Earth and Space Science

RESEARCH ARTICLE
10.1029/2020EA001179

Estuarine Forecasts at Daily Weather to Subseasonal Time Scales

Andrew C. Ross[1,2], Charles A. Stock[2], Keith W. Dixon[2], Marjorie A. M. Friedrichs[3], Raleigh R. Hood[4], Ming Li[4], Kathleen Pegion[5], Vincent Saba[6], and Gabriel A. Vecchi[7,8]

[1]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ, USA, [2]NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA, [3]Virginia Institute of Marine Science, William & Mary, Gloucester Point, VA, USA, [4]Horn Point Lab, University of Maryland Center for Environmental Science, Cambridge, MD, USA, [5]Atmospheric, Oceanic & Earth Sciences Department, George Mason University, Fairfax, VA, USA, [6]NOAA Northeast Fisheries Science Center, Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA, [7]Department of Geosciences, Princeton University, Princeton, NJ, USA, [8]Princeton Environmental Institute, Princeton University, Princeton, NJ, USA

Key Points:
- Skillful forecasts of estuarine temperature and salinity are possible beyond the few days typically considered
- Dissolved oxygen remains challenging to forecast
- Taking the mean of multiple forecasts driven by different atmospheric ensemble members improves the skill

Supporting Information:
- Supporting Information S1

Correspondence to:
A. C. Ross,
andrew.c.ross@noaa.gov

Citation:
Ross, A. C., Stock, C. A., Dixon, K. W., Friedrichs, M. A. M., Hood, R. R., Li, M., et al. (2020). Estuarine forecasts at daily weather to subseasonal time scales. *Earth and Space Science*, *7*, e2020EA001179. https://doi.org/10.1029/2020EA001179

Received 12 MAR 2020
Accepted 20 SEP 2020
Accepted article online 23 SEP 2020

Abstract Most present forecast systems for estuaries predict conditions for only a few days into the future. However, there are many reasons to expect that skillful estuarine forecasts are possible for longer time periods, including increasingly skillful extended atmospheric forecasts, the potential for lasting impacts of atmospheric forcing on estuarine conditions, and the predictability of tidal cycles. In this study, we test whether skillful estuarine forecasts are possible for up to 35 days into the future by combining an estuarine model of Chesapeake Bay with 35-day atmospheric forecasts from an operational weather model. When compared with both a hindcast simulation from the same estuarine model and with observations, the estuarine forecasts for surface water temperature are skillful up to about 2 weeks into the future, and the forecasts for bottom temperature, surface and bottom salinity, and density stratification are skillful for all or the majority of the forecast period. Bottom oxygen forecasts are skillful when compared to the model hindcast, but not when compared with observations. We also find that skill for all variables in the estuary can be improved by taking the mean of multiple estuarine forecasts driven by an ensemble of atmospheric forecasts. Finally, we examine the forecasts in detail using two case studies of extreme events, and we discuss opportunities for improving the forecast skill.

Plain Language Summary This paper evaluates a suite of forecasts for Chesapeake Bay water temperature, salinity, and dissolved oxygen created using a numerical model. By comparing the model forecasts with observations, we show that the model forecasts for temperature and salinity are more accurate than reference forecasts of previously observed conditions or the long-term mean; in other words, the forecasts are skillful. In general, the forecasts are skillful for at least 2 weeks into the future. Improvements to our forecasting system, such as predicting future river discharge into Chesapeake Bay, would likely improve the forecast skill even more. By showing that accurate, skillful forecasts are possible for a much longer time frame than previously considered, this paper takes an important step toward applying forecasts to improve water quality and fisheries management and to prepare for the impacts of extreme events like hurricanes and heat waves.

1. Introduction

Ocean model forecasts of water levels, temperature, salinity, and other properties for estuaries and similar coastal regions have primarily focused on lead times of a few days into the future. For example, in the United States, a number of Operational Forecast Systems provide guidance for temperature, salinity, water levels, and currents for the next 2 days for major coastal and estuarine regions including Chesapeake Bay (Lanerolle et al., 2011), the northern Gulf of Mexico (Wei et al., 2014), and San Francisco Bay (Peng et al., 2014). Experiments with both operational and research models have shown that short-term estuarine and coastal forecasts can help protect lives and property by predicting storm surges and inundation (Stanev et al., 2016) and by assisting with search and rescue operations (Cho et al., 2014) and can protect public and ecosystem health by forecasting the advection and dispersion of oil spills (Castanedo et al., 2006) and the development of harmful algal blooms (Brown et al., 2013). Forecasts for problems such as hypoxia can also make the public aware of the problem and its causes and solutions (Testa et al., 2017).

Although most estuarine ocean forecasts have focused on a few days of lead time, in the atmosphere, modern weather forecasts routinely have skill up to 10 days in advance as a result of substantial improvements to operational forecast models (Bauer et al., 2015). With additional improvements to these models, the extent of skillful weather forecasts may soon approach the estimated upper bound of weather-scale predictability of around 2 weeks (Zhang et al., 2019). Furthermore, recent modeling experiments have shown the ability to skillfully forecast weekly mean atmospheric conditions (but not daily weather variability) at subseasonal time scales of between 2 weeks to 2 months into the future (Li & Robertson, 2015; Pegion et al., 2019; Vitart, 2014). The lasting impact of initial land surface conditions (Koster et al., 2010, 2011) and the Madden-Julian oscillation of atmospheric convection over the tropical Indian and Pacific Oceans (Zhang, 2013) contribute to predictability at the subseasonal time scale. Skillful forecasts at even longer seasonal time scales of 3 months to 1 year are also possible (Baehr et al., 2015; Jia et al., 2015; MacLachlan et al., 2015). Much of the predictability at the seasonal time scale is driven by slow ocean modes, and skillful forecasts of monthly mean sea surface temperature (SST) have been produced for several large ocean regions (Hervieux et al., 2017; Hobday et al., 2016; Jacox et al., 2017; Siedlecki et al., 2016; Spillman & Alves, 2009; Stock et al., 2015).

There are multiple reasons to expect that skillful forecasts for estuaries are possible beyond the few days typically considered and potentially even beyond the 2-week limit for atmospheric weather forecasts. River discharge forcing is a major driver of variability in estuarine salinity and circulation; as one example, the relationship between the upstream length of saltwater intrusion in an estuary and the inflowing river discharge generally follows a power law (MacCready, 1999, 2007; Monismith et al., 2002). However, estuaries have a lagged response to river discharge forcing (e.g., Xu et al., 2012, found that modeled salinity lagged river discharge by 40 to 70 days in Chesapeake Bay), which implies a degree of future predictability from previously observed river discharge or upper estuary fluxes. Furthermore, tidal elevation and velocity follow nearly perfectly predictable cycles that likely confer a significant amount of predictability to tidally driven estuarine hydrodynamics. Finally, atmospheric forcing drives estuarine temperature and also has a role in modulating salinity and circulation (e.g., Li & Li, 2011), so the extended predictability and lasting impacts of atmospheric conditions are likely to transfer to predictability of estuarine conditions.

To test the limits of predictability of estuarine conditions, this study develops and tests a modeling system for subseasonal forecasting of conditions in Chesapeake Bay, a large coastal plain estuary in the Mid-Atlantic region of the United States. The model system combines an estuarine model with proven skill and routine use in the Chesapeake Bay research community (Da et al., 2018; Irby & Friedrichs, 2019; Irby et al., 2016, 2018; Scully, 2016; Xu et al., 2012) with atmospheric forecasts from an operational weather model (Pegion et al., 2019; Zhou et al., 2016, 2017; Zhu et al., 2018) (section 2). With this model system, we conduct an extensive set of retrospective forecast experiments in which the estuarine model is initialized with realistic conditions and then runs in forecast mode with forcing obtained from an atmospheric forecast. We use these forecast results to test whether temperature, salinity, and oxygen can be skillfully predicted using the model system (section 3). We also test whether the estuarine forecasts can be improved by generating and averaging multiple forecasts using multiple atmospheric model ensemble members-forecasts that represent uncertainty with slightly different initial conditions. Then, we examine model predictions for a few unique, high-impact events, and we discuss potential sources of future improvements to the model skill (section 4).

2. Methods

We tested whether skillful subseasonal forecasts for an estuary are possible by conducting forecast simulations using an estuarine model of Chesapeake Bay (ChesROMS) (Da et al., 2018; Xu et al., 2012) driven by forecasts from an atmospheric model from the SubX experiment (Pegion et al., 2019). We focused on predictability during April through August, when hypoxia (Bever et al., 2013; Hagy et al., 2004; Murphy et al., 2011; Officer et al., 1984; Taft et al., 1980) and other water quality issues (Glibert et al., 2001; Jacobs et al., 2014; Kaper et al., 1981; Mulholland et al., 2009) are common in the bay. We note that during these months, the predictability of the atmosphere over the United States is generally lower than during the late fall and winter seasons (DelSole et al., 2017; Pegion et al., 2019). The spring-summer focus of the experiment thus provides a high difficulty test case for subseasonal estuarine prediction, but one that is essential for intended applications.

2.1. Terminology

Because there are some differences between the terminologies commonly used by the atmospheric and ocean forecasting communities, we first clarify the meaning of some of the terms used in this paper. The goal of this paper is to assess the *skill* of our forecast model, which is a measure of the accuracy of the model forecasts compared to the accuracy of a (typically naive) reference forecast such as the long-term mean or the previous day's value (Murphy, 1988). Accuracy can be measured by metrics such as the root-mean-square error (RMSE), mean absolute error, or other metrics commonly used by ocean modelers to compare model simulations with observations. A forecast model is considered *skillful* if its forecasts have better accuracy than the reference forecasts.

Our estuarine *hindcast* simulation was designed to capture historically observed conditions; this follows the terminology commonly used by ocean modelers (Zhang et al., 2010). The hindcast was forced by best estimates of the boundary conditions for the atmosphere, ocean boundary, and river input, with these forcings varying during the duration of the hindcast; that is, the hindcast includes “future” information that would not even theoretically have been available to a prediction system. For example, the atmospheric forcing was obtained from a reanalysis that repeatedly assimilated observations to estimate the time-varying state of the atmosphere.

In our *retrospective forecast*, or *reforecast*, simulations, we used the ChesROMS estuarine model to forecast historical conditions using only data that could, in principle, have been available at the initialization times of the model simulations. For example, atmospheric forcing was obtained from a global weather model that was initialized at the same time as the estuarine model and subsequently ran in free-running forecast mode. The key property that makes these forecasts retrospective is that they were performed well after the verification time. For brevity, we also refer to the results from these experiments as simply *forecasts* in the text. Because these reforecast experiments capture the data constraints that a forecast model faces while providing an extended set of historical simulations for forecast evaluation, reforecast simulations can be used to estimate the accuracy and skill of a real or hypothetical *real-time forecast* system that forecasts future conditions naturally using only data that is presently available (Hamill et al., 2006).

We use the definition of *lead time* that is common in longer-range weather forecasting to refer to the time elapsed between when a forecast simulation was initialized and the earliest time when the forecast was valid. Under this definition, if a forecast simulation was initialized at 0000 1 January, the resulting forecast for a daily mean averaged over 0000–2359 1 January would be a forecast with 0-day lead, or a “Lead 0” forecast, and a “Lead 1” forecast from the same simulation would be a daily mean over 0000–2359 2 January.

2.2. ChesROMS Estuarine Model

The estuarine component of the model system is the Chesapeake Bay Regional Ocean Modeling System (ChesROMS) (Da et al., 2018; Xu et al., 2012). ChesROMS models the bay hydrodynamics on a 100 × 150 curvilinear horizontal grid (Figure 1), with a resolution of between 600 and 4,500 m, and with 20 terrain-following vertical layers. Additional details about the model configuration are presented in Da et al. (2018). The accuracy of the model at simulating velocity, tidal and nontidal elevation, temperature, and salinity over a range of time scales has been evaluated with hindcast simulations by Xu et al. (2012), Irby et al. (2016), and Da et al. (2018), so the hindcast accuracy will be only briefly examined in this manuscript. ChesROMS requires boundary conditions for the ocean boundary and for river discharge, which are discussed in section 2.5. ChesROMS is also driven by forcing from the atmosphere, including 3-hourly 2-m temperature and humidity, 10-m wind components, surface pressure, surface net longwave and shortwave radiation, and precipitation.

The ChesROMS estuarine model was coupled with the simple model for dissolved oxygen developed by Scully (2010, 2013, 2016). In this model, oxygen is treated as a passive tracer, oxygen concentrations in the uppermost model layer are set to saturation based on temperature and salinity at every time step, oxygen is removed (respired) at a rate of 1.4×10^{-4} mmole $\mathrm{m}^{-3}\,\mathrm{s}^{-1}$ that is constant at all times and everywhere in the estuarine portion of the model domain, and oxygen concentrations are not allowed to become negative. Oxygen concentration boundary conditions for both the open boundary and for river discharge are also set to saturation. This model, which has been termed the simple respiration model or constant respiration model (Bever et al., 2018; Irby et al., 2016), has been shown to perform comparably to complex biogeochemical models (Bever et al., 2013; Irby et al., 2016), and the computational cost savings make the large number

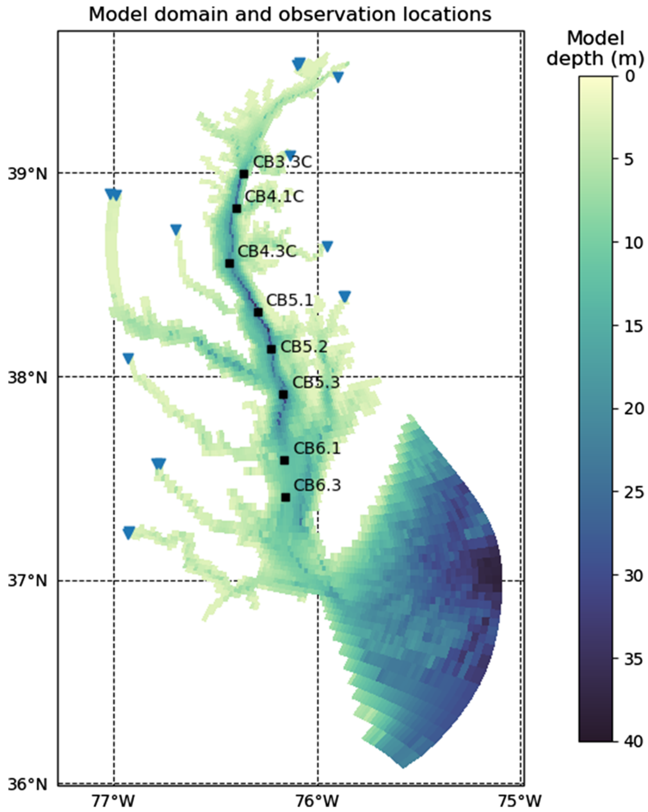

Figure 1. Map of the ChesROMS model domain, with the model bathymetry shaded for illustration. Black squares show the locations of observations used to evaluate the model experiments. Blue triangles show locations where river discharge was added.

of reforecast simulations needed for this study feasible. It should be noted that there are a few different variants of the simple respiration model that differ on the respiration rate, whether the rate varies seasonally or vertically, and how the surface concentration is set. The version we use is identical to the version in Scully (2016).

2.3. GEFS Atmospheric Model

We used forecasts from the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) (Zhou et al., 2016, 2017; Zhu et al., 2018) as atmospheric boundary conditions for our estuarine model reforecast simulations. The GEFS model output were obtained from the Subseasonal Experiment (SubX) data set (Pegion et al., 2019, doi:10.7916/D8PG249H), which contains subseasonal reforecast simulations from seven different models. We selected the GEFS model reforecasts for this study because the model output provided in the data set contains all variables necessary to drive the estuarine model, and the GEFS output also contains more ensemble members than other models in the data set.

The GEFS model reforecasts that we obtained from the SubX data set were initialized once weekly (every Wednesday) from 1999 to 2015. Each reforecast began with initial conditions obtained by assimilating observations and subsequently ran freely in forecast mode for 35 days. The SubX data set contains 11 GEFS ensemble members, each of which began with slightly different initial conditions. The output from each reforecast ensemble member was saved at daily resolution on a 1° grid, although the actual model was run at a higher resolution.

2.4. Atmospheric Forecast Bias and Drift Correction

We applied a simple bias and drift correction to the GEFS model output to prepare it for use as forcing for the ChesROMS estuarine model. This correction is necessary because weather and climate models commonly contain biases in their mean state, and, as a result of initializations that consistently deviate from the model mean state, they may also contain biases that are not stationary over time (Hermanson et al., 2018;

Stockdale, 1997). The method used in this study corrected the atmospheric model mean but did not adjust the variance or other moments of the distribution; this potential limitation is discussed in section 4.2. For simplicity, we refer to the bias and drift correction as only "bias correction" in the remainder of the text, with the intention that we are correcting for the mean difference between the GEFS data and the observations (bias) and are allowing this bias correction factor to be a function of the forecast lead time (drift).

To apply the correction, we first calculated a lead-dependent daily climatology for the ensemble mean of each GEFS variable needed to force ChesROMS using the method described in Pegion et al. (2019). The lead-dependent model climatology describes the model mean climate $\overline{M}$, averaged over 1999 to 2015 and for a given point, as a function of forecast initialization day of year i and forecast lead time l: $\overline{M}(i,l)$. Because GEFS reforecasts were initialized once weekly, the climatology for each combination of initialization day of year and lead is noisy because it was produced from the average of at most a few simulations. The method developed by Pegion et al. (2019) and applied in this study uses a triangular filter to produce a smooth lead-dependent climatology. After calculation, the GEFS lead-dependent climatology was bilinearly interpolated to a uniform 0.2° grid.

Next, to obtain a lead-dependent observed climatology $\overline{O}(i,l)$, we applied the same averaging method to reanalysis data from the North American Regional Reanalysis (NARR) (Mesinger et al., 2006), also interpolated to a uniform 0.2° grid and averaged to daily means. After determining the NARR daily climatology, each day of year in the NARR climatology was matched with possible combinations of initialization day and lead.

Finally, we defined a lead-dependent correction factor Δ for each variable and point:

$$\Delta(i,l) = \overline{O}(i,l) - \overline{M}(i,l) \tag{1}$$

and added this correction to the GEFS forecast for the given point, initialization date, and lead time. In this way, bias corrections were applied for each variable needed to drive ChesROMS. The u and v wind components were corrected simultaneously by calculating a correction factor for overall wind speed. For wind speed and precipitation, which should not be below 0, we used a ratio in Equation 1 and applied the correction by multiplying.

2.5. Model Forcing and Experiments

As a first step for our forecasting system, we used the ChesROMS model to run a hindcast simulation for 18 years (1998 to 2015). The first year was discarded afterward to eliminate the spinup period, leaving data from 1999 to 2015. In the hindcast simulation, the estuarine model was forced using 3-hourly atmospheric conditions from the NARR interpolated to the uniform 0.2° grid, as in other studies using ChesROMS (Xu et al., 2012; Scully, 2016). A hindcast simulation of the mechanistic Dynamic Land Ecosystem Model (DLEM) (Tian et al., 2015; Yang, Tian, Friedrichs, Hopkinson, et al., 2015; Yang, Tian, Friedrichs, Liu, et al., 2015) provided daily freshwater fluxes for ten rivers that drain into the Chesapeake (Figure 1). Using DLEM discharge matches the version of ChesROMS developed by Feng et al. (2015), although other studies using ChesROMS have used discharge observed at gauging stations. We used DLEM discharge as it allows potential future work to more easily switch to a full biogeochemical model (such as that in Feng et al., 2015) or to use model forecast river discharge. The temperature of the river discharge was set to the 1980–2011 climatological monthly mean values observed at gauging stations (Feng et al., 2015). Temperature and salinity conditions along the ocean boundary were set to radiation with nudging toward World Ocean Atlas climatology, and sea surface elevation was set using a nontidal component, derived from observations at two coastal stations, plus a tidal component calculated using seven harmonic constituents obtained from the Advanced Circulation (ADCIRC) model (Mukai et al., 2002). Boundary elevation and momentum used the Chapman (1985) and Flather (1976) conditions, respectively, and also included tidal currents calculated from the Mukai et al. (2002) data. These boundary conditions and other configuration choices not mentioned match those used in Da et al. (2018).

Restart files generated once per day during the estuarine hindcast simulation were subsequently used as initial conditions for retrospective forecast simulations. In these reforecast experiments, atmospheric forcing was obtained from the bias-corrected GEFS model reforecasts. River discharge was specified as a smoothed daily climatology averaged over 1980 to 2011, and ocean boundary conditions used the same sources as in the hindcast simulation, except the nontidal water level was fixed at the value obtained from the two coastal stations for the day of the model initialization. We also note that because the reforecast experiments were driven by daily averages from GEFS, the reforecasts will not capture diurnal variability driven by the atmosphere, and our analysis is primarily focused on daily averaged ChesROMS output. Although these simplifications may reduce the forecast skill, they are necessary in the absence of skillful models for forecasting river discharge and broader ocean boundary conditions. We will assess the forecast skill despite these simplifications in section 3, and we will discuss possible future improvements in section 4.2.

The reforecasts were initialized on the first Wednesday of each month (matching the GEFS model that was initialized on Wednesdays) from April to August of 1999 to 2015 and were run for 35 days. For each initialization date, we ran five separate estuarine model ensemble members, each with atmospheric forcing obtained from a different GEFS ensemble member. Only the atmospheric forcing differed between ensembles; each member used the same estuary initial conditions and river discharge and open boundary forcing. Overall, the retrospective forecast suite contains a total of 425 model runs and 14,875 days of model simulation.

2.6. Forecast Skill Assessment

We tested three different aspects of the model skill: how well the hindcast simulation compares with observations, how well the reforecast simulations compare with the hindcast experiment, and how well the reforecast simulations compare with observations. The comparison between the hindcast simulation and the observations is not the main focus of this paper because versions of ChesROMS have been compared with observations by Xu et al. (2012) and Irby et al. (2016), so the majority of the assessment of the hindcast simulation is provided in the supporting information (Tables S1–S5). We focus on the comparisons between the reforecast simulations and the hindcast and observations because these comparisons provide a baseline to test the skill of our model forecasts.

To compare the hindcast and reforecast skill with independent observations, we obtained observations of surface and bottom temperature, salinity, and dissolved oxygen from the Chesapeake Bay Program (CBP) data set (Chesapeake Bay Program, 2018). This data set contains instantaneous vertical profiles from over 100 locations in the bay, and in many locations the profiles were taken twice every month during the warm season. Although we evaluate the hindcast skill for 39 locations in the supporting information, in the main text we select data from eight locations that are representative of the center of the bay and are roughly evenly spaced apart (Figure 1). To match these instantaneous observations with the model forecasts, for each observation, we rounded the time of observation to the nearest hour and selected the model instantaneous value from the same date and hour. For surface variables, data from 2 m depth were selected from the observations, and the model output were interpolated to the same depth. For bottom variables, the deepest value was selected from each vertical profile and from the model output. Note that some error is potentially introduced to the bottom data comparison due to the model bathymetry and profile depths not matching exactly, either due to the inability of the model to resolve variations in bathymetry or to incomplete vertical profiles. We also used the CBP observations to assess the model skill at predicting density stratification. Density was calculated using the surface and bottom temperature and salinity observations, and stratification was defined as the difference between the bottom and surface density. To evaluate the model predictions of oxygen integrated over space, we compared the hypoxic volume predicted by the model (defined as the volume of water in the model with a dissolved oxygen concentration below 2 mg l^{-1}) with the hypoxic volume estimated from the CBP observations by Bever et al. (2013).

For a few qualitative comparisons, we also obtained observations of SST from the National Oceanic and Atmospheric Administration (NOAA) National Ocean Service monitoring station at Cambridge, MD. Hourly observations from this station were averaged to daily means and compared with the model daily means. Although this monitoring station and others in the bay provide fairly continuous observations of SST, we primarily use the CBP observations because the monitoring stations have shorter periods of record and several stations have evidence of instrument errors.

In addition to comparing the reforecasts with the CBP observations, we also compared the reforecast daily means with the ChesROMS hindcast daily means. Assessing forecast skill against the model hindcast augments the observation-based assessment in two ways. First, this comparison represents a situation where the forecast model is perfectly initialized and exactly captures the dynamics of the system, and thus it provides an isolated measure of the predictability limits imposed by the imperfect atmospheric forecasts and the climatological river discharge and open boundary conditions. Second, under the assumption that the hindcast simulation reasonably captures the dynamics of the system, comparing the forecasts against the

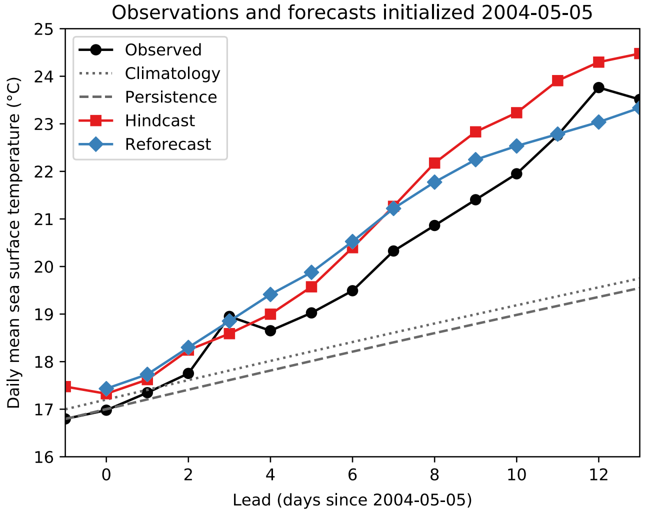

Figure 2. Example of SST observations, observation-based reference forecasts, and the model-based hindcast and reforecast predictions for the Cambridge, MD, station in May 2004.

hindcast provides a spatially and temporally continuous perspective on forecast skill that cannot be provided by patchy observational data. Due to these considerations, in section 3 we will begin by analyzing the forecast skill against the hindcast and then assess how the skill changes when the forecasts are compared with direct observations.

The skill of the forecasts was evaluated using the mean square error (MSE) skill score (Murphy, 1988; Murphy & Epstein, 1989), which compares the percent improvement of the MSE calculated for a set of forecasts and observations relative to the MSE for a set of reference forecasts and observations:

$$\text{Skill} = 100\% \times \left(1 - \frac{\text{MSE}_{\text{forecast}}}{\text{MSE}_{\text{reference}}}\right). \tag{2}$$

A skill score of 100% indicates a perfect improvement over the reference forecast (a perfect match to the observations), and a skill score of 0% indicates that the forecast is not any more accurate than the reference forecast. Positive skill scores indicate that the model is skillful, while negative skill scores, which are produced when $\text{MSE}_{\text{forecast}} > \text{MSE}_{\text{reference}}$, indicate an unskillful model. The means in Equation 2 were taken over time and, in most cases, over the eight CBP stations along the center of the bay (Figure 1) to provide integrative measures of forecast skill.

We considered two reference forecasts to assess the skill of the estuarine model with Equation 2: a forecast of climatological mean conditions for the given location and verification date and a forecast of persistence of the anomaly (the difference between the observed value and the climatological mean) observed at the given location averaged over the day before the forecast initialization. The ultimate choice of reference forecast is commonly the most accurate reference forecast, which tends to be persistence for short-term forecasts and climatology for long-term forecasts (Murphy, 1992). We found that climatology was typically the more accurate reference forecast for the variables and daily weather to subseasonal time scales examined here, so for brevity we present only skill scores calculated using the climatological reference forecast and provide scores calculated relative to the persistence forecast in the supporting information. Note that other reference forecasts are possible; for example, Murphy (1992) shows that persistence and climatology linearly combined based on autocorrelation is more accurate than either reference forecast alone. For both the comparison with the hindcast and with the observations, the climatological mean for the reference forecast was determined using the same method used to smooth the GEFS daily climatology.

To reduce the impact of systematic errors in the estuarine model, we applied a stationary bias correction (without correction for drift) to the estuarine hindcast and reforecast output when using the skill metric (Equation 2) to compare the model output with the observations. With this correction, we assess the model skill based on its capacity to simulate anomalies that are consistent with the observations, while allowing for the possibility of a simple mean offset. This correction is not strictly necessary but makes the skill score more meaningful by removing the impact of consistent model biases. A similar bias correction could be applied in a real-time forecasting context using the same model hindcast simulations, so the model-observation comparison remains fair. The hindcast and reforecast bias was corrected by subtracting the difference between the hindcast climatology and the observed climatology from the reforecasts. The reforecast climatology and mean biases are similar to the hindcast climatology and biases because the reforecasts were initialized from the hindcast and driven with bias-corrected atmospheric conditions and climatological conditions for other inputs; however, the reforecast climatology and biases may still differ from the hindcast, and the simple method applied here will not correct for this difference. Additionally, the stationary correction is not a function of forecast lead time and will not remove any changes in bias over time that may be present in the estuarine reforecasts. However, an advantage of this simple method that does not use the retrospective forecasts to calculate bias is that it avoids artificially increasing the skill of the forecasts.

To demonstrate the forecast evaluation described above graphically, we selected data for the first 2 weeks of the May 2004 forecast for the National Ocean Service monitoring station located at Cambridge, MD, a choice that allows a particularly clear illustration, and plotted the results in Figure 2. Because this example takes

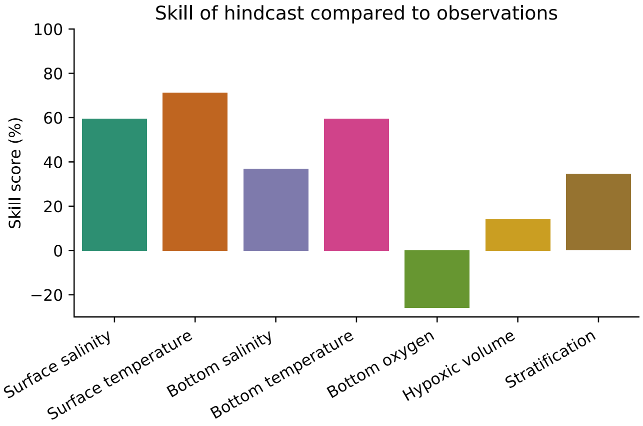

Figure 3. Skill scores comparing the hindcast simulation with the CBP observations, using the observed climatology as a reference.

place during spring, the reference forecasts of both climatology and anomaly persistence warm during the forecast period; however, persistence remains cooler than climatology, which is consistent with the modestly cooler than average conditions observed by the monitoring station during the day before the initialization of the forecast. Both the observations, hindcast, and reforecast indicate a rapid warming in the first two weeks, and the reforecast simulation is closer to the observations than the reference forecast for all but the first 3 days, which indicates a generally skillful forecast. Note that a mean bias exists in the reforecast even at Lead 0 because the hindcast simulation used for initialization is also biased. While we account for this bias via the simple correction discussed in the preceding paragraph, it could also be reduced by using data assimilation.

3. Results

Before evaluating forecasts produced using ChesROMS, we first evaluate the ChesROMS hindcast simulation during April through August of 1999 through 2015 (supporting information Tables S1–S5). For surface and bottom temperature and salinity, nearly all correlation coefficients are above 0.5, and the mean square and absolute errors are low relative to the modeled and observed means. The central portion of the bay in the model has a modest warm bias at the surface, and the majority of the bay is slightly too saline at both the surface and the bottom. These results are comparable to the evaluations of temperature and salinity in other studies that have used ChesROMS (Da et al., 2018; Irby et al., 2016; Xu et al., 2012). To predict oxygen, we coupled ChesROMS with the simple oxygen model developed by Scully (2016). This model does fairly well at capturing the seasonal cycle of oxygen, as indicated by the correlation coefficients in Table S5 that are mostly around 0.7 to 0.8 and RMSEs that are typically less than the observed standard deviations. However, the oxygen errors are high relative to the modeled and observed means, and some of this error stems from a low bias in the model. When hypoxic volume, the volume of water with an oxygen concentration below 2 mg L^{-1}, is calculated for the entire hindcast and compared with the data-derived estimates from Bever et al. (2013), we obtain an R^2 value of 0.86, which compares well with the R^2 value of 0.82 obtained by Scully (2016) for a longer hindcast simulation. We obtain a lower R^2 of 0.67 when the comparison is limited to the months from April to August.

In a forecasting context, it is important to consider how the errors of the model compare to the errors of a simple prediction that can be readily made without a model, such as a prediction of the seasonally varying long-term mean (i.e., compared to climatology) (section 2.6). In this comparison (Figure 3), the hindcast surface and bottom temperature and salinity predictions have much lower errors compared to the observations than the climatology does, as indicated by the large positive skill. However, the negative skill for bottom oxygen means that the error of the hindcast oxygen simulation is higher than the error of the long-term climatology; in other words, although the model predicts the regular seasonal cycle of oxygen fairly well (Table S5) it cannot reliably predict the anomalies relative to the seasonal cycle. This comparison, which requires the model to skillfully predict instantaneous oxygen measurements at eight discrete points in space (Figure 1), is a stringent test for a model with fairly coarse resolution and smooth bathymetry that predicts oxygen using a simple parameterization (section 2.2). We will discuss improvements to the model system that may improve the oxygen skill in section 4.2. Due to the substantial variability of oxygen in Chesapeake Bay over both time and space, we also expect that assessing predictions of oxygen averaged over time or space, rather than assessing instantaneous predictions for a few points, would result in positive skill. Indeed, the hindcast predictions of hypoxic volume do have a modest amount of skill compared to the estimated climatology of hypoxic volume (Figure 3). However, in the remainder of the results, we will continue with the pointwise assessment of bottom oxygen despite the apparent lack of skill when compared to instantaneous observations to keep the forecast assessment for oxygen comparable to the assessments of temperature and salinity. Comparing the pointwise oxygen forecasts with the hindcast may also still provide information about the potential forecastability of oxygen.

Next, we briefly examine the skill of the daily mean GEFS atmospheric temperature forecasts over Chesapeake Bay from the five ensemble members that we used to drive the estuarine model (Figure 4). Without

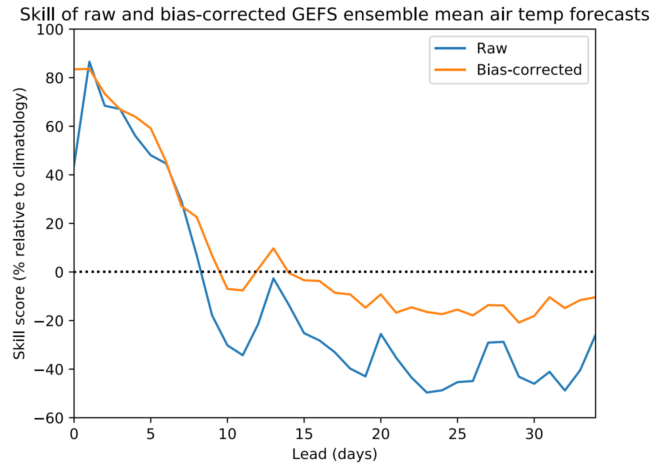

Figure 4. Skill score for raw (blue) and bias-corrected (orange) GEFS daily mean 2-m air temperature forecasts. Consistent with the model output used in this study, the forecasts are means of the first five GEFS ensemble members, and skill was only calculated for forecasts initialized between April and August and was averaged over a region representative of Chesapeake Bay (76°W, 38–39°N).

bias and drift correction, the five-member ensemble mean forecast remains skillful relative to climatology out to 7 days of lead time. Including bias correction makes a minor difference to the skill in the first week but does increase the length of the skillful forecast period to nine days. The raw forecast skill is lower for the first forecast day than for the second, which suggests the presence of an initial shock in the model as it adjusts to the assimilated initial conditions. With bias correction applied, this shock is removed, and the first forecast day is about as skillful as the second day. Additionally, note that the corrected atmospheric forecast skill in Figure 4 settles at a value below 0 for leads longer than 20 days; this occurs in part because the correction only modifies the forecast mean and does not correct the variance or other moments of the forecast distribution.

Even when only a single atmospheric ensemble member is used to drive the estuarine model, the estuarine surface temperature forecast skill evaluated against the hindcast simulation exceeds the atmospheric ensemble mean skill (comparing Figure 5 with Figure 4). In addition to surface temperature, all other variables can be skillfully forecast with at least 10 days of lead time, and surface salinity can be skillfully forecast for the entire period. Note that the forecast system has more difficulty exceeding the skill of a persistence forecast for surface salinity than the skill of a climatological forecast, but the forecasts are nevertheless skillful (Figure S1). Skill for bottom salinity and stratification and for surface temperature and bottom oxygen appear to be similar, suggesting connections between these variables. Surface salinity skill is higher than bottom salinity skill, which is consistent with numerical model hindcasts routinely simulating surface salinity more accurately than bottom salinity (Irby et al., 2016; Tables S2 and S4). Surface salinity variability in Chesapeake Bay is driven by tides and wind at shorter time scales and by previous river discharge at longer time scales (Xu et al., 2012), all factors with effects that can be well simulated by the model and well predicted in a forecasting context. Bottom salinity is also driven by factors including mixing and advection of shelf salinity (Lee & Lwiza, 2008) that are more difficult to model and to forecast. Furthermore, bottom salinity is less variable than

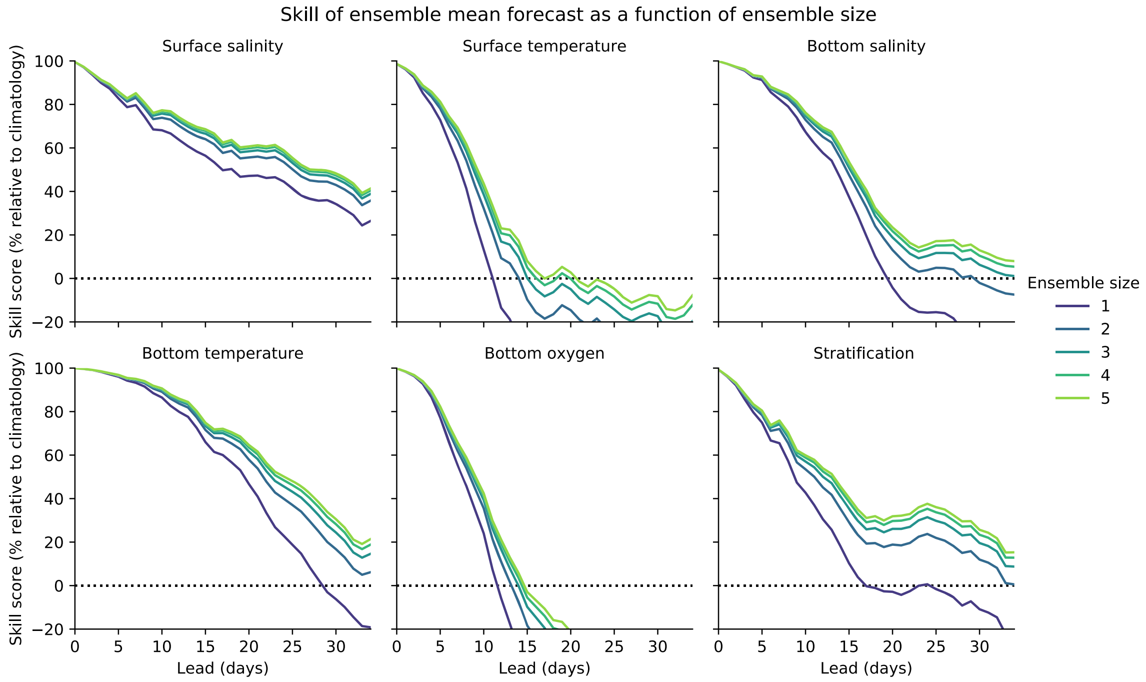

Figure 5. Skill of the ensemble mean estuarine model reforecasts evaluated against the hindcast simulation as a function of the size of the ensemble. The skill was calculated for the average MSE of forecasts from stations in the center of the bay (Figure 1) and compared against the MSE of a climatological reference forecast. For ensemble sizes between 1 and 4, the MSE was also averaged for ensemble mean forecasts from ensembles representing all possible combinations.

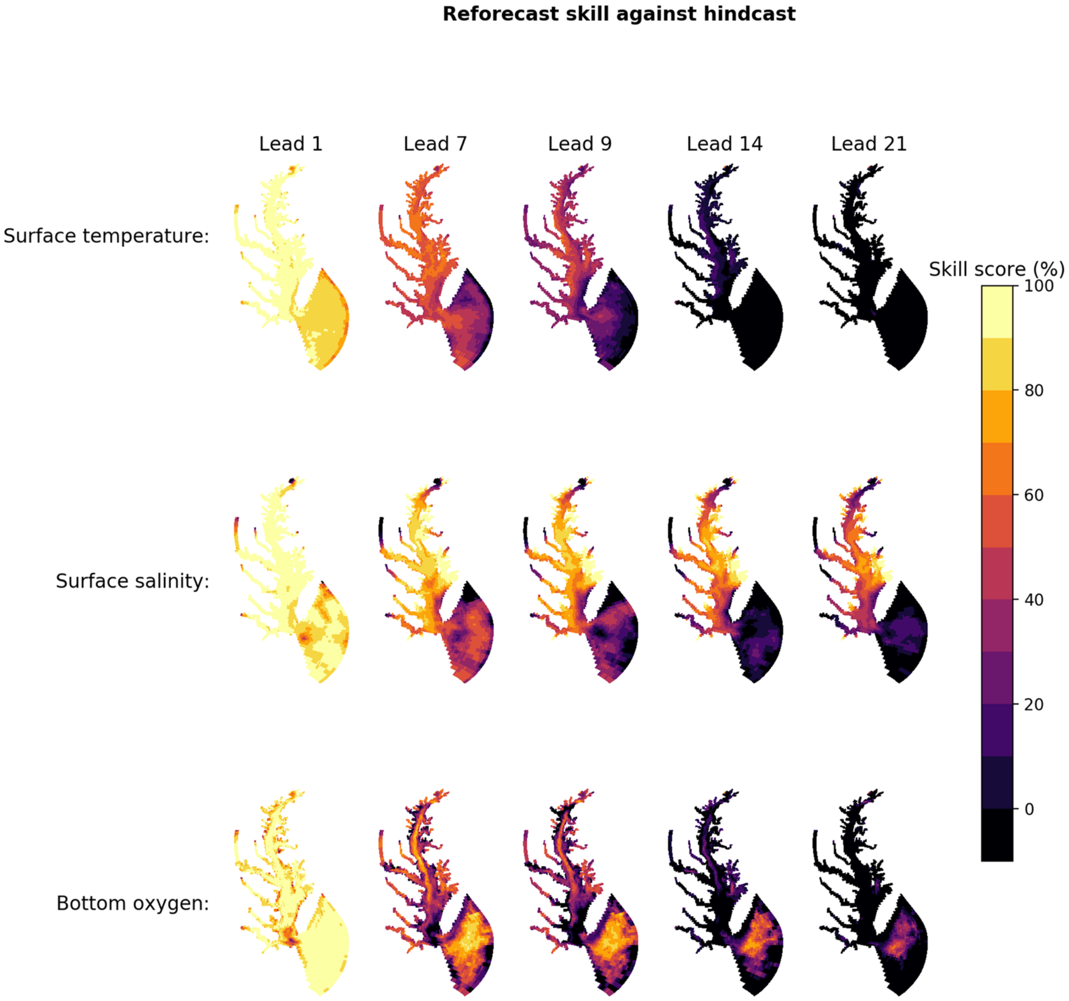

Figure 6. Skill of ensemble mean forecasts evaluated against the hindcast simulation at all model grid points. For this figure, surface values were taken from the uppermost model layer.

surface salinity, so a climatological reference forecast has a lower MSE for bottom salinity than for surface salinity, and therefore, model forecasts of bottom salinity must be more accurate to obtain a high skill score.

Adding an additional ensemble member simulation and using the ensemble mean as the forecast significantly increases the estuarine forecast skill and extends the skillful lead time by several days for most variables (Figure 5). The difference between the skill of the ensemble mean forecasts and the individual forecasts is especially high for surface temperature and stratification, two variables that have particularly strong responses to atmospheric forcing. The ensemble mean has a negligible difference in skill compared to the individual forecast only in roughly the first 5 days of the forecast period, which is not too surprising because (1) the atmospheric forecasts take some time to diverge from the perturbed initial conditions and (2) the estuarine simulations begin from the same initial conditions and take time to respond to the diverging atmospheric forcing. Although adding additional members to the ensemble always increased the skill of the ensemble mean, the marginal improvement from an additional ensemble member quickly diminishes, and the skill of a five-member ensemble was only modestly better than the skill of a three-member ensemble. Therefore, although we will focus on the five-member mean in the remainder of the results, we expect that a three-member ensemble would produce similar results and would be adequate in future studies if computational resources were limited.

When assessed against the hindcast simulation, the skill of the forecast SST is fairly uniformly distributed over the bay (Figure 6) with only slightly higher skill over the western side of the bay at intermediate lead times. The spatially uniform decline toward unskillful forecasts after around two weeks is consistent with

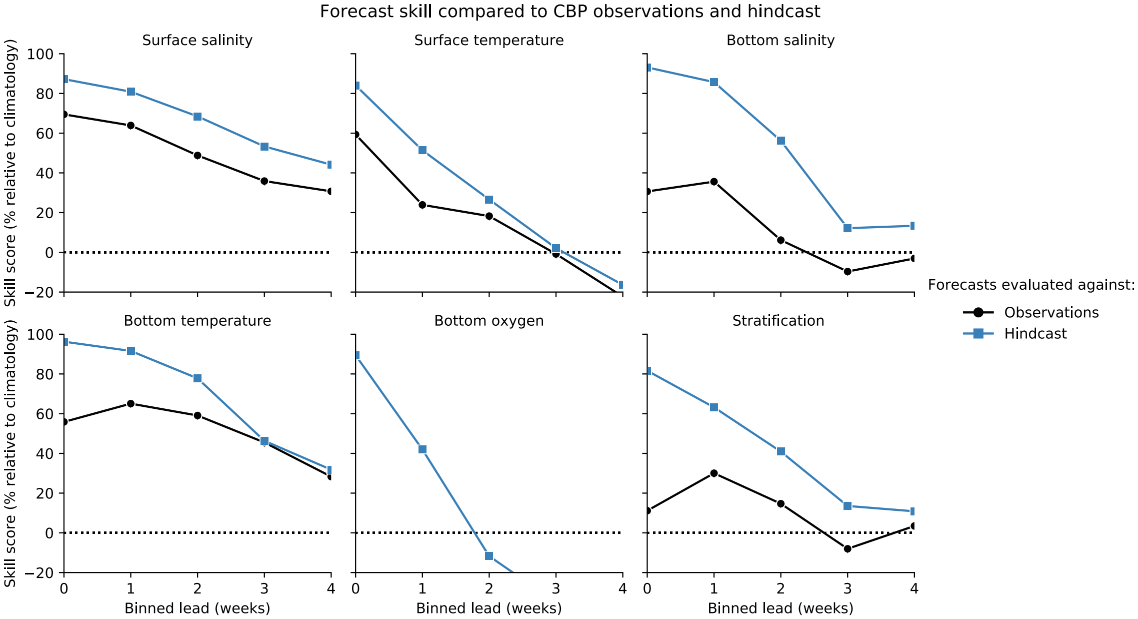

Figure 7. Skill of the ensemble mean forecasts evaluated against the hindcast simulation (blue) and against the CBP observations (black). For this comparison, the forecasts have been corrected for mean biases by comparing the hindcast and observed climatologies.

atmospheric forcing being the primary driver of SST in the bay. The surface salinity forecasts are skillful throughout the majority of the bay at short lead times. Skill is maintained over the full forecast window for the central portion of the bay but quickly drops to 0 near the mouths of the Susquehanna River and the other tributaries due to the use of climatological river discharge forcing in the reforecast simulations. Similarly, at longer lead times, the highest salinity skill is found near the eastern shore, the farthest from any of the river inputs and the open boundary. Aside from the first few days, skill for bottom oxygen is primarily concentrated in the deep central channel of the bay, the eastern shore, and in a few of the tributaries. Skill relative to climatology is also present outside of the bay along the shelf; however, this skill is an artifact of persistence (Figure S2).

The previous forecast evaluations have compared forecast skill against the hindcast simulation, so in Figure 7 we compare the forecasts with independent observations to get an estimate of real-world forecast skill. The reforecast-hindcast comparison in this figure also uses hourly model output that was subsampled to match the times of the observations to ensure a fair comparison; note that the reforecasts were forced with daily average atmospheric forecasts and will fail to capture some subdaily variability but will include variability due to tides. These results confirm that the skill identified by comparing with the hindcast is also present when comparing with observations, although naturally the skill when comparing with observations is somewhat lower. The only exception is for bottom oxygen, which is not skillful at any lead time when compared with the observations (skill scores for oxygen range from −28% to −41% and are not shown in Figure 7). Note that the forecast skill compared to the observations for Lead 0 is nearly the same as the hindcast skill compared to the observations in Figure 3, and in later leads the forecast skill generally declines. In other words, the skill of the forecasts is constrained by the skill of the hindcast that was run with the same estuary model and was used to initialize the forecasts.

4. Discussion

Overall, the results have shown that temperature, salinity, and stratification for the majority of Chesapeake Bay can be skillfully forecast at least 2 weeks in advance, which is beyond the extent of skillful atmospheric temperature forecasts from the GEFS model for the same region. Oxygen can also be forecast with skill when the forecasts are compared with the model hindcast, but not when they are compared with the observations. The forecast skill for all variables was improved by taking the mean of multiple estuarine model ensemble members driven by multiple atmospheric forecasts, although the improvement primarily occurred after the first 5 days and as the number of ensemble members was increased from one to three.

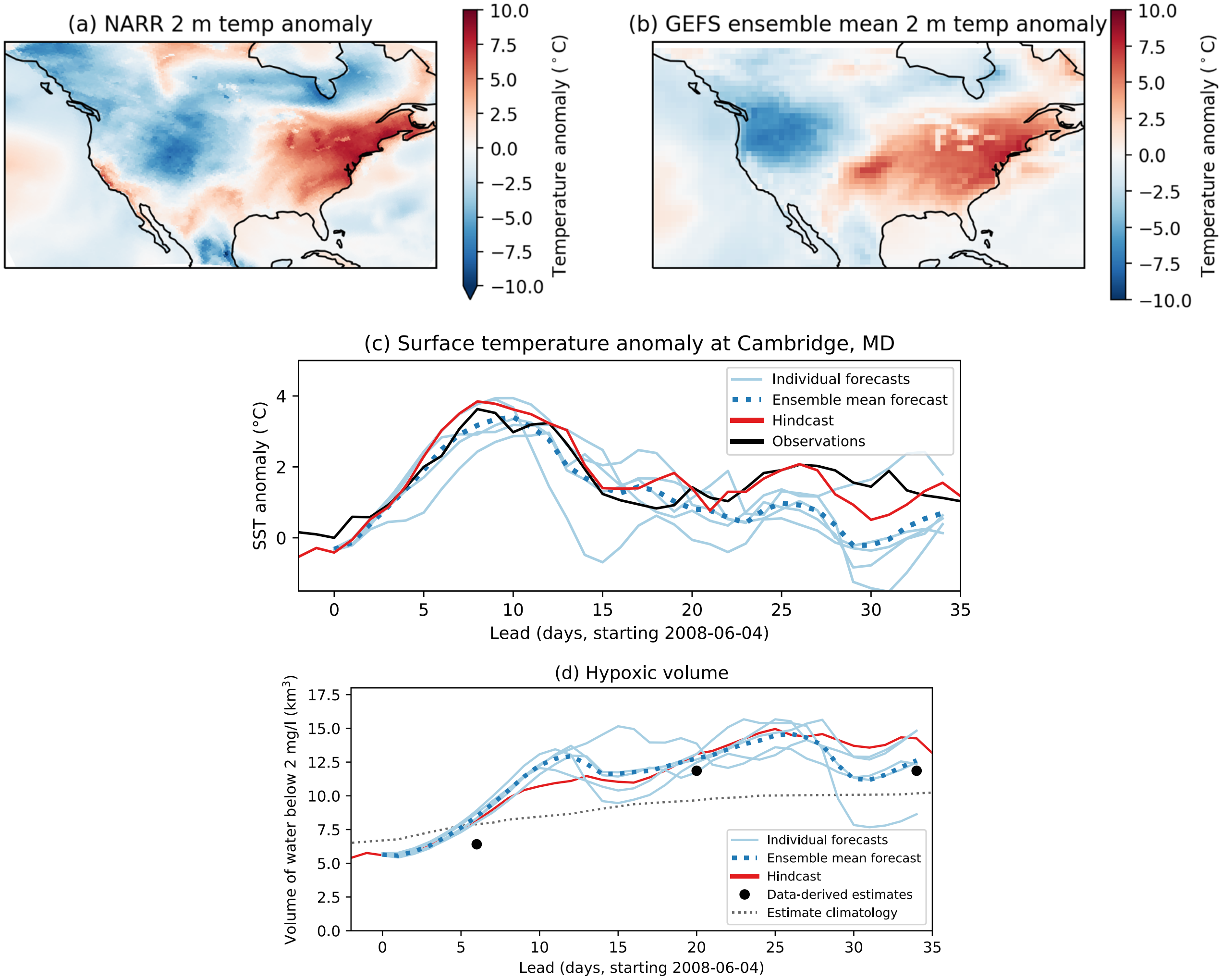

Figure 8. (a) NARR 2-m temperature anomaly for 9 June 2008. (b) GEFS five-member ensemble mean 2-m temperature anomaly forecast issued 4 June 2008 and valid 9 June 2008. (c and d) Observations or data-derived estimates (black), model hindcast (red), individual ensemble member forecasts (light blue), and ensemble mean forecast (dashed dark blue) for SST and baywide hypoxic volume, respectively. Hypoxic volume estimates are from Bever et al. (2013).

The forecast skill assessment presented so far was conducted at a high level by averaging over a long series of forecasts, so in this discussion we use two case studies to gain a perspective on the skill of individual forecasts and to discuss the strengths and weaknesses of the model forecast system (section 4.1). Then, we discuss modifications and additions to the forecast system that could improve the skill (section 4.2).

4.1. Forecast Case Studies

We chose two events for case studies that represent high-impact phenomena that occasionally affect the Chesapeake Bay region during the warm season: first, a heat wave that occurred over the eastern United States in June 2008, and second, a hurricane (Irene) that passed over the bay area in August 2011. These case studies also represent two types of extreme events with particularly promising forecastability at subseasonal to seasonal time scales (Vitart et al., 2019; Xiang et al., 2015).

4.1.1. June 2008

The first case study examines the heat wave that impacted the eastern United States in early June 2008. Over Chesapeake Bay, the warmest air temperature was observed on 9 June, and on this day temperatures of 5°C to 10°C above the 1999–2015 climatological mean were present over the entire eastern United States (Figure 8a). The GEFS weather forecasts used in this study were initialized on 4 June, and the ensemble mean 2-m temperature forecast for 9 June (Lead 5) was a close match to the reanalysis (Figure 8b). Consistent with the predictability of the atmospheric conditions during this event, SSTs were also accurately predicted by the forecast system (Figure 8c). Note that although the atmospheric temperature peaked on

Lead 5, the SST observed at the National Ocean Service station in Cambridge, MD, did not peak until Lead 8. The hindcast also peaked at Lead 8 at Cambridge, while the peak of the reforecast ensemble mean was delayed until Lead 10. This delay is consistent with the hypothesis that some of the extended skill of the estuarine forecasts is from the delayed, autocorrelated response of the estuary to predictable weather conditions. After the peak of the heat wave, both observed and hindcast temperatures gradually subsided. The ensemble mean forecast followed the hindcast closely until around Lead 20, when the forecast continued reverting to normal while the hindcast and observations leveled off at around 1°C above normal. The behavior of the forecasts is consistent with the general tendency for the mean SST forecast to approach the long-term climatology toward the end of the forecast period because of unskillful atmospheric forecasts that also fluctuate around climatology, the use of climatological temperatures for river input, and relaxation toward climatology at the boundary. Finally, the estimated baywide volume of hypoxic water increased significantly during and following the heat wave (Figure 8d), and every forecast ensemble member correctly predicted the shift from below-normal to above-normal hypoxic volume conditions, although the increase in hypoxic volume occurred too quickly in both the forecasts and the hindcast.

Although obtaining skill for this particular event is not too surprising because the maximum air temperatures occurred only a few days after the start of the forecast, recent research does suggest that heat waves in many regions may be predictable over even longer time scales. For example, Teng et al. (2013) identified an atmospheric pattern that typically precedes summer heat waves over the continental United States by 2 weeks, Lavaysse et al. (2019) found that model skill at forecasting European extreme temperatures extends to about 2 weeks, and McKinnon et al. (2016) found a pattern of Pacific Ocean SSTs that is predictive of warm temperatures in the Eastern US by up to 50 days in advance.

4.1.2. August 2011

Hurricane Irene passed near the mouth of Chesapeake Bay on 27–28 August 2011 and brought storm surge and inland flooding from heavy rain (Avila & Stewart, 2013). In addition to flooding, the storm had other impacts including increased oyster mortality due to reduced salinity in nearby Delaware Bay (Munroe et al., 2013) and increased turbidity in Chesapeake Bay (Palinkas et al., 2014; Xie et al., 2018). The storm also significantly cooled the surface water in Chesapeake Bay (Figure 9a) and aerated the bottom water (Figure 9b).

The GEFS and estuarine model reforecasts in this study were initialized on 3 August, long before Irene formed as a tropical storm on 20 August. Accordingly, the model forecasts did not precisely capture the storm and its impact. However, four of the five GEFS ensemble members used in this study did have a low-pressure system pass near the bay at some time during the second half of the forecast period (Figure 9c), while the remaining member (Member 5) produced a low to the northeast with a trailing cold front. Some of these events produced cooler surface waters similar to those observed following Irene (Figure 9a), although the timing of the events varied. Consistent with atmospheric uncertainty around the formation of a tropical cyclone as well around the timing of a cooling event earlier in the forecast period, the uncertainty in SSTs was high for the majority of the forecast period: The range of temperatures forecast by the ensemble spanned 3°C for much of the forecast, compared to the roughly 1° to 2° range in Figure 8c. During the first half of the forecast period, the ensemble also displayed some uncertainty about hypoxic volume, although all members predicted volumes above normal, which is consistent with the data-derived estimates from Bever et al. (2013) (Figure 9b). During the second half, the three ensemble members that forecast a low or cold front around the right time (Members 1, 3, and 5) also correctly forecast a significant aeration event and a shift to near- or below-normal hypoxia conditions, while the remaining two members forecast a continuation of severe hypoxia.

Although we only produced estuarine forecasts using the first five GEFS reforecasts initialized on 3 August 2011, we have examined the remaining six GEFS atmospheric reforecast members from 3 August and the GEFS reforecasts initialized on 10, 17, and 24 August. Of the remaining six members from 3 August, none predicted a storm making landfall near the bay like Member 3 in Figure 9c, although two members did produce a low offshore near the end of the forecast period. The reforecasts initialized on 10 August indicated a stronger probability of a cyclone near the bay: 4 of the 11 members produced a strong low over Cape Hatteras sometime between 22 August and 5 September, and several of the remaining members produced weaker lows or storms farther offshore. By 17 August, all but 2 of the 11 ensemble members correctly predicted a tropical or extratropical cyclone in the vicinity of the bay during the second or third week of

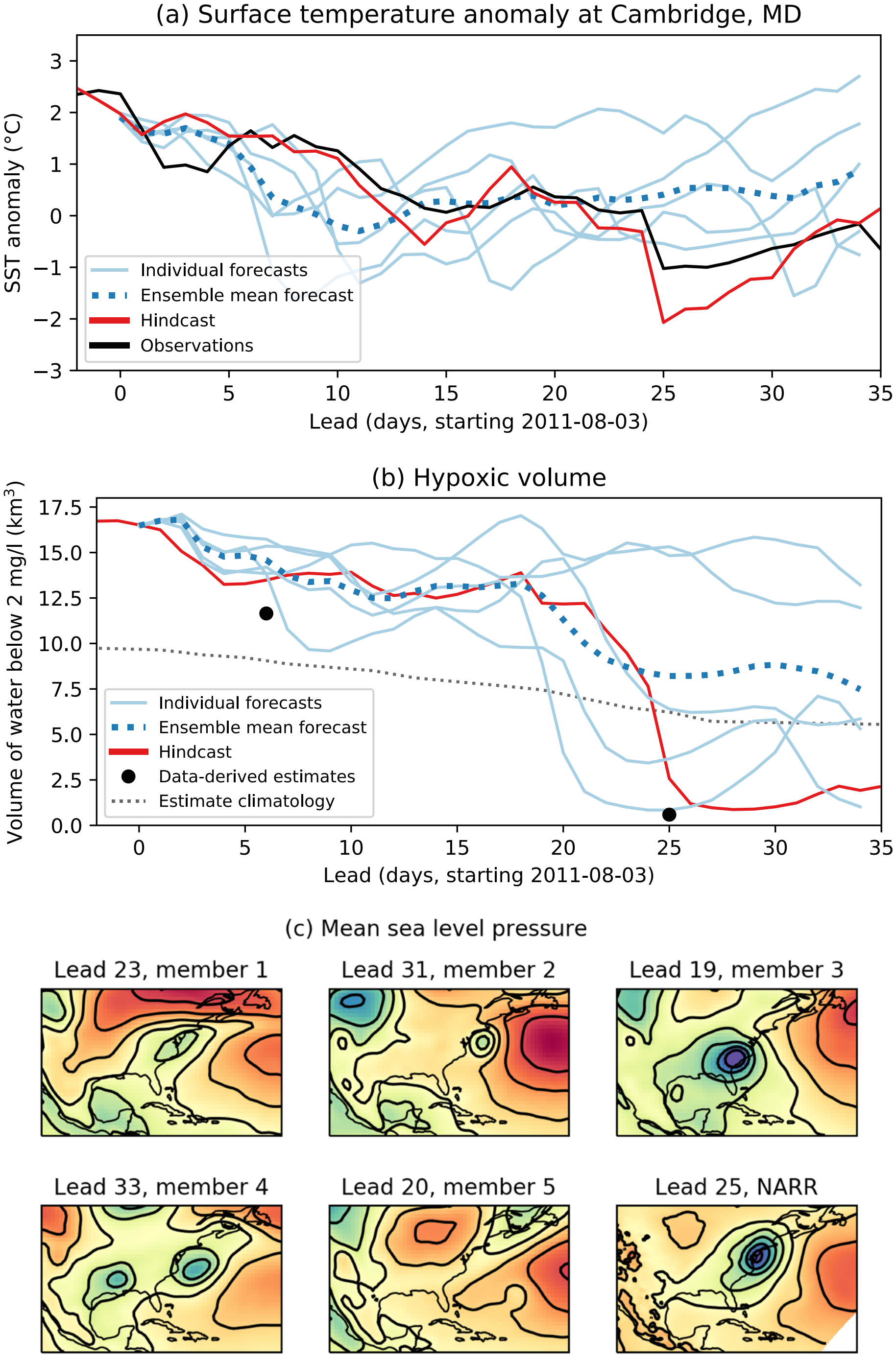

Figure 9. (a, b) Observations or data-derived estimates (black), model hindcast (red), individual ensemble member forecasts (light blue), and ensemble mean forecast (dashed dark blue) for SST and hypoxic volume, respectively. Hypoxic volume estimates are from Bever et al. (2013). (c) Mean sea level pressure forecasts for the five GEFS members (first five panels) and the NARR reanalysis (bottom right panel). All GEFS forecasts were initialized on the same date but are valid at different times. Red colors indicate high MSLP, and blue colors indicate low MSLP.

the forecast, although the exact timing still varied. In the reforecasts initialized 24 August, all 11 ensemble members correctly placed a strong tropical cyclone near the bay on 28 August. These results confirm the extended forecastability of Irene and highlight the importance of using multiple ensemble members to capture uncertainty about the formation, track, and timing of the storm.

Here we presented only a single case study that highlights the potential applications for estuarine forecasting in a situation where skillful tropical cyclone forecasts are possible. Although in general skillful deterministic prediction of the genesis and track of an individual tropical cyclone multiple weeks in advance remains an elusive challenge (Xiang et al., 2015), recent studies have shown that dynamical models do have skill at forecasting the statistics of tropical cyclones at subseasonal to seasonal time scales. For example, Li et al. (2016) showed that the GEFS model has skill at predicting tropical cyclone counts and intensity in the North Atlantic Ocean two weeks in advance. Xiang et al. (2015) demonstrated skill at 5- to 10-day lead predictions of individual tropical cyclones. Lee et al. (2018) found that for most ocean basins, the formation of tropical cyclones could be skillfully forecast in the first week by the majority of the six subseasonal models that they tested; however, forecast skill declined significantly in the second week and beyond. Vecchi et al. (2014) and Murakami et al. (2016) showed that skillful forecasts of regional tropical cyclone activity aggregated over a hurricane season are possible several months prior to the start of the season. However, beyond the single case study presented here, it remains to be seen whether these forecasts translate to skill at forecasting conditions for estuarine and coastal regions.

4.2. Potential Improvements to the Model System

In this study, we presented a proof of concept subseasonal estuarine forecast system for Chesapeake Bay that can skillfully forecast temperature, salinity, and potentially oxygen at least two weeks in advance in the majority of the bay. However, the modeling system included notable simplifications, such as not using forecast river discharge or data-assimilative initial conditions, that reduced the computational costs and model development time but may have also reduced the skill of the forecast system. In this section, we discuss modifications to the forecast system that could improve the forecast skill and make the forecasts more useful to managers and other stakeholders.

Although the hindcast used as initial conditions for the reforecast experiments had reasonable accuracy when compared to the observations, assimilating observations to improve the initial conditions would also improve the forecast skill, especially in the earlier part of the forecast (e.g., Hoffman et al., 2012). However, assimilating data in estuarine and coastal models, where large spatial and temporal variability is present, is more challenging than in global and regional ocean models where data assimilation is more common (Li et al., 2015; Stanev et al., 2016; Xu et al., 2002). Data assimilation is also computationally costly, and it would not be feasible to implement assimilation for the 18 years of simulation used in this study. Data assimilation is also not currently implemented in the operational forecast model for Chesapeake Bay, presumably for the same reasons.

To use the atmospheric forecasts to drive the estuarine model, we applied a simple bias correction and downscaling method that interpolated the forecasts to a higher resolution and corrected the mean bias of the forecasts by comparing with an atmospheric reanalysis. More sophisticated methods could improve the skill of the atmospheric and estuarine forecasts by also correcting other moments of the distribution, accounting for covariance between variables, and separating the differences between land and water and the differences in weather and climate across the bay, both of which are poorly resolved at the 1 degree resolution that the atmospheric forecasts were archived at. However, a more sophisticated downscaling method is not guaranteed to improve the forecasts; for example, some correction methods can produce erroneous results near land/sea boundaries (Lanzante et al., 2018).

Although the estuarine model configuration in the reforecast experiments used climatological river discharge, the hindcast from which the initial conditions were derived used realistic river discharge forcing, and so the reforecast simulations captured the effects of past river discharge anomalies (e.g., the predictable advection and dispersion of a recent freshet) but not future anomalies. Given that the 35-day forecasts are short compared to the bay residence time (which Du & Shen, 2016, estimated to be 180 days on average) and compared to the estimated 40 to 70 days by which bay mean salinity lags total river discharge (Xu et al., 2012), using climatological river discharge forcing for the 35-day forecast experiments is not likely to be a severe limitation in the mainstem bay. This hypothesis is supported by the presence of salinity skill for the majority of the bay at extended leads, except near the sources of river discharge (Figure 6). However, using skillfully forecast river discharge as input would still improve the salinity and stratification forecasts, and forecast discharge would be especially important to consider for seasonal forecasts longer than the 35-day forecasts tested here. Recently developed operational forecast models, such as the National Water Model, could be used in a real-time forecasting system. However, for retrospectively assessing the forecast skill, we are not aware of any river discharge models that have run reforecast simulations that cover the time period of this study. A possible alternative approach would be to use a simple water balance model to estimate runoff and river discharge using forecasts of precipitation and temperature over the Chesapeake Bay watershed (e.g., Muhling et al., 2018), although forecast skill for precipitation is significantly lower than for temperature (DelSole et al., 2017; Pegion et al., 2019) and it is possible that using a simple water balance method would not provide enough skill at forecasting river discharge to significantly increase the skill of the estuarine forecasts.

When comparing against the hindcast, the bottom oxygen forecasts had lower errors than the hindcast climatology or hindcast anomaly persistence. However, when compared against the CBP observations, the bottom oxygen forecasts were not skillful even with the simple, stationary bias correction included (Figure 7). The bottom oxygen in the hindcast was also not skillful (Figure 3). One possible cause of the low skill for oxygen concentration is the simple method used to model oxygen, which parameterizes all biogeochemical sinks of oxygen with a single, constant respiration rate. Utilizing a full biogeochemical model, such as the models in Da et al. (2018) or Irby et al. (2016), may provide improved forecast skill for oxygen by capturing the variability driven by biogeochemical processes. Supporting this theory, oxygen was forecast well during the two case studies (section 4.1) when oxygen variability was driven by extreme weather events, rather than by biogeochemical processes, that were captured by the atmospheric and estuarine models. Alternatively, allowing the respiration rate in the simple oxygen model to vary seasonally or over depth could improve the oxygen predictions (Bever et al., 2013). The significant spatial variability of dissolved oxygen may also complicate the skill assessment; in Figure 7, we assessed skill aggregated over eight discrete points along the deep central region of the bay, but forecast skill compared to the observations could be higher at other locations (e.g., Ross & Stock, 2019). The ChesROMS configuration also uses fairly coarse resolution compared to some other models of Chesapeake Bay (Irby et al., 2016), and the coarse resolution and resulting smooth bathymetry could limit the ability to correctly predict bottom oxygen. However, we tested skill at predicting hypoxic volume, which integrates oxygen concentrations over space, and found that although errors were reduced, the forecasts were still not skillful when compared to the observed climatology of hypoxic volume (not shown). We also found that skill compared to the hindcast was higher along the central channel than in other regions (Figure 6).

Although the hindcast and 35-day forecasts had reasonable accuracy, expanding the southeastern portion of the model domain (along which open boundary conditions were applied) to cover a larger portion of the shelf may improve the accuracy of the forecasts, especially if the forecasts were to be extended to longer lead times. In Chesapeake Bay, winds along the shelf drive a significant amount of subtidal variability in sea level and exchange between the estuary and the shelf (Wang, 1979a, 1979b; Wang & Elliott, 1978; Wong & Valle-Levinson, 2002), and dissolved oxygen is also advected up-estuary from the shelf and mouth in deep water (Li et al., 2015). These processes were not captured in the reforecast simulations because of the limited extent of the regional model domain and the use of fixed nontidal elevation and velocity boundary conditions. On the other hand, simulations with a coupled biogeochemical model by Da et al. (2018) have shown that even in the small ChesROMS domain, dissolved oxygen concentrations in the bay are only weakly sensitive to boundary conditions for dissolved inorganic nitrogen.

Finally, our results primarily examined the skill of the five-member ensemble average forecast and did not compare the accuracy of the distribution of the ensemble forecasts. With appropriate postprocessing, the information provided by the multiple ensemble members could be presented as a probabilistic forecast (Gneiting & Katzfuss, 2014), which is more informative and often more useful (Krzysztofowicz, 2001; Ramos et al., 2013). It would also be worth testing whether an ensemble of five simulations is also sufficient for probabilistic forecasts, or if more members are beneficial.

5. Conclusion

We conducted an extensive set of reforecast experiments for an estuary and found that temperature, salinity, and stratification could be skillfully forecast beyond the limit of weather-scale predictability of the atmosphere and beyond the short forecasts that have been previously developed for estuarine and coastal systems. After roughly the first five days, the skill of the forecasts was significantly improved by taking the ensemble mean of a series of estuarine forecasts produced from an ensemble of atmospheric forecasts. Given that extensive estuarine forecasts appear to be possible, it is worth exploring whether these forecasts can be implemented in management decisions and made informative to other forecast users. Further improvements to the forecast system, such as the production of skillful probabilistic forecasts, may also improve the usefulness of the model forecasts. Finally, although we expect that our results will generalize and apply to forecasts of other estuaries, it would be interesting to compare the predictability and forecast skill across multiple estuaries. Forecast skill may be even higher in some other regions, such as those with stronger atmospheric predictability or stronger forcing from predictable river discharge, and could be higher or lower in other types of estuaries, such as deep fjords, compared to the coastal plain estuary we studied.

Data Availability Statement

Estuarine model data are available online (https://doi.org/10.6084/m9.figshare.9893891.v5). All other data are publicly available from the sources cited in the text.

Acknowledgments
The authors thank Nathaniel Johnson and Jessica Luo for providing helpful reviews of a draft of this manuscript and three reviewers for providing additional helpful comments. This manuscript was prepared by the authors under Award NA18OAR4320123 from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce, and with funding from the Integrated Ecosystem Assessment program. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of the National Oceanic and Atmospheric Administration, or the U.S. Department of Commerce. NOAA/MAPP, ONR, NASA, and NOAA/NWS jointly provided coordinating support and led development of the SubX system. NOS, COMT, NCCOS, and NASA provided funding that supported development of the ChesROMS model.

References

Avila, L. A., & Stewart, S. R. (2013). Atlantic hurricane season of 2011. *Monthly Weather Review*, *141*, 2577–2596. https://doi.org/10.1175/MWR-D-12-00230.1

Baehr, J., Fröhlich, K., Botzet, M., Domeisen, D. I. V., Kornblueh, L., Notz, D., et al. (2015). The prediction of surface temperature in the new seasonal prediction system based on the MPI-ESM coupled climate model. *Climate Dynamics*, *44*(9), 2723–2735. https://doi.org/10.1007/s00382-014-2399-7

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Bever, A. J., Friedrichs, M. A. M., Friedrichs, C. T., & Scully, M. E. (2018). Estimating hypoxic volume in the Chesapeake Bay using two continuously sampled oxygen profiles. *Journal of Geophysical Research: Oceans*, *123*, 6392–6407. https://doi.org/10.1029/2018JC014129

Bever, A. J., Friedrichs, M. A. M., Friedrichs, C. T., Scully, M. E., & Lanerolle, L. W. J. (2013). Combining observations and numerical model results to improve estimates of hypoxic volume within the Chesapeake Bay, USA. *Journal of Geophysical Research: Oceans*, *118*, 4924–4944. https://doi.org/10.1002/jgrc.20331

Brown, C. W., Hood, R. R., Long, W., Jacobs, J., Ramers, D. L., Wazniak, C., et al. (2013). Ecological forecasting in Chesapeake Bay: Using a mechanistic-empirical modeling approach. *Journal of Marine Systems*, *125*, 113–125. https://doi.org/10.1016/j.jmarsys.2012.12.007

Castanedo, S., Medina, R., Losada, I. J., Vidal, C., Méndez, F. J., Osorio, A., et al. (2006). The prestige oil spill in Cantabria (Bay of Biscay). Part I: Operational forecasting system for quick response, risk assessment, and protection of natural resources. *Journal of Coastal Research*, *22*(6), 1474–1489. https://doi.org/10.2112/04-0364.1

Chapman, D. C. (1985). Numerical treatment of cross-shelf open boundaries in a barotropic coastal ocean model. *Journal of Physical Oceanography*, *15*(8), 1060–1075. https://doi.org/10.1175/1520-0485(1985)015<1060:NTOCSO>2.0.CO;2

Chesapeake Bay Program (2018). CBP water quality database (1984–present). https://www.chesapeakebay.net/what/downloads/cbp_water_quality_database_1984_present

Cho, K.-H., Li, Y., Wang, H., Park, K.-S., Choi, J.-Y., Shin, K.-I., & Kwon, J.-I. (2014). Development and validation of an operational search and rescue modeling system for the Yellow Sea and the East and South China Seas. *Journal of Atmospheric and Oceanic Technology*, *31*(1), 197–215. https://doi.org/10.1175/JTECH-D-13-00097.1

Da, F., Friedrichs, M. A. M., & St-Laurent, P. (2018). Impacts of atmospheric nitrogen deposition and coastal nitrogen fluxes on oxygen concentrations in Chesapeake Bay. *Journal of Geophysical Research: Oceans*, *123*, 5004–5025. https://doi.org/10.1029/2018JC014009

DelSole, T., Trenary, L., Tippett, M. K., & Pegion, K. (2017). Predictability of Week-3–4 average temperature and precipitation over the contiguous United States. *Journal of Climate*, *30*(10), 3499–3512. https://doi.org/10.1175/JCLI-D-16-0567.1

Du, J., & Shen, J. (2016). Water residence time in Chesapeake Bay for 1980–2012. *Journal of Marine Systems*, *164*, 101–111. https://doi.org/10.1016/j.jmarsys.2016.08.011

Feng, Y., Friedrichs, M. A., Wilkin, J., Tian, H., Yang, Q., Hofmann, E. E., et al. (2015). Chesapeake Bay nitrogen fluxes derived from a Land-Estuarine ocean biogeochemical modeling system: Model description, evaluation, and nitrogen budgets. *Journal of Geophysical Research: Biogeosciences*, *120*, 1666–1695. https://doi.org/10.1002/2015JG002931

Flather, R. A. (1976). A tidal model of the North-West European continental shelf. *Mémoires de la Société Royale des Sciences de Liége*, *10*, 141–164.

Glibert, P. M., Magnien, R., Lomas, M. W., Alexander, J., Fan, C., Haramoto, E., et al. (2001). Harmful algal blooms in the Chesapeake and Coastal Bays of Maryland, USA: Comparison of 1997, 1998, and 1999 events. *Estuaries*, *24*(6), 875. https://doi.org/10.2307/1353178

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*(1), 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Hagy, J. D., Boynton, W. R., Keefe, C. W., & Wood, K. V. (2004). Hypoxia in Chesapeake Bay, 1950–2001: Long-term change in relation to nutrient loading and river flow. *Estuaries*, *4*(4), 634–658. https://doi.org/10.1007/BF02907650

Hamill, T. M., Whitaker, J. S., & Mullen, S. L. (2006). Reforecasts: An important dataset for improving weather predictions. *Bulletin of the American Meteorological Society*, *87*(1), 33–46. https://doi.org/10.1175/BAMS-87-1-33

Hermanson, L., Ren, H.-L., Vellinga, M., Dunstone, N. D., Hyder, P., Ineson, S., et al. (2018). Different types of drifts in two seasonal forecast systems and their dependence on ENSO. *Climate Dynamics*, *51*(4), 1411–1426. https://doi.org/10.1007/s00382-017-3962-9

Hervieux, G., Alexander, M. A., Stock, C. A., Jacox, M. G., Pegion, K., Becker, E., et al. (2017). More reliable coastal SST forecasts from the North American multimodel ensemble. *Climate Dynamics*, *53*, 7153–7168. https://doi.org/10.1007/s00382-017-3652-7

Hobday, A. J., Spillman, C. M., Paige Eveson, J., & Hartog, J. R. (2016). Seasonal forecasting for decision support in marine fisheries and aquaculture. *Fisheries Oceanography*, *25*, 45–56. https://doi.org/10.1111/fog.12083

Hoffman, M. J., Miyoshi, T., Haine, T. W. N., Ide, K., Brown, C. W., & Murtugudde, R. (2012). An advanced data assimilation system for the Chesapeake Bay: Performance evaluation. *Journal of Atmospheric and Oceanic Technology*, *29*(10), 1542–1557. https://doi.org/10.1175/JTECH-D-11-00126.1

Irby, I. D., & Friedrichs, M. A. M. (2019). Evaluating confidence in the impact of regulatory nutrient reduction on Chesapeake Bay water quality. *Estuaries and Coasts*, *42*(1), 16–32. https://doi.org/10.1007/s12237-018-0440-5

Irby, I., Friedrichs, M. A., Da, F., & Hinson, K. (2018). The competing impacts of climate change and nutrient reductions on dissolved oxygen in Chesapeake Bay. *Biogeosciences*, *15*, 2649–2668. https://doi.org/10.5194/bg-15-2649-2018

Irby, I. D., Friedrichs, M. A., Friedrichs, C. T., Bever, A. J., Hood, R. R., Lanerolle, L. W., et al. (2016). Challenges associated with modeling low-oxygen waters in Chesapeake Bay: A multiple model comparison. *Biogeosciences*, *13*(7), 2011–2028. https://doi.org/10.5194/bg-13-2011-2016

Jacobs, J. M., Rhodes, M., Brown, C. W., Hood, R. R., Leight, A., Long, W., & Wood, R. (2014). Modeling and forecasting the distribution of *Vibrio vulnificus* in Chesapeake Bay. *Journal of Applied Microbiology*, *117*(5), 1312–1327. https://doi.org/10.1111/jam.12624

Jacox, M. G., Alexander, M. A., Stock, C. A., & Hervieux, G. (2017). On the skill of seasonal sea surface temperature forecasts in the California Current System and its connection to ENSO variability. *Climate Dynamics*, *53*, 7519–7533. https://doi.org/10.1007/s00382-017-3608-y

Jia, L., Yang, X., Vecchi, G. A., Gudgel, R. G., Delworth, T. L., Rosati, A., et al. (2015). Improved seasonal prediction of temperature and precipitation over land in a high-resolution GFDL climate model. *Journal of Climate*, *28*(5), 2044–2062. https://doi.org/10.1175/JCLI-D-14-00112.1

Kaper, J. B., Remmers, E. F., Lockman, H., & Colwell, R. R. (1981). Distribution of *Vibrio parahaemolyticus* in Chesapeake Bay during the summer season. *Estuaries*, *4*(4), 321–327. https://doi.org/10.2307/1352156

Koster, R. D., Mahanama, S. P. P., Yamada, T. J., Balsamo, G., Berg, A. A., Boisserie, M., et al. (2010). Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. *Geophysical Research Letters*, *37*, L02402. https://doi.org/10.1029/2009GL041677

Koster, R. D., Mahanama, S. P. P., Yamada, T. J., Balsamo, G., Berg, A. A., Boisserie, M., et al. (2011). The second phase of the global land–atmosphere coupling experiment: Soil moisture contributions to Subseasonal Forecast Skill. *Journal of Hydrometeorology*, *12*(5), 805–822. https://doi.org/10.1175/2011JHM1365.1

Krzysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, *249*(1–4), 2–9. https://doi.org/10.1016/s0022-1694(01)00420-6

Lanerolle, L. W. J., Patchen, R. C., & Aikman, F. I. I. I. (2011). The second generation Chesapeake Bay operational forecast system (CBOFS2): Model development and skill assessment (*NOAA Technical Report NOS CS 29*).

Lanzante, J. R., Dixon, K. W., Nath, M. J., Whitlock, C. E., & Adams-Smith, D. (2018). Some pitfalls in statistical downscaling of future climate. *Bulletin of the American Meteorological Society*, *99*(4), 791–803. https://doi.org/10.1175/BAMS-D-17-0046.1

Lavaysse, C., Naumann, G., Alfieri, L., Salamon, P., & Vogt, J. (2019). Predictability of the European heat and cold waves. *Climate Dynamics*, *52*(3–4), 2481–2495. https://doi.org/10.1007/s00382-018-4273-5

Lee, C.-Y., Camargo, S. J., Vitart, F., Sobel, A. H., & Tippett, M. K. (2018). Subseasonal tropical cyclone genesis prediction and MJO in the S2S dataset. *Weather and Forecasting*, *33*(4), 967–988. https://doi.org/10.1175/WAF-D-17-0165.1

Lee, Y. J., & Lwiza, K. M. M. (2008). Factors driving bottom salinity variability in the Chesapeake Bay. *Continental Shelf Research*, *28*(10–11), 1352–1362. https://doi.org/10.1016/j.csr.2008.03.016

Li, Y., & Li, M. (2011). Effects of winds on stratification and circulation in a partially mixed estuary. *Journal of Geophysical Research*, *116*, C12012. https://doi.org/10.1029/2010JC006893

Li, Y., Li, M., & Kemp, W. M. (2015). A budget analysis of bottom-water dissolved oxygen in Chesapeake Bay. *Estuaries and Coasts*, *38*, 2132–2148. https://doi.org/10.1007/s12237-014-9928-9

Li, Z., McWilliams, J. C., Ide, K., & Farrara, J. D. (2015). Coastal ocean data assimilation using a multi-scale three-dimensional variational scheme. *Ocean Dynamics*, *65*(7), 1001–1015. https://doi.org/10.1007/s10236-015-0850-x

Li, S., & Robertson, A. W. (2015). Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Monthly Weather Review*, *143*(7), 2871–2889. https://doi.org/10.1175/MWR-D-14-00277.1

Li, W., Wang, Z., & Peng, M. S. (2016). Evaluating tropical cyclone forecasts from the NCEP Global Ensemble Forecasting System (GEFS) reforecast version 2. *Weather and Forecasting*, *31*(3), 895–916. https://doi.org/10.1175/WAF-D-15-0176.1

MacCready, P. (1999). Estuarine adjustment to changes in river flow and tidal mixing. *Journal of Physical Oceanography*, *29*(4), 708–726. https://doi.org/10.1175/1520-0485(1999)029<0708:EATCIR>2.0.CO;2

MacCready, P. (2007). Estuarine adjustment. *Journal of Physical Oceanography*, *37*(8), 2133–2145. https://doi.org/10.1175/JPO3082.1

MacLachlan, C., Arribas, A., Peterson, K. A., Maidens, A., Fereday, D., Scaife, A. A., et al. (2015). Global seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, *141*(689), 1072–1084. https://doi.org/10.1002/qj.2396

McKinnon, K. A., Rhines, A., Tingley, M. P., & Huybers, P. (2016). Long-Lead predictions of eastern United States hot days from Pacific sea surface temperatures. *Nature Geoscience*, *9*, 389–394. https://doi.org/10.1038/ngeo2687

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., et al. (2006). North American regional reanalysis. *Bulletin of the American Meteorological Society*, *87*(3), 343–360. https://doi.org/10.1175/BAMS-87-3-343

Monismith, S. G., Kimmerer, W., Burau, J. R., & Stacey, M. T. (2002). Structure and flow-induced variability of the subtidal salinity field in Northern San Francisco Bay. *Journal of Physical Oceanography*, *32*(11), 3003–3019. https://doi.org/10.1175/1520-0485(2002)032<3003:SAFIVO>2.0.CO;2

Muhling, B. A., Gaitán, C. F., Stock, C. A., Saba, V. S., Tommasi, D., & Dixon, K. W. (2018). Potential salinity and temperature futures for the Chesapeake Bay using a statistical downscaling spatial disaggregation framework. *Estuaries and Coasts*, *41*, 349–372. https://doi.org/10.1007/s12237-017-0280-8

Mukai, A. Y., Westerink, J. J., Luettich, R. A., & Mark, D. (2002). Eastcoast 2001, a tidal constituent database for western North Atlantic, Gulf of Mexico, and Caribbean Sea (*ERDC/CHL-TR-02-24*). Vicksburg: US Army Corps of Engineers.

Mulholland, M. R., Morse, R. E., Boneillo, G. E., Bernhardt, P. W., Filippino, K. C., Procise, L. A., et al. (2009). Understanding causes and impacts of the dinoflagellate, *Cochlodinium polykrikoides*, blooms in the Chesapeake Bay. *Estuaries and Coasts*, *32*(4), 734–747. https://doi.org/10.1007/s12237-009-9169-5

Munroe, D., Tabatabai, A., Burt, I., Bushek, D., Powell, E. N., & Wilkin, J. (2013). Oyster mortality in Delaware Bay: Impacts and recovery from Hurricane Irene and Tropical Storm Lee. *Estuarine, Coastal and Shelf Science*, *135*, 209–219. https://doi.org/10.1016/j.ecss.2013.10.011

Murakami, H., Vecchi, G. A., Villarini, G., Delworth, T. L., Gudgel, R., Underwood, S., et al. (2016). Seasonal forecasts of major hurricanes and landfalling tropical cyclones using a high-resolution GFDL coupled climate model. *Journal of Climate*, *29*(22), 7977–7989. https://doi.org/10.1175/JCLI-D-16-0233.1

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, *116*(12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Murphy, A. H. (1992). Climatology, persistence, and their linear combination as standards of reference in skill scores. *Weather and Forecasting*, *7*(4), 692–698.

Murphy, A. H., & Epstein, E. S. (1989). Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, *117*(3), 572–582. https://doi.org/10.1175/1520-0493(1989)117<0572:SSACCI>2.0.CO;2

Murphy, R. R., Kemp, W. M., & Ball, W. P. (2011). Long-term trends in Chesapeake Bay seasonal hypoxia, stratification, and nutrient loading. *Estuaries and Coasts*, *34*, 1293–1309. https://doi.org/10.1007/s12237-011-9413-7

Officer, C. B., Biggs, R. B., Taft, J. L., Cronin, L. E., Tyler, M. A., & Boynton, W. R. (1984). Chesapeake Bay Anoxia: Origin, development, and significance. *Science*, *223*(4631), 22–27. https://doi.org/10.1126/science.223.4631.22

Palinkas, C. M., Halka, J. P., Li, M., Sanford, L. P., & Cheng, P. (2014). Sediment deposition from tropical storms in the upper Chesapeake Bay: Field observations and model simulations. *Continental Shelf Research*, *86*, 6–16. https://doi.org/10.1016/j.csr.2013.09.012

Pegion, K., Kirtman, B. P., Becker, E., Collins, D. C., LaJoie, E., Burgman, R., et al. (2019). The subseasonal experiment (SubX): A multi-model subseasonal prediction experiment. *Bulletin of the American Meteorological Society*, *100*, 2043–2060. https://doi.org/10.1175/BAMS-D-18-0270.1

Peng, M., Schmalz Jr, R. A., Zhang, A., & Aikman III, F. (2014). Towards the development of the National Ocean Service San Francisco Bay operational forecast system. *Journal of Marine Science and Engineering*, *2*(1), 247–286. https://doi.org/10.3390/jmse2010247

Ramos, M. H., Van Andel, S. J., & Pappenberger, F. (2013). Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, *17*(6), 2219–2232. https://doi.org/10.5194/hess-17-2219-2013

Ross, A. C., & Stock, C. A. (2019). An assessment of the predictability of column minimum dissolved oxygen concentrations in Chesapeake Bay using a machine learning model. *Estuarine, Coastal and Shelf Science*, *221*, 53–65. https://doi.org/10.1016/j.ecss.2019.03.007

Scully, M. E. (2010). Wind modulation of dissolved oxygen in Chesapeake Bay. *Estuaries and Coasts*, *33*(5), 1164–1175. https://doi.org/10.1007/s12237-010-9319-9

Scully, M. E. (2013). Physical controls on hypoxia in Chesapeake Bay: A numerical modeling study. *Journal of Geophysical Research: Oceans*, *118*, 1239–1256. https://doi.org/10.1002/jgrc.20138

Scully, M. E. (2016). The contribution of physical processes to inter-annual variations of hypoxia in Chesapeake Bay: A 30-Yr modeling study. *Limnology and Oceanography*, *61*(6), 2243–2260. https://doi.org/10.1002/lno.10372

Siedlecki, S. A., Kaplan, I. C., Hermann, A. J., Nguyen, T. T., Bond, N. A., Newton, J. A., et al. (2016). Experiments with seasonal forecasts of ocean conditions for the northern region of the California current upwelling system. *Scientific Reports*, *6*, 27203. https://doi.org/10.1038/srep27203

Spillman, C. M., & Alves, O. (2009). Dynamical seasonal prediction of summer sea surface temperatures in the Great Barrier Reef. *Coral Reefs*, *28*(1), 197–206. https://doi.org/10.1007/s00338-008-0438-8

Stanev, E. V., Schulz-Stellenfleth, J., Staneva, J., Grayek, S., Grashorn, S., Behrens, A., et al. (2016). Ocean forecasting for the German Bight: From regional to coastal scales. *Ocean Science*, *12*(5), 1105–1136. https://doi.org/10.5194/os-12-1105-2016

Stock, C. A., Pegion, K., Vecchi, G. A., Alexander, M. A., Tommasi, D., Bond, N. A., et al. (2015). Seasonal sea surface temperature anomaly prediction for coastal ecosystems. *Progress in Oceanography*, *137*, 219–236. https://doi.org/10.1016/j.pocean.2015.06.007

Stockdale, T. N. (1997). Coupled ocean–Atmosphere forecasts in the presence of climate drift. *Monthly Weather Review*, *125*(5), 809–818. https://doi.org/10.1175/1520-0493(1997)125<0809:COAFIT>2.0.CO;2

Taft, J. L., Taylor, W. R., Hartwig, E. O., & Loftus, R. (1980). Seasonal oxygen depletion in Chesapeake Bay. *Estuaries*, *3*(4), 242. https://doi.org/10.2307/1352079

Teng, H., Branstator, G., Wang, H., Meehl, G. A., & Washington, W. M. (2013). Probability of US heat waves affected by a subseasonal planetary wave pattern. *Nature Geoscience*, *6*(12), 1056–1061. https://doi.org/10.1038/ngeo1988

Testa, J. M., Clark, J. B., Dennison, W. C., Donovan, E. C., Fisher, A. W., Ni, W., et al. (2017). Ecological forecasting and the science of hypoxia in Chesapeake Bay. *BioScience*, *67*(7), 614–626. https://doi.org/10.1093/biosci/bix048

Tian, H., Yang, Q., Najjar, R. G., Ren, W., Friedrichs, M. A. M., Hopkinson, C. S., & Pan, S. (2015). Anthropogenic and climatic influences on carbon fluxes from eastern North America to the Atlantic Ocean: A process-based modeling study. *Journal of Geophysical Research: Biogeosciences*, *120*, 757–772. https://doi.org/10.1002/2014JG002760

Vecchi, G. A., Delworth, T., Gudgel, R., Kapnick, S., Rosati, A., Wittenberg, A. T., et al. (2014). On the seasonal forecasting of regional tropical cyclone activity. *Journal of Climate*, *27*(21), 7994–8016. https://doi.org/10.1175/JCLI-D-14-00158.1

Vitart, F. (2014). Evolution of ECMWF sub-seasonal forecast skill scores: Evolution of the ECMWF sub-seasonal forecast skill. *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 1889–1899. https://doi.org/10.1002/qj.2256

Vitart, F., Cunningham, C., DeFlorio, M., Dutra, E., Ferranti, L., Golding, B., et al. (2019). Sub-seasonal to seasonal prediction of weather extremes. In *Sub-seasonal to seasonal prediction* (pp. 365–386). Elsevier. https://doi.org/10.1016/B978-0-12-811714-9.00017-6

Wang, D.-P. (1979a). Subtidal sea level variations in the Chesapeake Bay and relations to atmospheric forcing. *Journal of Physical Oceanography*, *9*(2), 413–421. https://doi.org/10.1175/1520-0485(1979)009<0413:SSLVIT>2.0.CO;2

Wang, D.-P. (1979b). Wind-driven circulation in the Chesapeake Bay, winter, 1975. *Journal of Physical Oceanography*, *9*(3), 564–572. https://doi.org/10.1175/1520-0485(1979)009<0564:WDCITC>2.0.CO;2

Wang, D.-P., & Elliott, A. J. (1978). Non-tidal variability in the Chesapeake Bay and Potomac River: Evidence for non-local forcing. *Journal of Physical Oceanography*, *8*(2), 225–232. https://doi.org/10.1175/1520-0485(1978)008<0225:NTVITC>2.0.CO;2

Wei, E., Yang, Z., Chen, Y., Kelley, J. G. W., & Zhang, A. (2014). The northern Gulf of Mexico Operational Forecast System (NGOFS): Model development and skill assessment (*NOAA Technical Report NOS CS 33*).

Wong, K.-C., & Valle-Levinson, A. (2002). On the relative importance of the remote and local wind effects on the subtidal exchange at the entrance to the Chesapeake Bay. *Journal of Marine Research*, *60*, 477–498. https://doi.org/10.1357/002224002762231188

Xiang, B., Lin, S.-J., Zhao, M., Zhang, S., Vecchi, G., Li, T., et al. (2015). Beyond weather time-scale prediction for Hurricane Sandy and Super Typhoon Haiyan in a global climate model. *Monthly Weather Review*, *143*(2), 524–535. https://doi.org/10.1175/MWR-D-14-00227.1

Xie, X., Li, M., & Ni, W. (2018). Roles of wind-driven currents and surface waves in sediment resuspension and transport during a tropical storm. *Journal of Geophysical Research: Oceans*, *123*, 8638–8654. https://doi.org/10.1029/2018JC014104

Xu, J., Chao, S.-Y., Hood, R. R., & Wang, H. V. (2002). Assimilating high-resolution salinity data into a model of a partially mixed estuary. *Journal of Geophysical Research*, *107*(C7), 3074. https://doi.org/10.1029/2000JC000626

Xu, J., Long, W., Wiggert, J. D., Lanerolle, L. W. J., Brown, C. W., Murtugudde, R., & Hood, R. R. (2012). Climate forcing and salinity variability in Chesapeake Bay, USA. *Estuaries and Coasts*, *35*(1), 237–261. https://doi.org/10.1007/s12237-011-9423-5

Yang, Q., Tian, H., Friedrichs, M. A. M., Hopkinson, C. S., Lu, C., & Najjar, R. G. (2015). Increased nitrogen export from eastern North America to the Atlantic Ocean due to climatic and anthropogenic changes during 1901-2008. *Journal of Geophysical Research: Biogeosciences*, *120*, 1046–1068. https://doi.org/10.1002/2014JG002763

Yang, Q., Tian, H., Friedrichs, M. M., Liu, M., Li, X., & Yang, J. (2015). Hydrological responses to climate and land-use changes along the North American East Coast: A 110-year historical reconstruction. *Journal of the American Water Resources Association*, *51*(1), 47–67. https://doi.org/10.1111/jawr.12232

Zhang, C. (2013). Madden–Julian Oscillation: Bridging weather and climate. *Bulletin of the American Meteorological Society*, *94*(12), 1849–1870. https://doi.org/10.1175/BAMS-D-12-00026.1

Zhang, A., Hess, K. W., & Aikman, F. (2010). User-Based skill assessment techniques for operational hydrodynamic forecast systems. *Journal of Operational Oceanography*, *3*(2), 11–24. https://doi.org/10.1080/1755876X.2010.11020114

Zhang, F., Sun, Y. Q., Magnusson, L., Buizza, R., Lin, S.-J., Chen, J.-H., & Emanuel, K. (2019). What is the predictability limit of midlatitude weather? *Journal of the Atmospheric Sciences*, *76*(4), 1077–1091. https://doi.org/10.1175/JAS-D-18-0269.1

Zhou, X., Zhu, Y., Hou, D., & Kleist, D. (2016). A comparison of perturbations from an ensemble transform and an ensemble Kalman filter for the NCEP global ensemble forecast system. *Weather and Forecasting*, *31*(6), 2057–2074. https://doi.org/10.1175/WAF-D-16-0109.1

Zhou, X., Zhu, Y., Hou, D., Luo, Y., Peng, J., & Wobus, R. (2017). Performance of the new NCEP Global Ensemble Forecast System in a Parallel Experiment. *Weather and Forecasting*, *32*(5), 1989–2004. https://doi.org/10.1175/WAF-D-17-0023.1

Zhu, Y., Zhou, X., Li, W., Hou, D., Melhauser, C., Sinsky, E., et al. (2018). Toward the improvement of subseasonal prediction in the National Centers for Environmental Prediction Global Ensemble Forecast System. *Journal of Geophysical Research: Atmospheres*, *123*, 6732–6745. https://doi.org/10.1029/2018JD028506