Peer-reviewed version available at PLoS ONE 2018; doi:10.1371/journal.pone.0201255

Article

Variance Inflation Factor-Based Forward-Selection Method for Water-Quality Estimation via Combining Landsat TM, ETM+, and OLI/TIRS Images and Ancillary Environmental Data

Min-Cheng Tu [1,2,*], Patricia Smith [3] and Anthony Filippi [4,5]

[1] Department of Civil and Environmental Engineering, Villanova University, Villanova, PA 19085, USA
[2] Villanova Urban Stormwater Partnership (VUSP), Villanova University, Villanova, PA 19085, USA
[3] Department of Biological and Agricultural Engineering, Texas A&M University, College Station, TX 77843, USA; patti-smith@tamu.edu
[4] Department of Geography, Texas A&M University, College Station, TX 77843, USA; filippi@geos.tamu.edu
[5] Center for Geospatial Science, Applications and Technology (GEOSAT), Texas A&M University, College Station, TX 77843, USA

* Correspondence: min-cheng.tu@villanova.edu; Tel.: +1-610-519-6518

Abstract: A simple approach to enable water-management agencies employing free data to achieve the goal of using a single set of predictive equations for water-quality retrievals with satisfactory accuracy is proposed. Multiple regression-derived equations based on surface reflectance, band ratios, and environmental factors as predictor variables for concentrations of Total Suspended Solids (TSS), Total Nitrogen (TN), and Total Phosphorus (TP) were derived using a hybrid forward-selection method that considers Variance Inflation Factor (VIF) in the forward-selection process. Landsat TM, ETM+, and OLI/TIRS images were jointly utilized with environmental factors, such as wind speed and water surface temperature, to derive the single set of equations. The coefficients of determination of the best-fitting resultant equations varied from 0.62 to 0.79. Among all chosen predictor variables, ratio of reflectance of visible red (Band 3 for Landsat TM and ETM+, or Band 4 for Landsat OLI/TIRS) to visible blue (Band 1 for Landsat TM and ETM+, or Band 2 for Landsat OLI/TIRS) has a strong influence on the predictive power for TSS retrieval. Environmental factors including wind speed, remote sensing-derived water surface temperature, solar altitude, and time difference (in days) between the image acquisition and water sampling were found important in water-quality parameter estimation.

Keywords: Variance Inflation Factor; VIF; multiple regression; Landsat; Austin; Lady Bird Lake; water quality; environmental factor; energy flux; urban runoff

1. Introduction

Continuous monitoring of water quality is essential for the health and welfare of the people and ecosystems reliant upon them. Urbanization, agriculture, and other anthropogenic factors can alter water quality [1], and waiting to remediate until a change is clearly visible can be much more costly than early prevention. Despite this, the cost of adequate temporal and spatial physical measurements can potentially be prohibitive [2]. For example, the United States Geological Survey (USGS) regularly monitors water quality in Lady Bird Lake in Austin, Texas, USA; however, the frequency is only approximately twice per year at a single point near the outlet over the past decade [3]. Additionally, *in situ* measurements from year to year do not occur in the same months. As a result, it is difficult to distinguish whether a change in the water quality measured at a point is truly a long-term change or the result of a seasonal difference or recent event (e.g., a large precipitation

event) [4]. Additionally, it is impossible to evaluate the spatial variation in water quality from single-point measurements.

In recent decades, remote sensing has provided an alternative method for monitoring water quality in a spatially synoptic manner at a lower cost compared with extensive *in situ* measurement. Each water-column constituent exhibits a specific spectral response that can be observed by satellite- and aircraft-mounted remote sensors [5]. Suspended sediment usually exhibits strong backscattering of incident light [5], where the actual color depends on the terrestrial origin [6]. Colored dissolved organic matter (CDOM) is composed of algae, yellow substances, and organic plumes [5], and entails a broad-band solar-induced fluorescence over 490-530 nm [6]. Phytoplankton exhibits a volume reflectance (and water-leaving radiance) peak due to *chlorophyll-a*, with a well-defined Gaussian distribution around 685 nm [6].

For a particular wavelength, λ, the spectral radiance from the water observed vertically, known as the upwelling radiance, L_u, is given by

$$L_u(\lambda) = L_w(\lambda) + \Omega L_s(\lambda) \quad (1)$$

where L_w is the radiance reflected/backscattered by the water column, in-water constituents, and the bottom if the water column is optically shallow; L_s is the skylight radiance; and Ω is the ratio of radiance directly reflected by the water surface to L_s [7]. Note that the radiance observed by a satellite is composed of L_u, plus atmospheric interference; therefore, it requires atmospheric correction (discussed below). L_w, L_s, and Ω are influenced by a variety of factors. If the water column is sufficiently deep, bottom reflectance may be ignored, and L_w can be assumed to be a measure of the effects of water-column constituents alone. Atmospheric conditions (e.g., clear, cloudy, overcast) affect both Ω and L_s, whereas Ω can be further affected by wind speed in the form of surface ripples [7]. Wind speed has also been found to have some influence on water clarity [8].

Because of their higher capability to penetrate the water column, visible bands have conventionally been used to estimate water quality [5]. In addition, infrared bands have also shown significance in determining water-quality parameters in some studies [9, 10]. However, only near infrared wavelengths were used in these studies. Thermal infrared bands have not extensively been used in water-quality estimation.

Site-specific predictive models can be created to relate a number of band radiance measurements or derived reflectance values [5] to the water-quality parameter of interest by fitting the model to *in situ* water-quality measurements. Multiple regression analysis and artificial neural networks (ANNs) constitute two methods that are frequently used to generate such predictive models [5, 10, 11, 12].

In academia, satellite remote-sensing images have been increasingly available for water-quality determination. However, the popularity of this approach has not been extended to decision making by management agencies in general [13]. According to Schaeffer et al. [13], the reasons for this phenomenon include cost, product accuracy, data continuity, and programmatic support.

Cost is always a major constraint, as many water-management agencies have limited budgets [13]. Even though there are many free remote-sensing data sets available, such as the multispectral satellite images available from the Landsat program (e.g., Landsat Thematic Mapper (TM), Enhanced Thematic Mapper Plus (ETM+), and Operational Land Imager (OLI)/ Thermal Infrared Sensor (TIRS)) [14], MODIS [15], SeaWiFS [16], etc., the selection of images is predominately limited to moderate spatial-resolution images from the Landsat program, for example, for terrestrial pond/lake applications due to the finer spatial resolution of those data relative to other free remote-sensor image sources and the relatively small sizes/spatial extents of such features. Another aspect of the cost constraint is the cost to collect field water-sampling data, as the creation of empirical predictive models necessitates *in situ* water-quality data. Sometimes, due to cost, logistical, and other constraints, that means that a water-management agency can only resort to free water-quality data, such as those made available by the USGS. The downside, as noted above, is that spatio-temporal sampling density/data availability may be low. This drawback seriously limits the ability of a water-management agency to utilize free Landsat program data, for example, as the basis

of a water-quality monitoring program since the satellite images and corresponding *in situ* measurements must be acquired in a temporally proximal manner [17].

As a result, water-management agencies that resort to only using free remote-sensing resources often only have access to a limited number of useable satellite images for water-quality monitoring. Such a scenario often leads to the use of a single predictive model to determine water-quality information from satellite images. Nevertheless, many studies divide their analyses by season [18, 19] due to systemic seasonal differences in factors such as concentrations of color-producing substances (including phytoplankton), atmospheric disturbances [19], and solar zenith angle [20]. Some studies have shown that the predictive power of equations created without distinguishing by season is lower than it otherwise would be [21, 22].

Since the derived predictive equation is seasonally affected by the environment, a few studies have incorporated the influencing factors into predictive equation generation. One example is with the estimation of *chlorophyll-a* concentration. It has been known that phytoplankton growth is statistically significantly dependent on water temperature [23, 24]. Incorporating water temperature (derived from the satellite remote-sensor thermal band) in development of predictive equations has proven to be helpful in determining *chlorophyll-a* concentration [25]. However, this approach has not been investigated extensively. In this study, we consider additional environmental factors based on energy fluxes between a waterbody and the atmosphere. We posit that including these environmental factors in predictive equations not only increases prediction accuracy, but also facilitates the usage of a single set of predictive equations throughout different seasons. The direct benefit is that one can pool all observation data in creating equations, thus resulting in higher predictive power.

Programmatic support is also important to water-management agencies, according to Schaeffer et al. [13]. In most cases, local universities should be sufficient in providing support to water-management agencies. However, we posit that the methodology adopted for generating predictive models should entail model construction in a stepwise manner, such that most people with basic training could implement such methods can follow without much difficulty. For this reason, in choosing methodology implemented by water-management agencies, simple and well-understood methods such as multiple regressions should be weighed over more complex methods, such as ANNs.

Product accuracy is another major concern expressed by the water-management agencies [13]. Even though water-management agencies could utilize predictive models from peer-reviewed journals, such models may not yield high-accuracy estimates in a given application. Multiple regression analysis has been employed in many studies for its ease of application. However, for applications using this method, overfitting from multicollinearity can be a serious concern. Multicollinearity means that some of the explanatory variables in the multiple regression model are dependent on one another. The direct result from multicollinearity is that the standard error of coefficients of explanatory variables is inflated, which means that coefficients of the derived model are not reliable. Unfortunately, many past studies neither discussed the issue of multicollinearity, nor provided results of validation of the derived regression models [4, 5, 9, 17, 26, 27, 28, 29]. A common way to identify multicollinearity of a model is through the usage of indicators such as Akaike's Information Criteria [30], Mallow's Cp [31], PRESS [32], etc. However, such indicators apply to the whole model so all possible subsets of explanatory variables must be examined, and this approach becomes unattainable when the number of variables increases [33].

Other popular methods to identify multicollinearity include the deployment of a principal component analysis (PCA) or structural equation modeling (SEM) [33]. PCA creates orthogonal principal components, which are linear combination of variables, and a regression model can be created based on the orthogonal components in order to eliminate multicollinearity completely. Some studies show, however, that this methodology can result in a loss in explanatory power. Additionally, the main limitation of the PCA approach is rooted in the physical interpretation of the principal components. On the other hand, SEM accepts the existence of collinearity among explanatory variables and hypothesizes that a model exists among variables. Then all possible

combinations of causal links among variables are tested against the hypothesized model. Since SEM is not an exploratory technique, SEM is prone to inferential errors made during development and selection of the hypothetical models [33].

We propose utilizing the variation inflation factor (VIF) to minimize multicollinearity. Unlike other indicators described above, VIF is calculated for each predictor variable. VIF has been used in the field of remote sensing on a limited basis to check multicollinearity of results [34; 35]. Dubovyk et al. [36] used VIF to choose variables to enter into a logistic regression model. However, to our knowledge, VIF has not previously been employed in deriving predictive equations for water-quality parameters. Details regarding VIF computation and the methodology to include VIF in equation derivation is discussed below in the Methodology section.

The Landsat program constitutes a truly ideal free data-source candidate for water-management agencies, given the characteristics of the various Landsat sensors, as well its long-term data continuity. The Landsat program has maintained the longest uninterrupted satellite observation record of Earth from its beginning in 1970s. The Landsat program has employed several sensors over time including MSS, TM, ETM+, and OLI/TIRS (Landsat 8). Only a few water-quality studies have taken advantage of combining TM, ETM+, and OLI/TIRS datasets [37, 38, 39] even though these sensors have been shown to be compatible, as shown in Table 1 [39, 40, 41]. Note Table 1 shows only comparable bands among Landsat TM, ETM+, and OLI/TIRS sensors.

Due to the different band numbering in OLI/TIRS, in this study, numbering of bands will be based on TM/ETM+. For example, if Band 3 is noted, it means Band 3 for TM and ETM+, but Band 4 for OLI/TIRS.

Table 1. Band attributes of Landsat TM and ETM+ and OLI/TIRS sensors [39, 40, 41].

		Band 1	Band 2	Band 3	Band 4	Band 5	Band 6	Band 7	Band 8
TM	**Wavelength (μm)**	0.45-0.52	0.52-0.60	0.63-0.69	0.76-0.90	1.55-1.75	10.40-12.50	2.08-2.35	n/a
	Sensor spatial resolution (m)	30	30	30	30	30	60	30	n/a
ETM+	**Wavelength (μm)**	0.45-0.52	0.52-0.60	0.63-0.69	0.77-0.90	1.55-1.75	10.40-12.50	2.09-2.35	0.52-0.90
	Sensor spatial resolution (m)	30	30	30	30	30	60	30	15
OLI/TIRS	**Wavelength (μm)**	Band 2	Band 3	Band 4	Band 5	Band 6	Band 10	Band 7	Band 8
		0.45-0.51	0.53-0.59	0.64-0.67	0.85-0.88	1.57-1.65	10.60-11.19	2.11-2.29	0.50-0.68
	Sensor spatial resolution (m)	30	30	30	30	30	100	30	15

Based on the gaps in the research literature illustrated above, the objectives of this study were:

1. Incorporate environmental factors (such as temperature, wind speed, etc.) into a single set of predictive equations for remote-sensing water-quality parameter estimation; and

2. Increase model predictive power for a limnological water-quality parameter-estimation application by considering the effect of multicollinearity in model creation.

The goal of this study is to address all four concerns of utilizing satellite data in decision making by water-management agencies—i.e., cost, product accuracy, data continuity, and programmatic support. This study provides water-management agencies with a simple, easy-to-follow methodology for utilizing free observation data (from Landsat program, USGS, etc.) in order to address cost and programmatic-support issues for water-quality monitoring. The Landsat program guarantees long-term data continuity. The proposed methodology provides a single set of predictive equations; accuracy is maintained because all available data are consolidated for the creation of a single model. Also, consideration of multicollinearity increases the likelihood for acceptable estimation accuracy of the derived model in future water-quality parameter retrieval applications.

2. Methodology

2.1. Study Area

The population of City of Austin, Texas, USA has increased dramatically in recent decades, from 252,000 in 1970 to 926,000 in 2016 [42]. With significant population growth comes an increase in impervious area, higher runoff and lower water quality in local water bodies. Lady Bird Lake (formerly Town Lake), situated near the city center, provides an opportunity to remotely monitor water quality in an urban watershed (Figure 1). The lake, formed by damming the Colorado River, is maintained at an approximately constant level by the pass-through Longhorn Dam [43]. The surface area is ~173.6 hectares with a capacity of 905.1 ha-m. The mean depth is 6 meters, with a maximum depth over 11.7 meters [44].

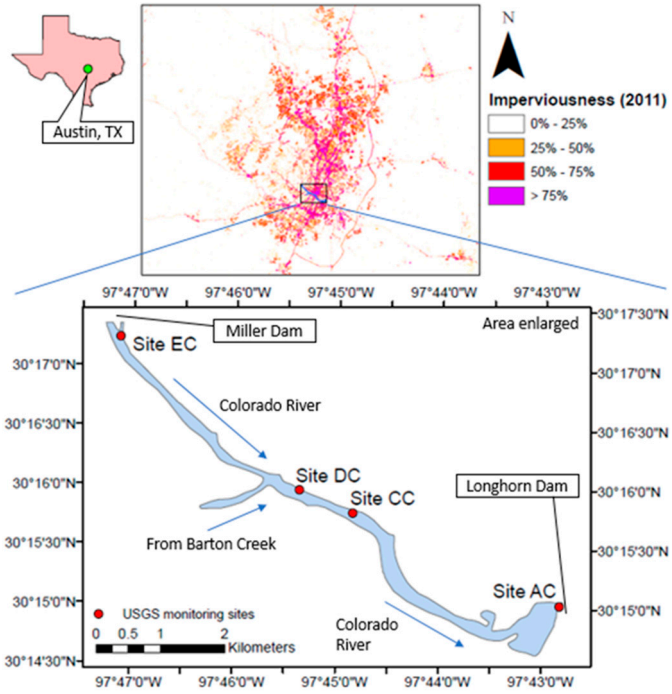


Figure 1. Locations of water-quality sampling stations (i.e., Sites AC, CC, DC, and EC) on Lady Bird Lake.

The USGS maintains a number of water-quality sampling stations on Lady Bird Lake, but only four of them, EC, DC, CC and AC (Figure 1), monitor the water-quality constituents of interest in this study within the time frame of available satellite images (i.e., 1983-2015) [3]. Table 2 provides basic information for these four sampling stations, including summary statistics for these water-quality quantities of interest—total suspended solids (TSS), total nitrogen (TN), and total phosphorus (TP)—derived from water-quality samples collected at a depth of 1 m. Secchi disc transparency, a pseudo-measure of turbidity, was measured in four locations when the samples of Table 2 were taken (Table 3). Secchi disc depths were much shallower than the average bottom depth of the lake (6 m); thus, bottom reflection is not observable from above the air-water interface for these cases. Therefore, contribution of bottom reflectance to the water-leaving radiance (Equation 1) can be ignored.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2017 doi:10.20944/preprints201710.0108.v1

Peer-reviewed version available at PLoS ONE 2018; doi:10.1371/journal.pone.0201255

Table 2. Summary statistics from *in situ* USGS water-quality stations in Lady Bird Lake, Texas, USA, over the time period 1983-2015.

USGS Water Quality Stations and Site Codes	Water-Quality Measures and USGS Parameter Code								
	TSS (mg/L) 00530*			TN (mg/L) 00600*			TP (mg/L) 00665*		
	# of Samples	Mean	Std. Dev.	# of Samples	Mean	Std. Dev.	# of Samples	Mean	Std. Dev.
EC 301712097470701**	7	4.57	4.24	11	0.58	0.22	4	0.015	0.0058
DC 301558097452201**	8	5.75	5.39	8	0.71	0.36	8	0.023	0.017
CC 301546097445101**	4	9.50	5.26	6	0.53	0.14	3	0.023	0.012
AC 301500097424801**	9	8.44	10.35	13	0.71	0.25	10	0.035	0.035
All	28	8.19	7.96	38	0.65	0.24	25	0.026	0.024

* Water-quality parameter code as assigned by USGS

** USGS station number

Table 3. Secchi disc transparency measurements for *in situ* USGS water-quality stations in Lady Bird Lake, Texas, USA, over the time period 1983-2015

Site Code	# of measurements	Mean (m)	Std. Dev. (m)
EC	11	2.22	0.86
DC	10	1.68	0.77
CC	8	1.23	0.62
AC	15	1.27	0.60

2.2. Selection of Satellite Images

Selection of Landsat TM, ETM+, and OLI + TIRS images [45] was based on several criteria. Images selected were cloud-free and were acquired within seven days of *in situ* water-quality measurements in Lady Bird Lake [10, 17]. In order to minimize the effects of spatio-temporally-close rainfall events, only images that entailed daily precipitation depths less than 1.25 cm (0.5 inch) observed between the dates of the selected images and their associated water-sampling dates (Table 4) were selected. This threshold rainfall depth is chosen based on the initial abstraction rainfall depth for a watershed with a runoff curve number of 80, since most of the urbanized area around Lady Bird Lake is residential [46]. Residential districts with small lot sizes (1/4 to 1/8 acre) have a curve number ranging from 61 to 92, depending on the soil hydrologic group [47]. Rainfall depth below this threshold is considered to generate insignificant runoff, and thus should have no marked effect on water quality in the lake.

Table 4. Dates of Landsat TM and ETM+ satellite images utilized and respective corresponding water-quality samples.

Sensor Name	Image Date	Water-Quality Sampling Date
Landsat 4 TM	January 9, 1983	January 6, 1983
Landsat 5 TM	August 18, 1985	August 20, 1985
Landsat 5 TM	January 15, 1988	January 19, 1988
Landsat 5 TM	April 20, 1988	April 19, 1988
Landsat 5 TM	July 25, 1988	July 27, 1988

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2017 doi:10.20944/preprints201710.0108.v1

Peer-reviewed version available at PLoS ONE 2018; doi:10.1371/journal.pone.0201255

Landsat 5 TM	March 6, 1989	February 27, 1989
Landsat 5 TM	April 7, 1989	April 12, 1989
Landsat 5 TM	August 5, 1992	August 10, 1992
Landsat 5 TM	July 24, 1999	July 22, 1999
Landsat 5 TM	December 20, 2001*	December 16, 2001
Landsat 7 ETM+	April 22, 2009	April 18, 2009
Landsat 5 TM	June 4, 2010	June 3, 2010
Landsat 7 ETM+	May 14, 2011	May 13, 2011
Landsat 8 (OLI + TIRS)	May 14, 2014	May 14, 2014
Landsat 8 (OLI + TIRS)	March 14, 2015	March 10, 2015

* Excluded from analysis due to issues with atmospheric correction.

2.3. Atmospheric Correction

2.3.1. FLAASH Theory

Surface reflectance values, ρ, corrected for path radiance, were derived using Fast Line-of-sight Atmospheric Analysis of Spectral Hypercube (FLAASH®) radiative transfer model [48, 49]. Pixel values were converted from a digital number (DN) to spectral radiance (L) following the Landsat Data User Handbook [41], and then corrected to surface reflectance using FLAASH. Note that

$$L = \left(\frac{A\rho}{1-\rho_e S}\right) + \left(\frac{B\rho_e}{1-\rho_e S}\right) + L_a \quad (2)$$

where L is the spectral radiance observed by the sensor; ρ is the "correct" surface reflectance for the pixel of interest; ρ_e is the average surface reflectance from the pixel of interest and the surrounding region; S is the spherical albedo of the atmosphere; L_α is the path radiance backscattered by the atmosphere; and A and B are coefficients dependent upon atmospheric and geometric conditions. The distinction between ρ and ρ_e accounts for adjacency/spatial-mixing effects [48, 49].

2.3.2. Image Pre-processing

Because water bodies such as Lady Bird Lake are often generally spectrally dark targets [50], remote-sensing reflectance from such areas is usually lower than the surrounding urban areas. With FLAASH, significant errors can occur when strong albedo contrasts exist among the materials in the scene [49]. To minimize this potential problem, a land mask was created and applied in order to exclude all surrounding land regions [51], leaving just the aquatic areas (i.e., Lady Bird Lake) for subsequent atmospheric-correction processing.

2.3.3. Determination of FLAASH Parameter Values

Two of the parameters required by FLAASH are: visibility and choice of atmospheric model. Visibility obtained from historical airport records [52] caused FLAASH to over-compensate in its correction of atmospheric effects and yield negative reflectance values. Therefore, the 2-band (K-T) aerosol retrieval method [49] with "urban" setting was used to estimate visibility. Ideally, selection of an atmospheric model is based on one of the following options, presented in order from most preferred to least preferred: known standard column water vapor amount, expected surface air temperature, or tabulated seasonal-latitude combinations [49]. Although there are atmospheric water-content products available [53], they do not cover all dates of interest in this research. Surface temperatures have been continuously recorded and archived by Camp Mabry Austin City Airport and Austin Bergstrom International Airport every hour over the past 30 years [52]. Therefore, atmospheric models were selected based on the surface air temperature at the time when each satellite image was acquired (Table 5). The initially-selected December 20, 2001 image was excluded from subsequent processing because it yielded negative reflectance values after FLAASH atmospheric correction.

Table 5. Selection of FLAASH atmospheric model based on measured surface air temperature.

Image Date	Surface Air Temperature (°C)	Chosen Atmospheric Model	Suggested Temperature for Model (°C) [49]
January 9, 1983	11	Sub-Arctic Summer	14
August 18, 1985	33	Tropical	27
January 15, 1988	10	Sub-Arctic Summer	14
April 20, 1988	23	Mid-Latitude Summer	21
July 25, 1988	31	Tropical	27
March 6, 1989	2	Mid-Latitude Winter	-1
April 7, 1989	25	Tropical	27
August 5, 1992	30	Tropical	27
July 24, 1999	32	Tropical	27
December 20, 2001	11	Sub-Arctic Summer	14
April 22, 2009	31	Tropical	27
June 4, 2010	31	Tropical	27
May 14, 2011	23	Mid-Latitude Summer	21
May 14, 2014	21	Mid-Latitude Summer	21
March 14, 2015	22	Mid-Latitude Summer	21

2.3.4. Determination of FLAASH Parameter Values

FLAASH should not be applied to thermal bands [49]; therefore, another atmospheric-correction method was applied to thermal bands. In particular, the single-band atmospheric-correction method described by Barsi et al. [54] was used. The methodology calculates atmospheric transmission and path radiance using MODTRAN [49], based on the atmospheric profiles generated by National Centers for Environmental Prediction (NCEP). Equation 3 provides the relationship between top-of-atmosphere radiance (L_{TOA}), the target radiance of kinetic temperature T (L_T), the path (upwelling) radiance (L_u), and the sky (downwelling) radiance (L_d):

$$L_{TOA} = \tau\varepsilon L_T + L_u + \tau(1-\varepsilon)L_d \quad (3)$$

In Equation 3, atmospheric transmission τ, path radiance L_u, and sky radiance L_d were obtained from the on-line calculator based on the atmospheric correction method of Barsi et al. [54]. Since water is a near-perfect blackbody, emissivity (ε) was set as 1 in this study. Emissivity and transmission are unitless, whereas radiance values are in units of W/m2·sr·μm.

The atmospheric profiles are only available after January 2000. For satellite images acquired prior to that, atmospheric profiles from "surrogate dates" in 2000 were used in this study. The surrogate date has nearly identical daily precipitation, temperature, and wind speed as the satellite image date. By choosing a surrogate date in such a manner, the atmospheric condition of the actual satellite image date and the surrogate date are expected to be similar. If more than two surrogate

dates were found based on the above criteria for one satellite image, the one that is temporally closest to the date in the year in which a given the satellite image was acquired was chosen. Table 6 provides the list of the satellite image dates, the corresponding surrogate dates, and daily meteorological parameters for both of them.

Table 6. Comparison between image and surrogate dates in atmospheric profile determinations.

	Image date weather parameters				Surrogate date weather parameters		
Satellite Date	**Daily rainfall (mm)**	**Daily Mean Temp (°C)**	**Daily mean wind speed (m/s)**	**Surrogate Date**	**Daily rainfall (mm)**	**Daily Mean Temp (°C)**	**Daily mean wind speed (m/s)**
Jan 9, 1983	0	11	3.1	**Dec 20, 2000**	0	11	3.1
Aug 18, 1985	0	31	3.6	**Aug 28, 2000**	0	32	3.6
Jan 15, 1988	0	9	2.8	**Nov 13, 2000**	0	9	3.6
Apr 20, 1988	0	21	3.6	**Apr 22, 2000**	0	20	3.4
Jul 25, 1988	0	30	3.1	**Jul 26, 2000**	0	31	3.2
Mar 6, 1989	0	3	5.8	**Dec 27, 2000**	0	3	4.1
Apr 7, 1989	0	22	2.8	**May 14, 2000**	0	23	2.8
Aug 5, 1992	0	29	3.1	**Aug 20, 2000**	0	30	3
Jul 24, 1999	0	29	1.7	**Jul 24, 2000**	0	29	1.6

2.3.5. Determining Surface Temperature from Landsat Thermal Bands

Target temperature (i.e., water surface temperature) was then derived after atmospheric correction according to equations provided in the Landsat Data User Manual [41]. For Landsat ETM+, the low-gain channel was used because the signal reflected from water-column constituents entail low signal strength. For Landsat TIRS, only band 10 was used because data from band 11 have been contaminated by a stray-light effect, and a remedy has not yet been found [55]. Bands 10 and 11 here are band numbering from Landsat TIRS.

2.3.6. Post-processing for Atmospherically-Corrected Surface Reflectance

Surface reflectance values at the water-quality stations were extracted from the FLAASH-corrected satellite images. Pixels located at the exact coordinates of the respective water-quality sampling stations are not necessarily the ideal pixels for which reflectance values should be extracted. Reasons for this include the geometric-offset error between the map coordinates of a pixel and actual corresponding *in situ* sampling planimetric locations; random surface debris; light unpredictably scattered or reflected into the instantaneous field-of-view (IFOV) of the sensor or onto the aquatic area of interest [56]; and optically shallow water near the sampling stations, yielding potentially confounding issues associated with bottom reflectance. To compensate for this, the search range was expanded to 90 m (i.e., a search neighborhood comprised of 3 x 3 image pixels, centered around the pixel located at the station coordinates). The pixel within this zone with the lowest value in band 5 was considered to contain the most information regarding water-column constituents [57]. If two pixels had the same band 5 values, the pixel closest to the coordinates of water-quality sampling location was selected.

2. Multiple Regression Analysis

Multiple regression equations were derived to predict constituent concentrations (TSS, TN, and TP, i.e. the dependent variables) from the predictor variables, such as band reflectance. The procedure for selection of predictor variables is delineated below.

The spectral bands and associated band ratios were all chosen as candidates for independent variables. Band ratios were included as independent variables in the regression analysis [10] because they are less apt to be influenced by lighting conditions [56].

Radiance data from the thermal bands (band 6 of Landsat TM and ETM+, and band 10 of Landsat TIRS) were converted to water surface temperature. As discussed, water temperature has been found to be related to phytoplankton concentration [23, 24], and thus, related to water quality [58]. However, in this study, most of the satellite image dates differ by several days compared with the closest corresponding actual water-quality sampling date; thus, the water surface temperature derived from the satellite images does not represent the actual water temperature at the time of water sampling.

Equation 4 considers the net energy fluxes between a waterbody and the atmosphere [59]:

$$NET = SWR_{net} - (LWR_{net} + LHF + SHF) \quad (4)$$

where NET is the net energy flux, SWRnet indicates the net short-wave radiation energy flux (Equation 5), LWRnet indicates the net long-wave radiation flux (Equations 6 and 7), LHF is the latent heat flux (Equation 8), and SHF is the sensible heat flux (Equation 9). These terms are calculated by the following equations [59]:

$$SWR_{net} = (1 - a)SWR_{down} \quad (5)$$

$$LWR_{net} \approx \varepsilon\sigma {T_s}^4 \left(0.39 - 0.05 {e_a}^{\frac{1}{2}}\right)(1 - 0.51C^2) + 4\varepsilon\sigma {T_s}^3 (T_s - T_a) \quad (6)$$

$$\text{and } C \approx 1.61\left(1 - \frac{SWR_{down}}{SWR_{cs}}\right) + 0.0019n \quad (7)$$

$$LHF = \rho L_e C_e U(Q_s - Q_a) \quad (8)$$

$$SHF = \rho C_p C_h U(T_s - T_a) \quad (9)$$

where a is the surface albedo (usually very low for water so $SWR_{net} \approx SWR_{down}$), ε is the surface emissivity (≈ 0.97), σ is the Stefan-Bolzman constant, T_s is the water surface temperature, T_a is the air temperature, e_a is the surface vapor pressure, C is the cloud cover index (Equation 7), SWR_{cs} is the clear-sky short wave radiation, n is the noon solar altitude, ρ is the density of air, L_e is the latent heat of evaporation, C_e is the turbulent exchange coefficient for latent heat, U is the wind speed, Q_s and Q_a are saturation specific humidity at the surface and at near-surface atmosphere, respectively, and C_h is the turbulent exchange coefficient for sensible heat.

Some of the variables in Equations 5 to 9 are known or can be reasonably assumed as constants (such as a, ε, σ, ρ, L_e, C_e, and C_h [60]). The surface vapor pressure, e_a, is dependent on water surface temperature [61]. Q_s and Q_a are both dependent on temperature as well [62]. The air temperature and noon solar altitude (T_a, and n respectively) can be obtained from the historical observation record. The water surface temperature T_s is obtained from thermal band data. That leaves only one variable unknown, which is the clear-sky short wave radiation SWR_{cs}. Calculating SWR_{cs} involves a complex procedure [63] so it is difficult to associate it with distinct environmental factor(s); thus, we did not consider it in evaluating heat flux in this study.

Assuming that the temperature change between the image date and the water-sampling date directly corresponds with the cumulative heat flux between the dates, the following variables are needed in order to account for the temperature change between the image-acquisition date and the water-sampling date [52]:

1. Time offset (in days) between the image date and the water-quality sampling date (positive offset means that the image date is later than the sampling date);
2. Water surface temperature (in K) derived from the thermal band;
3. Air temperature (in K): both instantaneous temperature at the time of satellite image acquisition, and daily mean air temperature between the image date and the water-quality sampling date are considered;
4. Wind speed (in m/s): both instantaneous wind speed at the time of satellite image acquisition and the daily mean wind speed between the image date and the water-quality sampling date are considered; and
5. Noon solar altitude (in degrees): the mean noon solar altitude between the image date and the water-quality sampling date.

Instantaneous temperature and wind speed were interpolated from the hourly historical data [52]. And further considering Equations 5 to 9, the full list of variables considered in the multiple regression process is provided in Table 7. A look-up table between variable abbreviations and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2017 doi:10.20944/preprints201710.0108.v1

Peer-reviewed version available at PLoS ONE 2018; doi:10.1371/journal.pone.0201255

variable descriptions is provided as Table 8. As described above, in this study, the band number is based on band-numbering scheme for TM and ETM+.

Table 7. Reflectance bands (i.e., band (B1), band 2 (B2), etc.) and ratios used in the variable-selection process.

Water constituent	# of valid observations	Initial predictor variables before p-threshold test
TSS	28	B1, B2, B3, B4, B2/B1, B3/B1, B4/B1, B3/B2, B4/B2, B4/B3, D_{off}, T_s, T_a, T_{mean}, T_s-T_a, T_s-T_{mean}, W, W_{mean}, Alt, Alt^2
TN	38	
TP	25	

Table 8. Look-up table for variable abbreviation and description of variables.

Variable abbreviations	Variable description
B1, B2, B3, B4	Reflectance value for Band 1, Band 2, Band 3, and Band 4, respectively.
D_{off}	Date offset between the image date and the water-quality sampling date
T_s	Water surface temperature derived from the remote-sensor thermal band
T_a	Instantaneous temperature at time of satellite image acquisition
T_{mean}	Daily mean air temperature between the image date and the water quality sampling date
W	Instantaneous wind speed at the time of satellite image acquisition
W_{mean}	Daily mean wind speed between the image date and the water quality sampling date
Alt	Mean noon solar altitude between the image date and the water-quality sampling date

Selection of predictor variables is based on a hybrid forward selection that considers the variation inflation factor (VIF). In conventional forward selection, variables are added to the regression one at a time, starting with no predictor variables being selected. The p-value threshold includes a predictor in the regression equation if its p-value is below a "probability to enter," and includes a predictor that will most improve the fit first (i.e., "forward"). A default value of 0.25 in JMP [64] was used for "probability to enter."

In addition to p-value, the variation inflation factor (VIF) was used to minimize multicollinearity of the model. Multicollinearity occurs when a predictor variable is a linear combination of other predictor variables in the model. The direct consequence of multicollinearity is that the error variance is inflated, which may result in low prediction power if the overfitted model is used with a new set of data. VIF is calculated as:

$$VIF_j = 1/(1 - R_j^2) \quad (10)$$

where R_j^2 is the multiple coefficient of determination between the j-th predictor variable of interest and the rest of the predictor variables. The rule of thumb to avoid serious multicollinearity is that all chosen predictor variables should have VIF less than 10 [65]. Unlike other criteria such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Mallow's Cp, VIF is generated for each predictor variable. Also, VIF has a suggested absolute criterion, whereas other criteria (AIC, BIC, Cp, etc.) provide only relative comparison between models.

We propose a novel approach that considers VIF while adding variables in forward selection. When a variable is added according to the rules of forward selection, VIFs of all included variables (including the one that is just added) are also checked. If VIFs are all below the threshold of 10, the newly-added variable is allowed, and the next variable is chosen according to the rule of forward selection. However, if any VIF is found to be larger than the threshold for any of the variables, the most recently-added variable is deleted and the selection procedure stops. Coefficients of variables,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2017 doi:10.20944/preprints201710.0108.v1

Peer-reviewed version available at PLoS ONE 2018; doi:10.1371/journal.pone.0201255

p-values, and VIF are dynamically recalculated when any variable is deleted from the model. The procedure is illustrated in Figure 2.

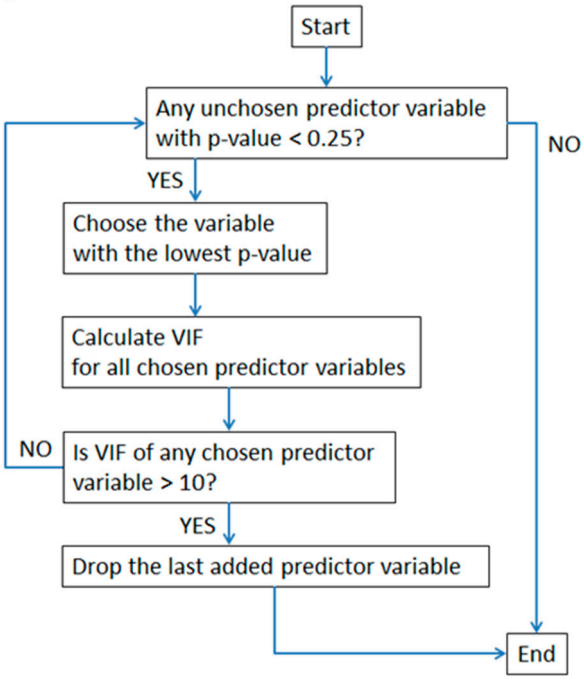


Figure 2. Flow chart of the hybrid forward-selection process for selecting predictor variables in multiple regression analysis.

The derived multiple linear equations were then validated by bootstrapping and Leave-One-Out Cross Validation (LOOCV). Bootstrapping and LOOCV are both resampling methodologies [66]. Bootstrapping assumes that samples (i.e., observations – sets of response and associated prediction variables in this case) represent the whole population, so a random resampling from the samples provides a prediction of what one expects to encounter (statistically) from unknown, future data. LOOCV leaves one sample out at a time and calibrates for the coefficients of predictor variables based on the rest of the observations. The left-out sample is used for validation. The package CARET in the software R [67] was used to perform validation by bootstrapping and LOOCV. For bootstrapping, 1000 trials were specified.

3. Results

The best-fitting regression equations chosen by the hybrid forward selection for each water-quality constituent (TSS, TN and TP) are provided in Table 9. The results in Table 9 include the predictor variables, importance of the predictor variable, associated regression coefficients and standard error, 95% confidence intervals for the regression coefficients, p-values, and VIF values for each of the response variables (TSS, TN, and TP). The importance values are calculated by dividing the change in R^2 when the variable of interest is dropped from the model by the overall R^2 when the variable of interest is included [68]. The sum of importance values of all variables does not equal to 1 since the importance is relative only. Note that the response variables are transformed to obtain a better regression fit, and the band-numbering convention is based on TM and ETM+ band numbers, as discussed previously.

Table 9. Best fitting multiple regression models for TSS, TN and TP using the hybrid forward selection considering VIF.

					Coefficient of predictor		Confidence Interval for coefficient			
Resp. Variable	**R2**	**Num. of Obs.**	**Pred. Variable**	**Imp. of Var.**	**Value**	**Std. Error**	**Lower 95%**	**Upper 95%**	**p**	**VIF**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2017 doi:10.20944/preprints201710.0108.v1

Peer-reviewed version available at PLoS ONE 2018; doi:10.1371/journal.pone.0201255

$\sqrt[2]{TSS}$	0.68	28	(intercept)	-	-0.67	0.50	-1.69	0.36	0.19	-
			B3/B1	0.93	1.67	0.24	1.16	2.17	<0.0001	1.21
			W	0.21	0.21	0.065	0.077	0.34	0.0034	1.08
			T_s-T_{mean}	0.04	0.038	0.027	-0.018	0.093	0.18	1.16
$\sqrt[2]{TN}$	0.62	38	(intercept)	-	4.36	0.91	2.50	6.21	<0.0001	-
			W_{mean}	0.39	-0.053	0.012	-0.078	-0.029	<0.0001	1.26
			T_s	0.32	-0.012	0.0031	-0.019	-0.0062	0.0003	1.31
			B1	0.18	4.50	1.50	1.44	7.55	0.0053	2.35
			B4/B1	0.11	-0.049	0.020	-0.090	-0.0089	0.018	2.73
			D_{off}	0.11	-0.013	0.0051	-0.023	-0.0021	0.020	1.18
			B2/B1	0.05	0.11	0.067	-0.030	0.24	0.12	4.50
$log(TP)$	0.79	25	(intercept)	-	26.40	9.57	6.30	46.50	0.013	-
			Alt^2	0.27	0.00049	0.00012	0.00024	0.00073	0.0005	7.13
			T_s	0.16	-0.11	0.034	-0.18	-0.039	0.0044	8.68
			D_{off}	0.06	0.050	0.024	0.00028	0.10	0.049	1.77
			B3/B2	0.06	1.16	0.57	-0.044	2.36	0.058	1.72
			W	0.04	-0.10	0.062	-0.23	0.026	0.11	3.06
			W_{mean}	0.04	-0.11	0.065	-0.24	0.032	0.12	1.84

The resulting multiple regression-based models are provided in Equations 11 to 13:

$$TSS = (-0.67 + 1.67 \cdot \frac{B3}{B1} + 0.21 \cdot W + 0.038 \cdot (T_s - T_{mean}))^2 \quad (11)$$

$$TN = (4.36 - 0.053 \cdot W_{mean} - 0.012 \cdot T_s + 4.5 \cdot B1 - 0.049 \cdot \frac{B4}{B1} - 0.013 \cdot D_{off} + 0.11 \cdot \frac{B2}{B1})^2 \quad (12)$$

$$TP = e^{(26.4+0.00049 \cdot Alt^2 - 0.11 \cdot T_s + 0.05 \cdot D_{off} + 1.16 \cdot \frac{B3}{B2} - 0.1 \cdot W - 0.11 \cdot W_{mean})} \quad (13)$$

Plots of the observed versus predicted concentrations of TSS, TN, and TP are plotted in Figures 3, 4, and 5, respectively. The 1:1 line is added to all three figures.

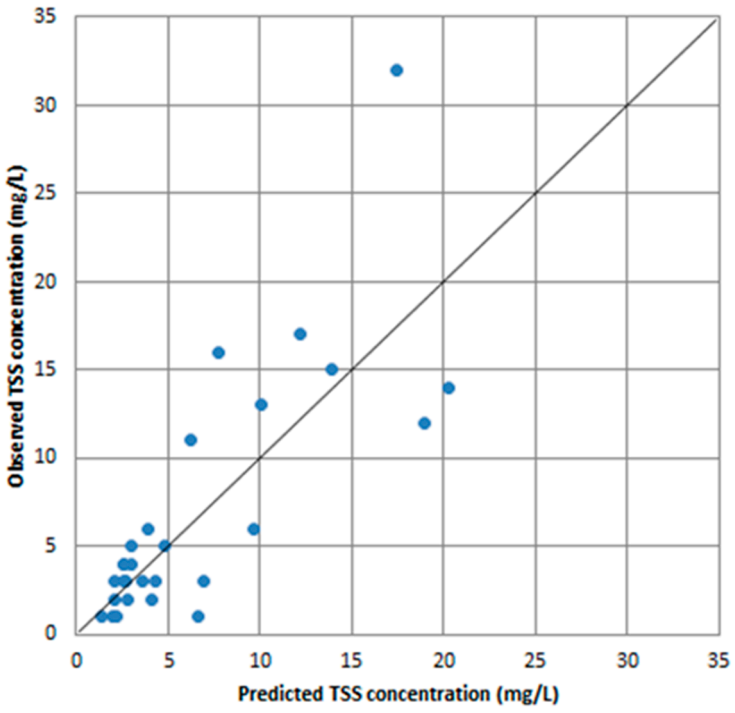


Figure 3. Observed versus predicted values for total suspended solids (TSS) (R2 = 0.68).

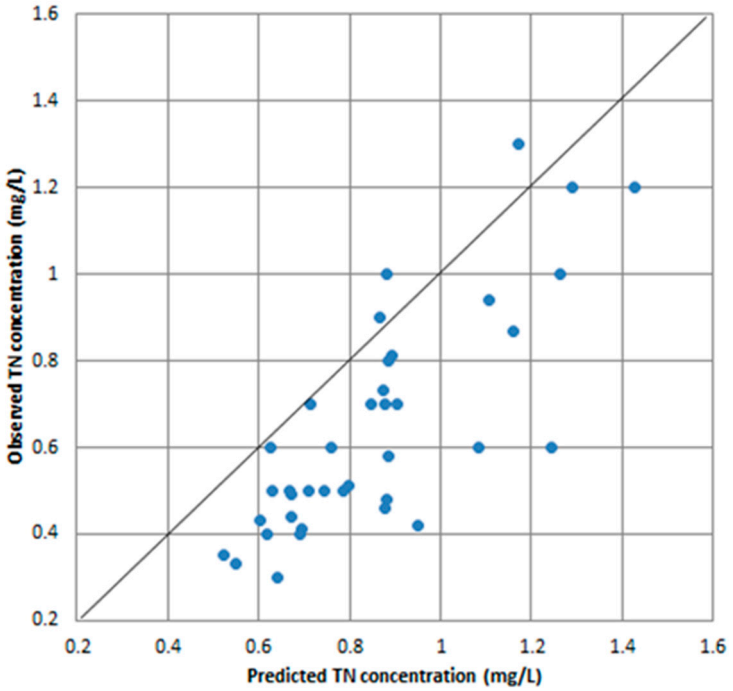


Figure 4. Observed versus predicted values for total nitrogen (TN) (R2 = 0.62).

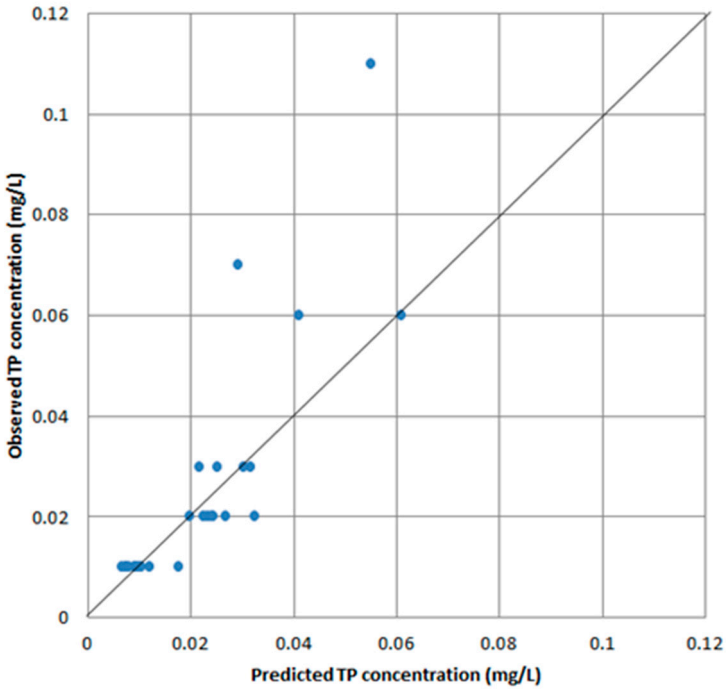


Figure 5. Observed versus predicted values for total phosphorus (TP) (R2 = 0.79).

The derived multiple linear predictive equations were validated by bootstrapping and LOOCV. The results are shown in Table 10. Table 10 shows that the equations have satisfactory predictive power for future, unknown data since validation R^2 values are above 0.5, except for weaker results regarding TN [69].

Table 10. Validation results of the predictive equations.

Equation	Calibration R^2	Validation R^2 from Bootstrapping	Validation R^2 from LOOCV
TSS	0.68	0.62	0.55
TN	0.62	0.46	0.47
TP	0.79	0.53	0.54

4. Discussion

The multiple linear equations derived from the regression analysis indicate that weather-related variables play an important role in predicting water-quality parameters. In fact, many weather variables bear more importance than the multispectral variables do. The relative importance of each variable is provided in Table 9. If all the weather variables are stripped from Table 9, the predictive variables related to Landsat bands alone provide only coefficients of determination, R^2, of 0.53, 0.26, and 0.36 for TSS, TN, and TP, respectively.

Kloiber et al. [10] found that both B1 and the ratio B3/B1 can be used to predict the Secchi disk transparency, which is closely related to TSS. From Kloiber et al. [10], the regression model containing B3/B1 and B1 predicted Secchi disk transparency with R^2 of 0.75. We also found B3/B1 as the dominant important variable in determining TSS concentrations, but did not find B1 as one of the significant prediction variables. Kloiber et al. [10] accrued a higher R^2 than our study since Kloiber et al. limited their *in situ* data collection to ±1 day from the corresponding satellite image acquisitions. In the current study, the predictive equation that includes B3/B1 alone has a R^2 of 0.53 for TSS because our available data only allows *in situ* samples to be ±7 days from satellite image acquisitions. Considering weather variables successfully boosted R^2 to 0.68, such that it was comparable with that of Kloiber et al. [10] (i.e., 0.75).

For TSS, we found the instantaneous wind speed, W, to be an important prediction variable. Since the instantaneous wind speed is chosen, instead of the daily mean wind speed between the image date and the water-quality sampling date (W_{mean}), it indicates that the instantaneous effect of wind (such as the surface ripple effect) is more important to TSS determination than the long-term heat-exchange effect. Even though the difference between the water surface temperature and the daily mean air temperature between the image date and the water-quality sampling date is selected as one of the prediction variables, it is of little importance in the model. It was chosen because the default forward-selection method has a lenient inclusion criterion ($p = 0.25$).

Dewidar and Khedr [9] determined that the band ratio B2/B1 is important in determining the TN concentration in brackish lagoons. However, the correlation between B2/B1 and TN was low in Dewindar and Khedr [9], with a correlation coefficient of 0.298. B2/B1 was also chosen by this study as one of the predictor variables, but B2/B1 still bears little predictive power as shown in Table 9. In contrast, the daily mean wind speed between the image date and the water-quality sampling date (W_{mean}) and water surface temperature (T_s) were determined to be the two most important predictor variables for TN prediction.

The high importance of water surface temperature T_s fortified the hypothesis that water temperature is related to the growth of microorganisms. The high importance of the daily mean wind speed between the image date and the water-quality sampling date (W_{mean}) and date difference (D_{off}) indicate that temperature change due to accumulated heat flux between the image date and sampling date is important. Referring to Equations 8 and 9, the mechanism involved should be the latent heat flux because latent heat flux (Equation 8) and sensible heat flux (Equation 9) are the only two components in the heat flux budget that involve wind speed. Latent heat flux is a main component of heat exchange between water and the atmosphere, and sensible heat plays a much lesser role [70].

As for TP, similar to TN prediction, the water surface temperature T_s still bears considerable importance. However, wind speed and D_{off} are not as important for estimating TP as it is for TN prediction, as the relative importance of variables in Table 9 indicates. It is intriguing that the square of noon solar altitude, Alt^2, has high importance in TP prediction. Referring to Equations 6 and 7, this implies that long wave radiation cooling correlates well with TP prediction. The weak

importance of wind speed and a strong importance of solar altitude for TP prediction jointly suggest that long wave radiative cooling constitutes the main process important for predicting TP concentration.

It is worth noting that, for TP prediction, optical multispectral data variables yield insignificant prediction power, and most of the prediction power is contributed by Alt^2 and T_s.

From the analyses above, it seems that determinations of TN and TP are influenced by quite different components of the net heat flux between air and water. For TN, latent heat flux seems to be the dominant factor, but long wave radiative cooling seems to be the dominant factor for TP. In other words, the difference in how the budget of cumulative heat flux between the image and sampling dates is constructed has different effects on water-quality constituents related to TN or TP. This can be reasoned by two effects:

1. TP concentration is highly correlated with *chlorophyll-a* [71], which is the photosynthetic pigment in algae or phytoplankton. However, optimal algae growth is at a depth within the water column [72], which means the transient change in heat flux due to wind speed at the surface will correlate poorly with algae growth.

2. Solar altitude affects not only long wave radiation, but also the penetration depth of light into the water column [73], which affects growth of phytoplankton, or algae, in water. Since the offset in correlation between *chlorophyll-a* concentration and TP concentration has seasonal variations [74], including solar altitude Alt (or the closely correlated square of solar altitude Alt^2) in the predictive model can account for the seasonal offset.

The inclusion of Alt^2 in the model probably signifies the combined effect of both effects.

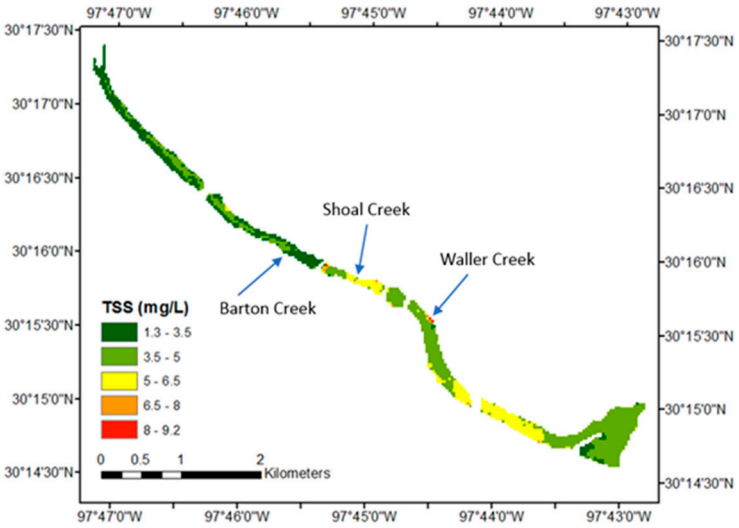
5. Field Application

To demonstrate the utility of water-quality monitoring by satellites via our proposed method, water-quality quantities for Lady Bird Lake on May 14, 2014 were estimated using Equations 11 to 13, respectively. This date was chosen because storms occurred the previous day and also the morning of the satellite flyover day before flyover time with cumulative rainfall depth of 27 mm, likely making it easier to discern the effect of urban stormwater runoff to the lake. Figures 6 to 8 give the respective predicted spatial distribution of TSS, TN, and TP concentrations.

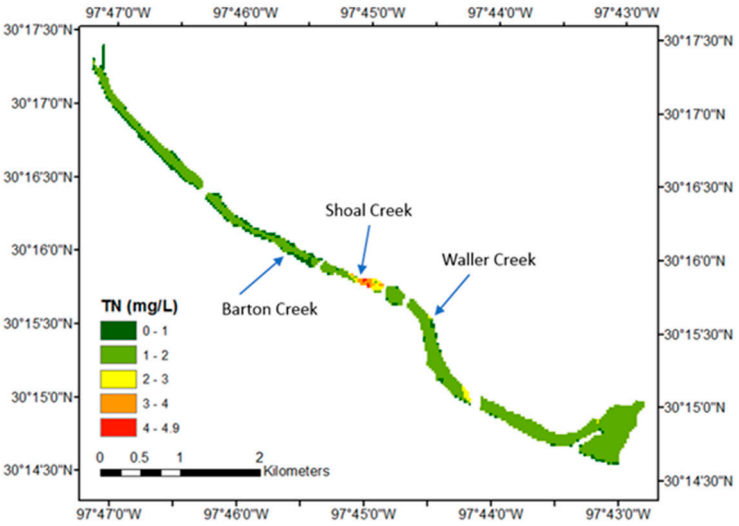
The water quality in the northwestern part of the lake is generally better than that in the southeastern extent, which is expected as a result of urban runoff. Lady Bird Lake has three major tributaries in the metropolitan Austin area: Barton Creek, Shoal Creek, and Waller Creek. The confluence points of the three streams are indicated in Figures 6-8. Barton Creek entails an extensive green belt around its riparian zone, and strict development regulations are in force because it is located within the Edwards Aquifer recharge zone [75]. As a result, there is no marked change in TSS, TN, and TP at the confluence point of Barton Creek, relative to proximal areas of the lake. To the contrary, the confluence points of Shoal Creek and Waller Creek show significant increase in TSS, TN, and TP. This illustrates the effects of amount of conservation efforts spent on each watershed. The influence of Shoal Creek is more visible in Figures 6-8 than that of Waller Creek because Shoal Creek has a larger drainage area [76]. Such details in spatial distribution can only be achieved via satellite-derived water-quality predictions and can serve as the precursor examination for more detailed water-quality examinations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2017 doi:10.20944/preprints201710.0108.v1

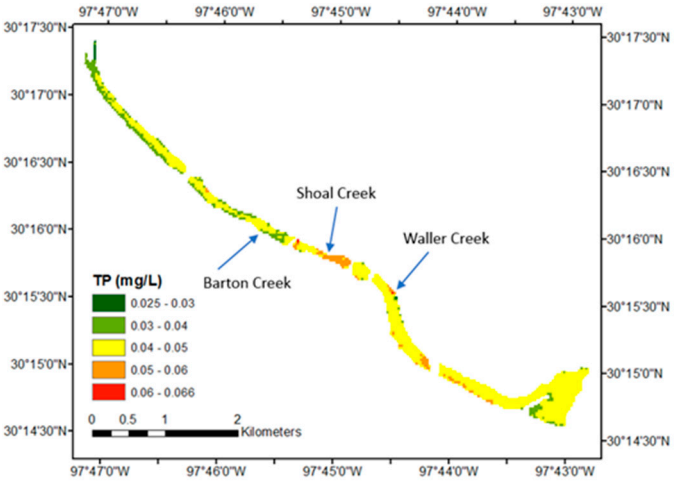
Peer-reviewed version available at PLoS ONE 2018; doi:10.1371/journal.pone.0201255



554
555 **Figure 6.** TSS concentrations for Lady Bird Lake, Austin, Texas, USA, May 14, 2014.
556



557
558 **Figure 7.** TN concentrations for Lady Bird Lake, Austin, Texas, USA, May 14, 2014.
559



560
561 **Figure 8.** TP concentrations for Lady Bird Lake, Austin, Texas, USA, May 14, 2014.

562 **6. Conclusions**

563 Multiple regression-derived equations using reflectance bands, band ratios, and environmental
564 factors as predictor variables for concentrations of TSS, TN, and TP, respectively, were derived using

a hybrid forward-selection method that considers VIF in the forward-selection process. Landsat TM, ETM+, and OLI/TIRS (Landsat 8) images were all used to derive the single set of equations. The coefficients of determination of the best-fitting resultant equations varied from 0.62 to 0.79. The predictive equations were also validated by bootstrapping and LOOCV with coefficients of determination in the range of 0.46 to 0.62.

Among all chosen predictor variables, B3/B1 has the strongest influence on the predictive power for TSS retrieval. The band ratio of B3/B1 was also selected by Kloiber et al. [10] in predicting Secchi disc transparency, indicating a correlation between Secchi disc transparency and TSS. Other reflectance bands and band ratios, such as B1, B2/B1, B4/B1, and B3/B2 are also influential in estimating TN and TP concentrations, but they are not dominant factors.

Environmental factors, such as wind speed and water surface temperature, were crucial in determination of water-quality parameters in this study. Inclusion of environmental factors allows usage of a single set of predictive equations across the seasons, as such predictive equations are innately adapted to the environmental changes for different seasons. The predictive equation will also likely to be more accurate because the pooling of all observation data.

The instantaneous wind speed, W, bears considerable importance in TSS determination, which is explained by wind-generated surface ripple effects. Water surface temperature T_s (derived from satellite remote-sensor thermal band image data) is important in determination of both TN and TP concentrations, as the growth of microorganisms in water is correlated with water nutrient concentrations.

However, the time offset between the satellite image-acquisition date and water-sampling date must be accounted for in water nutrient parameter (i.e., TN and TP) retrieval. The heat flux budget between air and the water surface was considered, and components in the budget equations were included in the forward-selection procedure. In additional to the predictor variables identified above, the daily mean wind speed between the image-acquisition date and water-sampling date (W_{mean}) and square of noon solar altitude (Alt^2) were identified as the most important predictor variables for TN and TP determinations, respectively. The time difference (in days) between the image-acquisition date and water-sampling date (D_{off}) was also chosen for TN and TP determination.

According to the heat flux budget equations, the inclusion of W_{mean}, T_s, and D_{off} indicates the dominance of latent heat flux in the determination of TN. On the other hand, the inclusion of Alt^2, D_{off}, and T_s in the TP model is an expression of the higher weight of long wave radiation cooling in TP estimation. Since *chlorophyll-a* concentration is highly correlated with TP concentration, we hypothesized that latent heat cooling is less important in TP determination because phytoplankton has the highest growth rate at a certain depth in water, which is less correlated with transient heat flux from evaporation at the surface.

The results showed that:

1. Environmental factors can constitute important ancillary variables in water-quality parameter estimation based on satellite remote-sensor images;
2. By including environmental factors, it is feasible to pool all observation data to create a single set of predictive equations, and use it to estimate water quality for all seasons;
3. A single set of predictive equations can be determined to retrieve year-round water-quality parameters (i.e., TSS, TN, and TP) with satisfactory accuracy from Landsat TM, ETM+, and OLI/TIRS imagery on the same lacustrine water body;
4. The derived predictive equations are robust enough to withstand the drastic change in the environment over 30+ years (1983 to 2015) while population in the metropolitan area almost tripled (from 373,000 in 1983 to 900,000 in 2015) over the same period of time [42]; and
5. Including VIF as part of the forward-selection process comprises a reliable methodology for choosing predictor variables.

In the future, the hybrid forward-selection method can be further refined to entail a stricter criterion for the inclusion of predictor variables. The default p=0.25 incurred inclusion of a few predictor variables that were not significant in the final selection of variables.

In addition, inclusion of ancillary environmental factors involving long-term averaging, such as average wind speed (W_{mean}), into the regression models demonstrated that it is possible to satisfactorily estimate water-quality parameters, even when a large temporal offset between satellite image-acquisition and *in situ* water sampling exists. Currently, the recommended longest temporal window between remote-sensor image-acquisition and water-sampling date is approximately seven days [17]. Since these environmental factors are part of the heat flux equations, including environmental factors in predictive equations means an active compensation in estimation error due to the temporal offset in collecting image and water-sample data. This hypothesis needs further testing as part of future research efforts.

References

[1] Kannel, P.R.; Lee, S.; Kanel, S.R.; Khan, S.P.; Lee, Y.-S. Spatial-temporal variation and comparative assessment of water qualities of urban river system: a case study of the river Bagmati (Nepal). *Environ. Monit. Assess*. **2007**, *129*, 433-459.

[2] Harmel, R.D.; King, K.W.; Haggard, B.E.; Wren, D.G.; Sheridan, J.M. Practical guidance for diacharge and water quality data collection on small watersheds. *Tran. ASABE* **2006**, *49(4)*, 937-948.

[3] USGS Water Data for the Nation. Available online: http://waterdata.usgs.gov/nwis (accessed on 9 October 2017).

[4] McCullough, I.M.; Loftin, C.S.; Sader, S.A. High-frequency remote monitoring of large lakes with MODIS 500 m imagery. *Remote Sens. Environ*. **2012**, *124*, 234-241.

[5] Liu, Y.; Islam, M.A.; Gao, J. Quantification of shallow water quality parameters by means of remote sensing. *Prog. Phys. Geog*. **2003**, *27(1)*, 24-43.

[6] Bukata, R.P. *Satellite Monitoring of Inland and Coastal Water Quality: Retrospection, Introspection, Future Directions*; Taylor and Francis Group: Boca Raton, FL, USA, 2005.

[7] Doxaran, D.; Froidefond, J.-M., Lavender, S.; Castaing, P. Spectral signature of highly turbid waters application with SPOT data to quantify suspended particulate matter concentrations. *Remote Sens. Environ*. **2002**, *81*, 149-161.

[8] Wu, G.; Leeuw, J.D.; Liu, Y. Understanding seasonal water clarity dynamics of Lake Dahuchi from in situ and remote sensing data. *Water Reso. Manage*. **2009**, *23*, 1849-1861.

[9] Dewidar, K.H.; Khedr, A. 2001. Water quality assessment with simultaneous Landsat-5 TM at Manzala Lagoon, Egypt. *Hydrobiologia* **2001**, *457*, 49-58.

[10] Kloiber, S.M.; Brezonik, P.L.; Olmanson, L.G.; Bauer, M.E. A procedure for regional lake water clarity assessment using Landsat multispectral data. *Remote Sens. Environ.* **2002**, *82*, 38-47.

[11] Kishino, M.; Tanaka, A.; Ishizaka, J. Retrieval of Chlorophyll a, suspended solids, and colored dissolved organic matter in Tokyo Bay using ASTER data. *Remote Sens. Environ*. **2005**, *99*, 66-74.

[12] Filippi, A.M. Derivative-neural spectroscopy for hyperspectral bathymetric inversion. *Prof. Geogr*. **2007**, *59(2)*, 236-255.

[13] Schaeffer, B.A.; Schaeffer, K.G.; Keith, D.; Lunetta, R.S.; Conmy, R.; Gould, R.W. Barriers to adopting satellite remote sensing for water quality management. *Int. J. Remote Sens*. **2013**, *34(21)*, 7534-7544.

[14] Free Landsat 7 Data Available from USGS. Available online: http://landsat.gsfc.nasa.gov/free-landsat-7-data-available-from-usgs/ (accessed on 29 September 2017).

[15] About MODIS: Moderate Resolution Imaging Spectroradiometer. Available online: https://modis.gsfc.nasa.gov/about/ (accessed on 29 September 2017).

[16] SeaWiFS Project. Available online: https://oceancolor.gsfc.nasa.gov/SeaWiFS/ (accessed on 29 September 2017).

[17] United States Geological Survey. *Predicting Water Quality by Relating Secchi-Disk Transparency and Chlorophyll A Measurements to Satellite Imagery for Michigan Inland Lakes, August 2002*; U.S. Geological Survey: Reston, VA, USA, 2004.

[18] Barrett, D.C.; Frazier, A.E. 2016. Automated method for monitoring water quality using Landsat imagery. *Water*. **2016**, *8*, 257.

[19] Kallio, K.; Attila, J.; Harma, P.; Koponen, S.; Pulliainen, J.; Hyytiainen, U.-M.; Pyhalahti, T. Landsat ETM+ images in the estimation of seasonal lake water quality in boreal river basins. *Environ. Manage*. **2008**, *42*, 511-522.

669 [20] Okin, G.S.; Gu, J. 2015. The impact of atmospheric conditions and instrument noise on atmospheric
670 correction and spectral mixture analysis of multispectral imagery. *Remote Sens. Environ*. **2015**, *164*, 130-141.
671 [21] Karakaya, N.; Evrendilek, F. 2011. Monitoring and validating spatio-temporal dynamics of
672 biogeochemical properties in Mersin Bay (Turkey) using Landsat ETM+. *Environ. Monit. Assess*. **2011**, *181*,
673 457-464.
674 [22] Lim, J.; Choi, M. Assessment of water quality based on Landsat 8 operational land imager associated
675 with human activities in Korea. *Environ. Monit. Assess*. **2015**, *187*, 384.
676 [23] Rasconi, S.; Gall, A.; Winter, K.; Kainz, M.J. Increasing water temperature triggers dominance of small
677 freshwater plankton. *PLOS ONE*. **2015**, *10(10)*, e0140449
678 [24] Voutilainen, A.; Jurvelius, J.; Lilja, J.; Viljanen, M.; Rahkola-Sorsa, M. Associating spatial patterns of
679 zooplankton abundance with water temperature, depth, planktivorous fish and chlorophyll. *Boreal Environ. Res*.
680 **2015**, *21*, 101-114.
681 [25] Bonansea, M.; Rodriguez, M.C.; Pinotti, L.; Ferrero, S. Using multi-temporal Landsat imagery and
682 linear mixed models for assessing water quality parameters in Río Tercero reservoir (Argentina). *Remote Sens.*
683 *Environ*. **2015**, *158*, 28-41.
684 [26] Alparslan, E.; Coskun, H.G.; Alganci, U. An investigation on water quality of Darlik Dam drinking
685 water using satellite images. *Thescientificworldj*. **2010**, *10*, 1293-1306.
686 [27] Wang, Y.; Xia, H.; Fu, J.; Sheng, G. Water quality change in reservoirs of Shenzhen, China: detection
687 using LANDSAT/TM data. *Sci. Total Environ*. **2004**, *328*, 195-206.
688 [28] Wu, G.; Leeuw, J.D.; Skidmore, A.K.; Prins, H.H.T.; Liu, Y. Comparison of MODIS and Landsat TM5
689 images for mapping tempo-spatial dynamics of Secchi disk depths in Poyang Lake National Nature Reserve,
690 China. *Int. J. Remote Sens*. **2008**, *29(8)*, 2183-2198.
691 [29] Zhou, Z.; Wang, X. Quantitative Remote Sensing Researches on Water Quality of the Weihe River
692 Based on SPOT-5 Imagery. Proceedings of Information Technology and Environmental System Sciences. 2008,
693 538-542.
694 [30] Lovric, M. *International Encyclopedia of Statistical Science*. Springer-Verlag: Berlin Heidelberg, Germany,
695 2011.
696 [31] Gilmour, S.G. The interpretation of Mallow's Cp-statistic. *J. R. Stat. Soc*. **1996**, *45(1)*, 49-56.
697 [32] Hair Jr., J.F.; Black, W.C.; Babin, B.J.; Anderson, R.E. *Multivariate Data Analysis A Global Perspective*, 7th
698 Ed. Pearson: Upper Saddle River, NJ, USA, 1998
699 [33] Graham, M.H. Confronting multicollinearity in ecological multiple regression. *Ecology*. **2003**, *84(11)*,
700 2809-2815.
701 [34] Yang, J.; Weisberg, P.J.; Bristow, N.A. 2012. Landsat remote sensing approaches for monitoring
702 long-term tree cover dynamics in semi-arid woodlands: Comparison of vegetation indices and spectral mixture
703 analysis. *Remote Sens. Environ*. **2012**, *119*, 62-71.
704 [35] Alsharif, A.A.A.; Pradhan, B. Urban sprawl analysis of Tripoli Metropolitan City (Libya) using remote
705 sensing data and multivariate logistic regression model. *J. Indian Soc. Remote Sens*. **2014**, *42(1)*, 149-163.
706 [36] Dubovyk, O.; Menz, G.; Conrad, C.; Kan, E.; Machwitz, M.; Khamzina, A. Spatio-temporal analyses of
707 cropland degradation in the irrigated lowlands of Uzbekistan using remote-sensing and logistic regression
708 modeling. *Environ Monit Assess*. **2012**, *185*, 4775-4790.
709 [37] Sun, D.; Hu, C.; Qiu, Z.; Shi, K. Estimating phycocyanin pigment concentration in productive inland
710 waters using Landsat measurements: A case study in Lake Dianchi. *Opt Express*. **2015**, *23(3)*, 3055-3074.
711 [38] Zheng, Z.; Li, Y.; Guo, Y.; Xu, Y.; Liu, G.; Du, C. Landsat-based long-term monitoring of total
712 suspended matter concentration pattern change in the wet season for Dongting Lake, China. *Remote Sens*. **2015**,
713 *7*, 13975-13999.
714 [39] Lymburner, L.; Botha, E.; Hestir, E.; Anstee, J.; Sagar, S.; Dekker, A.; Malthus, T. Landsat 8: Providing
715 continuity and increased precision for measuring multi-decadal time series of total suspended matter. *Remote*
716 *Sens. Environ*. **2016**, *185*, 108-118.
717 [40] Mishra, N.; Haque, M.O.; Leigh, L.; Aaron, D.; Helder, D.; Markham, B. Radiometric cross calibration
718 of Landsat 8 Operational Land Imager (OLI) and Landsat 7 Enhanced Thematic Mapper Plus (ETM+). *Remote*
719 *Sens*. **2014**, *6*, 12619-12638.
720 [41] Landsat Missions. Available online: http://landsat.usgs.gov/index.php (accessed on 9 October 2017).
721 [42] Demographic Data. Available online: http://www.austintexas.gov/page/demographic-data (accessed
722 on 2 October 2017).

[43] What are some general characteristics of each lake? Available online: https://www.austintexas.gov/faq/what-are-some-general-characteristics-each-lake (accessed on 2 October 2017).

[44] Texas Water Development Board. *Volumetric Survey of Lady Bird Lake*; Texas Water Development Board, Austin, TX, USA, 2009.

[45] USGS Global Visualization Viewer. Available online: http://glovis.usgs.gov/ (accessed on 9 October 2017).

[46] 2014 Land Use Inventory. Available online: ftp://ftp.ci.austin.tx.us/GIS-Data/planning/maps/Landuse%20Inventory%202014.pdf (accessed on 9 October 2017).

[47] Urban Hydrology for Small Watersheds (TR-55). Available online: https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1044171.pdf (accessed on 9 October 2017).

[48] Perkins, T.; Adler-Golden, S.; Matthew, M.; Berk, A.; Anderson, G.; Gardner, J. Retrieval of atmospheric properties from hyper- and multi-spectral imagery with the FLAASH atmospheric correction algorithm. *P. Soc. Photo-Opt. Ins*. **2005**, *5979*, 1-11.

[49] Exelis, Inc. *Atmospheric Correction Module: QUAC and FLAASH User's Guide*, Ver. 4.7. Exelis Inc.: Boulder, CO, USA, 2009.

[50] Siegel, D.A.; Wang, M.; Maritorena, S.; Robinson, W. Atmospheric correction of satellite ocean color imagery: the black pixel assumption. *Appl. Optics* **2000**, *39(21)*, 3582-3591.

[51] Hadjimitsis, D.G.; Clayton, C.R.; Hope, V.S. An assessment of the effectiveness of atmospheric correction algorithms through the remote sensing of some reservoirs. *Int. J. Remote Sens*. **2004**, *25(18)*, 3651-3674.

[52] National Weather Service Automated Surface Observing System. Available online: http://www.nws.noaa.gov/asos/ (accessed on 30 November 2016).

[53] MODIS Total Precipitable Water. Available online: https://modis.gsfc.nasa.gov/data/dataprod/mod05.php (accessed on 9 October 2017).

[54] Barsi, J.A.; Schott, J.R.; Palluconi, F.D., Hook, S.J. 2005. Validation of a Web-Based Atmospheric Correction Tool for Single Thermal Band Instruments, Proceedings of SPIE; 58820E1-58820E7.

[55] Barsi, J.A.; Schott, J.R.; Hook, S.J.; Raqueno, N.G.; Markham, B.L.; Radocinski, R.G. Landsat-8 Thermal Infrared Sensor (TIRS) vicarious radiometric calibration. *Remote Sens*. **2014**, *6*, 11607-11626.

[56] Jensen, J.R. *Remote Sensing of the Environment - an Earth Resource Perspective*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2007.

[57] Frazier, P.S.; Page, K.J. Water body detection and delineation with Landsat TM data. *Photogramm. Eng. Rem. S.* **2000**, *66(12)*, 1461-1467.

[58] Boyer, J.N.; Kelble, C.R.; Ortner, P.B.; Rudnick, D.T. Phytoplankton bloom status: Chlorophyll a biomass as an indicator of water quality condition in the southern estuaries of Florida, USA. *Ecol. Indic.* **2009**, *9(6)*, S56-S67.

[59] Kumar, B.P.; Vialard, J.; Lengaigne, M.; Murty, V.S.N.; McPhaden, M.J. TropFlux: air-sea fluxes for the global tropical oceans—description and evaluation. *Clim. Dynam*. **2012**, *38(7)*, 1521-1543.

[60] DeCosmo, J.; Katsaros, K.B.; Smith, S.D.; Anderson, R.J.; Oost, W.A.; Bumke, K.; Chadwick, H. Air-sea exchange of water vapor and sensible heat: the Humidity Exchange Over the Sea (HEXOS) results. *J. Geophys. Res*. **1996**, *101*, 12001-12016.

[61] Thomson, G.W. The Antoine equation for vapor-pressure data. *Chem. Rev*. **1946**, *38(1)*, 1-39.

[62] Stull, R.B. *Meteorology for Scientists and Engineers*, 3rd Ed.; Brooks Cole: Boston, MA, USA, 2011.

[63] Gubler, S.; Gruber, S.; Purves, R.S. Uncertainties of parameterized surface downward clear-sky shortwave and all-sky longwave radiation. *Atmos. Chem. Phys*. **2012**, *12*, 5077-5098.

[64] JMP Documentation. Available online: http://www.jmp.com/en_us/support/jmp-documentation.html (accessed on 9 October 2017).

[65] Chatterjee, S.; Simonoff, J.S. *Handbook of Regression Analysis*; Wiley Publication: Somerset, NJ, USA, 2013.

[66] Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; CRC Press: Boca Raton, FL, USA, 1993.

[67] Caret: Classification and Regression Training. Available online: https://cran.r-project.org/web/packages/caret/index.html (accessed on 9 October 2017).

[68] Judd, C.M.; McClelland, G.H.; Ryan, C.S. *Data Analysis: A Model Comparison Approach*; Routledge: New York, NY, USA, 2008.

[69] Moriasi, D.N.; Arnold, J.G.; Van Liew, M.W.; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *T. ASABE*. **2007**, *50(3)*, 885-900.

[70] Foltz, G.R.; McPhaden, M.J. Mixed layer heat balance on intraseasonal time scales in the Northwestern tropical Atlantic Ocean. *J. Clim*. **2005**, *18*, 4168-4184.

[71] Dillon, P.J.; Rigler, F.H. The phosphorus-chlorophyll relationship in lakes. *Limnol. Oceanogr.* **1974**, *19(5)*, 767-773.

[72] Sommer, U. Seasonal succession of phytoplankton in Lake Constance. *BioScience* **1985**, *35(6)*, 351-357.

[73] Bukata, R.P.; Jerome, J.H.; Kondratyev, K.Y.; Pozdnyakov, D.V. *Optical Properties and Remote Sensing of Coastal Waters*; CRC Press: Boca Raton, FL, USA, 1995; pp. 362.

[74] Marshall, C.T.; Peters, R.H. General patterns in the seasonal development of chlorophyll a for temperate lakes. *Limnol. Oceanogr.* **1989**, *34(5)*, 856-867.

[75] American Water Works Association. Conserving forest to protect water. *Opflow*. **2004**, *30(5), 1*, 4-7.

[76] City of Austin Watershed Regulation Areas. Available online: http://www.austintexas.gov/sites/default/files/files/Watershed/watershed_regs_map.pdf (accessed on 3 October 2017).