Stevens Flood Advisory System 2020 Ensemble Forecast Assessment: NY/NJ Harbor Area

Prepared by

Philip Orton, Ziyu Chen, Hoda El Safty, Mahmoud Ayyad,

Raju Datla, Jon Miller, Muhammad Hajj

Stevens Institute of Technology

November 2021

Executive Summary[1]

Coastal floods are among the world's most dangerous and damaging natural hazards, and forecasting is critical for evacuation and other emergency management decisions. Ensemble forecasts are increasingly being used to provide vital metrics such as the near-worst case flood area and depth (e.g., 95th percentile) in advance of a storm's arrival.

The Stevens Flood Advisory System (SFAS) is a coupled hydrologic-coastal forecast system, operational since 2007, and running in an ensemble mode since 2015. Forcing from a super-ensemble of US, European and Canadian atmospheric forecast ensembles (96 members total) is used to reflect the uncertainty in weather forecasts. Resulting data are available on a forecast website, sent to NOAA Weather Forecast Offices to help form their Total Water Level guidance, and used for launching street-scale simulations and forecasts for New York City metro area infrastructure. The modeling used for SFAS has been demonstrated with hindcast simulations to be capable of predicting time series or peak water levels based on atmospheric reanalysis data typically within 10-15% accuracy, including for Hurricanes Irene, Sandy and dozens of historical extreme events back to the 1700s. However, the ensemble SFAS forecasts have only seen a limited initial 4-month assessment for the winter 2015-2016 period.

The primary purposes of this 2020 forecast assessment are to quantitatively evaluate central forecasts and uncertainty ranges, identify any problem areas and chart a path toward future improvements, and create a baseline for which to compare future annual forecast assessments. We evaluate central forecasts and the 5^{th}-to-95^{th} percentile spread of coastal water level forecasts for the year 2020, which included Tropical Storm Isaias, the highest storm surge event since Hurricane Sandy. We quantify the central forecast accuracy using RMSE, and the spread using Coverage of Observation by forecast area of Uncertainty (COU), the latter being the percentage of the time that observed water level falls within the spread (a value of 90% being optimal). While this report is focused on only a set of five NY/NJ Harbor area stations that are representative of the harbor region's differing sub-embayments, future years' reports will expand out to regional stations.

The harbor wide RMSE on temporal maxima (peaks) for the highest five water level events of the year was 0.65 feet (13%) for lead times of 4 days, and was reduced down to 0.50 feet (10%) within one day of the peak. The RMSE for periods of time with >1.0 ft storm surge was 0.50 ft and for >1.5 ft surge was 0.65 ft. The uncertainty estimates (spread) of SFAS forecasts were very good, with harbor average COU of ~91% for cases with small surges, compared with an ideal value of 90%. Harbor-average COU averaged ~94% for the larger negative surges, and ~88% for the larger positive surges. RMSE was higher and COU lower for Isaias, and a separate assessment and peer-reviewed publication is being prepared that analyzes that storm's forecasts in greater detail.

[1] For correspondence or questions, email Philip Orton at porton@stevens.edu

1. Introduction

The Stevens Institute of Technology Flood Advisory System (SFAS) is the evolution of public coastal ocean forecasts from Stevens that have grown since the inception of forecasts for the New York Harbor Observing and Prediction System (NYHOPS) in 2006 (Bruno et al. 2006) and Storm Surge Warning System in 2010 (Georgas and Blumberg, 2010). It includes rainfall-driven hydrologic inputs, tides and storm surge. Ensemble total water level forecasts have been running since 2015 and are presently based upon 96 different meteorological forecasts, providing a central-estimate time series of water level with a 90% confidence interval (Georgas et al 2016; Jordi et al. 2019). Forecast graphics are posted with data access on the forecast webpage and interested users can sign up to be notified of impending flooding via email warnings.

As of late 2020, Stevens also provides probabilistic water level forecast data to the National Weather Service (NWS) Weather Forecast Offices at Upton and Mt. Holly. NWS has been using the Stevens Flood Advisory System as an important component of their storm forecast guidance development that serves the coastal New York, New Jersey and Connecticut region. They are now using the SFAS numeric forecast data in their Total Water Level forecast system to help inform their forecast guidance.

The modeling used for SFAS has been demonstrated with hindcast simulations to be capable of predicting time series or peak water levels based on atmospheric reanalysis data typically within 10-15% accuracy (RMS error), including for Hurricanes Irene (Orton et al. 2012), Sandy (Georgas et al. 2014; Orton et al. 2016; Jordi et al. 2019) and a suite of historical extreme events back to the 1700s (Orton et al. 2016). However, the ensemble SFAS forecasts have only seen a limited initial assessment for the winter 2015-2016 period (Georgas et al. 2016), and a broader assessment is warranted.

The purposes of this 2020 forecast assessment and report are to:

- Archive forecast data for future use
- Quantitatively evaluate central forecasts and uncertainty ranges
- Identify any problem areas and chart a path toward future improvements
- Create a baseline for which to compare future annual forecast assessments

To focus primarily on New York Metro area interests for this report, we assess a set of five NY/NJ Harbor area stations that are representative of the harbor region's differing sub-embayments. Future years' reports will expand out to regional stations. Below, in **Section 2** we review methods behind SFAS, in **Section 3** we recap the year's significant flood and surge events, in **Section 4** we report the results of our assessment, and in **Section 5** we provide some brief context, discussion and conclusions.

2. Methods

2.1. Stevens Flood Advisory System (SFAS)

The SFAS operational hydrologic-coastal ensemble prediction system forecasts water levels across the US Mid-Atlantic and Northeast (Georgas et al. 2016). The Stevens Estuarine and Coastal Ocean Model (sECOM) is used for hydrodynamic prediction, while the US Army Corps of Engineers (USACE) packages; Hydrologic Engineering Center's – Hydrologic Modeling System (HEC-HMS), and – River Analysis System (HEC-RAS) are used for hydrologic modeling of precipitation-runoff processes and for dendritic watershed and hydraulic calculations of water flow in channels, respectively. The system is forced by various data obtained from different resources such as: United States Geological Survey (USGS), National Oceanic and Atmospheric Administration (NOAA), including 15 NY/NJ Harbor-area water level stations installed and managed by Stevens Institute.

The sECOM model is a free-surface, hydrostatic, primitive equation model with terrain-following ("sigma") vertical coordinates with an orthogonal curvilinear Arakawa C-grid. A parallelized code version is used for rapid run times and efficient use of supercomputer resources (Jordi et al. 2017). A coupled rapid surface wind-wave model helps account for wave-current combined bed stress (Georgas, 2010) and explicit representation of the effects of wave steepness on wind stress (Taylor and Yelland, 2001; Orton et al. 2012).

For the central forecast region of New York Bight, the hydrodynamic modeling is performed using a nested application of sECOM (**Figure 1**). The New York Harbor Observing and Prediction System (NYHOPS) model domain encompasses continental shelf and estuary areas with 147×452 cells, a horizontal resolution from approximately 4.7 miles (7.5 km) at the open ocean boundary to less than 160 feet (50 m) in NY/NJ Harbor, and 10 vertical layers. The NYHOPS domain simulations are run in sECOM's three-dimensional mode and boundary conditions are applied at its offshore boundary (OBCs) and its interface with hydrologic models. The OBCs are a sum of (a) storm surge modeled on the Stevens Northwest Atlantic Prediction (SNAP) domain, (b) tides from the ADCIRC East Coast tide constituent database (Mukai et al. 2002), (c) a uniform ~0.5 ft (11 to 13 cm) cross-shore slope positive toward land (Georgas and Blumberg, 2010), and (d) a bias correction based on coastal tide gauge stations. Simulations span 108 h from present and are repeated every 6 hours with "00z", "06z", "12z" and "18z" (GMT times in hours) launch times.

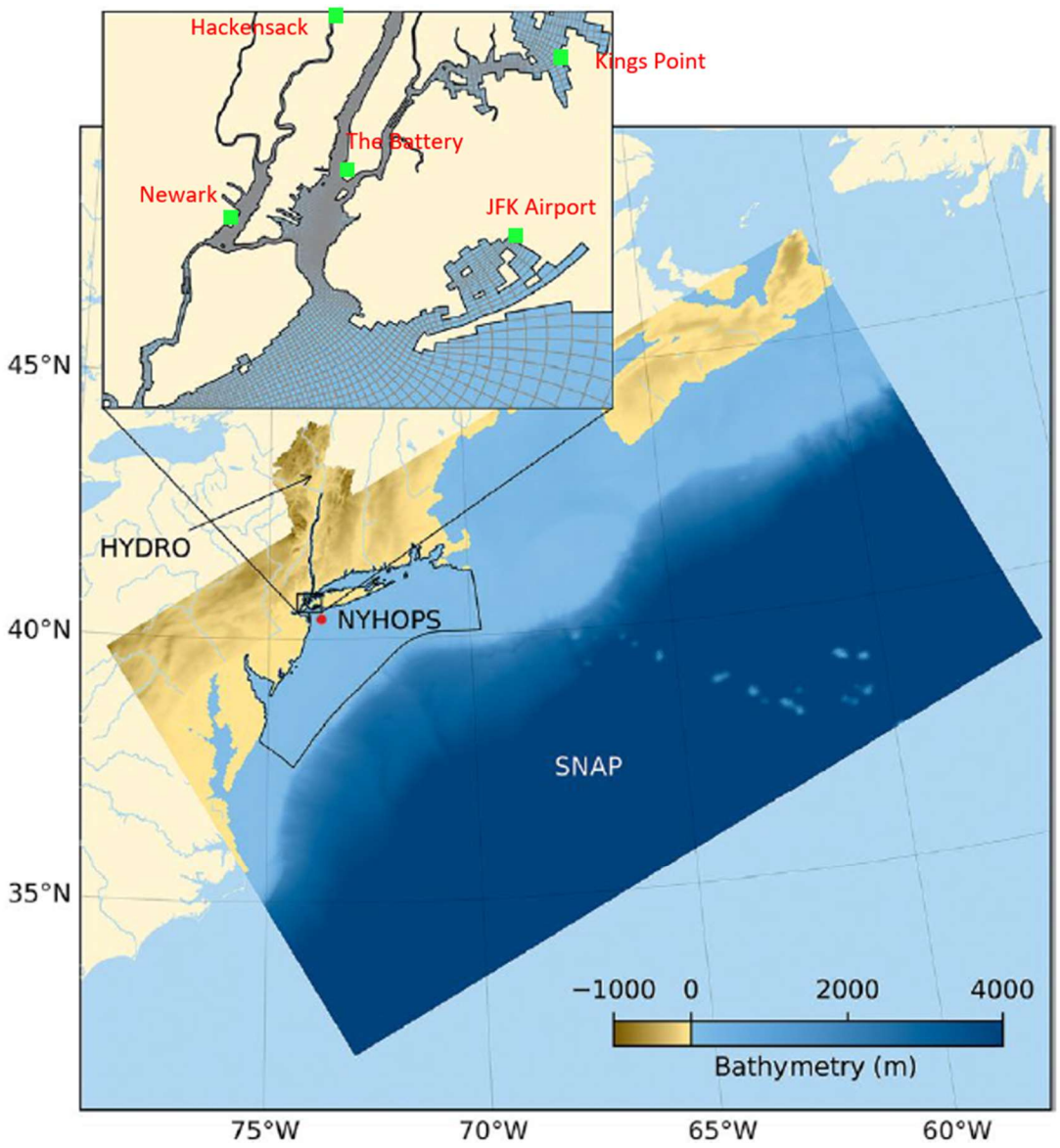

Figure 1: NYHOPS model domain linked to offshore SNAP model and inland Stevens HYDRO models. SFAS forecast locations assessed in this report include The Battery, Newark Bay ("Elizabeth Channel near Newark NJ" in SFAS), JFK Airport ("Bergen Basin at Jamaica Bay"), Kings Point and Hackensack River at Hackensack NJ.

SFAS uses an ensemble forecasting approach with sECOM and hydrologic simulations on each domain run with 96 different atmospheric forcing datasets. The ensemble meteorological forcing datasets are shown in **Table 1** and include the Global Forecast System and Global Ensemble Forecast System (GFS, GEFS; Han and Pan, 2011), North American Mesoscale forecast system (NAM; Skamarock et al. 2005), Canadian Meteorological Center (CMC) global ensemble prediction system (Charron et al. 2010), European Centre for Medium-Range Weather Forecasts (ECMWF; Buizza et al. 2007; ECMWF, 2018) ensemble and high-resolution member (ECMWF-HRES). Meteorological data are spatially and temporally interpolated using bicubic and cubic spline interpolation, respectively, to create hourly forcing fields on each domain's grid.

Two ensemble forecasts are produced within the SFAS system, the regional NYHOPS-E forecasts, which include tidal forcing in the hydrodynamic simulations, and NW Atlantic SNAP-Ex forecasts, for which the tides are added after the simulations are completed. Only NYHOPS-E forecasts are evaluated in this forecast assessment, given that SNAP-Ex forecasts are not available for harbor areas. Additional higher-resolution nested subgrid modeling for the 5th, 50th and 95th percentile scenarios is performed to map forecast flooding at critical infrastructure sites.

Table 1: Information on meteorological forcing datasets behind the ensemble

Product	**# members**	**original 2015 member code number**	**Spatial resolution (km)**	**Temporal resolution (h)**
GFS	1	23	25	3
GEFS	21	24-44	50	3
CMC (GEPS)	21	45-65	55	6
NAM	1	66	12	3
ECMWF-ENS	51	74-124	28	3
ECMWF-HRES	1	125	14	3

After every cycle of model simulations, the SFAS system splits the data into two periods; one hindcast day used for bias correction and weighting purposes and 4.5 forecasting days using the ensemble processing methods below. Hydrologic and hydrodynamic simulations, ensemble analyses and website graphics require 2 hours, and posting of timeseries water level forecasts occurs at about 02:00, 08:00, 14:00, and 20:00 hours GMT (03:00, 09:00, 15:00 and 21:00 EST) at http://www.

stevens.edu/SFAS. Results are then used for the PA subgrid simulations and posted to PAFAS about one hour later.

All model simulations and resulting SFAS and PAFAS forecasts are run at Stevens Institute in the Pharos Hyperscale Computing Facility, a 1,320-core Hewlett-Packard HPC built using HP Proliant servers, Mellanox FDR InfiniBand network and Seagate 2.2 PB Lustre-based storage facility, incorporating special-purpose modeling, database, and web presentation systems. The systems are housed in a custom data center designed to provide a high level of redundant power and cooling capacity to ensure uninterrupted operations.

2.1.2. Ensemble processing methods

SFAS uses a weighted average of the 96 members to find the central forecast (Georgas et al. 2016):

$$\eta_w = \sum_{j=1}^{m} w^{(j)} \eta^{(j)} \qquad (1)$$

Here, the superscript j denotes the ensemble number, m is the total number of ensembles, and w is the normalized weight which is defined by:

$$w^{(j)} = \frac{factor^{(j)}}{\sum_{j=1}^{m} factor^{(j)}} \qquad (2)$$

where $factor^{(j)}$ is the weight value:

$$factor^{(j)} = \frac{1}{\left(\left|\epsilon^{(j)}\right| + 0.05\right)\left(RMSE^{(j)} + 0.05\right)} \qquad (3)$$

Here, $\epsilon^{(j)}$ is the 24-hour hindcast mean bias and $RMSE^{(j)}$ the root mean squared error for member *j*.

$$RMSE^{(j)} = \sqrt{\frac{1}{N^{(j)}} \sum_{i=1}^{N^{(j)}} \left(\eta_{m\ i}^{(j)} - \eta_{oi}\right)^2} \qquad (4)$$

Thus, the normalized weights are estimated posterior model probabilities (Georgas et al. 2016).

The 5th and 95th percentiles are interpolated using the empirical distributions formed with all available ensemble members (equally weighted). The range of the ensemble represents atmospheric uncertainty across multiple ensembles but does not include any ocean modeling uncertainty. As a rough method for approximating the effect of additional ocean, wave and air-sea interaction uncertainties, the central forecast RMSE from the hindcast period is added to the 95th percentile and subtracted from the 5th percentile to broaden the spread.

2.2. Model performance quantification

In this report, we assess forecasts from SFAS NY/NJ Harbor-area stations for all available archived data in 2020 (all stormy periods and most of the year's non-storm periods). Five stations are representative of the major harbor embayments and tributaries -- The Battery (Manhattan), Newark Bay ("Elizabeth Channel near Newark NJ" in SFAS), JFK Airport ("Bergen Basin at Jamaica Bay"), Kings Point and Hackensack River at Hackensack NJ (**Figure 1**). We adopt the RMSE, and the Coverage of Observation by forecast area of Uncertainty (COU) as our performance metrics to measure the performance of the numerical model central forecast and uncertainty range with respect to the available observation,

RMSE is here computed in four ways: (a) As shown in Eq (4), for each 5.5-day simulation (hindcast and forecast period) of every ensemble member, (b) on the year's top-5 daily temporal maxima, (c) using all data within 6-hour bins of lead time for ensemble-based central forecasts, and (d) for ensemble-based central forecasts for each 6-hour forecast period separately (the "6-hour RMSE"), then presented as averages and 90% variability ranges. The latter computation used 50% overlapping bins, and for these the 6-hour average surge is also computed, to optimally capture cases with high positive or negative surges and avoid dividing one event into two bins. For example, Isaias' short intense storm surge that peaked at 1800h UTC would be divided into two bins with small average surges (one even including negative surge data points), if it were partitioned only into bins with the 00z, 06z, 12z, 18z borders.

`COU is computed as:

$$COU = \frac{n}{N} * 100\% \qquad (5)$$

Here, n is the number of cases where the observation falls within the forecast 90% confidence interval, and N is the total number of forecast time periods. Given that there are 4 forecasts per day, each with 108 hours, each with data every 10 minutes (6 data points), thus in a full year, N = 946000 (if there are no gaps in the observational data). Again, 6-hour, 50% overlapping bins were used for COU computations.

3. Recap of 2020 Storm Events and Forecast Reporting

Significant water level and storm surge events are summarized in **Table 2** and **Table 3** below. The year's highest water level occurred during Winter Storm Gail (**Table 2**). The year's peak 6-hour average for storm surge also occurred during Winter Storm Gail, at 2.46 ft, whereas Isaias took second at 2.20 ft (**Table 3**). The year's highest storm surge without temporal averaging at The Battery occurred during Isaias at 4.27 ft, which is also the largest surge since Hurricane Sandy. Fortunately, it occurred near low tide and total water level was not in the top-5 (**Table 2**).

Table 2: The top 5 peak water level events at The Battery station

Rank	Date	Peak water level (ft NAVD88)	6h-avg surge (ft)	Peak surge (ft)	Notes
1	12/17/2020	4.86 (minor[a])	0.85	3.28	winter storm Gail
2	4/10/2020	4.56 (minor[a])	0.46	1.19	king (perigean spring) tide
3	10/30/2020	4.56 (minor[a])	1.71	2.26	winter storm Zeta
4	5/9/2020	4.36	0.56	0.79	king (perigean spring) tide
5	12/25/2020	4.27	1.71	2.26	

a: The National Weather Service minor flood threshold (4.42 ft) was exceeded at The Battery, indicating minimal or no public property damage (e.g., shallow street flooding)

Table 3: The top 5 peak events for 6-hour average surge at The Battery station

Rank	Date	6h-avg surge (feet)	Peak surge (feet)	Peak water level (ft NAVD88)	Notes
1	12/17/2020	2.46	3.28	4.86 (minor)	winter storm Gail
2	8/4/2020 (Isaias)	2.20	4.27	3.84	tropical storm Isaias - peak storm surge since Sandy
3	4/4/2020	1.87	2.13	4.25	
4	10/30/2020	1.71	2.26	4.56 (minor)	winter storm Zeta
5	12/25/2020	1.71	2.26	4.27	

Three events caused an exceedance of the National Weather Service “minor flood” threshold of 4.42 ft NAVD88, and none the moderate flood threshold of 5.72 ft. The year’s top water levels were observed during either severe extratropical cyclones (e.g., winter storms Gail and Zeta) or king tides with small storm surges of only about one foot.

SFAS forecasts ran and the website reported the forecasts (as always) throughout the year, although there were no major flood events. An example of the online presentation of the SFAS forecasts is shown in **Figure 2**, demonstrating the forecast prior to passage of Isaias along with the perspective in hindcast. Observations (red dots) typically fall within the forecast 90% confidence interval (grey shading) but can deviate further from the central forecast (magenta line) in extreme cases. A separate peer-reviewed publication manuscript has been submitted with an assessment of flood forecasts for Isaias, to look more closely at these results (Ayyad et al. submitted).

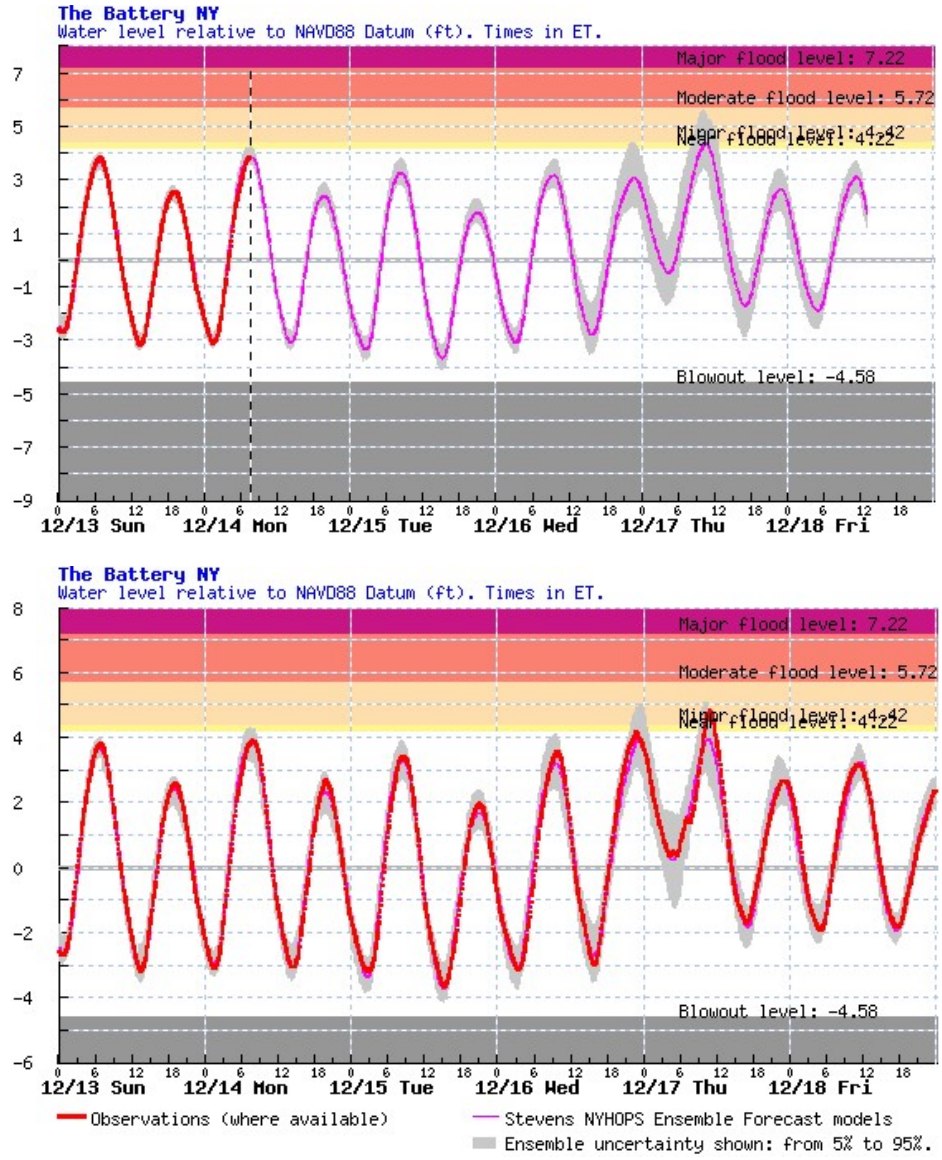

Figure 2: SFAS forecast time series of water level (top) before, and (bottom) after the passage of Winter Storm Gail, showing good accuracy several days in advance. Predicted peak water level was 4.30 ft NAVD88, and the observed water level came in at 4.86 ft (an -12% error). Times shown are Eastern Standard Time (ET).

SFAS forecasts ran and the website reported the forecasts (as always) throughout the year, although there were no major flood events. An example of the online presentation of the SFAS forecasts is shown in **Figure 2**, demonstrating the forecast prior to passage of Isaias along with the perspective in hindcast. Observations (red dots) typically fall within the forecast 90% confidence interval (grey shading) but can deviate further from the central forecast (magenta line) in extreme cases. A separate peer-reviewed publication manuscript has been submitted with an assessment of flood forecasts for Isaias, to look more closely at these results (Ayyad et al. submitted).

4. Results: Detailed Forecast Assessment

First, we review the forecast datasets being evaluated in this report, as well as any dropouts of ensemble members through the year. **Figure 3** depicts the time series RMSE at The Battery for all members and simulations across four forecast cycles a day. The Battery observational dataset is nearly complete (no gaps), so this station and figure is useful for observing cases where ensemble members failed, sometimes regularly, sometimes episodically.

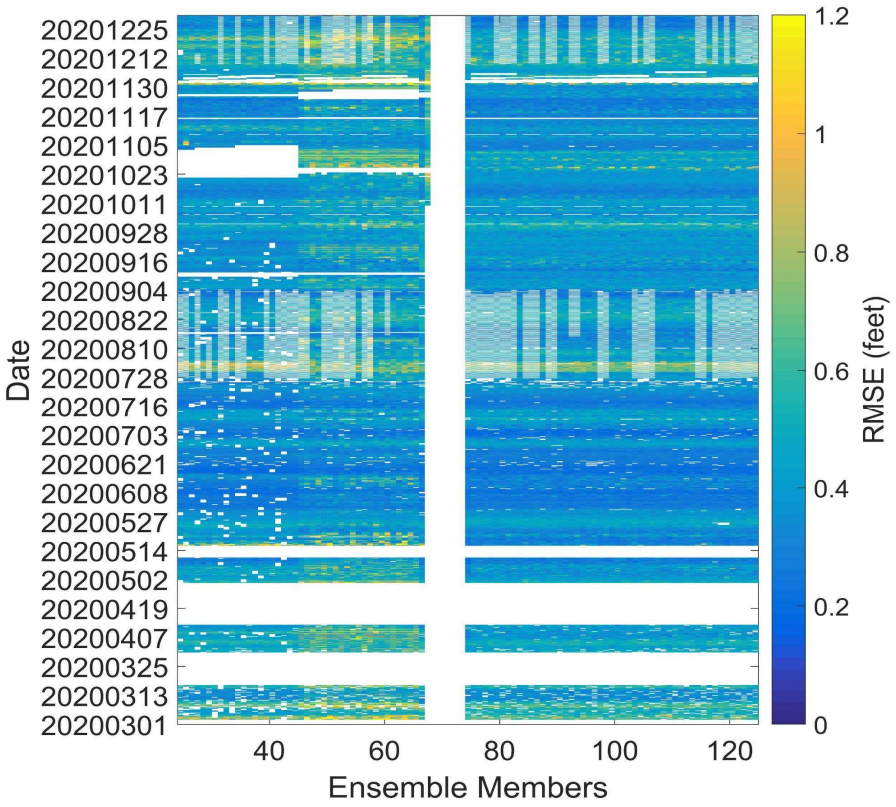

Figure 3: Assessment for individual ensemble members at The Battery station for 2020, after tide bias and mean bias corrections have been applied. The color shading is the RMSE of each member's forecast water level (relative to observed water level) across the simulation's full time range. The Y-axis is the forecast launch time, and the X-axis is the original 2015 code number for each of the NYHOPS Ensemble Members (**Table 1**).

The vertical column of white space denotes no simulations, as members 1-22 and 68-73 were discontinued due to poor accuracy or data unavailability. The white rows in early 2020 show periods with no storms where we did not set aside model time series data for all members, so we do not evaluate these periods in this report. From May 15 2020 onward, we have a system that backs up all data. Remaining white spaces or sporadic white flecks denote missing model data, which typically arose due to forcing data availability problems (e.g., with NOAA's NOMADS website) hardware failures (e.g. a computational node going down), or at times simple model or code crashes.

Typical RMSE for ensemble members is 0.2-0.4 ft, though the CMC members (#45-65) often have noticeably higher values in stormy periods (e.g., December 2020) than others. Some periods have higher RMSE values across all members. For example, in early August 2021, when tropical storm Isaias passed nearby and caused a rapid large storm surge, RMSE values were approximately 0.8-1.0 ft.

Overall, the figure demonstrates that there were at least 50 ensemble members through nearly the entire year, more than enough to form a probability distribution and estimate the 5th, 50th and 95th percentiles. Two exceptions with well below 50 members, visible as nearly complete horizontal bands of white in **Figure 3**, represent 6 cases out of the over 1100 forecast periods in 2020 where data were being backed up. One case was November 17, 06:00 and 12:00 UTC where we see no members (unknown reasons, and possibly just a missing backup). The other case began on December 2 18:00 which had 28 successful members, then December 3 00:00 UTC all simulations failed, and 06:00 UTC and 12:00 had only 3 successful members, before the system was brought back to full operation.

Focusing on the central forecast accuracy during extreme events, **Figure 4** shows the RMSE computed on peaks, across all five harbor stations. The harbor wide RMSE on temporal maxima (peaks) for the highest five water level events of the year was 0.65 feet (13% when errors are normalized) for lead times of 4 days, and was reduced down to 0.50 feet (10%) within one day of the peak.

Evaluating the forecasts at all times through the year, not just peaks during the highest water level events, the year-round forecast accuracy versus lead time is shown in Figure 5, with RMSE on the left and RMSE normalized by the mean higher-high water datum (MHHW) on the right. All cases in Figure 5 show improving accuracy as the date of the forecast water level gets closer (as lead time shrinks). Best performance, based on ordering of RMSE results, is Battery, Newark Bay, Kings Point, JFK Airport, then Hackensack River. However, for normalized RMSE the ordering is switched to Kings Point, Battery, Newark Bay, Hackensack, JFK Airport. Values of RMSE were from 0.25 to 0.55 ft across all stations and lead times, which equaled 4-9% when normalized by tide. RMSEs were reduced by about 20-35% from lead times of 105 hours down to 3 hours (e.g., from 0.35 to 0.24 ft at Battery, a 31% reduction).

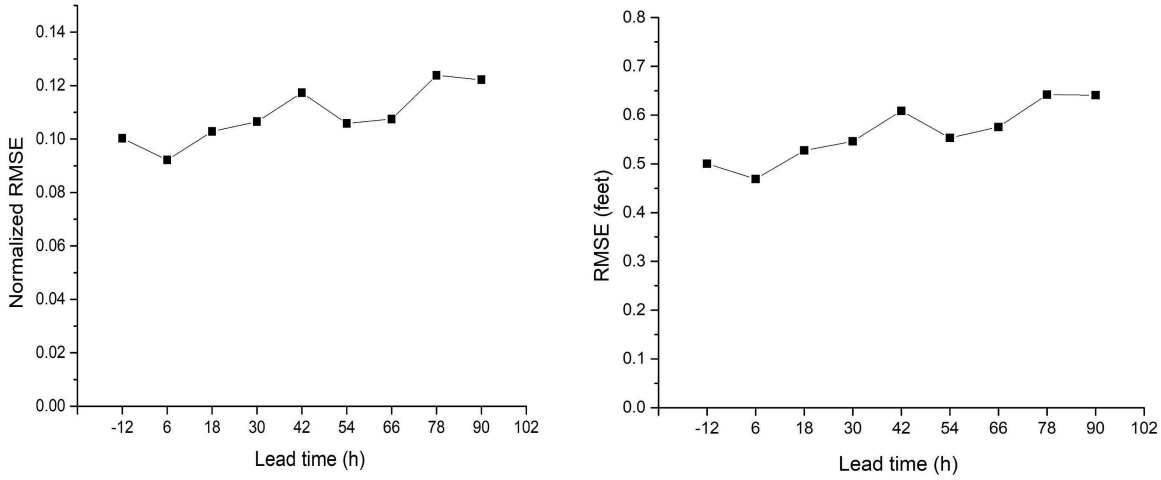

Figure 4: Assessment of accuracy of central forecasts for temporal maxima (peaks) for the top-5 events of each harbor station, (left) RMSE, (right) normalized RMSE where the error is normalized by the peak water level.

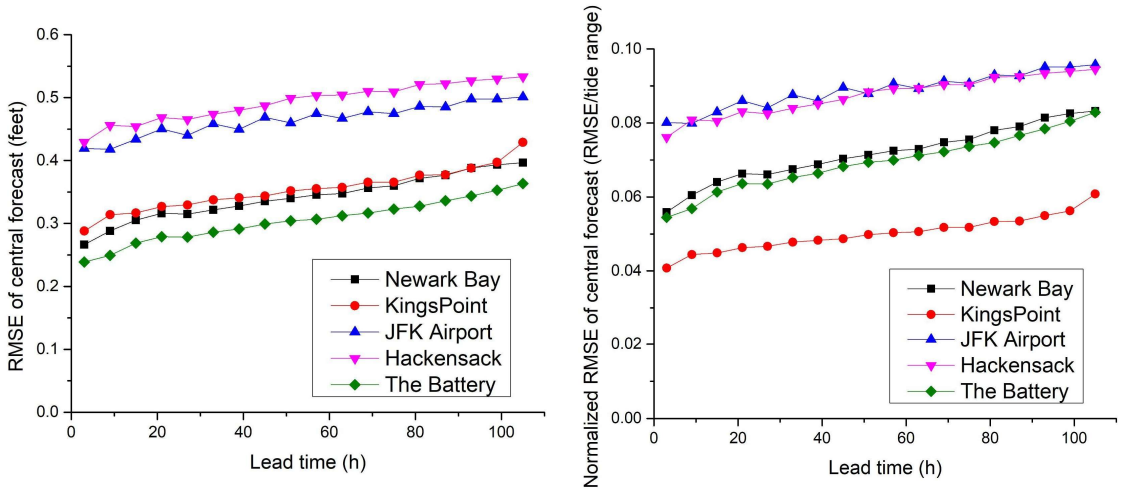

Figure 5: The SFAS weighted mean ensemble total water level forecasts against observations across the five representative harbor stations. (left) RMSE of the weighted mean against observed water level across the forecast time in 2020. (right) same, but normalized by each station's MHHW datum. The X-axis is the forecast lead hours.

The 6-hour RMSE mean and its 90% variability range versus lead time at each station is shown in **Figure 6**. While computed slightly differently (see detailed methods in **Section 2.3**), the mean of 6-hour RMSE shown here are very similar to the RMSE values shown in **Figure 5**. However, the addition of the variability range in this figure illustrates that RMSE for any 6-hour period can vary significantly from the mean.

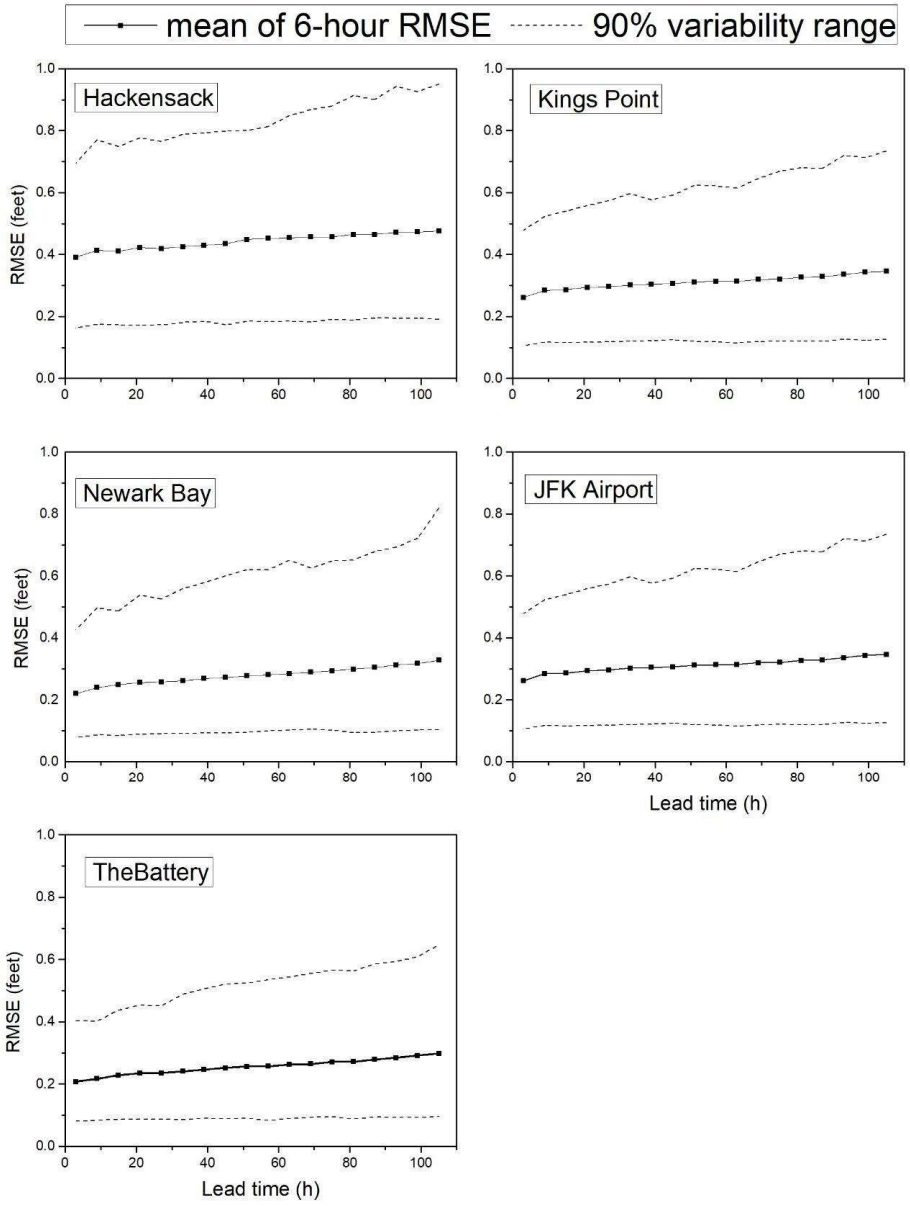

Figure 6: The weighted mean against observations across various forecast lead time range and its variability range for multi-stations comparison operated by the NYHOPS. The dot solid lines are the mean of the RMSE of the weighted average against observed water level for various forecast lead hour time ranges (e.g., 0-6 hours) for 2020. The dashed lines are their corresponding 90% variability range. The X-axis is the forecast lead hours.

The accuracy of central forecasts is particularly important during storm events. It is assessed in **Figure 7** where the mean 6-hour RMSE and its variability with storm surge is plotted for each station. Typically, RMSEs grow with surge amplitude (positive or negative). At Battery, for example, the RMSE for cases with >1.0 ft surge is 0.47 ft (90% variability from 0.13-1.15 ft) whereas for cases with >1.5 ft surge it is 0.62 (0.17-1.71 ft). Typical total water levels are 4-5 ft at Battery in these cases (**Table 3**), and therefore the relative error averages 12-15% during the year's worst storm surge events.

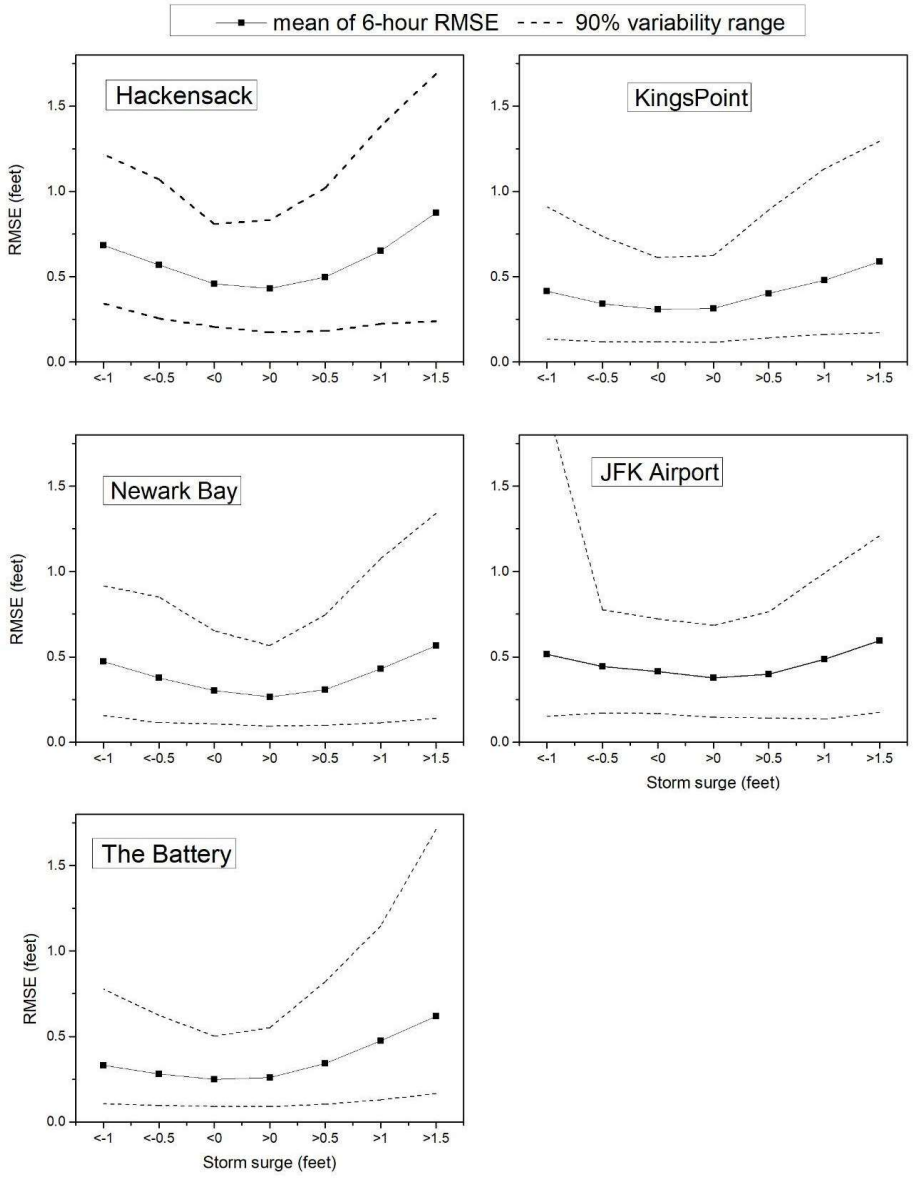

Figure 7: The RMSE of the weighted mean for various surge thresholds for the four harbor stations. RMSE and mean surge are computed in 6-hour periods, the black squares are the mean of RMSE results across all lead times and the dashed lines are the range within which 90 percent of the 6-hour RMSE values fall.

Overall, the year's average RMSE across the five stations for >1.0 ft surge was 0.50 ft and for >1.5 ft surge was 0.65 ft. JFK Airport has very little growth of RMSE with surge amplitude, suggesting that surges are fairly well-predicted, and tides are the source of error at that station. RMSE for >1 ft surge was 0.49 ft (0.14-0.99) and for >1.5 ft surge it was 0.59 ft (0.18-1.21).

Given the higher RMSE for all ensemble members during Tropical Storm Isaias seen in **Figure 3**, we also computed the RMSE versus storm surge without the data from Isaias for The Battery station (**Figure 8**). The results show clearly how Isaias worsened the RMSE results for the cases of high storm surge. With Isaias omitted, the RMSE values for cases of surge >1.5 ft are reduced to 0.53 (0.16-1.17 ft). For the Isaias forecasts, the larger errors resulted primarily from wind forecast errors and coarse wind forcing resolution (Ayyad et al. submitted).

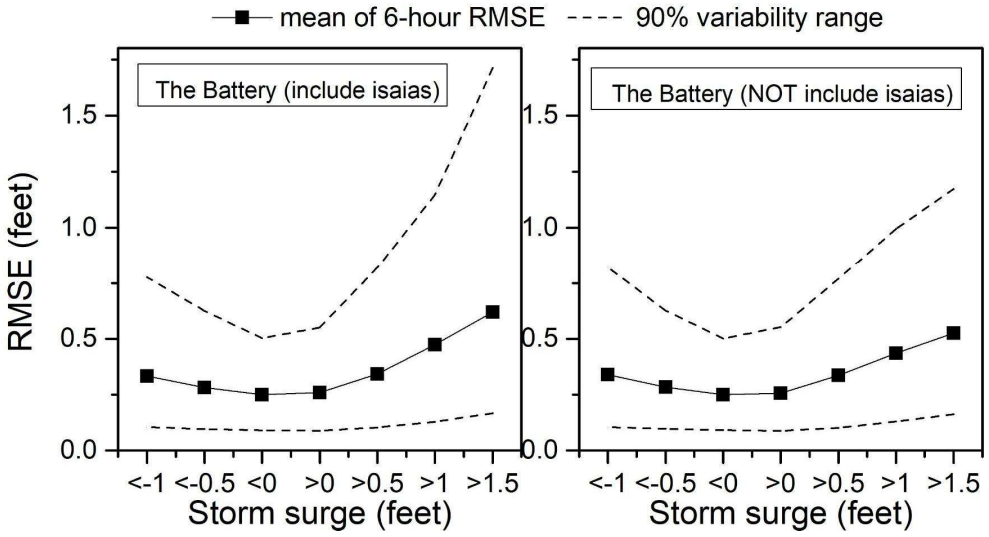

Figure 8: Same as Figure 6, but with the right panel omitting data from Isaias.

Evaluating the uncertainty estimates (spread) of SFAS forecasts (**Figure 9**), there is a general trend between surge and COU, representing a tendency toward slight overestimation of uncertainty for negative surges and (at times) slight underestimation for positive surges. Cases with no surge have approximately 90% COU (88-92%), meaning that the spread of the 90% confidence interval on forecast water levels is a very good approximation of uncertainty. Cases with a negative surge have a COU from 92-95%, indicating that the uncertainty is slightly overestimated. Cases with a positive surge have a COU from 82-95%, indicating that for some stations the spread can be slightly under- or over-estimated, but overall, it is accurate. For Isaias, the COU was even lower, at about 80% (Ayyad et al. submitted), furthering the trend shown here to a case of a fast storm with the year's highest peak instantaneous surge value.

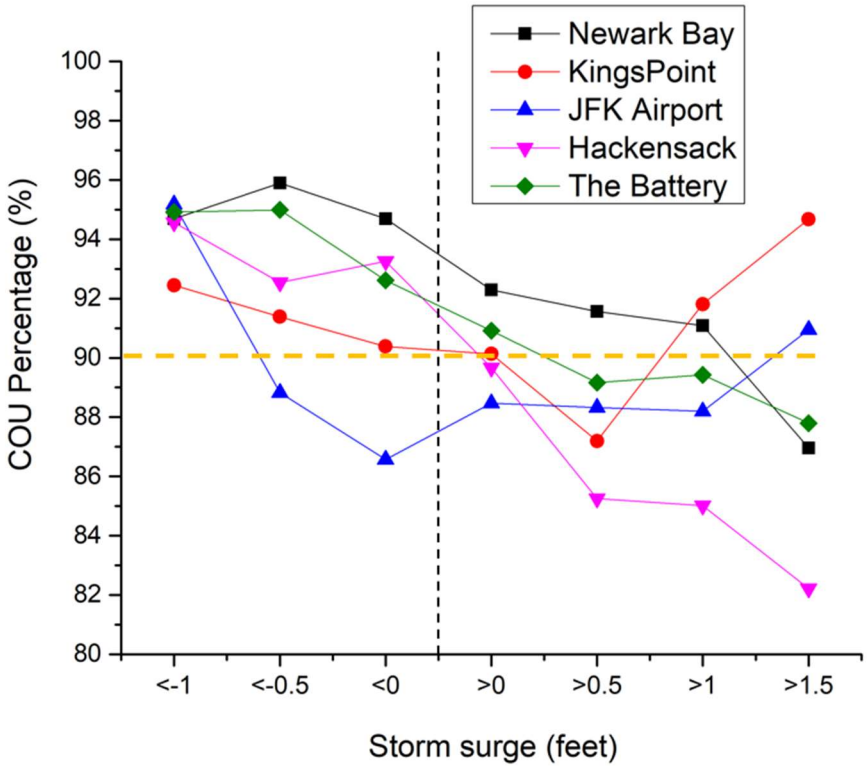

Figure 9: The Coverage of Observation by forecast area of uncertainty with surge threshold for the five stations. The orange dashed line is the 90% threshold, which is the ideal value of COU because it is the uncertainty range provided with the forecasts.

The impact of weighting on central forecast accuracy is assessed in **Figure 10**, showing very little impact in most cases and even some cases where accuracy was worsened. The mean for the "weighted mean method" is 0.286 ft and the mean for the "total mean method" (equal weighting) is 0.288 ft. Performance during storm surge events (positive or negative) also isn't significantly different. Therefore, when looking at the annual average, a weighted mean approach offers a negligible improvement over simple total mean.

5. Discussion and Conclusions

Forecast accuracy during the year's top five high water events was excellent, with RMSE of 0.65 ft for four-day lead times and 0.50 ft for those below one day (10-13%). Performance across the full year's datasets was also excellent for a coastal total water level forecast system, with an average RMSE across five harbor stations of ~0.45 ft (8%) for four-day lead times and ~0.35 ft (6%) for those below one day. A major system improvement that occurred in 2017 is that forecasts have since extended out 105 hours, compared to 81 hours in 2015-6.

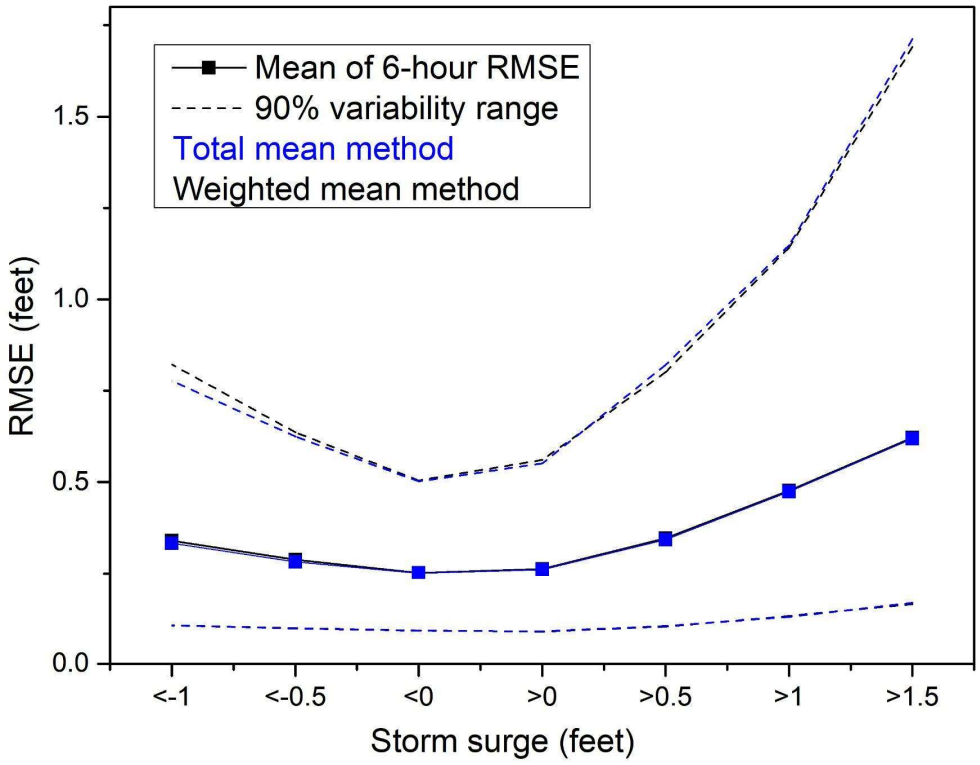

Figure 10: The comparison of weighted ensemble mean and total mean at the Battery station for 2020, across a range of storm surge bins.

During storm surge events, the forecast accuracy results for 2020 were generally similar to those reported for winter 2015-2016, except this computation includes forecasts over the 1 day longer lead time. Across several harbor stations, SFAS then averaged a 0.6 ft RMSE for >1.33 ft surge (Georgas et al. 2016). For the five stations quantified in 2020, the average RMSE across the five stations in **Figure 7** for >1.0 ft surge was 0.50 ft and for >1.5 ft surge was 0.65 ft. Interpolating these results to match the 2016 result (>1.33 ft) gives a RMSE value of 0.60 ft, the same result. When comparing on an apples-to-apples basis (looking at harbor stations for the same range of lead times and winter months only), system accuracy in storm surge events during 2020 is slightly improved relative to 2015-2016. For cases with storm surge greater than 1.33 ft, the RMSE was 0.44 ft, whereas it was 0.50 ft for the same in 2015-6 (Georgas et al. 2016). The improved accuracy is likely due to improvements in the system and its forcing data since 2016. These include improved CMC modeling and product resolution (improved from ~110 km to ~55 km resolution) and improved ECMWF-ENS and ECMWF-HRES modeling (but not product resolution - that is fixed with the price we pay for the products).

The uncertainty estimates (spread) of SFAS forecasts were very good, with harbor-average COU of ~91% for cases with small surges, compared with an ideal value

of 90%. Harbor-average COU averaged ~94% for the larger negative surges, and ~88% for the larger positive surges. The range of COU including all five stations separately is from 83 to 96%, still close to 90% and indicating reasonable estimates of uncertainty. A similar NOAA ensemble water level forecast system that runs only with GEFS and CMC ensembles showed COU values for 2018-2019 storms from 30-60%, compared with an ideal value of 80% for that system (which provides an 80% uncertainty range), revealing that uncertainty estimates with that system were always too small in storm events (Liu and Taylor, 2020). Experimental runs (not operational) with the addition of ECMWF-ENS in the ensemble resulted in little or no improvement, on average (Liu and Taylor, 2020).

5.1. Potential system improvements

Two concerns with SFAS were the lowered accuracy for central forecasts for tropical cyclone Isaias and the tendency toward underestimation of uncertainty for growing positive surges, and these two issues may be related to the same problem. A reduction in accuracy is expected for some tropical cyclones due to "resolution bias" if storms are small and/or fast so that the meteorological forcing (mostly 25-50 km and 3-hour resolution) doesn't resolve wind speed maxima. If the forecast water levels are consistently biased low, then both the central forecast can be biased low and the COU can be lowered further below the optimal value (because the observation more often falls outside the spread). Evidence of this was found in the deeper assessment of the Isaias forecasts, where resolution bias was a likely a factor in underestimation of the event's peak positive surge, negative surge and peak water level (Ayyad et al. submitted).

Further improvements in forecast accuracy and spread could be made by improving the system's use of some of the ensemble meteorological products. The CMC members clearly have less accuracy (**Figure 3**) and utilizing a higher temporal resolution of 3 hours could help reduce the RMSE. New forecast products are available with GEFS and GFS based on a more accurate FV3 model, and the system would likely be improved by switching to these products. However, limits on NOMADS data volumes make it a challenge to obtain and use these new products which are larger, higher-resolution datasets.

Another approach for improving the ensemble-based uncertainty computation would be to improve the estimated effect of ocean and air-sea interaction error. Presently the ocean uncertainty is represented only by widening the spread by hindcast RMSE times two (**Section 2.1.2**). A simple change could be made here to set it to be a standard function of wind or surge, ramping up for higher values of either. This could better represent the high uncertainties in waves and air-sea interaction (e.g., wind drag coefficient; Orton et al. 2012) and other aspects of oceanic modeling for extreme events.

A possible approach for reducing resolution bias would be to use atmospheric model resolution-dependent member weighting (see **Table 1** for resolutions). The

present weighting approach based on hindcast period results was shown in **Figure 10** to have little effect, which is unsurprising given that the prior day's conditions are often very different from the present and future days for which the forecast is used. A resolution-dependent weighting could be developed that unweights members that have too low a resolution to capture wind maxima. For such an approach, the system could be set up to automatically estimate the *required* wind resolution using the highest resolution member and analyzing wind velocity gradients in NY Bight.

5.2. Future work toward peer-reviewed publication of this report

As noted previously, this report is a first baseline effort for annual forecast assessments. More detailed analyses and comparisons will be performed using multiple years of data, and a manuscript will be prepared for submission to a peer-reviewed journal. Already, a paper is submitted on our detailed assessment of the Isaias forecasts (Ayyad et al. submitted). Some future planned analyses include:

- Quantitatively compare Stevens and NOAA 2020 ensemble forecasts (P-ETSS, P-SURGE), and seek any other possible forecasts for comparison to ours (e.g., CERA, UK - the system of Flowerdew and others)
- Compare forecast accuracy and spread using different subensembles (including a case using all members except ECMWF-ENS and ECMWF-HRES)
- Assess whether member dropouts shown as white flecks in **Figure 3** lead to forecast degradation. We can use the present datasets to evaluate this question.

With the advancements in backup accessibility and coding of identical processing in Matlab created herein, annual assessments for every future forecast year can be completed more rapidly.

For correspondence or questions, email Philip Orton at porton@stevens.edu

References

Ayyad, M., P. M. Orton, H. E. Safty, and M. R. Hajj (submitted), Assessment of a Super-Ensemble Forecast for Storm Tide and Resurgence from Tropical Cyclone Isaias, submitted to *Weather and Forecasting*.

Bruno, M.S., Blumberg, A.F. and Herrington, T.O., 2006. The urban ocean observatory-coastal ocean observations and forecasting in the New York Bight. In: Proceedings- Institute of Marine Engineering Science and Technology Part C, *Journal of Marine Science and Environment*, 4:31.

Georgas, N., Orton, P., Blumberg, A., Cohen, L., Zarrilli, D. and Yin, L., 2014. The impact of tidal phase on Hurricane Sandy's flooding around New York City and Long Island Sound. *Journal of Extreme Events*, *1*(01), p.1450006.

Georgas, N. and Blumberg, A.F., 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). In *Estuarine and Coastal Modeling (2009)* (pp. 660-685).

Georgas, N., Blumberg, A., Herrington, T., Wakeman, T., Saleh, F., Runnels, D., Jordi, A., Ying, K., Yin, L., Ramaswamy, V. and Yakubovskiy, A., 2016. The Stevens flood advisory system: Operational H3E flood forecasts for the greater New York/New Jersey metropolitan region. *Flood Risk Management and Response*, p.194.

Han, J. and Pan, H.L., 2011. Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Weather and Forecasting*, *26*(4), pp.520-533.

Jordi, A., Georgas, N., Blumberg, A., Yin, L., Chen, Z., Wang, Y., Schulte, J., Ramaswamy, V., Runnels, D. and Saleh, F., 2019. A next-generation coastal ocean operational system: Probabilistic flood forecasting at street scale. *Bulletin of the American Meteorological Society*, *100*(1), pp.41-54.

Liu, H., Taylor, A. and Kang, K., 2019, January. 3.8 LATEST DEVELOPMENT IN THE NWS'EXTRA-TROPICAL STORM SURGE MODEL, AND PROBABILISTIC EXTRA-TROPICAL STORM SURGE MODEL. In *99th American Meteorological Society Annual Meeting*.

Mukai, Ann Y., Joannas J. Westerink, Rick A. Luettich Jr, and David Mark. 2002. *Eastcoast 2001, a tidal constituent database for western North Atlantic, Gulf of Mexico, and Caribbean Sea*. DTIC Document.

Orton, P., Georgas, N., Blumberg, A. and Pullen, J., 2012. Detailed modeling of recent severe storm tides in estuaries of the New York City region. *Journal of Geophysical Research: Oceans*, *117*(C9).

Orton, P.M., Hall, T.M., Talke, S.A., Blumberg, A.F., Georgas, N. and Vinogradov, S., 2016. A validated tropical-extratropical flood hazard assessment for New York Harbor. *Journal of Geophysical Research: Oceans*, *121*(12), pp.8904-8929.