Improving the Management of Maritime Traffic in Southeast US Waters Using Machine Learning

Technical Report

Steven D. Meyers (smeyers@usf.edu)
Mark E. Luther (mluther@usf.edu)

Center for Maritime and Port Studies
College of Marine Science
University of South Florida

Highlights

- A machine learning algorithm was developed to predict high cross-currents above a threshold speed near Port Miami up to 48 hrs in advance as represented by HYCOM output variables
- Model accuracy was quantified by examining True and False Positive Rates, and True and False Negative Rates, and by constructing Receiver Operating Curves
- Initial predictor variables were downstream Gulf Stream frontal positions
- The highest cross-currents examined were 2 sd above the mean, and were predicted with ~90% TPR, and ~66% TNR
- Accuracy declined for lower current thresholds
- The addition of surface winds as a predictor increased the TPR and TNR for lower current thresholds, and decreased their False Positive and False Negative rates in all cases
- An initial evaluation of observational front data for use in an operational tool was performed

1. Background

The objective of this project is to create a cross-current prediction tool for the Port Miami shipping channel using machine learning methods. This report details efforts to develop such a tool based on output from the Hybrid Coordinate Ocean Model (HYCOM) system and observational data.

The data examined to date are: i) Acoustic Doppler Current Profiler (ADCP) current data from the vicinity of Port Miami. ii) the frontal position analysis provided by Fleet Numerical Meteorology and Oceanography Center (FNMOC) that includes the front positions throughout most of the North Atlantic and Gulf of Mexico.

1. The HYCOM

HYCOM is a global ocean circulation model (Bleck and Benjamin, 1993; Bleck and Boudra, 1981). There are two different versions of HYCOM, one run by NOAA, and another run by the US Navy. The horizontal Resolution is 1/12º, and the vertical structure is represented by 40 levels (interpolated) [0 2 4 6 8 10 12 15 20 25 30 35 40 45 50 60 70 80 90 100 125 150 200 250 300 350 400 500 600 700 800 900 1000 1250 1500 2000 2500 3000 4000 5000] meters. Output variables include water level, and temperature, salinity, and currents through the water column.

Output fields of surface elevation and surface currents from HYCOM was obtained from https://www.hycom.org/ at $\Delta t = 3$ hr time steps over the eastern Gulf of Mexico, the Florida Straits, and the eastern Florida coast (Figure 1) from January 1, 2018 through February 19, 2020. The model output variables had been interpolated to a fixed horizontal grid indexed (j, i). The grid point closest to the mouth of the Port Miami Channel was identified as (j_0, i_0). Model land boundaries used no-slip conditions and open boundaries use a sponge layer for energy dissipation. Additional details are provided by Bleck et al. (2001). The currents at (j_0, i_0) were significantly lower than those at (j_0, i_0+1) due to the former being strongly influenced by the no-slip boundary at (j_0, i_0-1). Values of the surface merdional current $v(j_0, i_0)$ were < 0 (southward) about 5% of the time. This was likely due to recirculation eddies, but these were not the focus of this study and not examined further.

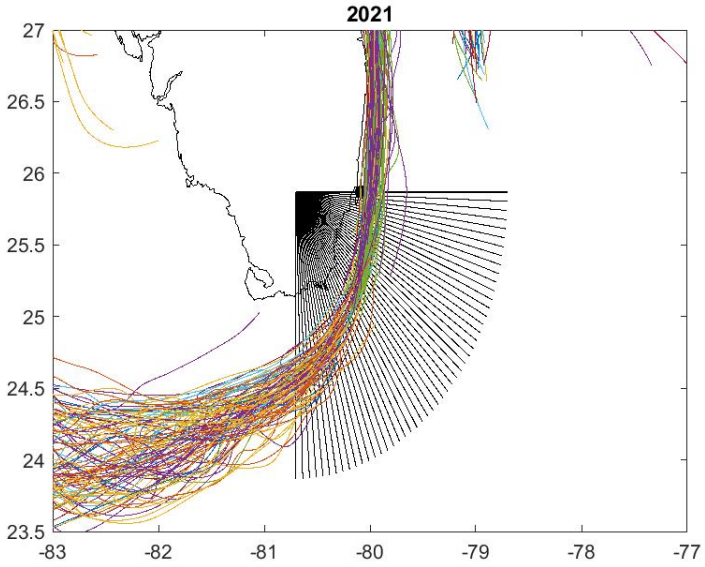

Figure 1. The polar projection used to define frontal position. Lines are the FNMOC frontal analysis for the first half of 2021. Colors are different dates.

The latitude (θ) and longitude (ϕ) of the Gulf Stream front positionv(θ, ϕ) at time step n, was determined using a polar projection along $k = 1, \ldots, N$ lines that captured the turning of the Florida Current (Fig. 1). The $k = N = 30$ line was zonal along the latitude of the Port Miami ADCP. The model current speed was interpolated to 100 points along each radial line using bicubic splines. The location of the maximum current speed along each radial line was determined for each model time-step. Differences between the position at time step n and the time-averaged position $\left\|(\theta_{kn}, \phi_{kn}) - (\bar{\theta}_k, \bar{\phi}_k)\right\| = \Delta\phi_{kn}$ were used to find i) the relation between $v(j_{Nn}, i_{Nn})$ and the front position (θ_{Nn}, ϕ_{Nn}), and ii) the time-lagged correlation of the front near the port $\Delta\phi_{Nn}$ and lagged frontal positions $\Delta\phi_{k(n-m)}$.

The wind fields used to drive the HYCOM surface stress vectors were from the Climate Forecast System (CFS) and transformed to the model grid. These gridded files were obtained from http://tds.hycom.org/thredds/catalog/datasets/force/ncep_cfsv2/netcdf/catalog.html over the same time period as the model elevation and current output. The zonal and meridional wind components (w_1, w_2) where each tested as a predictor variable. No benefit was found to including w_1 in the predictor set, so it will not be examined further in this report. The subsequent use of "wind" therefore only refers to the w_2 component.

2. Satellite Frontal Analysis

Frontal analyses provided by FNMOC were based on sea surface temperature (SST) satellite images composited over the course of a day or longer (Figure 2). Compositing was needed because cloud cover often obscures large portions of any individual image. Isotherms are identified and used to define the frontal position. The composited data are released every 2-4 days from 2015 through mid-2021. Each frontal segment is identified with the corresponding date of its analysis. The composite nature of these images removes any time association from the frontal position. Only the date is available, making these unsuitable for short-term (12-48 hr) predictions. Future work should investigate the use of the individual SST images, which have a date and time signature, as a predictor variable.

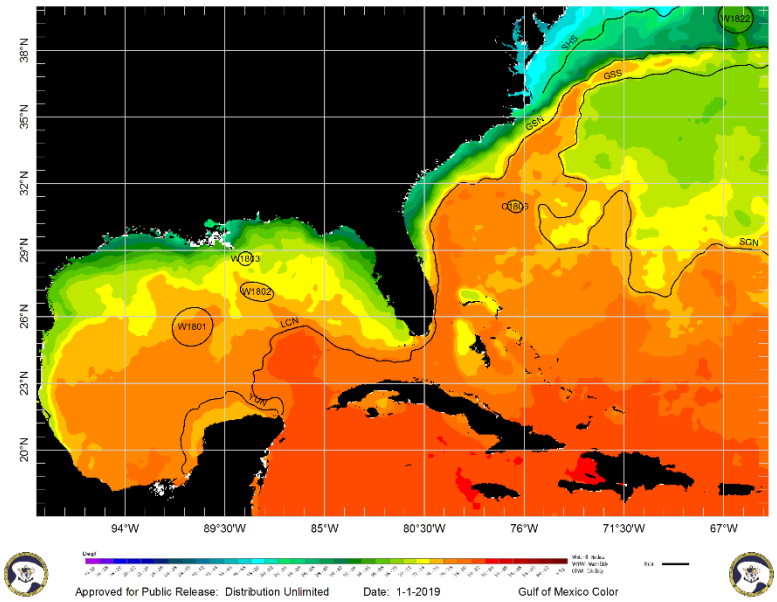

Figure 2. FNMOC front analysis for January 1, 2019. Colors are surface temperature from 26°-92°F.

3. Methods

3.1. HYCOM Front Position

Before spatial interpolation, the front was found about 8% of the time at (j_0, i_0+2) and was most common at i_0+3 and i_0+4, the last two being on either side of the mean front position for j_0 (Figure 3). Together, the latter represented about 76% of all front positions. During most of the time period examined, the front at j_0 was usually ±2 grid cells from its mean longitudinal position. The mean current speed and the maximum speed at (j_0, i_0) increased as the distance from the port $\Delta\phi_{Nn}$ decreased, with the exception when the front was at (j_0, i_0+1). However, this latter case only occurred in ~0.03% of the data, so this condition was not well-represented in the analysis.

The front position $\Delta\phi_{Nn}$ was correlated with its position further to the south determined by the polar projection,

$$\rho_{kN} = \frac{1}{2M\sigma_N\sigma_k}\sum_{m=-M}^{M}\left(\Delta\phi_{Nn}\Delta\phi_{k(n+m)}\right) \quad (1)$$

where σ represents the standard deviations of the frontal position for each time series, $k \leq N$, and M was chosen to encompass the expected time delay of maximum correlation based on the speed of meander propagation. A similar analysis was done for the correlation between current speed near the port and the frontal positions. The lagged correlation of meander propagation (1) was fairly strong, as is readily apparent in Figure 4. For a delay of ~12 hr, ρ peaked at 0.72 for k=25, and ρ peaked at 0.65 for k=20 for a delay of 24 hr. The effect of propagation was less clear in the correlation between $\Delta\phi$ and the current speed at (j_0, i_0) (Figure 4). The maximum $|\rho|$ was <0.3 at the locations of peak correlation for the meanders.

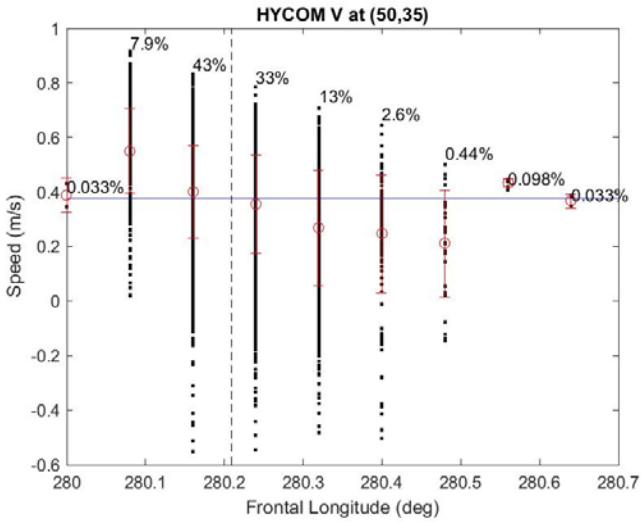

Figure 3. HYCOM meridional speed at (i_0, j_0) as a function of frontal position along j_0 (black). The mean (circle), and ± standard deviation (error bars). Long-term mean speed (blue) and long-term mean longitude of front (dashed). The percentage of time the front was identified at each location is indicated.

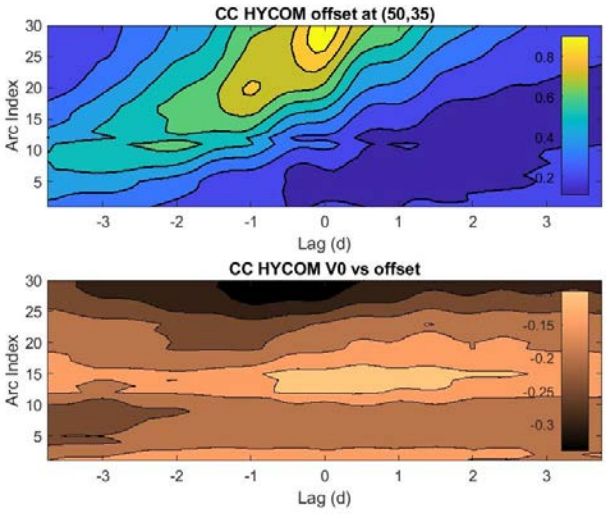

Figure 4. Lag correlation of $\Delta\phi_{k(n+m)}$ with (upper) $\Delta\phi_{Nn}$, and (lower) $v(j_0, i_0)$.

These initial results indicate: 1) the propagation of Gulf Stream meanders impact the frontal position at the port channel. 2) The frontal position impacts, but is not highly-correlated with, the channel cross-currents. 3) The impacts only occur for large negative frontal positions, indicating some threshold-type behavior. Machine learning algorithms are often used where conventional correlative techniques do not apply, but require the estimation of a boundary or threshold.

3.2 Wind

The wind at the time of the predicted current (time step n) and the same (j_0, i_0) grid point was also used as a predictor variable. Reasonably accurate forecast winds are routinely available up to 48 hours in advance over the United States and its nearshore waters. The winds used in this study represent such forecasts, though the values used were based on real hindcast and possibly had more accuracy than actual forecast wind vectors. This is an issue that would need to be addressed in future studies.

3.3. Logistic Regression

The findings in 3.1 might suggest useful predictions of currents $v(j_0, i_0)$ cannot be readily made solely using correlations with frontal positions to the south. However, extreme values of $v(j_0, i_0)$ were most commonly found when the frontal position was at (j_0, i_0+2), so some relation between currents and meander position existed, indicating a potential predictability that could be useful for forecasting high currents. Since such a forecast scheme could not be based on correlation, a categorical approach was explored. Machine learning algorithms have a high level of skill in such problems, so logistic regression was developed as a proto-type prediction tool.

Logistic regression (LR) is widely used to represent a dichotomous (2-valued) variable (y) that has a single transition between one value and the other (generally 0 and 1), dependent upon predictors $\boldsymbol{X}$ (Hilbe, 2016; Hosmer Jr et al., 2013). Here LR was used to predict whether or not $v(j_0, i_0)$ was ≥ a threshold value v^*. LR models yield the odds ratio of probability $0 \leq \pi \leq 1$ for y=1 as

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{i=1}^{N_v} \beta_i X_i = \boldsymbol{\beta} \cdot \boldsymbol{X} \tag{2}$$

where X is a set of N_v independent variables (alternatively called covariates or predictors), $\boldsymbol{\beta}$ is a vector of coefficients in this case determined by maximum likelihood. Inverting (2) yields the probability

$$\pi(y = 1|\boldsymbol{X}) = \frac{\exp(\boldsymbol{\beta} \cdot \boldsymbol{X})}{1 + \exp(\boldsymbol{\beta} \cdot \boldsymbol{X})} \tag{3}$$

In practice, a set of data $\mathcal{D} = \{\boldsymbol{X}, y\}$ of index $k = 1,\ldots,n$, is divided according to the value of y into two sets of size n_0 and n_1, respectively. The $\boldsymbol{\beta}$ are then determined, usually by maximizing the log-likelihood function

$$\arg\max_{\beta} \sum_{i=1}^{n} [y_i \log\ \pi_i + (1 - y_i)\ (1 - \log\ \pi_i)] \tag{4}$$

where the π_i carry the $\boldsymbol{\beta}$-dependence. A common issue that must often be addressed is unbalanced data, when $n_0 \gg n_1$, or the reverse, which can bias (4), resulting in poor estimates of the coefficients and degrade the fidelity of the model. See King and Zeng (2001) and Salas-Eljatib et al. (2018) for additional details. A similar issue arises when $\mathcal{D}$ contains clusters around one or more points in the data space (Merlo et al., 2006). Defining a subset of $\mathcal{D}$ using random subsampling is often employed in the case of

unbalanced data, whereas Tomek Link, Synthetic Minority Oversampling, and Neighborhood Cleaning are common solutions to clustered data (Elhassan and Aljurf, 2016; Guo and Wei, 2019). Here random subsampling was used to address the data imbalance as there was little clustering.

The result of LR (3) is a real value on the domain [0,1]. A threshold probability value is typically defined such that if $\pi < \pi_0$ then y is considered to equal 0, and y=1 when $\pi \geq \pi_0$. The most common selection for this threshold is π_0=0.5, for which rates of True Positive (TPR), False Positive (FPR), True Negative (TNR), and False Negative (FNR) were calculated for differing v^*. In this study π_0 was also allowed to vary, and the resulting changes in the TPR, and the FPR classifications were used to construct Receiver Operating Characteristic (ROC) curves, defined as TPR vs. FPR, and the Area Under Curve (AUC) measures of the ROC (Fawcett, 2006; Huang and Ling, 2005).

4. High-Current Prediction Model

4.1 Model Testing

Several speed thresholds $v^* = \bar{v} + z\sigma$, were used, where σ is the standard deviation of $v(j_0, i_0)$, z is a real number, and the mean meridional speed $\bar{v}(j_0, i_0)$ is adjacent to the port. The choices for z were 0, 0.5, 1.0, 1.5, and 2.0. For larger z the data became unbalanced. That is, the ratio of the number of values in the above threshold set was much smaller than the below threshold set. To eliminate this effect the larger of the two sets were randomly subsampled (without replacement) so that each set had the same number of points. The LR was then calculated. Rebalancing consistently yielded p<0.05. Subsampling of the original data was repeated 100 times, which was sufficient for the mean coefficient values to converge.

To test for overfitting, when the model hindcasts the training data but does not perform well for other data, a limited version of k-order cross-validation was applied (Aly, 2020; Pala and Atici, 2019). For each v^*, the indices corresponding to above and below threshold were divided as above. The two data sets were divided into k=5 sections of equal length, and each subsection was sequentially removed and the remaining data randomly subsampled to create sets of equal number. The average of the regression coefficients were then compared to those obtained regression on all the data. In all cases, the relative differences between the k-average and the full-data coefficients was < 2%, indicating there is little to no overfitting in the model.

4.2. Model Fidelity

The TPR, FPR, TNR, and FNR for π_0=0.5 were computed for each z with lags of 12, 24, and 48 hr (Table 1). This report focuses on results for the first two prediction windows. Predictors were frontal position alone, and frontal position in combination with wind, for a total of 20 different cases. For cases only based on frontal position, the TPR increased with z, more than doubling from $z = 0$ to $z = 2$, reaching about 90% for both the 12- and 24-hr forecasts (Table 1). In complement, the FNR decreased by a factor of almost 10, reaching a minimum < 10% for both lags. The FPR and TNR were essentially independent of z, being around 30% and 70% respectively. The inclusion of wind as a predictor produced a decrease in the TPR, but increased the TNR by about 10% for $z = 2$ for both lags. Wind also doubled the TPR for smaller z.

The ROC provided insights into the model response with the full range of π_0 (Figure 5). ROC curves in proximity to the upper-left corner of the domain (high TPR, low FPR) are considered to have high fidelity. Values of AUC range from 0 to 1, with the higher values generally considered an indication of an

accurate classification scheme. An AUC value of 0.5 indicates even probability of TP and FP, essentially random classification.

Cases with small z and no wind were close to the diagonal (random) with 0.6>AUC >0.5. Higher z had higher AUC, peaking over 0.8 for both lags. Adding wind to the predictor set had the largest impact on low z cases, with AUC now ~0.7, a nominal increase of about 20%. In contrast, the AUC for z=2 increased by ~7%.

Table 1. Confusion matrices for $v \geq v^* = \bar{v} + z\sigma$, using lagged frontal position south of the port, and lagged position and meridional winds, as indicated. 24-hr and 12-hr lags are examined. π_0=0.5.

	Format TPR FPR FNR TNR							
	Lagged Front				**Lagged Front+Wind**			
z	**24-hr**		**12-hr**		**24-hr**		**12-hr**	
0	35.7	32.8	39.6	29.7	69.3	38.3	68.7	36.2
	64.3	67.2	60.4	70.3	30.7	61.7	31.3	63.8
0.5	38.0	32.7	42.5	31.7	68.4	39.1	67.4	37.4
	62.	67.3	57.5	68.3	31.6	60.9	32.6	62.6
1.0	50.1	32.1	55.4	32.2	68.8	33.6	70.4	32.7
	49.8	67.9	44.6	67.8	31.2	66.4	29.6	67.3
1.5	67.1	33.0	74.7	33.5	73.9	29.4	81.5	29.1
	32.9	66.0	25.3	66.5	26.1	70.6	18.5	70.9
2	92.1	33.8	93.7	34.6	88.9	21.5	85.7	23.9
	7.9	66.2	6.3	65.4	11.1	78.5	14.3	76.1

These results demonstrate that meander propagation was the primary mechanism by which high currents occurred near Port Miami, but the inclusion of other factors, in this case wind, can sometimes be important. Machine learning is superior to correlative methods when considering occurrences above or below threshold. These results indicate a similar model based on observational data will be successful, but additional predictors may be needed as the model does not capture the full hydrodynamics near the port. Specifically, attention should be paid to the inclusion of tides, as represented by water levels from a nearby tidal station or the dominant tidal phase, in the next phase of this project.

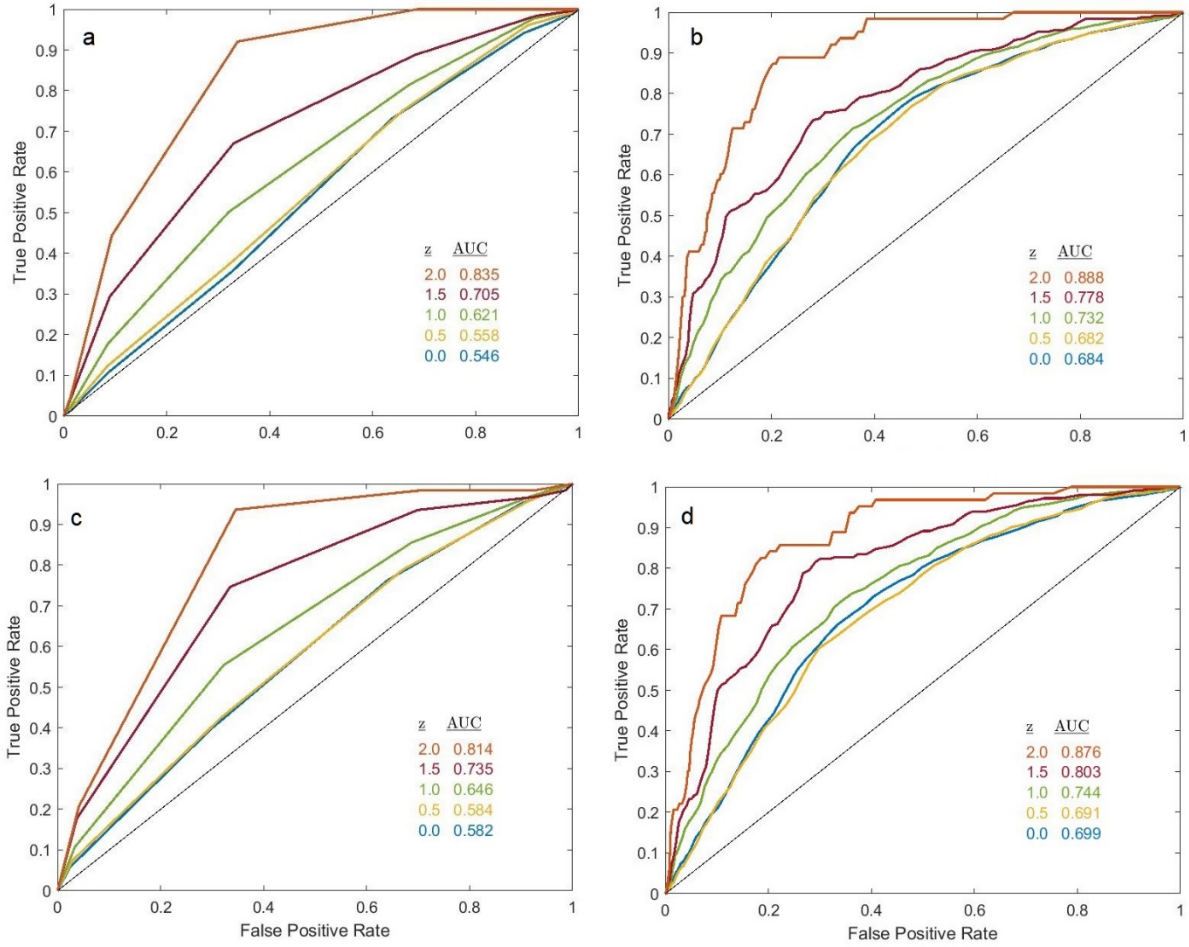

Figure 4. ROC curves and AUC values for varying *z* with a) 24-hr prediction using frontal position; c) 12-hr prediction using frontal position only as a predictor; b) 24-hr prediction using frontal position and wind as predictors; d) 12-hr prediction using frontal position and wind as predictors

5. Stakeholder engagement

Consulted with NOAA/CO-OPS (Chris Diveglio, Maritime Services Program Manager) and NOAA/NWS (Pablo Santos, Meteorologist In Charge, National Weather Service Office-Miami; Darren Wright, National Marine Program Manager; Brian LaMarre, Meteorologist In Charge, National Weather Service Office-Tampa Bay) on data availability. Contacted Laura DiBella, Executive Director, Florida Harbor Pilots Association, Capt. Sam Stephenson, Port Everglades Senior Harbor Pilot, and Capt. Jon Nitkin, Biscayne Pilots Association Senior Harbor Pilot, for status updates and input.

LITERATURE CITED

Aly, H.H., 2020. A novel approach for harmonic tidal currents constitutions forecasting using hybrid intelligent models based on clustering methodologies. Renewable Energy 147, 1554-1564.
Bleck, R., Benjamin, S.G., 1993. Regional weather prediction with a model combining terrain-following and isentropic coordinates. Part I: Model description. Monthly Weather Review 121, 1770-1785.
Bleck, R., Boudra, D.B., 1981. Initial Testing of a Numerical Ocean Circulation Model Using a Hybrid(Quasi-Isopycnic) Vertical Coordinate. Journal of Physical Oceanography 11, 755-770.
Bleck, R., Sun, S., Halliwell, G., 2001. Boundary conditions in HYCOM, p. 4.
Elhassan, T., Aljurf, M., 2016. Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. Global J Technol Optim S 1.
Fawcett, T., 2006. An introduction to ROC analysis. Pattern recognition letters 27, 861-874.
Guo, H., Wei, T., 2019. Logistic regression for imbalanced learning based on clustering. International Journal of Computational Science and Engineering 18, 54-64.
Hilbe, J.M., 2016. Practical guide to logistic regression. CRC Press.
Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied logistic regression. John Wiley & Sons.
Huang, J., Ling, C.X., 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering 17, 299-310.
King, G., Zeng, L., 2001. Logistic regression in rare events data. Political analysis 9, 137-163.
Merlo, J., Chaix, B., Ohlsson, H., Beckman, A., Johnell, K., Hjerpe, P., Råstam, L., Larsen, K., 2006. A brief conceptual tutorial of multilevel analysis in social epidemiology: using measures of clustering in multilevel logistic regression to investigate contextual phenomena. Journal of Epidemiology & Community Health 60, 290-297.
Pala, Z., Atici, R., 2019. Forecasting Sunspot Time Series Using Deep Learning Methods. Solar Physics 294, 50.
Salas-Eljatib, C., Fuentes-Ramirez, A., Gregoire, T.G., Altamirano, A., Yaitul, V., 2018. A study on the effects of unbalanced data when fitting logistic regression models in ecology. Ecological Indicators 85, 502-508.