Published in final edited form as:
J Mar Syst. 2009 February 20; 76(1-2): 151–161. doi:10.1016/j.jmarsys.2008.05.016.

Skill assessment for an operational algal bloom forecast system

Richard P. Stumpf[a,*], **Michelle C. Tomlinson**[a], **Julie A. Calkins**[b], **Barbara Kirkpatrick**[c], **Kathleen Fisher**[d], **Kate Nierenberg**[c], **Robert Currier**[c], and **Timothy T. Wynne**[e]

[a]NOAA, National Ocean Service, 1305 East-West Highway, 9th floor, Silver Spring, MD 20910, USA

[b]CSS, 1305 East-West Highway, Silver Spring, MD 20910, USA

[c]Mote Marine Laboratory, 1600 Ken Thompson Parkway, Sarasota, FL 34236, USA

[d]NOAA, Center for Operational Oceanographic Products and Services, 1305 East-West Highway, Silver Spring, MD 20910, USA

[e]IMSG, 1305 East-West Highway, Silver Spring, MD 20910, USA

Abstract

An operational forecast system for harmful algal blooms (HABs) in southwest Florida is analyzed for forecasting skill. The HABs, caused by the toxic dinoflagellate, *Karenia brevis*, lead to shellfish toxicity and to respiratory irritation. In addition to predicting new blooms and their extent, HAB forecasts are made twice weekly during a bloom event, using a combination of satellite derived image products, wind predictions, and a rule-based model derived from previous observations and research. These forecasts include: identification, intensification, transport, extent, and impact; the latter being the most significant to the public. Identification involves identifying new blooms as HABs and is validated against an operational monitoring program involving water sampling. Intensification forecasts, which are much less frequently made, can only be evaluated with satellite data on mono-specific blooms. Extent and transport forecasts of HABs are also evaluated against the water samples. Due to the resolution of the forecasts and available validation data, skill cannot be resolved at scales finer than 30 km. Initially, respiratory irritation forecasts were analyzed using anecdotal information, the only available data, which had a bias toward major respiratory events leading to a forecast accuracy exceeding 90%. When a systematic program of twice-daily observations from lifeguards was implemented, the forecast could be meaningfully assessed. The results show that the forecasts identify the occurrence of respiratory events at all lifeguard beaches 70% of the time. However, a high rate (80%) of false positive forecasts occurred at any given beach. As the forecasts were made at half to whole county level, the resolution of the validation data was reduced to county level, reducing false positives to 22% (accuracy of 78%). The study indicates the importance of systematic sampling, even when using qualitative descriptors, the use of validation resolution to evaluate forecast capabilities, and the need to match forecast and validation resolutions.

Keywords

Forecasts; Harmful algal bloom; *Karenia brevis*; Model; Red tide; Skill assessment; Transport; Florida; Remote sensing; Gulf of Mexico

[*]Corresponding author. Tel.: +1 301 713 3028x173. Richard.stumpf@noaa.gov (R.P. Stumpf).

1. Introduction

Harmful algal blooms (HABs), sometimes called “red tides”, pose significant hazards due to their production of toxins and/or negative ecological impacts (e.g., high biomass associated anoxia). In the case of the dinoflagellate *Karenia brevis*, which produces brevetoxin, the impacts include Neurotoxic Shellfish Poisoning (NSP), fish kills, marine mammal deaths, and health and economic losses resulting from respiratory irritation. Surface blooms of *K. brevis* produce an aerosolized toxin as a result of cell lysis due to wind and wave action. Onshore winds are then responsible for transporting the aerosol onto local beaches, which can induce respiratory irritation (coughing, nasal congestion, and throat irritation) in beachgoers and coastal residents. Asthmatics exposed to *K. brevis* aerosols during a one-hour walk have measurable changes in both symptoms and spirometry (a measure of lung function) (Fleming et al., 2005, 2007). In this same asthmatic cohort, symptoms persisted for 5 days after the one-hour exposure (Kirkpatrick et al., in press). The public health and associated economic impacts have been a major concern for the state of Florida, as it has annual blooms of *K. brevis* off its gulfside beaches, lasting months to over a year in duration.

Due to the recurrent problems associated with *K. brevis* blooms off Florida, a forecast system was developed through collaboration between the National Oceanic and Atmospheric Administration (NOAA) and the state of Florida (Fig. 1), with the forecast subsequently expanded to other Gulf states. Current bloom locations, future bloom locations, and areas of impacts are critical components of these forecasts. The concepts behind the forecast system for *K. brevis* have been presented in several papers, which include a combination of satellite imagery (Stumpf et al., 2003;Tomlinson et al., 2004;Wynne et al., 2005) and heuristic and numerical models (Tester et al., 1991;Lanerolle et al., 2006;Stumpf et al., 2008).

Most operational oceanographic forecasts are physical and involve water level and current predictions, while operational biological forecasts in the ocean are uncommon. Ecological forecasts, in general, are difficult because uncertainty and inherent stochasticity in the data, system, and models, lead to low information content from the forecasts (Clark et al., 2001). Biological forecasts outside the research realm are more limited. Some common examples involve predicting hypoxia, such as the annual forecasts of the Gulf of Mexico hypoxia zone (Scavia et al., 2003). Such models are effective because they are dependent on only a few inputs, such as nutrient loads and mixing. Other examples of ecological models include annual fishery yields (Scheuerell and Williams, 2005). Examples of the few real-time event based predictive models include: the occurrence of sea nettles in the Chesapeake Bay (Decker et al., 2007) which is based on the organisms’ preference for specific combinations of temperature and salinity as predicted through an existing hydrodynamic model; and coral bleaching which depends on anomalously high sea surface temperatures (SST) which can be found from satellite (Goreau and Hayes, 2005). A probabilistic approach developed for several HABs in Europe is the Harmful Algal Blooms Expert System (HABES). HABES has developed HAB predictions through the use of fuzzy logic (Blauw et al., 2006). Predictive capabilities are being developed for several areas which may lead to forecasts of the annual start or temporal occurrence of HAB impacts (McGillicuddy et al., 2005). Annual forecasts are comparatively simple to validate, however, the field logistics for event forecasts may be daunting due to problems associated with reducing the observational errors in order to achieve a robust estimate of the actual conditions.

For physical factors, validation can be made through the comparison of model results with time-series data provided by tide stations, current meters, or other moored instruments (Hess et al., 2003). For biological models, such continuous time-series are often unavailable for

validation and therefore standard time-series or statistical techniques are more difficult to apply. Water quality sampling schemes typically involve water samples collected at a few fixed locations at weekly or longer intervals, with additional stations sampled opportunistically. For algal blooms, validation data are often provided by discrete sampling efforts with low spatial and temporal resolution. The spatial distributions of HABs, in addition, are irregular and highly patchy, often with no spatial coherence beyond the limits of the bloom—often only a few tens of kilometers. As a result, one mooring or sampling site may not capture any information representative of the forecasted event.

Another difficulty in HAB forecasting is the enigmatic nature of bloom initiation. In the case of *K. brevis*, early warning is limited to locating the presence of a bloom offshore and predicting landfall. Locating a surface bloom of *K. brevis* can be accomplished via ocean color remote sensing. Satellite imagery, however, is limited by clouds and spatial resolution, as well as uncertainties in algorithms, which are not species-specific. Cell counts are required for species validation. While identification of *K. brevis* blooms from satellite imagery can reach 80% over certain time periods (Tomlinson et al., 2004), validation of the extent of blooms delineated from satellite imagery has not been conducted because the necessary in *situ* observations have rarely been collected. In addition, *K. brevis* blooms initiate as subsurface planktonic blooms, and hence new blooms are not readily located. Unfortunately, cell count measurements offshore are limited in availability. An exception exists for encysting cells, like *Alexandrium fundyense* in the Gulf of Maine. The location of the cyst bed prior to the bloom season provides a good estimate of the location for bloom initiation (McGillicuddy et al., 2005).

Validation and skill assessment are vital to forecasting, not only to determine model behavior, but also to identify needed improvements. This paper will examine a skill assessment of the forecasts made by an operational system (the HAB Forecast System in the eastern Gulf of Mexico) with available data. The results will examine the influence of characteristics of the validation data on the skill assessment, as well as issues in examining nominal and ordinal (i.e., non-quantitative) forecasts. In addition, the analysis will examine how variations in resolution and quality of both the forecast and validation data can influence skill assessment. Through this analysis, strengths and limitations of the current forecast system were determined and will be discussed in terms of validation methods.

2. Methods

Before discussing the model assessment methods applied to the HAB forecast system, a brief description of the forecasts and models used to produce them is necessary. The forecasts include a nowcast prediction of a new *K. brevis* bloom, followed by forecasts of intensification, transport, aerial extent and impact of an existing bloom at the coast. Of these, the most complex is the nowcast prediction, and the most significant, from a user perspective, is the (respiratory) impact at the coast.

2.1. The forecast system

2.1.1. Nowcast and identification—The nowcast prediction uses a heuristic model that depends on cell counts and ocean color satellite imagery, primarily from the Sea-Viewing Wide Field-of-view Sensor (SeaWiFS). SeaWiFS has a 1.1 km^2 pixel at nadir and is mapped at that resolution. Chlorophyll and chlorophyll anomaly products are generated according to the methods described by Tomlinson et al. (2004) and (Stumpf et al. 2000, 2003). The anomalies indicate new blooms (as well as the movement or change in extent of a bloom) making it an appropriate primary indicator for a true bloom-forming organism, such as *K. brevis*, that dominates the biomass during summer and fall (Vargo et al., 1987). Since the anomaly only highlights areas where chlorophyll concentration has increased, and is not

species-specific, the model uses a series of rules, summarized in Table 1, based on the knowledge of the ecology of *K. brevis* and the oceanography of the west Florida shelf. By this rule-based model, we determine whether an anomalous patch of increased chlorophyll is likely to be a new *K. brevis* bloom and then proceed to delineate the bloom extent from the anomaly and field measurements. An example of a delineated bloom is shown in Fig. 2, with red indicating a likely or confirmed *K. brevis* bloom, and yellow indicating blooms of other organisms. Once a bloom is established at the coast, a series of forecasts are produced biweekly. These include forecasts on intensification, transport, aerial extent, and beach impact.

2.1.2. Intensification—Intensification is defined as an increase in cell concentration to a higher level and is determined by wind speed and direction (with upwelling favoring intensification) and cell concentration at the coast (Table 2). Forecasted wind conditions from the Marine Weather Forecasts are used according to Lanerolle et al. (2006) and Stumpf et al. (2008).

2.1.3. Transport—Transport is determined from the predicted wind speed and direction. The present model uses a transport of 7% of the (Ekman adjusted) alongcoast wind vector to estimate the alongshore transport of these blooms, using a regional tuning based on the results of Tester et al. (1991) and Stumpf et al. (2003). An alongshore forecast is assumed to be along the coast, which is north/south for SW Florida and east/west for the northwest Florida coast (area of Cape San Blas). The magnitude of transport is not addressed unless a change in the bloom extent is predicted.

2.1.4. Extent and location—Extent is defined as the expansion of the bloom to new areas along the coast. Currently, forecasts are only produced by analysts at county to half-county level (nominally 30–60 km), therefore forecasts of extent describe whether the bloom is expected to expand into a new county (or portion of a county) to the north or south along Southwest Florida or to the east or west along the Florida Panhandle. The extent is closely linked to the transport and is forecasted in the same manner.

2.1.5. Impacts—Beach impacts are forecasts based on several factors: the transport of the bloom, the expected wind speed and direction, the concentration (cell counts) of the bloom at the coast and the location (proximity to shore) (Table 2). The most critical factors are presence of a bloom and wind direction. The current respiratory impact model was developed from the work of Milian et al. (2007) and is summarized in Table 3.

2.2. Validation data

The program is operational and not designed as an experiment, so validation is dependent on data collected by state monitoring programs, in which sampling varies in frequency and spatial resolution. Several types of data are used for validation. These are cell counts, satellite data on chlorophyll concentration, and respiratory irritation from both the lifeguard network and anecdotal reports (e.g., news media or personal communication).

2.2.1. Cell counts—Cell counts for total number of *K. brevis* are made from microscopic analysis of water samples overseen by the Florida Fish and Wildlife Research Institute (FWRI, 2008). Water samples are collected during an event, most often in the area having a potential bloom. Preemptive and monitoring sampling has also been made in areas adjacent to reported blooms or during seasons when blooms are expected. Samples are collected through state and volunteer monitoring programs, as well as research cruises. This information is required near shellfisheries during HABs and taken as deemed appropriate by state managers in other areas. The cell counts are grouped into categories set by the state of

Florida, to reduce uncertainty in actual cell count accuracy. These data are used for assessing transport and HAB presence.

The distribution of cell count samples from October 1, 2004 through February 28, 2007 was investigated to determine the frequency by which samples were collected along the segments of the coast (Fig. 1). Only samples taken in the Gulf of Mexico within three miles of the coast were used (small circles in Fig. 1). These samples capture the areas that would impact the beach. The coast of interest was defined by county boundaries including Pinellas in the north and Collier in the south (Fig. 1). Many of the sampling programs are supported by counties and this region is the normal area of bloom impact. Cell count samples were grouped by county and by equidistant portions of the coastline. Of the 305 km section of coast (length of the heavy black line along the coast in Fig. 1), five segments were chosen, each being 61 km long (alternate solid-open circles in Fig. 1 distinguish the segments The segments normalize effort to coastline, but the county aggregations identify variations that may be driven by sampling effort.

2.2.2. Satellite chlorophyll—Chlorophyll fields from SeaWiFS are used to identify transport and intensification. Within an anomaly that has been confirmed as a *K. brevis* bloom (Fig. 2), the chlorophyll concentration provides an estimate of *K. brevis* concentration (Tester et al., 1998). A change in chlorophyll concentration can then be used to estimate intensification (Stumpf et al., 2003).

2.2.3. Respiratory irritation—During the analysis of the first bloom season, respiratory irritation in most areas was identified from anecdotal verbal reports in county or state bulletins and the media, the same method as conducted by Fisher et al. (2006). In general, only the presence of respiratory irritation was reported, so a forecast of no irritation could not be validated. Starting in August 2006, the professional lifeguard corps in Sarasota County began twice-daily reports (approx. 10:00 and 15:00 local time) of the presence of respiratory irritation at six sites. In January 2007, two additional lifeguard sites were added in Manatee County (Fig. 1). Respiratory irritation is defined by the amount of coughing observed in addition to the personal conditions experienced by the lifeguard. The presence of people coughing is used as a proxy for respiratory irritation (cough, nasal congestion, throat irritation, chest tightness, wheezing, and shortness of breath). Coughing has been documented as a response to *K. brevis* aerosols in studies involving occupationally exposed workers, recreationally exposed beachgoers, and asthmatics (Backer et al., 2003, 2005; Fleming et al., 2005, 2007). Lifeguards are asked to ‘listen’ to the beachgoers for the presence and/or frequency of coughing. The symptoms observed by the lifeguards are reported at various levels of respiratory irritation as shown in Table 3. Besides the respiratory impact, the lifeguards also collect data on the surf condition, water clarity, presence of dead fish, and approximate wind direction. The lifeguards have four choices when recording HAB respiratory impact: none, slight, moderate, or high. We grouped moderate and high classes together, as these were the level at which impacts affect the general public. This data set was additionally used to examine the roles of resolution in the forecasts and the skill assessments.

2.2.4. Winds—As forecasted winds are obtained from other forecast systems and are critical to the impact forecasts, we performed an assessment of the forecasted winds against *in situ* standard meteorological wind records. This provided an understanding of the importance of the external forecasts to the forecasts of this system. The HAB bulletins are issued with a twice-daily wind forecast, which is adapted from the National Weather Service (NWS) marine forecast, and reports wind direction by either semi-octants (N, NNE, NE etc) or onshore/offshore. The National Data Buoy Center (NDBC) Coastal-Marine Automated Network (C-MAN) station at Venice Pier (station VENF1) records standard hourly

meteorological wind measurements with wind directions between 0 and 360°. To directly compare the datasets, the buoy winds were ranked into sixteen semi-octants with each direction occupying 22.5° on a wind rose, with the north semi-octant centered at 0° (348.75° to 11.25°). For the west coast of Florida, onshore winds were defined as winds blowing from 168.75° to 326.25° (S clock-wise to NW, Stumpf et al., 2003), which presumes the 330° to 150° orientation of the coast. We used the buoy data collected at 6:00 and 18:00 local time (Eastern Time Zone), to validate the twice-daily marine forecasts. When, on a given day, only one daily wind direction forecast was entered into the bulletin the same forecast was applied to both the morning and evening *in situ* data.

2.3. Skill assessment

The assessed skill depends on the accuracy, a measure of the agreement between the model prediction and truth, and precision, which is a measure of the variance of the prediction due to observational errors (Lynch et al., 2009-this issue). Therefore, in the determination of precision, we examined the change in misfit with change in resolution of both the forecast data and the observational data. To assess the nature in which the characteristics of the observational data drive the assessment of skill through model–data misfit, this study also compares skill assessment results for the first bloom season of the HAB forecast system with subsequent years following the availability of higher resolution lifeguard data (as described above).

Most of the forecasts involve categories comprising ordinal or nominal values (“high;”, “medium”, and “low”; “presence/absence”), and require non-parametric statistics. Accordingly, skill was determined either by percent correct (total accuracy) or through determination of user and producer accuracy (Story and Congalton, 1986). *User accuracy* is the percent ratio of correct forecasts of a specified condition to the total number of *forecasts* of that condition. Commission or user error (100%—user accuracy) indicates false positives, i.e., the forecasted condition was not observed. It is termed user error because this is the error the user sees—whether the forecast is right or wrong. *Producer accuracy* is the percent ratio of correct forecasts of a specified condition to the total number of *observations* of that condition. Omission or producer error (100%—producer accuracy) indicates the rate of false negatives, i.e., the specified condition was observed but was not forecast. Commission and omission errors are sometimes called Type 1 and Type 2 errors, respectively, drawing from the terminology of hypothesis testing.

The first skill assessment of the operational forecast system was performed for the period from 1 October, 2004 to 30 September, 2005 by Fisher et al. (2006). The Forecast program made a subsequent assessment, with equivalent methods, for the entire first bloom season through April 30, 2006, which included 193 bulletins with forecasts. That result is reported here. The total analysis in this paper includes findings for the first bloom season with subsequent analysis through February, 2007. The assessment measures the accuracy in which the operational system identified new blooms and their location and extent. Each forecast component (identification, transport, intensification, extent, and impact) was compared to all available data and information and marked as “confirmed true” or “confirmed false”. When deemed impossible to evaluate as a result of insufficient data, the forecast component was declared “unconfirmed”. Several factors influence the forecast and forecast skill: 1) uncertainty of the bloom location; 2) uncertainty and resolution of the forecast models; and 3) uncertainty, resolution, and completeness of the assessment validation data. These are discussed in Section 3.

3. Results

3.1. Nowcasts/identification

Initial identification of a bloom as a HAB is the most objective to validate. During the first operational year, six *K. brevis* HABs were identified and tracked, of which four were first identified through the forecast system and validated by sampling efforts. The remaining two were first identified through state sampling efforts and then tracked through the forecast system. In addition, four non-harmful blooms were accurately identified through the system and confirmed by state sampling. No false positive forecasts were identified. For HABs, the user accuracy was 100%, and the producer accuracy was 67%, with a corresponding 33% omission error owing to the two blooms found first in state sampling. The total accuracy was 80% (Table 4).

3.2. Forecasts

An analysis was performed for each forecast made within the bulletins. To reiterate, forecasts were made on the intensification, spatial extent, transport and impact. A summary of these results for the first bloom season is shown in Table 4. Overall, the transport and intensification forecasts have higher rates of assessability owing to the use of satellite imagery in combination with the field data.

3.2.1. Intensification—For the first operational bloom season, only 38% of the bulletins contained forecasts of intensification. Of these forecasts, 63% were assessable with 73% accuracy (Table 4). Intensification was based on satellite derived chlorophyll when blooms were considered predominately *K. brevis*. Insufficient field data (discussed in Section 4.1 below) exists to regularly forecast intensification.

3.2.2. Transport—Forecasts of transport were made in 83% of the bulletins. A 90% accuracy was calculated where 67% of the forecasts were assessable (Table 4). Most transport forecasts pertain to direction, with magnitude addressed only on rare occasions when a bloom will cross a county.

3.2.3. Extent—Spatial extent was forecasted the least, present in only 21% of the bulletins (Table 4). 56% of these forecasts were assessable with 77% accuracy.

3.2.4. Impact forecasts—Initially, the impact forecast was made for a range of conditions, and therefore broad in both extent and magnitude of impact. For example, a forecast might be for low to high impact over several counties. Preliminary accuracy assessment performed by Fisher et al. (2006) indicated difficulty in assessing these forecasts, as most conditions would, and did, validate the forecast. A forecasted impact of “very low to high” could be validated with any impact report other than “none”, which is why it was discontinued. To correct this (an immediate result of the skill assessment of Fisher et al., 2006), in 2006, the forecasts became more specific in order to highlight only the maximum impact expected. While “patchiness” was frequently part of the condition, a patchy “high” impact forecast must have at least one high value reported within the forecast region to be validated.

Based on anecdotal information from the local and state constituents, only 49% of the impact forecasts were assessable during the first bloom season (Table 4). During this time, 95% of the bulletins contained forecasts of coastal impacts, with 99% accuracy and 98% accuracy in which impacts were predicted to be moderate or high.

Unlike the anecdotal information, the lifeguard observations dataset provided a continuous, twice-daily record in the two counties: Sarasota (September 2006–March 2007) and Manatee (January 2007–March 2007). There were 80 lifeguard reports per week on average, with between zero and 40 reports per week of moderate or high respiratory impact (Fig. 3)to compare with the forecasts. This comparison also involves the more precise forecasts used starting in 2006. The continuous data allowed determination of both the user and producer accuracies. The accuracy changed distinctly from the first year analysis. In comparison with each lifeguard location, only 21% of the 567 "moderate–high" impact predictions were correct, the rest being "false positives" (Table 5). In contrast, 68% of "moderate–high" impacts observed were correctly forecasted, indicating a tendency for 32% false negatives (Table 5). This shows a strong tendency toward false positives, rather than false negatives.

3.3. Resolution of the validation data

The validity of extent and transport forecasts is influenced by the distribution of samples. This was investigated to better understand the frequency and spatial distribution of sampling efforts, as these datasets are essential for assessing forecasts of transport and extent. Observations were binned by the five equal-length (61 km) segments along the coast (Figs. 1 and 4). There is considerable variability in sampling between segments. Sampling depends on county and local departments and volunteers, which is not consistent. Sampling frequency also varied between years depending on timing and impact of the HABs. Overall, the median number of samples along this coast was 26 $week^{-1}$, which equates to less than 1 sample every 75 km of coast for each day (75 km d^{-1}) (Fig. 4, Table 6). Over the two years, 25% of the sampling frequency fell at zero for several segments, indicating that no samples were taken in that segment or county at least 25% of the time. For 75% of the time, all segments had 10 or less samples $week^{-1}$. With the regions, the most intense sampling occurred in Region 3, where the median resolution within a week in 2006 was 26 km d^{-1} and only 25% of the samples exceeded 18 $week^{-1}$ or 31 km d^{-1} (Fig. 4). Substantial differences in sampling occurred between 2005 and 2006 (Fig. 4). This was driven strongly by differences in segments 2 and 3, which include Sarasota County. A median of 2 $week^{-1}$ was observed for 2005 in segment 3, followed by 15 $week^{-1}$ in 2006. Increased sampling in Sarasota and Charlotte Counties drove the median for the entire coast from 13 $week^{-1}$ in 2005 to 40 $week^{-1}$ in 2006 (Table 6).

3.4. Influence of wind forecasts

Because the HAB Forecasting System is heavily reliant on the marine wind forecast, an assessment of the accuracy of the wind forecast is useful in determining skill and cause of misfit in the impact forecasts. Over the entire time period assessed (2004–2007) there was the potential for 1673 forecasts. Out of these, 1439 forecasts were made that did not describe the winds as "variable" and had corresponding data collected from the VENF1 station. Out of the 1439 forecasts, 495 predicted onshore winds, and 60% of these were confirmed as correct by measured winds (at 0600 and 1800) at the VENF1 station. An assessment of the wind forecasts during the period for which lifeguard data was available was also made. The wind forecasts were compared with the observed winds at 0600 and 1800 local time on 173 days. Of these, 71% of the onshore wind forecasts were correct (Table 7). With 29% of onshore wind forecasts incorrect (i.e. actual wind direction was offshore), it follows that a forecast for HAB respiratory impact based on false wind predictions would result in a false positive. From the wind forecast, there is the potential for 29% false positives.

To assess the accuracy on a site by site case, a subsequent analysis comparing actual winds and lifeguard impact reports was performed. We examined the proportion of high–moderate forecasts which we confirmed both with and without accounting for false onshore wind forecasts. At this fine resolution we observed 21% and 22% correct forecasts in both sets of

conditions, or ~78% false positives (Table 8). The forecasts are county-wide and are not resolved to individual beaches. In order to assess the forecasts at the same resolution, the lifeguard observations were grouped so that if any lifeguard reported an impact, the forecast was classified as correct (Table 8, columns 3 and 4). This county-scale resolution increases the correct forecasts to 78% with only correct winds and 73% with all winds.

In addition, seabreeze is anecdotally considered to influence the impact, and was therefore investigated. We analyzed the frequency of moderate–high respiratory impact as a function of time of day (i.e. morning or afternoon). Fig. 5 summarizes directional data collected at 0600 and 1800 local time from the VENF1 station from September 2006 to March 2007. The strong tendency for afternoon onshore winds is clear. This trend taken with the evidence that moderate–high impacts are twice as common in the afternoon (Table 9) suggests the importance of incorporating the seabreeze into the forecasts.

4. Discussion

4.1. Extent and transport forecasts

The forecast of extent and transport is dependent on the resolution and uncertainty of the input spatial field. With the data types available, this is problematic. The most intense sampling occurred in segment 3 in 2006, although one of the worst events in 30 years occurred in 2005. Segment 3 had but 1 week with more than 30 samples collected—the 40 samples equate to 1 sample per 10 km d^{-1}. Resolutions of more than 20 samples were rarely achieved, equating to one sample per 20 km d^{-1}, the maximum precision resolvable (Fig. 4). The result indicates a resolution of extent at the coast of only 30 km, corresponding to a transport at the coast of 30 km d^{-1}. How does the satellite imagery help this? The satellite imagery generally cannot resolve features within ~2 km (2 pixels) of the coast or features smaller than ~10 km^2 (9 pixels). Also, the satellite has a positional uncertainty of one pixel (km) as well as uncertainties on the exact boundary of the HAB and, with clouds, a sampling frequency of 1–5 days. With consecutive days of high quality images, 5 km d^{-1} may be achieved, but more typically the satellite does not offer better than 10–50 km d^{-1}. With these uncertainties in resolution of both data sets used in validation (the water sample and imagery), only large HABs, covering > 10–30 km of coast, can be located and validated. In contrast, the lifeguard data provides a consistent resolution under 10 km d^{-1}, a significant improvement in sampling rate over the standard water samples (Fig. 3).

4.2. Impact forecasts

The skill assessment depends on the degree of uncertainty, the resolution of the forecasts, the quality, and the spatial/temporal resolution of the validation data. The accuracy assessment for the first bloom season showed high user accuracy (a correct forecast); for all forecasts it was typically >90%. The change to a higher (and appropriate) resolution for impact forecasts and the use of the unbiased lifeguard data, led to a much different accuracy: <20% for events with moderate or high impact. The change in accuracy results from several factors. The low resolution of the forecasted condition certainly caused part of the error, as the condition resolution was modified as a result of the Fisher et al. (2006) study. However, the validation data set may have also played a role. For the first year, 42% of the forecasts were not assessable. Of these, it is likely that a majority had low or no impact, although this cannot be confirmed. As there were no routine official reports of impact, many of the impacts were defined by verbal reports, through email, newspaper accounts, etc. Informal reporting on the coast tends to note when there is a problem, rather than when there is not. If so, then the assessable forecasts are biased toward days with an impact. This suggests that bias in the validation data must be examined to determine whether it produces large errors in the assessment of accuracy. With the lifeguard data, observations were taken twice each day

regardless of conditions (Fig. 3), removing sampling bias. The nearly continuous sampling should better identify false positives—forecasts of impact when none occurred. This is consistent with the observed high rate of false positives. These results show that even a model with inherently high point-to-point misfit, as seen in the impact forecasts of the first year, can be assessed to be quite accurate when predictions and observations are spatially and temporally aggregated. The resolution has to be considered in the analysis; increasing the forecast resolution and data resolution resulted in a lower accuracy in subsequent years.

Using the lifeguard data, we get a sense for the influence of resolution on misfits in the forecasts (Table 8). The countywide forecast was assessed against observations at individual beaches. When the resolution of the lifeguard data was reduced to the county level, the accuracy of the forecast increased. *K. brevis* blooms are considered to be patchy, so the forecasts often noted that the impacts would be patchy, e.g., a forecast would be for “patchy moderate” impacts in a county. The results actually validate the defined county level forecast. Forecasts of “patchy moderate to high impacts” within the county are correct 78% of the time. The question of whether the impact will occur at a specific location and time (morning or afternoon) is only 20% accurate. The difference confirms that patchiness does occur and that the model is both correct in identifying “patchiness” and that it is inadequate at this time for higher resolution forecasts. The resolution of both the forecast and the validation data constrains the results of the skill assessment. Haeffner (1996) noted that increased detail in modeling increases commission (user error), exactly as found here. The ultimate user application is the latter, so that accuracy is important, although not achievable at this time.

4.3. Winds

The presence of respiratory impacts at the shore is largely a function of cell count patchiness and wind direction. One cause for inaccurate impact forecasts could be sensitivity to seabreeze. Seabreezes are included in some of the coastal marine forecasts, but they are not included in standard modeled forecast winds. The seabreeze is considered in some of the HAB forecasts, when seabreeze is clearly identified in the marine forecast, but this requires additional analysis for routine implementation. The tendency for a strong onshore seabreeze in the afternoon along with the evidence of a higher frequency of moderate to high impacts in the afternoon, indicate the need for the consideration of seabreeze effects when forecasting *K. brevis* impact.

4.4. Assessment of forecast utility

Finally, we should note another aspect of skill assessment that is vital to an operational system: the value of the forecast to the user community. The metrics may involve socio-economic analyses, but a minimum metric is whether the forecast is used by the target audience. Each week, analysts determined whether manager’s reports either referenced the forecasts or sampled in an area to verify a forecasted bloom. Between October 2004 and April 2006, the weekly bulletin information was used by managers 93% of the time. Bulletins were identified as high, medium or low priority based on the importance of the information they contained (high indicating that a management action is recommended). Of the 37 high priority bulletins that were released, 94% were utilized. Similar analyses would be critical as a part of the skill assessment used to maintain or improve an operational forecast system.

5. Conclusions

The assessment of skill of the operational forecasts depends on the ability of the model to forecast conditions at an appropriate resolution, within the constraints of the available

validation data. The availability of higher resolution validation data through the volunteer lifeguard program provided information necessary to identify limitations in our current *K. brevis* forecast of impacts. On the other hand, the assessment also indicates that limitations in the validation data will preclude evaluations of new models for transport or extent. As changes in both the validation system and forecast resolution are undertaken, consideration should be made as to whether comparisons can be made in an equivalent way as in previous years. Implementing or evaluating models at high resolution, such as the potential for 1-km resolution of HAB boundaries from satellite, may be unproductive without an appropriate validation system. A low resolution validation data set, however, may demonstrate the accuracy of the forecast model at larger scales than is desired for management purposes.

Anecdotal evidence suggests patchiness at a scale of a few kilometers. The difference in accuracy of the impacts between beach level and county level supports this evidence, indicating patchiness at scales of 10 km or finer. Data to resolve patchiness will be critical to providing more accurate forecasts at finer scales. The analysis performed here, using validation data of different resolution, further demonstrates how the assessment of skill changes depending on the combination of model resolution and validation data resolution. Highly resolved validation data sets can indicate limitations in a forecast, therefore indicating the limitations of the model at particular scales. Once the models achieve 10 km resolution, no further improvement will be quantifiable without improvements in the resolution of the validation data. The investment in validation should keep pace with the investment in improvements in the model. Ultimately the results need to be validated and maintained in a way to be usable and applicable to managers and the public.

Acknowledgments

We would like to thank the HAB analyst team (Allison Allen, Cristina Urizar, Zack Bronder, Lori Fenstermacher, Heidi Keller and Mark Vincent) at the NOAA Center for Operational Oceanographic Products and Services (CO-OPS) for producing the forecasts and initial assessment that made this analysis possible. We also acknowledge the Sarasota County Beach Patrol and the Manatee County Department of Public Safety, Marine Rescue Division and all of the beach lifeguards responsible for inputting data into the Beach Conditions Reporting System. Cell count data was provided courtesy of Cindy Heil and Earnest Truby of FWRI. Finally, we would like to thank the reviewers for their extensive review, which significantly improved the quality of this manuscript. This project was partially supported by the Centers for Disease Control, Center for Environmental Health, Agreement 07FED713955.

References

Backer LC, Fleming LE, Rowan A, Cheng Y, Benson J, Pierce RH, Zaias J, Bean J, Bossart GD, Johnson D, Quimbo R, Baden DG. Recreational exposure to aerosolized brevetoxins during Florida red tide events. Harmful Algae 2003;2:19–28. [PubMed: 19081765]

Backer LC, Kirkpatrick B, Fleming LE, Cheng YS, Pierce R, Bean JA, Clark R, Johnson D, Wanner A, Tamer R, Zhou Y, Baden DG. Occupational exposure to aerosolized brevetoxins during Florida red tide events: impacts on a healthy worker population. Environ Health Perspectives 2005;113:644–649.

Blauw AN, Anderson P, Estrada M, Johansen M, Laanemets J, Peperzak L, Purdie D, Raine R, Vahtera E. The use of fuzzy logic models for data analysis and modelling of European harmful algal blooms: results of the HABES project. African Journal of Marine Science 2006;28(2):365–369.

Clark JS, Carpenter SR, Barber M, Collins S, Dobson A, Foley JA, Lodge DM, Pascual M, Pielke R Jr, Pier W, Pringle C, Reid WV. Ecological forecasts: an emerging imperative. Science 2001;293:657–660. [PubMed: 11474103]

Decker MB, Brown CW, Hood RR, Purcell JE, Gross TF, Matanoski JC, Bannon RO, Setzler-Hamilton EM. Predicting the distribution of scyphomedusa *Chrysora quinquecirrha* in Chesapeake Bay. Marine Ecology Progress Series 2007;329:99–113.

Fisher KM, Allen AL, Keller HM, Bronder ZE, Fenstermacher LE, Vincent MS. Annual report of the Gulf of Mexico Harmful Algal Bloom Operational Forecast System (GOM HAB-OFS). NOAA Tech. Rep. 2006 NOS CO-OPS 047.

Fleming LE, Kirkpatrick B, Backer LC, Bean JA, Wanner A, Dalpra D, Tamer R, Zaias J, Cheng YS, Pierce R, Naar J, Abraham W, Clark R, Zhou Y, Henry MS, Johnson D, Van De Bogart G, Bossart GD, Harrington M, Baden DG, Fleming L. Initial evaluation of the effects of aerosolized Florida red tide toxins (brevetoxins) in persons with asthma. 2005. Environmental Health Perspectives 2005;113:650–657. [PubMed: 15866779]

Fleming LE, Kirkpatrick B, Backer LC, Bean JA, Wanner A, Reich A, Dalpra D, Zaias J, Cheng YS, Pierce R, Naar J, Abraham WM, Baden DG. Aerosolized red-tide toxins (brevetoxins) and asthma. Chest 2007;131(1):187–194. [PubMed: 17218574]

FWRI. Florida Fish and Wildlife Research Institute. 2008. http://research.myfwc.com

Goreau TJ, Hayes RL. Monitoring and calibrating sea surface temperature anomalies with satellite and in-situ data to study effects of weather extremes and climate changes on coral reefs. World Resource Review 2005;17(2):243–253.

Haeffner, JW. Modeling Biological Systems: Principles and Applications. New York: Chapman and Hall; 1996. p. 473

Hess, K.; Gross, TF.; Schmalz, RA.; Kelley, JGW.; Aikman, F., III; Wei, E.; Vincent, MS. NOS standards for evaluating operational nowcast and forecast hydrodynamic model systems. Silver Spring, MD: NOAA Technical Report NOS CS 17; 2003.

Kirkpatrick, B.; Bean, J.; Fleming, LE.; Backer, LC.; Akers, R.; Wanner, A.; Dalpra, D.; Nierenberg, K.; Reich, A.; Baden, D. Aerosolized red tide toxins (brevetoxins) and asthma: a 10 day follow up after 1 hour acute beach exposure. In: Moestrup, et al., editors. Proceedings of the 12th International Conference on Harmful Algae; International Society for Harmful Algae and Intergovernmental Oceanographic Commission of UNESCO; in press

Lanerolle LWJ, Tomlinson MC, Gross TF, Aikman F III, Stumpf RP, Kirkpatrick GJ, Pederson BA. Numerical investigation of the effects of upwelling on harmful algal blooms off the west Florida coast. Estuarine, Coastal and Shelf Science 2006;70:599–612.

Lynch DR, McGillicuddy Jr. DJ, Werner FE. Preface: Skill assessment for coupled biological/physical models of marine systems. Journal of Marine Systems 2009;76(1–3) (this issue) doi:10.1016/j.jmarsys.2008.05.002.

McGillicuddy DJ Jr, Anderson DM, Lynch DR, Townsend DW. Mechanisms regulating large-scale seasonal fluctuations in *Alexandrium fundyense* populations in the Gulf of Maine: results from a physical–biological model. Deep Sea Research II 2005;52:2698–2714.

Milian A, Nierenberg K, Fleming LE, Bean JA, Wanner A, Reich A, Backer LC, Jayroe D, Kirkpatrick B. Reported Asthma Symptom Intensity during Exposure to Aerosolized Florida Red Tide Toxins. Journal of Asthma 2007;44:583–587. [PubMed: 17885863]

Scavia D, Rabalais NN, Turner RE, Justic D, Wiseman WJ Jr. Predicting the response of Gulf of Mexico hypoxia to variations in Mississippi River nitrogen load. Limnology and Oceanography 2003;48(3):951–956.

Scheuerell MD, Williams JG. Forecasting climate induced changes in the survival of Snake River spring/summer Chinook salmon *(Oncorhynchus tshawytscha)*. Fisheries Oceanography 2005;14(6):448–457.

Story M, Congalton RG. Accuracy assessment: a user's perspective. Photogrammetric Engineering and Remote Sensing 1986;52(3):397–399.

Stumpf, RP.; Arnone, RA.; Gould, RW.; Martinolich, P.; Ransibrahmanakul, V.; Tester, PA.; Steward, RG.; Subramaniam, A.; Culver, M.; Pennock, JR. SeaWiFS ocean color data for US Southeast coastal waters. Proceedings of the Sixth International Conference on Remote Sensing for Marine and Coastal Environments; Ann Arbor, MI, USA: Veridian ERIM Intl.; 2000. p. 25-27.

Stumpf RP, Culver ME, Tester PA, Tomlinson M, Kirkpatrick GJ, Pederson BA, Truby E, Ransibrahmanakul V, Soracco M. Monitoring *Karenia brevis* blooms in the Gulf of Mexico using satellite ocean color imagery and other data. Harmful Algae 2003;2:147–160.

Stumpf RP, Litaker RW, Lanerolle L, Tester PA. Hydrodynamic accumulation of *Karenia* off the west coast of Florida. Continental Shelf Research 2008;28:189–213.

Tester PA, Stumpf RP, Vukovich FM, Fowler PK, Turner JT. An expatriate red tide bloom: transport, distribution, and persistence. Limnology and Oceanography 1991;36(5):1053–2061.

Tester, PA.; Stumpf, RP.; Steidinger, KA. Ocean color imagery: what is the minimum detection level for *Gymnodinium breve* blooms?. In: Ruguera, B.; Blanco, J.; Fernandez, ML.; Wyatt, T., editors. Harmful Algae. Xunta de Galicia and Intergovernmental Oceanographic Commission of UNESCO; Paris, France. 1998. p. 149-151.

Tomlinson MC, Stumpf RP, Ransibrahmanakul V, Truby EW, Kirkpatrick GJ, Pederson BA, Vargo GA, Heil CA. Evaluation of the use of SeaWiFS imagery for detecting *Karenia brevis* harmful algal blooms in the eastern Gulf of Mexico. Remote Sensing of Environment 2004;91(3–4):293–303.

Vargo GA, Carder KL, Gregg W, Shanley E, Heil C, Steidinger KA, Haddad KD. The potential contribution of primary production by red tides to the west Florida shelf ecosystem. Limnology and Oceanography 1987;32:762–767.

Wynne TT, Stumpf RP, Tomlinson MC, Ransibrahmanakul V, Villareal TA. Detecting *Karenia brevis* blooms and algal resuspension in the western Gulf of Mexico with satellite ocean color imagery. Harmful Algae 2005;4:992–1006.

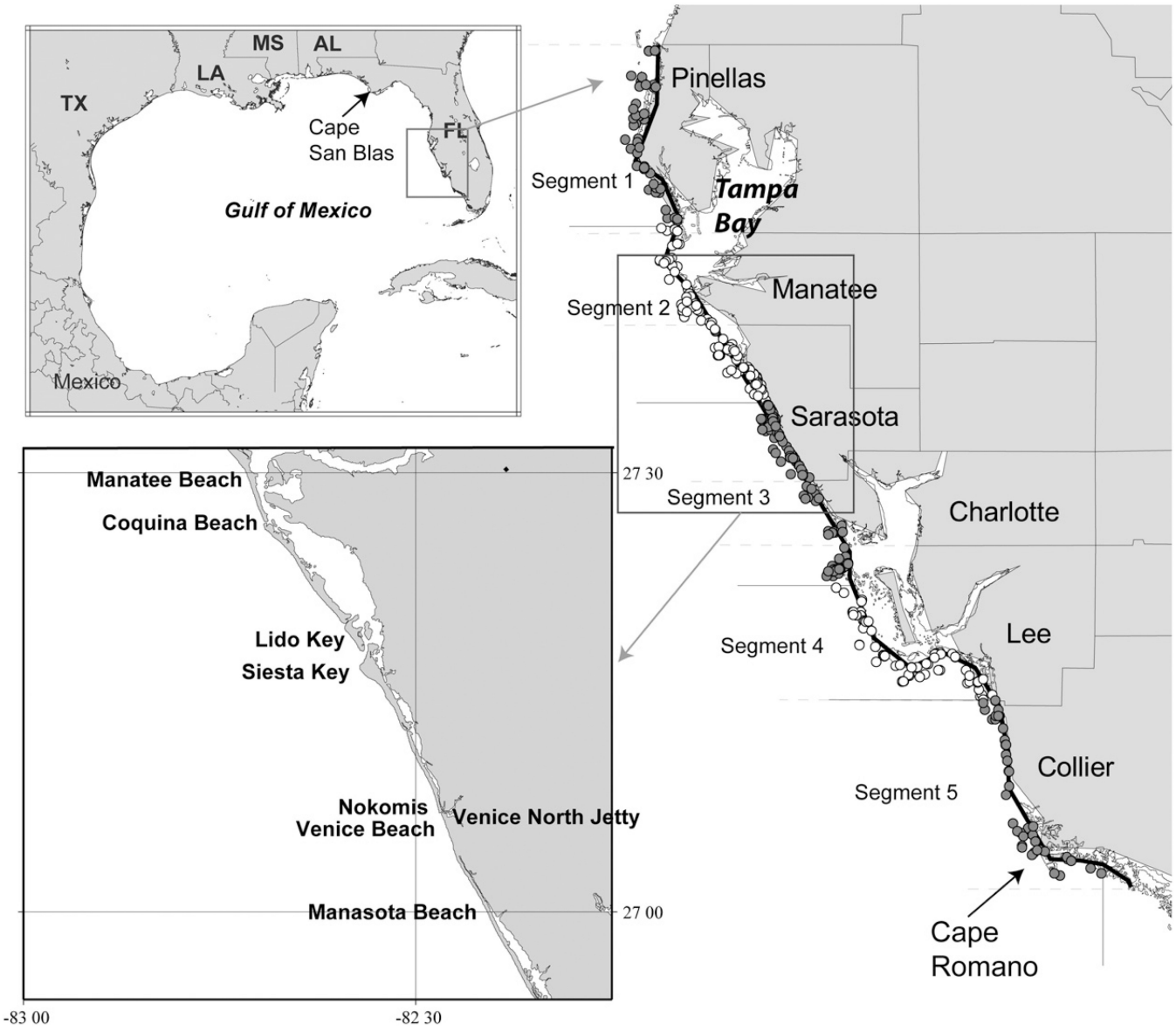

Fig. 1.

Map of the study area showing its location and the coastline from Pinellas County to Collier County, Florida, USA. Dashed lines show the extension of the county boundaries offshore, solid lines represent the boundaries for equidistant (61.4 km) segments along the coast (represented by solid black line). Circles represent the location of available *K. brevis* samples within 3 miles of the coast, from October 2004 to February 2007. These are coded dark or light to distinguish samples in adjacent equidistant segments. Inset shows the location of lifeguard beach stations.

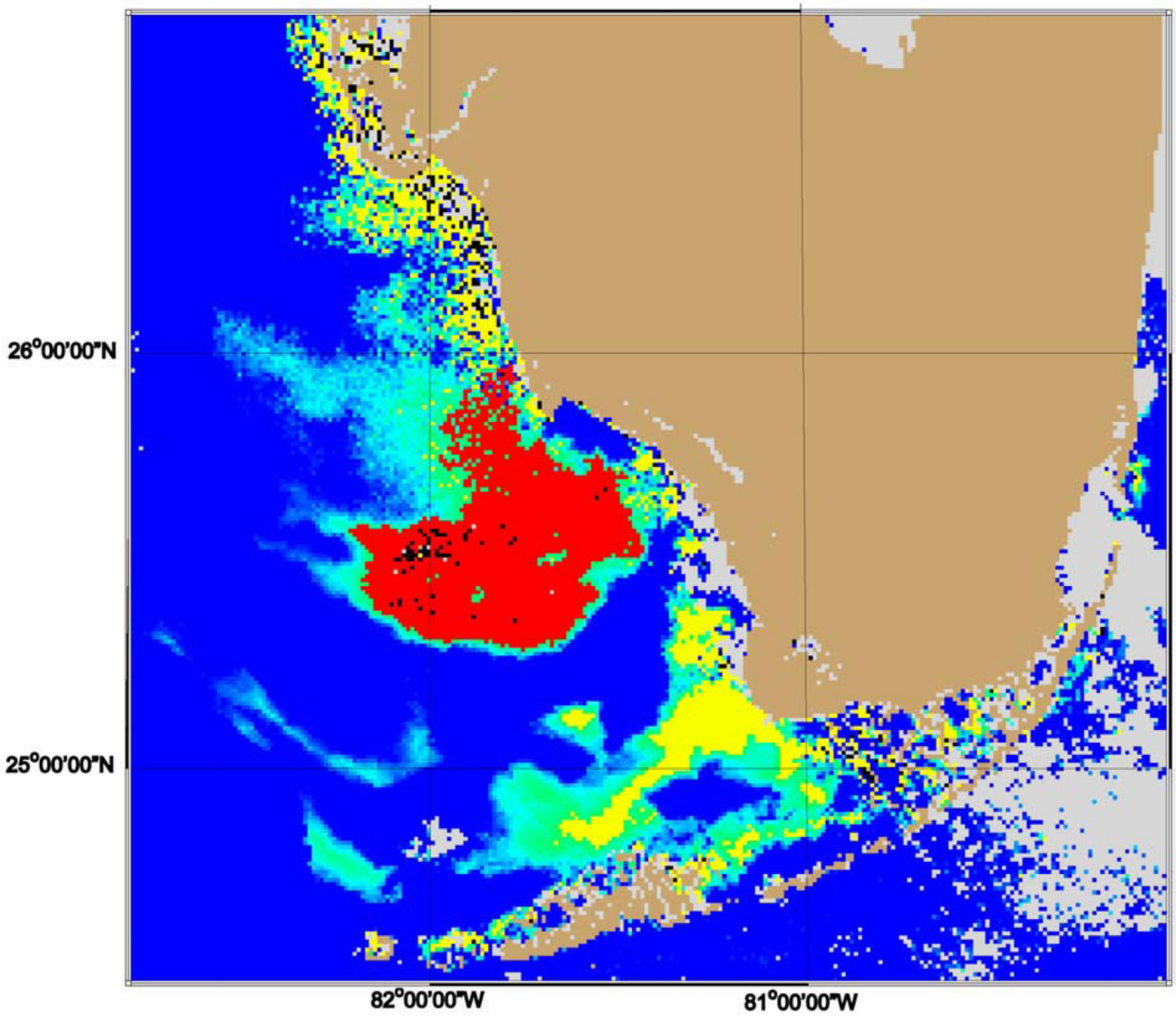

Fig. 2.

SeaWiFS satellite image from November 21, 2004. Yellow areas indicate where the chlorophyll anomaly based on Stumpf et al. (2003) exceeded 1 μgL^{-1} cyan/green show anomalies between 0 and 1, blue indicates no positive anomaly. Red represents locations of *K. brevis* blooms based on the criteria listed in Table 1. The yellow areas failed the criteria in Table 1 and are not considered to be due to *K. brevis*.

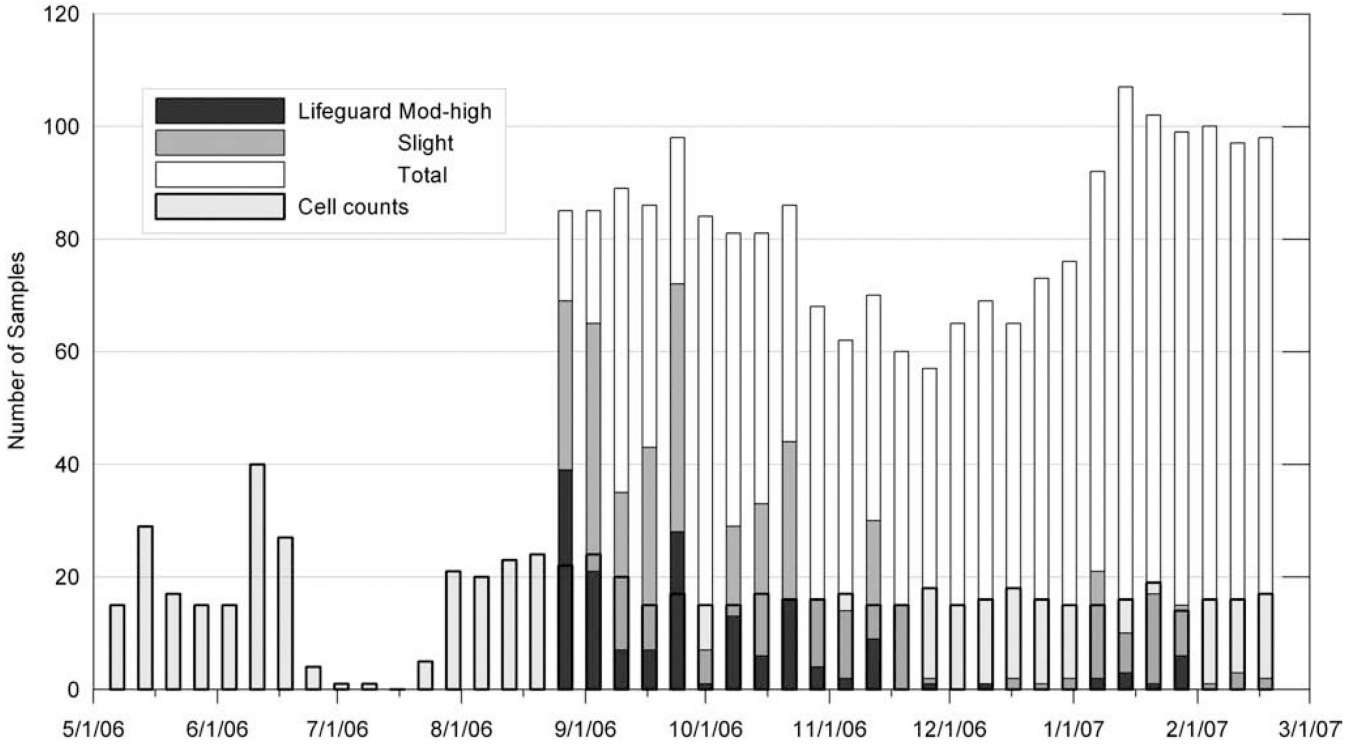

Fig. 3.

Comparison of the number of cell counts samples per week to lifeguard reports by week from Sept. 2006 to March 2007. The lifeguard reports also show the distribution of impacts, both the Slight impacts and the combined Moderate and High impact reports.

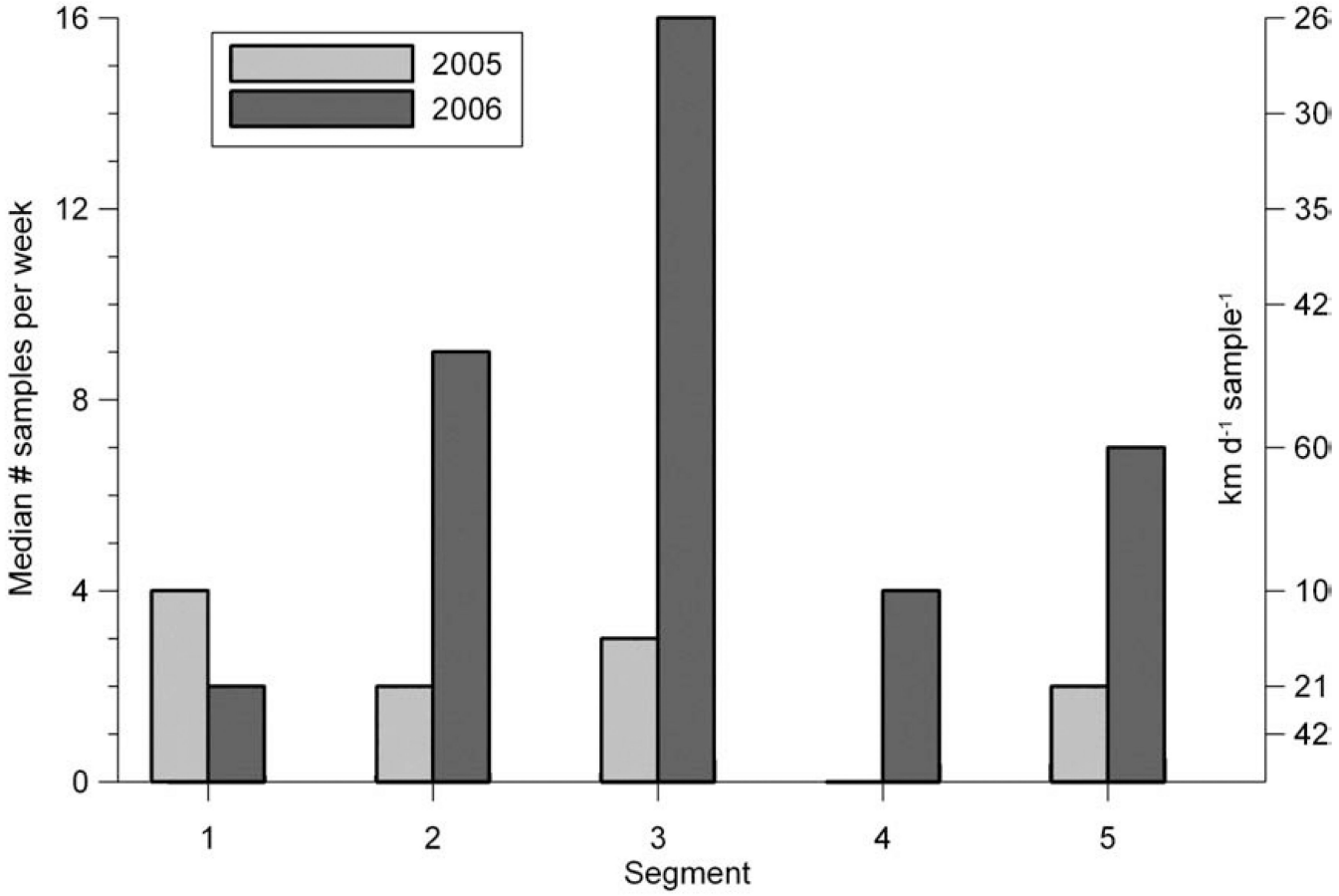

Fig. 4.

Median number of cell count samples per week by segment and the equivalent resolution in km d^{-1} per sample for transport (or km per sample per day for extent). Results shown for 2005 and 2006, and segments identified in Fig. 1.

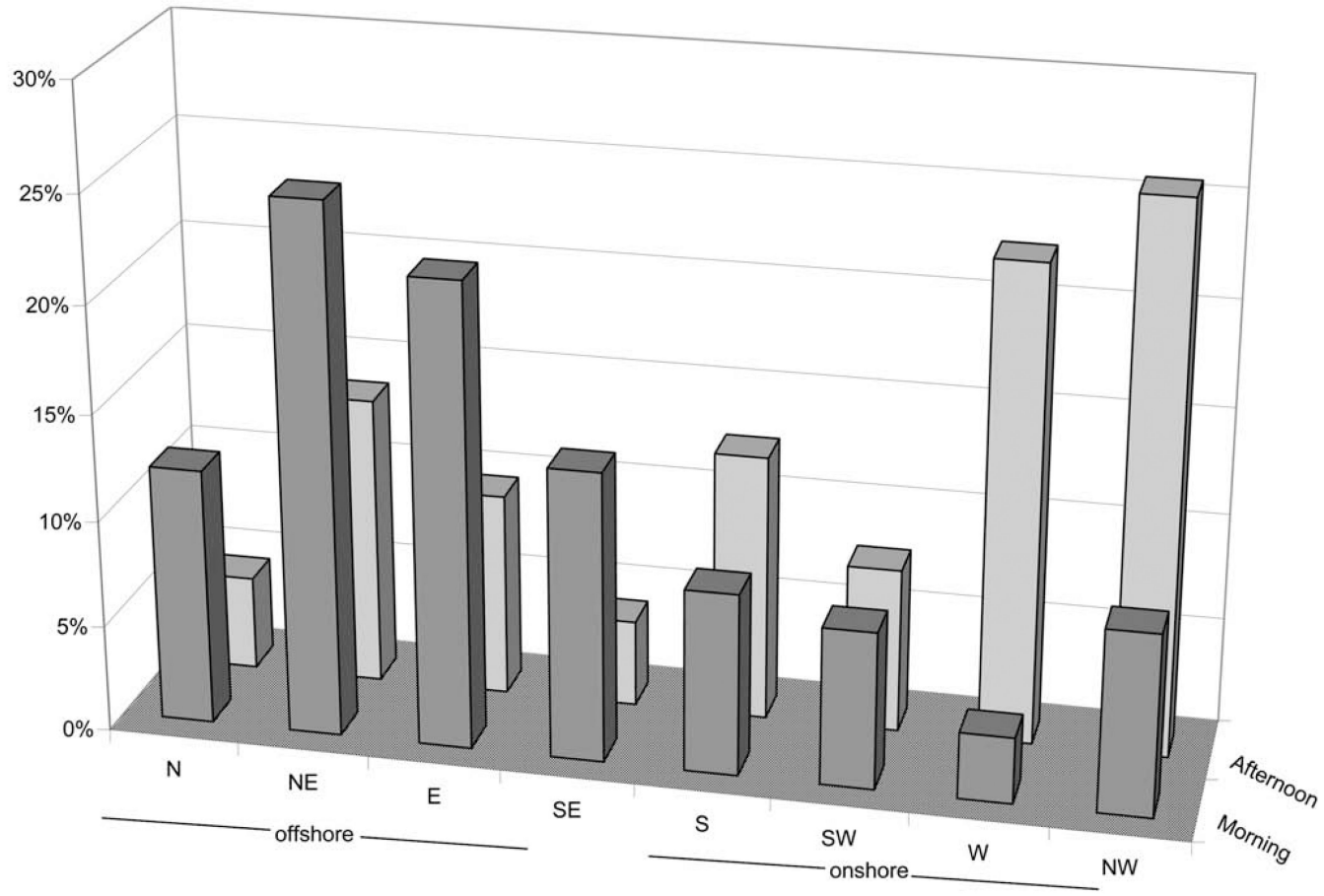

Fig. 5.
Distribution of octant wind directions as a function of time of day. Standard meteorological data from Venice Pier C-MAN Station (VENF1) from Sept. 2006 to March 2007.

Table 1

Nowcast/identification heuristic model

Chlorophyll anomaly	> 1 $\mu g L^{-1}$
Season	Aug–Jan (or during persistent HAB)
Geography	Pinellas to Collier Counties (unless know bloom is tracked outside this area)
Size	>30 km^2
Shape	Patch, not coast-wide
Upwelling/ winds	>20 km onshore transport
Respiratory	Impact reported with onshore winds
Cell counts	Used for subsequent confirmation

Table 2

K. brevis impact levels based on cell concentration and wind speed and direction, which are used to predict bloom intensification and impact

Concentration near coast (cells L^{-1})	**Offshore winds**	**Onshore light winds (≤3 m s^{-1}, ≤5 kts)**	**Onshore medium winds (3–8 m s^{-1}, 6–15 kts)**	**Onshore high winds (>8 m s^{-1}, >15 kts)**	**Expected impact**
None	None	None	None	None	None
Very low (< 10,000)	None	None	Very low	Very low	Beach impacts unlikely, exc. highly sensitive people (e.g. asthmatics)
Low(10,000–10^5)	Very low	Very low	Low	Low/moderate	Will not impact most people, shellfish harvesting closures likely.
Medium (10^5–10^6)	Very low	Low	Moderate	Moderate/high	Mild symptoms in beachgoers, presence of dead fish.
High (> 10^6)	Low	Moderate	High	High	Adverse respiratory symptoms in most people, discolored water, presence of dead fish, and shellfish harvesting closures.

K. brevis cell count levels are based on classes defined by the Florida Fish and Wildlife Research Institute.

J Mar Syst. Author manuscript; available in PMC 2010 July 12.

Table 3

Impact levels as defined by county lifeguard sampling program

Respiratory irritation report	In a 30 s sample
None	No coughing/sneezing heard in ~30 s
Slight	A few coughs/sneezes heard in ~30 s
Moderate	A cough/sneeze heard every ~5s
High	Coughing/sneezing heard almost continuously

Table 4

Forecasted bloom component accuracy and percent of assessable forecasts for period October 2004–April 2006

Forecast	Number of forecasts	Assessable (%)	Accuracy (%)
Identification	10	100	80
Transport	184	67	90
Extent	48	56	77
Intensification	84	63	73
Impact	209	49	99
Combined	525	58	89

Table 5

Comparison of lifeguard observations with HAB bulletin respiratory impact forecasts

		Lifeguard respiratory impact observations				
		High-moderate	**Slight**	**None**	**Total observations**	**User accuracy**
HAB bulletin respiratory	High-moderate	**116**	208	243	567	20.5%
impact	Low	29	**83**	355	467	17.8%
forecasts	None or very low	15	63	**673**	751	89.6%
	No forecast	13	36	280	329	
Total observations with a forecast		173	390	1551	*n*=2114	
Producer accuracy		67.6%	37.4%	61.4%		

The bulletin and lifeguard impact levels are defined in Tables 1 and 2, respectively. Bold on diagonal indicates correct forecasts.

J Mar Syst. Author manuscript; available in PMC 2010 July 12.

Table 6

Quartiles of samples week^{-1} for the coastline extending from Pinellas to Collier counties (Fig. 1)

Percentile	Entire time period		2005		2006	
	Samples week^{-1} for entire coast	Range of samples week^{-1} within segments	Samples week^{-1} for entire coast	Range of samples week^{-1} within segments	Samples week^{-1} for entire coast	Range of samples week^{-1} within segments
75%	39	2–8	24	1–6	48	2–18
50%	26	1–3	13	0.5–4	40	2–15
25%	11	0–2 0	0	0–0	28	0–10

Data from October 2004 to February 2007.

J Mar Syst. Author manuscript; available in PMC 2010 July 12.

Table 7

Comparison between observed and forecasted winds, from September 2006 to February 2007

Forecast/observed	Onshore	Offshore	N/A	Total	User accuracy
Onshore	61	25	0	86	70.9%
Offshore	44	168	3	215	78.1%
N/A	14	30	1	45	
Total	119	223	4	346	
Producer accuracy	51.3%	75.3%			

Table 8

A comparison of forecast user accuracy for only moderate–high reports using (1) all wind data and all lifeguard reports, (2) only correctly forecasted winds and all lifeguard reports, (3) all wind data and area level reports and (4) correctly forecasted winds and county level reports

	1	2	3	4
Forecast/ observed	**All winds/all reports (# reports)**	**Correct winds/ all reports (# reports)**	**All winds/ county level reports (# days)**	**Correct winds/ county level reports (# days)**
Forecast correct	116	115	33	32
Forecast total	567	521	45	41
Percent	20.5%	22.1%	73.3%	78%

Table 9

Morning and afternoon impacts

Samples/impact	No impact	Slight	Moderate–high
All (2114)	73.4% (1551)	18.4% (390)	8.2% (173)
Morning (1163)	78.8% (916)	15.4% (179)	5.8% (68)
Afternoon (951)	66.8% (635)	22.2% (211)	11.1% (105)