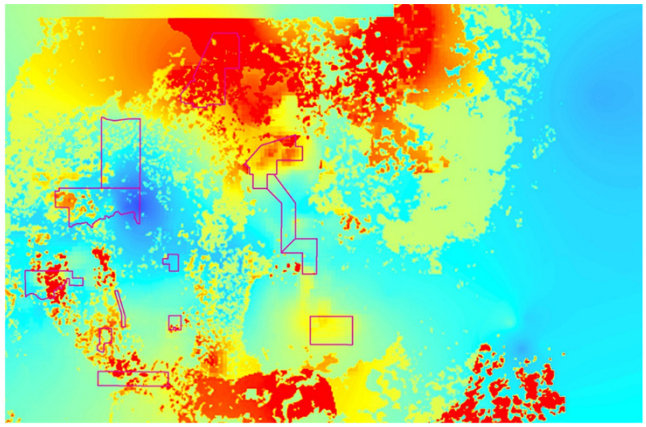
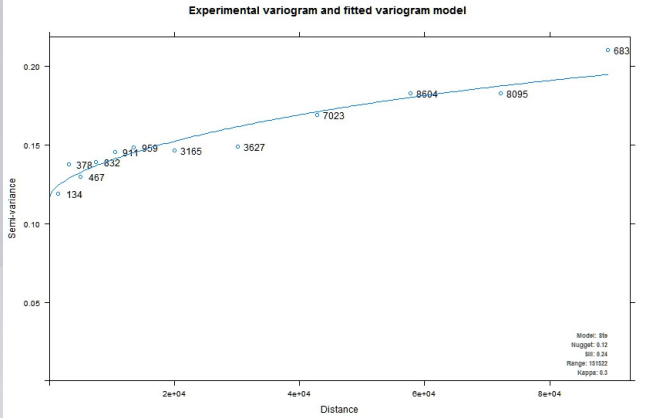
Regression Kriging Model for Improved Environmental Data Science Interpolation



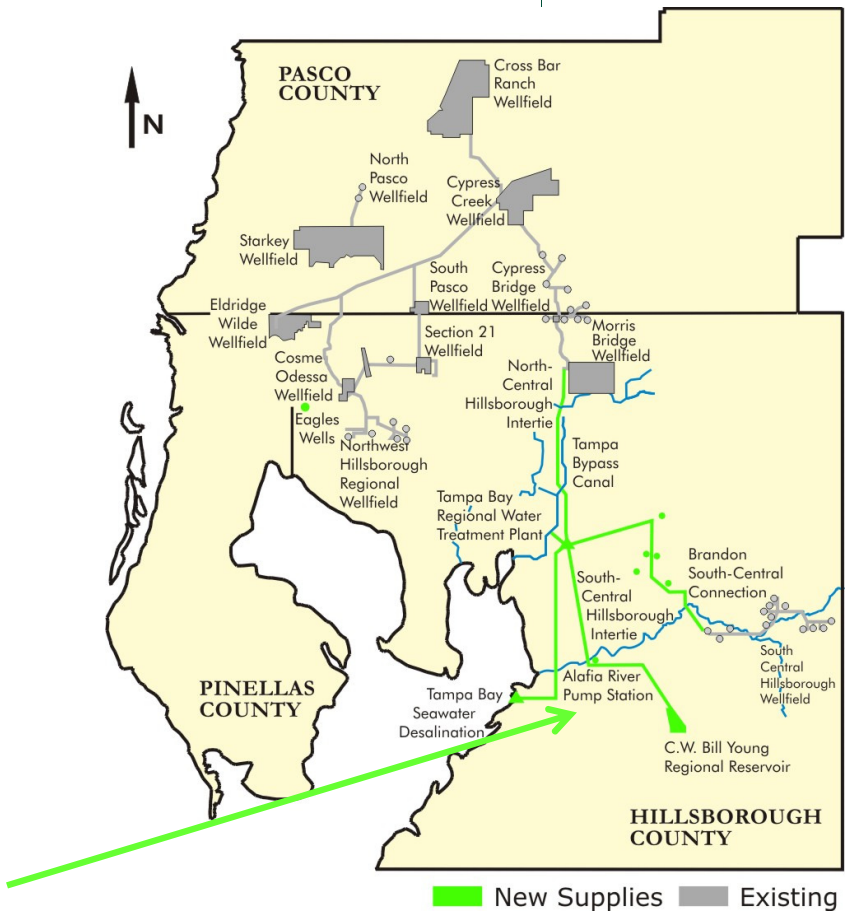
Dan Schmutz, MS
Chief Environmental Scientist



Orlando Machine Learning and Data Science Meetup
Melrose Center, Orlando
September 7, 2019 3:30 – 5:00 pm

Tampa Bay Water

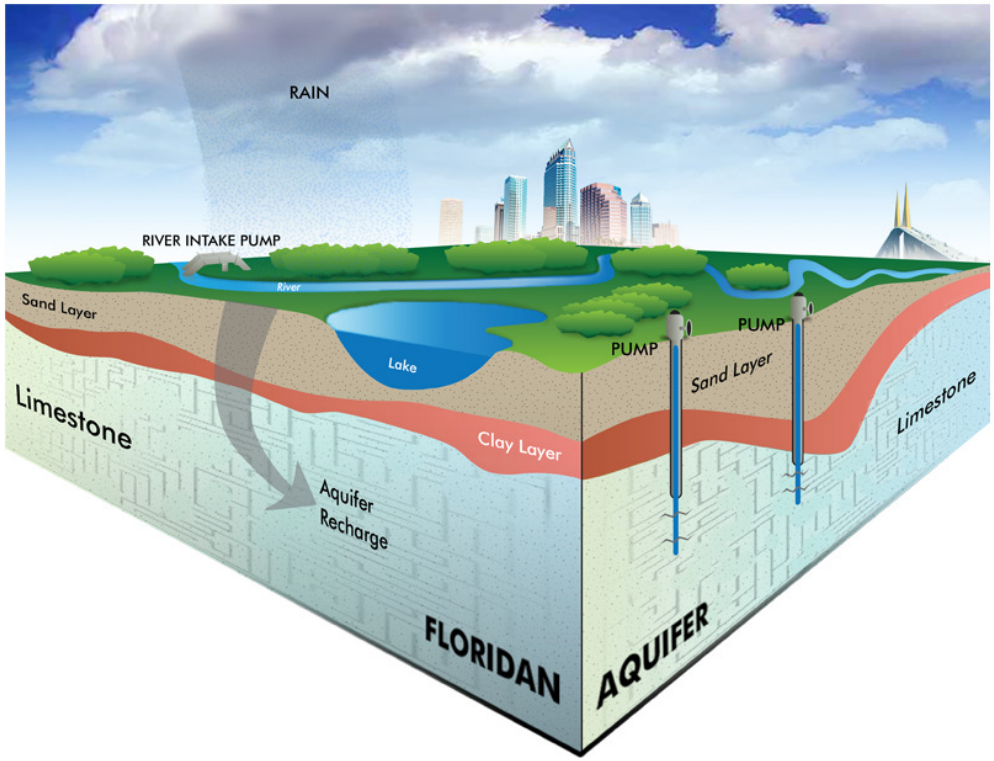
Tampa Bay Water is the largest wholesale drinking water supplier in Florida



Surface Water Sources

What lies beneath?

- Surficial Aquifer System
- Intermediate Aquifer/Confining Layer
  - Variable Confinement
- Upper Floridan Aquifer



https://www.tampabaywater.org/water-supply-source-groundwater

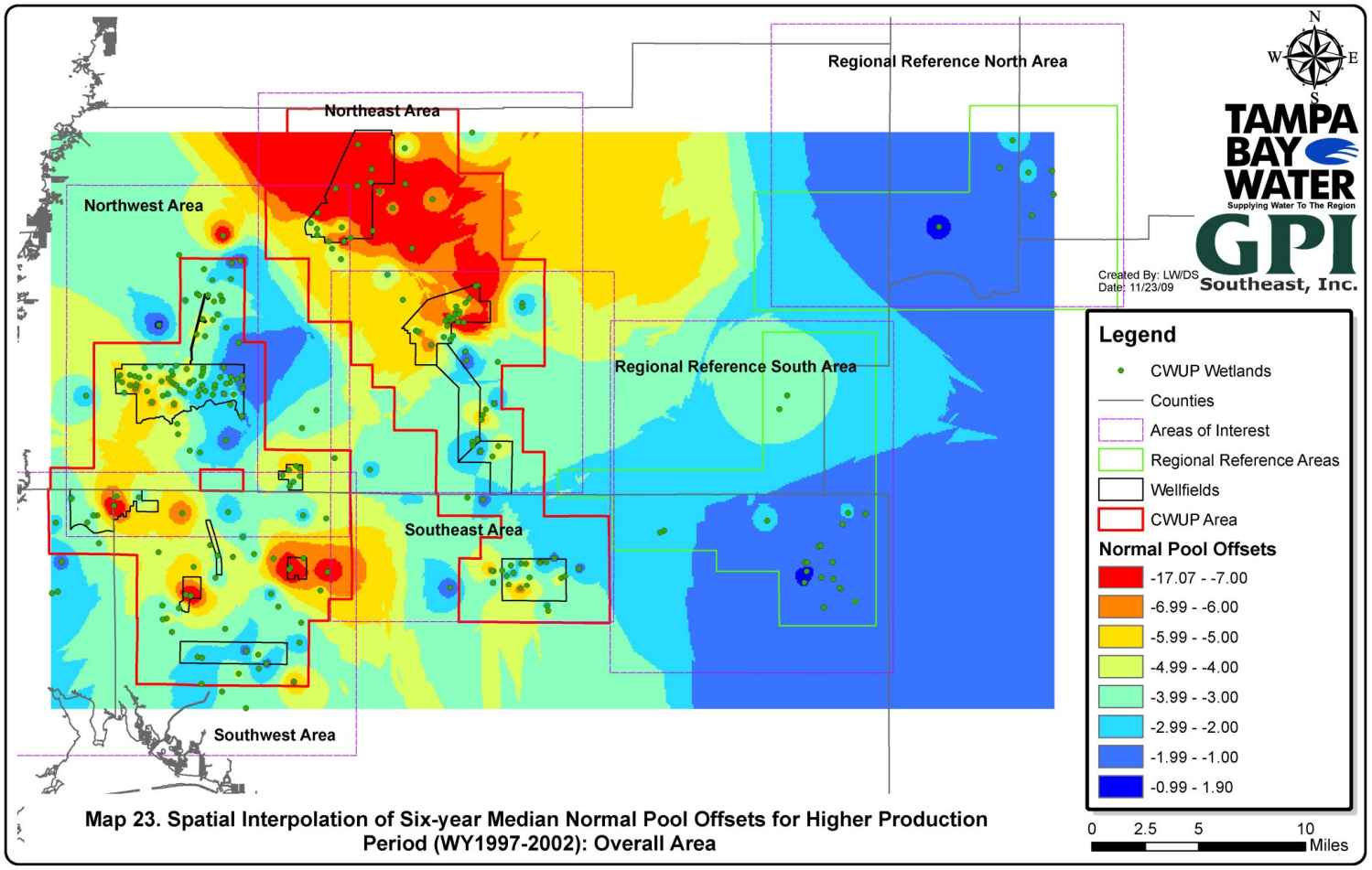
Historical Wetland/Lake Alterations

Cypress Wetland: No Change

Cypress Wetland: Severe Change

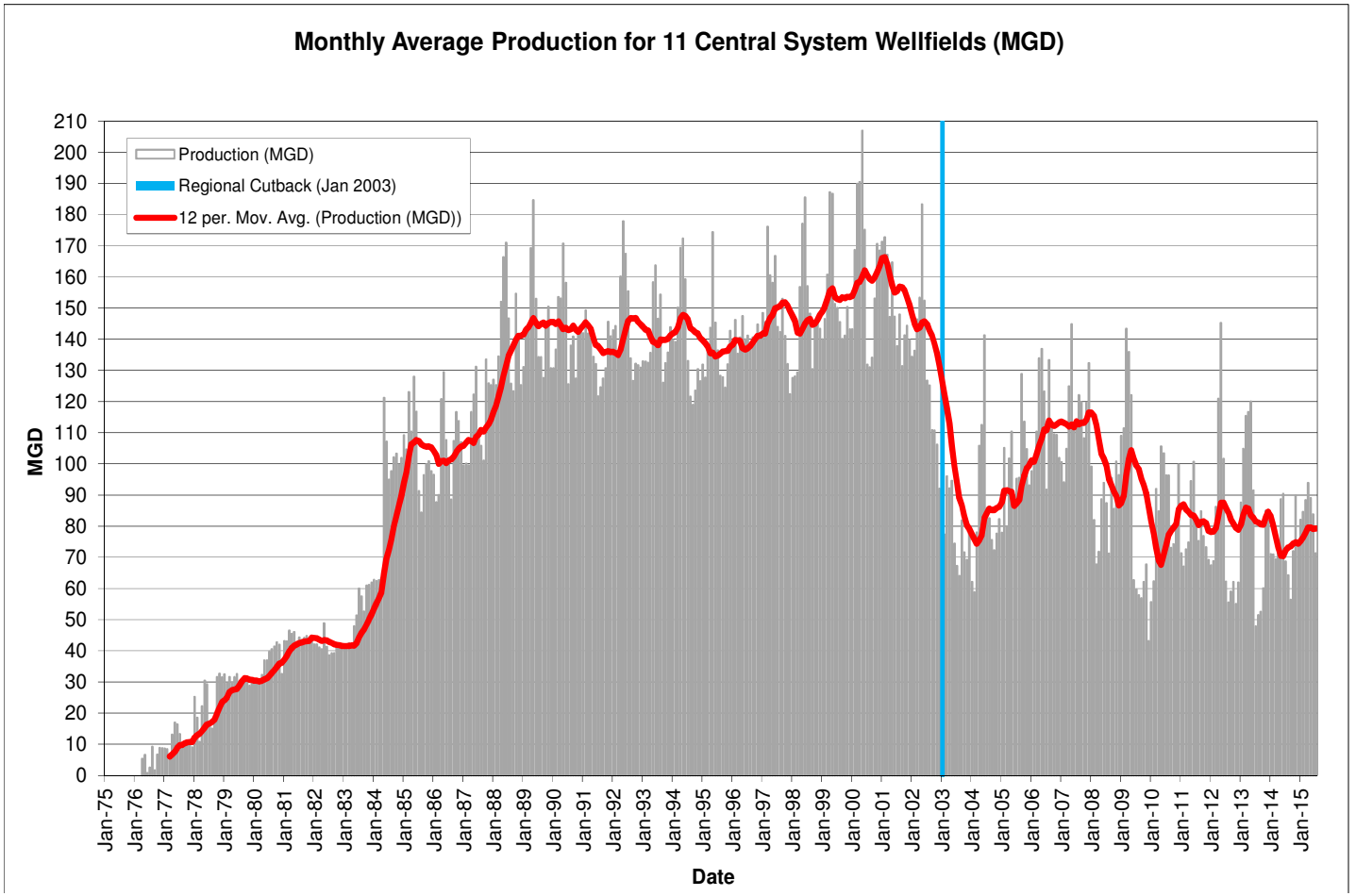
Historical Normal Pool Concept

Historical Wetland Water Level Drawdowns



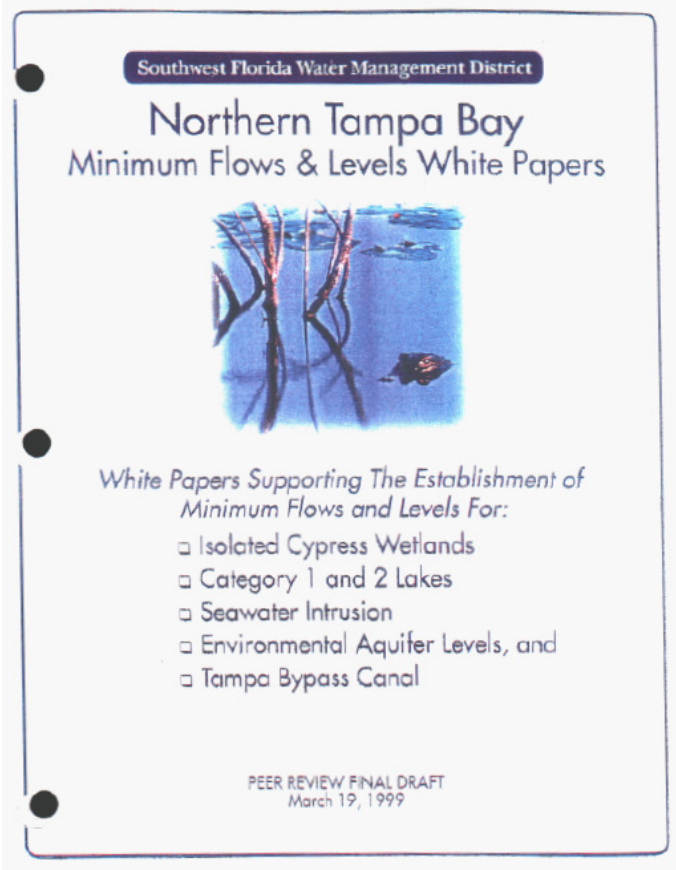
Map 23. Spatial Interpolation of Six-year Median Normal Pool Offsets for Higher Production Period (WY1997-2002): Overall Area

Regional Production Cutback (158 to below 120 MGD in Jan 2003)

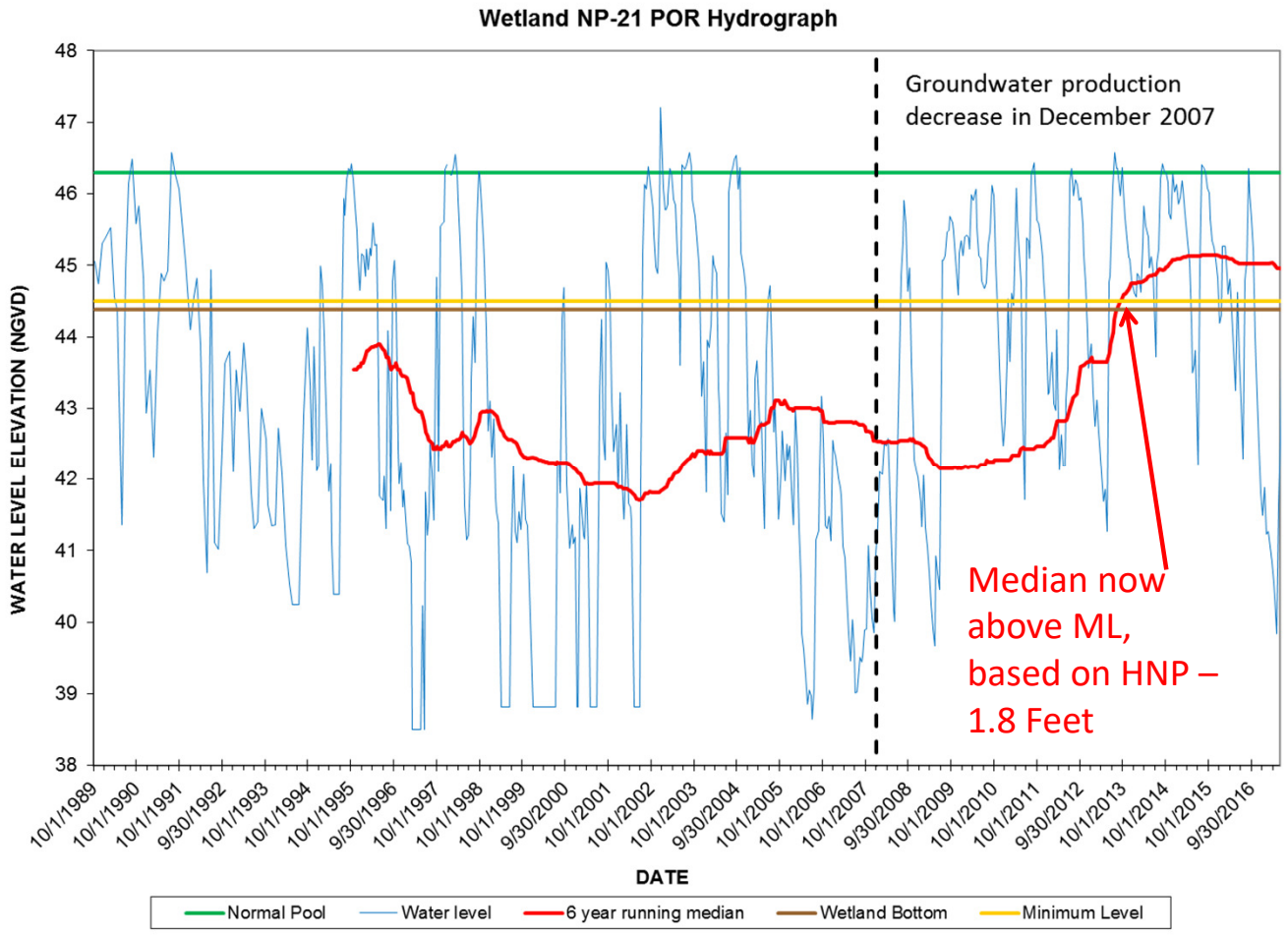


How do we know the affected lakes and wetlands have recovered?

- For isolated cypress wetlands and cypress-fringed lakes w/o structures, hydrologic recovery expected when long-term median water (P50) levels achieve surrogate Minimum Level of 1.8 feet below HNP (40D-8.623 FAC)
- Vegetative and hydric soils improvements expected to lag hydrologic recovery



Example of Hydrologic Recovery of Isolated Cypress (Mesic Soil Type)



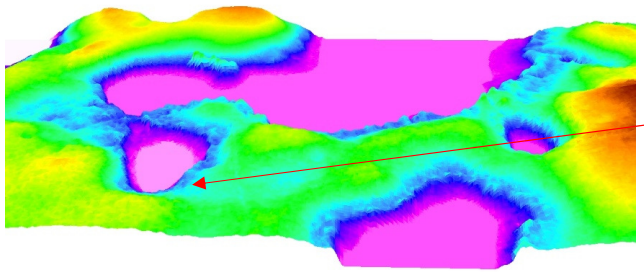
Tale of Two Wetland Types

Xeric-associated

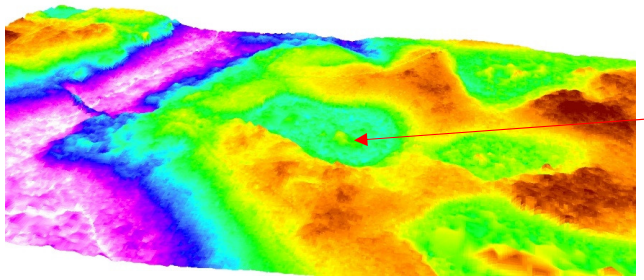
Mesic-associated

GPI

Xeric-Associated Wetlands and Landscape Position

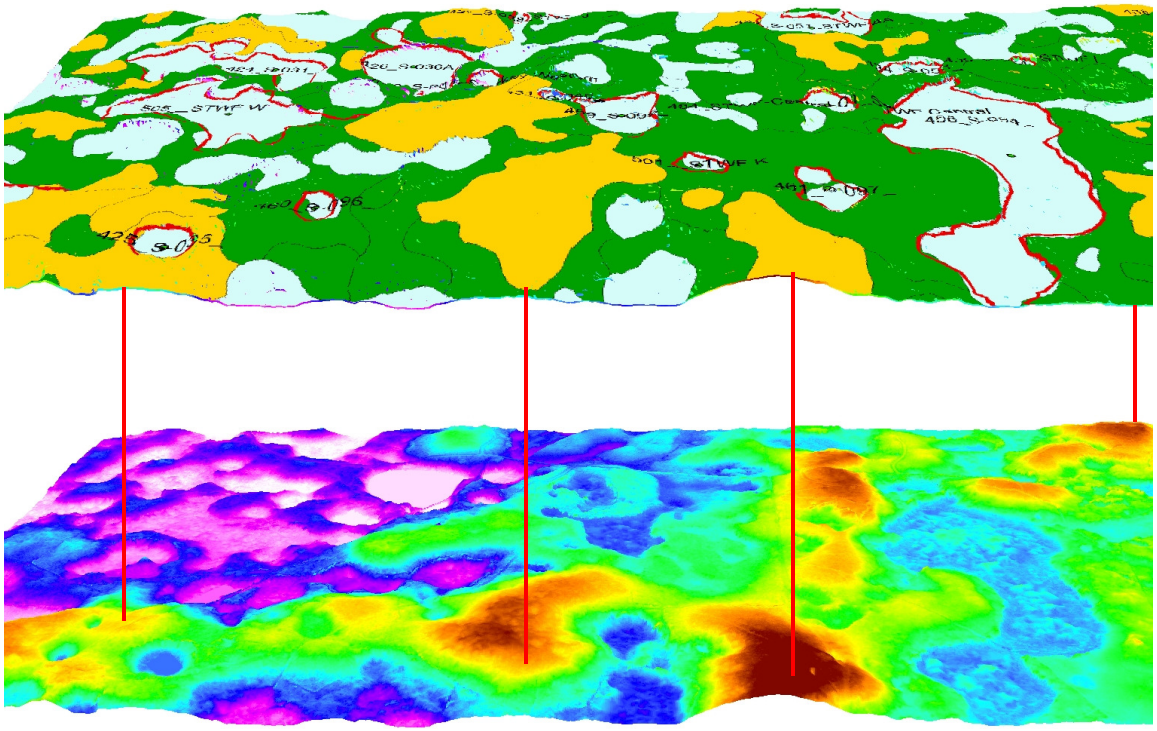


Xeric-associated wetlands (414_S-008 shown here) tend to be more isolated, occurring as deep, internally-drained sinkholes.

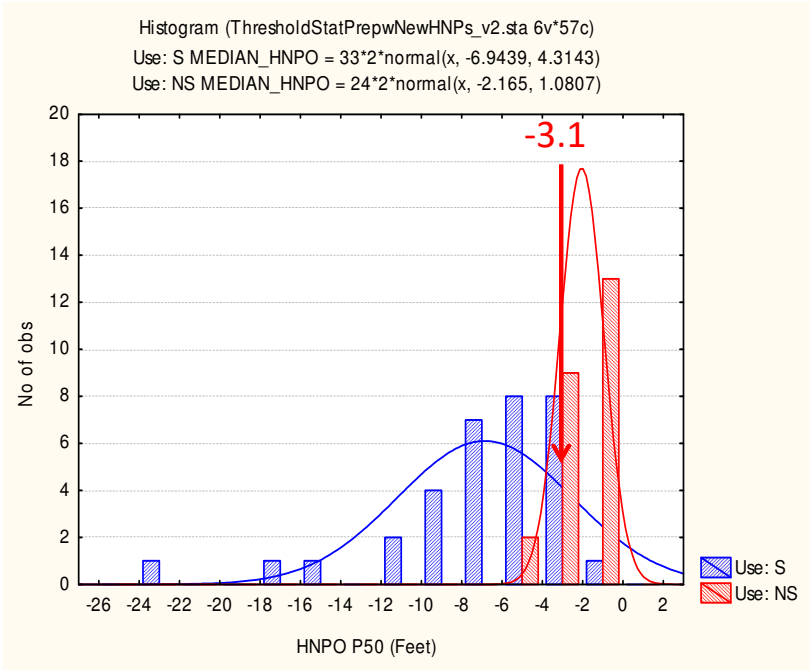


Mesic-associated wetlands (443_S-068 shown here) may tend to have surface water connections during wet seasons or wet years. Note low points between upstream and downstream wetlands.

Xeric Soils (Shown in Orange) Appear to Represent Local High Points (Ridges) in the Landscape; Mesic soils in Green

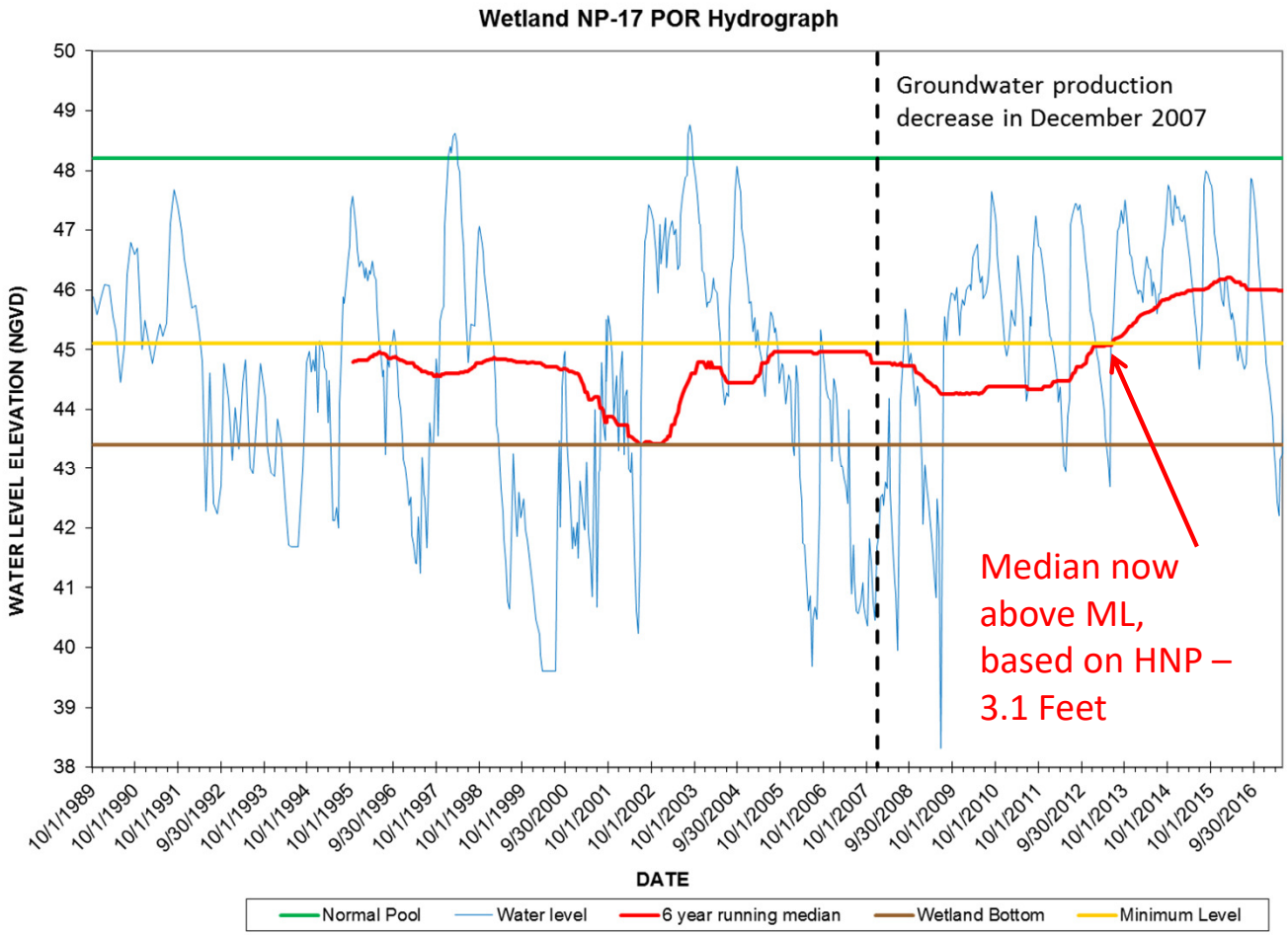


GPI Developed Xeric-associated Wetland/Lake ML as HNP-3.1 Feet

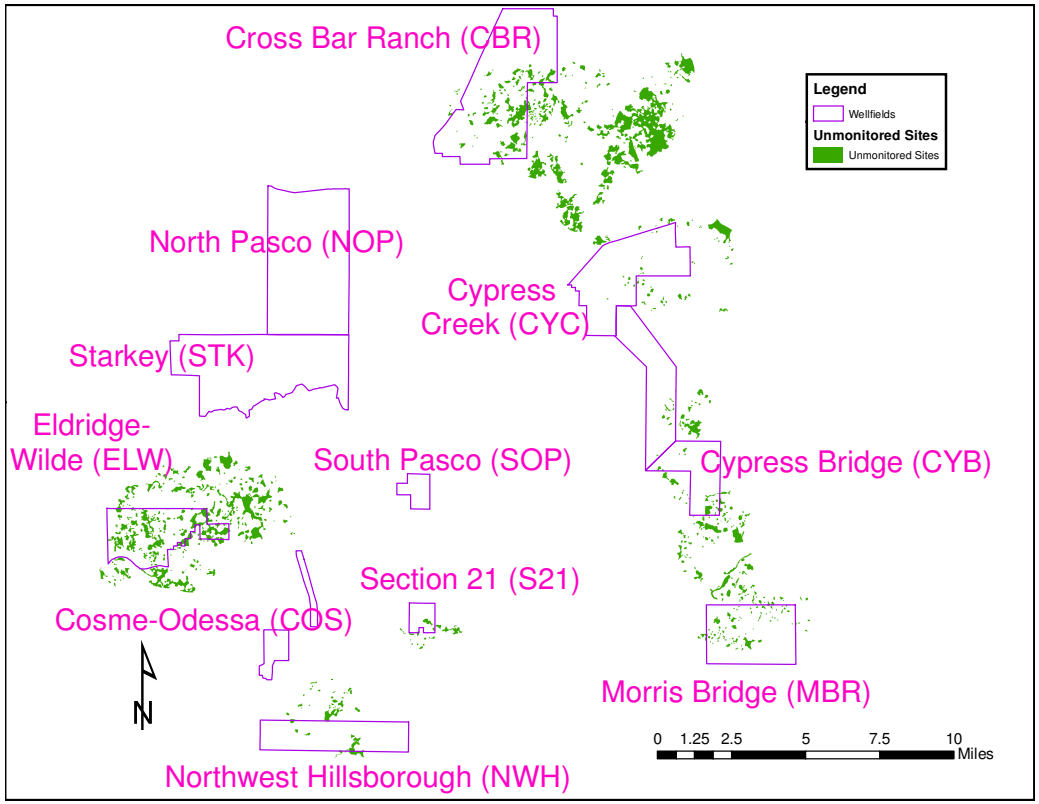


SWFWMD may perform a study to confirm these results and ultimately adopt a new Minimum Level for Xeric-associated sites

Example of Hydrologic Recovery of Isolated Cypress (Xeric Soil Type)



How do we assess the water level recovery of the Unmonitored Sites?



675 Wetlands and 9 Lakes in 2' SAS DDN with no water level monitoring

Approach

- Interpolation: using nearby sites to estimate conditions at unmonitored sites
  - Nearby in a statistical sense?
  - Nearby in a spatial sense?

GPI

Approach

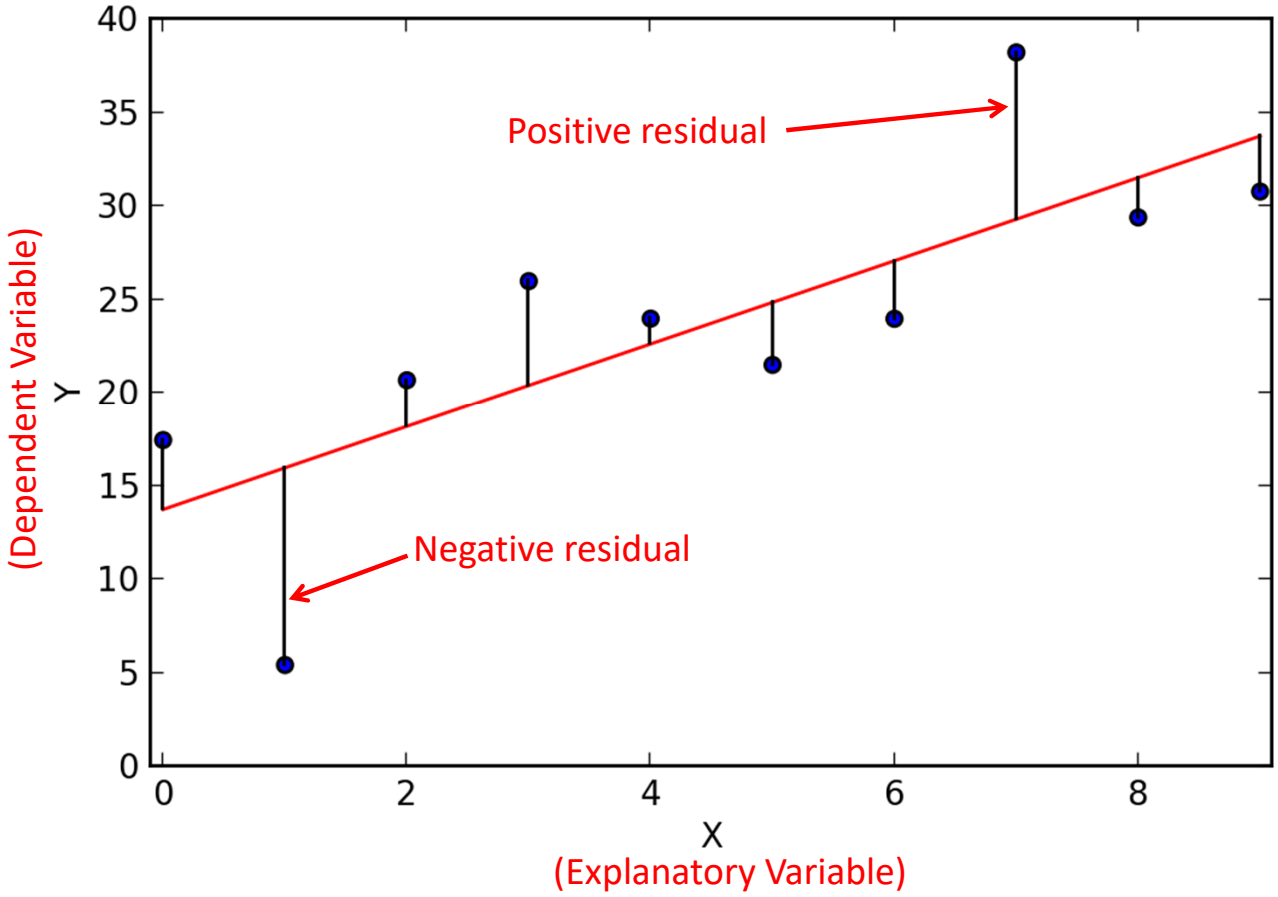
- Interpolation: using nearby sites to estimate conditions at unmonitored sites
  - Nearby in a statistical sense?
  - Nearby in a spatial sense?

Must I choose?

GPI

Review of Linear Regression

200-year old data mining technique!



GPI

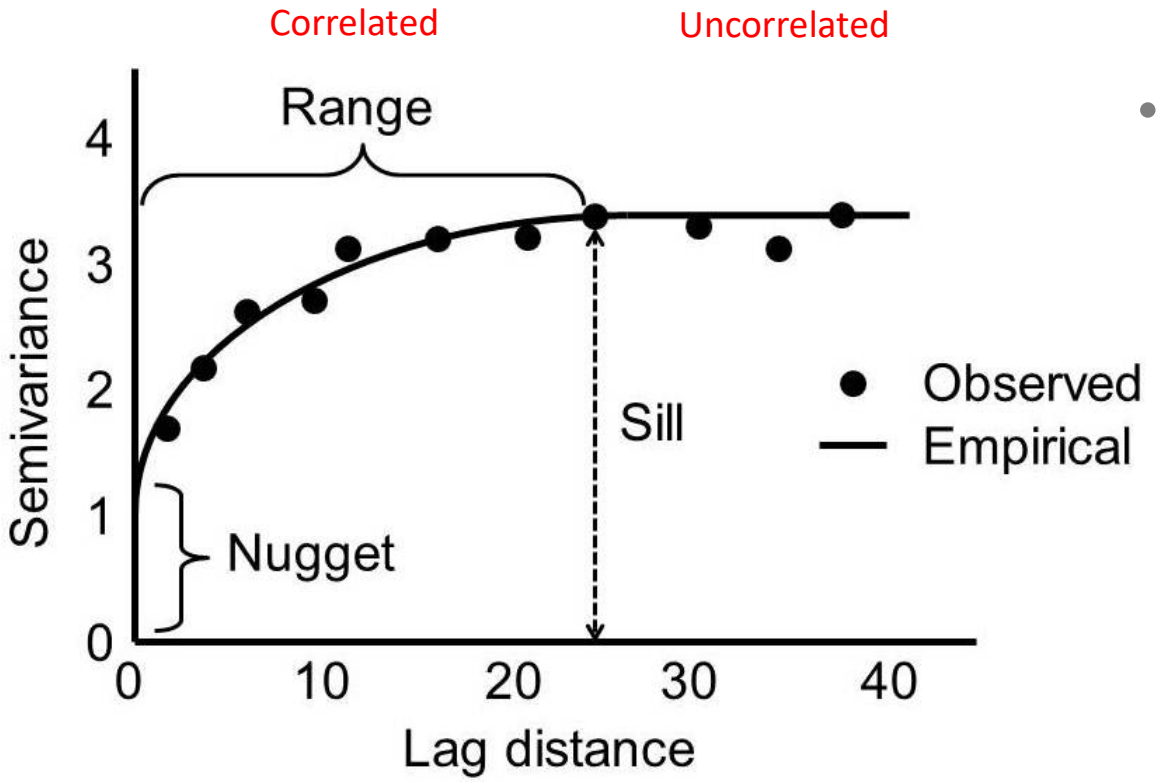
Assumptions of Linear Regression

- Linearity and additivity of relationship between x and y
- Statistical independence of errors
- Homoscedasticity (constant error variance)
- Normality of errors

GPI

Review of Kriging

70-year old data mining technique!



- Sophisticated spatial interpolation method that
  - accounts for spatial covariance in all data points, and
  - between all data points and the point to be estimated.

Asim Biswas and Bing Cheng Si (2013). Model Averaging for Semivariogram Model Parameters, Advances in Agrophysical Research, Prof. Stanisław Grundas (Ed.), InTech, DOI: 10.5772/52339. Available from: https://www.intechopen.com/books/advances-in-agrophysical-research/model-averaging-for-semivariogram-model-parameters

GPI

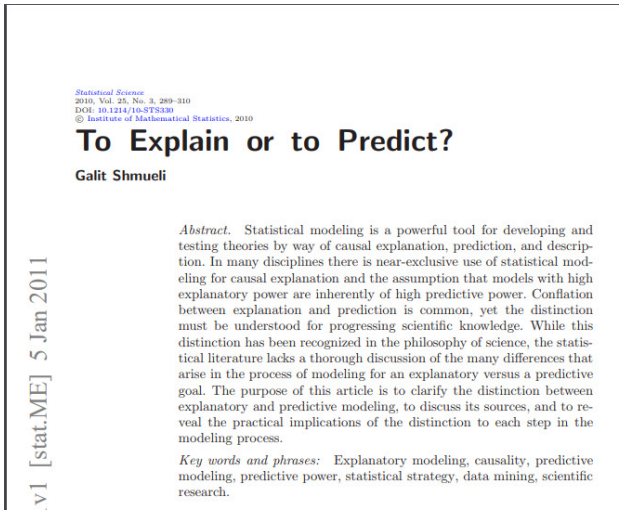
Assumptions of Ordinary Kriging

- Stationarity (constant mean and variance over study area)
- Isotropy (variogram structure same in all directions)

GPI

To explain or predict? (Shmueli 2010 worth reading)

- Statistical model-building can be useful for identifying important factors worthy of further experimental manipulation or further studies. However, there is no way to completely get around multicollinearity without a good conceptual model of system
- Prediction on the other hand requires no concern about multicollinearity, merely how well does the algorithm perform on out-of-sample (i.e., test data)



Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Abstract. Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description. In many disciplines there is near-exclusive use of statistical modeling for causal explanation and the assumption that models with high explanatory power are inherently of high predictive power. Conflation between explanation and prediction is common, yet the distinction must be understood for progressing scientific knowledge. While this distinction has been recognized in the philosophy of science, the statistical literature lacks a thorough discussion of the many differences that arise in the process of modeling for an explanatory versus a predictive goal. The purpose of this article is to clarify the distinction between explanatory and predictive modeling, to discuss its sources, and to reveal the practical implications of the distinction to each step in the modeling process.

Key words and phrases: Explanatory modeling, causality, predictive modeling, predictive power, statistical strategy, data mining, scientific research.

Again, must I choose?

Conceptual Diagram for Regression kriging Approach

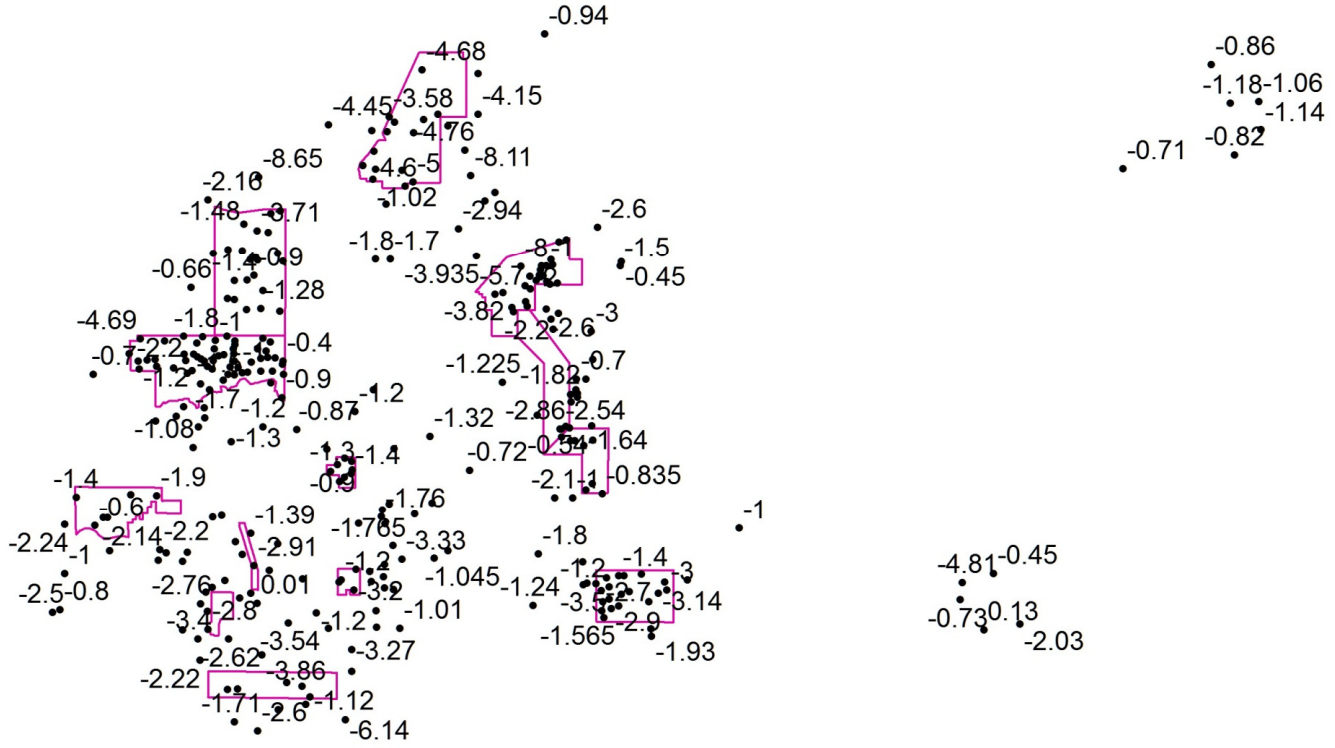
Identify most-likely multiple linear regression model, while avoiding overfitting

Perform kriging on residuals from multivariate model

Regression kriging model incorporates both aspatial and spatial trends for improved prediction

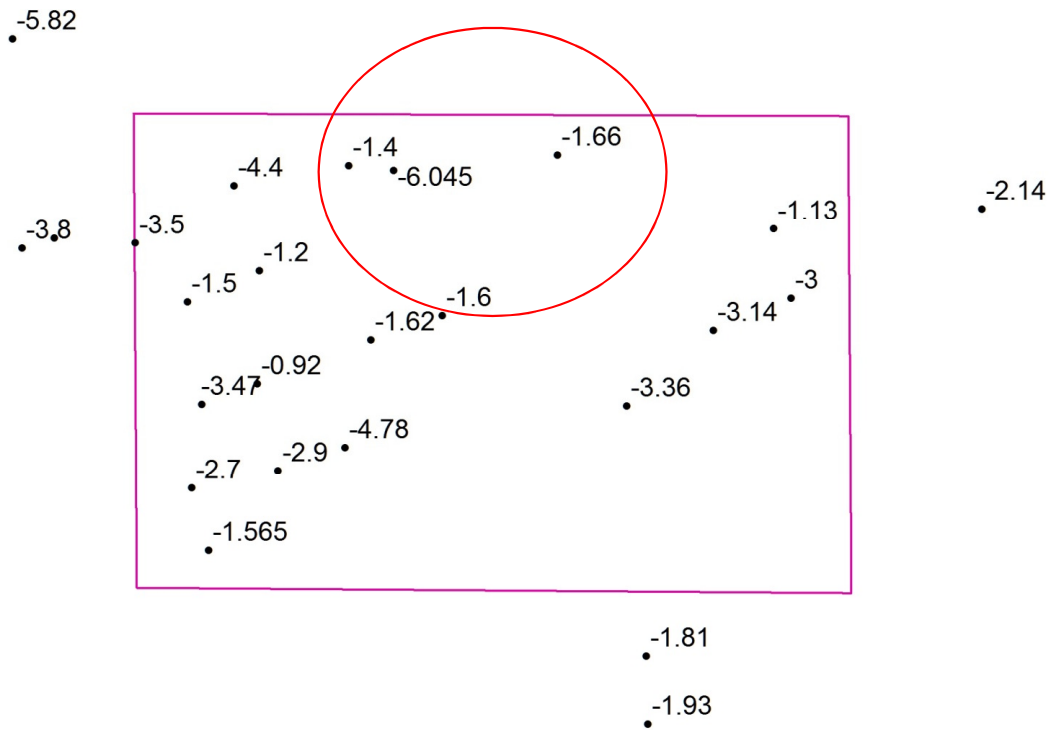
GPI

Sites with HNPs and water level data for 2008-2014 (N=309)

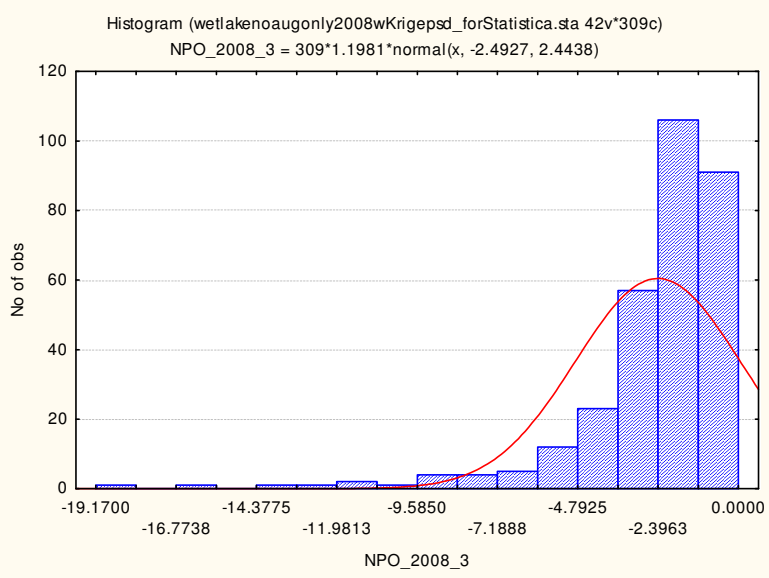


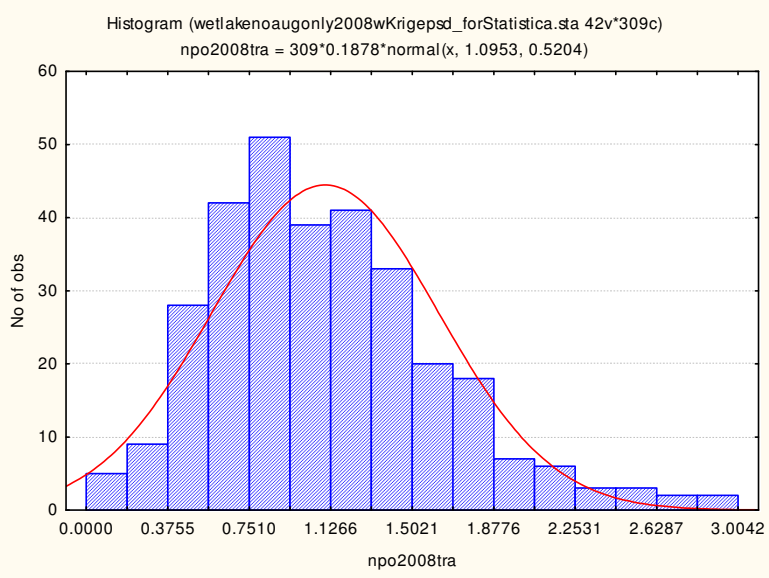
Large, Abrupt Changes in HNP Offsets: Example 1

Large, Abrupt Changes in HNPOs: Example 2



Transformed Dependent Variable (HPNO) to achieve normality

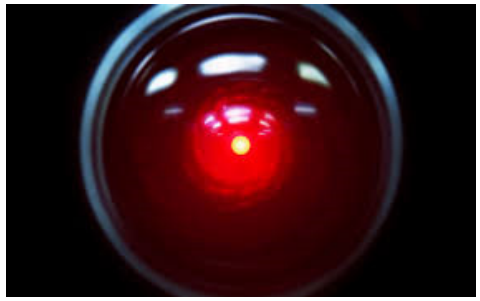




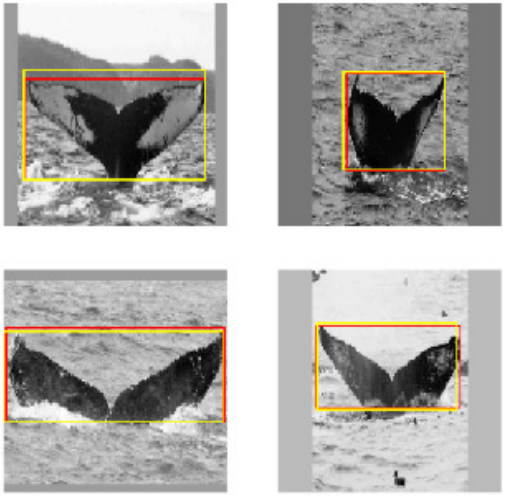
Transformation: LN((value*-1)+1)

Prior to transformation, adjusted 3 sites with very small positive HNPOs to 0:
490 (T-10), 91 (CNR-S9), and 703 (Pretty Lake)

What is machine learning?



- A field of computer science that develops algorithms that can learn from and make predictions on data
- Focus tends to be more on prediction than explanation
- Applied widely in business and scientific research
  - Product recommendations (e.g., Netflix, Amazon)
  - Price estimation (Zillow)
  - Google image segmentation and identification
  - Consumer credit scoring
  - Handwriting and object recognition
  - Drug development
  - Bioinformatics



GPI

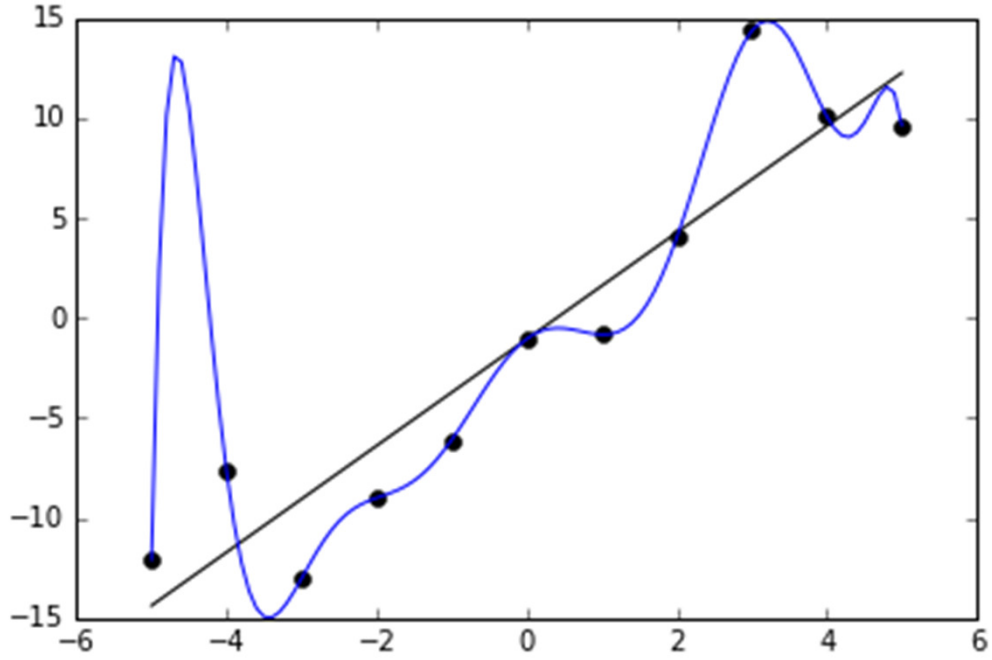
Some choices of algorithms

- Regression
- Instance-based (e.g., k-nearest neighbor)
- Regularization (e.g., ridge regression and glmnet)
- Decision Tree (e.g., classification and regression trees)
- Bayesian (e.g., naïve Bayes)
- Clustering (e.g., k-means)
- Association rules
- Artificial Neural Networks
- Deep Learning (e.g., convolutional neural networks)
- Dimensionality Reduction (e.g., multidimensional scaling)
- Ensemble (e.g., Random Forest and gradient boosting machines)

https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

GPI

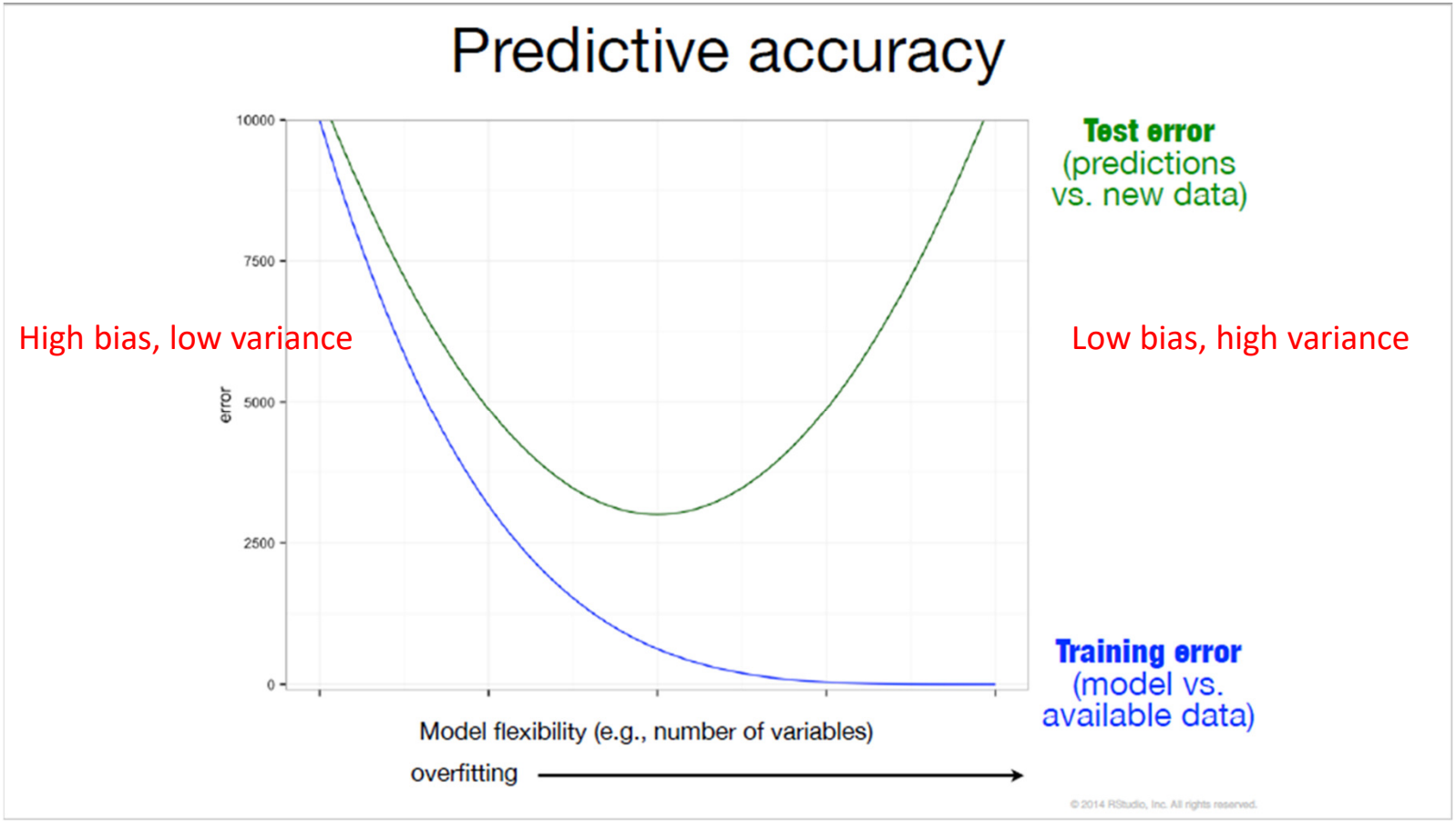
The most important concept in modeling and machine learning



Ghiles (2016): https://commons.wikimedia.org/wiki/File:Overfitted_Data.png

GPI

The most important concept in modeling and machine learning



A machine learning paradigm and approach

- Fill missing variable data or fit model with fewer variables
- Randomly divide data into train (e.g. 80%) and test (e.g., 20%) groups
- “Train” model (supervised learning) using all training data, avoiding overfitting through algorithm choice or cross validation
- “Tune” model if model has a hyperparameter*
- Evaluate final tuned model performance on test dataset
- Refit model for predictions using same settings for entire dataset
- Examine residuals and if spatial autocorrelation is present then krige residuals for regression kriging model

*higher level parameter that cannot be directly estimated from the data but is set by the practitioner using heuristics and then possibly tuned using cross validation to select an optimal setting.

GPI

Bayesian Information Criterion (BIC) Search for Best Aspatial Model (out of 2,048) using R package glmulti

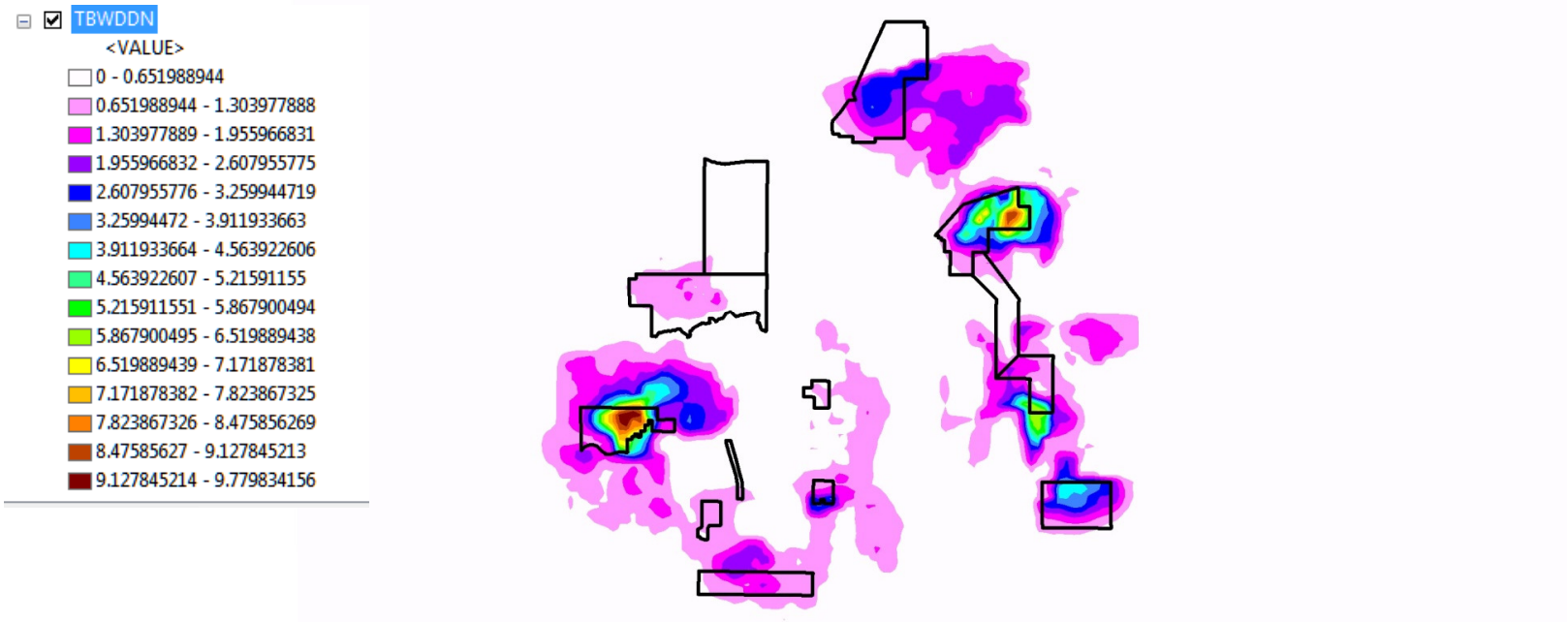
- Aquifer-related
  - SAS Drawdown
  - IA Thickness
  - Head Difference
- Soils-related
  - Xeric Ratio
  - Xeric Y/N
  - Soil Permeability
- Well-related
  - Distance to Near Well
  - Kernel Density of Wells
- Miscellaneous
  - Area Perimeter Ratio
  - Acres
  - Rainfall

lowest Bayesian Information Criterion (BIC) = most probable model given the dataset, avoids overfitting

```
glm1<-
glmulti(npo2008trans~TBWDDN+RAXericRat+RAXericYN+ACRES_RA+AREAPERATI+Distnearwe+Headdiffer+Soilperm+IAthicknes+Rain10NN+Kerneld,data=train,crit="bic",level=1,method='d')
```

GPI

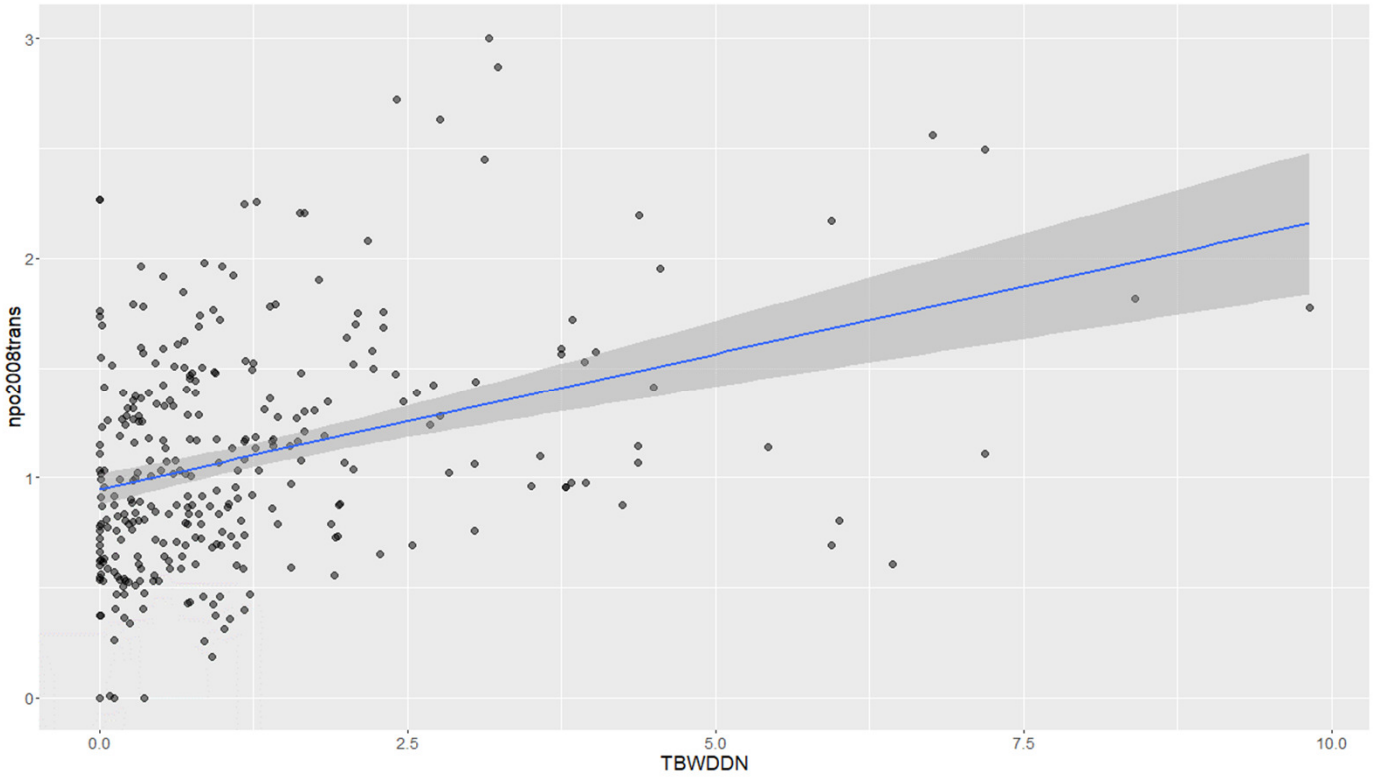
SASDDN (Drawdown)



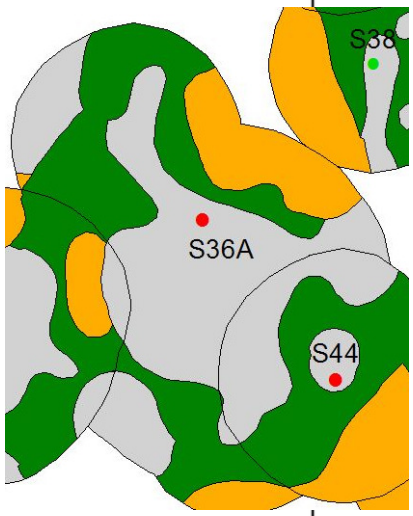
12 nearest neighbor Inverse Distance Squared Weighted Interpolation based on point file provided by Tampa Bay Water; represents the maximum of the "Historical Production and Scaled Pumpage" scenarios described in Tampa Bay Water (2013).

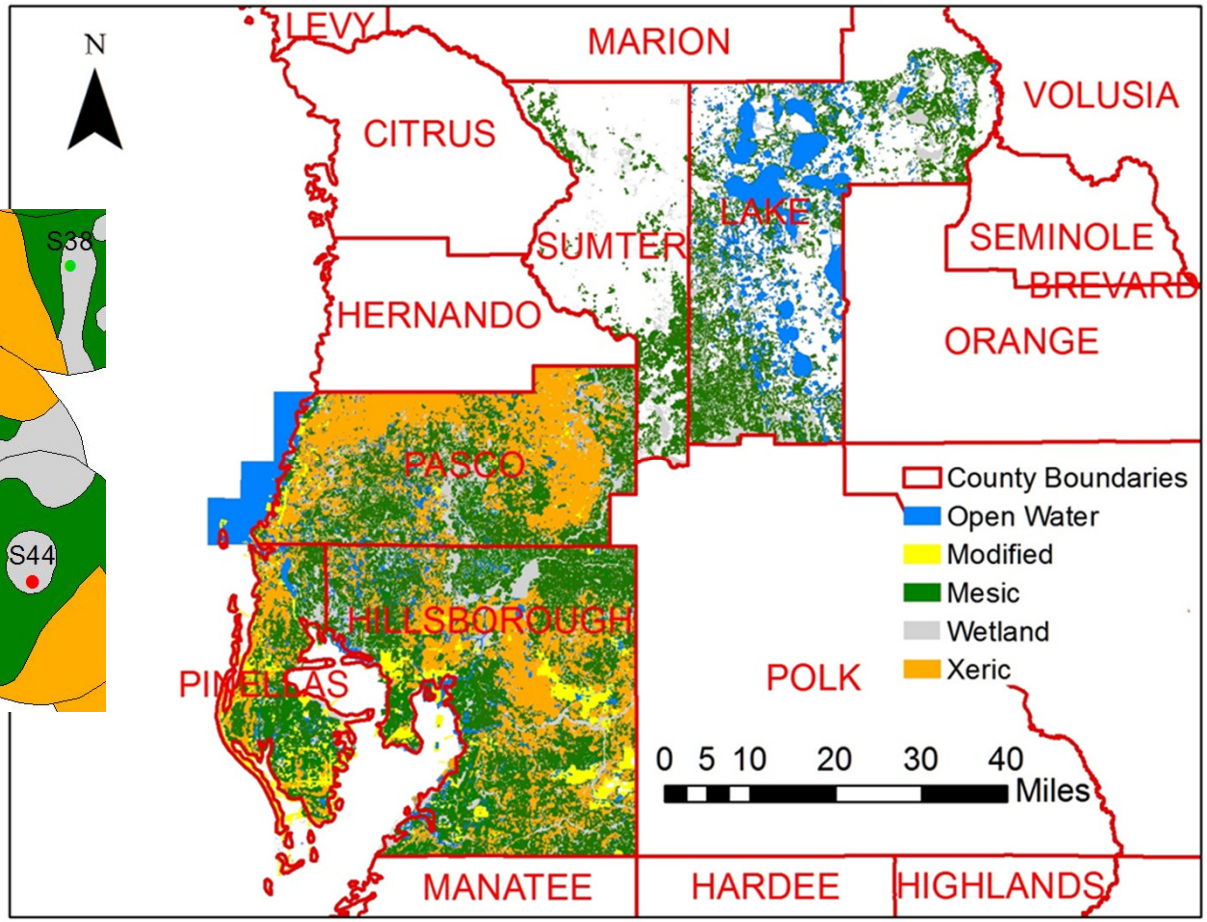
Tampa Bay Water. 2013. Defining Areas of Investigation for Recovery Analysis. Memorandum to Southwest Florida Water Management District. September, 2013.

NPOs vs. SASDDN

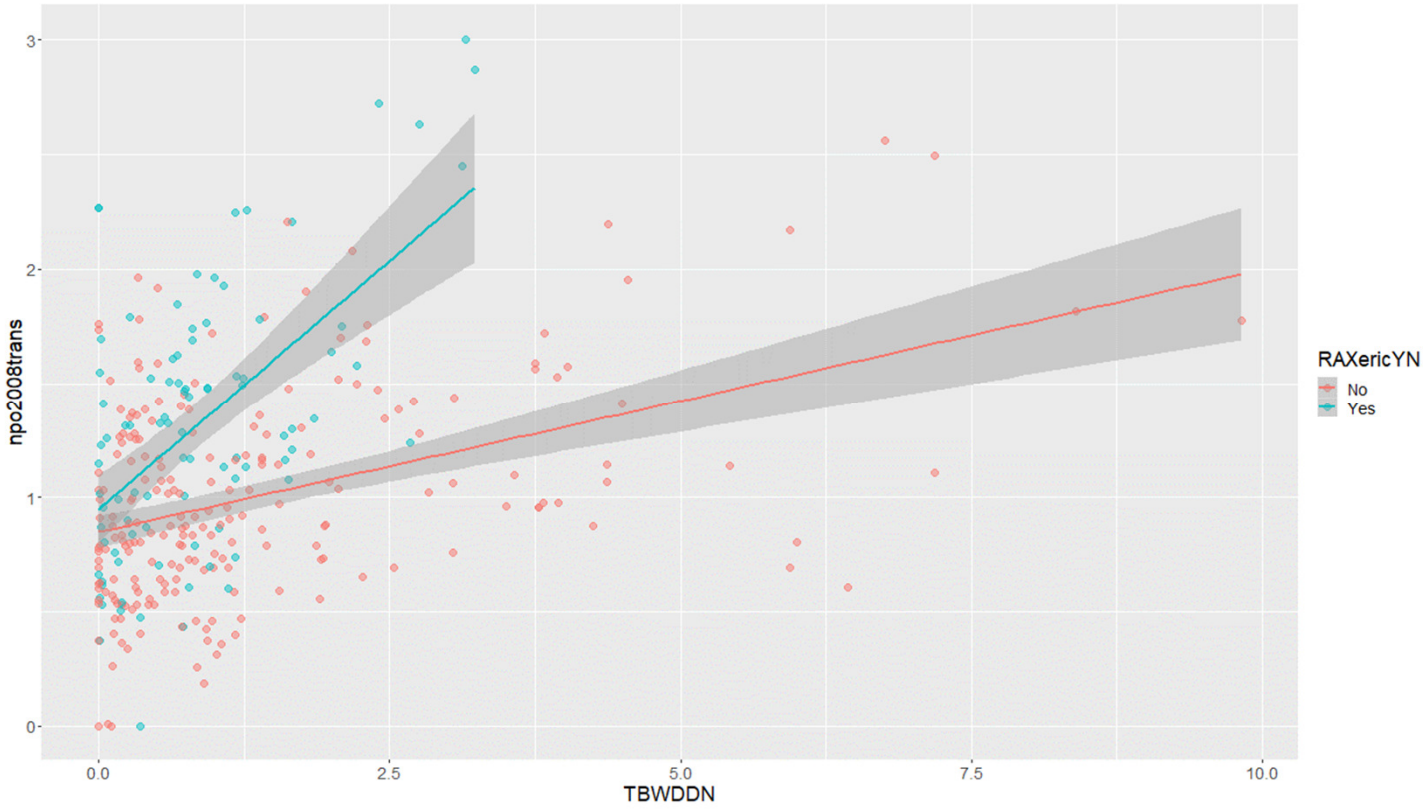


Classified Soils Layer from Xeric Recovery Analysis Project used to calculate Xeric Ratio (500-feet buffer)

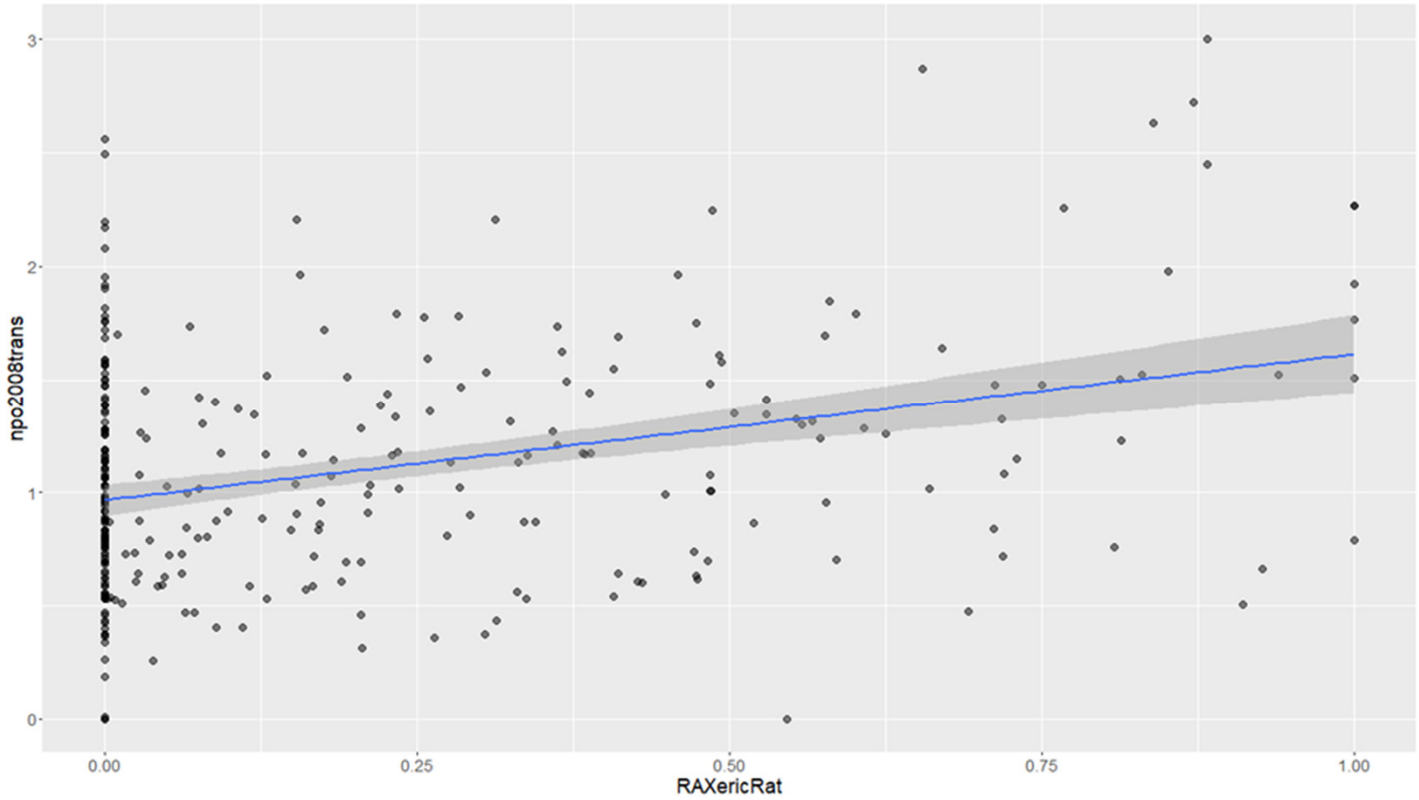




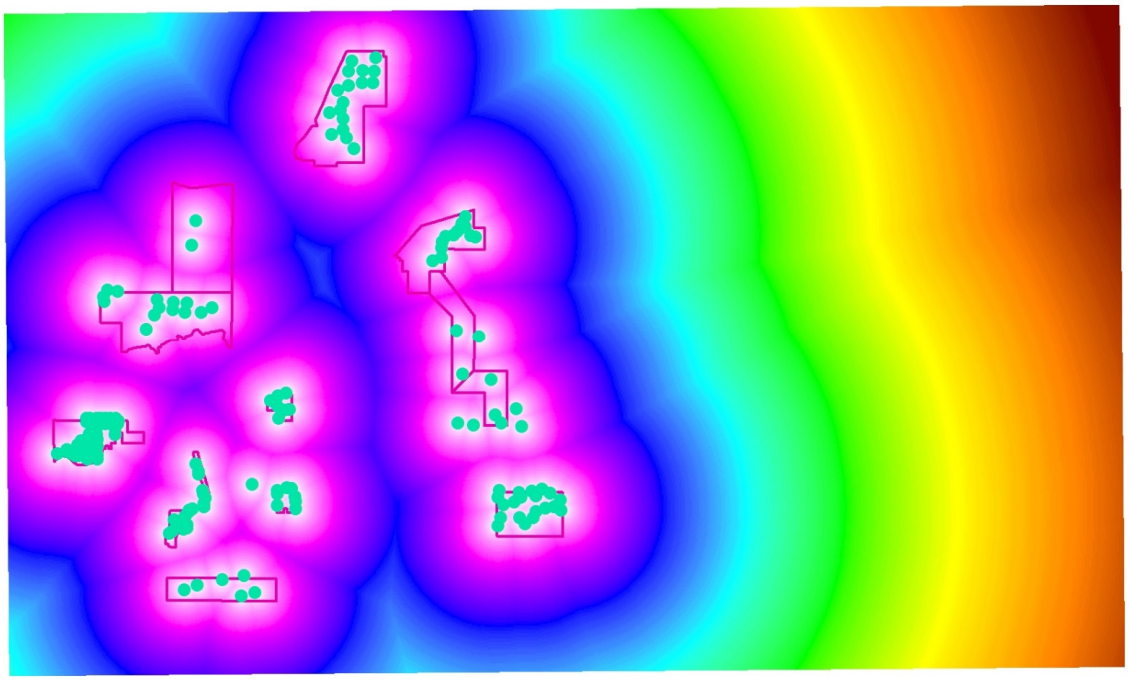
NPOs vs. SASDDN by Soil Type



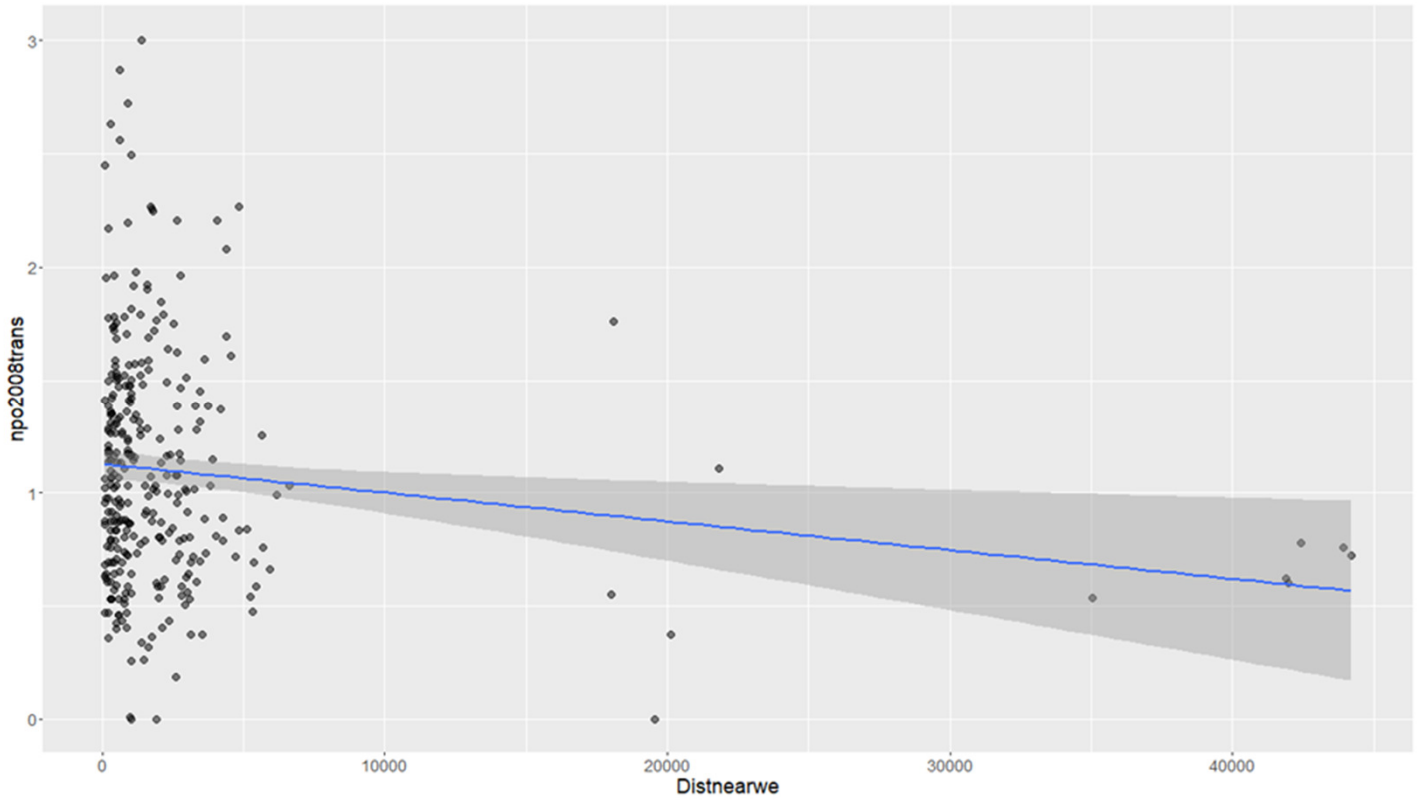
NPOs vs. RAXericRat



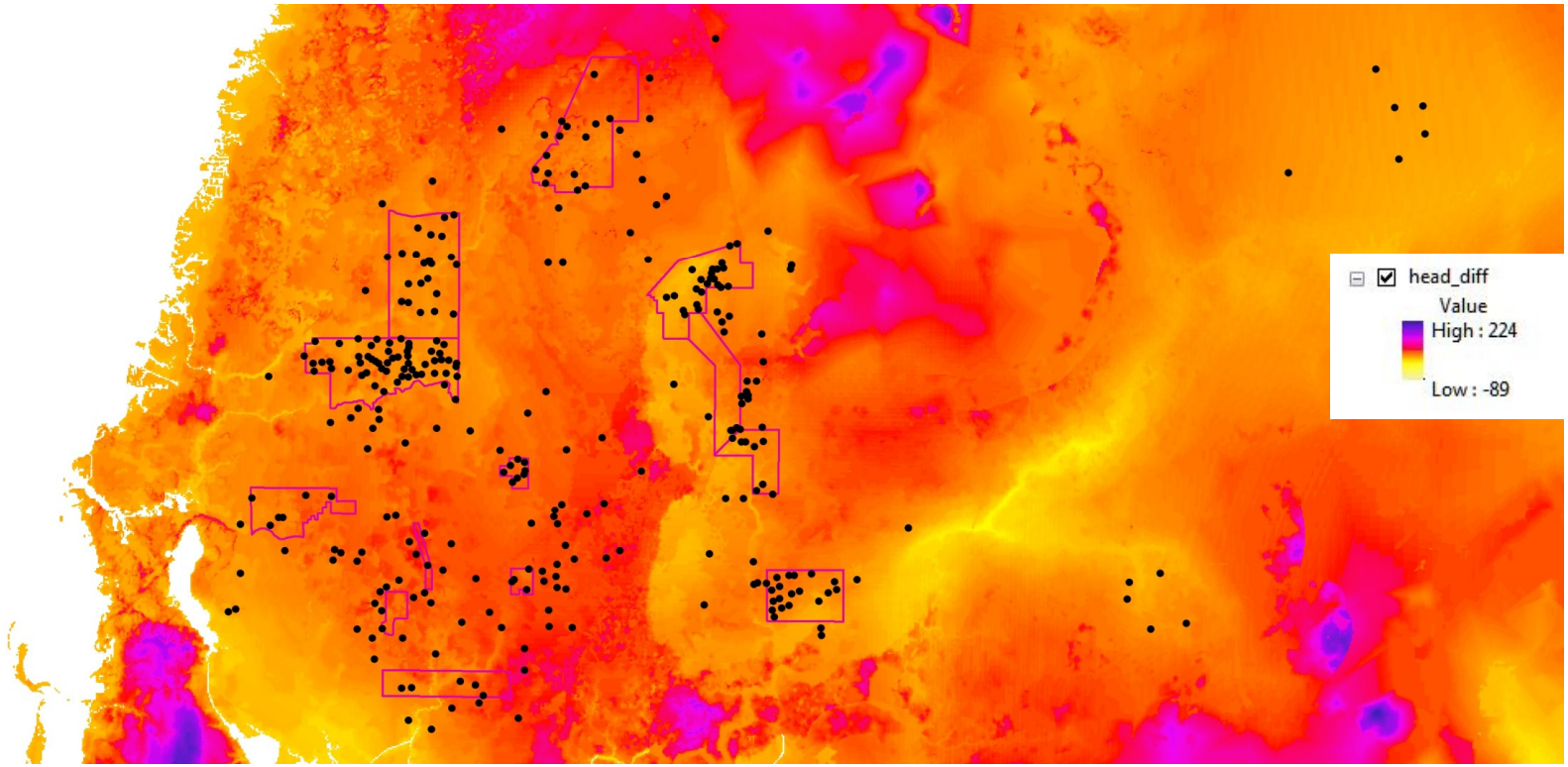
Distance to Nearest Well



NPOs vs. Distance to Nearest Well



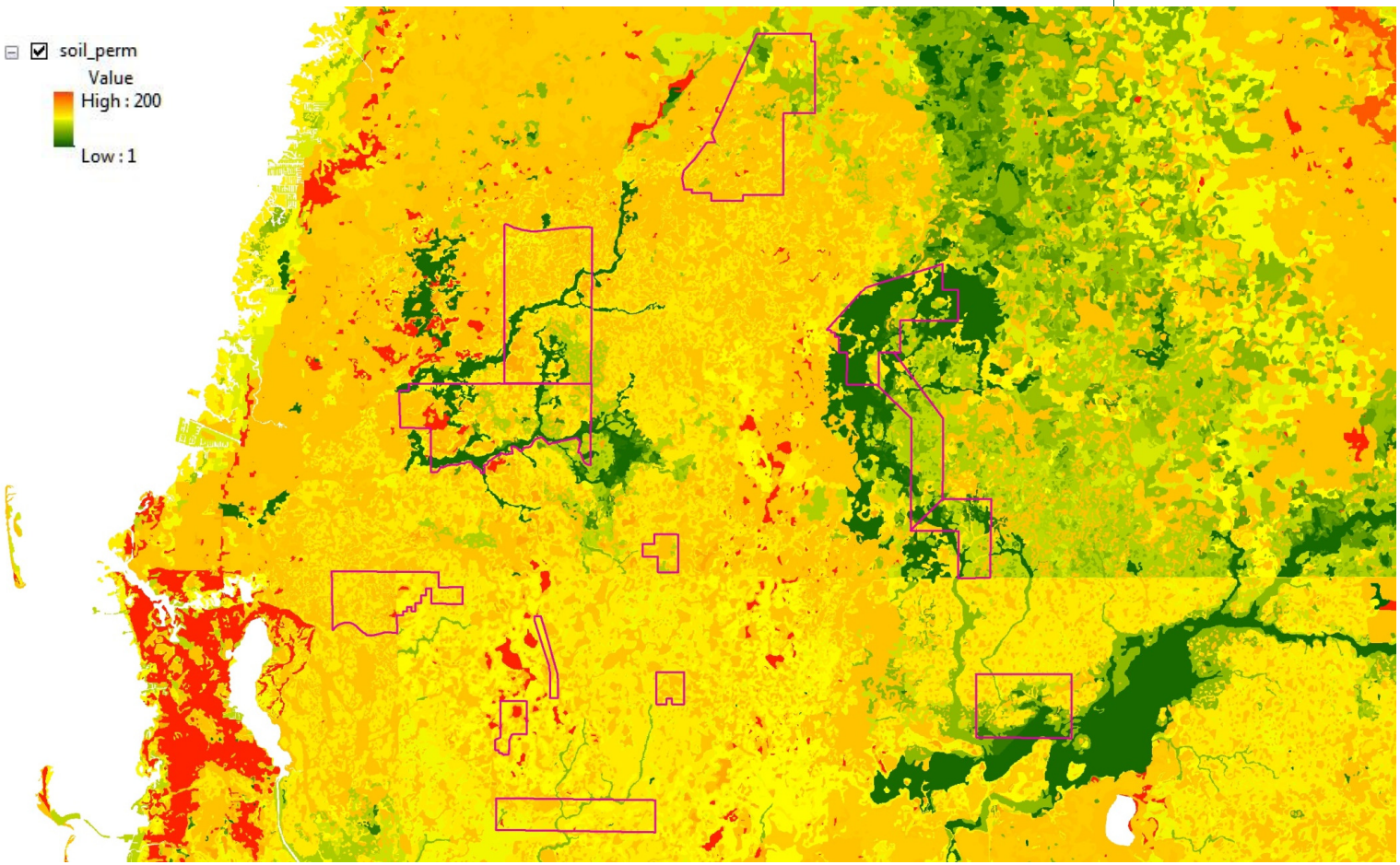
Head Difference



Florida Aquifer Vulnerability Assessment by FDEP

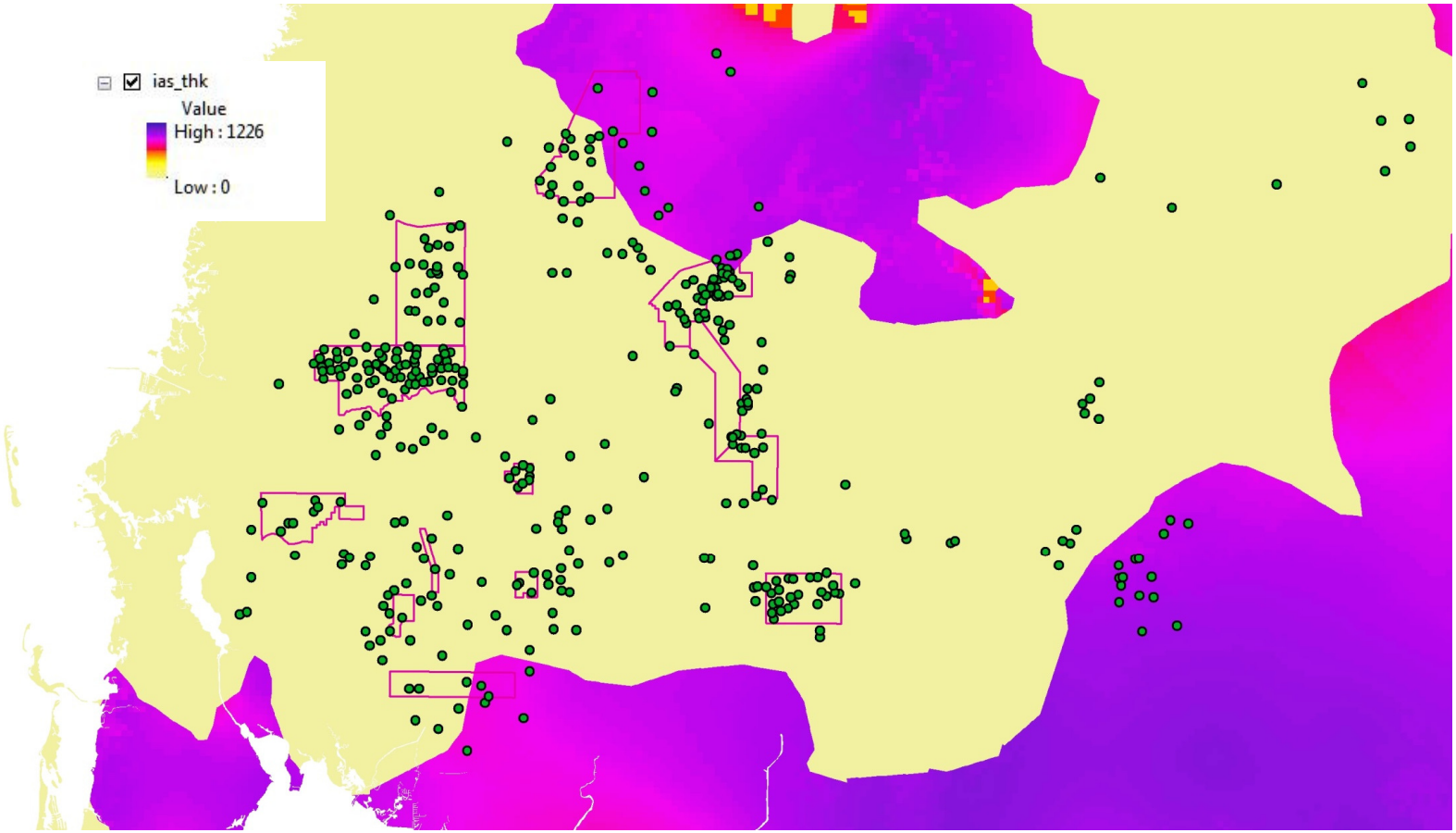
GPI

Soil Permeability (from FAVA)

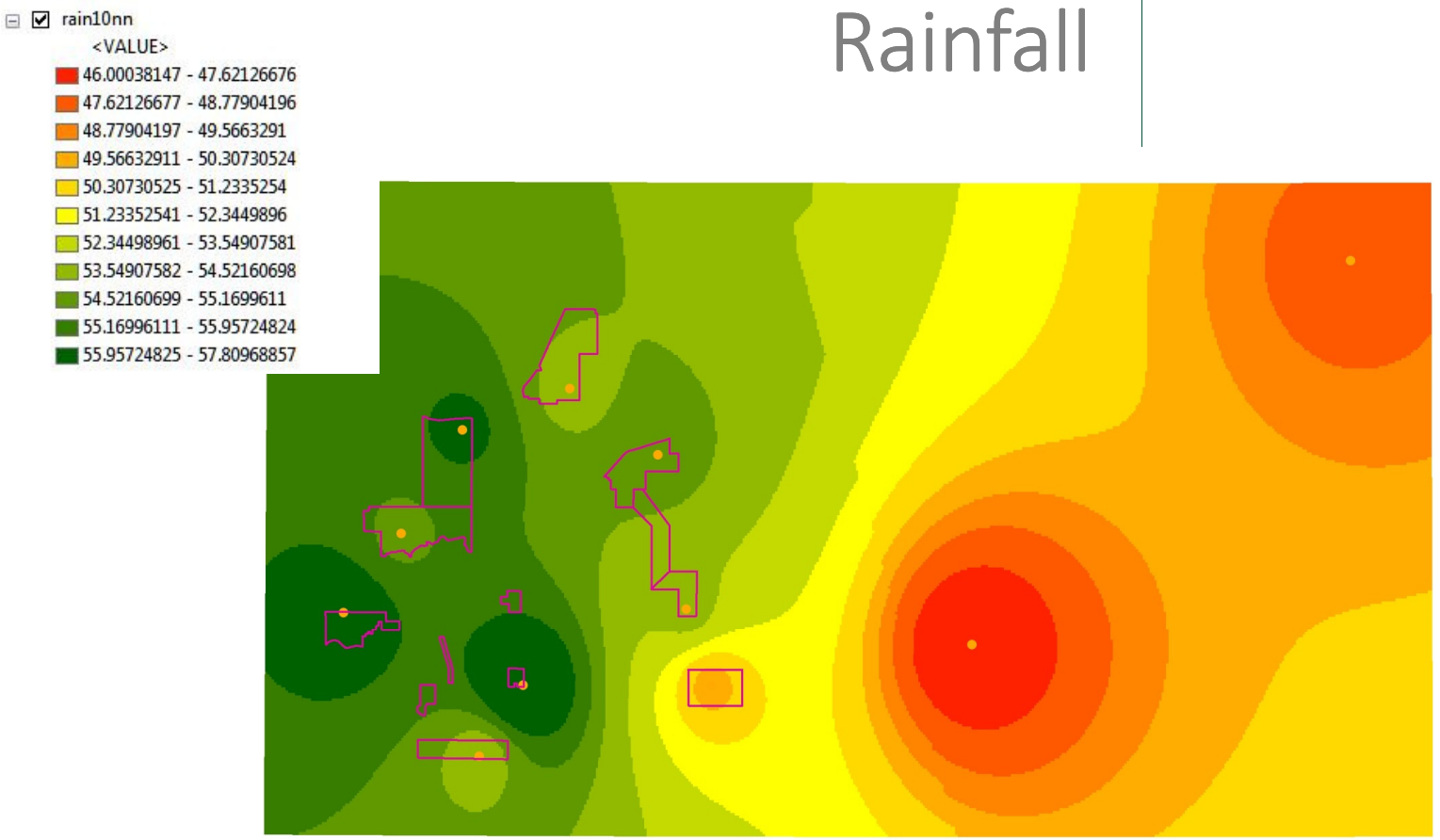


GPI

Intermediate Aquifer (IA) Thickness from FAVA

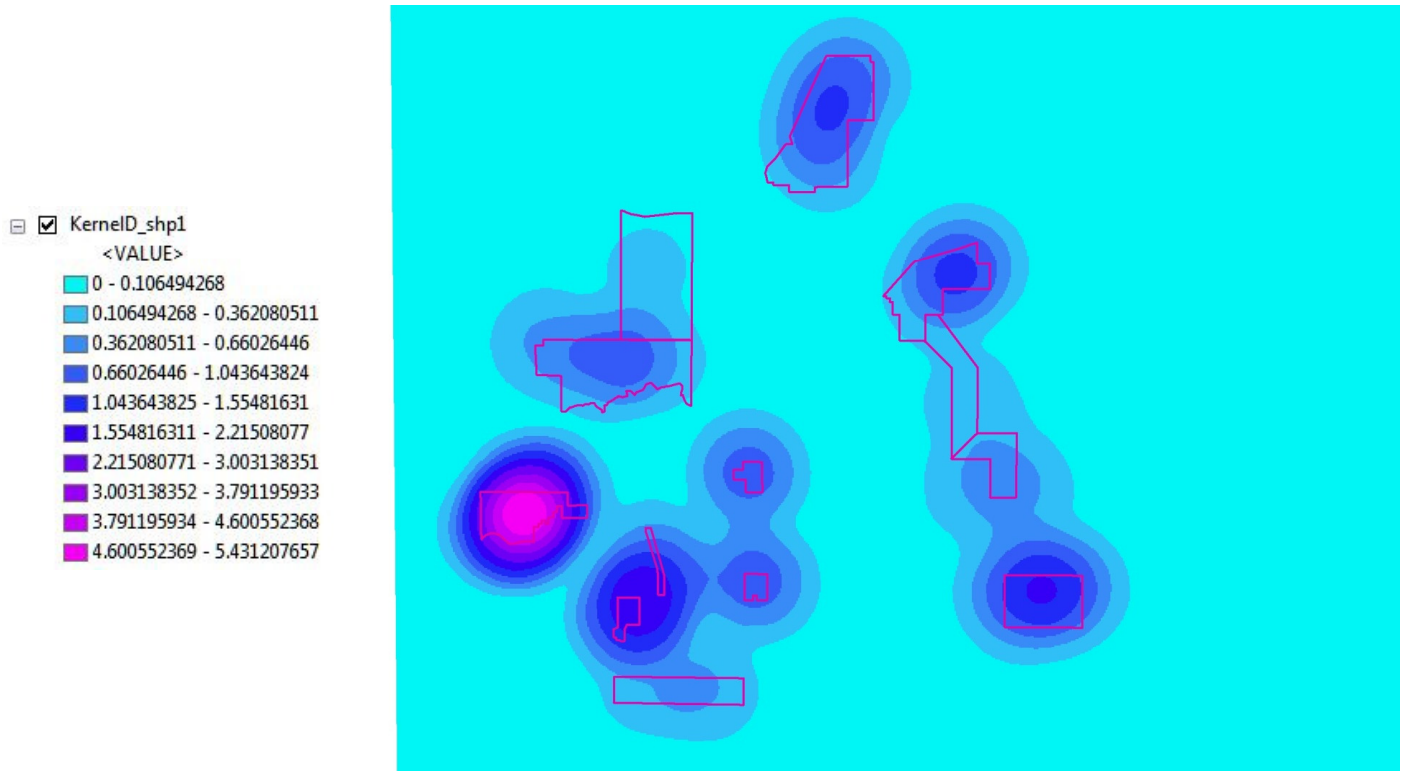


Rainfall



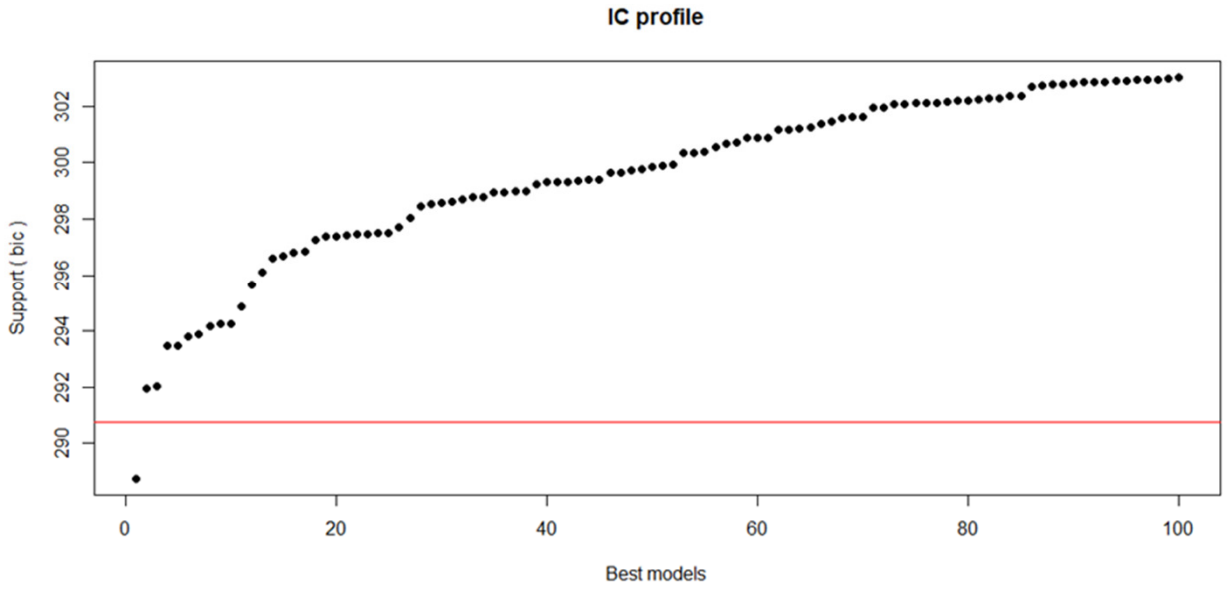
Mean rainfall for 2008-2014 from 11 gap-filled stations using 10 nn inv sq dw interpolation

Kernel Density of Wells



(BIC) Search Results: Highest Probability Model Selected

- SAS DDN
- Xeric Ratio



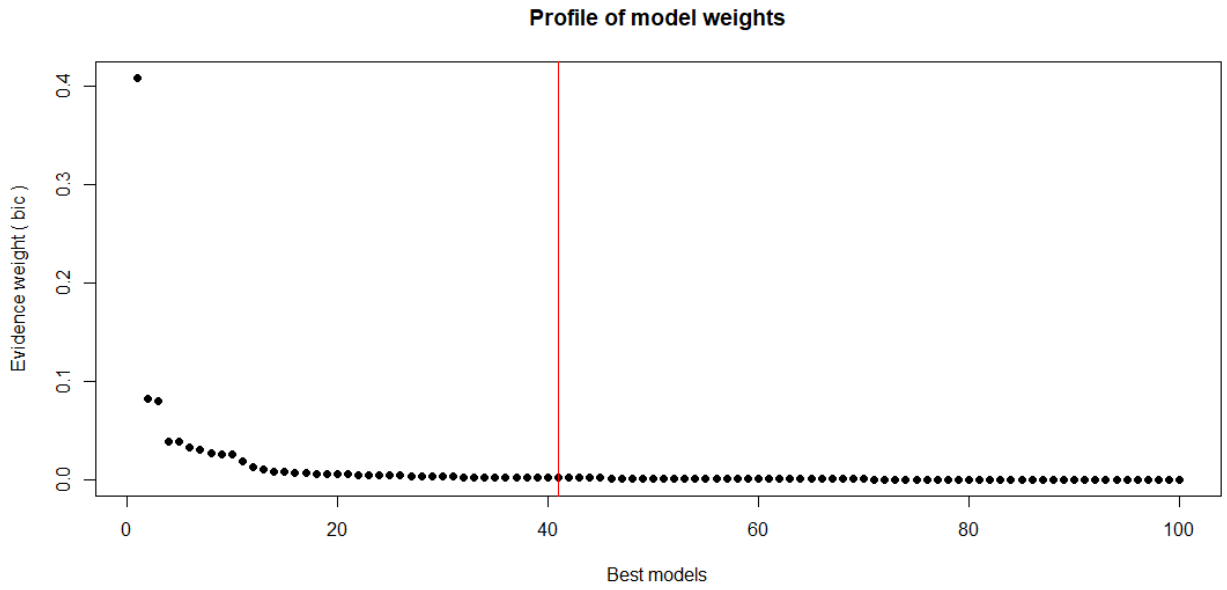

```
> print(glm1)
glmulti.analysis
Method: h / Fitting: glm / IC used: bic
Level: 1 / Marginality: FALSE
From 100 models:
Best IC: 288.748801297729
Best model:
[1] "npo2008trans ~ 1 + TBWDDN + RAXericRat"
Evidence weight: 0.408412586994249
Worst IC: 303.015488193147
1 models within 2 IC units.
40 models to reach 95% of evidence weight.
```

- Model selection here is essentially variable selection.
- Only 1 of 2,048 within 2 IC units (otherwise consider model-averaging).

GPI

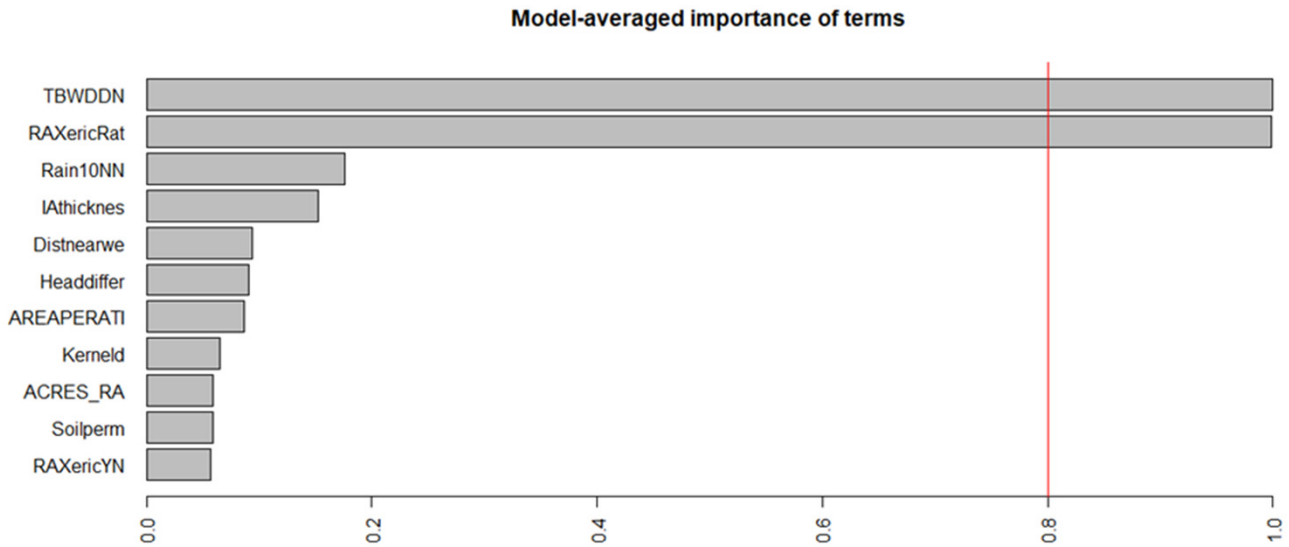
BIC-best model

- Model weight of best model is 41%, indicating 41% probability it is best model out of 2,048 evaluated

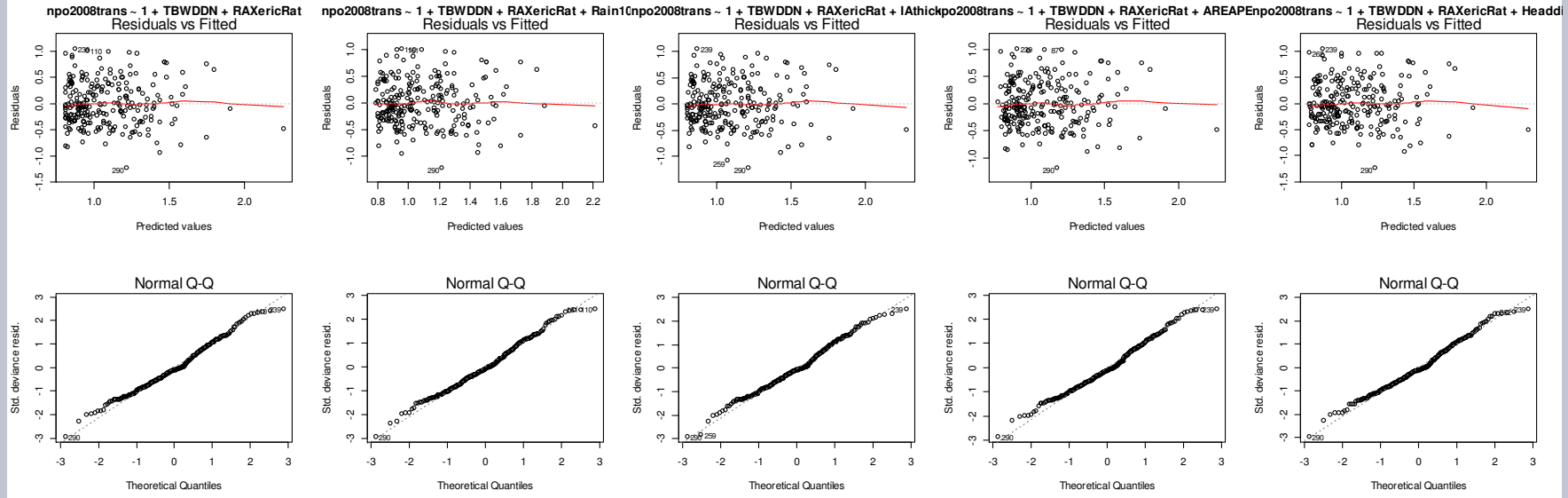


BIC-best model

- Sum of weights of models including each variable allows calculation of variable importance



Checking variance homogeneity and normality



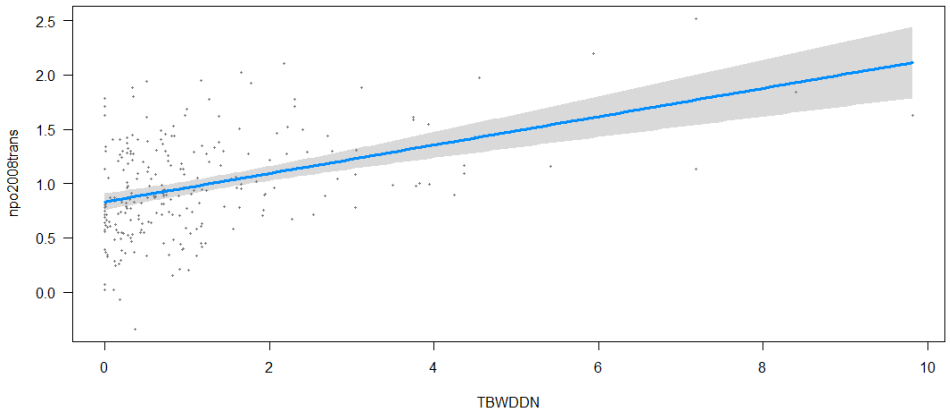
Top five models shown here, although only interested in the best in this case.

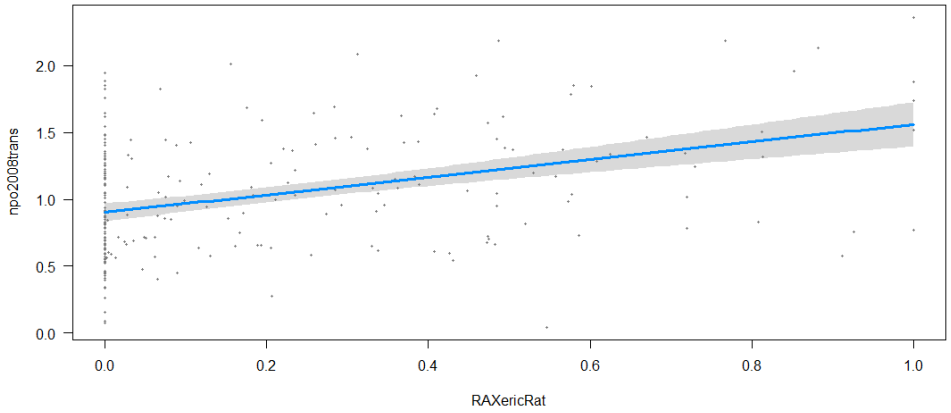
MLR performance on training data

```
lm2<-lm(npo2008trans ~ TBWDDN + RAXericRat,data=train)
> summary(lm2)

Call:
lm(formula = npo2008trans ~ TBWDDN + RAXericRat, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.21824 -0.30282 -0.04177  0.29377  1.04295

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.81027    0.04051  19.999  < 2e-16 ***
TBWDDN       0.13066    0.01896   6.892 4.65e-11 ***
RAXericRat   0.65979    0.10023   6.583 2.81e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4177 on 244 degrees of freedom
Multiple R-squared:  0.245,    Adjusted R-squared:  0.2388
F-statistic: 39.59 on 2 and 244 DF,  p-value: 1.287e-15
```

GPI

R package visreg allows visualization of partial residuals





GPI

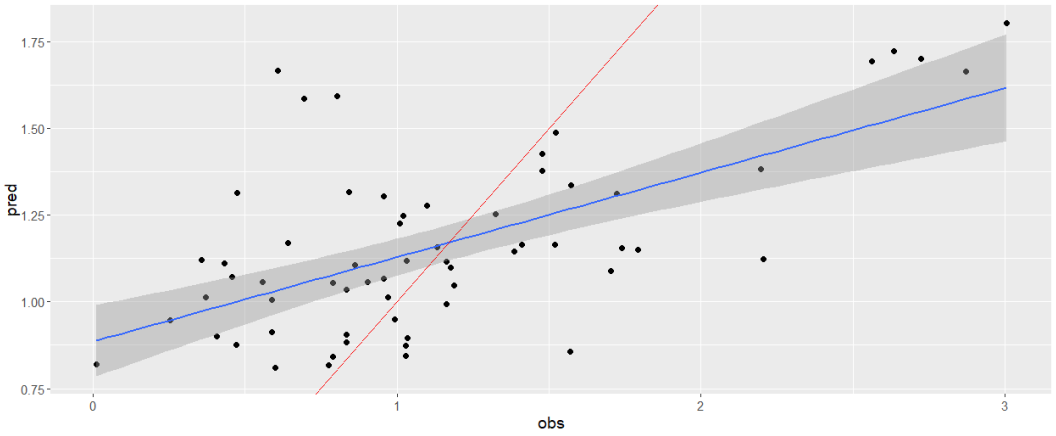
MLR performance on test data (using function defaultSummary from R package caret)

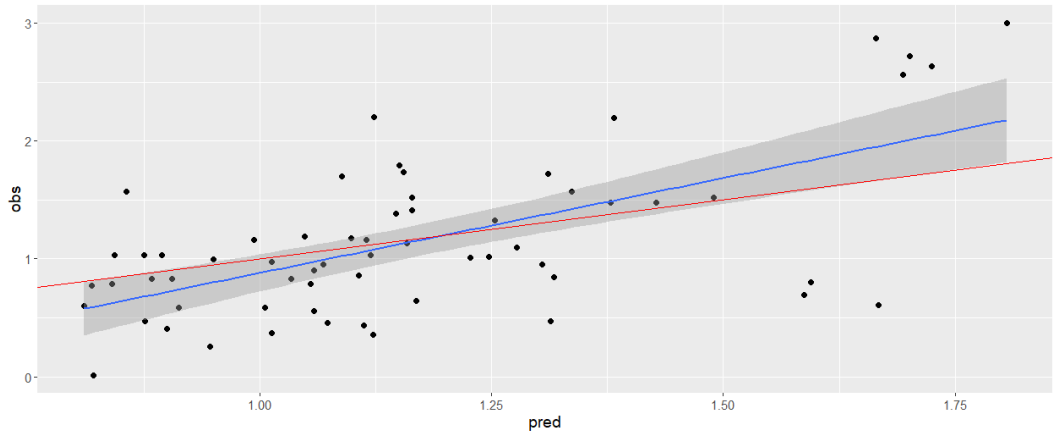
- Evaluating model performance
  - RMSE
  - Rsquared
  - Plot of predicted and observed
- Rsquared on test site higher than training set here
- May be nonrepresentative test set so we will evaluate using cross validation

```
> predtestlm2<-predict(lm2,newdata=test)
> lm2values<-data.frame(obs=test$npo2008trans,pred=predtestlm2)
> defaultSummary(lm2values)
     RMSE  Rsquared       MAE
0.5362219 0.3926008 0.4166636
```

GPI

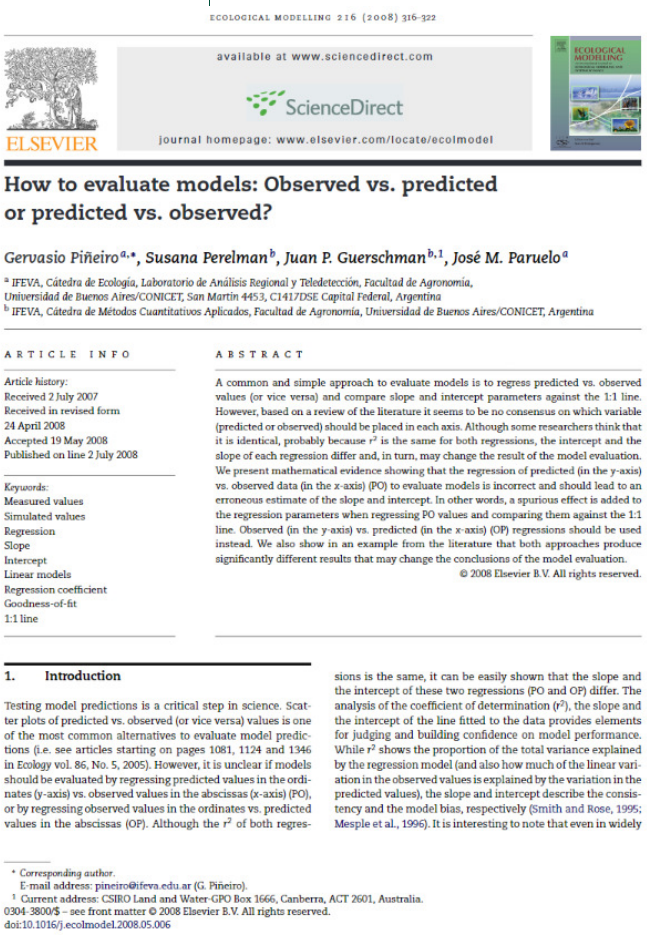
Pred vs. Obs or Obs vs. Pred?





Yes it matters!

- About half of publications get it wrong
- Obs (y) vs Pred (x) allows proper comparison with the 1:1 line



ECOLOGICAL MODELLING 216 (2008) 316–322

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel

ELSEVIER

How to evaluate models: Observed vs. predicted or predicted vs. observed?

Gervasio Piñeiro[a,*], *Susana Perelman*[b], *Juan P. Guerschman*[b,1], *José M. Paruelo*[a]

[a] IFEVA, *Cátedra de Ecología, Laboratorio de Análisis Regional y Teledetección, Facultad de Agronomía, Universidad de Buenos Aires/CONICET, San Martín 4453, C1417DSE Capital Federal, Argentina*

[b] IFEVA, *Cátedra de Métodos Cuantitativos Aplicados, Facultad de Agronomía, Universidad de Buenos Aires/CONICET, Argentina*

ARTICLE INFO

Article history:
Received 2 July 2007
Received in revised form
24 April 2008
Accepted 19 May 2008
Published on line 2 July 2008

Keywords:
Measured values
Simulated values
Regression
Slope
Intercept
Linear models
Regression coefficient
Goodness-of-fit
1:1 line

ABSTRACT

A common and simple approach to evaluate models is to regress predicted vs. observed values (or vice versa) and compare slope and intercept parameters against the 1:1 line. However, based on a review of the literature it seems to be no consensus on which variable (predicted or observed) should be placed in each axis. Although some researchers think that it is identical, probably because r^2 is the same for both regressions, the intercept and the slope of each regression differ and, in turn, may change the result of the model evaluation. We present mathematical evidence showing that the regression of predicted (in the y-axis) vs. observed data (in the x-axis) (PO) to evaluate models is incorrect and should lead to an erroneous estimate of the slope and intercept. In other words, a spurious effect is added to the regression parameters when regressing PO values and comparing them against the 1:1 line. Observed (in the y-axis) vs. predicted (in the x-axis) (OP) regressions should be used instead. We also show in an example from the literature that both approaches produce significantly different results that may change the conclusions of the model evaluation.

1. Introduction

Testing model predictions is a critical step in science. Scatter plots of predicted vs. observed (or vice versa) values is one of the most common alternatives to evaluate model predictions (i.e. see articles starting on pages 1081, 1124 and 1346 in *Ecology* vol. 86, No. 5, 2005). However, it is unclear if models should be evaluated by regressing predicted values in the ordinates (y-axis) vs. observed values in the abscissas (x-axis) (PO), or by regressing observed values in the ordinates vs. predicted values in the abscissas (OP). Although the r^2 of both regressions is the same, it can be easily shown that the slope and the intercept of these two regressions (PO and OP) differ. The analysis of the coefficient of determination (r^2), the slope and the intercept of the line fitted to the data provides elements for judging and building confidence on model performance. While r^2 shows the proportion of the total variance explained by the regression model (and also how much of the linear variation in the observed values is explained by the variation in the predicted values), the slope and intercept describe the consistency and the model bias, respectively (Smith and Rose, 1995; Mesple et al., 1996). It is interesting to note that even in widely

* *Corresponding author.*
E-mail address: pineiro@ifeva.edu.ar (G. Piñeiro).
[1] Current address: CSIRO Land and Water-GPO Box 1666, Canberra, ACT 2601, Australia.
0304-3800/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2008.05.006

MLR with all of data and 10-fold CV using R package caret

```
> data_ctrl <- trainControl(method = "cv", number = 10)
> model_caret <- train(npo2008trans ~ TBWDDN + RAXericRat,   # model to fit
+                      data = data,
+                      trControl = data_ctrl,               # folds
+                      method = "lm"                        # specifying regre
ssion model
+ )
> model_caret
Linear Regression

309 samples
  2 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 279, 277, 278, 279, 278, 280, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.4418519  0.2819027  0.3483197
```

So the 10-fold CV data results are between the train and the test. The results represent likely performance on future out-of-sample data (in our case—unmonitored sites).

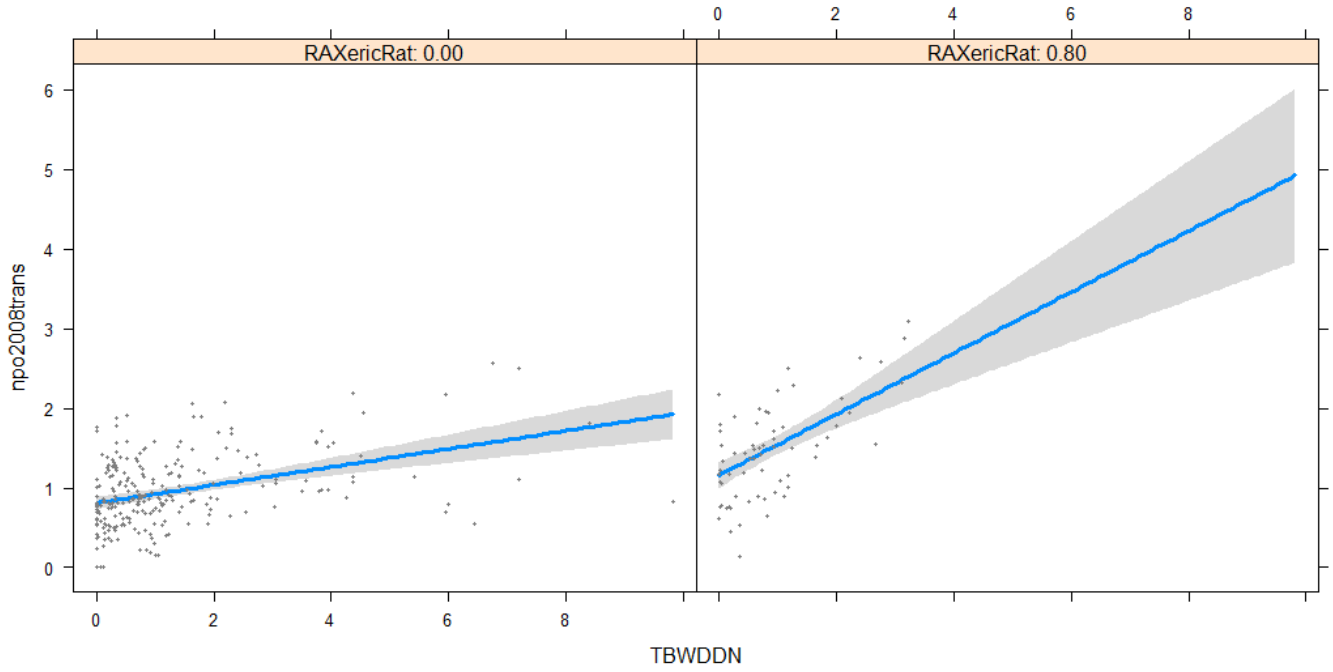
MLR with all of data and 10-fold CV, allowing interaction term

```
> # 10-fold CV with interaction term
> data_ctrl <- trainControl(method = "cv", number = 10)
> model_caret2 <- train(npo2008trans ~ TBWDDN * RAXericRat,   # model to fit
+                       data = data,
+                       trControl = data_ctrl,              # folds
+                       method = "lm"                       # specifying regression model
+ )
> model_caret2$finalModel


> model_caret2
Linear Regression

309 samples
  2 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 277, 278, 279, 278, 280, 280, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.4336405  0.3203806  0.3447678
```

Allowing our two BIC-selected variables to interact (essentially creating a third variable) has increased the CV-Rsquared by about 3% and lowered the RMSE by 0.01

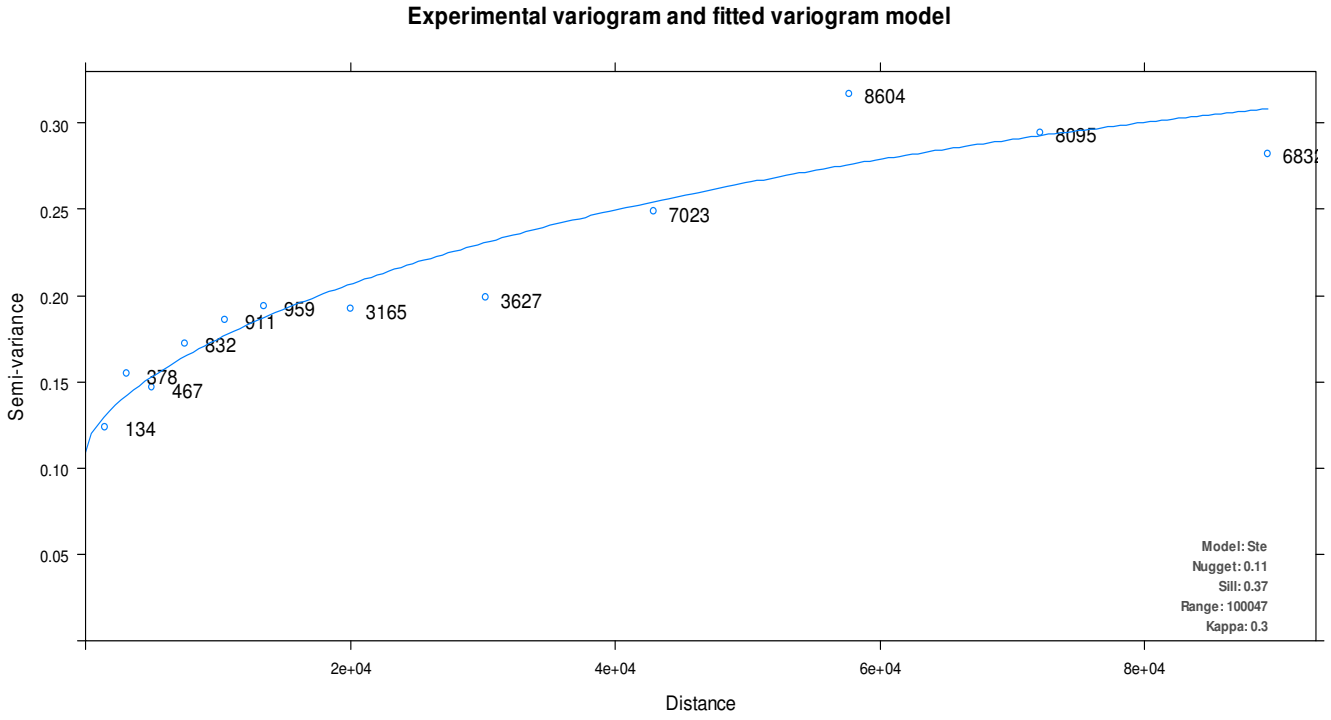
GPI

visreg plot showing interaction of SASDDN and Xeric Ratio



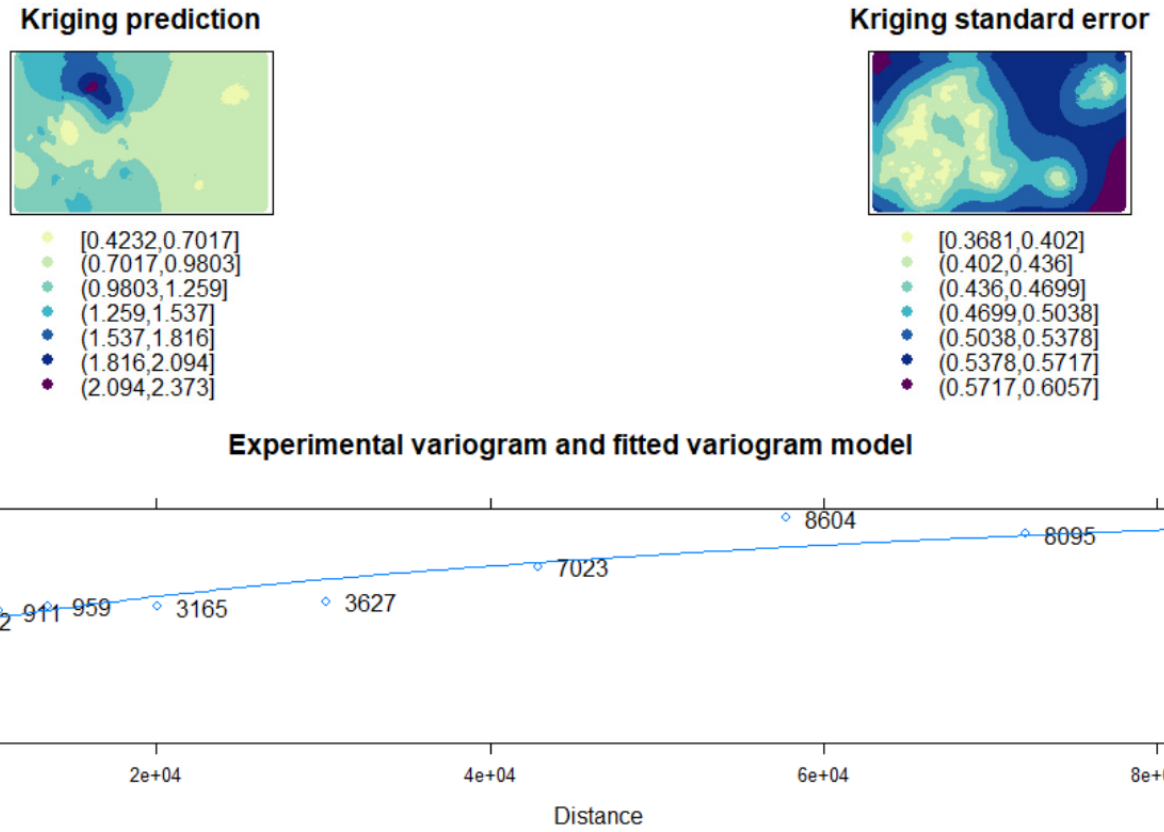
High xeric ratio sites (here shown as >80%) show lower water levels at same level of SAS drawdown

Ordinary Kriging using R package automap: ignore the auxiliary variables and look at prediction using location and values of dependent variable only




```
vrok=autofitVariogram(npo2008trans  ~ 1, modataspat)
plot(vrok)
```

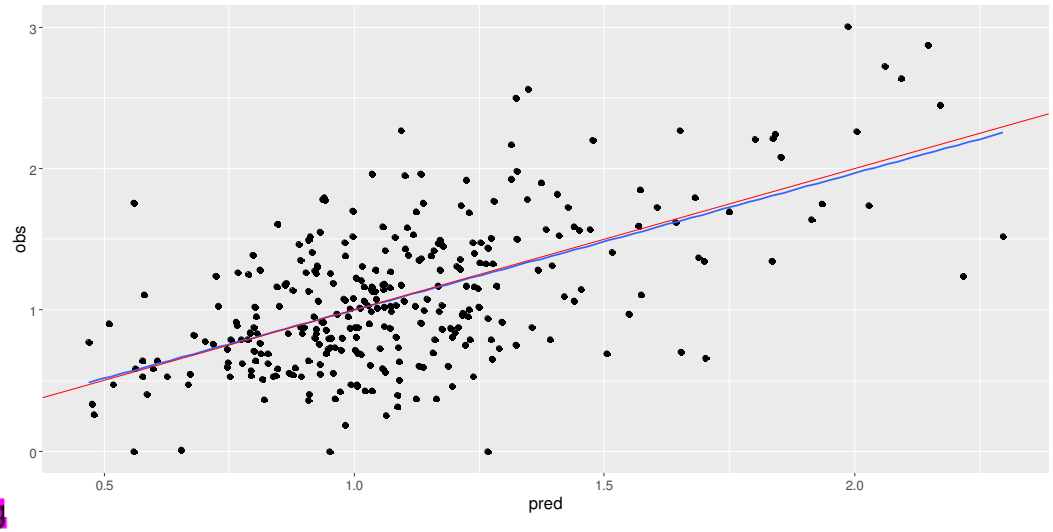
Ordinary Kriging provides prediction map and spatially-explicit standard error



plot(ok_krig_studyarea)

automap conveniently includes a leave one out cross validation routine

- Ordinary Kriging prediction is strong for this problem:
  - Rsquared 3% higher than best MLR (interaction)
  - RMSE 0.016 lower

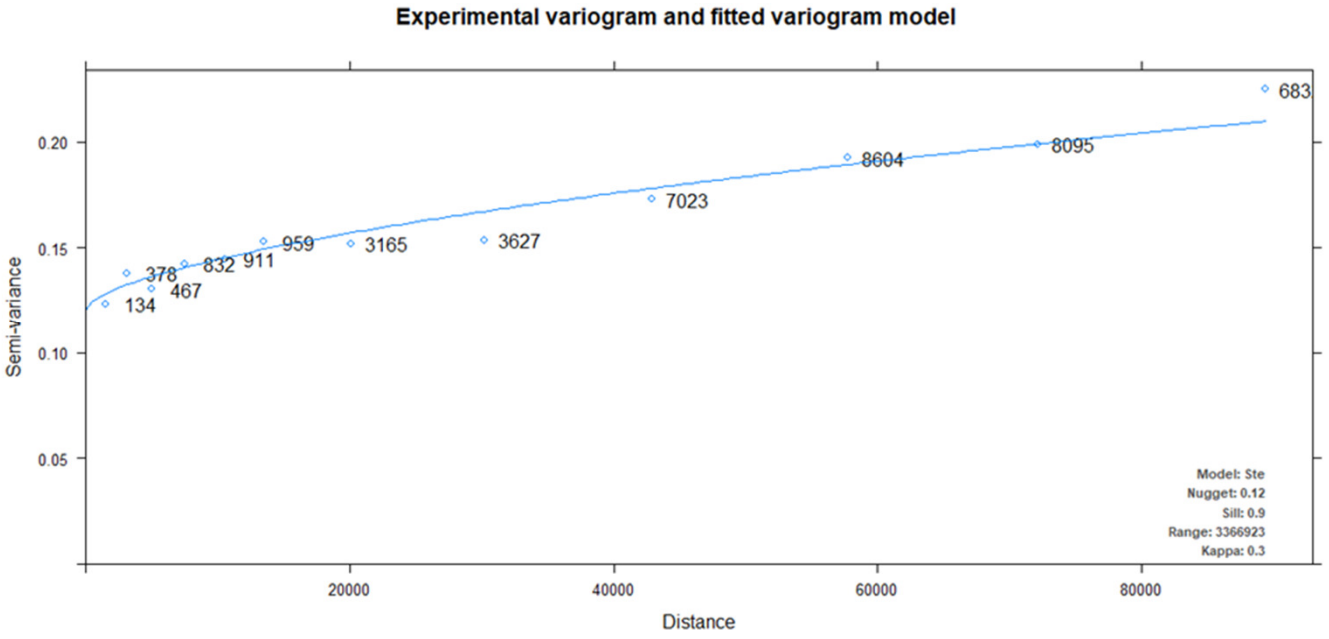



```
> defaultSummary(cvrkvaluesok)
     RMSE  Rsquared       MAE
0.4171600 0.3559196 0.3263445
Final results from Ordinary Kriging
```

kr.cv.ok=autoKrige.cv(npo2008trans ~ 1, modataspat)

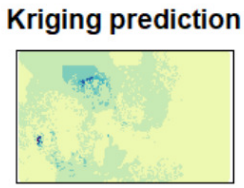
Regression Kriging: combining our best MLR and Kriging

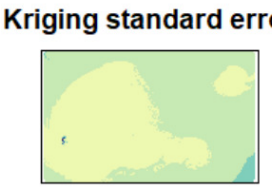
Not as much variance to explain with kriging because this is residual variance only

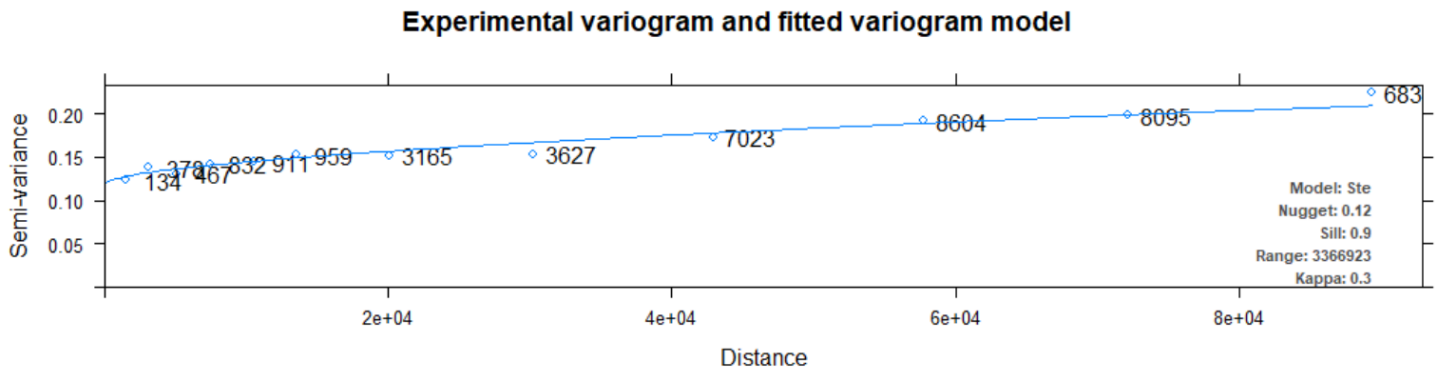


vr1=autofitVariogram(npo2008trans ~ 1+TBWDDN*RAXericRat, datark)
plot(vr1)

Regression Kriging Results





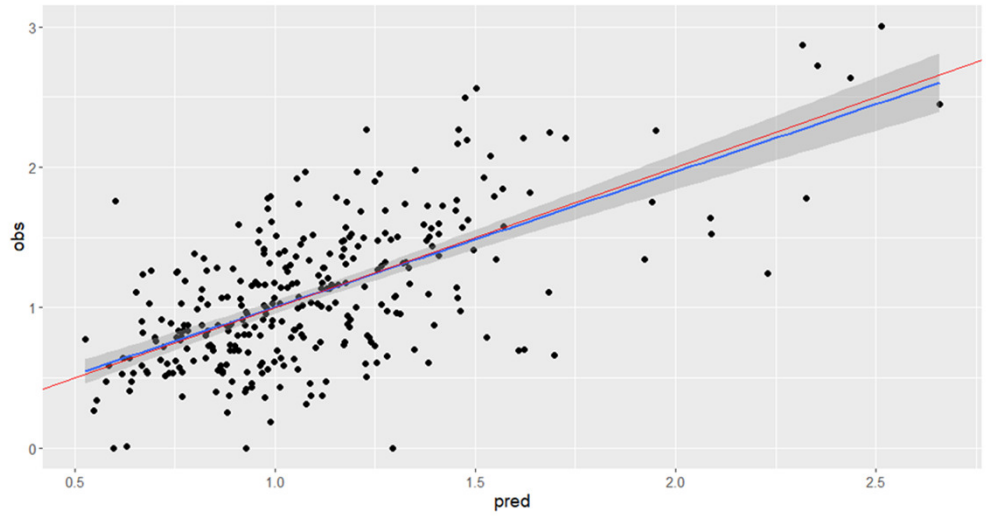


Plot(reg_krig_studyarea)

automap cross validation works for regression kriging too

- Regression Kriging prediction is better:
  - Rsquared 5% higher than Ordinary Kriging
  - RMSE 0.016 lower

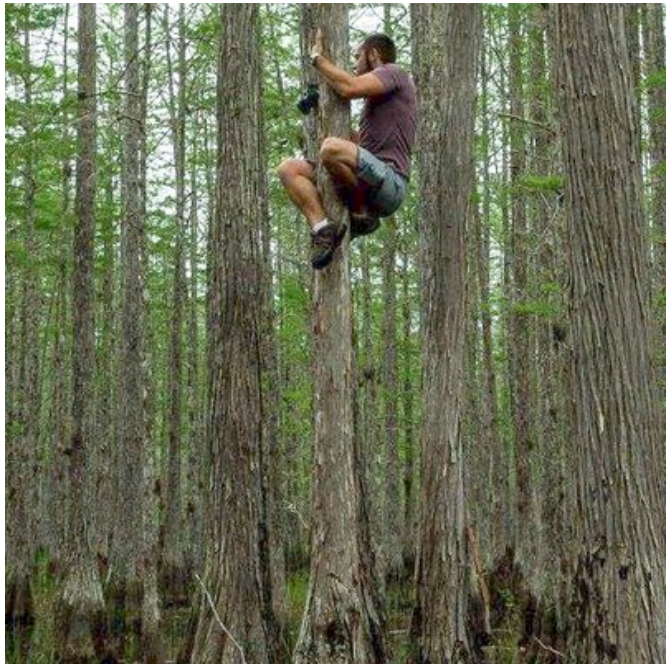
```
> defaultSummary(cvrkvalues)
     RMSE  Rsquared       MAE
0.4006853 0.4059031 0.3126121
```



kr.cv.ok=autoKrige.cv(npo2008trans ~ 1, modataspat)

Regression Kriging doesn't have to use "Regression"

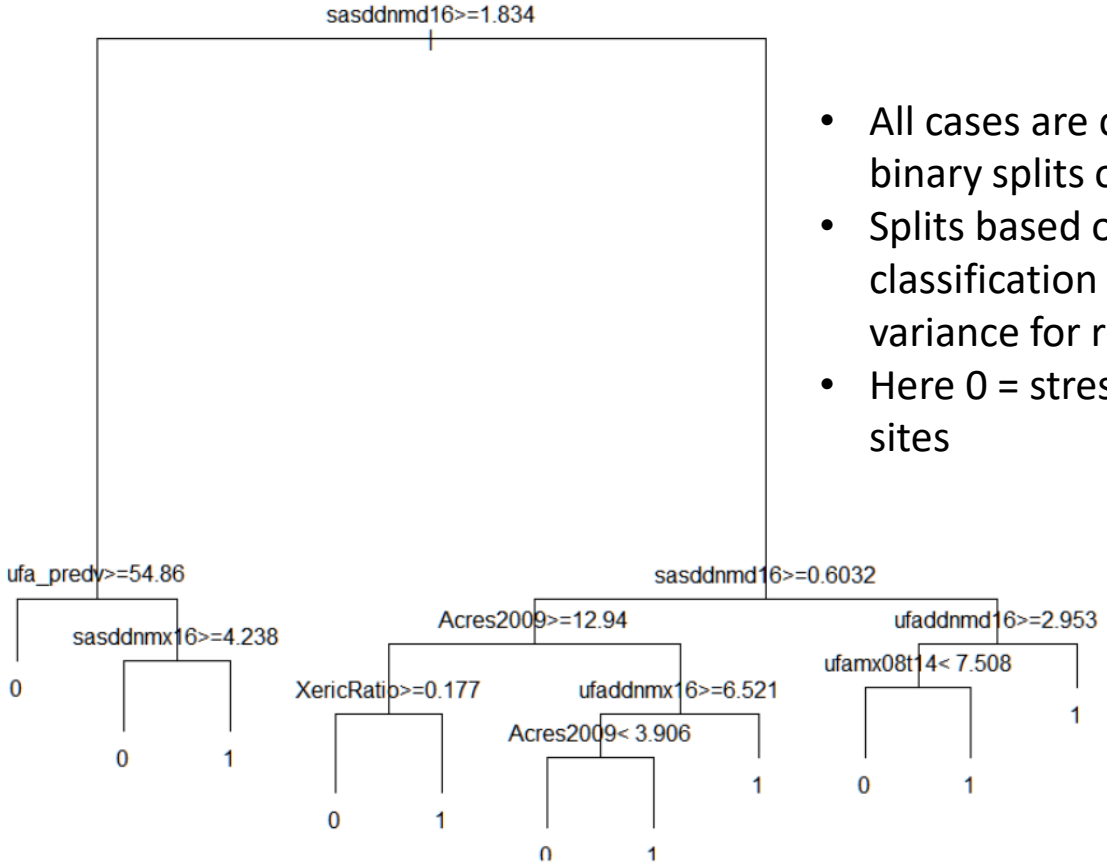
- You can substitute your favorite machine learning algorithm in place of regression
- Let's see what Random Forest can do



https://www.dannygoodding.com/About

A quick review…

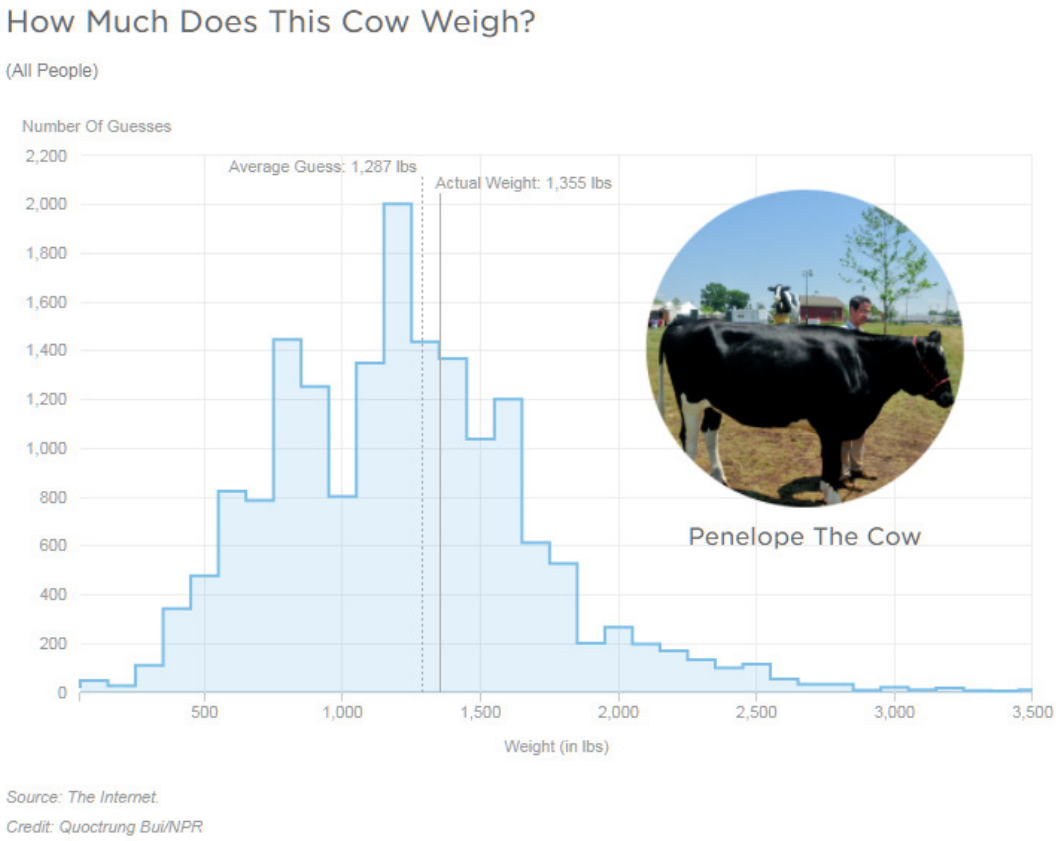
Example decision tree (CART)



- All cases are classified based on recursive binary splits of the variables.
- Splits based on maximizing node purity for classification problems or minimizing variance for regression
- Here 0 = stressed sites and 1 = unstressed sites

GPI

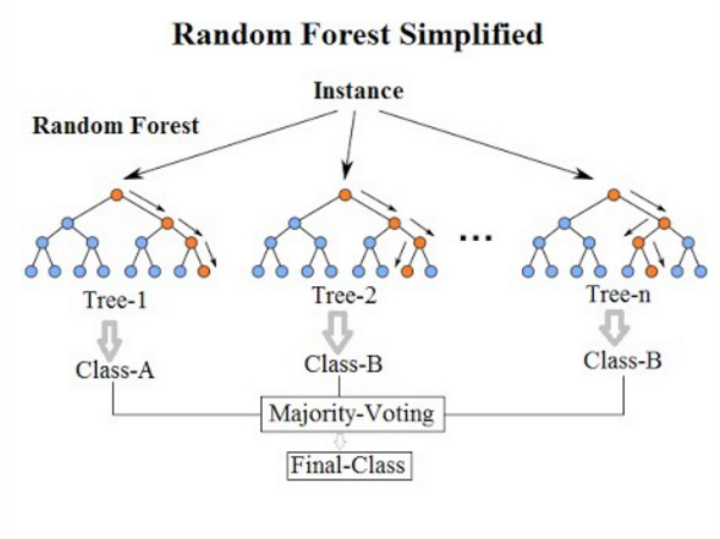
The wisdom of crowds in machine learning: ensemble techniques (bagging and boosting)



https://www.npr.org/sections/money/2015/08/07/429720443/17-205-people-guessed-the-weight-of-a-cow-heres-how-they-did

Random Forest

- Can be used for classification or regression
- Creates “forest” of decision trees using subset of cases and variables (bootstrap aggregation or “bagging”)
- Predictions are made by running case through all trees (e.g., 500) and taking the majority vote for classification
- Each tree is grown with only about 60% of the data, so the remaining data provides a conservative “out-of-bag” error estimate
- Robust to outliers and noise
- No statistical assumptions
- Handles datasets >1000 variables
- Avoids overfitting
- Minimal tuning required
- Limitations
  - Can’t extrapolate
  - May overestimate lows/underestimate highs



https://community.tibco.com/wiki/random-forest-template-tibco-spotfirer-wiki-page

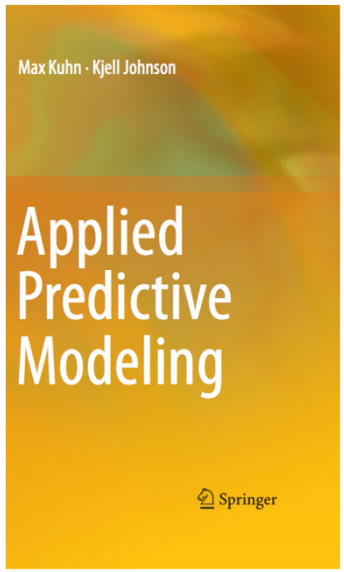
Random Forest (starting again with training data)

- Out of the box, no tuning, compared to MLR on training data:
  - OOB Rsquared 7% lower
  - RMSE 0.02 higher

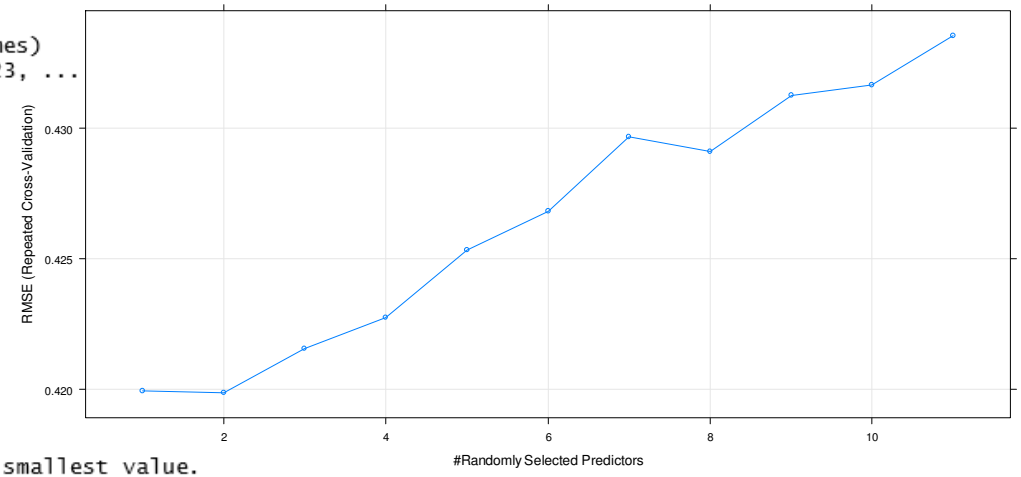
```
> print(rf1)
 Call:
  randomForest(formula = npo2008trans ~ TBWDDN + RAXericRat + RAXericYN +      ACRES_
RA + AREAPERATI + Distnearwe + Headdiffer + Soilperm +      IAthicknes + Rain10NN +
Kerneld, data = motrain)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 0.1865343
                    % Var explained: 18.3

> 0.1865343^0.5
[1] 0.4318962
```

GPI

Tuning hyperparameter mtry using R package caret (training data)




```
> metric<-"RMSE"
> set.seed(9)
> rf_gridsearch <- train(npo2008trans~TBWDDN+RAXericRat+RAXericYN+ACRES_RA+AR
EAPERATI+Distnearwe+Headdiffer+Soilperm+IAthicknes+Rain10NN+Kerneld,data=motr
ain, method="rf", metric=metric, tuneGrid=tunegrid, trControl=control)
> print(rf_gridsearch)
Random Forest

247 samples
 11 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 223, 222, 223, 222, 222, 223, ...
Resampling results across tuning parameters:

  mtry  RMSE       Rsquared   MAE
   1    0.4199242  0.2510676  0.3359679
   2    0.4198637  0.2417899  0.3348209
   3    0.4215435  0.2391825  0.3361488
   4    0.4227394  0.2372895  0.3371206
   5    0.4253147  0.2325872  0.3385885
   6    0.4268244  0.2299124  0.3395089
   7    0.4296466  0.2220067  0.3421373
   8    0.4291115  0.2251105  0.3413117
   9    0.4312362  0.2217112  0.3436414
  10    0.4316388  0.2216098  0.3430167
  11    0.4335335  0.2163469  0.3450305

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 2.
> plot(rf_gridsearch)
```



Notice the previous out-of-bag error for mtry=3 was conservative.

Hyperparameter ntree may also need "tuning" (training data)

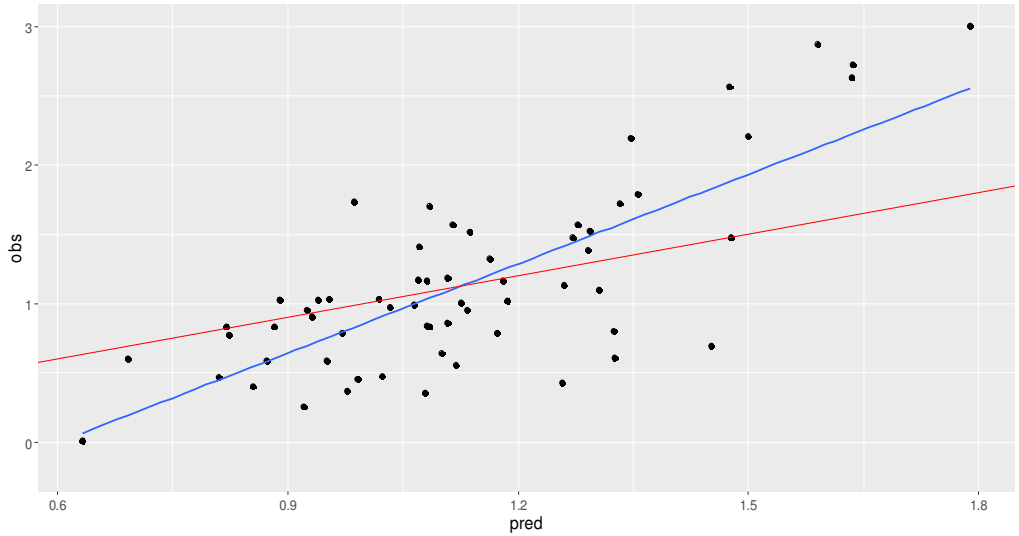
- Although there is debate about whether ntree affects performance beyond a certain size (e.g., default is 500 trees), I have found increasing to 10,000 usually helps
- Here OOB Rsquared bumps up 2.59% and RMSE down 0.006

```
rf1m2n10<-randomForest(npo2008trans~TBWDDN+RAXericRat+RAXericYN+ACRES_RA+AREAPERATI+
Distnearwe+Headdiffer+Soilperm+IAthicknes+Rain10NN+Kerneld,mtry=2,ntree=10000,data=m
otrain)
> print(rf1m2n10)

Call:
 randomForest(formula = npo2008trans ~ TBWDDN + RAXericRat + RAXericYN +      ACRES_
RA + AREAPERATI + Distnearwe + Headdiffer + Soilperm +      IAthicknes + Rain10NN +
Kerneld, data = motrain, mtry = 2,      ntree = 10000)
               Type of random forest: regression
                     Number of trees: 10000
No. of variables tried at each split: 2

          Mean of squared residuals: 0.1806239
                    % Var explained: 20.89

> 0.1806239^0.5
[1] 0.4249987
```

Random Forest on Test Data

```
> predtestrf1m2n10<-predict(rf1m2n10,newdata=motest)
> rf1values<-data.frame(obs=motest$npo2008trans,pred=predtestrf1m2n10)
> defaultSummary(rf1values)
     RMSE  Rsquared       MAE
0.5026328 0.5836772 0.3817792
```

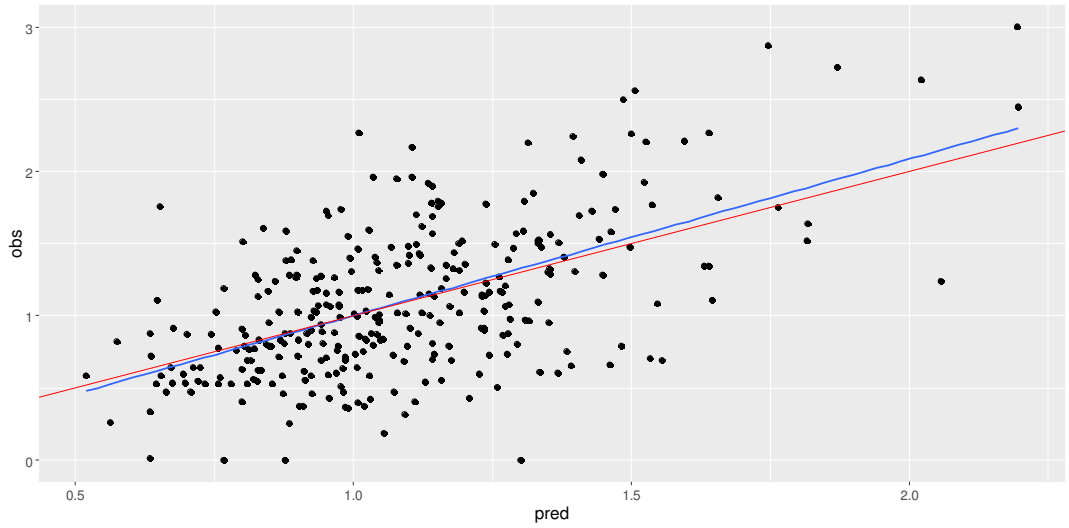
- Rsquared great on test data but RMSE has moved in the wrong direction
- Obs vs. Exp plot not following 1:1 line very well (model underpredicting at high values)
- Is this bias or chance result?



Refitting Random Forest to all data

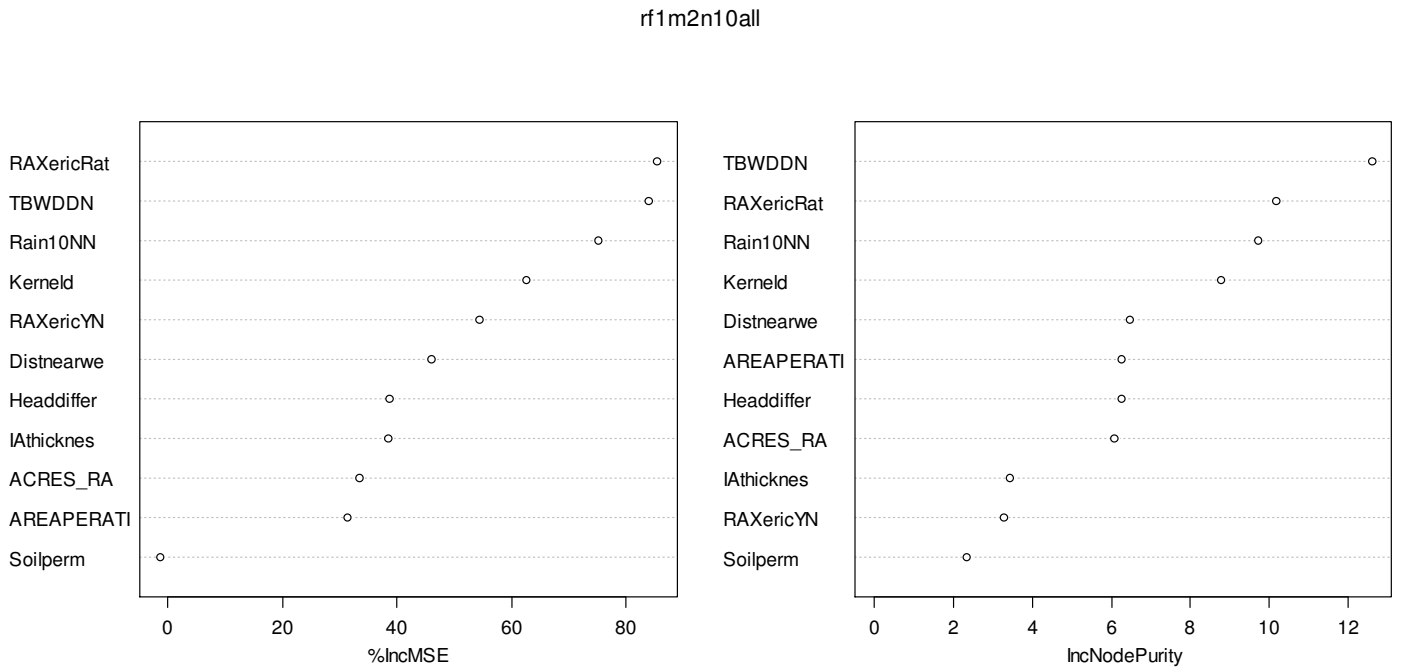
```
> set.seed(9)
> rf1m2n10all<-randomForest(npo2008trans~TBWDDN+RAXericRat+RAXericYN+ACRES_RA
+AREAPERATI+Distnearwe+Headdiffer+Soilperm+IAthicknes+Rain10NN+Kerneld,mtry=2
,ntree=10000,importance=TRUE,data=modata)
> print(rf1m2n10all)

Call:
 randomForest(formula = npo2008trans ~ TBWDDN + RAXericRat + RAXericYN +
ACRES_RA + AREAPERATI + Distnearwe + Headdiffer + Soilperm +      IAthicknes
+ Rain10NN + Kerneld, data = modata, mtry = 2,      ntree = 10000, importance
= TRUE)
               Type of random forest: regression
                     Number of trees: 10000
No. of variables tried at each split: 2

          Mean of squared residuals: 0.180084
                    % Var explained: 33.3
> 0.1800846^0.5
[1] 0.4243638
The final RF model for reporting results
```

Rsquared better and bias doesn't look to be an issue



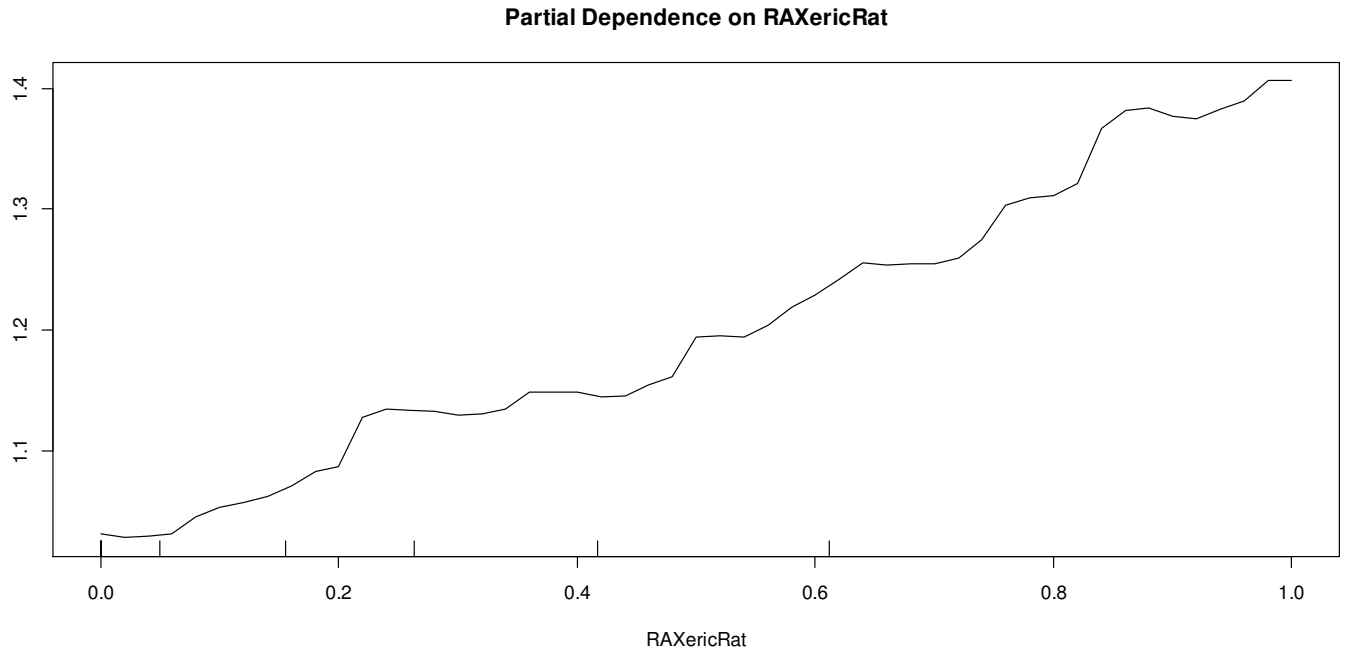
Peeking into the Random Forest– a “black box” model

Random Forest Variable Importance



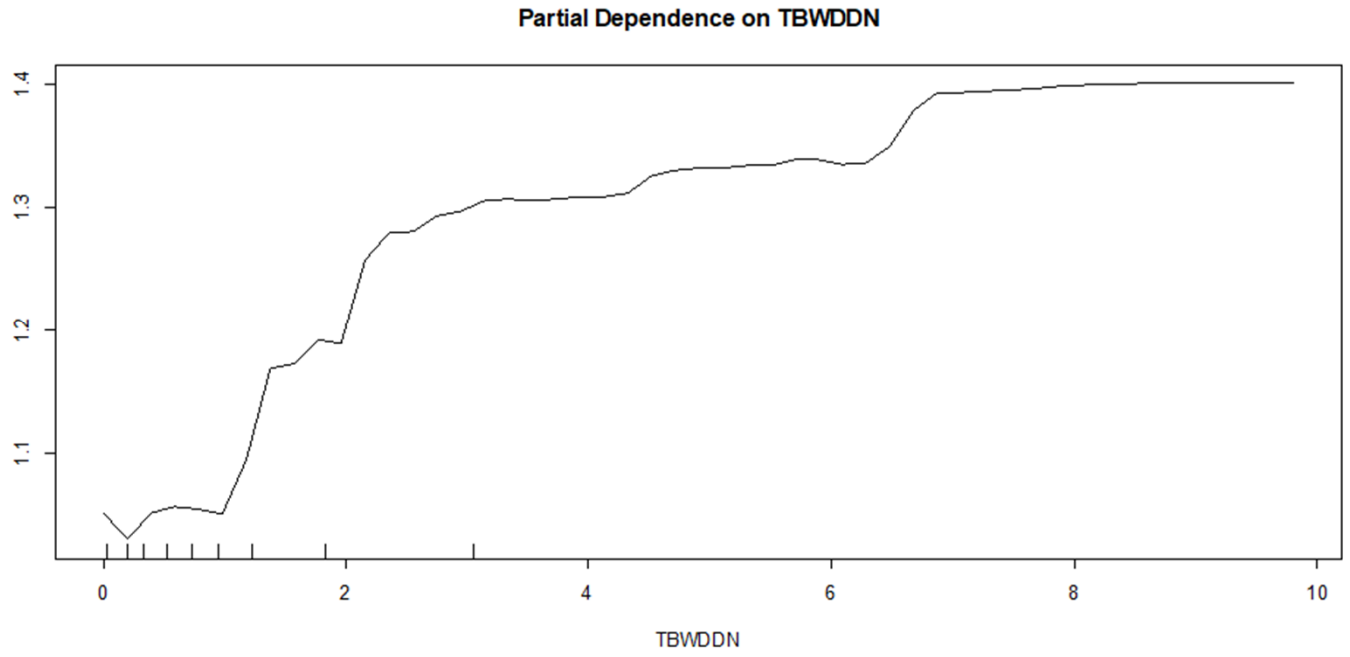
varImpPlot(rf1m2n10all)

Random Forest Partial Dependence Plot



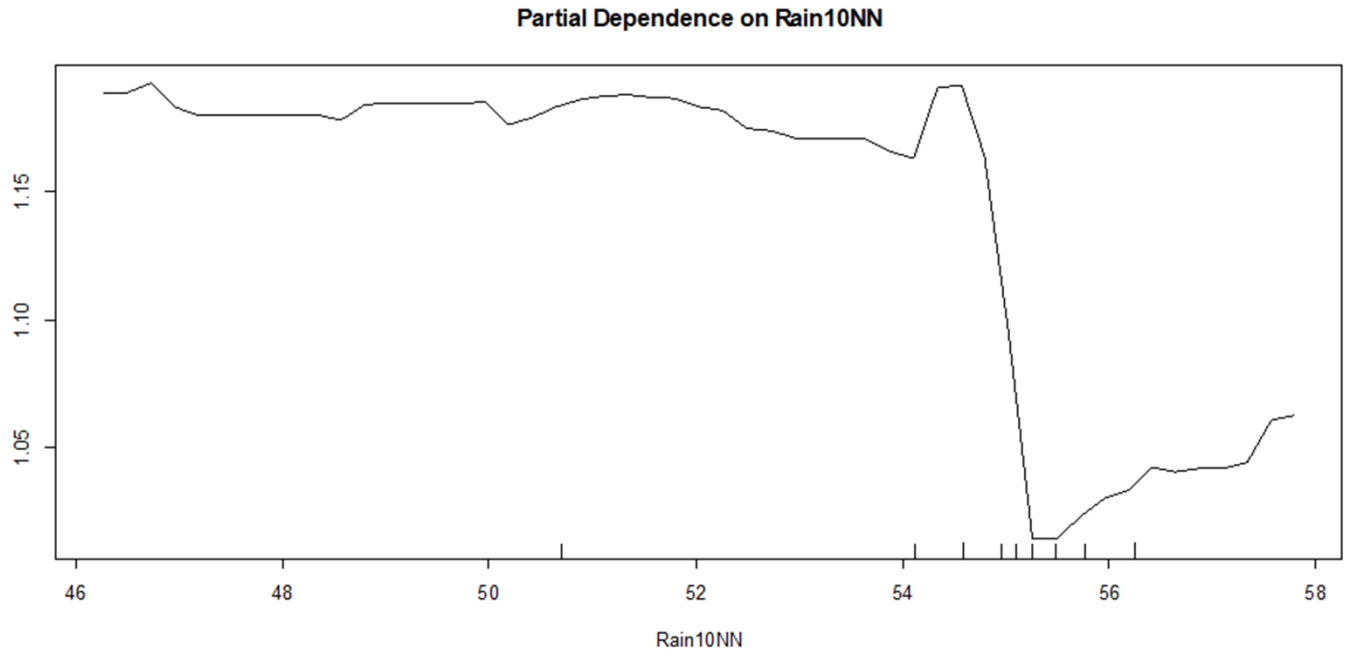
Increasing Xeric Ratio is associated with decreasing water levels (higher transformed NPOs)

Random Forest Partial Dependence Plot



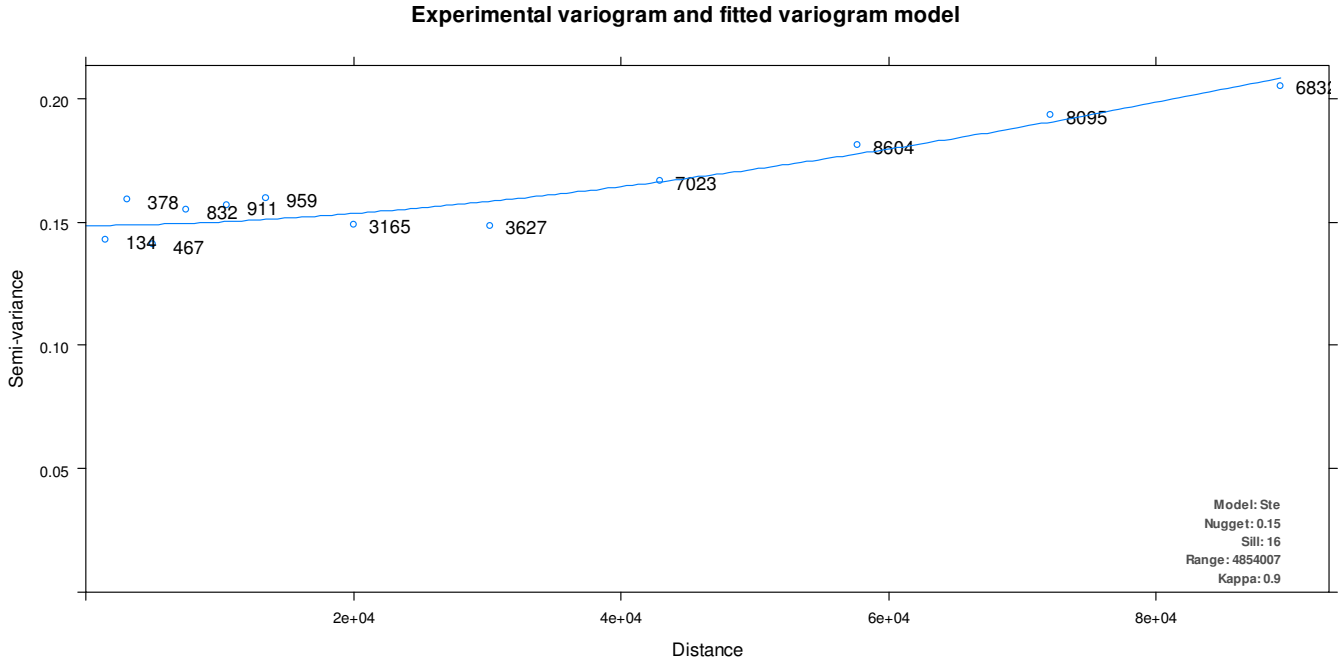
Increasing SAS DDN is associated with decreasing water levels (higher transformed NPOs)

Random Forest Partial Dependence Plot



The highest rainfall is associated with increasing water levels (lower transformed NPOs)

Residuals from Random Forest do show positive spatial autocorrelation, opening door to kriging




```
modataspat<-data.frame(modataeval,XCOORD=data$XCOORD,YCOORD=data$YCOORD)
coordinates(modataspat) = ~ XCOORD + YCOORD # create SpatialPointsDataFrame object from data
vrrf1=autofitVariogram(resid  ~ 1, modataspat)
```

Developed code to perform LOOCV for Random Forest Kriging

```
# loop to evaluate random forest kriging on loocv
set.seed(99)
xy<-modataspat # to be replaced with data to undergo LOOCV
n_train<-nrow(xy)
loocv_tmp <- matrix(NA, nrow = n_train)
for (k in 1:n_train) {
  train_xy <- xy[-k, ]
  test_xy <- xy[k, ]
  fitforest<-randomForest(npo2008trans~TBWDDN+RAXericRat+RAXericYN+ACRES_RA+AREAPERATI
+Distnearwe+Headdiffer+Soilperm+IAthicknes+Rain10NN+Kerneld,mtry=2,ntree=10000,data=train_xy)
  predforest<-predict(fitforest,newdata=test_xy)
  residsforest<-train_xy$npo2008trans-predict(fitforest)
  train_xyappend<-train_xy
  train_xyappend$resids<-residsforest
  fitted_models<-autoKrige(residsforest ~ 1, train_xyappend,test_xy)
  predictions<-fitted_models$krige_output$var1.pred
  loocv_tmp[k]<-predictions
  loocv_tmp[k]<-loocv_tmp[k]+predforest
  cat(k)
  cat(" ")
}
cat("\n")
cat("r2 between loocv predications and actual:")
cor(loocv_tmp,xy$npo2008trans)^2
cat("complete")
```

R package GSIF offers possibility of using random forest, but doesn't include built-in cross validation

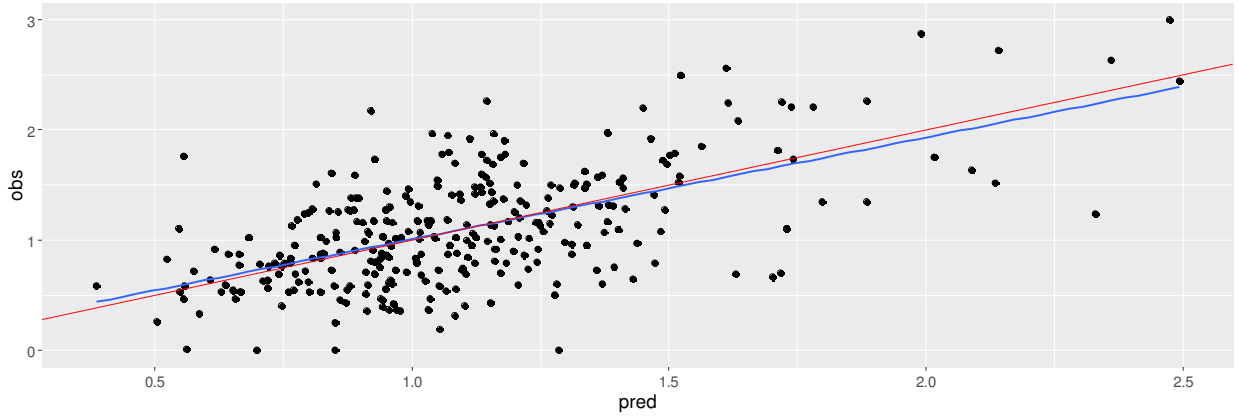
Holds out one point at a time from both random forest and kriging to represent out-of-sample evaluation

GPI

LOOCV Results for Random Forest Kriging

- With the kriging, Random Forest is still 4% Rsquared behind Regression Kriging and 0.014 RMSE worse

```
cvpredsrfm2n10k<-loocv_tmp
 > defaultSummary(cvrfvalues)
      RMSE   Rsquared        MAE
0.4147617 0.3652974 0.3236718
```



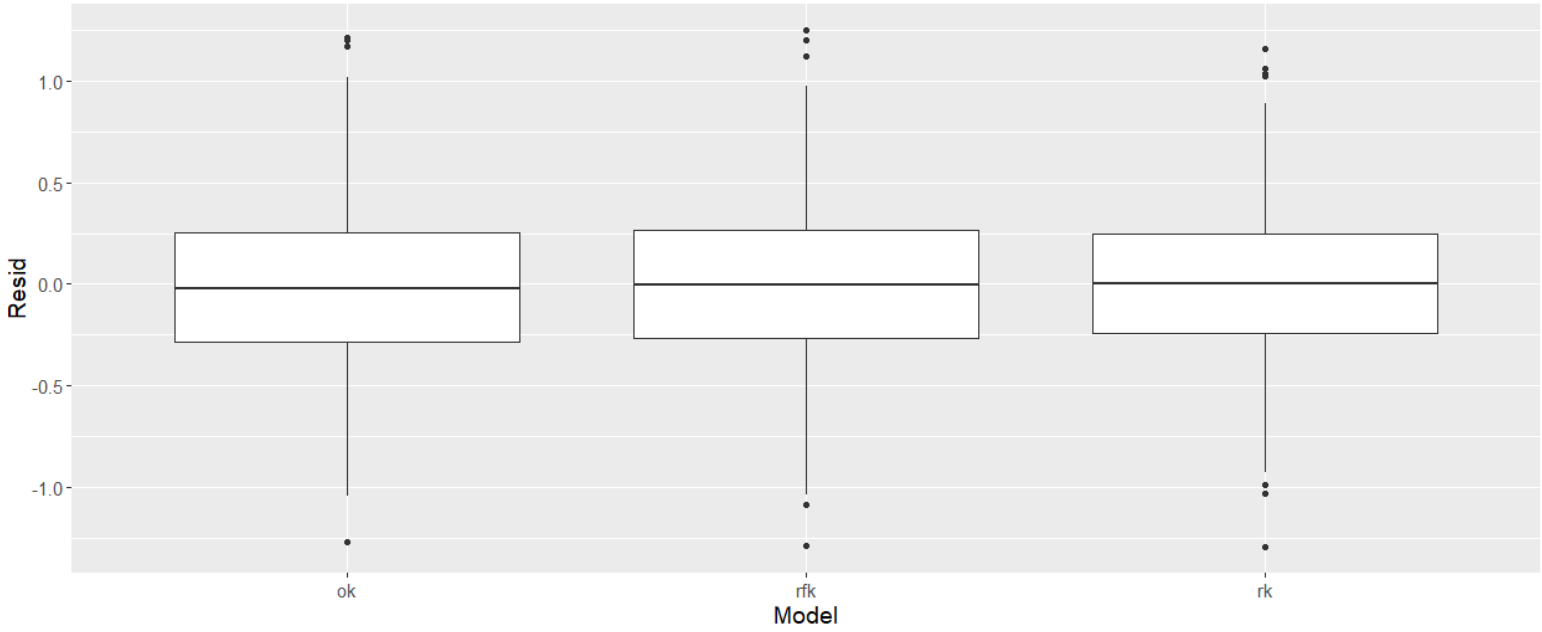
Summary of Models

- Linear Regression Kriging with just two variables and interaction the best
- Some problems are inherently linear and adding extra uninformative variables doesn't help

Model	RMSE	Rsquared	MAE	Evaluation
Linear Regression	0.442	0.282	0.348	10-fold CV
Linear Regression w/ Interaction	0.434	0.320	0.345	10-fold CV
Ordinary Kriging	0.417	0.356	0.326	LOOCV
Linear Regression Kriging	**0.401**	**0.406**	**0.313**	**LOOCV**
Random Forest	0.424	0.333	0.334	OOB
Random Forest Kriging	0.415	0.365	0.324	LOOCV

GPI

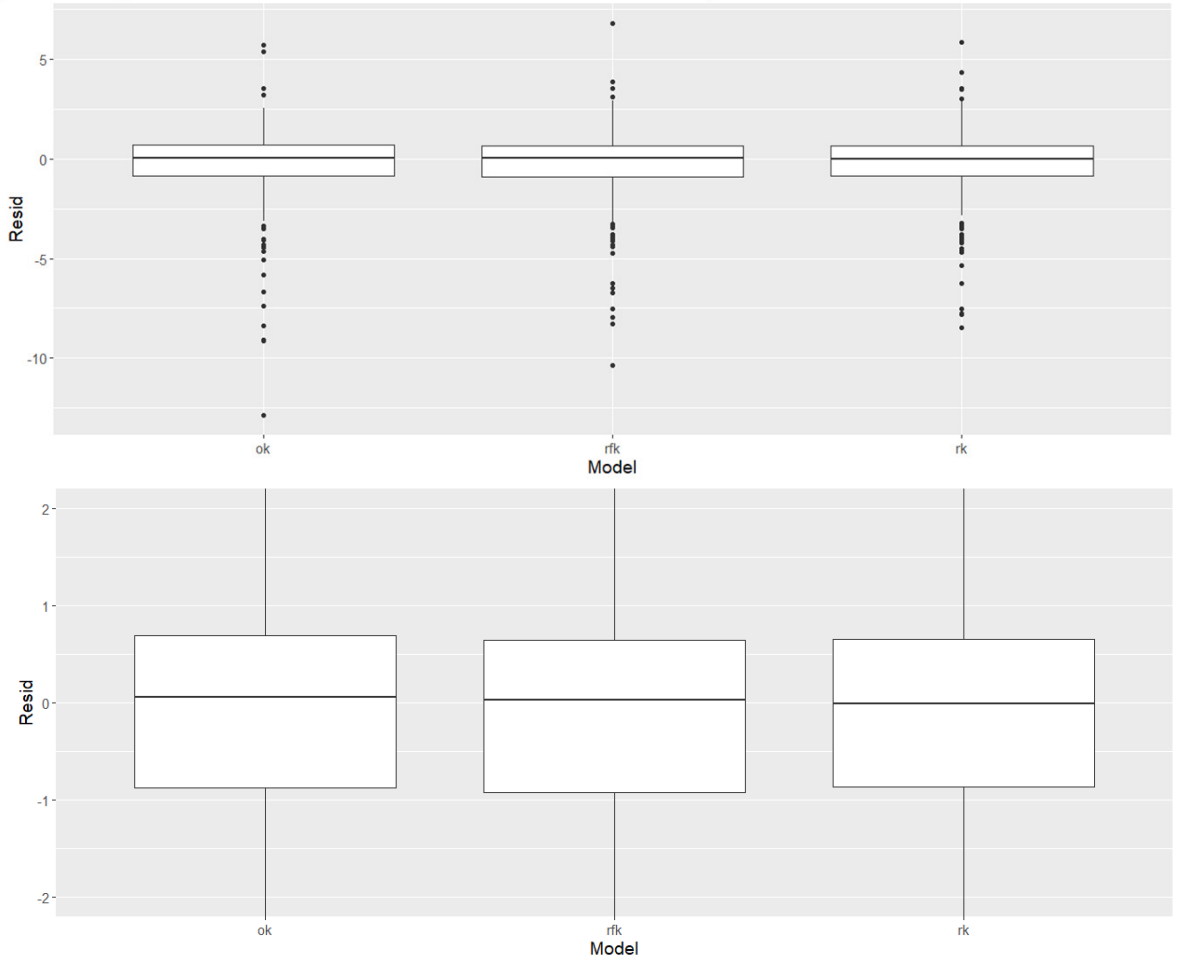
Residuals similar among top three models



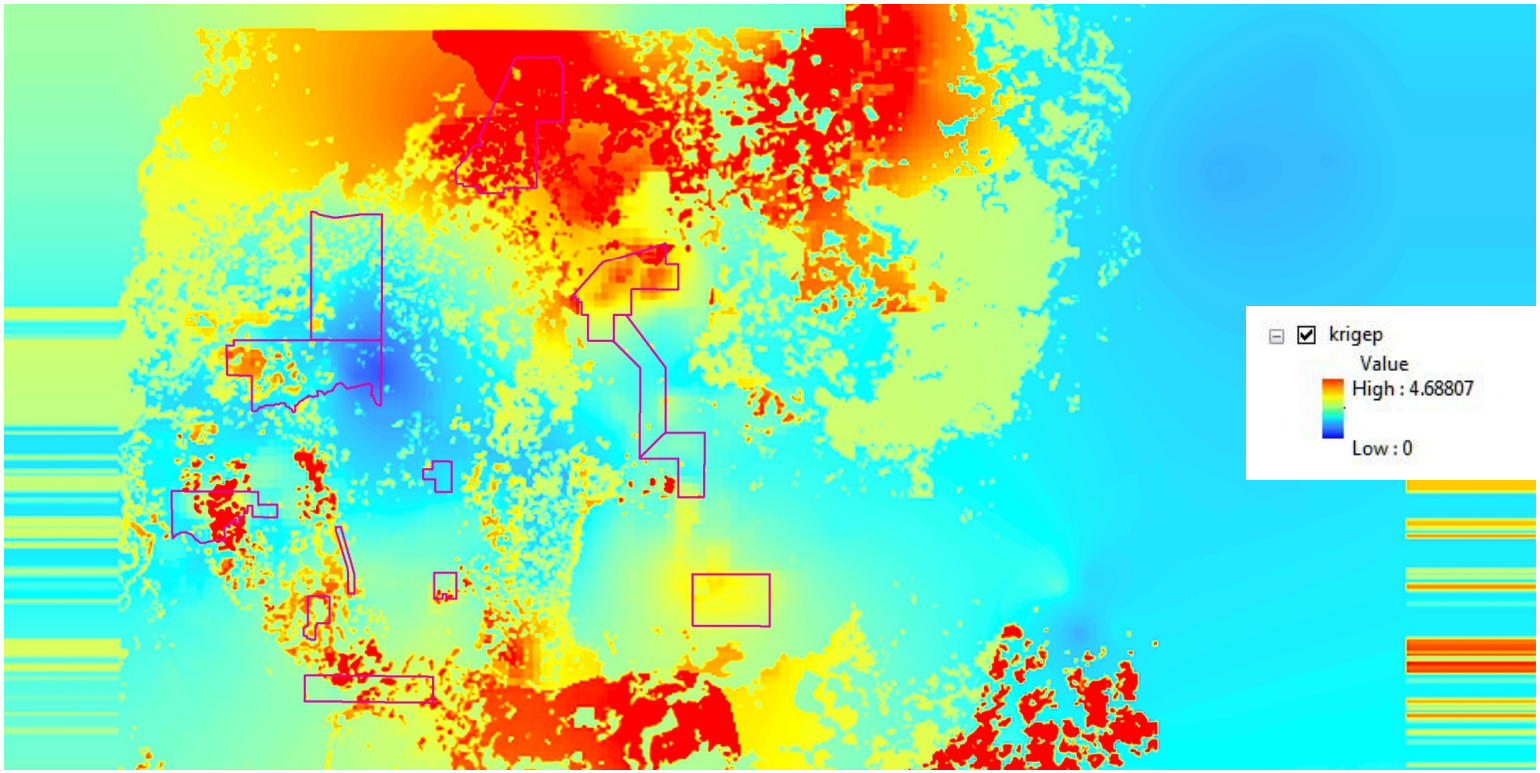
Regression kriging shows slightly tighter range and interquartile range of residuals.

Backtransformed residuals similar among top three models

- Similar distributions
- Narrower range in regression kriging
- LOOCV residuals provide estimate of performance on unmonitored sites (e.g., 62% of RK residuals within +/- 1 foot)

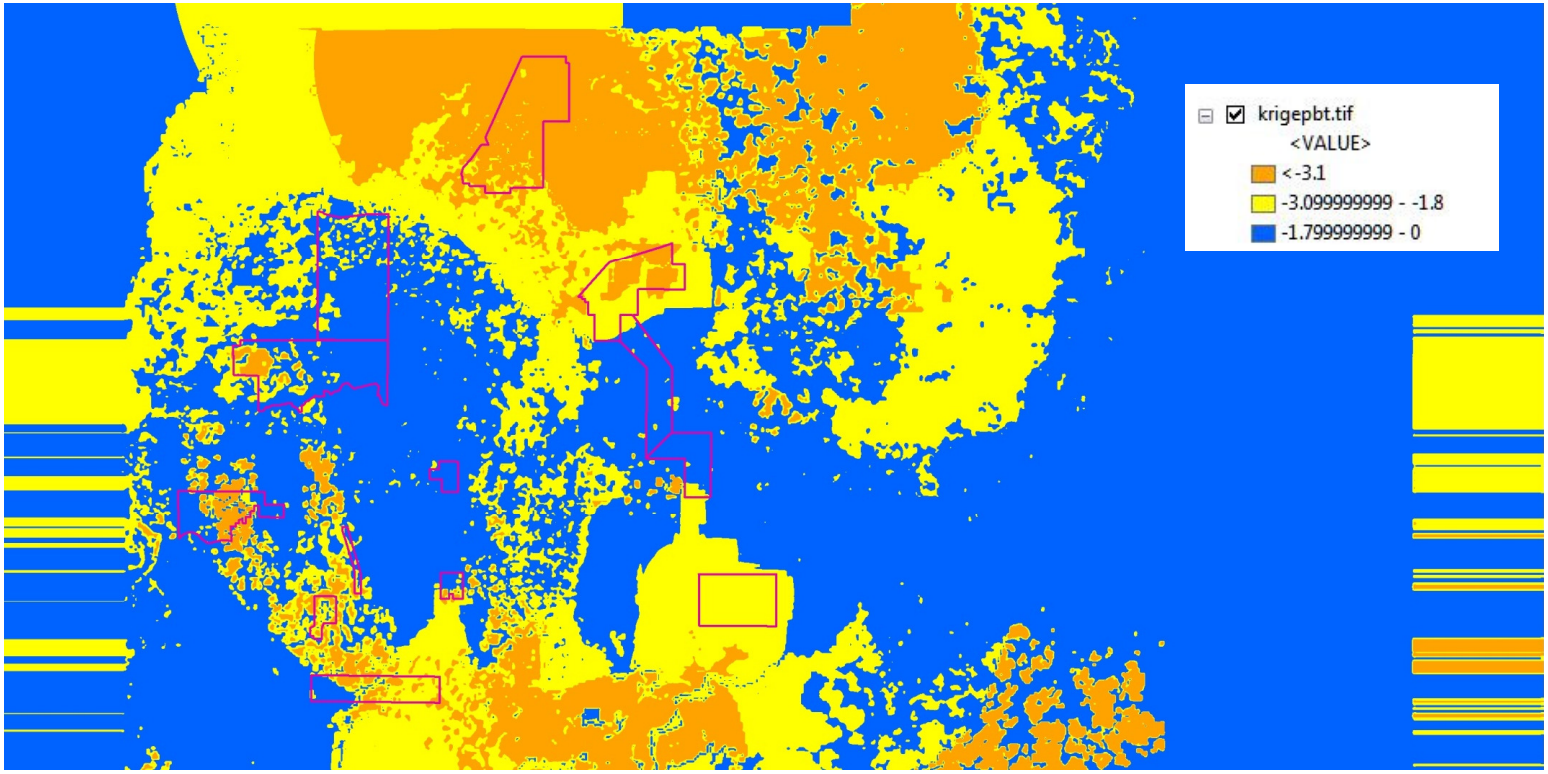


Regression Kriging Predictions (in transformed units)



GPI

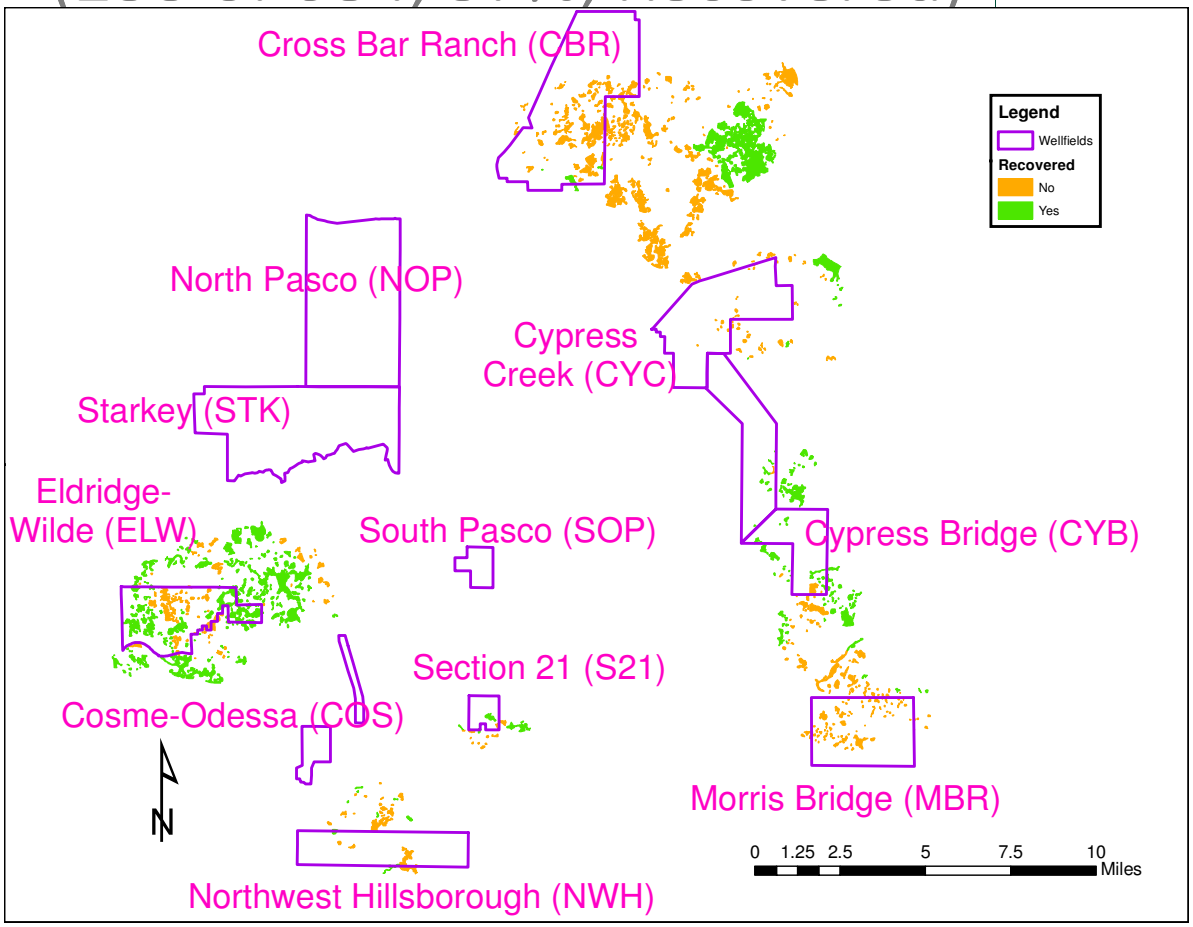
Regression Kriging Water Level Predictions Back-transformed to Feet



i.e., blue would meet mesic and xeric, yellow only meets xeric, and orange meets neither

GPI

Regression Kriging Interpolation to the Unmonitored Sites (253 of 684, 37%, Recovered)



Questions? Comments?

Please feel free to contact me by email or phone with questions and comments.

dschmutz@gpinet.com
813-765-0874