Spatially Explicit Predictors of Indicators of Water Quality: Example from Wadeable Streams in the U.S.

Mostafa Shirazi and Marc Weber
Western Ecology Division
U.S. Environmental Protection Agency
Corvallis, Oregon

Acknowledgement

This research is derived from four national-scale databases which were produced over decades by hundreds of staffs of government agencies and their collaborators from the academic and research institutions, efforts which inspire a heartfelt gratitude.

- NRCS National Soil Database SSURGO and STATSGO
- USGS National Land Cover Database (NLCD)
- USEPA National Aquatic Resource Database, Wadeable Stream Assessment (WSA).
- USEPA Ecoregions of the United States

Objective (1)

Spatially Explicit Prediction of Water Quality

- The Environmental Protection Agency (USEPA), in collaboration with the States, is required to assess and report the condition of surface waters nation-wide.

- The Wadeable Stream Assessment (WSA), is a nation-wide probability-based survey, which located a number of wadeable stream sites, applied uniform field procedures to each site, and collected data, which developed water quality indicators (WQI).

- Statistically valid inferences about all surface waters can be drawn from the sampled sites, but an explicit connection of WQI with the ecosystem and its land use, which influence water quality is missing.

- This study uses soil and land use predictors to establish explicit linkage between *WQI* of WSA watersheds and the Ecological Regions of the conterminous United States.

Objective (2)
Using SC+LC in predictive models

- Describe SSURGO and STATSGO soil *map units* by the mean Soil Characteristics (*SC*) of soil layers and soil components.

- Intersect soil map units with NLCD (*LC*) to produce a combined national *SC+LC* coverage: *SL map units.*

- Intersect *SL map units* with the following, and describe them identically by their mean *SC+LC*:
- 1- 1392 *WSA* “Probability” watersheds,
- 2- 451 “least Impacted, Reference” watersheds, and
- 3- 967 US Ecoregions.

- Use *SC+LC* in models as *independent predictors* of *WQI of a not-tested stream,* while using the remaining *WQI* as *co-predictors.*

Selected SC and LC of Soil Map Units and Water Quality indicators (WQI) of WSA Sites

- **Soil Characteristics *SC*** (57 *SC* from 281) Depth, Texture, Bulk Density, Available Water Capacity, Organic Matter, pH, CEC,…, Elevation, Slope, Air Temperature, Precipitation, Frost Free Days,…

- **Land Cover *LC*** (12 classes from 29), developed space, cropland, pastureland, deciduous forests, evergreen forests,…

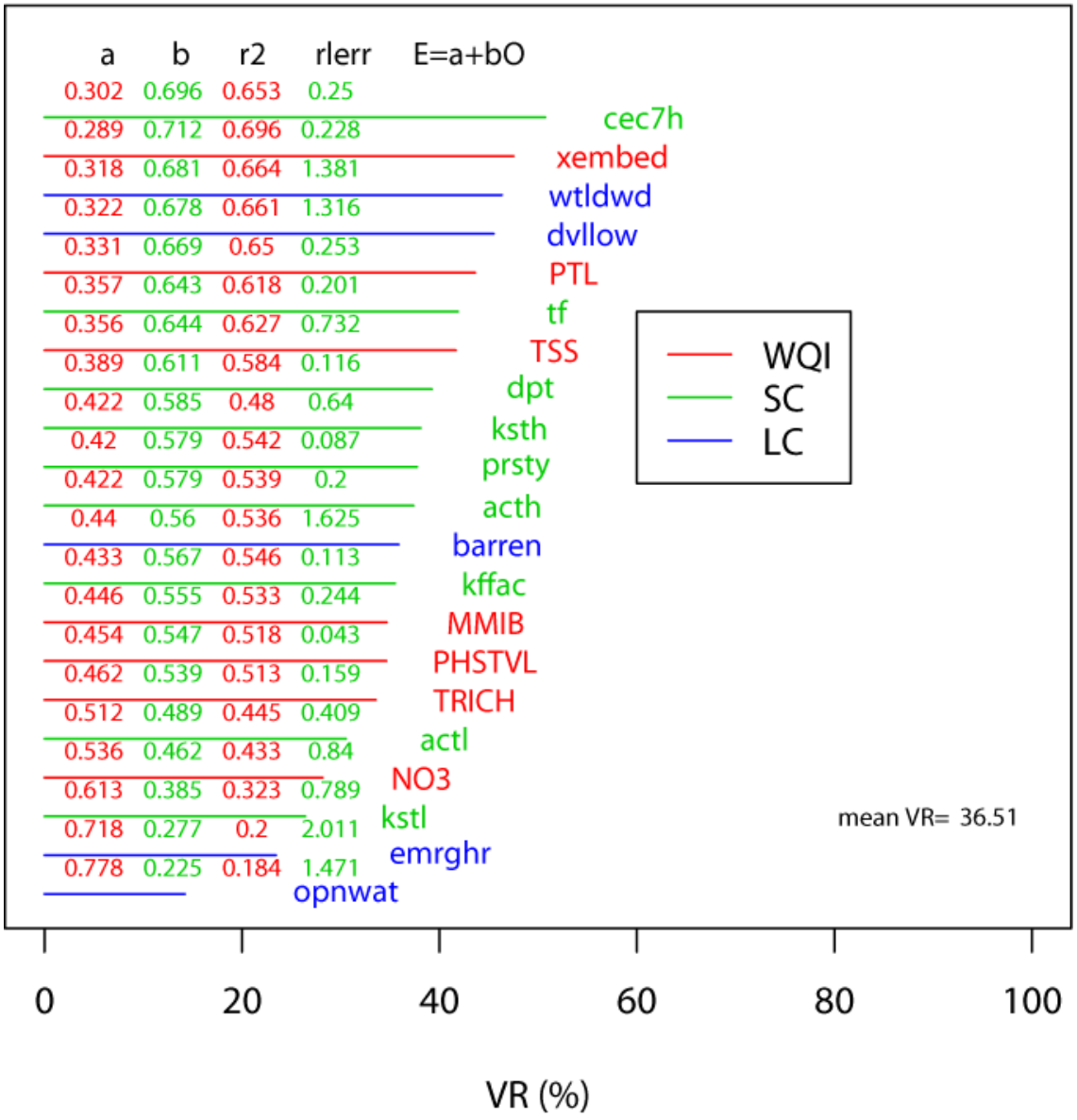
- **Stream Water Quality Indicators *WQI*** (17 items selected): pH, ANC, DOC, Stream Substrate, Taxa Richness, PTL, NO3, SO4, CL.

Models and Test Strategies (1)
Rank and Predictability of SC, LC, and WQI

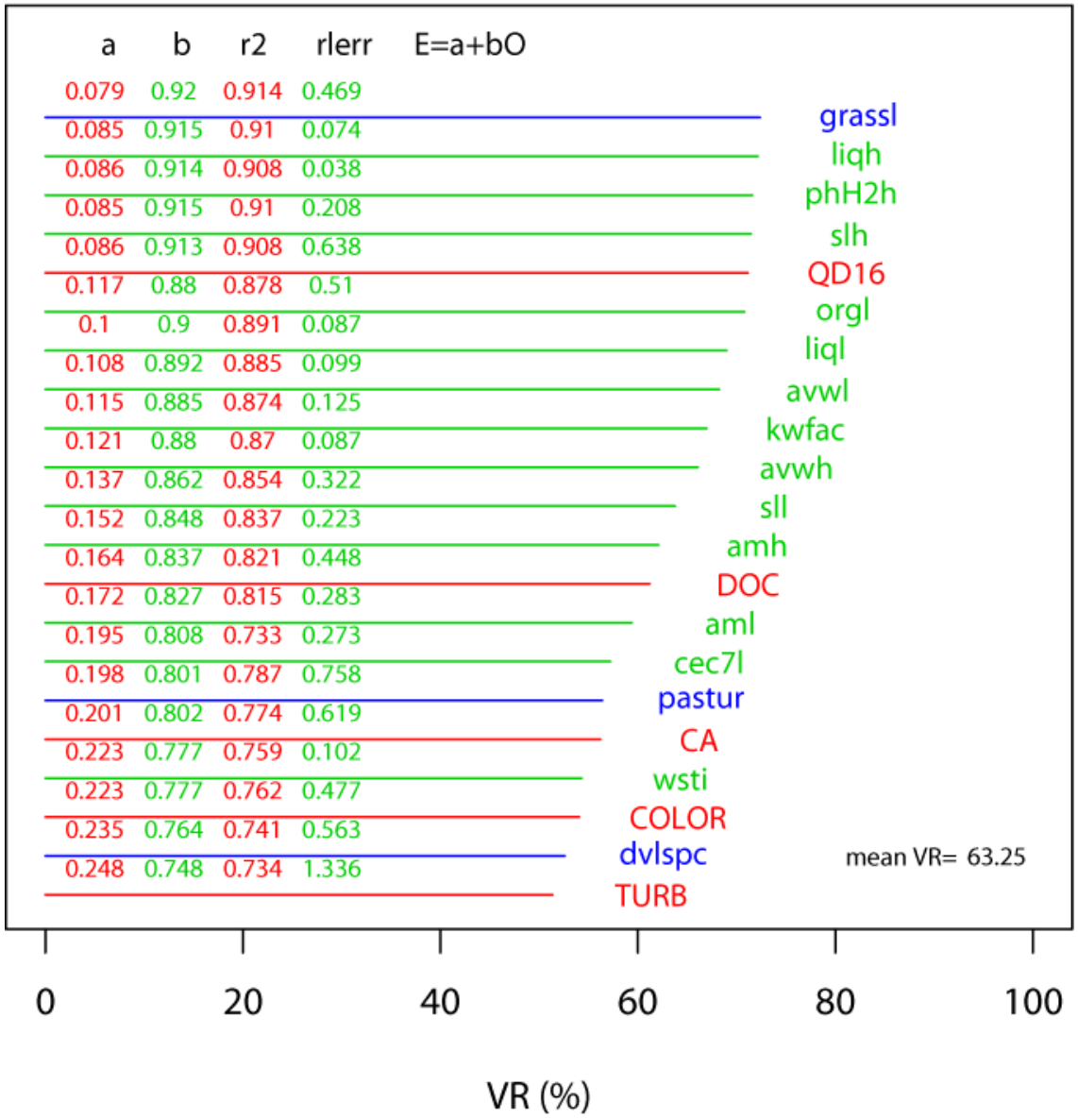
- Compare the sample (V_s) with the conditional variance (V_c), of each vector in the list *SC,LC,* and *WQI*, and the % variance reduction (*VR*) achieved because of vector's spatial relationships with the others not used from the list: ***VR= 100[1.0 – (V_c / V_s)$^{1/2}$].***

- Use one of the 1392 watersheds as a *set-aside,* build a model using the remaining 1391, and predict, in an extrapolation mode, each one of the *set-aside's SC, LC, and WQI.* Repeat 1392 times to calculate *relative standard error, and r^2.*

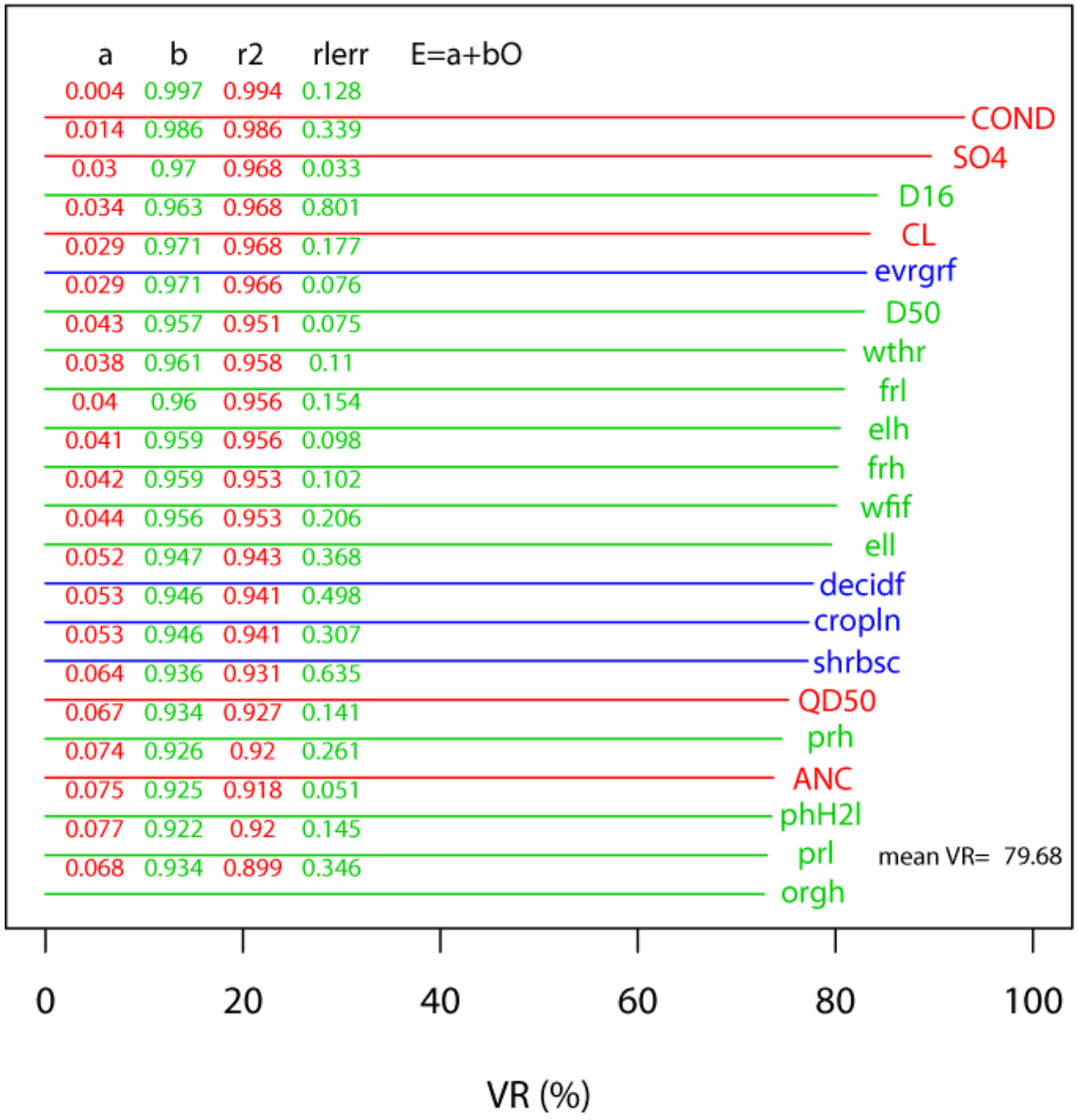
WQI, SC, LC


a b r2 rlerr E=a+bO
0.302 0.696 0.653 0.25
cec7h
0.289 0.712 0.696 0.228
xembed
0.318 0.681 0.664 1.381
wtldwd
0.322 0.678 0.661 1.316
dvllow
0.331 0.669 0.65 0.253
PTL
0.357 0.643 0.618 0.201
tf
0.356 0.644 0.627 0.732
TSS
0.389 0.611 0.584 0.116
dpt
0.422 0.585 0.48 0.64
ksth
0.42 0.579 0.542 0.087
prsty
0.422 0.579 0.539 0.2
acth
0.44 0.56 0.536 1.625
barren
0.433 0.567 0.546 0.113
kffac
0.446 0.555 0.533 0.244
MMIB
0.454 0.547 0.518 0.043
PHSTVL
0.462 0.539 0.513 0.159
TRICH
0.512 0.489 0.445 0.409
actl
0.536 0.462 0.433 0.84
NO3
0.613 0.385 0.323 0.789
kstl
0.718 0.277 0.2 2.011
emrghr
0.778 0.225 0.184 1.471
opnwat
WQI
SC
LC
mean VR= 36.51
0
20
40
60
80
100
VR (%)

WQI, SC, LC


a b r2 rlerr E=a+bO
0.079 0.92 0.914 0.469
0.085 0.915 0.91 0.074
0.086 0.914 0.908 0.038
0.085 0.915 0.91 0.208
0.086 0.913 0.908 0.638
0.117 0.88 0.878 0.51
0.1 0.9 0.891 0.087
0.108 0.892 0.885 0.099
0.115 0.885 0.874 0.125
0.121 0.88 0.87 0.087
0.137 0.862 0.854 0.322
0.152 0.848 0.837 0.223
0.164 0.837 0.821 0.448
0.172 0.827 0.815 0.283
0.195 0.808 0.733 0.273
0.198 0.801 0.787 0.758
0.201 0.802 0.774 0.619
0.223 0.777 0.759 0.102
0.223 0.777 0.762 0.477
0.235 0.764 0.741 0.563
0.248 0.748 0.734 1.336
grassl
liqh
phH2h
slh
QD16
orgl
liql
avwl
kwfac
avwh
sll
amh
DOC
aml
cec7l
pastur
CA
wsti
COLOR
dvlspc
TURB
mean VR= 63.25
0
20
40
60
80
100
VR (%)

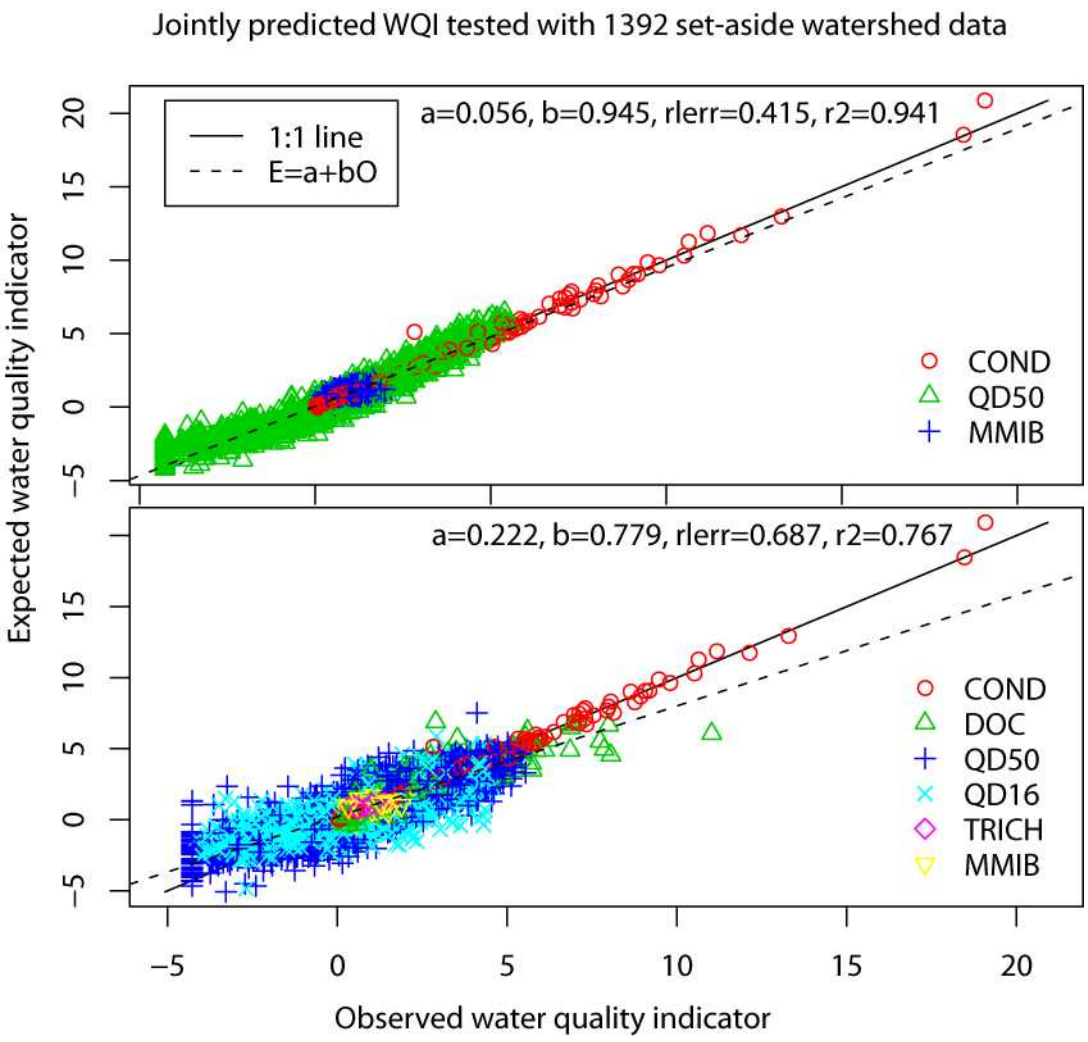
WQI, SC, LC


a b r2 rlerr E=a+bO
0.004 0.997 0.994 0.128
0.014 0.986 0.986 0.339
0.03 0.97 0.968 0.033
0.034 0.963 0.968 0.801
0.029 0.971 0.968 0.177
0.029 0.971 0.966 0.076
0.043 0.957 0.951 0.075
0.038 0.961 0.958 0.11
0.04 0.96 0.956 0.154
0.041 0.959 0.956 0.098
0.042 0.959 0.953 0.102
0.044 0.956 0.953 0.206
0.052 0.947 0.943 0.368
0.053 0.946 0.941 0.498
0.053 0.946 0.941 0.307
0.064 0.936 0.931 0.635
0.067 0.934 0.927 0.141
0.074 0.926 0.92 0.261
0.075 0.925 0.918 0.051
0.077 0.922 0.92 0.145
0.068 0.934 0.899 0.346
COND
SO4
D16
CL
evrgrf
D50
wthr
frl
elh
frh
wfif
ell
decidf
cropln
shrbsc
QD50
prh
ANC
phH2l
prl
orgh
mean VR= 79.68
0
20
40
60
80
100
VR (%)

Models and Test Strategies (2)
Joint Prediction of WQI (small and large VR)

- A weakly predicted *WQI* (mean *VR*=37%) improves in a joint prediction with one or more strongly predicted WQI (mean VR=80%).

- An example of joint prediction tested with data from1392 set-aside watershed sites (next slide).

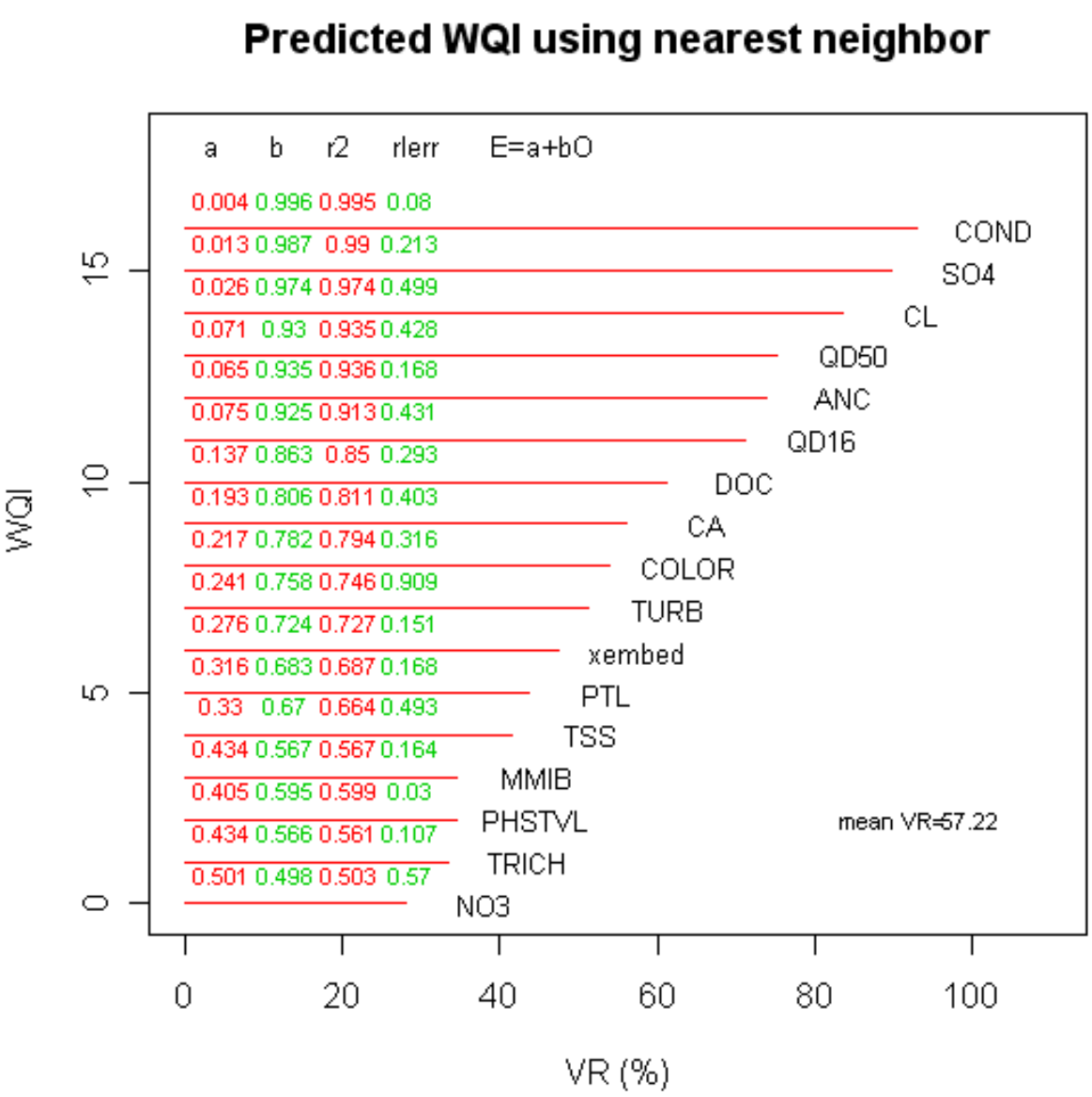

Jointly predicted WQI tested with 1392 set-aside watershed data
1:1 line
E=a+bO
a=0.056, b=0.945, rlerr=0.415, r2=0.941
COND
QD50
MMIB
a=0.222, b=0.779, rlerr=0.687, r2=0.767
COND
DOC
QD50
QD16
TRICH
MMIB
Expected water quality indicator
Observed water quality indicator
−5
0
5
10
15
20

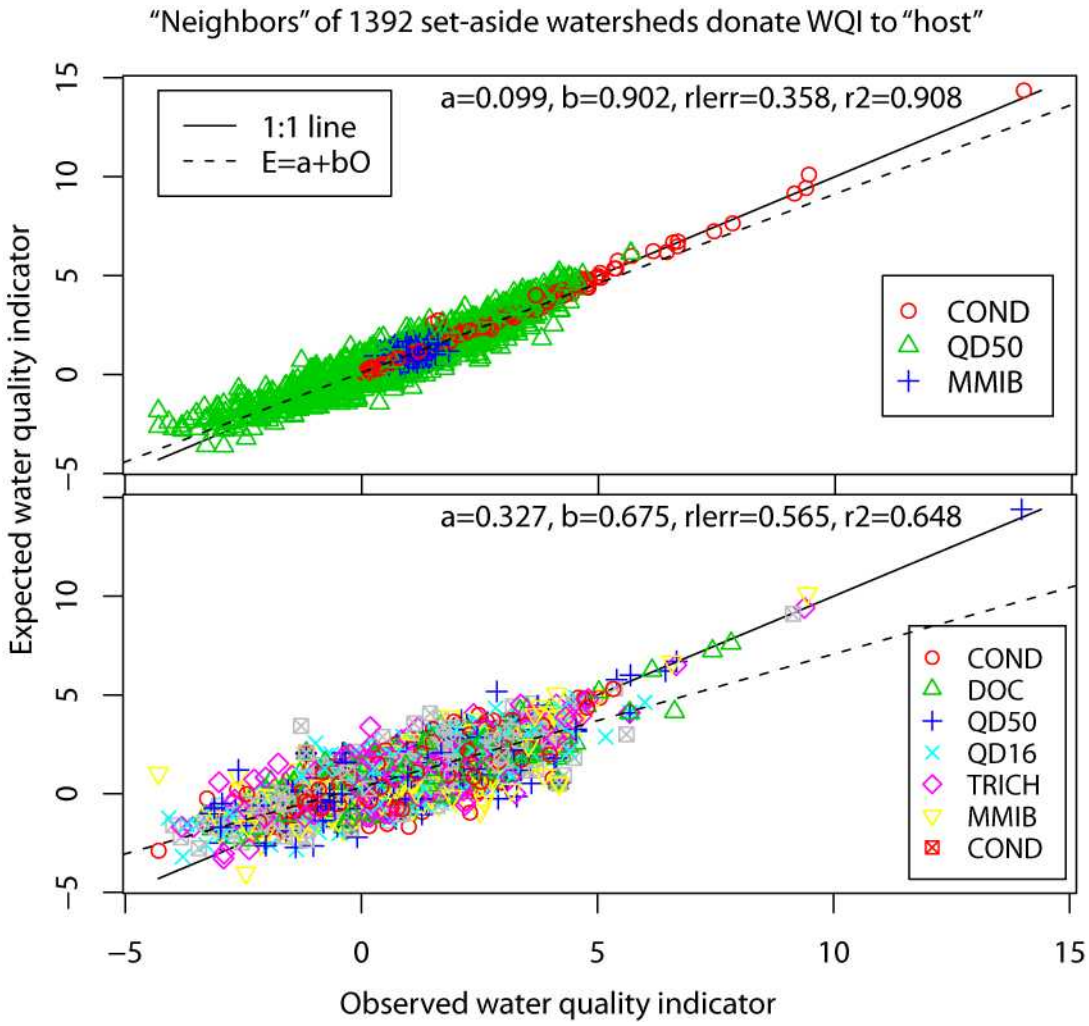
Models and Test Strategies (3)
Using Neighbors of a Set-aside Watershed

- **1-** Test models using 1392 probability sites with *WQI* obtained indirectly, from the "neighbor" watersheds of a set-aside "host" watershed.

- **2-** Locate the nearest neighbors of each host watershed using a sorted list of 1392 s*quared generalized distance* (*SL_sqrd*) of *SC+LC*.

- **3**- Predict *WQI* singly and jointly


Predicted WQI using nearest neighbor
a b r2 rlerr E=a+bO
0.004 0.996 0.995 0.08
0.013 0.987 0.99 0.213
0.026 0.974 0.974 0.499
0.071 0.93 0.935 0.428
0.065 0.935 0.936 0.168
0.075 0.925 0.913 0.431
0.137 0.863 0.85 0.293
0.193 0.806 0.811 0.403
0.217 0.782 0.794 0.316
0.241 0.758 0.746 0.909
0.276 0.724 0.727 0.151
0.316 0.683 0.687 0.168
0.33 0.67 0.664 0.493
0.434 0.567 0.567 0.164
0.405 0.595 0.599 0.03
0.434 0.566 0.561 0.107
0.501 0.498 0.503 0.57
COND
SO4
CL
QD50
ANC
QD16
DOC
CA
COLOR
TURB
xembed
PTL
TSS
MMIB
PHSTVL
TRICH
NO3
mean VR=57.22
WQI
0
5
10
15
0
20
40
60
80
100
VR (%)


"Neighbors" of 1392 set-aside watersheds donate WQI to "host"
a=0.099, b=0.902, rlerr=0.358, r2=0.908
1:1 line
E=a+bO
COND
QD50
MMIB
a=0.327, b=0.675, rlerr=0.565, r2=0.648
COND
DOC
QD50
QD16
TRICH
MMIB
COND
Expected water quality indicator
Observed water quality indicator
−5
0
5
10
15

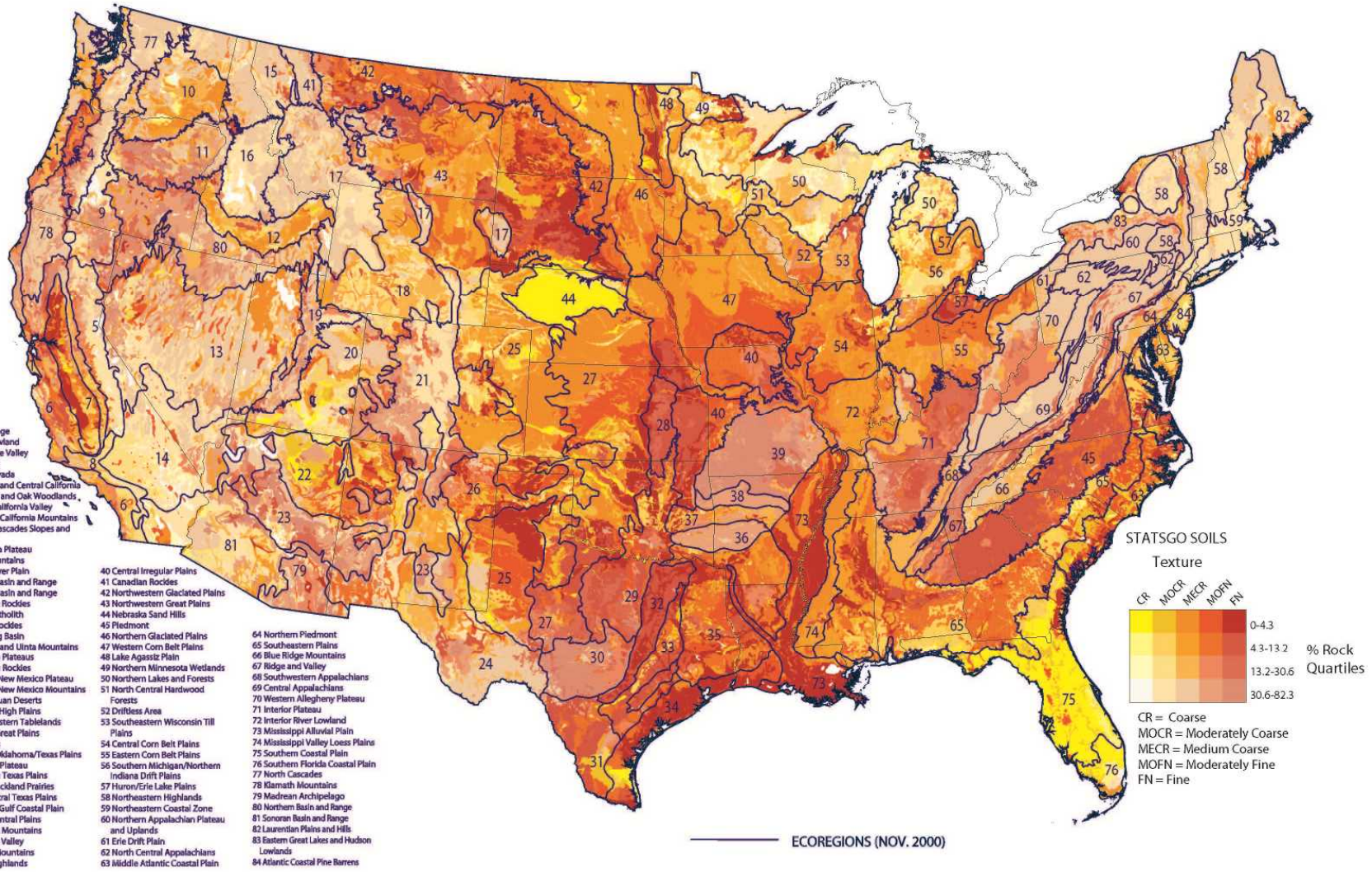
Linking WSA Watersheds and Ecoregions (1)

Example of quantification: Using 26 Soil Characteristics (SC)

- Ecological regions consist of ecosystems (forests, wetlands, deserts) which evolve under regional climate and geology, maintain particular water quality, sustain plant and animal communities, and variously invite human land use.

- Extensive mapped data (including land use), professional judgment, and interpretations determined the ecological boundaries in the U.S. for level *III* and level *IV* classifications, which are available online: (http://www.epa.gov/wed/pages/ecoregions.htm)

- A complete quantitative description of ecoregions is not feasible, but previous quantification, using, as a first approximation, 26 soil characteristics from STATSGO, variously matched boundaries of Level III ecoregions, for example, the soil texture map (next slide).


1 Coast Range
2 Puget Lowland
3 Willamette Valley
4 Cascades
5 Sierra Nevada
6 Southern and Central California Chaparral and Oak Woodlands
7 Central California Valley
8 Southern California Mountains
9 Eastern Cascades Slopes and Foothills
10 Columbia Plateau
11 Blue Mountains
12 Snake River Plain
13 Central Basin and Range
14 Mojave Basin and Range
15 Northern Rockies
16 Idaho Batholith
17 Middle Rockies
18 Wyoming Basin
19 Wasatch and Uinta Mountains
20 Colorado Plateaus
21 Southern Rockies
22 Arizona/New Mexico Plateau
23 Arizona/New Mexico Mountains
24 Chihuahuan Deserts
25 Western High Plains
26 Southwestern Tablelands
27 Central Great Plains
28 Flint Hills
29 Central Oklahoma/Texas Plains
30 Edwards Plateau
31 Southern Texas Plains
32 Texas Blackland Prairies
33 East Central Texas Plains
34 Western Gulf Coastal Plain
35 South Central Plains
36 Ouachita Mountains
37 Arkansas Valley
38 Boston Mountains
39 Ozark Highlands
40 Central Irregular Plains
41 Canadian Rockies
42 Northwestern Glaciated Plains
43 Northwestern Great Plains
44 Nebraska Sand Hills
45 Piedmont
46 Northern Glaciated Plains
47 Western Corn Belt Plains
48 Lake Agassiz Plain
49 Northern Minnesota Wetlands
50 Northern Lakes and Forests
51 North Central Hardwood Forests
52 Driftless Area
53 Southeastern Wisconsin Till Plains
54 Central Corn Belt Plains
55 Eastern Corn Belt Plains
56 Southern Michigan/Northern Indiana Drift Plains
57 Huron/Erie Lake Plains
58 Northeastern Highlands
59 Northeastern Coastal Zone
60 Northern Appalachian Plateau and Uplands
61 Erie Drift Plain
62 North Central Appalachians
63 Middle Atlantic Coastal Plain
64 Northern Piedmont
65 Southeastern Plains
66 Blue Ridge Mountains
67 Ridge and Valley
68 Southwestern Appalachians
69 Central Appalachians
70 Western Allegheny Plateau
71 Interior Plateau
72 Interior River Lowland
73 Mississippi Alluvial Plain
74 Mississippi Valley Loess Plains
75 Southern Coastal Plain
76 Southern Florida Coastal Plain
77 North Cascades
78 Klamath Mountains
79 Madrean Archipelago
80 Northern Basin and Range
81 Sonoran Basin and Range
82 Laurentian Plains and Hills
83 Eastern Great Lakes and Hudson Lowlands
84 Atlantic Coastal Pine Barrens
ECOREGIONS (NOV. 2000)
STATSGO SOILS
Texture
CR
MOCR
MECR
MOFN
FN
0-4.3
4.3-13.2
13.2-30.6
30.6-82.3
% Rock Quartiles
CR = Coarse
MOCR = Moderately Coarse
MECR = Medium Coarse
MOFN = Moderately Fine
FN = Fine

Linking WSA Watersheds and Ecoregions (2)

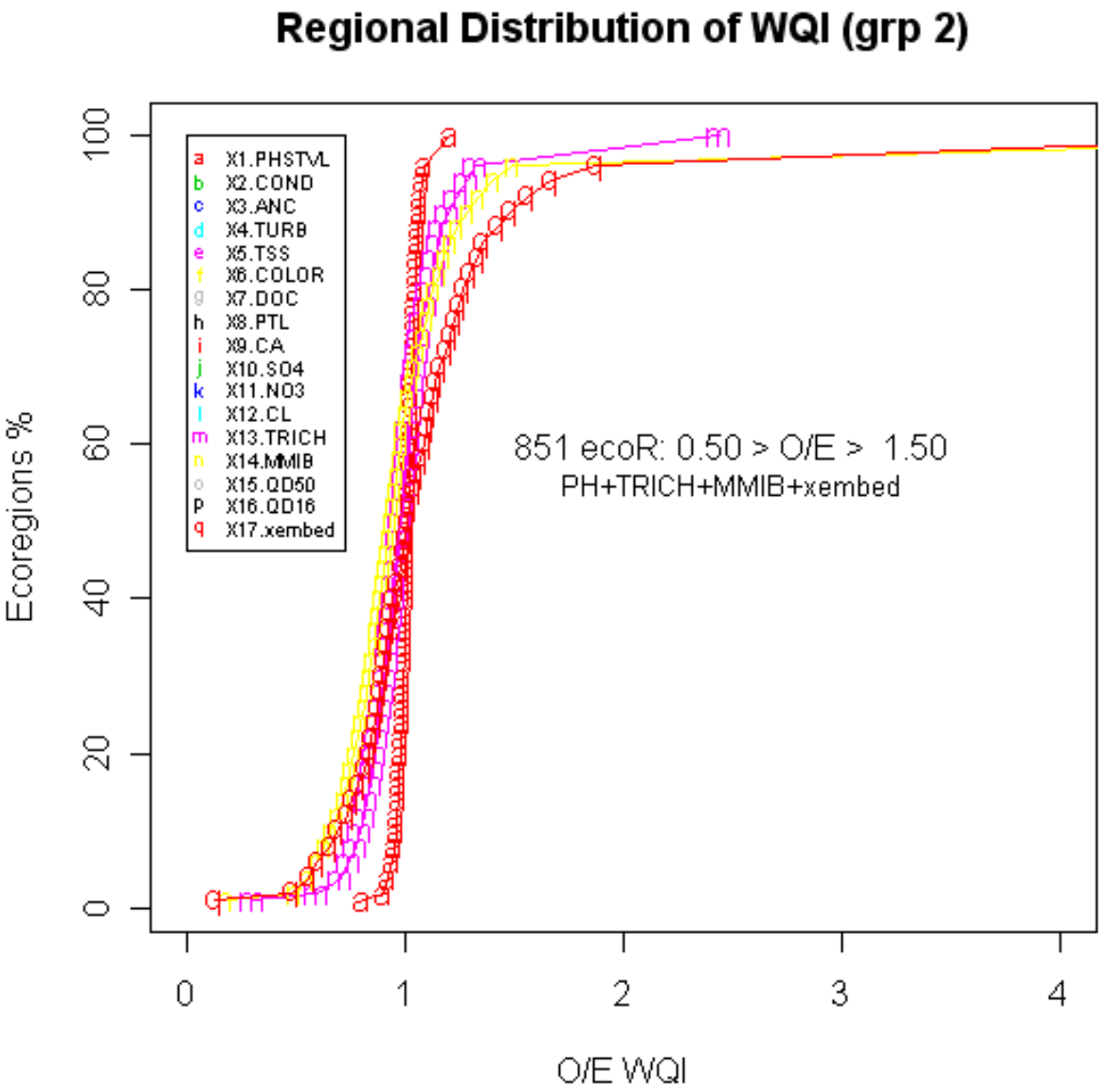
Expanded quantification using SL_sqrd(SC+LC)

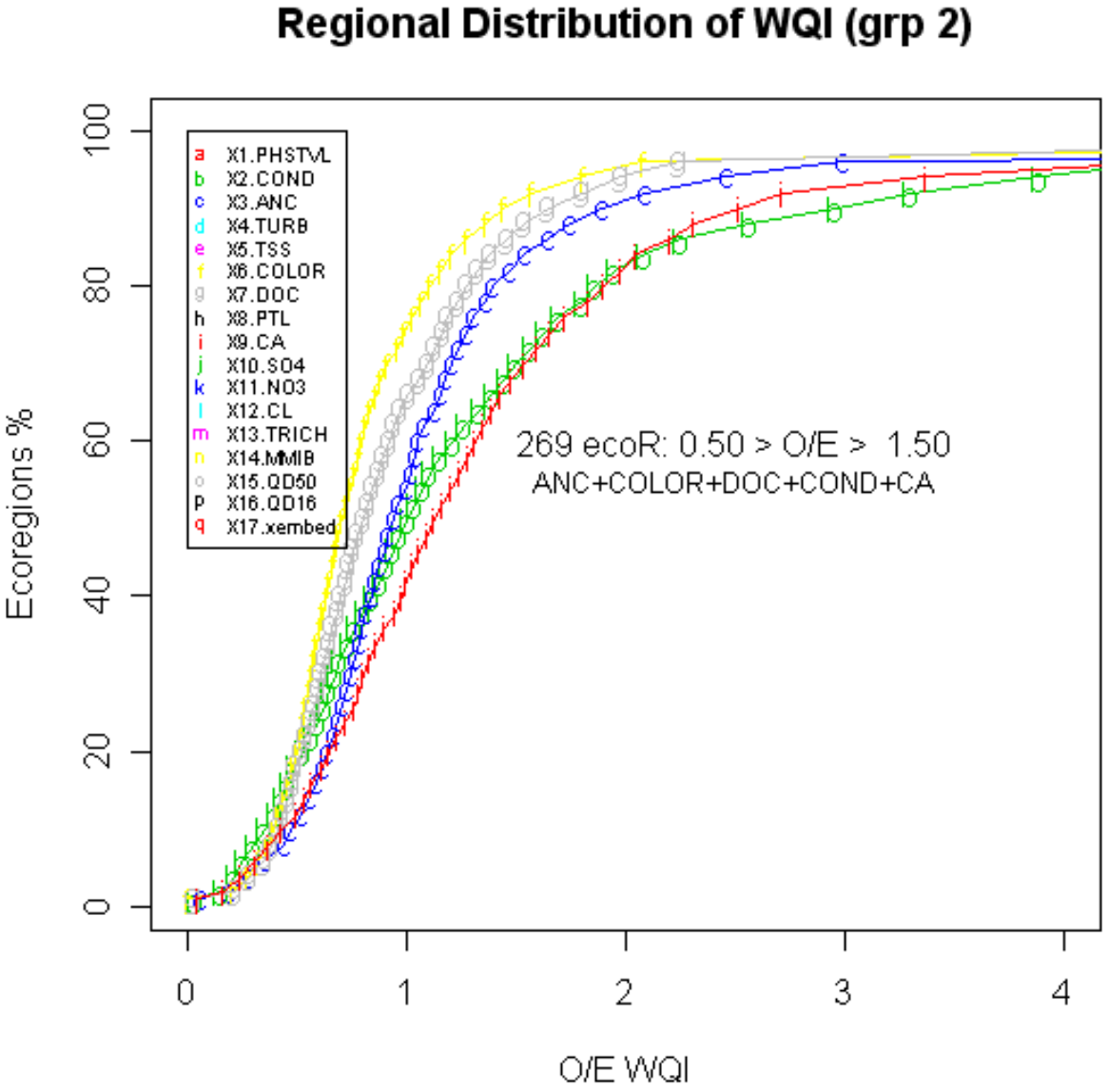
- Use a sorted, jointly calculated, list of *squared generalized distance SL_sqrd(SC+LC)* of ecoregions and probability sites, to locate nearest neighbor donors of a mean *WQI* of probability sites (*PRB)* to each ecoregion.

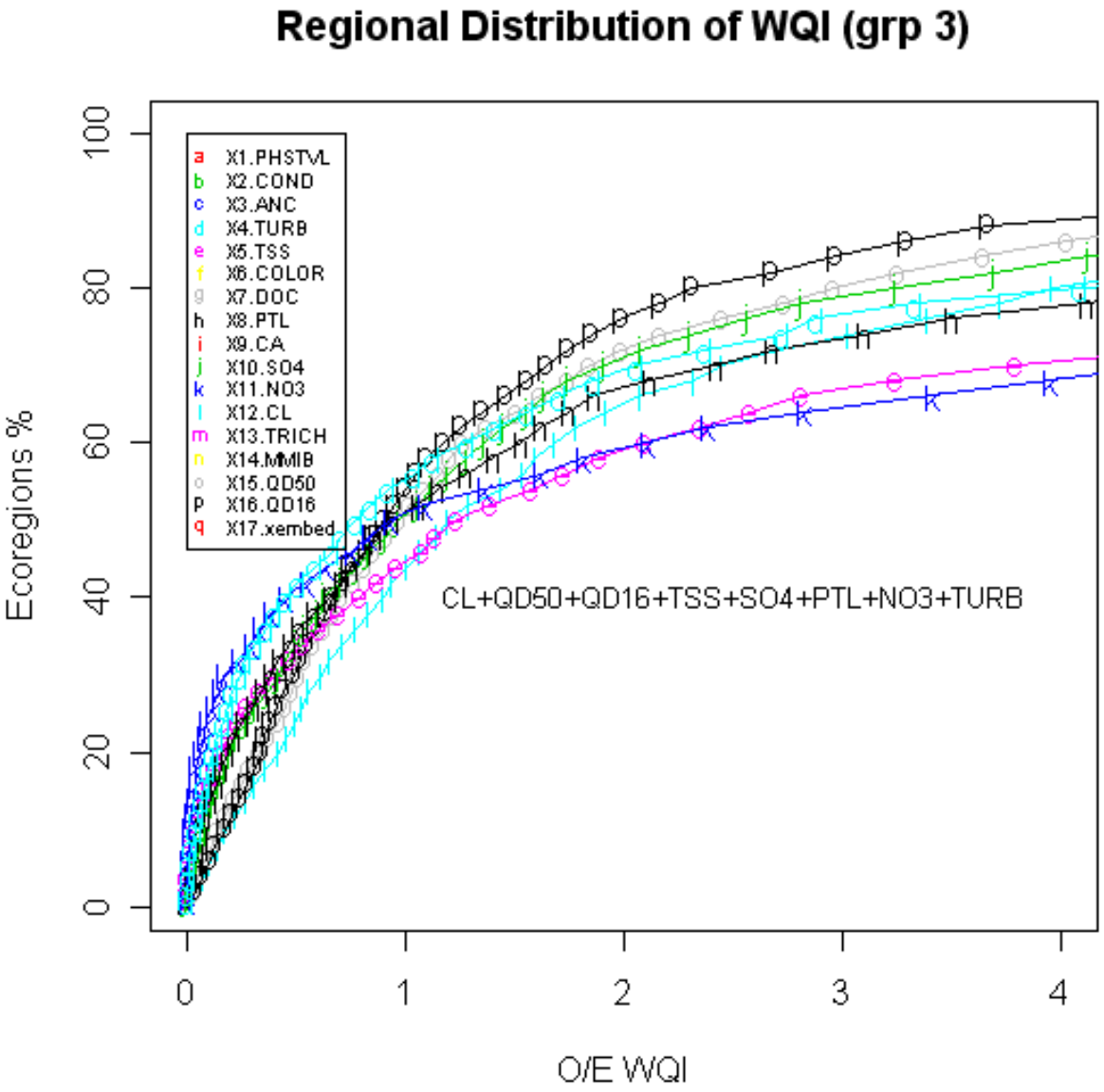
- Use sorted, jointly calculated, list of *SL_sqrd* of ecoregions and ”least-Impacted” reference sites to locate nearest neighbor donors of a mean *WQI of* reference sites (*REF*) to each ecoregion.

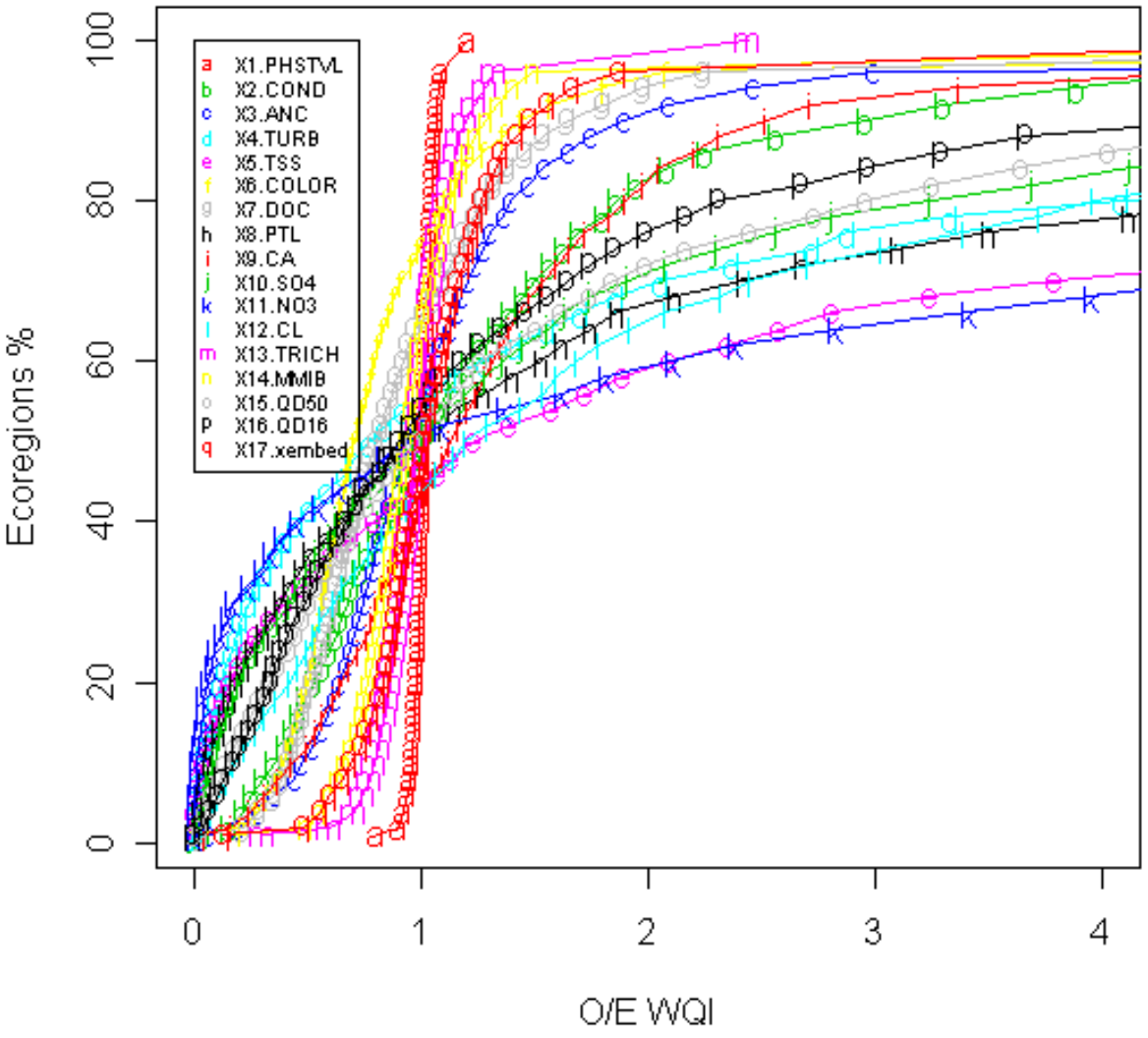
- Define each ecoregional *WQI* by the distribution of the ratio of its “*PRB*” and “*REF*” *WQI, that is,* “*PRB* / *REF* “. For example:

- a- “Least Impacted”: 0.50< *PRB / REF* <1.50
- b- Un-impacted: *PRB* / *REF* <0.50
- c- Impacted: *PRB* / *REF* >1.50


Regional Distribution of WQI (grp 2)
a X1.PHSTVL
b X2.COND
c X3.ANC
d X4.TURB
e X5.TSS
f X6.COLOR
g X7.DOC
h X8.PTL
i X9.CA
j X10.SO4
k X11.NO3
l X12.CL
m X13.TRICH
n X14.MMIB
o X15.QD50
p X16.QD16
q X17.xembed
851 ecoR: 0.50 > O/E > 1.50
PH+TRICH+MMIB+xembed
Ecoregions %
0
20
40
60
80
100
0
1
2
3
4
O/E WQI


Regional Distribution of WQI (grp 2)
a X1.PHSTVL
b X2.COND
c X3.ANC
d X4.TURB
e X5.TSS
f X6.COLOR
g X7.DOC
h X8.PTL
i X9.CA
j X10.SO4
k X11.NO3
l X12.CL
m X13.TRICH
n X14.MMIB
o X15.QD50
P X16.QD16
q X17.xembed
269 ecoR: 0.50 > O/E > 1.50
ANC+COLOR+DOC+COND+CA
Ecoregions %
0
20
40
60
80
100
0
1
2
3
4
O/E WQI


Regional Distribution of WQI (grp 3)
a X1.PHSTVL
b X2.COND
c X3.ANC
d X4.TURB
e X5.TSS
f X6.COLOR
g X7.DOC
h X8.PTL
i X9.CA
j X10.SO4
k X11.NO3
l X12.CL
m X13.TRICH
n X14.MMIB
o X15.QD50
P X16.QD16
q X17.xembed
CL+QD50+QD16+TSS+SO4+PTL+NO3+TURB
Ecoregions %
0
20
40
60
80
100
0
1
2
3
4
O/E WQI


Regional Distribution of WQI (1 to 17)
Ecoregions %
O/E WQI
0
20
40
60
80
100
0
1
2
3
4
a X1.PHSTVL
b X2.COND
c X3.ANC
d X4.TURB
e X5.TSS
f X6.COLOR
g X7.DOC
h X8.PTL
i X9.CA
j X10.SO4
k X11.NO3
l X12.CL
m X13.TRICH
n X14.MMIB
o X15.QD50
p X16.QD16
q X17.xembed

Conclusion

- Successful spatial prediction of WSA water quality (*WQI*), using *SC* and *LC* as independent predictors and *WQI* as co-predictors, was demonstrated by 1392 tests, using set-aside watersheds.

- Also successful was replacing set-aside “host” data in 1392 tests with data from the nearest neighbors of each host watershed, which supported transfer of *WQI* data to Ecoregions.

63 Water Quality indicators (*W*) , Soil Characteristics and land cover (*SL*) of Watersheds, and 1,392 observations summarized by 63 sample mean ***μ*** *and 63x63 covariance* Σ

Watershed *1*	*W11, W12, …, W1,17, SL1,18, …,SL1,51 …, SL1p*
Watershed *2*	*W21, W22, …, W2,17, SL2,18, …,SL2,51…, SL2p*
p = 17+34+12 = 63, n = 1,392	
Watershed *n*	*Wn1, Wn2, …, Wn,17, SLn,18, …SLn,51, …, SLnp*
Mean ***μ***	μ_1, μ_2, …., , μ_p
Variance ***σ***	σ_{11}, σ_{22}, …., , σ_{pp}
Covariance Σ	*p-dimensional and symmetric, with the diagonal =* ***σ***

Elements of the variance-covariance matrix Σ_{pxp} prior to splitting into *predictor* and *predicted* parts

σ_{11}	σ_{12}			$\sigma_{1,p-1}$	σ_{1p}
σ_{21}	σ_{22}			$\sigma_{2,p-1}$	σ_{2p}
σ_{p1}	σ_{p2}			$\sigma_{p,p-1}$	σ_{pp}

Partitioned $\boldsymbol{\Sigma_{pxp}}$ with $\boldsymbol{\Sigma_{1x1}}$ along with its covariance $\boldsymbol{\Sigma_{21}}$ and $\boldsymbol{\Sigma_{12}}$ isolated from the predictor variables $\boldsymbol{\Sigma_{22}}$

$\boldsymbol{\Sigma_{11}}=\sigma_{11}$	$\boldsymbol{\Sigma_{12}}=[\sigma_{12} \cdots \sigma_{1p}]$
$\boldsymbol{\Sigma_{21}}$	Predictors: $\boldsymbol{\Sigma_{22}}$

Regression with predictor origins=μ, conditional probability, Variance Reduction (VR), and Squred Distance (SL_sqrd)

- $x_1 = \mu_1 + b_2 (x_2 - \mu_2) + \cdots + b_p (x_p - \mu_p)$
- *x_1 differs from μ_1 by the degree of its correlation with all predictors, thus,x_1* = conditional expectation.
- $\boldsymbol{E}(x_1 \mid x_2 \cdots x_p) = \boldsymbol{\mu_1 + \Sigma_{12} \Sigma^{-1}_{22} (x_{p-1} - \mu_{p-1})}$
- *Similarly, Conditional Variance is:*
- $\boldsymbol{V_c}(x_1 \mid x_2, .., x_p) = \boldsymbol{\sigma_{11} - \Sigma_{12} \Sigma^{-1}_{22} \Sigma_{21}}$
- *Variance Reduction VR compares* $\boldsymbol{V_c}$ *with sample variance* $\boldsymbol{V_s}$
- $\boldsymbol{VR = 100[1.0 - (V_c / V_s)^{1/2}]}$
- $\boldsymbol{SL\_sqrd} = \boldsymbol{(x - \mu)' \Sigma^{-1} (x - \mu)}$