Assessing and understanding in-stream habitat suitability for invertebrate families, fish species and platypus of the Melbourne Region with the melbstreambiota package (version 0.1.2)

Christopher J. Walsh, Yung En Chee, Nick Bond, and Rhys Coleman

6 Apr 2023

Introduction

The models of macroinvertebrate family distributions among streams of the Melbourne region (south-eastern Australia) described by Walsh & Webb (2013) and Walsh & Webb (2016) (later refined by Walsh (2023)) allow prediction of occurrence of 59 macroinvertebrate families in all stream reaches of the region. Associated work developed similar habitat suitability models for 11 native fish species (Chee *et al.* 2020) and for platypus in streams of the region. Three models for platypus can be used in this package:

- female platypus, used by Coleman *et al.* (2022) and Chee *et al.* (2020) to support Melbourne Water's Healthy Waterway Strategy (Melbourne Water 2018)

- all platypus, used by Chee *et al.* (2020).

- a revised version of the all platypus model used by Coleman *et al.* (2022).

The models permit predictions under current conditions (2006, reflecting the time data used to build the models was collected) and under a range of scenarios (either changes in human influence or change in climate) by adjusting the values of predictor variables. Walsh (2023) further demonstrated that the ensemble predictions of the 59 macroinvertebrate family models produce reliable predictions of family richness, and used the models to derive an index of stream condition, LUMaR, that provides improved sensitivity to human disturbance, more consistently across the region than two commonly used indices: SIGNAL and AUSRIVAS. LUMaR combines sensitivity weightings like those used by SIGNAL Chessman (2003) with an observed:expected ratio such as used by AUSRIVAS (Simpson & Norris 2000).

The melbstreambiota package provides functions that allow calculation and mapping of a) the environmental variables used as predictors in the models, b) predictions of occurrence of each macroinvertebrate family, fish species and platypus, and c) ensemble predictions of LUMaR score and other indices under different scenarios. The package also permits the input of macroinvertebrate data collected from sites in the Melbourne region to calculate observed LUMaR score, to compare against the predicted score (this task can also be achieved using the web-based app at https://urbanstreams.net/tools/LUMaR/. This vignette provides examples of the application of the melbstreambiota functions for assessing and understanding stream macroinvertebrate assemblages of the region.

The package can be installed in R from https://github.com/cjbwalsh/melbstreambiota (see instructions in README.md below the list of files).

Which taxa have been modelled?

The 59 modelled macroinvertebrate families and their codes used in all analyses in the package are listed in the table taxon.classes:

- 13 caddisfly families (Trichoptera: Leptoceridae, QT25; Calamoceratidae, QT24; Atriplectididae, QT23; Philorheithridae, QT21; Calocidae, QT18; Helicopsychidae, QT17; Conoesucidae, QT15; Ecnomidae, QT08; Hydropsychidae, QT06; Philopotamidae, QT04; Hydroptilidae, QT03; Glossosomatidae, QT02; Hydrobiosidae, QT01)

- 4 stonefly families (Plecoptera: Notonemouridae, QP04; Gripopterygidae, QP03; Austroperlidae, QP02; Eustheniidae, QP01)

- 5 mayfly families (Ephemeroptera: Caenidae, QE08; Leptophlebiidae, QE06; Coloburiscidae, QE05; Oniscigastridae, QE03; Baetidae, QE02)

- 6 dragonfly and damsefly families (Odonata: Corduliidae, QO16; Aeshnidae, QO12; Synlestidae, QO08; Megapodagrionidae, QO07; Lestidae, QO05; Coenagrionidae, QO02)

- 1 dobsonfly family (Megaloptera: Corydalidae, QM01)

- 3 'true' bug families (Hemiptera: Notonectidae, QH67; Corixidae, QH65; Veliidae, QH56)

- 8 beetle families (Coleoptera: Ptilodactylidae, QC39; Psephenidae, QC37; Elmidae, QC34; Scirtidae, QC20; Hydraenidae, QC13; Hydrophilidae, QC11; Gyrinidae, QC10; Dytiscidae, QC09)

- 9 'true' fly families (Diptera: Tanypodinae, QDAE; Podonominae, QDAD; Empididae, QD35; Athericidae, QD22; Simuliidae, QD10; Ceratopogonidae, QD09; Culicidae, QD07; Dixidae, QD06; Tipulidae, QD01)

- 4 crustacean families (Decapoda and Amphipoda: Atyidae, OT01; Paramelitidae, OP06; Pontogeneiidae, OP03; Ceinidae, OP02)

- 4 snail families (Gastropoda: Physidae, KG08; Planorbidae, KG07; Ancylidae, KG06; Lymnaeidae, KG05)

- a leech family (Glossiphoniidae, LH01), and a flatworm family (Dugesiidae, IF61).

The 10 native fish species and two native lamprey species combined in a single model, and their codes—together with codes for the three platypus models: females only, all platypus (Chee *et al.* 2020), and all platypus (Coleman *et al.* 2022)—are listed in the table vertSpp (an object loaded with the package) (Table 1).

Table 1. Vertebrate species for which models are included in the melbstreambiota package

vertcode	species	authority	commonName
ANGUAUST	*Anguilla australis*	Richardson, 1841	Southern short-finned eel
GADOMARM	*Gadopsis marmoratus*	Richardson, 1848	River blackfish
GALABREV	*Galaxias brevipinnis*	Günther, 1866	Climbing galaxias
GALAMACU	*Galaxias maculatus*	(Jenyns, 1842)	Common galaxias
GALAORN	*Galaxias ornatus*	Castelnau 1873	Mountain galaxias
GALATRUT	*Galaxias truttaceus*	Valenciennes, 1846	Spotted galaxias
LAMPREYS	*Geotria australis*	Gray, 1851	Pouch lamprey[a]
LAMPREYS	*Mordacia mordax*	(Linnaeus 1758)	Australian lamprey[a]
NANNAUST	*Nannoperca australis*	Günther, 1861	Southern pygmy perch
PHILGRAN	*Philypnodon grandiceps*	(Krefft, 1864)	Flathead gudgeon
PSEUURVI	*Pseudaphritis urvillii*	(Valenciennes, 1832)	Tupong
RETRSEMO	*Retropinna semoni*	(Weber, 1895)	Australian Smelt
femPlaty	*Ornithorhynchus anatinus*	(Shaw, 1799)	Female platypus
allPlatyHWS	*Ornithorhynchus anatinus*	(Shaw, 1799)	Total platypus[b]
allPlatyColeman	*Ornithorhynchus anatinus*	(Shaw, 1799)	Total platypus[c]

[a]Lamprey species combined into a single model
[b]All platypus model used by Chee et al. 2020
[c]All platypus model used by Coleman et al. 2022

Mapping predictor variables

The 59 boosted-regression-tree models (Elith, Leathwick & Hastie 2008) of macroinvertebrate families are saved as model-objects in the list "bestModelsBugfams", and the fish and platypus models are saved as "bestModelsVerts", both saved in a system file that the package uses. The models predict the probability of occurrence of each taxon in a pair of rapid bioassessment samples (a standard sampling method used in the region since the early 1990s). All 59 macroinvertebrate models use the same 10 predictor variables.

```
require(melbstreambiota)
exampleSamppr <- collateBugSamppr(mwstreams[1,], SRI_48mth_weighted = 1)
names(exampleSamppr)[-1]

## [1] "AttImp_L9"              "AttForest_L35W1000"
## [3] "meanAnnQ_mm"            "CatchmentArea_km2_InclDams"
## [5] "mnAnnAirTm_deg"         "CatIgneous"
## [7] "SRI_48mth_weighted"     "nspring"
## [9] "nriff"                  "processN"
```

These predictor variables constitute: + 2 human-impact variables — attenuated imperviousness (AI, or "AttImp_L9" in "mwstreams"), as described by Walsh & Kunapo (2009), and attenuated forest cover (AF, or "AttForest_L35W1000" in mwstreams), as described by Walsh & Webb (2014);

- 4 variables that indicate physiographic variation across the region — catchment area (CatchmentArea_km2_InclDams), mean annual discharge depth (meanAnnQ_mm), mean air temperature (mnAnnAirTm_deg) and proportion of catchment with igneous geology (CatIgneous);

- A single variable that indicates temporal variation in stream flow — linearly-weighted 48-month antecedent discharge as a fraction of the long-term mean discharge (SRI_48mth_weighted: see below), and;

- 3 variables that indicate sample characteristics — number of riffle samples (nriff: 0, 1, or 2), number of spring samples (nspring: 0, 1 or 2), sample processing method (processN 0 = lab-sorted; 1 = field-sorted).

For fish models, the same core variables were used, together with the number of partial and full barriers as additional human-impact variables.

```
#The following command is used to build a predictor table to run fish models, setting
#antecedent rainfall (SRI...) at the long term average, and simulating the removal of
#of all fish barriers.  (Only a single row is extracted in order to list the column names)
exampleFishSamp <- collateSampleFP(mwstreams[1,], SRI_48mth_weighted = 1,
                                   barriersFromYear = FALSE,
                                   PartBarriersDS = 0, FullBarriersDS = 0,
                                   FishOrPlatypus = "fish")
names(exampleFishSamp)[-1]

## [1] "AttImp_L9"           "AttForest_L35W1000"
## [3] "meanAnnQ_mm"         "CatchmentArea_km2_InclDams"
## [5] "mnAnnAirTm_deg"      "CatIgneous"
## [7] "SRI_48_triang"       "PartBarriersDS"
## [9] "FullBarriersDS"
```

Note differences in the names of SRI and barriers variables compared to those in mwstreams. These variables are equivalent, but different names were used in building the models. Note also that fish (and platypus) models do not have predictor variables describing sample characteristics. The platypus models do not use fish barriers as predictor variables, but do have additional human impact variables describing the quantity of vegetation (vegBank) and large woody debris (LWDBank). Furthermore, the platypus models use forest cover in a 10-m buffer from the stream and 1-km upstream (AFb10L1000) rather than the exponentially weighted forest cover metric used by other models (AttForest_L35W1000).

```
#The same function can be used to build predictor tables for the platypus models

examplePlatSamp <- collateSampleFP(mwstreams[1,], SRI_48mth_weighted = 1,
                                   FishOrPlatypus = "platypus")

names(examplePlatSamp)[-1]

## [1] "AttImpMin4k_L9"      "AFb10L1000"
## [3] "meanAnnQ_mm"         "CatchmentArea_km2_InclDams"
## [5] "mnAnnAirTm_deg"      "CatIgneous"
## [7] "SRI_48_triang"       "LWDBank"
## [9] "vegBank"
```

An important criterion in the selection of predictor variables for the models was that they should be generalizable to all reaches in the region, thus allowing prediction of distributions in all reaches. The table "mwstreams" is saved as a data frame with the values of all predictor variables (and several others used for other models) for each of 8256 reaches in the region. This table can be joined to the simple features object "mw_stream_map" to create graduated maps indicating variation in each variable across the region using the function plotMWstreamsByVar (Fig. 1 shows 4 examples). SRI_48mth_weighted values for every subcatchment for every month from December 1984 are stored in the data frame sri48moW. This table is stored in the repository at https://osf.io/mcxrq/ (as sri48moW.rda), and updated annually. It should therefore contain SRI_48mth_weighted values until previous December in any year. The package saves this table to the user's cache, which may become out of date: check with this

command: `max(melbstreambiota::sri48moW$date)`. To update the table in your cache
run the function `update_sri.rda`.

```
par(mfrow = c(2,2))
plotMWstreamsByVar(mwstreams$meanAnnQ_mm, varName = "meanQ", nbreaks = 8, style = "fixed",
                   fixedBreaks = c(0,10,50,100,150,200,250,400,1000), legend.cex = 0.5)
title(main = "  A", adj = 0, line = -1.5)
plotMWstreamsByVar(mwstreams$AttImp_L9, varName = "AI", nbreaks = 7, style = "fixed",
                   fixedBreaks = c(0,0.001,0.005,0.01,0.02,0.05,0.1,0.7), rev = TRUE,
                   legend.cex = 0.5)
title(main = "  B", adj = 0, line = -1.5)
plotMWstreamsByVar(mwstreams$AttForest_L35W1000, varName = "AF", nbreaks = 5,
                   style = "fixed", fixedBreaks = seq(0,1,0.2), legend.cex = 0.5)
title(main = "  C", adj = 0, line = -1.5)
plotMWstreamsByVar(mwstreams$mnAnnAirTm_deg, varName = "AirTemp", nbreaks = 8,
                   style = "fixed", fixedBreaks = 8:16, rev = TRUE, legend.cex = 0.5)
title(main = "  D", adj = 0, line = -1.5)
```
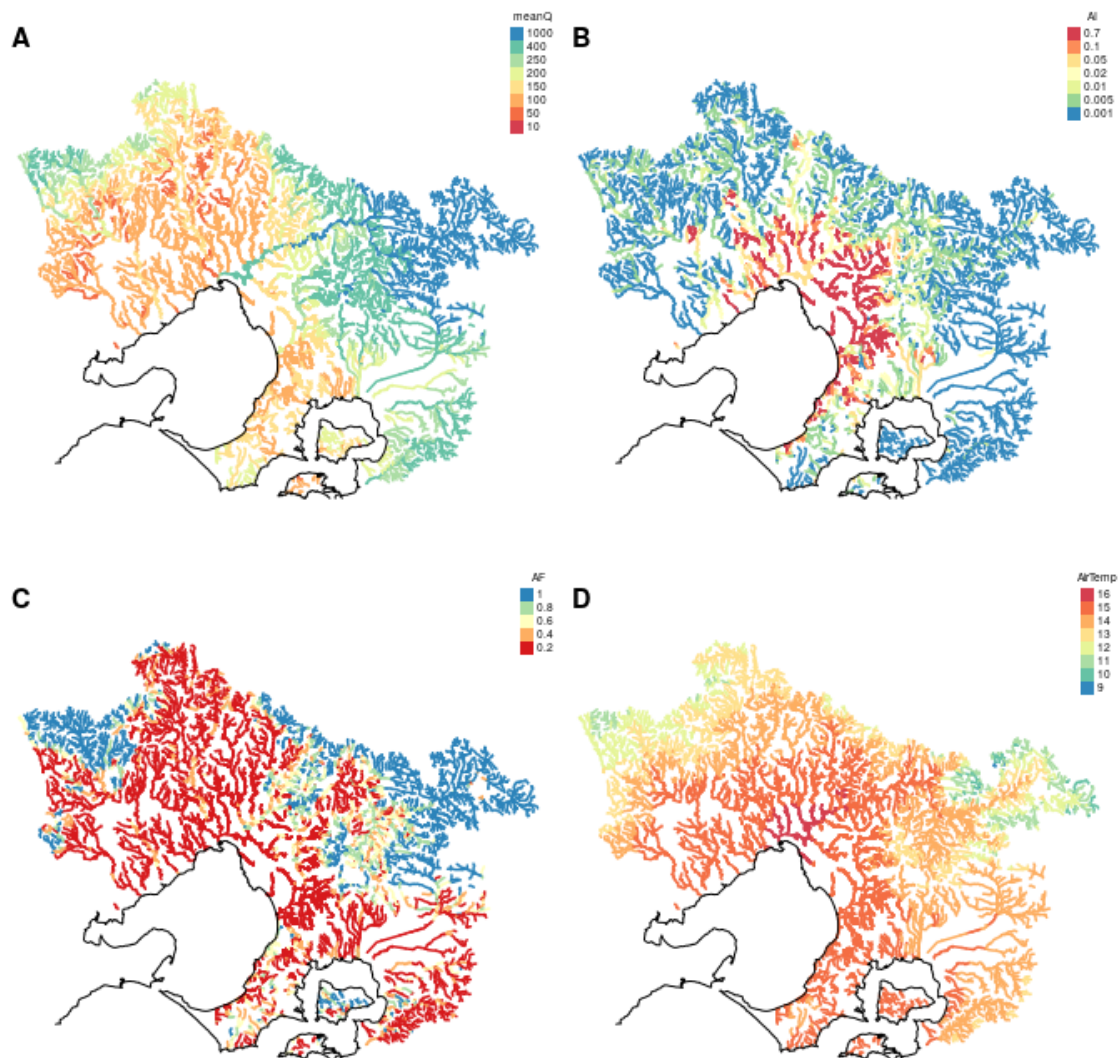

Fig. 1. Variation of 4 influential predictor variables across the stream network of the Melbourne region. A. meanQ (Mean annual discharge depth in mm), B. AI (proportion attenuated imperviousness), C. AF (proportion attenuated forest cover), D. MearAirT (Mean air temperature in deg C)

Mean Annual discharge depth (Fig. 1A) was commonly a strong predictor of distributions, and is indicative of the strong climatic gradient across the Melbourne region (Walsh & Webb 2014), although large rivers (such as the Yarra flowing from the east) transport high discharges through drier areas. AI (Fig. 1B) shows AI, which is highest in the small streams of the Melbourne Metropolitan area on small streams: it is noteworthy that two large rivers (The Yarra from the east and the Maribyrnong from the north) remain at low levels of AI as they flow into the metropolitan area. AF (Fig. 1C), shows the distribution of agriculture (areas of low AF) around the city, and the reservation of forested areas (high AF) in upland areas to the north-west and to the east. Mean air temperature is primarily driven by elevation: it is used in the models because parallel work has shown that air temperature in combination with the other physiographic and human impact variables used in the models is a strong predictor of stream temperature.

Alternatively, a subset of the full stream network can be drawn by providing a subset of subcs to be drawn. Two methods for deriving subsets are:

1. Use the subc names as a criterion for subsetting. The first 4 characters of each subc indicate one of 23 subcatchments (Table 2). All reaches in any one of these 23 subcatchments can be selected using the grep function (as shown in Fig. 2a for the Maribyrnong catchment).

2. Select all reaches upstream of a known subc, using the list of all upstream subcs in the package data object streamsubcs.allus2017.rda. Locations of subcatchments can be determined using quickMapR as described below. Fig.2 b shows the streams of the Merri Creek catchment upstream of subc YARR8590.

Table 2. subc codes begin with 4 characters specifying which of the 23 catchments/catchment groups they fall in.

Code	Catchment
BASS	Bass R
BUNY	Bunyip R
CARD	Cardinia Ck
DAND	Dandenong Ck
DEEP	Deep Ck (KooWeeRup Swamp)
ELST	Elster Ck
FREN	French Island streams
KANA	Kananook Ck
KORO	Kororoit Ck
LANG	Lang Lang R
LAVE	Laverton Ck
LITT	Little Ck
MAIN	Main Ck (Mornington Peninsula)
MARI	Maribyrnong R
MORD	Mordialloc Ck
MORP	Mornington Peninsula streams draining to Port Phillip
MORW	Mornington Peninsula streams draining to Westernport
SKEL	Skeleton Ck
TOOM	Toomuc Ck
WERR	Werribee R
WEST	Phillip Island streams
YALL	Yallock Ck
YARR	Yarra R

```
par(mfrow = c(1,2))
#Select Maribyrnong reaches by the catchment identifier in the subc names
mariSubcs <- mwstreams$subc[grep("MARI",mwstreams$subc)]
#Select Merri Creek reaches by identifying all reaches upstream of the most downstream reac
h "YARR8590"(Idenfitied by checking the map in GIS.
merriSubcs <- streamsubcs.allus2017$YARR8590
plotMWstreamsByVar(mwstreams$CatchmentArea_km2_InclDams, varName =
                    "Catchment area (sq km)", nbreaks = 8, style = "quantile",
                    legend.cex = 0.5, subcSubset = mariSubcs)
title(main = "  A. Maribyrnong River", adj = 0, line = -1.5)
plotMWstreamsByVar(mwstreams$CatchmentArea_km2_InclDams, varName =
                    "Catchment area (sq km)", nbreaks = 8, style = "quantile",
                    legend.cex = 0.5, subcSubset = merriSubcs)
title(main = "  B. Merri Ck", adj = 0, line = -1.5)
```
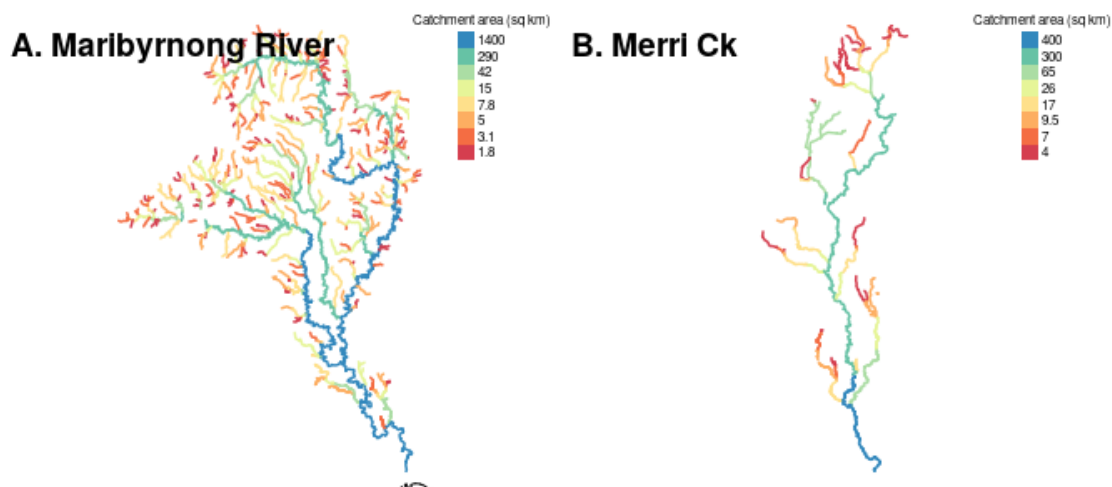

Fig. 2. Variation in catchment area for A) the Maribyrnong River catchment and B) the Merri Creek catchment

Predicting and mapping taxon distributions

The environmental variables in "mwstreams" permit prediction of taxon distributions under 2006 conditions and, with some manipulation, under alternative scenarios. Predictions can be made with the "bugModPred1" function, which requires a bugcode and a table of reaches, each with values for the 10 predictor variables of the models. Walsh & Webb (2016) classified families into a several sensitivity classes, and these (as modified by Walsh (2023)) are listed for each taxon in the taxon.classes table. This table also lists the bugcodes and full name for each family, as well as other details. The sensitivity classes are well illustrated by predicted distribution maps under 2006 conditions (Fig. 3).

```
QE06Pred <- bugModPred1("QE06", mwstreams)  #class A, a very sensitive family
QT25Pred <- bugModPred1("QT25", mwstreams)  #class B, a moderately sensitive family
QE08Pred <- bugModPred1("QE08", mwstreams)  #class D, sensitive to urban stormwater,
                                            #positively associated with forest loss
IF61Pred <- bugModPred1("IF61", mwstreams)  #class weedy
par(mfrow = c(2,2))
plotMWstreamsByVar(QE06Pred$pred1,
                varName = paste(taxon.classes$family[taxon.classes$fam == "QE06"],
                        "prob of occurrence"),
                nbreaks = 5, style = "fixed",
                fixedBreaks = seq(0,1,0.2), legend.cex = 0.5)
title(main = "  A", adj = 0, line = -1.5)
```

```
plotMWstreamsByVar(QT25Pred$pred1,
                   varName = paste(taxon.classes$family[taxon.classes$fam == "QT25"],
                                   "prob of occurrence"),
                   nbreaks = 5, style = "fixed",
                   fixedBreaks = seq(0,1,0.2), legend.cex = 0.5)
title(main = "  B", adj = 0, line = -1.5)
plotMWstreamsByVar(QE08Pred$pred1,
                   varName = paste(taxon.classes$family[taxon.classes$fam == "QE08"],
                                   "prob of occurrence"),
                   nbreaks = 5, style = "fixed",
                   fixedBreaks = seq(0,1,0.2), legend.cex = 0.5)
title(main = "  C", adj = 0, line = -1.5)
plotMWstreamsByVar(IF61Pred$pred1,
                   varName = paste(taxon.classes$family[taxon.classes$fam == "IF61"],
                                   "prob of occurrence"),
                   nbreaks = 5, style = "fixed",
                   fixedBreaks = seq(0,1,0.2), legend.cex = 0.5)
title(main = "  D", adj = 0, line = -1.5)
```
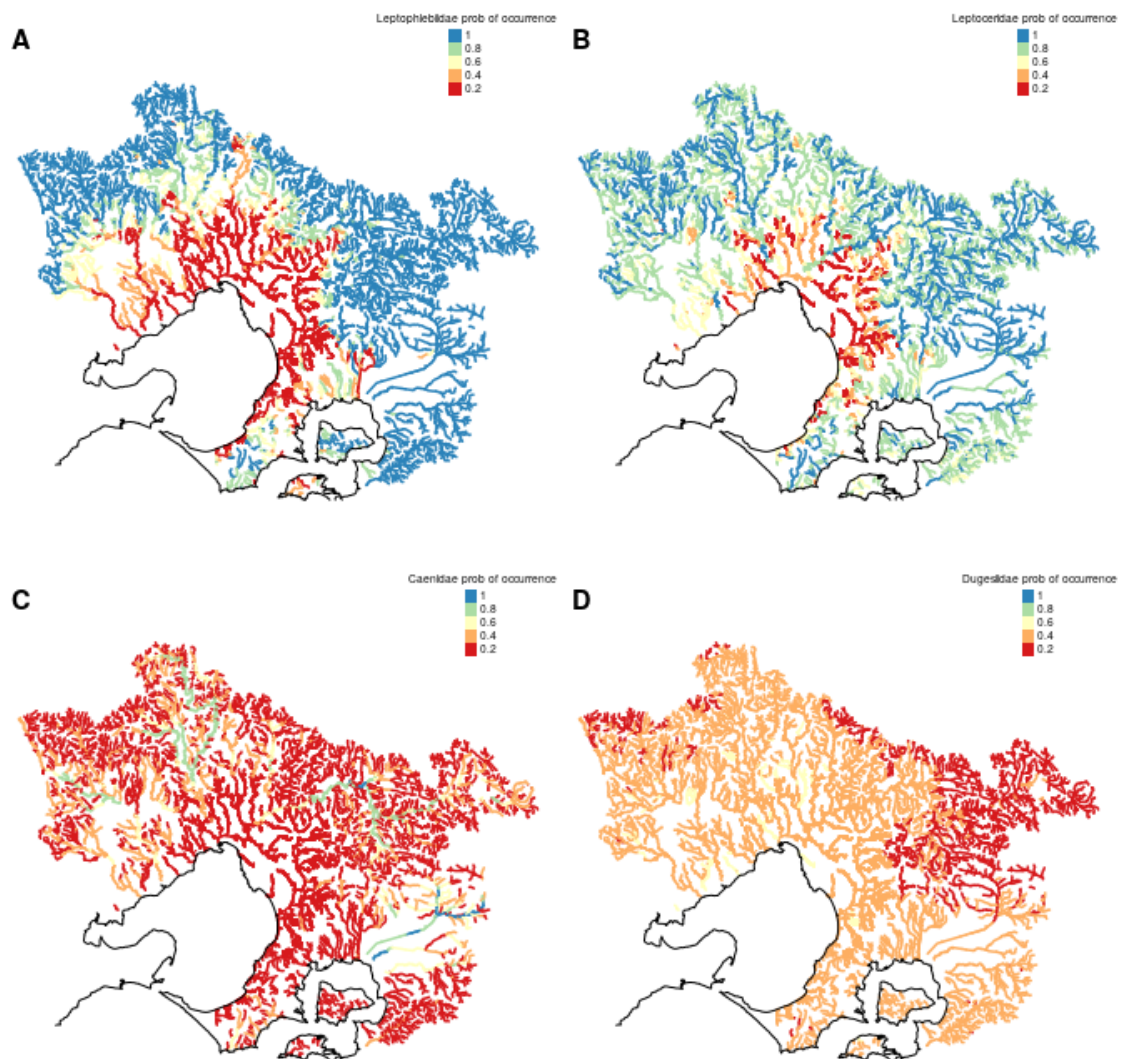

Fig. 3, Predicted probability of occurrence of 4 indicative families. A. Leptophlebiidae (QE08, a very sensitive family) B. Leptoceridae (QT25, a moderately sensitive family), C. Caenidae (QE08, sensitive to urban stormwater runoff, positively associated with forest loss), D. Lymnaeidae (KG05, a weedy family, positively associated with human impacts).

Very sensitive families such as leptophlebid mayflies (Fig. 3A) have very low probabilities of occurrence outside forested catchments, while moderately sensitive families such as leptocerid caddisflies (Fig. 3B) maintain high probabilities of occurrence under moderate levels of agricultural or urban impacts, but are much less likely to occur in highly urban streams. Caenid mayflies (Fig. 3C) are strongly negatively associated with urban stormwater runoff, but are positively associated with deforested, larger streams. Five of the 59 families are classed as weedy, such as dugesiid flatworms (Fig. 3D), which occur most commonly in the most degraded urban streams, and least commonly in forested catchments.

Similar predictions of occurrence for fish species and platypus can be also be mapped (Fig. 4).

```
mwstreamsFish <- collateSampleFP(mwstreams, SRI_48mth_weighted = 1, FishOrPlatypus = "fish"
,
                                 barriersFromYear = FALSE,
                                 PartBarriersDS = mwstreams$nPartBarriersDS_2016,
                                 FullBarriersDS = mwstreams$nFullBarriersDS_2014)
angaust2006 <- vertPred("ANGUAUST", mwstreamsFish)
galamacu2006 <- vertPred("GALAMACU", mwstreamsFish)
lampreys2006 <- vertPred("LAMPREYS", mwstreamsFish)
mwstreamsPlat <- collateSampleFP(mwstreams, SRI_48mth_weighted = 1, FishOrPlatypus = "platy
pus")
allplaty2006 <- vertPred("allPlatyColeman", mwstreamsPlat)
par(mfrow = c(2,2))
plotMWstreamsByVar(angaust2006,
                 varName = paste(vertSpp$species[vertSpp$vertcode == "ANGUAUST"],
                                 "prob of capture"),
                 nbreaks = 5, style = "fixed",
                 fixedBreaks = c(0,0.1,0.2,0.3,0.5,1), legend.cex = 0.5)
title(main = "  A", adj = 0, line = -1.5)
plotMWstreamsByVar(galamacu2006,
                 varName = paste(vertSpp$species[vertSpp$vertcode == "GALAMACU"],
                                 "prob of capture"),
                 nbreaks = 5, style = "fixed",
                 fixedBreaks = c(0,0.1,0.2,0.3,0.5,1), legend.cex = 0.5)
title(main = "  B", adj = 0, line = -1.5)
plotMWstreamsByVar(lampreys2006,
                 varName = "Lampreys (2 spp) prob of capture",
                 nbreaks = 5, style = "fixed",
                 fixedBreaks = c(0,0.1,0.2,0.3,0.5,1), legend.cex = 0.5)
title(main = "  C", adj = 0, line = -1.5)
plotMWstreamsByVar(allplaty2006,
                 varName = "Total platypus prob of capture",
                 nbreaks = 5, style = "fixed",
                 fixedBreaks = c(0,0.15,0.25,0.35,0.5,1), legend.cex = 0.5)
title(main = "  D", adj = 0, line = -1.5)
```
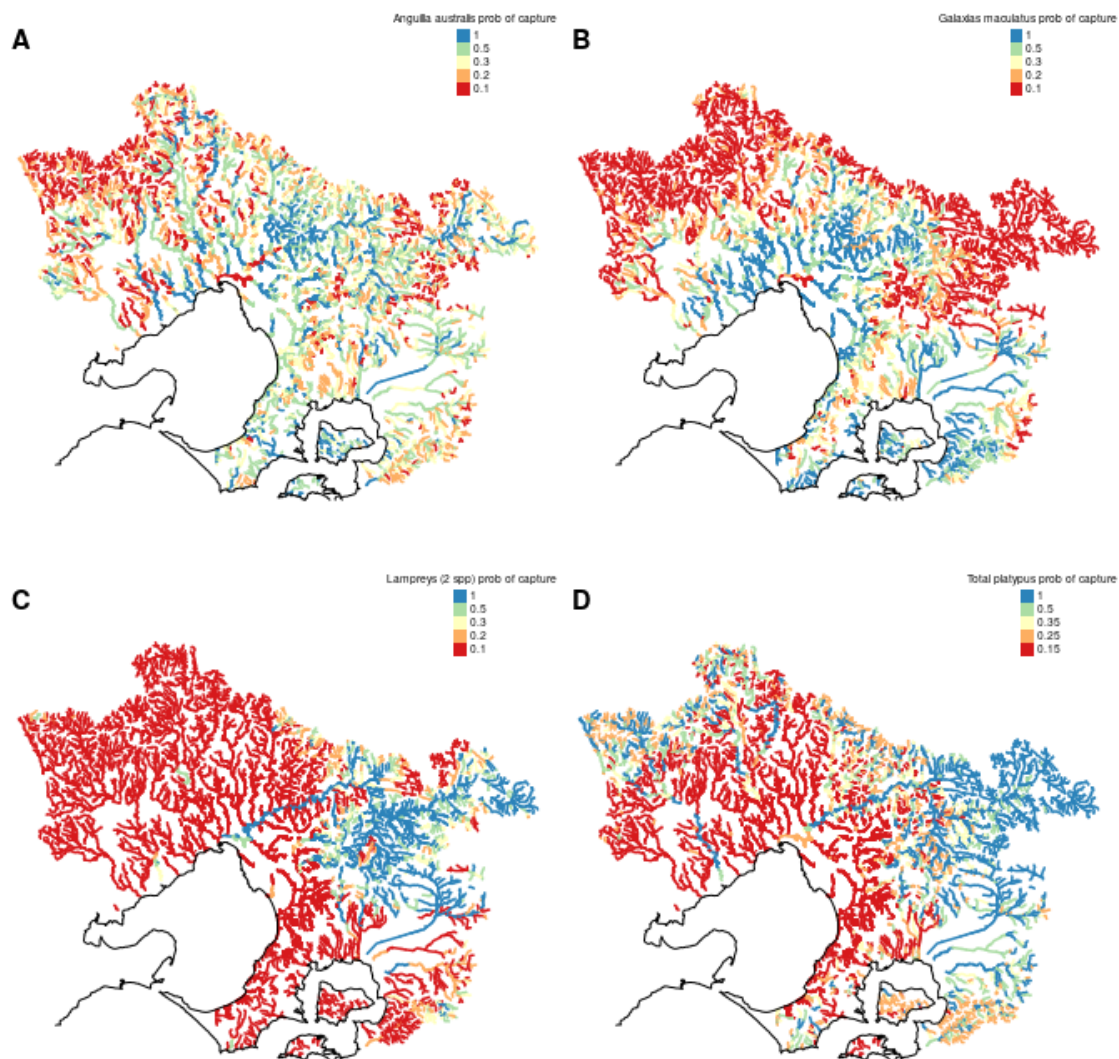
Fig. 4. Predicted probability of occurrence of: A. Southern short-finned eel B. common galaxias, C. Lampreys (two species combined), D. (total) Platypus (as modelled by Coleman et al. 2022)

Alternative scenarios can be explored for the distribution of each family by manipulating the environmental data. For example:

- No human impacts can be simulated by setting AI to zero and AF to 1. Such a prediction is the basis of the expected term in the LUMaR index (see below), and is calculated together with the current (2006) condition using the "bugModPred1" function.

- The mwstreams table contains four variables recalculating AF from 2006 conditions, assuming vegetated buffers along all streams. The effect of revegetating all streams with a (say) 40-m forested buffer can be predicted by replacing the values in "AttForest_L35W1000" with "Att.Forest_40mBuffer".

- Future urban expansion scenarios assuming standard stormwater-management practice have been developed by mapping areas of proposed development and assume AI in such developments is close to total imperviousness. The mwstreams table contains AI estimates for 2006, 2014 and 'ultimate' (assuming

development of all areas zoned for urban development within the urban growth boundary), with matching estimates for the 'min4k' variant of AI (AttImpMin4k_L9) that is used for predicting platypus distribution. The min4k variant is the minimum AI value in any reach within 4 km downstream of a site, as determined by Martin *et al.* (2014). Alternatively, the hypothesized effects of retaining stormwater to remove its impacts from streams could be simulated by setting AttImp_L9 to zero.

- The effects of climate change could be predicted by adding a set amount to mnAnnAirTm_deg. Long-term changes in mean annual discharge could also be simulated by changing the "meanAnnQ" variable. Calculation of such scenarios are in progress.

```
mwstreamsb40 <- mwstreams
#simulate a 40-m forested buffer along all streams
mwstreamsb40$AttForest_L35W1000 <- mwstreamsb40$Att.Forest_40mBuffer
QT25b40 <- bugModPred1("QT25", mwstreamsb40)
mwstreamsaiUlt <- mwstreams
mwstreamsaiUlt$AttImp_L9 <- mwstreamsaiUlt$AttImp_L9_Ultimate
QT25aiUlt <- bugModPred1("QT25", mwstreamsaiUlt)
mwstreamst2 <- mwstreams
mwstreamst2$mnAnnAirTm_deg <- mwstreamst2$mnAnnAirTm_deg + 2
#5020/8256 reaches now have mnAnnAirTm_deg > maximum value in mwstreams. To reduce
extrapolation, cap the absolute increase in air temperature to 0.5 degrees greater
than the maximum in mwstreams (this capping applies primarily to already-degraded
lowland streams)
mwstreamst2$mnAnnAirTm_deg[mwstreamst2$mnAnnAirTm_deg >
                          max(mwstreams$mnAnnAirTm_deg) + 0.5] <-
                max(mwstreams$mnAnnAirTm_deg) + 0.5
QT25t2 <- bugModPred1("QT25", mwstreamst2)
par(mfrow = c(2,2))
plotMWstreamsByVar(QT25Pred$pred1NHI,
                   varName = "P under no human impact",
                   nbreaks = 5, style = "fixed",
                   fixedBreaks = seq(0,1,0.2), legend.cex = 0.5)
title(main = "  A", adj = 0, line = -1.5)
plotMWstreamsByVar(QT25b40$pred1,
                   varName = "P with 40-m forested buffers",
                   nbreaks = 5, style = "fixed",
                   fixedBreaks = seq(0,1,0.2), legend.cex = 0.5)
title(main = "  B", adj = 0, line = -1.5)
plotMWstreamsByVar(QT25aiUlt$pred1,
                   varName = "P assuming 'ultimate' urban growth",
                   nbreaks = 5, style = "fixed",
                   fixedBreaks = seq(0,1,0.2), legend.cex = 0.5)
title(main = "  C", adj = 0, line = -1.5)
plotMWstreamsByVar(QT25t2$pred1,
                   varName = "P 2-deg warmer",
                   nbreaks = 5, style = "fixed",
                   fixedBreaks = seq(0,1,0.2), legend.cex = 0.5)
title(main = "  D", adj = 0, line = -1.5)
```
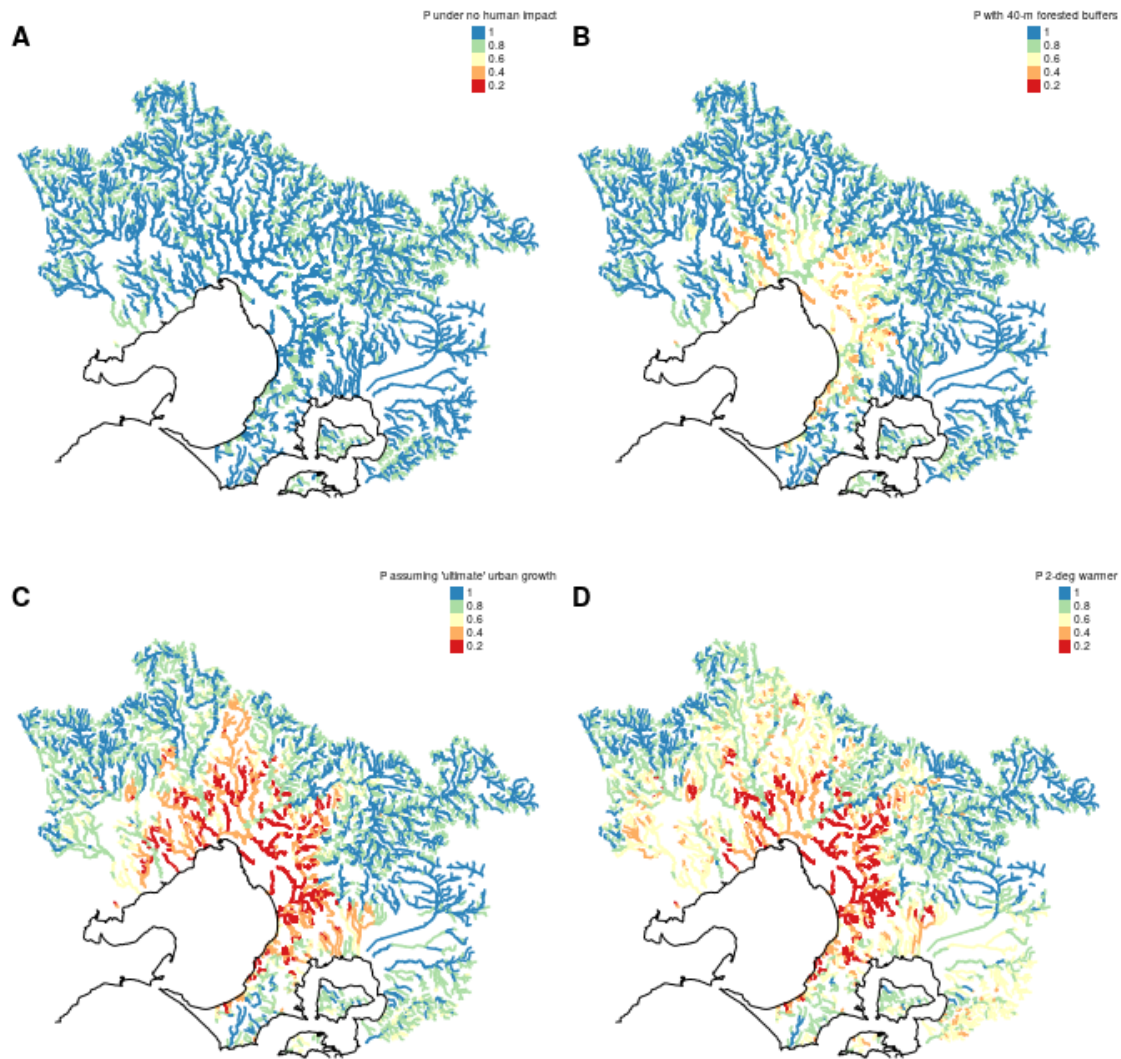
Fig. 5. Predicted probability of occurrence of Leptoceridae under four example scenarios. A. No human impacts, B. Reafforestation of 40-m buffers along all streams of the region, C. Removal of all urban stormwater impacts, D. Increase of 2 degrees in air temperature.

For a moderately sensitive family such as Leptoceridae, the models predict a high probability of occurrence in all streams in the absence of human impacts (Fig. 5A). Reafforestation of streams of the region with 40-m forested buffers is predicted to increase the probability of occurrence of Leptoceridae substantially in agricultural areas of the region, but result in little change in metropolitan streams (Fig. 5B). The increased urban stormwater impacts that would result from the ultimate expansion of Melbourne assuing business-as-usual stormwater practice is predicted to reduce the probability of leptocerid occurrence in streams to the north and west of the city that drain urban growth areas (Fig. 5C). An increase in mean air temperatures of 2°C is predicted to reduce the probability of occurrence in most streams of the region, particularly small streams (Fig. 5D).

Predicting and mapping ensemble predictions and indices

LUMaR is an index of stream condition. It can be calculated for pairs of macroinvertebrate samples collected from any reach in the region. Because every reach

has an estimated value for each of the predictor variables it is also possible to predict LUMaR score for each reach without a sample. Calculation of LUMaR with sample data will be considered below, but here, I describe the prediction of LUMaR based solely on the predictor variables. Predicted values of LUMaR under current (2006) conditions and under no human impact can be calculated as follows:

1. assemble predictions (current and no human impact) for all families using the "predCurrNHI59" function;

2. use the function "prob59ToPA" to convert the current predictions from probabilities to inferred presences or absences using the "pa.threshold" field in the taxon.classes table (this is necessary, as LUMaR requires its observed data to be presence (1) or absence (0))

3. use the predicted current presence-absence data frame, and the predicted no-human-impact data frame as the obs.table and exp.table arguments of the "lumar" function respectively to calculate LUMaR.

The "bugModPred59" function combines the above 3 steps and returns not only predicted LUMaR scores, but also SIGNAL2, and number of sensitive families. Each of these metrics is strongly correlated with human impacts across the region (Walsh (2023)). It should be noted that the SIGNAL2 value is based only on the 59 modelled taxa: normally SIGNAL2 for a sample is calculated using all of the taxa collected. If SIGNAL scores are of interest, then the function bugModPredSIGNAL() produces predictions of SIGNAL scores calculated using whole samples, not just the 59 modelled families. These predictions are based on boosted regression tree models of SIGNAL and SIGNAL2 using the same set of samples as used for the 59 family models.

The predicted index values calculated by "lumar" can be mapped in the same way as for the environmental variables and individual family predictions (Fig. 6). Fig 6B and C show predictions from bugModPredSIGNAL().

```
#calculate LUMaR (in 2006 and under no human impact) for all reaches the long way...
mwstreams59preds <- predCurrNHI59(mwstreams)
mwstreamsCurrpa <- prob59ToPA(mwstreams59preds$predCurr)
mwstreamsLumar <- lumar(mwstreamsCurrpa, mwstreams59preds$predNHI)
mwstreamsNHIpa <- prob59ToPA(mwstreams59preds$predNHI)
mwstreamsNHILumar <- lumar(mwstreamsNHIpa, mwstreams59preds$predNHI)
# and the same calculation the short way, also getting SIGNAL2 and nSensFams
mwstreamsIndices <- bugModPred59(mwstreams)$sampprPredsLumar
#check both methods give the same answer...
# # sum(mwstreamsIndices$lumar != mwstreamsLumar$lumar)  #should be zero if they are
re
#and for No human impact
mwstreamsNHI <- mwstreams
mwstreamsNHI$AttImp_L9 <- 0
mwstreamsNHI$AttForest_L35W1000 <- 1
mwstreamsNHIIndices <- bugModPred59(mwstreamsNHI)$sampprPredsLumar
#Use the bugModPredSIGNAL() function to make a matching prediction of SIGNAL2 scor
es
mwsSIG <- bugModPredSIGNAL(mwstreams)
par(mfrow = c(3,2))
plotMWstreamsByVar(mwstreamsLumar$lumar,
                   varName = "LUMaR (2006)",
```

```
                nbreaks = 5, style = "fixed",
                fixedBreaks = c(-1,0,0.25,0.5,0.75,1), legend.cex = 0.5)
title(main = "  A", adj = 0, line = -1.5)
plotMWstreamsByVar(mwstreamsNHILumar$lumar,
                varName = "LUMaR (no human impact)",
                nbreaks = 5, style = "fixed",
                fixedBreaks = c(-1,0,0.25,0.5,0.75,1), legend.cex = 0.5)
title(main = "  B", adj = 0, line = -1.5)
plotMWstreamsByVar(mwsSIG$SIGNAL2,
                varName = "SIGNAL2 (2006)",
                nbreaks = 6, style = "fixed",
                fixedBreaks =  c(0,3,3.75,4.5,5.25,6,8), legend.cex = 0.5)
title(main = "  C", adj = 0, line = -1.5)
plotMWstreamsByVar(mwsSIG$SIGNAL2nhi,
                varName = "SIGNAL2 (no human impact)",
                nbreaks = 6, style = "fixed",
                fixedBreaks = c(0,3,3.75,4.5,5.25,6,8), legend.cex = 0.5)
title(main = "  D", adj = 0, line = -1.5)
plotMWstreamsByVar(mwstreamsIndices$nSensFams,
                varName = "N Sens taxa (2006)",
                nbreaks = 6, style = "fixed",
                fixedBreaks = c(0,3,6,12,18,24,35), legend.cex = 0.5)
title(main = "  E", adj = 0, line = -1.5)
plotMWstreamsByVar(mwstreamsNHIIndices$nSensFams,
                varName = "N Sens taxa (no human impact)",
                nbreaks = 6, style = "fixed",
                fixedBreaks = c(0,3,6,12,18,24,35), legend.cex = 0.5)
title(main = "  F", adj = 0, line = -1.5)
```
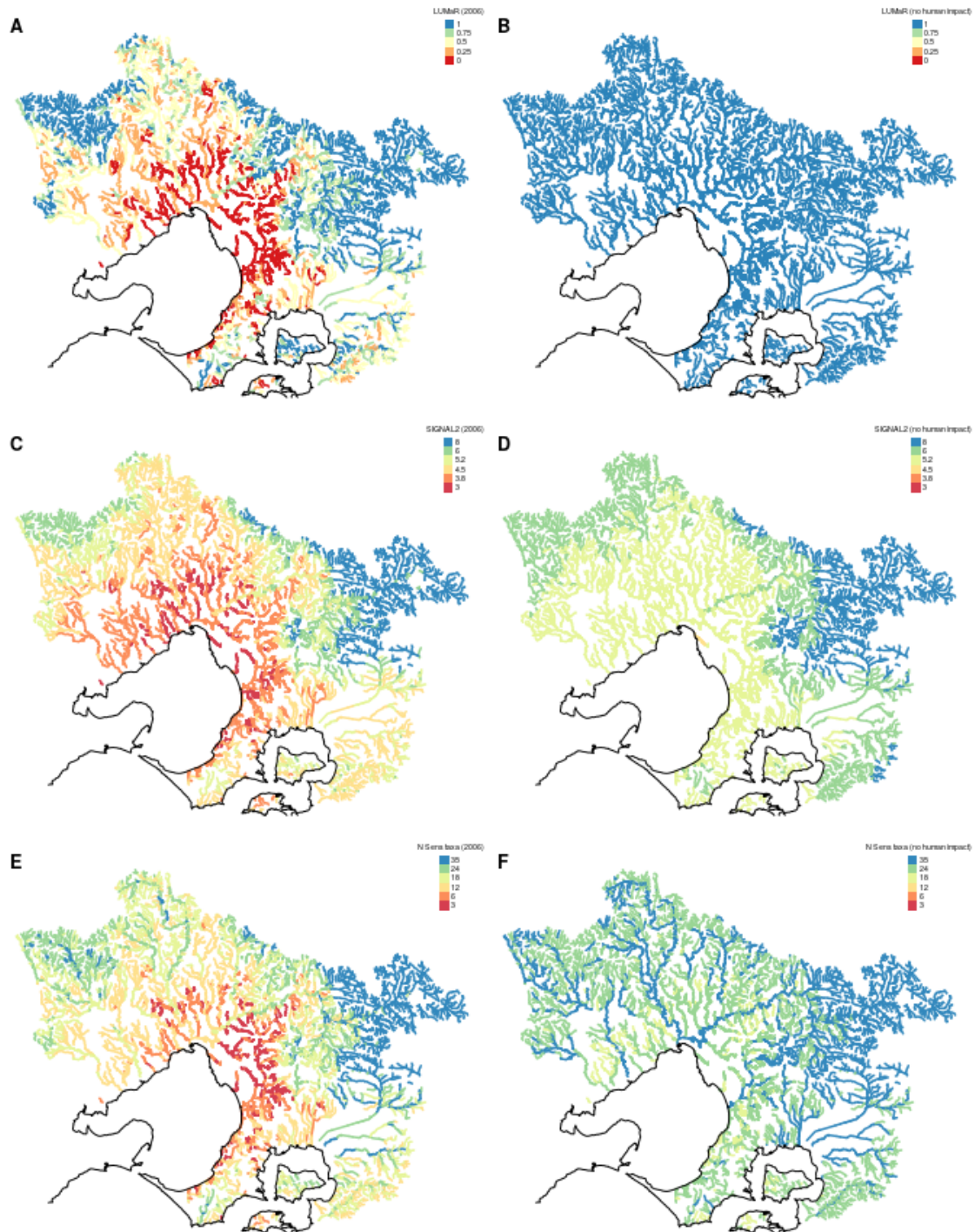
Fig. 6. Predicted LUMaR score (A) under 2006 conditions, (B) under no human impact, and similar contrasts for SIGNAL2 (C and D), and number of sensitive families (E and F)

Metropolitan streams (except the Yarra and the Maribyrnong) are associated with lowest LUMaR, SIGNAL2 scores and fewest sensitive taxa, and streams in cleared agricultural catchments also have lower LUMaR, SIGNAL2 scores and fewer sensitive taxa than upland forested streams (Fig 6A, C, E). While all three indicators are sensitive to human impacts, the predictions of the indicators under no human impact (Fig 6B, D, F) show the advantage of LUMaR over the other two, in that it is predicted to have an equivalent score of 1 under no human impacts across the region (Fig. 6B). In contrast,

drier, lowland streams are predicted to have fewer sensitive taxa (Fig. 6F), and resultantly lower SIGNAL2 scores (Fig. 6D) under no human impact.

Calculating LUMaR score from macroinvertebrate samples.

Calculating LUMaR scores from collected macroinvertebrate samples requires preprocessing of sample data to produce two tables.

1. A table with salient information about each samppr is also required. This table needs to provide a samppr code (matching those in the macroinvertebrate table [2, below]), a subcatchment code (subc), a date of sampling, and values for nriff (number of riffle samples), nspring (number of spring samples), and processN (0 = lab-sort; 1 = field-sort). With this information, the function "collateBugSamppr" can be used to compile predictor variables required for LUMaR calculation.

- Importing date data into R is often challenging, but if dates are formatted in excel, then importing the data using readxl::read_excel (see below) usually works without pain. Dates are only necessary to calculate antecedent discharge. If dates are unavailable, or values of "SRI_48mth_weighted" are known, this field can be supplied as an argument to "collateBugSamppr", instead of dates.

- Determining the subcatchment code for each samppr will require matching site locations to reaches, which is best done in GIS software using both the streams layer (the "mwstreams_map" object in this package), and its associated subcatchment map. (Both layers available in `mwstreams_subcs_coast.gpkg` from the melbstreambiota open science framework repository: https://osf.io/mcxrq/). If subcs need to be determined for only a few sites (or if you have more than a few, and a lot of time on your hands), this could potentially be achieved in R using the "quickmapr" package, using the following commands.

```
melmap <- quickmapr::qmap(mwstreams_map, mwcoast_map, colors = c("blue","black"))
quickmapr::ze(melmap) #to zoom
quickmapr::i(melmap) #to identify subc and other reach information
quickmapr::f(melmap) #to unzoom
?quickmapr #for more options
```

2. Macroinvertebrate taxa collected from two rapid bioassessment samples (EPA Victoria 2003) need to be combined to form each sample pair. If the data are arranged in a 3 column table (sample, taxon, abundance), then combining samples is as simple as creating a new column 'sample-pair', so that pairs of samples have the same sample-pair code (the melbstreambiota package calls these codes samppr). The unique sample-pairs in this table must match the list of sample-pairs in the sample-pair table. The "collateObsTable" function ignores the abundance field and reduces each sample pair to unique taxa. Taxa need to be identified by EPA Victoria bugcodes (used in calculation of [AUSRIVAS] (http://www.mdfrc.org.au/bugguide/resources/AUSRIVAS_Taxacodes.pdf)). It doesn't matter if taxa are identified to lower taxonomic levels than used by LUMaR. The 'collateBugSamppr' function uses only the first 4 characters of the bugcodes, thus automatically lumping to family level. It also removes unmodelled taxa and alters some taxonomic conventions to match the LUMaR conventions (with warnings). Alternatively taxa can be arranged in a taxoncode

(columns) by samppr (rows) matrix (with abundances or presence-absences in the data). The example excel file illustrates both formats.

```
#The sample pairs table imported from an excel file (8 sample pairs)
sampprs <- as.data.frame(readxl::read_excel(system.file("extdata","exampleData.xlsx",
                                            package = "melbstreambiota", mustWo
rk = TRUE),
                                    sheet = 1))
sampprs

##                     samppr stream    subc nriff nspring       date  process
## 1 BRS0015.2007.12.lab-sort    BRS YARR8202     1       2 2007-12-01 lab-sort
## 2 OLN0009.2007.12.lab-sort    OLN YARR8822     1       2 2007-12-01 lab-sort
## 3  BRS0015.2008.4.lab-sort    BRS YARR8202     1       0 2008-04-01 lab-sort
## 4  OLN0009.2008.4.lab-sort    OLN YARR8822     1       0 2008-04-01 lab-sort
## 5 BRS0015.2008.12.lab-sort    BRS YARR8202     1       2 2008-12-01 lab-sort
## 6 OLN0009.2008.12.lab-sort    OLN YARR8822     1       2 2008-12-01 lab-sort
## 7  BRS0015.2009.4.lab-sort    BRS YARR8202     1       0 2009-04-01 lab-sort
## 8  OLN0009.2009.4.lab-sort    OLN YARR8822     1       0 2009-04-01 lab-sort

#The matching macroinvertebrate data as a two-column dataset
bugData <- as.data.frame(readxl::read_excel(system.file("extdata","exampleData.xlsx",
                                            package = "melbstreambiota", mustWo
rk = TRUE),
                                    sheet = 2))
head(bugData)

##                   samppr bugcode abundance
## 1  OLN0009.2008.4.lab-sort    IB01   2.00000
## 2  BRS0015.2008.4.lab-sort    IF61   1.00000
## 3  BRS0015.2009.4.lab-sort    IF61   1.00000
## 4 BRS0015.2008.12.lab-sort    KG02  19.66667
## 5 OLN0009.2007.12.lab-sort    KG02  45.00000
## 6 OLN0009.2008.12.lab-sort    KG02  30.00000

#Sheet 3 of the excel file contains the same data in a samppr-by-bugcode matrix
```

With the two tables imported, calculating LUMaR scores of the samples is a four-step process:

1. Use the "collateBugSamppr" function to collate the predictor variables into the sample-pairs table

2. Use the "predCurrNHI59"" to calculate expected probabilities of occurrence in the absence of human impacts and format as required for the expTable argument of the "lumar" function

3. Format the macroinvertebrate data into a matching obsTable using the "collateObsTable" function

4. run the lumar function using the obsTable and the expTable.

```
sampprs <- collateBugSamppr(sampprs)
expTables <- predCurrNHI59(sampprs)
obsTable <- collateObsTable(bugData, sampprs)

## Your bugData does not include any entries for the following families that are used by LU
MaR.
## This may, of course, be correct, but check for any unexpected omissions, and check bugco
de/taxonomy.
##   QO12
##   OT01
```

```
##  QE02
##  QE08
##  QT24
##  QE05
##  QM01
##  QD06
##  QP01
##  QT17
##  QO05
##  QO07
##  QE03
##  OP06
##  QT04
##  QO08

## Warning: The following bugcodes were excluded as they are not used in LUMaR calculations
:
## LH05, LO03, LO05, LO06, LO08, LO11, MM99, OR25, QD12, QD23, QD24, QDAF, QDAJ, KG02, OP01
, QD89, KG09, QK01, QH52, QO21, IB01.

exampleLumar <- lumar(obsTable, expTables$predNHI)
exampleLumar[,1:3]  #just first 3 columns for display purposes

##                      samppr       lumar nSensFams
## 1 BRS0015.2007.12.lab-sort -0.07923222         4
## 2 OLN0009.2007.12.lab-sort  0.76537171        24
## 3  BRS0015.2008.4.lab-sort  0.08218193         6
## 4  OLN0009.2008.4.lab-sort  0.67667656        20
## 5 BRS0015.2008.12.lab-sort  0.17519829        10
## 6 OLN0009.2008.12.lab-sort  0.75338553        21
## 7  BRS0015.2009.4.lab-sort  0.25886586        10
## 8  OLN0009.2009.4.lab-sort  0.63658501        18
```

The lumar function returns a number of sub-indices that can be used to diagnose patterns in the LUMaR score, but the most important outputs are LUMaR score itself (lumar) and the number of sensitive families (nSensFam), both of which are strong indicators of stream health (Walsh 2017).

Furthermore, the function calcSIGNAL() can be used to calculate SIGNAL and SIGNAL2 scores from the supplied macroinvertebrate data. Unlike other functions in this package, it doesn't require samples be compiled into pairs of rapid bioassessment samples, although the input data.frame does require an ID field named samppr. (bugData). This function calculates SIGNAL using all taxa listed in each sample, not just the 59 modelled families. Input data format requirements are the same as for collateObsTable().

```
calcSIGNAL(bugData)

##                      samppr   SIGNAL  SIGNAL2
## 1 BRS0015.2007.12.lab-sort 4.000000 2.384615
## 2 BRS0015.2008.12.lab-sort 5.533333 3.812500
## 3  BRS0015.2008.4.lab-sort 4.352941 2.611111
## 4  BRS0015.2009.4.lab-sort 5.578947 3.650000
## 5 OLN0009.2007.12.lab-sort 6.322581 5.375000
## 6 OLN0009.2008.12.lab-sort 6.333333 5.714286
## 7  OLN0009.2008.4.lab-sort 6.440000 5.560000
## 8  OLN0009.2009.4.lab-sort 6.363636 5.500000
```

Diagnostic tools: interpreting the LUMaR results

The melbstreambiota package includes two primary diagnostic tools for interpreting LUMaR results.

1. The "sampprDiagnostic" function permits interpretation of an observed LUMaR score at an individual site. Fig. 7 shows the output for one of the sites in the example dataset (a sample pair from Brushy Creek in the eastern suburbs of Melbourne).

```
#Use the same sampprs, expTables and obsTable as above
#Use the same sampprs, expTables and obsTable as above
sampprDiagnostic(samppri = sampprs[3,],
                 obsTable = obsTable, currentPredTable = expTables$predCurr,
                 NHIPredTable = expTables$predNHI)
```
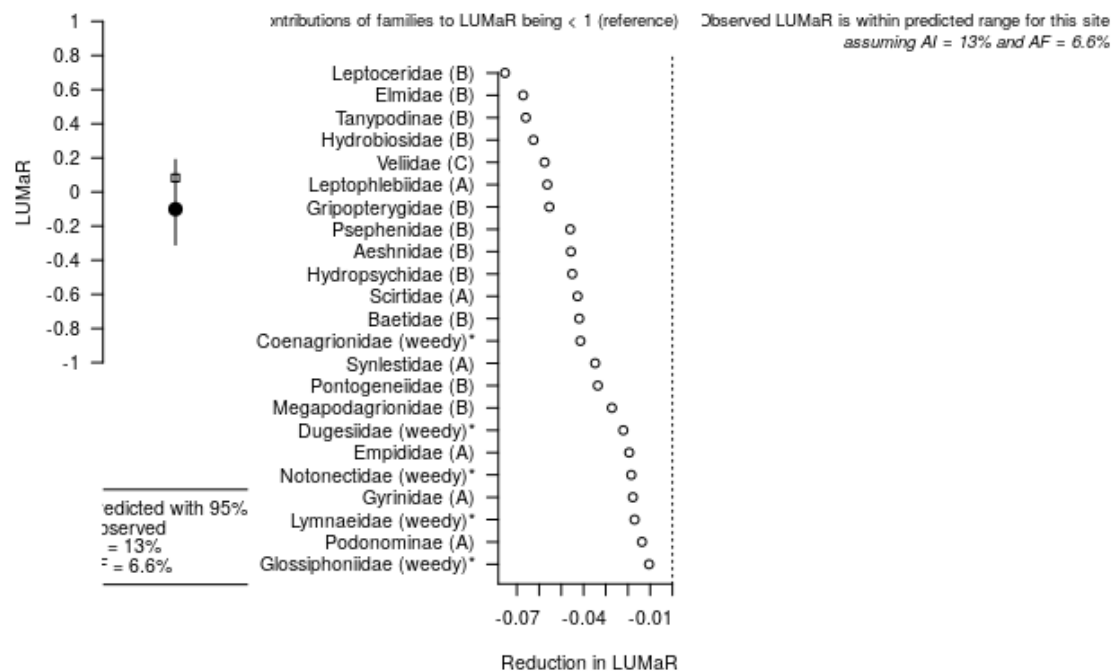

Fig. 7. Output of sampppDiagnostic(), comparing the observed LUMaR score to the predicted score for 2006 conditions, and listing the most influential taxa resulting in this sampple pair scoring less than would be expected in the absence of human impacts (reference). If the observed LUMaR score fell outside the predicted range for 2006, a second plot would be presented identifying taxa influencing that difference.

The output shows that the LUMaR score of this sample-pair falls within the predicted range for LUMaR at this site under 2006 conditions (plot on the left). The observed score of 0.08 is less than would be expected in the absence of human impacts (a score of >0.75 is predicted for such sites). The second plot shows the influence of the most influential taxa causing the LUMaR score to be less than reference condition (i.e. no human impacts). The taxa indicated by asterisks occurred in the sample but were unexpected, thus downweighting the LUMaR score. These five unexpected families were weedy (see Walsh (2023) for definitions of these sensitivity classes). Taxa without asterisks are those that were predicted to be present in the sample, but were not (thus downweighting the score). Leptocerid, hydrobiosid, hydropsychid caddisflies, tanypod midges, elmid and psephenid beetles, gripopterygid stoneflies (all moderately sensitive, class B), leptophlebiid mayflies and scirtid beetles (very sensistive, class A), and veliid bugs and atyid shrimp (sensitive to urban impacts, less so forest loss, class C) were the primary absences driving the lower score. The unexpected presence of several weedy species, and the unexpected absence of several moderately sensitive taxa in this sample-pair suggests that this site (Brushy Creek in suburban Mooroolbark) is degraded.
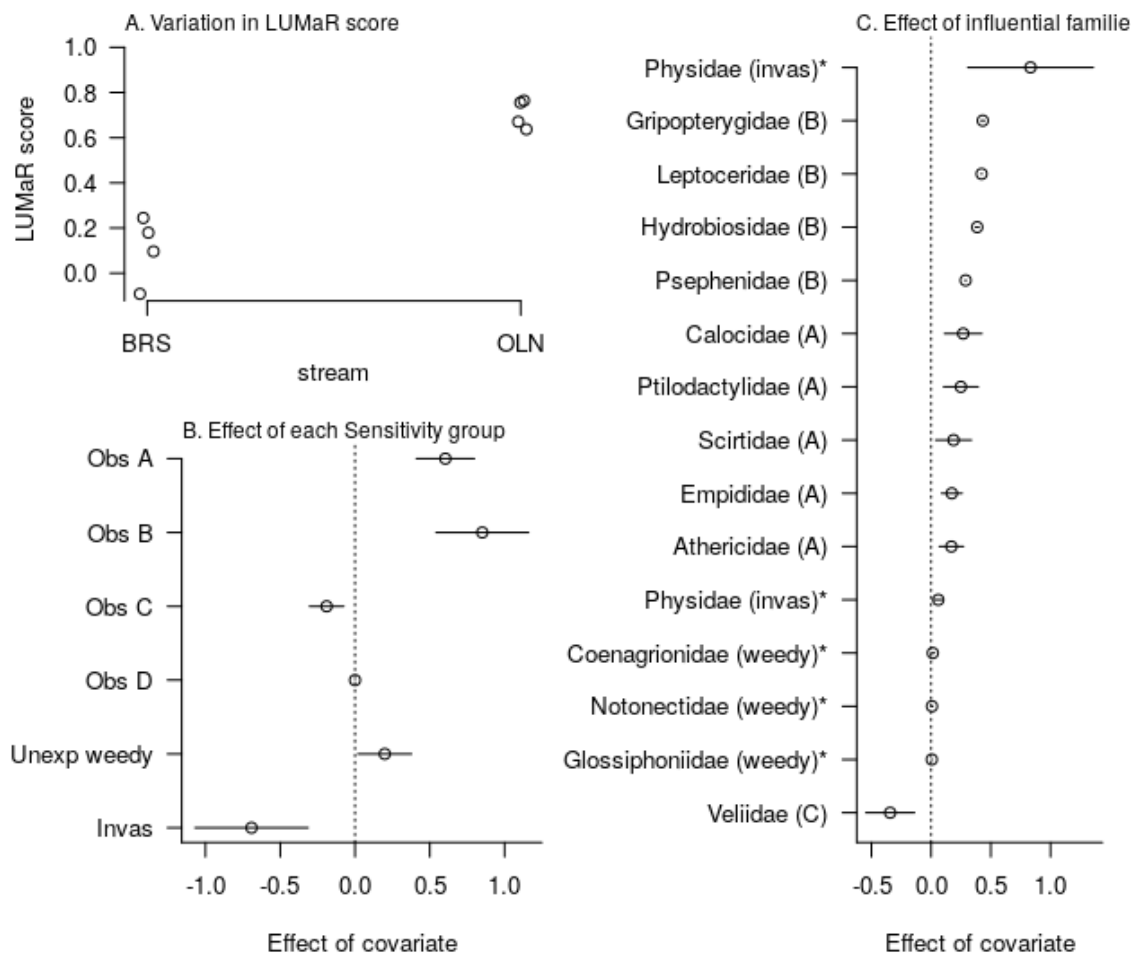
If the LUMaR score of the sample-pair was to fall outside the expected range under 2006 conditions, an additional plot would be plotted on the right, showing the influence of the most influential taxa driving that difference. Walsh (2023) illustrated and interpreted such an example.

2. "plotSensFamDiagnostics" is a diagnostic tool to aid understanding of differences in LUMaR between groups of sample-pairs of among sample-pairs that are from a gradient described by a covariate. In the example dataset in the melbstreambiota package, the eight sample pairs were collected from two streams (urban Brushy Creek, BRS, and forested Olinda Creek, OLN: see Burns *et al.* (2012) and Walsh, Fletcher & Burns (2012) for a comparison of the urban stormwater impacts of these streams). The LUMaR scores from Brushy Creek are substantialy lower than from Olinda Creek. The differences in LUMaR scores as defined by the covariate argument in "plotSensFamDiagnostics" are illustrated in the first plot produced by the function (Fig. 8)

```
#Use the same expTables, obsTable, exampleLumar, and stream covariate as above
# What taxa explain differences in LUMaR between the two streams (BRS and OLN)
stream <- factor(substr(exampleLumar$samppr,1,3), levels = c("BRS","OLN"))
famStreamEffects <- famCovariateEffects(obsTable, expTables$predNHI, stream)
sensStreamEffects <- sensGpCovariateEffects(exampleLumar, stream)
plotSensFamDiagnostics(exampleLumar, stream, covariateName = "stream",
                       sensLinEffects = sensStreamEffects,
                       famLinEffects = famStreamEffects)
```

Fig. 8. Output of plotSensFamDiagnostics(), comparing LUMaR scores in samples from Brushy Creek (BRS) and Olinda Creek (OLN).

The bottom left plot of the "plotSensFamDiagnostics" shows the relative importance of each sensitivity group (whether it was observed and expected, or observed but unexpected). In this example, the primary driver of OLN sample-pairs having higher LUMaR scores than BRS sample pairs is the absence of several expected sensitive taxa from BRS that were expected and found at OLN (Obs B) and the unexpected presence of the invasive snail *Physa* (Invas) at BRS. The plot on the right identifies the most influential sensitive families as Gripopterygidae, Leptoceridae, Hydrobiosidae and Psephenidae.

References

Burns, M.J., Fletcher, T.D., Walsh, C.J., Ladson, A.R. & Hatt, B.E. (2012) Hydrologic shortcomings of conventional urban stormwater management and opportunities for reform. *Landscape and Urban Planning*, **105**, 230–240.

Chee, Y.E., Coleman, R.A., RossRakesh, S. & Walsh, C.J. (2020) *Habitat Suitability Models, Scenarios and Quantitative Action Prioritisation (Using Zonation) for the Healthy Waterways Strategy 2018: A Resource Document*. Melbourne {{Waterway Research-Practice Partnership Report}} 20.3, The University of Melbourne, Melbourne.

Chessman, B.C. (1995) Rapid assessment of rivers using macroinvertebrates: A procedure based on habitat-specific sampling, family level identification and a biotic index. *Australian Journal of Ecology*, **20**, 122–129.

Chessman, B.C. (2003) New sensitivity grades for australian river macroinvertebrates. *Marine and Freshwater Research*, **54**, 95–103.

Coleman, R.A., Chee, Y.E., Bond, N.R., Weeks, A., Griffiths, J., Serena, M., Williams, G.A. & Walsh, C.J. (2022) Understanding and managing the interactive impacts of growth in urban land use and climate change on freshwater biota: A case study using the platypus (*Ornithorhynchus Anatinus*). *Global Change Biology*, **28**, 1287–1300.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.

EPA Victoria. (2003) *Rapid Bioassessment Methodology for Rivers and Streams*.

Martin, E.H., Walsh, C.J., Serena, M. & Webb, J.A. (2014) Urban stormwater runoff limits distribution of platypus. *Austral Ecology*, **39**, 337–345.

Melbourne Water. (2018) *Healthy Waterways Strategy 2018*. Melbourne Water, Melbourne.

Simpson, J. & Norris, R. (2000) Biological assessment of river quality: Development of AUSRIVAS models and outputs. *Assessing the biological quality of fresh waters: RIVPACS and other techniques. Proceedings of an international workshop held in Oxford, UK, on 16-18 september 1997*, pp. 125–142. Freshwater Biological Association (FBA).

Walsh, C.J. (2023) LUMaR: A sensitive macroinvertebrate index of stream condition combining observed:expected ratios and sensitivity weightings. *OSF pre-prints*.

Walsh, C.J., Fletcher, T.D. & Burns, M.J. (2012) Urban stormwater runoff: A new class of environmental flow problem. *PLOS ONE*, **7**, e45814.

Walsh, C.J. & Kunapo, J. (2009) The importance of upland flow paths in determining urban effects on stream ecosystems. *Journal of the North American Benthological Society*, **28**, 977–990.

Walsh, C.J. & Webb, J.A. (2013) *Predicting Stream Macroinvertebrate Assemblage Composition as a Function of Land Use, Physiography and Climate: A Guide for Strategic Planning for River and Water Management in the Melbourne Water Region. Melbourne Waterway Research Practice Partnership Report 13-1*. Department of Resource Management; Geography, The University of Melbourne, Melbourne.

Walsh, C.J. & Webb, J.A. (2014) Spatial weighting of land use and temporal weighting of antecedent discharge improves prediction of stream condition. *Landscape Ecology*, **29**, 1171–1185.

Walsh, C.J. & Webb, J.A. (2016) Interactive effects of urban stormwater drainage, land clearance, and flow regime on stream macroinvertebrate assemblages across a large metropolitan region. *Freshwater Science*, **35**, 324–339.