Reference Model Supporting Documentation
for CABIN Analytical Tools: Columbia Basin 2020

MODEL NAME:	Columbia Basin 2020
AUTHORS:	Stephanie Strachan
AFFILIATION:	Environment and Climate Change Canada
DATE:	July 2020
CONTACT(S):	Stephanie Strachan stephanie.strachan@canada.ca or Jolene Raggett, British Columbia Ministry of Environment and Climate Change Strategy Jolene.Raggett@gov.bc.ca

IMPORTANT NOTE: Reference Groups in the CABIN Analytical tools are labeled 1,2,3,4,5 and 6. These group numbers correspond to the reference groups in this document D,E,F, G, IJK and LM, respectively.

1. STUDY DESIGN AND SITE SELECTION

1.1 Model Purpose

The Columbia Basin 2020 model is an update to the Columbia-Okanagan 2010 Preliminary model (Gaber 2012) currently uploaded to the CABIN web application. The development of the preliminary model was initiated by Environment and Climate Change Canada (ECCC), BC Ministry of Environment and Climate Change Strategy (BC ENV), and Parks Canada in 2006 to assess streams exposed to a variety of disturbances in the basin such as forestry, mining, urban development and agriculture. This revision excludes the Okanagan Basin, which is now covered by a separate model (Strachan and Bennett 2018), and focuses solely on the Columbia Basin in Canada. This revision includes more than twice as many reference sites as well as a validation dataset that was not applied in the evaluation of the preliminary model. The reference data were collected through a collaborative effort among CABIN participants in the Columbia Basin.

1.2 Spatial and Temporal Scope

Spatial Scope: The Columbia Basin has 7 sub-basins with up to 6 stream orders (based on 1:50,000 scale). The Basin includes 5 ecoregions based on National ecoregion classification: Columbia Mountains and Highlands, Northern Continental Divide, Selkirk-Bitterroot Foothills, Southern Rocky Mountain trench, and Western Continental Ranges. The BC ecoregion classification identifies a 6th ecoregion, the Purcell Transitional Ranges, which is within the Columbia Mountains and Highlands in the National ecoregion classification. Potential reference sites were distributed among all ecoregions and stream orders as best as possible based on accessibility to sampling locations.

Temporal Scope: The model was built using data collected in the late summer and early fall within the Columbia Basin between 2003 and 2018. Collaborative sampling efforts between ECCC, BC ENV, and Parks Canada began in 2007 with the majority of sampling conducted in 2007 and 2008 and fewer sites sampled in subsequent years. The model also includes sites that were visited in multiple years to capture temporal variation.

1.3 Site Selection

Site selection was focused on minimally disturbed sites from the wide variety of landscape types and stream sizes across the Basin, however the approach to identifying reference sites evolved over the years of data collection (2003-2018) with new GIS techniques and availability and accessibility to landscape level data. Initially, the three government agencies involved in the Columbia Basin sampling design (i.e., ECCC, BC ENV, and Parks Canada) proposed reference site criteria based on criteria documented elsewhere (Table 1, Rosenberg et al. 1999, Davies 1994).

Table 1. Initial reference site criteria used to identify potential reference sites in the Columbia Basin.

Site selection variable	Criteria
Distance from lake or wetland	2 km downstream for small lakes (<5 km^2) and 5 km downstream for large lakes (>5 km^2)
Distance from culverts	> 500 m downstream of flow structures (includes dams, weirs and waterfalls) and >50 m upstream of flow structures
Point sources	Avoid point sources whenever possible otherwise site should be > 10 km downstream for small streams (stream order 1,2,3) and > 20 km for large streams (stream order 4,5,6)
Forest fire (Based on Minshall et al. 2003)	Intensity >50% catchment burned = 10 years post fire Intensity 20-50% of catchment = 5 years post fire Intensity <20% catchment burned = 1 year post fire If intensity can not be derived, assume 5 years post fire is sufficient to consider a reference site
Road density	<0.5 km/km^2

BC ENV also developed watershed-level and stream-level criteria in 2012 (Norris, 2012) to apply a reference site selection tool using GIS to aid in site selection (Table 2).

Between 2003 and 2018, a total of 289 potential reference samples were collected by the federal government, provincial government, non-government organizations (NGOs), and industry via consultants for inclusion in the updated model. Potential reference sites were identified from a total of 10 different studies based on the defined reference site criteria (Table 3).

There are 156 unique potential reference sites in this model. Of those unique sites, 49 sites were revisited at some frequency between 2003 and 2018 to capture temporal variation. Approximately 50% of the revisited sites were sampled more than twice and approximately 20% were revisited at least 5 times (Table 4). Repeated sites were included in this revision to account for the wet/dry year variation that may not be captured by the habitat variables measured on site.

Table 2. GIS-based reference site selection criteria established in 2012 to guide reconnaissance efforts for potential reference sites.

Watershed Criteria	Reference Site Selection Tool
Urbanization	<0.1%
Agriculture	<5%
Forestry	<10%
Road density	<0.5 km/km^2
Forest fire	<10%
Pine beetle infestation	<10% infested
Stream Criteria	
Downstream distance from waterbodies <5 km^2	>2 km
Downstream distance from waterbodies >5 km^2	>5 km
Downstream distance from flow structures	>500m
Upstream distance from flow structures	>50m
Upstream distance from road crossings	>50m upstream from any crossings
Downstream distance from road crossings	>500m downstream
Upstream distance from current and past producing mines within 100m of stream	>500m
Downstream distance from Mining	No streams downstream from MINFILE
Riparian Areas - Natural Vegetation within 30m of stream	No human impact within 30m of stream

Table 3. CABIN studies and number of potential reference sites used in the revision of the Columbia Basin model.

CABIN STUDY	Total
BC ENV - Kootenay Region	73
Industry - Beaver River Project - Masse	10
Industry - Bingay - Masse	3
NGO - Nature Conservancy of Canada Darkwoods Monitoring Program	9
NGO - CBWQ-St. Mary Watershed Group	4
NGO - CBWQ-Upper Columbia Watershed Group	4
NGO - CBWQ-Windermere Watershed Group	10
ECCC - Columbia Basin	68
ECCC - Federal-Provincial WQ Monitoring Stations	38
Parks Canada - Mountain Parks Biomonitoring	70
Grand Total	**289**

Table 4. Repeat reference sites and years sampled between 2003 and 2018.

	Year sampled (2003-2018)														
Site code	**03**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Total**
Ald-01			1	1											**2**
BEA01		1	1	1	1	1		1	1	1	1	1	1	1	**12**
BIN02						1						1			**2**
COL02			1	2				1	1			1			**6**
COL03			1	1					1			1			**4**
COL05C				1								1			**2**
COL06				1								1			**2**
Cup-01			1	1											**2**
DUN01			1			1									**2**
DUT01			1										1		**2**
ELK02			1						1						**2**
ELK03			1					1						1	**3**
ELK04			1			1							1		**3**
ELK05			1	1	1		1	1	1	1	2	1	1	1	**12**
FLT01			1										1		**2**
FLT02			1			1							1		**3**
FLT03			1								1				**2**
FLT04			1	1											**2**
FLT05			1								1				**2**
FOR02												1			**1**
ILL01		1	1	1	1	1		1	1	1	1	1	1	1	**12**
ILL02			1	1					1			1			**4**
K-K-07-42			1	1	1		1	1		1	1	1			**8**
KOO02	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**14**
KOO03				1	1			1		1			1		**5**
KOT04A							1		1						**2**
KOT05A							1		1						**2**
LAR04A						1			1				1		**3**
LAR04B						1							1		**2**
LAR05						1			1				1		**3**
MOY02				1	1			1		1			1		**5**
NAVEN01						1	1	1		1					**4**
NAWIN01					1	1	1	1	1	1					**6**
NAWIN04							1	1	1	1					**4**
NCCCUL01				1	1										**2**
NCCHDN01				1	1										**2**
NCCLAB01				1	1										**2**
NGSTM01						1		1	1	1					**4**
PAY01				1		1									**2**
STM01				1		1									**2**
STM03				1		1									**2**

	Year sampled (2003-2018)														
Site code	**03**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**Total**
STM04				1	1			1							**3**
Ven-02			1	1											**2**
WIG01			1						1						**2**
WIG02			1		1								1		**3**
WIG03			1					1	1	1	1	1	1	1	**8**
WIG04			1					1							**2**
WIG04a			1								1				**2**
WIG05			1								1				**2**
TOTAL	**1**	**3**	**27**	**24**	**13**	**16**	**8**	**16**	**17**	**12**	**11**	**13**	**15**	**6**	**182**

A summary of the spatial distribution of reference sites among sub-basins, ecoregions and stream order is provided in Table 5. Stream orders 2 to 5 were more frequently sampled. Stream order 1 sites were harder to access and suitable habitat was harder to find where streams were not ephemeral. Due to the location of urban and residential areas as well as the cumulative inputs in downstream systems, stream order 6 sites seldom met reference condition criteria. Future revisions of the model may include a detailed GIS analysis of the Columbia Basin to apply a Human Activity Gradient approach (Yates and Bailey 2010) to find additional reference sites on larger waterbodies. The spatial distribution across the Columbia Basin is illustrated in Figure 1.

Table 5. Spatial distribution of reference samples among ecoregions, sub-basins and stream orders.

	Stream Order						
Ecoregion/Sub-basin	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Columbia Mountains and Highlands	**7**	**27**	**38**	**62**	**12**	**6**	**152**
Central Columbia	3	3	11	20	5	1	43
Central Kootenay	1		6	6			13
Lower Kootenay		13	12	8			33
Slocan		3	2				5
Upper Columbia	3	8	7	28	7	5	58
Northern Continental Divide	**5**	**8**	**13**	**10**	**16**		**52**
Central Kootenay	2	5	11	8	14		40
Flathead	3	3	2	2	2		12
Selkirk-Bitterroot Foothills	**1**	**4**					**5**
Central Columbia	1	4					5
Southern Rocky Mountain Trench	**1**	**8**	**2**	**6**	**2**		**19**
Central Kootenay			1				1
Upper Columbia	1	8	1	6	2		18
Western Continental Ranges	**3**	**4**	**7**	**29**	**16**	**2**	**61**
Central Kootenay			1	2			3
Upper Columbia	3	3	5	13		1	25
Upper Kootenay		1	1	14	16	1	33
Grand Total	**17**	**51**	**60**	**107**	**46**	**8**	**289**

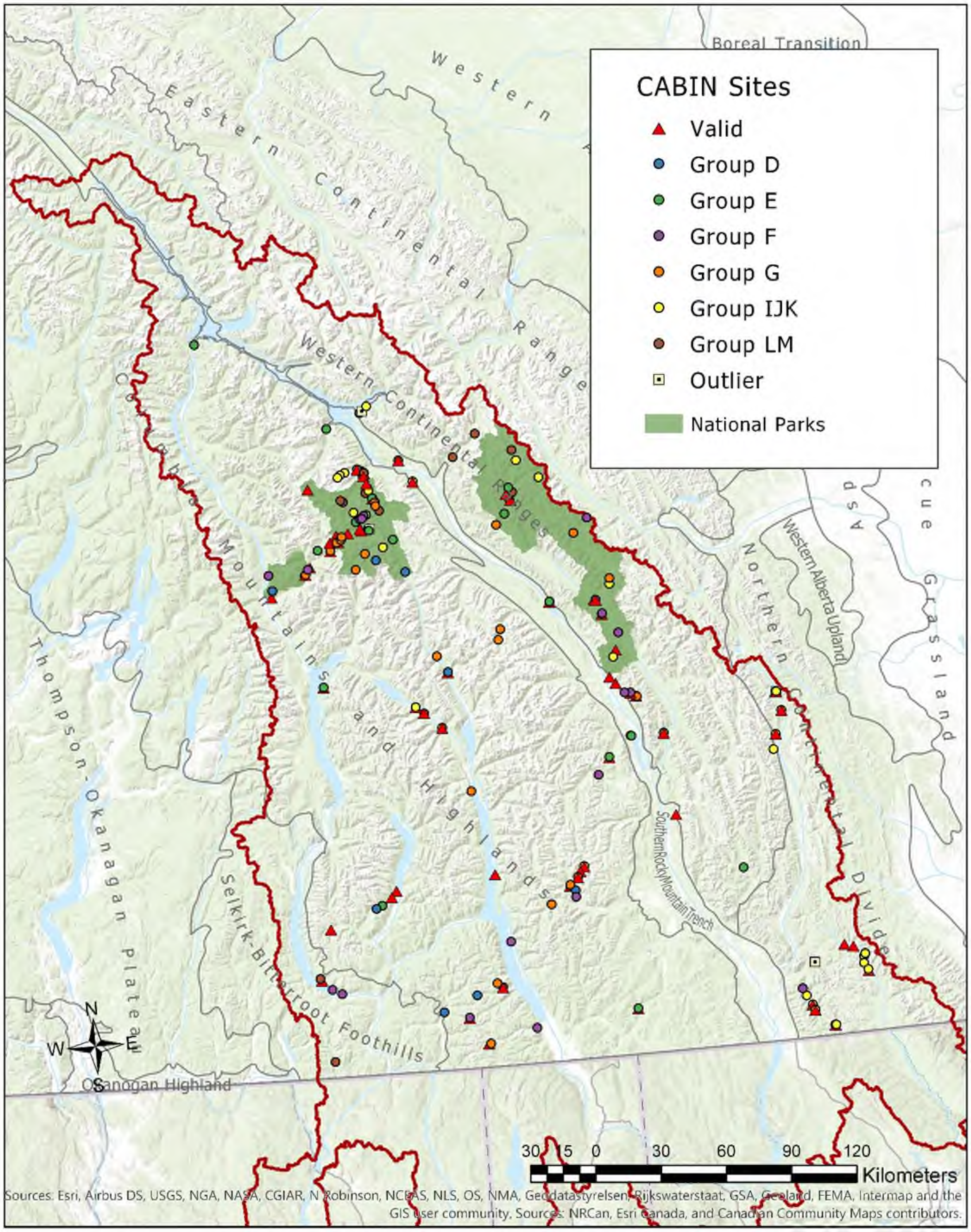

Figure 1. Distribution of reference sites among ecoregions (grey text labels and grey outline) in the Columbia River Basin (thick red outline). Circles represent sites used in the predictive model colour-coded by the biological group with outliers identified with squares. Triangles represent validation sites used to evaluate the model.

2. Exploring Benthic Group Structure

2.1 Cluster Analysis

The complete reference dataset of 289 samples was divided into a training dataset to build the model and a validation dataset to test the model. Approximately 30% of the reference sites were identified as validation data (N=84) and 70% as training data (N=205).

Group structure of the reduced training dataset (N=205) was analysed using SIMPROF in PRIMER6. SIMPROF tests for the significance of groups by assigning a significance factor (i.e., 1%). SIMPROF identified several significantly different groups (black lines in Figure 2), many of which had very few sites.

Small groups that branched off early were identified as outliers. In order to ensure a minimum group size of 10 sites, some groups in the dendrogram were subsequently rolled up to the parent branch to maintain the relationship with other sibling branches. In other words, if significant groups with more than 10 sites split off from the small group lower on the dendrogram, the entire branch was rolled up together. Alternatively, the small group was removed as outliers if only one or two sites were identified as unique sites. The final classification produced 6 groups and 7 outliers. One outlier site was completely unique from all the other reference sites. One outlier site had very low richness while two outliers had very high abundances as less than 5% of the sample was subsampled. Two remaining outliers had a unique composition of Diptera and Trichoptera.

SIMPER analysis was performed in PRIMER to compare within group similarity, among group dissimilarity as well as the contributing taxa based on this 6-group structure. The average within group similarity was 49%, with individual within group similarities ranging from 39.5% to 55.8% (Table 6). SIMPER illustrates that the taxa contributing to 80% of the similarity are common among many of the groups but the proportional composition of each of the major taxon groups differs among groups (Figure 3).

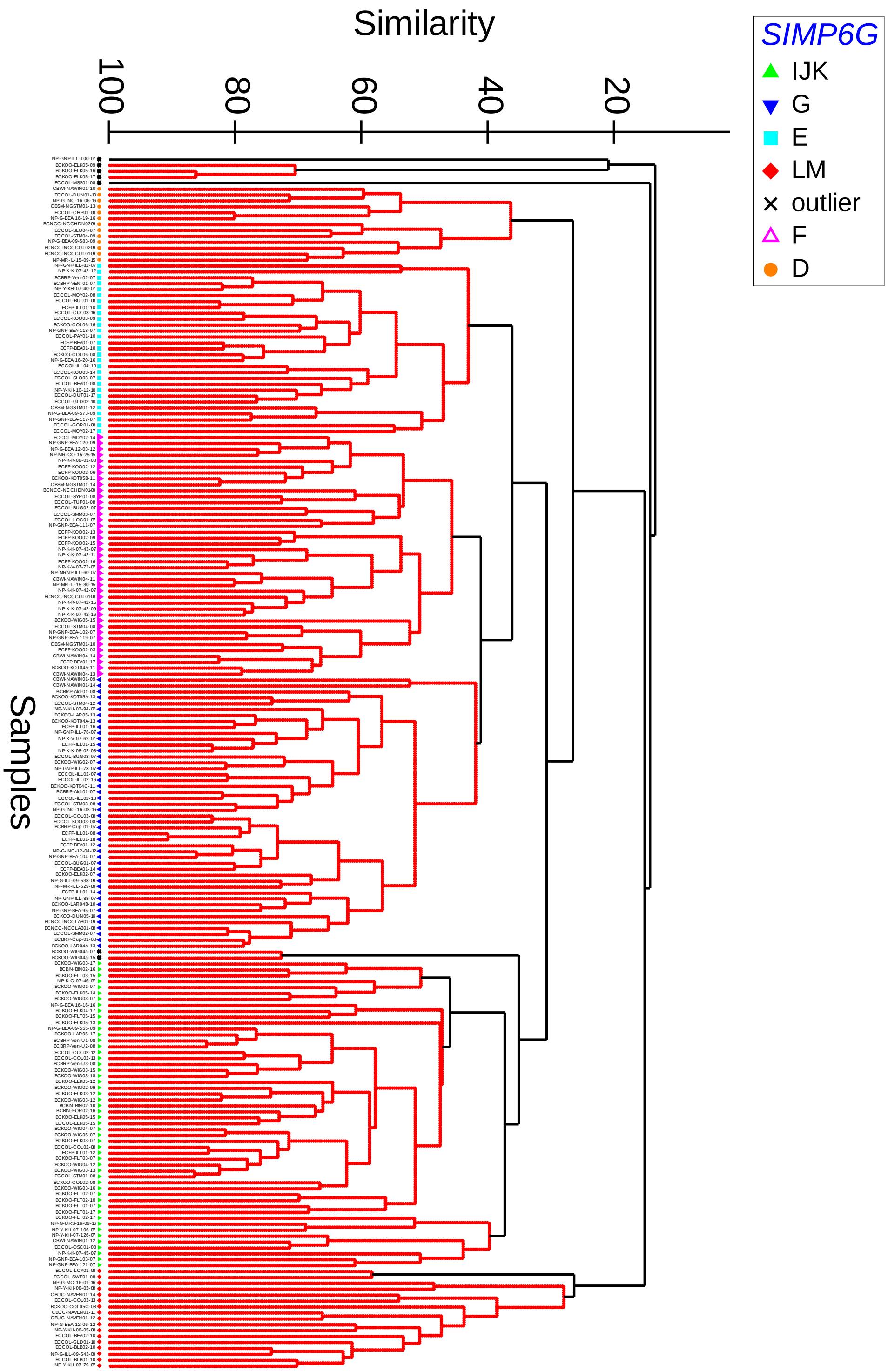

Figure 2. Dendrogram of the cluster analysis in PRIMER based on 123 family level taxa from 205 reference sites using a Bray-Curtis association and UPGMA hierarchical clustering. Black lines indicate significantly different groups based on SIMPROF with 1% threshold; red lines are not significantly different below the black parent branch. Some small but significantly different groups were rolled up to a parent branch for groups IJK and LM. Six reference groups are labeled by colour and symbols in the legend. Outliers are identified with an X.

Table 6. SIMPER analysis of the average within group similarity and the taxa contributing to 80% of similarity within groups.

Group	D	E	F	G	IJK	LM
Avg Sim	45.07	53.91	51.28	55.8	48.84	39.52
	Chironomidae	Heptageniidae	Heptageniidae	Baetidae	Heptageniidae	Heptageniidae
	Baetidae	Baetidae	Baetidae	Heptageniidae	Taeniopterygidae	Taeniopterygidae
	Heptageniidae	Taeniopterygidae	Taeniopterygidae	Ephemerellidae	Baetidae	Baetidae
		Chloroperlidae	Nemouridae	Nemouridae	Nemouridae	Chironomidae
		Ephemerellidae	Ephemerellidae		Chironomidae	Nemouridae
						Chloroperlidae

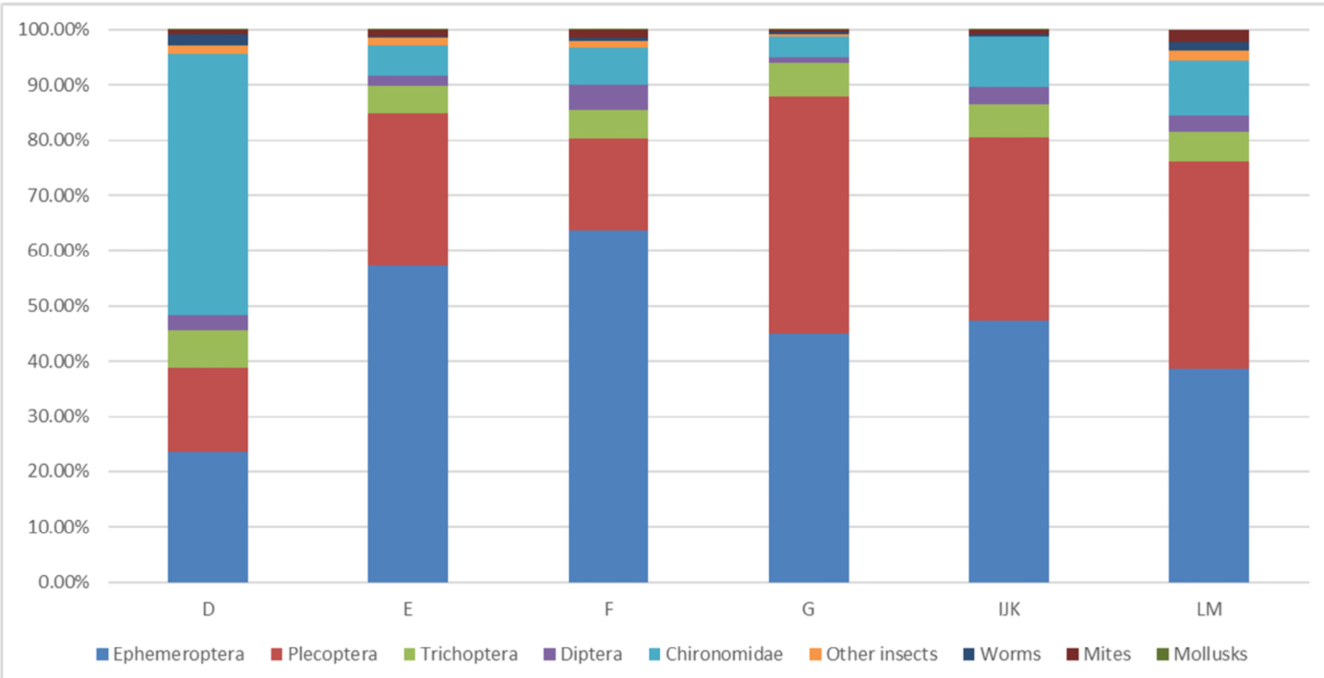

Figure 3. Proportional composition of major taxa groups

The average dissimilarity among groups was 75.2%. The most dissimilar groups were LM and IJK at 93.65% dissimilarity, while the lowest dissimilarity was between groups F and G, 58.82% (Table 7). Many of the taxa contributing to the differences among groups were similar, with abundance being a distinguishing difference between groups (Figure 4). The mean and standard deviation of select community metrics for each reference group are provided in Table 8.

Table 7. SIMPER analysis of the average dissimilarity between groups and the taxa contributing to 80% of the dissimilarity.

Avg Dissim	D	E	F	G	IJK
E	**76.06**				
	Chironomidae				
	Heptageniidae				
	Baetidae				
	Nemouridae				
	Ephemerellidae				
F	**68.59**	**67.02**			
	Chironomidae	Baetidae			
	Baetidae	Heptageniidae			
	Heptageniidae	Ephemerellidae			
	Nemouridae	Nemouridae			
	Ephemerellidae	Chironomidae			
G	**74.25**	**60.96**	**58.82**		
	Chironomidae	Taeniopterygidae	Baetidae		
	Taeniopterygidae	Heptageniidae	Taeniopterygidae		
	Heptageniidae	Baetidae	Heptageniidae		
	Baetidae	Nemouridae	Nemouridae		
	Nemouridae	Ephemerellidae	Ephemerellidae		
IJK	**74.79**	**82.22**	**65.09**	**82.88**	
	Heptageniidae	Heptageniidae	Heptageniidae	Heptageniidae	
	Taeniopterygidae	Taeniopterygidae	Taeniopterygidae	Taeniopterygidae	
	Chironomidae	Baetidae	Baetidae	Baetidae	
	Baetidae	Nemouridae	Nemouridae	Chironomidae	
	Nemouridae	Chironomidae	Chironomidae	Nemouridae	
			Ephemerellidae	Ephemerellidae	
LM	**86.8**	**67.9**	**85.73**	**83.48**	**93.65**
	Chironomidae	Heptageniidae	Baetidae	Heptageniidae	Heptageniidae
	Baetidae	Baetidae	Heptageniidae	Taeniopterygidae	Taeniopterygidae
	Nemouridae	Taeniopterygidae	Ephemerellidae	Baetidae	Baetidae
	Heptageniidae	Nemouridae	Nemouridae	Nemouridae	Nemouridae
	Ephemerellidae	Chloroperlidae	Chironomidae	Ephemerellidae	
		Chironomidae			

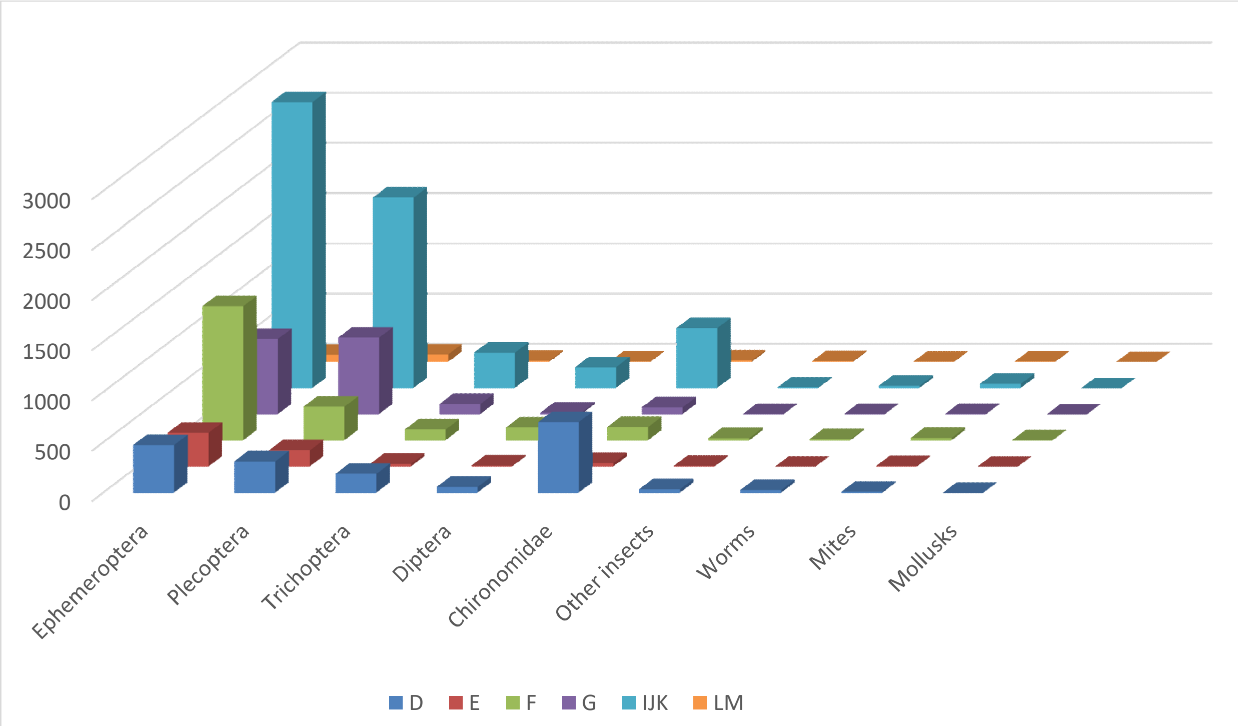

Figure 4. Average abundance of major taxonomic classification for each reference group.

Table 8. Mean and standard deviation of benthic community metrics of six reference groups.

	D (n=13)	**E** (n=29)	**F** (n=41)	**G** (n=46)	**IJK** (n=52)	**LM** (n=17)
Total Richness	16.2 ± 4.9	17.8 ± 4.5	18.15 ± 3.50	15.6 ± 3.2	17.6 ± 3.0	16.3 ± 6.1
Total Abundance	1829.48 ± 1127.75	597.17 ± 268.04	2124.08 ± 1031.80	1727.82 ± 771.79	5990.99 ± 3013.71	214.32 ± 171.45
Number of EPT taxa	11.2 ± 3.1	12.3 ± 2.3	12.2 ± 2.4	12.0 ± 2.0	12.4 ± 2.0	11.1 ± 3.2
% EPT	45.56% ± 21.51%	89.81% ± 7.67%	85.49% ± 10.70%	94.02% ± 5.73%	86.58% ± 10.96%	81.50% ± 15.83%
# E taxa	4.0 ± 0.9	3.7 ± 0.6	3.8 ± 0.6	3.6 ± 0.6	3.6 ± 0.6	3.7 ± 0.9
# P taxa	4.4 ± 1.3	5.5 ± 1.0	5.3 ±1.3	5.1 ± 1.2	5.6 ± 1.0	5.1 ± 1.2
# T taxa	2.9 ± 1.9	3.1 ± 1.5	3.1 ± 1.6	3.2 ± 1.2	3.2 ± 1.4	2.3 ± 1.6
% *Chironomidae*	47.28% ± 22.34%	5.59% ± 5.56%	6.75% ± 7.01%	3.84% ± 4.18%	9.14% ± 9.58%	9.84% ± 10.17%
% *Diptera*	2.87% ± 2.18%	1.87% ± 1.73%	4.58% ± 7.33%	1.03% ± 1.48%	3.05% ± 3.83%	2.99% ± 4.19%
% other insect taxa	1.42% ± 2.28%	1.24% ± 3.62%	1.13% ± 1.83%	0.32% ± 1.03%	0.11% ± 0.39%	1.89% ± 5.04%
% non-insect	2.87% ± 5.19%	1.48% ± 1.89%	2.06% ± 2.28%	0.79% ± 1.18%	1.12% ± 1.14%	3.78% ± 4.28%

3. Relating Habitat to Benthic Group structure

3.1 Determining candidate predictor variables for Discriminant Function Analysis

Nearly 200 habitat variables were available in the CABIN database for this dataset. Variables that were expected to respond to human disturbance such as water chemistry or land cover variables (e.g., related to agriculture, rangeland, developed land, exposed land, and forestry) were removed as potential predictors. The reduced list of candidate predictors consisted of 93 variables at a variety of scales. These habitat variables were not likely be affected by the human disturbances assessed by this model and have been shown to have some relationship to benthic communities based on the literature (i.e. Bailey et al. 2007, Collier 2008, Sandin and Johnson 2004, Marchant et al. 1997).

The habitat variables were grouped into five categories: Climate, Geography/Morphometry/ Topography, Channel, Landcover, and Bedrock Geology. The lengthy list was further reduced to remove potential multicollinearity through two different processes: Pearson correlations among variables and principle components analysis (PCA). Those variables that explained the most variation and were least correlated with other variables were retained for model development.

The list of 93 candidate predictors was reduced to a final list of 44 candidate predictors (Table 9). To provide a general summary of the different habitat among the reference groups, the mean and standard deviation of select habitat variables are provided in Table 10.

Table 9. Reduced list of candidate predictors

Geography/Morphometry/Topography	Climate
Latitude (dec deg)	Precip02_FEB (mm)
Longitude (dec deg)	Precip06_JUN (mm)
Altitude (masl)	Precip08_AUG (mm)
Drainage Area (km^2)	Precip10_OCT (mm)
Stream Density (m/km^2)	Temp03_MARmin (Degrees Celsius)
Elevation in u/s watershed, Max (m)	Temp06_JUN max (Degrees Celsius)
Slope in u/s watershed 30-50% (%)	Temp08_AUGmax (Degrees Celsius)
Slope in u/s watershed 50-60% (%)	Temp12_DECmin (Degrees Celsius)
Slope in u/s watershed, Avg (%)	
Slope in u/s watershed, Max (%)	**Channel**
Landcover	Depth, Max (cm)
Grassland (%)	% Canopy Coverage
Herb (%)	Dominant Streamside Veg (Category (1-4))
Rock/Rubble (%)	Coniferous (present/absent)
ShrubLow (%)	Grasses (present/absent)
ShrubTall (%)	Pools (present/absent)
SnowIce (%)	Runs (present/absent)
WetlandHerb (%)	Slope (m/m)
WetlandShrub (%)	Velocity, Max (m/s)
WetlandTreed (%)	Width, Bankfull (m)
Bedrock Geology	Dominant Substrate (Category (0-9))
Metamorphic (%)	Subdominant Substrate (Category (0-9))
Sedimentary (%)	Surrounding Material (Category (0-9))
Volcanic (%)	

Table 10. Mean and standard deviation of select habitat variables for each reference group.

	D (n=13)	**E** (n=29)	**F** (n=41)	**G** (n=46)	**IJK** (n=52)	**LM** (n=17)
Latitude (dec deg)	50.25 ± 0.92	50.82 ± 0.77	50.49 ± 0.73	50.75 ± 0.72	50.24 ± 0.96	51.22 ± 0.64
Longitude (dec deg)	-117.16 ± 0.67	-116.91 ± 0.87	-116.56 ± 0.76	-117.09 ± 0.75	-115.73 ± 1.19	-117.14 ± 0.62
Altitude (m)	1077.38 ± 408.83	915.21 ± 304.75	1111.22 ± 279.91	1006.59 ± 241.08	1322.42 ± 283.09	982.29 ± 238.47
Stream Order	3.0 ± 1.4	3.7 ± 1.3	3.6 ± 1.2	3.6 ± 1.1	3.1 ± 1.2	3.7 ± 1.2
BG-Sedimentary (%)	69.32 ± 39.17	91.26 ± 24.81	90.87 ± 24.77	88.92 ± 20.72	98.46 ± 8.11	86.84 ± 25.64
CH-Canopy Coverage (%)	1.00 ± 1.08	0.97 ± 0.87	1.54 ± 1.10	1.22 ± 0.94	1.06 ± 0.96	0.94 ± 0.97
CH-Depth-Avg (cm)	25.86 ± 18.09	32.29 ± 15.66	26.81 ± 13.79	27.53 ± 12.26	19.94 ± 8.61	25.49 ± 14.48
CH-Slope (m/m)	0.06 ± 0.08	0.05 ± 0.06	0.03 ± 0.04	0.03 ± 0.02	0.03 ± 0.03	0.04 ± 0.02
CH-Velocity-Avg (m/s)	0.47 ± 0.24	0.61 ± 0.17	0.53 ± 0.25	0.65 ± 0.31	0.58 ± 0.20	0.64 ± 0.27
CH-Width-Bankfull (m)	15.05 ± 14.68	23.36 ± 18.04	17.48 ± 16.07	22.07 ± 20.43	16.10 ± 13.05	32.54 ± 57.57
CH-Width-Wetted (m)	8.84 ± 9.67	14.37 ± 9.69	11.77 ± 13.51	14.76 ± 14.08	9.75 ± 7.74	13.60 ± 9.26
CL-Pcp02-FEB (mm)	108.70 ± 31.32	112.35 ± 43.33	88.30 ± 43.11	124.91 ± 38.70	83.23 ± 36.51	108.49 ± 41.61
CL-Pcp10-OCT (mm)	85.77 ± 23.93	94.41 ± 37.48	73.86 ± 35.19	103.05 ± 35.70	64.23 ± 33.58	94.27 ± 35.54
CL-Tmp08-AUGMax (°C)	17.81 ± 3.67	16.35 ± 3.31	18.38 ± 2.85	16.34 ± 2.33	17.35 ± 2.57	16.04 ± 3.49
CL-Tmp12-DECMin (°C)	-11.28 ± 2.41	-12.91 ± 1.59	-11.83 ± 4.93	-12.49 ± 1.59	-8.17 ± 10.03	-13.49 ± 2.12
LC-Grassland (%)	3.12 ± 7.29	4.74 ± 3.46	2.42 ± 3.72	4.35 ± 3.94	7.47 ± 6.30	7.29 ± 4.32
LC-Shrub low (%)	4.76 ± 3.24	4.93 ± 4.55	5.72 ± 2.61	3.50 ± 2.22	1.80 ± 1.50	3.75 ± 3.32
LC-Snow/ice (%)	4.36 ± 12.31	8.39 ± 9.84	3.27 ± 6.28	8.58 ± 8.39	3.79 ± 8.39	14.86 ± 16.40
LC-Water (%)	0.23 ± 0.51	0.24 ± 0.33	0.25 ± 0.39	0.28 ± 0.59	0.32 ± 0.59	0.28 ± 0.34
LC-Wetland shrub (%)	0.01 ± 0.01	0.02 ± 0.02	0.08 ± 0.13	0.01 ± 0.02	0.03 ± 0.07	0.04 ± 0.11
HY-Drainage Area (km^2)	215.94 ± 461.27	269.07 ± 346.96	271.08 ± 440.17	152.69 ± 249.74	100.10 ± 132.8	127.89 ± 95.64
SU-Dominant Substrate (category 0-9)	5.7 ± 1.7	6.0 ± 1.1	5.9 ± 1.3	6.3 ± 1.0	6.0 ± 0.8	6.3± 0.9
TO-Slope, Max% (%)	286.00 ± 129.33	479.28 ± 411.47	366.33 ± 330.01	371.47 ± 143.19	488.94 ± 542.33	392.17 ± 124.98

3.2 Stepwise DFA results

Both a backward and a forward stepwise discriminant function analysis (DFA) were performed in SYSTAT13 using the 6-group classification as the factor to compare which variables distinguished the reference groups and the resulting classification results. These analyses resulted in 12 or 15 predictor variables with individual group jackknifed classification rates ranging from 23% to 69%. It is important to obtain the best possible classification rates for individual groups for wide application of this model therefore attempts were made to improve groups with low classification rates. The CABIN Science Team recommends classification rates of at least 2x random (refer to CABIN Model Builders Guidance). In this case, the model has 6 groups therefore random classification probability would be 100/6=16.7%. Therefore, the recommended minimum classification rate for an individual group was 16.7%x2=33.4% for this model.

Various iterations of DFA models were investigated by substituting variables with low F values (limited contribution to discrimination) and or low tolerance scores (correlated with another variable) with other variables that might improve poor classification rates of some groups. A decrease in classification rate of a group that was very high was an acceptable trade off to improve a group that had a very low classification rate. The goal was to achieve a reliable prediction regardless of the predicted group. Based on numerous iterations with the available predictors, it was impossible to achieve a 33% classification rate for Group D with the candidate predictor variables. This model will have a difficult time predicting to this group due to its overlap or similarity to with Groups E and F (Figure 5).

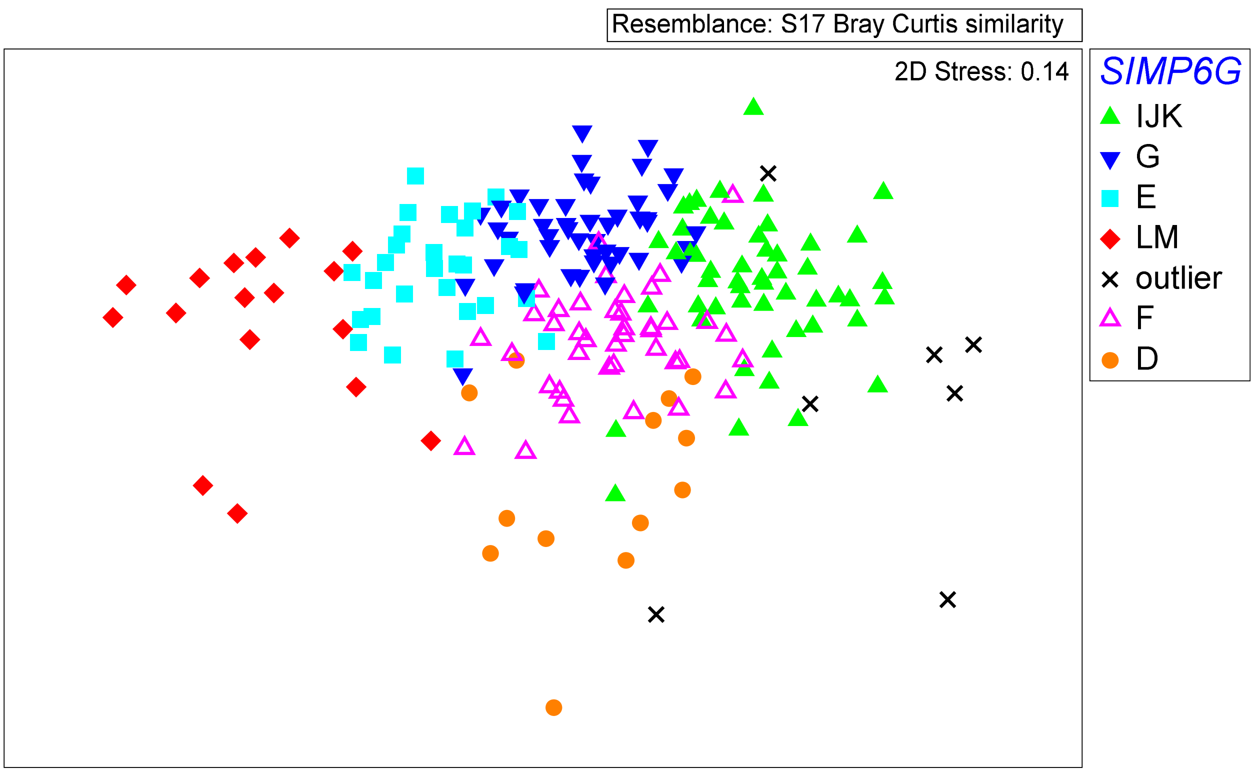

Figure 5. Multidimensional scaling (MDS) plot of 205 reference samples using Bray-Curtis similarity. Samples are labeled by different reference group symbols and colours; outliers are identified with an X.

4. Recommended Model

The recommended model is a 6 group model using 12 predictor variables (Table 12). The minimum group jackknifed classification rate was 31%, which is 1.8 x better than random (Table 13), similar to the forward stepwise model classification rates and better than the backwards stepwise model classification rates. A 33% jackknifed classification rate was not possible with the available predictor variables. Perhaps other environmental variables not considered here could improve the classification rate. In addition, more reference sites for Group D, which currently has only 13 reference sites, may provide a more predictable description of the biological community.

DFA output of the recommended Columbia Basin model

Table 12. DFA variable output from SYSTAT for the recommended model ($p<0.005$) and references for biological relevance for the predictor variables.

Variable	F-to-remove	Tolerance	Biological relevance
Longitude	4.991	0.281	Carter et al. 1996
Altitude	6.535	0.649	Namayandeh et al 2011, Corkum 1989
Bedrock Geology-Sedimentary (%)	2.073	0.534	Corkum 1989
Channel-Reach-Canopy Coverage (%)	3.065	0.632	Zimmerman and Death 2002
Channel-Slope (m/m)	2.276	0.749	Corkum 1989
Climate-Precipitation, OCT (mm)	2.409	0.311	U.S. EPA 2008, Gutiérrez-Fonseca et al 2018
Climate-Temperature, DEC min (deg C)	5.789	0.524	U.S. EPA 2008, Scrine et al. 2017
Hydrology-Drainage Area (km^2)	3.083	0.729	Corkum 1989, Vinson and Hawkins 2003
Landcover-Grassland (%)	5.497	0.755	Scotti et al. 2020
Landcover-Shrub low (%)	7.28	0.872	Scotti et al. 2020
Landcover-Water (%)	1.336	0.785	Scotti et al. 2020
Topography-Slope in u/s watershed, max (%)	0.675	0.7	Carter et al. 1996, Church 2002
Wilk's Lambda	0.268	Df (12, 5, 192)	
Approx. F-ratio	4.614	Df (60, 832)	

Table 13. DFA classification output from SYSTAT for the recommended model

Classification Matrix (Cases in row categories classified into columns)						
	D	E	F	G	IJK	LM	%correct
D (13)	**6**	2	4	0	1	0	46
E (29)	4	**13**	5	3	0	4	45
F (41)	6	3	**27**	3	2	0	66
G (46)	3	4	8	**22**	3	6	48
IJK (52)	3	1	1	6	**39**	2	75
LM (17)	1	0	2	2	0	**12**	71
Total	23	23	47	36	45	24	**60**
Jackknifed Classification Matrix							
	D	E	F	G	IJK	LM	%correct
D (13)	**4**	3	4	0	1	1	31
E (29)	5	**10**	4	4	0	6	34
F (41)	6	3	**25**	3	4	0	61
G (46)	3	5	9	**19**	3	7	41
IJK (52)	3	1	1	6	**39**	2	75
LM (17)	2	0	2	2	0	**11**	65
Total	23	22	45	34	47	27	**55**

4.1 Evaluating Model Performance

Type I Error Rate

Validation data were used to assess Type I error by running the model on these reference data as test sites. This approach determined the reference group to which each site most likely belongs and then compared the benthic community of the validation site to the predicted group of reference sites. The proportion of validation sites falling outside of a given threshold is the Type I error rate at that threshold.

A variety of thresholds were used to evaluate Type I error. Using a standard CABIN 90% confidence ellipse, a 10% Type I error rate is expected. However, the Type I error rate was much higher than expected and differed among reference groups. Based on the 90% ellipse, the Type I error rate for the validation sites was on average 59%.

The Type I error rate was further evaluated using 95% and 99% confidence ellipses (Table 14). It is important to note that 71 out of 84 validation sites (85%) fell in the standard CABIN assessment Band 2, between 90 and 99%, which is described by CABIN as Mildly Divergent. Often, sites that fall in Band 2, Mildly Divergent, do not require management action but instead continued monitoring to track change. The actions taken will depend on the monitoring project. Sites outside of the 99% ellipse, described by CABIN as Divergent, may require management action to determine the cause of the divergence and possibly remediation. The average Type I error rate was 39% at the 95% ellipse and 16% at the 99% ellipse (Table 14).

Table 14. Type I error rate of validation site assessment as predicted by the model across various thresholds (N=84).

	90% ellipse	95% ellipse	99% ellipse
D (n=10)	70%	50%	0%
E (n=11)	82%	55%	45%
F (n=18)	72%	56%	17%
G (n=19)	53%	42%	11%
IJK (n=21)	38%	33%	24%
LM (n=5)	40%	0%	0%
AVG	**59%**	**39%**	**16%**

Simpacted Dataset

The assessment of the Type II error rate requires that the validation benthic data be altered or simulated to represent an impacted community - a community that is no longer in the range of reference condition. These are referred to as "*SIMPACTED"* sites. The group predictions determined for the validation sites for the Type I error rate assessment apply here. The proportion of *SIMPACTED* validation sites that do not fall outside a given threshold is the Type II error rate at that threshold.

Three different *SIMPACTS* were applied for this evaluation to reflect typical disturbance from resource development (i.e., sedimentation, erosion, nutrient increases). A resource development tolerance score specific to BC was established based on all data in BC projects from more than 5500 sites in the CABIN database. Scores were based on correlations of taxa abundance with variables that are known to be influenced by resource development such as Turbidity/TSS, Conductivity, Temperature, Substrate. The tolerance score was used to determine the sensitive, insensitive and tolerant taxa and the degree of their tolerance or sensitivity (Table 15).

Table 15. BC specific family level tolerance scores for sensitive and tolerant taxa derived from correlations with substrate and water quality parameters influenced by resource development based on records in the CABIN database.

Sensitive (S)	Tol Score	Sensitive (S)	Tol Score	Tolerant (T)	Tol Score	Tolerant (T)	Tol Score
Ameletidae	-0.6875	Leuctridae	-0.5625	Acalyptonotidae	0.5	Hydropsychidae	0.625
Anisogammaridae	-0.375	Limnephilidae	-0.4375	Aeolosomatidae	0.375	Hydroptilidae	0.5
Asellidae	-0.5	Lumbriculidae	-0.5625	Aeshnidae	0.4375	Lebertiidae	0.8125
Athericidae	-0.4375	Perlidae	-0.6875	Ametropodidae	0.4375	Lepidostomatidae	0.5625
Baetidae	-0.5625	Rhyacophilidae	-0.5625	Apataniidae	0.5625	Leptoceridae	0.5625
Capniidae	-0.5	Sminthuridae	-0.4375	Aturidae	0.5625	Leptophlebiidae	0.5625
Chloroperlidae	-0.5625	Sperchontidae	-0.5625	Brachycentridae	0.625	Limnesiidae	0.5
Elmidae	-0.625	Taeniopterygidae	-0.4375	Ceratopogonidae	0.4375	Mideopsidae	0.4375
Empididae	-0.625	Torrenticolidae	-0.5625	Chironomidae	0.5625	Nemouridae	0.5625
Enchytraeidae	-0.375	Uenoidae	-0.375	Dixidae	0.5625	Perlodidae	0.5
Glossosomatidae	-0.5			Dytiscidae	0.4375	Psychodidae	0.375
Heptageniidae	-0.8125			Ephemerellidae	0.75	Simuliidae	0.5
Hydryphantidae	-0.5			Gammaridae	0.375	Tipulidae	0.625
Hygrobatidae	-0.5625			Hyalellidae	0.5		

Three different *SIMPACTS* were applied to the validation dataset:

1) S1: ABUNDANCE SIMPACT = a large increase in the abundance of tolerant taxa with a mild decrease in sensitive taxa abundance and limited loss of taxa (i.e. TolScore x sample abundance adding 75% of S abundance back to the community and increasing T abundance by a factor of 5)
2) S2: ABUNDANCE + RICHNESS SIMPACT = a moderate increase in tolerant taxa abundance with a loss of most sensitive taxa (i.e. TolScore x sample abundance adding 50% of S abundance back to the community and increasing T abundance by a factor of 2)
3) S3: RICHNESS SIMPACT = a decrease in tolerant taxa abundance and loss of all sensitive taxa (i.e. TolScore x sample abundance with no factors added therefore S taxa lost and slight decrease in T abundance)

There was no *SIMPACT* for species replacement. The *SIMPACTs* were based solely on the taxa that were found in the original community. The average change for the validation data and for each *SIMPACT* type for each model group is illustrated in Figure 6.

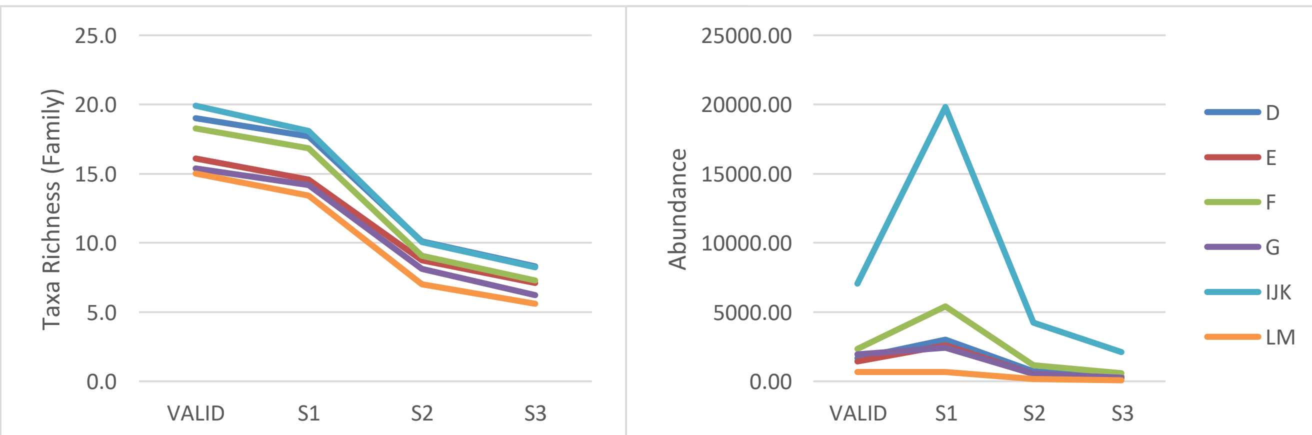

Figure 6. Change of richness and abundance to validation data predicted to each reference group based on three different *SIMPACTS*. VALID = Original Validation data, S1=Abundance SIMPACT, S2=Abundance + Richness SIMPACT and S3=Richness SIMPACT

Type II Error Rate

The validation sites were *SIMPACTED* in three different ways resulting in 252 tests for Type II error. Each site was plotted with each of the 3 different *SIMPACTED* communities in the same ordination and assessed for Type II error using a variety of confidence ellipses (Table 14). Type II error rates were very low. With the exception of Group D, the model was very good at detecting different abundance and richness *SIMPACTS* at all thresholds. On average, the model is able to detect decreases in taxa richness easily (S2 and S3). It has difficulty detecting changes in communities that are related to abundance with little change to taxa richness (S1), particularly with Group D and Group IJK which had a large abundance of Chironomidae (tolerant taxon) in the training dataset.

Table 16. Type II error rates of *SIMPACTED* site assessments as predicted by the model across various thresholds with 3 types of applied *SIMPACTS*.

	90% ellipse				95 % ellipse				99% ellipse			
	S1	S2	S3		S1	S2	S3		S1	S2	S3	
D (n=10)	30%	30%	30%		60%	40%	40%		100%	80%	70%	
E (n=11)	0%	0%	0%		0%	0%	0%		9%	0%	0%	
F (n=18)	0%	0%	0%		0%	0%	0%		17%	6%	0%	
G (n=19)	0%	0%	0%		11%	0%	0%		11%	0%	0%	
IJK (n=21)	0%	0%	0%		10%	10%	0%		48%	48%	24%	
LM (n=5)	0%	0%	0%		20%	40%	40%		80%	80%	60%	
AVG	**5%**	**5%**	**5%**		**17%**	**8%**	**7%**		**44%**	**36%**	**26%**	
Overall Average				**5%**				**10.7%**				**35.3%**

The model did a poor job of detecting *SIMPACTS* of any type with reference group D due to the fact that the training data for Group D are so variable and have a high proportion of tolerant taxa (Figure 7). The high error rates with Group D appear to be due to the difference in community composition between the Training dataset (n=13) and the Validation dataset (n=10) (Figure 7). While these data were deemed reference *a priori* by the reference site selection criteria, the communities are clearly different and therefore evaluated as divergent before any *SIMPACT* was applied (refer to Type I error in Table 14). The large biological variation in the Group D training data suggests that additional reference sites are needed to get a more precise description of Group D.

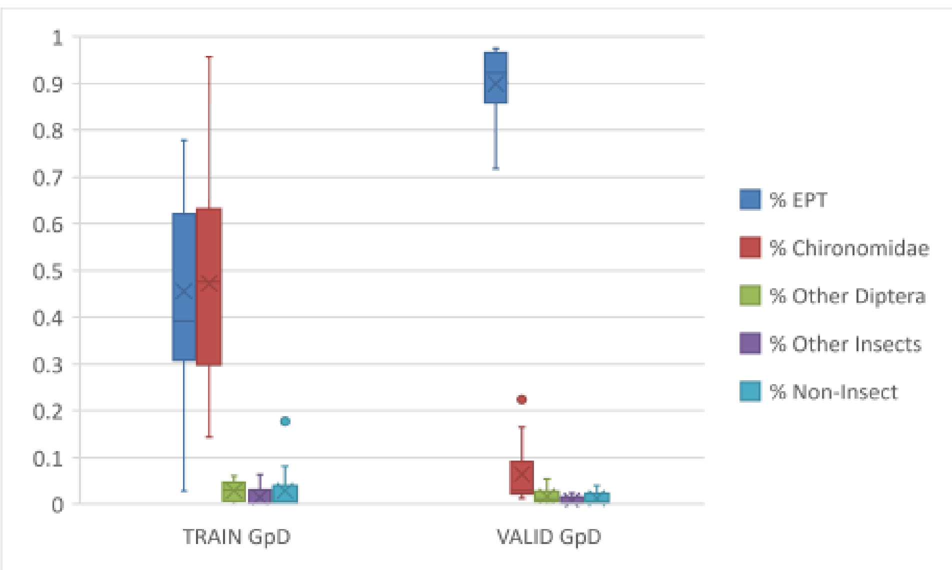

Figure 7. Comparison of average composition community metrics of Group D from the Training dataset (N=13) and those that were predicted to Group D from the Validation dataset (N=10).

A balance of Type I and Type II error rates for any assessment must be made. The standard ellipses for CABIN assessments (90%, 99% and 99.9%) define 4 assessment bands representing Similar to Reference, Mildly Divergent, Divergent, and Highly Divergent. Using the 90% threshold as the first assessment threshold, the Type I error rate for this model is very high (avg 59%) and the Type II error rate is very low (avg 5%), a rather extreme imbalance. A 95% threshold may be a more balanced first trigger of potential impact (i.e., Mildly Divergent) with an average Type I error rate of 39% and an average Type II error rate of 10.7%. At the 99% threshold, the next trigger for management action, the Type I and Type II error rates are slightly more balanced, with an average of 16% Type I error and 35.3% Type II error. It is at this level where test sites are assessed by CABIN as Divergent from reference condition and where one would most likely expect management action.

CONCLUSIONS AND RECOMMENDATION FOR MODEL USERS

- This model revision excludes the Okanagan basin that was previously included in the preliminary model (Gaber 2012). This model also includes a significant addition of new reference site data across the basin and encompasses the temporal variation observed between 2003 and 2018. With the available data, it was possible to divide the dataset into a training dataset to build the model and a validation dataset to test the model, which was not done previously.
- Six reference groups were identified in the Columbia Basin:
    1. Group D: Low abundance with a high proportion of Chironomidae. These sites tend to be steeper and narrower channels in smaller stream orders with smaller substrates and watersheds with lower sedimentary rock in the bedrock geology.
    2. Group E: Low abundance with a high proportion of EPT taxa.
    3. Group F: High abundance with a high richness and high proportion of Ephemeroptera taxa
    4. Group G: High abundance with a high proportion of EPT taxa.
    5. Group IJK: Very high abundance of all insect taxa, particularly Ephemeroptera.
    6. Group LM: Low abundance and greater proportion of non-insects relative to other groups.
- Columbia Basin model predictors include 4 variables measured on site (longitude, altitude, channel slope and canopy coverage) and 8 landscape level variables calculated after GPS coordinates are known (Bedrock Geology-Sedimentary%; Climate-Precipitation OCT and Temperature DEC Min; Hydrology-Drainage Area; Landcover-Grassland%, Shrub Low% and Water%; Topography-Max %Slope in the upstream watershed).
- The average DFA cross validation (jackknifed) classification success was 55% and the average resubstitution classification success was 60%. These are similar to other models uploaded to the CABIN web application.
- Caution is recommended with any test site predicted to Group D. The Group D classification success rate was lower than the recommended minimum standard for CABIN models. The biological communities overlap with Group E and F and are biologically variable. Given its small sample size (n=13), it is possible that the reference condition is not adequately described and therefore not well predicted. We recommend additional reference sites to get a more precise description of the natural variation of Group D.
- The validation dataset was used to assess Type I errors and was also *SIMPACTED* in three different ways to evaluate the kinds of disturbances the model can detect and Type II errors.
- On average, the model is able to detect decreases in taxa richness easily. It has a more difficult time detecting changes in communities that are related to abundance with little change to taxa richness, particularly with Group D and Group IJK.
- Based on Type I and Type II error rates established from a BC specific resource development tolerance score, it is recommended that the first (Mildly Divergent) ellipse for assessment using CABIN tools be modified from 90% to 95% confidence level in order to better balance Type I and Type II errors at the first assessment threshold. This recommendation assumes the functionality is eventually possible in the CABIN analytical tools. Until then, the user is cautioned that the 90% level results in high Type I errors.

References

Bailey, R.C., T.B. Reynoldson, A.G. Yates, J.L Bailey, and S. Linke. 2007. Integrating stream bioassessment and landscape ecology as a tool for land use planning. Freshwater Biology, 52: 908-917.

Carter J.L., Fend S.V. & Kennelly S.S. 1996. The relationships among three habitat scales and stream benthic invertebrate community structure. Freshwater Biology, 35, 109–124.

Church, M. 2002. Geomorphic thresholds in riverine landscapes. Freshwater Biology, 47: 541-557.

Collier, K.J. 2008. Temporal patterns in the stability, persistence and condition of stream macroinvertebrate communities: relationships with catchment land-use and regional climate. Freshwater Biology, 53: 603-616.

Corkum, L.D. 1989. Patterns of benthic invertebrate assemblages in rivers of northwestern North America. Freshwater Biology, 21: 191-205.

Davies, P. E. (Ed.) 1994. Monitoring River Health Initiative. River Bioassessment Manual. National River Processes and Management Program.(Freshwater Systems: Tasmania.)

Gaber, L. 2012. A predictive model for bioassessment of streams in coastal British Columbia using the reference condition approach: 2011. Report prepared by Water Protection and Sustainability Branch, B.C. Ministry of Environment, Victoria, B.C. 28p plus appendices.

Gutiérrez-Fonseca, P. E., Ramírez, A., & Pringle, C. M. (2018). Large-scale climatic phenomena drive fluctuations in macroinvertebrate assemblages in lowland tropical streams, Costa Rica: The importance of ENSO events in determining long-term (15y) patterns. *PloS one*, *13*(2), e0191781. https://doi.org/10.1371/journal.pone.0191781

Marchant, R., A. Hirst, R. H. Norris, R. Butcher, L. Metzeling, and D. Tiller. 1997. Classification and Prediction of Macroinvertebrate Assemblages from Running Waters in Victoria, Australia. Journal of the North American Benthological Society 16(3):664-681.

Minshall, W. 2003. Responses of stream benthic macroinvertebrates to fire. Forest Ecology and Management. 178:155-161.

Namayandeh, A. & R. Quinlan. 2011. Benthic Macroinvertebrate Communities in Arctic Lakes and Ponds of Central Nunavut, Canada, Arctic, Antarctic, and Alpine Research, 43(3):417-428.

Norris, S. 2012. British Columbia's Provincial Stream Biomonitoring Program Technical Documentation: GIS Tools for Reference Site Selection and Upstream Watershed Analysis. Draft report prepared by Hillcrest Geographics for Water Protection and Sustainability Branch, Ministry of Environment, Victoria, BC.

Rosenberg, D.M., Reynoldson, T.B., and Resh, V.H. 1999. Establishing reference conditions for benthic invertebrate monitoring in the Fraser River catchment, British Columbia, Canada. FRAP Report No. DOE-FRAP 1998-32. Fraser River Action Plan, Environment Canada, Vancouver, B.C.

Sandin, L., & Johnson, R. K. (2004). The importance of local and regional factors for the macroinvertebrate community structure in Swedish streams. Landscape Ecology, 19, 501–514. https://doi.org/10.1023/B:LAND.0000036116.44231.1c

Scotti, A, Füreder, L, Marsoner, T, Tappeiner, U, Stawinoga, AE, Bottarin, R. Effects of land cover type on community structure and functional traits of alpine stream benthic macroinvertebrates. *Freshwater Biology*. 2020; 65: 524– 539. https://doi.org/10.1111/fwb.13448

Scrine, J., M. Jochum, J.S. Ólafsson, and E.J. O'Gorman, 2017. Interactive effects of temperature and habitat complexity on freshwater communities. Ecol Evol. 7: 9333– 9346 https://doi.org/10.1002/ece3.3412

Strachan, S. and S. Bennett 2018. Reference model supporting documentation for CABIN Analytical Tools: Okanagan 2017. Prepared for CABIN web application by Environment and Climate Change Canada and Limnotek Research and Development Inc. 23 pg https://cabin-rcba.ec.gc.ca/cabin/login?culture=en-CA

U.S. Environmental Protection Agency (EPA). (2008) Climate change effects on stream and river biological indicators: A preliminary analysis. Global Change Research Program, National Center for Environmental Assessment, Washington, DC; EPA/600/R-07/085. Available from the National Technical Information Service, Springfield, VA, and online at http://www.epa.gov/ncea.

Vinson, M.R. and C.P. Hawkins. 2003. Broad-scale geographical patterns in local stream insect genera richness. Ecography, 26: 751-767. doi:10.1111/j.0906-7590.2003.03397.

Yates, A. and R.C. Bailey. 2010. Selecting objectively defined reference sites for stream bioassessment programs. Environmental Monitoring and Assessment 170:129-140.

Zimmermann, E. and R. Death. 2002. Effect of substrate stability and canopy cover on stream invertebrate communities. New Zealand Journal of Marine and Freshwater Research. 36. 537-545.

APPENDIX 1: DATA COLLECTION, ANALYSIS AND QUALITY ASSURANCE

A. Field Collection

CABIN Study Name	**BC MOE-Kootenay Region**	**EC-Columbia Basin**	**EC-Fed/Prov WQ Monitoring Stations**	**National Parks-Mountain Parks Biomonitoring**	**BC NGO-Nature Conservancy of Cda Darkwoods Aquatic Ass & Mon Prgm**
Agencies involved	BC Ministry of Environment and Climate Change Strategy	Environment and Climate Change Canada	Environment and Climate Change Canada	Parks Canada	Salmo Stream Keepers
Date range	2007-2018	2007-2018	2003-2018	2007-2018	2008-2009
Sampling season	Early Sept - Early Oct	Mid-Sept – Mid-Oct	Mid-Sept - Early Oct	Late Sept – Early Nov	October
# reference samples	73	68	38	70	9
Certified samplers (Y or N)	Y	Y	Y	Y	Y
Certified team leader (Y or N)	Y	Y	Y	Y	Y
400 um kicknet (Y or N)	Y	Y	Y	Y	Y
Preservative used	Ethanol	Formalin	Formalin	Formalin	Ethanol

CABIN Study Name	**BC-Beaver River Project - Masse**	**BC-Bingay - Masse**	**CBWQ-St. Mary**	**CBWQ-Upper Columbia**	**CBWQ-Windermere**
Agencies involved	Masse Environmental Consultants	Masse Environmental Consultants	St. Mary's Rural residents association	Wildsight Golden	Lake Windermere Ambassadors
Date range	2007-2008	2010-2016	2010-2014	2010-2014	2009-2014
Sampling season	Late Sept - Mid Oct	Mid-Sept – Mid-Oct	Late Sept - Early Oct	Mid-Late Sept	Late Aug - Early Nov
# reference samples	10	3	4	4	10
Certified samplers (Y or N)	Y	Y	Y	Y	Y
Certified team leader (Y or N)	Y	Y	Y	Y	Y
400 um kicknet (Y or N)	Y	Y	Y	Y	Y
Preservative used	Ethanol	Ethanol	Ethanol	Ethanol	Ethanol

B. Macroinvertebrate Identification

CABIN Study Name	**BC MOE-Kootenay Region**	**EC-Columbia Basin**	**EC-Fed/Prov WQ Monitoring Stations**	**National Parks-Mountain Parks Biomonitoring**	**BC NGO-Nature Conservancy of Cda Darkwoods Aquatic Ass & Mon Prgm**
Taxonomist	Cordillera Consulting	EcoAnalysts and Cordillera Consulting	EcoAnalysts and Cordillera Consulting	Cordillera Consulting, EcoAnlaysts and Living Streams	Lynn Wescott
Marchant Box used (Y or N)	Y	Y	Y	Y	Y
Subsample count	300	300	300	300	300
10% of reference samples sent to National Lab for QA	Y	Y	Y	Y	N
Reference Collection maintained	Y	Y	N	Y	N
CABIN Study Name	**BC-Beaver River Project - Masse**	**BC-Bingay - Masse**	**CBWQ-St. Mary**	**CBWQ-Upper Columbia**	**CBWQ-Windermere**
Taxonomist	Cordillera Consulting	Cordillera Consulting	Eco Analsyts (pre-2013) and Pina Viola (2013-2014)	Eco Analsyts (pre-2013) and Pina Viola (2013-2014)	Eco Analsyts (pre-2013) and Pina Viola (2013-2014)
Marchant Box used (Y or N)	Y	Y	Y	Y	Y
Subsample count	300	300	300	300	300
10% of reference samples sent to National Lab for QA	N	N	N	N	N
Reference Collection maintained	N	N	N	N	N

The data collection goes back to the early years of CABIN before the laboratory processing protocol and guidance for taxonomists was well established. As a result several taxa were identified and entered into CABIN that are not included in the development of a model. The excluded taxa are listed in Table A1-1 with the rationale for excluding them based on the CABIN protocol.

Table A1-1. Taxa exported from CABIN studies that were excluded from model analysis

Taxon	***Rationale***
Bosminidae	cladoceran (pelagic)
Candonidae	ostracod (too small for 400 kicknet reliably)
Chydoridae	ostracod (too small for 400 kicknet reliably)
Cyclocyprididae	ostracod (too small for 400 kicknet reliably)
Cyclopidae	copepoda (pelagic)
Cyprididae	ostracod (too small for 400 kicknet reliably)
Cytherideidae	ostracod (too small for 400 kicknet reliably)
Daphniidae	cladoceran (pelagic)
Hydridae	Colonial
Lumbricidae	terrestrial
Macrobiotidae	Tardigrada (pelagic)
Macrothricidae	cladoceran (pelagic)
Planariidae	in protocol to not count (therefore inconsistently analysed)

The CABIN database includes a linkage to the Integrated Taxonomic Information System (ITIS) to ensure consistency in nomenclature. It is well known among taxonomic experts that updates to taxonomy are delayed therefore some important updates must be acknowledged that are not yet current in ITIS. Tubificidae is now recognized as a subfamily of Naididae. Due to the delay in ITIS updates, it is flagged as "unverified" in the CABIN database.

C. GIS Analyses

All GIS data were generated by Adam Yates (University of Western Ontario). Watersheds were delineated using ArcGIS 10 ArcHdyro 2.0 (ESRI 2010). Delineations were based on 20 m resolution digital elevation models (DEM) and a 1:50,000 scale hydrological network. The DEM was subjected to pre-processing which "burned in" the stream network into the DEM and filled sinks to improve flow modeling. The corrected DEM was used to calculate flow direction and flow accumulation to carry out the terrain procession steps to model catchment areas (AcrHydro 2010). The delineated catchments were described using the GIS layers in the table below collected from publicly available sources.

Descriptor	**Scale/ Resolution**	**Source and method**
Basin Morphometry	*20 m*	*Area and perimeter were calculated from delineated catchments as described above*
Bedrock	*1:100,000*	*BC Ministry of Energy and Mines – BC Digital Geology Maps 2005 - http://www.empr.gov.bc.ca/Mining/Geoscience/PublicationsCatalogue/DigitalGeologyMaps/Pages/default.aspx* *Intersected with catchment boundaries using intersect function in ArcGIS (ESRI 2010)*
Climate	*7.5 km*	*Natural Resources Canada (contact: Dan McKenney – dan.mckenney@nrcan-rncan.gc.ca)* *Summarized using rasterized grids describing temperature normals from 1971-2000 giving long term monthly and annual averages of temperature and precipitation. Grids were used to generate average, minimum and maximum values for each catchment using Geospatial Modelling Environment v. 0.6.0.0 (Beyer 2012). Where catchments were completely contained within one grid cell, catchments were assigned the value of that cell.*
Hydrology	*1:50,000*	*www.geobase.ca – National Hydro Network* *Intersected with catchment boundaries using intersect function in ArcGIS (ESRI 2010)*
Land Use	*1:2,000,000*	*www.geobase.ca – Land Cover* *Intersected with catchment boundaries using intersect function in ArcGIS (ESRI 2010)*
Topography	*20 m*	*www.geobase.ca – Digital Elevation Data* *Described using 20 m DEM and the Geospatial Modeling Environment v. 0.6.0.0 (Beyer 2012) to describe the maximum and minimum elevation in each catchment. Percent slope was generated from the DEM using the slope function in ArcGIS (ESRI 2010) and classified into one of four groups based on the slope value for each grid cell (i.e. <30%, 30-50%, 50-60%, >60%). Areas of each class within each catchment were then calculated.*

D. Laboratory Analyses

Laboratory analyses for water quality samples are stored in CABIN but are not used as predictors in the development of the model. The laboratories and methods used varied for each CABIN study.

E. Statistical Analyses

Several software packages were used in the development of the Columbia model.

1. Excel- data manipulation and storage
2. PATN V.3.12 - classification and ordination of test sites for assessment
3. PRIMER 6 - classification, MDS ordination, ANOSIM, SIMPER
4. SYSTAT 13 - discriminant analysis and plotting BEAST assessments with probability ellipses