Operational Storm Surge Forecasting at the National Hurricane Center: The Case for Probabilistic Guidance and the Evaluation of Improved Storm Size Forecasts Used to Define the Wind Forcing

ANDREW B. PENNY,[a,b] LAURA ALAKA,[a,b] ARTHUR A. TAYLOR,[c] WILLIAM BOOTH,[a,b] MARK DEMARIA,[d] CODY FRITZ,[b] AND JAMIE RHOME[b]

[a] *University Corporation for Atmospheric Research/Cooperative Programs for the Advancement of Earth System Science, Boulder, Colorado*
[b] *National Hurricane Center, Miami, Florida*
[c] *Meteorological Development Laboratory, Silver Spring, Maryland*
[d] *Colorado State University/Cooperative Institute for Research in the Atmosphere, Fort Collins, Colorado*

(Manuscript received 2 December 2022, in final form 23 September 2023, accepted 8 October 2023)

ABSTRACT: The primary source of guidance used by the Storm Surge Unit (SSU) at the National Hurricane Center (NHC) for issuing storm surge watches and warnings is the Probabilistic Tropical Storm Surge model (P-Surge). P-Surge is an ensemble of Sea, Lake, and Overland Surges from Hurricanes (SLOSH) model forecasts that is generated based on historical error distributions from NHC official forecasts. A probabilistic framework is used for operational storm surge forecasting to account for uncertainty related to the tropical cyclone track and wind forcing. Previous studies have shown that the size of a storm's wind field is an important factor that can affect storm surge. A simple radius of maximum wind (RMW) prediction scheme was developed to forecast RMW based on NHC forecast parameters. Verification results indicate this scheme is an improvement over the RMW forecasts used by previous versions of P-Surge. To test the impact of the updated RMW forecasts in P-Surge, retrospective cases were selected from 25 storms from 2008 to 2020 that had an adequate number of observations. Evaluation of P-Surge forecasts using these improved RMW forecasts shows that the probability of detection is higher for most probability of exceedance thresholds. In addition, the forecast reliability is improved, and there is an increase in the number of high probability forecasts for extreme events at longer lead times. The improved RMW forecasts were recently incorporated into the operational version of P-Surge (v2.9), and serve as an important step toward extending the lead time of skillful and reliable storm surge forecasts.

KEYWORDS: Storm surges; Hurricanes/typhoons; Ensembles; Operational forecasting; Numerical weather prediction/forecasting

1. Introduction

The National Weather Service's National Hurricane Center (NHC) Storm Surge Unit (SSU) is responsible for issuing storm surge forecasts and storm surge watches (possibility of life-threatening inundation during the next 48 h) and warnings (danger of life-threatening inundation during the next 36 h) during landfalling tropical cyclones affecting the U.S. coastline. The SSU forecast responsibility includes saltwater only, which can be pushed miles inland and into rivers and bays due to the winds of a tropical cyclone. SSU storm surge forecast products are available approximately 48–60 h prior to the onset of hazardous conditions. Efforts are under way to increase the lead time to 72 h, which will provide state and local officials with additional guidance to make important evacuation decisions.

A key source of guidance for operational storm surge forecasting at the NHC comes from the Probabilistic Tropical Storm Surge model (P-Surge) (Taylor and Glahn 2008; Glahn et al. 2009; Gonzalez and Taylor 2018), which is developed and maintained by the Meteorological Development Laboratory (MDL). For storms threatening landfall in the U.S. Gulf and East Coast, P-Surge is run every 6 h. The P-Surge model relies on the Sea, Lake, and Overland Surges from Hurricanes (SLOSH) hydrodynamic model (https://vlab.noaa.gov/web/mdl/slosh), which was initially developed by MDL more than 40 years ago (Jelesnianski and Taylor 1973; Jelesnianski et al. 1992).

It has been well known for some time that storm surge is sensitive to a storm's track, intensity, size, and forward speed, as well as the coastal characteristics of the landfall area (e.g., bathymetry) (e.g., Irish et al. 2008; Rego and Li 2009, 2010; Fossell et al. 2017; Ramos-Valle et al. 2020; Xuan et al. 2021). Since there is often a great deal of uncertainty related to the meteorological forcing several days prior to landfall, the P-Surge ensemble attempts to account for this uncertainty using a probabilistic framework. The number of unique realizations can vary from about 200–1200 depending on the storm RMW and how the P-Surge ensemble tracks intersect the SLOSH basins.

An example from Hurricane Laura (2020) illustrates the utility of probabilistic track guidance and the danger of relying on a single-track deterministic forecast to assess storm surge risk (Fig. 1). Results from a single-track deterministic SLOSH forecast based on the NHC Official (OFCL) forecast from Advisory 23 (1200 UTC 25 August; ~42 h prior to

Corresponding author: Andrew B. Penny andrew.penny@noaa.gov

DOI: 10.1175/WAF-D-22-0209.1

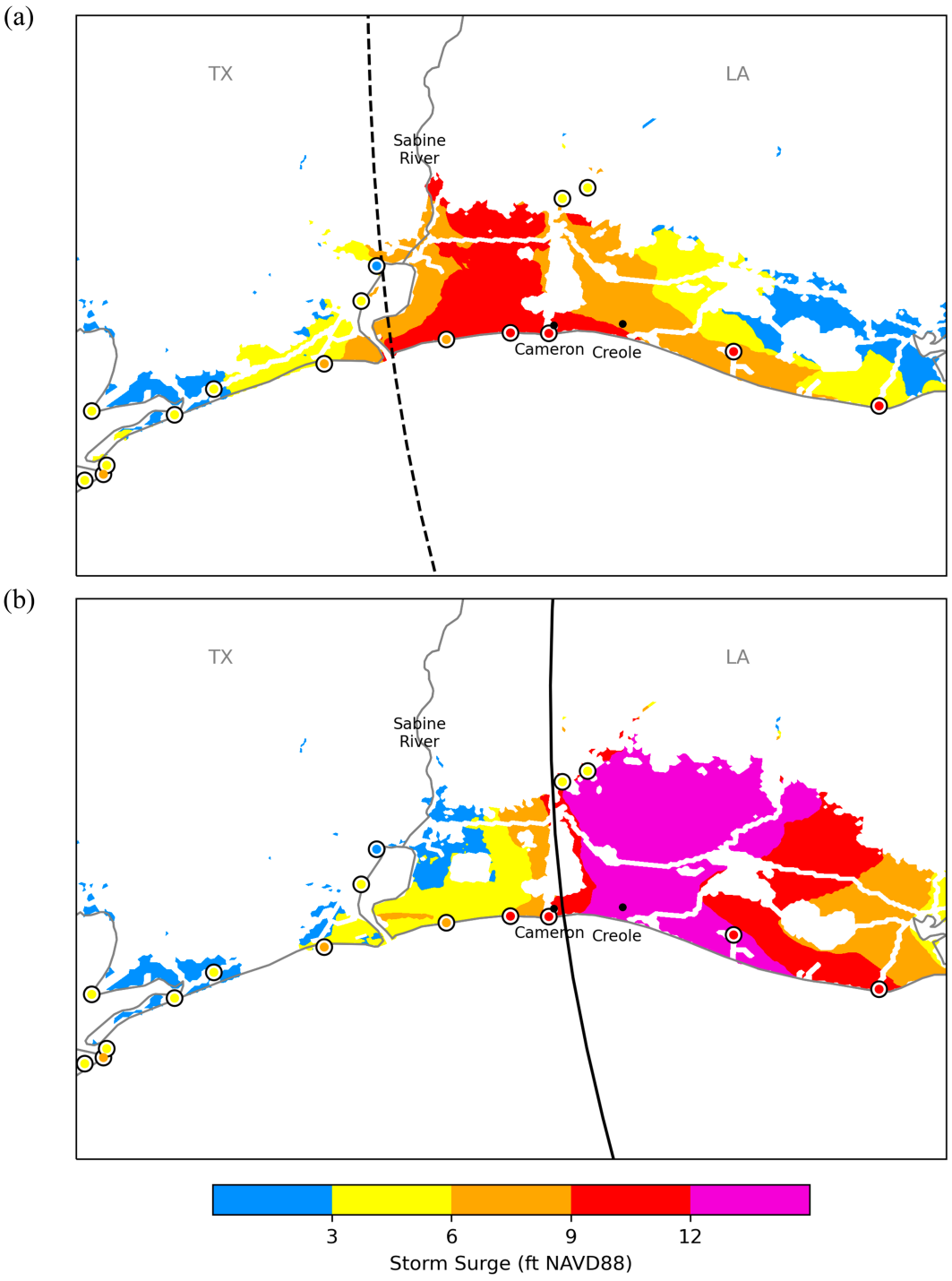

FIG. 1. Single-track (deterministic) SLOSH simulations for Hurricane Laura (2020) referenced to above datum (NAVD88) for (a) the NHC OFCL forecast (dashed line) initialized at 1200 UTC 25 Aug (~42 h prior to landfall) and (b) the NHC "best track" (solid line). Circles denote observations that are colored according to the legend.

landfall) are shown relative to the North American Vertical Datum of 1988 (NAVD88) in Fig. 1a. Based on this forecast, landfall was expected near the Texas–Louisiana state line and the area at greatest risk extended west of Cameron, Louisiana, to the Sabine River. In reality, Hurricane Laura made landfall closer to Cameron, or about 30 n mi (1 n mi = 1.852 km) east of the position suggested by the Advisory 23 forecast track, and was 30 kt (1 kt ≈ 0.51 m s^{-1}) stronger than

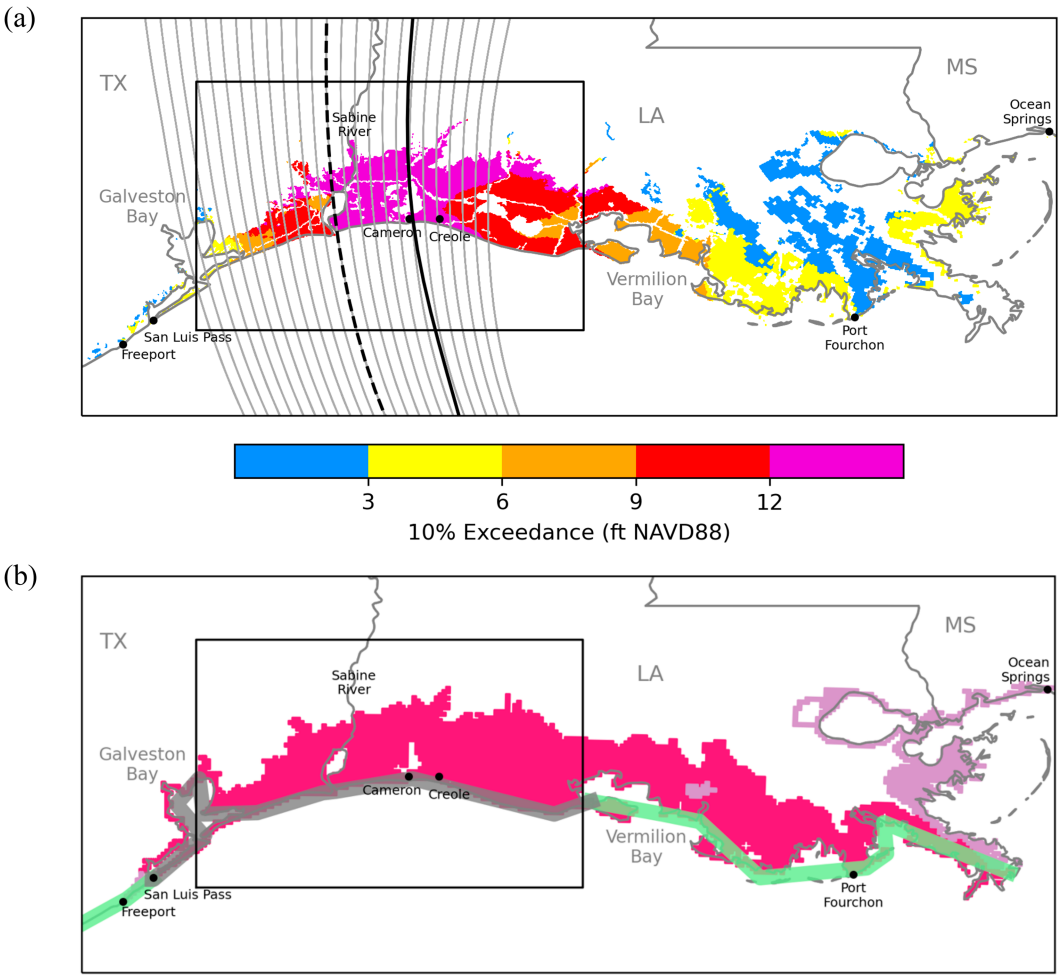

FIG. 2. (a) 10% exceedance values (shading; see legend) from the P-Surge ensemble forecast (tracks shown in thin solid lines) based on the 1200 UTC 25 Aug NHC OFCL forecast (dashed line) and (b) the corresponding storm surge watch (purple) and warning (magenta) issued with this advisory. The thick solid line in (a) denotes the NHC best track, and the lateral extents of the hurricane wind watch and hurricane wind warning along the coastline are indicated in (b) by the green and gray lines, respectively. The box-outlined area corresponds to the region shown in Fig. 1.

forecast (130 vs 100 kt). Observations and a single-track simulation based on the actual track (NHC best track) for Hurricane Laura (Fig. 1b) indicate the area most heavily impacted by storm surge was predominantly east of Cameron (Pasch et al. 2021).

In contrast to the deterministic forecast, the 10% exceedance guidance from the P-Surge forecast based on the 1200 UTC 25 August OFCL forecast (Fig. 2a) indicated there was a broad area at risk for storm surge greater than 9 ft (1 ft = 30.48 cm) above NAVD88 (hereinafter above NAVD88 is just given as NAVD88 for brevity) that extended from southeast Texas to Vermilion Bay, Louisiana, including the area most heavily affected. Based on the guidance available, a storm surge warning was issued with Advisory 23 (https://www.nhc.noaa.gov/archive/2020/al13/al132020.public.023.shtml) from San Luis Pass, Texas, to the mouth of the Mississippi River (magenta shading in Fig. 2b), while the coastal area from Freeport, Texas, to San Luis Pass and from the mouth of the Mississippi River to Ocean Springs, Mississippi, was placed under a storm surge watch (purple shading in Fig. 2b). This advisory also coincided with the issuance of hurricane and tropical storm wind warnings (see Fig. 2b).

While the Hurricane Laura example highlights the sensitivity to the landfall location, uncertainty related to the storm's size is also an important component of P-Surge because it affects the location and severity of storm surge (Irish et al. 2008; Zhang et al. 2008; Davis et al. 2010; Forbes et al. 2014; Mayo and Lin 2019). The sensitivity to storm size was demonstrated following the deadly storm surge associated with the very large wind field of Hurricane Katrina (2005) (e.g., Irish et al. 2008), and is considered in the NHC storm surge hazard mapping products (Zachry et al. 2015). In this paper, we discuss the changes made in P-Surge v2.9 (implemented operationally in May 2021; https://www.weather.gov/media/notification/pdf2/scn21-41p_surge2.9.pdf), which include the use of RMW

forecasts based on NHC OFCL forecast parameters to improve storm size information. Section 2 provides a description of SLOSH and the P-Surge ensemble as well as the methodology used to generate these RMW forecasts. Verification of the RMW forecasts is shown in section 3. Section 4 presents verification of P-Surge forecasts that use the improved RMW forecasts and discusses the observation-based verification procedures that were used to evaluate the v2.9 and other P-Surge upgrades. Additional discussion of these results is provided in section 5, along with the conclusions.

2. Method

a. SLOSH/P-Surge

SLOSH is a 2D hydrodynamic model that solves a simplified set of Navier–Stokes momentum equations using finite differencing on a semi-staggered Arakawa B-grid. The equations are integrated through the depth of the water column (Glahn et al. 2009). Bottom friction is accounted for via uniform bottom slip coefficients for each SLOSH basin, and dissipation is handled primarily through a vertical eddy viscosity coefficient. Pressure, Coriolis, and frictional forces are used to calculate the horizontal transport, which, along with a wetting and drying algorithm, determines the surge at each grid cell (Forbes et al. 2014). Water levels from tides are computed by using tidal constituents from the Advanced CIRCulation (ADCIRC) model (Szpilka et al. 2016), which are interpolated to each SLOSH grid (Haase et al. 2011; Fritz (Haase) et al. 2014).

SLOSH uses fixed computational grids, which, depending on the geometry of the coastline, are either polar, elliptical, or hyperbolic in shape. This allows for finer resolution near coastal areas of interest, and coarser resolution over the open ocean. While small-scale coastal features are generally better resolved using unstructured grids, a great deal of effort is spent retaining important subgrid-scale features in the SLOSH basins, such as channels, levees, etc. Arakawa C-grid cells are used to account for subgrid-scale features to simulate one-dimensional flow.

The SLOSH model is computationally efficient, allowing for thousands of simulations to be run within the 1-h operational time window on the NOAA Weather and Climate Operational Supercomputing System (WCOSS). The SLOSH model makes several simplifications to improve efficiency: the advection and baroclinic terms are ignored, and the drag coefficient for air and the eddy stress coefficient for water are held constant (Glahn et al. 2009), the latter of which is being actively addressed.

As Fig. 1 illustrates, a small change in the storm track can dramatically affect where the largest impacts from storm surge will occur. Thus, the largest source of error for any surge model generally comes from uncertainty related to the meteorological forcing. To identify coastal areas at risk to life-threatening storm surge, this uncertainty must be adequately accounted for. In general, this requires a large ensemble. While more sophisticated storm surge models exist, SLOSH was selected as the hydrodynamic model for P-Surge because it provides the best compromise between efficiency and accuracy that allows the NHC to achieve its operational forecasting objectives with finite computational resources.

To account for uncertainty in the NHC OFCL intensity forecast, P-Surge uses a three-member ensemble ("weak," "normal," and "strong"). The intensity of the medium-intensity storm is set to the NHC OFCL intensity forecast. A normal error distribution is assumed using the 5-yr average error to determine the perturbations for the weak and strong storms, which are centered at the 15th and 85th percentile of the intensity error distribution, respectively. In this way the ensemble members are designed to be representative storms of the lower 30%, middle 40%, and upper 30% of the assumed intensity error distribution. Therefore, the weights applied to the intensity ensemble members are 0.3, 0.4, and 0.3.

To account for uncertainty in the storm's forward speed, the ensemble includes seven tracks with different forward speeds. The perturbations are generated similarly to the intensity perturbations, except along-track error statistics are used. Forward speed affects the timing of landfall, which can either exacerbate or suppress the storm surge inundation depending on the magnitude and timing of the tidal cycle, and the duration of wind forcing near coastal areas.

The NHC OFCL cross-track errors are used to determine the spread in the cross-track direction as a function of lead time. The cross-track swath encompasses roughly 90% of the uncertainty based on the 5-yr average error. Therefore, tracks extend outside of the NHC track forecast cone, which is designed to account for ~67% of the average uncertainty (https://www.nhc.noaa.gov/aboutcone.shtml). Because of differences in bathymetry and coastal characteristics, which can affect inundation depth and extent (e.g., Weaver and Slinn 2010), it is vital to ensure that the entire coastal area threatened by storm surge is adequately sampled. The cross-track spacing is set so storm tracks are 1 RMW apart at 48-h lead time, which means more cross-track ensemble members are needed when the RMW is small.

P-Surge relies on the SLOSH parametric wind profile to define the atmospheric forcing for the SLOSH model forecasts. The parametric wind model specifies the wind speed v as a function of radius r and depends only on the radius, RMW, and maximum wind speed V_{max}:

$$v(r) = \frac{2.0 \times V_{max} \times \mathrm{RMW} \times r}{\mathrm{RMW}^2 + r^2}.$$

P-Surge v2.7, which was upgraded in May 2018 (https://www.weather.gov/media/notification/pdfs/scn18-36psurgeaaa.pdf), and all previous versions of P-Surge used the observed (i.e., NHC best track) intensity, latitude, forward speed, and minimum pressure to derive an initial RMW consistent with the SLOSH parametric wind profile during the "spinup" phase of the forecast (the 24-h period prior to the current synoptic time). However, the pressure–wind–RMW relationship used by P-Surge often resulted in an RMW that differed considerably from the NHC estimate. Since RMW forecasts are not provided as part of the NHC OFCL forecast, perturbations

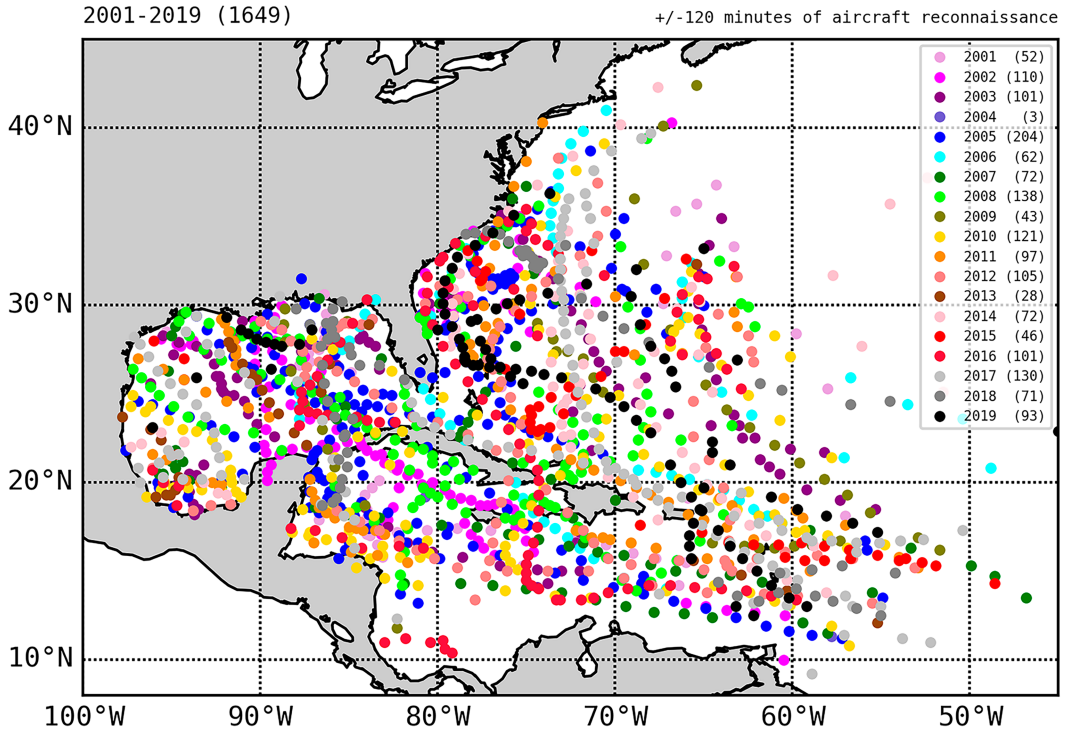

FIG. 3. Best-track storm locations at synoptic times within ±120 min of aircraft reconnaissance data. Circles are colored according to the year, and the number of cases included for each year is shown in the legend. These data were used to derive the relationship between RMW and other best-track parameters.

used to generate the three-member RMW ensemble (small, medium, and large) were based on the average error resulting from holding the RMW constant throughout the forecast (Taylor and Glahn 2008). These perturbations were a function only of the initial storm RMW and did not account for other factors that are known to be related to storm RMW (intensity, latitude, etc.).

P-Surge was upgraded to v2.8 in September 2020 (https://www.weather.gov/media/notification/pdf2/scn20-81p-surge2.8.pdf) to use the real-time best track intensity, latitude, forward speed, and RMW to derive an initial minimum pressure consistent with the SLOSH parametric wind profile during the "spinup" portion of the forecast. Results from the P-Surge v2.8 scientific evaluation indicated this approach led to more reliable storm surge probabilities (not shown), since storm surge forecasts appear to be more sensitive to small changes in RMW than minimum central pressure. While this upgrade led to better storm surge forecasts by improving the initial storm structure, the forecast methodology for RMW did not change for P-Surge v2.8. As a result, the RMW forecasts were often unrealistic, such that the three-member ensemble failed to encompass the NHC estimates of RMW.

b. RMW forecasts

Since storm surge prediction relies on accurately representing storm size, it was desirable to try to improve the method of forecasting RMW. To do this, the historical relationships between RMW and other best track parameters available in the NHC OFCL forecast were examined. Although RMW values were not quality controlled as part of the poststorm analysis until 2021, these data have been archived in the NHC best-track files since 2001 and are provided to the nearest 5 n mi. Since it is reasonable to assume that there is less uncertainty in the best-track RMW at times coincident with aircraft reconnaissance, the RMW dataset used in this study were limited to synoptic times within ±120 min of aircraft reconnaissance (i.e., center fix). In addition, only storms with a classification of tropical depression, tropical storm, hurricane, subtropical depression, or subtropical storm were included, which are the same set of classifications used to select cases included in the NHC verification reports (e.g., Cangialosi 2021). This yielded a sample size of 1649 cases in the Atlantic basin from 2001 to 2019 (Fig. 3).

Willoughby and Rahn (2004) used aircraft reconnaissance data from 23 hurricane seasons (1977–2000) to evaluate the Holland (1980) hurricane wind model and found that latitude and intensity could be used to predict RMW. However, their statistical relationship only explained ~25% of the variance. Vickery and Wadhera (2008) also used flight level data and H*Wind "snapshots" (Powell et al. 1998) to develop a statistical model for RMW based on the central pressure deficit and latitude. Knaff et al. (2015) also developed a linear regression model for RMW that depends on latitude and intensity to estimate tropical cyclone flight-level winds. While these previous studies focused on flight-level winds, the RMW in this study is obtained from the best-track dataset and is defined as the radius of the maximum 1-min averaged wind at 10-m elevation.

TABLE 1. Sample size N, coefficient of determination R^2, and coefficients a_0–a_6 used to forecast RMW as a function of forecast hour (label fhr) when all forecast wind radii are available. The predictors are listed under the coefficients a_1–a_6, where $\ln(\text{RMW}_0)$ is the natural logarithm of the initial RMW; $\ln(R_{34})$, $\ln(R_{50})$, and $\ln(R_{64})$ are the natural logarithms of the nonzero average 34-, 50-, and 64-kt forecast wind radii, respectively; $\ln(V_{\max})$ is the natural logarithm of forecast intensity; ϕ is the forecast latitude; and a_0 is the intercept. The regression coefficients were derived from Atlantic best-track data from 2001 to 2019 that were within ±120 min of aircraft reconnaissance (see the text for details).

fhr	N	R^2	a_0	a_1 $\ln(\text{RMW}_0)$	a_2 $\ln(R_{34})$	a_3 $\ln(R_{50})$	a_4 $\ln(R_{64})$	a_5 $\ln(V_{\max})$	a_6 ϕ
0	634	1.000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
12	631	0.749	3.1894	0.3524	0.1208	−0.1091	0.5862	−0.8070	0.0057
24	630	0.703	4.4373	0.1473	0.1045	−0.1112	0.7566	−1.0689	0.0061
36	620	0.695	4.9447	0.0784	0.1168	−0.1448	0.8246	−1.1709	0.0059
48	599	0.695	5.1818	0.0549	0.1335	−0.2345	0.8972	−1.2038	0.0063

Following the methodology of Willoughby and Rahn (2004), Vickery and Wadhera (2008), and Knaff et al. (2015), the relationship between RMW and other best-track variables was investigated to determine whether they could be used to predict RMW. Variables were included as predictors if there was a notable increase in the variance explained for any of the lead times being evaluated. While some predictors did not make a positive contribution at all lead times, we chose to keep the predictor selection constant for simplicity. After several sets of predictors were chosen for further evaluation, coefficients were derived for each year of forecasts from 2001 to 2019 by excluding data from the current year. For example, coefficients for 2018 included data from 2001 to 2017 and 2019. These coefficients were then used to compute sets of RMW forecasts from 2001 to 2019, which were evaluated in terms of error and bias to select the predictors and coefficients for operational forecasting. In addition to intensity and latitude, it was found that the RMW value at the start of the forecast (persistence) and the four-quadrant (nonzero) averages of the 34-, 50-, and 64-kt wind radii are also good predictors of RMW. A log-transformation of the predictand and most of the predictors (all except latitude) yielded the largest coefficient of determination R^2. This resulted in a power-law relationship for predicting RMW. Chavas and Knaff (2022) also use the 34-kt wind radii along with latitude and intensity to predict RMW at 10-m elevation. To estimate RMW, their model computes the loss of angular momentum that is expected when moving radially inward from the 34-kt wind radius to the RMW. Their methodology shows a lot of potential but was not available for a comparison at the time that this RMW prediction model was being developed.

Since wind radii forecasts are not always available (a storm's intensity may be below the corresponding wind speed threshold, or the lead time is beyond when wind radii forecast data are available), four sets of regression equations were constructed to predict RMW depending on which forecast wind radii are available (Tables 1–4). For example, the equation to predict RMW at the 36-h lead time when all wind radii forecast predictors are available is

$$\text{RMW}_t = \text{RMW}_0^{0.078} R_{34}^{0.117} R_{50}^{-0.145} R_{64}^{0.825} V_{\max}^{-1.171} e^{(4.9 + 0.006\phi)},$$

where RMW_0 is the initial RMW (n mi); R_{34}, R_{50}, and R_{64} are the average (nonzero) 34-, 50-, and 64-kt forecast wind radii (n mi), respectively; $V_{\max}$ is the forecast intensity (kt); and ϕ is the forecast latitude (°). The R^2 value is largest when all forecast wind radii are available, although the number of forecasts used to derive the coefficients is much smaller than when no wind radii data are used to predict RMW (i.e., cf. Tables 1 and 4).

Although wind radii estimates were not included in the post-storm analyses prior to 2004 (Cangialosi and Landsea 2016), there is reason to believe that the wind radii values listed in the best track are fairly reliable between 2001 and 2004 given the availability of QuikSCAT scatterometer data starting in 2000 (e.g., Knaff et al. 2021). Since 2001, the 34- and 50-kt wind radii forecasts were available out to the 72-h lead time. Between 2001 and 2018, the 64-kt wind radii forecasts were available out to the 36-h lead time, but beginning in 2019, they were extended to the 48-h lead time. Initial testing revealed there were often discontinuities in the predicted RMW values when the availability of the wind radii predictors changed during the forecast. This was especially an issue at the 84-h lead time when the 72-h RMW forecast that relied on 34- and 50-kt wind radii was not consistent with the 84-h forecast computed without wind radii data. To make the forecast RMW values more consistent, the 34- and 50-kt wind radii forecast values were held constant from 72 to 120 h if the forecast intensity was above the corresponding thresholds.

TABLE 2. As in Table 1, but when only the 34- and 50-kt forecast wind radii are available.

fhr	N	R^2	a_0	a_1 $\ln(\text{RMW}_0)$	a_2 $\ln(R_{34})$	a_3 $\ln(R_{50})$	a_5 $\ln(V_{\max})$	a_6 ϕ
0	969	1.000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
12	958	0.750	3.1131	0.3680	0.1589	0.4710	−0.9111	0.0068
24	926	0.693	4.1567	0.1834	0.2085	0.5873	−1.1841	0.0073
36	876	0.673	4.6694	0.1062	0.2330	0.6295	−1.3122	0.0074
48	808	0.655	4.9434	0.0459	0.3027	0.5828	−1.3675	0.0079
72	674	0.652	4.7906	0.0157	0.3953	0.5321	−1.3617	0.0067

TABLE 3. As in Table 1, but when only the 34-kt forecast wind radii are available.

fhr	N	R^2	a_0	a_1 ln(RMW$_0$)	a_2 ln(R_{34})	a_5 ln(V_{max})	a_6 ϕ
0	1551	1.000	0.0000	1.0000	0.0000	0.0000	0.0000
12	1484	0.755	2.6272	0.4230	0.6320	−0.9117	0.0064
24	1372	0.684	3.6525	0.2142	0.8222	−1.2158	0.0082
36	1227	0.653	4.2822	0.0884	0.9059	−1.3656	0.0091
48	1099	0.637	4.7700	−0.0042	0.9225	−1.4349	0.0102
72	862	0.633	4.7307	−0.0365	0.9153	−1.3882	0.0086

The first step to compute the RMW forecast is to determine which wind radii forecast data are available (if any), and select the corresponding regression equation. Prior to applying the coefficients, the NHC OFCL forecast parameters (latitude, intensity, and wind radii) are bias corrected based on verification results from the most-recent five years. After applying the bias correction and coefficients, a 3-point smoothing function is used to smooth the forecasts in time since discontinuities can arise if the input parameter availability changes during the forecast. A lower and upper limit of acceptable RMW forecast values is set to 5 and 120 n mi, respectively. The lower limit corresponds to the smallest RMW values listed in the NHC best track, and the upper limit was chosen by considering past storms that exhibited predominantly tropical characteristics.

While the along-track, cross-track, and intensity perturbations used to generate P-Surge ensemble members are computed based on assuming a normal error distribution, that assumption was not made for RMW. To generate the three-member P-Surge RMW ensemble, a cumulative error distribution was constructed from 10 years of retrospective forecasts using the RMW regression equations. Only forecasts initialized over water were included in the statistics, and to provide additional weight to more recent forecasts, the latest 5-yr period of forecasts was double counted in the cumulative error distribution. The ensemble perturbations are then generated based on the 15th (large), 50th (medium), and 85th (small) percentile of the cumulative distribution.

To calculate the total weight for each realization, the cross-track, along-track, intensity, and RMW weights for each realization are then multiplied together. It is important to note that different error statistics are used to generate the P-Surge ensemble perturbations for storms with an initial intensity of <50 kt, 50–95 kt, and >95 kt. Binning the error statistics based on the storm's initial intensity is a step toward better matching the ensemble spread with the expected uncertainty of the current forecast.

TABLE 4. As in Table 1, but when no forecast wind radii are available.

fhr	N	R^2	a_0	a_1 ln(RMW$_0$)	a_5 ln(V_{max})	a_6 ϕ
0	1649	1.000	0.0000	1.0000	0.0000	0.0000
12	1548	0.654	2.1633	0.6360	−0.3314	0.0154
24	1416	0.499	3.7884	0.3953	−0.5738	0.0219
36	1259	0.417	5.0213	0.1999	−0.7481	0.0276
48	1123	0.392	5.8092	0.0615	−0.8508	0.0318
72	875	0.394	6.3321	−0.0362	−0.9079	0.0343
96	691	0.411	6.6181	0.0041	−0.9599	0.0295
120	565	0.382	6.7073	−0.0028	−0.9478	0.0257

3. Verification of RMW forecasts

P-Surge track and RMW forecasts are shown in Fig. 4 for Hurricane Laura (2020) at 0600 UTC 22 August. It is apparent from Fig. 4b that the RMW forecast values from P-Surge v2.7 were much too small at the start of the forecast (~12 n mi vs 100 n mi in the NHC best track). By 48 h, the v2.7 RMWs were in better agreement with the best-track values. Laura had intensified considerably by this point and the RMW had decreased to about 25 n mi. Since P-Surge v2.8 is initialized using the best track, the RMWs from v2.8 (Fig. 4c) are in better agreement with the best-track RMW during the early stages of the forecast. However, the RMWs remain much too large throughout the forecast period. The RMWs from P-Surge v2.9 (Fig. 4d) do a much better job of encompassing the verifying RMW for this forecast, since the evolving characteristics of the storm are used to predict RMW.

Verification results from five years (2015–19) of retrospective RMW forecasts are shown for the medium-sized RMW ensemble members in Fig. 5. The v2.7 forecasts have the largest MAE (Fig. 5a), which is not surprising since, as discussed previously, the RMW values in v2.7 are derived from the less sophisticated method of using the pressure–wind–RMW relationship in the SLOSH model. The RMW forecasts from v2.8 are much improved from those of v2.7 during the early part of the forecast since the RMW forecast is initialized from the best-track value. By 60 h, the MAE of v2.8 is very similar to v2.7, likely because both versions use the same forecast method that depends only on the initial RMW. While the MAE of the v2.9 forecasts is very similar to v2.8 at 0 and 12 h, the v2.9 forecasts are considerably better than v2.8 throughout the remainder of the forecast period.

A comparison of the bias in the RMW forecasts (Fig. 5b) indicates that there is a considerable reduction in the negative bias from v2.7 to v2.8. The RMW forecasts of v2.7 were, on average, 15 n mi too small at the start of the forecast. The v2.8 and v2.9 RMW forecasts also exhibit a negative (small) bias, but to a far lesser degree. Relative to v2.8, the greatest improvement in bias for v2.9 occurs from 48 h onward.

It is important to note that changes to RMW also affect the outer wind profile (i.e., $r >$ RMW) since the SLOSH parametric wind model does not allow for the outer wind profile to be adjusted separately via a size parameter. To evaluate how the changes to RMW affected the outer wind profile, the 34-kt wind radii from the SLOSH parametric wind model were

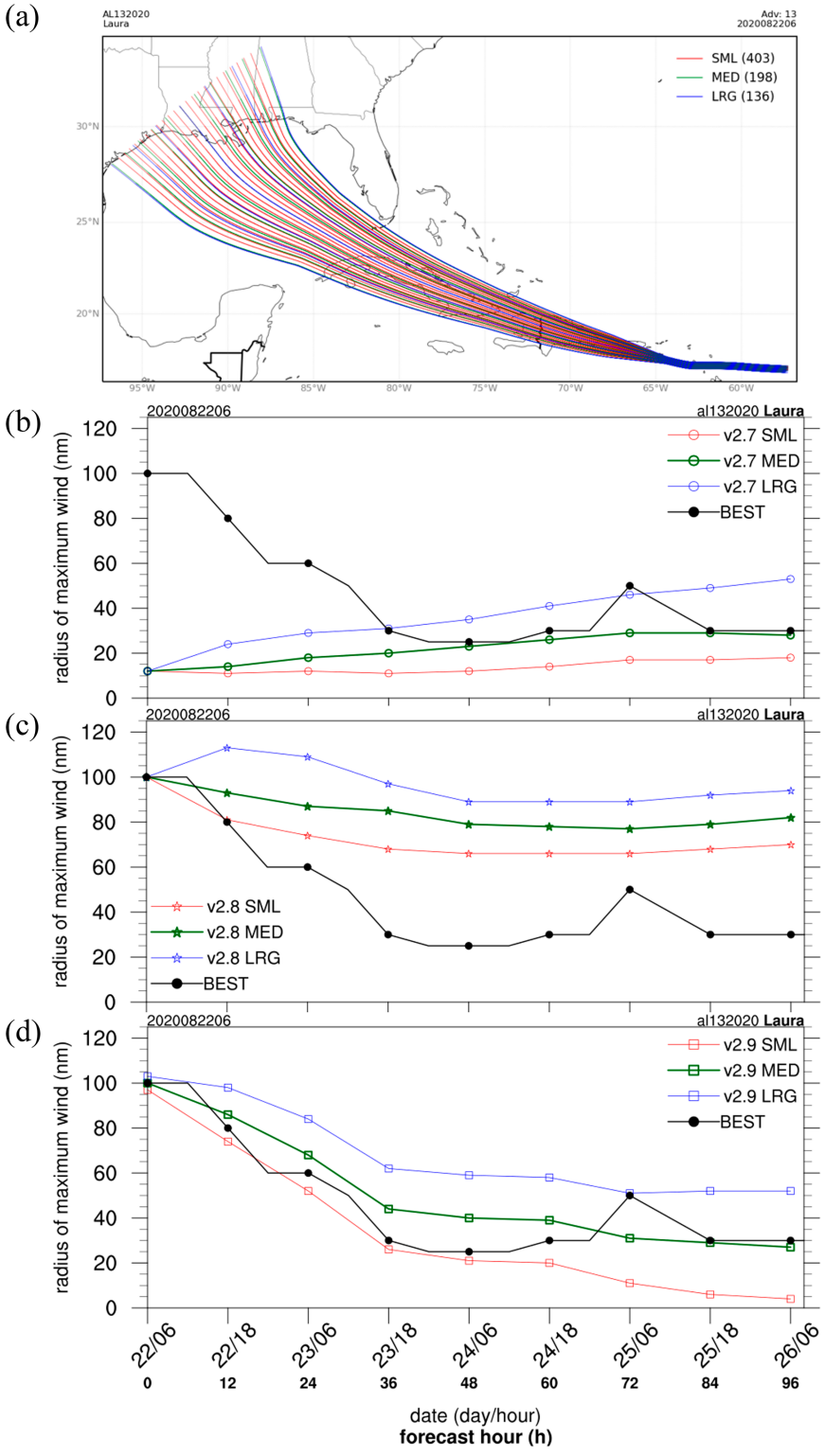

FIG. 4. Operational P-Surge (a) track and (b)–(d) RMW forecasts from different P-Surge versions for Hurricane Laura (2020) initialized on 0600 UTC 22 Aug 2020. The red, green, and blue track forecasts in (a) correspond to the small, medium, and large storms, respectively, that make up the P-Surge RMW ensemble. The number of tracks in each RMW group is shown in the legend in (a). The small-, medium-, and large-RMW forecasts are shown for P-Surge v2.7 in (b), P-Surge v2.8 in (c), and P-Surge v2.9 in (d), along with the verifying best-track RMW (black).

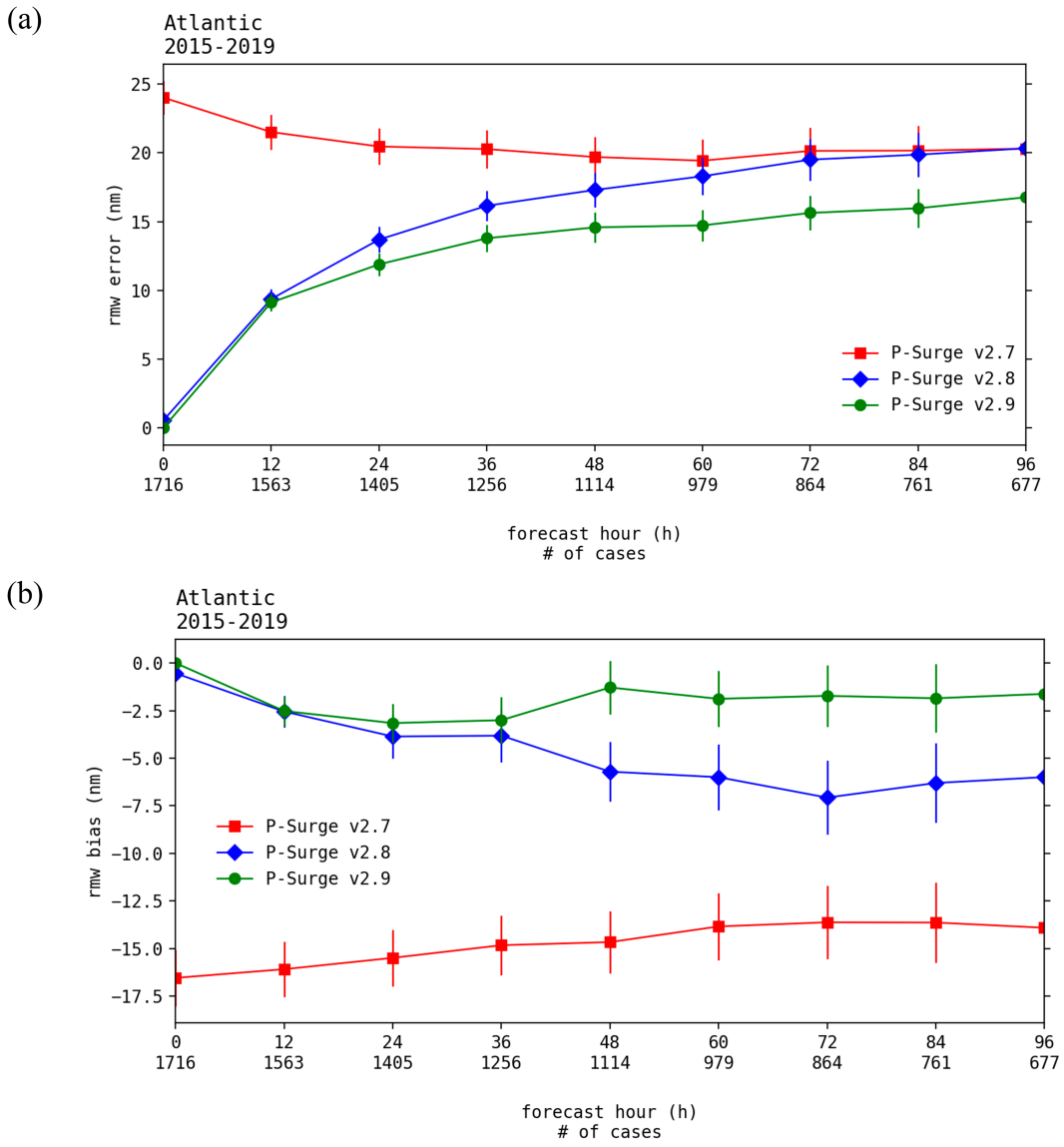

FIG. 5. (a) Mean absolute error and (b) bias of retrospective RMW forecasts from 2015 to 2019 from P-Surge v2.7 (red), v2.8 (blue), and v2.9 (green). Error bars indicate the 95% confidence interval.

compared with the four-quadrant (nonzero) average 34-kt wind radii from the best track for P-Surge v2.7, v2.8, and v2.9 (Fig. 6). Storm motion was not accounted for in the SLOSH parametric wind profiles, so the MAE and bias of the P-Surge versions should be compared in a relative sense. The MAE of the 34-kt wind radii (Fig. 6a) is smallest for v2.9 at all lead times. The MAE of v2.8 is similar to v2.9 initially but increases with lead time and is similar in magnitude to v2.7 by the 60-h lead time. In terms of bias (Fig. 6b), the 34-kt wind radii of v2.7 has a large negative bias at all lead times. The biases of v2.8 and v2.9 are also negative but are much smaller in magnitude. The magnitude of the negative bias for v2.9 is slightly larger than v2.8 at early lead times, but they are similar in magnitude from 48 to 96 h. Not surprisingly, the MAE and bias of the 34-kt wind radii are similar in pattern to the RMW MAE and bias (Fig. 5), indicating that the improvements to RMW also translate to the outer wind field for P-Surge v2.9.

4. Verification of probabilistic storm surge forecasts

To objectively evaluate the three different versions of P-Surge guidance (v2.7, v2.8, and v2.9), we designed an observationally based verification system using a 25-storm dataset from 2008 to 2020. The verification methodology presented here uses only in situ observations from the NOAA Center for Operational Oceanographic Products and Services (CO-OPS; NOAA/NOS CO-OPS 2021) tide stations and USGS (U.S. Geological Survey 2021) water level sensors that are predeployed prior to landfall. CO-OPS provides water level measurements at a fixed network

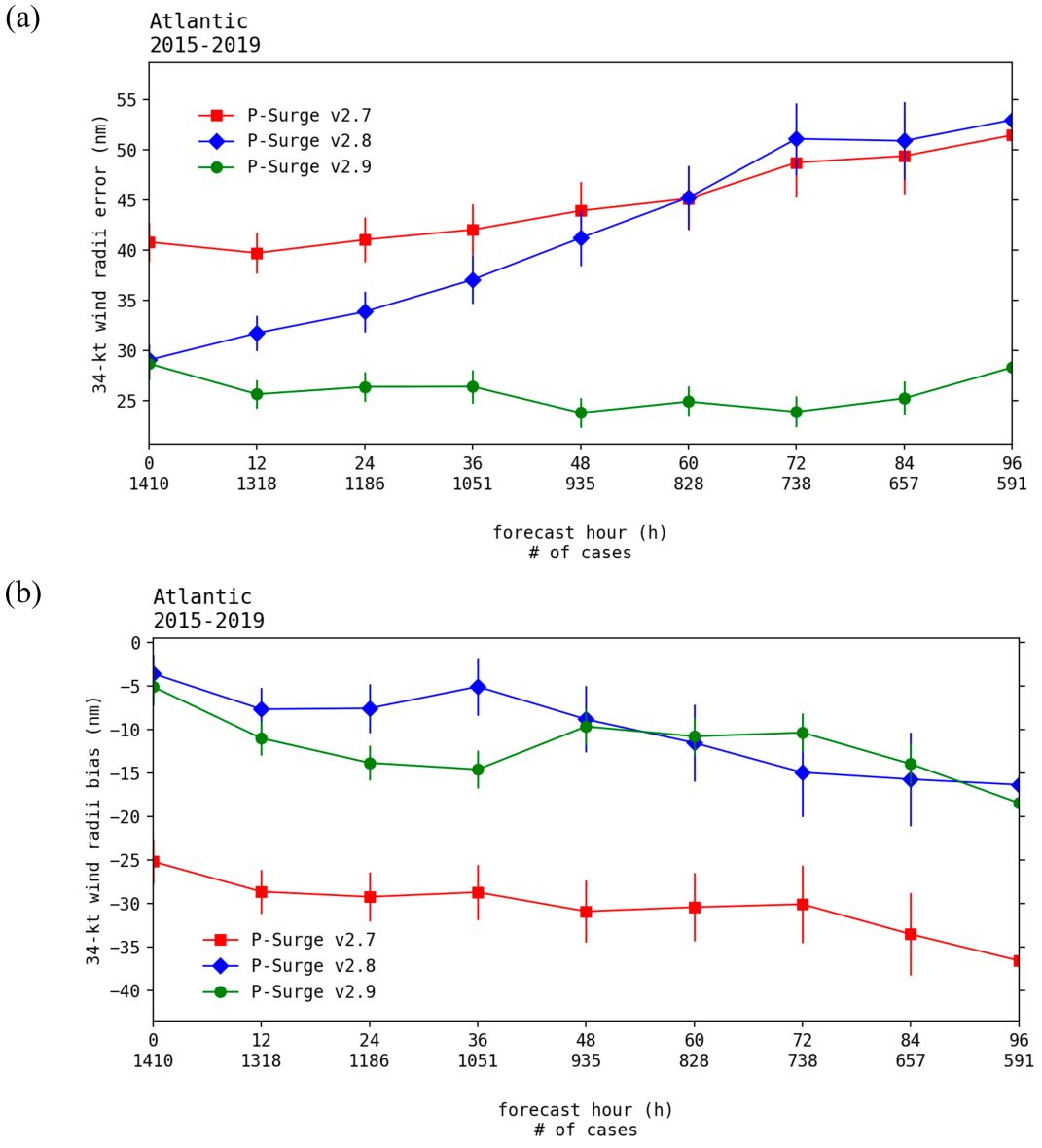

FIG. 6. (a) Mean absolute error and (b) bias of the average 34-kt wind radii based on retrospective RMW forecasts from 2015 to 2019 from P-Surge v2.7 (red), v2.8 (blue), and v2.9 (green). Error bars indicate the 95% confidence interval.

of coastal tide stations (https://tidesandcurrents.noaa.gov). The water level measurements are recorded every ~6 min and were retrieved from CO-OPS referenced to the NAVD88 datum, which allowed for a direct comparison with P-Surge output. The CO-OPS tide station data within ~300 km of the most-heavily impacted areas are included in the verification for an event. Coverage for the USGS water level sensor networks vary for each storm (https://stn.wim.usgs.gov/FEV/). These sensors record water level measurements every ~30 s referenced to the NAVD88 datum. Data were postprocessed by the USGS using a low-pass filter to remove wave noise.

Observations were collected for the duration of each event, and the peak observed value was identified (Table 5). A map and histogram of the observations used in this study are provided in Fig. 7. For context, the NHC generally considers inundation greater than 3 ft above ground level to be life threatening. With no published "ground level" height at each tidal station, the NWS Tropical Cyclone Products Directive recommends using the tidal datum Mean Higher High Water (MHHW) for measuring and communicating inundation at the immediate coastline (NWS 2023). While 0 ft MHHW may be the average highest high tide at a tidal station for a tidal day and serves as a threshold where inundation begins, a water level of 3 ft MHHW is a significant departure and suggests nearby flooding on normally dry ground. For most CO-OPS tidal stations used in this study, water level observations of

TABLE 5. List of forecasts included in the set of P-Surge retrospective forecasts for each case. Two sets of forecasts are analyzed for Irene (2011) and Matthew (2016) since multiple landfalls (or close approaches to land) occurred. Forecasts were selected for evaluation based on the forecast landfall time; forecasts with landfall times closest to the 72-, 60-, 48-, 36-, 24-, and 12-h lead times were included in the set of retrospectives. Forecast advisory numbers are listed for each forecast along with the assigned lead time prior to landfall. The time of landfall and storm intensity at landfall are from the HURDAT2 database (Landsea and Franklin 2013). The RMW values at landfall were interpolated from the NHC best track to the HURDAT2 landfall times. Maximum water level is the highest observation collected for verification relative to the NAVD88 datum. The number of observations used for each case is also listed in the table. Here, ATCF ID indicates Automated Tropical Cyclone Forecasting system identifier.

		Hours prior to landfall advisory No.									
Storm name	ATCF ID	72	60	48	36	24	12	Landfall (UTC time and date)	Intensity (kt)	RMW (n mi)	Max water level (ft above NAVD88)	No. of observations
Gustav	AL072008	21	23	25	27	29	31	1500 1 Sep	90	25	14.3	47
Ike	AL092008	38	40	41	43	45	47	0700 13 Sep	95	30	16.2	66
Irene[1]	AL092011	19	21	22	24	25	28	1200 27 Aug	75	45	8.8	68
Irene[2]	AL092011	22	24	25	27	29	31	0935 28 Aug	60	100	9.9	105
Isaac	AL092012	21	22	24	26	29	31	0800 29 Aug	70	40	14.1	109
Sandy	AL182012	20	22	24	26	28	29	2330 29 Oct	70	110	16.7	153
Hermine	AL092016	5	7	11	13	15	18	0530 2 Sep	70	25	9.6	9
Matthew[1]	AL142016	26	28	31	33	34	36	0000 7 Oct	115	15	8.4	47
Matthew[2]	AL142016	32	33	35	37	38	40	1500 UTC 8 Oct	75	25	8.7	109
Harvey	AL092017	12	14	16	19	20	22	0300 26 Aug	115	15	9.5	26
Irma	AL112017	34	36	39	42	44	46	1930 10 Sep	100	15	8.3	50
Nate	AL162017	5	7	8	10	12	14	0520 8 Oct	65	25	8.4	22
Florence	AL062018	50	52	54	55	57	59	1115 14 Sep	80	20	11.3	112
Gordon	AL072018	—	1	3	6	8	10	0315 5 Sep	60	20	4.1	9
Michael	AL142018	6	7	9	11	13	15	1730 UTC 10 Oct	140	10	15.6	27
Barry	AL022019	3	4	5	7	9	11	1500 13 Jul	65	40	8.9	14
Dorian	AL052019	40	42	45	47	49	51	1230 6 Sep	85	25	7.8	66
Cristobal	AL032020	15	16	18	20	22	24	2200 7 Jun	45	90	6.9	17
Hanna	AL082020	1	2	4	7	9	11	2200 25 Jul	80	20	6.6	22
Isaias	AL092020	17	19	21	23	25	26	0310 4 Aug	80	20	8.7	44
Laura	AL132020	18	20	22	24	26	28	0600 27 Aug	130	15	10.1	28
Marco	AL142020	—	—	10	13	15	17	1800 24 Aug	40	50	3.6	14
Sally	AL192020	6	8	10	14	17	19	0945 16 Sep	95	20	6.5	18
Beta	AL222020	8	9	11	12	13	16	0245 UTC 22 Sep	45	20	6.3	32
Delta	AL262020	10	11	14	16	18	20	2300 9 Oct	85	20	9.8	32
Zeta	AL282020	5	8	10	12	14	16	2100 28 Oct	100	25	9.2	17
Eta	AL292020	—	—	43	44	45	47	0920 12 Nov	45	30	4.7	7

6 ft NAVD88 are greater than 3 ft MHHW. Observations 6 ft NAVD88 and greater account for approximately 40% of the total observations in this study. Approximately 10% of all observations included in this verification study are greater than 9 ft NAVD88. It is important to note that the NHC does not currently provide a gridded analysis product that combines numerical simulations with storm surge observations as part of its poststorm assessment; otherwise, an analysis of this type could be used for verification purposes.

The quality of storm surge observations can vary considerably across storms. The peak inundation for an event is often missed by in situ networks due to sensors being spaced too far apart and the instrumentation can fail during extreme conditions. High-water marks determined from debris or stain lines can help deduce the peak inundation after an event and can be used to estimate how representative the sensor observations were of the peak values and areal extent. For example, the peak inundation observed by the CO-OPS network during Hurricane Laura (2020) was 9.2 ft MHHW at the Calcasieu Pass (Pasch et al. 2021), which is south of Cameron. However, a high-water mark survey conducted by the USGS in an area hardest hit by Laura's storm surge found evidence of water 17.1 ft above ground near Creole, Louisiana. Conversely, a sensor may be located much closer to the peak of an event. During Hurricane Michael, a USGS water level sensor deployed at Mexico Beach, Florida, reported a filtered water level of 15.55 ft NAVD88, which equates to approximately 14.7 ft MHHW. The wave action added to the damage, leaving Mexico Beach with many buildings reduced to their foundation (i.e., "slab cases"). High-water marks collected from the structures left behind provided additional evidence of 14 ft of inundation (Beven et al. 2019). Unfortunately, high-water mark surveys are not always possible if the area is inaccessible to survey crews, and although high-water mark assessments aim to measure the still water level, they may be contaminated by waves. Because of the subjectivity related to high-water marks, they were not included in this study.

For operational forecasting, P-Surge probabilities are assessed using two different approaches: (i) the probabilities of exceeding certain thresholds (e.g., >3, >6, and >9 ft

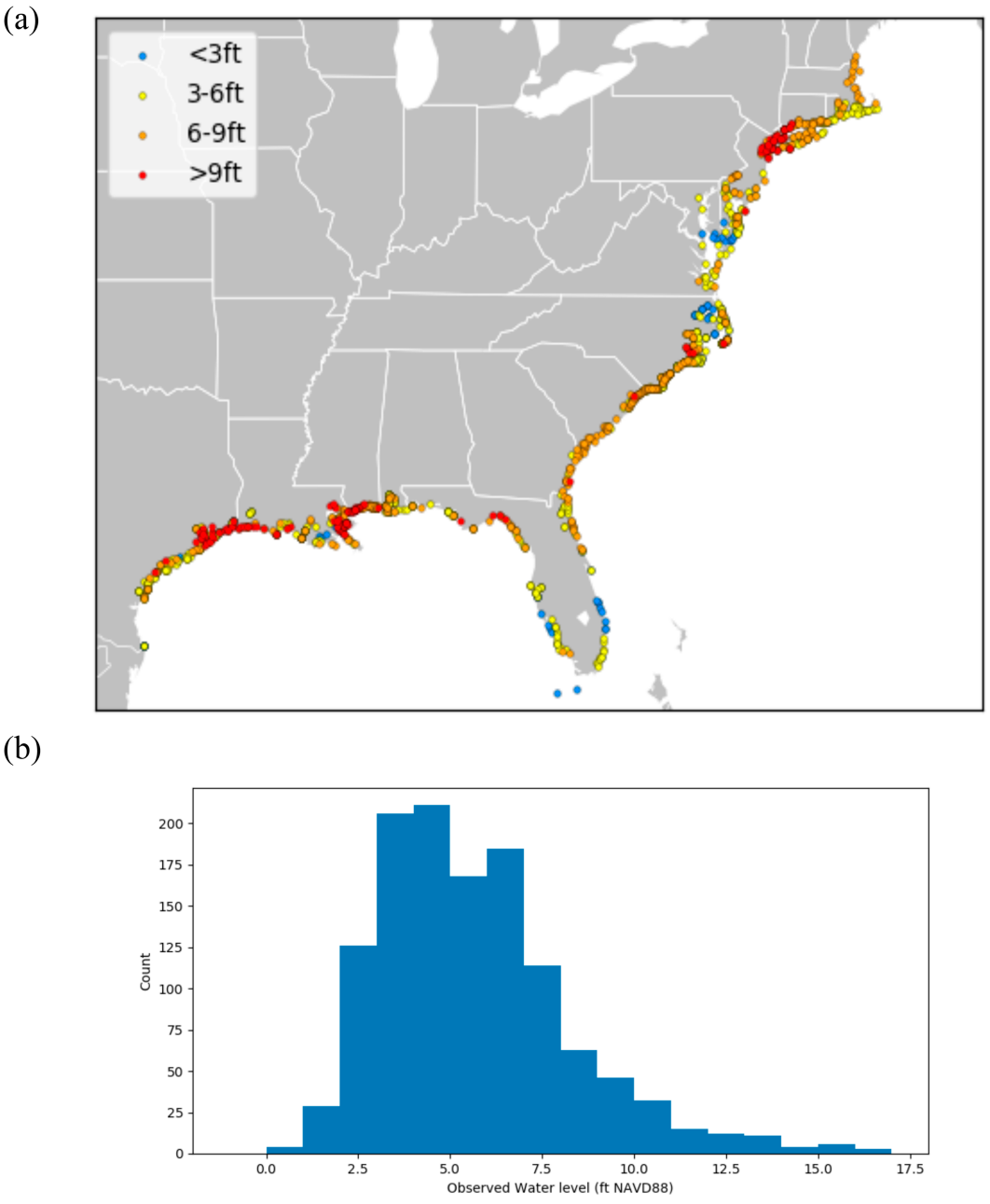

FIG. 7. Storm surge observations used for verification purposes in this study. Observations are from CO-OPS tide stations and USGS water level sensors. (a) Observation locations, colored according to the maximum observed value relative to the NAVD88 datum (see legend). (b) A histogram of the maximum observed values.

NAVD88) are used to assess vulnerability, especially at point specific locations such as levees or other critical infrastructure, and (ii) the storm surge values associated with certain probability of exceedance thresholds. The 10% probability of exceedance values (e.g., Fig. 2a) are considered to be a reasonable worst-case scenario at a given location and are used to generate the potential storm surge flooding map (not shown) and guide the storm surge watch/warning (e.g., Fig. 2b), which are public-facing products available on the NHC website. Because of the relevance to operational forecasting, the evaluation of model performance in this study will focus on the probabilities of exceeding 6 and 9 ft NAVD88, with an emphasis on the 10% exceedance threshold. To reduce errors introduced by datum conversions, the verification set includes observations and P-Surge products relative to the NAVD88 datum only. At present, NHC operational storm surge products do not include timing information. Therefore, for simplicity, the maximum forecast

values are verified against the peak observation at a given location irrespective of time.

Since the time of landfall is often different from one forecast to the next, the P-Surge retrospective runs were chosen based on the time of landfall inferred from the NHC OFCL forecasts, rather than when the forecasts were initialized relative to the actual landfall time. (i.e., the forecasts with landfall closest to the 0-, 12-, 24-, 36-, 48-, 60-, and 72-h lead times were included in the retrospective set for each event). This allows for an evaluation of the forecasts stratified by lead time. For most storms, seven retrospective P-Surge forecasts were conducted for each model configuration included in the evaluation. However, since landfall did not occur in all forecasts and storm formation sometimes occurred within 72 h of landfall, three cases do not have forecasts at all lead times prior to landfall. Furthermore, several additional runs were needed for storms that moved parallel to the coast (i.e., Hurricanes Irene and Matthew). For these cases, landfall was more loosely defined as the time/location of the storm's closest approach to land. Table 5 provides the advisory number corresponding to each lead time as well as the best-track intensity and RMW at landfall for each retrospective case.

An example of how the probability of exceeding 6 ft NAVD88 changes with lead time is shown in Fig. 8 from several P-Surge v2.9 forecasts for Hurricane Laura. At 72 h prior to landfall (Fig. 8a), probabilities between 0% and 20% cover a large area, from approximately Port Fourchon, Louisiana, to Galveston Bay, Texas. Probabilities of 20%–40% extend over western Louisiana, and there is a small area north of Vermilion Bay with probabilities between 40% and 60%. At 60 h prior to landfall (Fig. 8b), the 0%–20% and 20%–40% probability areas are similar to the forecast initialized 12 h prior. However, the area with 40%–60% probabilities has expanded and extends along the coast to near Cameron. By 24 h prior to landfall (Fig. 8c), the probabilities over western Louisiana have increased significantly; there is now a small area with 80%–100% probabilities near the coast between Cameron and Vermilion Bay. Figure 8 illustrates that at longer lead times, probabilities for extreme events are generally low, but will increase for some locations leading up to landfall. This is the nature of probabilistic forecasting. Beyond 72 h, probabilities are generally too low to be meaningful for decision makers, so the NHC advises users to consult storm surge composite risk maps (Zachry et al. 2015).

Point observations were matched to the 625-m gridded P-Surge output by identifying the maximum value from a "neighborhood" of 5 × 5 grid points surrounding the grid point associated with the observation's location. The forecast–observation pairs were used to create relative operating characteristic (ROC) diagrams (Mason 1982) to determine whether the forecast exceedances can discriminate between events and nonevents (i.e., between storm surge greater than 6 ft NAVD88 and storm surge less than 6 ft NAVD88). For this approach, a contingency table is defined with respect to each probability threshold. For example, when evaluating the probability of storm surge greater than 6 ft, the contingency table for the 10% probability of exceedance is defined as follows:

- forecast probability of exceedance $\geq$ 10% and observed storm surge $\geq$ 6 ft is a hit,
- forecast probability of exceedance $<$ 10% and observed storm surge $\geq$ 6 ft is a miss,
- forecast probability of exceedance $\geq$ 10% and observed storm surge $<$ 6 ft is a false alarm, and
- forecast probability of exceedance $<$ 10% and observed storm surge $<$ 6 ft is a correct negative.

The probability of detection (POD) and false alarm rate (FAR) can be calculated from the contingency table as POD = hits/(hits + misses) and FAR = false alarms/(false alarms + correct negatives). The POD–FAR pairs from all storms are plotted for increasing probability of exceedance thresholds to create ROC curves for each P-Surge version (Fig. 9). The ROC area under the curve (AUC) is used as a measure of skillfulness (e.g., Wilks 2006), where a perfect score is equal to 1 and a score below 0.5 is considered a "no skill" forecast. In general, the AUC increases as the lead time decreases, which is to be expected, since in general, the error of the input forecasts of track, intensity, and RMW increase with lead time. Improvements in skill for v2.8 relative to v2.7 are most notable at 12–36 h prior to landfall, but skill in both versions begins to drop off after 48 h as the POD decreases. Improvements in v2.9 over v2.8 and v2.7 are most evident 48–72 h prior to landfall when the POD is higher at comparable exceedance thresholds.

Figure 10 provides the 60- and 72-h ROC comparisons and indicates that P-Surge v2.9 has a greater skill (AUC) than v2.7 and v2.8 at both of these lead times. Confidence intervals were calculated at the 95% confidence level for the 10% and 30% exceedance thresholds following Wilks (2006), and reveal that the differences between the POD/FAR pairs of v2.9 and v2.7 are statistically significant. Looking at the 10% probability of exceeding 6 ft NAVD88 for 60 h prior to landfall (Fig. 10a), the POD for v2.9, v2.8, and v2.7 is 0.8, 0.75, and 0.65, respectively. The corresponding FAR pair is highest in v2.9, but it is important to consider that the NHC's mission of protecting life is more sensitive to misses than false alarms. While there is increased skill for v2.9 over the other versions at 72 h (Fig. 10b), the POD drops off quickly after the 10% probability of exceedance threshold, resulting in a reduction of the AUC relative to at 60 h. Not only does the AUC increase in v2.9 relative to v2.7 and v2.8, but the location of the various probability of exceedance thresholds changes on the ROC diagram between versions. For example, at 60 h the 20% probability of exceeding 6 ft NAVD88 in v2.9 has a similar FAR, but a higher POD when compared with the 10% probability of exceedance in v2.7 (Fig. 10a). The ROC curves, like those in Figs. 9 and 10, can help guide the SSU's decision to use a particular exceedance value when determining storm surge watches and warnings. Currently, the 10% exceedance value is used in an attempt to reduce overwarning (i.e., false alarms), without compromising the POD.

While the ROC diagram describes the ability of the model to discriminate between events and nonevents, the Threat Score (TS) indicates how well the "yes" forecasts correspond to observed events. The TS metric falls out of the contingency table discussed above:

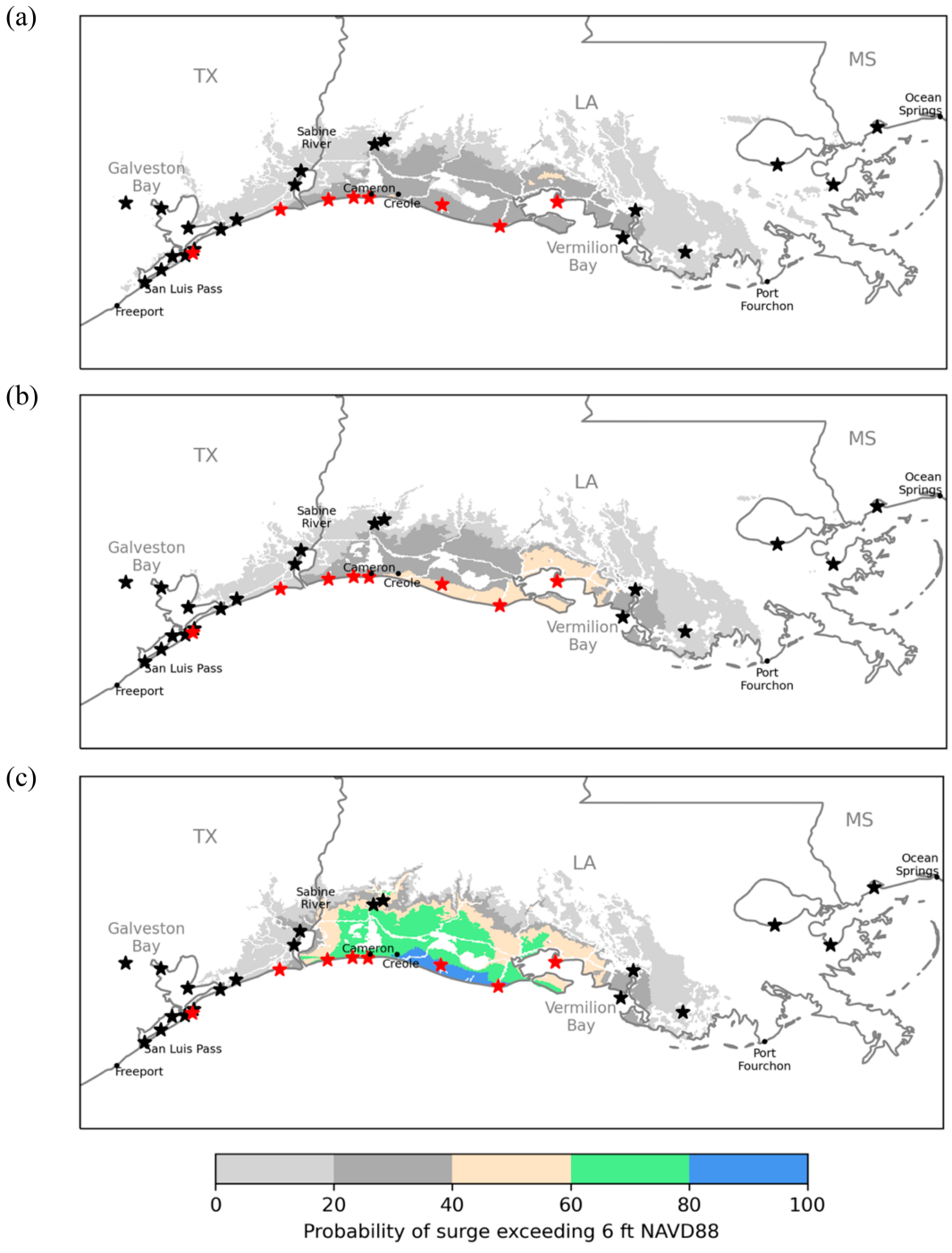

FIG. 8. Probability of exceeding 6 ft NAVD88 from P-Surge v2.9 forecasts for Hurricane Laura (2020) initialized at (a) 0600 UTC 24 Aug (~72 h prior to landfall), (b) 1800 UTC 24 Aug (~60 h prior to landfall), and (c) 0600 UTC 26 Aug (~24 h prior to landfall). Red and black stars respectively indicate observations greater than and less than 6 ft NAVD88.

$$\mathrm{TS} = \mathrm{hits}/(\mathrm{hits} + \mathrm{misses} + \mathrm{false\ alarms}),$$

and a perfect score is equal to 1. Figure 11 shows the 48-, 60-, and 72-h TS associated with the probability of storm surge exceeding 6 ft NAVD88. At these lead times, the v2.9 forecasts have the largest TS at almost all probability thresholds, followed by v2.8 and v2.7. A similar pattern exists at shorter lead times (not shown). The maximum TS for v2.9 occurs at the 30%, 20%, and 10% probability of exceedance for the 48-, 60- and 72-h lead times, respectively, while the maximum TS

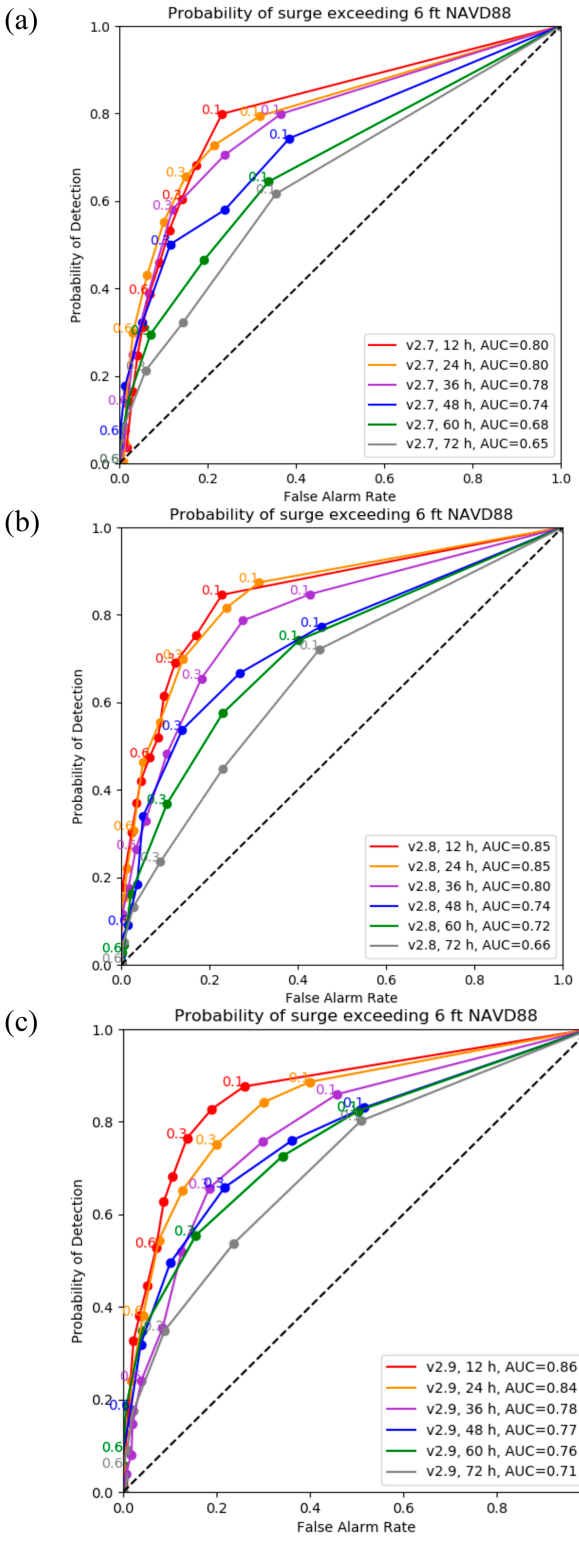

FIG. 9. Relative operating characteristic curves for various probability thresholds of exceeding the 6 ft NAVD88 (e.g., 0.1 = 10% probability of exceedance) for (a) P-Surge v2.7, (b) P-Surge v2.8, and (c) P-Surge v2.9. The ROC curves are colored by the lead times shown in the legend. The *x* axis is the false alarm rate, and the *y* axis is the probability of detection.

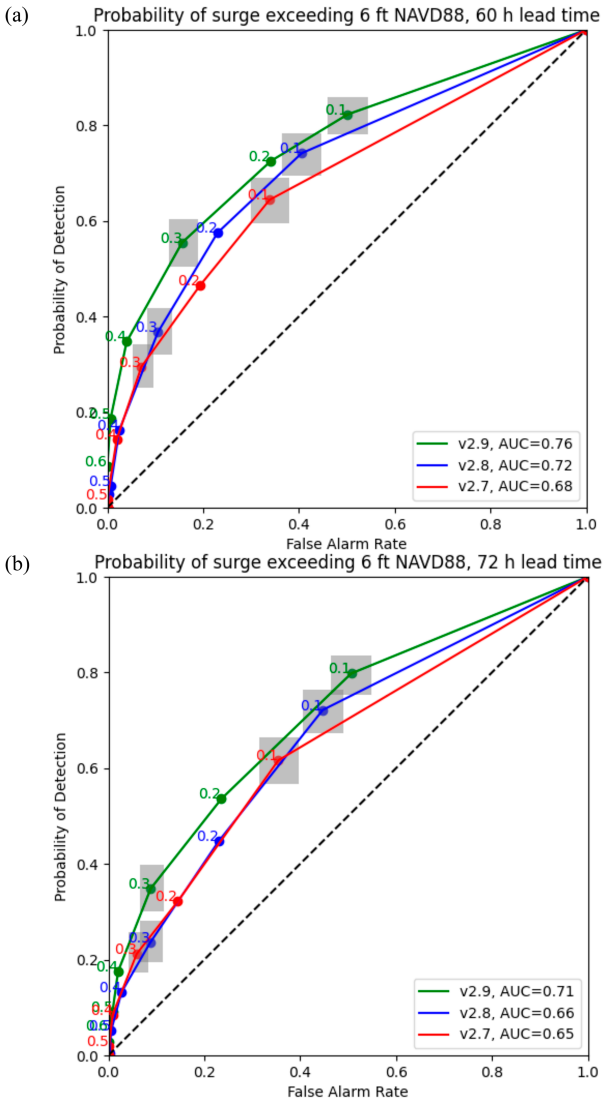

FIG. 10. As in Fig. 9, but results from each P-Surge version are shown on the same figure (see legend) for (a) 60 and (b) 72 h prior to landfall. Gray boxes indicate the 95% confidence interval (see the text for details).

for v2.7 is at the 10% probability threshold at each of the associated lead times. The improvement in the TS of v2.9 relative to v2.7 is a result of more hits and fewer misses, but v2.9 does exhibit more false alarms. For example, at the 60-h lead time (Fig. 11b) for the 30% probability of exceeding 6 ft NAVD88, the number of hits for v2.9 and v2.7 are 273 and 145, respectively.

Reliability diagrams are used to evaluate the usefulness of the forecast probabilities. Forecast probabilities of an event are binned (0%–20%, 20%–40%, 40%–60%, 60%–80%, and 80%–100%), then the corresponding observed relative frequency is calculated for each bin. Perfect reliability occurs when the observed relative frequency and forecast probabilities are equal, resulting in a line along the diagonal in the reliability

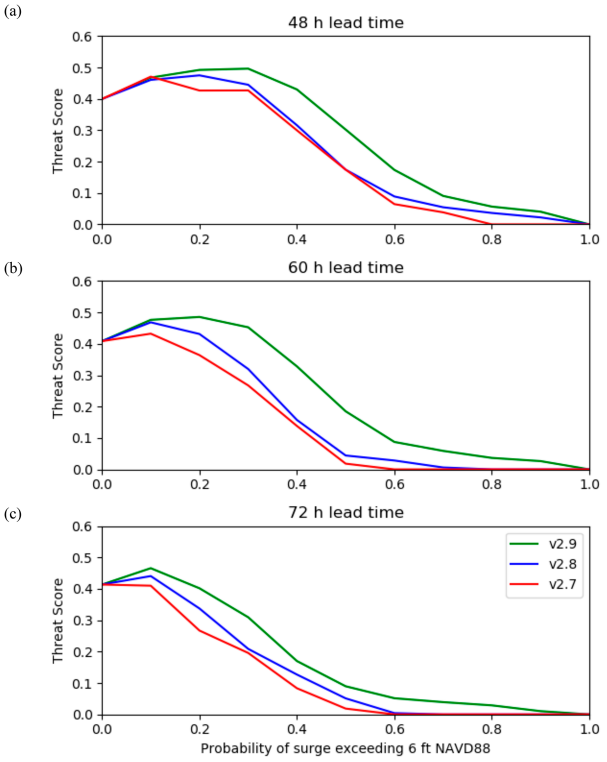

FIG. 11. Threat score for a range of probabilities of exceeding the 6 ft NAVD88 (e.g., 0.1 = 10% probability of exceedance) at (a) 48-, (b) 60-, and (c) 72-h lead times for P-Surge v2.7 (red), v2.8 (blue), and v2.9 (green).

diagram. Figure 12 shows reliability diagrams for both the probabilities of storm surge greater than 6 and 9 ft NAVD88 at 60 h prior to landfall. While all versions tend to underforecast (i.e., forecast probabilities are lower than the observed relative frequency), there is improvement in the v2.9 forecast reliability relative to v2.7 at 60 h prior to landfall. Note that v2.9 more frequently forecasts high probabilities of greater than 6 and 9 ft NAVD88 (see the bin counts at the bottom of each diagram in Fig. 12).

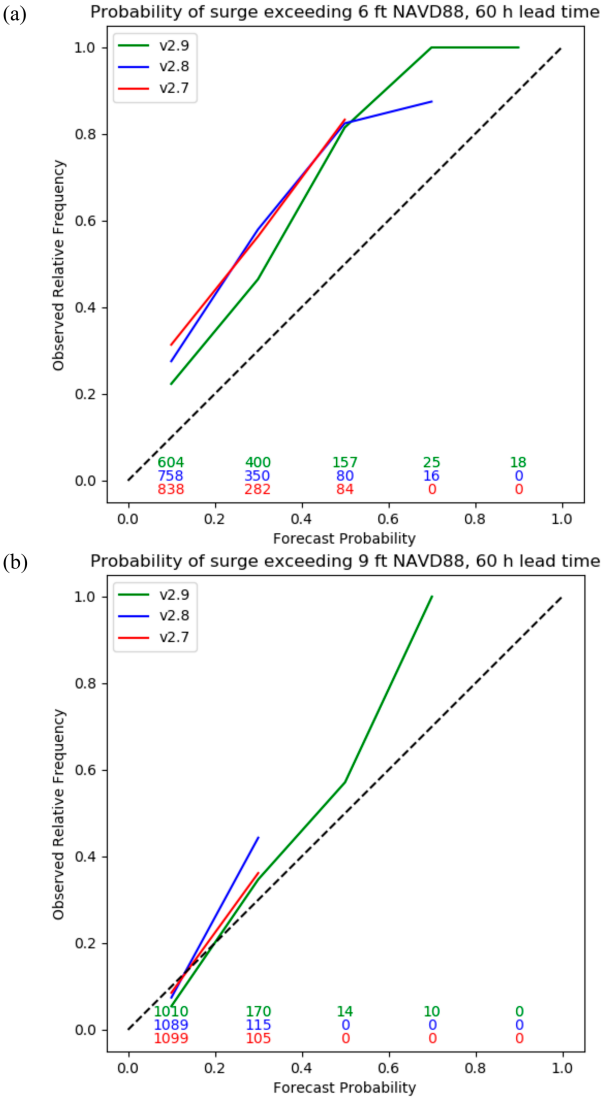

FIG. 12. Reliability diagrams for the probability of exceeding the (a) 6 ft NAVD88 and (b) 9 ft NAVD88 for forecasts 60 h prior to landfall from the different P-Surge configuration included in the evaluation (see legend).

5. Discussion and conclusions

Past studies have shown that storm size can have an important effect on the resulting storm surge (e.g., Irish et al. 2008). Previous versions of P-Surge relied on a relationship between storm intensity, latitude, forward speed, minimum pressure, and RMW that often resulted in RMW values that did not match up well with NHC best-track values. Based on forecasting experience and participation in U.S. Integrated Ocean Observing System (IOOS) community modeling testbeds (Kerr et al. 2013; Joyce et al. 2019), it became a priority to improve the storm size forecasts used by P-Surge. While the upgrade to P-Surge v2.8 did result in improvements to the storm structure at early lead times by using RMW values from the NHC best track to initialize the P-Surge forecast, there was little benefit at medium to extended lead times, since the RMW forecast methodology was not changed with the v2.8 upgrade.

To improve how storm size is represented in the P-Surge forecasts, regression equations for RMW were developed that use NHC OFCL forecast parameters as predictors, including latitude, intensity, and wind radii. Verification results from a 5-yr sample (2015–19) of retrospective RMW forecasts indicate that the RMW forecasts based on NHC OFCL forecast parameters are more accurate in terms of MAE and bias than the RMW forecasts used in previous versions of P-Surge (v2.8 and prior). Based on a comparison of the MAE of the average 34-kt wind radii, these RMW forecasts also result in an improvement to the SLOSH parametric wind profile that extends beyond the RMW.

To evaluate whether the improved RMW forecasts would translate to improvements in P-Surge, retrospective forecasts were selected for 25 storms between 2008 and 2020. Forecasts with landfall times closest to the forecast hours of 72, 60, 48, 36, 24, and 12 h were included in the analysis for each case. Observations from CO-OPS tide stations and USGS water level sensors were used to verify forecasts from three different versions of P-Surge: (i) RMW initialization based on the SLOSH parametric wind profile with RMW forecasts that were a function only of the initial RMW (P-Surge v2.7), (ii) initialization of RMW from the NHC best track with RMW forecasts that were a function only of the initial RMW (P-Surge v2.8), and (iii) initialization of RMW from the NHC best track with RMW forecasts based on NHC forecast parameters (P-Surge v2.9). P-Surge v2.9 forecasts have a higher POD for events greater than 6 ft NAVD88 for all probability of exceedance thresholds. Based on a comparison of the ROC area under the curve, this corresponds to an increase in skill relative to P-Surge v2.7 and v2.8. P-Surge v2.9 is also more reliable for forecasts greater than 6 ft NAVD88. In May 2021, the operational version of P-Surge was upgraded to incorporate the improved RMW forecasts.

An encouraging difference between P-Surge v2.9 and the previous versions included in the comparison is that v2.9 had a greater number of high-probability forecasts for life threatening events (greater than 6 and 9 ft NAVD88) at longer lead times (60–72 h). While false alarms are still an issue, the ability to forecast such events at longer lead times is a significant improvement and, it is hoped, will lead to an increase in forecaster confidence when a high-impact event is likely to occur. With effective messaging, it is perhaps more desirable that the POD increase at the expense of a higher FAR, rather than having a reduction in both the FAR and POD.

A major challenge related to evaluating the P-Surge forecasts is the limited number of observations that were available for verification. For some storms, there were only a handful of sensors within the area most heavily affected by storm surge (see Table 5). Clearly, a denser network of high-quality storm surge observations is needed. Therefore, the verification results pertaining to the peak forecast values need to be interpreted conservatively. Furthermore, this undersampling issue also made it impractical to verify the areal extent of the different probability of exceedance thresholds, especially since most of the observations were located very near the coast.

Efforts are under way to improve the skill and reliability of the P-Surge forecasts at longer lead times. For high-impact events, these improvements will provide additional time for emergency managers and the public to prepare for the approaching hazards and deal with evacuations. One avenue for improvement is to incorporate dynamic uncertainty information when generating the P-Surge ensemble. This information could be used directly to generate the ensemble, or in a hybrid manner similar to that described by DeMaria et al. (2013), where the degree of uncertainty in the current forecast determines which subset of the historical NHC OFCL error distribution to sample. This would allow computational resources to be focused on input parameters (e.g., storm intensity) with greater uncertainty.

Additionally, enhancements are needed so the outer region of the input wind profile can be adjusted independently of the RMW, and to enable SLOSH to represent storms that have an asymmetric wind structure. At present, the parametric wind profile used by SLOSH is not able to account for wind asymmetries apart from those related to a storm's forward motion, and the RMW regression forecasts often struggle with these types of storms, which typically occur during the early or late part of the hurricane season. For storms with an asymmetric wind structure, the SSU typically relies on the guidance from the Probabilistic Extratropical Storm Surge (P-ETSS) model (https://vlab.noaa.gov/web/mdl/petss; Liu and Taylor 2020), which uses input from the North American Ensemble Forecast System (NAEFS; Zhu et al. 2012).

Other avenues for potential improvement include implementing some of the newer SLOSH model grids in P-Surge, which cover a larger geographic extent than their predecessors and often have improved grid spacing in critical areas. Larger grids have been shown to improve the representation of large-scale (nonlocal) processes (e.g., geostrophic currents) (Blain et al. 1994; Kerr et al. 2013). There is also work under way to couple the SLOSH model to a simplified wave model. Accounting for waves is especially important for locations that lack a broad continental shelf (Joyce et al. 2019), such as Puerto Rico, the Virgin Islands, and southeast Florida. To incorporate either the wave model or newer SLOSH model grids into P-Surge, code optimization is needed to ensure the P-Surge forecasts are completed within the operational time window available on NOAA's supercomputer. In addition, testing is under way to incorporate nonuniform bottom slip coefficients (Manning's n values) (e.g., Zhang et al. 2012). This should provide more realistic surge values near the coast and limit the extent to which flooding from surge occurs inland, which is sometimes overpredicted in the current version.

Acknowledgments. We thank three anonymous reviewers for their constructive feedback during the review process. We also appreciate the time and effort spent by Michael Brennan, Jack Dostalek, Judy Ghirardelli, Wallace Hogsett, John Knaff, and Stephanie Stevenson for providing feedback on an earlier version of this paper. This research of the National Hurricane Center is supported in part by NOAA's Science Collaboration Program and administered by UCAR's Cooperative Programs for the Advancement of Earth System Science (CPAESS) under Award NA21OAR4310383.

Data availability statement. The NOAA/NOS CO-OPS and USGS water level data used in this study are available online (https://tidesandcurrents.noaa.gov and http://water.usgs.gov/floods/FEV/, respectively). Full output from the P-Surge ensemble simulations is too large to archive or to transfer, but output from select forecasts may be provided upon request.

REFERENCES

Beven, J. L., II, R. Berg, and A. Hagen, 2019: Tropical cyclone report: Hurricane Michael (7–11 October 2018). NHC Tech.

Rep. AL142018, 86 pp., https://www.nhc.noaa.gov/data/tcr/AL142018_Michael.pdf.

Blain, C. A., J. J. Westerink, and R. A. Luettich Jr., 1994: The influence of domain size on the response characteristics of a hurricane storm surge model. *J. Geophys. Res.*, **99**, 18 467–18 479, https://doi.org/10.1029/94JC01348.

Cangialosi, J. P., 2021: Forecast verification report: 2020 Hurricane Season. NHC Tech. Rep., 77 pp., http://www.nhc.noaa.gov/verification/pdfs/Verification_2020.pdf.

——, and C. W. Landsea, 2016: An examination of model and official National Hurricane Center tropical cyclone size forecasts. *Wea. Forecasting*, **31**, 1293–1300, https://doi.org/10.1175/WAF-D-15-0158.1.

Chavas, D. R., and J. A. Knaff, 2022: A simple model for predicting the tropical cyclone radius of maximum wind from outer size. *Wea. Forecasting*, **37**, 563–579, https://doi.org/10.1175/WAF-D-21-0103.1.

Davis, J. R., V. A. Paramygin, D. Forrest, and Y. P. Sheng, 2010: Toward the probabilistic simulation of storm surge and inundation in a limited-resource environment. *Mon. Wea. Rev.*, **138**, 2953–2974, https://doi.org/10.1175/2010MWR3136.1.

DeMaria, M., and Coauthors, 2013: Improvements to the operational tropical cyclone wind speed probability model. *Wea. Forecasting*, **28**, 586–602, https://doi.org/10.1175/WAF-D-12-00116.1.

Forbes, C., J. Rhome, C. Mattocks, and A. Taylor, 2014: Predicting the storm surge threat of Hurricane Sandy with the National Weather Service SLOSH model. *J. Mar. Sci. Eng.*, **2**, 437–476, https://doi.org/10.3390/jmse2020437.

Fossell, K. R., D. Ahijevych, R. E. Morss, C. Snyder, and C. Davis, 2017: The practical predictability of storm tide from tropical cyclones in the Gulf of Mexico. *Mon. Wea. Rev.*, **145**, 5103–5121, https://doi.org/10.1175/MWR-D-17-0051.1.

Fritz (Haase), A. T., A. A. Taylor, J. Wang, and J. C. Feyen, 2014: Tidal improvements to the SLOSH model. *12th Symp. on the Coastal Environment*, Atlanta, GA, Amer. Meteor. Soc., 4.1, https://ams.confex.com/ams/94Annual/webprogram/Paper235170.html.

Glahn, B., A. Taylor, N. Kurkowski, and W. A. Shaffer, 2009: The role of the SLOSH model in National Weather Service storm surge forecasting. *Natl. Wea. Dig.*, **33**, 3–14.

Gonzalez, T. D., and A. A. Taylor, 2018: Development of the NWS' probabilistic tropical storm surge model. *33rd Conf. on Hurricanes and Tropical Meteorology*, Ponte Vedra, FL, Amer. Meteor. Soc., 186, https://ams.confex.com/ams/33HURRICANE/webprogram/Paper340247.html .

Haase, A., J. Wang, A. Taylor, and J. Feyen, 2011: Coupling of tides and storm surge for operational modeling on the Florida coast. *Proc. 12th Int. Conf. on Estuarine and Coastal Modeling*, St. Augustine, FL, American Society of Civil Engineers, 230–238, https://doi.org/10.1061/9780784412411.00014.

Holland, G. J., 1980: An analytic model of the wind and pressure profiles in hurricanes. *Mon. Wea. Rev.*, **108**, 1212–1218, https://doi.org/10.1175/1520-0493(1980)108<1212:AAMOTW>2.0.CO;2.

Irish, J. L., D. T. Resio, and J. J. Ratcliff, 2008: The influence of storm size on hurricane surge. *J. Phys. Oceanogr.*, **38**, 2003–2013, https://doi.org/10.1175/2008JPO3727.1.

Jelesnianski, C. P., and A. D. Taylor, 1973: A preliminary view of storm surges before and after storm modifications. NOAA Tech. Memo. ERL WMPO-3, 33 pp.

——, J. Chen, and W. A. Shaffer, 1992: SLOSH: Sea, lake, and overland surges from hurricanes. NOAA Tech. Rep. NWS 48, 65 pp., https://repository.library.noaa.gov/view/noaa/7235/noaa_7235_DS1.pdf.

Joyce, B. R., J. Gonzalez-Lopez, A. J. Van der Westhuysen, D. Yang, W. J. Pringle, J. J. Westerink, and A. T. Cox, 2019: U.S. IOOS coastal and ocean modeling testbed: Hurricane-induced winds, waves, and surge for deep ocean, reef-fringed islands in the Caribbean. *J. Geophys. Res. Oceans*, **124**, 2876–2907, https://doi.org/10.1029/2018JC014687.

Kerr, P. C., and Coauthors, 2013: U.S. IOOS coastal and ocean modeling testbed: Inter-model evaluation of tides, waves, and hurricane surge in the Gulf of Mexico. *J. Geophys. Res. Oceans*, **118**, 5129–5172, https://doi.org/10.1002/jgrc.20376.

Knaff, J. A., S. P. Longmore, R. T. DeMaria, and D. A. Molenar, 2015: Improved tropical-cyclone flight-level wind estimates using routine infrared satellite reconnaissance. *J. Appl. Meteor. Climatol.*, **54**, 463–478, https://doi.org/10.1175/JAMC-D-14-0112.1.

——, and Coauthors, 2021: Estimating tropical cyclone surface winds: Current status, emerging technologies, historical evolution, and a look to the future. *Trop. Cyclone Res. Rev.*, **10**, 125–150, https://doi.org/10.1016/j.tcrr.2021.09.002.

Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, https://doi.org/10.1175/MWR-D-12-00254.1.

Liu, H., and A. Taylor, 2020: Latest developments in the NWS probabilistic extratropical storm surge model. *18th Symp. on the Coastal Environment*, Boston, MA, Amer. Meteor. Soc., 4.2, https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/370559.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

Mayo, T., and N. Lin, 2019: The effect of the surface wind field representation in the operational storm surge model of the National Hurricane Center. *Atmosphere*, **10**, 193, https://doi.org/10.3390/atmos10040193.

NOAA/NOS CO-OPS, 2021: Tides and currents. NOAA, accessed 12 January 2021, https://tidesandcurrents.noaa.gov.

NWS, 2023: Weather forecast office tropical cyclone products. Tropical Cyclone Weather Services Program, NWSPD 10-6, National Weather Service Instruction 10-601, 80 pp., https://www.nws.noaa.gov/directives/sym/pd01006001curr.pdf.

Pasch, R. J., R. Berg, D. P. Roberts, and P. P. Papin, 2021: Tropical cyclone report: Hurricane Laura (20–29 August 2020). NHC Tech. Rep. AL132020, 75 pp., https://www.nhc.noaa.gov/data/tcr/AL132020_Laura.pdf.

Powell, M. D., S. H. Houston, L. R. Amat, and N. Morisseau-Leroy, 1998: The HRD real-time hurricane wind analysis system. *J. Wind Eng. Ind. Aerodyn.*, **77–78**, 53–64, https://doi.org/10.1016/S0167-6105(98)00131-7.

Ramos-Valle, A. N., E. N. Curchitser, and C. L. Bruyère, 2020: Impact of tropical cyclone landfall angle on storm surge along the Mid-Atlantic Bight. *J. Geophys. Res. Atmos.*, **125**, e2019JD031796, https://doi.org/10.1029/2019JD031796.

Rego, J. L., and C. Li, 2009: On the importance of the forward speed of hurricanes in storm surge forecasting: A numerical study. *Geophys. Res. Lett.*, **36**, L07609, https://doi.org/10.1029/2008GL036953.

——, and ——, 2010: Nonlinear terms in storm surge predictions: Effect of tide and shelf geometry with case study from Hurricane Rita. *J. Geophys. Res.*, **115**, C06020, https://doi.org/10.1029/2009JC005285.

Szpilka, C., K. Dresback, R. Kolar, J. Feyen, and J. Wang, 2016: Improvements for the western North Atlantic, Caribbean and Gulf of Mexico ADCIRC tidal database (EC2015). *J. Mar. Sci. Eng.*, **4**, 72, https://doi.org/10.3390/jmse4040072.

Taylor, A. A., and B. Glahn, 2008: Probabilistic guidance for hurricane storm surge. *19th Conf. on Probability and Statistics*, New Orleans, LA, Amer. Meteor. Soc., 7.4, https://ams.confex.com/ams/88Annual/techprogram/paper_132793.htm.

U.S. Geological Survey, 2021: USGS flood event viewer: Providing hurricane and flood response data. Short-term network data portal, USGS, accessed 12 January 2021, http://water.usgs.gov/floods/FEV/.

Vickery, P. J., and D. Wadhera, 2008: Statistical models of Holland pressure profile parameter and radius to maximum winds of hurricanes from flight-level pressure and H*Wind data. *J. Appl. Meteor. Climatol.*, **47**, 2497–2517, https://doi.org/10.1175/2008JAMC1837.1.

Weaver, R. J., and D. N. Slinn, 2010: Influence of bathymetric fluctuations on coastal storm surge. *Coastal Eng.*, **57**, 62–70, https://doi.org/10.1016/j.coastaleng.2009.09.012.

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

Willoughby, H. E., and M. E. Rahn, 2004: Parametric representation of the primary hurricane vortex. Part I: Observations and evaluation of the Holland (1980) model. *Mon. Wea. Rev.*, **132**, 3033–3048, https://doi.org/10.1175/MWR2831.1.

Xuan, J., R. Ding, and F. Zhou, 2021: Storm surge risk under various strengths and translation speeds of landfalling tropical cyclones. *Environ. Res. Lett.*, **16**, 124055, https://doi.org/10.1088/1748-9326/ac3b78.

Zachry, B. C., W. J. Booth, J. R. Rhome, and T. M. Sharon, 2015: A national view of storm surge risk and inundation. *Wea. Climate Soc.*, **7**, 109–117, https://doi.org/10.1175/WCAS-D-14-00049.1.

Zhang, K., C. Xiao, and J. Shen, 2008: Comparison of the CEST and SLOSH models for storm surge flooding. *J. Coastal Res.*, **2008**, 489–499, https://doi.org/10.2112/06-0709.1.

——, H. Liu, Y. Li, H. Xu, J. Shen, J. Rhome, and T. J. Smith III, 2012: The role of mangroves in attenuating storm surges. *Estuarine Coastal Shelf Sci.*, **102–103**, 11–23, https://doi.org/10.1016/j.ecss.2012.02.021.

Zhu, Y., Z. Toth, R. Wobus, M. Wei, and B. Cui, 2012: NCEP global ensemble implementation news. May 2006 upgrade of the GEFS and first implementation of NAEFS systems, accessed 8 September 2022, https://www.emc.ncep.noaa.gov/gmb/ens/ens_imp_news.html.