Applying Automated Underway Ship Observations to Numerical Model Evaluation

SHAWN R. SMITH, KRISTEN BRIGGS, AND NICOLAS LOPEZ

Center for Ocean–Atmospheric Prediction Studies, Florida State University, Tallahassee, Florida

VASSILIKI KOURAFALOU

Rosenstiel School of Marine and Atmospheric Science, Department of Ocean Sciences, University of Miami, Miami, Florida

(Manuscript received 16 March 2015, in final form 19 November 2015)

ABSTRACT

Numerical models are used widely in the oceanic and atmospheric sciences to estimate and forecast conditions in the marine environment. Herein the application of in situ observations collected by automated instrumentation on ships at sampling rates ≤5 min is demonstrated as a means to evaluate numerical model analyses. Specific case studies use near-surface ocean observations collected by a merchant vessel, an ocean racing yacht, and select research vessels to evaluate various ocean analyses from the Hybrid Coordinate Ocean Model (HYCOM). Although some specific differences are identified between the observations and numerical model analyses, the purpose of these comparisons is to demonstrate the value of high-sampling-rate in situ observations collected on ships for numerical model evaluation.

1. Introduction

Numerical models are routinely used to estimate and forecast oceanic and atmospheric conditions. These models undergo continual changes (e.g., new model physics, improved data assimilation) that impact the model analyses and forecasts, and subsequently require these products to be evaluated for accuracy over a range of surface and subsurface features (e.g., winds, temperatures, currents, eddies, and thermohaline gradients). Studies (e.g., Scott et al. 2010) highlight a lack of consensus among different ocean general circulation models (OGCMs) in various predictions, often linked to differences in the air–sea exchange parameters derived by atmospheric reanalysis models (e.g., Smith et al. 2011). Scott et al. (2010) investigated the total kinetic energy derived from four separate OGCMs and found that at individual current meter moorings, the models differed not only from each other but also from the moored current meter records to which they were being compared. Smith et al. (2011) compared variations in turbulent heat fluxes and wind stress parameters from three reanalysis products to in situ and satellite-based flux products and found wide disagreement in the model fluxes. The need for accurate oceanic and atmospheric model forecasts continues to grow to support decision-making for industry (e.g., commercial fishing, offshore energy development) and managing risks (e.g., storm surge, harmful algae blooms, pollution) to coastal communities. Developing and improving numerical models can best be achieved using high-quality evaluation datasets.

Herein we demonstrate the application of in situ observations collected by automated instrumentation on ships at sampling rates ≤5 min as a means to evaluate numerical model analyses. The focus is on physical oceanographic parameters (velocity, salinity, and sea temperature); however, the techniques demonstrated could be applied using atmospheric, chemical, or biological measurements from similar vessels. The use of vessel-based observations to conduct model evaluation is certainly not without precedent. Sturges and Bozec (2013) examined a westward mean flow suggested in certain areas of the Gulf of Mexico by a long-term set of ship drift data (and a second, independent long-term set

Corresponding author address: Shawn R. Smith, Center for Ocean–Atmospheric Prediction Studies, Florida State University, P.O. Box 3062741, Tallahassee, FL 32306-2741.
E-mail: smith@coaps.fsu.edu

DOI: 10.1175/JTECH-D-15-0052.1

of in situ observations) and found that several numerical models that they investigated did not appear to capture the observed feature. Androulidakis and Kourafalou (2013) used research vessel observations to evaluate a high-resolution regional ocean model they were using to examine the transport and fate of Mississippi waters in the Gulf of Mexico when the river was experiencing flood outflow volumes. In general, root-mean-square errors were small (see their Fig. 5), indicating the model estimates of surface salinity and SST were consistent with the shipboard observations. Additionally, Smith et al. (2001) used automated meteorological observations from research vessels to identify major shortcomings in the air–sea fluxes in the NCEP–NCAR atmospheric model reanalysis.

The authors present three case studies that compare in situ observations from a merchant vessel, a racing yacht, and select research vessels to ocean analyses produced by the Hybrid Coordinate Ocean Model (HYCOM; Chassignet et al. 2009). HYCOM is used as the numerical model in this manuscript, but the techniques could be applied to other oceanic, as well as atmospheric, models. The case studies presented do not provide a comprehensive look at all ocean basins but focus on ocean regions where the individual ship's operations provide a unique comparison to the model.

Our goal is to demonstrate that automated underway observations collected by ships provide an excellent resource to evaluate numerical models, and the case studies shown are not intended to provide a comprehensive evaluation of HYCOM. The authors identify the strengths and limitations of each in situ data type and the comparison techniques used in each case study. Application of these data and the demonstrated techniques in future comprehensive model comparisons should benefit model developers by highlighting areas for improvement in models and allow users of model products to understand the strengths and limitations of the model fields presently available to the community.

2. Modeling and data

a. HYCOM

Three different applications of the HYCOM code (https://hycom.org/) are used for the validation case studies: the global (GLB-HYCOM; Chassignet et al. 2009), the regional Gulf of Mexico (GoM-HYCOM; Prasad and Hogan 2007; Kourafalou et al. 2009; Halliwell et al. 2009), and the nested northern Gulf of Mexico (NGoM-HYCOM; Schiller et al. 2011; Kourafalou and Androulidakis 2013; Androulidakis and Kourafalou 2013). The GLB and GoM models employ the Navy Coupled Ocean Data Assimilation (NCODA) system (Cummings 2005); NGoM is a free-running model. The GLB and GoM models run in real time, and their hindcast analyses are archived and accessed via the HYCOM THREDDS server maintained by the Florida State University (FSU) Center for Ocean–Atmospheric Prediction Studies (COAPS; http://hycom.org/dataserver/). Sea surface potential temperature, sea surface salinity, and ocean velocity (zonal and meridional) fields are extracted for the various comparisons in section 3.

The global GLB-HYCOM has a curvilinear $^1/_{12}^\circ$ grid covering 90°N–78°S and is forced by atmospheric parameters provided by the 0.5° coupled ocean–atmosphere Navy Operational Global Atmospheric Prediction System (NOGAPS; Hogan and Rosmond 1991; Rosmond 1992; Hogan and Brody 1993). We use daily archives of the Mercator grid–based portion of the GLB-HYCOM, which are limited to 47°N–78°S, for comparisons with data from the racing yacht and the merchant vessel.

The GoM-HYCOM provides hourly outputs on a terrain-following $^1/_{25}^\circ$ grid, which are used for evaluation of salinity predictions in the Gulf of Mexico. GoM-HYCOM is also forced by 0.5° NOGAPS and shares other attributes with GLB-HYCOM, especially in the treatment of riverine inputs, which influence the model's salinity fields, and the relaxation of sea surface salinity (SSS) to climatology. The major rivers are included and parameterized through a virtual salinity flux and monthly climatological values of river discharge.

The NGoM-HYCOM is nested within the GoM-HYCOM (Fig. 1d) and thus receives interactions of coastal/shelf dynamics with the basinwide flows (especially the Loop Current branch of the Gulf Stream system) through its boundaries. NGoM has double the horizontal resolution ($^1/_{50}^\circ$) of the outer GoM model and is also forced by higher-resolution atmospheric fields from the navy's 27-km-resolution Coupled Ocean–Atmospheric Mesoscale Prediction System (COAMPS; Hodur et al. 2002). Most importantly for this study, NGoM has a detailed parameterization of river plume dynamics, realistic salt and mass fluxes following Schiller and Kourafalou (2010), and no relaxation of SSS to climatology. Daily freshwater discharges are prescribed for 17 major rivers along the NGoM coastal zone.

The spatial and temporal sampling variations between the underway observations used herein and daily HYCOM output used for each comparison require interpolation of the respective HYCOM data to individual vessel data points. The MATLAB interp2 bilinear interpolation function (http://www.mathworks.com/help/matlab/ref/interp2.html#btyq8s0-2_1) is chosen for spatial interpolation using the function

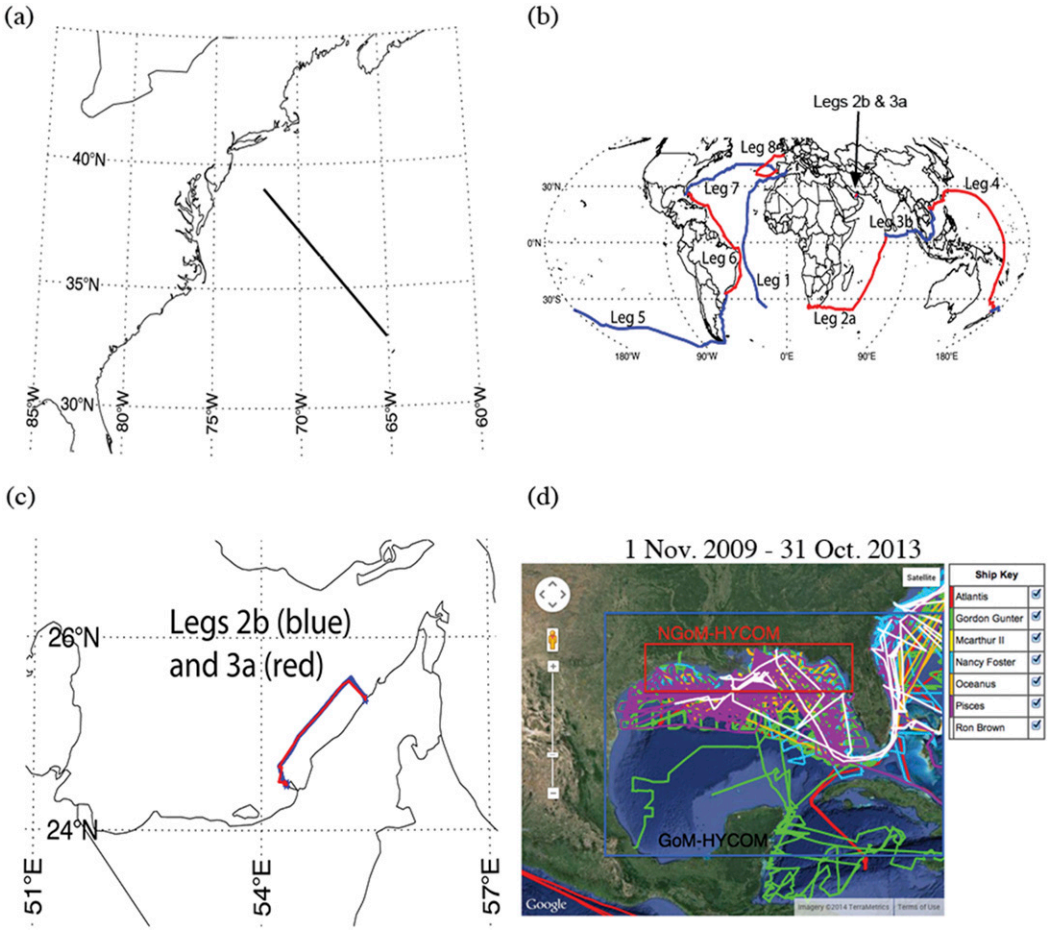

FIG. 1. Cruise maps for the ship observations used in the case studies. (a) A representative track by the M/V *Oleander* between Bermuda and New York. (b) *Mar Mostro* VOR 2011–12 race legs; race began at Alicante and officially ended at Galway (final two *Mar Mostro* legs not used and not shown). A broken mast during leg 1 resulted in a break in data between legs 1 and 2A. (c) *Mar Mostro* legs 2B and 3A, which were confined to the Persian Gulf. (d) Gulf of Mexico SAMOS vessel tracks from 1 Nov 2009 to 31 Oct 2013 showing available observations within the GoM-HYCOM (blue box) and the NGoM-HYCOM (red box) domains. Note that not all data along these tracks are used in the analysis (see text).

$$\mathrm{HY}(x,y) \approx \frac{1}{(x_2-x_1)(y_2-y_1)}\begin{bmatrix}\mathrm{HY}(x_1,y_1)(x_2-x)(y_2-y)+\mathrm{HY}(x_2,y_1)(x-x_1)(y_2-y)\\ +\mathrm{HY}(x_1,y_2)(x_2-x)(y-y_1)+\mathrm{HY}(x_2,y_2)(x-x_1)(y-y_1)\end{bmatrix}, \quad (1)$$

where $\mathrm{HY}(x, y)$ is the interpolated value at the desired x, y and $x_1 < x < x_2$ and $y_1 < y < y_2$. Temporally, each HYCOM product has daily analyses available and the 0000 UTC analysis data are interpolated spatially to the respective 5-min/10-s/1-min observations from the merchant/racing/research vessel (see next section) for each day along the cruise track. No time interpolation is applied. Hourly analyses from GoM-HYCOM are used to test the sensitivity of using only 0000 UTC model fields in the GoM analysis (see section 3c).

b. *Vessels*

Underway observations applied to evaluate HYCOM model results are obtained from three types of vessels: a merchant vessel, a racing yacht, and select oceanographic research vessels (Table 1). The first case study uses data from the Motor Vessel (M/V) *Oleander*, a container vessel operated by the Bermuda Container Line and instrumented by the Oleander Project since 1992. The *Oleander* provides near-surface and subsurface ocean measurements along weekly transects

TABLE 1. List of variables measured with automated instrumentation and available for each vessel. Not all parameters are used in this study.

Mar Mostro	M/V *Oleander*	SAMOS vessels
Latitude, longitude, measured current direction, measure current rate, SST, atmospheric pressure, true wind speed, true wind direction, boat speed, course, speed over ground, course over ground	Latitude, longitude, zonal current rate, meridional current rate, SST, SSS	All vessels—latitude, longitude, vessel speed and course over ground, Earth-relative wind direction and speed, atmospheric pressure, air temperature, relative humidity, SSS *McArthur II, Gordon Gunter, Ronald Brown, Nancy Foster, Oceanus, Atlantis*—SST, conductivity *Oceanus, Atlantis*—precipitation accumulation, rain rate *Ronald Brown, Oceanus, Atlantis*—shortwave radiation

from Bermuda to New Jersey (Rossby and Gottlieb 1998). Data for the second case study were collected by the *Mar Mostro* during the around-the-world Volvo Ocean Race (VOR) 2011–12 (http://www.volvooceanrace.com/en/home.html). The *Mar Mostro* is a 21.5-m sail-powered racing yacht that was manned by the Puma Ocean Racing team (sponsored by the Puma sports apparel company) and powered by Berg Propulsion of Sweden. Finally, select research vessels contributing data to the Shipboard Automated Meteorological and Oceanographic System (SAMOS) initiative (http://samos.coaps.fsu.edu/html/) provide sea surface salinity and sea temperature data for the third case study. The diversity of ships equipped to make underway ocean measurements and their variety of operating locations provides opportunities to demonstrate evaluation of model predictions from regional to global scales.

1) M/V *OLEANDER*

Acoustic Doppler current profiler (ADCP) observations collected by the M/V *Oleander* are obtained from the Oleander Project (http://po.msrc.sunysb.edu/Oleander/) at Stony Brook University in MATLAB format to examine the subsurface structure of the Gulf Stream in HYCOM. The Oleander Project (Rossby and Gottlieb 1998) maintains underway instrumentation to collect water temperature, salinity, and current observations. The typical *Oleander* transit cruise takes approximately three days, with about 1-day port stops at each end of the cruise (Fig. 1a). As a result, the *Oleander* transects the Gulf Stream on 70–80 cruises per year. ADCP measurements are taken approximately every 5 min on every cruise, so the Gulf Stream is well sampled. This makes the *Oleander* data an ideal candidate for studying HYCOM performance in this dynamic marine region.

Since 2005, the *Oleander* has been outfitted with a Teledyne RD Instruments 75-kHz Ocean Surveyor ADCP. At its optimal performance level, the ADCP can provide horizontal velocities to depths of ~800 m. The cruises for our comparative analyses were selected by our colleagues at the University of Rhode Island (URI) Graduate School of Oceanography to provide us with high-quality ADCP observations that have adequate coverage of the Gulf Stream to permit identification of the core of the current. This selection process resulted in 33 Gulf Stream transects (usually spanning 2 days each) between 16 February 2007 and 15 October 2008. Currents are sampled by the ADCP every 10 m at depths between 25 and 995 m, at approximate 5-min intervals. ADCP data are not detided; open-ocean tides in the vicinity of the Gulf Stream are of insufficient amplitude (generally 1 cm s^{-1} or less) to have significant impact on the comparisons herein. However, future users of ADCP data from vessels may wish to detide the data if appropriate for their detailed model evaluations.

For this comparison, we use the archives from the GLB-HYCOM, which contain 32 vertical layers, unevenly spaced, between the surface and the 5500-m depth. Of these levels, the 75- and 125-m depths are the only two levels matched between the HYCOM data and the *Oleander* ADCP data. To avoid interpolation in the z coordinate and thus minimize averaging error, 75 m was chosen for the GLB-HYCOM u (zonal) and v (meridional) interpolation to *Oleander* cruise tracks. The spatial and temporal sampling variations between 5-min interval ADCP observations from the *Oleander* and daily HYCOM output requires interpolation of the HYCOM data to individual vessel data points using the technique described in section 2a.

2) OCEAN RACING VESSEL *MAR MOSTRO*

Data from the *Mar Mostro* were provided to the authors postrace by Robert Hopkins Jr., team performance analyst and coach during the 2011–12 race. The

Mar Mostro dataset contains sea surface temperature, measured surface current speed and direction (both of which were derived from the vector difference of the vessel course and speed through water and the vessel course and speed over ground), and various navigational parameters (see Table 1). Geographically, the high-resolution (10-s sampling rate) data circled the globe between the latitudes of 58.86°S and 58.03°N. The race course (Fig. 1b) began at Alicante, Spain, and officially ended at Galway, Ireland. *Mar Mostro* sailed an additional (10th) 4-day private leg from Galway to Hönö Island, Sweden, postrace. The data were split into 10 legs prior to provision to the authors (with legs 2 and 3 each being split in two again). The version of the GLB-HYCOM used by the authors has a limited MATLAB-analyzable domain, as described in section 2a, only allowing analysis using legs 1–8 (Fig. 1b). SST data were collected from an Airmar depth/temperature sensor mounted in a through-hull configuration at about the 0.5-m depth. However, we suspect that while the vessel was sailing, the effective sampling depth of the sensor was a mixed sample of the first 0.2 m of the water column, owing to hull-induced surface water entrainment. Both measured current direction and measured current rate are defined by the vector difference of the vessel motion through water and with respect to the ground presentations. The speed and course over water are from a Nortek Doppler velocity logger, and the speed and course over ground are derived from a global positioning system (GPS). Ocean current velocity data were not detided prior to distribution, and this may account for some portion of the differences shown in our analyses. Again, future users of these data may wish to detide the data if required for their specific model evaluation. Occasional "bad" latitude/longitude measurements (0°N, 0°E) and anomalous "spikes" in current and SST data necessitated point removal via a tunable sigma-trimming window function. Details concerning *Mar Mostro* dataset point removal can be found in the appendix.

3) SAMOS RESEARCH VESSELS

The final case study examines SSS observations from research vessels participating in the SAMOS initiative and focuses on the Gulf of Mexico. SAMOS provides 1-min interval sampling of both atmospheric and oceanographic variables collected by 34 R/Vs, 7 of which routinely measure SSS in the Gulf of Mexico (Smith et al. 2009). Salinity data are extracted for the R/Vs *Pisces*, *Ronald Brown*, *McArthur II*, *Gordon Gunter*, *Nancy Foster*, *Oceanus*, and *Atlantis* within the GoM-HYCOM domain (Fig. 1d) for the period 1 November 2009–31 October 2013. Salinity data from approximately 70 cruises, typically lasting 4–10 days, are used. Ship routes are confined predominantly to the northern Gulf of Mexico, which allows for adequate sampling of this area, including the Mississippi River delta region (Fig. 1d). Temporally, the bulk of the observations were made between March and November (Fig. 2); five of the vessels are operated by NOAA, and these ships typically lay up during the winter.

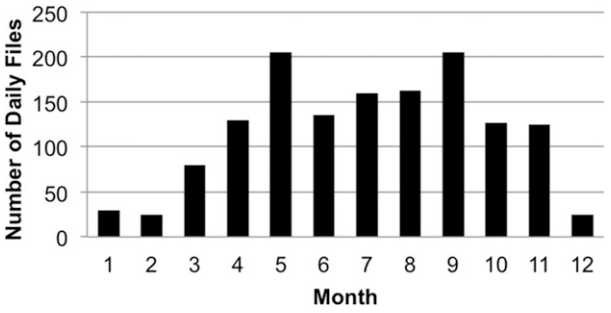

FIG. 2. Temporal distribution of SAMOS data between 2010 and 2013, showing that cruises in the Gulf of Mexico primarily occur between March and November annually.

Each SAMOS data record used in the comparison with HYCOM must have a salinity value that is flagged with a "Z" (good data) within the SAMOS's quality-control scheme (http://samos.coaps.fsu.edu/html/samos_quality_flag.php). Additionally, salinity values had to fall within a reasonable 5–50 psu range. Typical open-ocean salinity values are approximately 30–35 psu; shelf values near river-influenced areas are much lower. However, it is unlikely that the selected research vessels came close enough to the Louisiana coast to measure a salinity value below 5 psu. The low-limit constraint is applied to the dataset to reduce the chance of including data from a vessel that left its thermosalinograph running while entering and remaining in a port. No other modifications are made to the data from these research vessels. The spatial coverage of the accepted SAMOS salinity observations is most dense along the northern Gulf of Mexico coast because of the high frequency of ships entering and leaving Mobile Bay and Pascagoula, Mississippi.

3. Comparison case studies

a. *M/V* Oleander

1) 75-M CURRENTS

Current speed is calculated from interpolated GLB-HYCOM u and v data and from M/V *Oleander* u and v data using $\mathrm{spd}_k = \sqrt{u_k^2 + v_k^2}$, where k are *Oleander* data points; spd_k is HYCOM (*Oleander*) current speed

calculated at point k; and u_k and v_k are the HYCOM (*Oleander*) zonal and meridional current velocities, respectively, bilinearly interpolated to (measured at) point k. Box plots for u, v, and the speed of the current for all 33 *Oleander* cruises (Fig. 3) reveal that the *Oleander* data have a broader range of values than do those from the HYCOM. In particular, *Oleander* data maximums are clearly higher than those of the HYCOM for all three current variables, which is expected since the *Oleander* is sampling every 5 min versus the daily interval for GLB-HYCOM. On the other hand, variances appear quite similar for both platforms for u and v (as evidenced by interquartile ranges). Regarding the speed of the current, the variance appears only slightly larger and more positively skewed for the *Oleander* data, and the median value is approximately equal for both platforms. The mean velocity differences and root-mean-square errors between the GLB-HYCOM and *Oleander* 75-m u, v, and the current speed for each individual cruise (Table 2) reveal a slight negative spd bias (i.e., more positively skewed for the *Oleander* data) on 19 of the 33 cruises. The majority of the u and v RMSE and all spd RMSE are within 0.5 ms^{-1}. In terms of these statistics alone, the GLB-HYCOM performs well, overall, in predicting the strength of the Gulf Stream at the 75-m depth.

Subtle differences between the *Oleander* ADCP data and the GLB-HYCOM u and v velocity vectors do appear when we examine individual cruises (Fig. 4). The 17–18 November 2007 cruise (Fig. 4a) shows some areas of fairly good speed agreement even though the directions are not in total agreement between the HYCOM and the *Oleander* velocity vectors. The 29–31 March 2008 cruise (Fig. 4b) is a case in which the speed and direction of the currents are similar between HYCOM and *Oleander* in the southeast portion of the cruise track, where currents are small, but exhibit larger differences (particularly in direction) in the currents when the *Oleander* crosses the two main eddies along the northwest half of the track. The discrepancies may indicate a difference in the location or intensity of these eddies in the model as compared to the location or intensity of the eddies in the *Oleander* observations. Overall, the 29–31 March 2008 case highlights a tendency, which we note throughout our comparisons, toward some disagreement on the strength and/or location or shape of eddies between the GLB-HYCOM and the *Oleander* ADCP data. In general, locations where the current is strongest exhibit the largest differences in 75-m currents. The sea surface elevation (SSH; Fig. 4) suggests these differences occur where the vessel crosses the sharp gradients or eddies associated with the Gulf Stream.

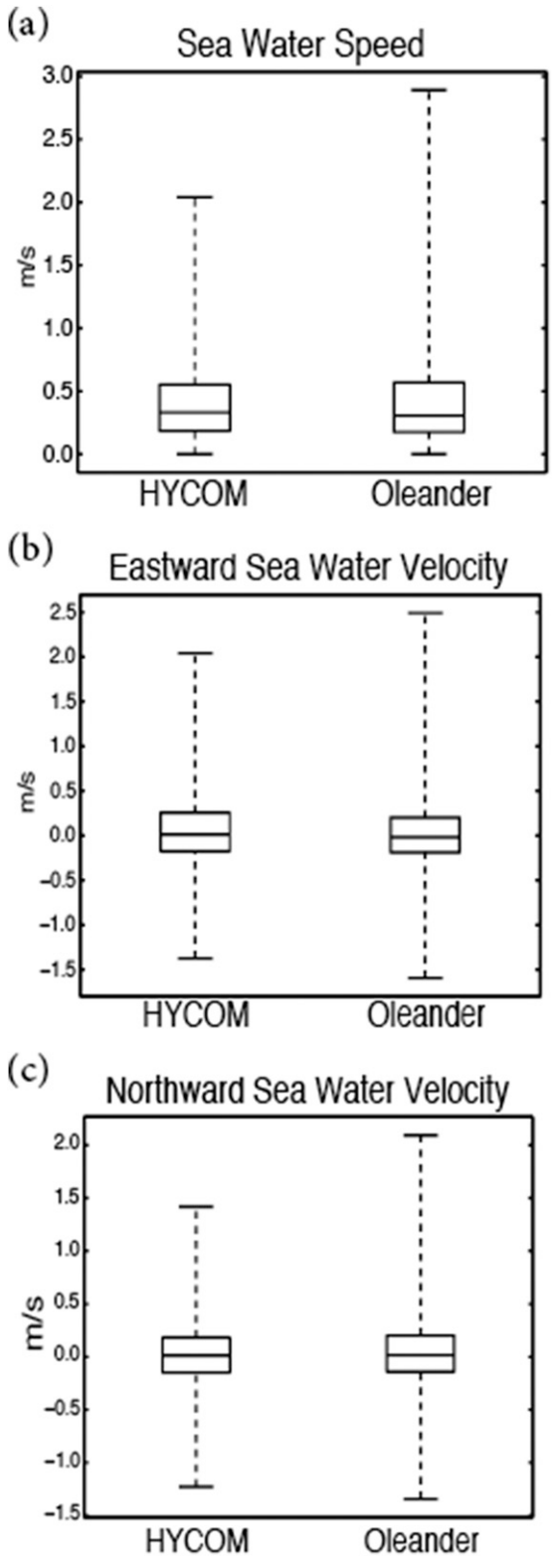

FIG. 3. Box plots for (a) 75-m seawater speed, (b) 75-m seawater eastward velocity, and (c) 75-m seawater northward velocity for GLB-HYCOM vs the *Oleander* for 33 *Oleander* cruises, spanning February 2007 through October 2008. Lower/upper box edges represent 25th/75th percentiles. Whiskers represent population maximums/minimums (presumed valid).

TABLE 2. Comparison statistics for GLB-HYCOM vs *Oleander* for 75-m u, 75-m v, and spd of 75-m current.

	u (ms^{-1})		v (ms^{-1})		spd (ms^{-1})	
Cruise ending	Mean bias	RMSE	Mean bias	RMSE	Mean bias	RMSE
18 Feb 2007	−0.07	0.37	0.13	0.34	−0.03	0.33
3 Jun	−0.13	0.53	0.19	0.55	−0.07	0.46
10 Jun	0.04	0.35	−0.01	0.28	−0.03	0.36
15 Jul	0.00	0.43	0.03	0.31	−0.05	0.32
22 Jul	0.08	0.52	−0.05	0.26	0.01	0.45
26 Jul	0.05	0.53	−0.09	0.33	−0.02	0.50
2 Aug	0.05	0.58	−0.05	0.22	0.01	0.48
6 Sep	0.10	0.32	−0.13	0.27	−0.06	0.30
30 Sep	0.18	0.47	−0.18	0.50	0.17	0.44
1 Nov	0.07	0.42	−0.14	0.40	−0.03	0.41
18 Nov	0.11	0.38	−0.11	0.36	0.07	0.28
17 Feb 2008	−0.12	0.45	−0.06	0.23	0.01	0.40
16 Mar	−0.07	0.42	−0.04	0.37	−0.09	0.41
31 Mar	−0.01	0.43	0.04	0.47	−0.12	0.46
13 Apr	0.13	0.47	−0.06	0.51	−0.17	0.49
24 Apr	0.12	0.49	−0.21	0.53	−0.13	0.49
4 May	0.11	0.51	−0.07	0.37	−0.13	0.37
8 May	0.14	0.49	−0.20	0.42	−0.21	0.40
25 May	0.01	0.33	−0.01	0.37	−0.06	0.38
15 Jun	−0.02	0.46	−0.04	0.30	−0.13	0.39
27 Jul	0.03	0.29	−0.02	0.32	−0.09	0.32
31 Jul	0.03	0.30	−0.01	0.33	−0.09	0.34
3 Aug	−0.03	0.39	0.03	0.36	0.04	0.43
7 Aug	0.09	0.40	−0.13	0.37	0.04	0.36
10 Aug	−0.11	0.29	0.05	0.35	0.10	0.27
14 Aug	0.19	0.40	−0.17	0.42	0.07	0.36
27 Aug	0.06	0.27	−0.02	0.26	0.04	0.26
31 Aug	0.03	0.28	0.04	0.30	−0.02	0.27
14 Sep	−0.09	0.36	0.05	0.29	−0.03	0.37
18 Sep	0.17	0.28	−0.30	0.63	−0.08	0.34
2 Oct	0.08	0.47	−0.11	0.34	−0.04	0.47
5 Oct	−0.04	0.34	0.07	0.41	−0.07	0.30
16 Oct	−0.17	0.59	0.12	0.29	−0.21	0.52

2) IDENTIFICATION OF GULF STREAM CORE

As discussed in Howe et al. (2009), various methodologies exist for defining the core of the Gulf Stream (e.g., Meinen and Luther 2003; Meinen et al. 2009). A primary factor in choosing a method is the type of observation being considered. Since the M/V *Oleander* provides ADCP data, we have chosen to define the Gulf Stream core as the location where maximum seawater velocity at the 75-m depth (i.e., 75m $|uv|_{\max}$) occurs along the *Oleander* track, for both the *Oleander* and the interpolated GLB-HYCOM datasets. Notably, this choice agrees well with GLB-HYCOM data, since the 75-m depth is one of the 32 levels on which seawater velocity is explicitly defined by the model. This depth was also determined by our URI colleagues to be a depth at which the ADCP consistently measured accurate currents. Our approach is similar to a study of the Kuroshio Extension by Howe et al. (2009) in which ADCP data were averaged over the 100- to 300-m depth range and gridded horizontally to a 5-km grid. The core was then identified at the location of maximum velocity within the averaged, regridded domain. Howe et al. (2009) opted not to use a single depth for their core definition to reduce the influence of noise in the data. In our case, because we are comparing ADCP data with regularly gridded model data, we choose to use a single depth on the model's native grid to avoid introducing averaging error into the model data. The only other matched level between the GLB-HYCOM and the *Oleander* data is the 125-m depth level; therefore, any averaging between levels other than 75 and 125 m would have involved additional interpolation and thus more uncertainty in the comparison. Further, while the *Oleander* ADCP levels are evenly spaced every 10 m, GLB-HYCOM depth levels are spaced farther apart in the subsurface region of interest for Gulf Stream core identification, increasing from 25 m apart between the

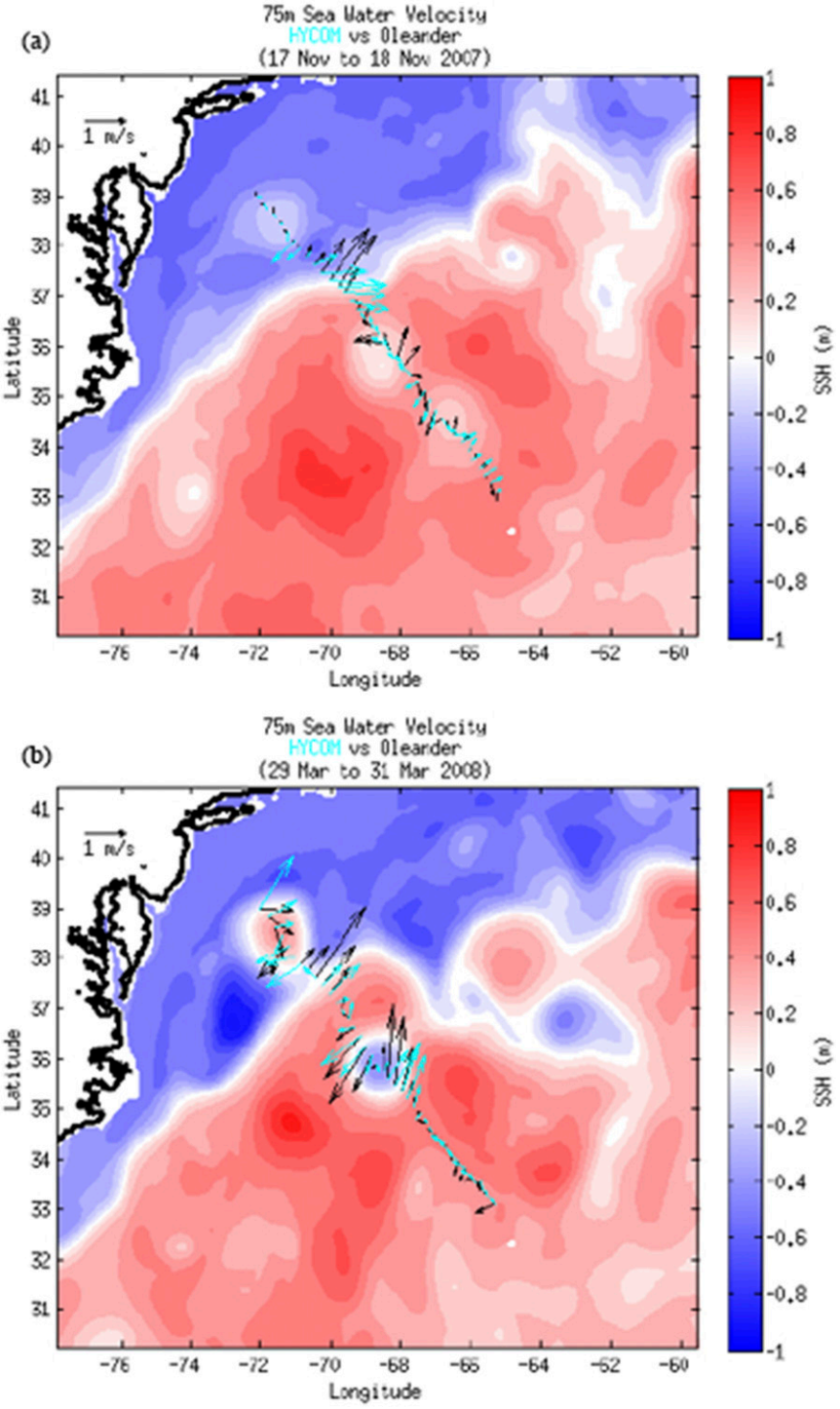

FIG. 4. Comparisons of seawater velocity vectors at the 75-m depth between the GLB-HYCOM (cyan) vs *Oleander* (black) for the cruises ending (a) 18 Nov 2007 and (b) 31 Mar 2008. Velocity vectors are plotted at every 10th point along the *Oleander* track and overlay the cruise-averaged GLB-HYCOM SSH. SSH is contoured every 0.1 m with the magnitude noted by the color bar.

50- and 150-m depth levels to 50 m apart between the 150- and 300-m depth and even greater below the 300-m depth level. A core identification scheme that involves averaging velocities over a range of depths therefore seemed a less robust option in this particular case study.

Examination of the position and magnitude of the Gulf Stream core at 75 m provides an index of the performance of the HYCOM. With few exceptions, the comparisons between the position and direction of the core agree well between the *Oleander* velocity data and the GLB-HYCOM interpolated data, and core speeds are slightly lower in the GLB-HYCOM fields, with a mean difference of almost $-1\,\mathrm{m\,s^{-1}}$ (Fig. 5). Broadly speaking, mean differences in core location of ~4 and ~3 km for latitude and longitude, respectively, as compared to the Gulf Stream being typically ~100 km wide indicate that the GLB-HYCOM core location deviations are well within reason, especially when considering the ~9-km gridpoint spacing available from the GLB-HYCOM. The mean directional difference is 29°, representing only 16% of the maximum difference possible between two core directions of ±180°, since direction is a polar quantity. As such, two of the three spikes in the core direction plot (Fig. 5d) appear much larger than they actually are; only the 25 May 2008 data points diverge as greatly as perceived, with a separation of ~170°.

Using Fig. 4, we focus on two notable cases from the core identification exercise (Fig. 5). In the first example, the cruise ending 18 November 2007 (Fig. 4a), we see that although the core location and speed of the core may agree quite well—differences of only 0.1° latitude and −0.2° longitude and $-0.15\,\mathrm{m\,s^{-1}}$ core speed—the direction of the core can diverge significantly between the two platforms, in this case a separation of ~43°. The cruise ending 31 March 2008 (Fig. 4b) is an example of poor agreement of core latitude, longitude, and speed, but good agreement of core direction. While the *Oleander* core in this case is identified around 37°N, 70.5°W, the GLB-HYCOM places the core at around 39°N, 72°W. Referring to the SSH (Fig. 4b), we note the *Oleander* core location corresponds with the position of the gradient evident in the GLB-HYCOM SSH, and the HYCOM core location corresponds with the northern edge of the eddy just northwest of this boundary. Using a different methodology to define the core (e.g., averaging velocities over a depth representative of the core) might change the identified core locations for either or both of the datasets and may be preferred for other model evaluations, but our choice to use a single depth is intended to avoid any averaging ambiguity that might result from mismatched data depths. Although we use a simple index for Gulf Stream core identification, the analysis reveals that accurately predicting the location and strength of the meandering Gulf Stream core can be challenging when eddies are present. The meandering of the actual Gulf Stream core is a dynamic process, and it varies in both time and space over several scales. Some small intrinsic error can thus be anticipated between the GLB-HYCOM and *Oleander* Gulf Stream core positions for two reasons: 1) the *Oleander* data are finer spatially, as the GLB-HYCOM data resolution is 1/12° longitude (approximately 9 km), whereas the *Oleander* data are typically recorded every 1 or 2 km; and 2) the *Oleander* data are much finer temporally, as the GLB-HYCOM provides instantaneous daily values, whereas the *Oleander* ADCP values are sampled once every few minutes. These sampling differences likely contribute to the differences identified between the GLB-HYCOM and the *Oleander* Gulf Stream core speed and direction.

b. *Ocean racing vessel* Mar Mostro

1) SURFACE CURRENTS

Eastward and northward components of surface currents, u and v, are calculated from *Mar Mostro*'s current magnitude and direction using $u_k = \mathrm{spd}_k \sin(\mathrm{dir}_k)$ and $v_k = \mathrm{spd}_k \cos(\mathrm{dir}_k)$, respectively, where k are *Mar Mostro* data points; u_k and v_k are the zonal and meridional current velocities, respectively, calculated at point k; and spd_k and dir_k are the current magnitude and direction, respectively, measured at point k. Similar to the technique described in section 3a, the magnitude (speed) of the current is calculated from interpolated GLB-HYCOM u and v data using $\mathrm{spd}_k = \sqrt{u_k^2 + v_k^2}$, where k are *Mar Mostro* data points; spd_k is HYCOM current speed, calculated at point k; and u_k and v_k are the HYCOM zonal and meridional current velocities, respectively, bilinearly interpolated to point k.

Box plots comparing *Mar Mostro* u and v and GLB-HYCOM interpolated u and v for the entire race (Figs. 6c and 6d) show greater variability in the *Mar Mostro* data [see also interquartile range (IQR) statistics in Table 3]. This variability is, naturally, mirrored in the *Mar Mostro* surface seawater speeds (Fig. 6b; Table 3). Some portion of this difference is plausibly explained by the difference in temporal sampling of the *Mar Mostro* data (generally at 10-s intervals) versus the once-daily HYCOM fields. Spatially, the GLB-HYCOM data can also be considered effectively "smoothed" as compared to *Mar Mostro* data, since the bilinear interpolation of the HYCOM data relies on the comparatively coarse 1/12° grid points. Additionally, *Mar Mostro*'s Nortek velocity logger, which is used with GPS to determine the current measurements [section 2b(2)], did not sample at a constant depth. When the vessel was upright and at

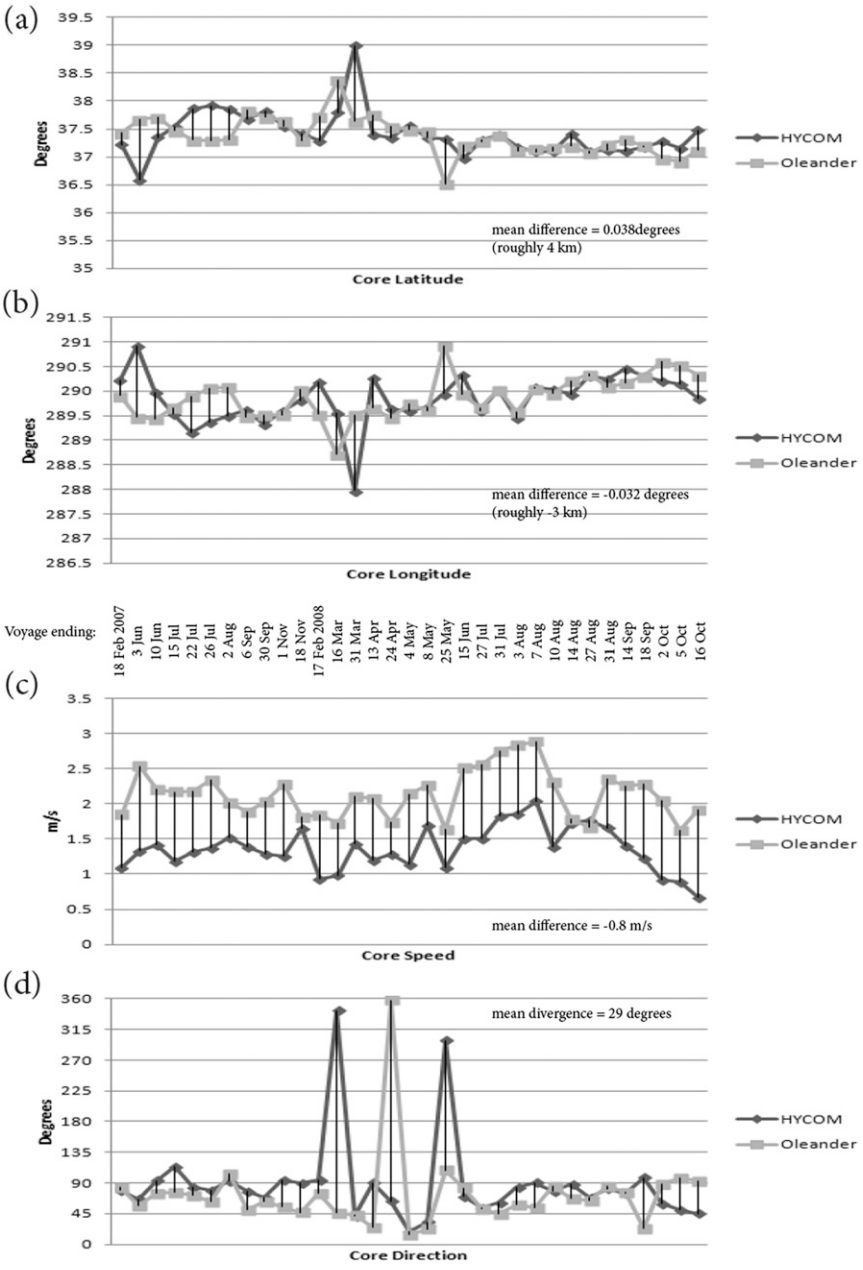

FIG. 5. Gulf Stream core (a) latitude, (b) longitude, (c) speed, and (d) direction identified from the GLB-HYCOM (black) and the *Oleander* (gray) 75-m velocity data for 33 *Oleander* cruises, with the mean differences annotated (for GLB-HYCOM − *Oleander*).

rest, the sensor lay at the 4.5-m depth. When the vessel was sailing at a typical 20° heel, with an additional 40° windward cant, the sensor lay at the 1.8-m depth. On the other hand, the GLB-HYCOM surface velocity data used were calculated for zero depth with results influenced by the model's upper-layer thickness. It follows that this variability in *Mar Mostro* measurement depth likely also contributed to the apparent noisiness of the *Mar Mostro* u and v data, as compared to the HYCOM u and v data. A different approach would have been to apply some factor of normalization, such as adjusting the current measurements to a common depth using known measurement depth data; however, the actual depth of the Nortek on the *Mar Mostro* was not recorded from one sample to the next. Another factor to keep in mind is the flow distortions induced by the hull and bulb of the *Mar Mostro* and their associated wave trains. These factors may contribute to the greater variance noted in

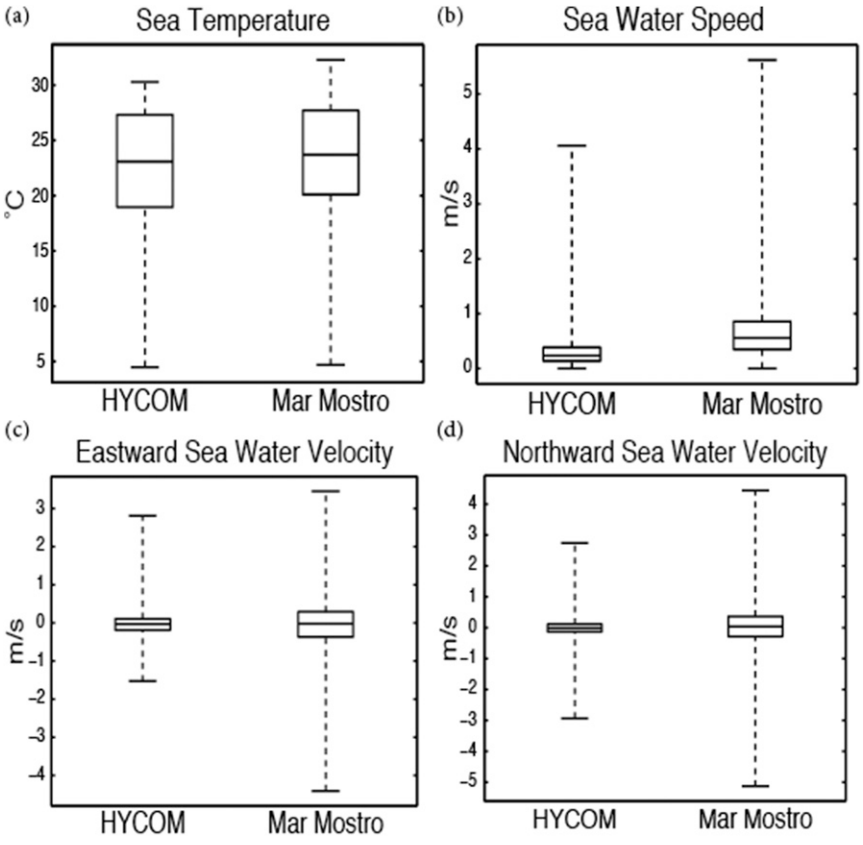

FIG. 6. Box plots for (a) SST, (b) surface seawater speed, (c) surface seawater eastward velocity, and (d) surface seawater northward velocity for the *Mar Mostro* vs the GLB-HYCOM, for the entirety of the *Mar Mostro* VOR 2011–12 race (legs 1–8). Lower/upper box edges represent 25th/75th percentiles. Because efforts were made to remove suspected faulty data in the *Mar Mostro* dataset (see appendix) and the GLB-HYCOM values are assumed to be valid, whiskers represent valid maximums and minimums.

the vessel current measurements as compared to the GLB-HYCOM (Figs. 6b–d).

However, the likelihood exists that not all of the discrepancy between the two platforms is explained by temporal/depth differences and flow distortion. The *Mar Mostro* data also exhibit a much broader range of u and v values overall, as evidenced by the most extreme data points (whiskers in Figs. 6c and 6d). Although speed

TABLE 3. GLB-HYCOM (H) vs *Mar Mostro* (MM) comparison statistics for all data and each leg (with date ranges) for surface u, surface v, spd of surface current, and SST. Negative difference in IQR implies lesser variance in the GLB-HYCOM.

	u (m s^{-1})	v (m s^{-1})	spd (m s^{-1})	SST (°C)
$IQR_H - IQR_{MM}$ (for all data)	−0.36	−0.41	−0.26	0.75
$Median_H - median_{MM}$ (for all data)	−0.01	−0.04	−0.32	−0.61
Mean biases				
All data	0.01	−0.05	−0.37	−0.68
Leg 1 (6–25 Nov 2011)	0.01	0.06	−0.37	−0.69
Leg 2a (11–27 Dec 2011)	0.01	−0.07	−0.41	−0.84
Leg 2b (4 Jan 2012)	−0.18	−0.15	−0.54	−0.70
Leg 3a (14 Jan 2012)	−0.05	−0.19	−0.44	−1.14
Leg 3b (22 Jan–4 Feb 2012)	−0.03	−0.12	−0.30	−0.78
Leg 4 (19 Feb–17 Mar 2012)	−0.08	0.09	−0.42	−0.62
Leg 5 (17 Mar–6 Apr 2012)	0.03	−0.00	−0.32	−0.19
Leg 6 (22 Apr–9 May 2012)	0.10	−0.17	−0.44	−0.68
Leg 7 (20–31 May 2012)	−0.08	−0.26	−0.18	−1.01
Leg 8 (10–14 Jun 2012)	0.35	0.18	−0.47	−1.08

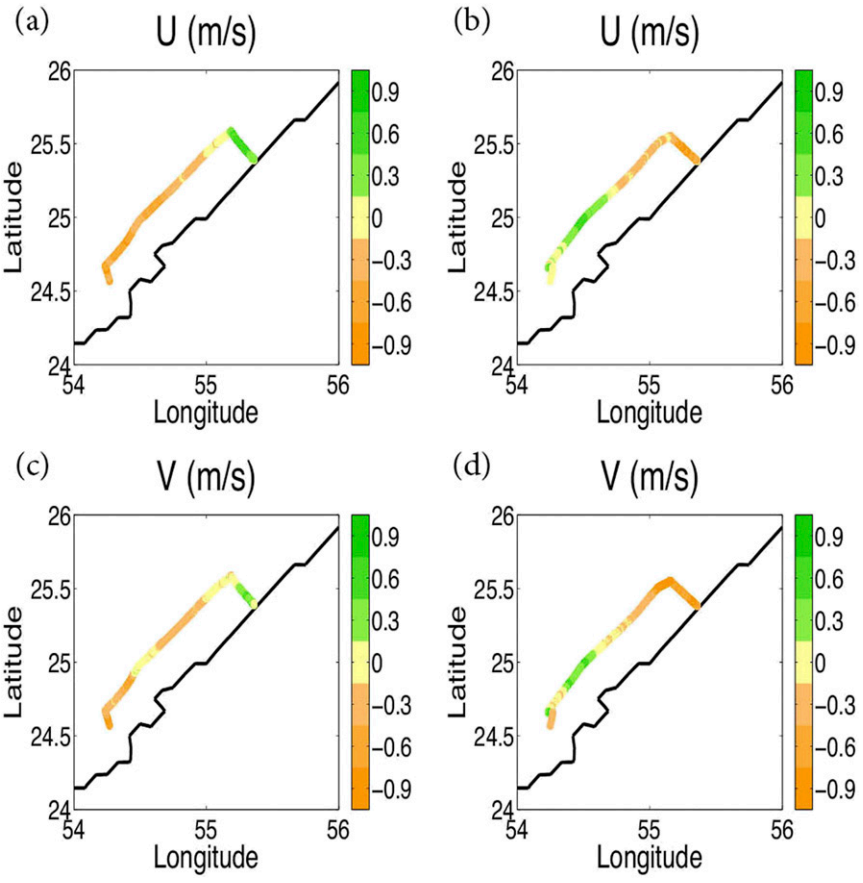

FIG. 7. Difference plots of (GLB-HYCOM − *Mar Mostro*) for (a) leg 2b surface seawater eastward velocity (U), (b) leg 3a surface seawater eastward velocity (U), (c) leg 2b surface seawater northward velocity (V), and (d) leg 3a surface seawater northward velocity (V) plotted along the cruise track.

measurement errors can occur when a vessel is sailing at speeds > 20 kts ($1\,\text{kt} = 0.51\,\text{m s}^{-1}$), as the *Mar Mostro* sometimes did during the VOR, the Nortek speeds were calibrated and checked very closely by team Puma and it was determined that at a sailing speed of 20 kt, the current error was typically about 0.4 kt (roughly $0.2\,\text{m s}^{-1}$) at most (R. Hopkins Jr. 2012, personal communication). During the entire VOR, the *Mar Mostro* was sailing at or above speeds of 20 kt roughly 13% of the time. Although this slight instrumental inaccuracy at high speeds is perhaps factored into the differences shown in maximum current speed between the *Mar Mostro* and the GLB-HYCOM (Fig. 6b), other internal biases in either dataset may also contribute to the differences.

Overall biases in the estimations of u and v by the GLB-HYCOM are small as compared to the *Mar Mostro* (Table 3), even though there is a tendency toward underestimation of current speed along all tracks (with a $-0.37\,\text{m s}^{-1}$ average bias in speed for all legs). Although velocity comparisons along ship tracks are challenging, as there are differences between the spatial and temporal variability captured by model and ship observations, the comparison between GLB-HYCOM and the *Mar Mostro* data does reveal larger biases in some legs.

For example, legs 2b (completed on 4 January 2012) and 3a (completed on 14 January 2012), which are both confined to the southeastern edge of the Persian Gulf (Fig. 1c), have speed biases of $-0.54\,\text{m s}^{-1}$ and $-0.44\,\text{m s}^{-1}$, respectively (Table 3). Examining the velocity differences along the cruise tracks (Fig. 7) reveals notable changes between the legs. Leg 2b (Figs. 7a and 7c) has mostly negative differences (except near the coast), while leg 3b (Figs. 7b and 7d) shows a clear split of between positive and negative differences near the center of the track (at ~25.1°N). The changes, occurring over only 10 days between the legs, may be the result of small-scale variations in Persian Gulf circulation features.

The Persian Gulf has one of the highest salinities globally (~40 psu), with strong seasonal variability, and

some of the speed differences in legs 2b and 3a may result from ocean features more easily defined by salinity. Land inputs of freshwater are limited, and salinity exchanges are controlled through the narrow Strait of Hormuz between the Persian Gulf and the Arabian Sea. Details of the exchanges are thus challenging for a global model to resolve. It is also worth noting that satellite-based salinity measurements are not assimilated into the GLB-HYCOM, as is the current state of the art, but this paradigm is gradually changing as these satellite-based salinity measurements become available. As a result, salinity is much harder to model, especially in regions with limited CTD or Argo observations, and limited observation of the exchanges through the Strait of Hormuz perhaps contributes to the surface current differences noted between legs 2b and 3a and the GLB-HYCOM. Furthermore, the relaxation of sea surface salinity to climatology in the GLB-HYCOM may smooth out short-term gradients, which may be associated with eddies noted in Persian Gulf satellite imagery (Reynolds 1993). More precisely, a basinwide circulation present in the spring and summer months has been shown, in a finescale (~7 km) numerical study using in situ observations for comparison, to dissolve into a network of mesoscale eddies during autumn and winter (Kämpf and Sadrinasab 2006). Although the GLB-HYCOM resolves spatial scales similar to Kämpf and Sadrinasab (2006), accurate prediction of eddy variability in a regional sea would benefit from using a higher-resolution regional model. For example, Thoppil and Hogan (2010) identified mesoscale eddy features in a 1-km regional HYCOM study in the Persian Gulf.

Dynamic ocean regions (e.g., major currents), marginal seas (e.g., the Persian Gulf), and freshwater mixing (e.g., river input to the ocean) are all possible features where evaluation of models using shipboard data may have limitations; however, the examples presented herein do show that shipboard observations can help identify if/where predictions from a global model might be less accurate. These highlighted regions are prime candidates for further analysis for the purposes of observational planning, model evaluation, and/or model fine-tuning or nesting. Additionally, these regions are often traversed by vessels equipped with research-grade meteorological and oceanographic instrumentation.

2) SEA SURFACE TEMPERATURE

Box plots for both the *Mar Mostro* and the interpolated GLB-HYCOM sea surface temperature for the entirety of the VOR (Fig. 6a) suggest several appreciable differences between the two datasets. Although population minimums are roughly equal (4.7°C for *Mar Mostro* data and 4.5°C for GLB-HYCOM data), the *Mar Mostro* population maximum is 2°C greater than that for GLB-HYCOM (32.3° and 30.3°C, respectively). Interquartile range comparison also indicates slightly higher variance in the GLB-HYCOM SST data and an overall shift toward lower values of SST (Table 3). Median values of GLB-HYCOM and *Mar Mostro* SST are 23.1° and 23.7°C, respectively. Since distributions are nonnormal (negatively skewed), a Wilcoxon rank-sum (WRS) test is performed to examine the difference in medians for the SST datasets: a resultant p value too small to represent (i.e., near zero) indicates unequal medians between GLB-HYCOM and *Mar Mostro* SST at the 1% significance level (WRS = 1.088×10^{12}, $\alpha < 0.01$ two tailed). Quantitatively, ~83% of GLB-HYCOM interpolated SST values are cooler than their *Mar Mostro* counterparts, with roughly 28% of those (or roughly 23% of the total data) showing a difference of −1°C or greater. Recalling the variations in sampling depth when the *Mar Mostro* is under sail [see section 3b(1)], it is estimated that, although the sensor was installed 0.5 m below the waterline when the vessel is at rest, data likely represent a mixed sample of the first 0.2 m of the water column while the vessel is sailing. Compare this nominal 0.2-m depth to GLB-HYCOM, wherein the top marine layer in the model was 1 m thick and data were then outputted at standard Levitus depth levels (http://iridl.ldeo.columbia.edu/SOURCES/LEVITUS94/), the topmost of which (0.0 m) was used for this study. The authors anticipate that some portion of the differences in SST are the result of mixing caused by the vessel motion and how mixing is represented in the uppermost model layer, but quantifying the contribution to the differences requires, at minimum, a deeper understanding of the flow along hull of the *Mar Mostro*. When additional information is available (such as along-hull flow statistics), uncertainty in SST biases (whether the result of mixing or other factors) might be quantified and SST values from one of the datasets could be normalized to the other for better model evaluation; however, this would require additional information (e.g., flow model results of the vessel hull) to be included in the analysis. Although such flow modeling is beyond the scope of this study, the authors encourage additional information on vessel instrument siting and flow characteristics to be made available with the SST, and other in situ, observations to refine model-to-data comparisons.

Analysis of the SST biases for each leg consistently shows SST to be cooler in the GLB-HYCOM data as compared to the *Mar Mostro* observations (Table 3). The leg with the smallest bias, leg 5 (17 March 2012–6 April 2012), is notably the only leg primarily in an oceanographically temperate zone; the remaining legs

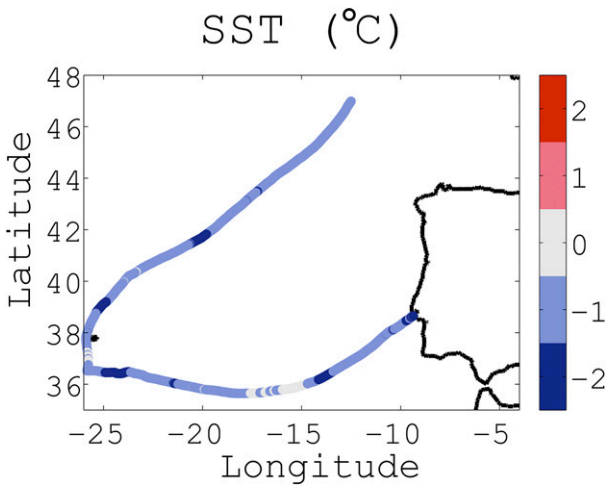

FIG. 8. Difference plot of (GLB-HYCOM − *Mar Mostro*) for SST from leg 8 of the *Mar Mostro* race taken during VOR 2011–12 (plotted along track). Leg 8 originated at Lisbon, Portugal, and officially ended at Lorient, France, but because of domain limitations for the GLB-HYCOM analysis used herein, our analysis omits the terminating portion of this leg.

occur in tropical/subtropical waters. Although limited in scope, the comparison to the *Mar Mostro* data raises the question of whether the GLB-HYCOM physics are better tuned to these temperate waters. The leg with the highest bias, leg 3a, notably occurs in the Persian Gulf, which is hypothesized in the preceding section to be a region where the GLB-HYCOM may not resolve mesoscale features (e.g., eddies). For the remaining two legs with a bias larger than −1—namely, legs 7 (20–31 May 2012) and 8 (10–14 Jun 2012)—geography does not immediately appear to suggest an explanation. This is demonstrated off the coast of Spain (Fig. 8), where only the northernmost of the four apparent clusters of much cooler GLB-HYCOM water (between about 41° and 42° north, showing a difference of around −2°C) appears to be near major ocean currents—the North Atlantic and the Canary—revealing that the coastal–offshore interaction within small spatial scale may pose some challenges for the GLB-HYCOM. Alternatively, the GLB-HYCOM archives are available only at 0000 UTC, which may impose an inherent nighttime cooler signal, when compared with daytime vessel observations. With that caveat in mind, the mean bias is recalculated for leg 8 using only data points that fall inside a 2300–0100 UTC time window, to represent nighttime off the coast of Spain where leg 8 occurred. In this case, mean SST bias lowered from −1.08° to −0.91°C, which is still one of the higher biases of all legs (Table 3) but which difference does at least suggest a mitigation of daytime heating effects on the near-surface ocean. The authors note that another strength of automated vessel data is that it allows users to select the appropriate portion of the diurnal cycle within the data for their model evaluations.

More generally speaking, there is some evidence that discrepancies appear magnified and/or more variable when the vessel was closer to land and/or in marginal seas. The GLB-HYCOM analyses may not be as representative in these regions, or the limited sample we compare may not fully capture the range of variation in SST in these regions. Tides may also factor into the differences in these regions. The authors emphasize that selective temporal comparisons (as shown for leg 8 above) or the use of nested or regional models might improve comparisons to the observed fields. The final case study using research vessel data provides an example of the role regional models play in enclosed or marginal seas.

c. *SAMOS research vessels*

An accurate comparison of the 1-min interval research vessel observations from SAMOS to the salinity data from one of the Gulf of Mexico HYCOM (GoM-HYCOM) model grids requires a precise interpolation in time and space. A 2D linear interpolation [Eq. (1)] is used to match the daily GoM-HYCOM data to the location and time for each available SAMOS salinity observation. SSS typically has a low variance with respect to a single day, so the comparison uses SSS values found in the regional model's 0000 UTC file associated with the date of each ship observation. A complementary comparison that matched ship SSS values to hourly analyzed SSS from the GoM-HYCOM resulted in an average difference between hourly and daily interpolations of only 0.6 psu, so only the comparisons to the daily 0000 UTC model values are shown herein. Although salinity values are reported at a depth of 3–5 m for each research vessel (from the surface down to the seawater intake port), for this comparison the salinity measurements are assumed to be at the surface and no interpolation in the vertical direction is performed. Unknown variations of salinity between intake depth and the surface is another limitation when comparing in situ data to near-surface model values.

Once interpolation is completed, each SAMOS salinity value can be differenced from the corresponding model-predicted values and then the differences can be averaged on a specified grid. In this case, the differences are averaged in half-degree bins, but a different bin size could be used as long as it is coarser than the resolution of the model output being evaluated. The time frame of data used to grid these differences could be on the order of years, or even months, since there is a high frequency of cruises in the northern Gulf of Mexico. In our case

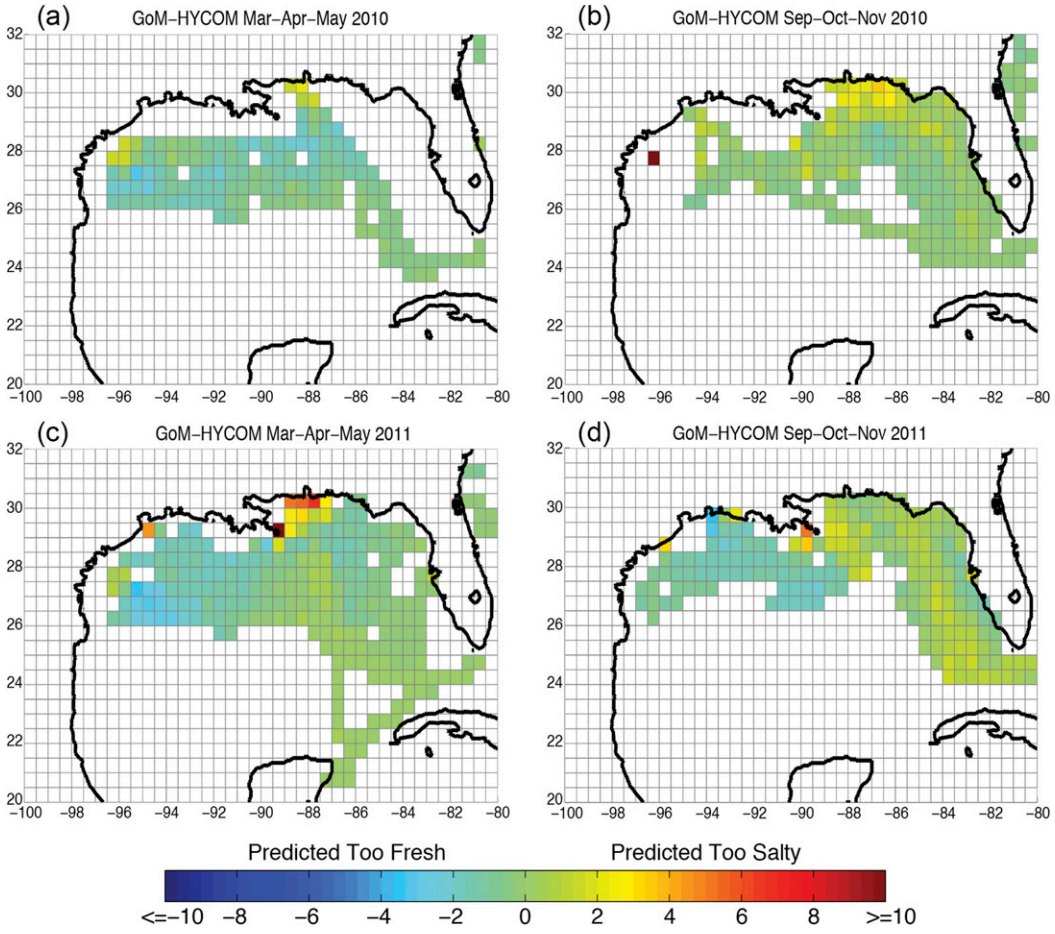

FIG. 9. Bin-averaged (0.5°) SSS differences (psu) between GoM-HYCOM and SAMOS ship observations for (a) spring 2010, (b) fall 2010, (c) spring 2011, and (d) fall 2011. Red values suggest that the model predicts salinity values that are higher than those measured by the ship.

study, the binned SAMOS-to-model differences are first examined for the 1/25° GoM-HYCOM and then for the 1/50° NGoM-HYCOM. Evaluating 2 yr of spring (March–May) and fall (September–November) differences using both models reveals that there is some seasonality in the differences (Figs. 9 and 10). This is consistent with the Mississippi River discharge also displaying seasonality in its streamflow, as river outflow climatologically peaks in the spring and typically reaches a minimum in fall months [see also 10-yr daily discharge time series in Kourafalou and Androulidakis (2013), their Fig. 2].

Both models generally agree very well with the data, as noted by near-zero differences (between −2 and 2 psu in Figs. 9 and 10). This provides a level of confidence in the seasonally averaged model predictions. The highest differences (salinity overprediction by the models) occur to the northeast of the Mississippi River delta in the averaged values for the spring of 2011. Although the difference is only a few salinity units in both models (Figs. 9c and 10c), the result is unexpected, since the two models each have different river inputs and river treatments. Examination of the discharge conditions for this period [see discharge time series in Androulidakis and Kourafalou (2013)] reveals that daily values of Mississippi River discharge (employed by the NGoM model) dropped to about 20 000 $m^3 s^{-1}$ (well below the 10-yr average discharge of 30 000 $m^3 s^{-1}$) in April, before quickly rising to the unprecedented flood peak of close to 45 000 $m^3 s^{-1}$ in May. The low value is actually very close to the climatological monthly mean employed by the GoM model for this particular period. The high salinities resulting from the low Mississippi discharge thus influenced the seasonal average in both models, an effect that was not captured by the research vessel data.

Comparisons over specific cruise tracks are also performed for a more detailed evaluation. An example

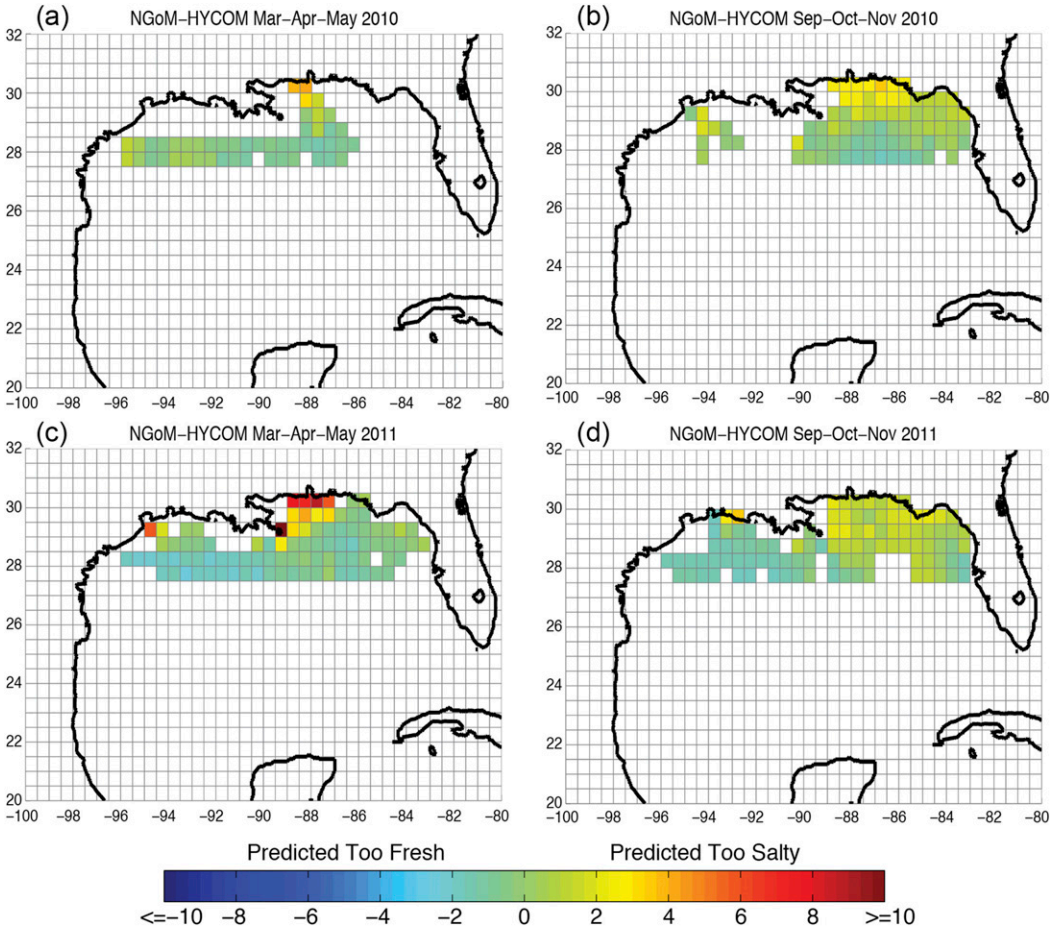

FIG. 10. As in Fig. 9, but differences use NGoM-HYCOM and SAMOS observations. Note that the domain of the NGoM-HYCOM does not extend as far south as the domain of the GoM-HYCOM.

is given in Fig. 11, where three individual cruises provided by SAMOS allow for examination of the regional variations in model performance in the northern Gulf of Mexico (Fig. 11). This examination compares all three HYCOM-based models (each having different horizontal grid resolutions) discussed herein: GLB (1/12°), GoM (1/25°), and NGoM (1/50°) HYCOM. Near the Mississippi River delta (2 October 2011 cruise), the NGoM salinity captures not only the salinity values, but also the short-term variability recorded by the SAMOS research vessel (Fig. 11d). The results shown in Fig. 11 (particularly Fig. 11d) can also be seen in a comparison of a few grid boxes near the Mississippi River mouth in the fall 2011 plots of Figs. 9 and 10. For the western track (Fig. 11c), NGoM again captures the drop in salinity near the coast (indicating the presence of low-salinity waters measured by the ship at this time). In both periods, the coarser models with climatological river inputs and SSS relaxation to climatology differ from the research vessel observations. The eastern track comparison (Fig. 11b) indicates that when salinity does not vary much (samples are outside the Mississippi River plume region), all models are in relative agreement with the SAMOS data. A small drop in salinity at locations near the Florida Panhandle that is not captured by the models is most likely due to local river input (which is climatological for all models; hence, short-term variability is not included).

4. Strengths and limitations

Employing vessel-based ocean observations for model validation requires rigorous examination of the observation method: what is the precision or reliability of the instrument; have independent checks verifying the accuracy of the data been performed; are there any conditions under which the data may be compromised and, if so, what can be done to minimize the effects, etc. For example, most automated underway salinity

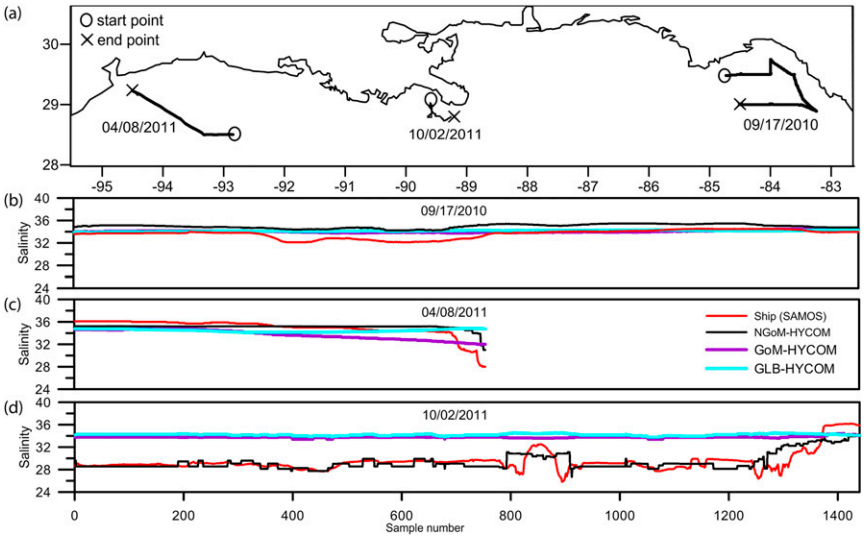

FIG. 11. Individual cruise tracks (a) from SAMOS research vessels from which underway salinity data are compared to the global ($^1/_{12}°$), GoM ($^1/_{25}°$), and NGoM ($^1/_{50}°$) HYCOM for (b) 17 Sep 2010, (c) 8 Apr 2011, and (d) 2 Oct 2011. Sample number represents sequential points along the cruise tracks; the start points and endpoints for the samples are marked by "O" and "X", respectively.

measurements from ships do not undergo postsampling calibration against bottle-salinity samples. When possible, we have provided information on instruments and data quality control for the data used herein [section 2b, section 3b(1), appendix], but metadata on instrument precision or data quality are not available for every vessel. This may limit the accuracy of vessel-to-model comparison in some cases, but vessel measurements are still useful for evaluating variability instead of exact values. Recent programs have focused on improving access not only to the vessel observations but also the enhanced metadata and quality control to meet the accuracy requirements for model evaluation.

Programs focusing on expanding access to underway atmospheric and ocean observations include, but are not limited to, SAMOS (Smith et al. 2012), OceanScope (SCOR/IAPSO 2014), the Joint Archive for Shipboard ADCP (http://ilikai.soest.hawaii.edu/sadcp/main_inv.html), and the Global Ocean Surface Underway Data (http://www.gosud.org/) project. The SAMOS initiative (http://samos.coaps.fsu.edu) provides direct access to all their quality-controlled vessel data and detailed instrumental metadata via web, FTP, and THREDDS services. Additionally, SAMOS data are readily available from the National Centers for Environmental Information (NCEI; Smith et al. 2009). Data from the M/V *Oleander* are accessible online (http://po.msrc.sunysb.edu/Oleander/). It is noted that although Volvo Ocean racing is, at present, no longer supporting high-frequency measurements like those from the *Mar Mostro*, this past experience has shown the value of deploying ocean sensors on future global racing yachts. Additionally, the *Mar Mostro* data used in this analysis are now available to the public via NCEI (Hopkins et al. 2015).

Another limitation of vessel-to-model comparisons is that global model outputs are unlikely to be available on the exact temporal or spatial scale that is represented by ship observations, so some amount of uncertainty will always exist when trying to match a point value from a ship to a gridcell average from a model. Having access to the high temporal sampling from ships allows users to select a representative sample from the ship data to compare to the model and to assess some of the errors in their representations of the model grid using the in situ data. Furthermore, users may wish to consider alternative methods to the bilinear approach used herein to match model and ship data. Examples include the Wilmott skill score (Willmott 1981) and the Pearson coefficient (Pearson 1903) for all data points in a study area, as well as the root-mean-square error for each ship track between in situ and modeled time series (see Androulidakis and Kourafalou 2013; Kourafalou and Androulidakis 2013). In summary, careful selection and processing of both in situ and model datasets can overcome the challenges noted above.

In fact, the availability of vessel data in areas where modelers may be working to resolve key features (i.e., the meandering and eddy-generating Gulf Stream and the isolated Persian Gulf) makes them an ideal component of model validation. Additionally, vessels can go

where moored buoys are scarce (e.g., Southern, South Atlantic, and South Pacific Oceans) and where other in situ observing platforms may not be practical (e.g., profiling floats are rarely used in the Gulf of Mexico or shallower marginal seas). Vessels using automated instrumentation often provide data with very high temporal and spatial resolution, which makes these data ideal for locating any small-scale dynamic ocean features (e.g., salinity and temperature fronts) that one may wish to simulate using numerical models. In the case of research vessels, SAMOS ships also frequently operate along coastlines or on continental shelves, providing detailed observations that can be used to enhance ocean model performance in the challenging transition areas between the open ocean and the coastal zones. It is noted that, with the exception of the M/V *Oleander*, repeat transects by ships over nearly the exact same region are rare, so vessel data may need to be collected over a long sample period to enable drawing conclusions from any model-to-ship comparison. It is also noted that transects for vessels sometimes favor seasonal operations (e.g., SAMOS observations in the Gulf of Mexico; Fig. 2), limiting model evaluation for some periods and features of interest.

Automated vessel observations of the types presented in this manuscript are not associated with operational weather or ocean modeling. As such, these data are not routinely transmitted via the typical data communication channels that feed into operational models (e.g., the Global Telecommunications System) and do not enter the models' data assimilation stream. Therefore, a strength of these high-sampling-rate vessel observations is that they can provide an independent evaluation of data assimilative, predictive (near–real time) numerical models.

5. Conclusions

Three case studies compare automated in situ observations from ocean-going vessels to numerical model output. We employ three different applications of the community HYCOM code (global, Gulf of Mexico, and northern Gulf of Mexico) that have increasingly higher resolution and certain differences in data assimilation and coastal physics. This allows a broad display of our methodology, which is applicable to a wide range of atmospheric and oceanic model outputs. Platforms and parameters that are examined include SST and subsurface currents measured by a merchant vessel, SST and surface ocean currents measured by an ocean racing yacht, and sea surface salinity measured by oceanographic research vessels. Each of the comparisons reveals broad agreements and also differences between the observations and the model-estimated fields for each parameter.

The analyses are presented to demonstrate the high value that automated underway observations from vessels have for evaluating numerical model output. Such measurements can also be incorporated into emerging methodologies for observing-system planning in both regional seas and the open ocean, through observing system simulation experiments (OSSEs; Halliwell et al. 2014). The use of high-resolution ship observations could also be applied to assess model forecast fields as a means to assess forecast errors. The techniques presented are only examples, and alternate methods to match ship to model data (beyond the bilinear approach used herein) could be explored to meet the needs of individual validation projects. Although this study presents some clear similarities and differences between the analyses from various HYCOM model experiments and the vessel data, the authors noted known methodology limitations. This study does not attempt a conclusive evaluation of model performance but offers specific examples over a variety of dynamical regions to demonstrate the data's potential. Taking advantage of high-quality, high-temporal-resolution observations from a variety of vessels using techniques similar to the examples shown herein will provide model developers and users with the tools to evaluate model-derived analyses and forecast products.

Acknowledgments. Support for this work at COAPS is provided by the HYCOM (Grant ONRN00014-09-1-0587) and SAMOS projects. SAMOS is funded by the NOAA Climate Program Office, Climate Observation Division, via the Northern Gulf Institute administered by Mississippi State University (Cooperative Agreement NA11OAR4320199) and the National Science Foundation (NSF) Oceanographic Instrumentation and Technical Services program (Award OCE-0917685). Thanks to K. Donohue and S. Fontana at URI for providing access to the high-quality M/V *Oleander* ADCP data and recommendations on the data's application to Gulf Stream analysis. V. Kourafalou was partially funded by NOAA (Cooperative Agreements NA13OAR4830224; NA10OAR4320143) and NSF (Award OCE-0929651). We thank H. Kang (UM/RSMAS) for the preparation of NGoM-HYCOM model archives.

APPENDIX

Quality Control of *Mar Mostro* Observations

Errors in the *Mar Mostro* dataset necessitate the removal of points by several means. For the case in which either the SST, or both the current speed and direction,

TABLE A1. Number of points used or removed for *Mar Mostro* vs number of defined points for GLB-HYCOM analysis (because of coastline proximity, GLB-HYCOM features undefined data at some *Mar Mostro* lat/lon reference points). Statistics provided for SST, current velocity (csp), and current direction (cdr).

	Leg 1	Leg 2a	Leg 2b	Leg 3a	Leg 3b	Leg 4	Leg 5	Leg 6	Leg 7	Leg 8
Total provided	163 818	134 703	45 876	30 596	113 148	179 430	172 284	147 887	97 277	44 227
Equal to NaN in original file	46	30	0	0	46	40	26	38	6	6
Removed for bad position	27	4	0	0	27	22	5	1	1	6
Removed for SST = 0	0	0	0	0	0	40	589	0	0	0
Removed as spikes (SST)	35	17	0	0	17	40	56	1	7	3
Total SST used *Mar Mostro* (after removals)	163 710	134 652	45 876	30 596	113 058	179 288	171 608	147 847	97 263	44 212
Total SST used HYCOM	160 419	129 006	24 384	17 566	104 205	166 944	166 367	144 664	96 412	36 241
Removed for csp = cdr = 0	427	310	183	185	404	413	836	315	57	54
Removed as spikes (csp/cdr)	1152	458	0	0	257	859	801	0	62	7
Total csp/cdr used *Mar Mostro* (after removals)	162 166	133 901	45 693	30 411	112 414	178 096	170 616	147 533	97 151	44 154
Total csp/cdr used HYCOM	160 419	129 006	24 384	17 566	104 205	166 944	166 367	144 664	96 412	36 241

or both latitude and longitude are exactly zero, the offending data are reassigned as not a number (NaN), effectively removing them from the series so that they will not interfere with any statistical measures or the continuity of plotted data. Once these obviously faulty data points are removed, the SST and current speed data are subjected to a tunable, moving-window, sigma-trimming function to identify more subtle, suspected outliers (spikes). The function, composed by one of the authors, is defined as $\mathrm{dfr}_i = \mathrm{abs}(d_i - \mu_{d\_w}) > \mathrm{num}_\sigma \sigma_{d\_w}$, where dfr_i is a data point determined for removal at series location i if the right-hand side of the equality is true; d_i denotes the data at point i; μ_{d_w} denotes the mean of data d within the user-defined window w; num_σ denotes the user-chosen number of sigmas to use as a threshold for point removal; and σ_{d_w} denotes the standard deviation of data d within (user defined) w. To define w, the function relies on user input declaring the size of desired "windows," or divisions, of the full time series. This choice is typically accomplished by declaring a window size of some percentage times the full sample size, for example, 0.01 × length(data) to accomplish 100 approximately even windows (the probability existed that the final window in the series would not be exactly equal in size to the remainder of the windows because of random sample size). The size of the window and the number of sigmas used to describe the threshold for point removal is tuned in each case until optimal results are realized by visually examining the results. These identified data points are finally reassigned as NaNs, again effectively removing them so they do not interfere with any statistics or data plots. Table A1 summarizes the total number of points used for each leg and includes totals for the number of points removed per leg, per method.

REFERENCES

Androulidakis, Y. S., and V. H. Kourafalou, 2013: On the processes that influence the transport and fate of Mississippi waters under flooding outflow conditions. *Ocean Dyn.*, **63**, 143–164, doi:10.1007/s10236-012-0587-8.

Chassignet, E. P., and Coauthors, 2009: US GODAE: Global ocean prediction with the HYbrid Coordinate Ocean Model (HYCOM). *Oceanography*, **22** (2), 64–75, doi:10.5670/oceanog.2009.39.

Cummings, J., 2005: NCODA status: NRL coupled ocean data assimilation. *Ninth HYCOM Consortium Meeting*, Miami, FL, National Ocean Partnership Program, 13 pp. [Available online at http://hycom.org/attachments/084_J.Cummings(2).pdf.]

Halliwell, G. R., Jr., A. Barth, R. H. Weisberg, P. B. Hogan, O. M. Smedstad, and J. Cummings, 2009: Impact of GODAE products on nested HYCOM simulations on the West Florida Shelf. *Ocean Dyn.*, **59**, 139–155, doi:10.1007/s10236-008-0173-2.

——, A. Srinivasan, V. H. Kourafalou, H. Yang, D. Willey, and M. Le Hénaff, 2014: Rigorous evaluation of a fraternal twin ocean OSSE system for the open Gulf of Mexico. *J. Atmos. Oceanic Technol.*, **31**, 105–130, doi:10.1175/JTECH-D-13-00011.1.

Hodur, R. M., X. Hong, J. D. Doyle, J. Pullen, J. Cummings, P. Martin, and M. A. Rennick, 2002: The Coupled Ocean/Atmosphere Mesoscale Prediction System (COAMPS). *Oceanography*, **15** (1), 88–98, doi:10.5670/oceanog.2002.39.

Hogan, T. F., and T. E. Rosmond, 1991: The description of the Navy Operational Global Atmospheric Prediction System. *Mon. Wea. Rev.*, **119**, 1786–1815, doi:10.1175/1520-0493(1991)119<1786:TDOTNO>2.0.CO;2.

——, and L. R. Brody, 1993: Sensitivity studies of the navy global forecast model parameterizations and evaluations and evaluation of improvements to NOGAPS. *Mon. Wea. Rev.*, **121**, 2373–2395, doi:10.1175/1520-0493(1993)121<2373:SSOTNG>2.0.CO;2.

Hopkins, R., T. Addis, and the Puma Ocean Racing Team, 2015: Sea surface temperature (SST) and surface current data collected from the Mar Mostro during the around-the-world Volvo Ocean Race (VOR) from 2011-11-05 to 2012-07-12,

version 1.1. NCEI Accession 0130694, NOAA National Centers for Environmental Information, accessed 18 August 2015, http://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.nodc:0130694.

Howe, P. J., K. A. Donohue, and D. R. Watts, 2009: Stream-coordinate structure and variability of the Kuroshio Extension. *Deep-Sea Res. I*, **56**, 1093–1116, doi:10.1016/j.dsr.2009.03.007.

Kämpf, J., and M. Sadrinasab, 2006: The circulation of the Persian Gulf: A numerical study. *Ocean Sci.*, **2**, 27–41, doi:10.5194/os-2-27-2006.

Kourafalou, V. H., and Y. S. Androulidakis, 2013: Influence of Mississippi River induced circulation on the Deepwater Horizon oil spill transport. *J. Geophys. Res. Oceans*, **118**, 3823–3842, doi:10.1002/jgrc.20272.

——, G. Peng, H. Kang, P. J. Hogan, O. M. Smedstad, and R. H. Weisberg, 2009: Evaluation of Global Ocean Data Assimilation Experiment products on South Florida nested simulations with the Hybrid Coordinate Ocean Model. *Ocean Dyn.*, **59**, 47–66, doi:10.1007/s10236-008-0160-7.

Meinen, C. S., and D. S. Luther, 2003: Comparison of methods of estimating mean synoptic structure in ''stream coordinates'' reference frames with an example from the Antarctic Circumpolar Current. *Deep-Sea Res. I*, **50**, 201–220, doi:10.1016/S0967-0637(02)00168-1.

——, ——, and M. O. Baringer, 2009: Structure, transport, and potential vorticity of the Gulf Stream at 68°W: Revisiting older data sets with new techniques. *Deep-Sea Res. I*, **56**, 41–60, doi:10.1016/j.dsr.2008.07.010.

Pearson, K., 1903: Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Philos. Trans. Roy. Soc. London*, **200**, 1–66, doi:10.1098/rsta.1903.0001.

Prasad, T. G., and P. J. Hogan, 2007: Upper-ocean response to Hurricane Ivan in a 1/25° nested Gulf of Mexico HYCOM. *J. Geophys. Res.*, **112**, C04013, doi:10.1029/2006JC003695.

Reynolds, R. M., 1993: Physical oceanography of the Persian Gulf, Strait of Hormuz, and the Gulf of Oman—Results from the *Mt. Mitchell* expedition. *Mar. Pollut. Bull.*, **27**, 35–59, doi:10.1016/0025-326X(93)90007-7.

Rosmond, T. E., 1992: The design and testing of the Navy Operational Global and Atmospheric Prediction System. *Wea. Forecasting*, **7**, 262–272, doi:10.1175/1520-0434(1992)007<0262:TDATOT>2.0.CO;2.

Rossby, T., and E. Gottlieb, 1998: The Oleander Project: Monitoring the variability of the Gulf Stream and adjacent waters between New Jersey and Bermuda. *Bull. Amer. Meteor. Soc.*, **79**, 5–18, doi:10.1175/1520-0477(1998)079<0005:TOPMTV>2.0.CO;2.

Schiller, R. V., and V. H. Kourafalou, 2010: Modeling river plume dynamics with the HYbrid Coordinate Ocean Model. *Ocean Modell.*, **33**, 101–117, doi:10.1016/j.ocemod.2009.12.005.

——, ——, P. Hogan, and N. D. Walker, 2011: The dynamics of the Mississippi River plume: Impact of topography, wind and offshore forcing on the fate of plume waters. *J. Geophys. Res.*, **116**, C06029, doi:10.1029/2010JC006883.

SCOR/IAPSO, 2014: OceanScope: A proposed partnership between the maritime industries and the ocean observing community to monitor the global ocean water column. Rep. of SCOR/IAPSO Working Group 133, 86 pp. [Available online at http://www.scor-int.org/Publications/OceanScope_Final_report.pdf.]

Scott, R. B., B. K. Arbic, E. P. Chassignet, A. C. Coward, M. Maltrud, W. J. Merryfield, A. Srinivasan, and A. Varghese, 2010: Total kinetic energy in four global eddying ocean circulation models and over 5000 current meter records. *Ocean Modell.*, **32**, 157–169, doi:10.1016/j.ocemod.2010.01.005.

Smith, S. R., D. M. Legler, and K. V. Verzone, 2001: Quantifying uncertainties in NCEP reanalyses using high-quality research vessel observations. *J. Climate*, **14**, 4062–4072, doi:10.1175/1520-0442(2001)014<4062:QUINRU>2.0.CO;2.

——, J. J. Rolph, K. Briggs, and M. A. Bourassa, 2009: Quality-controlled underway oceanographic and meteorological data from the Center for Ocean-Atmospheric Predictions Center (COAPS)—Shipboard Automated Meteorological and Oceanographic System (SAMOS). NOAA/National Oceanographic Data Center, accessed 15 November 2015, doi:10.7289/V5QJ7F8R.

——, P. J. Hughes, and M. A. Bourassa, 2011: A comparison of nine monthly air–sea flux products. *Int. J. Climatol.*, **31**, 1002–1027, doi:10.1002/joc.2225.

——, M. A. Bourassa, and D. L. Jackson, 2012: Supporting satellite research with data collected by vessels. *Sea Technol.*, **53**, 21–24.

Sturges, W., and A. Bozec, 2013: A puzzling disagreement between observations and numerical models in the central Gulf of Mexico. *J. Phys. Oceanogr.*, **43**, 2673–2681, doi:10.1175/JPO-D-13-081.1.

Thoppil, P. G., and P. J. Hogan, 2010: A modeling study of circulation and eddies in the Persian Gulf. *J. Phys. Oceanogr.*, **40**, 2122–2134, doi:10.1175/2010JPO4227.1.

Willmott, C. J., 1981: On the validation of models. *Phys. Geogr.*, **2**, 184–194, doi:10.1080/02723646.1981.10642213.