Online Supplement

Smoothed Hockey-stick Model

The piecewise linear model is computationally unstable because of the discontinuity in the model's first derivative Qian and Richardson (1997). The discontinuous first-order derivative

$$\frac{dy}{dx} = \begin{cases} \beta_0 & \text{when } \log(Chla) \leq \phi \\ \beta_1 & \text{when } \log(Chla) > \phi \end{cases}$$

can be approximated by a general logistic model:

$$\frac{dy}{dx} = \beta_0 + \frac{\beta_1 - \beta_0}{1 + e^{-\frac{\log(Chla)-\phi}{\lambda}}}$$

where λ is a "smoothness'' parameter determining the shape of the curve. Integrating the logistic function, we have a continuous version of the hockey stick model:

$$y = \beta_0 + \beta_1(x - \phi) + (\beta_1 - \beta_0)\lambda \log\left(1 + e^{\frac{x-\phi}{\lambda}}\right) + \epsilon$$

If the change point version of the hockey stick model is a "broken" stick model, the smoothed model is a "bent'' stick model (Chiu, et al. 2006). A numerical issue arises because of the introduction of the smoothness parameter λ. The bent stick model can be reduced to a linear model through two venues. One is through the estimated change point ϕ (at one of the two ends of the $\log(Chla)$ range and the other is through λ (when $\lambda \to \infty$). In other words, ϕ and λ cannot be empirically estimated simultaneously. To avoid the ambiguity, we set λ to 0.01 of the observed $\log(Chla)$ range.

Temperature Effect

Because MC is a biological product, its concentration is likely affected by temperature. The temperature effect can be included by adding a linear term to the hockey stick model:

$$\log(MC) = \beta_0 + \beta_1(\log(Chla) - \phi) + (\beta_1 - \beta_0)\lambda \log\left(1 + e^{\frac{\log(Chla)-\phi}{\lambda}}\right) + \beta_2(T - 20) + \epsilon$$

This term reflects the typical treatment of the temperature effect in a biological process, that is adding a multiple of $\theta^{(T-20)}$. In this case, our model in MC concentration scale is the exponential of the hockey-stick model times $e^{\beta_2(T-20)} = (e^{\beta_2})^{(T-20)}$. That is, the temperature constant is $\theta = e^{\beta_2}$.

Bayesian Sequential Updating

When examined separately, the model for each year in our BHM can be seen as a Bayesian nonlinear regression model with an informative prior. When making short-term forecasting, we further simplify the model to construct independent priors for each model parameter using the normal-inverse-gamma conjugate family of priors. Parameters of the prior can be estimated based on the hierarchical model fitted using MCMC through the method of moments. At each sampling event with a year, we fit the Bayesian nonlinear model using data from the most recent data (incremental or cumulative). We describe the general setup here.

For a nonlinear regression model of $y = f(\boldsymbol{X}, \boldsymbol{\beta}) + \varepsilon$, where $\boldsymbol{\beta}$ is the coefficient vector, we use independent priors for each coefficient β_j:

$$
\begin{array}{rcl}
\beta_j & \sim & N(\mu_j, \sigma_j^2) \\
\mu_j|\sigma_j^2 & \sim & N(\mu_{0j}, \sigma_j^2/\nu_{0j}) \\
\sigma_j^2 & \sim & IG(\alpha_0, \beta_0)
\end{array}
$$

The prior parameters $\mu_{0j}, \nu_{0j}, \alpha_0$, and β_0 can be estimated based on the posterior MCMC samples of the hyper-parameters estimated in the BHM using data from previous years. Once the model is fit using the current sampling data, the posterior distributions of μ_j and σ_j^2 (MCMC samples) can be used to summarize the joint distribution of μ_i, σ_j^2 using the normal-IG distribution (using the method of moments). This approach allows us to incorporate data from previous years to represent annual variation in the model coefficients, as well as the short-term (seasonal) changes in the same coefficients.

Computational Details

We used R and Stan as the main programming tool. The computer programs are summarized in 11 separate RMarkdown files, including:

- Data importing, data cleaning, and exploratory data analysis
- Fitting Bayesian hierarchical model
- Model Evaluation: fitting subsets of data
- Summarizing model results and estimating exceedance probabilities
- Bayesian hierarchical model without the last two years of data
- Sequential updating, and
- Verifying model's predictive accuracy -- predicting last two years MC sequentially based on priors developed without the last two years of data.

In addition, figures used in the paper were drawn using functions from R package `ggplot2`. Code for `ggplot2` plots are included in a `.org` file.

References

- Chiu, G.; Lockhart, R.; Routledge, R. (2006) Bent-cable regression theory and applications. *Journal of the American Statistical Association*, 101, 542-553.
- Qian S.S. and Richardson, C.J. (1997) Estimating the long-term phosphorus accretion rate in the Everglades: A Bayesian approach with risk assessment. *Water Resources Research*, 33(7), 1681-1688.