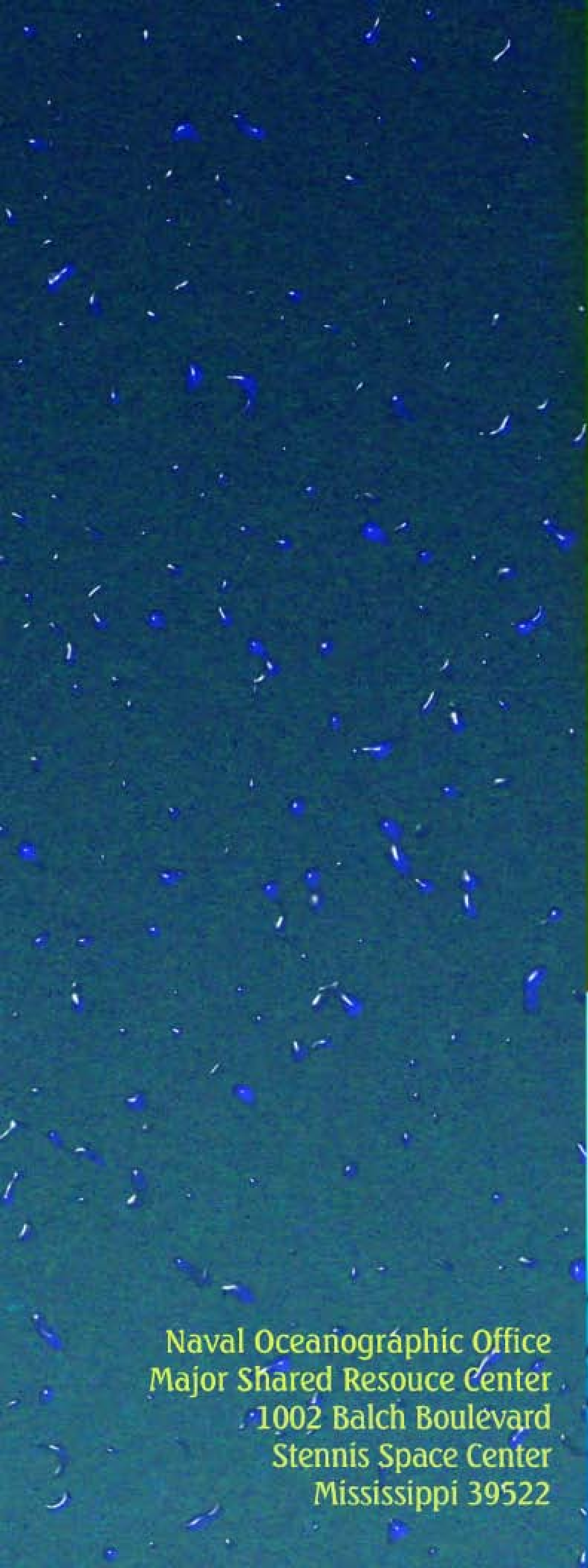
CIS^{2}E 05
Int'l Joint Conferences
on Computer, Information
and Systems Sciences,
and Engineering
December 10 - 20
Online Conference
http://cisse2005.org

CollaborateCom 2005
Collaborative Networking,
Applications and Worksharing
December 19 - 21
San Jose, CA
www.collaboratecom.org

SD WEST 2006
Software Development
West Conference and Expo
March 13 - 17
Santa Clara, CA
www.sdexpo.com

SAC 2006
21st Annual ACM Symposium
on Applied Computing
April 23 - 27
Dijon, France
www.acm.org/conferences/sac/sac2006/

HPC 2006
High Performance
Computing Symposium
April 2 - 6
Huntsville, AL
www.caip.rutgers.edu/hpc2006/



Naval Oceanographic Office
Major Shared Resouce Center
1002 Balch Boulevard
Stennis Space Center
Mississippi 39522