Verification of a Multimodel Storm Surge Ensemble around New York City and Long Island for the Cool Season

TOM DI LIBERTO* AND BRIAN A. COLLE

School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York

NICKITAS GEORGAS AND ALAN F. BLUMBERG

Stevens Institute of Technology, Hoboken, New Jersey

ARTHUR A. TAYLOR

Meteorological Development Laboratory, NOAA/NWS, Office of Science and Technology, Silver Spring, Maryland

(Manuscript received 21 November 2010, in final form 27 June 2011)

ABSTRACT

Three real-time storm surge forecasting systems [the eight-member Stony Brook ensemble (SBSS), the Stevens Institute of Technology's New York Harbor Observing and Prediction System (SIT-NYHOPS), and the NOAA Extratropical Storm Surge (NOAA-ET) model] are verified for 74 available days during the 2007–08 and 2008–09 cool seasons for five stations around the New York City–Long Island region. For the raw storm surge forecasts, the SIT-NYHOPS model has the lowest root-mean-square errors (RMSEs) on average, while the NOAA-ET has the largest RMSEs after hour 24 as a result of a relatively large negative surge bias. The SIT-NYHOPS and SBSS also have a slight negative surge bias after hour 24. Many of the underpredicted surges in the SBSS ensemble are associated with large waves at an offshore buoy, thus illustrating the potential importance of nearshore wave breaking (radiation stresses) on the surge predictions. A bias correction using the last 5 days of predictions (BC) removes most of the surge bias in the NOAA-ET model, with the NOAA-ET-BC having a similar level of accuracy as the SIT-NYHOPS-BC for positive surges. A multimodel surge ensemble (ENS-3) comprising the SBSS control member, SIT-NYHOPS, and NOAA-ET models has a better degree of deterministic accuracy than any individual member. Probabilistically, the ALL ensemble (eight SBSS members, SIT-NYHOPS, and NOAA-ET) is underdispersed and does not improve after applying a bias correction. The ENS-3 improves the Brier skill score (BSS) relative to the best deterministic member (SIT-NYHOPS), and the ENS-3 has a larger BSS and better reliability than the SBSS and ALL ensembles, thus illustrating the benefits of a multimodel storm surge ensemble.

1. Introduction

a. Background

Storm surge is a major hazard for those coastal areas exposed to tropical and/or extratropical cyclones. Storm surge is defined as the rise of seawater above the astronomical tide prediction as a result of surface winds around the storm and relatively low surface pressure (Glickman 2000), with winds having the largest impact on the surge. About 260 km^2 of the New York City (NYC), New York, area is at risk for storm surge flooding by a 100-yr storm event (Bowman et al. 2005), so it is important to accurately forecast these surge events. A surge of 0.6 m can cause minor flooding during a high tide at Battery Park in NYC, and thus this threshold often leads to a Coastal Flood Advisory being issued by the National Weather Service around south Manhattan (Colle et al. 2010). However, surges as low as 0.2–0.3 m can cause coastal flood advisory conditions, especially during a spring high tide in vulnerable locations such as the back bays of the south

[*] Current affiliation: NOAA/CPC/Wyle, Camp Springs, Maryland.

Corresponding author address: Dr. Brian A. Colle, School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY 11794-5000.
E-mail: brian.colle@stonybrook.edu

DOI: 10.1175/WAF-D-10-05055.1

shore of Long Island, western Long Island Sound, and southern Queens (J. S. Tongue 2011, NYC National Weather Service, personal communication).

There are a number of ocean models used to forecast storm surge in the coastal zone. Historically, most have been applied to landfalling hurricanes. For example, the National Weather Service (NWS) utilizes the Sea, Lake and Overland Surges (SLOSH) model developed by the Meteorological Development Laboratory of the NWS (Jelesnianski et al. 1992) for its hurricane surge predictions. SLOSH ingests surface pressure, size of the storm, surface winds, cyclone track, and forward speed to estimate storm surge heights for one of approximately 40 SLOSH basins along the U.S. east coast, including the New York City metropolitan region.

Other ocean models have also been used to simulate storm surge events for tropical cyclones. For example, Shen et al. (2005) used the Advanced Three-Dimensional Circulation Model for Coastal Ocean Hydrodynamics (ADCIRC; Westerink et al. 1993) to hindcast the surge for Hurricane Isabel (2003) in the Chesapeake Bay. Utilizing a simplified wind model similar to SLOSH to calculate the atmospheric forcing for ADCIRC, the authors found that the simulated surge (1.9–2.5 m) was within 0.3 m of the observed, with an RMS error during the event of 0.19 m. Westerink et al. (2008) used ADCIRC to successfully hindcast the surges associated with Hurricanes Betsy (1965) and Andrew (1992) in southern Louisiana, and the model was within 10% at most gauge stations. Colle et al. (2008) used ADCIRC forced by a mesoscale meteorological model (the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model, MM5) at 12-km grid spacing to realistically simulate the surge from Tropical Storm Floyd (1999) impacting the NYC-Long Island (NYC-LI) region to within ~0.1 m of the observed. Weisberg and Zheng (2006) used the Finite Volume Coastal Ocean Model (FVCOM) in the Tampa Bay, Florida, area and found that the storm surge for that region is sensitive to storm track and storm speed, with slower-moving storms having twice the storm surge of faster-moving cyclones.

Storm surge from nontropical (extratropical) cyclones generates different challenges than those from hurricanes. The assumed parametric wind field in SLOSH, which is perturbed based on the forward motion of the storm, is less accurate due to large asymmetries in the surface wind and pressure fields for midlatitude cyclones and tropical storms that undergo an extratropical transition (Atallah and Bosart 2003; Colle 2003). As a result, the NWS developed the Extratropical Storm Surge (NOAA-ET) model during the mid-1990s (Burroughs and Shaffer 1997, Blier et al. 1997). The NOAA-ET model is a vertically averaged barotropic ocean model that includes a specified bottom friction (Tilburg and Garvine 2004), similar to the SLOSH model. The NOAA-ET model is forced with the surface wind and pressure fields from the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) model. Also, in addition to the U.S. west coast, the Gulf of Mexico, and Alaska, the NOAA-ET model is run along the U.S. east coast on a large grid that covers much of the western Atlantic and for about 4 days, since extratropical surge events can persist at a location for more than 24 h (Blier et al. 1997). Tilburg and Garvine (2004) verified that the NOAA-ET model and a simple linear regression model for surge forecasts at Atlantic City, New Jersey, from 1997 to 1998. They found that the short-term (5 day) anomaly-corrected NOAA-ET model explained 79% of the total observed subtidal frequency of the water level as compared to the 74% of the anomaly-corrected regression model. Blier et al. (1997) used the NOAA-ET model along with winds and sea level pressure from the GFS analyses to hindcast storm surges along the southern coast of Alaska. They found that the model realistically predicted the timing of the surge in a longer-duration event (>24 h) (6 October 1992) and the magnitude to within 0.8 m, although for a short-duration surge event (~12 h) (20 August 1993) the model did not have a peak surge, failing to capture the surge event at all and underpredicting the observed surge by ~1.0 m. This deficiency may have been from the low temporal and spatial resolutions of the GFS wind fields, which could not resolve the coastally enhanced winds along the Alaskan coast.

Over the course of several years, Blumberg et al. (1999) used a shallow-water derivative of the Princeton Ocean Model (POM) to forecast storm surge in the New York Bight. They found that the mean error in the total water level was ~11% of the local maximum range and the mean correlation coefficient between the data and model was 0.94 from October 1994 to September 1995.

b. Motivation

There has been limited evaluation of operational storm surge models for coastal southern New England and New York. Colle et al. (2008) successfully simulated the surge at the Battery in NYC to within 5%–10% of the observed from the December 1992 nor'easter using ADCIRC forced with 12-km MM5 observed winds. Although these results are encouraging, existing storm surge models need to be verified for more events. In addition, such storm surge hindcasts can be more accurate than real-time storm surge forecasts if the hindcasts are based on models forced with analysis winds, pressures, etc. With the relatively large NYC-LI population and a complex coastal geometry prone to storm surge flooding, it is important to investigate the

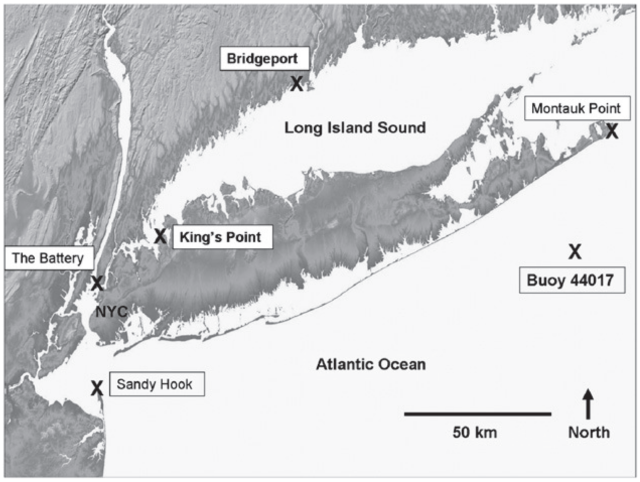

FIG. 1. Spatial map showing the locations of the five water level sites around Long Island used for storm surge verification.

skill of storm surge forecast models over a longer verification period than just a few surge events.

Although it is important that these storm surge models can predict the severe flooding events, it is first useful to compare the models against all types of events. There are not enough moderate coastal flooding events in the last several years at NYC to verify the models with statistical significance, but this study will include a large sample of storm surges to 0.4 m, which can cause problems during a high tide around parts of NYC and Long Island.

There are currently several modeling systems from various institutions [NOAA-ET, the Stevens Institute of Technology's New York Harbor Observing and Prediction System (SIT-NYHOPS), and Stony Brook University] that forecast storm surge in real time for the NYC-LI region; however, there has not been any formal intercomparison of these models. Also, the Stony Brook storm surge (SBSS) system is run in an ensemble (eight member) configuration, so the benefits of the SBSS need to be quantified over using any single model as well as using all three (multimodel) operational models together. An ensemble with combined physics and initial condition perturbations has been shown to improve atmospheric predictions as compared to the individual members on average (Eckel and Mass 2005). In addition, a multimodel atmospheric ensemble can further improve the forecast over an ensemble of just one model (Mylne et al. 2002; Woodcock and Engel 2005). Thus, it is hypothesized that an ensemble of storm surge models using various atmospheric models as forcing should perform better on average than a single atmospheric model used as forcing for a single ocean model.

This paper will address the following questions:

- How well do the ADCIRC, SIT-NYHOPS, and NOAA-ET models predict storm surge during one-and-a-half cool seasons around NYC-LI, which includes several extratropical cyclone events?
- What is the potential contribution to surge errors from the simulated atmospheric forcing as compared to uncertainties between surge models?
- What is the potential benefit of using a multimodel ensemble versus a single model for storm surge prediction?

Section 2 describes the data and methods used in the analysis. The cool-season deterministic verification of the ADCIRC, NOAA-ET, and SIT-NYHOPS forecast systems and the ensemble of all three models are presented in section 3, while section 4 summarizes the probabilistic verification. Summary and conclusions are highlighted in section 5.

2. Data and methods

a. Storm surge modeling systems

Three operational, real-time modeling systems were evaluated around NYC-LI over one-and-a-half cool-season periods during which the forecasts from the modeling systems were available (74 forecasts from November 2007 to March 2008 and from October 2008 to December 2008). Operational storm surge forecasts were obtained from the SBSS, SIT-NYHOPS, and NOAA-ET modeling systems. Figure 1 shows the five stations around NYC-LI used in the verification analysis. Table 1 lists the important configurations for the three modeling systems.

The ADCIRC model used by the SBSS simulates water elevation and currents for the domain along the U.S. east coast to offshore over the continental shelf (Fig. 2a) by solving the generalized wave continuity equation. It was run in a two-dimensional (barotropic) configuration on a grid that has triangular finite elements (~108 000 nodes) ranging from a 70-km grid-node separation offshore to about 10 m around parts of Long Island and NYC (Colle et al. 2008), with the highest resolution around the coast, specifically New York Harbor. The *M2*, *K1*, *O1*, *N2*, and *S2* tidal constituents obtained from an ADCIRC global model are applied along the boundary during the entire run, as in Westerink et al. (1993).

The SBSS ensemble surge system consists of eight separate ADCIRC members run every 0000 UTC to hour 48 with the surface wind and pressure from either the Weather Research and Forecasting Model (WRF; Skamarock et al. 2005) or the MM5 (Grell et al. 1995). The MM5 and WRF are run with two domain resolutions (36- and 12-km grid spacing) starting each day at 0000 UTC. The 36-km grid covers from the Rocky

TABLE 1. Description of three storm surge forecasting systems for the NYC-LI region.

	Storm surge forecasting systems		
Institution	Atmospheric forcing	Ocean model	Start time (UTC)
Stony Brook	Five MM5 and three WRF members	ADCIRC	0000
Stevens Institute of Technology	NCEP-NAM model	ECOM	0500
NOAA	NCEP-GFS model	NOAA-ET model	0000

Mountains to well off the U.S. east coast (Jones et al. 2007), which provides boundary conditions for a one-way nested 12-km grid covering the northeast United States. The 12-km grid is used to force the ADCIRC model (Fig. 2a). The ADCIRC members are ''hot started'' each day by using the previous day's 24-h forecast as the initial water level for the new forecast in order to reduce spinup issues; however, observations are not assimilated to improve this first guess given the general lack of water level observations across the domain. To avoid model errors continually being used as initial conditions, the ensemble is cold started approximately every 2 weeks, during which time there is a 1-day spinup period with tidal forcing before applying the MM5 or WRF winds. The five MM5 members and three WRF members used in the ensemble are run with varying initial conditions and model physics (Table 2). Four different atmospheric analyses are used for the initial and boundary conditions [the Nonhydrostatic Mesoscale Model component of the WRF (WRF-NMM), the Navy Operational Global Atmospheric Prediction System (NOGAPS), GFS, and the Canadian Global Model], four different microphysics schemes [Simple Ice (Dudhia 1989), Reisner (Reisner et al. 1998), Ferrier (Ferrier et al. 2002), and the WRF single-moment three-class (WSM3; Hong et al. 2004)], four different planetary boundary layer (PBL) schemes [the Medium-Range Forecast (MRF; Hong and Pan 1996), Mellor–Yamada (Janjić 2002), Blackadar (Zhang and Anthes 1982), and Yonsei University (YSU; Hong et al. 2006)], two radiation packages [for cloud radiation, the Chemistry–Climate Models (CCM2; Hack et al. 1993), and the Rapid Radiative Transfer Model (RRTM; Mlawer et al. 1997)], and three different convective parameterizations [Grell (Grell 1993), Betts–Miller (Betts and Miller 1993), and Kain–Fritsch (Kain 2004)]. The first member (9a) in Table 2 is the control member for the SBSS ensemble in the subsequent discussion, since it uses the MM5 member that has been run for several years over the northeast United States (Colle et al. 2003; Jones et al. 2007).

The SIT-NYHOPS hydrodynamic model used for this paper comprises the forecasting module of NYHOPS (Bruno et al. 2006; Georgas 2010). The model is based on an updated version of the Estuarine, Coastal, and Ocean Model (ECOM), itself a derivative of POM. ECOM is a three-dimensional, time-dependent model that solves the primitive shallow-water equations regarding the conservation of mass, momentum, heat, and salt (Blumberg et al. 1999). The version of the model used in NYHOPS is driven by tides, surface wind, surface heat fluxes, and river and stream inflows, as well as freshwater and heat fluxes from 280 point sources [sewage treatment and thermal power plants; Georgas (2010)]. The 3-hourly forecasts by the 12-km resolution version of the NCEP North American Model (NAM) are used for the surface wind stress. The domain extends from the coast of Maryland in the south to Nantucket, Massachusetts, to the north and includes the Hudson River up to the Troy Dam (Fig. 2b). The resolution of the finite-difference numerical grid runs from 50 m inside the rivers to 11 km at the southern edge of the offshore boundary. Tidal forcing is provided at the offshore boundary by the *M2*, *S2*, *N2*, *K2*, *O1*, *K1*, *Q1*, *M4*, and *M6* tide and overtide constituents extracted from the East Coast 2001 database (Mukai et al. 2002). The SIT-NYHOPS system is run in a three-dimensional configuration, with 10 sigma levels scaled to the local water column depth. In forecast mode, the model performs a 72-h simulation starting each day at midnight eastern standard time (EST; 0500 UTC). This includes a 24-h hindcast of the previous day by using the 6-hourly NAM analyses (Bruno et al. 2006; Georgas 2010).

The NOAA-ET model is based on the same quasi-linear depth-integrated shallow-water equations as the SLOSH model (Burroughs and Shaffer 1997; Jelesnianski et al. 1992). However, the NOAA-ET simulations do not use a parametric wind model as in SLOSH, but rather the GFS model for hourly wind and pressure forcing. Figure 2c shows the model domain. Unlike the SBSS and SIT-NYHOPS modeling systems, the NOAA-ET model only predicts storm surge and does not include tides (Burroughs and Shaffer 1997). The NOAA-ET model uses a reference sea level to determine the storm surge produced with the surface wind and pressure forcing. The storm surge levels are then added to the National Ocean Service (NOS) predicted tides above the North American Vertical Datum of 1988 (NAVD-88; Zilkoski et al. 1992) to obtain the total water levels.

Observed storm surge levels were determined by subtracting the predicted astronomical tide from the total

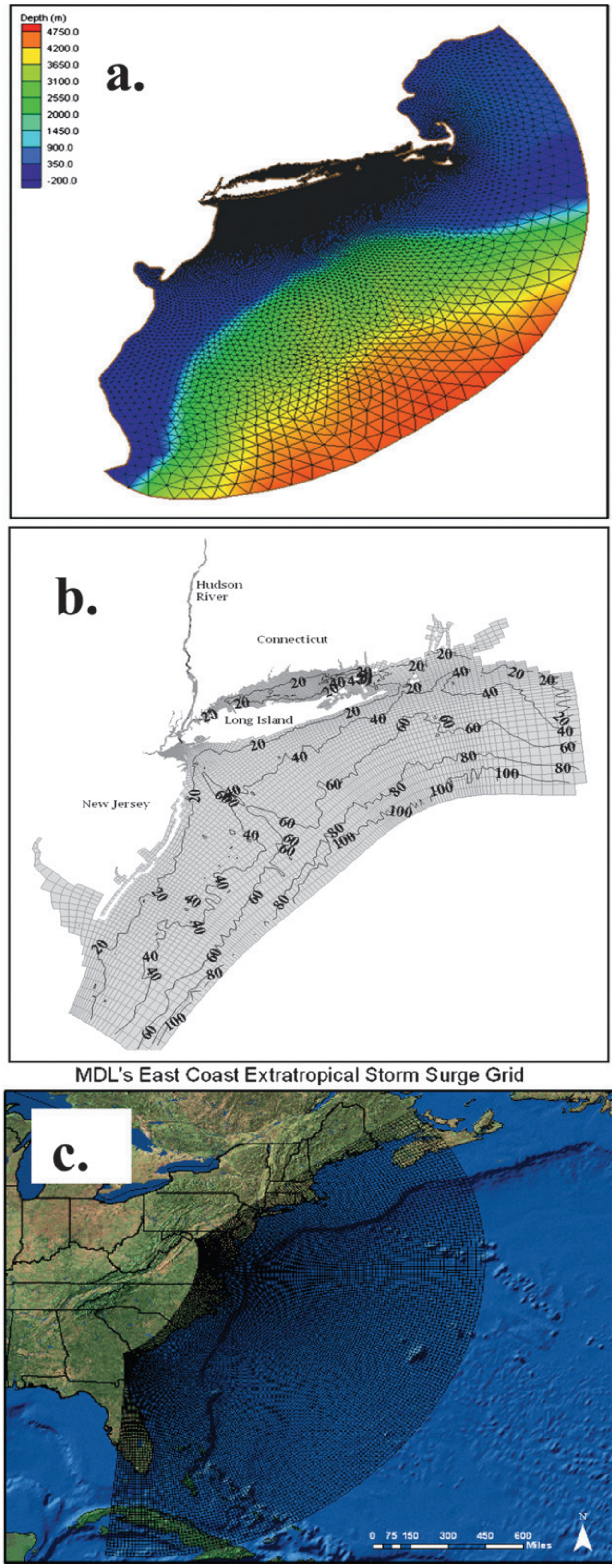

FIG. 2. (a) Domain used in the SBSS real-time storm surge forecasting ensemble. (b) The grid used by SIT-NYHOPS (from Georgas and Blumberg 2010). (c) The NOAA-ET model East Coast grid.

(measured) water levels at each station. The tide was derived from the T-tide program (Pawlowicz et al. 2002), which performs a harmonic analysis on the observed water level: in this study the November 2007–March 2008/October 2008–December 2008 NOAA observed water levels at each individual station around NYC-LI. The derived tides were used, since the astronomical tides provided by NOAA are referenced to the 1983–2001 National Tidal Datum Epoch, and there are other tidal variations from dredging and other shoreline changes.

b. Postprocessing and verification approach

Several error metrics were used to determine the accuracy of each modeling system based on five stations around NYC-LI (Fig. 1). The RMSE in Eq. (1) and the mean error [ME in Eq. (2)'s "bias"] were used in the deterministic verification:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2} \quad \text{and} \tag{1}$$

$$\mathrm{ME} = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k), \tag{2}$$

where y_k is the forecasted value, o_k is the observed value at the same time, and n is the total number of values. A forecast time was included in the ME and RMSE calculations if either the observation or any ensemble member (or average of select members for ensemble mean) exceeded a given threshold.

To determine how well the SBSS ensemble system and all models together performed probabilistically, the Brier score (BS) and Brier skill score (BSS) were calculated from

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{n}(t_k - w_k)^2 \tag{3}$$

and

$$\mathrm{BSS} = 1 - \left(\frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}\right), \tag{4}$$

where t_k refers to the probability forecast, w_k refers to whether the event (observed exceeding a particular threshold at a station) occurred ($w_k = 1$) or not ($w_k = 0$), and $\mathrm{BS}_{\mathrm{ref}}$ is the reference BS to compare the ensemble with, which is typically climatology or some other model. In essence, the Brier score is the mean square error of the probability forecasts (Wilks 2006). Furthermore, the Brier score was separated into reliability (REL), resolution

TABLE 2. Description of the atmospheric SBSS ensemble members, including models used, microphysical schemes, PBL schemes, radiation schemes, cumulus schemes, and initial conditions.

			SBSS model atmospheric ensemble members			
Members	Model	Microphysics	PBL scheme	Radiation	Cumulus	Initial conditions
9a	MM5	Simple Ice	MRF	Cloud radiation	Grell	WRF-NMM
BMMY	MM5	Simple Ice	MY	CCM2	Betts–Miller	GFS
GRBLK	MM5	Simple Ice	Blackadar	CCM2	Grell	NOGAPS
K2MRF	MM5	Reisner	MRF	Cloud radiation	Kain–Fritsch	GFS
K2MY	MM5	Simple Ice	MY	CCM2	Kain–Fritsch	Canadian model
221	WRF	Ferrier	YSU	RRTM	Kain–Fritsch	WRF-NMM
GFS	WRF	Ferrier	YSU	RRTM	Grell	GFS
NOG	WRF	WSM3	YSU	RRTM	Betts–Miller	NOGAPS

(RES), and uncertainty (UNC), where BS = REL − RES + UNC (Wilks 2006). Rank histograms (Talagrand diagrams) were also calculated for several stations across the New York metropolitan region to evaluate the dispersion of the ensemble members (as in Wilks 2006). Small random noise (∼0.02 m) was added to the observed surges in order to represent the uncertainty in the observed values in these diagrams (Hamill 2001).

In addition to evaluating the individual storm surge systems (SBSS, SIT-NYHOPS, and NOAA-ET), two ensembles were created by using all surge models and the SBSS ensemble (multimodel or the ALL ensemble), as well as a three-member ensemble of storm surge models (ENS-3): the SBSS control member and the NOAA-ET and SIT-NYHOPS models. The surge verification was hourly to be consistent with the availability of the NOAA-ET model. Also, since the SIT-NYHOPS began its forecasts at 0500 UTC, while SBSS and NOAA-ET began at 0000 UTC, all models were compared for the same forecast hour (i.e., 1–48 h) rather than the same time of day (i.e., 0600–0600 UTC). As a result, during the first 5 h (0000–0500 UTC), the ALL and ENS-3 ensembles consisted of only the SBSS and NOAA-ET members. The models were also only intercompared for 74 days in which all three models had data (listed in DiLiberto 2009). The verification was done for five stations across NYC-LI: The Battery, NYC; Sandy Hook, New Jersey; King's Point, New York; Bridgeport, Connecticut; and Montauk Point, New York see Fig. 1).

The members in the three ensemble systems were "bias corrected" with the approach used for the operational NOAA-ET. For each member, the predicted surge is adjusted by averaging the surge error for the previous 5 days of 1–24-h forecasts and subtracting this error from the latest prediction [an "anomaly correction," as described in Tilburg and Garvine (2004)]. A similar approach has been applied to atmospheric ensembles using the previous 7–14 days of forecasts (Eckel and Mass 2005; Jones et al. 2007). Although this 5-day approach may not necessarily improve the forecast, this adjustment will still be referred to as a bias correction given that some forecasts/members are improved. The ALL–BC and ENS-3–BC ensembles verified below include members after applying the 5-day bias correction.

Finally, in order to test for statistical significance, a bootstrapping approach was used to resample the data and obtain proper confidence intervals around the means (Zwiers 1990). For each parameter (e.g., RMSE of surge error), a new sample of the same size was obtained by randomly selecting from the original sample and allowing for repeated selections. The mean was calculated and this process was repeated 1000 times. The 90% confidence intervals around the mean were determined by finding the 5th and 95th percentiles of the means of all 1000 resamples. If the confidence intervals of two particular samples did not overlap, then they were considered to be significantly different at the 90% level.

3. Deterministic verification results

a. No bias correction

The surge predictions were binned in 12-h intervals from 1 to 48 h in order to increase the sample size. Figure 3a shows the ME for storm surge versus forecast hour averaged for all five stations. During the 1–12-h forecast, there is little bias (ME) in the SIT-NYHOPS results, a slight negative bias (−0.03 to −0.04 m) in the SBSS control member (SBSS CTL), and a larger negative bias (0.08–0.09 m) in the raw NOAA-ET data. The SBSS CTL ME has little bias by 13–24 h and is similar to the SIT, while the raw NOAA-ET negative bias becomes slight larger (−0.10 m). The SIT-NYHOPS and SBSS simulations develop slight negative biases (−0.03 to −0.04 m) by 25–48 h of the forecast, while the NOAA-ET negative bias increases to −0.12 m. Because all members develop a negative bias, the ALL ensemble is also negative and it is similar to the SBSS ensemble mean given the large number of SBSS members in the ALL dataset (Fig. 3b). The ENS-3 mean has a negative bias

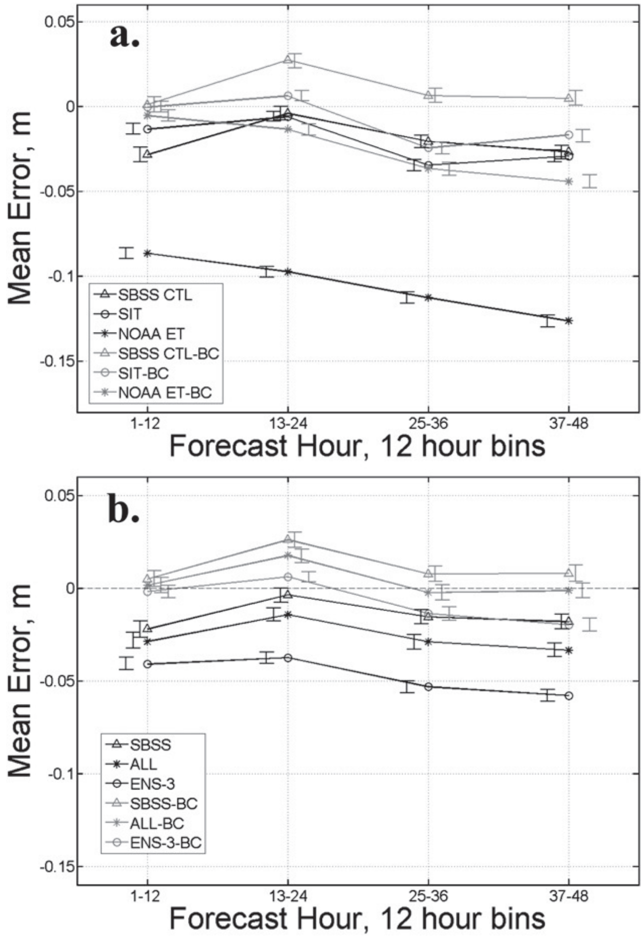

FIG. 3. (a) Mean error (m) in predicted storm surge vs forecast hour averaged over 12-h periods and the five stations throughout southern New England before and after bias correction (BC). (b) As in (a), but for the SBSS, ALL, and ENS-3 ensemble means.

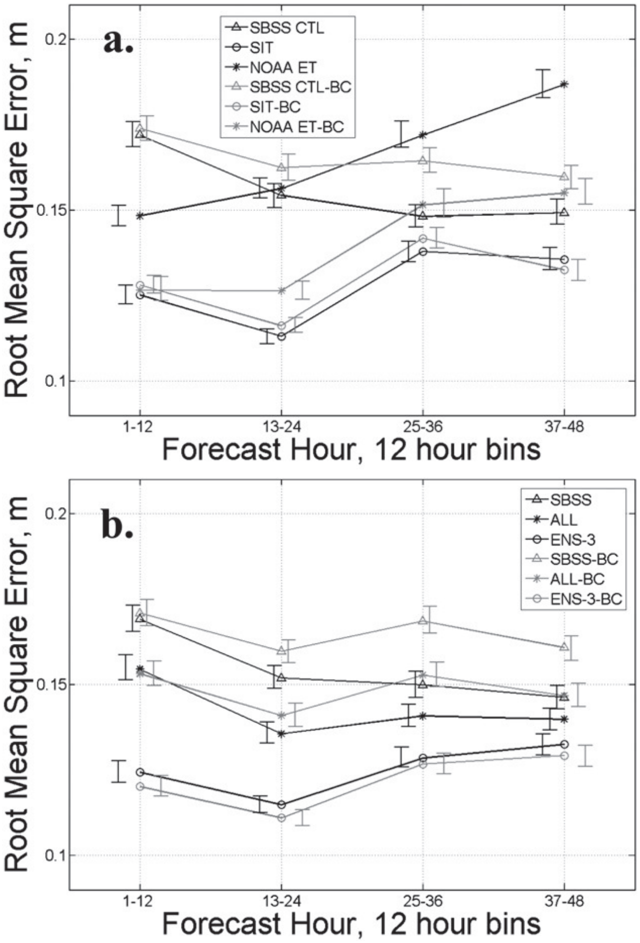

FIG. 4. As in Fig. 3, but for RMSE (m).

(−0.06 to −0.07 m between hours 25 and 48 h), since one-third of the weight is from the negatively biased NOAA-ET member.

Figure 4a shows the RMSE for the SBSS CTL, SIT-NYHOPS, and NOAA-ET members. At hours 1–12, the SBSS CTL has the largest RMSEs (~0.17 m), while the NOAA-ET has slightly lower RMSEs (~0.15 m), and SIT performs the best with RMSEs around 0.13 m. By hours 13–24, the SBSS CTL has errors similar to those of NOAA-ET (~0.16 m), while the SIT-NYHOPS system has the lowest RMSEs on average (~0.12–0.13 m). After hour 24, the NOAA-ET errors (~0.17 m) are the largest of all members and increase with forecast hour, while the SIT-NYHOPS RMSEs (~0.13 m) are smaller than the SBSS CTL and SBSS ensemble mean (~0.15 m) but are similar to the ALL ensemble (Fig. 4b). The ENS-3 has slightly smaller errors than the SIT-NYHOPS, but this difference is only significant at the 90% level between 25 and 36 h; however, this does illustrate that the three-member ensemble can outperform the best individual member for a deterministic verification.

The ME and RMSE for the individual SBSS members are clustered (within 0.03 m of each other; see Fig. 5). During hours 1–12, all SBSS members show more negative MEs and larger RMSEs than for hours greater than 12, which is likely due to initialization errors in the SBSS ensemble. From hours 37–48, the K2MY and GRBLK models (Table 2), with errors ~−0.01 m, are less biased than the other ensemble members at the 90% confidence interval. There is some suggestion that the WRF members (gray lines) have a slightly more negative bias at all hours, but the results are not statistically significant. Across all stations, the SBSS ensemble mean and control member have slightly lower RMSEs than do the other SBSS members, but the differences are not statistically significant. Generally, the GRBLK has the largest error for 13–24 h, while the BMMY has an error that is larger (significant at the 90% level) than most other members from 25 to 36 h.

Interestingly, the model errors for the SBSS and ALL ensembles do not increase with increasing forecast lead

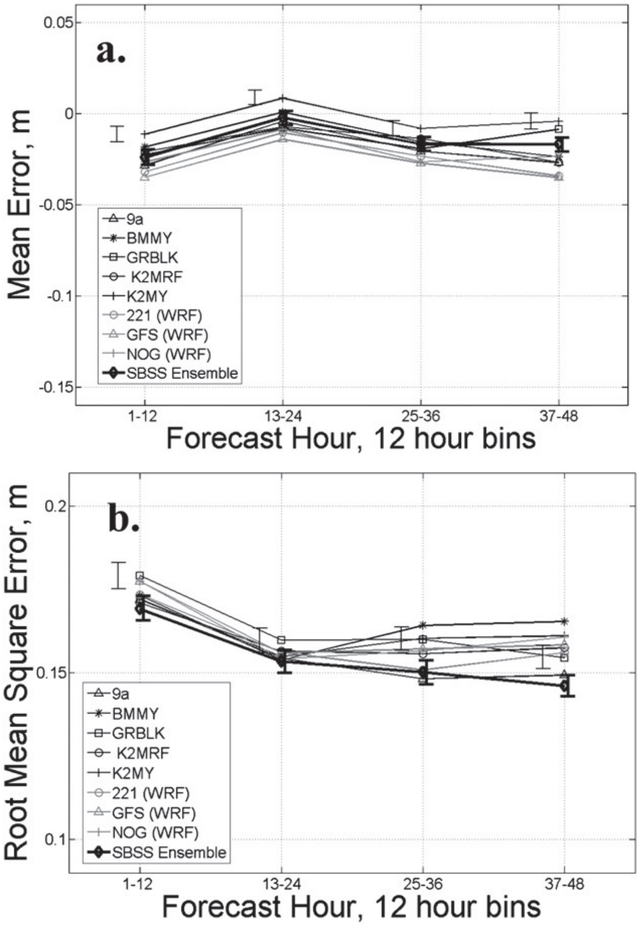

FIG. 5. (a) Mean surge error (m) for each member and ensemble mean (SBSS) of the Stony Brook storm surge ensemble. (b) As in (a), but with RMSE.

time (Fig. 4b). Meanwhile, the NOAA-ET model and, to a lesser degree, the SIT-NYHOPS between 1 and 48 h show increasing RMSEs with forecast hour (Fig. 4a), which is expected given that atmospheric wind and pressure errors typically increase with greater lead time (as will be shown in the next section). This is not surprising given the performance of the SBSS individual members shown in Fig. 5. This suggests that use of the previous day's forecast to initialize the SBSS members had a detrimental impact early in the forecast.

To understand the spatial distribution of errors across the region, the MEs and RMSEs are calculated at each station for the same 13–48-h forecast period (Fig. 6). The first 12 h of the forecast was not included to minimize any spinup issues. The SIT-NYHOPS model has the smallest ME at the Battery and Sandy Hook stations, while the SBSS had the smallest bias at Kings Point and Montauk. The ALL ensemble had the smallest ME at Bridgeport. The largest negative MEs in the NOAA-ET (−0.12 to −0.13 m) were at Bridgeport and King's Point, which may be in part from the using the GFS, which could not resolve the surface winds over Long Island Sound. The SIT-NYHOPS and ENS-3 have comparable RMSEs at most stations, which are smaller than the other models and the ALL mean. The ENS-3 is more accurate than SIT-NYHOPS at Sandy Hook, which is significant at the 90% level. Overall, the largest errors among all of the models are within Long Island Sound at Kings Point and Bridgeport, which is likely because the tides are larger there, and so part of the RMSE is the result of the tide not being perfectly resolved by the tidal models (Georgas and Blumberg 2010).

The NOAA-ET, SIT-NYHOPS, and SBSS CTL simulations, and the combined ensemble, were also evaluated for different surge thresholds between 13 and 48 h to determine how their performance behaves during small and large events (Fig. 7). The ME and RMSE were calculated at each station when either the model or observation exceeded the following surge thresholds: >0.4, >0.3, >0.2, >0.1, >0, <0, ≤0.1, ≤0.2, ≤0.3, and ≤0.4 m), so the model performance can be assessed. The five stations were combined to increase the sample size. For example, there are 720 h when the observations or SBSS CTL forecasts are less than −0.4 m and 455 h when greater than 0.4 m. For the negative surge events (water less than tide or less than 0 m), the SBSS CTL has a positive ME (~0.05 m; see Fig. 7a), while the SIT-NYHOPS has a −0.02 m bias and the NOAA-ET has a negative error (~0.11 m). This positive error increases for the SBSS CTL and SIT-NYHOPS as the negative surge increases, so that for events ≤0.4 m, the MEs for the SBSS CTL and SIT-NYHOPS are 0.13 and 0.04 m, respectively. In contrast, the NOAA-ET negative bias increases to −0.13 m for events ≤0.4 m. Given the clustering among members, the ALL ensemble has a similar mean (bias) error as the SBSS and SIT-NYHOPS (not shown). The RMSE errors are the smallest for SIT-NYHOPS for surge events <0 m.

For the positive surge events (water level greater than tide), the SBSS CTL has an ME that is more negative for larger storm surge events (Fig. 7a), reaching −0.22 m for surges >0.4 m. Meanwhile, the NOAA-ET mean error remains negative and fairly constant with increasing storm surge (−0.13 to −0.15 m). The SIT-NYHOPS performs the best at all stations for positive surges (Figs. 7a and 7b), with the SIT-NYHOPS having an ME for >0.4 m surge of about −0.09 m and an RMSE of ~0.22 m for the 13–48-h forecast.

Although there were only five events (~40 h) with >0.6-m surges (coastal flooding threshold for the Battery), the results averaged for 1–48-h forecasts are similar to the >0.4-m threshold above. More specifically, all members have a negative ME for a >0.6-m surge (not shown), with SBSS the largest (−0.18 m) and the SIT

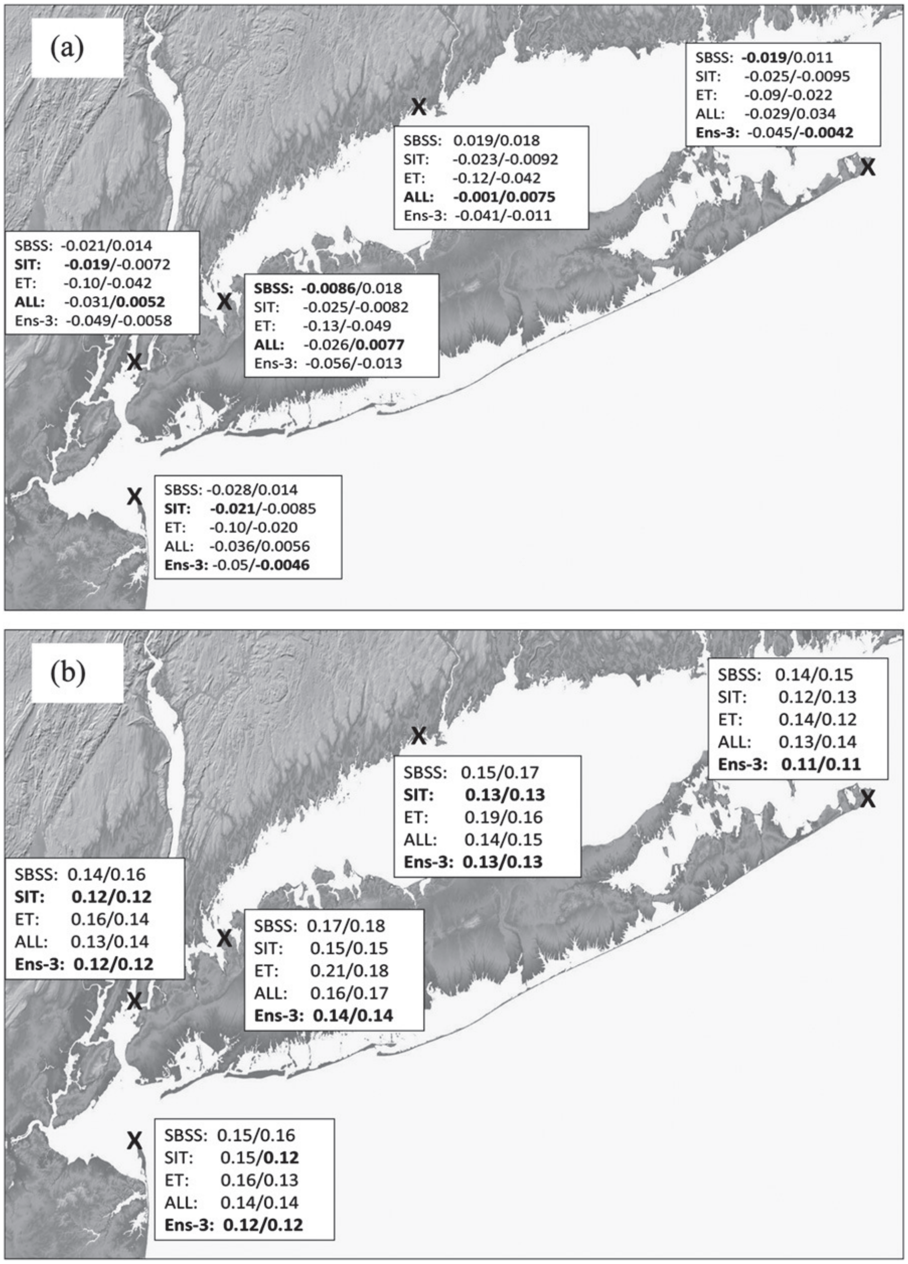

FIG. 6. (a) Mean errors (m) before (left number) and after (right number) bias correction for the individual stations around Long Island averaged between 12 and 48 h for the SBSS control member, SIT-NYHOPS, NOAA-ET, ALL, and ENS-3. (b) As in (a), but for RMSE. The boldface numbers highlight which member or ensemble mean performed best.

and NOAA-ET simulations having similar errors around −0.11 m. For these more major events, the SIT-NYHOPS model has the lowest RMSE (0.21 m) among all three of the modeling systems, which is similar to the result for ENS-3 (0.22 m).

b. Relationship of raw surge errors to wind errors

While there are greater differences between the various storm surge modeling systems, the MEs and RMSEs of the raw MM5 and WRF members in the SBSS ensemble are clustered. To illustrate the relationship between some of the surge errors and the surface wind predictions over the water, the surface winds in the NCEP NAM, as well as the Stony Brook WRF and MM5 members, were verified at the closest available buoy (see 44017 in Fig. 1). The model data were bilinearly interpolated to the observation point, in this case the buoy, using surrounding grid points. Due to missing archived wind data in both the NCEP-NAM and the SBSS ensemble simulations, comparisons are made for only 50 days, in which the

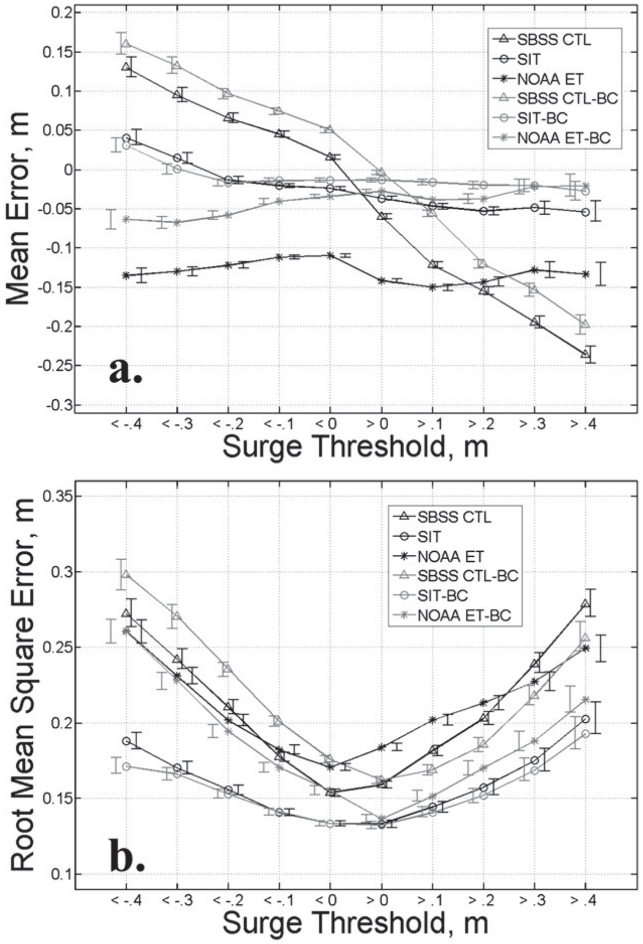

FIG. 7. (a) Storm surge mean error (m) across 5 stations for each of 10 different storm surge thresholds for the SBSS control, SIT-NYHOPS, and NOAA-ET before and after bias correction. For each ensemble member, the data were binned depending on whether the observation or the model met the threshold. (b) As in (a), but for RMSE.

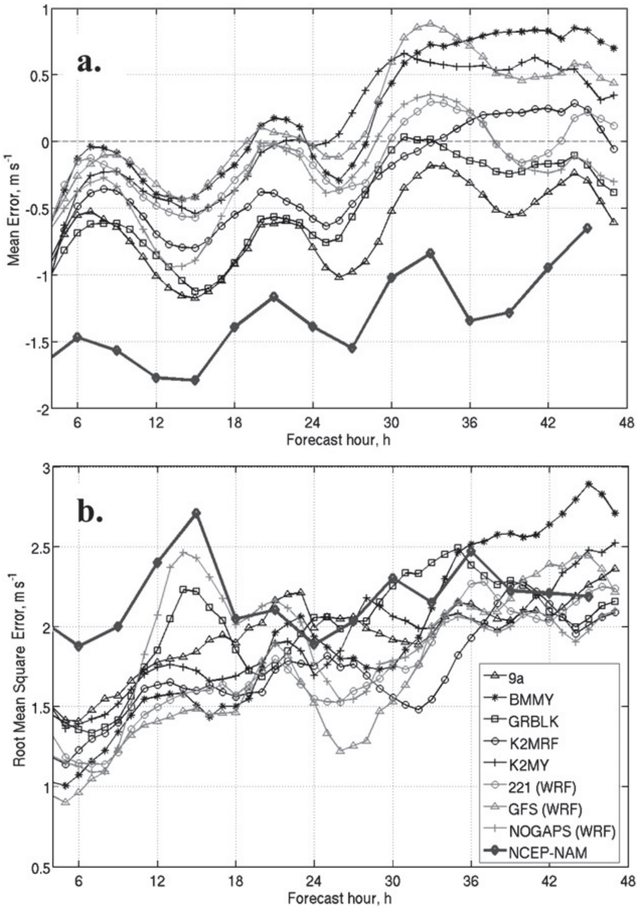

FIG. 8. (a) Wind speed mean error (m s^{-1}) for all eight members of the SBSS ensemble and the NCEP-NAM averaged over 48 h of the forecast. (b) As in (a), but for RMSE. The data before hour 5 were unavailable.

average wind speed and its standard deviation at the buoy are ~8.9 and 3.8 m s^{-1}, respectively. The average RMS differences in wind direction between members are generally less than 10° throughout the 48-h forecast (not shown), similar to the results in Jones et al. (2007; see their Fig. 2d). Thus, the focus is on the relationship between the predicted wind speed and storm surge error. In general, Jones et al. (2007) also showed that wind speeds among the atmospheric members of the SBSS ensemble tend to be underdispersed. The surface wind errors are more negative (~0.5 m s^{-1} lower) in the NAM than in the MM5–WRF (Fig. 8a), yet the SIT-NYHOPS mean errors are similar to the SBSS control results for many forecast hours (Fig. 3a). The NAM also has one of the largest RMSEs for surface wind speeds (Fig. 8b); however, the SIT-NYHOPS ocean model (which uses the NAM forcing) has lower RMSEs than do the SBSS ensemble members (Figs. 4b and 5b). Interestingly, even though model wind RMSEs increase with forecast hour (Fig. 8a), SIT-NYHOPS and the SBSS control member RMSEs are flat after hour 24 (Fig. 4a). However, unlike the SBSS and NOAA-ET, when the SIT-NYHOPS errors are binned every 3 h (not shown), the SIT-NYHOPS surge errors have a diurnal pattern that is similar to that of the NAM wind errors shown in Fig. 8a, although the surge errors are lagged by several hours (not shown). This suggests that the pattern of diurnal wind errors has a larger influence on the SIT-NYHOPS results. Closer inspection of the surface stress formulations used in ADCIRC (Garratt 1977), and SIT-NYHOPS (Large and Pond 1982), revealed that SIT-NYHOPS has 15%–25% less stress than ADCIRC for wind speeds 11–25 m s^{-1}, with similar results for other winds. Thus, the differences in water level predictions between ADCIRC and SIT-NYHOPS are likely not the result of the surface stress formulations, but rather from the ocean model setup (spinup procedures, model three-dimensionality, etc.). The ADCIRC forecasts errors are relatively clustered even with some diversity in the wind errors (Fig. 8), which

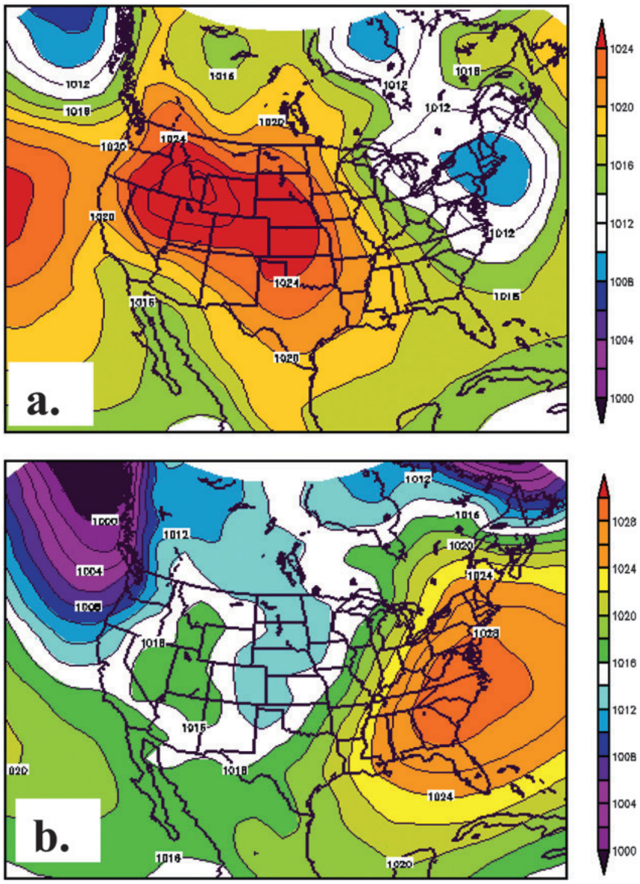

FIG. 9. (a) NCEP–NCAR reanalysis SLP composite (shaded, hPa) at the times for the unique dates between the 10 largest negative storm surge error, day-2 (24–48 h) forecasts for SBSS control members at the Battery, Sandy Hook, and King's Point. (b) As in (a), but the 10 largest positive surge errors.

suggests that to increase the diversity of the storm surge predictions, different ocean models should be used in an ensemble.

c. Spatial weather patterns favoring SBSS biases

It was hypothesized that there are certain weather patterns that yield larger negative and positive errors in the SBSS system. A composite was constructed using the NCEP–National Center for Atmospheric Research (NCAR) daily reanalysis data (Kistler et al. 2001) showing sea level pressure (SLP) for the unique dates among the 10 most negative (15 total dates) and positive (13 total dates) averaged 25–48-h forecast error days in the SBSS control member at the Battery; Sandy Hook, New Jersey; and King's Point, New York (Figs. 9a and 9b). The composite of the largest negative error days ($\leq$0.22 m) for the control SBSS member has a cyclone located just off of Cape Cod, Massachusetts, at the time of the largest surge error (Fig. 9a). This is a familiar setup for large surge events for NYC as a nor'easter cyclone tracks along the coast (Colle et al. 2010). The northeasterly flow with the cyclone helps push water along the mid-Atlantic coast.

In contrast, the SLP composite of the largest positive errors (>0.096 m) in the SBSS shows the presence of a large ~1030-hPa high pressure area off the coast of Virginia and North Carolina (Fig. 9b). Thus, the southwesterly winds over the northeast United States result in water transported away from the coast and the storm surges are negative. These results suggest that the model has a difficult time capturing the movement of water away from the coast during these events.

One possible reason for a negative surge on storm days is the impacts of waves on storm surge, which is neglected in the operational models. Westerink et al. (2008) mentioned that waves positively enhanced storm surges by 0.6–1.2 m during Hurricane Katrina across Louisiana. To illustrate the potential impacts of waves on the verification of surge results, significant wave heights at buoy 44017 were averaged daily over the same 74 days as the surge verification and compared with the corresponding day 2 (25–48 h) surge error of the SBSS ensemble mean at Montauk Point (Fig. 10). When there are larger wave heights, the surge error tends to be more negative, with a correlation coefficient (r) of -0.38. This occurs more when the waves are relatively large (>2.5 m). It is well known that waves can increase storm surge through the momentum they impart into the water column ("radiation stress") through wave breaking (Resio and Westerink 2008). This illustrates the potential importance of coupled ocean–wave models for surge prediction. Although, outliers (large positive model errors) in Fig. 10 emphasize that other factors, including wind speed, wind direction, and sea level pressure error, can be as or more influential than waves.

d. Impacts of bias correction

The above raw ensemble results suggest that surge models may have biases, which can result from a number of sources (errors in predicted tides in SIT-NYHOPS and SBSS, grid resolution, surface and bottom stress formulations, waves, meteorological wind and pressure biases, initialization schemes, and vertical datum calibration issues). A 5-day anomaly bias correction was applied to each ensemble member as described in section 2b. This correction removed most of the negative bias in the NOAA-ET member at all forecast hours (Fig. 3a). There is also a slight positive ME (0–0.2 m) for the SBSS CTL member from 1 to 48 h (Fig. 3a); therefore, the 5-day correction results in a slight positive bias (~0.03 m) during the 1–48-h period in the SBSS and ALL ensembles (Fig. 3b). The NOAA-ET MEs were improved at all thresholds (Fig. 7a), while the SBSS and SIT-NYHOPS

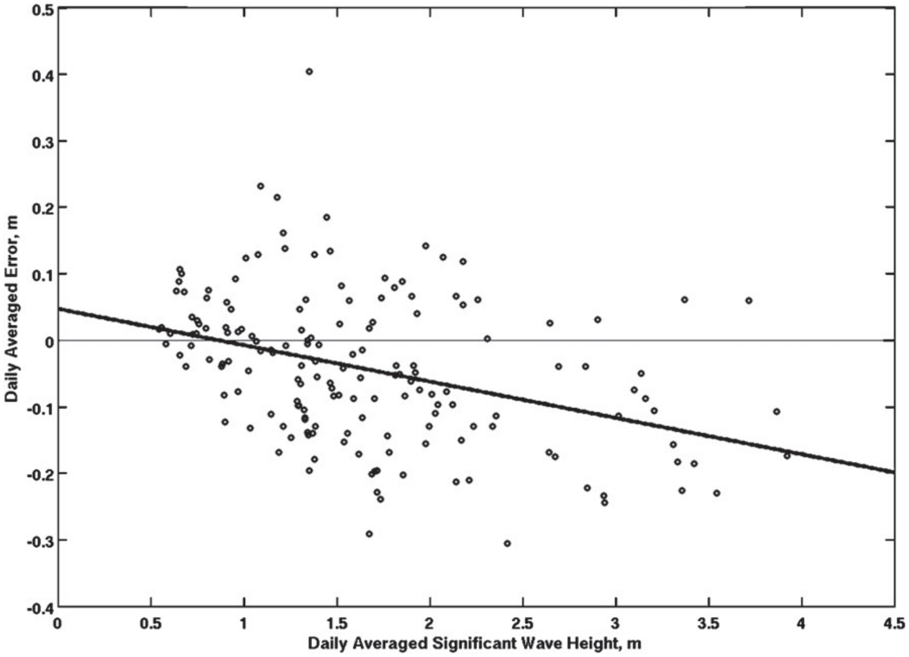

FIG. 10. Averaged daily significant wave height at buoy 44017 compared with daily averaged mean error for the SBSS control member at Montauk Point for the same corresponding days. See Fig. 1 for buoy 44017's location.

biases were reduced only for the positive surges. The MEs for the SBSS negative surges increased after bias correction. Closer examination revealed that there were many cases in which the bias correction added the wrong (sign) correction during these offshore flow (negative surge) conditions (Figs. 7a and 7b), since these periods sometimes occur after a positive surge period with onshore flow near a front or cyclone.

The bias correction reduced the RMSEs for the NOAA-ET model at all times (Fig. 4a), but it resulted in little change in the SIT-NYHOPS and an increase in errors for the SBSS. As a result, the SIT-NYHOPS and NOAA-ET have similar accuracy after bias correction for positive surges, while the bias-corrected SBSS has the largest errors. The SBSS used a water level cycling from the previous day's surge forecast to initialize, and this likely increased the error for this member. Also, the largest SBSS RMSEs are for the negative surges, which the bias correction could not fix as easily.

All ensemble members were averaged after bias correction (ALL–BC), which can be compared to the no bias correction ensemble (ALL). A three-member ensemble of bias-corrected SBSS CTL, SIT-NYHOPS, and NOAA-ET members was also created (ENS-3–BC). For hours 1–24 (Fig. 3b), there was no improvement in the mean error after bias correction for the SBSS ensemble, since bias correction resulted in a slight positive ME (~0.02 m), while there is less bias in the ENS-3–BC and ALL–BC results. After hour 24, the ALL–BC run has a near-zero bias, which is better than the other ensemble averages at the 90% significant level. The ensemble mean with the lowest RMSE after bias correction is the ENS-3–BC (Fig. 4b), but the ENS-3–BC RMSEs are only slightly less than those of SIT-NYHOPS–BC and thus are not significant at the 90% level. The reason the ENS-3–BC model is outperforming the ALL–BC deterministically is that most of the ALL–BC simulation is weighted from the SBSS members, which are clustered and have less accuracy than the SIT-NYHOPS and NOAA-ET members after bias correction.

4. Probabilistic verification results

An important objective of an ensemble is to improve the probabilistic skill. Ideally, each of the equally weighted ensemble members should have a similar level of accuracy. Figure 11a shows the percentage of the time that each member of the SBSS–BC ensemble, as well as the two ensemble means (ALL–BC and ENS-3–BC), performed the best (lowest RMSE) for the 13–48-h forecasts. Because the comparison is between the ensembles means and their individual members, the SIT-NYHOPS used in Fig. 11 has been shifted so that all models are compared over the same forecast period. Even so, the SIT-NYHOPS and NOAA-ET models after bias correction are the best, with a similar percentage on average (15%–30%), which is greater than any of the eight SBSS ensemble members (2%–7%). The SIT-NYHOPS

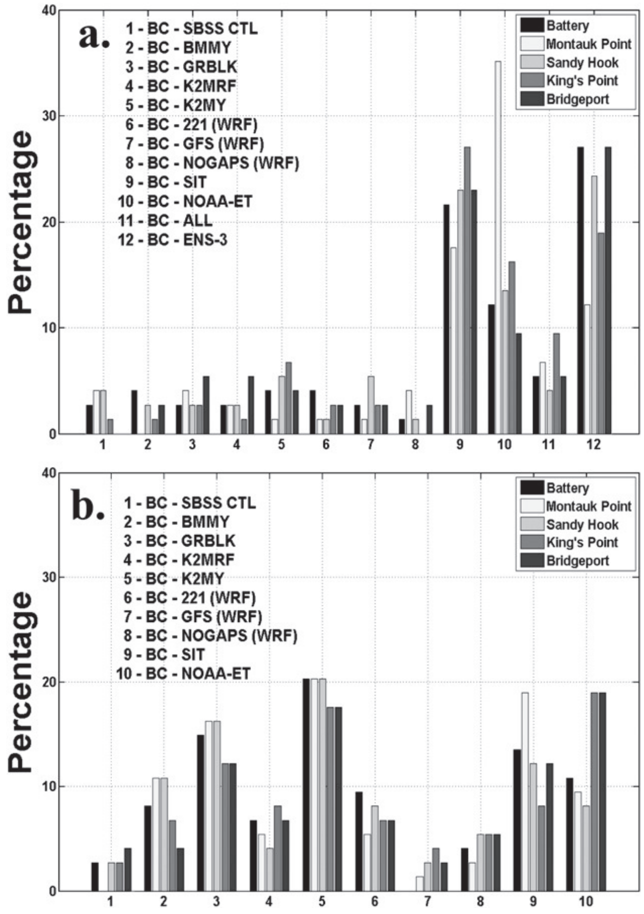

FIG. 11. (a) Percentage of time after bias correction that each member in the SBSS ensemble, SIT-NYHOPS, NOAA-ET, ALL, and ENS-3 performed the best (lowest 12–48-h-averaged RMSE) in storm surge forecasts at each station. (b) As in (a), but for the worst member for the SBSS, SIT-NYHOPS, and NOAA-ET. There is no mean ensemble plotted for (b), since the mean cannot be worst.

is the best, more than the NOAA-ET at King's Point, while NOAA-ET is the best member most often at Montauk. The clustering of the SBSS member errors (Fig. 6) is consistent with the percentage best being spread nearly equally among its members. The ENS-3–BC is best also nearly the same as in the NOAA-ET and SIT-NYHOPS models, since two-thirds of the ENS-3 mean originates from these two models. On the other hand, the ALL–BC is more weighted by the SBSS ensemble, and thus it is best only slightly more than each of the SBSS members.

The number of times that each member after bias correction is the worst was also calculated (Fig. 11b). For the SBSS ensemble, the GRBLK and K2MY members are worst members a larger percentage of the time (12%–20%) than the other SBSS members, while the GFS-WRF and SBSS control member are rarely the worst members (<5%). The three WRF members for the SBSS ensemble are the worst less often on average than are the MM5 members, except for the SBSS control. The SIT-NYHOPS after bias correction has one of the top three worst member percentages at the Battery, Sandy Hook, and Montauk stations (13%–19%), while the NOAA-ET has a similar large percentage at the Long Island Sound stations (19%–20%). Inspection of several individual surge events revealed that the SIT-NYHOPS tended to have 5%–10% larger surge maxima and minima than did the SBSS members (not shown), which helps the SIT-NYHOPS if the timing of the surge is predicted well. As a result, the difference between the SIT-NYHOPS during its best forecasts and the other members during the same periods is the largest of any of the models. The SIT-NYHOPS and ENS-3–BC also finishes as second or third best member more than the other models. On the other hand, after bias correction, timing errors combined with this larger surge variance in the SIT-NYHOPS simulation resulted in many forecasts in which it was the worst. However, if the days in which the SIT-NYHOPS are worst are compared with the days in which the other models are worst, the SIT-NYHOPS still had the lowest RMSEs. There are many forecasts in which the SIT-NYHOPS's errors are slightly larger than the other models, while the other models have much larger errors than SIT-NYHOPS when they are the worst (not shown).

The rank histogram for the ALL ensemble for all stations between 12 and 48 h is L shaped (Fig. 12), which denotes a negatively biased ensemble that is underdispersed. Thus, the observation falls outside the ensemble envelope a majority of the time (~60% for all stations). These results are consistent with the clustering of the SBSS deterministic verification above (Fig. 4a). After applying a bias correction, the ensemble is more U shaped, which is less biased, but there is still underdispersion, with the observation falling outside the ensemble ~53% the time (Fig. 12). Additional postprocessing (calibration) would be needed to reduce this underdispersion, such as Bayesian modeling averaging (Hoeting et al. 1999), which has been shown to work well for temperature (Wilson et al. 2007) and precipitation (Sloughter et al. 2007) ensemble forecasts.

To determine the probabilistic skill of the ensemble, the BS and BSS were calculated for each ensemble using Eqs. (3) and (4) above, respectively. Each of these scores is based on whether the observed and 12–48-h forecast surges exceeded a set positive surge threshold (>0, >0.1, >0.2, >0.3, >0.4 m; see Fig. 13). The BS in the SBSS, ALL, and ENS-3 ensembles decreases (improves) as the positive surge becomes larger (Fig. 13a), which suggests that the probabilistic accuracy increases for the larger surge events. For all thresholds except >0.4 m,

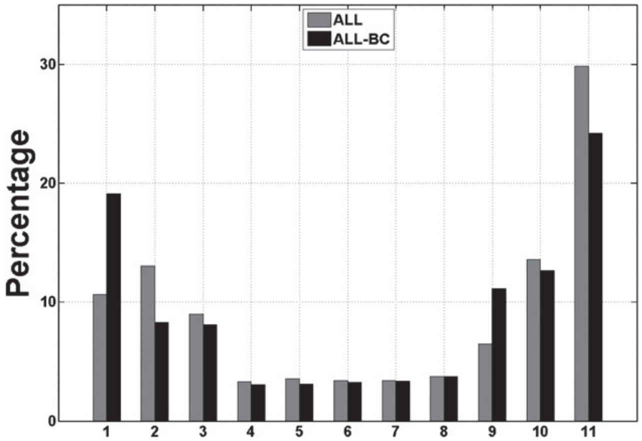

FIG. 12. Rank (Talagrand) histogram for all five stations using the ALL ensemble (gray) as well as the bias-corrected ensemble (black).

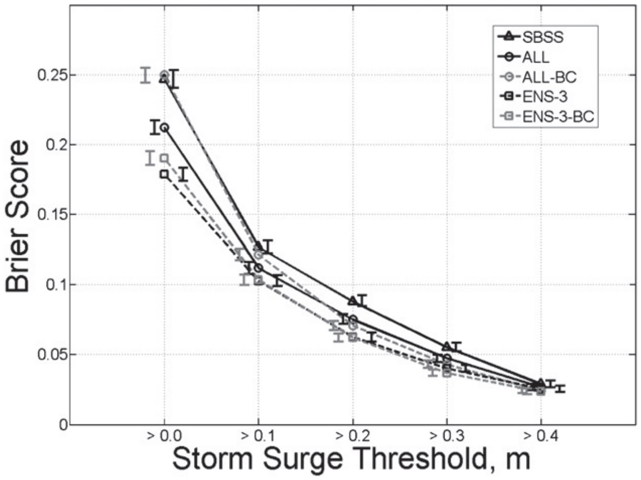

FIG. 13. BSs calculated for the SBSS, ALL, ALL-BC, ENS-3, and ENS-3-BC ensembles for five positive surge thresholds. Table 3 shows the breakdown of the BS for the >0.3-m surge.

the ALL ensemble has more accuracy than does the SBSS ensemble, while the ENS-3 is better (smaller BS) than the ALL ensemble. There is little improvement in the BS after bias correction for the ALL–BC and ENS-3–BC ensembles for the >0- and >0.1-m surges. Overall, these results illustrate that a multimodel ensemble constructed using only a few different surge models can yield more probabilistic accuracy on average than one surge model ensemble (SBSS) that uses different atmospheric forcings. If the SBSS ensemble members were just as skillful as the SIT-NYHOPS and NOAA-ET members, then some of the benefits of including different atmospheric winds and pressures may have been more fully realized for this period of study.

The BSS was calculated relative to the best bias-corrected member in the ALL ensemble (SIT-NYHOPS member). There is positive skill across all thresholds for the ALL and ENS-3 ensembles (Fig. 14a), while there is only statistically significant positive skill for the SBSS ensemble for the >0.1-m threshold. The ENS-3 has greater skill than the ALL ensemble at all thresholds, but the ENS-3 improvement for the >0.3-m thresholds is only significant at the 90% level given the smaller sample for these larger surges. There is some improvement in the BSS after bias correction is applied for the >0.2-m threshold, but it is not significant at the 90% level.

The BS was broken down into its reliability and resolution (Table 3). Since the uncertainty is greater than the BS for all ensembles, it can be seen that the ensembles have a positive BSS compared to climatology, which is defined as the uncertainty term (UNC above) in the Brier score (Fig. 14b). All ensembles except ALL–BC and SBSS have a positive BSS when climatology is used as a reference for cases >0 m. The ENS-3 outperforms the ALL ensembles for the >0-, 0.1-, 0.2-, and 0.3-m thresholds, and there is little difference after bias correction (not statistically significant).

Figure 15 shows reliability diagrams for surge events >0.3 m. The horizontal line refers to a climatological relative frequency of a >0.3-m surge event, which represents an ensemble having no resolution. The top diagonal line represents an ensemble having perfect reliability, while the bottom diagonal represents no reliability. For probabilities less than 80%, both the SBSS and ALL ensembles have smaller probabilities than are observed. When the SBSS (ALL) ensemble predicts an event to occur 25% (30%) of the time, the event actually occurs ~55% (70%) of the time for the SBSS (ALL) ensemble. This issue continues for moderate-probability (>0.3 m) events (0.4–0.7). For higher probabilities (>0.8), both ensembles have forecast probabilities roughly equal to the observed probability. There is a slight improvement in the reliability after applying bias correction to the members of the ALL ensemble. The ENS-3 ensemble is the most reliable of all ensembles; thus, this combined with the BSS results in Fig. 13 illustrates the benefit of running a multimodel surge ensemble system. Bias correction helps make the ENS-3–BC ensemble even more reliable than the raw, with the predicted probabilities agreeing well with the observed relative frequency. For a slightly larger storm surge event (>0.4 m), the ensembles perform similarly although the sample size is smaller (not shown).

5. Discussion and conclusions

For 74 days from November 2007 to March 2008 and from October to December 2008, an eight-member Stony

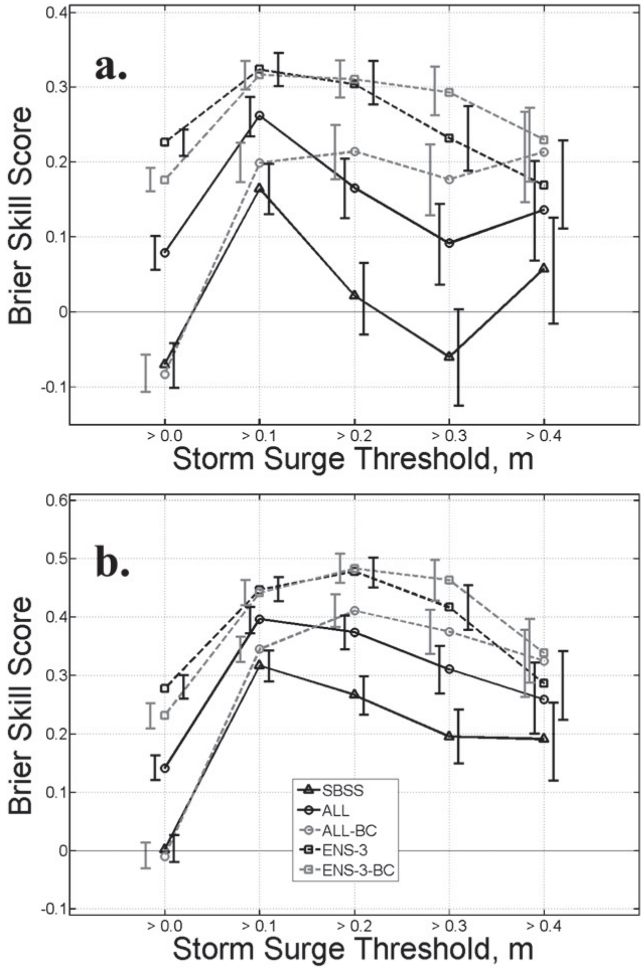

FIG. 14. As in Fig. 13, but for the BSS calculated relative to the (a) SIT-NYHOPS model and (b) climatology.

TABLE 3. BSS components: reliability (REL), resolution (RES), uncertainty (UNC), and BS for five ensemble configurations: SBSS, ALL, ALL–BC, ENS-3, and ENS-3–BC ensembles for the >0.3-m surge threshold.

	REL	RES	UNC	BS
SBSS	0.005	0.019	0.069	0.056
ALL	0.011	0.032	0.069	0.047
ALL–BC	0.007	0.032	0.069	0.043
ENS-3	0.002	0.030	0.069	0.040
ENS-3–BC	0.001	0.033	0.069	0.037

Brook storm surge (SBSS) ensemble, Stevens Institute hydrodynamic model (SIT-NYHOPS), and NOAA extratropical surge model (NOAA-ET) were verified for five stations around New York City and Long Island. The ensemble of all members (ALL), as well as a three-member ensemble (ENS-3: SBSS control member, SIT-NYHOPS, and NOAA-ET), were also evaluated for this same time period. The SBSS ensemble consists of the MM5 and WRF atmospheric members.

The raw NOAA-ET simulation has the largest negative bias (−0.12 m), while the SIT-NYHOPS and SBSS control members also have a slight negative surge bias after hour 24. In addition, the ME of the SBSS members does not increase with forecast time and the RMSE is highest for hours 1–12, which suggests that the use of the previous day's forecast as the initialization for the SBSS members has a negative impact early in the forecast. Many of the underpredicted surges in the SBSS ensemble are associated with high wave heights at an offshore buoy, thus illustrating the potential importance of nearshore waves (radiation stresses) on the surge predictions. A 5-day bias correction removes most of the surge bias in the NOAA-ET model, thus bringing it closer to the best deterministic model, SIT-NYHOPS, in terms of accuracy (RMSE) for positive surges. Conversely, this bias correction did not generally improve the SIT-NYHOPS and SBSS members. The multimodel surge ensemble after bias correction (ENS-3–BC) has the best deterministic accuracy as compared to all of the ensemble members (except SIT-NYHOPS, which is similar) and the ALL ensemble between hours 1 and 48. The ALL ensemble has accuracy that is less than for NOAA-ET and SIT-NYHOPS, since it is weighted more by the less accurate SBSS members.

Probabilistically, the raw ALL ensemble is biased given the L-shaped rank histogram, and after bias correction the ALL ensemble is underdispersed (U shaped), with the storm surge observations falling outside the ensemble ~53% of the time. Many of the atmospheric members have similar wind accuracies on average, and this lack of

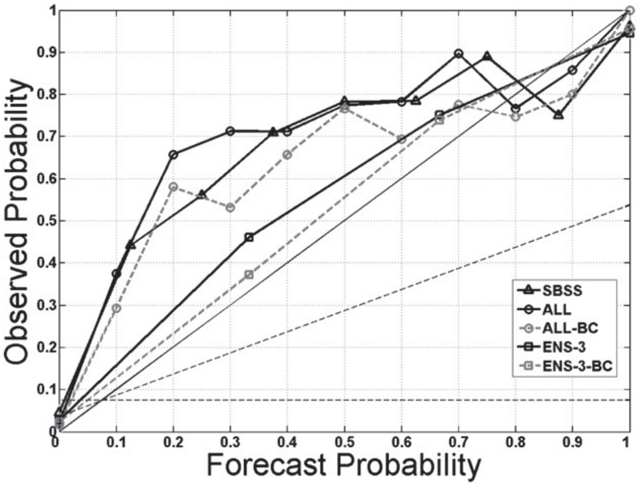

FIG. 15. Reliability diagram for >0.3-m threshold showing the SBSS, ALL, ALL-BC, ENS-3, and ENS-3-BC ensembles. The 1:1 line represents perfect reliability while the lower diagonal dashed line represents the line of ''no skill'' in a probability forecast.

wind dispersion is likely leading to some of this ensemble underdispersion. Brier scores (BSs) and Brier skill scores (BSSs) calculated for surge thresholds >0, >0.1, >0.2, >0.3, and >0.4 m shows the benefits of including ensemble members that use different ocean models. The BSS was calculated relative to the best deterministic member after ‘‘bias correction’’ (SIT-NYHOPS–BC). The ALL ensemble improves upon the SBSS in terms of BS and BSS even though 80% of the members of the ALL ensemble consist of the SBSS. The ENS-3 has the best BSs and BSSs for all surge thresholds compared to the ALL ensemble. The bias correction had only a minor improvement (not statistically significant) on the BSs and BSSs for the ALL–BC and ENS-3–BC simulations. Reliability diagrams for surge events >0.3 m shows that the SBSS and ALL ensembles are underconfident for forecast probabilities of less than 0.8. The ENS-3 increases the reliability even though the ensemble is only three members, and the ENS-3-BC has a nearly perfect reliability.

These results illustrate that there are biases in these storm surge models that may or may not be corrected with postprocessing, like the one used operationally for NOAA-ET surge forecasts and adopted here (5-day anomaly BC). For example, ADCIRC in the SBSS seems to have trouble moving water in and out of the region. This is seen through a composite of SLP for unique dates between the 10 largest positive and negative error days for SBSS control members at the Battery, Sandy Hook, and King’s Point. ADCIRC tends to underpredict cyclone events while overpredicting offshore (negative surge) events. This is not from a low MM5 or WRF bias in surface wind speed or the surface drag formation (SIT-NYHOPS applies a larger drag), but it may be the result of running ADCIRC in two dimensions. The SBSS model uses ADCIRC in a 2D configuration, which parameterizes the bottom stress based on a depth-averaged velocity within the water column. SIT-NYHOPS uses a 3D ocean model (POM type) that includes a bottom layer velocity, which is used to calculate the bottom stress. Weisberg and Zheng (2008) found that 2D models overestimate bottom stress compared to 3D models, which leads to an underestimation of surge heights in 2D ocean models. Another potential impact neglected in all models is the influence of wave radiation. It is shown that larger wave heights at an offshore buoy tend to correspond to larger negative error days for SBSS predictions at the coast.

Overall, these results have shown the benefits of using a multimodel surge prediction system. Both the probabilistic skill and reliability are improved by using a few different storm surge modeling systems simultaneously. Thus, operational forecasters would benefit if surge predictions from different ocean models could be combined in real time, rather than using just one surge model and an atmospheric ensemble. Individual model predictions are also continuously improving. For example, since July 2010, the SIT-NYHOPS model includes offshore surge forcing from NOAA-ET at its continental shelf break open boundary. This nesting of SIT-NYHOPS into NOAA-ET was found to further reduce surge RMS errors at the coast significantly (not shown). Meanwhile, additional research is needed exploring different bias-correction methods, coupled wave–storm surge impacts, and storm surge ensembles for more extreme surge events.

Acknowledgments. We thank the three anonymous reviewers for their comments and suggestions that helped improve the manuscript. This publication is a resulting product from project NYSG R/CCP-13 funded under Award NA07OAR4170010 from the National Sea Grant College Program of the U.S. Department of Commerce’s National Oceanic and Atmospheric Administration, to the Research Foundation of the State University of New York on behalf of New York Sea Grant. The statements, findings, conclusions, views, and recommendations are those of the authors and do not necessarily reflect the views of any of those organizations.

REFERENCES

Atallah, E. H., and L. F. Bosart, 2003: The extratropical transition and precipitation distribution of Hurricane Floyd (1999). *Mon. Wea. Rev.,* **131,** 1063–1081.

Betts, A. K., and M. J. Miller, 1993: The Betts–Miller scheme. *The Representation of Cumulus Convection in Numerical Models, Meteor. Monogr.,* No. 46, Amer. Meteor. Soc., 107–121.

Blier, W., S. Keefe, W. Shaffer, and S. Kim, 1997: Storm surge in the region of western Alaska. *Mon. Wea. Rev.,* **125,** 3094–3108.

Blumberg, A. F., L. A. Khan, and J. P. St. John, 1999: Three-dimensional hydrodynamic model of New York Harbor region. *J. Hydrol. Eng.,* **125,** 799–816.

Bowman, M. J., B. A. Colle, R. Flood, D. Hill, R. E. Wilson, F. Buonaiuto, P. Cheng, and Y. Zheng, 2005: Hydrologic feasibility of storm surge barriers to protect the metropolitan New York–New Jersey region. Marine Sciences Research Center Tech. Rep., Stony Brook University, 28 pp.

Bruno, M. S., A. F. Blumber, and T. O. Herrington, 2006: The urban ocean observatory – coastal ocean observations and forecasting in the New York Bight. *J. Mar. Sci. Environ.,* **C4,** 31–39.

Burroughs, L. B., and W. A. Shaffer, 1997: East Coast extratropical storm surge and beach erosion guidance. NWS Tech. Procedures Bull. 436, National Oceanic and Atmospheric Administration, 24 pp.

Colle, B. A., 2003: Numerical simulations of the extratropical transition of Floyd (1999): Structural evolution and responsible mechanisms for the heavy rainfall over the northeast United States. *Mon. Wea. Rev.,* **131,** 2905–2926.

——, J. B. Olson, and J. S. Tongue, 2003: Multiseason verification of the MM5. Part I: Comparison with the Eta Model over the central and eastern United States and impact of MM5 resolution. *Wea. Forecasting,* **18,** 431–457.

——, F. Buonaiuto, M. J. Bowman, R. E. Wilson, R. Flood, R. Hunter, A. Mintz, and D. Hill, 2008: Simulations of past cyclone events to explore New York City's vulnerability to coastal flooding and storm surge model capabilities. *Bull. Amer. Meteor. Soc.,* **89,** 829–841.

——, K. Rojowsky, and F. Buonaiuto, 2010: New York City storm surges: Climatology and an analysis of the wind and cyclone evolution. *J. Appl. Meteor. Climatol.,* **49,** 85–100.

Di Liberto, T., 2009: Verification of a storm surge modeling system for the New York City–Long Island region. M.S. thesis, School of Marine and Atmospheric Sciences, Stony Brook University, 135 pp. [Available from MSRC, Stony Brook University, Stony Brook, NY 11794-5000.]

Dudhia, J., 1989: Numerical study of convection observed during the Winter Monsoon Experiment using mesoscale two-dimensional model. *J. Atmos. Sci.,* **46,** 3077–3107.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective short-range ensemble forecasting. *Wea. Forecasting,* **20,** 328–350.

Ferrier, B., Y. Jin, Y. Lin, T. Black, E. Rogers, and G. DiMego, 2002: Implementation of a new grid-scale cloud and precipitation scheme in the NCEP Eta model. Preprints, *19th Conf. on Weather Analysis and Forecasting/15th Conf. on Numerical Weather Prediction,* San Antonio, TX, Amer. Meteor. Soc., 10.1. [Available online at http://ams.confex.com/ams/SLS_WAF_NWP/techprogram/paper_47241.htm.]

Garratt, J. R., 1977: Review of drag coefficients over oceans and continents. *Mon. Wea. Rev.,* **105,** 915–929.

Georgas, N., 2010: Establishing confidence in marine forecast systems: The design of a high fidelity marine forecast model for the NY/NJ Harbor Estuary and its adjoining waters. Ph.D. dissertation, Stevens Institute of Technology, 272 pp.

——, and A. F. Blumberg, 2010: Establishing confidence in marine forecast systems: The design and skill assessment of the New York Harbor Observation and Prediction System, version 3 (NYHOPS v3). *Estuarine and Coastal Modeling: Proceedings of the Eleventh International Conference,* M. L. Spaulding, Ed., American Society of Civil Engineers, 660–685.

Glickman, T. S., 2000: *Glossary of Meteorology*. 2nd ed. Amer. Meteor. Soc., 855 pp.

Grell, G. A., 1993: Prognostic evaluation of assumptions used by cumulus parameterizations. *Mon. Wea. Rev.,* **121,** 764–787.

——, J. Dudhia, and D. R. Stauffer, 1995: A description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5). NCAR Tech. Note NCAR/TN-398+STR, 122 pp.

Hack, J. J., B. A. Boville, B. P. Briegleb, J. T. Kiehl, P. J. Rasch, and D. L. Williamson, 1993: Description of the NCAR Community Climate Model (CCM2). NCAR Tech. Note NCAR/TN-382+STR, 108 pp.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial (with discussion). *Stat. Sci.,* **14,** 382–401.

Hong, S.-Y., and H.-L. Pan, 1996: Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.,* **124,** 2322–2339.

——, J. Dudhia, and S.-H. Chen, 2004: A revised approach to ice-microphysical processes for the bulk parameterization of cloud and precipitation. *Mon. Wea. Rev.,* **132,** 103–120.

——, Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.,* **134,** 2318–2341.

Janjić, Z. I., 2002: Nonsingular implementation of the Mellor-Yamada level 2.5 scheme in the NCEP Meso Model. NCEP Office Note 437, 61 pp.

Jelesnianski, C., J. Chen, and W. Shaffer, 1992: SLOSH: Sea, Lake, and Overland Surges from Hurricanes. NOAA Tech. Rep. NWS 48, 71 pp.

Jones, M., B. A. Colle, and J. Tongue, 2007: Evaluation of a short-range ensemble forecast system over the northeast United States. *Wea. Forecasting,* **22,** 36–55.

Kain, J. S., 2004: The Kain–Fritsch convective parameterization: An update. *J. Appl. Meteor.,* **43,** 170–181.

Kistler, R., and Coauthors, 2001: The NCEP–NCAR 50-Year Reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteor. Soc.,* **82,** 247–268.

Large, W. G., and S. Pond, 1982: Sensible and latent heat flux measurements over the ocean. *J. Phys. Oceanogr.,* **12,** 464–482.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmosphere: RRTM, a validate correlated-k model for the longwave. *J. Geophys. Res.,* **102,** 16 663–16 682.

Mukai, A., J. Westerink, R. Luettich, and D. Mark, 2002: Eastcoast 2001: A tidal constituent database for the western North Atlantic, Gulf of Mexico and Caribbean Sea. Coastal and Hydraulics Laboratory Tech. Rep. ERDC/CHL TR-02-24, U.S. Army Engineer Research and Development Center, 210 pp.

Mylne, K., R. Evans, and R. Clark, 2002: Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Quart. J. Roy. Meteor. Soc.,* **128,** 361–384.

Pawlowicz, R., B. Beardsley, and S. Lentz, 2002: Classical tidal harmonic analysis including error estimates in MATLAB using T_TIDE. *Comput. Geosci.,* **28,** 929–937.

Reisner, J., R. M. Rasmussen, and R. T. Bruintjes, 1998: Explicit forecasting of supercooled liquid water in winter storms using the MM5 mesoscale model. *Quart. J. Roy. Meteor. Soc.,* **124,** 1071–1107.

Resio, D. T., and J. J. Westerink, 2008: Hurricanes and the physics of surges. *Phys. Today,* **61,** 33–38.

Shen, J., W. Gong, and H. Wang, 2005: Simulation of Hurricane Isabel using the Advanced Circulation Model (ADCIRC). *Hurricane Isabel in Perspective,* K. G. Sellner, Ed., Chesapeake Research Consortium, 107–116.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech. Note NCAR/TN-468+STR, 113 pp.

Sloughter, J. M., A. E. Raftery, and T. Gneiting, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 3209–3220.

Tilburg, C. E., and R. W. Garvine, 2004: A simple model for coastal sea level prediction. *Wea. Forecasting,* **19,** 511–519.

Weisberg, R. H., and L. Zheng, 2006: Hurricane storm surge simulations for Tampa Bay. *Estuaries Coasts,* **29,** 899–913.

——, and ——, 2008: Hurricane storm surge simulations comparing three-dimensional with two-dimensional formulations based on an Ivan-like storm over the Tampa Bay, Florida region. *J. Geophys. Res.,* **113,** C12001, doi:10.1029/2008JC005115.

Westerink, J. J., R. A. Luettich Jr., and N. W. Scheffner, 1993: ADCIRC: An advanced three-dimensional circulation model for shelves coasts and estuaries. Report 3: Development of a tidal constituent data base for the western North Atlantic and Gulf of Mexico. Dredging Research Program Tech. Rep. DRP-92-6, U.S. Army Engineers Waterways Experiment Station, 154 pp.

——, and Coauthors, 2008: A basin- to channel-scale unstructured grid hurricane storm surge model applied to southern Louisiana. *Mon. Wea. Rev.,* **136,** 833–864.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian Ensemble Prediction System using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 1364–1385.

Woodcock, F., and C. Engel, 2005: Operational consensus forecasts. *Wea. Forecasting,* **20,** 101–111.

Zhang, D., and R. A. Anthes, 1982: A high-resolution model of the planetary boundary layer—Sensitivity tests and comparisons with SESAME-79 data. *J. Appl. Meteor.,* **21,** 1594–1609.

Zilkoski, D. B., J. H. Richards, and G. M. Young, 1992: Results of the general adjustment of the North American Vertical Datum of 1988. *Surv. Land Info. Syst.,* **52,** 133–149.

Zwiers, F. W., 1990: The effect of serial correlation on statistical inferences made with resampling procedures. *J. Climate,* **3,** 1452–1461.