Modeling the temperature-nitrate relationship in the coastal upwelling domain of the California Current

Daniel M. Palacios,[1,2] Elliott L. Hazen,[1,2] Isaac D. Schroeder,[1,2] and Steven J. Bograd[2]

Received 9 January 2013; revised 22 April 2013; accepted 23 April 2013.

[1] Given the importance of nitrate in sustaining high primary production and fishery yields in eastern boundary current ecosystems, it is desirable to know the amounts of this nutrient reaching the euphotic zone through the upwelling process. Because such measurements are not routinely available, we developed predictive models of water-column (0–200 m) nitrate based on temperature for a region of the California Current System (30–47°N) within 50 km from the coast. Prediction was done using generalized additive models based on a compilation of 37,607 observations collected over the period 1959–2004 and validated with a separate set of 6430 observations for the period 2005–2011. A temperature-only model had relatively high explanatory power (explained deviance, $D^2 = 71.6\%$) but contained important depth, latitudinal, and seasonal biases. A model incorporating salinity in addition to temperature ($D^2 = 91.2\%$) corrected for the latitudinal and depth biases but not the seasonal bias. The best model included oxygen, temperature, and salinity ($D^2 = 96.6\%$) and adequately predicted nitrate temporal behavior at two widely separated locations (44°39.1′N and 32°54.6′N) with slight or no bias [root-mean-square error (RMSE) = 2.39 and 0.40 μ*M*, respectively). For situations when only temperature is available, a model including depth, month, and latitude as proxy covariates corrects some of the biases, but it had lower predictive skill (RMSE = 2.50 and 5.22 μ*M*, respectively). The results of this study have applications for the proxy derivation of nitrate availability for primary producers (phytoplankton, macroalgae) in upwelling regions and for biogeochemical and ecosystem modeling studies.

Citation: Palacios, D. M., E. L. Hazen, I. D. Schroeder, and S. J. Bograd (2013), Modeling the temperature-nitrate relationship in the coastal upwelling domain of the California Current, *J. Geophys. Res. Oceans*, *118*, doi:10.1002/jgrc.20216.

1. Introduction

[2] Injection of nutrients into the upper layer of the ocean through the process of wind-induced coastal upwelling results in elevated primary production in eastern boundary currents. Although direct measurement of the upwelling process remains elusive, the coastal upwelling index (UI; defined as the magnitude of the offshore component of the Ekman transport, in meter cube per second per 100 m of coastline) was devised in the 1970s as a large-scale estimate of the amount of water upwelled from the base of the Ekman layer [*Bakun*, 1973; *Schwing et al.*, 1996]. However, UI is derived exclusively from geostrophic estimates of wind stress over the ocean and does not contain information on the properties of the water being upwelled. In order to obtain a more direct indication of the potential for upwelled waters to sustain high biological production, it is desirable to know the amounts of the major nutrients such as nitrate reaching the euphotic zone through the upwelling process.

[3] Routine, direct measurements of nitrate, however, are scarce for large regions of the world oceans, and therefore statistical prediction is by necessity undertaken by exploiting the relationship between nitrate and other more widely available measurements such as temperature [e.g., *Kamykowski and Zentara*, 1986; *Garside and Garside*, 1995; *Louanchi and Najjar*, 2000]. The advent of satellite remote sensing in the 1980s and 1990s enabled the production of large-scale maps of surface nutrient concentrations estimated from satellite-derived sea-surface temperature measurements [*Traganza et al.*, 1983; *Dugdale et al.*, 1989, 1997; *Sathyendranath et al.*, 1991; *Morin et al.*, 1993; *Goes et al.*, 1999; *Kamykowski et al.*, 2002; *Henson et al.*, 2003], an area of research still active today [*Silió-Calzada et al.*, 2008; *Steinhoff et al.*, 2010; *Sarangi*, 2011].

Additional supporting information may be found in the online version of this article.

[1]Cooperative Institute for Marine Ecosystems and Climate, Institute of Marine Sciences, Division of Physical and Biological Sciences, University of California, Santa Cruz, California, USA.

[2]Environmental Research Division, Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Pacific Grove, California, USA.

Corresponding author: D. M. Palacios, Environmental Research Division, Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 1352 Lighthouse Avenue, Pacific Grove, CA 93940-2097, USA. (Daniel.Palacios@noaa.gov)

2169-9275/13/10.1002/jgrc.20216

[4] A typical temperature-nitrate (*T-N*) scatterplot shows a strong inverse relationship, such that even a simple linear fit can be statistically significant and have a high coefficient of determination (R^2). Upon closer inspection, however, the shape of the relationship may not necessarily be linear, and it may display a broad scatter. A widely used approach for nitrate prediction has been to fit polynomial expansions of temperature, typically quadratic [*Chavez et al.*, 1996; *Louanchi and Najjar*, 2000] or cubic [*Kamykowski and Zentara*, 1986; *Switzer et al.*, 2003], to account for potential curvature in the relationship within a linear regression framework. Curve fitting with a sigmoid function has also been used as a form of nonlinear regression [*Sarangi*, 2011]. The inclusion of covariates such as salinity, chlorophyll concentration, or wind speed in the models tends to further increase the amount of nitrate variance explained [e.g., *Roy*, 1991; *Garside and Garside*, 1995; *Goes et al.*, 2000; *Sarangi*, 2011].

[5] In contrast to least-squares based methods, more recently developed approaches based on local fitting of nonparametric smooth functions of the predictors, such as generalized additive models (GAMs) [*Hastie and Tibshirani*, 1990; *Wood*, 2006], provide responses that more closely follow the shape of nonlinear relationships, and therefore have the potential to capture more complex aspects of the *T-N* relationship. While GAMs do not provide the familiar empirical algorithms estimated by parametric models (i.e., a regression equation), for many applications this is not a requirement since GAMs can be reconstructed from the smoothing function and basis dimension used to fit the model.

[6] Our focus here is on the prediction of water-column (0–200 m) nitrate in the California Current System (CCS) within the ~50 km region off the coast that is most directly affected by the upwelling process. For this purpose, we build a series of GAMs using a historical data set spanning the period 1959–2004. The resulting models are assessed though residual diagnostics, and their predictive power are evaluated relative to a set of independent observations for the period 2005–2011. As an application, nitrate time series are predicted and evaluated at two sites in the CCS to examine the ability of the models to represent their temporal behavior. To conclude, the strengths and limitations of each model are discussed.

2. Data Sources

[7] The CCS is one of the world's most intensely sampled ocean ecosystems. Multiple field programs and process studies have been conducted in the CCS by both academic and government institutions [e.g., *Bograd et al.*, 2003; *Huyer et al.* 2007; *Peña and Bograd*, 2007; *Checkley and Barth*, 2009]. We obtained hydrographic data for the CCS covering the region 30–47°N, 126–116°W from four of these programs: (1) the California Cooperative Oceanic Fisheries Investigations program (CalCOFI; 7 January 1969 to 29 January 2011; available at http://www.calcofi.org/), (2) the U.S. GLOBEC Northeast Pacific Long-Term Observation Program (LTOP; 19 September 1997 to 1 September 2004; available at http://globec.whoi.edu/jg/dir/globec/nep/ccs/ltop/), (3) the Coastal Ocean Processes/Wind Events and Shelf Transport program (CoOP/WEST; 1 June 2000 to 20 January 2003; http://ccs.ucsd.edu/coop/west/), and (4) NOAA's Newport Hydrographic Line program (NH-Line; May 1996 to present; http://www.nwfsc.noaa.gov/oceanconditions).

[8] In addition, we queried historical records from other programs in the World Ocean Database 2009 (WOD09; available at http://www.nodc.noaa.gov/) [*Boyer et al.*, 2009] for this same region, yielding observations for the period 8 July 1959 to 25 July 1985. The WOD09 extract contained data collected in the Monterey Bay area by Hopkins Marine Station (HMS) of Stanford University (14 October 1969 to 22 April 1974) and by Moss Landing Marine Laboratories (MLML) of California State University under the auspices of California Sea Grant (two time periods: 13 April 1970 to 17 December 1971 and 3 January 1975 to 7 December 1976). It also contained observations from the Coastal Transition Zone experiment (CTZ; 30 April to 19 May 1987) off northern California. Observations in the WOD09 for which the accompanying metadata was insufficient to determine the specific program under which they were collected were grouped together under "other." CalCOFI data contained in the WOD09 were excluded to avoid duplication.

[9] Because the focus of our effort was on the waters directly influenced by coastal upwelling, we limited the extracts to observations in the upper 0–200 m in the water column occurring within a strip 50 km from the coastline. For this purpose the great-circle distance between each observation and the nearest coastline was computed using the high-resolution level of the Global Self-consistent, Hierarchical, High-resolution Shoreline Database (available at http://www.ngdc.noaa.gov/mgg/shorelines/gshhs.html) [*Wessel and Smith*, 1996]. For each source, we retained observations that contained at least three variables: temperature, salinity, and nitrate (in addition to date, depth, and geographic location). If other variables, such as oxygen, phosphate, or silicate were available, these were also retained. The nitrate-to-phosphate (*N:P*) and the silicate-to-nitrate (*Si:N*) ratios were computed for observations containing these measurements. These variables were used at the different stages of data screening as filtering criteria. Tabular and graphical summaries of the temporal and spatial coverage of the observations used in this study are given in Table 1 and Figure 1.

3. Data Preparation

[10] Considering the wide variation in data sources, time periods, and sampling protocols in the initial data compilation, quality control was conducted to identify and deal with potentially problematic observations, as is commonly performed on hydrographic databases prior to analysis [e.g., *Louanchi and Najjar*, 2000]. We implemented the following steps for data screening, outlier identification, and potential transformation.

3.1. Data Screening

[11] The initial data compilation contained 54,487 observations in the upper 200 m and within 50 km from the coast. It consisted of 12 columns for date, latitude, longitude, observation depth, temperature, nitrate, salinity, oxygen, phosphate, silicate, *N:P* ratio, and *Si:N* ratio (the

Table 1. Sources of Hydrographic Data, Number of Stations, Number of Observations, and Dates in the Cleaned Data Set (See Section 3 for Details on Data Preparation)

Program[a]	Number of Stations	Number of Observations	Start Date	End Date
CalCOFI	2731	27,544	7 Jan 1969	29 Jan 2011
HMS	321	1755	14 Oct 1969	22 Apr 1974
MLML	432	1826	13 Apr 1970	7 Dec 1976
CTZ	11	118	30 Apr 1987	18 May 1987
LTOP	362	3370	15 Nov 1997	1 Sep 2004
WEST	637	4424	1 Jun 2000	20 Jan 2003
NH-Line	102	102	16 Nov 1997	8 Jun 2011
Other	970	4898	8 Jul 1959	25 Jul 1985
Total	5566	44,037	8 Jul 1959	8 Jun 2011

[a] See section 2 for details of the programs.

abbreviated names for these variables, as used in the models, are defined in Table 2). The first step in data screening consisted of removing duplicate entries ($n = 84$). The second step involved removing observations with measured values of nitrate, phosphate, or silicate equal to zero ($n = 3814$). The rationale for this being that once a nutrient becomes depleted any temperature-nutrient relationship breaks down and for these observations the temperature at

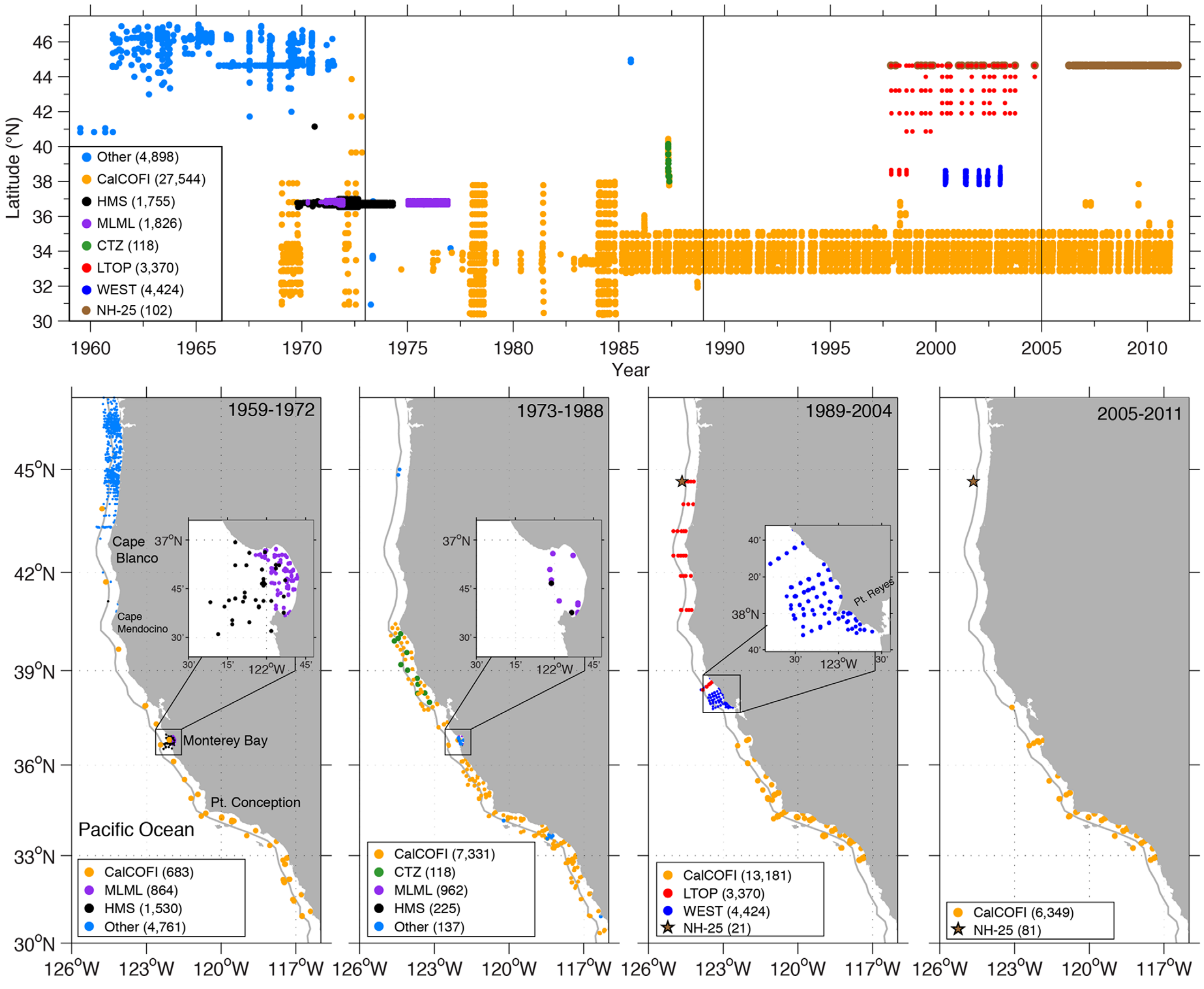

Figure 1. (top) Time-latitude plot showing the coverage of the hydrographic data over the 53 year period 1959–2011, colored by major program. The number of observations collected by each program is indicated in parentheses. Vertical lines indicate the breaks used in the subperiods in the bottom. (bottom) Maps showing the location of hydrographic stations in the cleaned data set for the 53 year period 1959–2011, colored by major program. Data are split into three subperiods for presentation. The number of observations in each subperiod is indicated in parentheses.

Table 2. Descriptive Statistics for the Cleaned Variables[a]

Variable	Abbreviation	n	Range	Mean	sd	Skewness	Kurtosis	CV
Nitrate (μM)	N	44,037	[0.02, 37.70]	14.48	10.33	0.10	1.70	71.30
Temperature (°C)	T	44,037	[6.09, 20.62]	11.14	2.28	0.78	3.52	20.49
Salinity (psu)	S	44,037	[31.060, 34.365]	33.51	0.48	−1.61	6.26	1.42
Oxygen (mL/L)	O	35,786	[0.79, 9.28]	4.41	1.52	−0.07	1.93	34.46
Phosphate (μM)	P	39,378	[0.01, 3.22]	1.30	0.67	0.09	1.84	51.65
Silicate (μM)	Si	42,830	[0.100, 67.410]	18.27	12.89	0.60	2.50	70.55
$N{:}P$ ratio	$N{:}P$	39,378	[0.051, 60.000]	8.94	4.28	−0.78	2.8	47.88
$Si{:}N$ ratio	$Si{:}N$	42,830	[0.015, 60.000]	2.73	4.98	4.34	24.30	182.46
Obs. depth (m)	Z	44,037	[0, 200]	56.51	51.44	1.06	3.32	91.02
Month	M	44,037	[1, 12]	5.63	3.23	0.15	1.92	57.42
Latitude (°N)	L	44,037	[30.342, 47.000]	36.63	4.14	1.15	3.05	11.31

[a] The variables used in data screening and in GAM model specification, sample size (n), range, mean, standard deviation (sd), skewness, kurtosis, and coefficient of variation (CV). Note that the second column provides the abbreviated names of the variables used in the text.

which depletion occurred is unknown (cf., the "nitrate depletion temperature," defined as the intercept in a *T-N* scatterplot at the temperature axis) [see *Switzer et al*., 2003]. The third step involved removing out-of-range values in temperature ($5 < T < 25$°C), salinity ($30 < S < 35$), nitrate ($0 < N < 50$ μM), phosphate ($0 < P < 10$ μM), silicate ($0 < Si < 150$ μM), $N{:}P$ ratio ($0 < N{:}P < 60$), and $Si{:}N$ ratio ($0 < Si{:}N < 60$) for typical oceanic waters in the upper 200 m of the CCS ($n = 2279$).

3.2. Outlier Analysis

3.2.1. Univariate Outliers

[12] Rather than the traditional boxplot, we used Cleveland dotplots as an efficient graphical tool to visualize univariate outliers. In a Cleveland dotplot, the row number of an observation is plotted against the observation value, and points that stick out on either side are potential outliers [*Zuur et al*. 2009]. We generated Cleveland dotplots for temperature, nitrate, salinity, and oxygen using latitude and depth as sort variables for the rows, and regarded as outliers observations that departed noticeably from the point cloud (likely caused by measurement or data entry error). The combined univariate outliers from these plots ($n = 120$) were excluded from the data set.

3.2.2. Outliers in the Temperature-Salinity Relationship

[13] Unusual relationships in the temperature-salinity scatterplot, evident as "tendrils" emanating from the main point cloud, were explored. In all cases, these observations ($n = 194$) were isolated to individual profiles and may have been the result of extreme environmental conditions (such as strong El Niño events in 1970–1971, 1981, 1997–1998) or instrumental error. These observations were excluded from the analyses.

3.2.3. Outliers in the Nitrate-Phosphate Relationship

[14] A nitrate-phosphate scatterplot for those observations containing both measurements ($n = 43{,}331$) revealed that most of the data set had a well-behaved linear relationship. However, a small cluster with a much steeper relationship (i.e., low nitrate, high phosphate conditions) was readily evident. Further investigation indicated that these observations belonged to a unique water mass found off Oregon and Washington during the period 1961–1970. This anomalous water mass occupied much of the water column, and it was generally characterized by relatively cool ($T < 10$°C), salty ($S > 32$), low-nitrate ($N < 6.5$ μM), high-phosphate ($P > 1$ μM), and low-oxygen ($O < 6$ mL/L) conditions. While a full study of this water mass remains to be conducted, it was judged that these observations should be removed prior to analysis because their characteristics were not representative of the conditions encountered in this area at any other time in the data set. To accomplish this, a linear regression of phosphate on nitrate was fitted to the nonanomalous observations, and the 95% prediction band (simultaneous, protected against multiple observations with Scheffé adjustment) for this regression was applied to the entire data set as the basis for identifying and excluding outliers ($n = 1914$) in the nitrate-phosphate relationship.

3.2.4. Multivariate Outliers

[15] The final step in the outlier analysis took advantage of concurrent measurements in addition to temperature and nitrate available for several of the data sources. Multivariate outliers were identified following the technique outlined in *Tabachnik and Fidell* [1989, 2001]. Briefly, the Mahalanobis distance was computed for each observation from the centroid of the n-dimensional point cloud, and a χ^2 test was used to identify points with distances greater than the critical value of the distribution [with $\alpha = 0.001$ and degrees of freedom (df) equal to the number of variables included in the particular combination of data points being considered]. This procedure was implemented on two subsets of the data, one containing concurrent observations for the eight variables temperature, nitrate, salinity, oxygen, phosphate, silicate, $N{:}P$ ratio, and $Si{:}N$ ratio ($n = 35{,}949$), and another containing observations for which only the three variables temperature, nitrate, and salinity were available ($n = 10{,}138$).

[16] To a large extent the multivariate outliers identified in these two subsets ($n = 1847$ and $n = 198$, respectively) occurred near the margins of the n-dimensional point cloud as gleaned from inspection of the respective property-property plots. Closer examination revealed that they could be grouped into three types. Type 1 outliers ($n = 926$) were mostly in the upper 50 m and in the southern half of the latitudinal range ($L < 38$°N). These waters had a wide range in temperature ($8.25 < T < 24.1$°C), a high salinity ($S > 33.15$), were low in nitrate ($N < 20$ μM), and high in oxygen ($O > 5$ mL/L). Type 2 outliers ($n = 146$)

corresponded to a small cluster found at middepths (20–200 m) and mostly in the northern half of the study area ($L > 36^\circ$N). These waters were cold ($T < 8.25^\circ$C) and had a high salinity ($S > 33.15$), high nitrate ($N > 20$ μM), and low oxygen ($O < 4$ mL/L). Type 3 outliers ($n = 770$) were found mostly in the upper 50 m and occurred primarily in the northern half of the study area ($L > 36^\circ$N). These waters had a wide range of temperature ($6 < T < 17^\circ$C), a low salinity ($S < 33.15$), low nitrate ($N < 20$ μM), and high oxygen ($O > 4.5$ mL/L).

[17] Thus, these outliers identified the most extreme nutrient relationships in multivariate space and were separable by water mass characteristics, latitude, and depth. The procedure also eliminated observations with salinities lower than 31.060, which are influenced by riverine discharge (e.g., the area of influence of the Columbia River plume as well as other rivers discharging along the Oregon and California coasts). For these reasons, we felt justified to remove these outliers from data compilation.

3.3. Final Data Set

[18] The cleaned data set contained 44,037 observations and spanned the period 8 July 1959 to 8 June 2011 (Table 1 and Figure 1). Descriptive statistics for this data set, including range, mean, standard deviation, skewness, kurtosis, and coefficient of variation are provided in Table 2. We also examined the univariate distribution of the main variables (nitrate, temperature, salinity, oxygen, depth, month, and latitude) as well as the bivariate relationships among them and these are shown in Figure 2. Pearson correlation coefficients among variables are presented in Table 3. Phosphate and silicate covaried strongly with nitrate ($r = 0.98$ and 0.95, respectively) but were also highly collinear with each other ($r = 0.95$) as well as with oxygen ($r = -0.92$ and -0.87, respectively) and with temperature ($r = -0.83$ and -0.82, respectively); therefore, they were not included as potential explanatory variables in the models.

[19] Both the descriptive metrics and the graphical exploration (Table 2 and Figure 2) indicated some degree of nonnormality in the primary variables used in modeling (N, T, S, O); however common data transformations (i.e., log, square root, inverse) did not improve the descriptive metrics or the distributions. (Although normality can be achieved with more drastic transformations like the Box-Cox or the rank, difficulties with interpretability and back

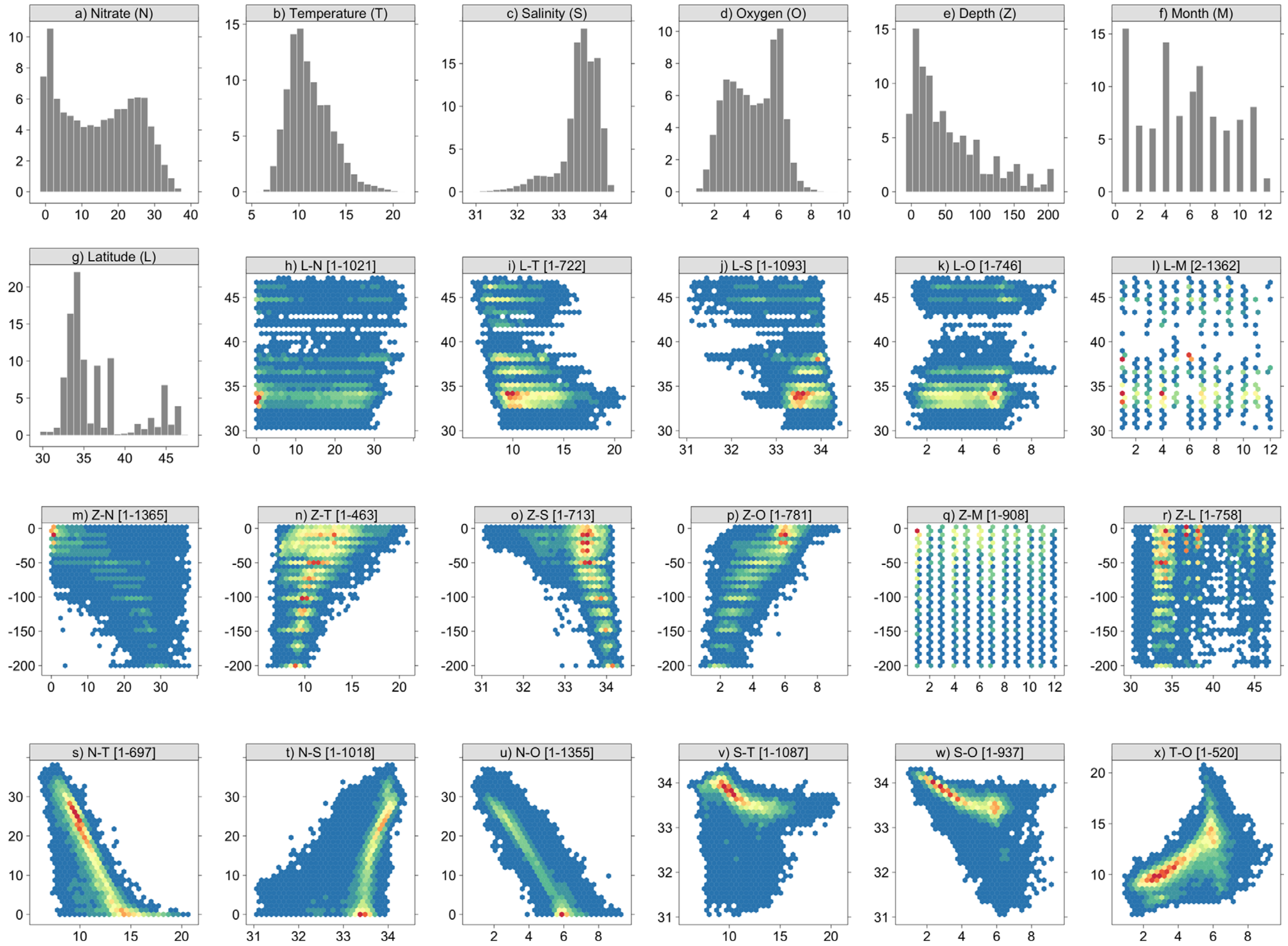

Figure 2. (a–g) Univariate frequency histograms ($n = 22$ bins) for variables used in GAM modeling after screening and outlier analysis. (h–x) Binned bivariate scatterplots (hexagonal bins, $n = 30$) with color shade indicating the density of observations (blue = low, red = high) for selected variable combinations. The range of the number of observations per bin is given above each figure. Axis units are as in Table 2.

Table 3. Correlation Matrix (Pairwise Pearson Correlation Coefficients) of Main Variables in the Compiled Data Set Considered in Data Screening and in GAM Model Specification

	N	*T*	*S*	*O*	*P*	*Si*	*Z*	*M*	*L*
N	1.000								
T	−0.824	1.000							
S	0.620	−0.255	1.000						
O	−0.929	0.659	−0.663	1.000					
P	0.983	−0.832	0.556	−0.921	1.000				
Si	0.951	−0.819	0.529	−0.868	0.951	1.000			
Z	0.651	−0.483	0.520	−0.802	0.673	0.572	1.000		
M	0.059	0.031	0.028	−0.024	0.044	0.064	−0.004	1.000	
L	0.089	−0.428	−0.528	0.157	0.125	0.258	−0.226	0.083	1.000

transformation made them less applicable for our purpose). Therefore, we used the variables untransformed and note that some nonnormality remained in the final models (see supporting information for additional diagnostics of model residuals). We also point out that after screening and deletion of outliers the final data set reflected waters of oceanic character and therefore the inferences drawn from the models in the following sections are only applicable to the ranges and variable combinations in the data analyzed here.

4. Nitrate Patterns in the Coastal Upwelling Domain of the CCS

[20] The large observational data set compiled here allowed us to examine the patterns of biologically available nitrate in the water column in the coastal upwelling domain of the CCS over a period of five decades. As occurs throughout the world ocean, nitrate concentrations in the study area were high at depth and became progressively lower in the upper levels of the water column (Figures 2m and 3a). Nitrate increased with latitude, especially at depth ($Z > 50$ m) (Figure 3a). In waters $Z < 50$ m nitrate tended to peak at 37–40°N and then decrease.

[21] Nitrate had a strong inverse relationship with temperature and oxygen and a direct one with salinity (Figures 2s–2u and 4a). These relationships were reasonably linear and tight at $Z > 50$ m, but they developed a "hockey stick" shape and displayed wide scatter at shallower depths, especially where $N < 5$ μ*M*, $T > 14.5$°C, $S < 33.25$, and $O > 6$ mL/L (Figures 3b–3d and 4a).

[22] In terms of temporal patterns, nitrate followed a seasonal cycle with a broad peak in the spring and summer months (Figure 4b). This cycle occurred at all depths although it was less marked in very shallow ($Z < 10$ m) and very deep ($Z > 100$ m) waters. There was also evidence of decadal trends over the period 1959–2011, with nitrate decreasing from 1959 to the late 1980s and then increasing through the present time (Figure 4c). This trend occurred at all depths, but, like the seasonal pattern, it was most marked at middepths ($10 < Z < 100$ m) (Figure 4c).

[23] Progressively lower nitrate and oxygen concentrations toward the southern half of the study area ($L < 37$°N; Figures 2h, 2k, and 3a) indicated that denitrified waters are upwelled in this region relative to the waters upstream. However, the fact that the observed nitrate concentrations were always high at low oxygen levels indicated that denitrification generally did not occur in the upper 200 m anywhere in the data set.

5. Modeling Approach

[24] Although global methods (i.e., linear regression, including polynomial expansions) have been widely used to model the *T-N* relationship [*Kamykowski and Zentara*, 1986; *Garside and Garside*, 1995; *Chavez et al*., 1996; *Louanchi and Najjar*, 2000], our focus here is on local methods (i.e., GAMs), which have the potential to better capture nonlinearities in the relationship that may be related to regional gradients (e.g., water mass distributions, upwelling intensity). A brief introduction to the GAM methodology is given in section 5.1 followed by details of variable and smoothness selection in section 5.2. Models quantifying the *T-N* relationship are built in sections 5.3, starting from a *T*-only model to more complex models that further explore the nonlinear dependence of nitrate on other influential variables and that address some of the shortcomings of the simpler models.

5.1. General Form of a GAM

[25] GAMs are a nonparametric extension of the more familiar generalized linear models (GLMs). Instead of assuming a priori any rigid parametric form, GAMs represent the relationship between the response and the explanatory variables by smooth functions, which can take virtually any form [*Hastie and Tibshirani*, 1990; *Wood*, 2006]. Because they implement a local, data-driven regression, GAMs can be used to quantitatively explore complex relationships when little is known about the underlying mechanisms responsible for generating the observations. Like GLMs, GAMs are estimated using the method of penalized likelihood and the discrepancy between the observations and the estimated mean is measured using deviance residuals (expressed as the percent of deviance explained, D^2) [*Hastie and Tibshirani*, 1990; *Wood*, 2006].

[26] The general form of a GAM is:

$$g(\mu_i) = \beta + \sum_{j=1}^{p} f_j(X_i) + \varepsilon_i \qquad (1)$$

where the function $g(\mu)$ is a link function relating the mean of the response variable given the explanatory variables, $\mu = E(Y\imath X_1, \ldots, X_P)$, to the additive predictor $\beta + \sum f_j\ (X_i)$.

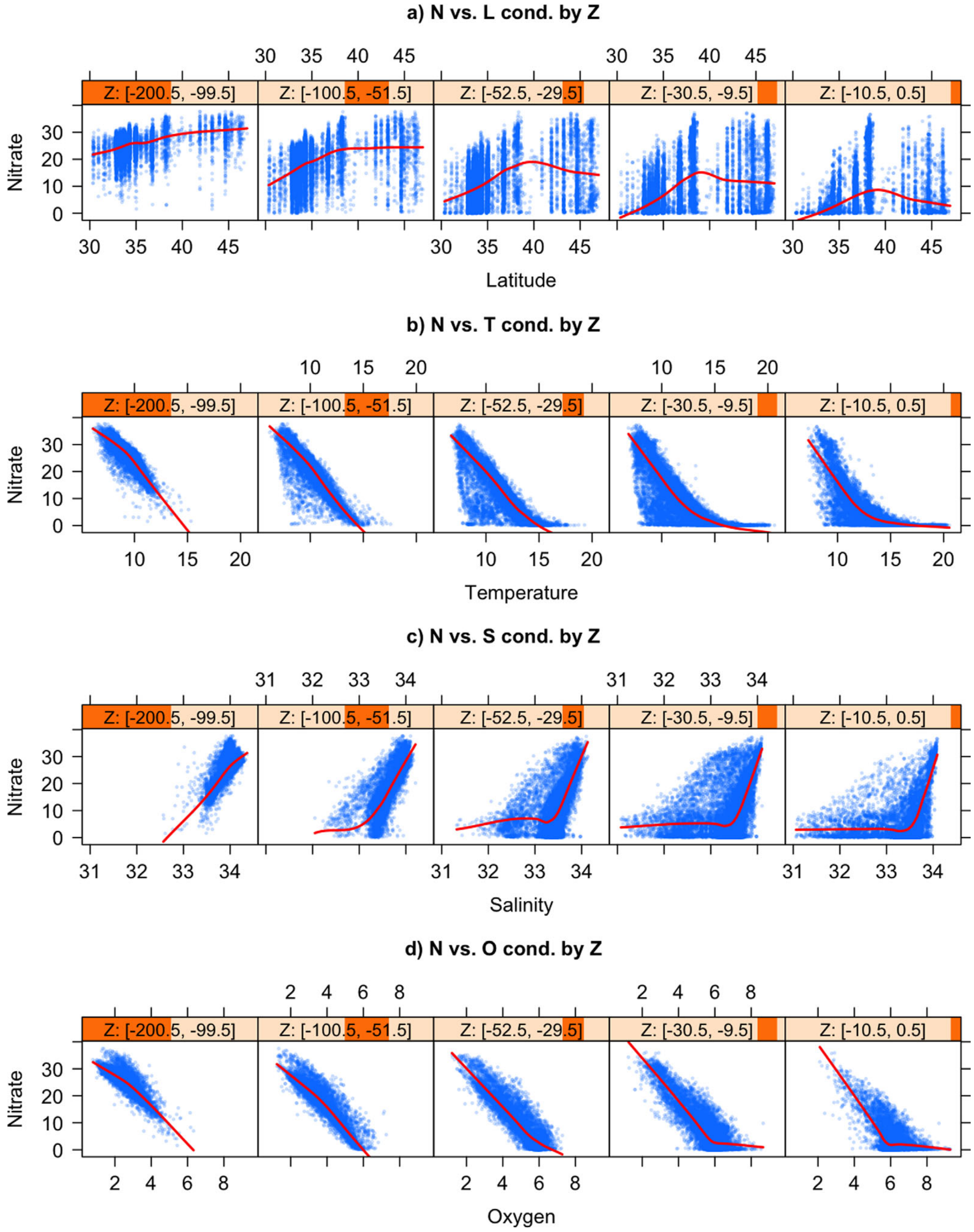

Figure 3. Trellis scatterplots of N against L, T, S, and O conditioned by Z, with observations grouped into five levels having roughly the same number of counts. The depth levels (in meter) are displayed on the strip above each figure, both numerically (bracketed intervals) and visually (darker shaded sections of the strip). Red curve in all figures is a loess scatterplot smoother (degree = 2, span = 3/4) intended to guide the eye through the point cloud.

The term β represents any strictly parametric component in the model (e.g., the intercept), while components $f_j\ (X_i)$ in the additive predictor are specified as nonparametric smooth functions of the explanatory variables, and ε_i are independent and identically distributed normal random variables [*Hastie and Tibshirani*, 1990; *Wood*, 2006].

[27] Model fitting was carried out in the R environment version 2.15.1 [*Ihaka and Gentleman*, 1996; *R Core Team*, 2012] using the "mixed GAM computation vehicle" (mgcv) library version 1.7–22 [*Wood*, 2006]. The mgcv library implements an automatic selection of the smoothing parameters associated with each smooth term, based on generalized cross-validation (GCV). Simply put, cross validation involves leaving one of the data points out, fitting the model to the remaining data, and then calculating the square difference between those points and the fitted model (smaller differences mean better models). This procedure is repeated for all data points and for several amounts of smoothing (and hence several values of df for each term). The GCV score reflects the overall balance between the gains obtained by increasing the amount of smoothing, and the costs in terms of increasing the number of df [*Wood*,

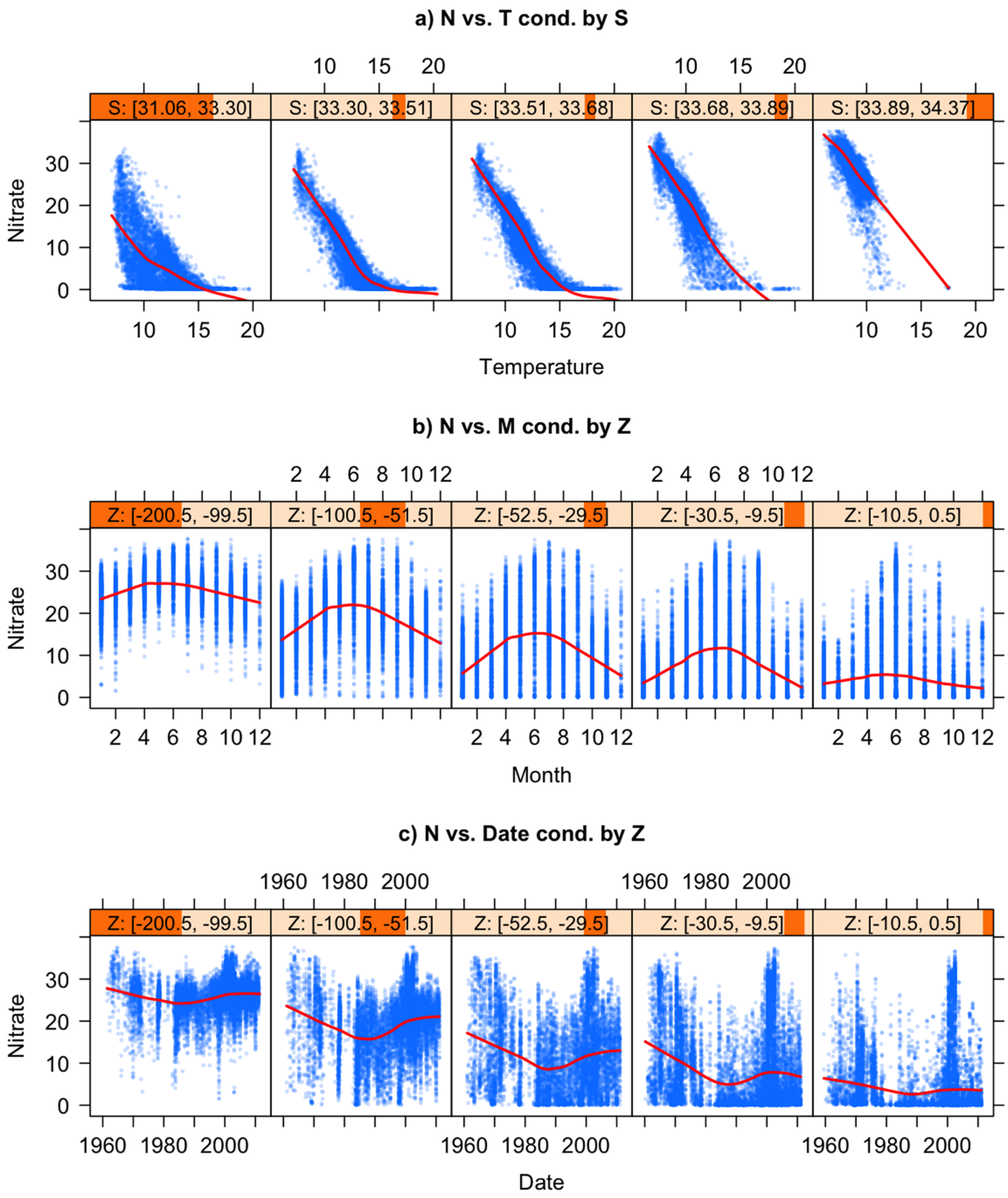

Figure 4. Trellis scatterplots of (a) N against T conditioned by S, (b) N against M conditioned by Z, and (c) N against date conditioned by Z, with observations grouped into five levels having roughly the same number of counts. Other plot details are as in Figure 3.

2006]. The maximum number of df of each smooth term (and hence the complexity of the relationship) must be set initially by the user, but the fitting procedure subsequently “downgrades” them to minimize the GCV score of the entire model [*Wood*, 2006]. Optimal smoothness for each term can be automatically achieved by initially fitting a model with restricted maximum likelihood (REML) instead of GCV, but the user must still decide if the functional responses obtained with this method are an appropriate representation of the underlying processes, and make adjustments to the df as necessary [*Wood*, 2006].

5.2. Model Specification and Selection

[28] The data set was divided into training and testing portions using the following scheme: observations for the period 1959–2004 ($n = 37{,}612$) were used for model training while data for the period 2005–2011 ($n = 6430$) were saved for model testing. This allowed us to evaluate the ability of the models to predict new nitrate observations while also allowing us to assess the models’ representation of temporal behavior at selected locations over the testing period (see section 6). As mentioned in section 3.3, nitrate concentration was treated as a continuous normal random variable, and thus GAMs were built using the identity function as the link function (i.e., $N \sim$ Gaussian). Thin-plate splines were used as the basis for the smooth function for the explanatory variables except for month (M), which was treated as a cyclical variable using the cyclic cubic regression spline.

[29] Selection of the smoothing parameter associated with each term involved specifying the basis dimension, k, which amounts to setting the maximum possible df allowed for each term (the actual effective df for each term are then

Table 4. Assessment of Single-Term GAM Models Used to Determine the Relative Importance of the Explanatory Variables Based on GCV Score, AIC, Drop in Deviance, and the Percent Deviance Explained (D^2)[a]

Model $N \sim$	df	p Value	GCV Score	AIC	Resid. Deviance	Deviance Drop	D^2 (%)
Intercept	1.00	<0.001	97.41	218,630.8	2,861,313		0
L	5.61	<0.001	96.83	217,713.1	2,843,158	18,155	0.6
M	4.85	<0.001	91.21	215,952.3	2,677,900	183,413	6.4
Z	5.82	<0.001	45.43	195,471.6	1,333,519	1,527,794	53.4
S	6.00	<0.001	41.08	192,523.2	1,206,170	1,655,143	57.8
T	5.96	<0.001	27.13	180,319.1	796,152	2,065,161	72.2
O	5.97	<0.001	12.07	156,542.4	354,410	2,506,903	87.6

[a]For each model, the drop in deviance is relative to the null model (i.e., intercept-only model in first row). Variables are sorted in increasing order of importance based on these criteria.

estimated from the data by GCV), and the value for gamma, γ, that minimizes the GCV scores. Univariate GAMs with $k = 10$ and $\gamma = 1$ (the mgcv defaults) were initially specified to assess the relative importance of the individual explanatory variables in predicting nitrate. For the final multivariate models, gamma was set to $\gamma = 1.4$, as suggested by *Wood* [2006] to avoid overfitting, while an initial estimate of the optimal smoothness for each term was obtained by fitting the models with REML. These estimates of k were subsequently adjusted downward in cases where the functional responses appeared too "wiggly," and the final models were fit with GCV.

[30] Assessment of the univariate GAMs was based on drop in deviance, D^2, GCV score, and Akaike's information criterion (AIC). These criteria indicated that the explanatory variables should be entered into candidate multivariate models in the following order: *O, T, S, Z, M*, and *L* (Table 4). Interactions among variables (i.e., when the conditional dependence between the response and an explanatory variable changes with the values of a third variable) were explored visually using Trellis displays [*Fuentes et al*., 2011]. Clear interactions between *T* and *Z* (Figure 3b) and *T* and *S* (Figure 4a) were identified from these displays, and they were entered in multivariate models using tensor products constructed with cubic regression splines as the smooth function [*Wood*, 2006].

[31] As the last step, candidate multivariate models were checked to ensure that none of the terms included were superfluous. Term selection was implemented using the "shrinkage" method, which adds a shrinkage parameter to the smoothing penalty such that under heavier penalization redundant terms are reduced to the zero function and thereby "selected out" of a model [*Marra and Wood*, 2011]. None of the terms in the models of interest specified in sections 5.3 were dropped by this selection procedure.

5.3. A Simple Model for Nitrate

[32] The first GAM model explored was one with a single smooth term for temperature:

$$N_i = \beta_0 + f(T_i) + \varepsilon_i \tag{2}$$

[33] Automatic smoothness selection with REML yielded 8.94 df for the smooth term but the functional response contained unrealistic "bumps" and "wiggles," so for the final fitting with GCV the basis dimension was constrained to $k = 8$. The model results are given in Table 5. This model used 1 parametric df for the intercept and 6.32 effective df for the smooth term, for a total of 7.32 df. The smooth term was highly significant (p value < 0.001), with fit statistics $D^2 = 71.6\%$, GCV score $= 30.27$, and AIC $= 234965.3$.

[34] The functional response of nitrate to temperature consisted of a sigmoid curve steeply descending from high nitrate concentrations at low temperatures and with a long tail at low nitrate concentrations and high temperatures (Figure 5a). The partial residuals (dots colored by depth in Figure 5a) indicated that much of the original scatter in the *T-N* relationship (e.g., Figures 2s and 3b) remained in the model and, further, that it had a strong tendency to

Table 5. Final GAM Model Specifications, Estimated Terms, and Fit Statistics Reported by mgcv[a]

	$N{\sim}T$	$N{\sim}T \times S$	$N{\sim}O + (T \times S)$[b]	$N{\sim}(T \times Z) + M + L$
n	37,607	37,607	29,378	37,607
P value	T (<0.001)	$T \times S$ (<0.001)	O (<0.001) $T \times S$ (<0.001)	$T \times Z$ (<0.001) M (<0.001) L (<0.001)
Eff. df	T (6.32)	$T \times S$ (21.12)	O (6.27) $T \times S$ (23.83)	$T \times Z$ (19.51) M (4.95) L (6.99)
Tot. df	7.32	22.12	31.10	32.45
D^2	71.6%	91.2%	96.6%	87.2%
GCV score	30.27	9.45	3.29	13.69
AIC	234965.3	191167.0	118366.7	205112.4

[a]Total df include 1 df for the intercept plus the effective df for the smooth terms.
[b]The number of observations was lower for this model due to missing *O* values.

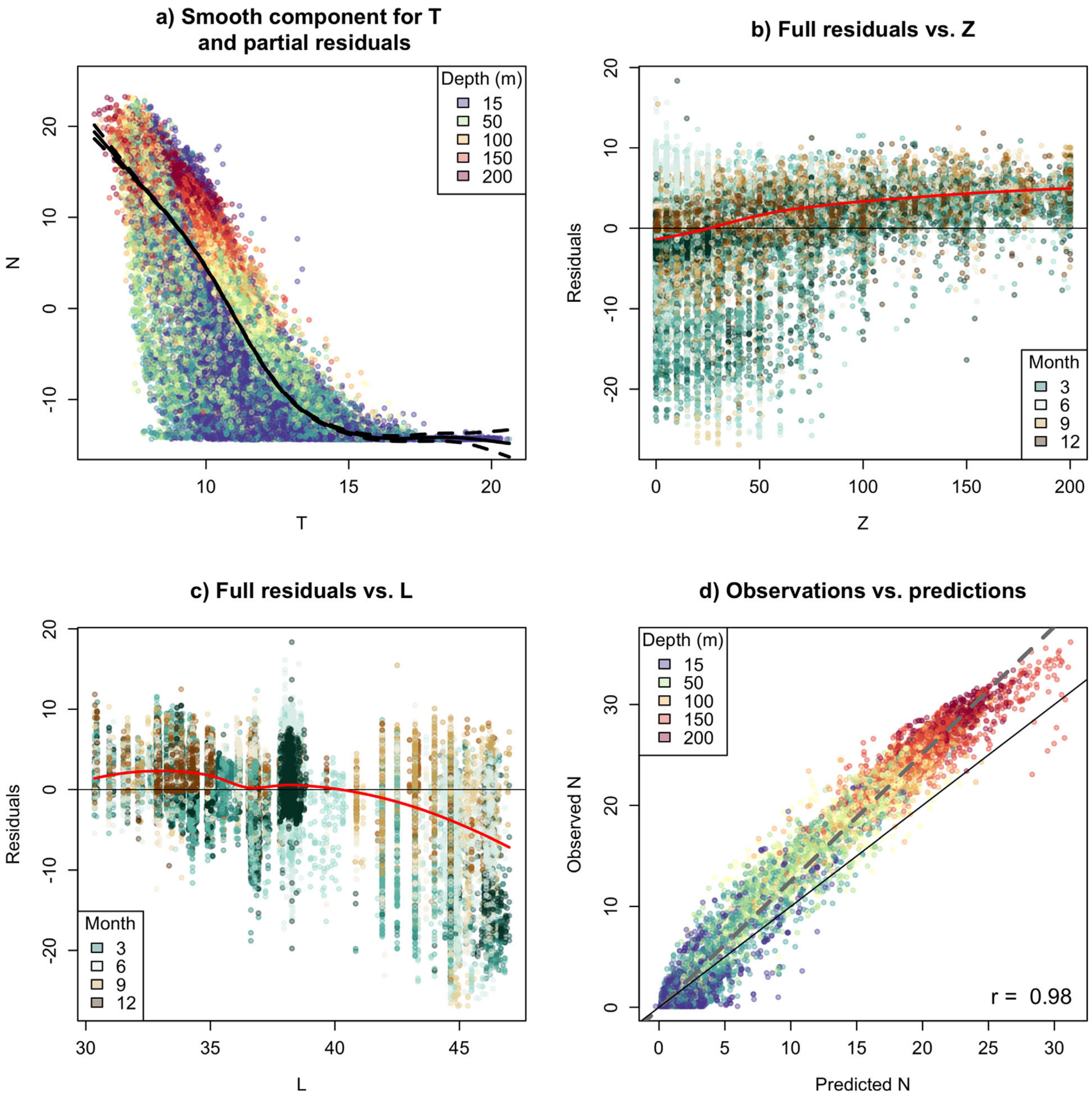

Figure 5. (a) Estimated effects (solid black curve) at the scale of the linear predictor for a simple GAM based on *T* for the training data set (1959–2004, $n = 37{,}607$). The 95% confidence limits (strictly Bayesian credible intervals) are shown as dashed black lines. Dots are the partial residuals, colored by *Z*. (b) Model residuals versus *Z* (colored by *M*). (c) Model residuals versus *L* (colored by *M*). Red curve in Figures 5b and 5c is a loess scatterplot smoother (degree = 2, span = 3/4). (d) Scatterplot of observed versus predicted values (colored by *Z*) by a simple GAM based on *T* for the testing data set (2005–2011, $n = 6430$). Dashed gray line is the linear fit and black line is the 1:1 line.

overpredict nitrate at shallow depths ($Z < 50$ m) and to underpredict at deeper depths (Figures 5a and 5b). Biases were also evident in the residuals in relation to month and latitude, with the model tending to overpredict in the early part of the year (January-August) and at the high latitudes (Figures 5b and 5c). Additional residual diagnostics for this model indicated a noticeable departure from normality and significant heterogeneity relative to the linear predictor (Figure S1 in supporting information).

[35] The performance of the model at predicting new observations was evaluated using the testing set. These results are presented in Table 6. Although the correlation between the observed and the predicted nitrate was very high ($r = 0.98$), a scatterplot showed that the slope of the observed versus predicted values was significantly different from a 1:1 relationship (Figure 5d), such that the model increasingly underpredicted the higher nitrate concentrations observed at depth. Thus, we concluded from the

Table 6. Metrics of the Performance of the Final GAM Models in Predicting Nitrate Observations From the Testing Data Set[a]

	$N{\sim}T$	$N{\sim}T \times S$	$N{\sim}O + (T \times S)$[b]	$N{\sim}(T \times Z) + M + L$
n	6430	6430	6408	6430
r	0.98	0.97	0.99	0.98
CV obs.	70.47	70.47	70.28	70.47
CV pred.	68.57	71.82	72.56	69.72
Mean ratio obs.: pred.	1.20	0.95	0.89	0.66
Range|obs.-pred.|(μM)	[1.34×10^{-4}, 12.05]	[1.21×10^{-4}, 10.43]	[4.35×10^{-4}, 9.05]	[3.60×10^{-4}, 11.37]
Mean|obs.-pred.|(μM)	3.51	2.06	1.02	2.14
Sd|obs.-pred.|(μM)	2.27	1.51	1.10	1.69

[a] Row wise these are: the sample size (n), the correlation coefficient between observations and predictions (r), the coefficient of variation for observations and for predictions, the mean ratio of observations to predictions, and the range, mean and standard deviation (sd) of the absolute value of the difference between observations and predictions.

[b] The number of observations was lower for this model due to missing O values.

diagnostics of both the training and testing steps that a T-only model was inadequate for the coastal upwelling domain of the CCS (but it may be adequate for a sufficiently small area with a limited depth range).

5.4. A Model With Temperature and Salinity

[36] Since salinity is the variable most commonly measured with temperature, we considered it useful to construct a model with these two variables to examine the performance of salinity as an explanatory variable. As mentioned in section 5.2, a marked interaction between T and S was detected in the data, and initial assessment indicated that a model including this interaction provided a better fit than a model with only the main effects. Using the previously defined training and testing data sets, the model was specified as:

$$N_i = \beta_0 + f(T_i \times S_i) + \varepsilon_i \qquad (3)$$

[37] Where the smooth function $f(T \times S)$ is a tensor product constructed with a cubic regression spline as the basis. The estimate of the optimal smoothness for this term obtained by initially fitting the model with REML (21.3 effective df) yielded a reasonable functional response, so for the final fitting with GCV the basis dimension was set to $k = 5$ for both T and S. (For tensor product smooths the upper limit of the df is given by the product of k values provided for each marginal smooth less one, which is lost to the identifiability constraint on the smooth, *Wood* [2006]). This model used 1 parametric df for the intercept and 21.12 effective df for the $T \times S$ smooth term, for a total of 22.12 df. The smooth term was highly significant (p value < 0.001). The addition of salinity increased this model's D^2 to 91.2% and lowered both the GCV score and the AIC to 9.45 and 191,167, respectively (Table 5).

[38] The functional response of nitrate in $T \times S$ space (contoured and colored surface in Figure 6a) showed a rapid and uniform decrease at $T < 12$°C and $S > 32.5$ that then leveled out at higher temperatures and lower salinities. In addition to this large-scale gradient, a local maximum in nitrate was predicted at low temperature and low salinity (Figure 6a), corresponding to estuarine influences at the northern part of the study area. This model largely corrected the depth and latitude biases in the T-only model, indicating that nitrate levels in the CCS are highly dependent on water mass, but the tendency to overpredict in the early part of the year remained (Figures 6b and 6c). Additional diagnostics of the model's residuals indicated an improvement toward normality and a closer fit, but significant heterogeneity remained at low and intermediate values of the linear predictor (Figure S2 in supporting information).

[39] The performance of this model at predicting new observations was evaluated using the testing set (Table 6). A scatterplot of observed versus predicted nitrate indicated that although the slope was closer to a 1:1 relationship than for the T-only model (Figure 6d), the shape of the relationship had some curvature, leading to underprediction at intermediate nitrate levels and slight overprediction at both low and high nitrate levels (Figure 6d).

5.5. An Expanded Model With Oxygen, Temperature, and Salinity

[40] Including additional hydrographic variables in the models can yield functional relationships useful in elucidating relevant patterns and processes while also improving the fit and predictive power. Of all the variables considered, the initial univariate GAMs (Table 4) indicated that oxygen was the single most important variable in predicting nitrate. Therefore, we fitted and evaluated models including the three variables oxygen, temperature, and salinity and their interactions. The following model provided the best fit:

$$N_i = \beta_0 + f_1(O_i) + f_2(T_i \times S_i) + \varepsilon_i \qquad (4)$$

[41] Automatic smoothness selection with REML suggested 8.28 effective df for the O smooth term but the functional response appeared slightly overfitted, so for the final fitting with GCV the basis dimension for this term was constrained to $k = 8$. The basis dimension for the $T \times S$ term was maintained as $k = 5$ for both T and S, as in the TS model in the previous section. This model used a total of 31.1 df (1 for the intercept, 6.27 for the O term, and 23.83 for the $T \times S$ interaction), with all smooth terms being highly significant (p values < 0.001). The fit statistics for this model were the best of any model, with $D^2 = 96.6\%$, GCV score $= 3.29$, and AIC $= 118366.7$ (Table 5).

[42] The smooth term for oxygen, the strongest effect, consisted of a sigmoid curve descending from high nitrate at the lowest oxygen concentrations to low nitrate at the higher oxygen concentrations (Figure 7a). This smooth captured the well-known influence of oxygen levels on

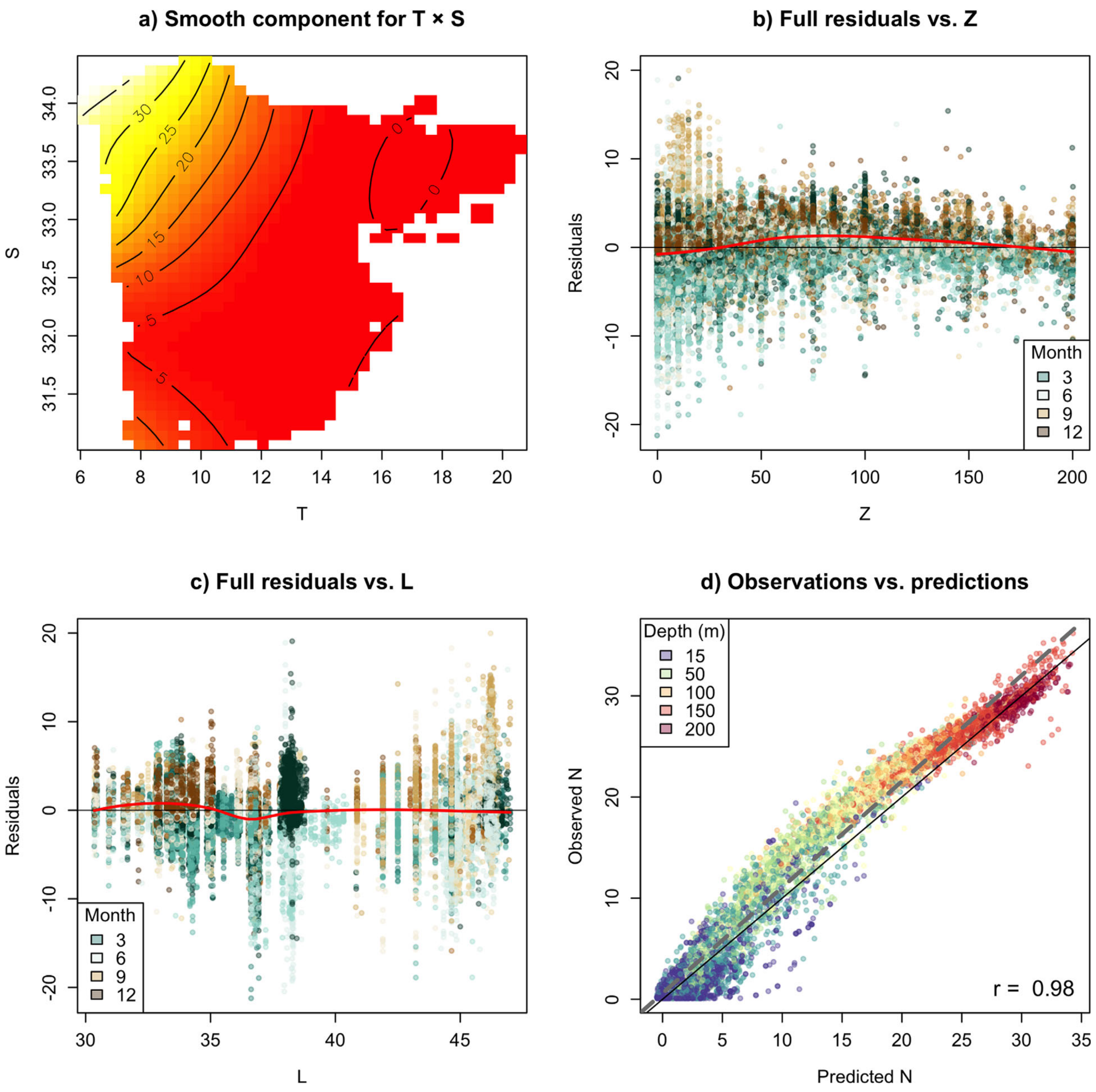

Figure 6. (a) Estimated effects (contoured surface) at the scale of the linear predictor for the $T \times S$ interaction term for a GAM based on T and S for the training data set (1959–2004, $n = 37{,}607$). (b) Model residuals versus Z (colored by M). (c) Model residuals versus L (colored by M). Red curve in Figures 6b and 6c is a loess scatterplot smoother (degree = 2, span = 3/4). (d) Scatterplot of observed versus predicted values (colored by Z) by a GAM based on T and S for the testing data set (2005–2011, $n = 6430$). Dashed gray line is the linear fit and black line is the 1:1 line.

nitrate uptake, respiration, and microbial processes in the water column [e.g., *Sarmiento and Gruber*, 2009], as evidenced by the depth-dependent pattern in the partial residuals (colored dots in Figure 7a). The functional response of nitrate to the $T \times S$ term was similar to that of the *TS* model, but temperature tended to be a stronger driver of the relationship than salinity in this model, especially at high nitrate values (Figure 7b). This model also predicted the local maximum in nitrate at low temperature and low salinity (Figure 7b) described in the previous section. No latitudinal bias was apparent in the model's residuals, and the seasonal bias was also largely corrected (Figure 7c). Further diagnostics of the model's residuals indicated that, while present at low levels, the issues of nonconstant variance, nonlinearity and departure from normality in the simpler models were greatly reduced with the expanded model and were no longer of concern (Figures S3 and S4 in supporting information).

[43] The performance of this model at predicting new observations was evaluated using the testing set (Table 6). This model achieved the highest correlation between the observed and the predicted nitrate ($r = 0.99$). More importantly, the scatterplot showed that the slope of the relationship was indistinguishable from a 1:1 relationship and that

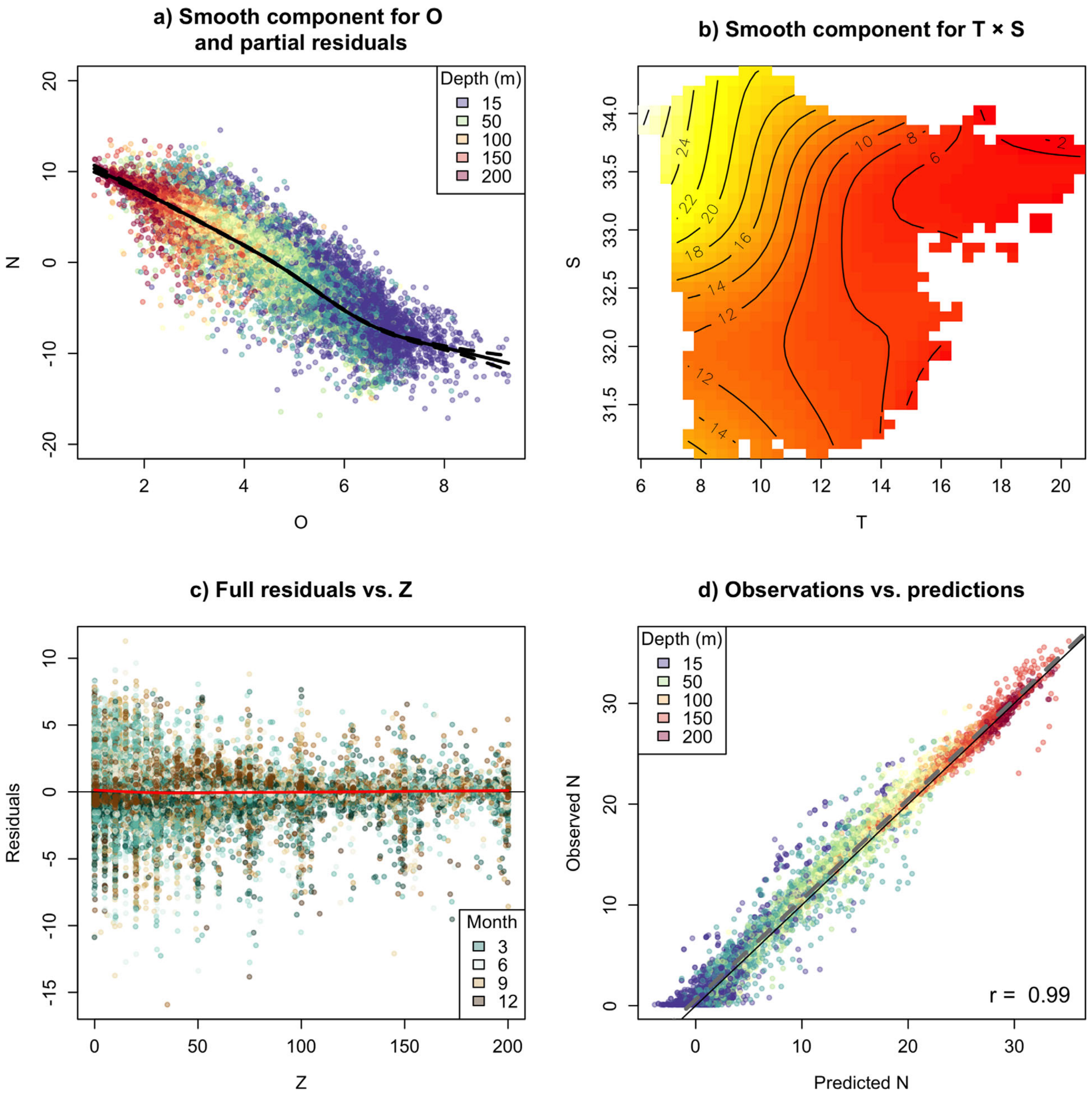

Figure 7. (a) Estimated effects (solid black curve) at the scale of the linear predictor for the *O* smooth term for a GAM based on *O*, *T*, and *S* for the training data set (1959–2004, $n = 29{,}378$). The 95% confidence limits (strictly Bayesian credible intervals) are shown as dashed black lines. Dots are the partial residuals, colored by *Z*. (b) Estimated effects (contoured surface) at the scale of the linear predictor for the $T \times S$ interaction. (c) Model residuals versus *Z* (colored by *M*). Red curve is a loess scatterplot smoother (degree = 2, span = 3/4). (d) Scatterplot of observed versus predicted values (colored by *Z*) by a GAM based on *O*, *T*, and *S* for the testing data set (2005–2011, $n = 6408$). Dashed gray line is the linear fit and black line is the 1:1 line.

there was very little bias relative to depth in the predictions (Figure 7d).

5.6. A Proxy Model With Temperature, Depth, Month, and Latitude

[44] The prediction biases in depth, latitude, and season detected in the *T*-only model (as well as the strong structure in the residuals) were adequately corrected by including oxygen and salinity, indicating that these three variables captured the dominant physical and biological processes driving nitrate variability in the upper 200 m of the CCS. However, despite its shortcomings, a *T*-only model may be the only option in applications for which no other properties are available. Thus, we explored models that incorporated depth, month, and latitude as proxies for the spatial and temporal processes accounted for by the *OTS* model.

Using the previously defined training and testing data sets, the best fitting model was:

$$N_i = \beta_0 + f_1(T_i \times Z_i) + f_2(M_i) + f_3(L_i) + \varepsilon_i \qquad (5)$$

[45] Automatic smoothness selection with REML suggested 19.76 effective df for the $T \times Z$ interaction, 7.76 df for the M term and 8.46 df for the L term. However, the functional responses for both M and L appeared slightly overfitted, so for the final fitting with GCV the basis dimension for these terms was constrained to $k = 7$ and 8, respectively, while the basis dimensions for the $T \times Z$ term were set to $k = 6$ and 4, respectively. This model used 1 parametric df for the intercept, 19.51 df for the $T \times Z$ interaction, 4.95 df for the month and 6.99 df for the latitude smooth term, for a total of 32.45 df. All three smooth terms were highly significant (p values < 0.001). The fit statistics were similar to those of the TS model ($D^2 = 87.2\%$, GCV score $= 13.69$, AIC $= 205112.4$; see Table 5).

[46] The functional response of nitrate in $T \times Z$ space showed a rapid decrease at $T < 14°\text{C}$ at all depths, indicating that temperature was a stronger driver of the relationship than depth, and it then leveled out at the higher temperatures that occurred at $Z < 100$ m (Figure 8a). The smooth term for month (Figure 8b) predicted negative nitrate values from December through May and positive values from July through November, indicating the periods when this term corrected for overprediction and underprediction, respectively, relative to the T-only model. However, this term only made a modest contribution to nitrate prediction, as evidenced also by the wide scatter in the partial residuals (Figure 8b). Finally, the response of nitrate to

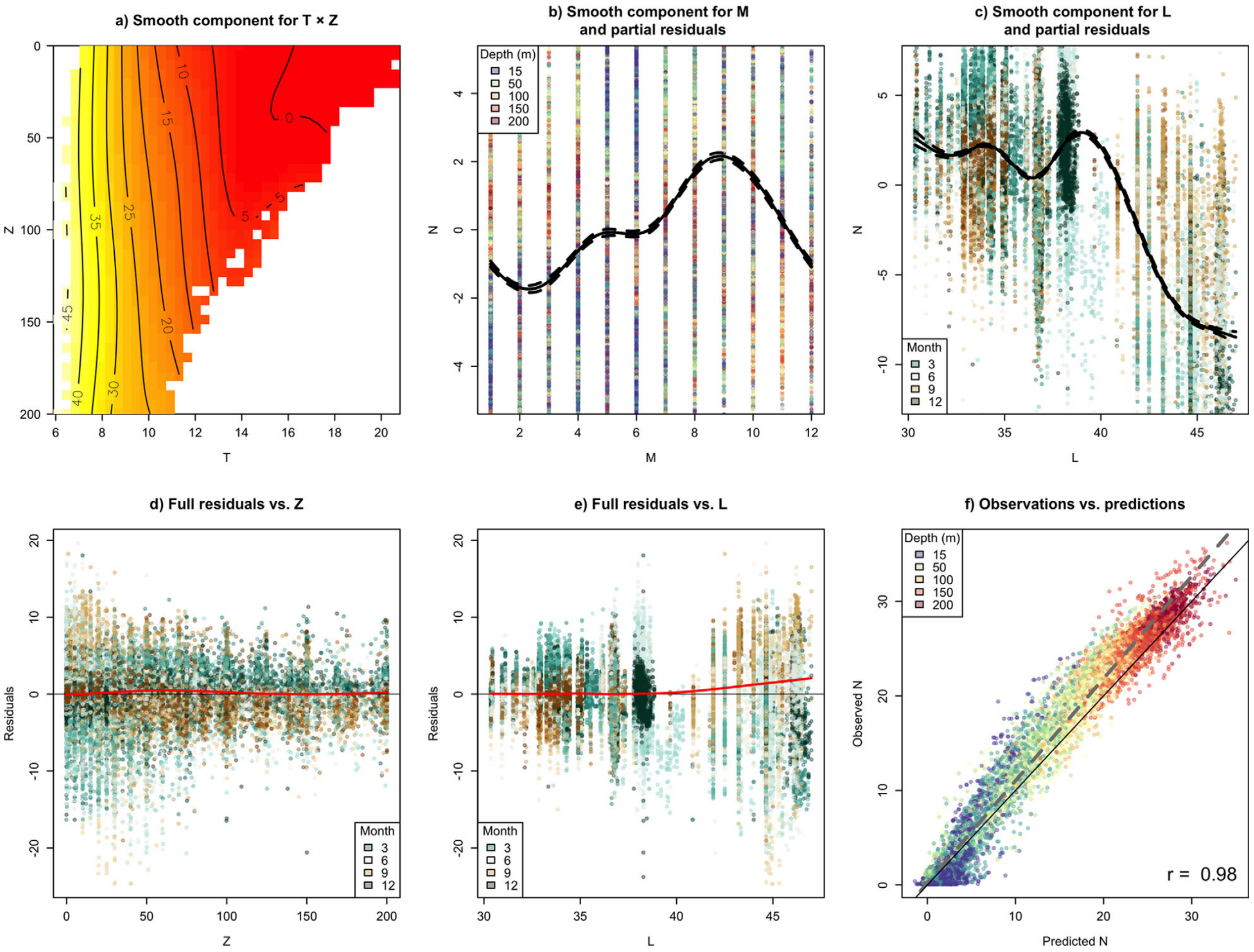

Figure 8. (a) Estimated effects (contoured surface) at the scale of the linear predictor for the $T \times Z$ interaction term for a GAM based on T, Z, M, and L for the training data set (1959–2004, $n = 37{,}607$). (b and c) Estimated effects (solid black curves) at the scale of the linear predictor for the M and L smooth terms, respectively, for a GAM based on T, Z, M, and L for the training data set (1959–2004, $n = 37{,}607$). The 95% confidence limits (strictly Bayesian credible intervals) are shown as dashed black lines. Dots are the partial residuals, colored by Z in Figure 8b or by M in Figure 8c (note that the y axis has been constrained in these two plots to highlight the functional shape of the smooth functions). (d) Model residuals versus Z (colored by M). (e) Model residuals versus L (colored by M). Red curve in Figures 8d and 8e is a loess scatterplot smoother (degree $= 2$, span $= 3/4$). (f) Scatterplot of observed versus predicted values (colored by Z) for a GAM based on T, Z, M, and L for the testing data set (2005–2011, $n = 6430$). Dashed gray line is the linear fit and black line is the 1:1 line.

latitude was described by a large-scale gradient, with nitrate decreasing with increasing latitude (Figure 8c). This term corrected for overprediction in the *T*-only model, especially at the higher latitudes. The two local maxima near 33–35°N and 38–40°N along this gradient (Figure 8c) probably indicate localized areas of increased nitrate supply due to intense upwelling centered around Point Conception and Cape Mendocino, respectively.

[47] The model's residuals indicated that the depth and latitudinal bias were largely corrected but there was evidence that some seasonal bias still remained (Figures 8d and 8e), probably because of the wide scatter in the relationship of this variable with nitrate (Figures 4b and 8b). Further examination of the residuals indicated an improvement over the *T*-only model, with levels of nonnormality and heterogeneity similar to those of the *TS* model (Figure S5 in supporting information).

[48] The performance of this model at predicting new observations was evaluated using the testing set, with most metrics being very similar to those of the *TS* model (Table 6). A scatterplot of observed versus predicted nitrate had a slope that departed moderately from the 1:1 relationship model and the shape of the relationship had a slight curvature (Figure 8f), leading to underprediction at intermediate nitrate levels and to overprediction at both low and high nitrate levels (Figure 8f).

6. Application: Predicting Nitrate Time Series

[49] We conducted further evaluation of three of the models developed in section 5 (*T*-only, *OTS*, and *TZML* models) to investigate the temporal behavior of predicted nitrate at seasonal and interannual scales. For this purpose, we inspected time series of predicted nitrate at 150 m depth from the testing data set (2005–2011) and assessed their skill relative to the observed series (via the root-mean-square error, RMSE) for two locations, one on the NH-Line and the other in the CalCOFI domain. These locations corresponded to station NH-25 (44°39.1′N, 124°39′W), located 46 km from shore on the NH-Line, and CalCOFI station 93.3.28 (station 28 on line 93.3; 32°54.6′N, 117°23.4′W), located 13 km from shore.

[50] Mean nitrate concentration at NH-25 (31.51 µ*M*) was appreciably higher than at CalCOFI station 93.28 (25.80 µ*M*), but otherwise both stations exhibited a similar seasonal cycle, with lowest levels in winter and fall and highest levels in spring and summer. The predicted time series by all three models captured the observed seasonality at both sites (Figures 9a and 9b). They also captured important interannual variations as seen in the high value in 2008 (spring for CalCOFI and summer for NH-25) and the low value in the winter of 2010 (Figures 9a and 9b). The high values of nitrate for both CalCOFI and NH-25 in spring/summer of 2008 were likely associated with the La Niña event of 2007–2008, when cool sea surface temperatures, strong upwelling, and high primary production prevailed in the CCS [*McClatchie et al*., 2009]. Conversely, the low nitrate values in winter 2010 occurred during the short and weak El Niño event of 2009–2010 [*Bjorkstedt et al*., 2011].

[51] However, only the *OTS* model had a high predictive skill (RMSE = 2.39 µ*M* and 0.40 µ*M*, respectively for NH-25 and CalCOFI station 93.3.28). The *T*-only model had the lowest predictive skill (RMSE = 4.13 and 6.39 µ*M*, respectively for NH-25 and CalCOFI station 93.3.28). The

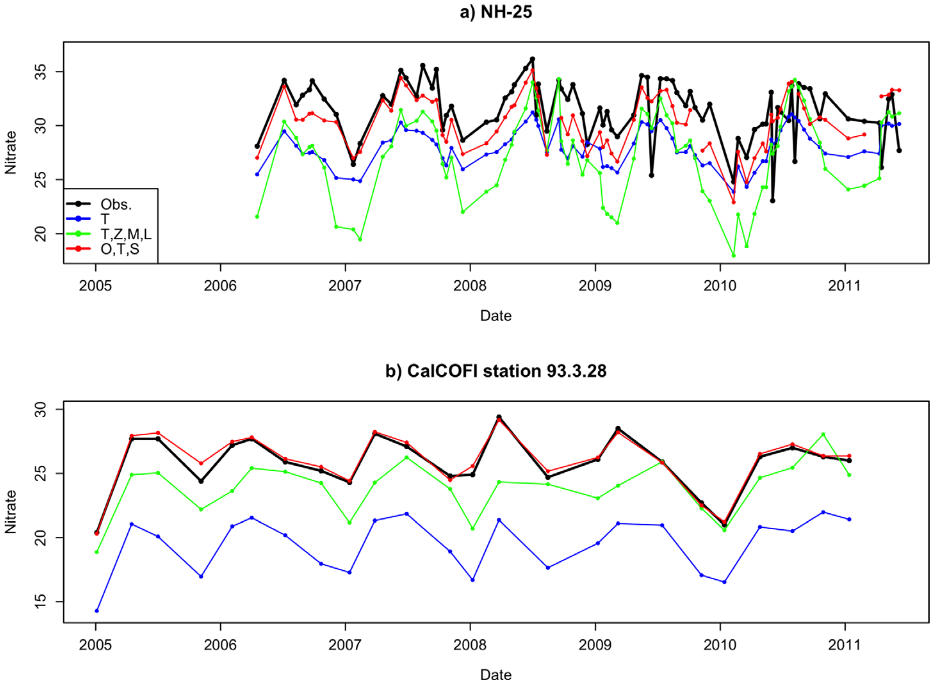

Figure 9. Time series of observed and predicted *N* in the testing data set (2005–2011) at 150 m depth for: (a) hydrographic station NH-25 located off Newport, Oregon, and (b) CalCOFI station 93.3.28 located near San Diego, California. For the two locations the observation data (black line) are compared to three predictive GAM models: *T*-only (blue line); *T*, *Z*, *M* and *L* (green line); and *O*, *T*, and *S* (red line). RMSE values for each of the predicted series are given in the text.

proxy *TZML* model significantly improved the skill for the CalCOFI station (RMSE = 2.50 μ*M*) but not for NH-25 (RMSE = 5.22 μ*M*).

[52] In spite of the corrective steps taken to minimize the biases in the *T*-only model with the *TZML* proxy model, it still underpredicted nitrate at both stations (Figures 9a and 9b). This is not surprising, considering the wide scatter of the residuals relative to the fitted responses for both month and latitude (Figures 8b and 8c). In contrast, the *OTS* model showed good agreement with the observations, although the slight underprediction at NH-25 (2.39 μ*M* RMSE) suggested that a regional bias remained, most likely due to the majority of the training data coming from southern California (see Table 1 and Figures 1 and 2), where nitrate levels tended to be less variable (Figure 3a).

7. Discussion and Conclusions

[53] The flexible GAM framework provided optimal nonlinear functional responses describing large-scale vertical and latitudinal gradients in nitrate in the CCS, as well as more complex interactions related to water mass distributions that could not be easily replicated with parametric models. However, the usefulness of GAMs to predict nitrate at larger spatial scales (basinwide to global) has not been evaluated, and future studies should compare their performance relative to traditional regression approaches, which have relied on implementing separate subregional models in order to improve predictive power [e.g., *Chavez et al*., 1996; *Goes et al*., 2000; *Louanchi and Najjar*, 2000].

[54] The four models considered here showed progressive improvement in the quality of the fit, in the residuals and in the predictive skill. A *T*-only model had a number of shortcomings due to strong depth and latitudinal biases, but it may be the only option when no other properties are measured. For localized studies with restricted depth ranges (e.g., in the nearshore) a *T*-only model is probably adequate, but for ecosystem-wide studies where different water masses are expected to occur, the inclusion of proxy variables containing the spatial structure underlying these processes may help alleviate some of these problems (alternatively, separate *T*-only models could be fitted to stratified subsets of the data). A model that included salinity in addition to temperature corrected much of the latitudinal and depth biases. Finally, a model with oxygen, temperature, and salinity provided the most unbiased predictions, indicating that, at a minimum, these three variables are necessary to adequately describe the spatial and temporal processes driving nitrate variability throughout the CCS. Although it is possible to use density as a variable that combines temperature and salinity, modeling nitrate explicitly as a function of *T* and *S* had the added benefit of allowing us to visualize water masses in *TS* space, which was helpful in understanding the spatial distribution of nitrate. For studies where this is not necessary, modeling nitrate as a function of density would be more parsimonious from a modeling perspective.

[55] Given the strong covariation of nitrate and temperature in the water column, it is not surprising that all the models explored had a good agreement between the response and the fitted values ($D^2 \sim 71.6$–96.6%; Table 5), as has been widely reported in the literature. Similarly, the overall ability of the GAMs to predict nitrate observations in the testing data set was very high ($r \sim 0.97$–0.99; Table 6). However, examination of the residuals and the skill at predicting individual time series exposed important biases in most models. The *OTS* model was the only one that yielded residuals that approached normality and that contained no significant structure (see supporting information). Seasonal bias was present in all models to varying degrees, and it was the least tractable source of bias due to the wide scatter in the relationship between nitrate and month. To some extent, these issues arise because of the varying degree of non-normality present in the input variables (Figures 2a–2d). But if the goal is to produce unbiased nitrate predictions, this study highlights the value of thorough assessment of residuals as a tool for model improvement over other metrics of model performance.

[56] Accurate estimation of nitrate concentration in the euphotic zone from proxy variables has several applications. For example, a nitrate inventory for the water column combined with information about thermocline depth and water-column stratification [*Palacios et al*., 2004] could be used to derive a more direct index of the biological utility of upwelled waters relative to existing wind-derived upwelling indices like the UI. This is not only relevant for the monitoring of primary production mediated by phytoplankton [*Saba et al*., 2011] and benthic macroalgae [*Broitman and Kinlan*, 2006] but also for the estimation of secondary production and fisheries yields [*Friedland et al*., 2012]. Biogeochemical and ecosystem numerical models that use nitrogen as currency [e.g., *Powell et al*., 2006; *Doney et al*., 2009; *Somes et al*., 2010] could also benefit from the statistical relationship between nitrate and variables like temperature, salinity, depth, and latitude, to constrain their nitrate budgets and thus obtain improved estimates of phytoplankton primary production and biomass.

[57] Finally, large-scale prediction of surface nitrate has relied on satellite measurements of sea surface temperature and chlorophyll concentration [e.g., *Goes et al*., 1999; *Kamykowski et al*., 2002, *Silió-Calzada et al*., 2008, *Sarangi*, 2011]. Given the improvement provided by our GAM models that included salinity over the *T*-only model, it is expected that the incorporation of satellite-measured sea surface salinity, which only recently became available [*Lagerloef*, 2012], into these efforts will result in improved maps of surface nitrate at regional and global scales.

[58] **Acknowledgments.** This study was initially motivated by discussions with Frank Schwing and Roy Mendelssohn. The LTOP nutrient data were collected by Pat Wheeler (College of Oceanic and Atmospheric Sciences, Oregon State University) and are distributed through the U.S. GLOBEC Data System. John Largier (Bodega Marine Laboratory, University of California at Davis) and Richard Dugdale and Al Marchi (Romberg Tiburon Center, San Francisco State University) kindly provided access to the CoOP/WEST data set. William T. Peterson (NOAA/NMFS, Northwest Fisheries Science Center) and Jay Peterson (Cooperative Institute for Marine Resource Studies, Oregon State University) generously contributed the NH-25 data to this project. We thank Xuemei Qiu for programming assistance. Support came from a grant from the U.S. GLOBEC Northeast Pacific program and from the NOAA Fisheries and the Environment (FATE) program. This is U.S. GLOBEC contribution No. 729. The manuscript benefited from comments by Andrew W. Leising, Paul C. Fiedler, Richard C. Dugdale, and two anonymous reviewers.

References

Bakun, A. (1973), Coastal upwelling indices, west coast of North America, 1946–71, U.S. Department of Commerce, *NOAA Tech. Report*, *NMFS-SSRF-671*, Natl. Oceanic Atmos. Admin.,Washington, D. C.

Bjorkstedt, E. P., et al. (2011), State of the California Currrent, 2010–2011: Regionally variable responses to a strong (but fleeting?) La Niña, *CalCOFI Rep., 52*, 36–68.

Bograd, S. J., D. A. Checkley, and W. S. Wooster (2003), CalCOFI: A half century of physical, chemical, and biological research in the California Current System, *Deep Sea Res. Part II*, *50*(14-16), 2349–2353, doi:10.1016/S0967-0645(03)00122-X.

Boyer, T. P., et al. (2009), World Ocean Database 2009, edited by S. Levitus, *NOAA Atlas NESDIS 66*, pp. 216, U.S. Gov. Print. Off., Washington D. C.

Broitman, B. R., and B. P. Kinlan (2006), Spatial scales of benthic and pelagic producer biomass in a coastal upwelling ecosystem, *Mar. Ecol. Prog. Ser.*, *327*, 15–25, doi:10.3354/meps327015.

Chavez, F. P., S. K. Service, and S. E. Buttrey (1996), Temperature-nitrate relationships in the central and eastern tropical Pacific, *J. Geophys. Res.*, *101*(C9), 20,553–20,563, doi:10.1029/96JC01943.

Checkley, D. M., and J. A. Barth (2009), Patterns and processes in the California Current System, *Progr. Oceanogr.*, *83*, 49–64, doi:10.1016/j.pocean.2009.07.028.

Doney, S. C., I. Lima, J. K. Moore, K. Lindsay, M. J. Behrenfeld, T. K. Westberry, N. Mahowald, D. M. Glover, and T. Takahashi (2009), Skill metrics for confronting global upper ocean ecosystem-biogeochemistry models against field and remote sensing data, *J. Mar. Syst.*, *76*(1-2), 95–112, doi:10.1016/j.jmarsys.2008.05.015.

Dugdale, R. C., A. Morel, A. Bricaud, and F. P. Wilkerson (1989), Modeling new production in upwelling centers: A case study of modeling new production from remotely sensed temperature and color, *J. Geophys. Res.*, *94*(C12), 18,119–18,132, doi:10.1029/JC094iC12p18119.

Dugdale, R. C., C. O. Davis, and F. P. Wilkerson (1997), Assessment of new production at the upwelling center at Point Conception, California, using nitrate estimated from remotely sensed sea surface temperature, *J. Geophys. Res.*, *102*(C4), 8573–8585, doi:10.1029/96JC02136.

Friedland, K. D., C. Stock, K. F. Drinkwater, J. S. Link, R. T. Leaf, B. V. Shank, J. M. Rose, C. H. Pilskaln, and M. J. Fogarty (2012), Pathways between primary production and fisheries yields of large marine ecosystems, *PLoS One*, *7*(1), e28945, doi:10.1371/journal.pone.0028945.t003.

Fuentes, M., B. Xi, and W. S. Cleveland (2011), Trellis display for modeling data from designed experiments, *Stat. Anal. Data Min.*, *4*, 133–145, doi:10.1002/sam.10102.

Garside, C., and J. C. Garside (1995), Euphotic-zone nutrient algorithms for the NABE and EqPac study sites, *Deep Sea Res. Part II*, *42*(2-3), 335–347, doi:10.1016/0967-0645(95)00026-M.

Goes, J. I., T. Saino, H. Oaku, and D. L. Jiang (1999), A method for estimating sea surface nitrate concentrations from remotely sensed SST and chlorophyll a—A case study for the North Pacific Ocean using OCTS/ADEOS data, *IEEE Trans. Geosci. Remote Sens.*, *37*(3), 1633–1644, doi:10.1109/36.763279.

Goes, J. I., T. Saino, H. Oaku, J. Ishizaka, C. S. Wong, and Y. Nojiri (2000), Basin scale estimates of sea surface nitrate and new production from remotely sensed sea surface temperature and chlorophyll, *Geophys. Res. Lett.*, *27*(9), 1263–1266, doi:10.1029/1999GL002353.

Hastie, T. J., and R. J. Tibshirani (1990), *Generalized Additive Models, Monographs on Statistics and Applied Probability 43*, Chapman and Hall/CRC, Boca Raton, Fla.

Henson, S. A., R. Sanders, J. T. Allen, I. S. Robinson, and L. Brown (2003), Seasonal constraints on the estimation of new production from space using temperature-nitrate relationships, *Geophys. Res. Lett.*, *30*(17), 1912, doi:10.1029/2003GL017982.

Huyer, A., P. A. Wheeler, P. T. Strub, R. L. Smith, R. Letelier, and P. M. Kosro (2007), The Newport line off Oregon–Studies in the North East Pacific, *Prog. Oceanogr.*, *75*(2), 126–160, doi:10.1016/j.pocean.2007.08.003.

Ihaka, R., and R. Gentleman (1996), R: A language for data analysis and graphics, *J. Comput. Graphical Stat.*, *5*, 299–314, doi:10.1080/10618600.1996.10474713.

Kamykowski, D., and S. -J. Zentara (1986), Predicting plant nutrient concentrations from temperature and sigma-t in the upper kilometer of the world ocean, *Deep Sea Res.*, *33*(1), 89–105, doi:10.1016/0198-0149(86)90109-3.

Kamykowski, D., S. -J. Zentara, J. M. Morrison, and A. C. Switzer (2002), Dynamic global patterns of nitrate, phosphate, silicate, and iron availability and phytoplankton community composition from remote sensing data, Global *Biogeochem. Cycles*, *16*(4), 1077, doi:10.1029/2001GB001640.

Lagerloef, G. (2012), Satellite mission monitors ocean surface salinity. *Eos Trans. AGU*, *93*(25), 233–234.

Louanchi, F., and R. G. Najjar (2000), A global monthly climatology of phosphate, nitrate, and silicate in the upper ocean: Spring-summer export production and shallow remineralization, *Global Biogeochem. Cycles*, *14*(3), 957–978, doi:10.1029/1999GB001215.

Marra, G., and S. N. Wood (2011), Practical variable selection for generalized additive models, *Comput. Stat. Data Anal.*, *55* (7), 2372–2387, doi:10.1016/j.csda.2011.02.004.

McClatchie, et al. (2009), State of the California Currrent, spring 2008–2009: Cold conditions drive regional differences in coastal production, *CalCOFI Rep.*, *50*, 43–68.

Morin, P., M. V. M. Wafar, and P. Le Corre (1993), Estimation of nitrate flux in a tidal front from satellite-derived temperature data, *J. Geophys. Res.*, *98*(C3), 4689–4695, doi:10.1029/92JC02445.

Palacios, D. M., S. J. Bograd, R. Mendelssohn, and F. B. Schwing (2004), Long-term and seasonal trends in stratification in the California Current, 1950–1993, *J. Geophys. Res.*, *109*, C10016, doi:10.1029/2004JC002380.

Peña, M. A., and S. J. Bograd (2007), Time series of the Northeast Pacific, *Prog. Oceanogr.*, *75*(2), 115–119, doi:10.1016/j.pocean.2007.08.008.

Powell, T. M., C. V. W. Lewis, E. N. Curchitser, D. B. Haidvogel, A. J. Hermann, and E. L. Dobbins (2006), Results from a three-dimensional, nested biological-physical model of the California Current System and comparisons with statistics from satellite imagery, *J. Geophys. Res.*, *111*, C07018, doi:10.1029/2004JC002506.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna. [Available at http://www.R-project.org/, accessed on 3 June 2013.].

Roy, C. (1991), Relations entre sels nutritifs et chlorophylle: Une approche exploratoire, in *Pêcheries Ouest Africaines: Cariabilité, Instabilité et Changement*, edited by P. Cury and C. Roy, pp. 390–394, Editions de l'ORSTOM, Paris.

Saba, V. S., et al. (2011), An evaluation of ocean color model estimates of marine primary productivity in coastal and pelagic regions across the globe, *Biogeosciences*, *8*(2), 489–503, doi:10.5194/bg-8–489-2011.

Sarangi, R. K. (2011), Remote-sensing-based estimation of surface nitrate and its variability in the southern peninsular Indian waters, *Int. J. Oceanogr.*, Article ID 172731, doi:10.1155/2011/172731.

Sarmiento, J. L., and N. Gruber (2009), *Ocean Biogeochemical Dynamics*, Princeton Univ. Press, Princeton, N. J.

Sathyendranath, S., T. Platt, E. P. W. Horne, W. G. Harrison, O. Ulloa, R. Outerbridge, and N. Hoepffner (1991), Estimation of new production in the ocean by compound remote sensing, *Nature*, *353*(6340), 129–133, doi:10.1038/353129a0.

Schwing, F., M. O'Farrell, J. Steger, and K. Baltz, (1996), Coastal upwelling indices, west coast of North America, 1946–1995, U.S. Dep. of Commerce, NOAA Tech. Memo. NOAA-TM-NMFS-SWFSC-231, La Jolla, Calif.

Silió-Calzada, A., A. Bricaud, and B. Gentili (2008), Estimates of sea surface nitrate concentrations from sea surface temperature and chlorophyll concentration in upwelling areas: A case study for the Benguela system, *Remote Sens. Environ.*, *112*(6), 3173–3180, doi:10.1016/j.rse.2008.03.014.

Somes, C. J., A. Schmittner, E. D. Galbraith, M. F. Lehmann, M. A. Altabet, J. P. Montoya, R. M. Letelier, A. C. Mix, A. Bourbonnais, and M. Eby (2010), Simulating the global distribution of nitrogen isotopes in the ocean, *Global Biogeochem. Cycles*, *24*, GB4019, doi:10.1029/2009GB003767.

Steinhoff, T., T. Friedrich, S. E. Hartman, A. Oschlies, D. W. R. Wallace, and A. Körtzinger (2010), Estimating mixed layer nitrate in the North Atlantic Ocean, *Biogeosciences*, *7*(3), 95–807, doi:10.5194/bg-7–795-2010.

Switzer, A. C., D. Kamykowski, and S. -J. Zentara (2003), Mapping nitrate in the global ocean using remotely sensed sea surface temperature, *J. Geophys. Res.*, *108*(C8), 3280, doi:10.1029/2000JC000444.

Tabachnik, B. G., and L. S. Fidell (1989), *Using Multivariate Statistics*, Harper and Row, New York, N. Y.

Tabachnik, B. G., and L. S. Fidell (2001), *Using Multivariate Statistics*, 4th ed., Allyn and Bacon, Boston, Mass.

Traganza, E. D., V. M. Silva, D. M. Austin, W. L. Hanson, and S. H. Bonsink (1983), Nutrient mapping and recurrence of coastal upwelling centers by satellite remote sensing: Its implication to primary production and the sediment record, in *Coastal Upwelling, Part A*, edited by E. Suess and J. Thiede, pp. 61–83, Plenum, New York, N. Y.

Wessel, P., and W. H. F. Smith (1996), A Global self-consistent, hierarchical, high-resolution shoreline Database, *J. Geophys. Res.*, *101*(B4), 8741–8743, doi:10.1029/96JB00104.

Wood, S. N. (2006), Generalized Additive Models: An introduction with R, in *Texts in Statistical Science*, Chapman and Hall/CRC, Boca Raton, Fla.

Zuur, A. F., E. N. Ieno, and C. S. Elphick (2009), A protocol for data exploration to avoid common statistical problems, *Methods Ecol. Evol.*, *1*(1), 3–14, doi:10.1111/j.2041-210X.2009.00001.x.