MR. HAIKUN XU (Orcid ID : 0000-0002-1269-6012)

Article type : Original Article

Evaluating the utility of the Gulf Stream Index for predicting recruitment of Southern New England-Mid Atlantic yellowtail flounder

HAIKUN XU[1a*], TIMOTHY J. MILLER[2], SULTAN HAMEED[1], LARRY A. ALADE[2], JANET A. NYE[1]

[1] *School of Marine and Atmospheric Science, Stony Brook University, Stony Brook, New York, USA*

[2] *NOAA Northeast Fisheries Science Center, Woods Hole, Massachusetts, USA*

[a] *Current address: School of Aquatic and Fishery Sciences, University of Washington, Seattle, Washington, USA*

Revision submitted to *Fisheries Oceanography*

* *Corresponding author address*:

Haikun Xu, School of Aquatic and Fishery Sciences,

University of Washington, Seattle, WA 98195

E-Mail: haikunxu@uw.edu

Running title: Gulf Stream influences yellowtail flounder recruitment

Abstract

The justification for incorporating environmental effects into fisheries stock assessment models has been investigated and debated for a long time. Recently, a state-space age-structured assessment model which includes the stochastic change in the environmental covariate over time and its effect on recruitment was developed for Southern New England-Mid Atlantic yellowtail flounder (*Limanda ferruginea*). In this paper, we first investigated the correlations of environmental covariates with Southern New England-Mid Atlantic yellowtail flounder recruitment deviations. The covariate that was most strongly correlated with the recruitment deviations was then incorporated into the state-space model and alternative effects on the stock-recruit relationship were estimated and compared. For the model that performed best as measured by Akaike information criterion, we also compared the estimates and predictions of various population attributes and biological reference points with those from an otherwise identical model without the environmental covariate in the stock-recruit function. We found that the estimates of population parameters are similar for the two models but the predictions differed substantially. To evaluate which model provided more reliable predictions, we quantitatively compared the prediction skill of the two models by generating two series of retrospective predictions. Comparison of the retrospective prediction pattern suggested that from an average point of view, the environmentally-explicit model can provide more accurate near-term recruitment predictions especially the one-year ahead recruitment prediction. While for a specific near-term recruitment prediction from the environmentally-explicit model, the accuracy of which is largely determined by the accuracy of the corresponding environment prediction the model provides.

Key Words: state-space model; Gulf Stream Index; recruitment; stock assessment; Southern New England-Mid Atlantic yellowtail flounder; retrospective prediction; prediction skill

INTRODUCTION

Whether we should incorporate environmental drivers explicitly into fisheries stock assessment models has been investigated and debated for a long time (Walters & Collie, 1988, Haltuch & Punt, 2011, Punt et al., 2014, Szuwalski & Hollowed, 2016). Several recent studies have shown that the environment cannot be ignored to better understand some of the factors influencing changes in stock productivity of many fish populations (Vert-pre et al., 2013, Essington et al., 2015, Szuwalski et al., 2015), but incorporating environmental drivers into stock assessment models remains elusive. However, Miller et al. (2016) recently developed a state-space age-structured assessment model that allows for environmental covariates in the stock-recruit function. According to model fit and retrospective pattern, they concluded that incorporating the effect of Mid-Atlantic cold pool dynamics on Southern New England-Mid Atlantic (SNEMA) yellowtail flounder recruitment can improve model performance.

The physical environment in the SNEMA region is highly dynamic owing to variability in both atmospheric and oceanographic processes. The North Atlantic Oscillation or NAO is an atmospheric process known to have a profound effect on water temperature, storm tracks and northern North Atlantic ecosystems (Drinkwater et al., 2003). The NAO has a lagged effect on surface and bottom water temperature in the Northeast US as the signal propagates from the Labrador Sea (Mountain, 2012, Xu et al., 2015). In the Northeast US, two current systems collide where cold water emanating from the Labrador Current to the north and warm water moving from the south and east in the Gulf Stream meet (Greene et al., 2013). The position of the north wall of the Gulf Stream is the best leading indicator of the relative strength of cold Labrador slope water and warm subtropical water and is highly correlated with temperature on the shelf (Nye et al. 2011). In the SNEMA region, the Mid-Atlantic cold pool is a distinct remnant cold winter water at depth occurring from late spring to early fall, formed as a result of the strong seasonal thermocline in the SNEMA region (Houghton et al., 1982).

Determining the cause of the low recruitment since the 1990’s was argued to be one of the main sources of uncertainty in the most recent SNEMA yellowtail flounder benchmark assessment (NEFSC, 2012). The persistent low recruitment since the 1990’s resulted in the exploration of two recruitment scenarios in the benchmark assessment to account for the notable drop in stock productivity. The first scenario assumed that unfavorable environmental conditions reduced stock productivity significantly since the 1990's such that the stock was considered rebuilt (albeit at a low level) and not overfished. By contrast, the second scenario also accounted for greater historical recruitments prior to the 1990's such that the stock was considered overfished. Therefore, making clear what processes are responsible for the recruitment drop since the 1990's will be invaluable to improving current understanding of the population dynamics and determining the stock status of SNEMA yellowtail flounder.

Recruitment of SNEMA yellowtail flounder may be dependent on temperature condition during the early life stages. SNEMA yellowtail flounder usually spawn in spring and early summer, with a peak in May (NEFSC, 2012). Their fertilized eggs float at the surface for about 2 months, then larvae metamorphosis occurs and juveniles settle to the bottom of the continental shelf (Sullivan et al., 2000). Both field observations (Sullivan et al., 2005, Sullivan et al., 2000) and modeling studies (Miller et al., 2016) have shown that recruitment of SNEMA yellowtail flounder is closely related to the dynamics of the Mid-Atlantic cold pool. In the field, Sullivan et al. (2000) found that the SNEMA stock heavily relies on the cold bottom water in the Mid-Atlantic cold pool as a thermal refuge in summer when water temperature reaches the annual maximum. Later on, Sullivan et al. (2005) also found that stronger young-of-the-year cohorts were observed with colder and longer-lasting Mid-Atlantic cold pools.

Based on the survey evidence from the field, incorporating the Mid-Atlantic cold pool dynamics in SNEMA yellowtail flounder stock assessment model was investigated in the last benchmark assessment, attempting to explain the low productivity level since the 1990's (NEFSC, 2012). The Cold Pool Index (CPI), defined as the first principle component of the Mid-Atlantic cold pool temperature and area matrix, was chosen in the study to represent the thermal condition in the Mid-Atlantic cold pool. A negative correlation was found between the CPI and the recruitment deviations from the Beverton-Holt stock-recruit function. Also,the CPI-incorporated Beverton-Holt stock-recruit function was found to fit data better than the traditional Beverton-Holt stock-recruit function without any environmental covariate. Although this preliminary analysis demonstrated the negative effect of cold pool temperature on SNEMA yellowtail flounder recruitment, the CPI was not accepted in the baseline run in the last benchmark assessment as the low productivity level since 1990 could not be fully explained by the CPI alone (NEFSC, 2012).

After the last benchmark assessment, the effect of CPI on SNEMA yellowtail flounder recruitment was further investigated in a new state-space age-structured assessment model (Miller et al., 2016). State-space models have the advantage of separately modeling time-varying stochastic processes and observation errors, and have recently become increasingly popular due to the developments of software packages that can efficiently handle such models (Nielsen & Berg, 2014). This state-space assessment model allows CPI effects on recruitment, and assumes stochastic changes of the CPI over time and accounts for errors in the annual CPI observations (Miller et al., 2016). Comparison of the state-space models with and without CPI effects on recruitment indicated that the former had lower AIC and provided less retrospective patterns (Mohn, 1999) in terminal year estimates of population attributes. This study further emphasized the importance of the environment in modulating SNEMA yellowtail flounder recruitment.

In addition to understanding stock productivity and determining stock status, another goal in fisheries stock assessment is to predict stock biomass trajectories under various harvest scenarios (Quinn & Deriso, 1999, Haddon, 2010). Prediction skill is a term popularly used in climate science referring to the ability of a model in predicting climate variables (Boer et al., 2013). It is usually assessed by generating a series of historical climate predictions and comparing them with the corresponding observations (Meehl et al., 2009). Although a good prediction skill in historical predictions does not necessarily guarantee a good prediction skill for the future, the historical prediction skill can inform us about the uncertainty in model predictions for the future. In fisheries stock assessment, retrospective analysis is often done to evaluate the systematic bias in population estimates in the terminal year when additional years of data are added (Mohn, 1999). Borrowing the idea of model prediction skill from climate science, a series of retrospective prediction can also be generated in a similar way for fisheries stock assessment models to evaluate the skill of models in predicting population attributes. Indeed, Brooks and Legault (2015) recently have used the idea of the retrospective prediction to evaluate the predictive performance of New England groundfish stock assessment models, although their retrospective prediction scheme is different from that typically used in climate science.

The first objective of this paper was to examine the correlation of various atmospheric and oceanographic covariates with SNEMA yellowtail flounder recruitment deviations. Until now, the examples of incorporating environmental effects directly into fisheries stock assessment and management are still very limited (but see Schirripa, 2007, Hill et al., 2011, and Miller et al., 2016). Thus, our second objective was to comprehensively compare the estimates and predictions from the state-space assessment models with and without the most strongly correlated climate process in the stock-recruit function. This comparison provided suggestions for future fisheries studies that incorporate environmental effects into stock assessment models.

DATA AND METHODS

The correlation of various large-scale atmospheric and oceanographic climate indices with annual deviations in recruitment for the SNEMA yellowtail flounder stock were examined. Hydrology and ecosystem dynamics on the Northeast US Continental Shelf have been known to be affected by the NAO - the dominant and most influential atmospheric oscillation mode in the North Atlantic (Drinkwater et al., 2003, Mountain, 2012). The NAO index represents the scaled pressure difference between the two pressure centers of the NAO, namely the Azores high (AH) pressure center and the Icelandic low (IL) pressure center (Hameed & Piontkovski, 2004). The large-scale atmospheric indices investigated in this study include the winter NAO index from (http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml) as well as the pressures, latitudes, and longitudes of the IL pressure and AH pressure centers from Hameed and Piontkovski (2004). In this study, the reason for including various indices related to the two NAO pressure centers instead of only considering the NAO index is that treating the two pressure centers as two separate systems can potentially explain a larger portion of the NAO-induced variance (Hameed & Piontkovski, 2004). The oceanographic indices investigated in this study include two Gulf Stream related indices (Joyce & Zhang, 2010, Taylor & Stephens, 1998) and the previously investigated CPI. The two Gulf Stream related indices are 1) the Gulf Stream Index (GSI), calculated by using water temperature at 200 m depth (Joyce & Zhang, 2010); and 2) the Gulf Stream North Wall (GSNW), calculated by using sea surface temperature observation (Taylor & Stephens, 1998). The two indices differ in the data source but both quantify the latitudinal position of the Gulf Stream: one at the surface and one at 200 m depth. Particularly, the GSI was shown to be a good indicator of bottom temperature condition in the SNEMA region (Nye et al., 2011).

Assuming that the recruitment deviations from fitting to the Beverton-Holt stock-recruit function are at least partially related to environmental processes, the Pearson cross-correlations between the recruitment deviations in log-space and environmental indices that lead recruitment by zero to two years were calculated. The lead time was designed to account for the delayed effects of some large-scale climate processes on the local environment in the SNEMA region. The recruitment and spawning stock biomass (SSB) time series used to fit to the Beverton-Holt stock-recruit function were extracted from the baseline run in the most recent benchmark assessment (NEFSC, 2012). Recognizing that a significant correlation does not necessarily indicate causation (Hilborn, 2016) and the time series used in the correlation analysis are from stock assessment models which are subject to various sources of uncertainty and bias (Brooks & Deroba, 2015), we also incorporated the most significantly correlated environmental indices internally in the state-space assessment model to compare model performance with respect to AIC and retrospective bias. Following the method in Burnham and Anderson (2002), the Akaike weight was also calculated for each model using AIC.

As in Miller et al. (2016), the environmental covariate (x) at time t is modeled as a random walk for t > 2:

$$x_t | x_{t-1} \sim \mathrm{N}\left(x_{t-1}, \sigma_x^2\right) \quad (1)$$

and the observation of which is

$$y_t | x_t \sim \mathrm{N}\left(x_t, \sigma_y^2\right) \quad (2)$$

Essentially, the incorporated environmental covariate is a random walk process with white noise. In this state-space assessment model, the environmental covariate can be incorporated into the stock-recruit function and therefore adjust the expected recruitment at time t:

$$logN_{t,1} | x_{t-1}, SSB_{t-1} \sim \mathrm{N}\left(g, \sigma_{N_1}^2\right) \quad (3)$$

where g is an environmentally-explicit Beverton-Holt stock-recruit function. Throughout this paper, recruitment ($N_{t,1}$) is used to refer to the abundance of age-1 fish unless otherwise noted. The environmental covariate (x) and abundance-at-age (N) are both random-effect variables and estimated in ADMB based on empirical Bayes (Fournier et al., 2012). The state-space assessment model fitted to the data between 1973-2011, including three abundance indices from bottom trawl surveys, two spawning stock indices from ichthyoplankton surveys, commercial catch, and annual age composition observations from the three bottom trawl surveys and the commercial catch (see Miller et al. 2016).

Miller et al. (2016) found that performance of the state-space assessment model was improved by including CPI effect on recruitment. As the CPI was hypothesized to affect the carrying capacity for pre-recruits, CPI was modeled as a “limiting factor” in the Beverton-Holt stock-recruit function (see Iles and Beverton 1998). However, Iles and Beverton (1998) also considered effects of the environment on spawner density and(or) mortality (controlling factor) and on pre-recruit mortality and(or) growth (masking factor) (Fry, 1971, Neill et al., 1994). To evaluate the sensitivity of the state-space assessment model to the form of the environmentally-explicit Beverton-Holt stock-recruit function, we also incorporated the most strongly correlated environmental covariate into the Beverton-Holt stock-recruit function as a controlling and masking factor.

After finding the best fitting environmentally-explicit stock-recruit function for SNEMA yellowtail flounder, we compared the estimates and predictions of three population attributes (recruitment, SSB, and fully-selected fishing mortality (F)) and two biological reference points (maximum sustainable yield (MSY) and SSB_{MSY}) provided by the two models with and without the environmental effect on recruitment. Both models made five-year predictions for years 2012-2016 under the assumption that future F is at the level that produce the MSY (F_{MSY}). To evaluate which model can provide more reliable population predictions for 2012-2016, the prediction skill of the two models were compared by generating a series of retrospective predictions for each model.

In fisheries stock assessments, assessing model performance by generating a series of retrospective assessments is not a new idea. Terminal year population estimates are of key importance to stock status determination and harvest management, but are usually subject to retrospective bias (Mohn, 1999, Legault, 2009). Retrospective bias arises due to misspecification in stock assessment models (Legault, 2009) and is usually evaluated in the corresponding retrospective pattern generated through refitting the model to the data after removing (peeling) its terminal year sequentially for several times (Mohn, 1999). Similar to assessing the retrospective bias by generating a series of retrospective peels, the prediction skill of stock assessment models can also be assessed by generating a series of retrospective predictions using the “true” F during the prediction years and then comparing the retrospective predictions of population attributes with the corresponding “true” values. The “true” values in this paper are defined as the estimates from the assessment using the full data from 1973 to 2011. The “true” F rather than “true” catch was specified in predictions considering that the "true" catch can exceed the estimated population biomass and result in negative population abundance (Brooks & Legault, 2015). As this paper is focused on understanding environmental effects on SNEMA yellowtail flounder recruitment, we chose recruitment as the target population attribute in retrospective predictions. For each model, 13 retrospective three-year predictions were generated in a way similar to retrospective peeling: first, the state-space model fitted to the data between 1973-2008 with three years (2009-2012) recruitment predicted; then, the state-space model fitted to the data between 1973-2007 with three years (2008-2011) recruitment predicted; repeated this process in the same manner until the state-space model fitted to the data between 1973-1996 with three years (1997-1999) recruitment predicted. The mean relative difference (MRD) and mean absolute relative difference (MARD) of the 13 retrospective recruitment predictions from the "true" recruitment were calculated for each prediction lead time (from one year to three years) to quantitatively compare the retrospective prediction skill among the candidate models. The MRD and MARD for prediction lead year t were calculated as

$$\mathrm{MRD_t} = \frac{1}{13}\sum_{i=1996}^{2008}\frac{\theta_{i,t}-\Theta_{i+t}}{\Theta_{i+t}} \quad (4)$$

$$\mathrm{MARD_t} = \frac{1}{13}\sum_{i=1996}^{2008}\frac{|\theta_{i,t}-\Theta_{i+t}|}{\Theta_{i+t}} \quad (5)$$

respectively. $\theta_{i,t}$ is the t^{th} recruitment prediction from the state-space model fitted to the data up to year i, and Θ_{i+t} is the corresponding "true" recruitment the full data in year $i+t$.

RESULTS

Environmental drivers of recruitment deviations

While the correlation of recruitment deviations with the CPI was significant and stronger than with any atmospheric indices, the strongest correlations were observed with the two Gulf Stream related indices, especially the GSI (Table 1; Fig. 1). The GSI and recruitment deviations were negatively correlated with a lag of one year (Table 1). In other words, the latitudinal position of the Gulf Stream negatively impacted the abundance of age 1 fish one year later.

The GSI and CPI were then separately incorporated in the state-space assessment model in the Beverton-Holt stock-recruit function either as a limiting, controlling, or masking factor. Based on AIC, the GSI-incorporated models performed better than the CPI-incorporated models (Table 2), which is in agreement with the stronger correlation of GSI with recruitment deviations. Moreover, among the GSI-incorporated models, the one with GSI assumed to be a limiting factor performed best and more than two times more likely to be the best model than is the second-best one with GSI assumed to be a controlling factor.

Myers (1998) noticed that only 1 out of 47 environmental-recruitment correlations was used in routine stock assessments at the time of his study, and moreover, most environmental-recruitment correlations broke down when more years of data were added. As another measure of model performance, we compared the retrospective AIC values from the best model above and from the best model Miller et al. (2016) found (with the CPI assumed to be a limiting factor). Consistently in all seven peels, the GSI-incorporated model had smaller AIC values than the CPI-incorporated model (Table 3), indicating that the GSI-incorporated model consistently outperformed the CPI-incorporated model over time. The retrospective estimates of the environmental link parameter showed that the sign and degree of the GSI effect on recruitment were also consistent as addition years of data were included (Table 3). The two models were also compared with respect to the Mohn's ρ, which was defined in this study as the mean of the seven relative differences in each terminal year. Compared to the CPI-incorporated model, the GSI-incorporated model had larger Mohn's ρ for all three population attributes while the differences in Mohn's ρ are negligible for SB and F (Table 4).

Effects of the GSI on predicting recruitment

The estimated stock-recruit function in the two models with (R(SSB&GSI)) and without (R(SSB)) the GSI effect on recruitment was first compared. When recruitment is solely a function of SSB, the recruitment expected from a given SSB is always constant. However, when recruitment is also a function of environment and the environmental effect is strong, recruitment can vary dramatically with the environment (Fig. 2). Given that R(SSB) and R(SSB&GSI) differ in the stock-recruit function, the estimates and especially predictions of population attributes and biological reference points provided by the two models are expected to be different. It is important to note that R(SSB) was treated as a base model in this study to evaluate the consequences of incorporating an environmental covariate into a stock assessment model, it however was not considered in the last benchmark assessment because the stock-recruit relationship was not used (NEFSC, 2012).

As for recruitment, R(SSB) and R(SSB&GSI) provided similar estimates before 2011 (except in some individual years such as 1975-1980) but notably different five-year predictions for 2012-2016 (Fig. 3a). Although under the same harvest scenario, R(SSB) predicted that future recruitment will be increasingly higher while R(SSB&GSI) predicted that future recruitment will be persistently lower than that estimated in the terminal year. The SSB estimates provided by the two models were also similar and the SSB predictions provided by the two models were also notably different (Fig. 3b). Specifically. R(SSB) provided higher SSB predictions than R(SSB&GSI), primarily due to higher recruitment predictions from R(SSB). The recent unfavorable environmental conditions negatively affected recruitment and resulted in a decreasing stock size is predicted by R(SSB&GSI) for the next five years. By contrast, the higher SSB predicted by R(SSB) provides an optimistic view that the stock size will slowly rebuild over the next five years. Same as recruitment and SSB, F was also estimated to be similar in R(SSB) and R(SSB&GSI) (Fig. 3c).

While R(SSB) and R(SSB&GSI) provided similar F estimates, the estimated F_{MSY} from the two models were notably different (Fig. 3c). Specifically, the F_{MSY} estimate from R(SSB) is very close to the reference point from the most recent benchmark assessment ($F_{40\%}$), but notably smaller than that from R(SSB&GSI). Both MSY and SSB_{MSY} in this state-space assessment model are functions of the incorporated environmental covariate, so their estimates from R(SSB&GSI) varied annually with the GSI. The MSY estimates from R(SSB&GSI) were relatively low since the 1990's (Fig. 3d) as unfavorable environmental conditions (indicated by high GSI values) were more frequent during that time (Fig. 1). Note that the MSY estimates from R(SSB) is time-invariant also due to the fact that the terminal year weight-at-age and selectivity were used in the calculation of MSY for all previous years. In this case, only using a constant set of values for these allows us to look at how MSY varies annually just due to the annual fluctuation in the environment. As for $\log(SSB/SSB_{MSY})$, both the estimates and predictions from the two models differ substantially (Fig. 3e). When the GSI values were low (i.e., environmental conditions were favorable) before 1990, the stock was estimated to be more productive and therefore $\log(SSB/SSB_{MSY})$ was estimated to be lower from R(SSB&GSI) than from R(SSB). Conversely, when the GSI values were high (i.e., environmental conditions were unfavorable) after 1990, the stock was estimated to be less productive and therefore $\log(SSB/SSB_{MSY})$ was estimated to be higher from R(SSB&GSI) than from R(SSB).

Overall, both models over-predicted recruitment after 1996, except in a few years between 2003-2006 (Fig. 4). The recruitment predictions from R(SSB&GSI) were generally higher than those from R(SSB) when the terminal year GSI values were lower than the long-term average and vice versa, as a result of the negative correlation between the GSI and recruitment deviations. As expected, the recruitment predictions from either model become more biased (larger MRD) and less accurate (larger MARD) as prediction lead time increases (Table 5). Generally speaking, incorporating the GSI into the stock-recruit function improved the accuracy of recruitment predictions as suggested by a smaller MARD for R(SSB&GSI). Also, it reduced the bias in recruitment predictions as suggested by a smaller MRD for R(SSB&GSI). The importance of the incorporation to recruitment predictions is most pronounced in the first prediction year and finally becomes negligible in the third prediction year.

We compared each retrospective prediction pair from the two models and found that the relative performance of the two models in predicting recruitment had dramatic year-to-year fluctuations and neither model consistently outperformed the other in predicting recruitment (Fig. 4). Although both MRD and MARD are smaller for R(SSB&GSI), the comparison indicates that R(SSB&GSI) fails to provide better recruitment prediction in all 13 retrospective prediction cases. Because the GSI is modeled as a random walk, the best future prediction is the same as the estimate in the last observed year of data. However, the GSI had very large interannual fluctuation relative to the long-term average, so the annual

GSI prediction could be very different from the “true” GSI. As a consequence, the recruitment predictions from R(SSB&GSI) could also be very different from the “true” recruitment. For example, if the GSI did not change notably in the three prediction years (e.g., 2000-2002), the GSI and its effect on recruitment were found to be more accurately predicted by R(SSB&GSI). In this case, the recruitment predictions from R(SSB&GSI) (the bold solid line starting in 2000) were more similar to the “true” recruitments than the predictions from R(SSB) which did not account for the unfavorable environmental conditions (the bold dashed line starting in 2000). In contrast, if the GSI changed dramatically in the three prediction years (e.g., 2006-2008), the GSI and its effect on recruitment were poorly predicted by R(SSB&GSI). In this case, the recruitment predictions from R(SSB&GSI) (the bolded solid line starting in 2006) were further from the “true” recruitments than the predictions from R(SSB) (the bold dashed line starting in 2006).

DISCUSSION

This paper uses the state-space age-structured assessment model from Miller et al. (2016) to explore other environmental covariates to explain SNEMA yellowtail flounder recruitment variability and builds on it to evaluate alternative effects of environmental covariates within the Beverton-Holt stock-recruit relationship. Furthermore, we explore the ability of environmental covariates to improve prediction of future recruitments. Specifically, we incorporated indicators of climate variability directly into the stock-recruit function as a limiting, controlling, and masking factor, respectively. Compared to the model without any environmental covariate, the model with GSI as a limiting factor performed better with respect to AIC and provided recruitment predictions that were closer to the “true” recruitments estimated from the full data with respect to both MRD and MARD. However, the recruitment predictions provided by the model with GSI were not closer to the “true” recruitments in every single retrospective prediction case. Indeed, we found that recruitment predictions from the model with GSI can be further from the “true” values when GSI predictions from that model are far away from the “true” GSI. Therefore, we suggest to treat the environmentally-explicit model as an alternative model, instead of the best and only model, to be considered in population prediction and stock management. The model with GSI as a limiting factor strongly suggested that the recent low productivity of the SNEMA yellowtail flounder can be explained by the unfavorable environmental conditions result from a northward shift of the Gulf Stream, and also, the stock has been rebuilt relative to the current productivity level.

Many environmental indicators including the CPI were correlated with the recruitment deviations taken from the stock assessment, but the correlation was strongest for the GSI. The CPI is a local-scale environmental index representing bottom temperature condition in the Mid-Atlantic Cold Pool (NEFSC, 2012) while the GSI is a basin-scale environmental index representing the latitudinal anomalies of the Gulf Stream path (Joyce & Zhang, 2010). Comparison of model fits suggests that local bottom temperature in the Mid-Atlantic Cold Pool should not be the only environmental factor affecting SNEMA yellowtail flounder recruitment. Indeed, the basin-scale GSI also indicates some other shelf physical/biological conditions that potentially affect the recruitment: (1) shelf SST condition (Gawarkiewicz et al., 2012), which affects the physiology of SNEMA yellowtail flounder during the early pelagic phase; (2) shelf current and eddy conditions, which affect larval transport and retention on the continental shelf (Hare & Cowen, 1996); and (3) shelf primary production condition in spring (Saba et al., 2015), which affects food availability to the larvae. Hallett et al. (2004) argued that large-scale climate indices contain information on several local processes, so potentially they could better predict ecological processes compared to local weather conditions when a mechanistic understanding of how local environment influences a biological process is lacking. We hypothesize that the better performance of the GSI than the CPI in explaining SNEMA yellowtail flounder recruitment is due to the aggregation of factors beyond the bottom temperature in the Mid-Atlantic cold pool that affects recruitment.

Of the alternative type of environmental effects in the Beverton-Holt stock-recruit function, we found the “limiting factor” assumption where the carrying capacity of pre-recruits is regulated by the GSI to perform best for SNEMA yellowtail flounder (Iles & Beverton, 1998, Neill et al., 1994). We noticed that the differences in AIC between the models with different forms of the stock-recruit functions (limiting, controlling, or masking factor) were smaller than those between different environmental covariates incorporated in the stock-recruit function (GSI or CPI). The retrospective AIC pattern showed that the GSI-incorporated model was consistently better than the CPI-incorporated model in all seven retrospective peels. According to Mohn's ρ, another important metric of model performance, the retrospective biases in the estimates of recruitment, SSB, and F from the GSI-incorporated model were minimally larger than those from the CPI-incorporated model.

Even though the GSI provided a lower AIC than the CPI, it is important to also assess how the inclusion of GSI impacts model estimates and predictions. Comparison suggested that R(SSB) and R(SSB&GSI) provided similar estimates of recruitment, SSB, and F, but distinct estimates of biological reference points and predictions of recruitment and SSB. Indeed, the GSI affected the expected recruitment from the stock-recruit function. Although recruitment estimates from the two models were similar, R(SSB&GSI) still had a much smaller AIC as the deviations between the estimated and expected recruitment were generally closer to zero when the GSI effect on recruitment was included. When there are catch and survey data, recruitment estimates are informed by them and are also constrained by a penalty term to not be far from the stock-recruit function expected. However, in the prediction period when no fisheries data are available, the best recruitment predictions are from the stock-recruit function. Specifically, recruitment predictions from R(SSB) are based solely on SSB while those from R(SSB&GSI) are also profoundly affected by year-to-year fluctuations in the GSI. Since the only difference between R(SSB) and R(SSB&GSI) lies in the stock-recruit function, whether includes GSI effects on recruitment prediction is the only possible source responsible for the large differences between the predictions of recruitment and SSB from the two models.

In the first prediction year (i.e., 2012), the SSB predictions from R(SSB) and R(SSB&GSI) are not differentiable, although the recruitment predictions from R(SSB) and R(SSB&GSI) have been notably different. Since few SNEMA yellowtail flounder can be mature at age 1, recruitment minimally impacts SSB in the first prediction year. In the second prediction year (i.e., 2013), the difference in recruitment prediction propagates to age 2 at which maturity reaches 0.5, leading to divergent SSB predictions from the two models. This divergence in SSB predictions then propagates back to recruitment one year later (i.e., 2014) through the stock-recruit function, resulting in an even lower recruitment prediction from R(SSB&GSI), as the high predicted GSI has already led to a lower recruitment prediction in 2014. The increasingly different predictions of recruitment and SSB from the two models clearly show the importance of selecting the most appropriate stock-recruit function to near-term population predictions.

Interesting enough, a higher F_{MSY} was estimated by R(SSB&GSI) than R(SSB) and the recruitment estimates (when GSI equal 0) from R(SSB&GSI) were also higher than from R(SSB). It indicates that the stock was estimated to be more productive at low SSB levels when the GSI effect is included in the stock-recruit function. The reason for which, we suspect, is that the recent low recruitments from R(SSB&GSI) were partly attributed to unfavorable environmental conditions, instead of being solely attributed to reduced SSB as those from R(SSB). In other words, when environmental conditions are neutral, the productivity and consequently F_{MSY} of SNEMA yellowtail flounder estimated from R(SSB&GSI) should be higher than R(SSB) in which the environmental effect on recruitment is not included.

As shown earlier, the five-year recruitment predictions from R(SSB) and R(SSB&GSI) differ substantially. To evaluate which model can provide more reliable recruitment predictions, we generated 13 retrospective predictions for each model and then compared those predictions with the “true” recruitments which were defined in this study as the estimates from the full data. Generally speaking, the differences between the predicted and “true” recruitments are smaller in R(SSB&GSI), especially in the first prediction year in which recruitment prediction is a function of the relatively reliable GSI observation in the last year of observations. The second and third recruitment predictions are functions of the increasingly unreliable GSI predictions from R(SSB&GSI), so the inclusion of GSI effect on recruitment leads to relatively small improvement in those recruitment predictions. As expected, comparison of either MRD or MARD indicated that incorporating the GSI into the stock-recruit function cannot reduce the difference between the predicted and “true” recruitment beyond a lead time of two years. In addition, we also made year-by-year comparison of the retrospective recruitment predictions from the two models to evaluate whether R(SSB&GSI) consistently outperformed R(SSB) over time in predicting recruitment. The year-by-year comparison suggested that the accuracy of recruitment predictions from R(SSB&GSI) is largely dependent on the accuracy of those years’ GSI prediction from R(SSB&GSI). Because the GSI had large interannual fluctuations relative to the long-term average, its predictions from a random walk model in R(SSB&GSI) could be considerably biased. If it happens, the recruitment predictions from R(SSB&GSI) are also expected to be considerably biased, albeit in the opposite direction. However, if the environmental covariate incorporated into the stock-recruit function is a low-frequency decadal oscillation such as the Pacific Decadal Oscillation, the random walk model is more likely to provide reliable near-term predictions for the environmental covariate. However, another problem can arise when incorporating a low-frequency decadal oscillation into the stock-recruit function. Haltuch and Punt (2011) found that when fisheries data are relativity short in time compared to the period of the incorporated decadal oscillation, stock assessment models are not able to correctly tell whether the incorporated environmental process has a significant effect on recruitment or not.

This study evaluated the skill of the state-space assessment model in predicting recruitment via implementing the retrospective prediction scheme that is popularly used in the climate science community. However, the skill of an assessment model in predicting recruitment needs to be interpreted differently from that of a climate model in predicting climate variables. Generally speaking, the predicted climate variable such as sea surface temperature can be observed through either *in situ* or remote sensing method. Therefore, the prediction skill can be evaluated by comparing model predictions with the corresponding observations that are relatively credible. By contrast, some population attributes predicted (e.g., recruitment and SSB) by stock assessment models are inherently unobservable in the field. This study evaluated the skill of the state-space model in predicting recruitment by comparing model predictions with the corresponding estimates from the full data, which are model output and less accurate than direct climate observations. As a result, a good skill in predicting recruitment does not necessarily equals accurate recruitment predictions. For instance, a high prediction skill can possibly exist when retrospective recruitment predictions and recruitment estimates from the full data are both systematically biased in the same direction to a large extent. Thus, like Mohn's ρ, the retrospective prediction skill is only one metric of model performance and should be evaluated together with other model diagnostics.

In brief, this paper provides two major suggestions for future fisheries studies that incorporate environmental effects into stock assessment models. First, care should be taken even when the model with an environmental covariate fits data better. When the environmental covariate is poorly predicted, the model with that environmental covariate can provide less accurate predictions than the model without any environmental covariate. In future work, alternative time series models for an environmental covariate should also be considered in the state-space model to potentially improve its predictive performance. For instance, a stationary autoregressive process of order greater than 1 has been found to be robust in predicting the GSI and its effect on silver hake distribution for the near-term (Davis et al., 2017). Second, analyzing the retrospective prediction pattern to quantitatively evaluate model prediction skill is recommended before making management decisions based on model predictions, in the same way that retrospective pattern is analyzed as a regular procedure in stock assessments to evaluate the biases in terminal year model estimate.

ACKNOWLEDGEMENT

We thank Sarah Gaichas, Chris Legault, two anonymous reviewers, and the editor for providing constructive comments and suggestions that improve the quality of this paper. This work was supported by a NOAA FATE Grant NA12OAR4320071 to Nye and Hameed. The authors have no conflict of interest to declare

REFERENCES

Boer, G., Kharin, V. and Merryfield, W. (2013) Decadal predictability and forecast skill. *Climate dynamics,* **41,** 1817-1833.

Brooks, E.N. and Deroba, J.J. (2015) When “data” are not data: the pitfalls of post hoc analyses that use stock assessment model output. *Canadian Journal of Fisheries and Aquatic Sciences,* **72,** 634-641.

Brooks, E.N. and Legault, C.M. (2015) Retrospective forecasting—evaluating performance of stock projections for New England groundfish stocks. *Canadian Journal of Fisheries and Aquatic Sciences,* **73,** 935-950.

Davis, X.J., Joyce, T.M. and Kwon, Y.-O. (2017) Prediction of silver hake distribution on the Northeast US shelf based on the Gulf Stream path index. *Continental Shelf Research,* **138,** 51-64.

Drinkwater, K.F., Belgrano, A., Borja, A., Conversi, A., Edwards, M., Greene, C.H., Ottersen, G., Pershing, A.J. and Walker, H. (2003) the response of marine ecosystems to climate variability associated with the North Atlantic Oscillation. *Geophysical Monograph Series,* **134,** 211-234.

Essington, T.E., Moriarty, P.E., Froehlich, H.E., Hodgson, E.E., Koehn, L.E., Oken, K.L., Siple, M.C. and Stawitz, C.C. (2015) Fishing amplifies forage fish population collapses. *Proceedings of the National Academy of Sciences,* **112,** 6648-6652.

Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M.N., Nielsen, A. and Sibert, J. (2012) AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software,* **27,** 233-249.

Fry, F. (1971) 1 The effect of environmental factors on the physiology of fish. *Fish physiology,* **6,** 1-98.

Gawarkiewicz, G.G., Todd, R.E., Plueddemann, A.J., Andres, M. and Manning, J.P. (2012) Direct interaction between the Gulf Stream and the shelfbreak south of New England. *Scientific reports,* **2**.

Greene, C.H., Meyer-Gutbrod, E., Monger, B.C., McGarry, L.P., Pershing, A.J., Belkin, I.M., Fratantoni, P.S., Mountain, D.G., Pickart, R.S. and Proshutinsky, A. (2013) Remote climate forcing of decadal-scale regime shifts in Northwest Atlantic shelf ecosystems. *Limnol. Oceanogr,* **58,** 803-816.

Haddon, M. (2010) *Modelling and quantitative methods in fisheries.* CRC press.

Hallett, T., Coulson, T., Pilkington, J., Clutton-Brock, T., Pemberton, J. and Grenfell, B. (2004) Why large-scale climate indices seem to predict ecological processes better than local weather. *Nature,* **430,** 71-75.

Haltuch, M.A. and Punt, A.E. (2011) The promises and pitfalls of including decadal-scale climate forcing of recruitment in groundfish stock assessment. *Canadian Journal of Fisheries and Aquatic Sciences,* **68,** 912-926.

Hameed, S. and Piontkovski, S. (2004) The dominant influence of the Icelandic Low on the position of the Gulf Stream northwall. *Geophysical research letters,* **31**.

Hare, J.A. and Cowen, R.K. (1996) Transport mechanisms of larval and pelagic juvenile bluefish (Pomatomus saltatrix) from South Atlantic Bight spawning grounds to Middle Atlantic Bight nursery habitats. *Limnology and Oceanography,* **41,** 1264-1280.

Hilborn, R. (2016) Correlation and causation in fisheries and watershed management. *Fisheries,* **41,** 18-25.

Hill, K.T., Crone, P.R., Lo, N.C., Macewicz, B.J., Dorval, E., McDaniel, J.D. and Gu, Y. (2011) Assessment of the Pacific sardine resource in 2011 for US management in 2012. *NOAA Technical Memorandum NMFS-SWFSC,* **487**.

Houghton, R.W., Schlitz, R., Beardsley, R.C., Butman, B. and Chamberlin, J.L. (1982) The Middle Atlantic Bight cold pool: Evolution of the temperature structure during summer 1979. *Journal of Physical Oceanography,* **12,** 1019-1029.

Iles, T. and Beverton, R. (1998) Stock, recruitment and moderating processes in flatfish. *Journal of Sea Research,* **39,** 41-55.

Joyce, T.M. and Zhang, R. (2010) On the Path of the Gulf Stream and the Atlantic Meridional Overturning Circulation. *Journal of Climate,* **23**.

Legault, C.M. (2009) Report of the retrospective working group. *NOAA NMFS Northeast Fisheries Science Center Reference Document***,** 09-01.

Meehl, G.A., Goddard, L., Murphy, J., Stouffer, R.J., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M.A., Greene, A.M. and Hawkins, E. (2009) Decadal prediction: can it be skillful? *Bulletin of the American Meteorological Society,* **90,** 1467-1485.

Miller, T.J., Hare, J.A. and Alade, L.A. (2016) A state-space approach to incorporating environmental effects on recruitment in an age-structured assessment model with an application to Southern New England yellowtail flounder. *Canadian Journal of Fisheries and Aquatic Sciences,* **73,** 1261-1270.

Mohn, R. (1999) The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. *ICES Journal of Marine Science: Journal du Conseil,* **56,** 473-488.

Mountain, D.G. (2012) Labrador slope water entering the Gulf of Maine—response to the North Atlantic Oscillation. *Continental Shelf Research,* **47,** 150-155.

Myers, R.A. (1998) When do environment–recruitment correlations work? *Reviews in Fish Biology and Fisheries,* **8,** 285-305.

NEFSC (2012) 54th Northeast Regional Stock Assessment Workshop (54th SAW) Assessment Report. *US Dept Commer, Northeast Fish Sci Cent Ref Doc. 12-18; 600 p*.

Neill, W.H., Miller, J.M., Van Der Veer, H.W. and WINEMIllER, K.O. (1994) Ecophysiology of marine fish recruitment: a conceptual framework for understanding interannual variability. *Netherlands Journal of Sea Research,* **32,** 135-152.

Nielsen, A. and Berg, C.W. (2014) Estimation of time-varying selectivity in stock assessments using state-space models. *Fisheries Research,* **158,** 96-101.

Nye, J.A., Joyce, T.M., Kwon, Y.O. and Link, J.S. (2011) Silver hake tracks changes in Northwest Atlantic circulation. *Nature communications,* **2,** 412.

Punt, A.E., A'mar, T., Bond, N.A., Butterworth, D.S., de Moor, C.L., De Oliveira, J.A., Haltuch, M.A., Hollowed, A.B. and Szuwalski, C. (2014) Fisheries management under climate and environmental uncertainty: control rules and performance simulation. *ICES Journal of Marine Science,* **71,** 2208-2220.

Quinn, T.J. and Deriso, R.B. (1999) *Quantitative fish dynamics.* Oxford University Press.

Saba, V.S., Hyde, K.J., Rebuck, N.D., Friedland, K.D., Hare, J.A., Kahru, M. and Fogarty, M.J. (2015) Physical associations to spring phytoplankton biomass interannual variability in the US Northeast Continental Shelf. *Journal of Geophysical Research: Biogeosciences,* **120,** 205-220.

Schirripa, M.J. (2007) Status of the sablefish resource off the continental US Pacific Coast in 2007. *Pacific Fishery Management Council, Portland, OR*.

Sullivan, M.C., Cowen, R.K., Able, K.W. and Fahay, M.P. (2000) Spatial scaling of recruitment in four continental shelf fishes. *Marine Ecology Progress Series,* **207,** 141-154.

Sullivan, M.C., Cowen, R.K. and Steves, B.P. (2005) Evidence for atmosphere–ocean forcing of yellowtail flounder (Limanda ferruginea) recruitment in the Middle Atlantic Bight. *Fisheries Oceanography,* **14,** 386-399.

Szuwalski, C.S. and Hollowed, A.B. (2016) Climate change and non-stationary population processes in fisheries management. *ICES Journal of Marine Science,* **73,** 1297-1305.

Szuwalski, C.S., Vert-Pre, K.A., Punt, A.E., Branch, T.A. and Hilborn, R. (2015) Examining common assumptions about recruitment: a meta-analysis of recruitment dynamics for worldwide marine fisheries. *Fish and Fisheries,* **16,** 633-648.

Taylor, A.H. and Stephens, J.A. (1998) The North Atlantic oscillation and the latitude of the Gulf Stream. *Tellus A,* **50,** 134-142.

Vert-pre, K.A., Amoroso, R.O., Jensen, O.P. and Hilborn, R. (2013) Frequency and intensity of productivity regime shifts in marine fish stocks. *Proceedings of the National Academy of Sciences,* **110,** 1779-1784.

Walters, C.J. and Collie, J.S. (1988) Is research on environmental factors useful to fisheries management? *Canadian Journal of Fisheries and Aquatic Sciences,* **45,** 1848-1854.

Xu, H., Kim, H.-M., Nye, J.A. and Hameed, S. (2015) Impacts of the North Atlantic Oscillation on sea surface temperature on the Northeast US Continental Shelf. *Continental Shelf Research,* **105,** 60-66.

TABLES

Table 1. The Pearson correlation coefficients between SNEMA yellowtail flounder recruitment deviations and various environmental indices with a lag of one to three years. Positive lags mean the environment leads the recruitment. IL stands for Icelandic low and AH stands for Azores high. Only correlations that are significant at the 95% confidence level are shown and the coefficient marked in bold represents the correlation is significant at the 99% confidence level.

Environmental index	Lag 1	Lag 2	Lag 3
IL Pressure	0.38		
IL Longitude			
IL Latitude			
AH Pressure			
AH Longitude			-0.36

AH Latitude			-0.38
NAO		-0.37	-0.38
GSI	**-0.52**		
GSNW	-0.41		
CPI	-0.39		

Table 2. The fits of the state-space assessment model with different environmentally-explicit stock-recruit functions. These fits are compared according to AIC and Akaike weight, which are shown in column three and four, respectively.

Model	Stock-recruit function	ΔAIC	ω (AIC)
R(SSB)	$\frac{SSB}{b+aSSB}$	13.89	0.00
R ($CPI_{limiting}$, SSB)	$\frac{SSB}{b+aSSBe^{cCPI}}$	5.40	0.04
R ($CPI_{masking}$, SSB)	$\frac{SSB}{be^{cCPI}+aSSB}$	4.53	0.06
R ($CPI_{controlling}$, SSB)	$\frac{SSB}{b+aSSB}e^{cCPI}$	3.93	0.08
R ($GSI_{limiting}$, SSB)	$\frac{SSB}{b+aSSBe^{cGSI}}$	0.00	0.55
R ($GSI_{masking}$, SSB)	$\frac{SSB}{be^{cGSI}+aSSB}$	4.94	0.05
R ($GSI_{controlling}$, SSB)	$\frac{SSB}{b+aSSB}e^{cGSI}$	1.67	0.24

Table 3. The retrospective difference in AIC between the model with the GSI as a limiting factor and that with the CPI as a limiting factor (column 2) as well as the retrospective estimates of the environmental link parameter and the associated standard deviation from the model with the GSI as a limiting factor. Positive difference in AIC corresponds to the GSI-incorporated model outperforms the CPI-incorporated model and vice versa.

Peel	AIC(CPI) - AIC(GSI)	c (sd)
0	5.40	1.53 (0.37)
1	5.17	1.52 (0.37)
2	5.48	1.53 (0.37)
3	5.55	1.51 (0.36)
4	7.31	1.50 (0.34)
5	12.08	1.46 (0.29)
6	10.51	1.51 (0.32)

Table 4. The Mohn's ρ of SSB, F, and recruitment from the state-space assessment model with and without environmental (GSI or CPI) effect on recruitment. A smaller Mohn's ρ (absolute value) corresponds to less retrospective bias and consequently better model performance.

Model	ρ (SSB)	ρ (F)	ρ (R)
R(SSB)	0.11	-0.14	0.24
R ($CPI_{limiting}$, SSB)	0.11	-0.14	0.22
R ($GSI_{limiting}$, SSB)	0.14	-0.16	0.36

Table 5. The mean relative difference and mean absolute relative difference in the near-term recruitment predictions from R(SSB) and R(SSB&GSI).

Prediction lead time	1 year	2 years	3 years
MRD - R(SSB)	1.23	1.53	1.68
MRD - R(SSB&GSI)	0.89	1.26	1.59
MARD - R(SSB)	1.45	1.73	1.76
MARD - R(SSB&GSI)	1.04	1.50	1.77

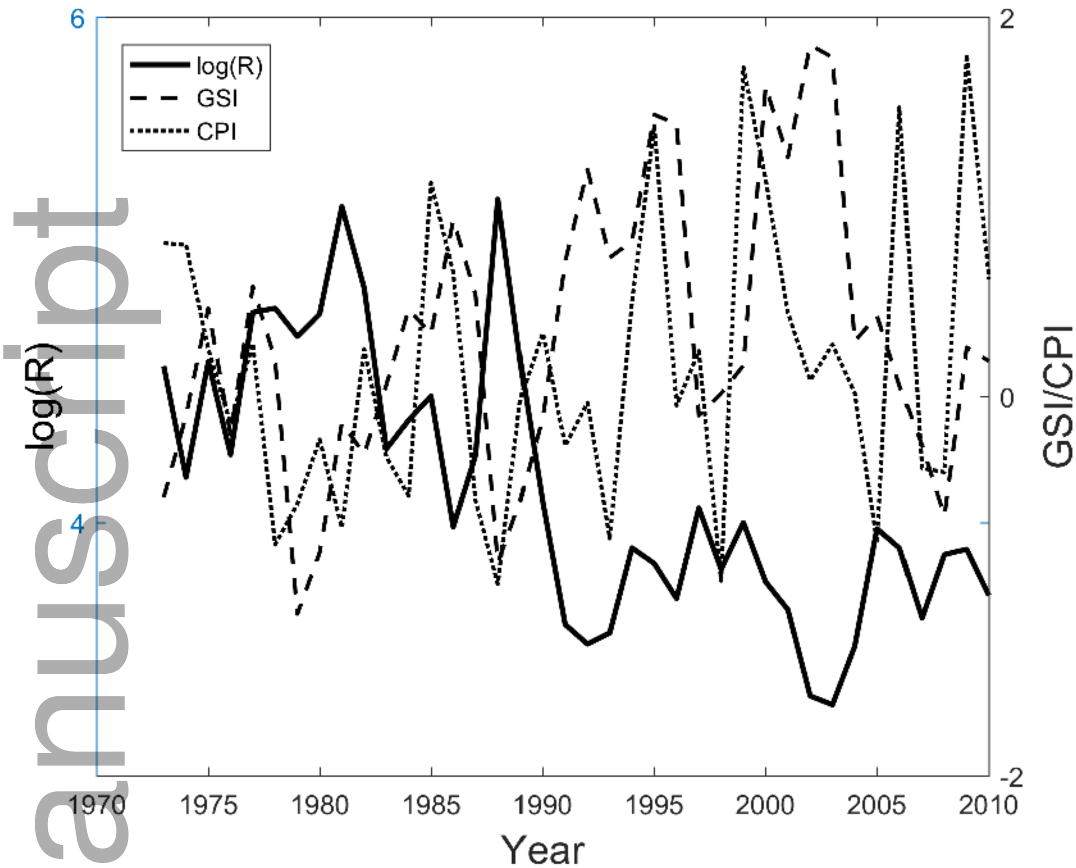
FIGURE CAPTIONS

Figure 1. Compare the natural log of SNEMA yellowtail flounder recruitment with the two most related environmental indices (GSI and CPI). The recruitment time series is from the baseline run in the most recent benchmark assessment (NEFSC, 2012) and the CPI is scaled to have the same variance as the GSI for easier comparison. Both environmental indices in this figure have been shifted one year backward to account for the one-year lag between them with recruitment.

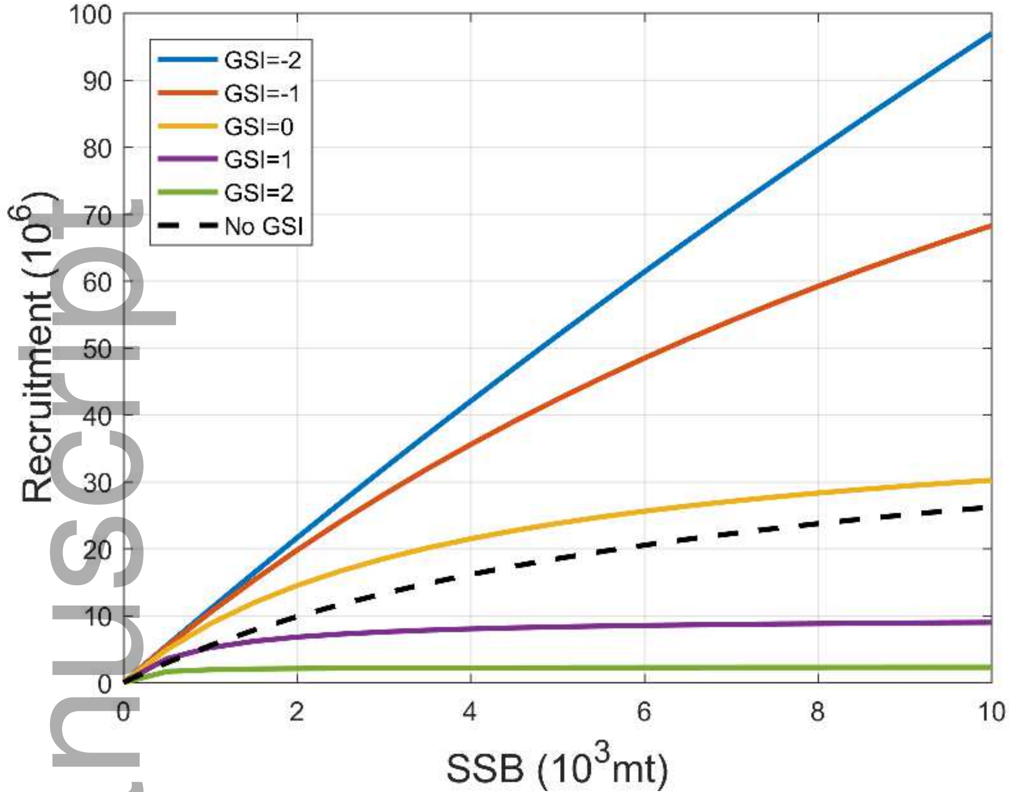
Figure 2. The estimated Beverton-Holt stock-recruit function from R(SSB) (dashed line) in comparison to those from R(SSB&GSI) (solid lines) under various GSI values.

Figure 3. The first row shows the estimated and predicted recruitment (a) and SSB from the two comparing models under the F_{MSY} harvest scenario. The second row shows the estimated F and F_{MSY} (c) and the estimated MSY (d) from the two comparing models. The third row shows the estimated and predicted log (SSB/SSB_{MSY}) from the two comparing models under the F_{MSY} harvest scenario (e). In this figure, the color dashed lines and vertical error bars represent the 95% confidence interval, and the black vertical dashed lines mark last year of data.

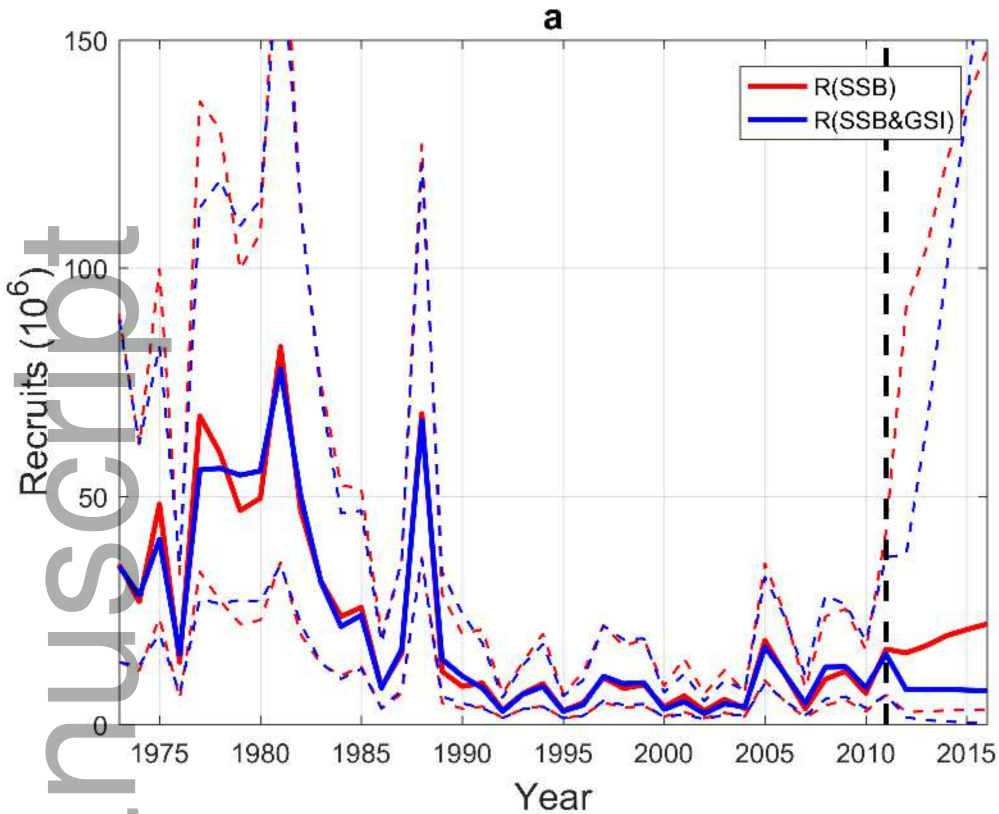
Figure 4. The retrospective prediction patterns from R(SSB) (dashed lines) and R(SSB&GSI) (solid lines). The two black lines represent the “true” recruitment from each model, i.e., the recruitment estimates when performing the assessment on full data. The dots are the terminal year recruitment estimates and color of lines and dots represents the last year of data available for each retrospective prediction, from dark blue (1996) to dark red (2008). The two pairs of recruitment predictions mentioned in the discussion are highlighted as bold color lines.



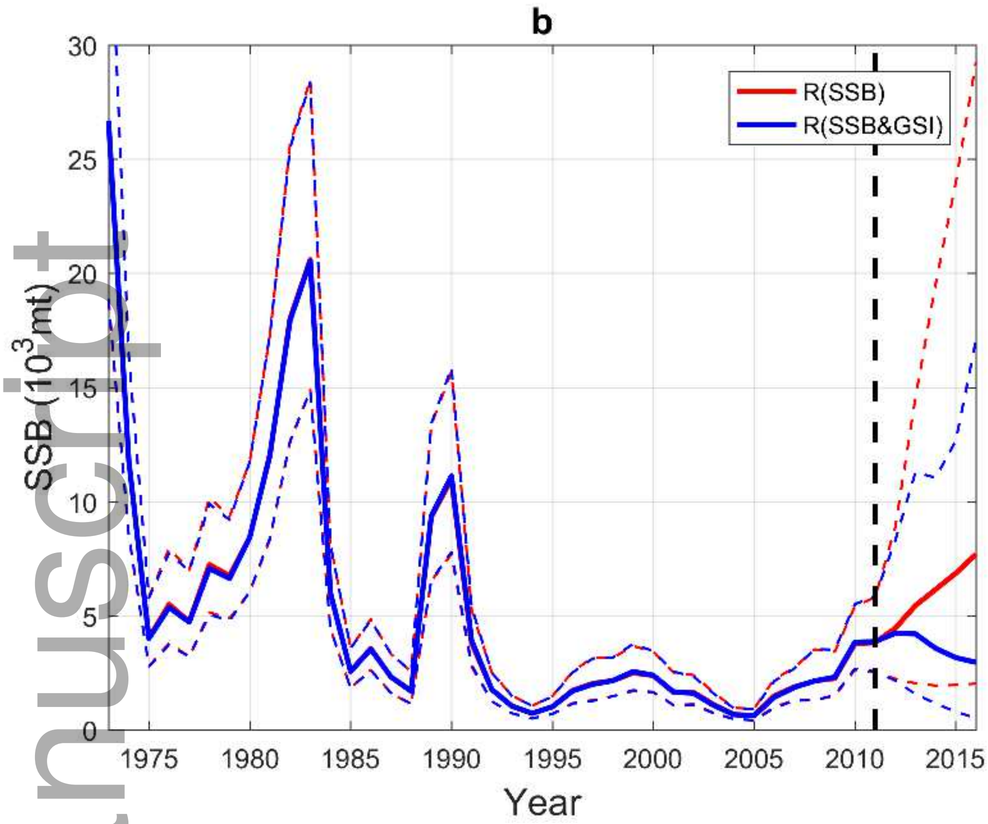
fog_12236_f1.tif



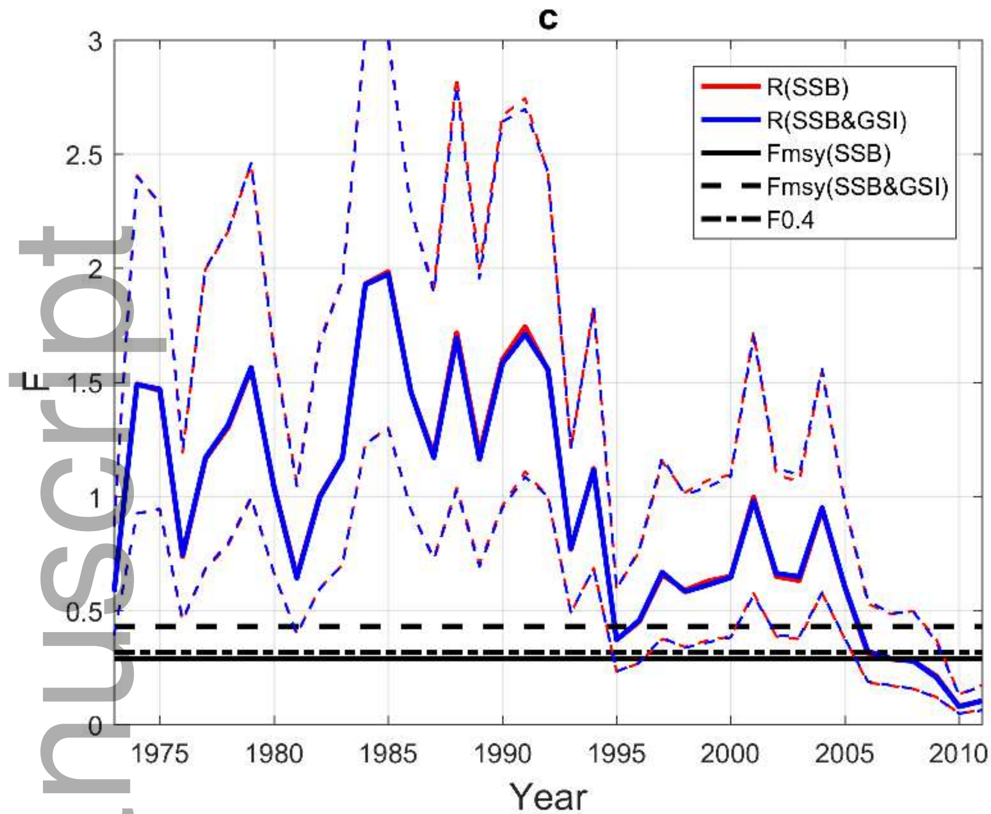
fog_12236_f2.tif



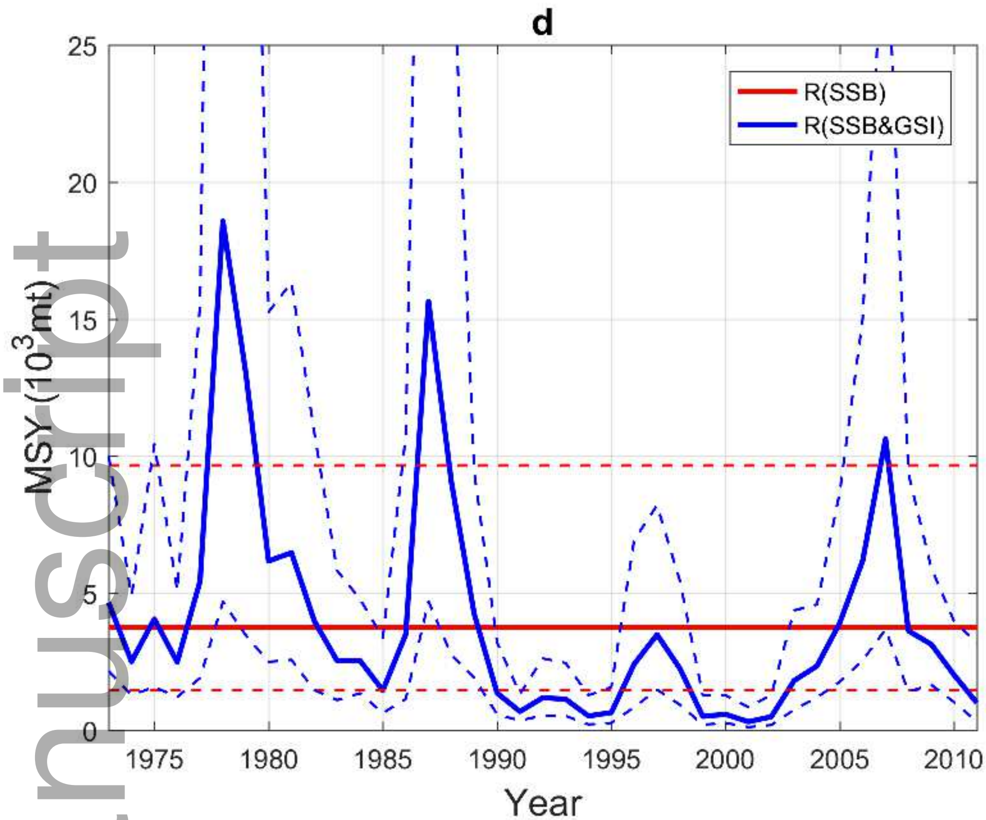
fog_12236_f3a.tif



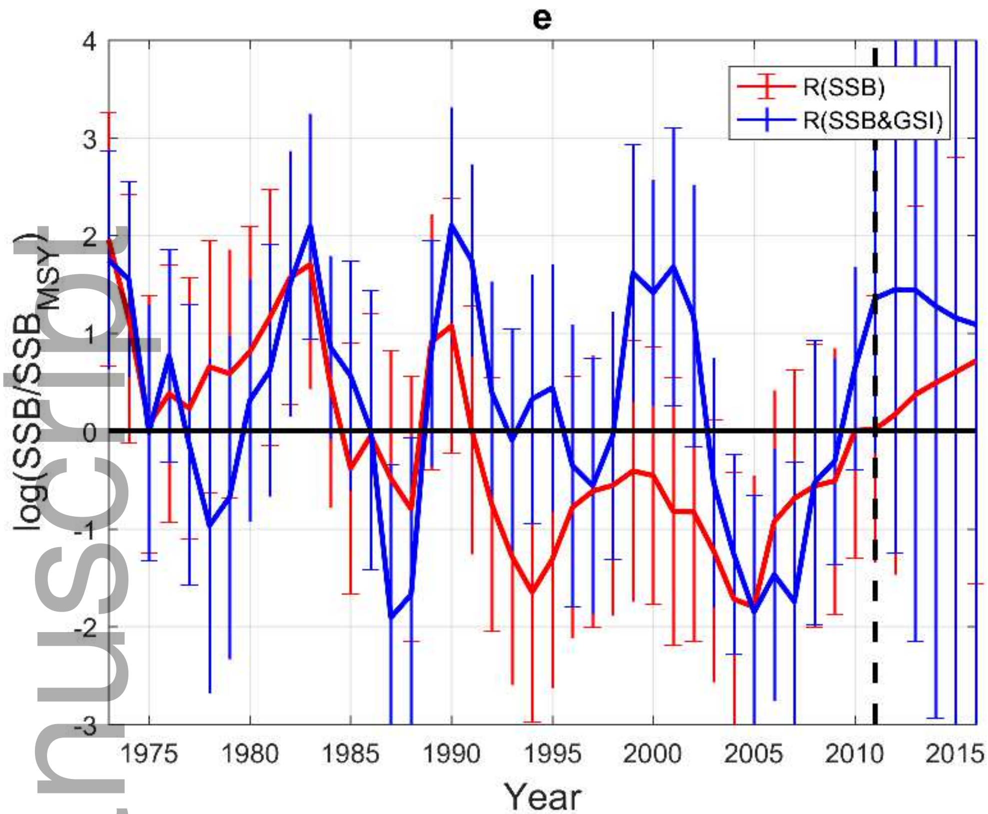
fog_12236_f3b.tif



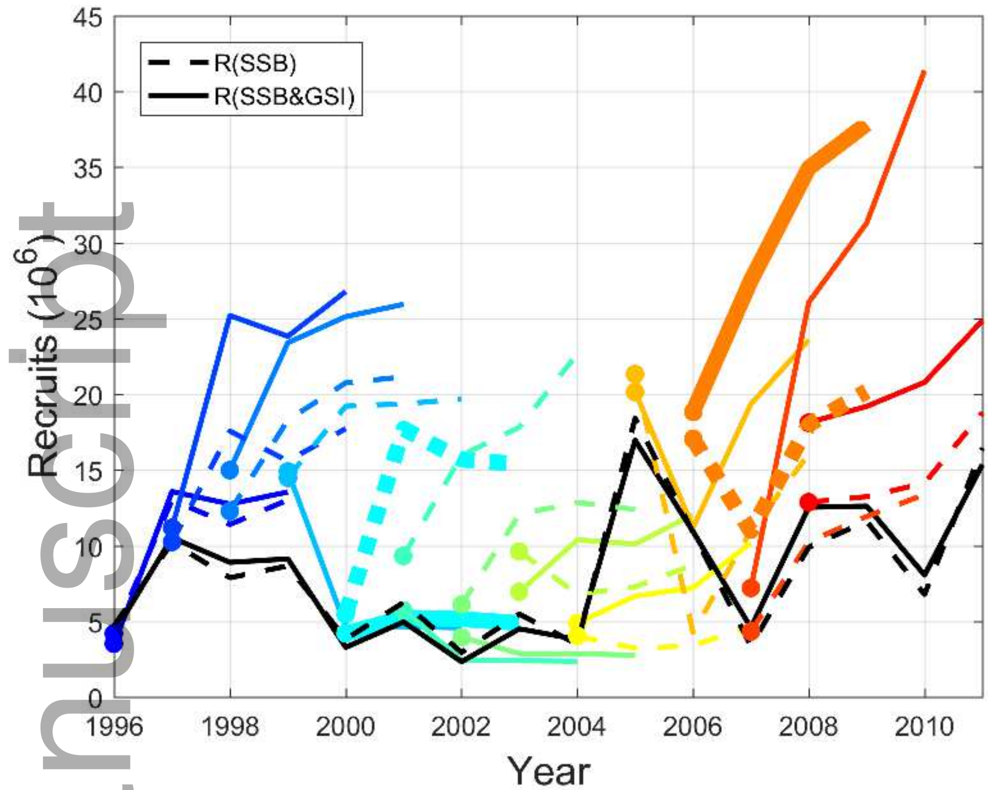
fog_12236_f3c.tif



fog_12236_f3d.tif



fog_12236_f3e.tif



fog_12236_f4.tif