Project # 1 – Estimating standard errors using incomplete time series

Objective:
Develop an approach to estimate standard errors for the area-corrected abundance estimates of Peary caribou.

Data:
An example of the raw data is provided in Table 1 for Banks Island which is located in the Northwest Territories. Surveys were performed from the 1970s to 2014. The area surveyed (sampling transects) differed among surveys. The percent of the area surveyed represents the total area covered by sampling transects. The abundance estimate is the product of the average density across sampling transects (animals/km) and the area sampled. These surveys were repeated throughout the Canadian High Artic (Fig. 1) from the 1970s to 2014 and all of these data will be used in the analyses. Densities were not consistently extrapolated to the same area across survey years. To ensure consistency, survey areas were recalculated using a land mask that was generated from the CanVec dataset which is an open source digital cartographic reference product produced by Natural Resources Canada (http://open.canada.ca/data/en/dataset/83d45149-35e9-46e8-bab4-6f3d124a481c). The North Pole Azimuthal Equal Distance projection was used to estimate areas that were consistently surveyed. The standardized abundances are hereafter referred to as area-corrected abundance estimates.

Possible Approaches:
- Using a subset of the data (Banks Island), develop a preliminary statistical model with standard errors (SEs) as the response and predictors such as abundance and coverage (Table 1). Use the final model to derive SEs for the area-corrected abundance estimates for Banks Island.
- Write a program in R or SAS and repeat this procedure for the remaining islands such as Prince of Wales and Ellesmere (Fig. 1).

Table 1. Data derived from sampling surveys of Banks Island. The area-corrected population estimates have been extrapolated to an area of 70,579.70 km^2.

Year	Area Sampled (km2)	% Area Surveyed	Abundance Estimate	Coverage	SEs	95% CI	Area-corrected Abundance Estimate
1970	38804	1.8188769	5300	50			9640
1971	74332.659	0.9495113	10327	15.5			9806
1972	74332.659	0.9495113	12098	15.5			11487
1982	70,582	0.9999674	11066	20.14286		2927	11066
1985	70582	0.9999674	5765	16.75	910		5765
1987	70582	0.9999674	5517	7.34	660		5517
1989	70,028	1.0078783	3185	20	334		3210
1991	70,028	1.0078783	936	10	151		943
1992	70,583	0.9999532	1469	30	132.5		1469

1994	70,583	0.9999532	812	30	132		812
1998	70,583	0.9999532	566	20	60		566
2001	70,585	0.9999249	1142	20	155.4		1142
2005	70,583	0.9999532	1180	20	142.5		1180
2010	70,579.70	1	1347	20		343	1347
2014	70,579.70	1	2725	20		830	2725

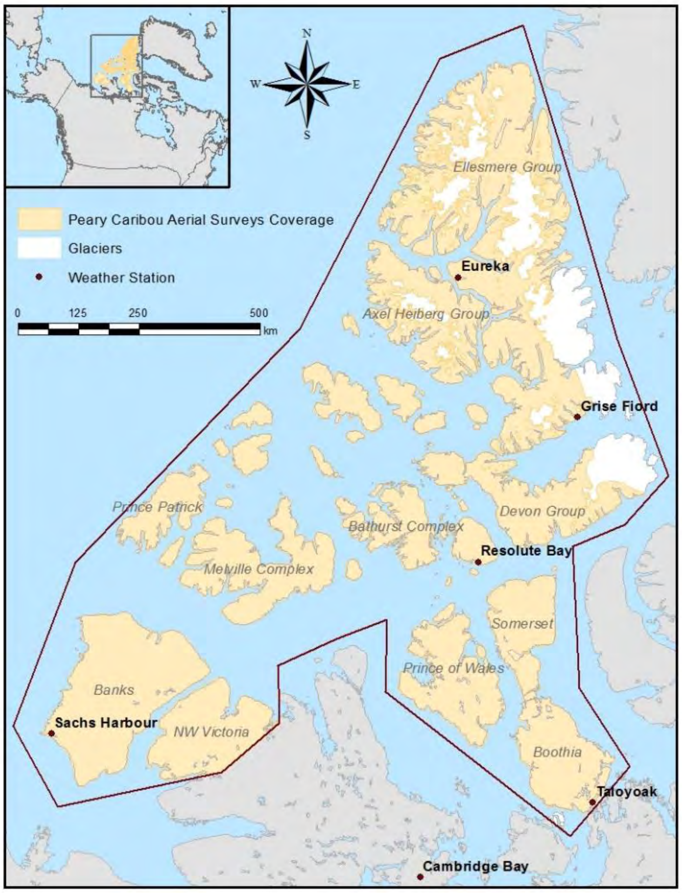

Figure 1. Survey areas.

Project # 2 – Developing data imputation techniques for time series of Peary caribou abundance

Objective:
Develop a data imputation approach to infill missing data resulting from non-systematic surveys for each of the four delineated populations (Fig. 1). The imputed time series will be summed up across islands which will provide estimates of population size.

Possible Approaches:

- Using a subset of the data (Banks Island and NW Victoria), develop statistical model(s) to fill in missing data e.g., Banks Island 1980 and NW Victoria Island 1982 (Table 1).
- Abundance data from neighbouring islands (e.g., Banks 1985) can be used to predict missing values for NW Victoria Island. Use the final model (s) to predict missing data.
- To assess the effects of imputation an R program will be provided which compares the rate of change from the infilled time series to the original time series. If possible, provide additional methods of assessing the effects of imputation.
- Write a program in R or SAS and repeat this procedure for the remaining islands such as Prince of Wales and Ellesmere (Fig. 1).

Data:
Surveys were performed from the 1970s to 2014. The area surveyed differed among surveys. The percent of the area surveyed represents the total area covered by sampling transects. The abundance estimate is the product of the average density across sampling transects (animals/km) and the area sampled. These surveys were repeated throughout the Canadian High Artic from the 1970s to 2014 which were delineated into four populations (Fig. 1). Abundances were not consistently extrapolated to the same area across survey years. All of the abundances were standardized to the same survey area and hereafter they are referred to as area-corrected abundances. These were used to reconstruct the time series of abundances for each population. The surveys were used to reconstruct the time series of abundances for each population. The reconstructed surveys were summed for of the four delineated populations (Fig. 1). For example, the Banks Island sub-population was not surveyed in 1980, but NW-Victoria Island was sampled. The assumption of no animals on Banks Island for 1980 can bias estimates of population trends and in an effort to reduce this bias an infilling procedure was used to complete the time series with an approximate population size. Miller et al. 2005 used population estimates from neighbouring islands to fill in historical data gaps. For each sub-population (e.g., island), the data gaps were infilled through imputation by: 1) log-transforming the original time series, 2) building a regression model, and 3) imputing an abundance estimate for missing years following the sequence of steps below. The final time series of abundance for each of the delineated populations are shown in Table 1 (an example of the Banks Island and NW-Victoria population) with the infilled values highlighted in yellow.

Fig. 1 The delineated Peary caribou populations are: 1) Banks Island-NW Victoria, 2) Price of Wales-Somerset-Boothia, 3) Western Queen Elizabeth Islands, and 4) Eastern Queen Elizabeth Islands.

Table 1. A summary of Peary caribou abundance estimates for the Banks Island and NW Victoria Island (Minto Inlet) population. The highlighted values were infilled using linear interpolation which has been described above. The summed abundances were an estimate of population size.

Year	Banks	NW-Victoria	Population Size
	Abundance	Abundance	
1970	9699		9699
1971	9866		9866
1972	11558		11558
1973-1979			
1980	5421	9202	14623
1981			
1982	11134	2456	13590
1983			
1984			
1985	5800	1715	7515
1986			
1987	5551	3825	9376
1988			
1989	3230	1062	4292
1990			
1991	949	836	1785
1992	1478	741	2219
1993	2604	271	2875
1994	817	52	869
1995			
1996			

1997			
1998	569	158	727
1999			
2000			
2001	1542	472	2014
2002			
2003			
2004			
2005	1187	160	1347
2006			
2007			
2008			
2009			
2010	1355	299	1654
2011			
2012			
2013			
2014	2742	53	2795

Reference:

Frank L. Miller, Samuel J. Barry & Wendy A. Calvert. 2005. Conservation of Peary caribou based on a recalculation of the 1961 aerial survey on the Queen Elizabeth Islands, Arctic Canada. Rangifer, Special Issue No. 16: 65-75.

Project # 3 – Towards improving Fluxmaster: statistical model to estimate nutrient loadings

Objective:

Fluxmaster is a five parameter program SAS program which has been developed by the USGS (http://water.usgs.gov/nawqa/sparrow/) to estimate nutrient loads (e.g., Total phosphorous in kg/year) input into a river system. Examine existing SAS code and implement in R. If possible provide recommendations on the existing statistical model on model fit and bias.

Possible Approaches:

- A simple data set will be provided to test the SAS code. Examine standard errors (preferred at <50%) and Bias (observed/predicted) (preferred at <50%).
- Take the SAS code and start programming the model in R. Suggest improvements.

Data:

A subset of data will be provided from the Red-Assiniboine River Basin (RARB) which spans portions of Canada (Manitoba and Saskatchewan) and the United States (U.S., North Dakota, South Dakota and Minnesota) (Fig. 1). These data will include constituents of loads which are estimated from streamflow and nutrient concentrations. Fluxmaster, which is a SAS program developed by the USGS (http://water.usgs.gov/nawqa/sparrow/), is used to combine these two sets of time-series data and generate a detrended mean annual load for each water quality monitoring station (Fig. 2).

Fig. 1 Red, Souris, Assiniboine and Qu'Appelle River Basins.

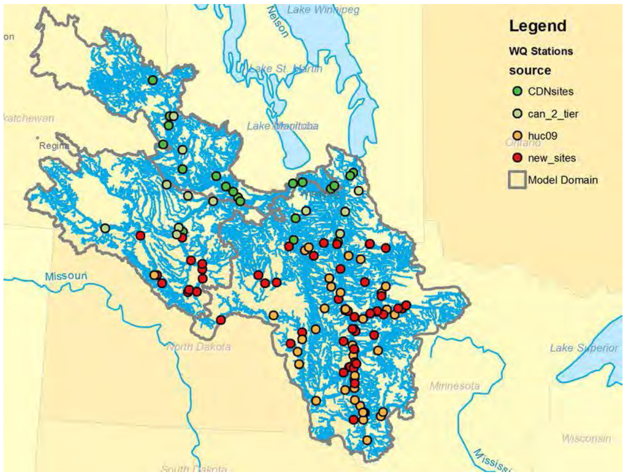

Fig. 2 Map of water quality monitoring stations from the Red and Assiniboine basins.

Fluxmaster:

The program implements regression methods and computes detrended long-term mean annual loads normalized to a base year (2002 for the Red-Assiniboine) (Fig. 3). The use of detrended mean annual helps compensate for differences in length of the period of record and frequency of monitoring data among sites. It also minimizes the inherent variability introduced by year-to-year variations in rainfall facilitating the identification of environmental factors that affect loading over long periods. The water-quality model (Equation 1) relates the logarithm of concentration *ct*, at time *t*, to: the logarithm of daily flow *qt*, a decimal time term to represent trend, *Tt*, sine and cosine functions of decimal time to account for seasonal variation, and a model residual, *et*,

$$ct = b0 + bq\, q + bT\, Tt + bs \sin(2\pi Tt) + bc \cos(2\pi Tt) + et$$

where *b0*, *bq*, *bT*, *bs*, and *bc* are coefficients estimated for each site by a ordinary least squares method or, if some of the *ct* measurements are censored, by the adjusted maximum likelihood method , and *et* is assumed to be independent and normally distributed.

Detrended flow is then estimated using a flow model with the form

$$q_t = a_0 + a_T T_t + a_s \sin(2\pi T_t) + a_c \cos(2\pi T_t) + u_t$$

where *a0*, *aT*, *as*, and *ac* are model parameters estimated using the maximum likelihood SAS Autoreg procedure (SAS Institute Inc., 2004), and *ut* is a model residual that is assumed to be correlated across time according to a 30-day lag autoregressive model.

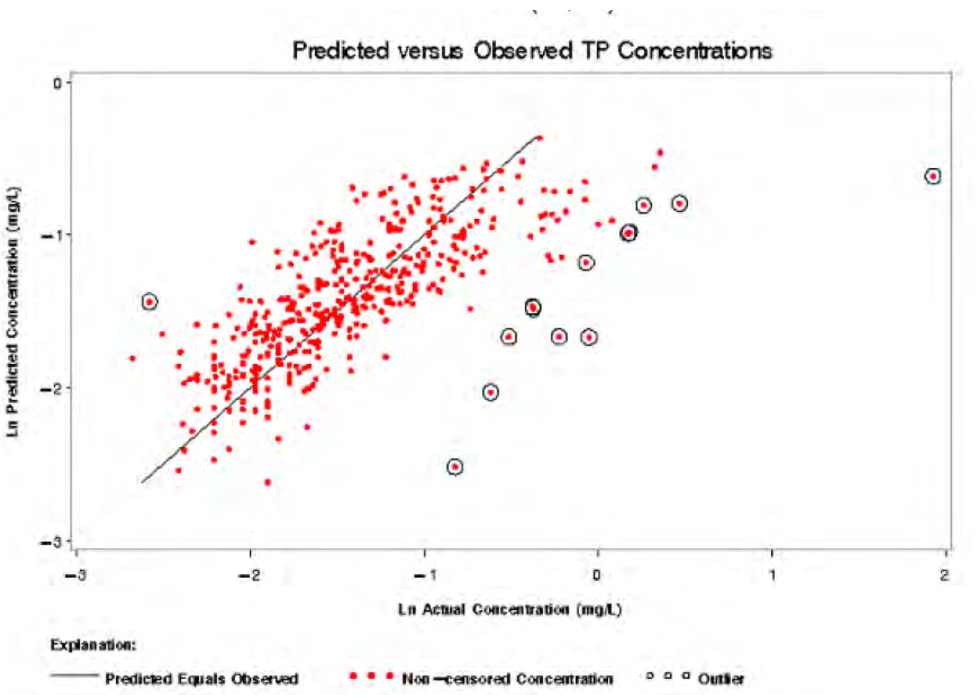

Fig. 3 Output from Fluxmaster (y-axis) compared to observed concentrations (x-axis).

Project # 4 – Has cancer rate decreased in fish throughout the Canadian Great Lakes?

Objective:
Develop a statistical approach to perform a statistical test to determine if the cancer rates in the exposed sites located throughout the Canadian Great Lake are greater than the rates at the reference sites.

Possible Approaches:

- An access database will be provided and SAS can be used to read in the data (e.g., proc sql). The data query should include tumour rates and fish length/age which can be used as covariate at the exposed and reference sites.
- Develop a simple test for a single exposed and reference site. Expand the analyses for the remaining sites possibly using bootstrapping techniques (some code will be provided) and Bayesian methods.

Data:
Data will be provided from survey sites across the Canadian Great Lakes (Fig. 1) in an access database. Queries will be used to extract exposed and reference sites. Table 1 has a summary of the exposed and reference sites. Additional data include fish length which should be used as a covariate and pre-cancer tumours which can develop into cancer tumours.

Table 1. AOC locations (exposed sites) and their associated reference locations with the years sampled for brown bullhead tumor studies (from west to east).

AOC Location (exposed)	Reference Site	Years Sampled
Detroit River	Peche Isle	2002
Wheatley Harbour	Hillman Marsh, Port Rowan	2002, 2006
Niagara River	Point Abino	2004, 2008
Hamilton Harbour	Jordan Harbour	2001, 2005, 2007
Toronto	Frenchman's Bay	2003, 2006
Bay of Quinte	Prince Edward Bay, Deseronto	2004, 2005
St. Lawrence River	Morrisburg	2004, 2005

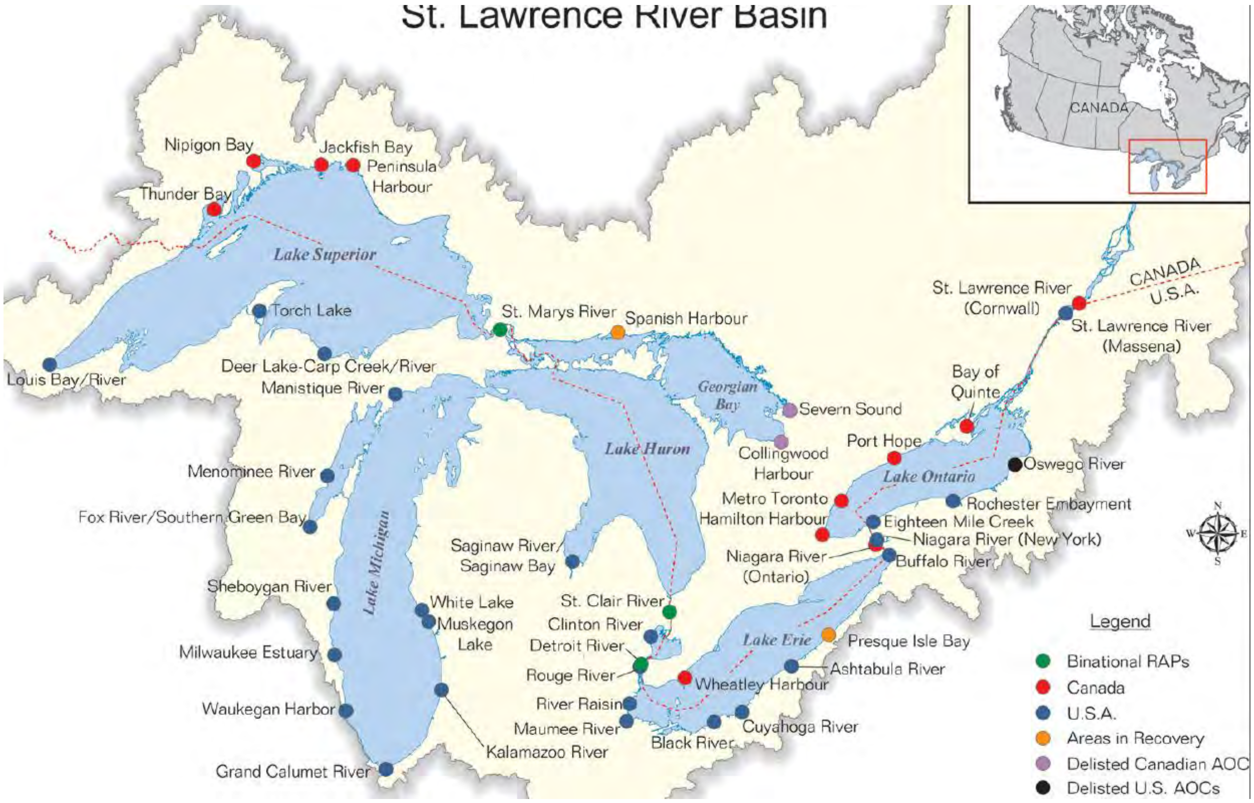

Fig. 1 The Great Lakes Areas of Concern (exposed sites).

Project # 5– Towards optimizing survey design to estimate nutrient loads

Objective:
Develop a simulation study that examines the effects of survey design on estimates of nutrient loading computed from co-located estimates of nutrient concentration and river flow.

Possible Approaches:

- Use the provided data to develop a statistical model that examines the effects of river flow, land use, and precipitation patterns.
- Use the final statistical model (s) to simulate the effects of removing sampling sites or adding sampling sites.

Data:
Data will be provided from the Bay of Quinte, Great Lakes Area of Concern which has been listed in the Great Lakes Water Quality Agreement (Fig. 1). The data will consist of water quality stations were nutrients (phosphorous and nitrogen) have been measured, river flow information, land use (e.g., crop versus urban), precipitation patterns, and agricultural practices.

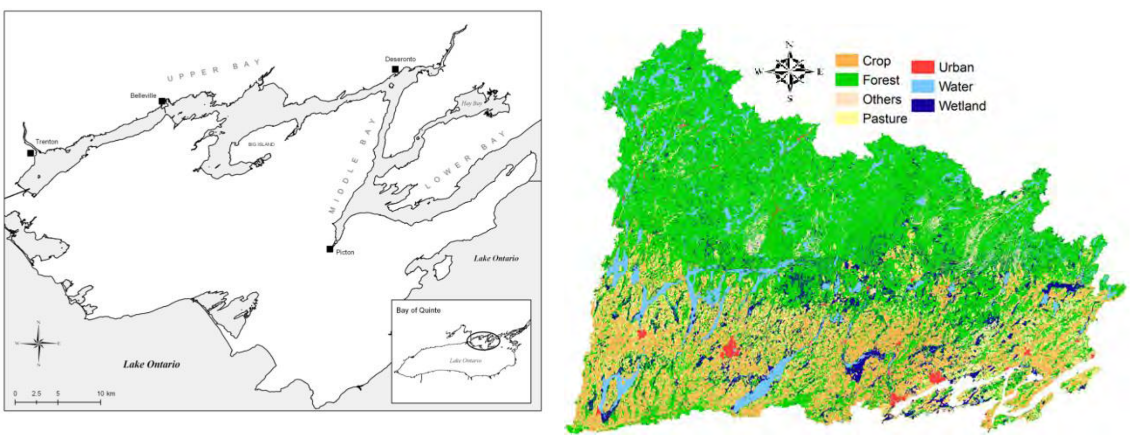

Fig. 1 Bay of Quinte Area of Concern (A) with landuse shown (B).

Project # 6 – Spatial analyses of input and output patterns on nutrients to identify hotspots

Objective:

Compare the spatial patterns of input data with modelled output to identify nutrient hotspots.

Possible Approaches:

- Examine spatial patterns using Neighborhood Nearest Index (does not account for spatial autocorrelation) and compare with Moran's I or Gary's global statistics which account for spatial autocorrelation.
- Additional local measures of spatial autocorrelation such ass Gettis-Ord Gi* statistic which is a local indicator of spatial patterns can also be estimated.
- Try applying mapping techniques to depict hotspots such as spatial ellipses and kernel density estimation.
- Chow, J. (2013, June 24). Visualising crime hotspots in England and Wales using {ggmap}. Retrieved from: http://www.r-bloggers.com/visualising-crime-hotspots-in-england-and-wales-using-ggmap-2/.
- DiMaggio, C. (2013). *P9489 Practicals and exercises. Part III: Spatial analysis in R*. Retrieved from: http://www.columbia.edu/~cjd11/charles_dimaggio/DIRE/resources/R/practicalsBookNoAns.pdf.
- Nelson, J. (2011, October 25). Mapping hotspots with R: The GAM. Retrieved from: http://www.r-bloggers.com/mapping-hotspots-with-r-the-gam/.

Data:

Data will be provided from the Bay of Quinte, Great Lakes Area of Concern which has been listed in the Great Lakes Water Quality Agreement (Fig. 1). The data will consist of water quality stations were nutrients (phosphorous and nitrogen) have been measured, river flow information, landuse (e.g., crop versus urban), precipitation patterns, and agricultural practices.

Phosphorus patterns were modelled using SPARROW (SPAtially Referenced Regressions On Watershed attributes) is a SAS program developed by the USGS (http://water.usgs.gov/nawqa/sparrow/) and it is a technique for relating water-quality measurements made at a network of monitoring stations to attributes of the watersheds containing the stations (Fig. 2).

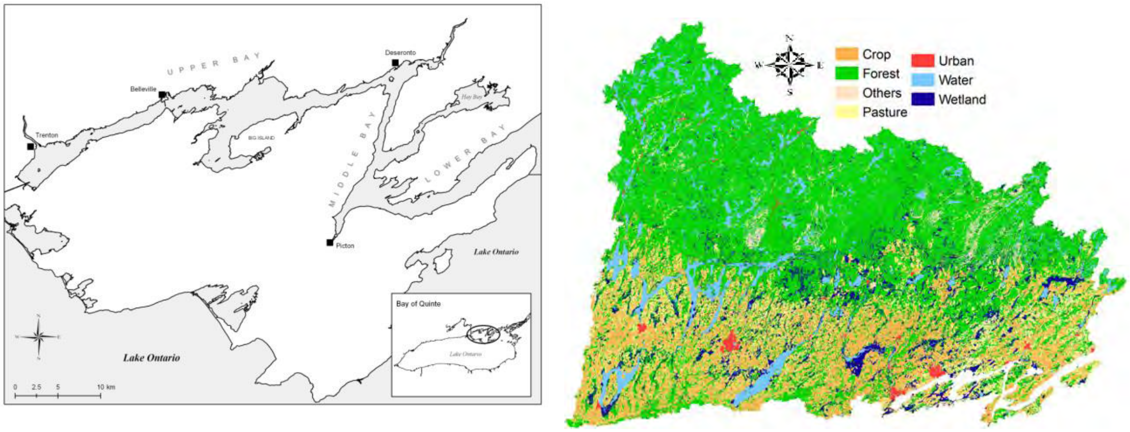

Fig. 1 Bay of Quinte Area of Concern (A) with landuse shown (B).

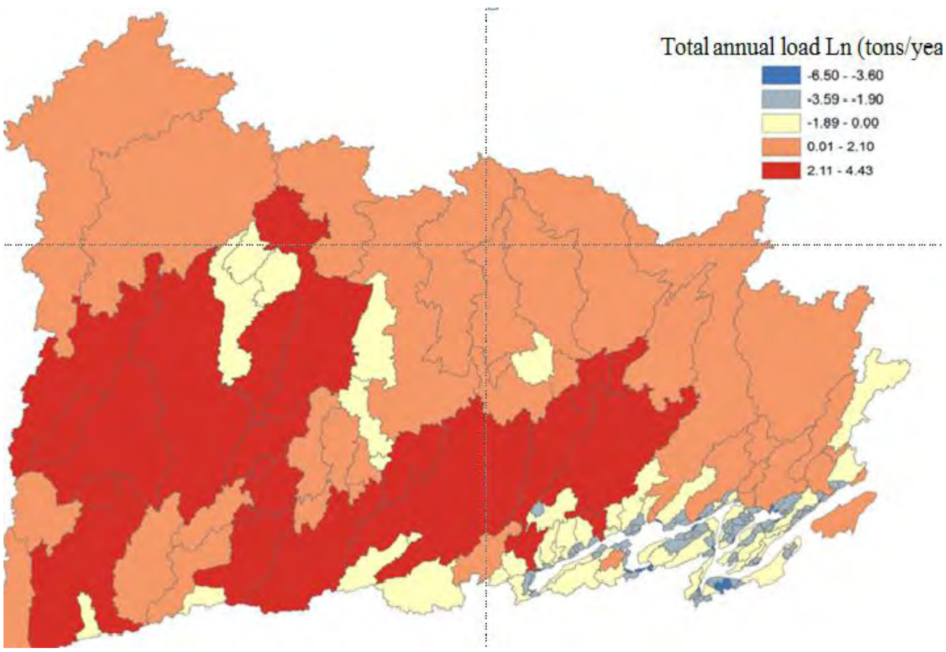

Fig. 2 Modelled output of phosphorous loads.

Project # 7 – Code conversion from WinBugs to R for a Bayesian calibration of a watershed model

<u>Objective:</u>

Convert code from an existing program that fits a non-linear regression in WinBugs (http://www.mrc-bsu.cam.ac.uk/software/bugs/) to R. This is a watershed model called SPARROW (<u>SPA</u>tially <u>R</u>eferenced <u>R</u>egressions <u>O</u>n <u>W</u>atershed attributes) and it is a technique for relating water-quality measurements made at a network of monitoring stations to attributes of the watersheds containing the stations (Fig. 2).

<u>Possible Approaches:</u>

- The input data are part of the code and the actual model is about 10 lines of code.
- Examine the input data and then run the model line by line. An illustration of the WinBugs code will be done either in class or remotely (e.g., webex).
- https://cran.r-project.org/web/packages/gibbs.met/gibbs.met.pdf

<u>Data:</u>

Input data and code will be provided from the Bay of Quinte, Great Lakes Area of Concern which has been listed in the Great Lakes Water Quality Agreement (Fig. 1). The data will consist of water quality stations were nutrients (phosphorous and nitrogen) have been measured, river flow information, landuse (e.g., crop versus urban), precipitation patterns, and agricultural practices.

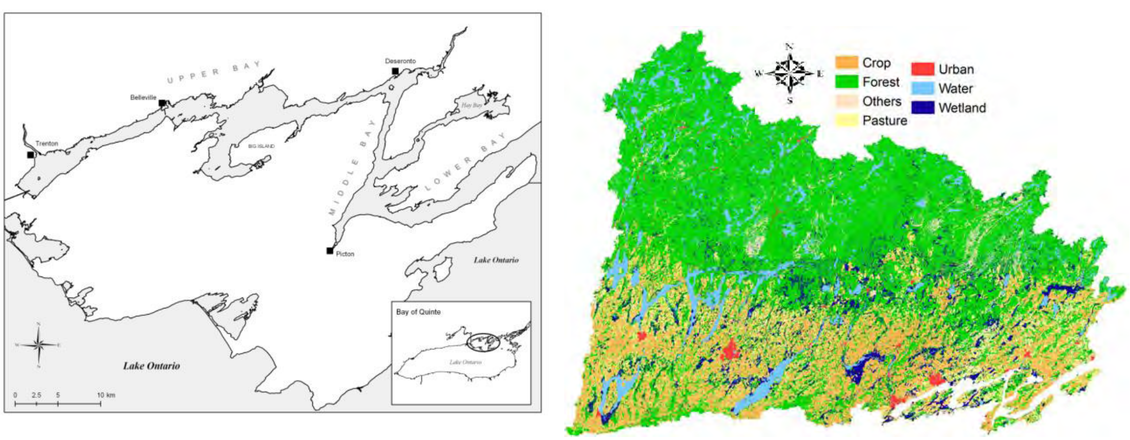

Fig. 1 Bay of Quinte Area of Concern (A) with landuse shown (B).

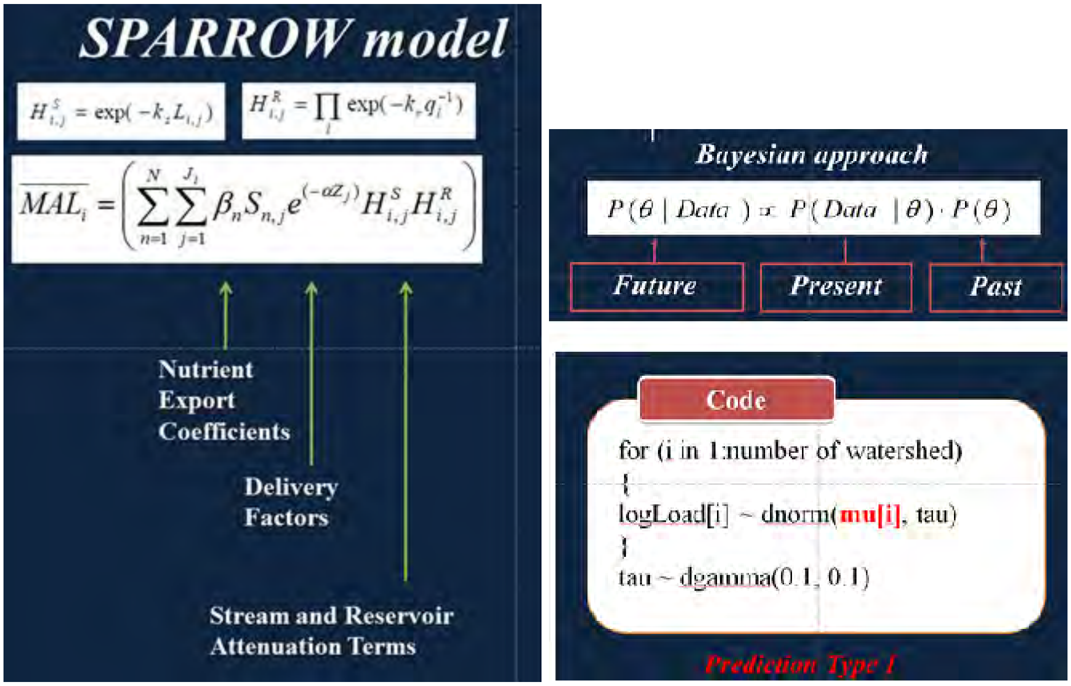

Fig. 2 SPARROW model structure that fits a non-linear regression once the watershed is divided up into smaller spatially homogenous areas called catchments and predicts nutrient loads. A Bayesian approach has been used to incorporate prediction errors.