GIScience & Remote Sensing

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgrs20

Iterative spatial leave-one-out cross-validation and gap-filling based data augmentation for supervised learning applications in marine remote sensing

Andy Stock & Ajit Subramaniam

To cite this article: Andy Stock & Ajit Subramaniam (2022) Iterative spatial leave-one-out cross-validation and gap-filling based data augmentation for supervised learning applications in marine remote sensing, GIScience & Remote Sensing, 59:1, 1281-1300, DOI: 10.1080/15481603.2022.2107113

To link to this article: https://doi.org/10.1080/15481603.2022.2107113

Published online: 22 Aug 2022.

GISCIENCE & REMOTE SENSING
2022, VOL. 59, NO. 1, 1281–1300
https://doi.org/10.1080/15481603.2022.2107113

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Iterative spatial leave-one-out cross-validation and gap-filling based data augmentation for supervised learning applications in marine remote sensing

Andy Stock[a,b] and Ajit Subramaniam[a]

[a]Lamont-Doherty Earth Observatory, The Earth Institute, Columbia University, Palisades, New York, USA; [b]Institute for Resources, Environment and Sustainability, University of British Columbia, Vancouver, British Columbia, Canada

ABSTRACT

In marine remote sensing, supervised learning can link variables measured in-situ near the ocean surface to variables that can be measured from space. However, the in-situ data used for training and validating such empirical satellite algorithms are often spatially auto-correlated and clustered, giving rise to various statistical challenges such as overfitting to spatial structures. Furthermore, co-located in-situ and satellite measurements are rare in the oceans because of the cost of data collection from research vessels and frequent cloud cover. We propose two methods to mitigate these challenges. The first method builds on spatial leave-one-out cross-validation (SLOOCV), an approach designed to provide sound error estimates when data are spatially auto-correlated by enforcing a minimum separation distance between training and test observations. However, estimating this distance may be impossible with sparse and spatially clustered data. We hence propose to iterate and integrate error estimates over a range of separation distances (iSLOOCV). To address the often-small size of labeled data sets based on marine in-situ data, we tested if increasing the number of observations for algorithm training by means of cloud-filling algorithms for marine satellite data improved predictions. The potential of these two methods is demonstrated by developing empirical algorithms for mapping the proportions of seven diagnostic pigments (DPs) that serve as proxies for phytoplankton community composition in the northern Gulf of Mexico. We estimated the prediction accuracy of 13 algorithms with iSLOOCV, using various sets of satellite data products as input, and found adequate algorithms for 4 of the 7 DPs. Random forests combining ocean color and environmental variables as input had the lowest prediction errors overall. Correlations between predictions and observations estimated by iSLOOCV ranged from 0.69 to 0.85 and mean absolute errors from 0.02 to 0.13. Daily maps and longer-term composites of these DPs were broadly consistent with previously published results. Overall, errors increased when extrapolating over larger distances, highlighting how iSLOOCV can illuminate changes in algorithm performance based on sub-regional data coverage. Generating larger training sets by prior gap-filling substantially improved all error measures for 3 of the 7 DPs, with mixed results for the others. Therefore, data augmentation by gap-filling of satellite data should not be used as a default approach but can be a useful tool when supervised learning applications are suspected to be limited by the size of the training set.

ARTICLE HISTORY
Received 16 March 2022
Accepted 22 July 2022

KEYWORDS
Ocean color; phytoplankton; Gulf of Mexico; accuracy; autocorrelation; gap-filling

1. Introduction

Satellite remote sensing allows for mapping of physical and biological phenomena at high temporal resolution and broad spatial scales (Kerr and Ostrovsky 2003). In marine applications, supervised learning – from linear regression to deep neural networks – often serves to map variables measured in-situ based on variables that can be measured from space. Approaches used for this purpose include linear regression, generalized additive models, random forests, and neural networks (e.g. Chen et al. 2019; Doerffer and Schiller 2007; Hieronymi, Müller, and Doerffer 2017; Hu et al. 2018; Keiner and Brown 1999; Liu et al. 2021; O'Reilly et al. 1998; Stock 2015; Xi et al. 2020). However, the in-situ data used for training and validating such empirical satellite algorithms are rarely randomly distributed in space and time, creating statistical pitfalls for supervised learning (Stock 2022). For example, marine labeled data are often spatially autocorrelated and clustered, for example, along ship tracks and near phenomena of interest like river plumes. Consequently, the data may not be independent, a core assumption of standard approaches for the training and validation of supervised learning algorithms. Ignoring dependence structures when validating statistical models can lead to an underestimation of

CONTACT Andy Stock andy.stock@ubc.ca

prediction errors and to the selection of too flexible models (Roberts et al. 2017). These problems are exacerbated in marine research because in-situ data sets available for the oceans tend to be small, because data collection from research vessels is time-consuming and expensive, and because cloud cover often prevents matching in-situ measurements with co-located satellite observations. Beyond the statistical limitations arising from small data sets, frequent cloud cover may leave whole sub-regions of a study area and rare conditions underrepresented in the data, posing a major barrier to the training and validation of supervised-learning based models (Stock 2022).

The objective of this study is to address such challenges of small, spatially clustered and autocorrelated labeled data sets. For this purpose, we propose two new computational approaches focused on error estimation, model selection (the choice between different statistical models to minimize prediction errors), and data augmentation (increasing the amount of available data) by gap-filling. The first approach builds on spatial leave-one-out cross-validation (SLOOCV), a method for error estimation and model selection when data are spatially auto-correlated (Le Rest et al. 2014; Le Rest, Pinaud, and Bretagnolle 2013). However, the standard SLOOCV algorithm relies on the calculation of residual variograms, which can be misleading when data are spatially clustered (Bel et al. 2009) or when flexible machine learning models are used (Roberts et al. 2017). We therefore adjusted the standard SLOOCV algorithm to avoid the calculation of residual variograms by iterating over a range of distances (iterative, or iSLOOCV).

The second approach mitigates the typically small size and limited spatial coverage of marine labeled data sets. Because satellite data often have large gaps caused by clouds, the size of the available data set can be increased by means of gap-filling algorithms (e.g. Alvera-Azcárate et al. 2007; Barth et al. 2020; Hilborn and Costa 2018; Liu et al. 2019; Saulquin, Gohin, and Fanton D' Andon 2018; Stock et al. 2020). On the one hand, prior gap-filling could improve predictive models by creating many additional matchups for training. On the other hand, reconstructing pixel values where no satellite observations exist can introduce additional errors compared to direct satellite measurements. For example, phytoplankton communities inside and outside of mesoscale eddies can differ (Soja-Woźniak et al. 2020), and such differences can be obscured beyond reconstruction by high cloud cover lasting several days. It is not clear *a priori* if the advantages of a larger data set for model training would outweigh additional errors introduced by gap-filling. We hence tested if including additional in-situ observations matched with reconstructed satellite data in the training of empirical algorithms can improve their prediction accuracy.

This study demonstrates the potential of these new methods by mapping phytoplankton diagnostic pigments (DPs) that serve as biomarkers for different phytoplankton types in the northern Gulf of Mexico (NGOM). Chlorophyll *a* concentration, a proxy for phytoplankton biomass, is a widely available standard satellite data product (McClain 2009). However, phytoplankton primary production and carbon fixation – and hence, their biogeochemical and ecological functions – depend also on community composition (Chakraborty, Lohrenz, and Gundersen 2017; Quere et al. 2005). Researchers have already proposed many algorithms for satellite mapping of different aspects of phytoplankton community composition (IOCCG 2014; Mouw et al. 2017), including several algorithms for mapping DPs (e.g. Bracher et al. 2015b; Moisan et al. 2017; Pan et al. 2010). Despite these advances, the development and accuracy assessment of satellite algorithms for mapping DPs remain challenging (Bracher et al. 2017; Stock and Subramaniam 2020), especially in coastal regions – where monitoring would be especially important, because human uses and pressures are concentrated at the coasts (Stock et al. 2018b). Phytoplankton community composition is correlated with environmental variables such as light availability and SST (Mouw, Ciochetto, and Yoder 2019). We thus combined SLOOCV with an ecological satellite-mapping approach (Raitsos et al. 2008), i.e. we mapped the DPs based on satellite-based spectral and environmental variables. We trained and validated various statistical models, including both widely used approaches such as artificial neural networks and models that are theoretically suitable but have been less frequently used in ocean color remote sensing, like boosted regression trees. Given a lack of DP algorithms optimized for the NGOM, we used the best-performing algorithms identified here to generate

daily maps, 8-day, monthly and annual composites, as well as seasonal climatologies. These data are available for download in GEOTIF format (see Section "Data availability").

2. Materials and methods

2.1 Study area

The NGOM is a region facing substantial environmental changes and risks. Over the 21st century, the NGOM's physical climate will warm considerably (Biasutti et al. 2012). Offshore oil extraction poses risks to the Gulf's marine biota and ecosystem services (Beyer et al. 2016; Ozhan, Parsons, and Bargu 2014). Riverine inputs of nutrients and stratification lead to seasonal hypoxic conditions in a large "dead zone," with substantial reduction of opportunities for demersal fishing (Rabalais, Turner, and Wiseman 2002). From a remote sensing perspective, the NGOM covers a wide range of biogeochemical and bio-optical conditions, from eutrophic coastal waters to oligotrophic offshore waters (Martínez-López and Zavala-Hidalgo 2009; Müller-Karger et al. 1991; Xue et al. 2013. Because of these characteristics, and the resulting high spatiotemporal variability of phytoplankton dynamics and optical water properties, standard ocean color algorithms for the global oceans can have considerable absolute errors when applied in the NGOM (e.g.; Nababan et al. 2011).

2.2 In-situ data

We combined HPLC (high-performance liquid chromatography) data for 2003–2018 from two sources: Kramer and Siegel (2019) and SeaBASS (Werdell et al. 2003; Werdell and Bailey 2002). We extracted concentrations of seven diagnostic pigments (DPs) as response variables: 19′-butanoyloxyfucoxanthin (But.fuco), 19′-hexanoyloxyfucoxanthin (Hex.fuco), alloxanthin (Allo), fucoxanthin (Fuco), peridinin (Perid), zeaxanthin (Zea) and total chlorophyll b (Chl.b). These seven DPs are widely used to characterize phytoplankton community composition (Mouw et al. 2017; Uitz et al. 2006; Vidussi et al. 2001). We removed observations made within 10 km of land according to GSHHS full-resolution shorelines (Wessel and Smith 1996) to mitigate potential effects of stray light and extreme nearshore conditions such as ephemeral turbidity due to resuspension of sediment. We also removed observations made at depths greater than 10 m. If there were multiple observations within the first 10 m of the water column for a location and time, we only retained the observation closest to the surface.

While many empirical satellite algorithms for mapping DPs predict absolute concentrations (e.g. Bracher et al. 2015b), we were primarily interested in predicting phytoplankton community composition. Absolute DP concentrations, however, reflect both phytoplankton biomass and community composition. Many studies interested primarily in community composition predict phytoplankton size classes or functional types based on weighted relative concentrations (i.e. percentage made up by each of the DPs; Hirata et al. 2011; Mouw et al. 2017). However, this conversion benefits from locally derived weights and involves major uncertainties (Chase et al. 2020). No local weights were available for the Gulf of Mexico, and we could not find published evidence that global weights (e.g. Uitz et al. 2006) would be adequate in this region. Thus, relative concentrations of the DPs were used as response variables serving as proxies for community composition. The relative concentrations were calculated by dividing each pigment's absolute concentration by the sum of all seven pigments' concentrations, S_{DP} (Vidussi et al. 2001):

$$r_X = \frac{c_X}{S_{DP}} \quad \text{(Eq.1)}$$

$$S_{DP} = c_{But.fuco} + c_{Hex.fuco} + c_{Allo} + c_{Fuco} + c_{Perid} + c_{Zea} + c_{Chl.b} \quad \text{(Eq.2)}$$

where c_X is the HPLC-measured concentration of the pigment indicated by the subscript and r_X is the corresponding relative concentration that we predicted as indicators of community composition. The linear correlation between in-situ Chl *a* and S_{DP} was 0.97, indicating a high consistency of the various pigment measurements. Histograms of in-situ relative concentrations for locations with matching satellite data are shown in Fig. S1.

2.3 Satellite data

Predictors were daily satellite data products for 2003–2018 from various sources (Table 1). Following El Hourany et al. (2019), Xi et al. (2021), and Xi et al. (2020), we obtained most satellite data from the GlobColour project (version 4.1; Fanton D' Andon et al. 2009; Maritorena et al. 2010). GlobColour merges ocean color data from several sensors (SeaWiFS, MERIS, MODIS-Aqua, VIIRS, and OLCI-A), allowing for a larger number of matchups for algorithm training and testing. To ensure the best spatial coverage, we only included GlobColour data products that merged data from all available sensors for the given time. The spatial resolution was 4 km and all data from other sources were resampled to the same grid as the GlobColour data. We acknowledge that other multi-sensor data sets, such as OC-CCI (Ocean Color Climate Change Initiative) data, have also been successfully used for mapping phytoplankton pigments (Gittings et al. 2019; Sun et al. 2019).

Based on the GlobColour remote sensing reflectances (RRS), we calculated band ratios as additional predictors; for example, these variables allow the distinction of optical water types (Le et al. 2014). Finally, given its well-established statistical relationship to Chl *a* concentration, we included the maximum blue-to-green band ratio *R* as a predictor:

$$R = log_{10}((max(RRS443, RRS490))/RRS555) \quad \text{(Eq.3)}$$

In addition, we downloaded multi-instrument, optimally interpolated sea surface temperature data (JPL MUR MEaSUREs Project 2015), sea level anomaly data (E.U. Copernicus Marine Service 2019), and wind speed and stress data from IFREMER (Institut Français de Recherche pour l'Exploitation de la Mer; wwz.ifremer.fr). The wind data products were based on QuikSCAT for 2003–2007 (CERSAT at Ifremer 2019a), and from ASCAT on METOP-A for 2008–2018 (CERSAT at Ifremer 2019b). These environmental variables were chosen based on previous research using supervised learning to map aspects of phytoplankton community composition at different spatial scales and based on associations with biologically relevant phenomena. For example, remotely sensed sea surface height indicates the

Table 1. Satellite data and derived data used as predictors and their sources.

Abbreviation	Variable	Res.	Online source/comments	References
CHL CHL_GSM	Chlorophyll *a* Chlorophyll *a* (GSM model)	4 km daily	GlobColour data downloaded from ftp://ftp.hermes.acri.fr between March 23rd and 30 March 2019.	(Fanton D' Andon et al. 2009; Maritorena et al. 2010)
Rrs(412)	Remote sensing reflectance at 412 nm			
Rrs(443)	Remote sensing reflectance at 443 nm			
Rrs(490)	Remote sensing reflectance at 490 nm			
Rrs(555)	Remote sensing reflectance at 555 nm			
Rrs(670)	Remote sensing reflectance at 670 nm			
KD490	Attenuation coefficient at 490 nm			
KDPAR	Attenuation coefficient of PAR			
ZSD	Secchi depth			
ZEU	Euphotic zone depth			
BBP	Particulate backscattering coefficient at 443 nm			
CDM	Absorption coefficient of colored dissolved and detrital organic matter at 443 nm			
R412_443 etc.	Pairwise reflectance band ratios	4 km daily	Calculated for all pairs of Rrs(λ) products (e.g. 412/443, 412/490, . . ., 443/490, 443/555, . . .).	-
R	Maximum blue-to-green band ratio	4 km daily	Calculated as described in Section 2.3.	
SST	Sea surface temperature	0.01° daily	Optimally interpolated data downloaded from ftp://podaac-ftp.jpl.nasa.gov/ on 30 March 2019.	(JPL MUR MEaSUREs Project 2015)
SLA	Sea level anomaly	0.25° daily	Downloaded from my.cmems-du.eu/Core/SEALEVEL_GLO_PHY_L4_REP_OBSERVATIONS_008_047/dataset-duacs-rep-global-merged-allsat-phy-l4/ on 30 March 2019.	(E.U. Copernicus Marine Service 2019)
WV WS	Wind speed Wind stress	0.25° daily	Downloaded from ftp://ftp.ifremer.fr/ifremer/ cersat/products/gridded/MWF/L3/QuikSCAT/ Daily/Netcdf/ (2003–2007) and ftp://ftp.ifremer.fr/ifremer/cersat/products/gridded/mwf-ascat/data/daily/Netcdf/ (2008–2018) on 30 March 2019.	2003–07: (CERSAT at Ifremer 2019a) 2008–18: (CERSAT \at Ifremer 2019b)

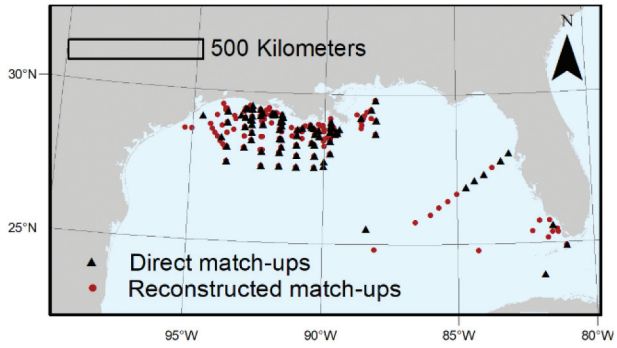

Figure 1. Direct matchups of in-situ pigment and satellite data (n = 130), and reconstructed match-ups (additional in-situ observations with satellite data reconstructed by a gap-filling algorithm; n = 219; see Section 2.7).

spatial extent of the Loop Current (Otis et al. 2019). Finally, we included day-of-year as a predictor to account for any seasonal patterns in the data.

While it has been recommended to use a 3 h temporal window and a spatial window consisting of few pixels at the sensor's native resolution for matching satellite data with in-situ data (Bailey and Werdell 2006), the trade-off between a tight spatiotemporal match and the number of match-ups has led past research mapping different aspects of phytoplankton community composition from space to relax these criteria (e.g. Bracher et al. 2015b; Raitsos et al. 2008). We followed these examples and matched in-situ with satellite observations using a same-calendar-day temporal window. Spatially, we used a 4-pixel window at 4 km resolution, and bilinearly interpolated the extracted value because there are strong land-sea gradients in coastal parts of our study area (Stock and Subramaniam 2020). With these criteria, we obtained 130 matchups of the satellite data with in-situ measurements of the DPs (Figure 1). The match-ups covered diverse biooptical and environmental conditions (Fig. S2). Some algorithms used all predictors, and some used only a subset (e.g. only RRS). To address co-linear predictors (Fig. S3), variable selection was integrated in the algorithms using all predictors, e.g. by regularization, by using principal components calculated from the original predictors (following Bracher et al. 2015b; Xi et al. 2020), or by bagging and boosting based algorithms that are insensitive to high dimensionality and colinear predictors (random forests and boosted regression trees; Belgiu and Drăgu 2016; Dormann et al. 2013).

2.4 Supervised learning algorithms

For each of the 7 DPs, we compared how accurately 13 empirical algorithms predicted relative concentrations for previously unseen data that were spatially separated from all training data (Table 2).

Pan et al. (2013), (2010) proposed that the broad-scale spatial distribution of DPs can be approximated as a cubic polynomial function of remote sensing reflectance band ratios (algorithm PAN). Following this example, we fit (least squares) a function of the form

$$log_{10}(Y_{DP}) = a_0 + a_1 r + a_2 r^2 + a_3 r^3 \qquad \text{(Eq.4)}$$

where r is a band ratio and Y_{DP} is the relative concentration of each pigment. Using the closest bands available in the GlobColour data, we tested both

$$r = log_{10}\left(\frac{RRS490}{RRS555}\right) \qquad \text{(Eq.5)}$$

and

$$r = log_{10}\left(\frac{RRS490}{RRS670}\right) \qquad \text{(Eq.6)}$$

For each pigment, we chose the ratio that led to the best least-squares fit of the polynomial. For predicting zeaxanthin Pan et al. (2013), (2010) modified r based on

Table 2. Overview of empirical algorithms tested in this study. "All" predictors means that all variables listed in Table 1 were provided as input and collinearity was addressed by dimensionality reduction methods (e.g. regularization or using principal components as predictors instead of the original variables).

Algorithm	Model	Predictors
PAN	Regionally fitted polynomial band-ratio algorithm	Band ratios: Either 490/555 or 490/670
ANN5	Artificial neural network, 1 hidden layer with 5 nodes	All
ANN10	Artificial neural network, 1 hidden layer with 10 nodes	All
ANN20	Artificial neural network, 1 hidden layer with 20 nodes	All
ANN4 + 4	Artificial neural network, 2 hidden layers with 4 nodes each	All
RF	Random forest	All
PCRLIN	Linear principal component (PC) regression	PC/EOF scores of Rrs(λ)
PCRALL	Linear principal component (PC) regression	PC/EOF scores of all predictors
PCRRF	Random forest with principal components as predictors	PC/EOF scores of all predictors
BRT1	Boosted regression trees, interaction depth 1	All
BRT2	Boosted regression trees, interaction depth 2	All
BRT3	Boosted regression trees, interaction depth 3	All

SST. However, on our data, incorporating SST – while optimizing model fit – led to outlier predictions that resulted in large mean error estimates, and we hence did not include SST in our model. Furthermore, it is important to note that in contrast to the original algorithm, we predicted relative concentrations of the pigments, not absolute concentrations. Consequently, the algorithm tested here follows the idea but is not an exact reproduction of Pan et al.'s methods.

Many sophisticated neural-network based ocean color algorithms have been developed in recent years for different purposes (e.g. Hieronymi, Müller, and Doerffer 2017; Pahlevan et al. 2020; Ruescas et al. 2018), but given the relatively small size of our data set, we used simpler model structures. Following Keiner and Brown (1999) and González Vilas, Spyrakos, and Torres Palenzuela (2011), we trained ANNs with 5 (ANN5), 10 (ANN10) and 20 (ANN20) nodes in a single hidden layer, as well as 4 nodes each in 2 hidden layers (ANN4 + 4). All ANNs were trained by means of stochastic gradient descent using the R package *ANN2* (Lammers 2020), with L2 regularization and hyperbolic-tangent activation functions. We tested different multipliers for the L2 penalty. Because ANNs are sensitive to initial, randomly chosen parameters, we repeated training each ANN five times for each penalty multiplier, while withholding 20% of the data for error estimation. We chose the ANN that achieved the smallest mean squared error on the withheld data.

Random forests (algorithm RF; Breiman 2001) are an increasingly popular model choice in remote sensing applications of supervised learning (Belgiu and Drăgu 2016). We generated random forests with 300 trees and various proportions of predictors considered at each split with the R package *randomForest* (Liaw and Wiener 2002). The proportion of predictors considered was selected based on the out-of-bag error. We trained boosted regression trees with the R package *gbm* (Greenwell, Boehmke, and Cunningham 2019) and a learning rate of 0.001, a bag fraction of 75%, and interaction depths from 1 to 3 (models BRT1, BRT2, BRT3). For each tested interaction depth, we chose the optimal number of trees using the algorithm of Elith, Leathwick, and Hastie (2008).

Finally, given the relatively large number of potential predictors and correlations between some predictors (e.g. RRS in neighboring bands), we tested three models that used principal component (PC) scores of the predictors as input. We first used a linear multiple regression for this purpose (PCRLIN), following the methods described by Bracher et al. (2015a) and Xi et al. (2020): PCs of remote sensing reflectances were calculated from the matchups, and the number of PCs chosen to include in the models based on the Akaike Information Criterion. We also tested principal components of all original predictors instead of only RRS (PCRALL), and random forests instead of a linear model (PCRRF).

2.5 Iterative Spatial Leave-One-Out Cross-Validation (iSLOOCV)

Cross-validation (CV) estimates a statistical model's prediction error by repeatedly splitting the data into folds. Each fold serves as a test set for an algorithm trained on all other folds. The resulting error estimates are then averaged. This approach reduces the reliance of the error estimate on the sample drawn for testing, yielding more reliable estimates (Lyons et al. 2018). In marine remote sensing, a split into training and test sets is most often made randomly (Bracher et al. 2015a; Hirata et al. 2011; Raitsos et al. 2008; Xi et al. 2020), which assumes that observations are independent (Arlot and Celisse 2010). However, marine labeled data (including the data used in this study) are often spatially autocorrelated and clustered in space and time (Figure 2). Therefore, observations that are randomly selected for validation may not be independent of the observations in the training set, which in turn can lead to an underestimation of prediction errors (Pohjankukka et al. 2017; Stock 2022). This problem can be overcome by ensuring that training and test sets are sufficiently separated in geographic space, in time, or in predictor space, depending on the data and application. There are two main cross-validation strategies for such situations. Spatial block CV splits the data into folds based on geographical blocks (Roberts et al. 2017; Stock et al. 2018a; Stock and Subramaniam 2020; Valavi et al. 2019). However, the choice of block size and shape can be challenging. In contrast, spatial leave-one-out CV (SLOOCV; Le Rest et al. 2014; Le Rest, Pinaud, and Bretagnolle 2013) modifies leave-one-out cross-validation, where each observation is held-out as a test set once, while the training set in each step consists of all other observations. SLOOCV

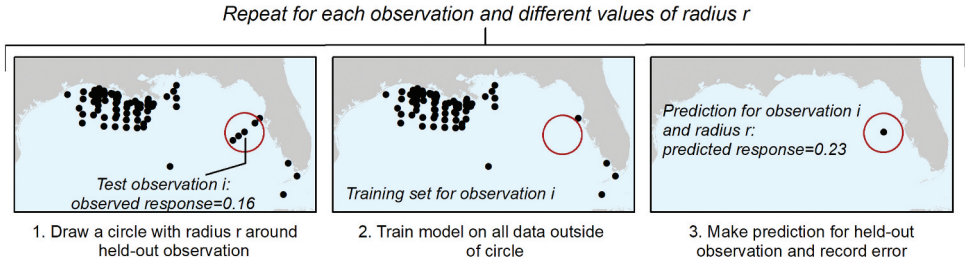

Figure 2. Spatial leave-one-out cross-validation. The iterative version proposed in this study explores how error estimates change as a function of the circle's radius.

adjusts this strategy to avoid the effects of spatial autocorrelation by excluding all observations located within a circle around each test observation from model training (Figure 2). Error estimates are hence based on models that have been trained using only data that are at least the circle's radius, *r*, away from the test observation. Le Rest et al. (2014) suggest using the distance at which the residuals of a model fitted to the full data set become independent. They tested this recommendation on a relatively large data set that was randomly distributed in space. However, residuals can be misleading if the model is over-fitting to the spatial structure of the data (Roberts et al. 2017), and our smaller data set with highly uneven spatial coverage did not allow the reliable estimation of variograms and the de-correlation range.

We hence used an iterative version of SLOOCV (in the following, iSLOOCV) for the validation of empirical satellite algorithms for the oceans. Instead of choosing a fixed distance *r*, we iterated over values from 0.1 km to 200 km. At r = 0.1 km, the test observation itself as well and close-by matchups like measurements made in the same location at different days are excluded from algorithm training. At r = 200 km, a large sub-region around the test observation is excluded from training. The following pseudo-code summarizes the iSLOOCV algorithm for a given statistical model:

(1) For *r* in 0.1 km to 200 km:
  a. For each observation o_i in the set of matchups $O = \{o_1, o_2, \ldots, o_n\}$:
    i. Calculate distances $d_i(o_j)$ between o_i and all o_j
    ii. Create training set $O_{train} = \{o_j: o_j \in O$ and $d_i(o_j) \geq r\}$
    iii. Train and tune model (see Section 2.4) with O_{train}
    iv. Predict response $\hat{y}_{i,r}$ for o_i
  b. Calculate error measures *e(r)* based on differences between $\hat{y}_{i,r}$ and true value y_i of the response.
(2) Calculate average error over the range of *r*: $\bar{e} = e(r)dr/(\max(r) - \min(r))$

2.6 Error measures

We calculated the following error measures (step 1.b in the pseudo-code above), for simplicity omitting subscripts for the radius *r*:

- Linear correlation: *COR*
- Mean absolute error: $MAE = Mean(|(y_i - \hat{y}_i)|)$, $i = 1 \ldots n$
- Root mean squared error: $RMSE = \sqrt{Mean\left((y_i - \hat{y}_i)^2\right)}$, $i = 1 \ldots n$
- Mean error: $ME = Mean(\hat{y}_i - y_i)$, $i = 1 \ldots n$
- Median error: $MDE = Median(\hat{y}_i - y_i)$, $i = 1 \ldots n$
- Median percentage difference: $MDPD = 100 * Median\left(\frac{|y_i - \hat{y}_i|}{y_i}\right)$, $i = 1 \ldots n$

All error estimates for DP algorithms reported in this study were calculated by means of iSLOOCV.

2.7 Data augmentation

To increase the number of matchups, we filled data gaps separately for all predictors using three algorithms: Linear temporal interpolation (LTI), data-interpolating empirical orthogonal functions (DINEOF), and spatiotemporal Kriging (STKR; Fig. S4). Stock et al. (2020) found these algorithms to produce solid reconstructions of pixels obscured by clouds in 3-day composites of Chl *a* for the Gulf of Mexico;

furthermore, these algorithms interpolate in time, which is important given the higher temporal resolution of the data in this study. The LTI algorithm simply interpolated between the closest prior and following observation for each pixel. DINEOF (Alvera-Azcárate et al. 2005; Beckers and Rixen 2003) was run using the software provided by GHER (2016). To achieve acceptable computation times, we split the 16 years of satellite data into four non-overlapping, 4-year subsets. STKR was implemented using the R packages *spacetime* (Pebesma 2012) and *gstat* (Gräler, Pebesma, and Heuvelink 2016; Pebesma 2004). We constructed empirical variograms for a sample of over 50,000 pixels from satellite data distributed evenly over the year 2017 (Fig. S4), fitted theoretical variogram models (manually fine-tuning the fitting process), and interpolated pixels with missing data using the 300 closest data points. The GlobColour product contained two Chl *a* products (see Table 1): CHL (based on empirical band ratio algorithms) and CHL_GSM (based on the semi-analytical Garver-Siegel -Maritorena algorithm; Maritorena, Siegel, and Peterson 2002). Following a comparison of the reconstructed CHL and CHL_GSM values to in-situ observations of Chl *a* (Table S1), we chose spatiotemporal Kriging to fill data gaps in our satellite data, yielding 219 additional, reconstructed matchups (Figure 1). Reconstructed matchups were optionally used for algorithm training, but not for validation.

2.8 Selected models and mapping

Among all tested empirical algorithms, we selected one algorithm for each response based on the iSLOOCV results. We considered both the average errors over threshold distances *r* from 0.1 km to 200 km, and plots of distance-specific errors *e(r)* versus the radius *r*. However, to ensure an acceptable accuracy of predictions, we only selected final models for which the linear correlation between predicted and observed values in the iSLOOCV was ≥0.6, and for which there was negligible bias (*ME* and *MDE* close to zero). As the required accuracy for data products depends on the specific application (Agumya and Hunter 2002), these criteria are intended as a minimum requirement because data products with larger errors are unlikely to be useful for further applications. Among algorithms fulfilling these criteria, we qualitatively considered all error measures, as well as how the estimated errors changed with increasing SLOOCV radius. Once we selected one algorithm for each response variable for which the criteria above could be met, we trained the algorithm on the full data set. We then used it to create daily maps for the period 2003–2018. Finally, we averaged the daily maps into 8-day, monthly, and annual composites, and monthly and seasonal climatologies. To illustrate seasonal dynamics, we extracted time series for four selected locations (GC600: 27.36°N, 90.56°W; Central GOM: 26.00°N, 90.00°W; Tampa: 27.50°N, 82.90°W; LATEX: 29.00°N, −93.50°W) using 5 × 5 pixel windows.

3. Results

3.1 Error estimation

For most algorithms and response variables, prediction errors increased and the correlation between predicted and observed values decreased with higher threshold distance *r* in the iSLOOCV (Figures 3, S6). Furthermore, as *r* increases, in-situ locations are one-by-one removed from the training set in order of their distance to the test observation, resulting in minor changes of the training set. Hence, fluctuating lines in Figure 3 (as exhibited by ANN10) indicate that the trained models were sensitive to particularities of the sample or random aspects of model training.

We identified at least one and often several algorithms fulfilling the quality criteria (see Section 2.8) for four of the seven DPs: But.fuco, Hex.fuco, Fuco and Zea. For these four DPs, the best achieved linear correlations ranged from 0.85 (Fuco) to 0.73 (Zea). MDPDs ranged from 26% (Fuco) to 82% (But.fuco). In most cases, the best error statistics were achieved by random forests or boosted regression trees; the polynomial algorithm by Pan et al. (2013), (2010) worked best for Hex.fuco according to all error measures. For the other three DPs, none of the algorithms achieved an adequate correlation between predicted and observed values. We hence did not select final models and present no maps for these three DPs. A complete list of all tested algorithms' cross-validated errors is provided in the Supplementary Materials (Tables S2-S8).

Among the algorithms using principal components as predictors, using a random forest (PCRRF) instead of multiple linear regression improved predictions. The first principal component (PC1, 43% of variance)

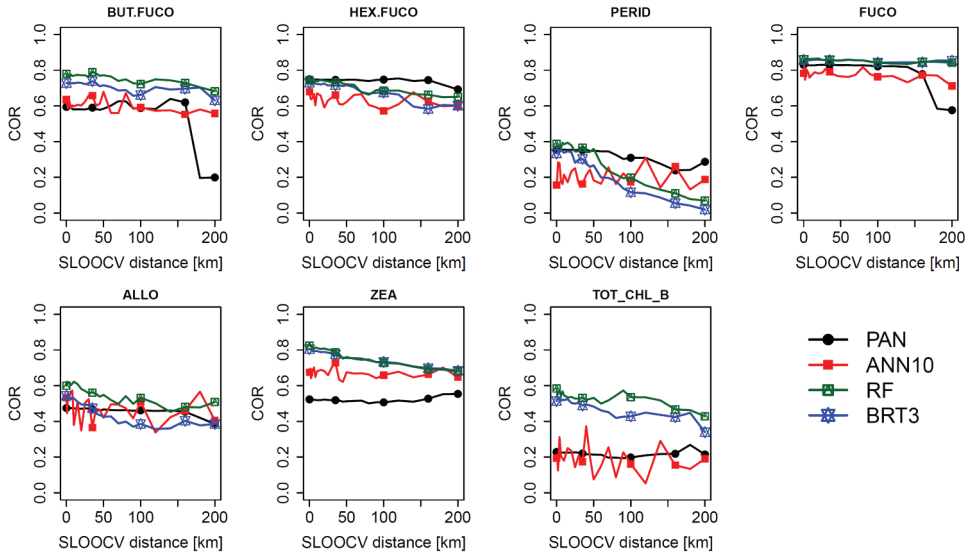

Figure 3. Cross-validated correlations between predicted and observed values as a function of threshold distance *r* used in the iterative spatial leave-one-out cross-validation (iSLOOCV), shown for selected algorithms. Models were trained using direct and gap-filled matchups for Perid, Allo, Zea, and Chl *b*, but only direct matchups for the other pigments. The mean of the curves' y-values corresponds to the single-number correlation over 0.1 km – 200 km reported in Tables 3–6 and S2-S8.

represented variables related to water transparency (Chl, ZEU, Kd490, etc.), PC2 (22%) was related to RRS (dominated by 443, 490). PC3 (12%) and PC4 (7%) were related to environmental variables – PC3 dominated by wind and PC4 by SLA. PC1 was the most important predictor in PCRRF algorithms across pigments.

3.2 Data augmentation by gap-filling

Increasing the number of observations for algorithm training by including data reconstructed by a gap-filling algorithm had mixed positive and negative effects, depending on the response variable and the algorithm (Table 4, Tables S2-S8). Including reconstructed observations in algorithm training improved all error measures for Zea, Allo and Chl.b. It also improved the correlation for Perid substantially, while resulting in small increases of other error measures for this DP. However, of the four pigments for which training with reconstructed matchups improved predictions, only algorithms for Zea met the basic quality criteria for justifying further applications and analyses (see Section 2.8). Gap-filling was therefore used in the training of the final algorithms for Zea, but not for But.fuco, Hex.fuco and Fuco.

Table 3. Best achieved error statistics averaged over 0.1 km – 200 km distance thresholds in the spatial leave-one-out cross-validation by the tested empirical algorithms, and the model which achieved the best value for each response variable and error measure. A "+" behind the model abbreviation indicates that the best error was achieved when using gap-filled data for training, in addition to direct matchups.

Response	COR	MAE	RMSE	MDPD
But.fuco	0.79 BRT3 +	0.02 RF +	0.02 RF	82% RF
Hex.fuco	0.74 PAN	0.06 PAN	0.11 PAN	52% PAN +
Fuco	0.85 BRT2	0.12 PCRRF	0.15 RF	26% PCRRF +
Zea	0.73 RF +	0.11 RF +	0.14 RF +	49% RF +
Allo	0.52 RF +	0.02 PCRRF +	0.02 PCRRF +	66% RF +
Perid	0.30 PAN +	0.03 PAN	0.05 BRT3	71% BRT3
Chl.b	0.51 RF +	0.06 RF +	0.07 RF+	41% RF +

Table 4. Relative difference between best error statistics of models trained on direct matchups and of models trained on direct and reconstructed matchups. Values <0 mean a decrease of the measure if including reconstructed matchups. Abbreviations as in Table 3.

Response	COR	MAE	RMSE	MDPD
But.fuco	+7%	−5%	+0%	+4%
Hex.fuco	−0%	+10%	+4%	−3%
Fuco	−1%	+1%	+3%	−1%
Zea	+2%	−4%	−2%	−3%
Allo	+10%	−5%	−2%	−0%
Perid	+37%	+3%	+0%	+4%
Chl.b	+40%	−6%	−8%	−5%

3.3 Model selection and predictions

Several algorithms with similar error statistics existed for each DP, and no algorithm worked best for all DPs. We chose to use random forests for the four DPs

Table 5. iSLOOCV errors of final models (all random forests) used to generate DP maps for further analyses. Abbreviations as in Table 3, and ME: mean error; MDE: median error.

Response	Gap-filling	COR	MAE	RMSE	ME	MDE	MDPD
But.fuco	No	0.74	0.02	0.02	0	0	82%
Hex.fuco	No	0.69	0.07	0.11	−0.01	−0.02	61%
Fuco	No	0.85	0.13	0.16	0.00	0.00	26%
Zea	Yes	0.73	0.11	0.14	−0.01	−0.02	49%

Table 6. Percent difference between random forests' error statistics and best statistics achieved by any model. For example, if we had chosen the final But.fuco algorithm based on RMSE alone (i.e. ignoring all other error statistics), the correlation would have been 6% higher, and the MAE 5% lower (but other error statistics would have been worse, as the random forest was the best model according to these). Abbreviations as in Table 3.

Response	COR	MAE	RMSE	MDPD
But.fuco	6%	5%	best	best
Hex.fuco	7%	19%	7%	17%
Fuco	0%	1%	best	1%
Zea	2%	5%	2%	4%

where the basic quality criteria (see Section 2.8) were met, because random forests worked well across all DPs and error measures. These models either achieved the lowest cross-validated errors or came close to the best ones for But.fuco, Fuco, and Zea (Tables 5, 6, S2-S8). Figure 4 shows example 8-day composites generated with these algorithms. For Hex.fuco, the polynomial regression following Pan et al. (2010), (2013) had better error statistics overall, yet to maintain consistency between algorithms, random forests were used as final models to create daily maps of all four DPs. However, we provide Hex.fuco maps generated with the PAN algorithm for download in addition to the maps generated with random forests (see "Data availability").

3.4 Spatial and summer-winter patterns of diagnostic pigments

In summer, Fuco made up the largest proportion in nearshore waters up to 20 km from the coastline, with some geographic variation (Figures 5, 6). Zea also made a notable contribution, constituting over 30% of the DPs in nearshore waters, and was the dominant pigment further offshore. Hex.fuco accounted for only a small fraction of the total pigment concentration in nearshore waters, but for about 25% offshore. But.fuco made up only few percent of the diagnostic pigment throughout the study area, increasing

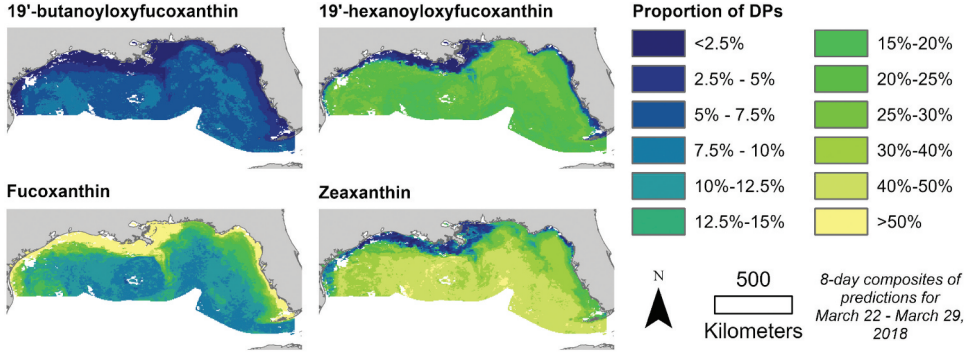

Figure 4. Example of predicted relative pigment concentrations averaged over 8 days (March 22–29, 2018).

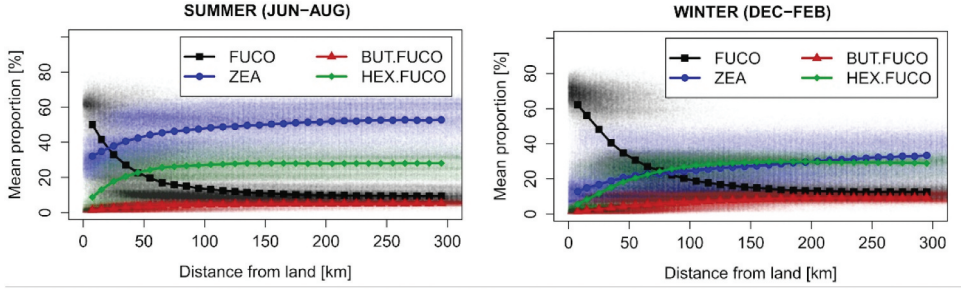

Figure 5. Seasonal land-sea gradients of the fractions of diagnostic pigments. Lines are averages across the study area (2003–2018); clouds of semi-transparent dots represent the point density of daily predictions for individual pixels during this period. See Fig. S4 for spring and autumn gradients.

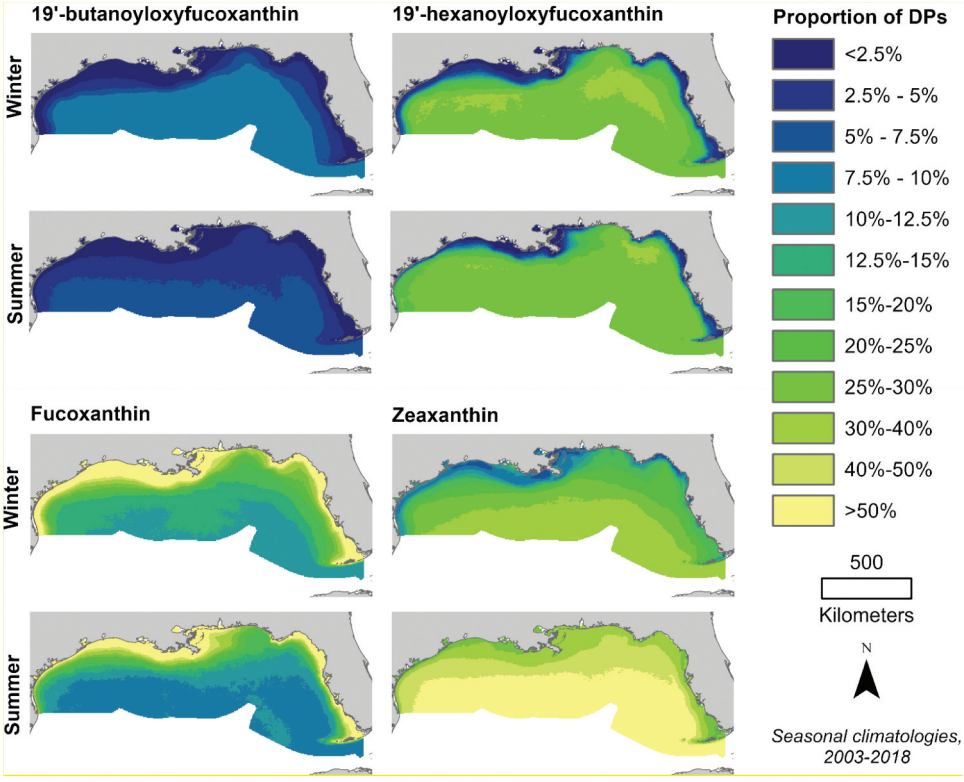

Figure 6. Seasonal mean fraction of diagnostic pigments for 2003–2018. See Fig. S8 for spring and autumn.

slightly when moving offshore. In winter, Fuco was the dominant pigment within the first 80 km from the coastline, again with some geographic variation. Hex.fuco and Zea were the dominant pigments further offshore, with (when averaged throughout the study area) similar proportions that increased offshore. The proportion of But.fuco was on average higher than in summer, but still low. Spring and autumn climatologies showed primarily transitions between winter and summer conditions (Figs. S7, S8). It is important to keep in mind that the numbers presented are the proportion of the sum of seven DPs, whereas we present results for only four, because the available data did not support sufficiently accurate algorithms for the other three.

Time series of daily predictions for selected locations reflected these seasonal patterns (Figure 7). The daily predictions had high short-term variability, likely resulting from the sometimes large but unbiased prediction errors (such that subsequent observations may have overestimation followed by underestimation or vice versa). In particular, the proportion of Fuco fluctuated strongly at the coastal stations over short time periods. These fluctuations were mostly countered by opposing fluctuations in Zea and Hex.fuco (Fig. S9). The time series also exhibited rare and short-lived peaks of fucoxanthin, with high concentrations typical of coastal locations and occurring across seasons. These peaks occurred at times in which high-chlorophyll waters reached far offshore (Fig. S10). While coastal water advection is more common, the predicted Fuco peaks coincided with the highest satellite measured Chl *a* concentrations in the study period for the two offshore locations (Fig. S11). While less visible in the composite-based Figure 7, these peaks were in daily data accompanied in reductions of other DPs' proportions (Fig. S12). For example, from July 4th to 7th 2009, Fuco at GC600 increased by 0.3, while Zea decreased by 0.1, Hex.fuco decreased by 0.2, and But.fuco stayed almost the same (all numbers rounded).

4. Discussion

4.1 Cross-validation for satellite mapping

The challenges of predictive modeling when data are not independent have long been discussed by statisticians, e.g. in the context of time series analysis (Arlot and Celisse 2010; Opsomer, Wang, and Yang 2001).

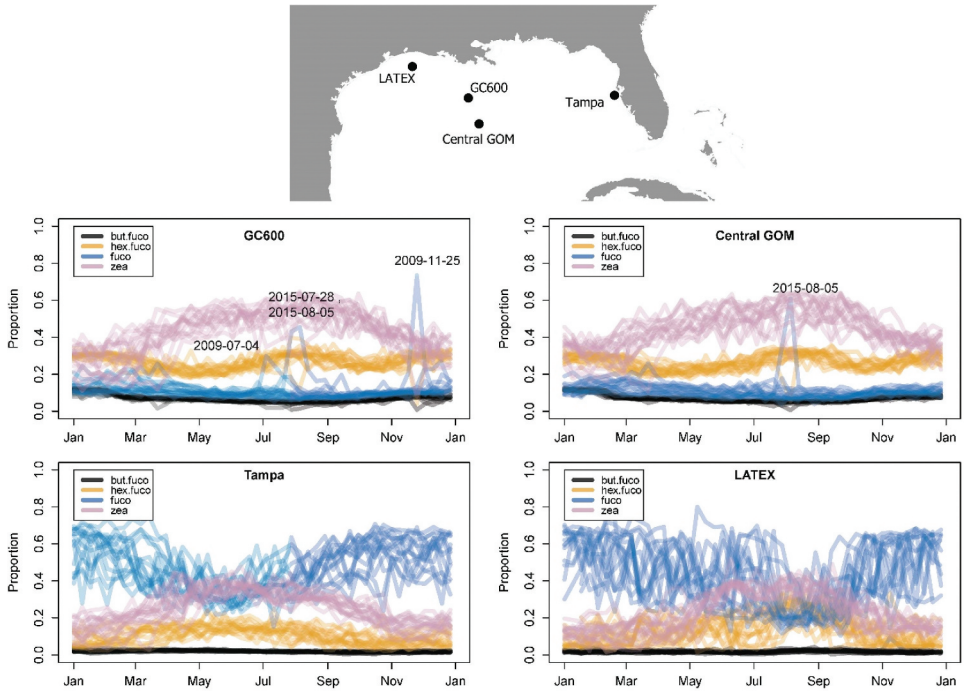

Figure 7. Time series (daily predictions) of diagnostic pigments for four selected locations. Each individual line represents a year from 2003–2018. Rare peaks of fucoxanthin that occur at the two offshore locations are marked with the date.

Recently, the effects of non-independent data have received renewed interest in spatial statistics, especially because more complex machine learning methods are prone to over-fitting to dependence structures (Gregr et al. 2019; Roberts et al. 2017; Stock et al. 2018b). Such overfitting cannot be detected by validation methods relying on randomly held-out observations, and could be especially problematic for ocean remote sensing, because marine in-situ data are often clustered along cruise tracks or phenomena of interest. For example, Stock and Subramaniam (2020) found that error estimates from spatial block CV were considerably larger than error estimates from 5-fold CV and supported different conclusions about which algorithms were accurate enough for further applications. Yet the choice of statistical designs for algorithm validation is rarely justified in the biological ocean remote sensing literature. Recent discussions of and progress in algorithm validation have focused on the collection of high-quality in-situ data following shared protocols, the mismatch of spatial scale between in-situ samples and satellite observations, and new sources of in-situ data such as Argo floats (Bracher et al. 2014; Brewin et al. 2016; Dierssen et al. 2020; Groom et al. 2019; IOCCG 2014; Ruddick et al. 2019; Wojtasiewicz et al. 2018). These are crucial aspects of satellite algorithm validation, yet the statistical consequences of the spatial distribution of labeled data for supervised learning require similar attention (Stock 2022).

This study demonstrated the use of spatial leave-one-out cross-validation (SLOOCV) for validating and selecting empirical satellite algorithms with data that were spatially clustered. Like spatial block cross-validation, SLOOCV enforces a spatial separation of the data used for training and testing models, and avoids the often challenging definition of spatial blocks (Roberts et al. 2017; Stock and Subramaniam 2020). However, the original SLOOCV method uses a single radius, the range of auto-correlation in the residuals, within which training data are omitted around each test observation (Le Rest et al. 2014; Le Rest, Pinaud, and Bretagnolle 2013). This range can be impossible to estimate on sparse and clustered data sets that are common in marine research based on in-situ measurements. We therefore conducted SLOOCV iterating over a range of distances as opposed to a single distance. We proposed an iterative version (iSLOOCV) to explore how the estimated error changed as a function of distance and used the average

error over the whole range to compare the algorithms. This approach bypasses the need to choose a single radius and allowed for the selection of algorithms that worked well both in locations well-covered by training data and when making predictions for locations farther away. Overall, prediction errors increased with larger radius, but only moderately, suggesting that the algorithms were not overfitting to spatial structures. A possible explanation for the moderate increase of errors at larger radii is that spatially clustered in-situ measurements were sometimes made in different years or seasons. In dynamic marine systems, data collected in the same geographic location but at different times can nevertheless represent a large variety of environmental and bio-optical conditions, and thus be less correlated than suggested by their spatial proximity. Indeed, separating the data used for model training and validation in time or in predictor space can be preferable to spatial separation, depending on the characteristics of the study system and data and on the model's intended application (Roberts et al. 2017). Adaptation of iSLOOCV to separating training and test data based on spatiotemporal distances or in predictor space is a promising direction for future research.

4.2 Data augmentation by gap-filling

The availability of sufficiently large labeled data sets for supervised learning applications in remote sensing is a widespread problem and can be addressed from various angles such as semi-supervised learning (Liu et al. 2017). Here, we proposed a direct, simple approach to data augmentation that matched in-situ observations with reconstructed, gap-free satellite data. This approach increased the number of available matchups from 130 to 349. At the same time, reconstructed pixels can have larger errors than direct satellite retrievals or errors with a different distribution. This additional noise could counteract possible improvements of prediction accuracy gained from a larger training set. In this study, including matchups of in-situ measurements with reconstructed satellite imagery in algorithm training had mixed effects on prediction errors. It led to a considerable reduction of prediction errors according to all measures for three of the seven DPs, including one of the four DPs for which pre-set accuracy criteria for further analyses were met. Thus, while we cannot recommend increasing the number of matchups by means of gap-filling as a default procedure, the results show that in situations where researchers are concerned about the size of the labeled data set for algorithm training, increasing the number of matchups by means of gap-filling algorithms can be helpful.

4.3 Quality of selected algorithms and generated data products

While the focus of this study was on methodological aspects, we provide generated maps for download and report detailed error statistics to allow potential users of the algorithms and data to judge their fitness for potential applications in marine science. For example, our results show that fine-scale predictions (i.e. daily data for individual pixels) could have large errors, resulting in high short-term variability. However, all algorithms had negligible bias and regional-scale spatial distributions as well as seasonal patterns were adequately predicted. Hence, the generated data products should be primarily used for broad-scale and longer-term analyses, and potential users should carefully consider their fitness for the intended application. The final algorithms' errors were similar to those of other published ocean color algorithms for the NGOM. For example, Le et al. (2014) report relative errors of 40%-60%, and R^2 between 0.52 and 0.65 (i.e. linear correlations between 0.72 and 0.81) for a Chl *a* algorithm for the Louisiana shelf – an easier prediction task because of the smaller study area and well-established correlations between chlorophyll concentrations and ocean color variables. The four DP algorithms clearing the quality criteria from Section 2.8 achieved median absolute percentage differences between 26% and 82% and linear correlations between 0.72 and 0.85. The predictions exhibited negligible bias; therefore, averaging over multiple daily images or areas encompassing several pixels could further increase accuracy (as random errors cancel each other out). Several algorithms achieved similar error statistics for each of these DPs, suggesting that the achieved accuracy is close to what is possible with the currently available data. Rare and short-lived high offshore fucoxanthin concentrations predicted by the algorithms were associated with unusually high chlorophyll *a* concentration in these locations, reflecting oceanographic conditions normally associated with phytoplankton typical of coastal waters; however, lacking

in-situ data for these situations, we cannot conclude that the spikes reflect real changes in the phytoplankton community. Overall, at broad spatial scales, our results were qualitatively consistent with the findings of previous field campaigns focusing on phytoplankton communities in the NGOM (Chakraborty, Lohrenz, and Gundersen 2017; Chakraborty and Lohrenz 2015; Lambert, Bianchia, and Santschi 1998; Qian et al. 2003). A detailed qualitative comparison is provided in Appendix A1. These results demonstrate the potential of iterative SLOOCV to select adequate supervised learning-based algorithms for satellite mapping applications with relatively small, spatially autocorrelated and unevenly distributed data sets.

Despite these broad-scale similarities between the predicted spatial distributions of diagnostic pigments and independent field campaigns investigating phytoplankton community composition, it is important to keep in mind that pigments are imperfect proxies for phytoplankton types. Yet overall, HPLC is among the most common and quality-controlled methods available. Four broad taxonomic groups of phytoplankton can be reliably distinguished based on their pigment signatures as described by HPLC data, and several of the individual pigments mapped here can serve as useful proxies for these groups: for example, Fuco for diatoms, Hex.fuco for haptophytes, and Zea for cyanobacteria (Kramer and Siegel 2019). Locally, more phytoplankton groups can be distinguished based on HPLC data (Kramer, Siegel, and Graff 2020; Kramer and Siegel 2019). However, distinguishing more detailed groups requires data on more pigments than those for which we found adequate algorithms. Together, these limitations of pigment-based phytoplankton community characterization and satellite retrievals of relevant pigments suggest that only broad phytoplankton classes can be distinguished from space by linking HPLC data with multi-spectral reflectances and environmental variables.

5. Summary and conclusions

(1) Spatial leave-one-out cross-validation that iterates over a range of distances separating training and test observations allows the validation and selection of algorithms based on small, spatially clustered data sets, without the need to choose a fixed separation distance *a priori*. It also provides insights into how errors change as the distance from locations with data increases, and into over-fitting as the training set shrinks with increasing separation distance.

(2) Gap-filling methods can be used to increase the number of matchups between satellite and in-situ measurements. The benefits of more matchups for training supervised learning algorithms sometimes, but not always, outweigh additional errors introduced. Data augmentation by gap-filling is hence worth testing in applications where a small matchup data set is suspected to be the limiting factor for supervised learning.

(3) Regionally optimized supervised learning algorithms for remote sensing of diagnostic phytoplankton pigments achieved adequate accuracy for four out of seven diagnostic pigments, suggesting that some, but not all relevant, broad phytoplankton classes can be distinguished from space based on multi-spectral satellite and environmental data in the northern Gulf of Mexico.

Data availability

Diagnostic pigment maps generated in this study are available for download under CC BY 4.0 license (DOI: 10.17632/nvxy6bd4hm.1). All source code is publicly available under MIT license (https://doi.org/10.6084/m9.figshare.13011557.v3). The original data used in this study are publicly available for download from the sources provided in Table 1.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Gulf of Mexico Research Initiative's "Ecosystem Impacts of Oil and Gas Inputs to the Gulf" (ECOGIG) program, NASA OBB grant NNX16AAJ08G and NSF OCE 1737128. ASt was also supported by an Earth Institute Postdoctoral Fellowship at Columbia University, and a Liber Ero Postdoctoral Fellowship and MEOPAR Postdoctoral Award at the University of British Columbia.

References

Agumya, A., and G. J. Hunter. 2002. "Responding to the Consequences of Uncertainty in Geographical Data." *International Journal of GeographersInformation Science* 16 (5): 405–417. doi:10.1080/13658810210137031.

Alvera-Azcárate, A., A. Barth, M. Rixen, and J. M. Beckers. 2005. "Reconstruction of Incomplete Oceanographic Data Sets Using Empirical Orthogonal Functions: Application to the Adriatic Sea Surface Temperature." *Ocean Model* 9 (4): 325–346. doi:10.1016/j.ocemod.2004.08.001.

Alvera-Azcárate, A., A. Barth, J. M. Beckers, and R. H. Weisberg. 2007. "Multivariate Reconstruction of Missing Data in Sea Surface Temperature, Chlorophyll, and Wind Satellite Fields." *Journal of Geophysics Research Ocean* 112: 1–11. doi:10.1029/2006JC003660.

Arlot, S., and A. Celisse. 2010. "A Survey of cross-validation Procedures for Model Selection." *Statistical Surveys* 4 (none): 40–79. doi:10.1214/09-SS054.

at Ifremer, C. E. R. S. A. T. 2019a. SeaWinds on QuikSCAT Level 4 Gridded Mean Wind Fields [WWW Document]. URL http://products.cersat.fr/details/?id=CER_WND_GLO_1D_025_MWF_QS (accessed 3.30.19).

at Ifremer, C. E. R. S. A. T. 2019b. ASCAT on METOP-A Level 4 Daily Gridded Mean Wind Fields in 0.25° Geographical Grid [WWW Document]. URL http://products.cersat.fr/details/?id=CER_WND_GLO_1D_025_ASCAT (accessed 3.30.19).

Bailey, S. W., and P. J. Werdell. 2006. "A multi-sensor Approach for the on-orbit Validation of Ocean Color Satellite Data Products." *Remote Sensing of Environment* 102 (1–2): 12–23. doi:10.1016/j.rse.2006.01.015.

Barth, A., A. Alvera-Azcárate, M. Licer, and J. M. Beckers. 2020. "DINCAE 1.0: A Convolutional Neural Network with Error Estimates to Reconstruct Sea Surface Temperature Satellite Observations." *Geoscientific Model Development* 13 (3): 1609–1622. doi:10.5194/gmd-13-1609-2020.

Beckers, J. M., and M. Rixen. 2003. "EOF Calculations and Data Filling from Incomplete Oceanographic Datasets." *Journal of Atmospheric and Oceanic Technology* 20 (12): 1839–1856. doi:10.1175/1520-0426(2003)020<1839:ECADFF>2.0.CO;2.

Bel, L., D. Allard, J. M. Laurent, R. Cheddadi, and A. Bar-Hen. 2009. "CART Algorithm for Spatial Data: Application to Environmental and Ecological Data." *Computational Statistics & Data Analysis* 53 (8): 3082–3093. doi:10.1016/j.csda.2008.09.012.

Belgiu, M., and L. Drăgu. 2016. "Random Forest in Remote Sensing: A Review of Applications and Future Directions." *ISPRS Journal of Photogrammetry and Remote Sensing* 114: 24–31. doi:10.1016/j.isprsjprs.2016.01.011.

Beyer, J., H. C. Trannum, T. Bakke, P. V. Hodson, and T. K. Collier. 2016. "Environmental Effects of the Deepwater Horizon Oil Spill: A Review." *Marine Pollution Bulletin* 110 (1): 28–51. doi:10.1016/j.marpolbul.2016.06.027.

Biasutti, M., A. H. Sobel, S. J. Camargo, and T. T. Creyts. 2012. "Projected Changes in the Physical Climate of the Gulf Coast and Caribbean." *Climatic Change* 112 (3–4): 819–845. doi:10.1007/s10584-011-0254-y.

Bracher, A., N. Hardman-mountford, T. Hirata, S. Bernard, E. Boss, A. Bricaud, V. Brotas, et al., 2014. Report on IOCCG Workshop Phytoplankton Composition from Space: Towards a Validation Strategy for Satellite Algorithms, Nasa/Tm–2015-217528.

Bracher, A., N. Hardman-mountford, T. Hirata, S. Bernard, E. Boss, A. Bricaud, V. Brotas, et al. 2015a. "Phytoplankton Composition from Space : Towards a Validation Strategy for Satellite Algorithms."

Bracher, A., M. H. Taylor, B. Taylor, T. Dinter, R. Röttgers, and F. Steinmetz. 2015b. "Using Empirical Orthogonal Functions Derived from remote-sensing Reflectance for the Prediction of Phytoplankton Pigment Concentrations." *Ocean Science* 11 (1): 139–158. doi:10.5194/os-11-139-2015.

Bracher, A., H. A. Bouman, R. J. W. Brewin, A. Bricaud, V. Brotas, A. M. Ciotti, L. Clementson, et al. 2017. "Obtaining Phytoplankton Diversity from Ocean Color: A Scientific Roadmap for Future Development." *Frontiers in Marine Science* 4: 1–15. doi:10.3389/fmars.2017.00055.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. doi:10.1023/A:1010933404324.

Brewin, R. J. W., G. Dall'Olmo, S. Pardo, V. van Dongen-Vogels, and E. S. Boss. 2016. "Underway Spectrophotometry along the Atlantic Meridional Transect Reveals High Performance in Satellite Chlorophyll Retrievals." *Remote Sensing of Environment* 183: 82–97. doi:10.1016/j.rse.2016.05.005.

Chakraborty, S., and S. Lohrenz. 2015. "Phytoplankton Community Structure in the River-influenced Continental Margin of the Northern Gulf of Mexico." *Marine Ecology Progress Series* 521: 31–47. doi:10.3354/meps11107.

Chakraborty, S., S. E. Lohrenz, and K. Gundersen. 2017. "Photophysiological and Light Absorption Properties of Phytoplankton Communities in the river-dominated Margin of the Northern Gulf of Mexico." *Journal of Geophysical Research: Oceans* 122 (6): 4922–4938. doi:10.1002/2015JC011516.

Chase, A. P., S. J. Kramer, N. Haëntjens, E. S. Boss, L. Karp-Boss, M. Edmondson, and J. R. Graff. 2020. "Evaluation of Diagnostic Pigments to Estimate Phytoplankton Size Classes." *Limnology and Oceanography: Methods* 18 (10): 570–584. doi:10.1002/lom3.10385.

Chen, S., C. Hu, B. B. Barnes, R. Wanninkhof, W. J. Cai, L. Barbero, and D. Pierrot. 2019. "A Machine Learning Approach to Estimate Surface Ocean pCO2 from Satellite Measurements." *Remote Sensing of Environment* 228: 203–226. doi:10.1016/j.rse.2019.04.019.

Dierssen, H., A. Bracher, V. Brando, H. Loisel, and K. Ruddick. 2020. "Data Needs for Hyperspectral Detection of Algal Diversity across the Globe." *Oceanography* 33 (1): 74–79. doi:10.5670/oceanog.2020.111.

Doerffer, R., and H. Schiller. 2007. "The MERIS Case 2 Water Algorithm." *International Journal of Remote Sensing* 28 (3–4): 517–535. doi:10.1080/01431160600821127.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. J. R. G. Marquéz, et al. 2013. "Collinearity: A Review of Methods to Deal with It and A Simulation Study Evaluating Their Performance." *Ecography (Cop.)* 36 (1): 27–46. doi:10.1111/j.1600-0587.2012.07348.x.

El Hourany, R., M. Abboud-Abi Saab, G. Faour, O. Aumont, M. Crépon, and S. Thiria. 2019. "Estimation of Secondary Phytoplankton Pigments from Satellite Observations Using Self-Organizing Maps (Soms)." *Journal of Geophysical Research: Oceans* 124 (2): 1357–1378. doi:10.1029/2018JC014450.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. "A Working Guide to Boosted Regression Trees." *Journal of Animal Ecology* 77 (4): 802–813. doi:10.1111/j.1365-2656.2008.01390.x.

E.U. Copernicus Marine Service. 2019. Global Ocean Gridded L4 Sea Surface Heights and Derived Variables Reprocessed. Product Identifier: SEALEVEL_GLO_PHY_L4_REP_OBSERVATIONS_008_047 [WWW Document]. http://marine.copernicus.eu/ (accessed 3.30.19).

Fanton D' Andon, O., A. Mangin, S. Lavender, D. Antoine, S. Maritorena, A. Morel, G. Barrot, J. Demaria, and S. Pinnock, 2009. GlobColour - the European Service for Ocean Colour, Proceedings of the 2009 IEEE International Geoscience & Remote Sensing Symposium. 10.1029/2006JC004007

GHER. 2016. DINEOF [WWW Document]. URL http://modb.oce.ulg.ac.be/mediawiki/index.php/DINEOF (accessed 2.27.19).

Gittings, J. A., R. J. W. Brewin, D. E. Raitsos, M. Kheireddine, M. Ouhssain, B. H. Jones, and I. Hoteit. 2019. "Remotely Sensing Phytoplankton Size Structure in the Red Sea. Remote Sens." *Environ* 234: 111387. doi:10.1016/j.rse.2019.111387.

González Vilas, L., E. Spyrakos, and J. M. Torres Palenzuela. 2011. "Neural Network Estimation of Chlorophyll a from MERIS Full Resolution Data for the Coastal Waters of Galician Rias (NW Spain). Remote Sens." *Environ* 115: 524–535. doi:10.1016/j.rse.2010.09.021.

Gräler, B., E. Pebesma, and G. Heuvelink. 2016. "Spatio-temporal Geostatistics Using Gstat." *The R Journal* 8 (1): 204–218. doi:10.1007/978-3-319-17885-1.

Greenwell, B., B. Boehmke, and J. Cunningham. 2019. Gbm: Generalized Boosted Regression Models. R Package Version 2.1.5.

Gregr, E. J., D. M. Palacios, A. Thompson, and K. M. A. Chan. 2019. "Why Less Complexity Produces Better Forecasts: An Independent Data Evaluation of Kelp Habitat Models." *Ecography (Cop.)* 42 (3): 428–443. doi:10.1111/ecog.03470.

Groom, S. B., S. Sathyendranath, Y. Ban, S. Bernard, B. Brewin, V. Brotas, C. Brockmann, et al. 2019. "Satellite Ocean Colour: Current Status and Future Perspective." *Frontiers in Marine Science* 6. doi:10.3389/fmars.2019.00485.

Hieronymi, M., D. Müller, and R. Doerffer. 2017. "The OLCI Neural Network Swarm (ONNS): A bio-geo-optical Algorithm for Open Ocean and Coastal Waters." *Frontiers in Marine Science* 4: 1–18. doi:10.3389/fmars.2017.00140.

Hilborn, A., and M. Costa. 2018. "Applications of DINEOF to satellite-derived chlorophyll-a from a Productive Coastal Region." *Remote Sensing* 10 (9): 11–13. doi:10.3390/rs10091449.

Hirata, T., N. J. Hardman-Mountford, R. J. W. Brewin, J. Aiken, R. Barlow, K. Suzuki, and T. Isada, et al. 2011. "Synoptic Relationships between Surface Chlorophyll-*a* and Diagnostic Pigments Specific to Phytoplankton Functional Types." *Biogeosciences* 8 (2): 311–327. doi:10.5194/bg-8-311-2011.

Hu, S., H. Liu, W. Zhao, T. Shi, Z. Hu, Q. Li, and G. Wu. 2018. "Comparison of Machine Learning Techniques in Inferring Phytoplankton Size Classes." *Remote Sensing* 10 (3): 191. doi:10.3390/rs10030191.

IOCCG. 2014. "Phytoplankton Functional Types from Space." In *Reports of the International Ocean-Colour Coordinating Group (IOCCG)*, edited by S. Sathyendranath, 154. Dartmouth, Canada: International Ocean-Colour Coordinating Group.

JPL MUR MEaSUREs Project. 2015. GHRSST Level 4 MUR Global Foundation Sea Surface Temperature Analysi v4.1. WWW Document. doi: 10.5067/GHGMR-4FJ04.

Keiner, L. E., and C. W. Brown. 1999. "Estimating Oceanic Chlorophyll Concentrations with Neural Networks." *International Journal of Remote Sensing* 20 (1): 189–194. doi:10.1080/014311699213695.

Kerr, J. T., and M. Ostrovsky. 2003. "From Space to Species: Ecological Applications for Remote Sensing." *Trends in Ecology & Evolution* 18 (6): 299–305. doi:10.1016/S0169-5347(03)00071-5.

Kramer, S. J., and D. A. Siegel. 2019. "How Can Phytoplankton Pigments Be Best Used to Characterize Surface Ocean Phytoplankton Groups for Ocean Color Remote Sensing Algorithms?" *Journal of Geophysical Research: Oceans* 124 (11): 7557–7574. doi:10.1029/2019JC015604.

Kramer, S. J., D. A. Siegel, and J. R. Graff. 2020. "Phytoplankton Community Composition Determined from Co-variability among Phytoplankton Pigments from the NAAMES Field Campaign." *Frontiers in Marine Science* 7: 1–15. doi:10.3389/fmars.2020.00215.

Lambert, C. D., T. S. Bianchia, and P. H. Santschi. 1998. "Cross-shelf Changes in Phytoplankton Community Composition in the Gulf of Mexico (Texas shelf/slope): The Use of Plant Pigments as Biomarkers." *Continental Shelf Research* 19 (1): 1–21. doi:10.1016/S0278-4343(98)00075-2.

Lammers, B. 2020. "ANN2: Artificial Neural Networks for Anomaly Detection." R package version 2.3.4. https://CRAN.R-project.org/package=ANN2

Le, C., J. C. Lehrter, C. Hu, M. C. Murrell, and L. Qi. 2014. "Spatiotemporal chlorophyll-a Dynamics on the Louisiana Continental Shelf Derived from a Dual Satellite Imagery Algorithm." *Journal of Geophysics Research Ocean* 175: 238. doi:10.1038/175238c0.

Le Rest, K., D. Pinaud, and V. Bretagnolle. 2013. "Accounting for Spatial Autocorrelation from Model Selection to Statistical Inference: Application to a National Survey of a Diurnal Raptor." *Ecological Informatics* 14: 17–24. doi:10.1016/j.ecoinf.2012.11.008.

Le Rest, K., D. Pinaud, P. Monestiez, J. Chadoeuf, and V. Bretagnolle. 2014. "Spatial leave-one-out cross-validation for Variable Selection in the Presence of Spatial Autocorrelation." *Global Ecology and Biogeography* 23 (7): 811–820. doi:10.1111/geb.12161.

Liaw, A., and M. Wiener. 2002. "Classification and Regression by randomForest." *R News* 2: 18–22.

Liu, H., T. Shi, Y. Chen, J. Wang, T. Fei, and G. Wu. 2017. "Improving Spectral Estimation of Soil Organic Carbon Content through semi-supervised Regression." *Remote Sensing* 9: 4–8. doi:10.3390/rs9010029.

Liu, X., M. Wang, M. Wang, X. Liu, and M. Wang. 2019. "Filling the Gaps of Missing Data in the Merged VIIRS SNPP/NOAA-20 Ocean Color Product Using the DINEOF Method." *Remote Sensing* 11 (2): 178. doi:10.3390/rs11020178.

Liu, H., Q. Li, Y. Bai, C. Yang, J. Wang, Q. Zhou, S. Hu, T. Shi, X. Liao, and G. Wu. 2021. "Improving Satellite Retrieval of Oceanic Particulate Organic Carbon Concentrations Using Machine Learning Methods." *Remote Sensing of Environment* 256: 112316. doi:10.1016/j.rse.2021.112316.

Lyons, M. B., D. A. Keith, S. R. Phinn, T. J. Mason, and J. Elith. 2018. "A Comparison of Resampling Methods for Remote Sensing Classification and Accuracy Assessment." *Remote Sensing of Environment* 208: 145–153. doi:10.1016/j.rse.2018.02.026.

Maritorena, S., D. A. Siegel, and A. R. Peterson. 2002. "Optimization of a Semianalytical Ocean Color Model for global-scale Applications." *Applied Optics* 41 (15): 2705–2714. doi:10.1364/AO.41.002705.

Maritorena, S., O. H. F. D'Andon, A. Mangin, and D. A. Siegel. 2010. "Merged Satellite Ocean Color Data Products Using a bio-optical Model: Characteristics, Benefits and Issues." *Remote Sensing of Environment* 114 (8): 1791–1804. doi:10.1016/J.RSE.2010.04.002.

Martínez-López, B., and J. Zavala-Hidalgo. 2009. "Seasonal and Interannual Variability of cross-shelf Transports of Chlorophyll in the Gulf of Mexico." *Journal of Marine Systems* 77 (1–2): 1–20. doi:10.1016/j.jmarsys.2008.10.002.

McClain, C. R. 2009. "A Decade of Ocean Color Observations." *Annual Review of Marine Science* 1 (1): 19–42. doi:10.1146/annurev.marine.010908.163650.

Moisan, T. A., K. M. Rufty, J. R. Moisan, and M. A. Linkswiler. 2017. "Satellite Observations of Phytoplankton Functional Type Spatial Distributions, Phenology, Diversity, and Ecotones." *Frontiers in Marine Science* 4: 1–24. doi:10.3389/fmars.2017.00189.

Mouw, C. B., N. J. Hardman-mountford, S. Alvain, A. Bracher, R. J. W. Brewin, A. Bricaud, A. M. Ciotti, et al. 2017. "A Consumer ' S Guide to Satellite Remote Sensing of Multiple Phytoplankton Groups in the Global Ocean." *Frontiers in Marine Science* 4: 41. doi:10.3389/fmars.2017.00041.

Mouw, C. B., A. B. Ciochetto, and J. A. Yoder. 2019. "A Satellite Assessment of Environmental Controls of Phytoplankton Community Size Structure Global Biogeochemical Cycles." *Global Biogeochemical Cycles* 33 (5): 540–558. doi:10.1029/2018GB006118.

Müller-Karger, F. E., J. J. Walsh, R. H. Evans, and M. B. Meyers. 1991. "On the Seasonal Phytoplankton Concentration and Sea Surface Temperature Cycles of the Gulf of Mexico as Determined by Satellites." *Journal of Geophysical Research* 96 (C7): 12645. doi:10.1029/91JC00787.

Nababan, B., F. E. Muller-Karger, C. Hu, and D. C. Biggs. 2011. "Chlorophyll Variability in the Northeastern Gulf of Mexico." *International Journal of Remote Sensing* 32 (23): 8373–8391. doi:10.1080/01431161.2010.542192.

Nair, A., S. Sathyendranath, T. Platt, J. Morales, V. Stuart, M. Forget, E. Devred, and H. Bouman. 2008. "Remote Sensing of Phytoplankton Functional Types." *Remote Sensing of Environment* 112 (8): 3366–3375. doi:10.1016/j.rse.2008.01.021.

O'Reilly, J. E., S. Maritorena, B. G. Mitchell, D. A. Siegel, K. L. Carder, S. A. Garver, M. Kahru, and C. McClain. 1998. "Ocean Color Chlorophyll Algorithms for SeaWiFS." *Journal of Geophysical Research: Oceans* 103 (C11): 24937–24953. doi:10.1029/98JC02160.

Opsomer, J., Y. Wang, and Y. Yang. 2001. "Nonparametric Regression with Correlated Errors." *Statistical Science* 16 (2): 134–153. doi:10.1214/ss/1009213287.

Otis, D. B., M. Le Hénaff, V. H. Kourafalou, L. McEachron, and F. E. Muller-Karger. 2019. "Mississippi River and Campeche Bank (Gulf of Mexico) Episodes of cross-shelf Export of Coastal Waters Observed with Satellites." *Remote Sensing* 1. doi:10.3390/RS11060723.

Ozhan, K., M. L. Parsons, and S. Bargu. 2014. "How Were Phytoplankton Affected by the Deepwater Horizon Oil Spill?" *Bioscience* 64 (9): 829–836. doi:10.1093/biosci/biu117.

Pahlevan, N., B. Smith, J. Schalles, C. Binding, Z. Cao, R. Ma, K. Alikas, et al. 2020. "Seamless Retrievals of chlorophyll-A from Sentinel-2 (MSI) and Sentinel-3 (OLCI) in Inland and Coastal Waters: A machine-learning Approach." *Remote Sensing of Environment* 240: 111604. doi:10.1016/j.rse.2019.111604.

Pan, X., A. Mannino, M. E. Russ, S. B. Hooker, and L. W. Harding. 2010. "Remote Sensing of Phytoplankton Pigment Distribution in the United States Northeast Coast." *Remote Sensing of Environment* 114 (11): 2403–2416. doi:10.1016/j.rse.2010.05.015.

Pan, X., G. T. F. Wong, T. Ho, F. Shiah, and H. Liu. 2013. "Remote Sensing of Environment Remote Sensing of Picophytoplankton Distribution in the Northern South China Sea. Remote Sens." *Environ* 128: 162–175. doi:10.1016/j.rse.2012.10.014.

Pebesma, E. J. 2004. "Multivariable Geostatistics in S: The Gstat Package." *Computers & Geosciences* 30 (7): 683–691. doi:10.1016/j.cageo.2004.03.012.

Pebesma, E. 2012. "Spacetime: Spatio-Temporal Data in R." *Journal of Statistical Software* 51 (7): 1–30. doi:10.18637/jss.v051.i07.

Pohjankukka, J., T. Pahikkala, P. Nevalainen, and J. Heikkonen. 2017. "Estimating the Prediction Performance of Spatial Models via Spatial k-fold Cross Validation." *International Journal of Geographical Information Science* 31 (10): 2001–2019. doi:10.1080/13658816.2017.1346255.

Qian, Y., A. E. Jochens, M. C. Kennicutt, and D. C. Biggs. 2003. "Spatial and Temporal Variability of Phytoplankton Biomass and Community Structure over the Continental Margin of the Northeast Gulf of Mexico Based on Pigment Analysis." *Continental Shelf Research* 23 (1): 1–17. doi:10.1016/S0278-4343(02)00173-5.

Quere, C., H. Le, S. P. Colin Prentice, I. Buitenhuis, E. T. Aumont, O. Bopp, L. Claustre, et al. 2005. "Ecosystem Dynamics Based on Plankton Functional Types for Global Ocean Biogeochemistry Models." *Global Change Biology* 051013014052005. doi:10.1111/j.1365-2486.2005.1004.x.

Rabalais, N. N., R. E. Turner, and W. J. Wiseman. 2002. "Gulf of Mexico Hypoxia, A.K.A. "The Dead Zone." *Annual Review of Ecology and Systematics* 33 (1): 235–263. doi:10.1146/annurev.ecolsys.33.010802.150513.

Raitsos, D. E., S. J. Lavender, C. D. Maravelias, J. Haralabous, A. J. Richardson, and P. C. Reid. 2008. "Identifying Four Phytoplankton Functional Types from Space: An Ecological Approach." *Limnology and Oceanography* 53 (2): 605–613. doi:10.4319/lo.2008.53.2.0605.

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, et al. 2017. "Cross-validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography (Cop.)* 40 (8): 913–929. doi:10.1111/ecog.02881.

Ruddick, K. G., K. Voss, A. C. Banks, E. Boss, A. Castagna, R. Frouin, M. Hieronymi, et al. 2019. "A Review of Protocols for Fiducial Reference Measurements of Downwelling Irradiance for the Validation of Satellite Remote Sensing Data over Water." *Remote Sensing* 11. doi:10.3390/rs11151742.

Ruescas, A. B., M. Hieronymi, G. Mateo-Garcia, S. Koponen, K. Kallio, and G. Camps-Valls. 2018. "Machine Learning Regression Approaches for Colored Dissolved Organic Matter (CDOM) Retrieval with S2-MSI and S3-OLCI Simulated Data." *Remote Sensing* 10 (5): 1–25. doi:10.3390/rs10050786.

Saulquin, B., F. Gohin, and O. Fanton D' Andon. 2018. "Interpolated Fields of satellite-derived multi-algorithm chlorophyll-a Estimates at Global and European Scales in the Frame of the European Copernicus-Marine Environment Monitoring Service." *Journal of Operational Oceanography* 1–11. doi:10.1080/1755876X.2018.1552358.

Soja-Woźniak, M., L. Laiolo, M. E. Baird, R. Matear, L. Clementson, T. Schroeder, M. A. Doblin, and I. M. Suthers. 2020. "Effect of Phytoplankton Community Size Structure on remote-sensing Reflectance and Chlorophyll a Products." *Journal of Marine Systems* 211: 103400. doi:10.1016/j.jmarsys.2020.103400.

Stock, A. 2015. "Satellite Mapping of Baltic Sea Secchi Depth with Multiple Regression Models." *International Journal of Applied Earth Observation and Geoinformation* 40. doi:10.1016/j.jag.2015.04.002.

Stock, A., L. B. Crowder, B. S. Halpern, and F. Micheli. 2018a. "Uncertainty Analysis and Robust Areas of High and Low Modeled Human Impact on the Global Oceans." *Conservation Biology* 32 (6): 1368–1379. doi:10.1111/cobi.13141.

Stock, A., A. J. Haupt, M. E. Mach, and F. Micheli. 2018b. "Mapping Ecological Indicators of Human Impact with Statistical and Machine Learning Methods: Tests on the California Coast." *Ecologica Information* 48. doi:10.1016/j.ecoinf.2018.07.007.

Stock, A., and A. Subramaniam. 2020. "Accuracy of Empirical Satellite Algorithms for Mapping Phytoplankton Diagnostic Pigments in the Open Ocean: A Supervised Learning Perspective." *Frontiers in Marine Science* 7. doi:10.3389/fmars.2020.00599.

Stock, A., A. Subramaniam, G. L. Van Dijken, L. M. Wedding, K. R. Arrigo, M. M. Mills, M. A. Cameron, and F. Micheli. 2020. "Comparison of cloud-filling Algorithms for Marine Satellite Data." *Remote Sensing* 12 (20): 3313. doi:10.3390/rs12203313.

Stock, A. 2022. "Spatiotemporal Distribution of Marine Labeled Data Can Bias the Validation and Selection of Supervised Learning Algorithms." *ISPRS Journal of Photogrammetry and Remote Sensing* 187: 46–60. doi:10.1016/j.isprsjprs.2022.02.023.

Sun, X., F. Shen, R. J. W. Brewin, D. Liu, and R. Tang. 2019. "Twenty-Year Variations in Satellite-Derived Chlorophyll-a and Phytoplankton Size in the Bohai Sea and Yellow Sea." *Journal of Geophysical Research: Oceans* 124 (12): 8887–8912. doi:10.1029/2019JC015552.

Uitz, J., H. Claustre, A. Morel, and S. B. Hooker. 2006. "Vertical Distribution of Phytoplankton Communities in Open Ocean: An Assessment Based on Surface Chlorophyll." *Journal of Geophysics Research Ocean* 111. doi:10.1029/2005JC003207.

Valavi, R., J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita. 2019. "Block CV: An R Package for Generating Spatially or Environmentally Separated Folds for K -fold cross-validation of Species Distribution Models." *Methods in Ecology and Evolution* 10 (2): 225–232. doi:10.1111/2041-210X.13107.

Vidussi, F., H. Claustre, B. B. Manca, A. Luchetta, and J.-C. Marty. 2001. "Phytoplankton Pigment Distribution in Relation to Upper Thermocline Circulation in the Eastern Mediterranean Sea during Winter." *Journal of Geophysical Research: Oceans* 106 (C9): 19939–19956. doi:10.1029/1999JC000308.

Werdell, P. J., and S. W. Bailey. 2002. *The SeaWiFS bio-optical Archive and Storage System (Seabass): Current Architecture and Implementation*. Greenbelt, MD: Goddard Space Flight Center.

Werdell, P. J., S. Bailey, G. Fargion, C. Pietras, K. Knobelspiesse, G. Feldman, and C. McClain. 2003. *Unique Data Repository Facilitates Ocean Color Satellite Validation*, 377–387. Vol. 84. Washington. DC: Eos.

Wessel, P., and W. H. F. Smith. 1996. "A Global, self-consistent, Hierarchical, high-resolution Shoreline Database." *Journal of Geophysical Research: Solid Earth* 101 (B4): 8741–8743. doi:10.1029/96JB00104.

Wojtasiewicz, B., N. J. Hardman-Mountford, D. Antoine, F. Dufois, D. Slawinski, and T. W. Trull. 2018. "Use of bio-optical Profiling Float Data in Validation of Ocean Colour Satellite Products in a Remote Ocean Region." *Remote Sensing of Environment* 209: 275–290. doi:10.1016/j.rse.2018.02.057.

Xi, H., S. N. Losa, A. Mangin, M. A. Soppa, P. Garnesson, J. Demaria, Y. Liu, O. H. F. D'Andon, and A. Bracher. 2020. "Global Retrieval of Phytoplankton Functional Types Based

on Empirical Orthogonal Functions Using CMEMS GlobColour Merged Products and Further Extension to OLCI Data." *Remote Sensing of Environment* 240: 111704. doi:10.1016/j.rse.2020.111704.

Xi, H., S. N. Losa, A. Mangin, P. Garnesson, M. Bretagnon, J. Demaria, M. A. Soppa, O. Hembise Fanton D' Andon, and A. Bracher. 2021. "Global Chlorophylla Concentrations of Phytoplankton Functional Types with Detailed Uncertainty Assessment Using Multisensor Ocean Color and Sea Surface Temperature Satellite Products." *Journal of Geophysical Research: Oceans* 126 (5): 1–27. doi:10.1029/2020JC017127.

Xue, Z., R. He, K. Fennel, W. J. Cai, S. Lohrenz, and C. Hopkinson. 2013. "Modeling Ocean Circulation and Biogeochemical Variability in the Gulf of Mexico." *Biogeosciences* 10: 7219–7234. doi:https://doi.org/10.5194/bg-10-7219-2013

Appendices

Appendix A1: Comparison of generated diagnostic pigment maps to field campaigns investigating phytoplankton dynamics in the northern Gulf of Mexico

At broad spatial scales, our predicted spatial distributions of four diagnostic pigments were qualitatively consistent with previous in-situ phytoplankton research in the NGOM. Chakraborty, Lohrenz, and Gundersen (2017) report that nearshore waters are dominated by diatoms, with exceptions in summer when cyanobacteria and prochlorophytes can be dominant. Correspondingly, we found Fuco (a pigment characteristic of diatoms, although relationships between diagnostic pigments and phytoplankton types can be ambiguous; Nair et al. 2008) to be the dominant pigment in nearshore waters in winter; in summer, Zea (a pigment characteristic of cyanobacteria) accounted for a large fraction of the diagnostic pigments in nearshore waters. Offshore, Chakraborty, Lohrenz, and Gundersen (2017) found mixed communities, with haptophytes often being a major taxon. Correspondingly, our algorithms predicted that Hex.fuco (a pigment characteristic of haptophytes) could be dominant in winter, still making up around ¼ of the diagnostic pigments in summer.

Chakraborty and Lohrenz (2015) found diatoms to be dominant in inner and mid-shelf waters of the NGOM, especially in winter and spring, and still accounting for >30% of Chl a in summer and fall. Zea was the dominant pigment further offshore. These spatial and seasonal results are reflected in our maps of relative Fuco concentrations.

Qian et al. (2003) found prymnesiophytes to be dominant in much of the northeastern Gulf, with increasing relative abundance offshore. The exception were waters near the mouth of the Mississippi, were prymnesiophytes accounted for less of the Chl a. They also found diatoms primarily in nearshore waters. While our predicted concentrations of Hex.fuco (prymnesiophyes) and Fuco (diatoms) qualitatively reflect these broad-scale spatial trends, our estimated Hex.fuco concentrations were overall lower, and Fuco concentrations higher, than expected based on this field campaign. Furthermore, Qian et al. (2003) found the highest abundance of diatoms on the outer shelf to occur in summer, whereas our algorithms predicted the highest offshore concentrations of fucoxanthin in winter. However, this result is consistent with findings from other field campaigns (Chakraborty and Lohrenz 2015). Our algorithms' predictions of increasing relative Zea concentrations from coastal to offshore waters are consistent with a spatial trend described for prokaryotes, and our predicted low relative concentrations of But.fuco are consistent with overall low relative abundance of pelagophytes reported by Qian et al. (2003).

Lambert, Bianchia, and Santschi (1998) found cyanobacteria to be abundant in both coastal and offshore waters, diatoms (Fuco) to be abundant over the continental shelf, and pelagophytes (But.fuco) and prymnesiophytes (Hex.fuco) to be more abundant in slope waters. Our algorithms similarly predicted high concentrations of Fuco in coastal waters, high summer Zea concentrations throughout the NGOM, and high Hex.fuco concentrations offshore. They also predicted that But.fuco made up only a small fraction of the diagnostic pigments in most situations. Accordingly, Lambert et al. found pelagophytes to make up at most 20%, and often less, of the phytoplankton community at their sampling stations, and that But.fuco occurred in comparatively small concentrations even then. Our satellite-based results are hence also consistent with the findings of Lambert et al.'s field campaign.