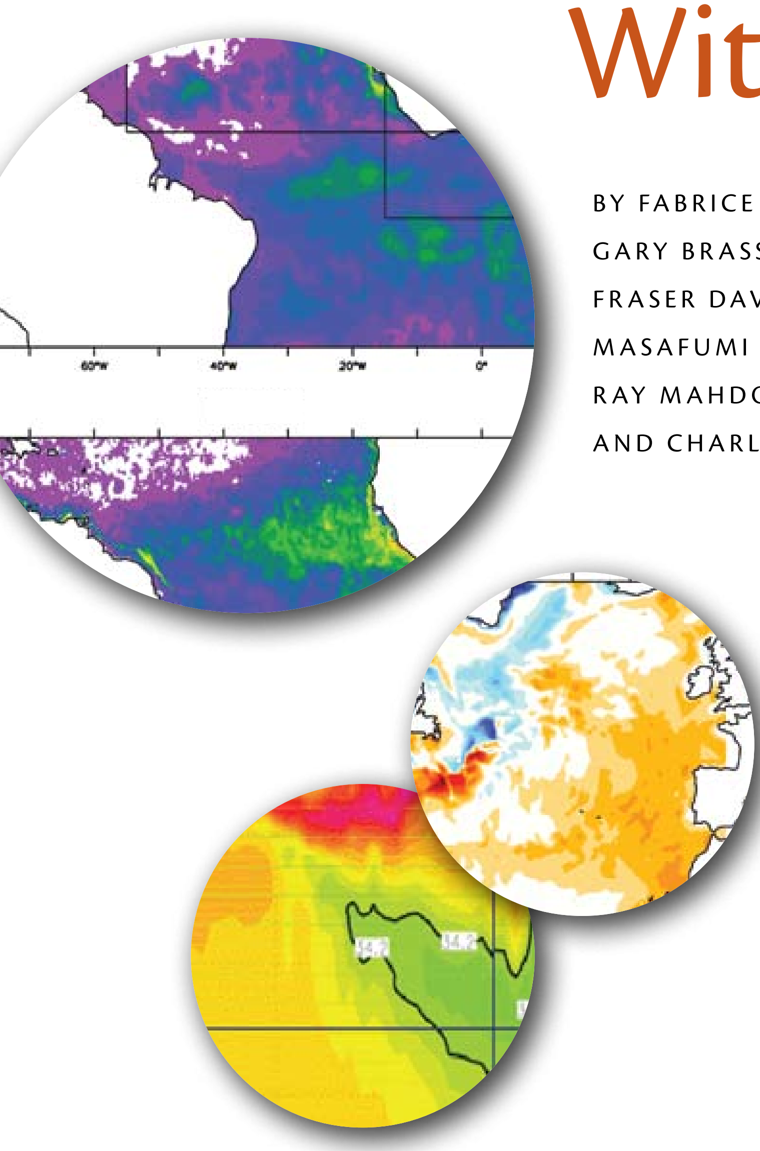
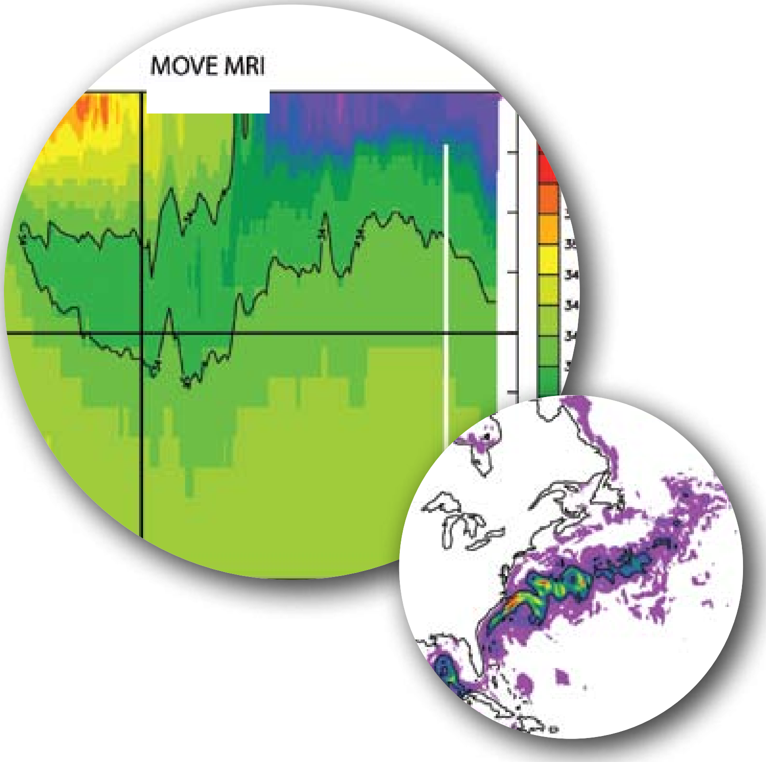
Validation and Intercomparison Studies Within GODAE

BY FABRICE HERNANDEZ, LAURENT BERTINO, GARY BRASSINGTON, ERIC CHASSIGNET, JAMES CUMMINGS, FRASER DAVIDSON, MARIE DRÉVILLON, GILLES GARRIC, MASAFUMI KAMACHI, JEAN-MICHEL LELLOUCHE, RAY MAHDON, MATTHEW J. MARTIN, ANDRY RATSIMANDRESY, AND CHARLY REGNIER





ABSTRACT. During the Global Ocean Data Assimilation Experiment, seven international operational centers participated in a dedicated modeling system intercomparison exercise from February to April 2008. The objectives were: (1) to show GODAE global-ocean and basin-scale forecasting systems of different countries in routine interaction and continuous operation, (2) to assess the quality and perform scientific validation of the ocean analyses and the forecasting performance of each system, and (3) to learn from this exercise in order to increase interoperability and collaboration in real time. The validation methodology has steadily improved through several validation experiments and projects performed within the operational oceanography community. It relies on common approaches and standardization of outputs, with a set of diagnostics based on fully detailed metrics that characterize its strengths and weaknesses, but also provide error levels for ocean estimates. The ocean forecasting systems provide daily fields of mesoscale water mass distribution and ocean circulation, with an option for sea-ice variations. We present a subset of the intercomparisons performed over different areas, showing general ocean circulation in agreement with known patterns. We also present some accuracy assessments through comparison with observed data.

INTRODUCTION

In the MERSEA (Marine Environment and Security for the European Area) Strand1 European Union (EU) framework, a first attempt to intercompare eddy-permitting, basin-scale ocean data assimilating systems was conducted. Hindcasts originating from the different systems were intercompared using climatology and historical high-quality ocean data sets (i.e., World Ocean Circulation Experiment [WOCE] sections) as a reference (Crosnier et al., 2006). In parallel, ocean forecasting systems were developed and improved in several countries as part of, and associated with, the Global Ocean Data Assimilation Experiment (GODAE) community effort (for a detailed description of GODAE, see Bell et al., this issue). While operating continuously, these systems were monitored and compared to a predetermined quality standard. Outputs were either dedicated reanalyses or long simulations, or operational hindcasts, nowcasts, and forecasts. Feedback from these scientific evaluations were used to improve the systems' components—ocean model parameterization, forcing, or assimilation methodology.

There are more constraints on assessment of ocean analysis and forecasting systems than on validations normally performed for academic projects. The ocean assessments must be performed in real time with practical operational constraints such as computer resources, storage capacity, and availability of reference values (e.g., independent ocean data). Because outputs rely on continuously changing information (e.g., availability of input such as atmospheric forcing fields, in real time), monitoring and assessment of the validation procedures are mandatory.

Outputs from operational systems are also used for commercial and other applications (e.g., oil spills, water-quality assessments, marine security). Thus, the assessment methodology must reflect user requirements. Different applications require different levels of accuracy. For instance, an ocean model that may be satisfactory for general ocean study may not provide sufficient information for search and rescue activities.

The assessment of data assimilation systems is more focused on accuracy than on overall quality. In other words, where a certain level of quality is sought in pure modeling research (e.g., Is there deep convection? Has Labrador Sea water formed? Is there a Gulf Stream overshoot, an acceptable meridional heat transport, and meridional overturning circulation?), assimilation experiments are tested on "realistic representation" where reference data are used to directly quantify error levels. A comprehensive error budget is also required for proper assessment of data assimilation results. Assimilation schemes are more or less guided by background and observation errors, and the most sophisticated schemes provide robust forecast error estimates (Brasseur, 2006). It is, then, necessary to verify model error assumptions against dedicated error validation procedures.

In this context, using different model configurations and data assimilation methods, operational oceanography teams have tried to develop their tools for assessing the quality of outputs, and they have started to provide "error bars" to users. Thanks to GODAE, these initiatives could be shared at the international level (Smith, 2006). A special intercomparison exercise was performed at the beginning of 2008 as a GODAE project. The objectives were to: (1) demonstrate GODAE systems in operation, (2) share expertise and design validation tools and metrics endorsed by all GODAE operational centers, and (3) evaluate the overall scientific quality of the different GODAE operational systems. During the exercise, most operational centers worldwide delivered daily ocean products. These included BLUElink> (Australia), Hybrid Coordinate Ocean Model (HYCOM; United States), Meteorological Research Institute (MRI, which developed the multivariate ocean variational estimation, (MOVE/MRI.COM; Japan), Mercator Océan (France), Forecasting Ocean Assimilation Model (FOAM; United Kingdom), Canada-Newfoundland Operational Oceanography Forecast System (C-NOOFS; Canada), and Towards an Operational Prediction System for the North Atlantic European Coastal Zones (TOPAZ; Norway). The next section outlines the assessment methodology. It is followed by a section summarizing the assessment and intercomparisons performed during GODAE and, particularly, presents a subset of results from the special intercomparison project.

METHODOLOGY OF VALIDATION AND INTERCOMPARISON

The validation methodology performed during the MERSEA Strand1 project (2003–2004), and endorsed at the GODAE level by non-European data assimilation teams, was enhanced during the EU MERSEA Integrated Project (2004–2008). Aspects of the validation methodology specified during the course of this project were: (1) Perform the validation continuously, and thereby stimulate data processing and archiving centers to provide observations in real time. (2) Apply diagnostics that offer robust scientific evaluation of each system, and select the most suitable diagnostics among those applied in research mode. (3) Evaluate both operational system performance and product quality, taking user requirements into account (usually from short-term to seasonal time-scale applications); (4) Encourage consistency of assessment among the different forecasting centers, applying similar diagnostics to the different systems, thus strengthening the overall assessment management activity through central team expertise. (5) Take advantage of this consistency to allow intercomparison of the operational systems, and thus design and implement technical architecture that allows robust exchanges, interconnections, and interoperability among these systems.

The two last points clearly promote the use of methods and technologies that encourage exchanges and intercomparisons among the different forecasting centers. They also provide impetus for consistently implementing interoperable activities such as ensemble forecasting.

From 2003 to 2008, four test periods allowed different teams to improve the validation and intercomparison methodology. The first intercomparison in the North Atlantic Ocean and Mediterranean Sea during the MERSEA Strand1 project involved five "eddy-resolving" systems (Crosnier et al., 2006; Crosnier and Le Provost, 2007). Conclusions from these test periods permitted: (1) evaluation of system faults, (2) improvements to the different systems, and (3) refinements to the assessment methodology.

Validation procedures developed during MERSEA Strand1 have benefitted operational center developments. Validation procedures were used to verify improvements during upgrades of their systems not only by validating a new system against the former one but also by using validation results from the MERSEA Strand1 intercomparison with other forecasting centers to quantify the new system's overall improvements.

The GODAE intercomparison project carried out in 2008 involved the seven

---

Fabrice Hernandez *(fabrice.hernandez@mercator-ocean.fr) is Senior Scientist, Mercator Océan, Ramonville-Saint-Agne, France.* **Laurent Bertino** *is Research Scientist, Nansen Environment and Remote Sensing Center, Bergen, Norway.* **Gary Brassington** *is Research Scientist, Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, Australia.* **Eric Chassignet** *is Professor and Director, Center for Ocean-Atmospheric Prediction Studies, Florida State University, Tallahassee, Florida, USA.* **James Cummings** *is Oceanographer, Naval Research Laboratory, Monterey, California, USA.* **Fraser Davidson** *is Research Scientist, Department of Fisheries and Oceans, St. John's, Newfoundland, Canada.* **Marie Drévillon** *is Research Scientist, Mercator Océan, Ramonville-Saint-Agne, France.* **Gilles Garric** *is Research Scientist, Mercator Océan, Ramonville-Saint-Agne, France.* **Masafumi Kamachi** *is Head, Second Laboratory, Department of Physical Oceanography, Meteorological Institute, Tsukuba, Japan.* **Jean-Michel Lellouche** *is Research Scientist, Mercator Océan, Ramonville-Saint-Agne, France.* **Ray Mahdon** *is Ocean Modelling Scientist, Met Office, Exeter, UK.* **Matthew J. Martin** *is Ocean Data Assimilation Scientist, Met Office, Exeter, UK.* **Andry Ratsimandresy** *is Research Scientist, Department of Fisheries and Oceans, St. John's, Newfoundland, Canada.* **Charly Regnier** *is Scientific Engineer, Mercator Océan, Ramonville-Saint-Agne, France.*

global and basin-scale "eddy-permitting" to "eddy-resolving" systems listed earlier. Its focus was on scientific assessment, and it provided a way to improve the assessment methodology for the world ocean, and also offered a first opportunity to intercompare recent systems such as BLUElink>, MOVE/MRI.COM, and C-NOOFS.

The assessment methodology proposed for the GODAE intercomparison project was a direct outcome of previous validation work. Following Crosnier and Le Provost (2007), the assessment had two aspects. First, *the philosophy*: apply a set of basic principles for assessing the quality of MERSEA/GODAE products and systems through a collaborative partnership. Second, *the methodology*: use a set of tools for computing diagnostics and have a set of reference standards for assessing the quality of the products. Both tools and standards must be shareable and usable among the different MERSEA/GODAE members and systems. Both tools and standards should be subject to upgrades and improvements.

The following set of principles was adopted for the assessment:

- Consistency: verifying that system outputs are consistent with current knowledge of ocean circulation and climatologies
- Quality (or accuracy of the hindcast/nowcast): quantifying the differences between the systems' "best results" (analysis) and sea truth, as estimated from observations, preferably using independent observations (not assimilated)
- Performance (or accuracy of the forecast): quantifying the short-term forecast capacity of each system (i.e., answering the question: Does the forecasting system perform better than persistence and better than climatology?)

> "THE VALIDATION METHODOLOGY HAS STEADILY IMPROVED THROUGH SEVERAL VALIDATION EXPERIMENTS AND PROJECTS PERFORMED WITHIN THE OPERATIONAL OCEANOGRAPHY COMMUNITY."

A fourth principle was proposed—to verify and take into account the interest and relevance of system outputs for customers, and catch intermediate- or end-user feedbacks:

- Benefit: end-user assessment of the quality level that must be reached before the products are useful for an application

This validation methodology was built using "metrics"— mathematical tools that compute scalar measures from system outputs and compare them to "references" (e.g., climatology, observations). The metrics provide equivalent quantities extracted from the different systems for the same geographic locations. Applied to different forecasting systems, they provide homogeneous and consistent sets of quantities that can be compared without reference to the specific configuration of each system (e.g., horizontal resolution, vertical discretization).

"Shareability" was the second important aspect of the validation methodology. It allowed each forecasting center to perform intercomparison and validation independently, using results from other centers. Metrics, computed in a standardized way, were stored by each center in order to be available for others. The NetCDF file format using the COARDS-CF (Cooperative Ocean/Atmosphere Research Data Service-Climate and Forecast) convention was chosen, allowing time aggregation, easy and flexible manipulation, and self-consistent metadata representation. Distribution relied on Internet communication protocols, basically through FTP. However, more user-friendly communication technologies based on Open-source Project for a Network Data Access Protocol (OPeNDAP) servers that can be visualized through a Live Access Server (LAS), using Dynamic Quick View portals or with similar clients, have now been widely adopted (see Blower et al., this issue). In practice, these technologies allow each forecasting center to compute a considerable amount of diagnostics stored on the local servers of other centers. The total set of validation data does not need to be centralized, which would require large storage capacities. Instead, for a given diagnostic, one can specifically gather the information spread it across the different centers, as shown during the MERSEA Strand 1 project.

Metrics were defined in four types, or "classes," described below. The consistency and quality of each system could be deduced, or intercompared, from Class 1, 2, and 3 metrics. A system's performance could be addressed using Class 4 metrics. The "benefit" could also be addressed using a set of Class 1, 2, 3, and 4 metrics. However, new "user-oriented" metrics might need to be defined to fully address the latter.

Class 1 Metrics

Class 1 metrics aim to provide a general overview of ocean and sea-ice dynamics from the different systems. Ocean and sea-ice model variables corresponding to different horizontal and vertical native grids are interpolated into a common set of horizontal and vertical grids over different regions of the world ocean. Horizontal resolution is selected for an eddy-permitting description of the ocean, whatever the original grid resolution of the different systems (Table 1). Vertical resolution uses the following principal depths: 0, 30, 50, 100, 200, 400, 700, 1000, 1500, 2000, 2500, and 3000 m. These depths do not aim to fully monitor ocean water mass variability, but are a compromise on the storage capacity needed for the world ocean overview.

Class 1 diagnostics present two- and three-dimensional fields. The two-dimensional fields are sea surface height (SSH), wind stress, solar and net heat fluxes, total freshwater fluxes, and mixed-layer depth (MLD); the three-dimensional fields are temperature, salinity, and currents. The diagnostics also present two-dimensional sea-ice variables for mid- and high-latitude areas: concentration, ice and snow thickness, and velocity. Class 1 metrics (i.e., daily means) can be used as "instantaneous" estimates of ocean mesoscale circulation for direct comparison to

Table 1. Description of regional NetCDF Class 1 files. Names, limits and gridding, type of geographical projections, and specific features for each Class 1 area.

	Name	Horizontal resolution		Type of projection	Geographical limits	Specific points
North Atlantic	NAT	1/6°	787 x 597	Mercator	0°–70°N 100°W–31°E	Baltic and Caribbean seas, European shelves, Gulf of Mexico. Sea ice variables.
South Atlantic	SAT	1/6°	601 x 453	Mercator	60°S–0°S 70°W–30°E	Drake Passage, Agulhas Current
Tropical Atlantic	TAT	1/4°	421 x 163	Mercator	20°S–20°N 90°W–15°E	Caribbean seas, Gulf of Guinea
North Pacific	NPA	1/6°	1099 x 518	Mercator	0°–65°N 100°E–77°W	Japan, China seas, Panama. Sea ice variables.
South Pacific	SPA	1/6°	1141 x 453	Mercator	60°S–0° 100°E–70°W	Circum-Australia area
Tropical Pacific	TPA	1/4°	801 x 163	Mercator	20°S–20°N 90°E–70°W	Indonesian seas and straits
Indian Ocean	IND	1/6°	601 x 458	Mercator	20°E–120°E 40°S–31°N	Mozambique Channel, Red and Arabian seas, Bay of Bengal
Arctic Ocean	ARC	12.5 km	609 x 881	Stereo Polar	180°W–180°E 34°N < λ < 90°N	North Atlantic Subpolar Gyre, Baltic, Bering, and Okhotsk seas. Sea ice variables.
Southern Ocean	ACC	1/4°	1441 x 937	Mercator	89°S–35°S 0°–360°E	Antarctic Circumpolar Current system, Ross and Weddell ice caps. Sea ice variables.
Mediterranean and Black seas	MED	1/8°	385 x 187	Mercator	6°E–42°W 30N–48°N	Dedicated resolution for the Mediterranean and Black seas.
Global	GLO	1/2°	721 x 359	Regular	180°W–180°E 89°S–90°N	Overview of the world ocean. Sea ice variables.

observed quantities, for example, maps of satellite sea surface temperature (SST), satellite altimetry SSH, and dynamic height from synoptic hydrographic data sets. When time-averaged, Class 1 metrics allow "consistency assessment" (i.e., comparison to climatologies or ocean circulation patterns described in the literature).

From Class 1 files, in a given area, one can develop time series of any variable (e.g., temperature, MLD, wind stress) or derived quantities (e.g., eddy kinetic energy), study ocean variability at different depths for different variables (e.g., Hovmuller diagrams of SSH, SST, and Empirical Orthogonal Mode decompositions), and compare these data to equivalent observational data sets (e.g., altimetry sea level maps, SST maps). Spectral analyses can also be performed.

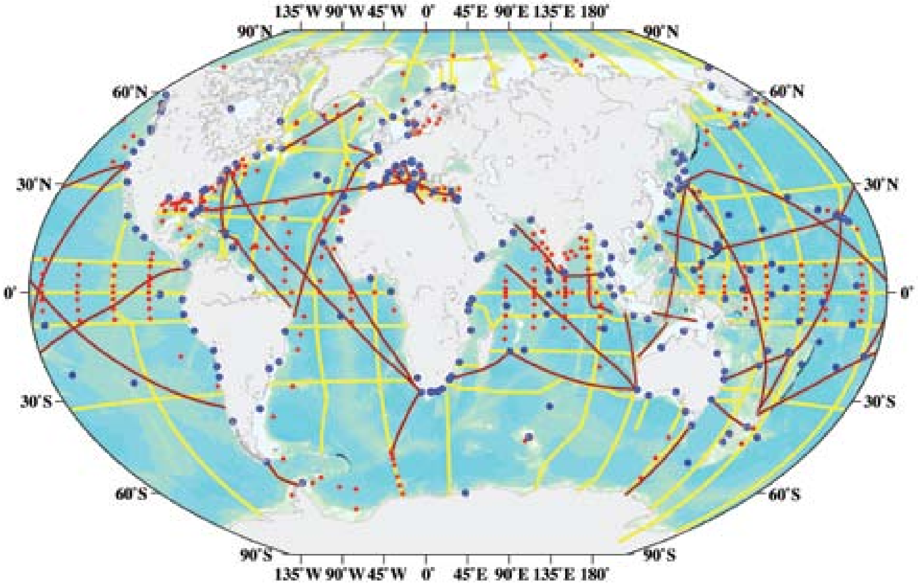


Figure 1. Locations of Class 2 metrics. Yellow = straight sections. Brown = expendable bathythermograph (XBT) sections. Blue = tide gauges. Red = other moorings.

Class 2 Metrics

Like Class 1, Class 2 metrics were designed to monitor operational system outputs, but in a complementary way. Wherever higher horizontal and vertical resolutions are required, Class 2 metrics provide virtual moorings or sections of the model domain. These specifically chosen sections and moorings represent a reduced amount of stored data and reduce the need to store full three-dimensional fields of the full system domains at high resolution.

Class 2 metrics essentially gather temperature, salinity, currents, and SSH along chosen section tracks and moorings. Sections were sampled horizontally every 10–15 km with a vertical resolution chosen to match standard levels as in the Levitus World Ocean Atlas 2005 (WOA05; Locarnini et al., 2006) or the Generalized Digital Environment Model (GDEM3.0) (Teague et al., 1990) climatologies—78 levels starting with 2-m resolution near the surface, increasing to 250-m resolution in bottom layers.

These sections and moorings were defined to match well-observed regions in order to allow validation, for example, using data from tide gauges, tropical moorings, and current meter moorings that are transmitting in real time, or are serviced regularly. Sections match the most frequently visited expendable bathythermograph (XBT) Ship of Opportunity Program (SOOP) lines during 2000–2005 as well as the main WOCE and CLIVAR (Climate Variability and Predictability program) repeat sections (Figure 1). Because Class 2 metrics are directly computed online during the model runs for some systems, their total number is a compromise between computer-time resources and overall description of the world ocean.

Class 2 metrics provide finer knowledge of ocean dynamics and water properties for comparison with in situ or remote-sensing observations. Consistency assessments compare time-averaged Class 2 sections to climatologies, while XBT sections can be compared in near-real time to daily Class sections, allowing the accuracy of daily products to be inferred. Sea-level time series can be validated against tide gauge data. Statistics can also be produced and compared with satellite and other time-series data.

Class 3 Metrics

Class 3 metrics are derived physical quantities computed using the model variables on the model native grids at each time step that, therefore, cannot be derived from Class 1 or Class 2 metrics. Typical Class 3 diagnostics are integrated quantities such as daily volume transport through chosen sections, which may coincide geographically with Class 2 sections. Typical Class 3 metrics are: (1) volume transports across chosen sections and total transports as well as, sometimes, transports split into potential temperature, density, salinity, or depth classes; (2) heat transport across sections or meridional heat transport (global, basin-wide) computed similarly to volume transport; and (3) the Overturning Streamfunction (global, basin-wide) as a function of latitude and depth, potential temperature, or potential density.

Class 4 Metrics

Class 1, 2, and 3 metrics can be applied to any field produced by the forecasting system (hindcasts, nowcasts, or forecasts). Class 4 metrics aim to measure the performance of the forecasting system, its capability to describe the ocean (in hindcast mode), as well as its forecasting skill (analysis and forecast mode). Class 4 metrics are natural (model-observation) joint products that may be output from the assimilation systems.

Class 4 metrics are limited here to "observational space," with observations chosen (preferably independent of those used during the assimilation procedure) and compared at all stages of the hindcast-analysis-forecast cycle. Class 4 metrics provide a series of statistics for differences between data and model values for several fields produced by the operational systems, along with observational comparisons to climatology and persistence forecasts. Thus, one can quantify differences in "forecasting skill" for any variable; for example, a given SSH observation from a tide gauge can be compared to model estimates on that day, or any forecast performed from previous assimilation cycles, or against climatology, or against observation persistence at any time lead. All of these values can be brought together to provide error statistics.

Any data can be used in Class 4 metrics, provided equivalent information can be computed from the model variables. MERSEA Strand 1 compared sea level assessments to satellite altimetry. The MERSEA Integrated Project compared in situ temperature and salinity profiles, sea level from tide gauges, and also satellite sea ice concentrations using Class 4-like metrics. Figure 2 provides an example of sea-ice

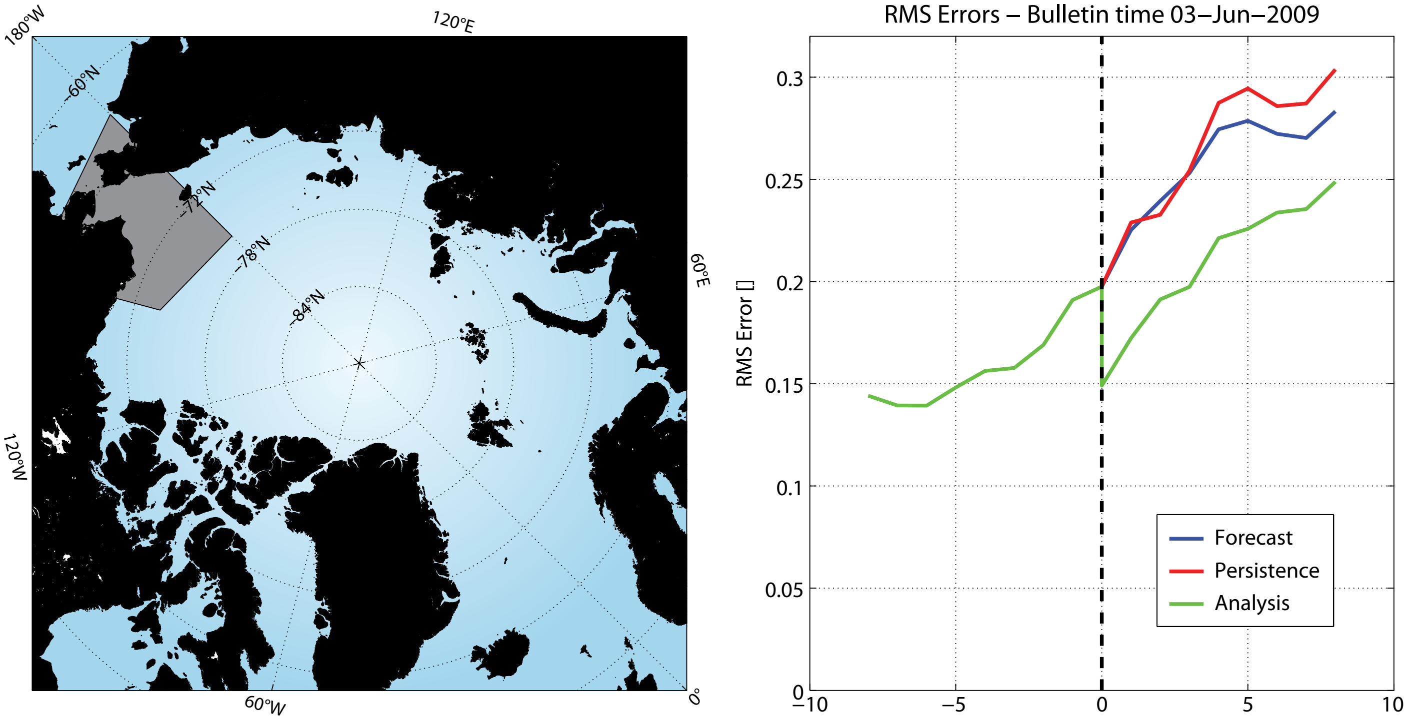


Figure 2. Sea-ice performance diagnostic in the Arctic Ocean. August 2006 to February 2007 root mean square daily differences of sea-ice concentration between Special Sensor Microwave Imager (SSM/I) observed products and the TOPAZ forecasting system are computed for different outputs—analysis and forecasts (1 to 15 days ahead). RMS differences are computed in geographical boxes (left panel, the Bering Strait box), then the averaged performance from hindcast (5 days back) to forecasting 15 days ahead are plotted for analysis, persistence, and forecast (right panel).

concentration performance diagnostics computed with the TOPAZ operational system. The sea-ice concentration error, as given by the root mean square (RMS) difference between satellite data and model estimates in a given area, allows quantification of the overall behavior of the system, including the quality of hindcasts, absolute error of forecasts, and forecasting skill relative to persistence.

Again, once model values and equivalent observation quantities are obtained, it is possible to diagnose forecasting system performance on any derived quantity. For instance, from modeled and observed temperature and salinity at any depth, one can plot the θ-S diagrams for observed data and climatology, as well as for model three-day forecasts, associated persistence, or analyses. These plots allow one to infer which water masses are well represented in real-time estimates, in predictions, or using climatology or persistence approaches.

Note that performance and forecast skill can also be inferred in the "model space" by computing statistics on misfits, residuals, and forecast minus analysis. Some examples are given in Cummings et al. (this issue).

THE GODAE SPECIAL INTERCOMPARISON PROJECT

The seven ocean forecasting systems that participated in the 2008 intercomparison were all "eddy-permitting" or "eddy-resolving," providing daily estimates and forecasts in real time over the global ocean, or regionally. The three-month period February–April 2008 was chosen because it was the first possible period during which all seven ocean forecasting centers could provide daily averaged hindcast estimates for the Class 1 metrics. The three-month intercomparison period was agreed upon to allow a first assessment from daily to monthly temporal scales. Each system provided estimates over the different regions of their model domains. For each region in Table 1, at least three systems can be intercompared for some Class 1 variables and derived quantities. The assessment of consistency and accuracy relied on real-time in situ or satellite data gathered at the same time in the framework of GODAE.

In this issue, Dombrowsky et al. describe the BLUElink>, FOAM, Mercator Océan, and HYCOM global systems; Hurlburt et al. describe the MOVE/MRI.COM system, the Mercator Océan high-resolution North Atlantic system, the TOPAZ Arctic system, and the C-NOOFS Northwest Atlantic system; and Cummings et al. describe the data assimilation methodology applied by each system. Table 2 summarizes system characteristics. These diverse systems use four types of ocean models, can be global or regional, have different vertical discretizations, are eddy-permitting to eddy-resolving, are coupled or not with sea-ice models, employ different air-sea flux representations, and use different assimilation techniques.

The North Atlantic basin allows comparison of three global systems (HYCOM, FOAM, and Mercator Océan), and three regional systems (TOPAZ, C-NOOFS, and the high-resolution 1/12° Mercator system). The two Mercator Océan systems differ only in their horizontal resolution (1/4° versus 1/12°), permitting inference of the impact of higher resolution. All other systems are 1/4° eddy-permitting, except the eddy-resolving 1/12° HYCOM global system.

A water mass analysis was computed for all systems as part of the consistency assessment. Figure 3 shows comparisons of temperature with WOA05 Levitus Climatology (Locarnini et al., 2006) at 30-m depth in February 2008. A general difference pattern shows Labrador Current and East Greenland Current waters colder by 2°C, waters warmer by 0.5°C at the eastern boundary, anomalies of 1–2°C in the North Sea and southeast of Iceland, and a pronounced difference (larger than 2°C) along the African tropical coasts. These results represent real 2008 differences against the climatology. The cold signature in the Labrador Current is associated in most cases with a warmer signature in the inner Labrador Sea, which could indicate a lateral shift or a temperature change in the Labrador Current waters. A negative/positive pattern oriented north/south along the Gulf Stream could be observed in most of the monthly differences. The position, shape, and thermal content of the Gulf Stream could all be responsible for this difference against climatology; however, the monthly signature of the Gulf Stream was partially different from one system to the other. FOAM and Mercator Océan systems showed similar features, quite different from HYCOM, TOPAZ, and C-NOOFS. Note that the ocean interior exhibits a 0.5–1°C colder difference with HYCOM, and the opposite with FOAM. The TOPAZ and Mercator Océan systems have lower warm anomalies. Further accuracy assessment should be made using in situ observation comparisons, as proposed with Class 4 metrics.

SSTs of the forecasting systems were compared with the high-resolution SST product Operational Sea Surface Temperature and Sea Ice Analysis

Table 2. Description of each ocean analysis and forecasting system

System/ assimilation	Ocean model	Configuration	Atmospheric forcing	Assimilated data	Hindcast/ Forecast window
BLUElink>/ BODAS	MOM4.0d	OFAM; Global; 1/10° horizontal resolution around Australia (90°–180°E, south of 17°N); 47 z-levels	6-hour surface fluxes from Bureau of Meteorology	In-situ data from GTS and Coriolis + US GODAE; All available altimetric along-track SLA; Autralian tide gauge sea level; AMSR-E SST	Twice weekly analysis, 9-day hindcast and 7-day forecast
HYCOM/ NCODA	HYCOM; Los Alamos Ice model (CICE); KPP mixed layer model	Global; 1/12°cos (lat) resolution; hybrid vertical levels 32 σ layers	FNOC NOGAPS 0.5° surface fluxes, except NOGAPS 1° precipitation	T/S in-situ data; SST from satellite and in-situ data; Sea ice from SSM/I; Altimetric along-track SLA and SST used to produce T/S synthetic profiles	Daily analysis, 5-day hindcast, 5-day forecast
TOPAZ/ En KF	HYCOM 2.1.03; EVP; Thermodynamic ice model; KPP mixed layer model	North Atlantic (15°S) and Arctic (Bering Strait); 11–16 km resolution; hybrid vertical levels 22 σ layers	6-hour surface fluxes from ECMWF	In situ T/S from Coriolis; weekly altimetric SLA maps from AVISO; RTG SST; Sea Ice concentration; Sea ice drift	Weekly analysis, 1 week back in time, 10-day forecast
Mercator Océan global PSY3V2/ SAM2V1	NEMO 1.09; LIM2 sea ice model	ORCA025; Global, 1/4° x 1/4°cos(lat); 50 z-levels, partial steps	Daily mean; ECMWF; Bulk CLIO	In situ T/S from Coriolis; Jason-1, GFO, EnviSAT along-track SSH; RTG SST	Weekly analysis, 2 weeks back in time hindcast and 2 weeks forecasts
Mercator Océan North Atlantic PSY2V3/ SAM2V1	NEMO 1.09; LIM2 sea ice model	NATL12; Regional Atlantic 80°N–20°S; 1/12° x 1/12°cos(lat); 50 z-levels, partial steps	id	id	id
FOAM/ FOAM	NEMO 1.09; LIM2 sea ice model	Same as Mercator Océan PSY3V2	UK Met. 6-hour surface flux	In situ T/S from GTS; Jason-1, GFO and EnviSAT along-track SLA; SST OSTIA (GHRSST); Sea Ice concentration from SSM/I	Daily analysis, 5-day forecast
C-NOOFS/ No assimilation	NEMO 1.09	Regional Northwest Atlantic 103°–27°W, 26°–86°N; 1/4° x 1/4°cos(lat); Nested into Mercator Océan Global PSY3V1; 46 z-levels, partial steps	Hourly Environment Canada surface fluxes (33-km resolution)	Weekly condition of Mercator Océan PSY3 global system	Daily forecast
MOVE/MRI.COM-NP/ MOVE/MRI	MRI.COM; EVP Thermodynamic ice model	North Pacific nested into the global system: 15°S–65°N, 100°E–75°W; 1/2°x 1/2°; 54 levels σ-z hybrid coordinates	6-hour surface fluxes from JMA operational outputs	GTS in situ profiles; Jason-1, Envisat along-track SSH; MGDSST	Daily analysis; 1/3-month hindcast

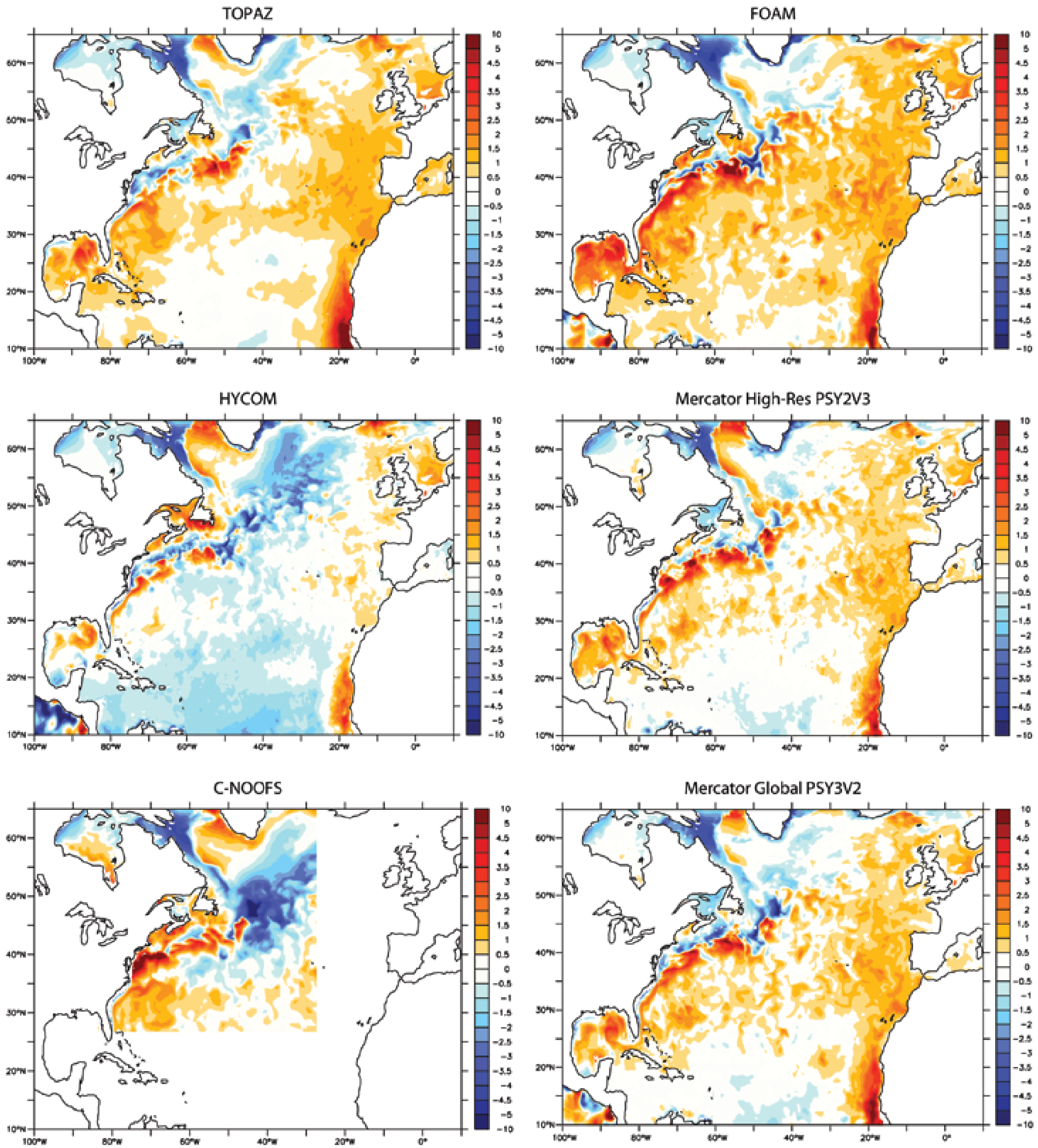


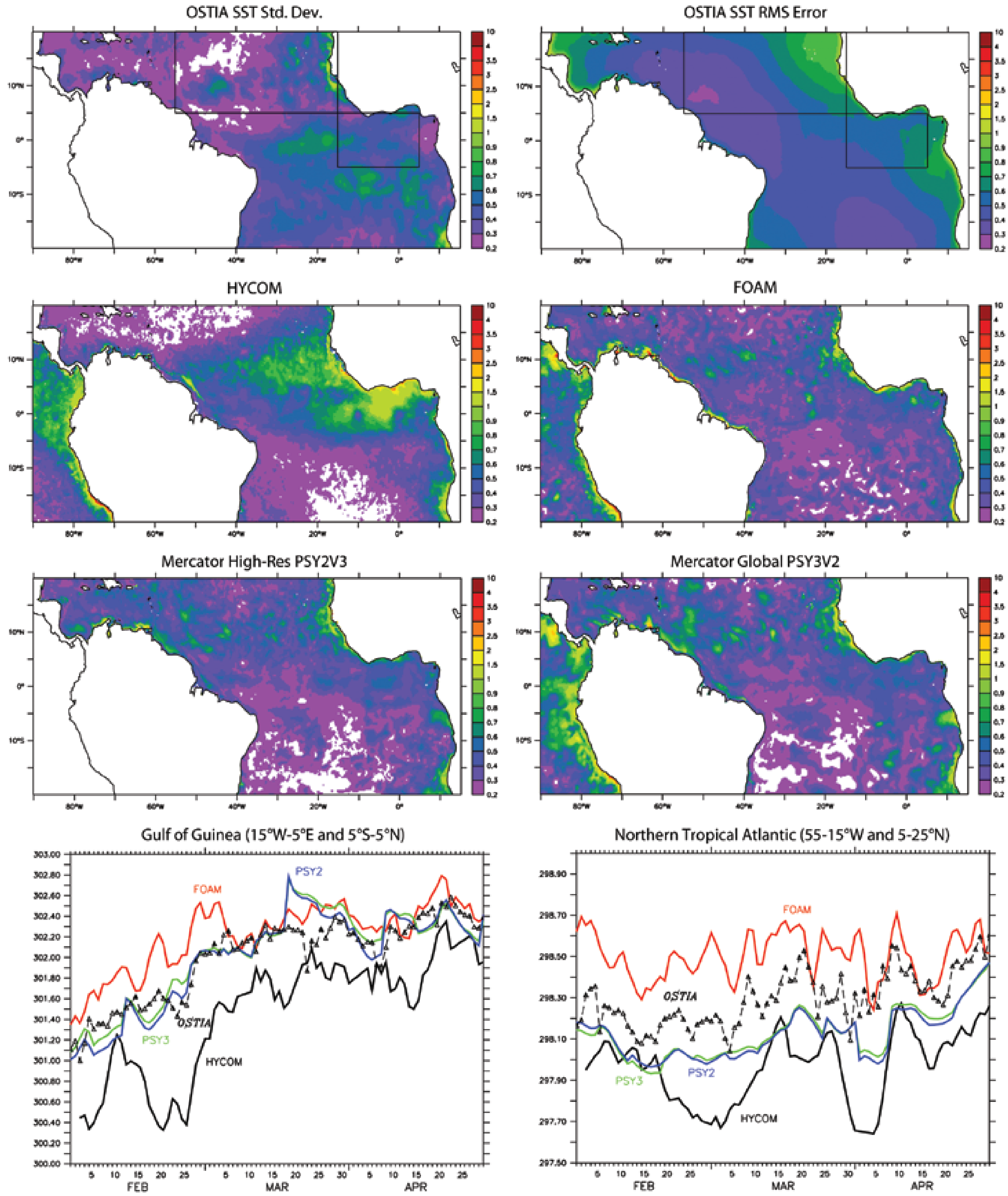
Figure 3. Temperature at 30 m in the North Atlantic area. Monthly mean differences with respect to World Ocean Atlas 2005 (Locarini et al., 2006) climatology in February 2008. Units are +5/-5 in Kelvin.

(OSTIA; see Donlon et al., this issue, for information on OSTIA and other high-resolution SST products developed during GODAE). The OSTIA product's high resolution (~ 6 km) highlighted errors in the eddy field, demonstrating, in the North Atlantic Ocean, that high-resolution systems such as HYCOM and Mercator Océan provide a better description of the eddy field, even when large-scale biases remain in these systems. Comparison with high-resolution SST observations also identified differences with respect to the climatology coming from interannual variability. At the surface, on average in the North Atlantic (not shown), HYCOM was colder (~ 0.2 K) and FOAM was warmer (~ 0.4 K) than OSTIA. These anomalies were reduced in April, indicating that the systems could have been in a "spin-up" process, where assimilation was still reducing surface waters' temperature biases. But the primary use of real-time observed products was intended to provide error levels for ocean forecasting products. Figure 4 illustrates this for the tropical Atlantic. The spatial distributions of RMS differences with OSTIA show that errors in the forecasting systems are not correlated with SST variability itself during the three-month test period. In other words, where SST changed, the systems were able to represent them. For the whole period, FOAM offered slightly better RMS differences compared to the Mercator Océan global product. Because the OSTIA SST product is assimilated into FOAM, this result is expected. HYCOM generally had higher discrepancies north of 5°S on the eastern side of the basin, but the lowest differences north of 10°N on the western side. Box average statistics confirmed that FOAM was slightly warmer than OSTIA SST in the Gulf of Guinea, and warmer in the northern part of the tropical Atlantic, in particular, until April. HYCOM appeared generally too cold. The two Mercator Océan systems showed similar behavior in box averages (with little impact from coarse versus higher horizontal resolution). Differences never exceeded 1°C RMS. All four systems more or less matched the temporal SST changes. Mercator Océan SST changes appeared spatially smoother than those of FOAM and HYCOM, possibly due to the assimilation of low-resolution NCEP/Reynolds SST.

Analysis of heat content showed a wider spectrum of results, where one system could be accurate in one area and present the largest biases in another. The main outcomes at this stage were that water masses present large differences, and all discrepancies (i.e., the level of accuracy) were quantified for all regions. However, to identify the causes of discrepancies both between the forecasting systems and against observations, dedicated analyses need to be carried out by looking at heat transports, air-sea fluxes, ocean mixing and other variables. The roles of each type of model, assimilation technique, or data set used for assimilation all need to be assessed as causes of the discrepancies. Because Class 1, 2, and 3 metrics allow analysis of heat and buoyancy fluxes as well as heat transports, a dedicated analysis of MLD (based on a temperature criterion difference of 0.2°C from SST) showed large discrepancies among operational system outputs. Again, air-sea fluxes, vertical mixing in the upper ocean layers, and differences in assimilation schemes could all play a role in causing these differences.

Figure 5 is a comparison of water masses at depth. Class 2 sections from the regional MOVE/MRI.COM (Multivariate Ocean Variational Estimation/Meteorological Research Institute Community Ocean Model) and the FOAM and Mercator global systems for a given day (April 3, 2008) were compared to salinity reference data from two cruises (April and September 2000). The objectives were to assess the consistency of North Pacific Intermediate Water distributions and to qualitatively identify the main biases at that depth. No large discrepancies were identified. The three assimilation methods, although different, reduced the main biases of these non-eddy-resolving systems, in particular, model errors caused by erroneous vertical/lateral mixing schemes (known problems of most ocean models), and thus they provide three reliable solutions. The distributions of North Pacific Intermediate Waters were similar to observations, although too shallow. Near the surface, at 45°N, the low-salinity patterns showed considerable differences, indicating that a more careful analysis of the mixed layer (freshwater fluxes, mixing, impact of assimilation) would be helpful. Similarly, in the subtropical gyre, representations of North Pacific Tropical Waters with high salinities from 0–200-m depth from the equator toward 30°N were not very realistic.

A careful analysis of the mean kinetic energy (KE) for the three-month period in different areas identified differences in wind-driven circulation patterns. Quality assessment was performed using independent current products, such as Surcouf (Larnicol et al., 2006) or Ocean Surface Current and Analysis in Real time (OSCAR; Johnson et al., 2007), that


OSTIA SST Std. Dev.

OSTIA SST RMS Error

HYCOM

FOAM

Mercator High-Res PSY2V3

Mercator Global PSY3V2

Gulf of Guinea (15°W-5°E and 5°S-5°N)

Northern Tropical Atlantic (55-15°W and 5-25°N)

Figure 4. Statistics of sea surface temperature (SST) differences from February 1 to April 30, 2008, in the tropical Atlantic. Top left: Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) SST standard deviation. Top right: OSTIA SST RMS errors. As labelled, HYCOM (HYbrid Coordinate Ocean Model), FOAM (Forecast Ocean Assimilation Model), and Mercator Océan systems' RMS differences with respect to OSTIA SST are given. Units are 0.2–5 Kelvin. Bottom panels: Daily time series of box-averaged SST (Kelvin) from February to April 2008. HYCOM = black. FOAM = red. Mercator Océan PSY2 = blue. Mercator Océan PSY3 = green. The bottom left plots are for a box-limited area in the Gulf of Guinea (15°W–5°E and 5°S-5°N), and the bottom right plots are for a box-limited area in the northern tropical Atlantic (55°–15°W and 5°–25°N). The two areas are plotted in the top left panel.

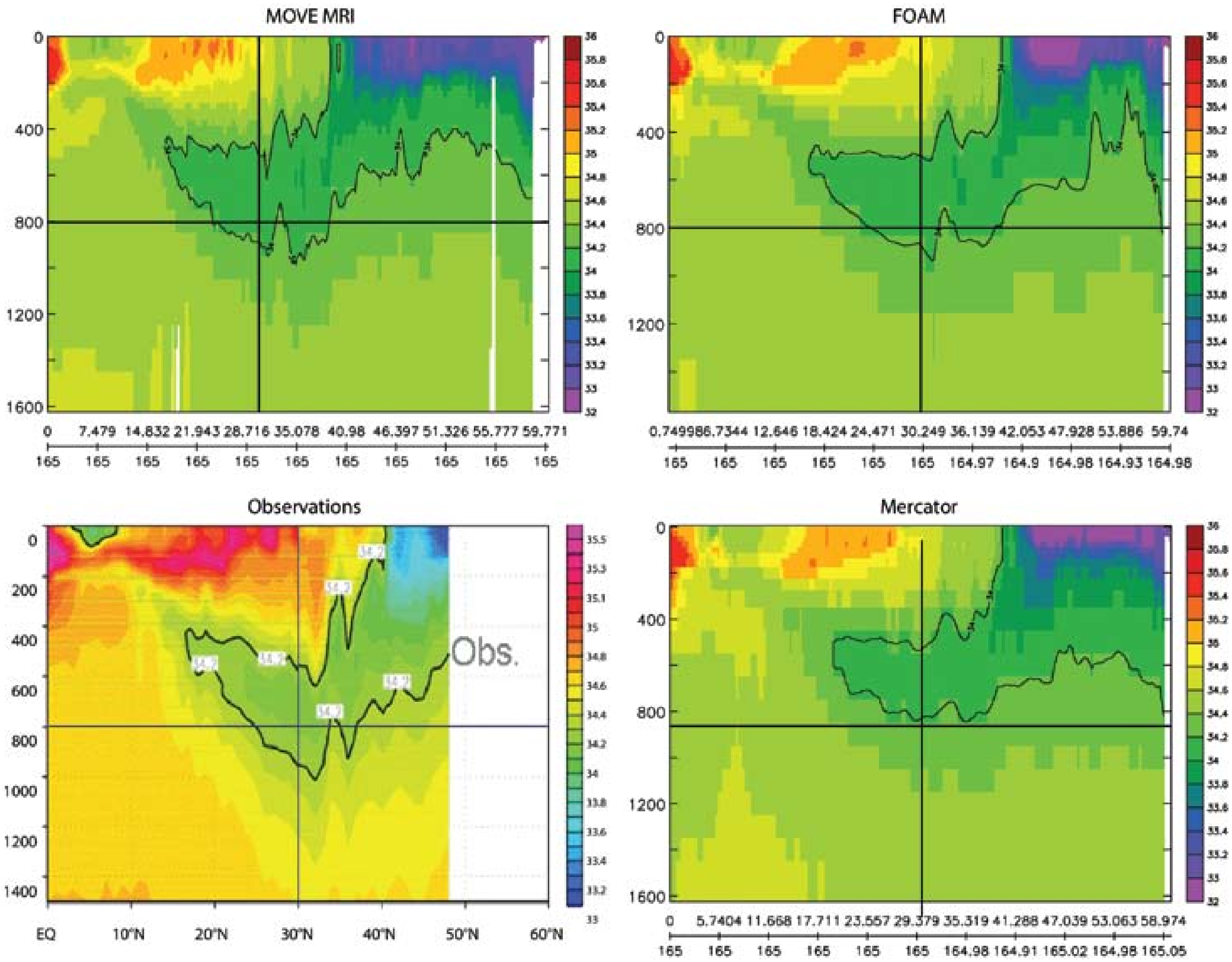
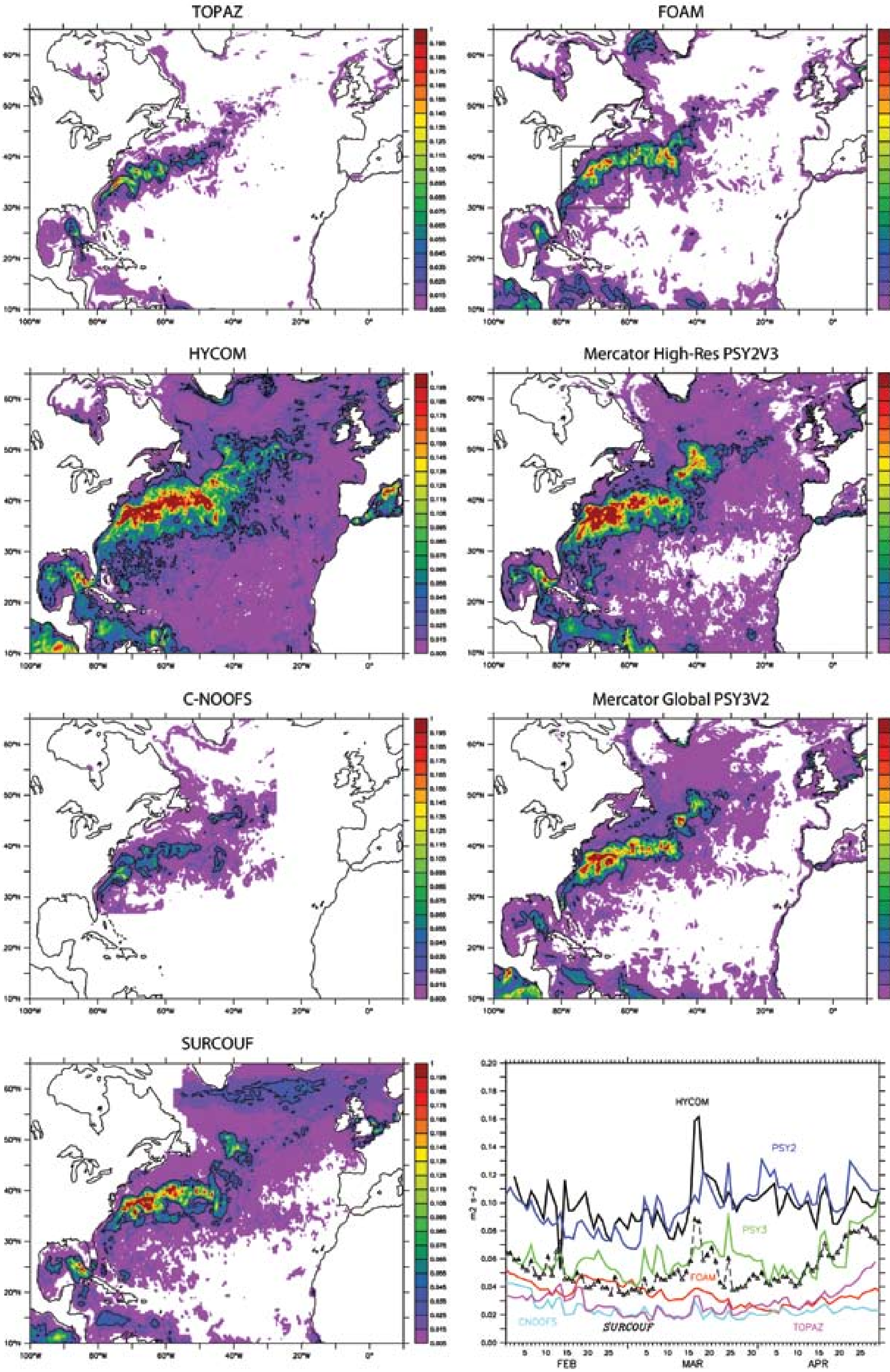


Figure 5. Salinity section comparison (latitude, depth, in practical salinity units) in the North Pacific Ocean at 165°E. Snapshots of MOVE/MRI.COM, FOAM, and Mercator Océan PSY3 Class 2 sections on April 3, 2008, are compared to a composite of CTD data obtained in April 2000. The 34.2 psu isohaline is represented.

routinely provide ocean currents near the surface. These products are deduced from satellite altimetry, using geostrophic assumptions, and Ekman current estimates from numerical weather predictions or satellite scatterometry. Thus, scales of Surcouf currents used in the comparisons are similar to altimeter-derived information, typically 20–30 km.

Comparisons showed that the HYCOM and Mercator Océan high-resolution systems offered higher mean energy and more consistent patterns than the other systems, confirming the positive impact of the finer horizontal grid, especially within interior or eastern boundary currents. Figure 6 presents eddy field turbulence, as deduced from eddy kinetic energy (EKE), for Surcouf and the six forecasting systems in the North Atlantic area during the three-month period. Conclusions are similar to the mean current analysis. The FOAM and Mercator Océan global systems give eddy field statistics similar to Surcouf, and Mercator Océan follows most of Surcouf time changes. EKE levels were computed daily in the Gulf Stream box described in Figure 6. Levels are higher but rather similar for the HYCOM and Mercator Océan high-resolution systems, with important variability. C-NOOFS and TOPAZ show EKE at half the levels of Surcouf EKE. This discrepancy may be due to satellite altimetry assimilation, lacking in C-NOOFS, while TOPAZ



TOPAZ

FOAM

HYCOM

Mercator High-Res PSY2V3

C-NOOFS

Mercator Global PSY3V2

SURCOUF

Figure 6. Averaged eddy kinetic energy (EKE) computed from February to April 2008 at the surface for the six forecasting systems, and the Surcouf surface current daily products. The contour line corresponds to 300 cm^2 s^{-2}. Energy levels from 50 to 2000 cm^2 s^{-2} are shaded. Bottom right panel: Daily time series of box-averaged eddy kinetic energy in a box-limited area around the Gulf Stream (80°–60°W and 30°–42°N, represented in the FOAM map). HYCOM = black). FOAM = red. Mercator Océan PSY2 = blue.. Mercator Océan PSY3 = green). TOPAZ = magenta. C-NOOFS = cyan. Surcouf EKE is plotted in a thin black line with symbols. Units are in m^2 s^{-2}.

ensemble averaging could impact averaged energy levels (about 20% less EKE than individual members, especially in low-energy areas).

A dedicated analysis was performed in the Indonesian through-flow area, where the BLUElink> system, with its high horizontal resolution, could also be compared. Results show that high resolution is key for representing transports through straits and for jets associated with Rossby waves in tropical bands. The impact of bottom friction and drag effects on mixing is clearly evident in the Indonesian marginal seas, where Pacific and Indian waters are mixed differently by FOAM, Mercator, BLUElink>, and HYCOM. Salinity differences in this area could also be caused by mixing effects, although runoff or precipitation errors should not be neglected, and further characterization is needed. Water mass distributions and thermocline depths certainly differ in the eastern Indian Ocean downstream of the Indonesian through flow. However, even if ocean model parameterization errors are present, assimilation of temperature and salinity data can also reduce the discrepancies and corresponding errors.

DISCUSSION AND CONCLUSIONS

The GODAE program provided the opportunity to develop an assessment methodology for ocean operational forecasting systems. Most of the operational centers dealing with open ocean, eddy-permitting model and assimilation techniques have been involved in the design and implementation of these assessment tools dedicated to providing scientific validation of forecasting systems and their products.

Consistency, quality, and performance of the operational forecasting systems have been evaluated through several projects. Metrics are now implemented in most of the GODAE ocean forecasting systems, and tools for communicating and exchanging these metrics have been adopted by most of these centers. The GODAE intercomparison project demonstrates the usefulness of these developments. In the near future, this assessment architecture is expected to be endorsed by the technical panels of the World Meteorological Organization-Intergovernmental Oceanographic Commission Joint Technical Commission for Oceanography and Marine Meteorology.

The GODAE intercomparison project's main outcomes are:

1. Intercomparisons of several systems were conducted that can represent any area of the world ocean.
2. The ocean forecasting systems offer a variety of model (ocean and sea-ice) configurations, and different assimilation techniques and observations were used.
3. Considering the short three-month period for the intercomparison study, ocean dynamics were found to be "consistent" in all systems—the general wind driven circulation was satisfactorily represented for the boreal winter season, and the assessment over different ocean basins (not fully described in this paper) indicated that thermohaline circulation and water mass distribution were also reasonably represented.
4. The systems were eddy-resolving or eddy-permitting; their day-to-day representations of eddy fields varied, but, statistically, the ocean variability was similar among the systems. There are regional discrepancies that need further analysis with consideration for all system components (e.g., numerical weather prediction forcing, ocean modeling, assimilation techniques, data used).

One important aspect of operational validation not emphasized here is the key role played by observations (a comprehensive presentation of the global ocean observing system is provided by Clark et al., this issue). There is no doubt that the validation methodology could only be successfully applied because observational projects promoted by GODAE, such as Argo (Roemmich and Argo Steering Team, this issue) and Global High-Resolution Sea Surface Temperature (GHRSST; Donlon et al., this issue), provide sufficient data in real time. Moreover, the data archiving and assembly centers such as US-GODAE and Coriolis, among others, play an important role, providing for the robust distribution of data in real time.

The set of tools and metrics that GODAE leaves as a legacy should be linked with several applications and activities. The CLIVAR Global Synthesis and Observation Panel aims to develop added value to long-term simulations and reanalyses of the ocean, setting up dedicated diagnostics over different basins. Coupled ocean-atmosphere models used for seasonal and longer forecasting are also validated using specific diagnostics. These different communities can benefit from GODAE metrics in order to verify consistency with operational ocean systems. These longer simulations are often performed inside operational centers using the same ocean model configuration used for short-term predictions. Observing system experiments based on an operational model configuration can also rely on GODAE metrics for assisting impact studies (see review by Oke et al., this issue). As downscaling becomes a more systematic approach to forecasting in coastal areas and regional seas, this set of GODAE metrics can be used in two ways: (1) to aid in the design of dedicated metrics for coastal areas following the methodology presented in this paper or by applying GODAE metrics directly to regional systems, thus allowing intercomparison between large-scale and downscaled operational systems (see De Mey et al., this issue), and (2) by the biogeochemical community, where most biogeochemical models are now being coupled to the physical ocean forecasting systems described here.

The GODAE intercomparison project focused mainly on scientific validation of operational products. The end-to-end system assessment, as tested in the MERSEA integrated project, was not emphasized among GODAE activities, that is, to design, implement, and test diagnostics that monitor the robustness of daily operations and detect failure and faults that could impact the quality of operational ocean estimates and forecasts in real time (e.g., lack of a particular set of satellite data during a given day, or most recent atmospheric forcing). However, in the near future, new European initiatives in the framework of the Global Monitoring for Environment and Security program, like the MyOcean project that attempts to define the future of operational oceanography in Europe, will endorse the need for the GODAE validation methodology.

ACKNOWLEDGEMENTS

The European Commission provided financial support to develop TOPAZ systems in the DIADEM/TOPAZ projects, and to develop FOAM, Mercator Océan, MFS, and TOPAZ systems in MERSEA Strand1 (Contract no. EVK3-CT-2002-00089) or the Mersea Integrated Project (Contract no. SIP3-CT-2003-502885) and BOSS4GMES projects. TOPAZ uses a CPU grant from the Norwegian supercomputing project (NOTUR). The French Mercator Océan systems were implemented with financial support from the Région Midi Pyrénées. The US effort was supported by the project US GODAE: Global Ocean Prediction with the HYbrid Coordinate Ocean Model (HYCOM), funded under the National Oceanographic Partnership Program, by the Office of Naval Research 6.1 project Global Remote Littoral Forcing via Deep Water Pathways under program element 601153N, and by grants of computer time from the US Department of Defense High Performance Computing Modernization Program (HPCMP). The Australian BLUElink> science and technical team was supported by the Bureau of Meteorology, CSIRO, Wealth from Oceans National Research Flagship, and the Royal Australian Navy. The Japanese GODAE effort was supported by JMA/MRI and by Category 7 of the MEXT RR2002 Project for Sustainable Coexistence of Humanity, Nature, and the Earth.

REFERENCES

Brasseur, P. 2006. Ocean data assimilation using sequential methods based on Kalman filter. Pp. 371–316 in *GODAE Summer School in "Ocean Weather Forecasting: An Integrated View of Oceanography."* E.P. Chassignet and J. Verron, eds, Springer, Lallonde les Maures, France.

Crosnier, L., and C. Le Provost. 2007. Inter-comparing five forecast operational systems in the North Atlantic and Mediterranean basins: The MERSEA-strand1 methodology. *Journal of Marine Systems* 65:354–375.

Crosnier, L., C. Le Provost, and MERSEA Strand1 team, 2006. Internal metrics definition for operational forecast systems inter-comparison: Examples in the North Atlantic and Mediterranean Sea. Pp. 455–465 in *GODAE Summer School in "Ocean Weather Forecasting: An Integrated View of Oceanography."* E.P. Chassignet and J. Verron, eds, Springer, Lallonde les Maures, France.

Johnson, E.S., F. Bonjean, G.S.E. Lagerloef, J.T. Gunn, and G.T. Mitchum. 2007. Validation and error analysis of OSCAR sea surface currents. *Journal of Atmospheric and Oceanic Technology* 24:688–701.

Larnicol, G., S. Guinehut, M.-H. Rio, M. Drévillon, Y. Faugère, and G. Nicolas. 2006. The Global Observed Ocean Products of the French Mercator project. *International Symposium on Radar Altimetry: 15 Years of Altimetry*, ESA/CNES.

Locarnini, R.A., A.V. Mishonov, J.I. Antonov, T.P. Boyer, and H.E. Garcia. 2006. *World Ocean Atlas 2005, Volume 1: Temperature*. S. Levitus, ed., NOAA Atlas NESDIS 51, US Government Printing Office, 182 pp.

Smith, N. 2006. Perspective from the Global Ocean Data Assimilation Experiment. Pp. 1–17 in *GODAE Summer school in "Ocean Weather Forecasting: An integrated view of oceanography."* E.P. Chassignet and J. Verron, eds, Springer, Lallonde les Maures, France.

Teague, W.J., M.J. Carron, and P.J. Hogan. 1990. A comparison between the generalized digital environmental model and Levitus climatologies. *Journal of Geophysical Research* 95:7,167–7,183.