Appendix 2: Lake Washington Analysis

Here we show the GAM and DLM analysis for the Lake Washington dataset. For complete code, see the full quarto document

Data Processing

In Lake WA the major change ecosystem change is related to sewage in the lake – adding wastewater plants in the 1960s reduced phosphorous (Figure S1) and in turn, blue green algae (Figure S2). We can see this by plotting the raw data:

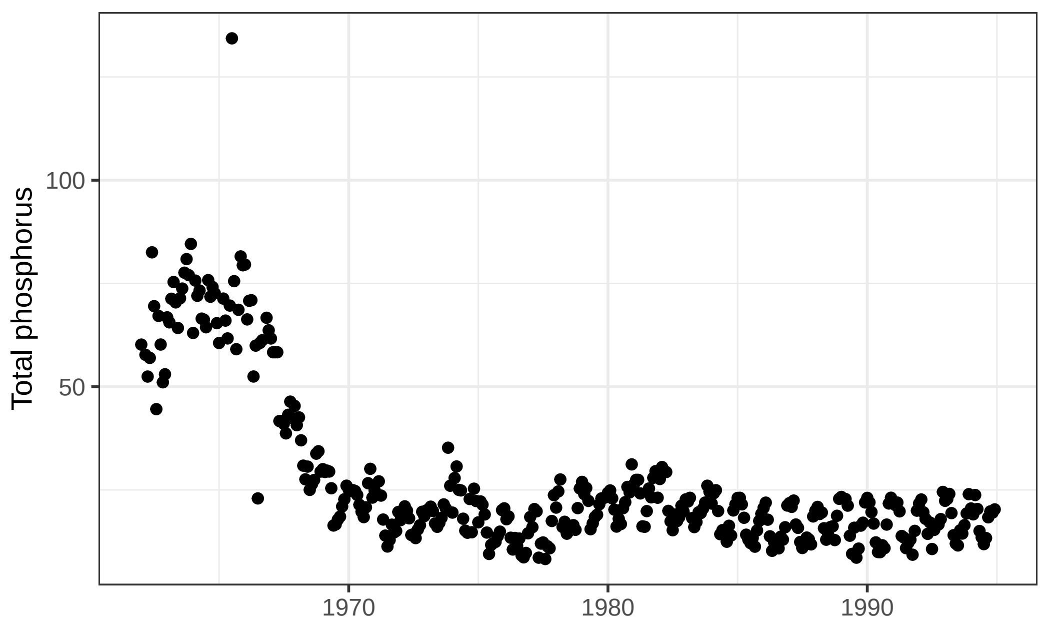

Figure S1: Time series of total phosphorus in Lake Washington

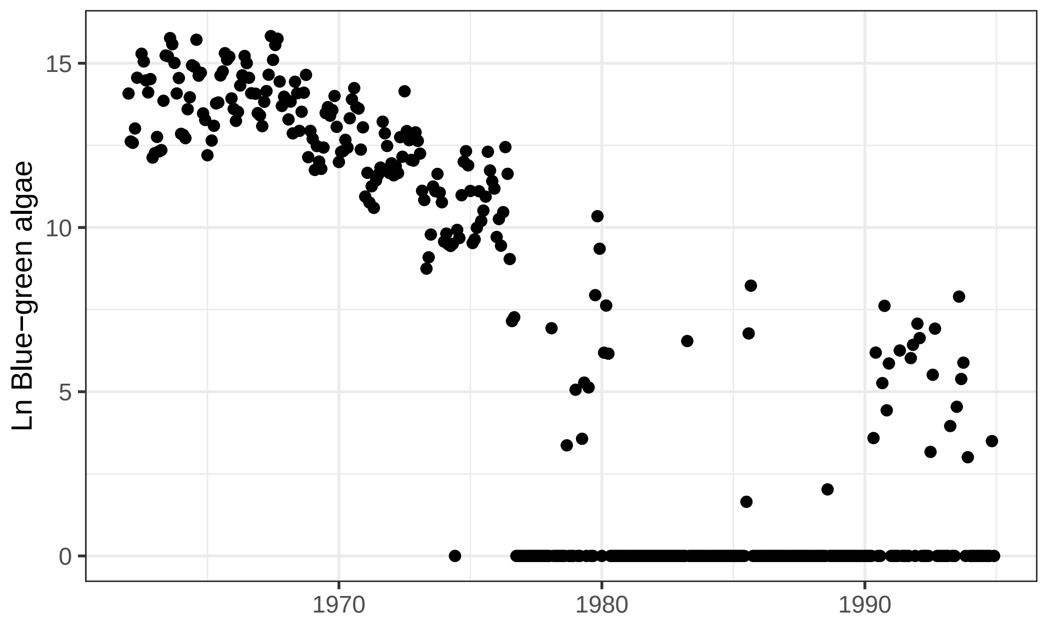

Figure S2: Time series of bluegreen algae (cyanobacteria) abundance in Lake Washington

We can then use the MARSS package to fit a DLM. We'll filter the data to only use counts before 1974, and log transform densities. We also z-score the covariate. This is a good example of imperfect data – there's a couple of missing observations in the covariate, which we can't have in a DLM or GAM – so let's just fill them in with a spline. We can do this as follows:

```
dat <- dplyr::filter(dat, Year < 1974) %>%
  dplyr::mutate(y = log(Bluegreens), zp = as.numeric(scale(TP)))

interpolated <- spline(dat$zp, xout = 1:nrow(dat))$y
dat$zp[which(is.na(dat$zp))] <- interpolated[which(is.na(dat$zp))]
```

The response data exhibit a strong seasonal cycle – and it's a good idea to de-season that to focus on the phosphorous effect. Here we de-season the data:

```
# Fill in a few missing values
dat$y_interp <- dat$y
interpolated <- spline(dat$y, xout = 1:nrow(dat))$y
missing <- which(is.na(dat$y_interp))
dat$y_interp[missing] <- interpolated[missing]
```

```r
dat <- dplyr::group_by(dat, Month) %>%
  dplyr::mutate(month_mean = mean(y,na.rm=T)) %>%
  dplyr::ungroup() %>%
  dplyr::mutate(y_adj = y_interp - month_mean)
```

Last, we should look at the lags between driver and response relationships. The CCF here indicates a 2-4 month lag, where TP and algae are more strongly associated with one another.

```r
ccf(dat$y_adj, dat$zp, main="CCF")
```

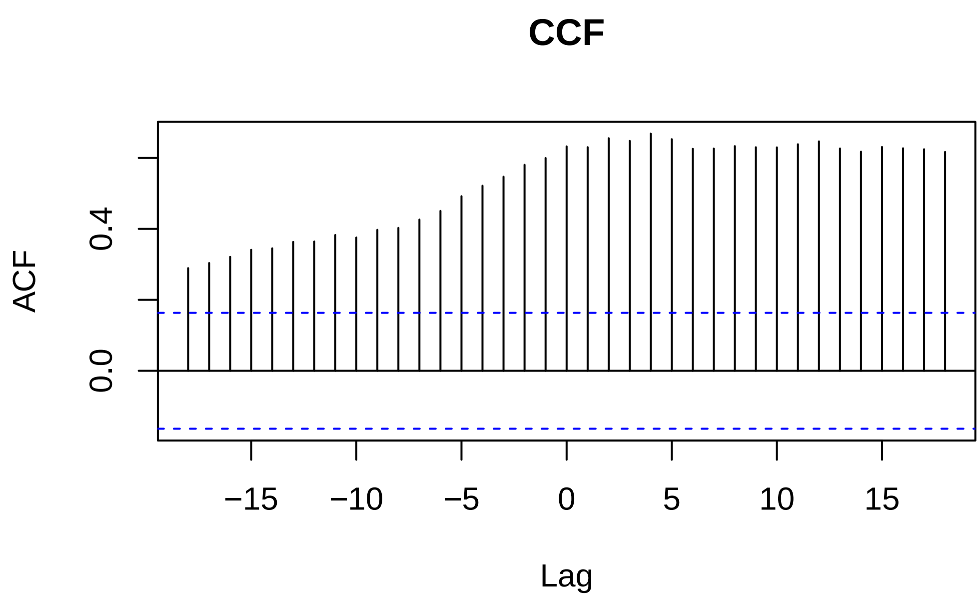

Figure S3: A CCF (Cross-Correlation Function) plot that represents the correlation between total phosphorus and blugreen algae abundance as a function of the lag (time difference) between them.

We'll use a lag of 4 months, which we can apply as follows:

```r
lag_n <- 5
dat$lagged_zp <- c(rep(NA, lag_n), dat$zp[1:(nrow(dat)-lag_n)])
dat <- dat[-c(1:lag_n),]
```

DLMs

First, we need to define the model – this block is basically copied from the MARSS book, the salmon survival case study

```
m <- 2
TT <- nrow(dat)
B <- diag(m) ## 2x2; Identity
U <- matrix(0, nrow = m, ncol = 1) ## 2x1; both elements = 0
Q <- matrix(list(0), m, m) ## 2x2; all 0 for now
diag(Q)[1] <- 0.0000001
diag(Q)[2] <- c("q.beta")
#diag(Q) <- c("q.alpha", "q.beta") ## 2x2; diag = (q1,q2)
Z <- array(NA, c(1, m, TT)) ## NxMxT; empty for now
Z[1, 1, ] <- rep(1, TT) ## Nx1; 1's for intercept
Z[1, 2, ] <- dat$lagged_zp ## Nx1; predictor variable
A <- matrix(0) ## 1x1; scalar = 0
R <- matrix("r") ## 1x1; scalar = r
## only need starting values for regr parameters
inits_list <- list(x0 = matrix(c(0, 0), nrow = m))
## list of model matrices & vectors
mod_list <- list(B = B, U = U, Q = Q, Z = Z, A = A, R = R)
# convert response to matrix
dat_mat <- matrix(dat$y_adj, nrow = 1)
# fit the model -- crank up the maxit to ensure convergence
dlm_1 <- MARSS(dat_mat, inits = inits_list, model = mod_list,
control = list(maxit=4000), method="TMB")
```

```
MARSS fit is
Estimation method: TMB
Estimation converged in 18 iterations.
Log-likelihood: -155.082
AIC: 318.1641   AICc: 318.4626

         Estimate
R.r         0.226
Q.q.beta    0.235
x0.X1       0.498
x0.X2       1.422
Initial states (x0) defined at t=0
```

```
Standard errors have not been calculated.
Use MARSSparamCIs to compute CIs and bias estimates.
```

Let's put the state estimates back into the original dataframe and do some plotting. The slope is generally positive, which is what we expect for the hyposthesized relationship between bluegreen algae and total phosphorus – and exhibits clear variation through time

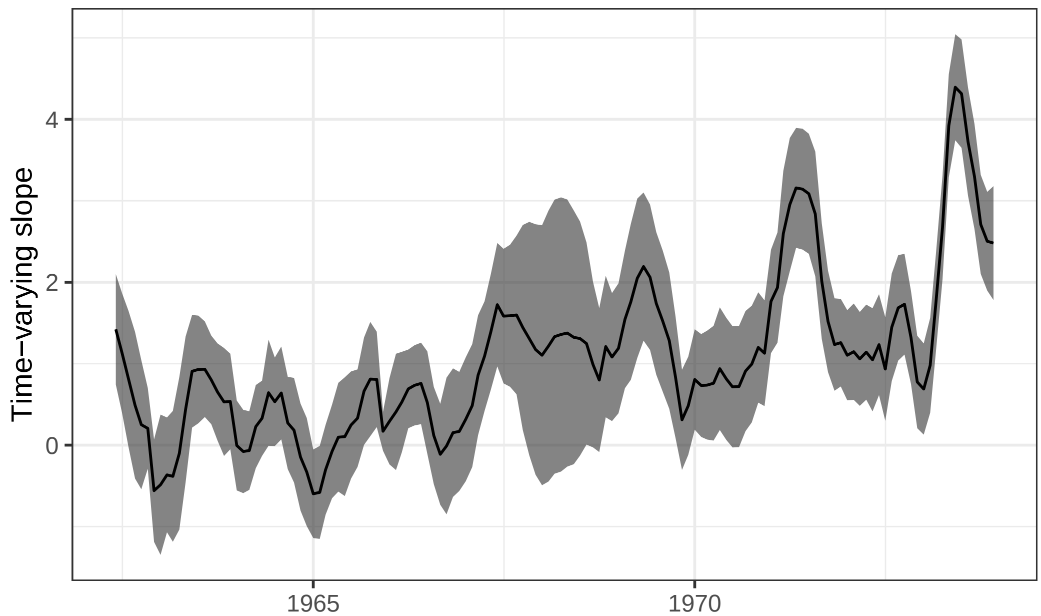

Figure S4: Time-varying slope of the relationship between bluegreen algae (response) and total phosphorus (driver) from a DLM

We can also plot the fitted values from our model.

GAMs

We can fit a GAM model 3 ways, 1) so that the smoother is a function of phosphorus which represents a nonlinear relationship between total phosphorus and bluegreen algae that is not temporally structured thus the relationship is stationary and fixed through time. This is how GAMs are typically structured for ecological analyses. 2) A temporally structured model so that the smooth function of phosphorus varies by time, and 3) a smooth on date as a driver of cyanobacteria where the smooth through time varies by phosphorus level. We show the model predicted cyanobacteria trend through time using these three modeling approaches. While they are quite similar, the predictions diverge sustantially at the end of the time series.

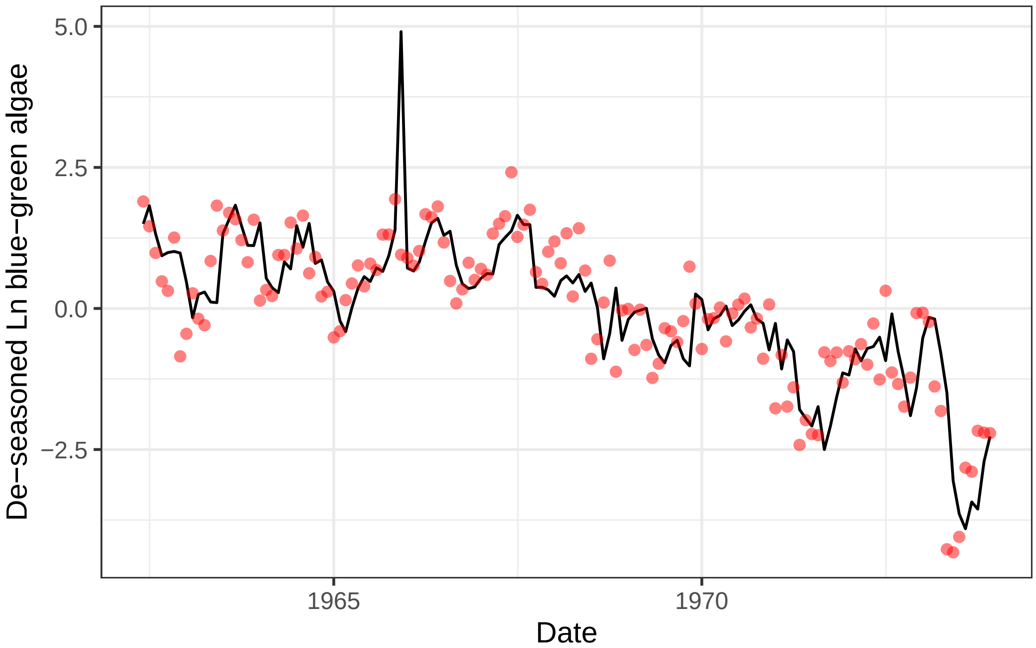

Figure S5: Fit of the DLM model with time-varying slope (black line) to observed bluegreen algae abundance (red points).

```
dat$n_date <- as.numeric(as.factor(dat$date))

gam_cr1 <- gam(y_adj ~ s(lagged_zp, bs = "cr"), data = dat) #1
gam_cr2 <- gam(y_adj ~ s(lagged_zp, by = n_date, bs = "cr"), data = dat) #2
gam_cr3 <- gam(y_adj ~ s(n_date, by = lagged_zp, bs = "cr"), data = dat) #3
```

There are a few more analysis approaches for running GAMs including the choice of smoother function (Figure S6 & Figure S7) and the number of knots, or the "wiggliness" (Figure S8) of the smoothed relationship. We can plot these different results and include the results of the DLM for comparison, note that 'cr' and 'tp' are so similar they are difficult to discern from the plot. We can do a similar comparison for the choice of k. Here we compare a four different numbers of knots (Figure S8).

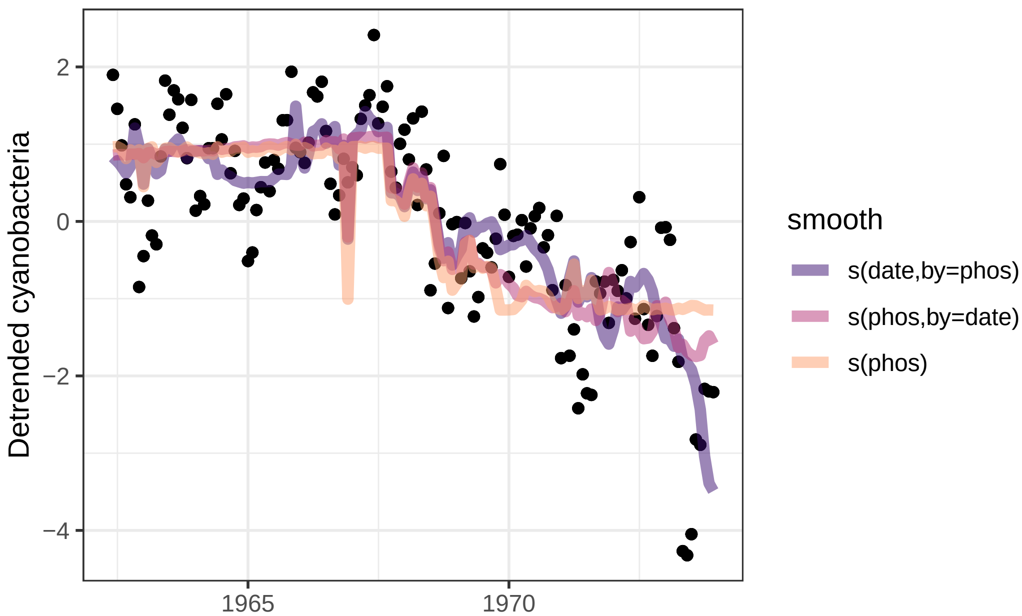

Figure S6: Fit of the GAM models (lines) to observed bluegreen algae (cyanobacteria) abundance (black points). Colors denote different specifications of the smoother functions.

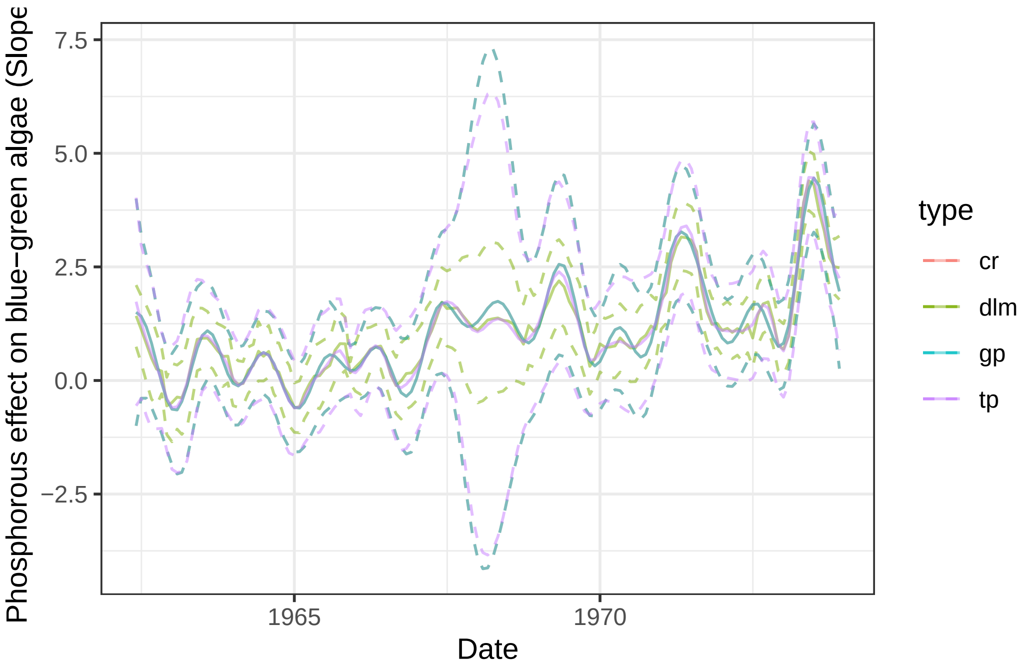

Figure S7: Effect of the choice of smoother function on estimates of time-varying slope parameter where color denotes model type. 3 GAM smoother types (cr, tp, gp) are contrasted with a DLM (dlm).

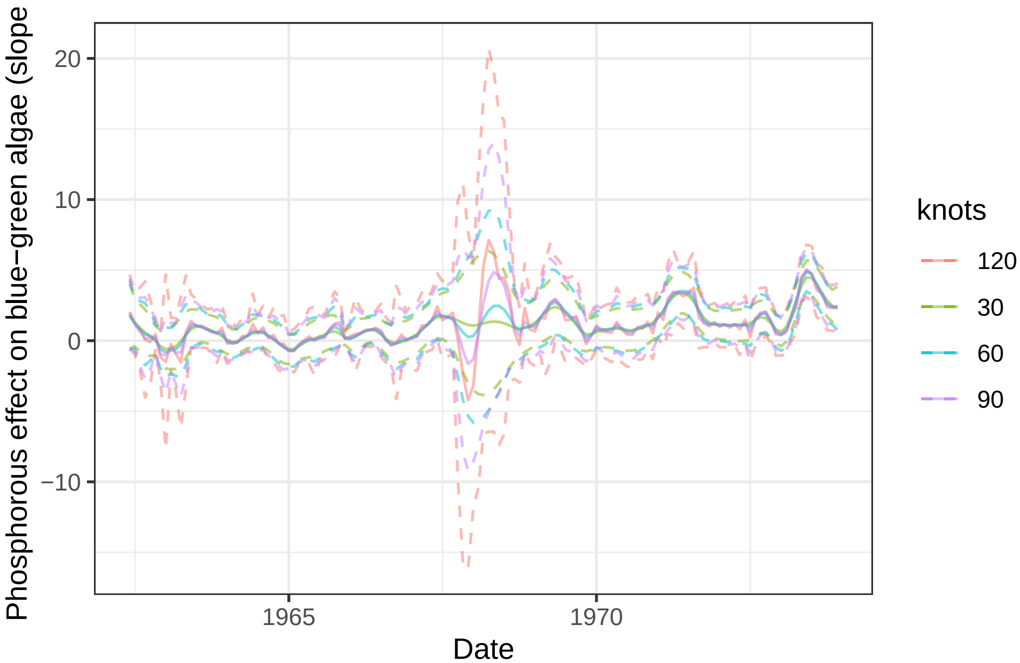

Figure S8: Effect of smoothing degrees of freedom on the standard errors of the estimates of time-varying slope parameter where color denotes model type. We show 4 distinct knots (colors)