Bayesian Methods for Regional Eutrophication Models using the Nutrient Criteria Database

E. Conrad Lamon III

Nicholas School of the Environment and Earth Sciences

Duke University

Craig A. Stow

Department of Environmental Health Sciences

University of South Carolina

NICHOLAS SCHOOL OF THE ENVIRONMENT AND EARTH SCIENCES

DUKE UNIVERSITY

This research is funded by

U.S. EPA - Science To Achieve Results (STAR) Program

Grant # RD-83088701-0

Goals and Objectives

- Use modern classification and regression trees and hierarchical Bayesian techniques to link multiple environmental stressors to biological responses and quantify uncertainty in model predictions and parameters.

Guidance for TMDL model selection (NRC 2001)

- report prediction uncertainty
- be consistent with the amount of data available
- flexible enough to permit updates and improvements

Overview

- Goals and Objectives
- Approach
- Preliminary Findings
- Next Steps

Approach

Methods

- Classification And Regression Trees (CART),
- it's Bayesian analogue, BCART
- a recently developed enhancement to the BCART procedure, which includes BCART as a model subclass, known as Bayesian Treed (BTREED) models, and
- Bayesian Hierarchical Models

Tree based methods

- are a flexible approach useful for variable subset selection,
- when the analyst suspects global non-linearity,
- and cannot (or does not want to) specify the functional form of possible interactions *a priori*.

Bayesian Treed models

- Bayesian Hierarchical model to:
  - Select subsets on $X \rightarrow X_s$
  - Fit linear models to these subsets X_s
- Tree structured models
  - "ANOVA in Reverse"

- "Leaves" contain linear models, not just a mean (like in CART models)

Bayesian Treed model search

- MCMC used to stochastically search for high posterior probability trees *T*.
- Metropolis –Hastings algorithm simulates a Markov chain with limiting distribution $p(T|Y,X)$
- Chipman, George and McColloch, 2000, JASA.

  http://gsbwww.uchicago.edu/fac/robert.mcculloch/research/papers/index.html

BTREED Models

- Were used by Lamon and Stow, 2004, Water Research, 38(11): 2764-2774.
- Used with EPA Nutrient Criteria Database Freeman et al, 2005, *in Review*, Environmetrics
- Used with Finnish Lakes data, Lamon and Malve, 2005, in prep.

Data

- Response variables may be
  - either continuous (such as biological indices of abundance) or
  - discrete (such as designated use attainment classes).

EPA NCD: response variable is $\log_{10}$ Chlorophyll *a* concentration.

Data

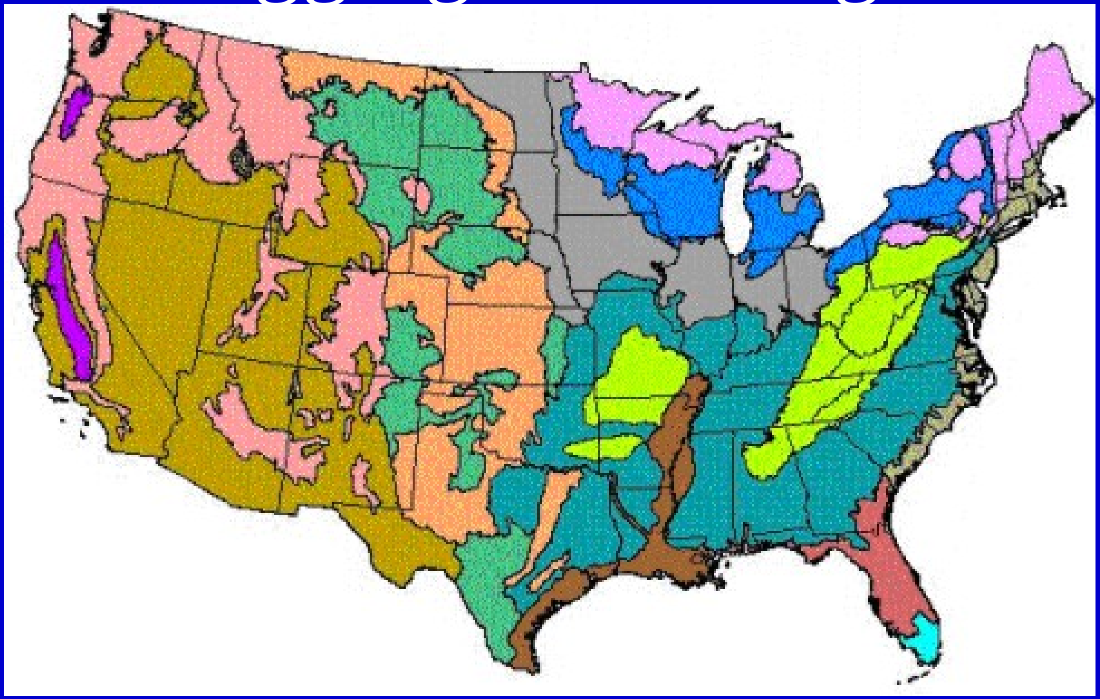
Predictor variables in tree based methods may also be continuous or discrete, and may include :

source agency, basin, sub-watersheds, states, EPA regions, latitude and longitude, and many continuous predictors related to water chemistry, water use, discharges or pollutant loading.

For the NCD, we are using ecoregion, waterbody type (lake or res.), Month, ***TNsrc, Chlasrc*** in the tree, and

$\log_{10}$TP and $\log_{10}$TN in the endnode LM's.

Aggregate Ecoregions



I – Willamette and central valley

II - Western Forested Mountains

III – Xeric west

IV Great Plains Grass & Shrublands

V South Central Cult. Great Plains

VI Corn Belt & N. great plains

VII - Mostly Glaciated Dairy Region

VIII - Nutrient Poor Largely Glaciated Upper Midwest and Northeast

IX - Southeastern Temperate Forested Plains and Hills

X - Texas-Louisiana Coastal &Mississippi All. Plains

XI - Central and Eastern Forested Uplands

XII - Southern Coastal Plain

XIII - Southern Florida Coastal Plain

XIV - Eastern Coastal Plain

Data

- > 656,000 observations in the NCD (!)
- 98,169 have non-missing TP measurements
- For these observations, four methods of chlorophyll determination were used (STORET 32211, 32209, 32210, 32230)
- Three methods of nitrogen determination existed for these observation (TKN 00625, TN 00600 and NO2NO3+TON 00630+00605)

Data

- The different methods were combined into one variable while creating new categorical variables to keep track of the source method.
- Chlsrc = a for STORET 32211

  b 32209

  c 32210

  d 32230
- TNsrc = a for TKN

  b TN

  c NO2NO3+TON

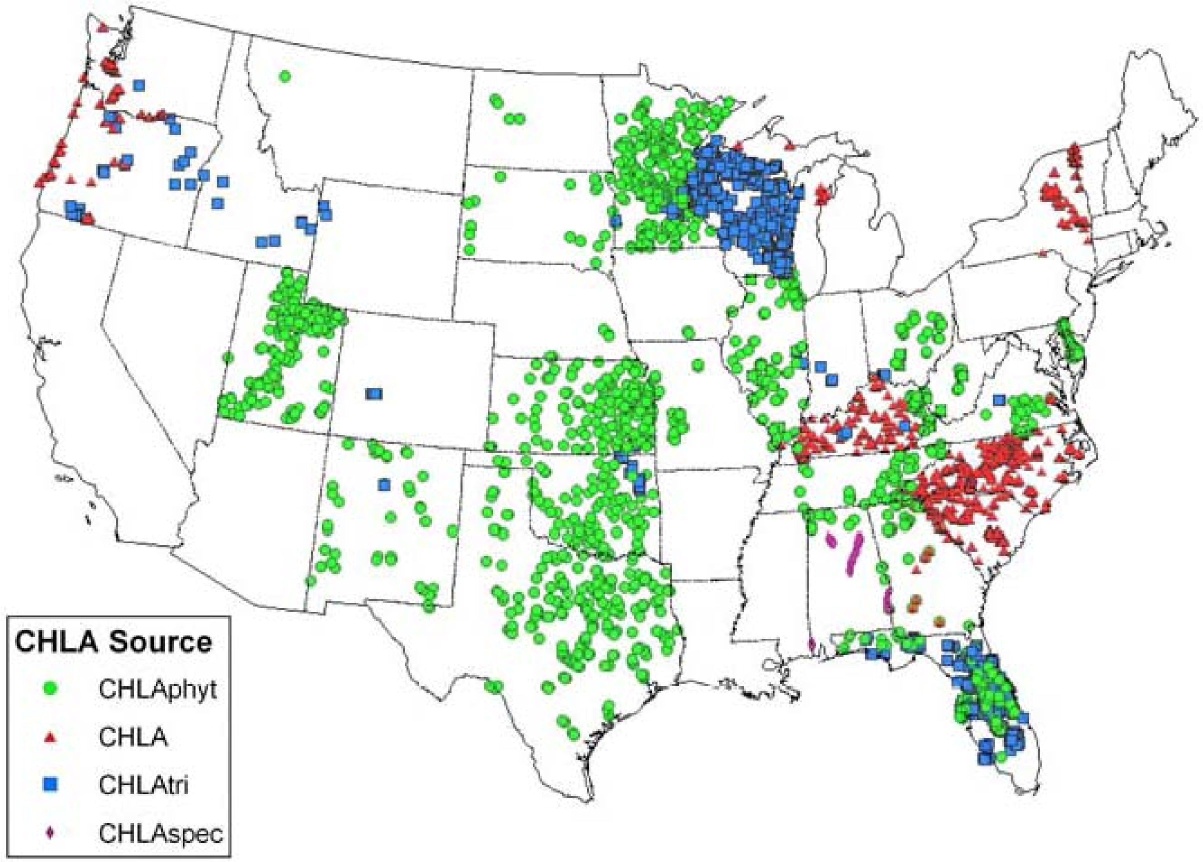


Figure 1 Geographical distribution of Chlorophyll by source type, CHLAphyt, n = 18968; CHLA, n = 10945; CHLAtri, n = 62563; CHLAspec, n = 5691 (total n = 98167)

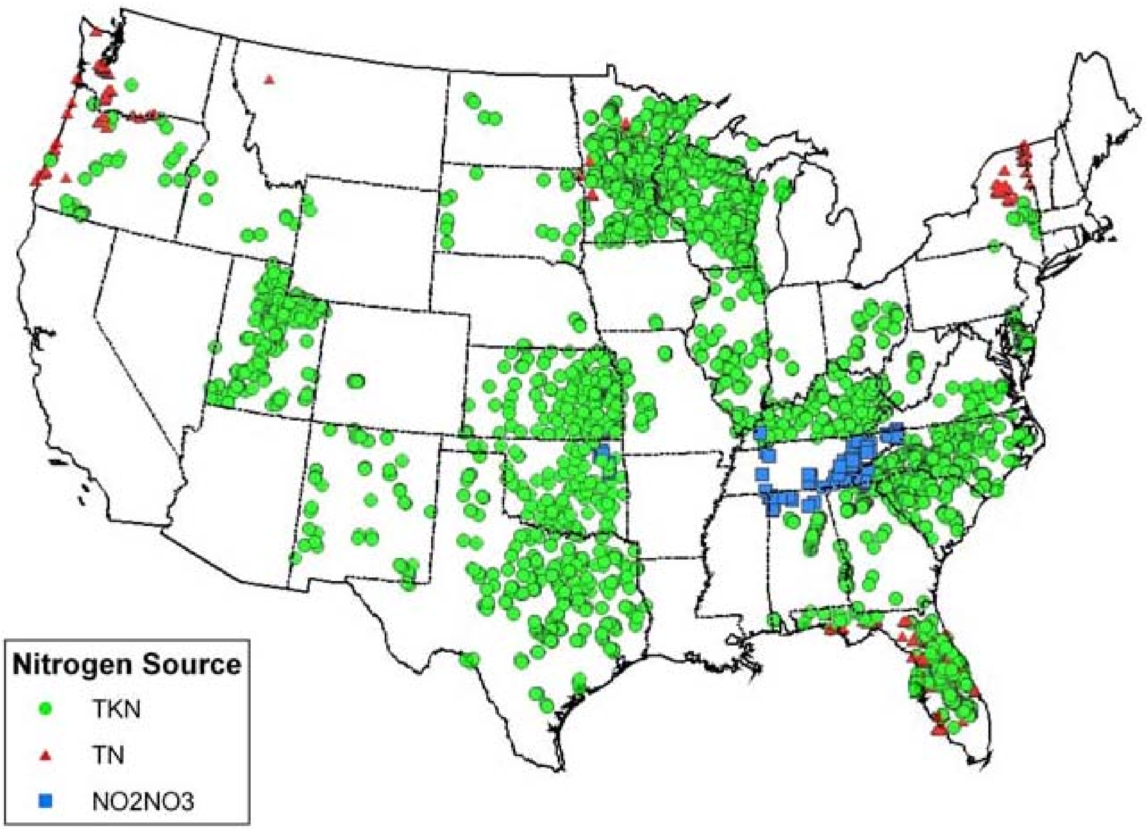


Figure 2 Geographical distribution of Nitrogen by source type, TKN, n = 46241; TN, n = 49792; NO2NO3, n = 2134 (total n = 98167)

This research is funded by
U.S. EPA - Science To Achieve Results (STAR) Program
Grant # RD-83088701-0

Preliminary Findings

Freeman, Lamon and Stow, 2005. Regional Nutrient-Chlorophyll Relationships in Lakes and Reservoirs: a Bayesian TREED Model Approach, *in review*, *Environmetrics*.

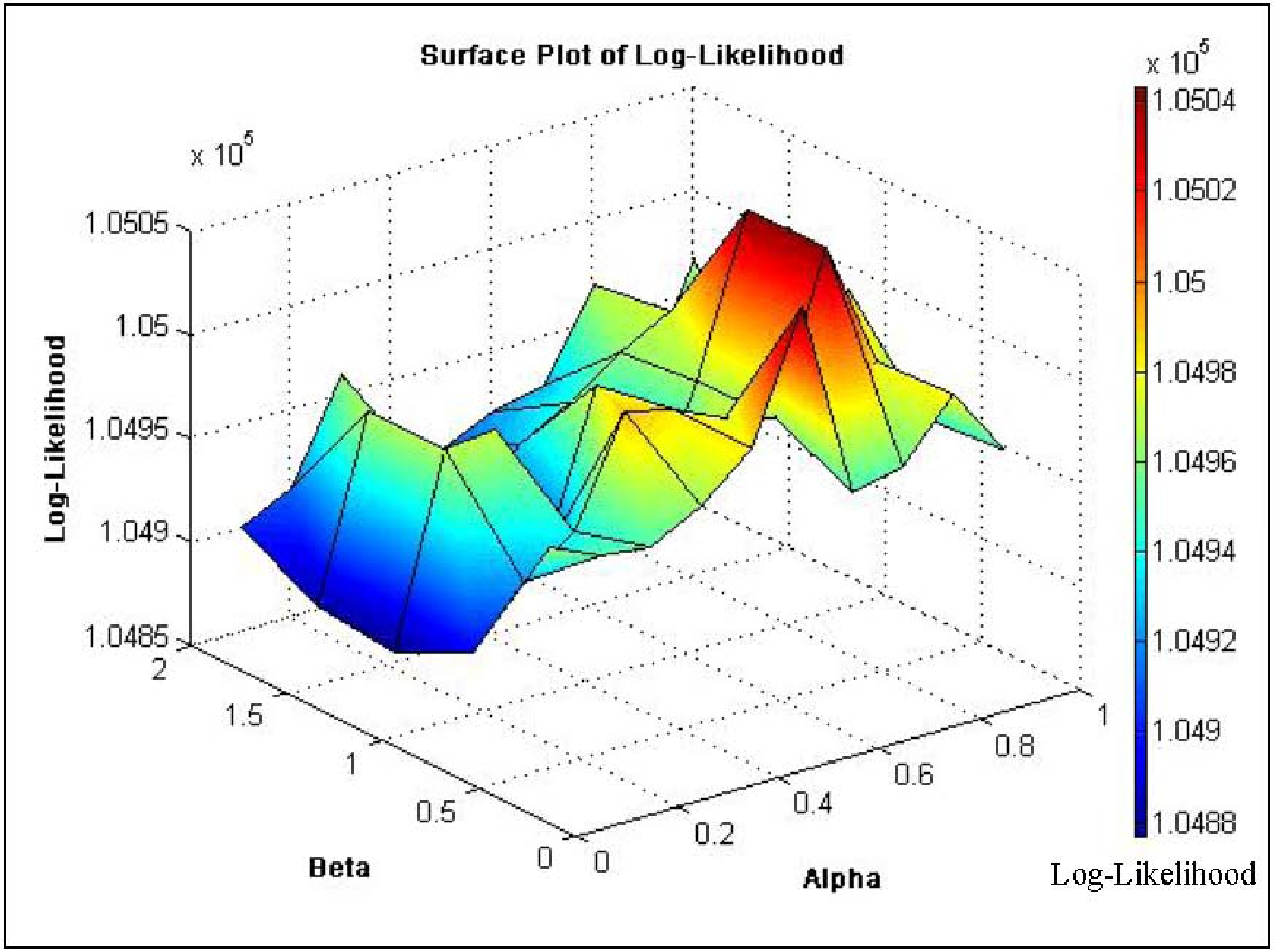


Figure 5 Surface plot of alpha, beta versus log-likelihood (alpha grid from 0.2 to 1.0, step 0.1, beta grid from 0.4 to 2.0, step 0.4). Maximum Log-Likelihood of 105044.6025 at alpha = 0.6, beta = 0.4, tree size 103. Second highest Log-Likelihood of 105041.4123 at alpha = 0.8, beta = 0.8, and at alpha = 0.8, beta = 1.2, both tree size 119.

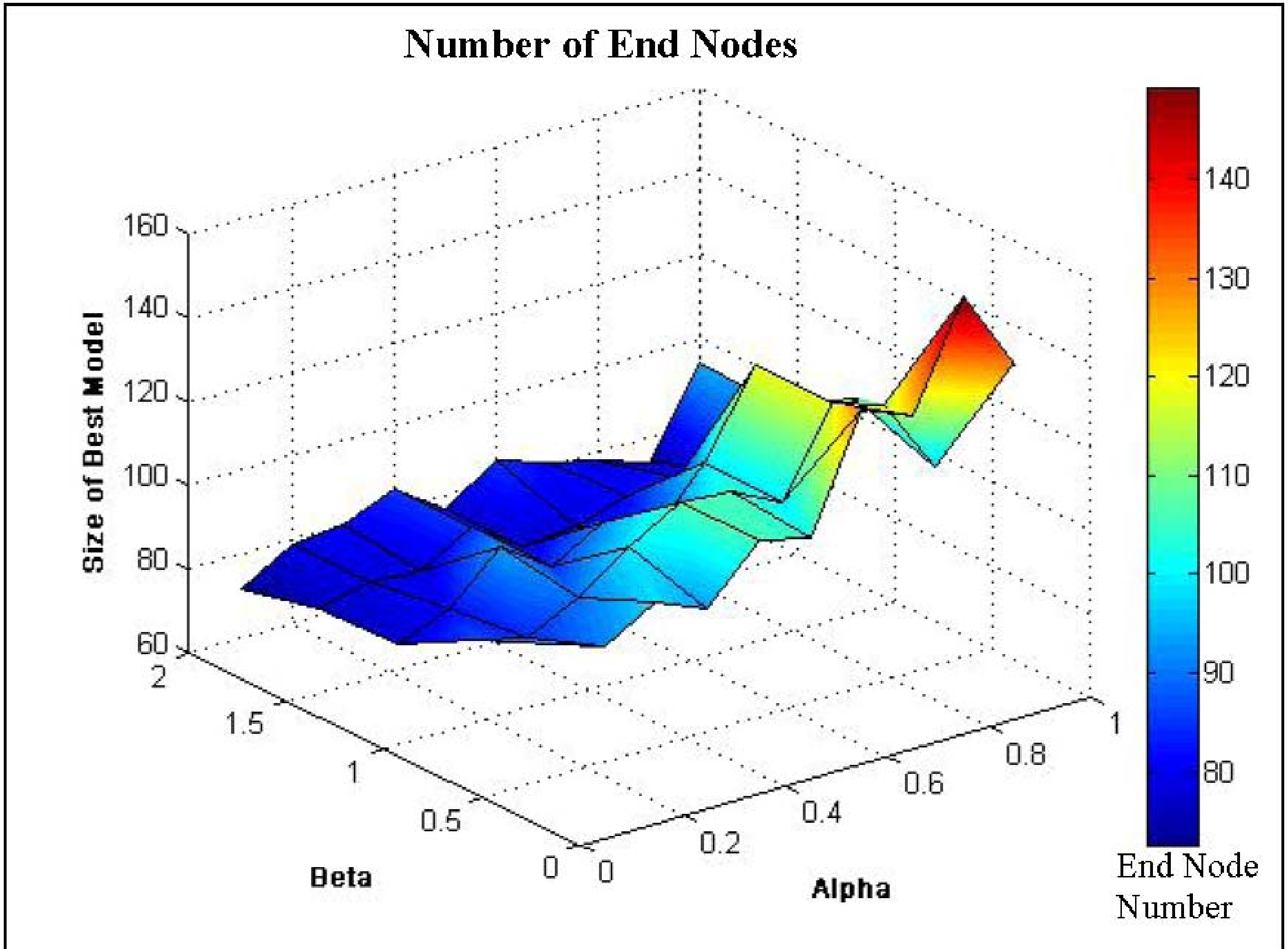


Figure 6 Surface plot of alpha, beta number of end nodes in TREE (alpha grid from 0.2 to 1.0, step 0.1, beta grid from 0.4 to 2.0, step 0.4)

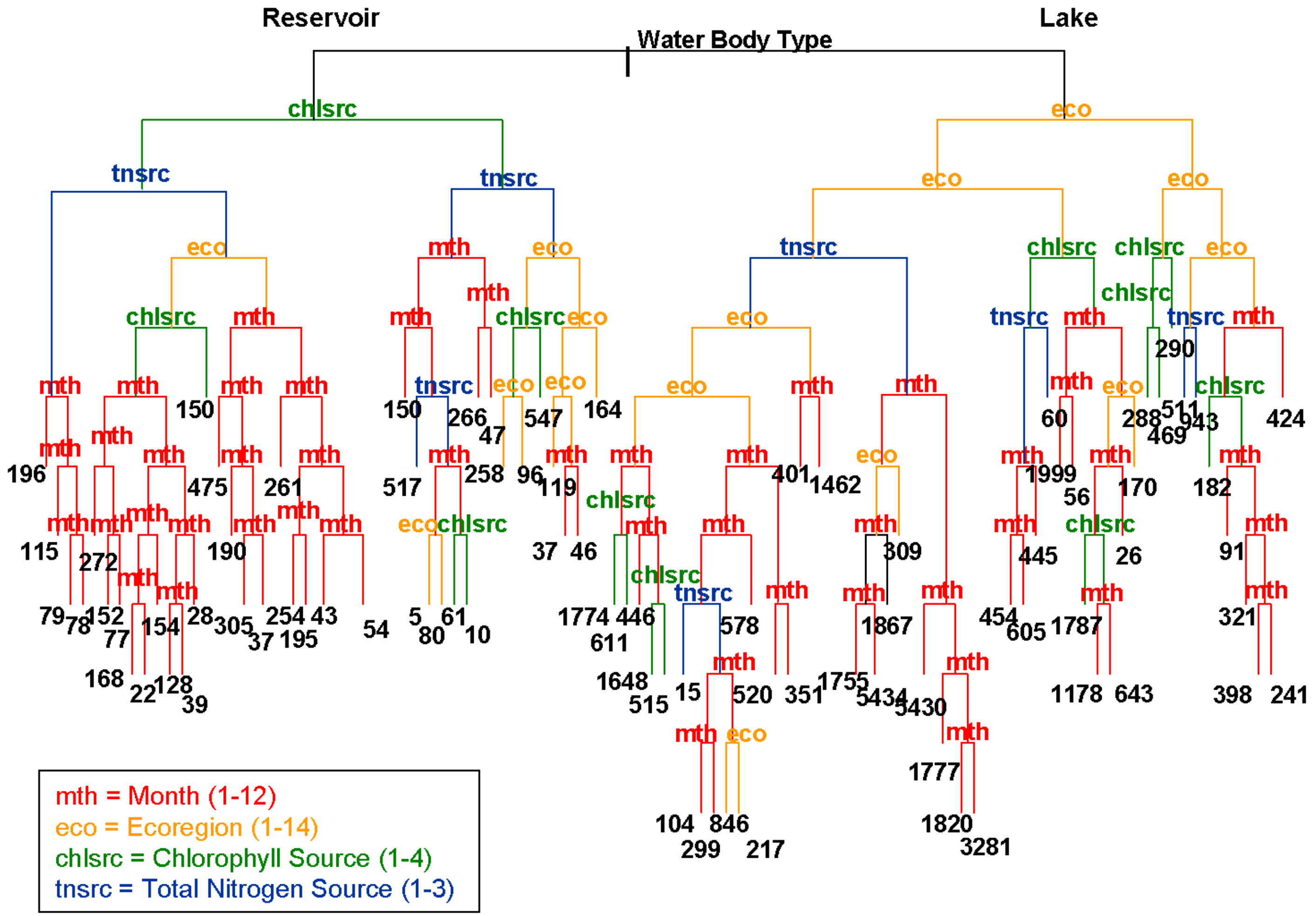
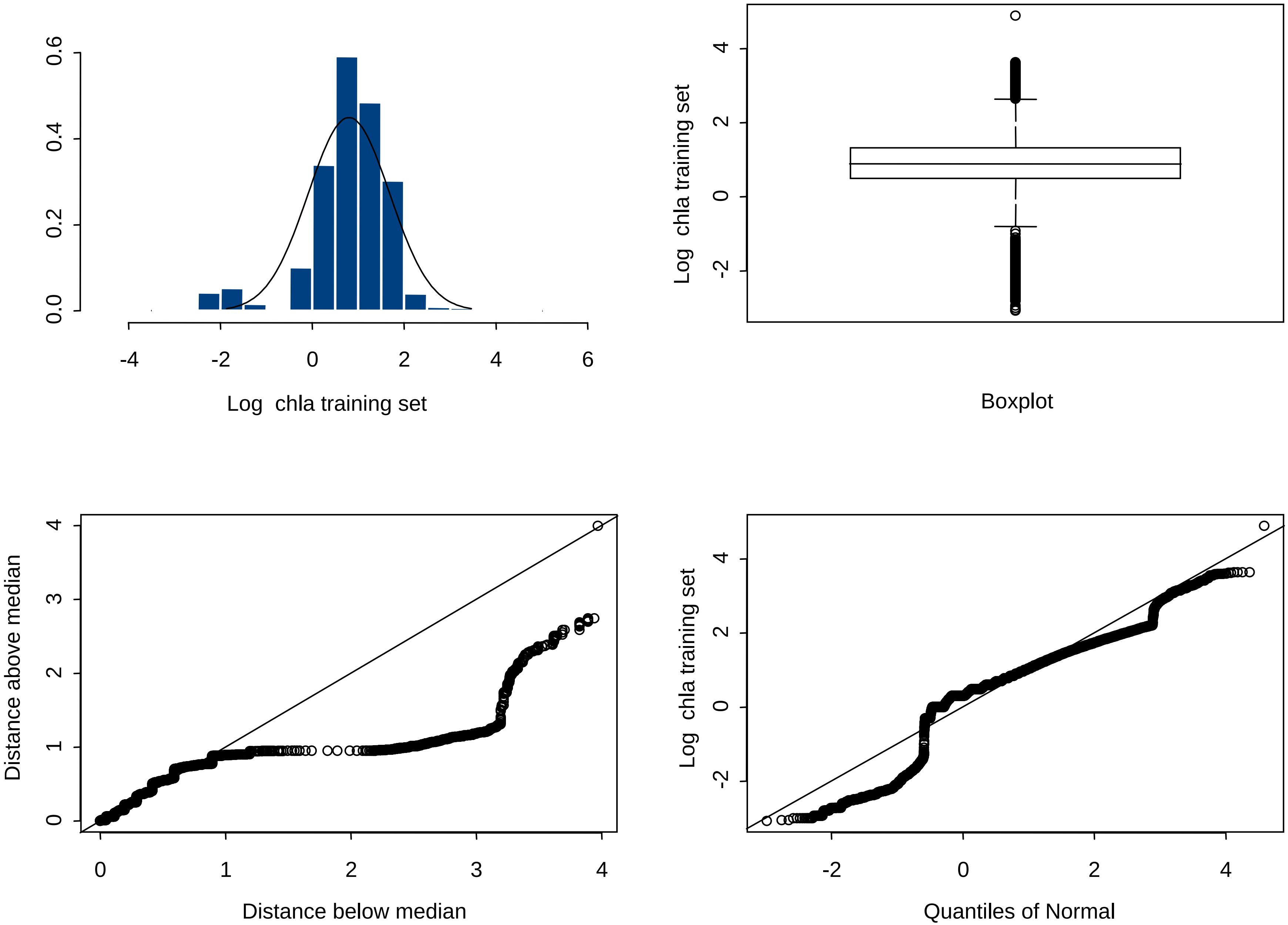


Figure 7 Dendrogram of BTREED Highest Log-Likelihood Model (alpha = 0.6, beta = 0.4, Log-Likelihood = 105044.6). Zero nodes removed in plot; the number of observations are given below each terminal node

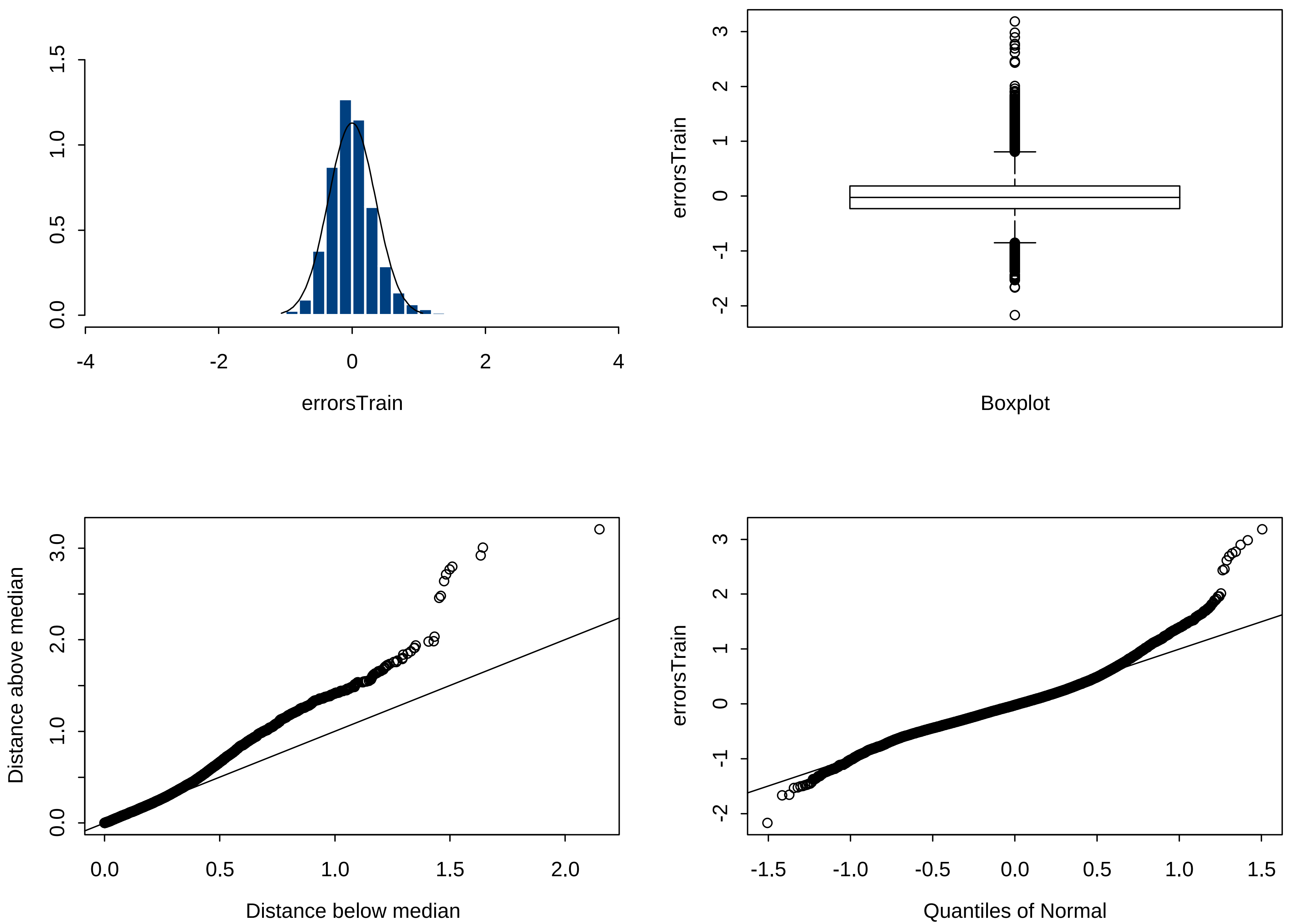
Model fit and “out of sample” predictive ability

Table 4 Mean squared error (MSE) and median absolute deviation (MAD) of training data and test data fits, and Relative Efficiency for four highest log-likelihood Bayesian TREED models

Alpha	Beta	Log-Likelihood	MAD Error Training	MAD Error Test	MSE Error Training	MSE Error Test	Relative Efficiency
0.6	0.4	105044.6	0.20873	0.21207	0.12478	0.12751	1.02196
0.8	0.8	105041.4	0.20954	0.21253	0.12498	0.12776	1.02223
0.8	1.2	105041.1	0.20954	0.21253	0.12498	0.12776	1.02223
0.4	0.8	104990.4	0.20765	0.21230	0.12527	0.12795	1.02140



0.0
0.2
0.4
0.6
-4
-2
0
2
4
6
Log chla training set
Log chla training set
4
2
0
-2
Boxplot
Distance above median
0
1
2
3
4
Distance below median
Log chla training set
-2
0
2
4
Quantiles of Normal



0.0
0.5
1.0
1.5
-4
-2
0
2
4
errorsTrain
errorsTrain
-2
-1
0
1
2
3
Boxplot
Distance above median
0.0
1.0
2.0
3.0
0.0
0.5
1.0
1.5
2.0
Distance below median
errorsTrain
-2
-1
0
1
2
3
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Quantiles of Normal

Geomorphological typology of Finnish Lakes
Finnish Environment Institute (SYKE)

Lake Type	Name	Details
I	Large, non-humic lakes	SA > 4000 Ha, color < 30
II	Large, humic lakes	SA > 4000 Ha, color > 30
III	Medium and small, non-humic lakes	SA: 50 – 4000 Ha, color < 30
IV	Medium Area, humic deep lakes	SA: 500 – 4000 Ha, color: 30-90, D > 3 m
V	Small, humic, deep lakes	SA: 50 – 500 Ha, color: 30-90, D > 3 m
VI	Deep, very humic lakes	Color > 90, D > 3 m
VII	Shallow, non-humic lakes	Color < 30, D < 3 m
VIII	Shallow, humic lakes	Color: 30-90, D < 3 m
IX	Shallow, very humic lakes	Color > 90, D < 3 m

Finnish Lake Data

The response variable is chlorophyll *a* ($\mu g\ L^{-1}$), a surrogate for algal biomass.

The tree portion includes the variables :

- altitude (m),
- latitude (decimal degrees),
- surface area (km^2),
- mean depth (m), and
- color ($mg\ Pt\ L^{-1}$).

Predictor variables used in the endnode models were :

- total nitrogen (TN, $\mu g\ L^{-1}$) and
- total phosphorus (TP, $\mu g\ L^{-1}$).

We log (base *e*) transformed TP, TN and Chla for use in fitting the endnode regressions, then took the annual averages by lake, providing 280 growing season lake-wide averages.

Model 2

lnChla ~

Lake Type

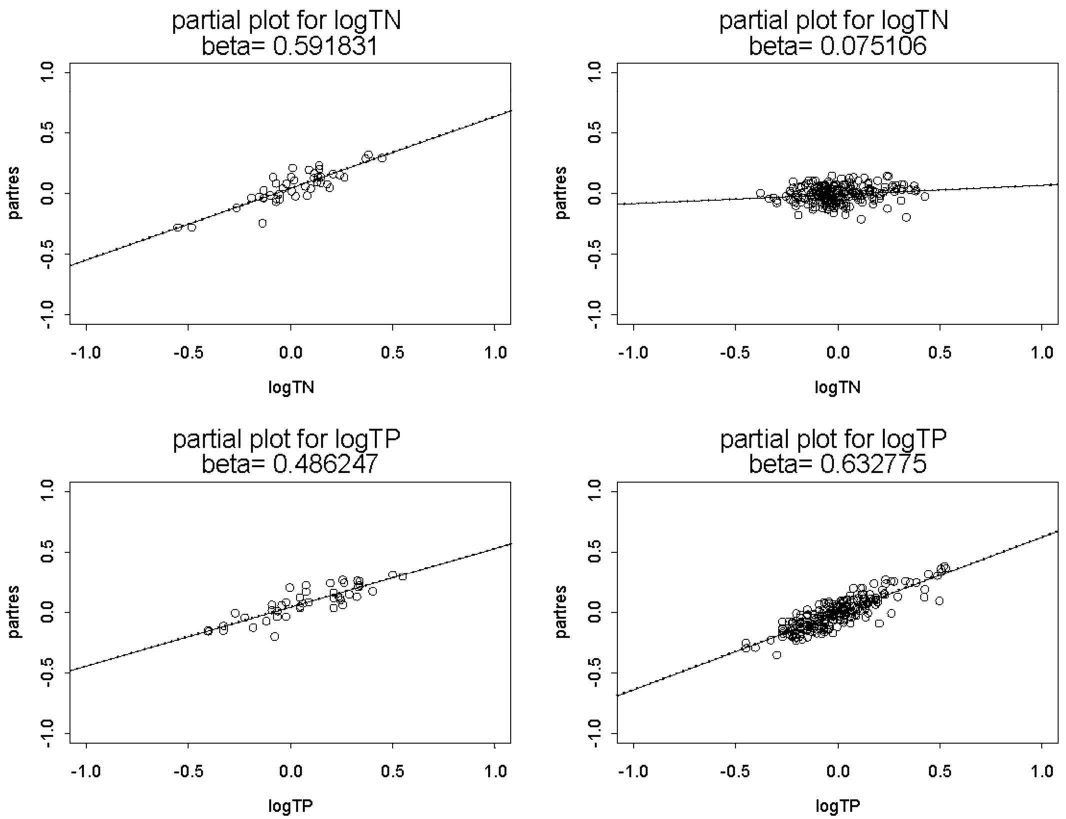
Types 3,5,8,9

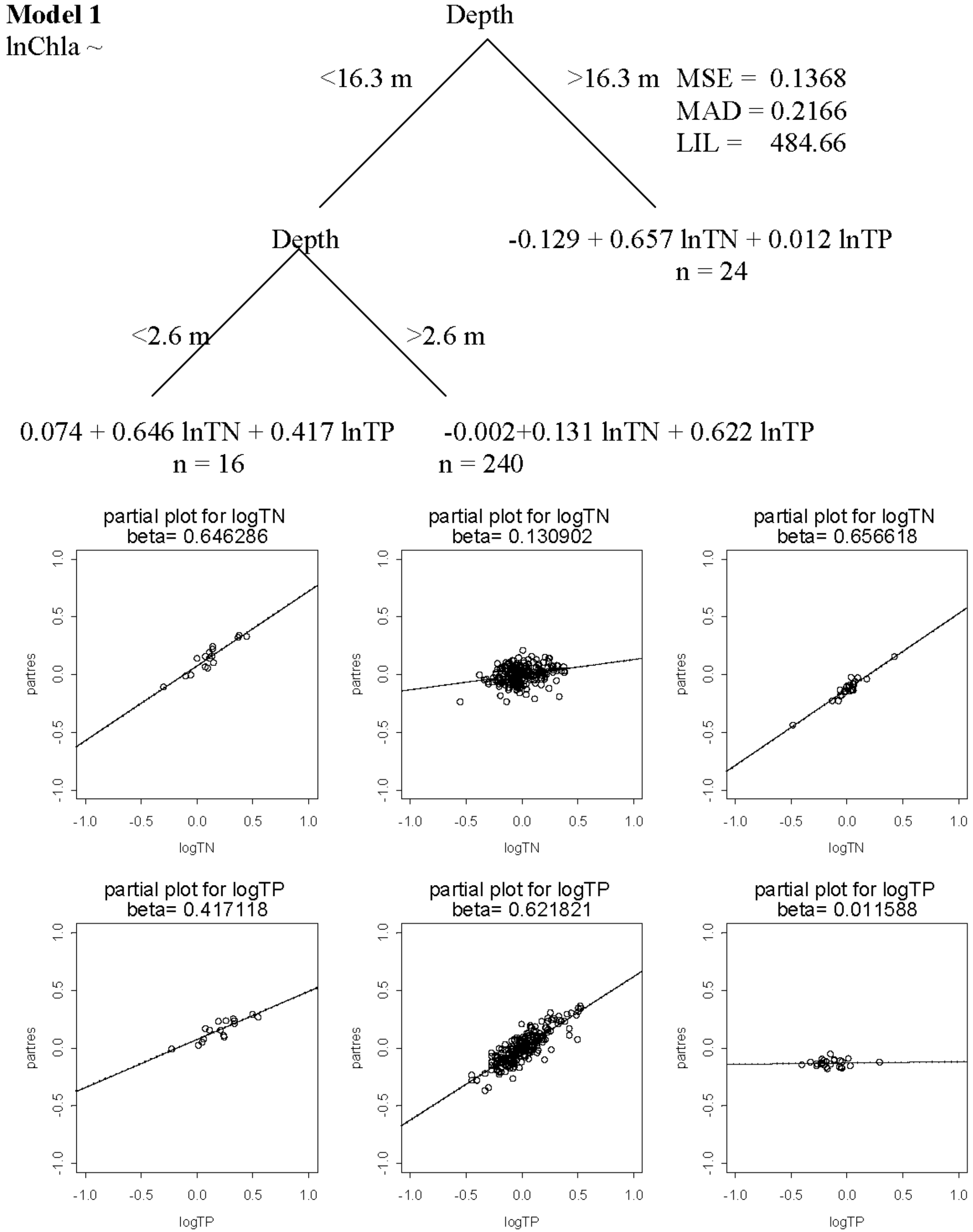
Types1,2,4,6,7

MSE = 0.1370
MAD = 0.2461
LIL = 483.2

$0.043 + 0.592 \ln TN + 0.486 \ln TP$
n = 43

$-0.008 + 0.075 \ln TN + 0.633 \ln TP$
n = 237





Model 1
lnChla ~
Depth
<16.3 m
>16.3 m
MSE = 0.1368
MAD = 0.2166
LIL = 484.66
Depth
-0.129 + 0.657 lnTN + 0.012 lnTP
n = 24
<2.6 m
>2.6 m
0.074 + 0.646 lnTN + 0.417 lnTP
n = 16
-0.002+0.131 lnTN + 0.622 lnTP
n = 240
partial plot for logTN
beta= 0.646286
partial plot for logTN
beta= 0.130902
partial plot for logTN
beta= 0.656618
partial plot for logTP
beta= 0.417118
partial plot for logTP
beta= 0.621821
partial plot for logTP
beta= 0.011588
partres
logTN
logTP

Next Steps

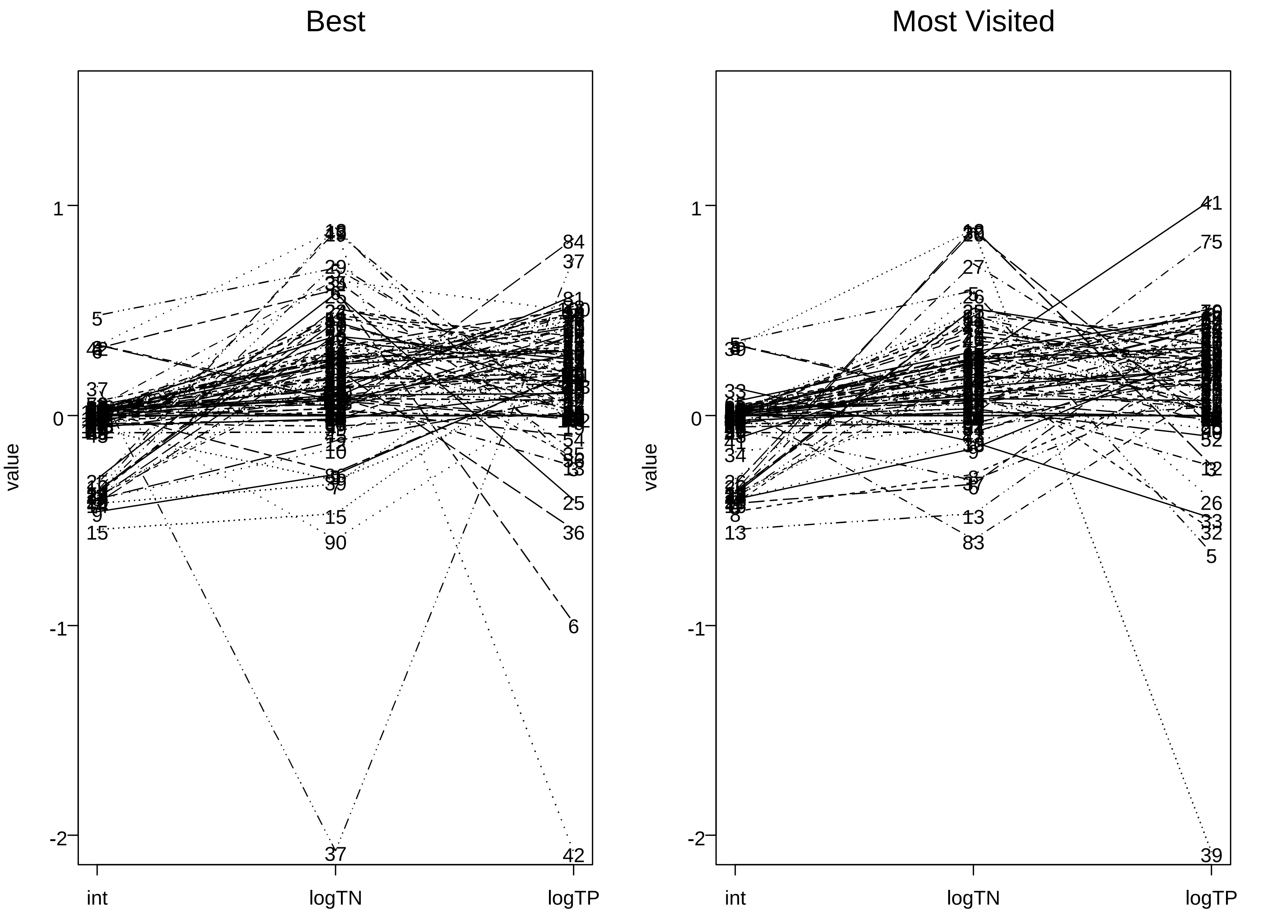
Next Steps

- More predictor variables for endnode models
- Use resultant tree structures to identify important hierarchical structure (Ecoregion, Chlasrc and TNsrc, seasonality, etc.)
- Explore these structures with other Hierarchical Bayesian methods
- Non-linear specification? Spline basis functions (HBM), or in leaf model or inclusion of all predictors in tree
- Tools
- Collaborations

Thanks!

This research is funded by
U.S. EPA - Science To Achieve Results (STAR) Program
Grant # RD-83088701-0

- EPA STAR program for funding.
- Hugh Chipman, Univ. of Waterloo and Robert McColloch, University of Chicago for BCART/BTREED computer code.
- Anindita Das, Angelina Freeman, Jessica Rury, Boknam Lee


Best
Most Visited
value
int
logTN
logTP

Results using new scale prior

α	β	LIL	msetr	madtr	msetest	madtest	leaves	Rel.Eff.
0.50	1.50	104012.5	0.1290	0.2123	0.1305	0.2154	59	1.012
0.65	1.00	**104050.1**	0.1291	**0.2115**	0.1303	0.2154	56	1.009
0.85	2.00	104027.9	0.1298	0.2130	0.1313	0.2177	**48**	1.012
0.90	0.50	104013.5	0.1295	0.2121	**0.1300**	0.2155	76	1.004
0.90	1.00	104043.4	**0.1289**	0.2128	0.1302	**0.2150**	61	1.010

Bayesian TREED Model search (specifics)

$$p(Y \mid X, T) = \int p(Y|X, \theta, T) p(\theta, T) d\theta$$

$$= \prod_{i=1}^{b} \int \prod_{j=1}^{n_i} p(y_{ij} \mid x_{ij}, \theta_i) p(\theta_i) d\theta_i$$

Eq (1)

Bayesian TREED Model search (specifics)

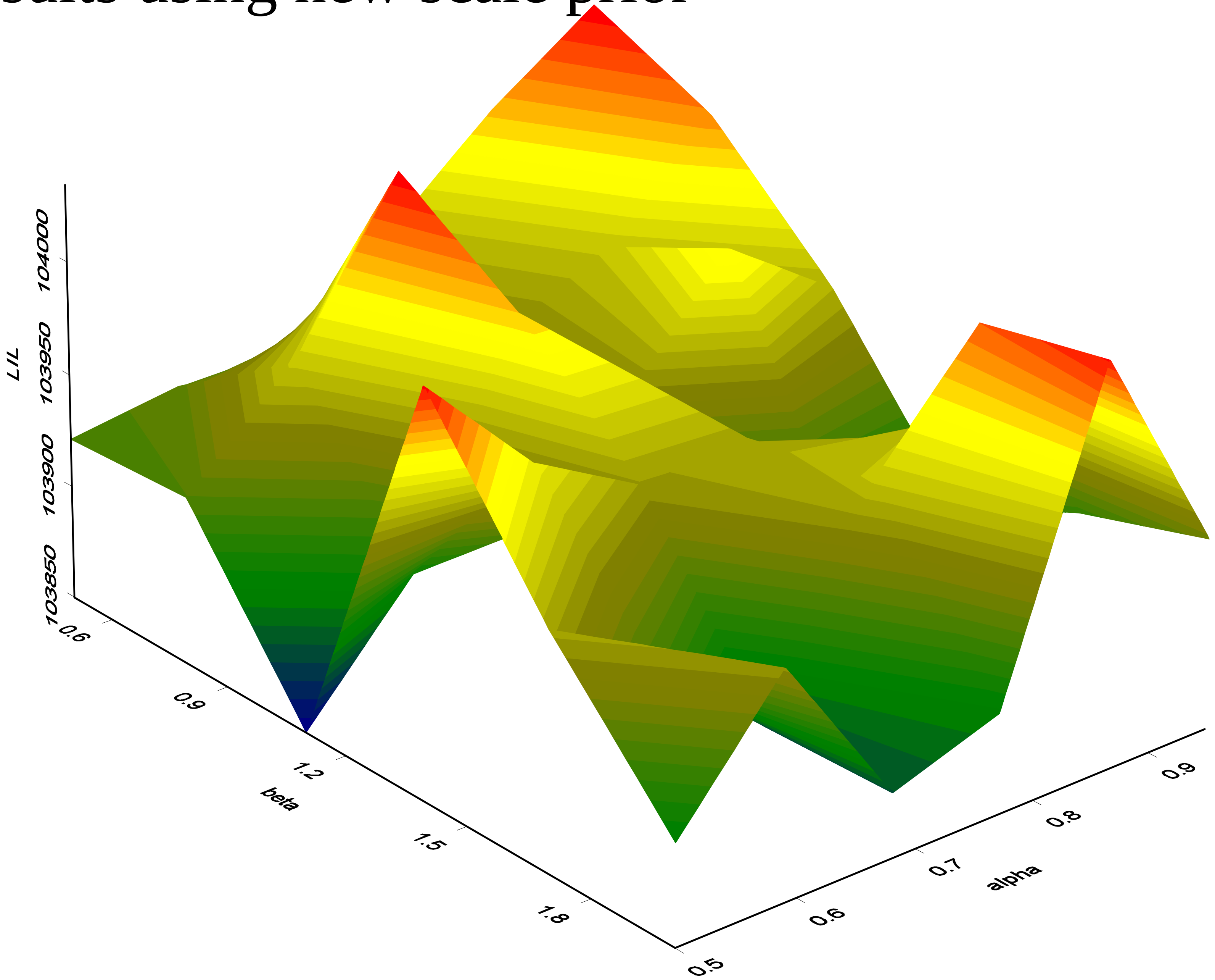
Start with initial tree T^0, iteratively simulate the transitions from T^i to T^{i+1} by two steps:

1. Generate a candidate value T^* with probability distribution $q(T^i, T^*)$.
2. Set $T^{i+1} = T^*$ with probability

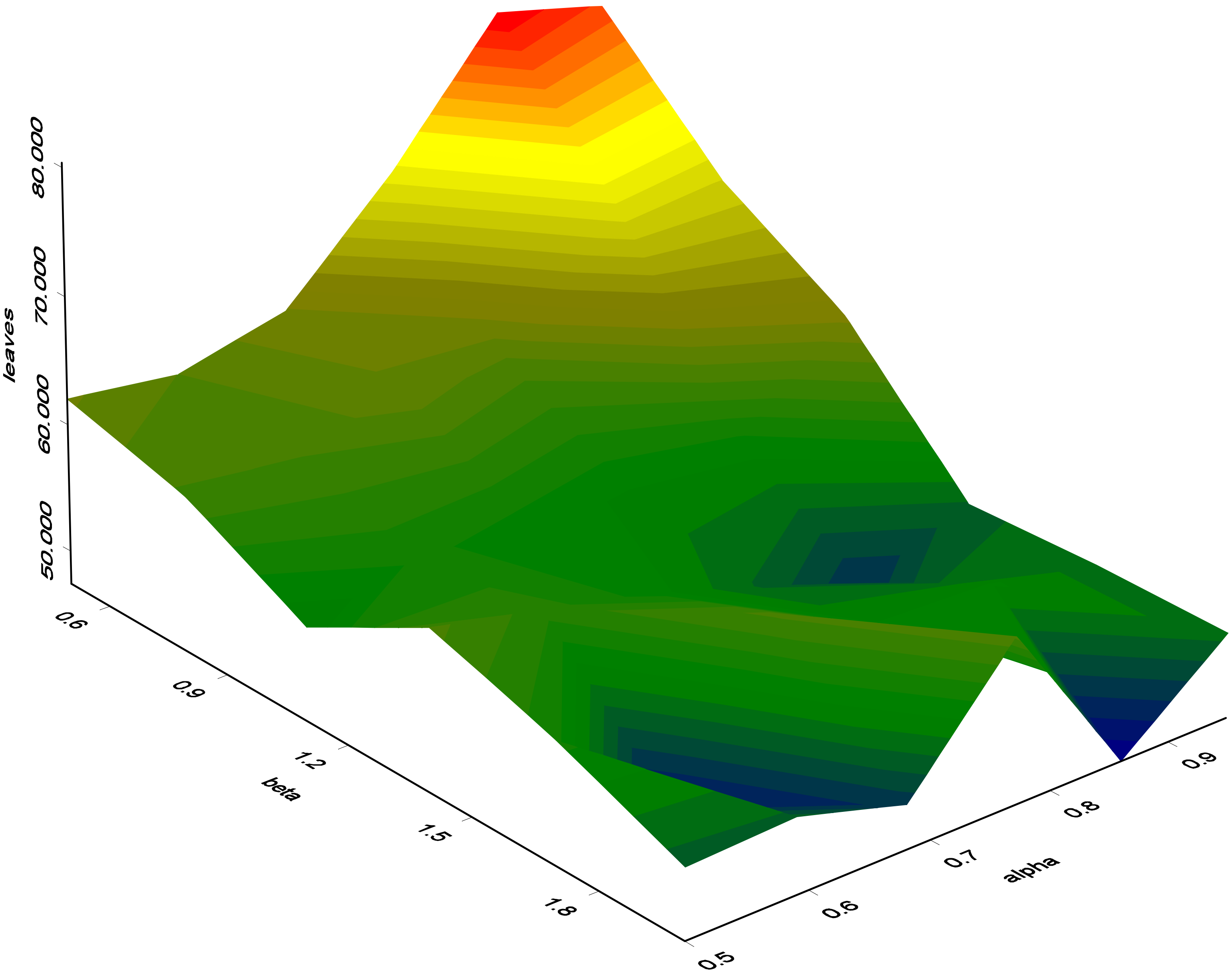
$$\alpha(T^{i+1}, T^*) = \min\left\{\frac{q(T^*, T^i)\,p(Y \mid X, T^*)\,p(T^*)}{q(T^i, T^*)\,p(Y \mid X, T^i)\,p(T^i)}, 1\right\}$$

Else set $T^{i+1} = T^i$

Results using new scale prior



Results using new scale prior



Lakes and ponds, water body type = 5