Murray-Stoker, D, and K. Murray-Stoker. Consistent metacommunity structure despite inconsistent drivers of assembly at the continental scale.

Data files and R script for all analyses described in the main text.

Authors
David Murray-Stoker
Department of Ecology and Evolutionary Biology
University of Toronto, Toronto
Ontario M5S 3B2, Canada
E-mail: dstoker92@gmail.com

Kelly Murray-Stoker
Department of Ecology and Evolutionary Biology
University of Toronto, Toronto
Ontario M5S 3B2, Canada
E-mail: kmmuray14@gmail.com

Author Notes

The provided data were derived from the National Rivers and Streams Assessment (NRSA) conducted by the United States Environmental Protection Agency during the summers of 2008-2009 (USEPA 2016). Although the data are provided as a supplement to the manuscript to allow for re-analysis by readers and a resource for reviewers and examiners, any re-use for further publication should cite the original source (USEPA 2016).

We removed excess columns containing data unrelated to the aim of the manuscript, and some columns were renamed after careful examination of the metadata provided by the USEPA. We renamed columns due to personal coding preferences (e.g., primarily lowercase column names, more informative and intuitive variable names) and to keep track of variables included in function calls; this had the added benefit of understanding analysis summaries without having to frequently consult the USEPA metadata.

Citation for the data:
U.S. Environmental Protection Agency (USEPA). (2016). National Rivers and Streams Assessment 2008-2009 (siteinfo_0.csv, land.csv, phabmed.txt, nrsa0809bentcts.csv, and corresponding metadata files). Available from U.S. EPA web page: https://www.epa.gov/national-aquatic-resource-surveys/data-national-aquatic-resource-surveys. Date accessed: 2017-06-21

File List
NRSA-landscape.csv
NRSA-physical_habitat.csv
NRSA-water_chemistry.csv
NRSA-invertebrates

NRSA-landscape.csv

Landscape data for the sites survey during the NRSA 2008-2009. Descriptions are derived from the metadata provided by the USEPA for each data file. Original variable names are provided in square brackets. Data were collated from land.csv.

UID = unique site visit ID [UID]
mean.annual.flow = mean annual flow in cubic feet per second (cfs) at bottom of flowline as computed by Unit Runoff Method [MAFLOWU]
dam.count = count of dams located within NHDPlus 10 km navigated subcatchment [NHD10_DAMS_CNT]
dam.density = density of dams located within NHDPlus 10 km navigated subcatchment [NHD10_DAMS_DEN]
basin.area = area of NHDPlus navigated cumulative basin in square kilometers [NHDWAT_AREA_SQKM]
max.basin.elevation = maximum elevation value in meters (NED 30m resolution) within NHDPlus navigated cumulative basin [NHDWAT_NEDELEV_MAX]
mean.basin.elevation = mean elevation value in meters (NED 30m resolution) within NHDPlus navigated cumulative basin [NHDWAT_ELEV]
min.basin.elevation = minimum elevation value in meters (NED 30m resolution) within NHDPlus navigated cumulative basin [NHDWAT_NEDELEV_MIN]
range.basin.elevation = ange of elevation values in meters (NED 30m resolution) within NHDPlus navigated cumulative basin [NHDWAT_NEDELEV_RANGE]
pct.canopy.cover = mean percent tree canopy (NLCD2001 tree canopy) within NHDPlus navigated cumulative basin [NHDWAT_PCT_CANOPY]
pct.ISC = mean percent impervious surface (NLCD2006 Impervious Surface) within NHDPlus navigated cumulative basin [NHDWAT_PCT_IMPERV]
tmax.annual = average maximum annual temperature in degrees Celsius (PRISM 1997-2000) at the sample site [TMAX_ANN]
tmax.july = average maximum July temperature in degrees Celsius (PRISM 1997-2000) at the sample site [TMAX_JULY]
tmin.annual = average minimum annual temperature in degrees Celsius (PRISM 1997-2000) at the sample site [TMIN_ANN]
tmin.july = average minimum July temperature in degrees Celsius (PRISM 1997-2000) at the sample site [TMIN_JULY]
pct.ag = % of NHDPlus navigated cumulative basin area classified as Pasture/Hay (81) or Row Crops (82) in NLCD
pct.urb = % of NHDPlus navigated cumulative basin area classified as either Low Intensity Residential (21), High Intensity Residential (22), Commercial/Industrial/Transportation (23), or Developed High Intensity (24) in NLCD
pct.for = % of NHDPlus navigated cumulative basin area classified as Deciduous Forest (41), Evergreen Forest (42), or Mixed Forest (43) in NLCD
pct.wet = % of NHDPlus navigated cumulative basin area classified as Woody Wetlands (90), or Emergent Herbaceous Wetlands (95) in NLCD
pct.shrub = % of NHDPlus navigated cumulative basin area classified as Shrub/Scrub (52) or Grasslands/Herbaceous (71) in NLCD

NRSA-physical_habitat.csv

Physical habitat data for the sites survey during the NRSA 2008-2009. Descriptions are derived from the metadata provided by the USEPA for each data file. Original variable names are provided in square brackets. Data collated from siteinfo_0.txt and phablow.txt.

UID = unique site visit ID [UID]
site.ID = site identification code [SITE_ID]
site.lat = X-site GPS latitude decimal degrees [XLAT_DD]
site.long = X-site GPS longitude decimal degrees [XLON_DD]
strahler.order = Strahler stream order from RF3 stream data
ecoregion = NARS 9-level reporting region (2015), based on aggregated Omernik Level III ecoregions: CPL = Coastal Plains; NAP = Northern Appalachians; NPL = Northern Plains; SAP = Southern Appalachians; SPL = Southern Plains; TPL = Temperate Plains; UMW = Upper Midwest; WMT = Western Mountains; XER = Xeric West [AGGR_ECO9_2015]
HUC8 = 8-digit HUC catalog unit number [HUC8]
reach.length = sample reach length (m) [REACHLEN]
bed.stability = $\log_{10}$(Streambed Critical Diameter-at Bankfull - mm) [LDMB_BW5]
pct.bedrock = bed surface % bedrock [PCT_BDRK]
pct.bigrock = bed surface % larger than gravel (64mm) [PCT_BIGR]
pct.hardpan = bed surface % hardpan [PCT_HP]
pct.fine = bed surface % fines < 0.06mm [PCT_FN]
pct.sand = bed surface percent sand or smaller (< 2.0mm) [PCT_SAFN]
pct.fine.gravel = bed surface fine gravel or smaller (< 16mm) [PCT_SFGF]
pct.coarse.gravel = bed surface coarse gravel or smaller (< 64mm) [PCT_SFG]
pct.riffle = percent fast water habitat [PCT_FAST]
pct.pool = percent of reach with pool or glide habitat [PCT_SLOW]
mean.thalweg.depth = mean thalweg depth (cm), converted from raw data in m [XDEPTH_CM]
mean.width = mean wetted width (m) [XWIDTH]
sinuosity = sinuosity of sample reach [SINU]
LWD.reach = wood volume (m^3) per 100m channel [LWDEQVOLM100]
embeddedness = mean streambed embeddedness (%) [EMBED]
ALG.cover = filamentous algae Mean areal cover [XFC_ALG]
AQM.cover = aquatic macrophyte mean areal cover [XFC_AQM]
LWD.cover = large woody debris areal cover [XFC_LWD]
NAT.cover = sum of non-anthropogenic fish areal cover types [XFC_NAT]

NRSA-water_chemistry.csv

Water chemistry data for the sites survey during the NRSA 2008-2009. Descriptions are derived from the metadata provided by the USEPA for each data file. Original variable names are provided in square brackets. Data were collated from chem.txt.

UID = unique site visit ID [UID]
Al = total dissolved aluminum mg/L [AL]
ANC = acid neutralizing capacity ueq/L [ANC]
Ca = calcium mg/L [CA]
Cl = chloride mg/L [CL]
cond = specific conductance µS/cm [COND]
DOC = dissolved organic carbon mg/L [DOC]
K = potassium mg/L [K]
Mg = magnesium mg/L [MG]
sodium = sodium mg/L [SODIUM]
NH4 = ammonia as nitrogen mg/L [NH4]
NO2 = nitrite as nitrogen mg N/L [NO2]
NO3 = nitrate as nitrogen mg N/L [NO3]
total.N = total nitrogen µg/L [NTL]
pH.lab = laboratory measured pH [PHLAB]
total.P = total phosphorous µg/L [PTL]
SiO2 = silica mg/L [SIO2]
SO4 = sulfate mg/L [SO4]
TSS = total suspended solids mg/L [TSS]
turb = turbity ntu [TURB]

NRSA-invertebrates.csv

Invertebrate data for the sites survey during the NRSA 2008-2009. Data were transformed in R using the reshape2 package in order to have site-by-taxa community matrix: rows represent sites and each column represents individual taxa. Taxonomic names are all exactly as provided in the "TAXA_ID" column in the original data, except in lowercase instead of uppercase. Renaming of taxa (i.e., making all names lowercase) was done in Microsoft Excel.

The site-by-taxa matrix was formatted by using the following code in R:

```
require(reshape2)
site.by.taxa.matrix <- dcast(nrsa0809bentctsmet.csv, UID ~ TARGET_TAXON, sum,
                             value.var = "TOTAL300")
```

Data were collated from nrsa0809bentctsmet.txt.