remote sensing

MDPI

Article

Comparison of Machine Learning Techniques in Inferring Phytoplankton Size Classes

Shuibo Hu [1,2], Huizeng Liu [1,3,*], Wenjing Zhao [4], Tiezhu Shi [1], Zhongwen Hu [1], Qingquan Li [1] and Guofeng Wu [1,2,*]

[1] Key Laboratory for Geo-Environmental Monitoring of Coastal Zone of the National Administration of Surveying, Mapping and GeoInformation & Shenzhen Key Laboratory of Spatial Smart Sensing and Services & Research Institute for Smart Cities & Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China; hsb514@163.com (S.H.); tiezhushi@szu.edu.cn (T.S.); zwhoo@szu.edu.cn (Z.H.); liqq@szu.edu.cn (Q.L.)

[2] College of Life Sciences and Oceanography, Shenzhen University, Shenzhen 518060, China

[3] Department of Geography, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong, China

[4] South China Institute of Environmental Sciences, the Ministry of Environmental Protection of RPC, Guangzhou 510535, China; wenjing_jingjing@163.com

[*] Correspondence: HuizengLiu@life.hkbu.edu.hk (H.L.); guofeng.wu@szu.edu.cn (G.W.)

Received: 15 January 2018; Accepted: 5 March 2018; Published: 8 March 2018

Abstract: The size of phytoplankton not only influences its physiology, metabolic rates and marine food web, but also serves as an indicator of phytoplankton functional roles in ecological and biogeochemical processes. Therefore, some algorithms have been developed to infer the synoptic distribution of phytoplankton cell size, denoted as phytoplankton size classes (PSCs), in surface ocean waters, by the means of remotely sensed variables. This study, using the NASA bio-Optical Marine Algorithm Data set (NOMAD) high performance liquid chromatography (HPLC) database, and satellite match-ups, aimed to compare the effectiveness of modeling techniques, including partial least square (PLS), artificial neural networks (ANN), support vector machine (SVM) and random forests (RF), and feature selection techniques, including genetic algorithm (GA), successive projection algorithm (SPA) and recursive feature elimination based on support vector machine (SVM-RFE), for inferring PSCs from remote sensing data. Results showed that: (1) SVM-RFE worked better in selecting sensitive features; (2) RF performed better than PLS, ANN and SVM in calibrating PSCs retrieval models; (3) machine learning techniques produced better performance than the chlorophyll-a based three-component method; (4) sea surface temperature, wind stress, and spectral curvature derived from the remote sensing reflectance at 490, 510, and 555 nm were among the most sensitive features to PSCs; and (5) the combination of SVM-RFE feature selection techniques and random forests regression was recommended for inferring PSCs. This study demonstrated the effectiveness of machine learning techniques in selecting sensitive features and calibrating models for PSCs estimations with remote sensing.

Keywords: phytoplankton size classes; machine learning; feature selection; random forest; remote sensing

1. Introduction

Phytoplankton plays a critical role in ocean ecosystems and the global carbon cycle via carbon fixation during photosynthesis, and they account for up to 50% of the total primary production on Earth [1]. More importantly, phytoplankton can serve as a “biological pump” by moving fixed carbon into deep ocean [2]. One of the factors affecting carbon fixation and sinking rate is the size of the phytoplankton cell [3]. Phytoplankton size structures, expressed as phytoplankton size classes (PSCs), are divided into three size classes: microplankton (>20 μm), nanoplankton (2–20 μm), and picoplankton (<2 μm) [4]. PSCs are found to closely relate to phytoplankton functional types (PFTs) [5]. Therefore, synoptic mapping of PSCs has been recently targeted by ocean color remote sensing community [6].

The mechanism of PSCs retrieval from ocean color data lies in the fact that PSCs are closely associated with the phytoplankton abundance as well as inherent and apparent optical properties in waters [6–8]. Hitherto, several PSCs retrieval approaches have been developed [8], including abundance-based, spectral-based, and statistical-based ones [6]. Abundance-based methods employ chlorophyll *a* concentrations to infer phytoplankton size structure, because large phytoplankton cells are generally associated with high biomass and small cells with low biomass [7]. Spectral-based methods depend on the spectral shape of either phytoplankton absorption or particulate backscattering [9–12].

Machine learning techniques have also successfully applied to estimate PSCs and PFTs. For examples, Raitsos et al. [13] and Brewin et al. [8] applied artificial neural networks to retrieve PFTs and PSCs from bio-optical, spatial, temporal, and physical features; Organelli et al. [14] calibrated a partial least squares (PLS) models with in situ particulate absorption coefficients for PSCs retrieval in the Mediterranean Sea. In addition to the original spectral features of waters, their spectral derivatives and indices were also applied in model development, and feature selection techniques were applied to identify sensitive features for PSCs and PFTs estimations. For examples, Torrecilla et al. [15] found, through the sensitivity test in cluster analysis, that the second derivatives of original spectral features of waters worked better than band ratios and original features in discriminating phytoplankton pigment assemblages; and Li et al. [16] used support vector machine recursive feature elimination (SVM-RFE) to select sensitive spectral features and then applied them to develop PSCs estimation models with SVM regression.

Considering the successful applications of abovementioned methods, machine learning techniques may be promising for PSCs mapping over global oceans, while statistical PSCs modeling mainly involves sensitive feature selection and model calibration. However, few comprehensive comparisons of machine learning techniques for inferring PSCs with remote sensing data were carried out in the literature. Therefore, this study, using the NOMAD HPLC database and satellite-derived products, aimed to compare the effectiveness of machine learning techniques for selecting useful spectral features and developing PSCs retrieval models. Three feature selection algorithms, including genetic algorithm, successive projection algorithm and SVM-RFE, and four modeling techniques, including PLS, ANN, SVM and random forests, were tested. The results from this study would be a good reference for statistical estimations of PSCs with remote sensing techniques.

2. Materials and Methods

2.1. In-Situ Pigments

The global pigment data of the NOMAD HPLC database (https://seabass.gsfc.nasa.gov) was used in this study, and it includes high quality data selected for ocean color algorithm development [17]. This database includes 4459 pigment measurements in sea surface waters between 1991 and 2007, among which 3118 were sampled within the SeaWiFS acquisition period. The Diagnostic Pigment Analysis method, which was proposed by Vidussi et al. [18] and extended by Uitz et al. [19] and Hirata et al. [11], was applied to calculate the fractions of picoplankton (Fp), nanoplankton (Fn), and microplankton (Fm) [8], and it employed seven diagnostic pigments, including fucoxanthin, peridinin, 19′-hexanoyloxyfucoxanthin, 19′-butanoyloxyfucoxanthin, alloxanthin, total chlorophyll *b* and zeaxanthin, to infer PSCs. To control the quality of pigments measurement, only the samples with in situ measured chlorophyll *a* (Chl *a*) concentration exceeding three standard deviations of mean covariance for the sum of accessory pigments were discarded, because they were found to be highly correlated and covaried in a quasi-linear manner [20]. Finally, a total of 2871 samples were used in this study.

2.2. Satellite Data

SeaWiFS ocean color data were used in this study for inferring PSCs, because most of the NOMAD samples were collected during SeaWiFS acquisition period. The level three 8-day composite products (9 × 9 km^2 resolution) of SeaWiFS remote sensing reflectance (Rrs) at 412, 443, 490, 510, 555, and 670 nm, phytoplankton absorption coefficient at 443 nm (aph_443), particulate backscattering coefficient at 443 nm (bbp_443), near surface Chl *a*, photosynthetically active radiation (PAR) and optical aerosol thickness at 865 nm (T865) were downloaded from NASA Oceancolor website (https://oceancolor.gsfc.nasa.gov/). The Chl *a* concentration was estimated with the OCx band ratio algorithm merged with color index [21], and aph_443 and bbp_443 were retrieved with Generalized Inherent Optical Property [22].

Besides ocean color data, remotely sensed sea surface temperature (SST) and wind stress products were also incorporated in PSCs modelling. The nighttime Advanced Very High Resolution Radiometer (AVHRR) Pathfinder 5.0 daily of sea surface temperature data (4 × 4 km^2 resolution) were obtained from the NOAA website (https://data.nodc.noaa.gov/pathfinder/), and weekly composites of mean wind stress data derived from ER-2 (1° × 1° resolution) and QuikSCAT (0.5° × 0.5° resolution) were obtained from ftp://ftp.ifremer.fr/ifremer/cersat/products/gridded.

2.3. Procedure for Matching Satellite and In Situ data

To maximize the number of match-ups of in situ and satellite data, the in situ HPLC database was matched to SeWiFS level three 8-day products. A 3 × 3 window matching procedure adopted by Hu et al. [23] was used to extract pixel values around each sampling station. The mean of valid values within the window was calculated for each satellite-sensed parameter. The pixels with low quality were identified considering quality control flags or high T865 values, and removed from further analyses. The match-ups with incomplete satellite-sensed parameters were identified and eliminated. To further control the quality of match-ups, in situ measured vs. SeaWiFS-derived $R_{rs}(443)$ as well as in situ measured vs. SeaWiFS-derived Chl *a* concentrations were also used to reduce the error caused by temporal gaps. The match-ups with any SeaWiFS-derived parameter exceeding three mean covariance for in situ measured ones were eliminated. Finally, a total of 725 match-ups were selected in this study.

2.4. Feature Selection Techniques

Feature selection is a process of selecting a subset of relevant and sensitive features for model development, and it is helpful to simplify model and enhances generalization by reducing overfitting. Feature selection techniques are generally divided into filters, wrappers, and embedded methods, according to the relationship between feature selection and the modeling process [24]. Filters are independent from model calibration, and they often applied as a pre-processing method; however, they may fail to select the most sensitive features [24]. Wrappers consider modeling techniques as a black box, and just rely on the prediction performance to evaluate the usefulness of subsets of features [25]. Embedded methods perform feature selection in model training process, and select the features contributing the most to model accuracy. Two wrapper methods, including genetic algorithm (GA) and successive projection algorithm (SPA), and one embedded method, i.e., recursive feature selection (RFE), were applied in this study.

GA, a popular heuristic optimization technique, uses a probabilistic and non-local search process inspired by Darwin's natural selection theory [26]. In GA, a subset of features is encoded into a binary string called chromosomes, in which a binary code (1 or 0) for each feature represents selected or not. It starts with a set of randomly initiated chromosomes, selects chromosomes resulting in higher prediction performance, and then generates next generations with the selected population through crossover and mutation. The evolution was stopped at 50 generations, and the selected features of the fittest chromosome were identified. The prediction performance of each subset of features was evaluated using PLS regression. The selection process was repeated 100 times, and the selection frequencies of features were calculated and ranked in decreasing order to be successively adopted in modeling [27].

SPA, a forward selection technique, is devised to minimize the collinearity between selected features [28], and it uses a simple projection operation in vector space to obtain a subset of features with minimal collinearity. SPA starts with one feature, and then sequentially selects new feature having the maximum projection value on the orthogonal subspace of previously selected features [29]. This process is continued until the prediction performance obtained by PLS regression does not increase. In implementation, each feature was tested iteratively to be the first feature selected, and the sequential selection process was carried out for each initial feature. And, the subset with which PLS regression produces the lowest prediction error was selected as the optimal features.

RFE, a backward selection technique, first trains a model with all features, removes the feature contributing the least to model, and then retrains the model with the remaining features [30]. The procedure is repeated sequentially, and the number of features is determined according to the highest prediction performance. In this study, RFE based on support vector machine regression (SVM-RFE) was implemented. Feature contribution was evaluated according to the absolute of weight change with and without taking the feature into calculation. The weight value was calculated as: $\mathrm{W}^2 = (\sum \alpha * y * K(x,\ x))^2$ [31], where α are the support vectors, y is a variable to be modeled, and *K(x, x)* is the kernel function.

2.5. Model Development

Four supervised machine learning techniques, including PLS, ANN, SVM, and random forests (RF), were applied to calibrate remote sensing-based PSCs retrieval models.

PLS is a popular multivariate analysis technique in spectral modeling [32]. It holds a similar structure to principal component regression; however, it takes dependent variables into account when generating latent variables. Specifically, PLS regression searches for a set of latent variables that perform a simultaneous decomposition of independent and dependent variables with the constraint that these components explain the covariance between them as much as possible. Therefore, it can reduce dimensionality and computation time while avoiding multi-collinearity. SIMPLS [33] was used to implement PLS regression in this study, and leave-one-out cross-validation was used to determine optimal number of latent variables.

ANN is based on a collection of connected artificial neurons, and it stimulates human learning processes through establishing and reinforcing the connections between input and output [34]. Generally, ANN consists of input, hidden, and output layers, each with a set of interconnected neurons, and transforming from input data into output values. The connections between neurons were represented by a weight matrix, which represents the linkage between the input and output data. One popular ANN algorithm, feed forward neural networks with back-propagation, was implemented with 10 hidden neurons in this study. To train the ANN, the connection weight matrix was initiated with an arbitrary one. The output values were compared with the training data, and the estimation error was calculated. The estimation error was back-propagated to the network, and GA algorithm was used to optimize the weight matrix.

SVM regression is a kernel-based learning algorithm, it first maps training data into a new hyperspace using a kernel function, and then constructs an optimal hyper plane fitting training data [35]. The major advantage of SVM lies in its complex fitting ability for non-linear data. Least squares SVM (LSSVM) is a modified version of SVM, which solves multivariate modeling by applying least squares error in training error function. In LSSVM, a linear estimation is done in a kernel-induced feature space ($y\ =\ \omega^T\phi(x) +\ b$), where $\phi(x)$ denotes the feature map. Therefore, LSSVM has a more simplified training process than SVM. In this study, LSSVM was implemented using the LS-SVM lab toolbox [36], and also the most popular radial basis function (RBF, exp $(-\|x_i - x\|^2/2\sigma^2)$ because of its adaptability to non-linear data, where σ^2. is the width of Gaussian function.

RF is a tree-based non-parametric ensemble learning technique. It produces a final prediction by combining predictions from many individual decision trees [37], which are created by drawing a subset of training samples through a bagging approach. In RF, about two thirds of the samples (referred to as in-bag samples) are used to train the trees. The remaining samples (referred to as out-of-bag samples) are used in the internal cross-validation, which produces an error estimate called out-of-bag error. Therefore, RF are robust to noise and resistant to overfitting. The RF was implemented using the package provided by [38], with 500 trees in this study.

2.6. Modeling Framework

Three main steps are need to develop a statistical PSCs algorithm [16]: (1) constructing input features for selection; (2) selecting sensitive features using feature selection techniques; and (3) calibrating models using statistical methods. In this study, four types of data were used, including ocean color (R_{rs} at 412, 443, 490, 510, 555, and 670 nm, band indices, a_{ph}_443 and b_{bp}_443), temporal (month), biological (Chl *a*) and physical (PAR, SST, and wind stress) data. In addition to six original R_{rs}, six continuum-removed spectra and twenty spectral curvature were also constructed [16], which were denoted CR(λ) and CV(λ_1, λ_2, λ_3) with $\lambda_1 < \lambda_2 < \lambda_3$, respectively. Thus, a total of 39 features were incorporated into feature selections. The modeling framework used in this study is illustrated in Figure 1:

(1) Two-thirds of the samples (483 samples) were randomly selected from NOMAD match-ups as a training dataset, and the left 241 samples were used as a validation dataset. Each input feature and phytoplankton size class was transformed to be dimensionless by standardization using the mean and standard deviation of training set.

(2) Sensitive features were selected for each phytoplankton size class (i.e., Fm, Fn and Fp) using GA, SPA, and SVM-RFE, respectively. During feature selection, 10-fold cross validation was implemented for model selection, and Akaike information criterion [39] was used to select optimal features.

(3) Statistical models were calibrated for each phytoplankton size class by using above-selected features with PLS, SVM, ANN and RF, respectively, and 10-fold cross validation was carried out to evaluate calibration performance.

(4) After one model was calibrated with the training dataset, the validation dataset was used to test its performance.

(5) To ensure the robustness of results, steps (1)–(4) were repeated 100 times, and the calibration and validation results from each iteration were compiled together for assessing model performance.

2.7. Result Comparison and Interpretation

To evaluate model calibration and validation performance, the coefficient of determination (R^2), root mean square error (RMSE), mean absolute percentage error (MAPE), and relative RMSE (RRMSE) were calculated from the 10-fold cross validation and validation results [40].

For evaluating the sensitive features selected for each phytoplankton size class by each feature selection technique, the mean number of selected features over 100 iterations and the selected frequency for each feature were statistically calculated, and the features with selected frequency over 50% were identified as frequently selected features and analyzed. For each phytoplankton size class, the features frequently selected by at least two feature selection techniques were considered as useful ones. The correlations between these useful features and three PSCs were analyzed to figure out the relationships between these variables.

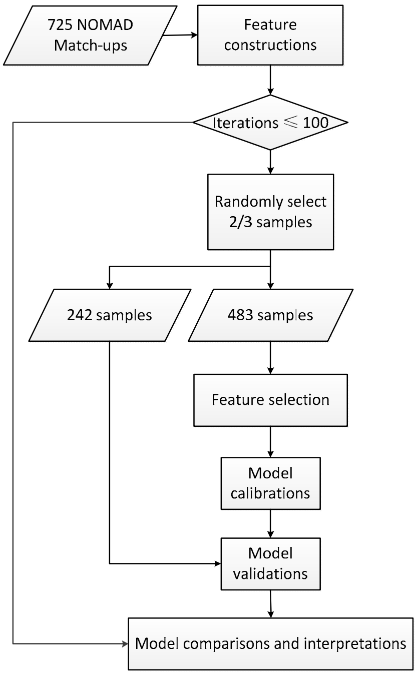

Figure 1. Framework for model development.

3. Results

3.1. Microphytoplankton

The cross validation and independent validation results for Fm from the combinations of three feature selection algorithms and four modeling techniques, as well as three baseline methods are shown in Table 1. Cross and independent validations produced generally consistent performances. The models calibrated with the features selected by GA, SPA and SVM-RFE, and all the features showed no obvious difference. However, the prediction performances of these four modeling techniques showed a clear pattern (RF > SVM > ANN > PLS) with RF being the best and PLS the worst, irrespective of feature selection algorithms. RF explained 76–78% of the variation of validation dataset with a RMSE of 0.13–0.14, and SVM, ANN, and PLS explained 73–75%, 72–73% and 64–65% of the variation, respectively.

Table 1. Model performance of cross validation (CV) and independent validation (V) for microplankton size fractions obtained by different combinations of feature selection and modelling techniques. Feature selection techniques included genetic algorithm (GA), successive projection algorithm (SPA), and recursive feature elimination based on support vector machine regression (SVM-RFE). And, the modeling techniques included partial least square (PLS) regression, artificial neural network (ANN), support vector machine (SVM), and random forests (RF).

Feature Selection	Modeling Techniques	R^2_{CV}	$RMSE_{CV}$	$RRMSE_{CV}$	R^2_V	$RMSE_V$	$RRMSE_V$
GA	PLS	0.66	0.16	40.13%	0.65	0.16	40.59%
	ANN	0.72	0.15	36.56%	0.72	0.15	36.675
	SVM	0.75	0.14	35.04%	0.74	0.14	35.43%
	RF	**0.77**	**0.13**	**33.54%**	**0.77**	**0.13**	**33.47%**
SPA	PLS	0.66	0.16	40.17%	0.65	0.16	40.72%
	ANN	0.72	0.15	36.75%	0.72	0.15	36.94%
	SVM	0.74	0.14	35.22%	0.73	0.14	36.01%
	RF	**0.76**	**0.14**	**34.06%**	**0.76**	**0.14**	**33.86%**

Table 1. *Cont.*

Feature Selection	Modeling Techniques	R^2_{CV}	$RMSE_{CV}$	$RRMSE_{CV}$	R^2_V	$RMSE_V$	$RRMSE_V$
SVM-RFE	PLS	0.64	0.16	41.13%	0.64	0.17	41.11%
	ANN	0.72	0.15	36.50%	0.72	0.15	36.68%
	SVM	0.75	0.14	34.55%	0.74	0.14	35.82%
	RF	**0.77**	**0.13**	**33.17%**	**0.78**	**0.13**	**33.14%**
– [a]	PLS	0.65	0.17	40.62%	0.65	0.16	40.49%
	ANN	0.72	0.15	36.49%	0.73	0.14	36.18%
	SVM	0.75	0.14	34.58%	0.75	0.14	34.95%
	RF	**0.77**	**0.13**	**33.53%**	**0.77**	**0.13**	**33.41%**
– [b]	Three-component	0.50	0.20	50.33%	0.49	0.20	50.78%
SVM-RFE [c]	SVM	0.61	0.17	44.72%	0.59	0.18	44.59%

[a] refers to the first baseline method, in which all of the features were used in modeling; [b] indicates the second baseline method, in which only chlorophyll *a* concentration was used; and [c] is the third baseline method, in which only ocean color data were input into SVM-RFE for feature selection.

Figure 2 illustrates and compares the results for microplankton obtained by RF models calibrated using the features from SVM-RFE, SVM models calibrated using the features from SVM-RFE, three-component model and SVM model based on only ocean color features. The former two models showed obvious better performance than the latter two. The RF models, with a R^2 of 0.78, a RMSE of 0.13 and a regression line of y = 0.85x + 0.06, produced similar results with those of SVM models calibrated using the features from SVM-RFE (R^2_V = 0.74, RMSE = 0.14 and a regression line: y = 0.89x + 0.04). The three-component method produced the worst result with a slope of 0.69 and a bias of 0.17 for its regression line, overestimating low fractions, underestimating high ones, and producing scattering estimations for middle values. The mean parameter values for three-component model are shown in Table 2. The SVM model based on only ocean color data produced the worse result with R^2_V = 0.59, RMSE = 0.18, and RRMSE = 44.59%. Figure 2d also proved the advantage of RF methods, with more samples peaking at low relative errors.

Table 2. Mean parameter values obtained for three-component method from the dataset used in this study.

Population	Maximum Chl *a* for Given Population	Initial Slope
Combined nano- and picoplankton	0.766 mg/m^3 ($C^m_{p,n}$)	1.009 ($S_{p,n}$)
Picoplankton	0.102 mg/m^3 (C^m_p)	6.791 (S_p)

The sensitive features and their selected frequencies in 100 iterations are listed in Table 3. On average, GA, SPA, and SVM-RFE selected 10.83, 10.05, and 9.29 features in each modeling with eight, eight, and nine sensitive features identified, respectively. All of the three feature selection algorithms identified SST, CV(490, 510, 555) and wind stress as useful features. Both GA and SPA selected Chl *a*, a_{ph}_443 and CR(443) as sensitive features, and both SPA and SVM-RFE selected CV(443, 490, 510).

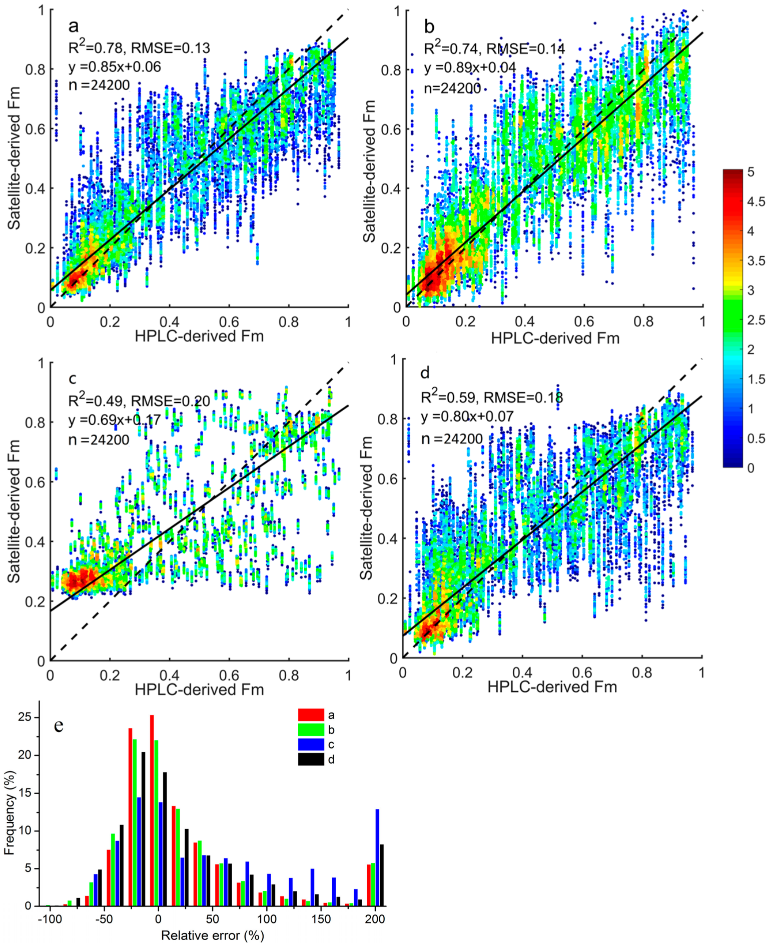

Figure 2. Scatter plots of satellite-derived versus high performance liquid chromatography (HPLC) microplankton size fractions (Fm): (**a**) random forests using features selected with SVM-RFE, (**b**) SVM using features selected with SVM-RFE, (**c**) SVM using ocean color features selected with SVM-RFE, and (**d**) three-component method. The dashed line is a 1:1 line, and the solid line is a regression line. Plot (**e**) shows the frequency distributions of their relative errors, and the numbers along the color ramp indicates the pixel density after log transformation (y = ln(x)).

Table 3. Sensitive features for retrieving microphytoplankton size fractions selected, respectively, by genetic algorithm (GA), successive projection algorithm (SPM) and recursive feature elimination based on support vector machine regression (SVM-RFE).

GA		**SPA**		**SVM-RFE**	
Features	**Frequency**	**Features**	**Frequency**	**Features**	**Frequency**
Chl-a	100	**CV(490, 510, 555)**	100	PAR	100
Wind stress	100	Chl-a	100	Month	99
SST	100	**SST**	100	**SST**	98
CV(490, 510, 555)	99	**Wind stress**	97	**CV(490, 510, 555)**	97
CR(510)	85	CV(443, 490, 510)	75	CV(443, 490, 555)	85
CR(443)	64	a_{ph}_443	61	R_{rs}(510)	83
R_{rs}(670)	53	CR(555)	52	R_{rs}(490)	65
a_{ph}_443	53	CR(443)	51	**Wind stress**	56
				CV(443, 490, 510)	53

Note: the features selected by all of the three algorithms are shown in bold.

3.2. Nanoplankton

The cross validation and independent validation results for nanoplankton size classes obtained by different methods are shown in Table 4. Similar to the results of Fm, the cross validation and independent validation for Fn produced generally consistent performances. Irrespective of modeling technique, the models calibrated with the features from SVM-RFE performed the best, followed by GA and SPA. The models calibrated with all features produced slightly worse performance than those with the features selected by SVM-RFE. The prediction performance of these four modeling techniques showed a consistent pattern (RF > SVM > ANN > PLS), with RF being the best, SVM second, and PLS the worst. RF explained 49–56% of the variation of validation dataset with a RMSE of 0.11–0.12, while SVM, ANN and PLS explained 43–50%, 39–45%, and 26–31% of the variation, respectively.

Table 4. Model performances of cross validation (CV) and independent validation (V) for nanoplankton size fractions obtained by different combinations of feature selection and modelling techniques. Feature selection techniques included genetic algorithm (GA), successive projection algorithm (SPA), and recursive feature elimination based on support vector machine regression (SVM-RFE). The modeling techniques included partial least square (PLS) regression, artificial neural network (ANN), support vector machine (SVM), and random forests (RF).

Feature Selection	Modeling Techniques	R^2_{CV}	$RMSE_{CV}$	$RRMSE_{CV}$	R^2_{V}	$RMSE_{V}$	$RRMSE_{V}$
GA	PLS	0.34	0.14	45.48%	0.31	0.14	46.12%
	ANN	0.44	0.13	42.58%	0.43	0.13	42.67%
	SVM	0.49	0.12	40.50%	0.47	0.12	41.18%
	RF	**0.54**	**0.12**	**38.75%**	**0.53**	**0.12**	**38.67%**
SPA	PLS	0.34	0.14	45.60%	0.31	0.14	46.38%
	ANN	0.41	0.13	43.72%	0.39	0.13	43.93%
	SVM	0.46	0.13	41.90%	0.43	0.13	42.54%
	RF	**0.50**	**0.12**	**40.39%**	**0.49**	**0.12**	**40.40%**
SVM-RFE	PLS	0.27	0.15	48.53%	0.26	0.15	48.80%
	ANN	0.46	0.13	41.92%	0.45	0.13	41.91%
	SVM	0.52	0.12	39.62%	0.48	0.12	40.74%
	RF	**0.56**	**0.11**	**37.82%**	**0.56**	**0.11**	**37.73%**
–[a]	PLS	0.32	0.14	46.38%	0.31	0.14	46.29%
	ANN	0.45	0.13	42.06%	0.45	0.13	41.86%
	SVM	0.52	0.12	39.59%	0.50	0.12	39.83%
	RF	**0.54**	**0.12**	**38.58%**	**0.54**	**0.12**	**38.48%**
–[b]	Three-component	0.18	0.16	50.32%	0.17	0.16	52.17%
SVM-RFE[c]	SVM	0.41	0.13	43.57%	0.37	0.13	44.76%

Note: [a] refers to the first baseline method, in which all of the features were used in modeling; [b] indicates the second baseline method, in which only chlorophyll-a concentration was used; and [c] is the third baseline method, in which only ocean color data were input into SVM-RFE for feature selection.

The sensitive features for nanoplankton size fractions and their selected frequencies in 100 iterations are listed in Table 5. On average, GA, SPA and SVM-RFE selected 9.03, 7.79, and 8.77 features in each modeling with eight, five and eight sensitive features identified, respectively. All of the three feature selection algorithms identified month and CV(490, 510, 555) as useful features. Both GA and SPA selected Chl *a*, CR(490) and a_{ph}_443 as sensitive features, while both GA and SVM-RFE selected wind stress and R_{rs}(490).

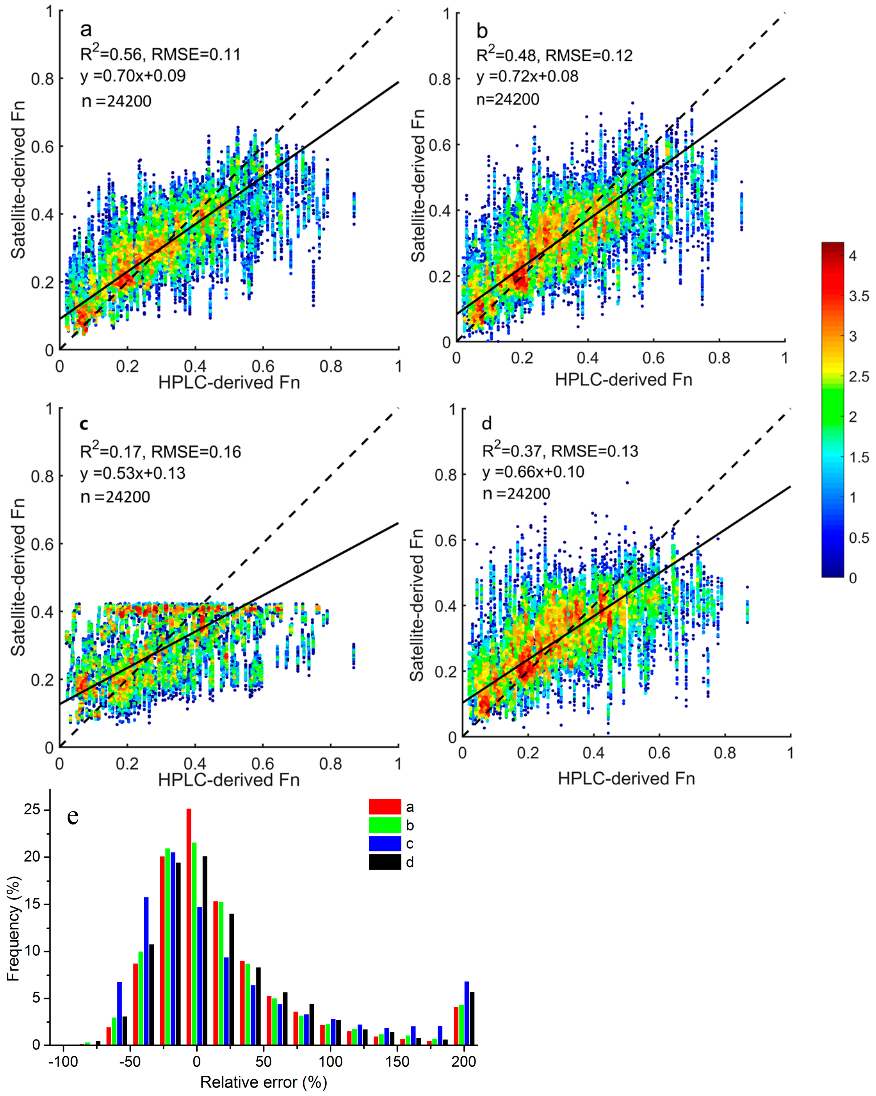

Figure 3. Scatter plots of satellite-derived versus high performance liquid chromatography (HPLC) nanoplankton size fractions (Fn): (**a**) random forests using features selected with SVM-RFE, (**b**) SVM using features selected with SVM-RFE, (**c**) SVM using ocean color features selected with SVM-RFE, and (**d**) three-component method. The dashed line is a 1:1 line, and the solid is a regression line. Plot (**e**) shows the frequency distributions of their relative errors, and the numbers along the color ramp indicates the pixel density after log transformation (y = ln(x)).

Table 5. Sensitive features for retrieving nanoplankton size fractions selected, respectively, by genetic algorithm (GA), successive projection algorithm (SPM) and recursive feature elimination based on support vector machine regression (SVM-RFE).

GA		SPA		SVM-RFE	
Features	**Frequency**	**Features**	**Frequency**	**Features**	**Frequency**
Month	100	CR(490)	100	R_{rs}(490)	100
Chl-a	100	CV(490, 510, 555)	100	Month	100
CR(490)	95	Chl-a	100	PAR	100
CV(490, 510, 555)	89	a_{ph}_443	97	Wind stress	99
R_{rs}(412)	86	Month	64	CV(443, 490, 555)	95
a_{ph}_443	74			SST	95
Wind stress	62			CV(490, 510, 555)	92
R_{rs}(490)	59			CR(555)	64

Note: The features selected by all of the three algorithms are shown in bold.

3.3. Picoplankton

All of the cross validation and independent validation results for picoplankton size classes obtained by different methods are shown in Table 6. The cross validation and independent validation obtained by each method for Fp produced generally consistent performance. For each modeling technique, the models calibrated features from GA, SPA, and SVM-RFE, and all features produced similar results, in which SVM-RFE showed slightly better performance. The prediction performance of these four modeling techniques showed a consistent pattern (RF > SVM > ANN > PLS) with RF being the best, SVM the second and PLS the worst. RF explained 80–82% of the variation of validation dataset with a RMSE of 0.10–0.11, while SVM, ANN, and PLS explained 77–80%, 76–77%, and 73% of the variation, respectively.

Table 6. Model performances of cross validation (CV) and independent validation (V) for picoplankton size fractions obtained by different combinations of feature selection and modelling techniques. Feature selection techniques included genetic algorithm (GA), successive projection algorithm (SPA) and recursive feature elimination based on support vector machine regression (SVM-RFE). The modeling techniques included partial least square (PLS) regression, artificial neural network (ANN), support vector machine (SVM), and random forests (RF).

Feature Selection	Modeling Techniques	R^2_{CV}	$RMSE_{CV}$	$RRMSE_{CV}$	R^2_V	$RMSE_V$	$RRMSE_V$
GA	PLS	0.74	0.12	38.02	0.73	0.12	37.94
	ANN	0.77	0.12	36.55	0.77	0.12	36.38
	SVM	0.79	0.11	35.15	0.78	0.11	35.31
	RF	**0.80**	**0.11**	**34.51**	**0.81**	**0.11**	**34.21**
SPA	PLS	0.74	0.12	37.98	0.73	0.12	37.94
	ANN	0.76	0.12	37.01	0.76	0.12	36.76
	SVM	0.78	0.11	35.97	0.77	0.12	36.17
	RF	**0.80**	**0.11**	**35.18**	**0.80**	**0.11**	**34.90**
SVM-RFE	PLS	0.72	0.13	38.80	0.73	0.13	38.27
	ANN	0.77	0.12	36.13	0.77	0.12	35.77
	SVM	0.80	0.11	33.71	0.80	0.11	34.14
	RF	**0.82**	**0.11**	**33.45**	**0.82**	**0.10**	**33.09**
_ [a]	PLS	0.73	0.13	38.68	0.73	0.13	38.30
	ANN	0.76	0.12	36.92	0.76	0.12	36.36
	SVM	0.79	0.11	34.56	0.79	0.11	34.63
	RF	**0.80**	**0.11**	**34.77**	**0.80**	**0.11**	**34.39**
_ [b]	Three-component	0.50	0.18	54.38	0.50	0.18	0.54
SVM-RFE [c]	SVM	0.67	0.14	44.68	0.65	0.14	45.35

Note: [a] refers to the first baseline method, in which all of the features were used in modeling; [b] indicates the second baseline method, in which only chlorophyll *a* concentration was used; and [c] is the third baseline method, in which only ocean color data were input into SVM-RFE for feature selection.

Figure 4 illustrates and compares the results for Fp obtained by RF model calibrated with the features from SVM-RFE, SVM models calibrated using the features from SVM-RFE, the three-component model, and the SVM model based on only ocean color features. All of these four methods tended to underestimate high Fp. Comparatively, the three-component method produced the worst result, producing high deviations from 1:1 line especially for low values (Figure 4b). The SVM model based on only ocean color data produced acceptable performance with $R^2_V = 0.65$, RMSE = 0.14, and a regression line of y = 0.82 + 0.06. It, however, obviously overestimated very low fractions. RF model performed the best with a R^2_V of 0.82, a RMSE of 0.10, and a $RRMSE_V$ of 33.09%. Similar and slightly worse results ($R^2_V = 0.80$, RMSE = 0.11 and a regression line: y = 0.91x + 0.03) were obtained by SVM models calibrated using the features selected by SVM-RFE. Figure 4d also demonstrates the advantage of RF, producing the highest frequency at low relative errors; however, about 10–18% of samples were estimated with over 200% relative errors.

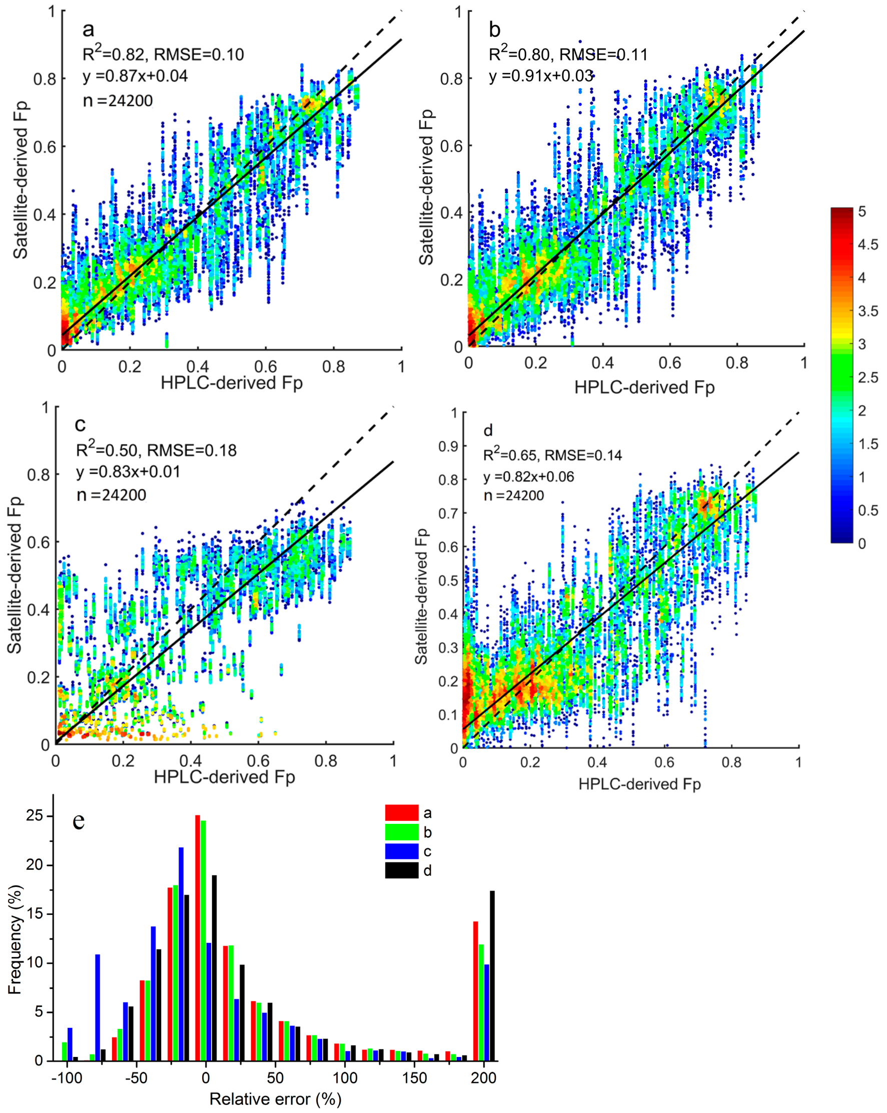

Figure 4. Scatter plots of satellite-derived versus high performance liquid chromatography (HPLC) picoplankton size fractions (Fp): (**a**) random forests using features selected with SVM-RFE, (**b**) SVM using features selected with SVM-RFE, (**c**) SVM using ocean color features selected with SVM-RFE, and (**d**) three-component method. The dash line is 1:1 line, and the solid is regression line. Plot (**e**) shows the frequency distributions of their relative errors, and the numbers along the color ramp indicates the pixel density after log transformation (y = ln(x)).

The sensitive features for picoplankton size fractions and their selected frequencies in 100 iterations are listed in Table 7. On average, GA, SPA, and SVM-RFE selected 9.29, 7.32, and 10.15 features in each modeling with seven, five, and 10 sensitive features identified, respectively. All of the three feature selection algorithms identified wind stress, CV(490, 510, 555) and SST as useful features. Both GA and SPA selected CR(490) as sensitive features, while both GA and SVM-RFE selected R_{rs}(412) and CV(443, 490, 555).

Table 7. Sensitive features for retrieving picoplankton size fractions selected, respectively, by genetic algorithm (GA), successive projection algorithm (SPM), and recursive feature elimination based on support vector machine regression (SVM-RFE).

GA		SPA		SVM-RFE	
Features	**Frequency**	**Features**	**Frequency**	**Features**	**Frequency**
Wind stress	100	SST	100	CV(490, 510, 555)	100
SST	100	CR(490)	97	Month	100
CR(490)	97	Wind stress	96	PAR	100
CV(490, 510, 555)	73	CV(490, 510, 555)	85	SST	100
Month	64	CV(443, 490, 510)	58	Wind stress	97
CV(443, 490, 555)	57			CV(443, 490, 555)	74
$R_{\rm rs}$(412)	53			$R_{\rm rs}$(412)	71
				CR(555)	68
				CR(510)	62
				$R_{\rm rs}$(490)	54

Note: The features selected by all of the three algorithms are shown in bold.

3.4. Correlation Analysis

The correlations of the selected features against PSCs varied greatly (Table 8). There were medium to high correlations between each ocean color feature and some other ocean color features, except for CV(443, 490, 510). For example, R_{rs}(412) was highly correlated to R_{rs}(412) and CR(510), and moderately correlated to CR(490) and CV(443, 490, 555). The Chl *a* concentration was highly correlated to a_{ph}_443, and moderately correlated to R_{rs}(412), CV(443, 490, 555), and CV(490, 510, 555). The Fm was moderately correlated to R_{rs}(412), CV(443, 490, 555), CV(490, 510, 555), Chl *a*, SST, and Fn, and highly and negatively correlated to Fp with a correlation coefficient of 0.80. The Fn was only found to be moderately correlated to CV(490, 510, 555) with a correlation coefficient of 0.50. There existed moderate correlations of Fp against R_{rs}(412), R_{rs}(490), CR(510), CV(443, 490, 555), and SST.

Table 8. Correlation matrix showing the relationships of some selected features against phytoplankton size classes.

	1. R_{rs}(412)	**2. R_{rs}(490)**	**3. CR(490)**	**4. CR(510)**	**5. CV(443, 490, 510)**	**6. CV(443, 490, 555)**	**7. CV(490, 510, 555)**	**8. Month**	**9. $a_{ph_}$443**	**10. Chl-a**	**11. Wind Stress**	**12. SST**	**13. Fm**	**14. Fn**	**15. Fp**
1.	1.00														
2.	0.85	1.00													
3.	−0.56	−0.38	1.00												
4.	−0.81	−0.56	0.79	1.00											
5.	−0.35	−0.34	0.24	0.27	1.00										
6.	0.67	0.47	−0.23	−0.44	0.23	1.00									
7.	0.36	0.24	0.04	−0.10	−0.12	0.65	1.00								
8.	−0.08	−0.10	−0.02	0.01	0.06	−0.01	−0.10	1.00							
9.	−0.36	−0.24	−0.18	0.05	−0.19	−0.65	−0.62	0.07	1.00						
10.	−0.53	−0.33	0.03	0.22	0.01	−0.77	−0.77	0.04	0.84	1.00					
11.	−0.16	−0.21	0.06	0.14	0.13	−0.03	0.11	−0.18	−0.05	−0.05	1.00				
12.	0.42	0.41	−0.31	−0.44	−0.14	0.17	−0.12	0.31	−0.02	−0.07	−0.41	1.00			
13.	−0.64	−0.49	0.29	0.48	0.07	−0.69	−0.51	−0.07	0.46	0.62	0.10	−0.50	1.00		
14.	0.02	0.01	0.21	0.19	0.05	0.30	0.50	−0.10	−0.36	−0.46	0.12	−0.13	−0.51	1.00	
15.	0.72	0.57	−0.48	−0.68	−0.11	0.58	0.23	0.15	−0.28	−0.40	−0.20	0.67	−0.80	−0.11	1.00

Note: r $\geq$ 0.50 are shown in bold.

4. Discussion

This study demonstrated the effectiveness of machine learning techniques in inferring phytoplankton size classes from satellite-sensed data. A stable rank order was found among four modeling techniques (RF > SVM > ANN > PLS) in terms of their prediction performance for inferring PSCs. RF performed the best, and was slightly better than SVM, followed by ANN. The better performance achieved by RF could be attributed to the combination of multiple diverse individual decision trees. Additionally, the out-of-bag error estimation used in RF modeling guarantees its generalization and resistance to overfitting [37,41]. The good results obtained by SVM should lie in its complex fitting properties, even for non-linear data, through RBF kernel mapping [42]. Some studies indicated that RF performed better than SVM, while contrary results could also be found in the literature [41,43,44]. This might be explained by the fact that SVM worked better for small sample size, while RF was thought to be more stable and reliable for large and high dimensional datasets. Considering the growing numbers of PSCs databases [8], RF may be a better choice for PSCs estimations globally. However, PLS was not recommended due to its obvious lower accuracy.

This study emphasized the usefulness of feature selection algorithms in selecting sensitive features for PSC inferring. Although the models calibrated with all variables produced equivalent prediction performance, feature selection techniques could dramatically reduce model complexity by selecting a few sensitive features. The importance of ecological variables for PSCs estimations were highlighted, and SST, wind stress and temporal variables were among the features frequently selected, because these factors could directly affect phytoplankton growth and reproduction [13]. Additionally, the less accurate predictions obtained by the models calibrated with only ocean color data also justified their importance. The geographic information was not incorporated in modeling following the suggestion by Raitsos and Lavender [13], since a global PSCs database covering oceans worldwide is still not available.

The SVM-RFE used in this study showed its advantage over GA and SPA in selecting sensitive PSCs features. This might be explained by: (1) as an embedded method, the feature ranking in SVM-RFE was more effective, because it directly evaluates feature importance according to its contribution to the model [30]; and (2) SVM was more effective than PLS regression [42,45], which was used as the learning machine in GA and SPA. Theoretically, SVM can also be used as a learning machine in GA; however, its efficiency would be reduced further, since it demands constant iterations to optimize initial random chromosomes to identify sensitive features [26,27]. Comparatively, SPA produced the worst performance, which might be explained by the fewer features it selected for modeling. Moreover SPA was designed to eliminate collinearities among features, which may exclude some useful and include some non-correlated but uninformative variables [27]. For example, CV(443, 490, 510), selected by SPA both for microplankton and picoplankton, produced very low correlation to other features as well as PSCs.

This study indicated that the three-component method based on Chl *a* alone could not achieve accurate estimations of PSCs. Similar result was also found by Alvain et al. [46] for PFTs. However, Chl-a still plays important roles in PSCs retrieval [13]. Even though Chl *a* was not included in modeling, the features closely related to it might be selected instead. For examples, spectral curvatures, like CV(443, 490, 555) and CV(490, 510, 555), could be used to estimate Chl *a* [47]; and a_{ph}_443, instead of Chl *a*, was also used in some abundance-based models [8]. Moreover, the results shown in Table 7 also supported this statement, since many prominent features were closely related to Chl *a*. The mean parameters for the three-component model in our study was different from those of Brewin et al. [7], which could be explained by the different dataset used to fit models [48]. Considering the parameters obtained from different studies (refers to Table 4.2 in [6]), the parameters obtained in this study (Table 2) should be valid.

We found that microplankton and picoplankton fractions could be estimated accurately from space, and inferring nanoplankton fractions still appeared to be challenging. This was consistent with the findings by Li et al. [16]. Such results could be explained by several reasons: (1) the microplankton and picoplankton abundances tend to increase and decrease monotonically, respectively, as a function of total Chl *a*, while the Fn appears to first increase and then decrease with Chl *a* [6,49,50]; (2) several features have been found moderately correlated to Fm and Fp, while only one with Fn; (3) Fn shows narrow range of variations than Fm and Fp, which might partially account for its low determination of coefficients; and (4) the determination deviations introduced by HPLC should also be considered, since some types phytoplankton could be found both in picoplankton and nanoplankton [6,51]. The samples with relative errors >200% for Fp were largely accounted for by those with low picoplankton fractions (<2%), since a tiny absolute deviation would also tend to result in a large relative error.

5. Conclusions

Three feature selection algorithms (GA, SPA, and SVM-RFE) were applied to select the features sensitive to phytoplankton size classes from a total of 39 ocean color, biological, physical, and temporal variables, and four modeling techniques (PLS, ANN, SVM, and RF) were used to calibrate PSCs retrieval models. The embedded feature selection method, SVM-RFE, worked better than the other two wrapper methods (i.e., GA and SPA), and random forests produced the highest prediction performance. Therefore, the combination of SVM-RFE and RF in further applications was recommended. Moreover, besides popular ocean color data, the satellite-sensed ecological factors were found useful for PSCs inferring.

Acknowledgments: The research was supported by the National Key R&D Program of China (No. 2017YFC0506200), Basic Research Program of Shenzhen Science and Technology Innovation Committee (No. JCYJ20151117105543692), National Natural Science Foundation of China grants (Grant Nos. 41606199, 41506202), The General Financial Grant from the China Postdoctoral Science Foundation (No. 2016M592521), the Open Project Program of State Key Laboratory of Tropical Oceanography, South China Sea Institute of Oceanology, Chinese Academy of Sciences (Project No. LTO1611), and Shenzhen Future Industry Development Funding Program (No. 201507211219247860). We thank NASA GSFC for providing in situ data and SeaWiFS products, thank NOAA NODC for providing SST data, and thank CERSAT for providing wind stress products.

Author Contributions: All authors conceived and designed the study. Shuibo Hu, Huizeng Liu, and Guofeng Wu made substantial contributions to experiments design. Shuibo Hu, Huizeng Liu, Wenjing Zhao and Tiezhu Shi implemented the experiments. All authors discussed the basic structure of the manuscript. Shuibo Hu and Huizeng Liu finished the first draft. Guofeng Wu, Zhongwen Hu, and Qingquan Li reviewed and edited the draft. All authors read and approved the submitted manuscript, agreed to be listed, and accepted the version for publication.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Field, C.B.; Behrenfeld, M.J.; Randerson, J.T.; Falkowski, P. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **1998**, *281*, 237–240. [CrossRef] [PubMed]
2. Longhurst, A.R.; Glen Harrison, W. The biological pump: Profiles of plankton production and consumption in the upper ocean. *Prog. Oceanogr.* **1989**, *22*, 47–123. [CrossRef]
3. Boyd, P.; Newton, P. Evidence of the potential influence of planktonic community structure on the interannual variability of particulate organic carbon flux. *Deep Sea Res.* **1995**, *42*, 619–639. [CrossRef]
4. Sieburth, J.M.; Smetacek, V.; Lenz, J. Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions. *Limnol. Oceanogr.* **1978**, *23*, 1256–1263. [CrossRef]
5. Nair, A.; Sathyendranath, S.; Platt, T.; Morales, J.; Stuart, V.; Forget, M.-H.; Devred, E.; Bouman, H. Remote sensing of phytoplankton functional types. *Remote Sens. Environ.* **2008**, *112*, 3366–3375. [CrossRef]
6. Bracher, A.; Bouman, H.A.; Brewin, R.J.; Bricaud, A.; Brotas, V.; Ciotti, A.M.; Hardman-Mountford, N.J. Obtaining phytoplankton diversity from ocean color: A scientific roadmap for future development. *Front. Mar. Sci.* **2017**, *4*, 55. [CrossRef]
7. Brewin, R.J.; Sathyendranath, S.; Hirata, T.; Lavender, S.J.; Barciela, R.M.; Hardman-Mountford, N.J. A three-component model of phytoplankton size class for the Atlantic Ocean. *Ecol. Model.* **2010**, *221*, 1472–1483. [CrossRef]

8. Brewin, R.J.; Hardman-Mountford, N.J.; Lavender, S.J.; Raitsos, D.E.; Hirata, T.; Uitz, J.; Devred, E.; Bricaud, A.; Ciotti, A.; Gentili, B. An intercomparison of bio-optical techniques for detecting dominant phytoplankton size class from satellite remote sensing. *Remote Sens. Environ.* **2011**, *115*, 325–339. [CrossRef]
9. Uitz, J.; Huot, Y.; Bruyant, F.; Babin, M.; Claustre, H. Relating phytoplankton photophysiological properties to community structure on large scales. *Limnol. Oceanogr.* **2008**, *53*, 614–630.
10. Kostadinov, T.; Siegel, D.; Maritorena, S. Retrieval of the particle size distribution from satellite ocean color observations. *J. Geophys. Res.* **2009**, *114*. [CrossRef]
11. Hirata, T.; Aiken, J.; Hardman-Mountford, N.; Smyth, T.; Barlow, R. An absorption model to determine phytoplankton size classes from satellite ocean colour. *Remote Sens. Environ.* **2008**, *112*, 3153–3159. [CrossRef]
12. Lin, J.; Cao, W.; Zhou, W.; Sun, Z.; Xu, Z.; Wang, G.; Hu, S. Novel method for quantifying the cell size of marine phytoplankton based on optical measurements. *Opt. Express* **2014**, *22*, 10467–10476. [CrossRef] [PubMed]
13. Raitsos, D.E.; Lavender, S.J.; Maravelias, C.D.; Haralabous, J.; Richardson, A.J.; Reid, P.C. Identifying four phytoplankton functional types from space: An ecological approach. *Limnol. Oceanogr.* **2008**, *53*, 605–613. [CrossRef]
14. Organelli, E.; Bricaud, A.; Antoine, D.; Uitz, J. Multivariate approach for the retrieval of phytoplankton size structure from measured light absorption spectra in the mediterranean sea (boussole site). *Appl. Opt.* **2013**, *52*, 2257–2273. [CrossRef] [PubMed]
15. Torrecilla, E.; Stramski, D.; Reynolds, R.A.; Millán-Núñez, E.; Piera, J. Cluster analysis of hyperspectral optical data for discriminating phytoplankton pigment assemblages in the open ocean. *Remote Sens. Environ.* **2011**, *115*, 2578–2593. [CrossRef]
16. Li, Z.; Li, L.; Song, K.; Cassar, N. Estimation of phytoplankton size fractions based on spectral features of remote sensing ocean color data. *J. Geophys. Res.* **2013**, *118*, 1445–1458. [CrossRef]
17. Werdell, J.; Bailey, S. An improved bio-optical data set for ocean color algorithm development and satellite data product variation. *Remote Sens. Environ.* **2005**, *98*, 122–140. [CrossRef]
18. Vidussi, F.; Claustre, H.; Manca, B.; Luchetta, A.; Marty, J. Phytoplankton pigment distribution in relation to upper thermocline circulation in the eastern mediterranean sea during winter. *J. Geophys. Res.* **2001**, *106*, 19939–19956. [CrossRef]
19. Uitz, J.; Claustre, H.; Morel, A.; Hooker, S.B. Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *J. Geophys. Res.* **2006**, *111*. [CrossRef]
20. Mouw, C.; Yoder, J. Optical determination of phytoplankton size composition from global seawifs imagery. *J. Geophys. Res.* **2010**, *115*. [CrossRef]
21. Hu, C.; Lee, Z.; Franz, B. Chlorophyll algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. *J. Geophys. Res.* **2012**, *117*. [CrossRef]
22. Werdell, P.J.; Franz, B.A.; Bailey, S.W.; Feldman, G.C.; Boss, E.; Brando, V.E.; Dowell, M.; Hirata, T.; Lavender, S.J.; Lee, Z. Generalized ocean color inversion model for retrieving marine inherent optical properties. *Appl. Opt.* **2013**, *52*, 2019–2037. [CrossRef] [PubMed]
23. Hu, S.; Cao, W.; Wang, G.; Xu, Z.; Lin, J.; Zhao, W.; Yang, Y.; Zhou, W.; Sun, Z.; Yao, L. Comparison of Meris, Modis, Seawifs-derived particulate organic carbon, and in situ measurements in the South China Sea. *Int. J. Remote Sens.* **2016**, *37*, 1585–1600. [CrossRef]
24. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
25. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
26. Leardi, R.; Gonzalez, A.L. Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemom. Intell. Lab.* **1998**, *41*, 195–207. [CrossRef]
27. Shi, T.; Chen, Y.; Liu, H.; Wang, J.; Wu, G. Soil organic carbon content estimation with laboratory-based visible-near-infrared reflectance spectroscopy: Feature selection. *Appl. Spectrosc.* **2014**, *68*, 831–837. [CrossRef] [PubMed]
28. Araújo, M.C.U.; Saldanha, T.C.B.; Galvao, R.K.H.; Yoneyama, T.; Chame, H.C.; Visani, V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom. Intell. Lab.* **2001**, *57*, 65–73. [CrossRef]
29. Wang, J.; Shi, T.; Liu, H.; Wu, G. Successive projections algorithm-based three-band vegetation index for foliar phosphorus estimation. *Ecol. Indic.* **2016**, *67*, 12–20. [CrossRef]
30. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]

31. Bazi, Y.; Melgani, F. Toward an optimal SVM classification system for hyperspectral remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 3374–3385. [CrossRef]
32. Wold, S.; Martens, H.; Wold, H. The Multivariate Calibration Problem in Chemistry Solved by the PLS Method. In *Matrix Pencils*; Springer: Berlin/Heidelberg, Germany, 1983; pp. 286–293.
33. de Jong, S. Simpls: An alternative approach to partial least squares regression. *Chemom. Intell. Lab.* **1993**, *18*, 251–263. [CrossRef]
34. Were, K.; Bui, D.T.; Dick, O.B.; Singh, B.R. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an afromontane landscape. *Ecol. Indic.* **2015**, *52*, 394–403. [CrossRef]
35. Drucker, H.; Burges, C.J.; Kaufman, L.; Smola, A.J.; Vapnik, V. Support vector regression machines. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1997; pp. 155–161.
36. Pelckmans, K.; Suykens, J.A.; Van Gestel, T.; De Brabanter, J.; Lukas, L.; Hamers, B.; De Moor, B.; Vandewalle, J. *Ls-Svmlab: A MATLAB/C Toolbox for Least Squares Support Vector Machines*; KULeuven-ESAT: Leuven, Belgium, 2002.
37. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
38. Breiman, L.; Cutler, A.; Liaw, A.; Wiener, M. Package'randomforest. Available online: http://stat.www.berkeley.edu/~breiman/RandomForests (accessed on 05 March 2018).
39. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*; Springer: Berlin, Germany, 1998; pp. 199–213.
40. Wu, G.; Cui, L.; Duan, H.; Fei, T.; Liu, Y. An approach for developing landsat-5 TM-based retrieval models of suspended particulate matter concentration with the assistance of modis. *ISPRS J. Photogramm.* **2013**, *85*, 84–92. [CrossRef]
41. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm.* **2016**, *114*, 24–31. [CrossRef]
42. Liu, H.; Shi, T.; Chen, Y.; Wang, J.; Fei, T.; Wu, G. Improving spectral estimation of soil organic carbon content through semi-supervised regression. *Remote Sens.* **2017**, *9*, 29. [CrossRef]
43. Chan, J.C.-W.; Paelinckx, D. Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sens. Environ.* **2008**, *112*, 2999–3011. [CrossRef]
44. Yoo, S.; Im, J.; Wagner, J.E. Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landsc. Urban Plan.* **2012**, *107*, 293–306. [CrossRef]
45. Xu, S.; Lu, B.; Baldea, M.; Edgar, T.F.; Nixon, M. An improved variable selection method for support vector regression in nir spectral modeling. *J. Process Control* **2017**. [CrossRef]
46. Alvain, S.; Moulin, C.; Dandonneau, Y.; Bréon, F.-M. Remote sensing of phytoplankton groups in case 1 waters from global seawifs imagery. *Deep Sea Res.* **2005**, *52*, 1989–2004. [CrossRef]
47. Lee, Z.; Carder, K.L. Band-ratio or spectral-curvature algorithms for satellite remote sensing? *Appl. Opt.* **2000**, *39*, 4377–4380. [CrossRef] [PubMed]
48. Brotas, V.; Brewin, R.J.W.; Sá, C.; Brito, A.C.; Silva, A.; Mendes, C.R.; Diniz, T.; Kaufmann, M.; Tarran, G.; Groom, S.B.; et al. Deriving phytoplankton size classes from satellite data: Validation along a trophic gradient in the eastern atlantic ocean. *Remote Sens. Environ.* **2013**, *134*, 66–77. [CrossRef]
49. Sammartino, M.; Di Cicco, A.; Marullo, S.; Santoleri, R. Spatio-temporal variability of micro-, nano-and pico-phytoplankton in the Mediterranean Sea from satellite ocean colour data of SeaWiFS. *Ocean Sci.* **2015**, *11*, 759. [CrossRef]
50. Di Cicco, A.; Sammartino, M.; Marullo, S.; Santoleri, R. Regional empirical algorithms for an improved identification of Phytoplankton Functional Types and Size Classes in the Mediterranean Sea using satellite data. *Front. Mar. Sci.* **2017**, *4*, 126. [CrossRef]
51. Brewin, R.J.W.; Sathyendranath, S.; Lange, P.K.; Tilstone, G. Comparison of two methods to derive the size-structure of natural populations of phytoplankton. *Deep Sea Res.* **2014**, *85*, 72–79. [CrossRef]