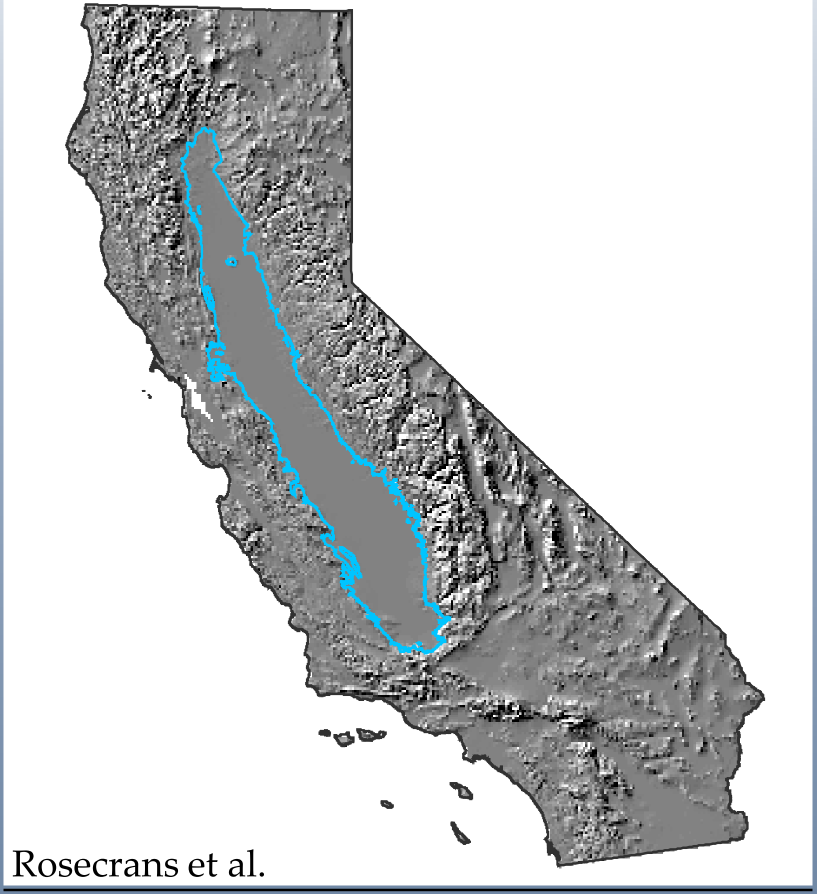
A Hybrid Boosted Regression Tree Model to Predict and Visualize Nitrate Concentration Throughout the Central Valley Aquifer

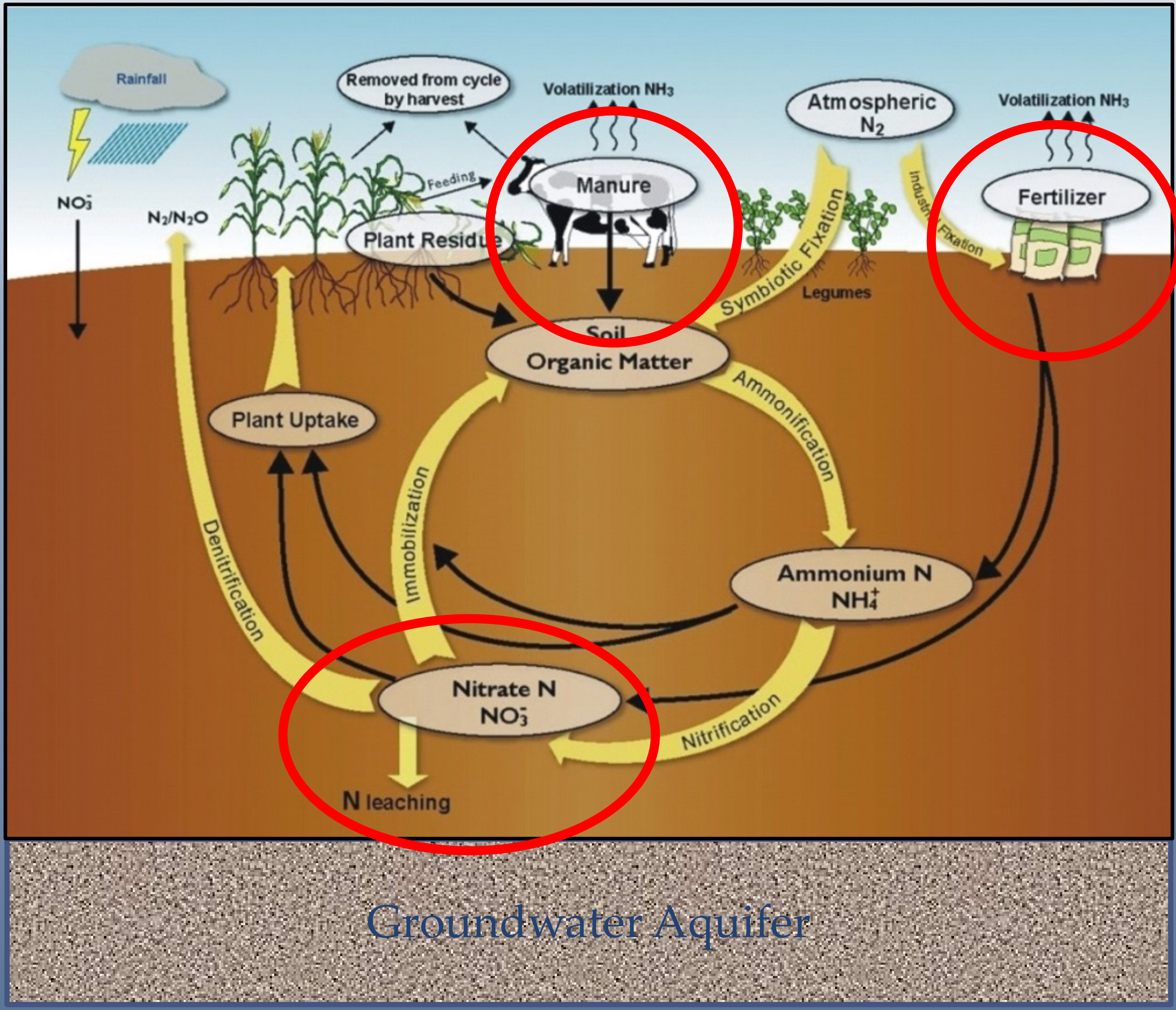
Katherine M. Ransom, Bernard T. Nolan, Jon Traum, Claudia C. Faunt, Andrew M. Bell, Jo Ann M. Gronberg, David C. Wheeler, Celia Rosecrans, Bryant Jurgens, Gregory E. Schwarz, Kenneth Belitz, Sandra Eberts, George Kourakos, and Thomas Harter

Study Goals and Overview

- To map groundwater nitrate concentration "wall to wall and top to bottom"
- Gain understanding of the system
- Groundwater age, field scale nitrogen input, oxidation/reduction potential
- Boosted Regression Trees



Nitrate in Groundwater - Sources

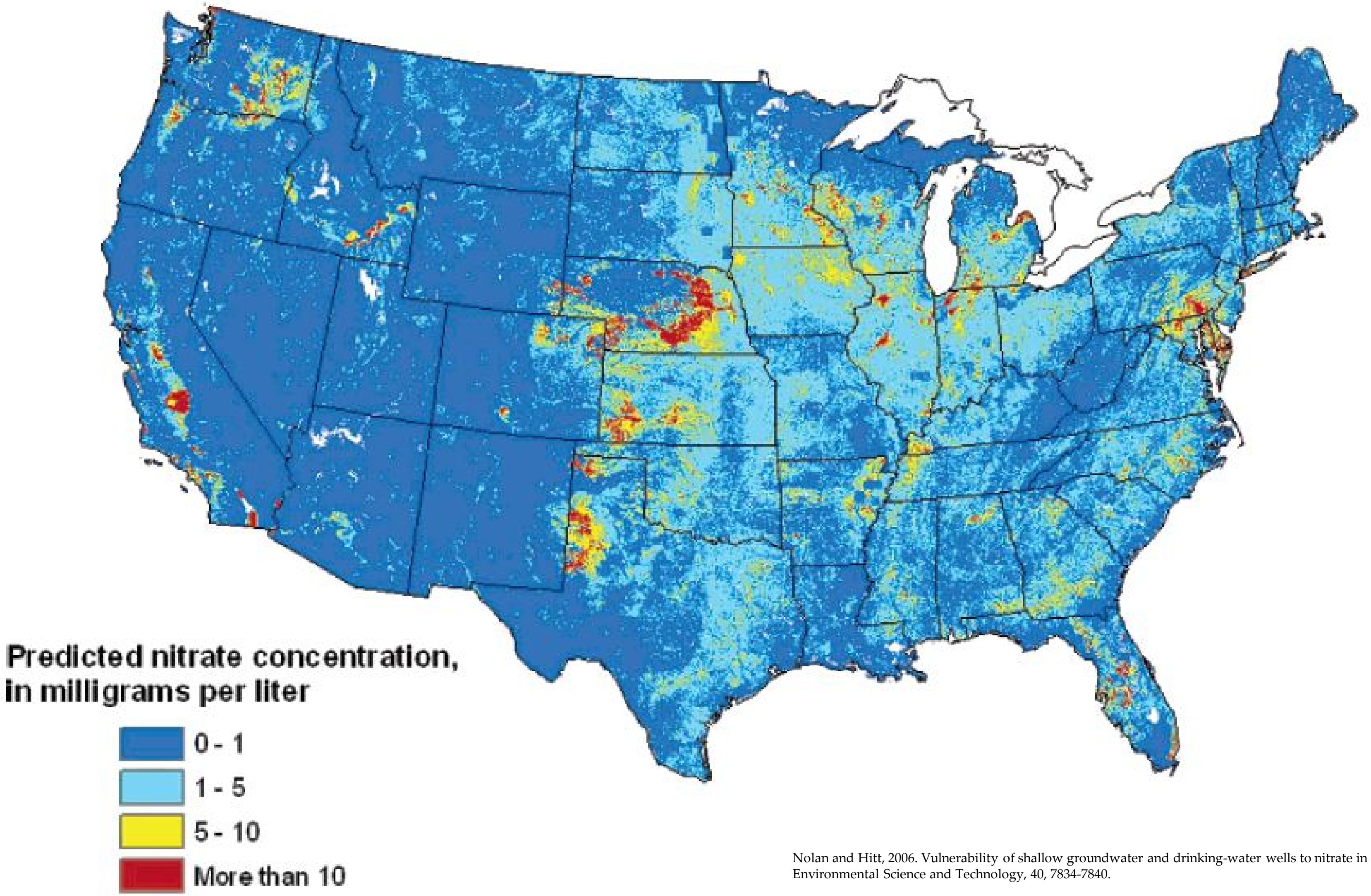


Domestic wastewater is a potential source in rural and urban areas from septic tanks or leaky sewer lines (Bremer and Harter, 2012, and Viers et al., 2012).

Natural sources (organic matter decay) contributes a minimal amount.

*Nitrogen Cycle image: Modified from University of Wisconsin Integrated Pest and Crop Management, shown on http://fyi.uwex.edu/discoveryfarms/page/6/.

Nitrate in Groundwater - US



Nolan and Hitt, 2006. Vulnerability of shallow groundwater and drinking-water wells to nitrate in the United States, Environmental Science and Technology, 40, 7834-7840.

Nitrate in Groundwater – Models

Authors	Scale	Method(s)
Nolan, Hitt, and Ruddy, 2002	National	Logistic Regression
Nolan and Hitt, 2006	National	Non-linear Regression
Nolan et al., 2014	Central Valley	Logistic Regression, Random Forest
Nolan, Fienen, and Lorenz, 2015	Central Valley	Boosted Regression Trees, Bayesian Networks, Artificial Neural Networks
Ransom et al., 2017	**Central Valley**	**Boosted Regression Trees**

Nolan, Hitt, and Ruddy, 2002. Probability of Nitrate Contamination of Recently Recharged Groundwaters in the Conterminous United States, Environmental Science and Technology, 36 (10), 2138-2145.

Nolan and Hitt, 2006. Vulnerability of Shallow Groundwater and Drinking-Water Wells to Nitrate in the United States, Environmental Science and Technology, 40 (24), 7834-7840.

Nolan et al., 2014. Modeling Nitrate at Domestic and Public-Supply Well Depths in the Central Valley, California, Environmental Science and Technology, 48 (10), 5643-5651.

Nolan et al., 2015. A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA, Journal of Hydrology, 531, 902-911.

Ransom et al., 2017. A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA, Science of the Total Environment, 601-602, 1160-1172.

Building on Previous Work

Hybrid Approach

- Oxidation/reduction potential
- Groundwater age
- Nitrogen loading – field scale

3D map

- Predictions mapped at depth
- Interpolation between predictions

Machine Learning for Nitrate

Pros

- Relations need not be linear or follow a particular data distribution
- Screens large numbers of variables
- Handles missing data
- Results not affected by collinearity
- Automatically incorporates interactions and thresholds
- Useful for inference

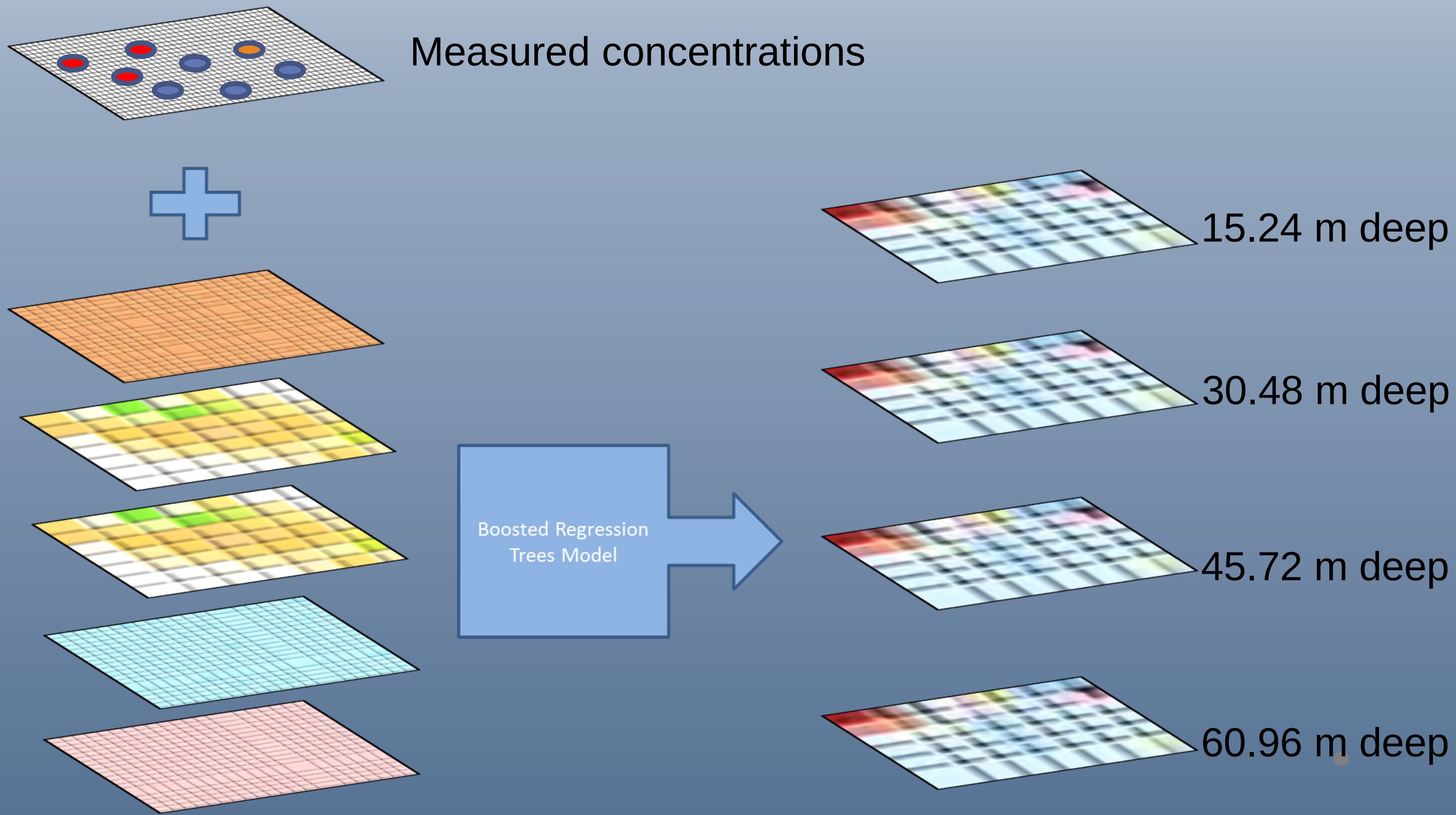
Cons

- Overfitting the data
- Model is harder to interpret
- Perceived as “black box”

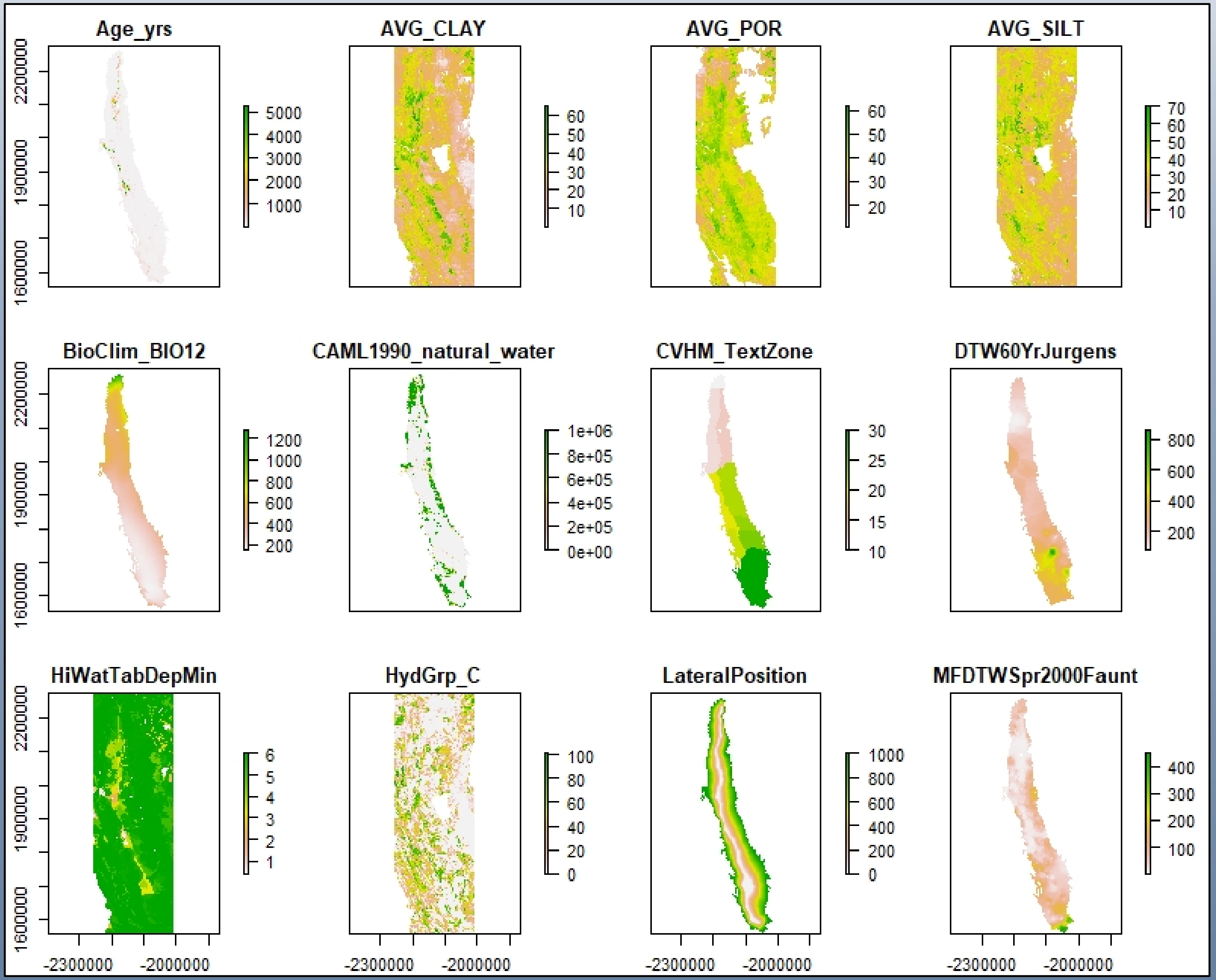
Modified from: B.T. Nolan, 2017

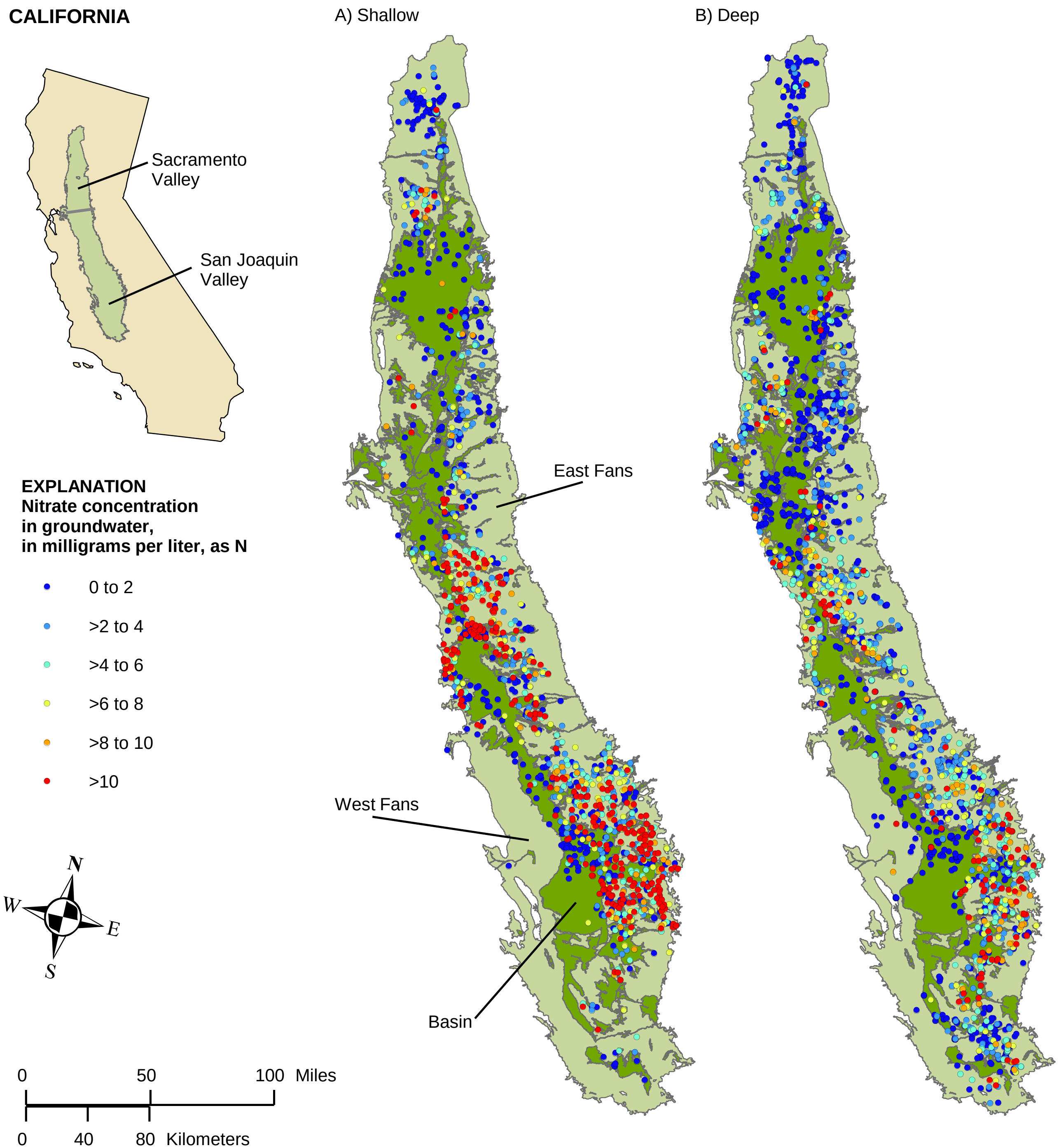
Statistical Methods - Workflow

- Predictor variables attributed to wells, 145 total
- Boosted regression tree modeling
- Predictors ranked based on importance (variable reduction routine)
- Top 25 variables kept for final
- Predictions made at 17 depths, 3D map created



Well Data and Predictor Variables





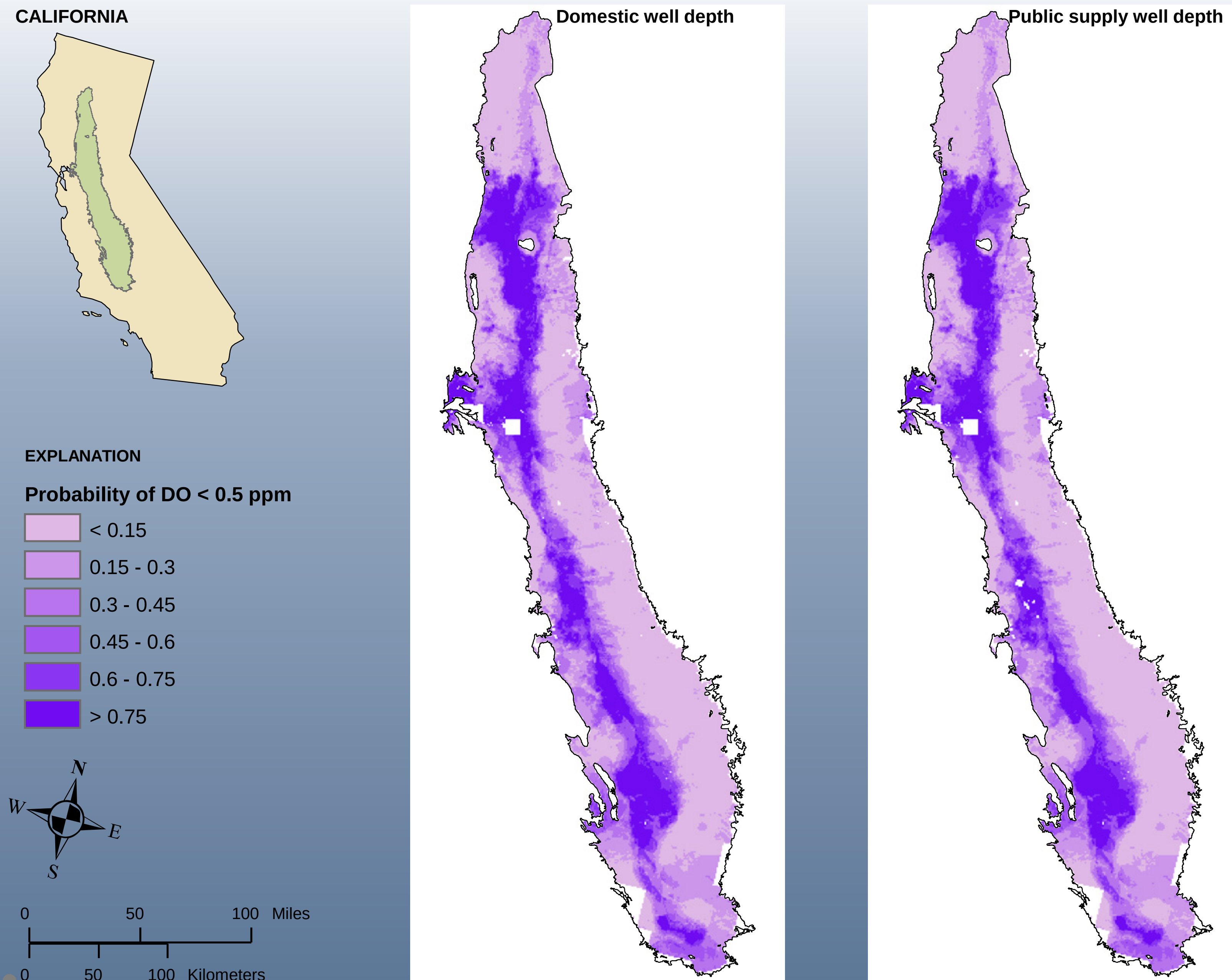
3508 Training wells (shown)

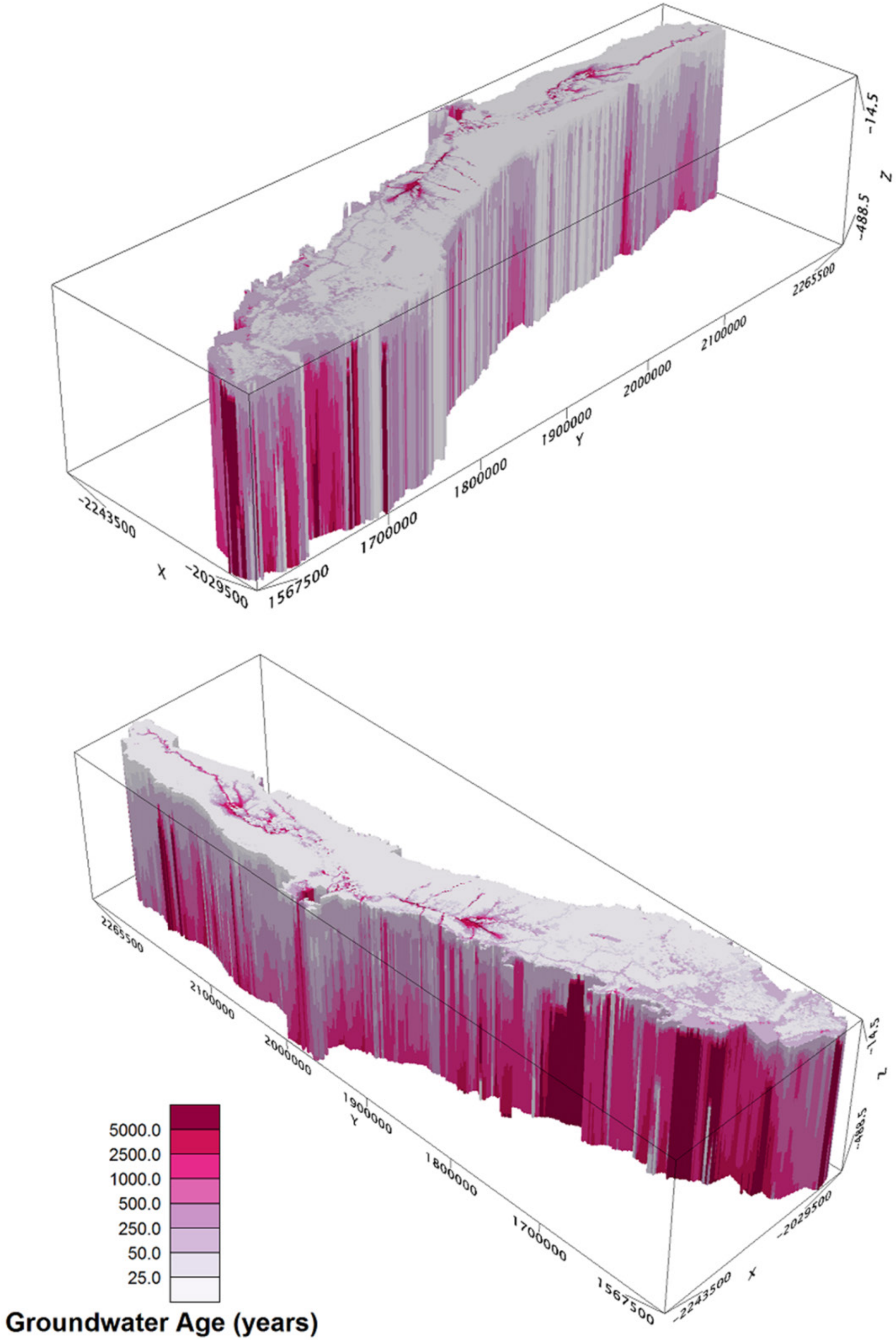
Shallow:
1400 wells
Domestic wells
180 ft/54.9 m
27% exceedance

Deep:
2108 wells
Public wells
400 ft/121.9 m
6% exceedance

1662 “Hold-out” wells (not shown)

Probability of Anoxic Condition





MODFLOW/MODPATH Estimates of Groundwater Age with Depth

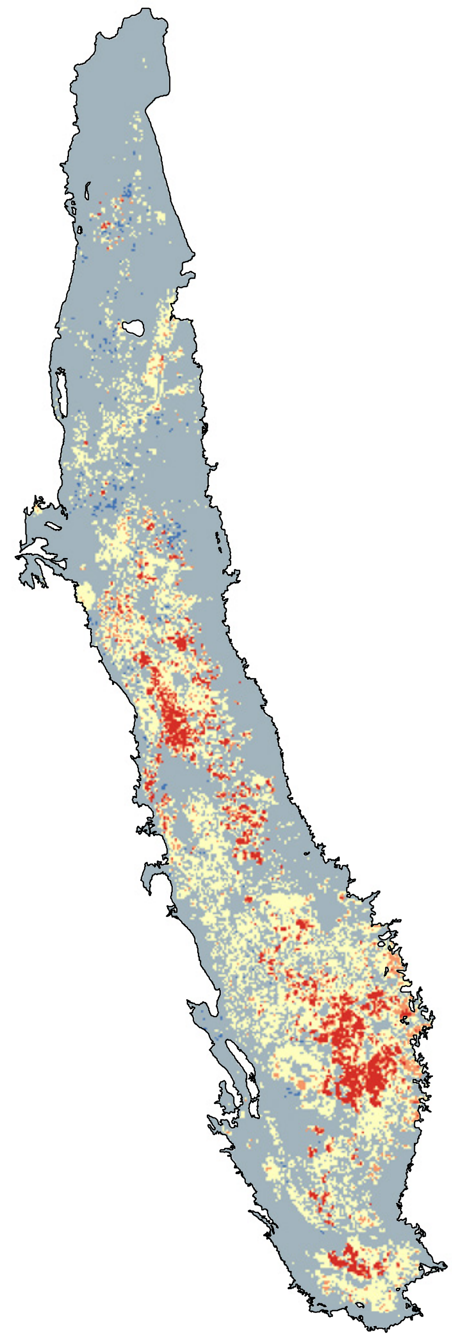
- Key component not included in previous models.
- “Proxies” such as well depth or depth to water.

Estimates from: Central Valley Hydrologic Model, Faunt, C. C. (2009). *Groundwater availability of the Central Valley Aquifer, California.* Professional Paper 1766, U.S. Geological Survey.

Field-Scale Nitrogen Leaching Flux - 1975

Based on nearly 200 land use types, including 60 crop types.

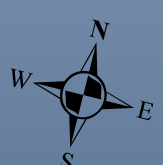
Available for 1945, 1960, 1975, 1990, and 2005.



EXPLANATION

Unsaturated zone nitrogen leaching flux to groundwater, 1975

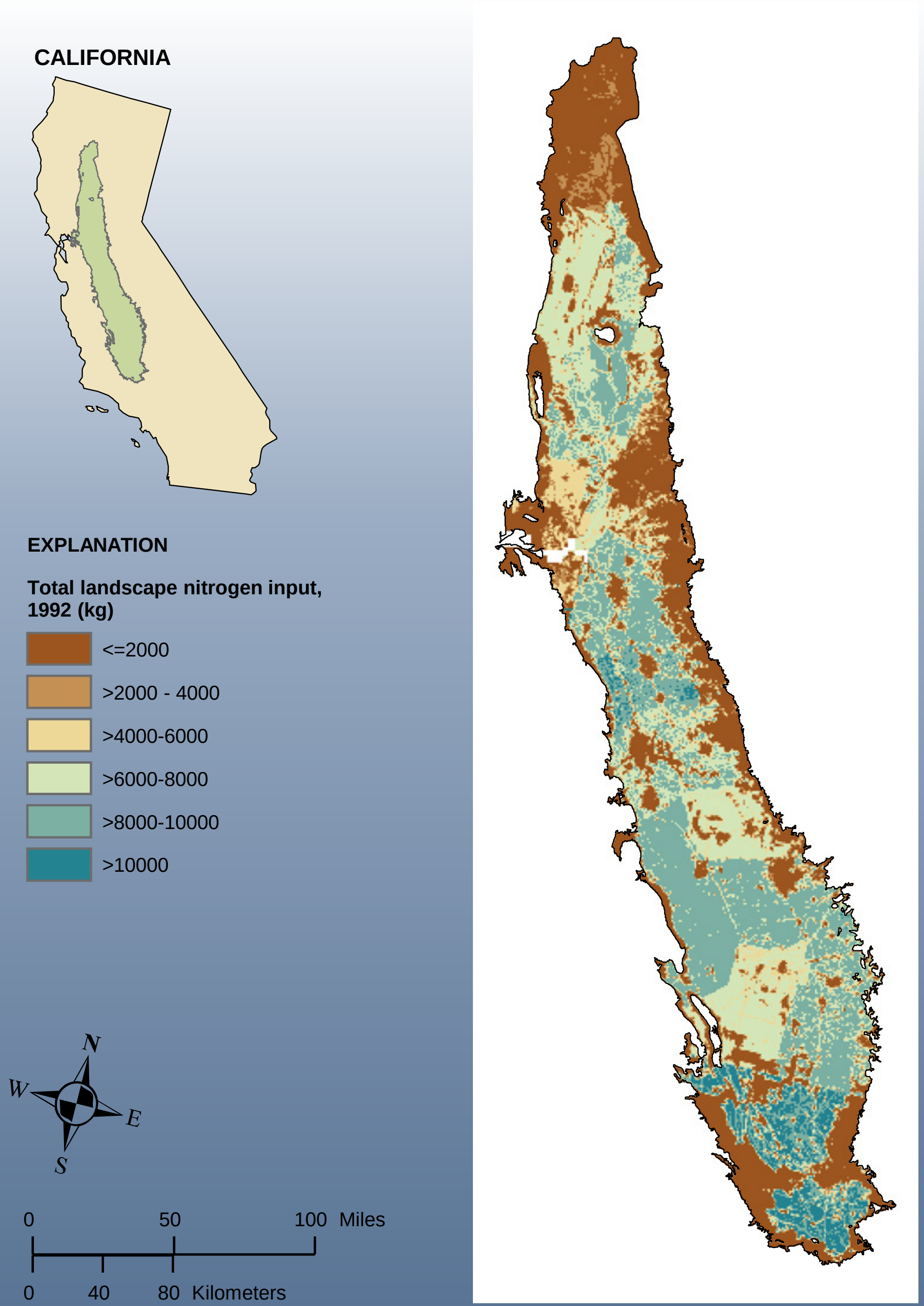
- < 4
- 4 - 6
- 6 - 8
- 8 - 10
- > 10



0 50 100 Miles

0 40 80 Kilometers

County-Scale Nitrogen Input



Statistical Methods - Software

Variable Processing

ArcGIS
ESRI

Modeling and Prediction

R

RStudio

Packages

- caret
- gbm
- raster
- sensitivity
- boot

3D Visualization

GEOSOFT
Oasis montaj

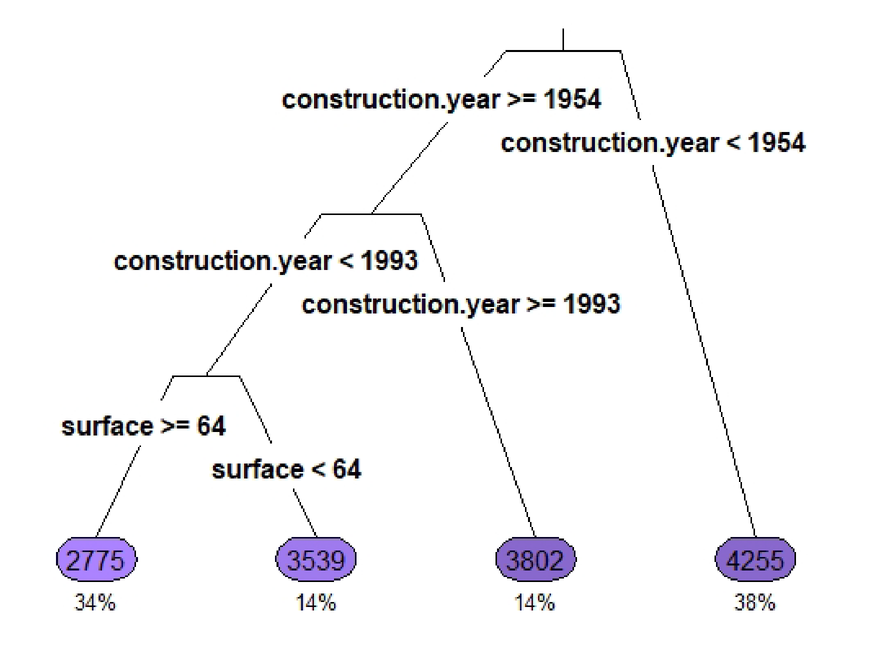
Statistical Methods - Boosted Regression Trees

- aka Gradient Boosting Machine
- An ensemble method: collection of many small models (boosting)
- Based on classification trees
- Each new tree built on the residuals of the previous tree (gradient)
- Randomness added by subsampling data
- Trees controlled by tuning aka metaparameters

Example Apartments Dataset

	m2.price	construction.year	surface	floor	no.rooms
1	5897	1953	25	3	1
2	1818	1992	143	9	5
3	3643	1937	56	1	2
4	3517	1995	93	7	3
5	3013	1992	144	6	5
6	5795	1926	61	6	2
7	2983	1970	127	8	5
8	2346	1985	105	8	4
9	4745	1928	145	6	6
10	4284	1949	112	9	4

Simple Regression Tree

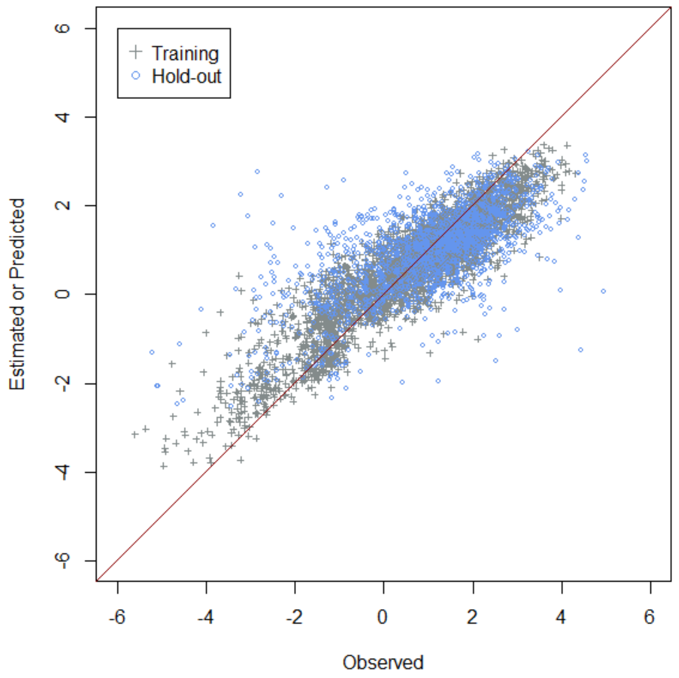


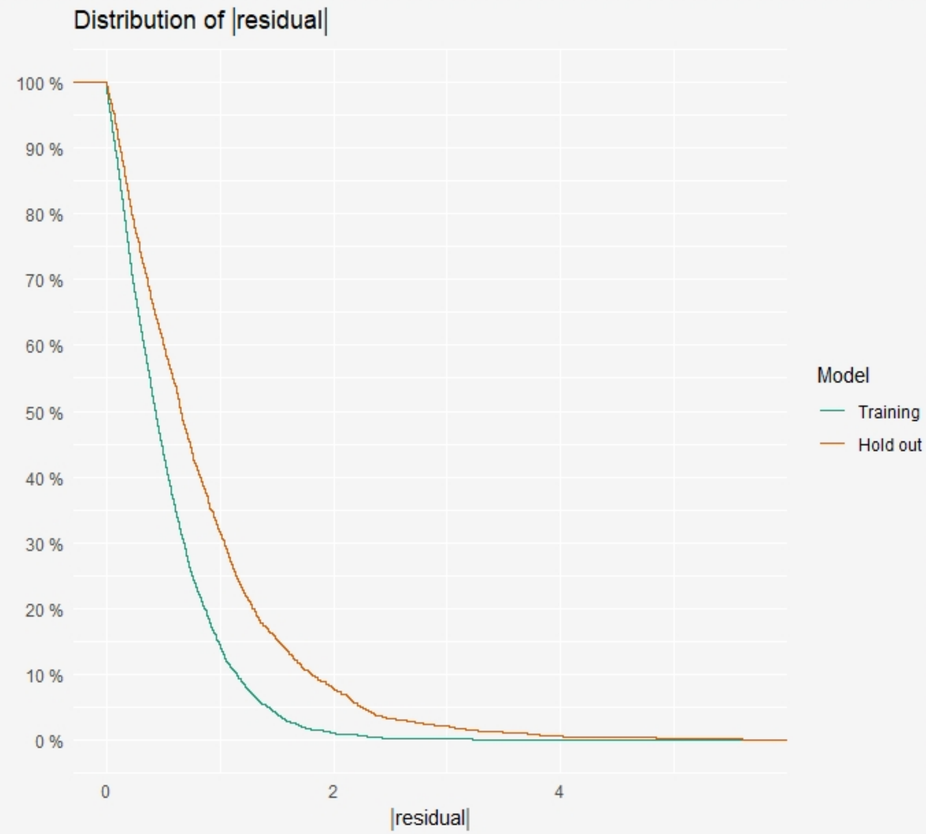
Results – Model Performance

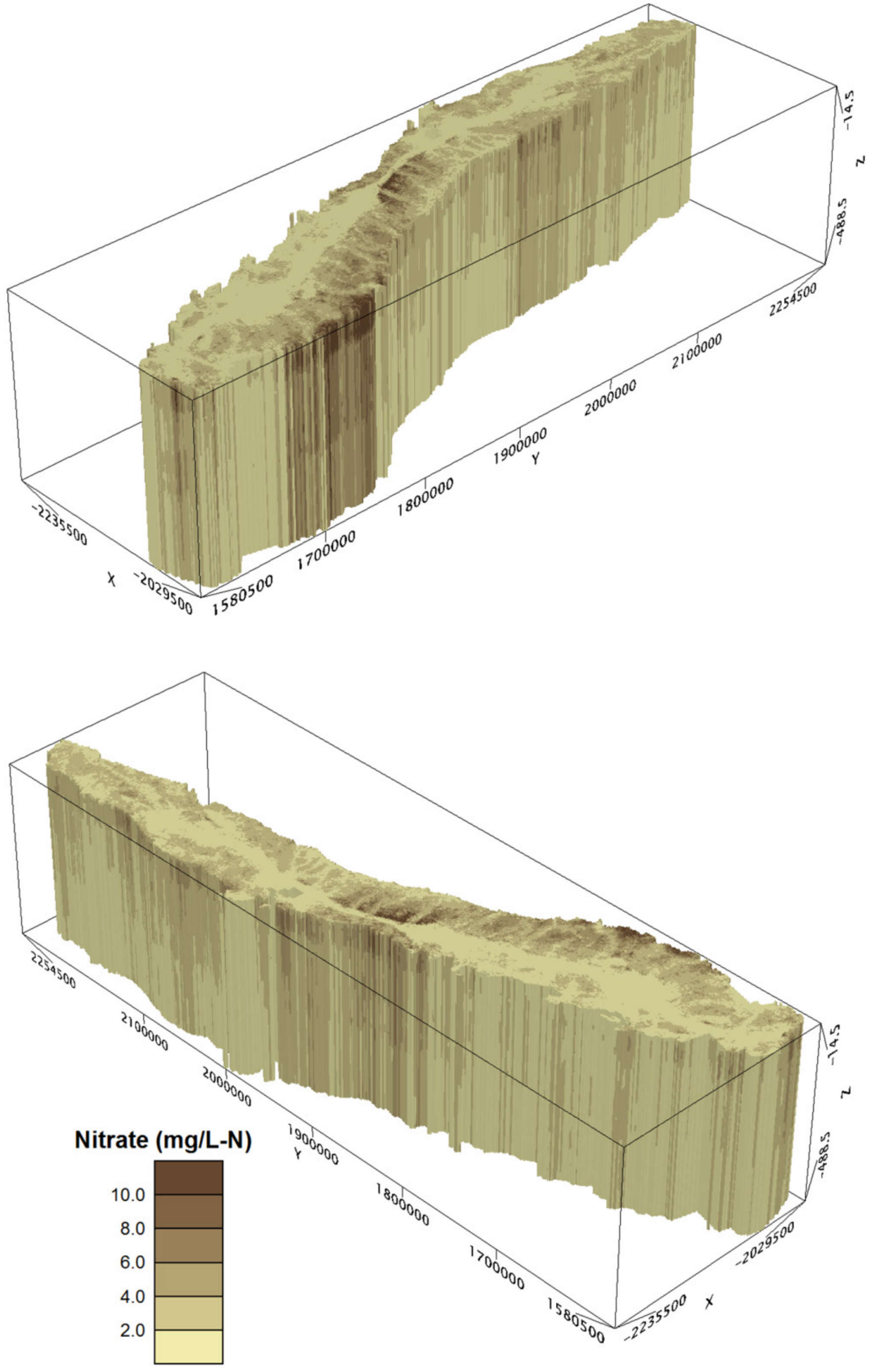
Training RMSE: 0.705
Training R^2: 0.825

Hold-out RMSE: 1.132
Hold-out R^2: 0.443

Residual Comparison



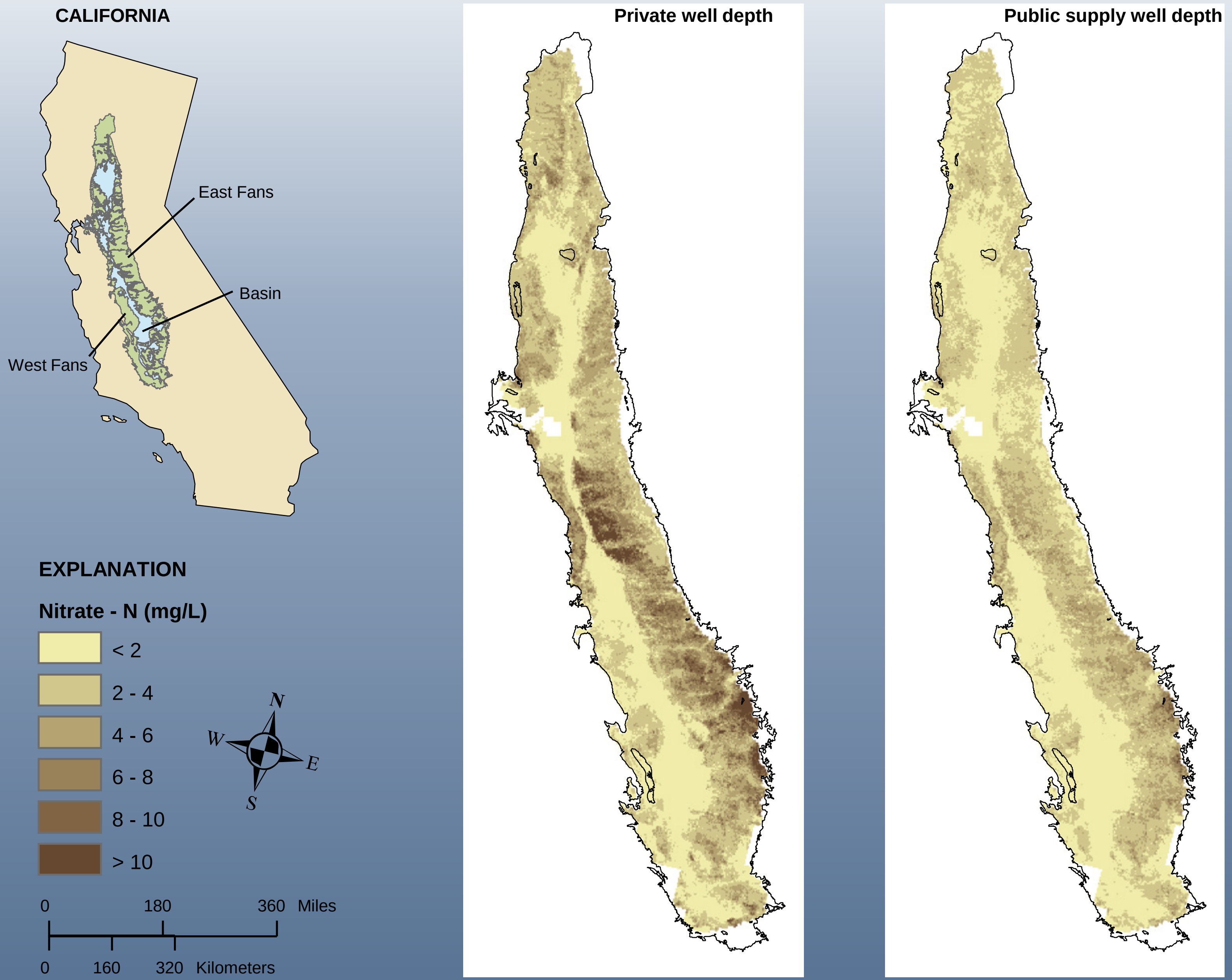




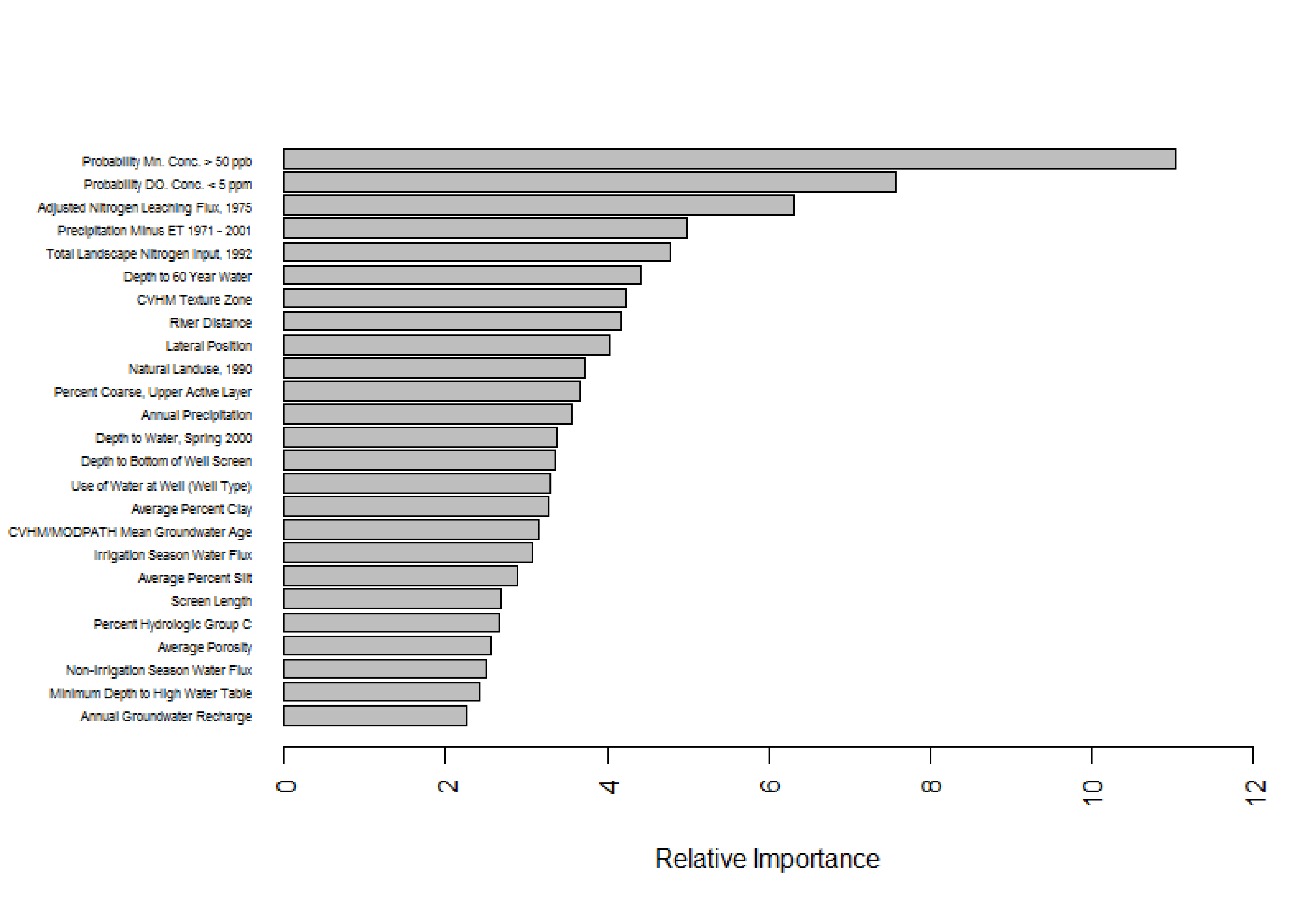
Results – Oasis Montaj 3D map

- To 1600 ft below ground surface
- 17 predicted layers
- Linear interpolation
- 1 m vertical resolution

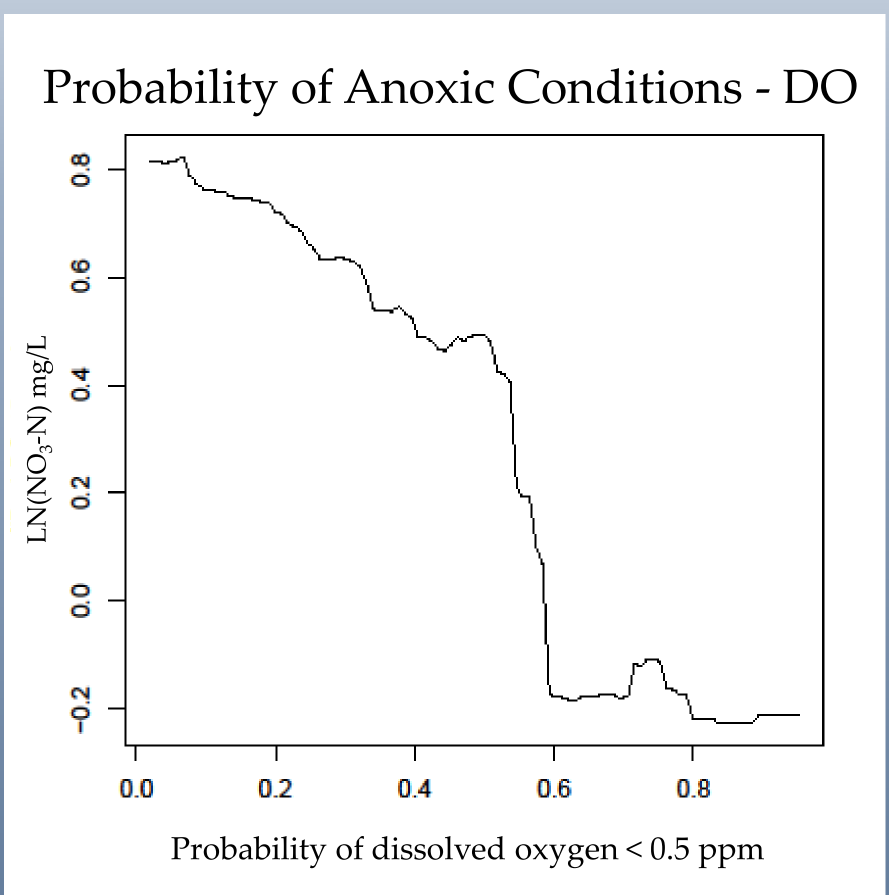
Results – Predictions at Specified Depths

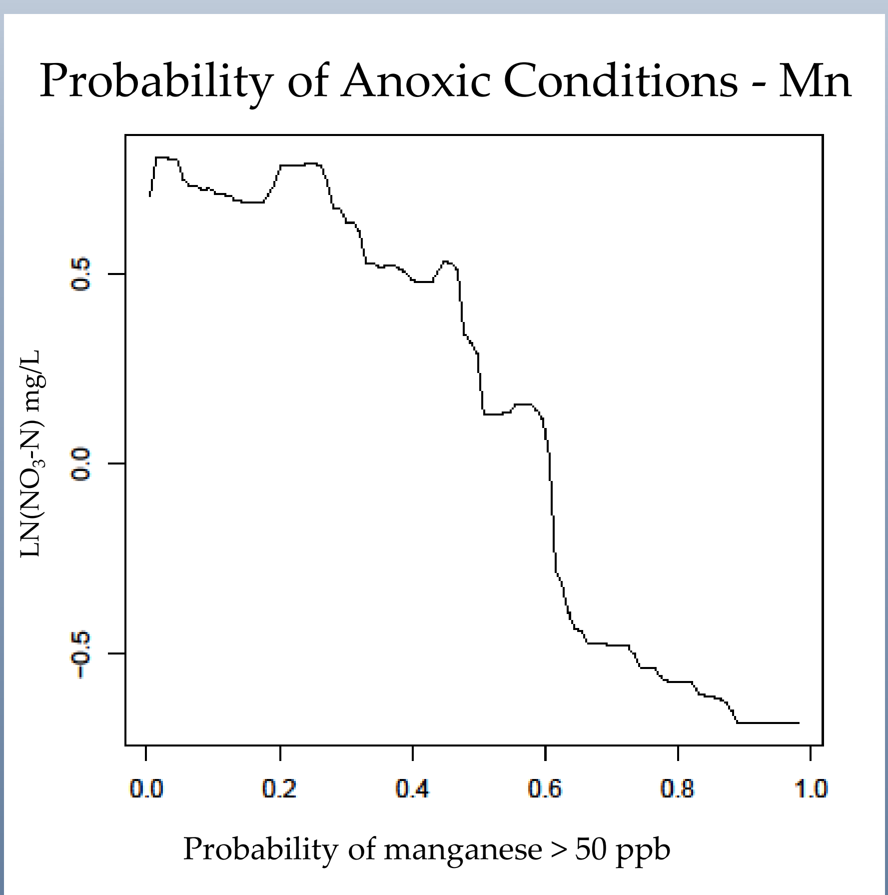


Secondary Results - Importance Ranking

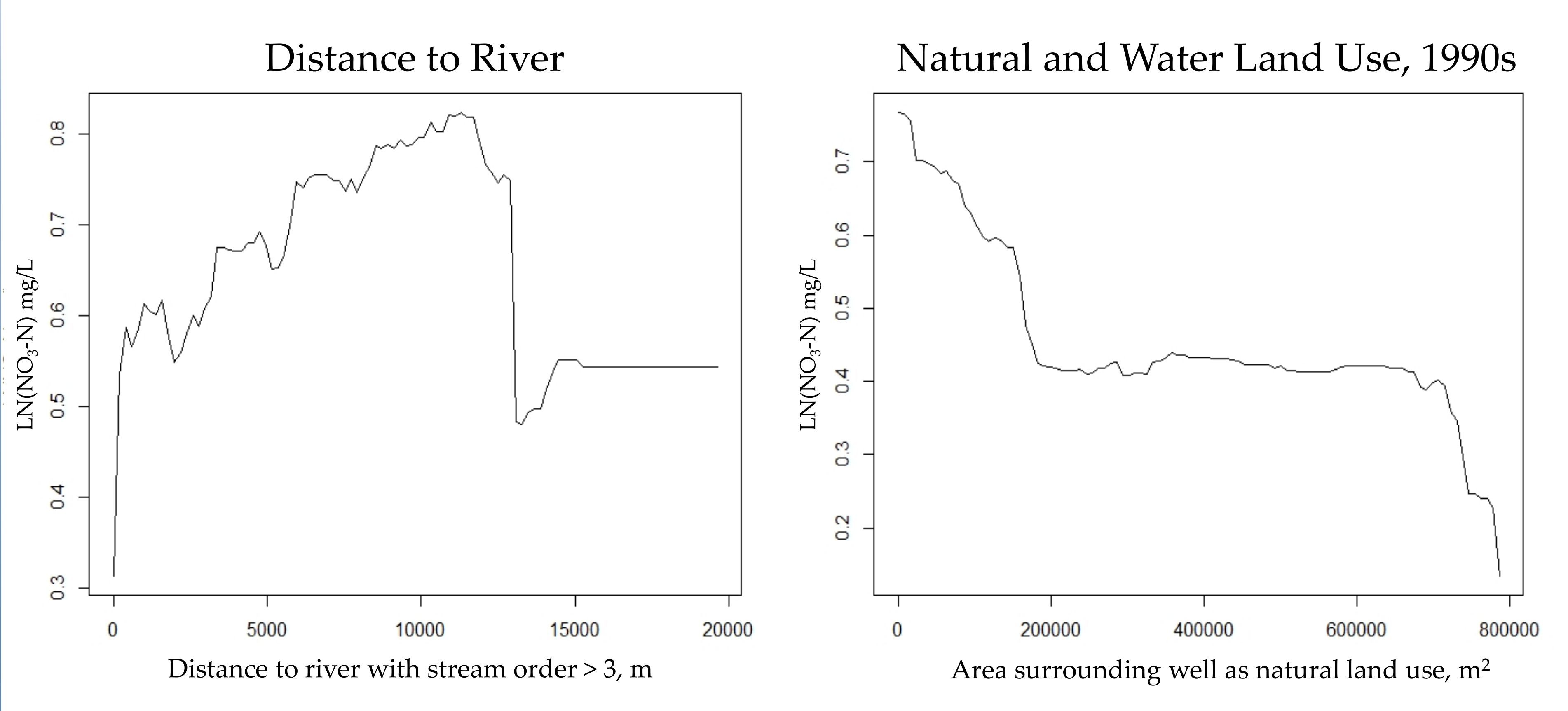


Secondary Results – Partial Dependency Plots

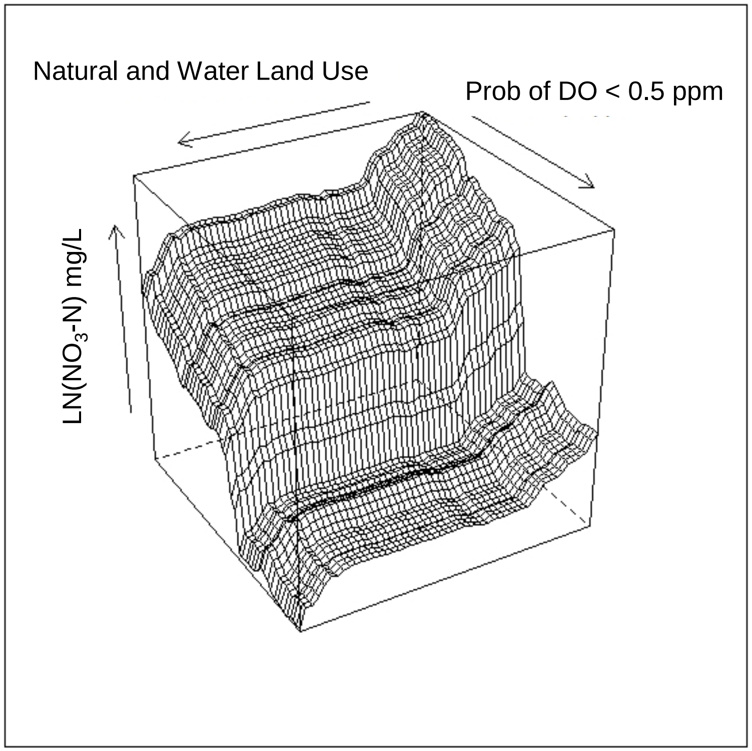


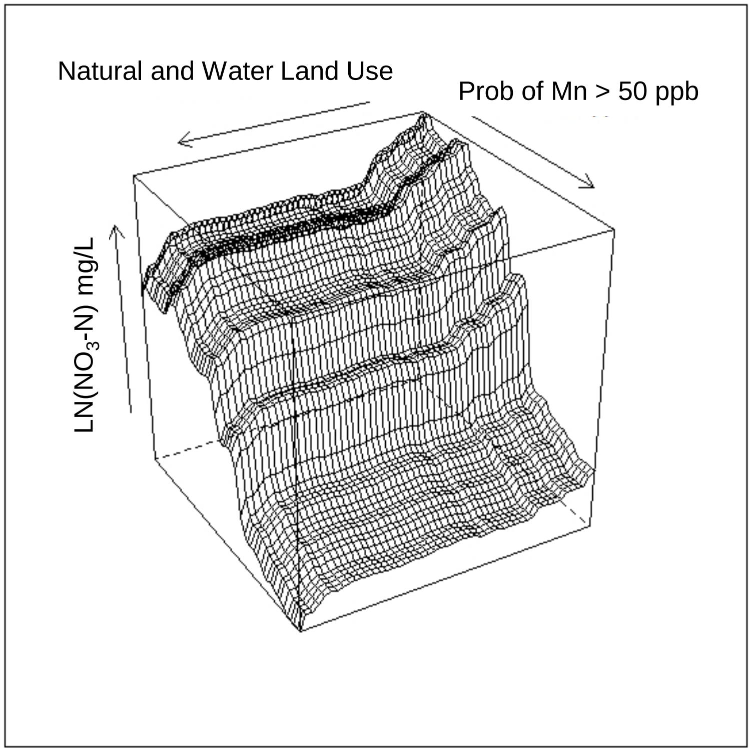


Secondary Results – Partial Dependency Plots



Secondary Results – Partial Dependency Plots



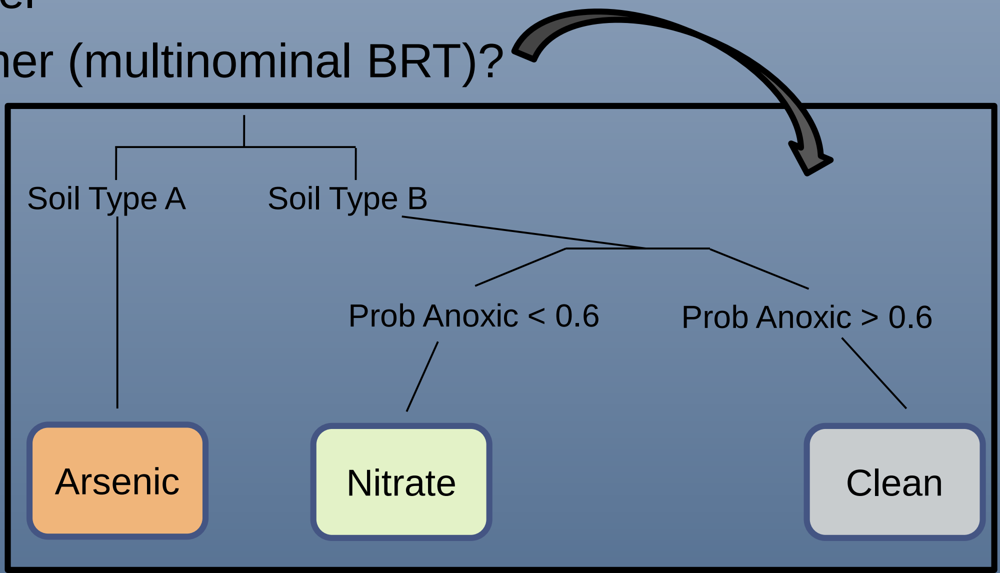


Summary and Conclusions

- Mapped nitrate tended to decrease with depth
- Alluvial fans region had higher nitrate concentrations than basin subregion
- Anoxic conditions highly related to nitrate concentration
- Patterns on partial plots make intuitive sense
- Coming soon: updated national nitrate and arsenic maps

Locating High Risk Domestic Wells

- Cookie cutter national models (updated or current) for full coverage
- Use estimates from current national arsenic model (Ayotte et al., 2017)
- Develop new California specific model
- Consider multiple constituents together (multinominal BRT)?
- Nitrate, arsenic, uranium, others?
- Overlay with well locations



Reference: Estimating the High-Arsenic Domestic-Well Population in the Conterminous United States, Ayotte et al., Environmental Science and Technology , 2017, 51 (21) pg. 12442 – 12454. https://pubs.acs.org/doi/10.1021/acs.est.7b02881

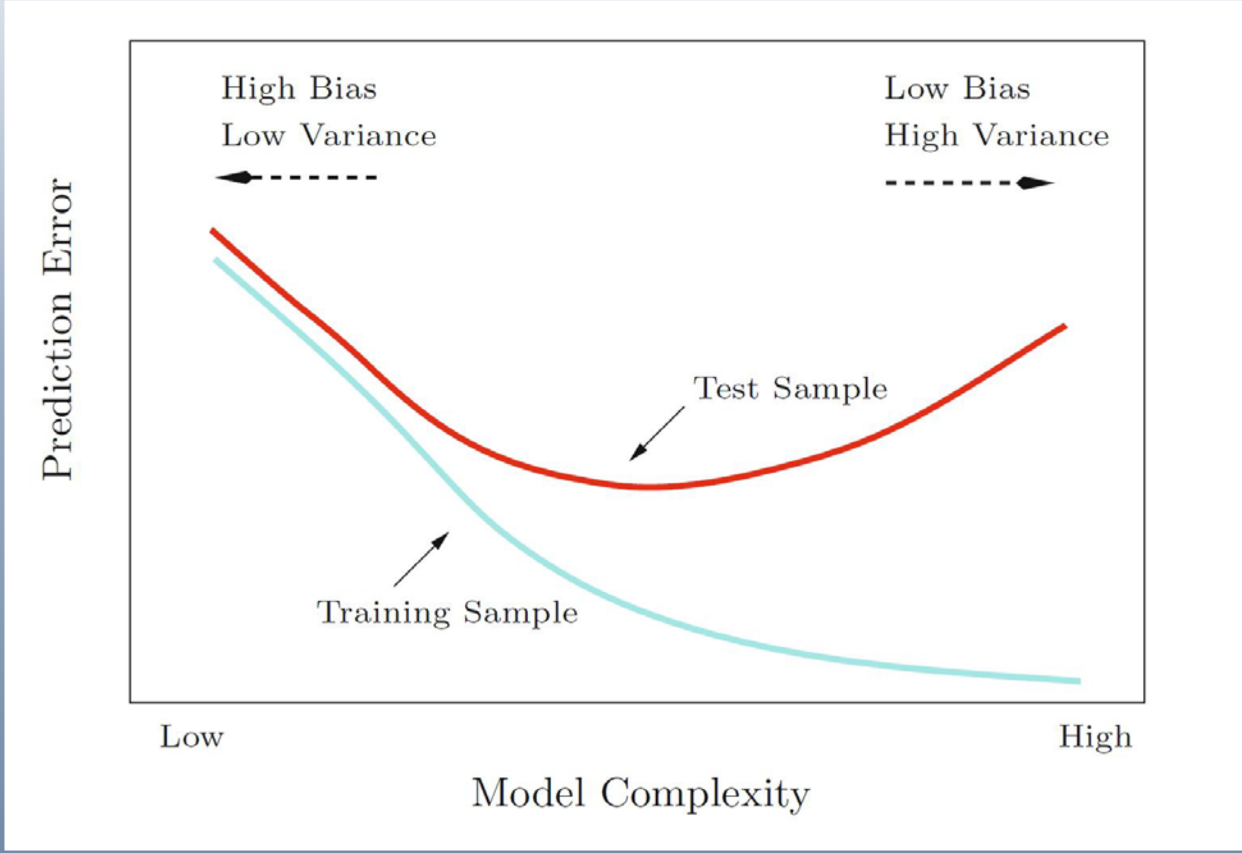
Questions?

Article available at:
https://www.sciencedirect.com/science/article/pii/S0048969717313013?via%3Dihub

Data raster grids available at:
https://www.sciencebase.gov/catalog/item/58c1d920e4b014cc3a3d3b63

Appendix

Statistical Methods – Cross Validation

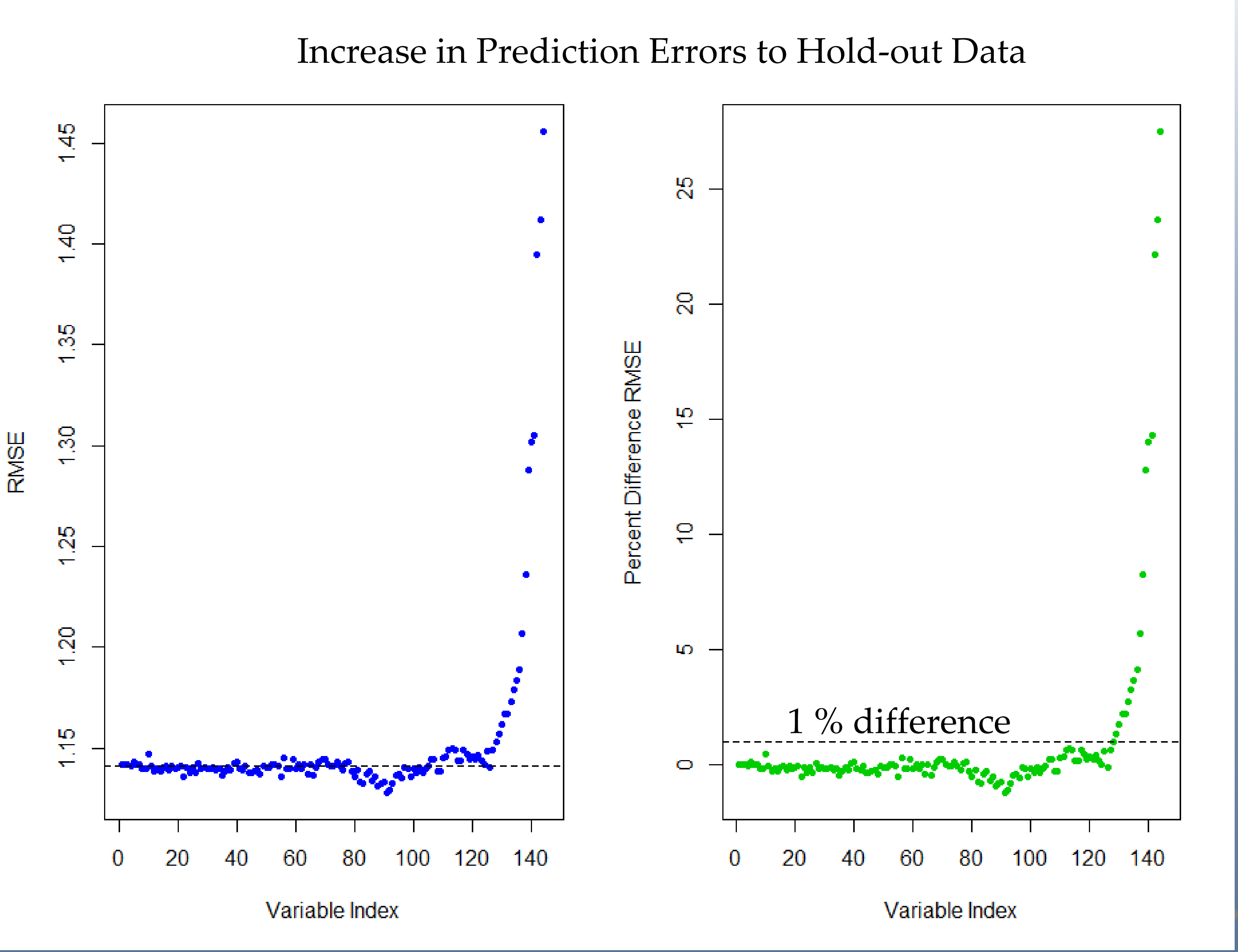


Metaparameters: interaction depth, shrinkage, number of trees, size of terminal nodes

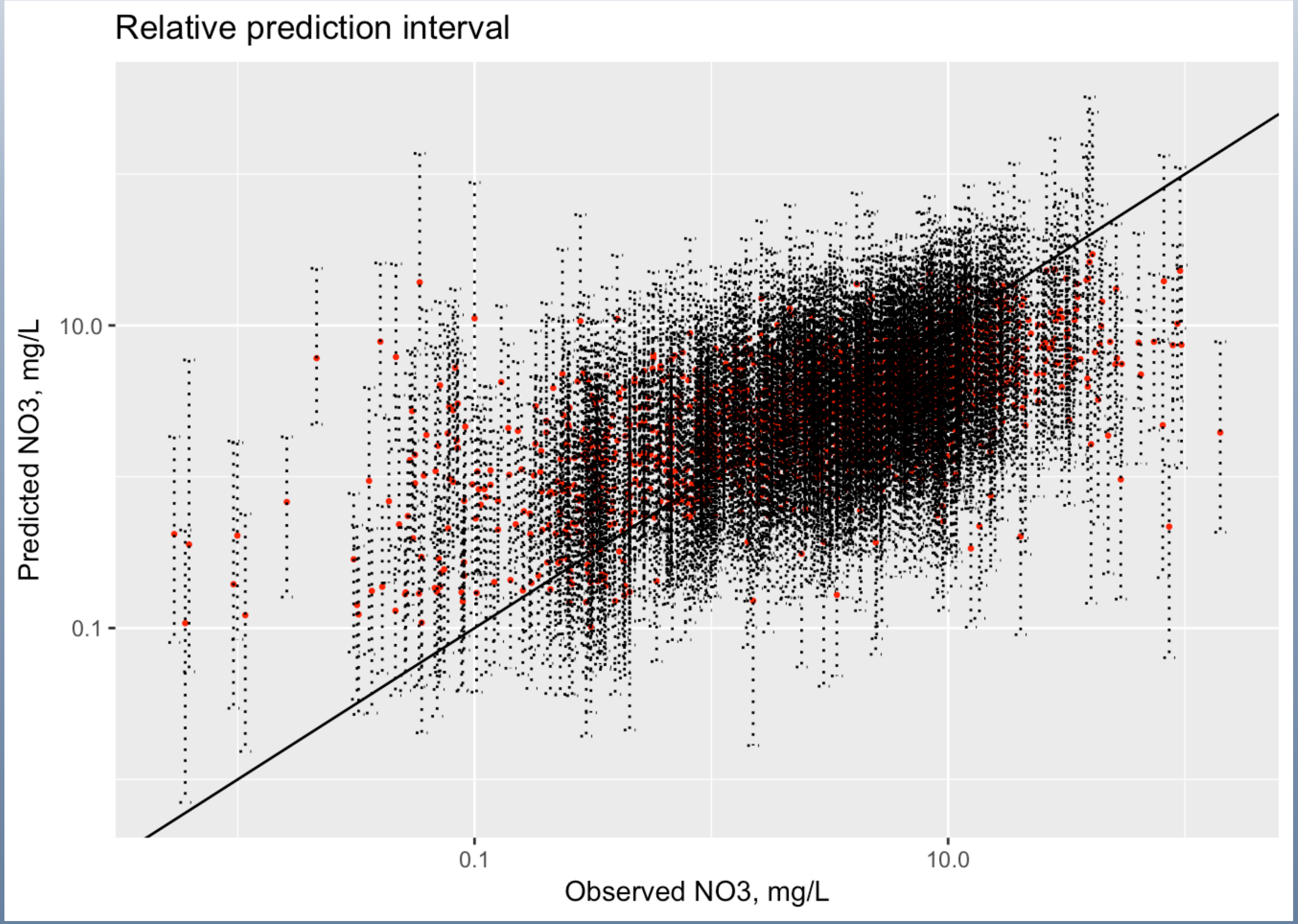
CV tuning addresses over fit by limiting model complexity

Credit: Hastie et al., 2009. The Elements of Statistical Learning.

Statistical Methods - Variable Reduction



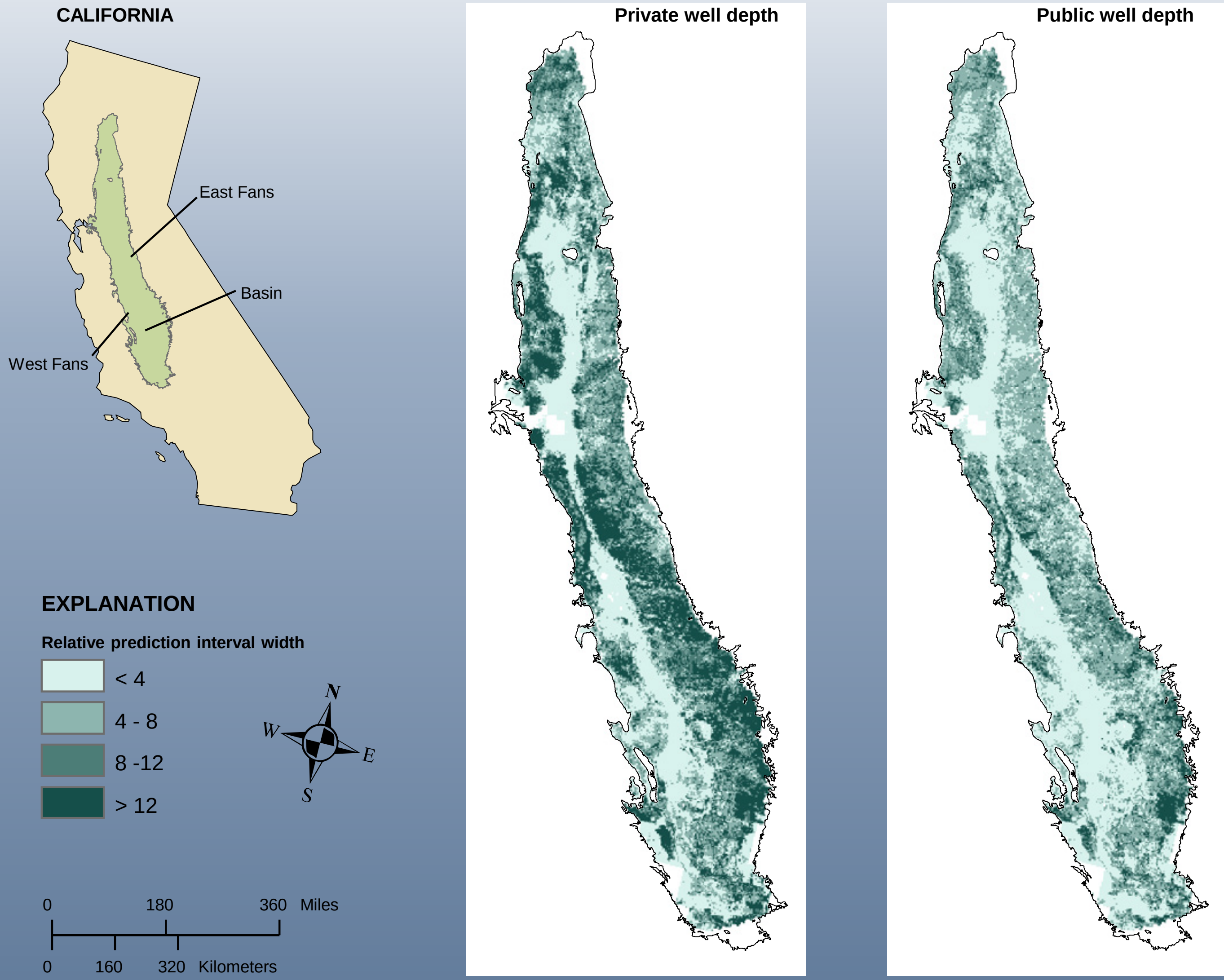
Results – Prediction Intervals



199 models made with bootstrapped sets of the training data

199 predictions made to hold-out data

Results – Prediction Interval Width



Results – Sobol Sensitivity Indices