Rationale and diagnostics --- Routines CO flux

Emma

Feb 2016

This document outlines the regression modelling of fully resolved temporal series of all routines sites (Pasqua, Wascana, Diefenbaker, Buffalo Pound, Last Mountain, Katepwa, Crooked). Pasqua series started in 2006, for the others CO flux calculations were available from the 1990s. Missing data were not replaced by extrapolation. Variable selection was based on results from previous analyses, indicating that on an annual resolution, climate, pH, and for between-lake differences, production, were the most important drivers of flux.

Since analysis has previously been done at an annual resolution, with lake-specific effects identified, both factors were added as random errors into the models. This effectively means that the model looks into the sub-annual variation in all data, given the effect of Year and Lake on the intercept.

pH has been shown to be the dominant control over CO flux, owing to its use in the calculations and the range over which it varies in the study sites. Therefore, the modeling process was divided into two steps. The primary model was designed to model pH (pH ~ .), and the secondary model to look into the detail of the CO ~ pH relationship. The latter was split into two steps, one model with flux ~ pH only (CO ~ pH), and a second with the residuals of this model against the rest of the parameters (res ~ .).

Further, to investigate whether any lake significantly departed from the global (i.e. "all-data") trend between predictor and response, the modelling at the outset included lake-specific smooths for each predictor. Where terms were insignificant, the predictor was defaulted to the global option. In marginally significant cases, anova methods for gam objects were used to test whether the more complicated model was justifiable.

### Variable selection

In order to avoid covariance and interpolation/introducing more NAs, some variables used previously were dropped. Further, owing to the within-year focus of the model (and taking out Year effects), explicitly annual-scale variables such as ice-off were dropped.

The Southern Oscillation Index (SOI) and Pacific Decadal Oscillation (PDO) were selected as climate variables owing to their significance in previous analyses. Furthermore, since SOI and PDO are known to interact (low flow especially during negative SOI and positive PDO; Bonsal and Shabbar 2008, Shabbar and Yu 2012), they were modeled as a tensor product. [Peter -- this from Rich's stuff -- I remember you said that this area may be different from the more widely established SOI-PDO pattern so if you could comment on this that would be great]

Several measures of lake production were collinear, including NPP-GPP-R, and P-N. GPP and N were selected for the model, N primarily as these lakes are largely N-limited and therefore a unique effect of N (rather than P) on production could be more expected. (However, the fine detail of *total* measures of N or P with respect to production or biological demand may be debated)

The variables selected were: chl a, oxygen, GPP (production/metabolism); SOI~PDO (climate); and DOC (since DOC can be metabolised and released as CO); lake + year (random effects). Shrinkage was allowed within the multivariable models to minimise variables with no effect. Chl a, N, and DOC were log10-transformed to control variance.

## Model results

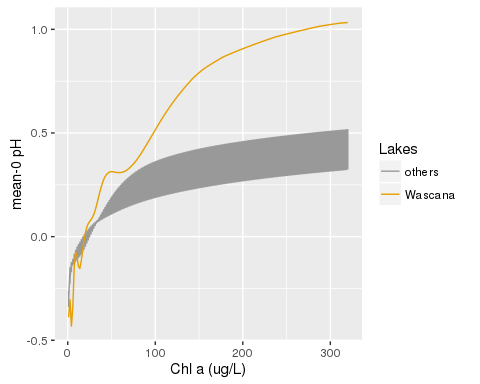
### Regression model for CO ~ pH and res ~ .

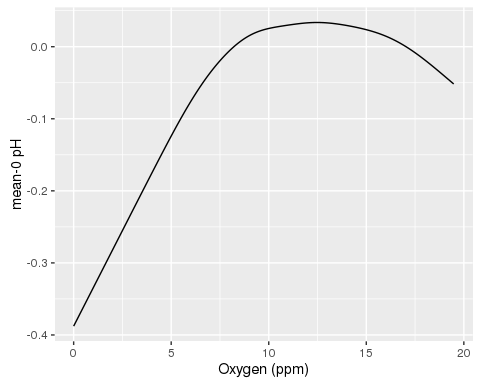
pH explained 97% of the CO flux with all lake-specific smooths significant. Differences in the smooths mostly occurred at the lower end of pH, where some lakes had higher-than mean effect on flux for a given pH (Katepwa, Last Mountain, Pasqua), and others lower (Buffalo Pound, Diefenbaker). Only marginal differences were found for Wascana and Crooked.

As unexplained deviance was very small (3%), the model of the residuals was trivial in biological significance. Further, said model explained only 8.6% of the residual variation and therefore 0.25% of all variation in CO flux. Chl a, PDO~SOI, and N were significant, with a higher-than-mean-effect on CO flux of N for Pasqua (i.e. the more N, the higher the flux compared with the overall trend).

### Regression model for pH ~ .

The pH model explained 42% of deviance. Out of all variables, GPP (shrunk out of the model), N (p < 0.1) and DOC (p < 0.1) were marginally or completely insignificant. As for previous analyses on an annual resolution, climate had the largest effect on pH, however production as chl a also had a significant effect in contrast to results from annual-resolution models. Furthermore, lake-specific effects were found for Wascana and Buffalo Pound, the former more significant than the latter (higher-than-mean-effect on pH).

  
Under ca 10ppm, decreasing oxygen had a negative effect on pH (holding all other variables at their mean).



## Note on diel data

Night-time flux for Wascana, based on running gas exchange equations with mean wind, and using altitude instead of pressure (initial quick run), and calculating DIC from conductivity (DIC or alkalinity not measured by sonde), indicates that no significant release of CO occurs at night to offset daytime influx. pH stays above 8.6. The most positive flux occurred in late September when both day- and night-time pH dropped below 8.6. These initial data suggest that it is plausible to assume that between-day/week fluctuations exceed diurnal fluctuations. However, if the lakes have a long ice-free period through September into November with sinking temperature and production, larger effluxes can occur.

## Model summaries

Summary of pH ~ .

##   
## Family: Scaled t(2.119,0.229)   
## Link function: identity   
##   
## Formula:  
## pH\_surface ~ s(log10(Chl\_a\_ug\_L)) + s(log10(Chl\_a\_ug\_L), by = Lake,   
## m = 1) + s(GPP\_h) + s(log10(TDN\_ug\_L)) + s(log10(DOC\_mg\_L)) +   
## s(Oxygen\_ppm) + te(PDO, SOI) + s(Lake, Year, bs = "re", by = dummy)  
##   
## Parametric coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 8.94617 0.02423 369.3 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df Chi.sq p-value   
## s(log10(Chl\_a\_ug\_L)) 9.770e-01 9 118.593 1.85e-12 \*\*\*  
## s(log10(Chl\_a\_ug\_L)):LakeK 8.217e-05 8 0.000 0.79873   
## s(log10(Chl\_a\_ug\_L)):LakeL 5.400e-05 8 0.000 0.50965   
## s(log10(Chl\_a\_ug\_L)):LakeB 2.777e+00 8 14.654 0.00513 \*\*   
## s(log10(Chl\_a\_ug\_L)):LakeC 5.953e-04 8 0.001 0.35388   
## s(log10(Chl\_a\_ug\_L)):LakeD 1.169e-04 8 0.000 0.40401   
## s(log10(Chl\_a\_ug\_L)):LakeWW 5.506e+00 8 76.557 4.37e-06 \*\*\*  
## s(log10(Chl\_a\_ug\_L)):LakeP 5.327e-01 8 0.871 0.19673   
## s(GPP\_h) 1.607e-04 9 0.000 0.33706   
## s(log10(TDN\_ug\_L)) 6.303e-01 9 8.739 0.06335 .   
## s(log10(DOC\_mg\_L)) 9.255e-01 9 5.416 0.08959 .   
## s(Oxygen\_ppm) 2.642e+00 9 29.425 0.00054 \*\*\*  
## te(PDO,SOI) 1.606e+01 24 754.485 < 2e-16 \*\*\*  
## s(Lake,Year):dummy 7.128e+01 94 278.275 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.288 Deviance explained = 41.9%  
## -REML = 461.09 Scale est. = 1 n = 749

Summary of CO2 ~ pH

##   
## Family: gaussian   
## Link function: identity   
##   
## Formula:  
## co2Flux ~ s(pH\_surface) + s(pH\_surface, by = Lake, m = 1) + s(Lake,   
## Year, bs = "re")  
##   
## Parametric coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.7608 0.7298 -10.63 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df F p-value   
## s(pH\_surface) 8.337 9 215.268 < 2e-16 \*\*\*  
## s(pH\_surface):LakeK 6.950 8 4.820 1.25e-05 \*\*\*  
## s(pH\_surface):LakeL 7.072 8 12.837 2.01e-15 \*\*\*  
## s(pH\_surface):LakeB 7.011 8 11.007 1.56e-15 \*\*\*  
## s(pH\_surface):LakeC 1.093 8 0.166 0.01760 \*   
## s(pH\_surface):LakeD 6.028 8 7.628 1.61e-10 \*\*\*  
## s(pH\_surface):LakeWW 3.489 8 1.175 0.00149 \*\*   
## s(pH\_surface):LakeP 6.857 8 6.750 6.64e-06 \*\*\*  
## s(Lake,Year) 69.465 94 3.057 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.965 Deviance explained = 97%  
## -REML = 2978 Scale est. = 86.084 n = 775

Summary of res ~ .

##   
## Family: Scaled t(2,3.54)   
## Link function: identity   
##   
## Formula:  
## res ~ s(log10(Chl\_a\_ug\_L)) + s(log10(Chl\_a\_ug\_L), by = Lake,   
## m = 1) + s(GPP\_h) + s(log10(TDN\_ug\_L)) + s(log10(DOC\_mg\_L)) +   
## s(Oxygen\_ppm) + te(PDO, SOI) + s(Lake, Year, bs = "re", by = dummy)  
##   
## Parametric coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -0.0443 0.1863 -0.238 0.812  
##   
## Approximate significance of smooth terms:  
## edf Ref.df Chi.sq p-value   
## s(log10(Chl\_a\_ug\_L)) 3.3051638 9 44.765 6.48e-11 \*\*\*  
## s(log10(Chl\_a\_ug\_L)):LakeK 0.0001993 8 0.000 0.801623   
## s(log10(Chl\_a\_ug\_L)):LakeL 0.0007864 8 0.001 0.276297   
## s(log10(Chl\_a\_ug\_L)):LakeB 0.0001454 8 0.000 1.000000   
## s(log10(Chl\_a\_ug\_L)):LakeC 0.0001395 8 0.000 1.000000   
## s(log10(Chl\_a\_ug\_L)):LakeD 0.0003402 8 0.000 0.425728   
## s(log10(Chl\_a\_ug\_L)):LakeWW 0.0002609 8 0.000 0.791917   
## s(log10(Chl\_a\_ug\_L)):LakeP 2.7652597 8 7.207 0.022826 \*   
## s(GPP\_h) 0.0553528 9 0.067 0.268912   
## s(log10(TDN\_ug\_L)) 3.2821943 9 23.500 3.04e-05 \*\*\*  
## s(log10(DOC\_mg\_L)) 0.5416733 9 1.213 0.142550   
## s(Oxygen\_ppm) 0.0009724 9 0.001 0.396108   
## te(PDO,SOI) 3.4660424 24 17.231 0.000437 \*\*\*  
## s(Lake,Year):dummy 9.6128905 88 11.571 0.084736 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.0186 Deviance explained = 8.1%  
## -REML = 2274.9 Scale est. = 1 n = 693