**An improved approach for using nutrients to predict chlorophyll *a* concentrations in lakes of the United States**

Jeffrey W. Hollister1,\*, W. Bryan Milstead1, Betty J. Kreakie1, and Richard B. Moore 2

1 Atlantic Ecology Division, National Health and Environmental Effects Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Narragansett, RI, U.S.A.

2 New Hampshire-Vermont Water Science Center, U.S. Geological Survey, Pembroke, New Hampshire.

\* E-mail: [hollister.jeff@epa.gov](mailto:hollister.jeff@epa.gov)

# Abstract

Chlorphyll *a* (Chl *a*)concentrations in lakes, ponds, and reservoirs are strongly related to the concentrations of nitrogen (N) and phosphorus (P) as well as the ratio (N/P) of these two nutrients. Many research studies have developed models to predict Chl *a* from these well-known relationships with these nutrients. Over time the thinking on whether lentic ecosystems were N or P limited has changed and the currently accepted thinking is that phytoplankton productivity is controlled by both N and P; however, the response to N, P, and N/P is not linear. Many efforts to model Chl *a* from N and P are linear or suffer from other limitations such as a small sample size, or limited geographic extent. To build on these past studies, we use data from the US EPA’s National Lakes Assessment and use principle components regression to better account for the response of Chl *a* to N, P and N/P. We found MORE STUFF HERE. Our models were able to predict Chl *a* MORE STUFF HERE. Models that use N, P and N/P are simple to implement, improve on past efforts by accounting for the non-linear nature of the Chl *a*  response to N, P, and N/P, and provide accurate estimates across many types of lakes, in many different geographic settings.

# Introduction

Chlorohpyll *a* (Chl *a)* is an oft used indicator of phytoplankton abundance and provides significant information about the trophic status of lentic ecosystems, the possible presence of harmful algae, the overall health of the system, and the delivery and availability of lake ecosystem services (Hambrook Berkman and Canova, 2007). This richness of information available from estimates of Chl *a* make it a primary focus of many limnological studies and the ability to accurately model Chl *a* has received significant attention in the literature (Crowley et al., 2012; Dillon and Rigler, 1974; Dimberg et al., 2013; Elser et al., 1990, 2007; Prairie et al., 1989; Schindler, 1977; Stow and Cha, 2013). Many factors determine Chl *a*  concentration in lakes and include the following: the concentration of nutrients, in particular, Total Nitrogen (TN) and Total Phosphorus (TP), carbon, light availability, planktivorus zooplankton and fish, and temperature (Kaiser et al., 1994; Maberly et al., 2002).

It was long believed that lentic systems were phosphorus limited and that phosphorus was the primary determinant of phytoplankton abundance; however in recent years that supposition has been challenged by several studies (Dzialowski et al., 2005; Maberly et al., 2002). Now it is believed that phytoplankton abundance is co-limited by both nitrogen and phosphorus (Elser et al., 1990, 2007); however, nutrients are not the only factors driving phytoplankton biomass and Chl *a*. Factors such light availability, presence and abundance of grazers, and carbon levels all play important roles. Yet, of these many determinants, nutrients have the most promise for management; thus, modeling efforts have focused on nutrients.

The many modeling studies focusing on nutrients and Chl *a* can be grouped into broad categories that represent the form of the empirical relationships and the geographic extent of the study (Table 1). While many of these studies have been successfully used to model Chl *a*, they do suffer from a number of limitations. First, using TN and TP together as dependent variables in a linear model violates the regression assumption that TN and TP are independent. In other words, TN and TP are highly collinear (Figure 1). In the case of ordinary least squares (OLS) regression, using collinear predictor variables, results in unstable coeffecients and often produces poorer predictions than regression based on non-collinear predictors (Næs and Mevik, 2001). Second, many of these studies show the relationships between nutrients and Chl *a* for a relatively small sample of lakes. Lastly, very few studies, especially in the United States, have been able to examine nutrient-Chl *a* relationships across broad geographic regions. In spite of these limitations, the previously mentioned studies have laid a strong foundation for future modeling work relating nutrients and Chl *a*.

The goal of this paper is to build on the past studies and address many of the limitations outlined above. We plan to use principal components regression to avoid the problems associated with multicollinearity and use a large dataset that is representative of lakes across broad regions. We expect to find that principal components regression better accounts for the Chl *a* variability from TN, TP and TN/TP than do simple OLS regressions. Furthermore, we provide access to the data and scripts used for the study.

# Methods

None of the many studies mentioned (Table 1) are able to account for all of the issues outlined above. This study does so by utilizing a dataset that is geographically broad and relatively large. Additionally, we develop a model which does not suffer from multicollinearity of the dependent variables. The data and scripts outlined below are freely available as supplements to this paper.

## Study Area and Data

Our study area for this research is the conterminous United States. The specific focus of our work is lakes, ponds and reservoirs of the conterminous United States that are represented in the NHD Plus, version 1.0 and are larger than XXX hectares (USEPA and USGS, 2005). The NHD Plus serves as the sample frame for the USEPA’s National Lakes Assessments (NLA) for 2007 and 2012 (Figure 2). The NLA is a probabilistic survey of lakes, ponds and reservoirs conducted every 5 years by US EPA’s National Aquatic Resource Surveys. These surveys collect physical, chemical, and habitat data. For this study, we use the TN, TP, and Chl *a* data from the 2012 NLA (n = 969) to train our models and TN, TP, and Chl *a* from the 2007 NLA (n = 1152) to test the accuracy of the model predictions. Ecoregions were also used to develop the 2007 NLA and these are modified from the XXX (NEED REF).

## Principal Components Regression

Most published models of the relationships between Chl *a* and TN and TP rely on OLS regressions of the raw and/or log transformed nutrient concentrations. While these regressions do fit the data quite well, they suffer from multi-collinearity of the predictor variables or don’t account for the role that TN:TP ratio plays. To account for this, we have chosen to use principal components regression (REFERENCES). Fully describing principal components regression is outside the scope of this study, but readers should consult (REFERENCES) for more details. Further, the exact analysis described here, may be repeated with scripts (Supplement 1) and data (Supplement 2).

In short, we construct a data matrix of ln(TN), ln(TP), and ln(TN/TP) and conduct a principal components analysis on that matrix. The resultant components, principal components 1, 2, and 3 (PC1, PC2, PC3), are orthogonal and no long collinear. We use these three components as predictor variables in a simple OLS regression of the form:

[1]

We use the 2012 data as a training dataset to calculate the principal components as well as parameterize the OLS regression.

*Accuracy Assessment*

To test if there was any improvement in predictions with our model versus traditional methods, we validate both our model and traditional models with NLA data from 2007. To estimate accuracy of each of the models, we used linear regression of the predicted 2007 Chl *a* vs the measured 2007 Chl *a* values to assess accuracy (Hollister et al., 2004). To compare the different models, we see which of the predicted vs. measured Chl *a* regressions had an *R2* closest to one, a slope (*β*1) closest to one, an intercept (*β*0) closest to zero and the lowest Root Mean Square Error (RMSE). Using this method we compare our Principal Components regression to the three most commonly used regressions to predict Chl *a* from TN and TP:

[2]

[3]

[4]

Each of the models are parameterized with the 2012 NLA data, and validated with the 2007 NLA data.

# Results and Discussion

The TN, TP, and Chl *a* conditions in the lakes sampled in both 2007 and 2012 NL NLA varied considerably across the US (Table 2). On average, the highest TN and TP values were seen in the Northern Plains (NPL) ecoregion for both years of sampling. The highest Chl *a* was in the Temperate Plains (TPL) in 2007 and in the Southern Plains (SPL) in 2012. The lowest mean TN and Chl *a* was in the Western Mountains (WMT) for both years. Lowest average Chl *a* is found in the Northern Apppalachians (NAP).

*Principal Components Regression*

Fig 3. PCA Plots

*Accuracy Assessment*

Fig 4. Predicted Chl *a* vs. Observed Chl *a* for PC Regression, TN, TP and TN+TP

*Regional Differences*

Fig 5. Scatterplot of TN, TP, and Chl *a* showing skew in Northern Plains (NPL) ecoregion

*Final Model*

Formulas: With NPL and Without NPL

# References

Crowley, K.F., McNeil, B.E., Lovett, G.M., Canham, C.D., Driscoll, C.T., Rustad, L.E., Denny, E., Hallett, R.A., Arthur, M.A., Boggs, J.L., et al. (2012). Do Nutrient Limitation Patterns Shift from Nitrogen Toward Phosphorus with Increasing Nitrogen Deposition Across the Northeastern United States? Ecosystems *15*, 940–957.

Dillon, P.J., and Rigler, F.H. (1974). The phosphorus-chlorophyll relationship in lakes. Limnol Ocean. *19*, 767–773.

Dimberg, P.H., Hytteborn, J.K., and Bryhn, A.C. (2013). Predicting median monthly chlorophyll-a concentrations. Limnol. - Ecol. Manag. Inland Waters *43*, 169–176.

Dzialowski, A.R., Wang, S.-H., Lim, N.-C., Spotts, W.W., and Huggins, D.G. (2005). Nutrient limitation of phytoplankton growth in central plains reservoirs, USA. J. Plankton Res. *27*, 587–595.

Elser, J.J., Marzolf, E.R., and Goldman, C.R. (1990). Phosphorus and Nitrogen Limitation of Phytoplankton Growth in the Freshwaters of North America: A Review and Critique of Experimental Enrichments. Can. J. Fish. Aquat. Sci. *47*, 1468–1477.

Elser, J.J., Bracken, M.E.S., Cleland, E.E., Gruner, D.S., Harpole, W.S., Hillebrand, H., Ngai, J.T., Seabloom, E.W., Shurin, J.B., and Smith, J.E. (2007). Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. Ecol. Lett. *10*, 1135–1142.

Hambrook Berkman, J., and Canova, M. (2007). Algal biomass indicators (ver. 1.0): U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chap. A7, section 7.4,.

Hollister, J.W., Gonzalez, M.L., Paul, J.F., August, P.V., and Copeland, J.L. (2004). Assessing the accuracy of National Land Cover Dataset area estimates at multiple spatial extents. Photogramm. Eng. Remote Sens. *70*, 405–414.

Kaiser, M.S., Speckman, P.L., and Jones, J.R. (1994). Statistical Models for Limiting Nutrient Relations in Inland Waters. J. Am. Stat. Assoc. *89*, 410–423.

Maberly, S.C., King, L., Dent, M.M., Jones, R.I., and Gibson, C.E. (2002). Nutrient limitation of phytoplankton and periphyton growth in upland lakes. Freshw. Biol. *47*, 2136–2152.

Næs, T., and Mevik, B.-H. (2001). Understanding the collinearity problem in regression and discriminant analysis. J. Chemom. *15*, 413–426.

Prairie, Y.T., Duarte, C.M., and Kalff, J. (1989). Unifying Nutrient–Chlorophyll Relationships in Lakes. Can. J. Fish. Aquat. Sci. *46*, 1176–1182.

Schindler, D.W. (1977). Evolution of phosphorus limitation in lakes. Science *195*, 260–262.

Stow, C.A., and Cha, Y. (2013). Are Chlorophyll a–Total Phosphorus Correlations Useful for Inference and Prediction? Environ. Sci. Technol. *47*, 3768–3773.

USEPA (2009). National Lakes Assessment: A Collaborative Survey of the Nation’s Lakes. (Washington, D.C.).

USEPA, and USGS (2005). National Hydrography Dataset Plus - NHD Plus.

Table 1. Selection of previously published relationships between total nitrogen, total phosphorus, and chlorophyll *a*

|  |  |  |  |
| --- | --- | --- | --- |
| **Citation** | **Empirical**  **Form(s)** | **Study**  **Area** | **Sample**  **Size** |
| Dillon and Rigler (1974) (Dillon and Rigler, 1974) | Log (Chl *a*) = Log (TP) | Southern Ontario, Canada | 19 |
| Prairie et al. (1989) (Prairie et al., 1989) | Log (Chl *a*) = Log (TP)  Log (Chl *a*) = Log (TN)  Log (Chl *a*) = Log (TP) + Log (TN) | US, Canada, Europe | 133 |
|  |  |  |  |
|  |  |  |  |

Table 2. Nutrient and Chl *a* summary statistics for 2007 and 2012 National Lakes Assessments by Ecoregion

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | **Total Nitrogen** | | **Total Phosphorus** | | **Chl  *a*** | |
| **2007** | **Ecoregion** | **N** | **Mean** | **Std. Dev** | **Mean** | **Std. Dev** | **Mean** | **Std. Dev** |
| Coastal Plains (CPL) | 119 | 1042.87 | 942.23 | 78.00 | 85.10 | 45.27 | 50.64 |
| Northern Appalachians (NAP) | 122 | 409.27 | 494.71 | 18.66 | 38.03 | 7.70 | 15.58 |
| Northern Plains (NPL) | 67 | 3712.45 | 4355.38 | 355.48 | 434.01 | 48.25 | 97.04 |
| Southern Appalachians (SAP) | 131 | 503.31 | 500.25 | 35.68 | 55.15 | 20.38 | 35.83 |
| Southern Plains (SPL) | 145 | 2012.65 | 3894.55 | 190.32 | 329.97 | 45.92 | 103.67 |
| Temperate Plains (TPL) | 157 | 1826.83 | 1802.70 | 179.54 | 262.75 | 56.55 | 80.57 |
| Upper Mid-West (UMW) | 163 | 847.68 | 1261.13 | 33.49 | 112.16 | 15.62 | 74.07 |
| Western Mountains (WMT) | 165 | 332.76 | 399.59 | 28.36 | 61.46 | 5.88 | 17.98 |
| Xeric (XER) | 88 | 961.52 | 1082.48 | 203.67 | 598.72 | 15.45 | 29.73 |
| **2012** | Coastal Plains (CPL) | 134 | 1220.17 | 1393.55 | 127.64 | 316.31 | 37.50 | 53.23 |
| Northern Appalachians (NAP) | 88 | 432.67 | 396.86 | 29.45 | 36.55 | 13.08 | 28.94 |
| Northern Plains (NPL) | 78 | 2935.68 | 6134.46 | 334.95 | 596.52 | 37.53 | 68.53 |
| Southern Appalachians (SAP) | 90 | 510.37 | 531.63 | 52.77 | 87.60 | 17.72 | 22.43 |
| Southern Plains (SPL) | 69 | 2897.26 | 6654.55 | 262.26 | 496.75 | 53.82 | 91.14 |
| Temperate Plains (TPL) | 140 | 1637.16 | 1193.41 | 177.63 | 224.84 | 46.35 | 62.44 |
| Upper Mid-West (UMW) | 120 | 851.17 | 713.05 | 41.78 | 55.15 | 13.39 | 21.13 |
| Western Mountains (WMT) | 168 | 396.71 | 504.60 | 54.61 | 93.21 | 7.51 | 21.45 |
| Xeric (XER) | 79 | 830.11 | 1054.13 | 132.16 | 181.00 | 26.23 | 91.96 |

Figure Captions

Figure 1. Scatterplot showing linear relationship between total nitrogen and total phosphorus for lakes sampled in 2007 and 2012 as part of the US Environmental Protection Agency’s National Lakes Assessment.

Figure 2. Map of United States, eco-regions and lakes sampled in 2007 and 2012 as part of the US Environmental Protection Agency’s National Lakes Assessment.

Figure 3. Plot of Principal Components 1 and Principal Components 2. Colors represent trophic status as defined by Chl *a* cutoffs from the National Lakes Assessment (USEPA, 2009).

Figure 4. Predicted Chl *a* vs Observed Chl *a* for formulas 1 – 4.

Figure 5. Scatterplots of TN and TP versus Chl *a*. Color