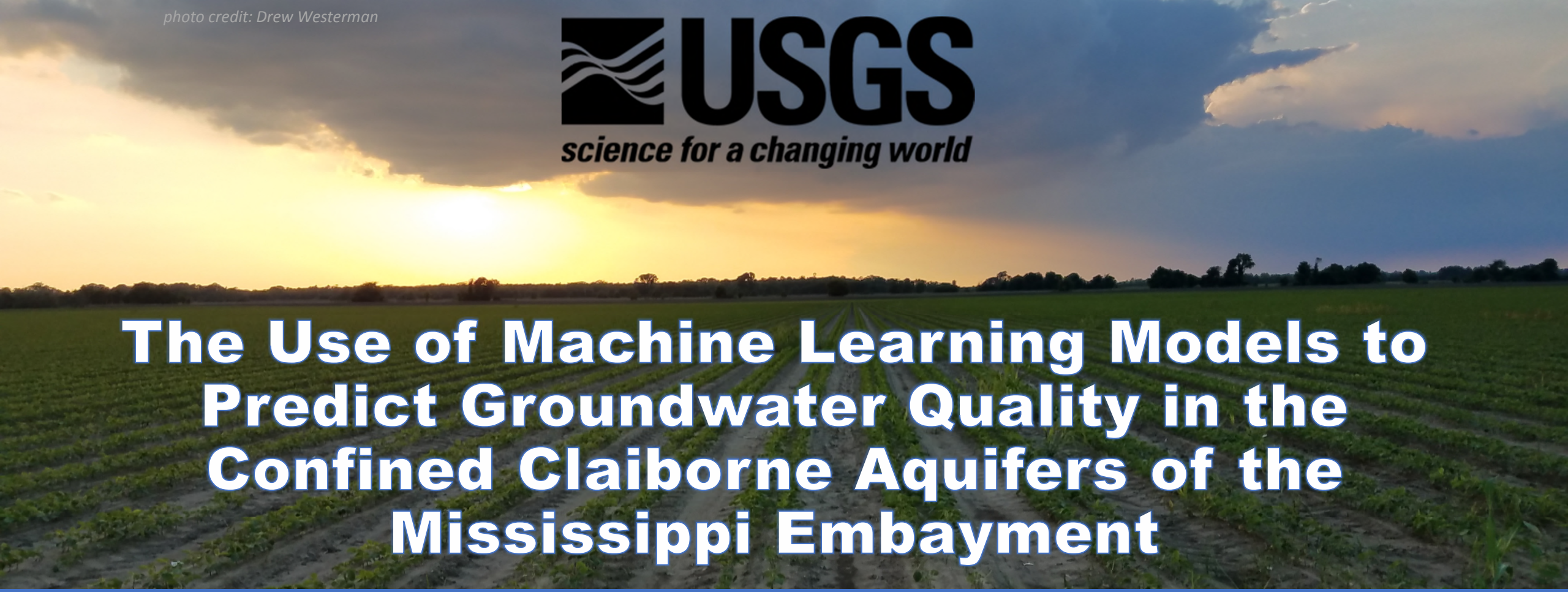


The Use of Machine Learning Models to Predict Groundwater Quality in the Confined Claiborne Aquifers of the Mississippi Embayment

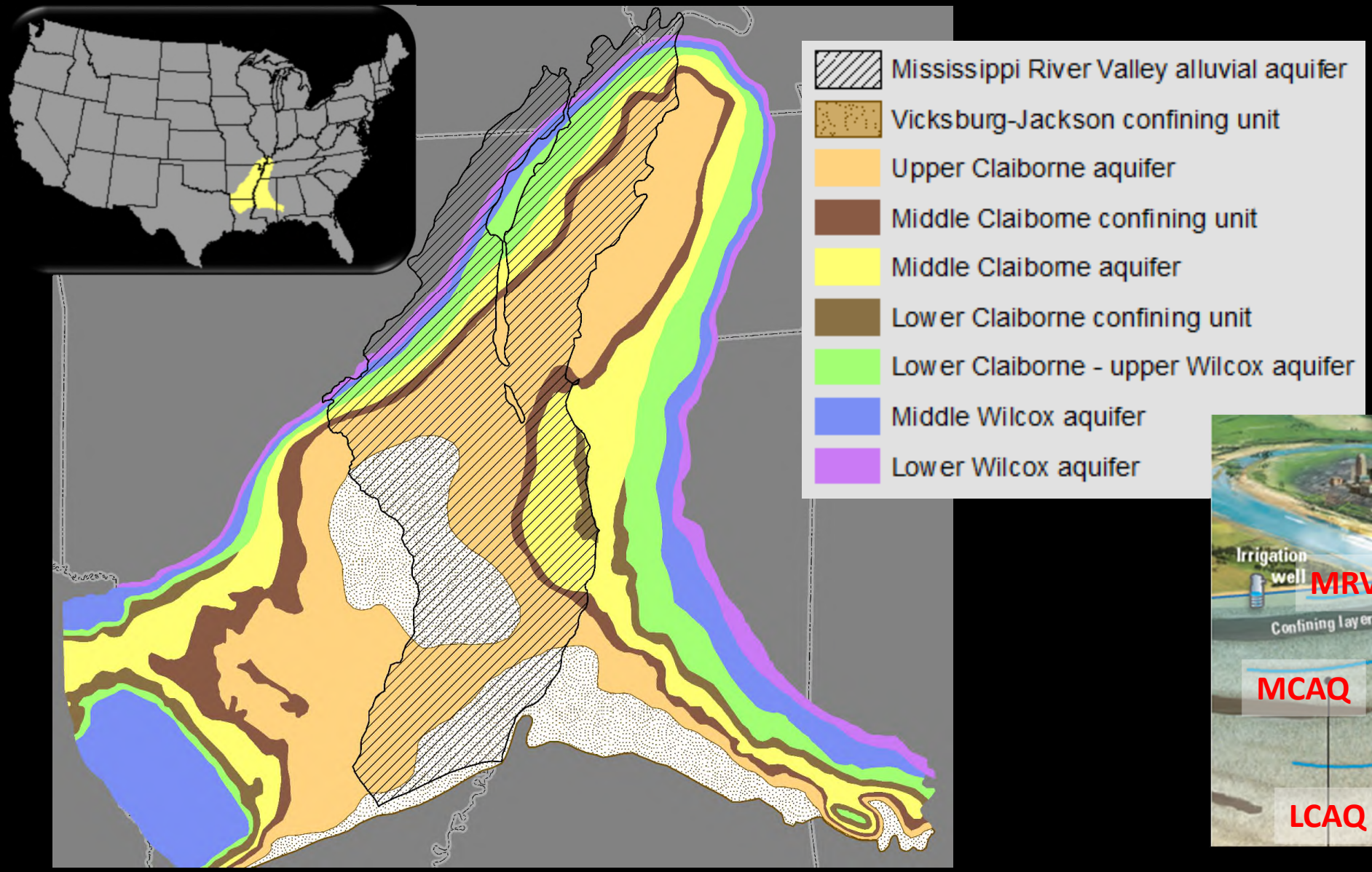
Katherine Knierim

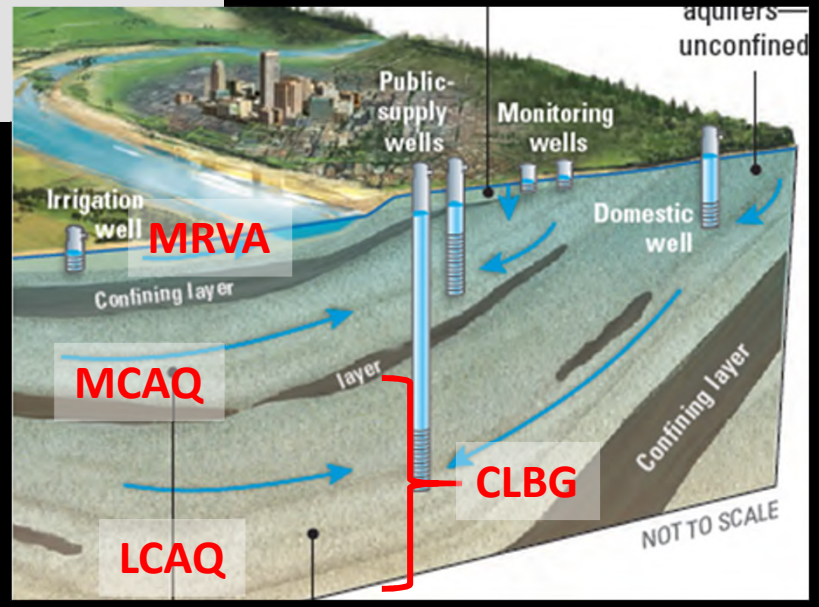
James Kingsbury, Connor Haugh

Lower Mississippi-Gulf Water Science Center

National Water Quality Assessment Project – Intensive Principal Aquifer Analysis

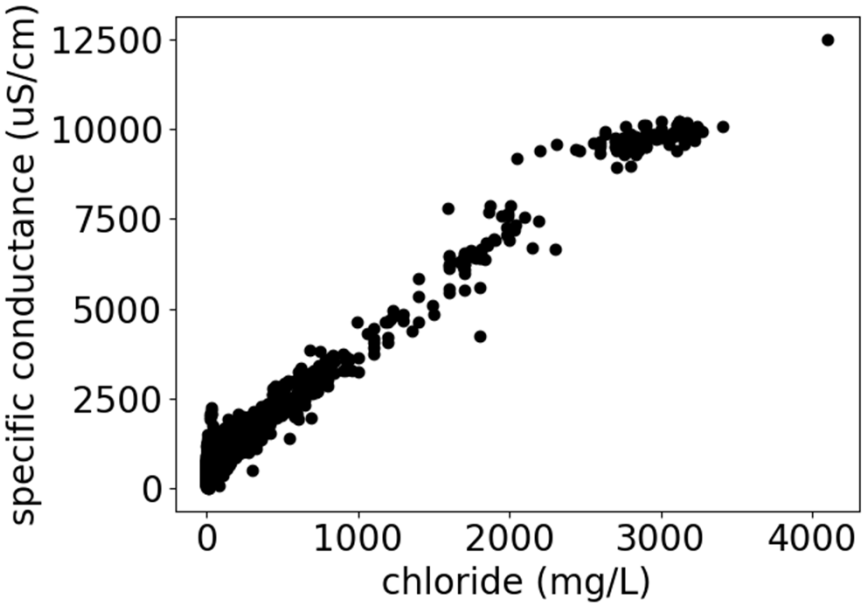
Mississippi River Valley Alluvial & Mississippi Embayment Aquifers

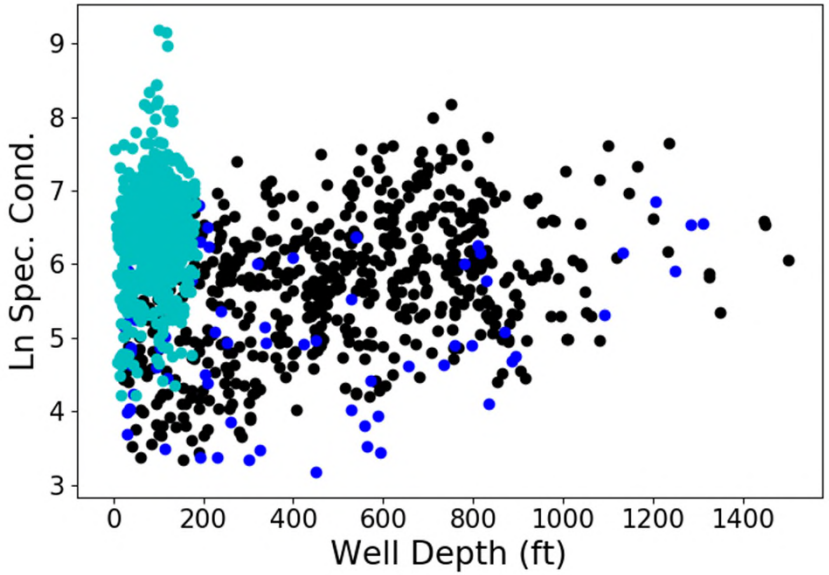


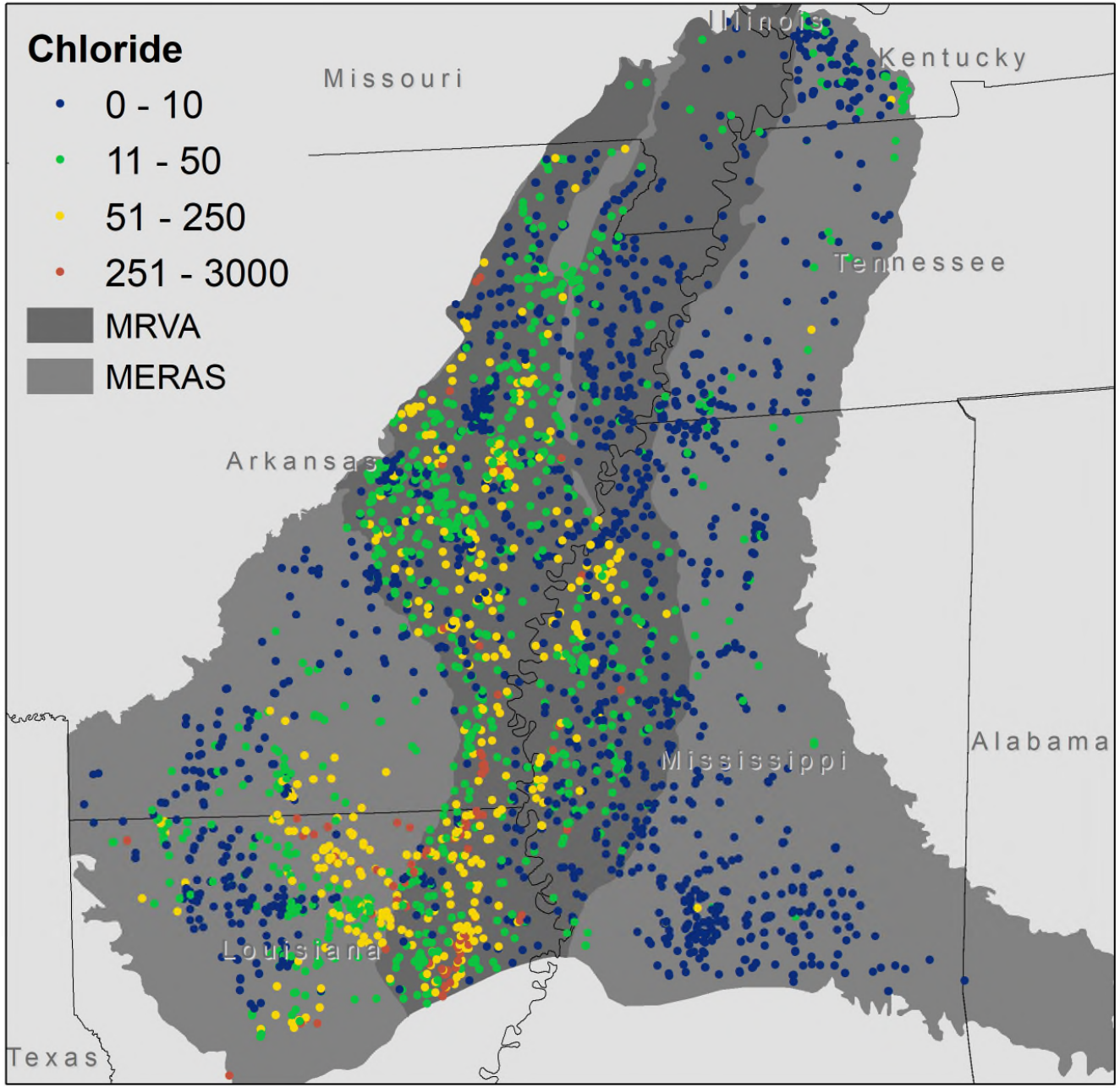


Preliminary Information—Subject to Revision.Not for Citation or Distribution

MERAS Groundwater Quality – salinity

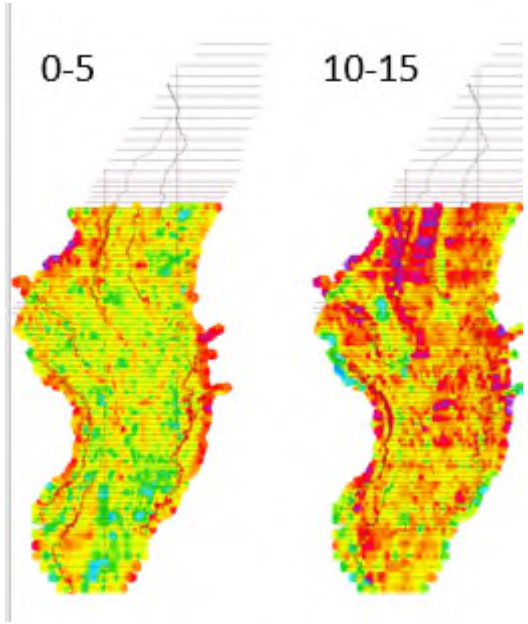


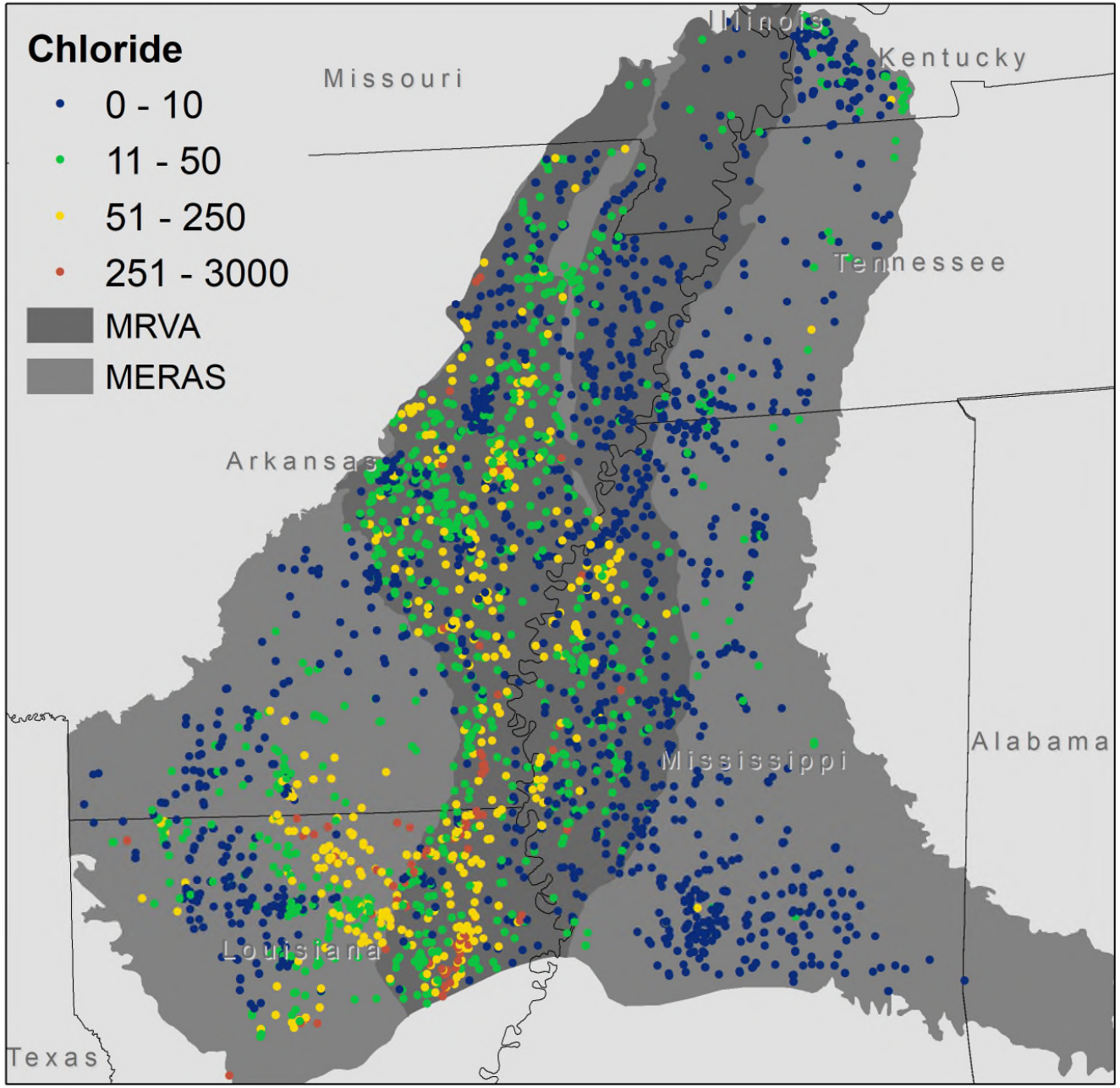




MERAS Groundwater Quality – salinity

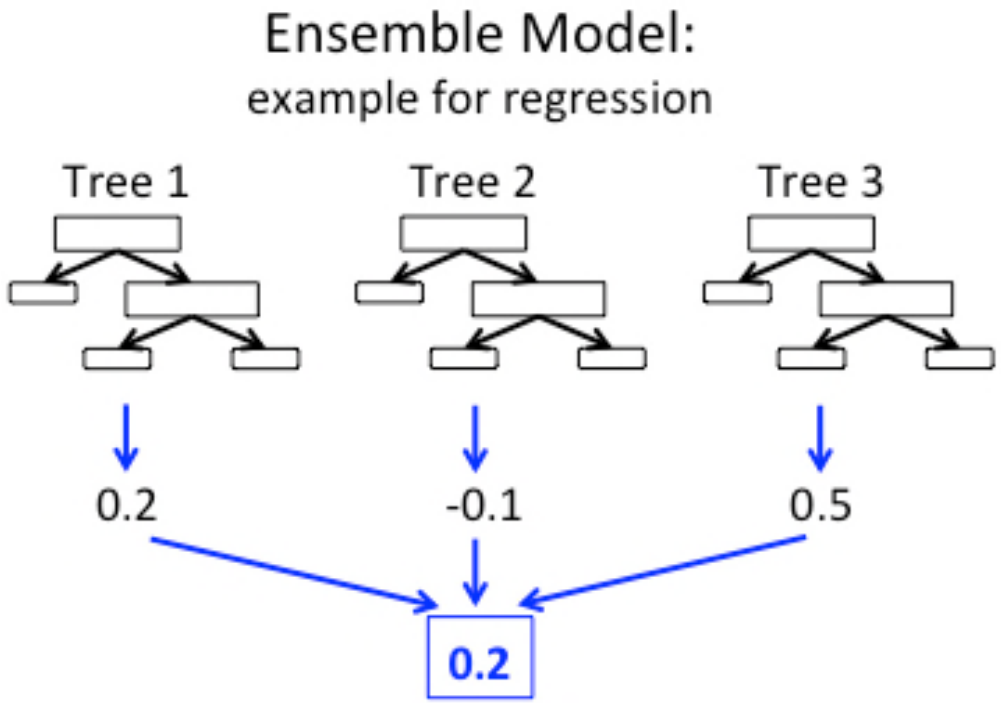
Airborne Electromagnetic Survey





Machine-Learning Methods: Boosted Regression Trees

ensemble method - build lots of simple trees (weak learners)
boosting - use residual of previous tree



PROS: non-linear, non-monotonic, no hypothesis testing assumptions, handles missing data, can use many predictors

CONS: susceptible to over-fitting, can be harder to interpret (black box?)

Model Hyperparameters

complexity

- interaction depth (how deep to split trees)
- minimum observations per node
- learning rate (how much of previous tree to use)
- number of trees

Solution: find simpler models with similar predictive performance

Preliminary Information—Subject to Revision. Not for Citation or Distribution

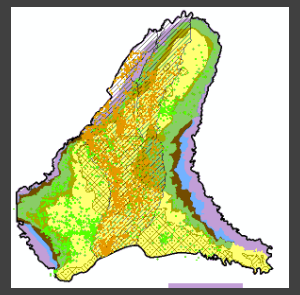
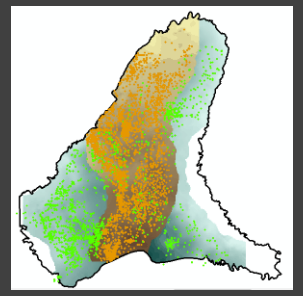
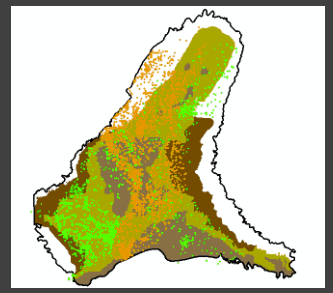
Modeling 3D WQ in Mississippi Embayment

Machine learning to model and map groundwater quality

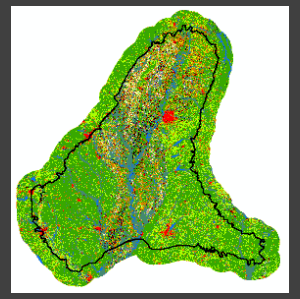
Enables groundwater quality to be placed into a groundwater-flow system context

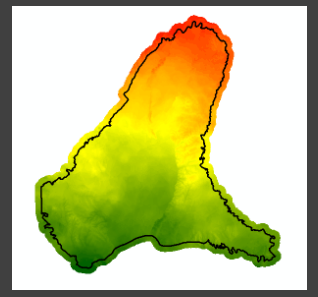
EXPLANATORY VARIABLES (400+)

WELL GEOMETRY

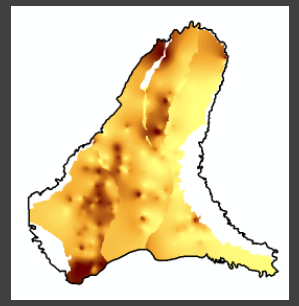




SURFACE VARIABLES





GW MODEL VARIABLES



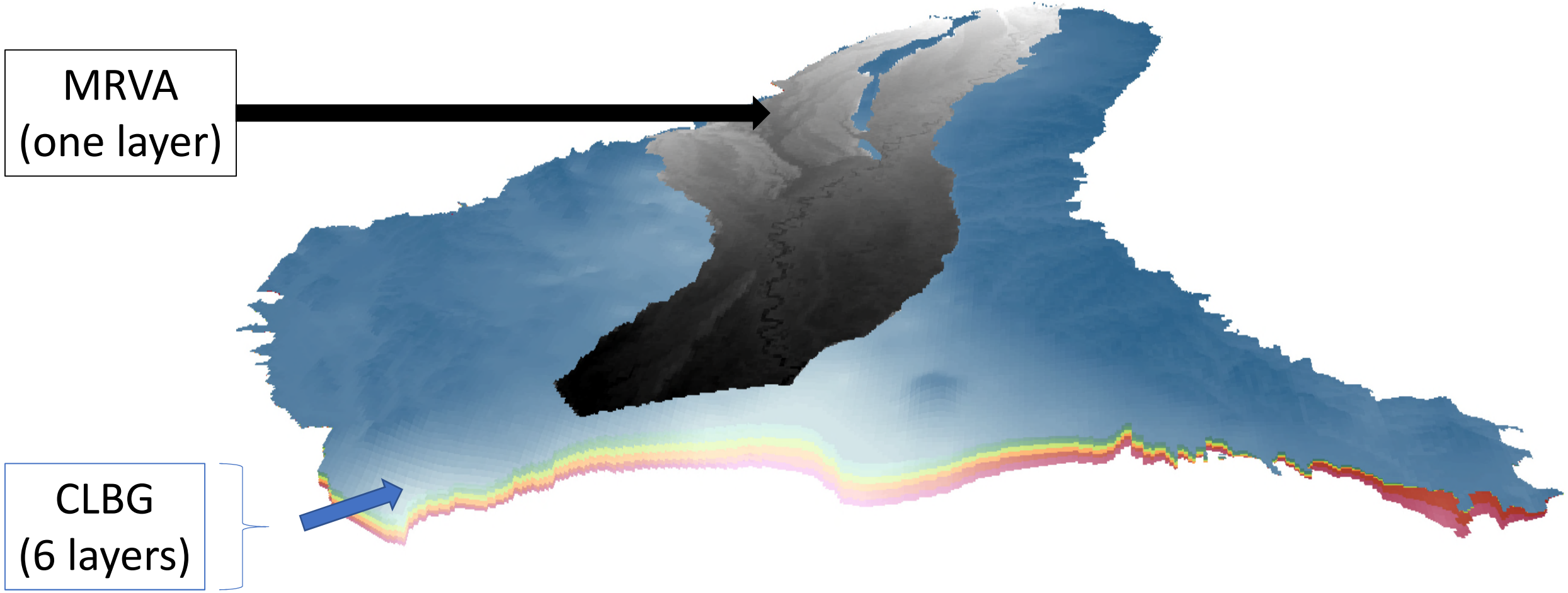
Machine Learning:
Boosted Regression Tree

RESPONSE VARIABLES
SC, Cl, pH, redox, Mn, As

Map GW WQ across MRVA, MCAQ, and LCAQ

Preliminary Information—Subject to Revision.
Not for Citation or Distribution

Final Prediction Framework



Modeling 3D WQ in Mississippi Embayment

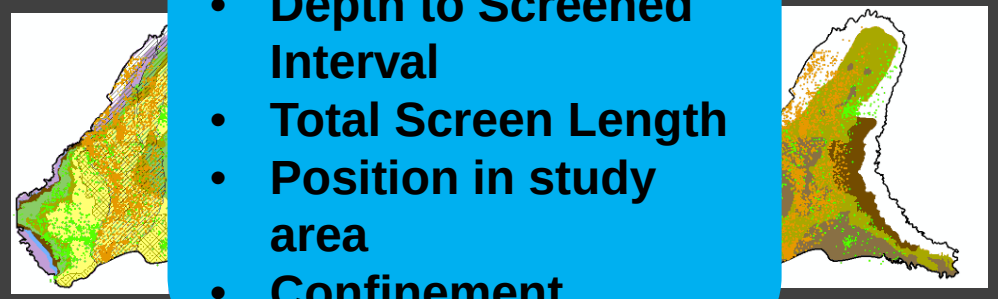
Machine learning to model and map groundwater quality

Enables groundwater quality to be placed into a groundwater-flow system context

EXPLA (400+)

- **Aquifer**
- **Depth to Screened Interval**
- **Total Screen Length**
- **Position in study area**
- **Confinement**

WELL GEOMETRY



SURFACE VARIABLES

GW MODEL VARIABLES

Machine Learning: Boosted Regression Tree

RESPONSE VARIABLES
SC, Cl, pH, redox, Mn, As

Map GW WQ across MRVA, MCAQ, and LCAQ

Preliminary Information—Subject to Revision.
Not for Citation or Distribution

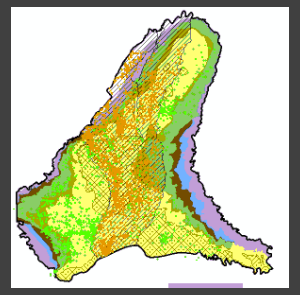
Modeling 3D WQ in Mississippi Embayment

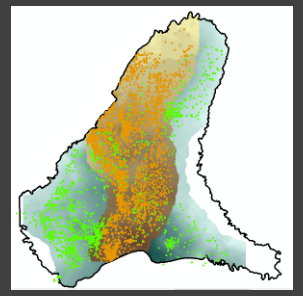
Machine learning to model and map groundwater quality

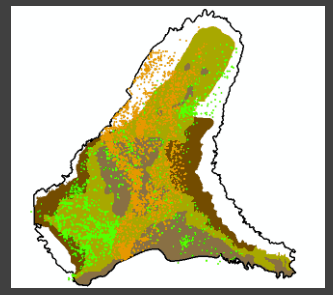
Enables groundwater quality to be placed into a groundwater-flow system context

EXPLANATORY VARIABLES (400+)

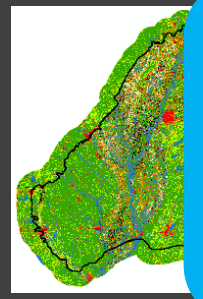
WELL GEOMETRY



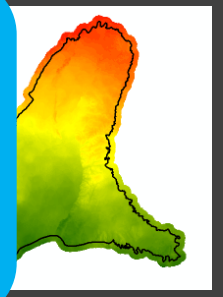




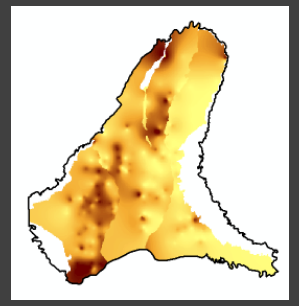
SURFACE VARIABLES



- **Land Use**
- **Soils (phys. params)**
- **Soil (chemistry)**
- **Climate**
- **Recharge**



GW MODEL VARIABLES



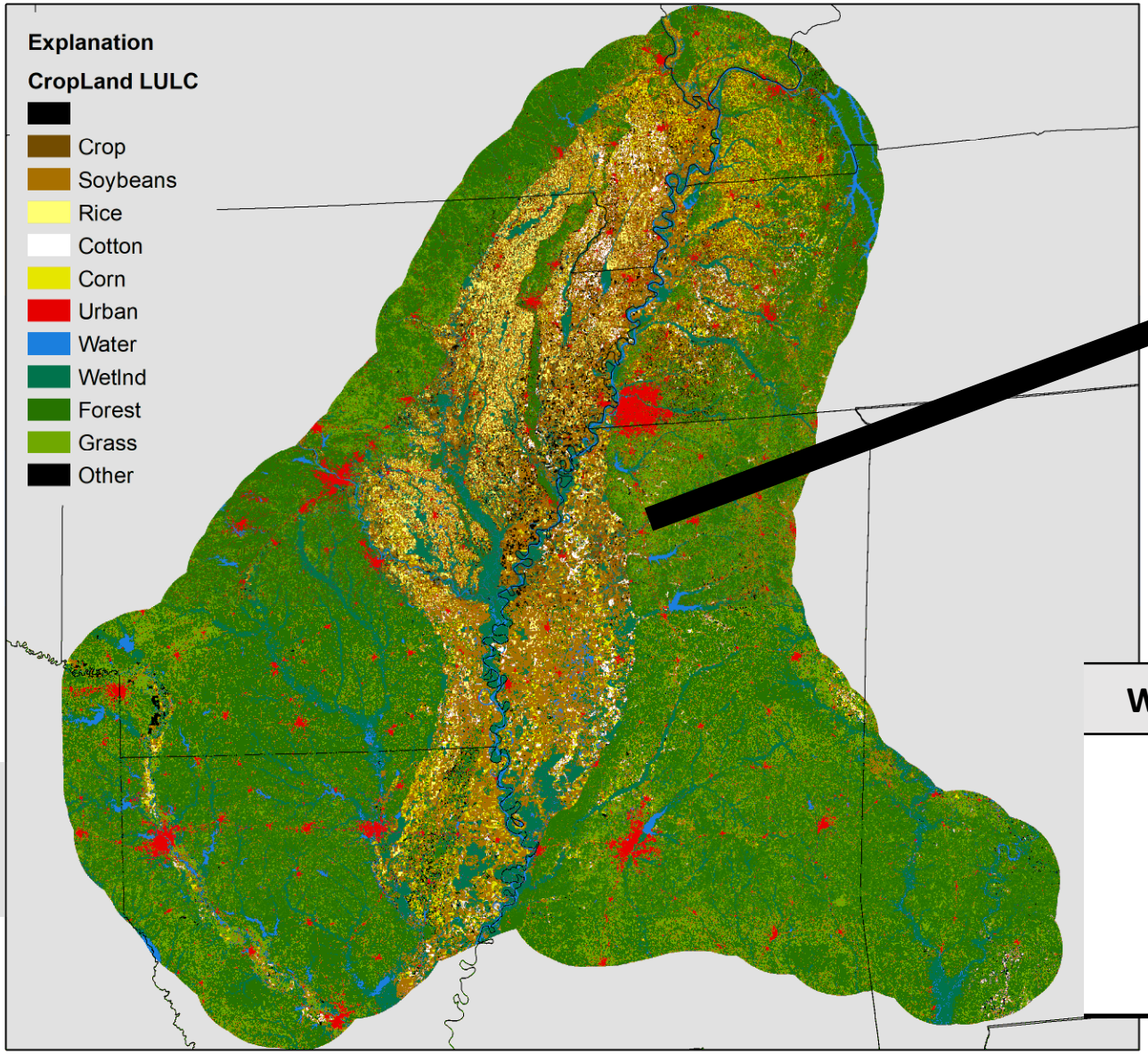
Machine Learning:
Boosted Regression Tree

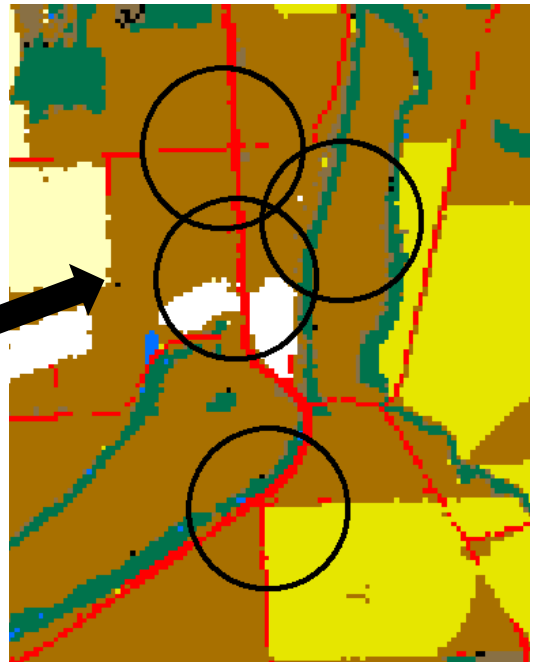
RESPONSE VARIABLES
SC, Cl, pH, redox, Mn, As

Map GW WQ across MRVA, MCAQ, and LCAQ

Preliminary Information—Subject to Revision.
Not for Citation or Distribution

Surface Variable Example: Land Use/Land Cover





https://github.com/brclark-usgs/zonepy

Well	Soybeans	Corn	Cotton	Wetlnd	Crop	Urban
1	92	-	-	-	-	8
2	72	9	-	13	6	-
3	56	27	-	7	1	9
4	61	-	25	5	1	8

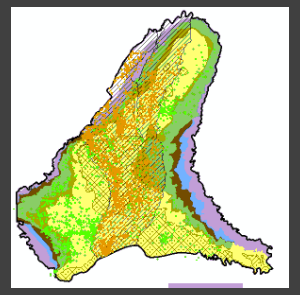
Modeling 3D WQ in Mississippi Embayment

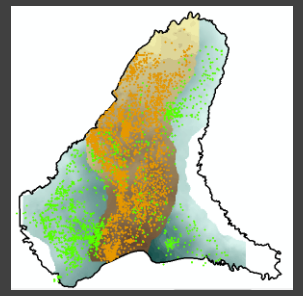
Machine learning to model and map groundwater quality

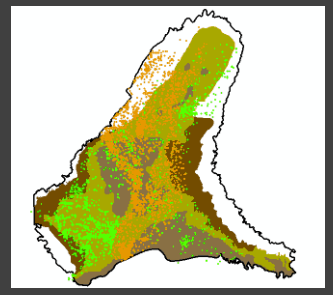
Enables groundwater quality to be placed into a groundwater-flow system context

EXPLANATORY VARIABLES (400+)

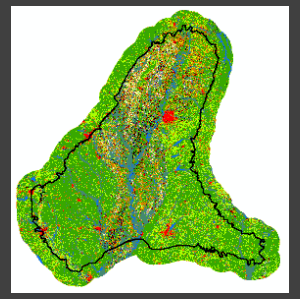
WELL GEOMETRY

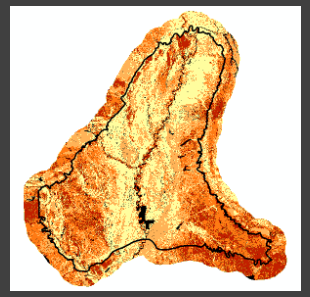


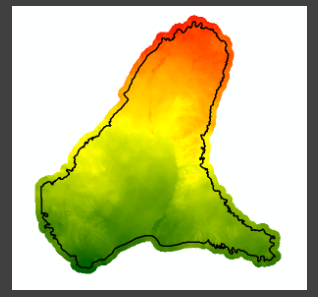




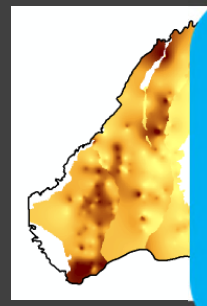
SURFACE VARIABLES







GW MODEL VARIABLES



- Travel Time
- Heads
- Flux
- Water Use
- Groundwater Age

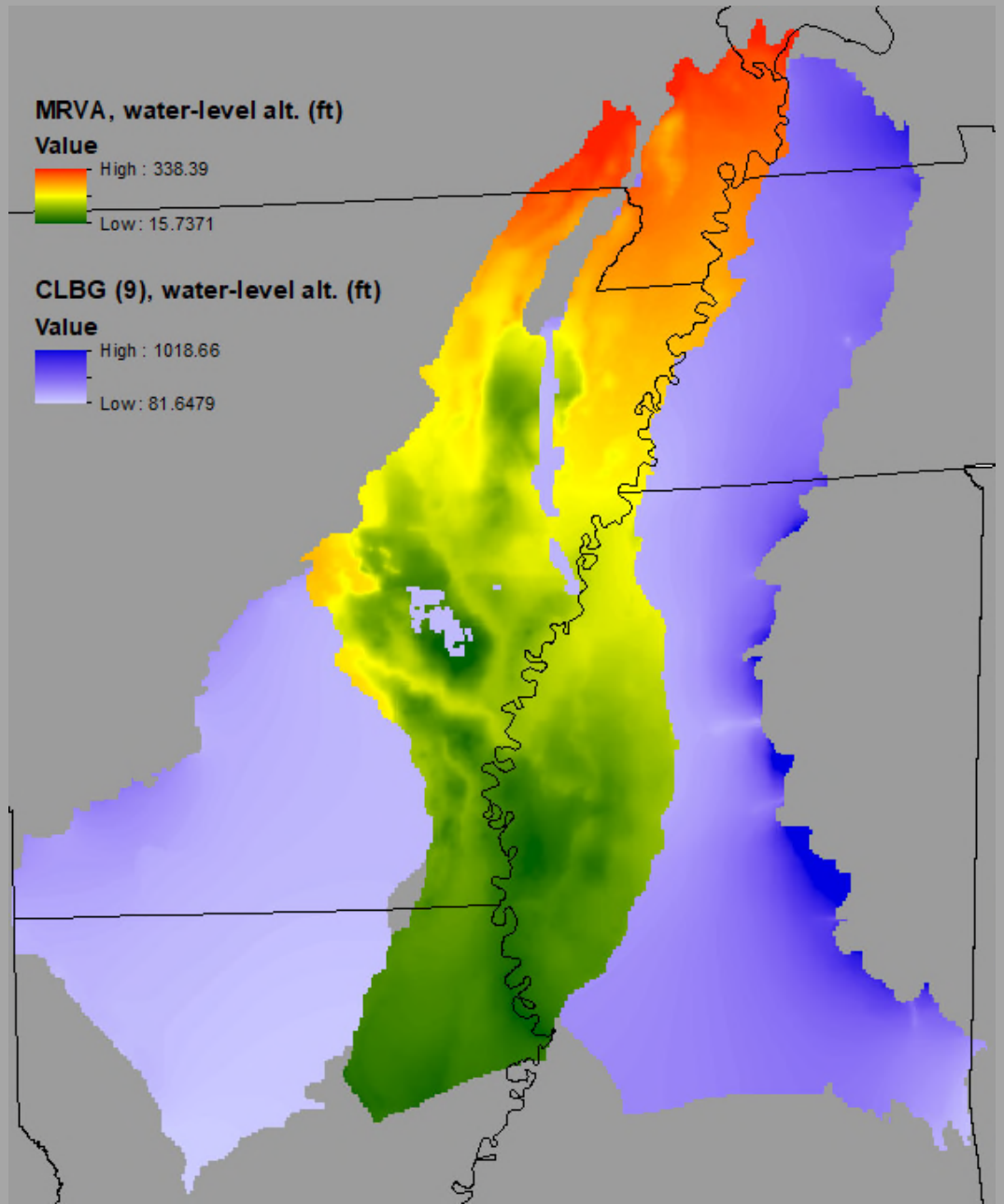
Machine Learning: Boosted Regression Tree

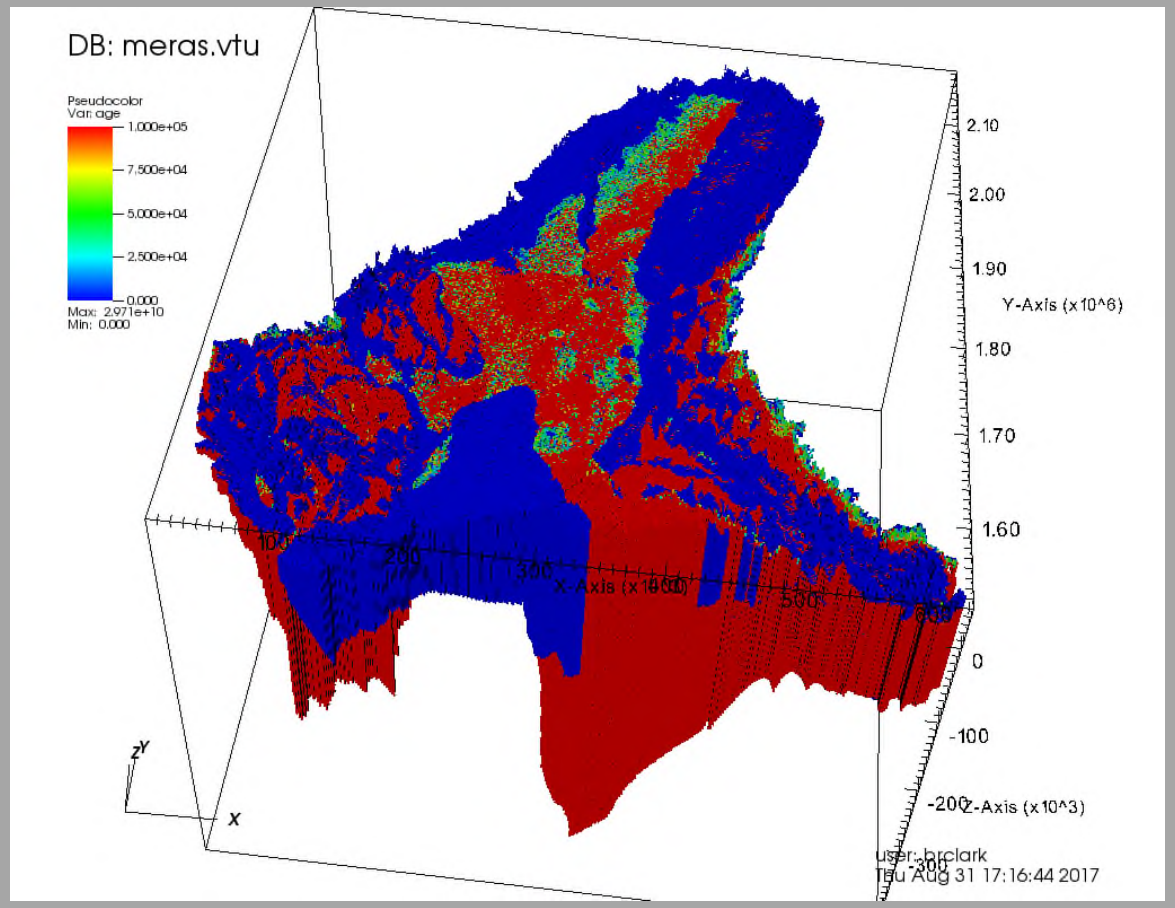
RESPONSE VARIABLES
SC, Cl, pH, redox, Mn, As

Map GW WQ across MRVA, MCAQ, and LCAQ

Preliminary Information—Subject to Revision.
Not for Citation or Distribution

GW Model Variables: GW Levels and Age

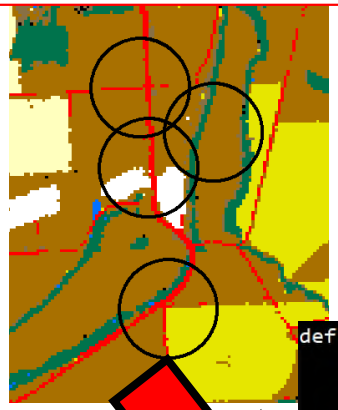




Particles added to the GW flow model → time of travel approximates (relative) groundwater age

Workflow – automation via Python & R

Zonepy → loop thru exp. var. rasters and attribute all wells



Python (Glob) → merge multiple text files into one attribute table (400+ variables)

```python
def JoinVars(varsloc, fldname, calcstat, delcols=None):
    # loop thru texts in folder
    texts = []
    for f in glob.iglob(os.path.join(varsloc,'*.csv'))
        texts.append(f)
    # setup empty lists for dfs
    dfs = []
    mod_dfs = []

    # loop thru texts to extract columns
    for t in texts:
        param = os.path.splitext(os.path.basename(t))[
        print(param)
        data = pd.read_table(t, sep = ',', dtype = {fl
        # print(data.head(5))
        # print(data.shape)
        cols = data.columns.tolist()
        if 'gw' in param: #gw model variables
            newdf = data.copy()
            print(newdf.shape)
            newdf.set_index(fldname, inplace=True)
            # print(newdf.head(5))
            dfs.append(newdf)
            mod_dfs.append(newdf)
```

- bg_grav.csv
- bg_iso.csv
- gSur_ph1to1h2o_r.csv
- gw_ang.csv
- gw_dtw.csv
- gw_h2v.csv
- gw_hd.csv
- gw_lk4.csv
- gw_mag.csv
- gw_qz.csv
- gw_rch.csv
- gw_wu.csv
- lu_c15_10.csv
- lulc_n74_8.csv
- modis.csv
- mohp1DSD.csv
- mohp1LP.csv
- mohp2DSD.csv
- mohp2LP.csv
- mohp3DSD.csv
- mohp3LP.csv
- mohp4DSD.csv
- mohp4LP.csv
- mohp5DSD.csv
- mohp5LP.csv

Python (pandas) & R (corr) → remove variables with linear correlation $r^2 > 0.8$

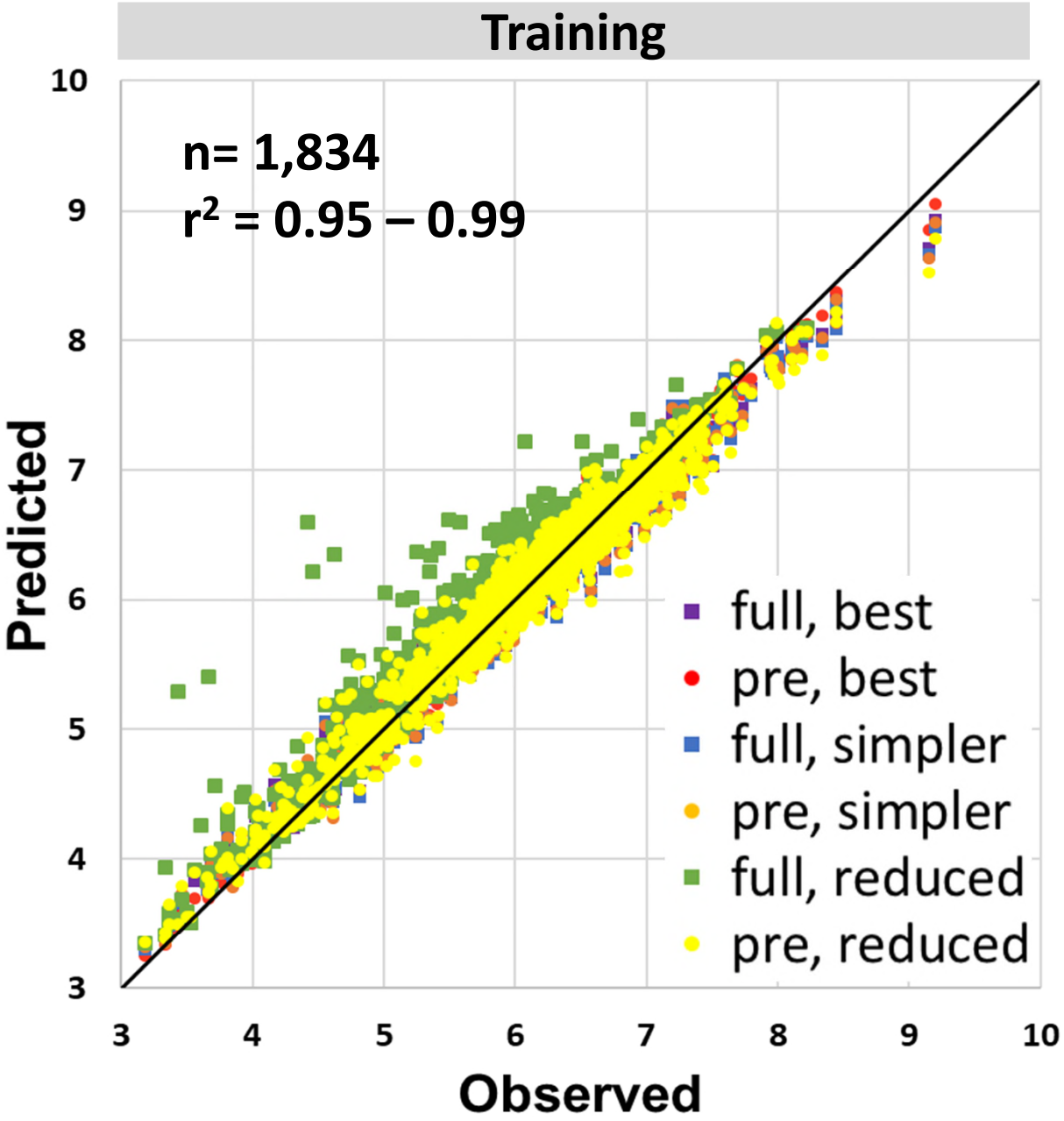
	C	D	E	F	G
	bg_iso	gSur_ph1t	ang_rng	ang_slp	ang_med
	7.260073	6.280295	160.1024	-0.0014	-8.07866
	10.89028	6.12266	71.98197	-0.00052	-65.1159
	6.692144	7.117763	18.58083	1.09E-05	-3.71036
AR008_330302091212401	5.032942	9.4683	5.410837	32.62897	0.000396	-44.4697
AR008_330341091300901	8.025515	12.00828	5.665025	46.47678	-0.00099	-146.896
AR008_330626091352901	8.512645	12.89164	5.665025	20.37122	9.73E-05	-87.3965
AR008_330703091132301	12.10899	16.00603	7.117763	17.32036	-0.00051	-26.7025
AR008_330918091261501	15.8318	20.30731	6.821675	165.5763	0.00066	-128.276
AR008_331255091074001	3.614657	7.920826	7.117763	24.39902	-4.39E-05	16.07767
AR008_331406091114201	12.48824	17.04513	7.117763	19.81951	0.000624	13.75939
AR008_331408091095301	6.424456	10.89895	7.117763	26.28498	0.001032	18.41809
AR008_331415091170801	23.44107	28	.117763	64.23013	0.002183	14.29748

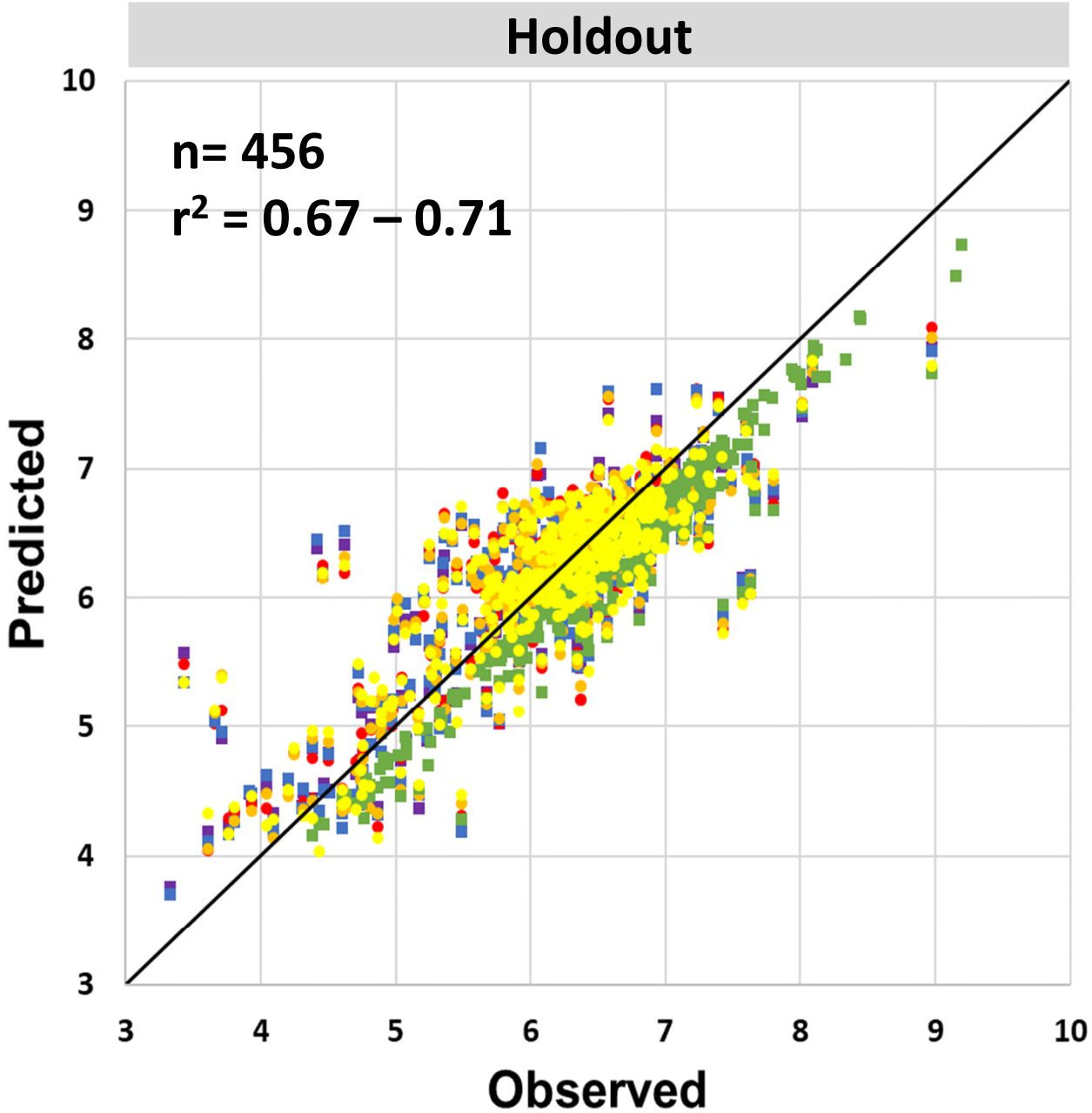
R (parallel & caret) → Train models on YETI USGS supercomputer in parallel

```
/lustre is online
/cxfs/projects is

Docs: https://gitl
```

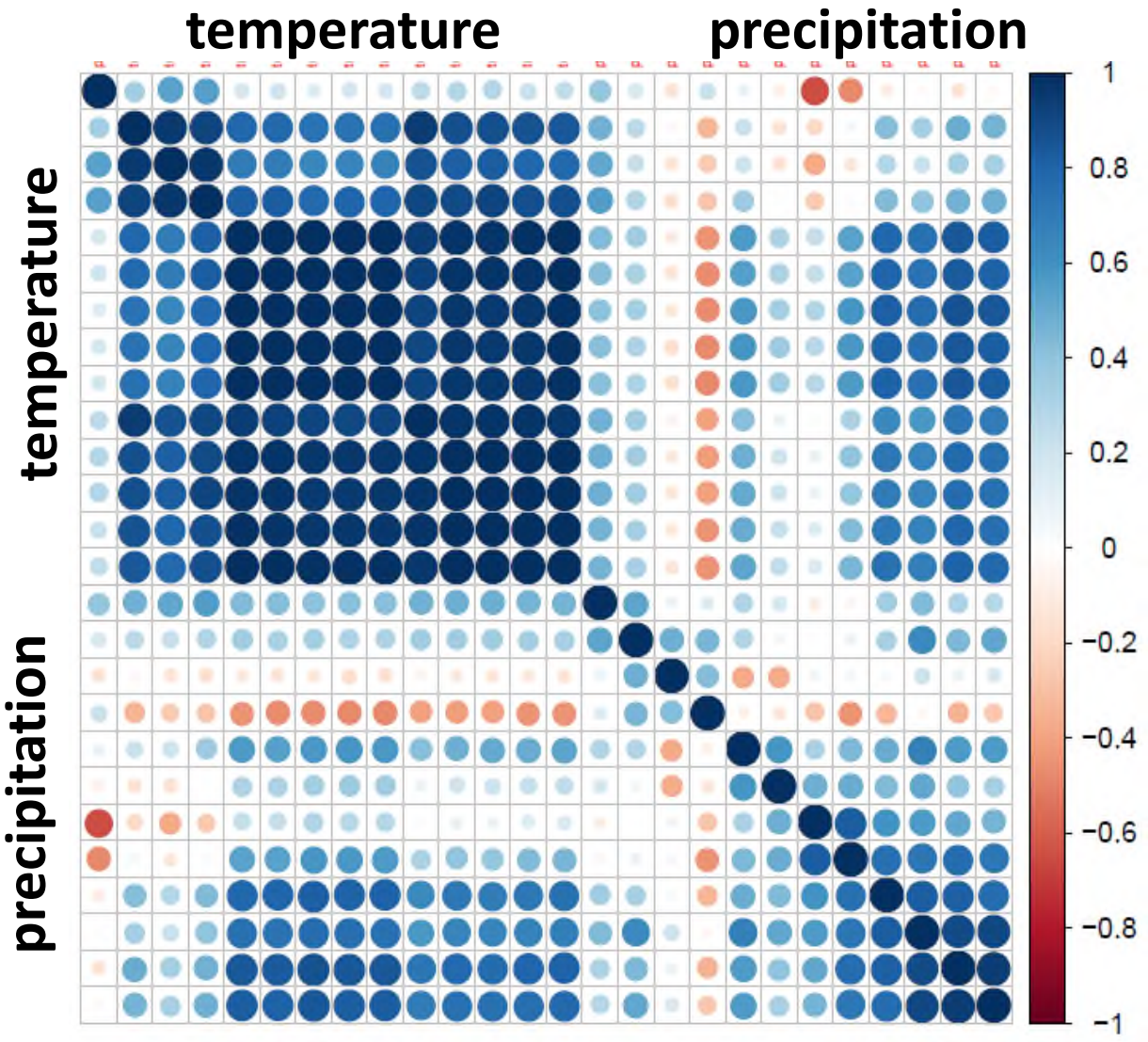
Specific Conductance Models

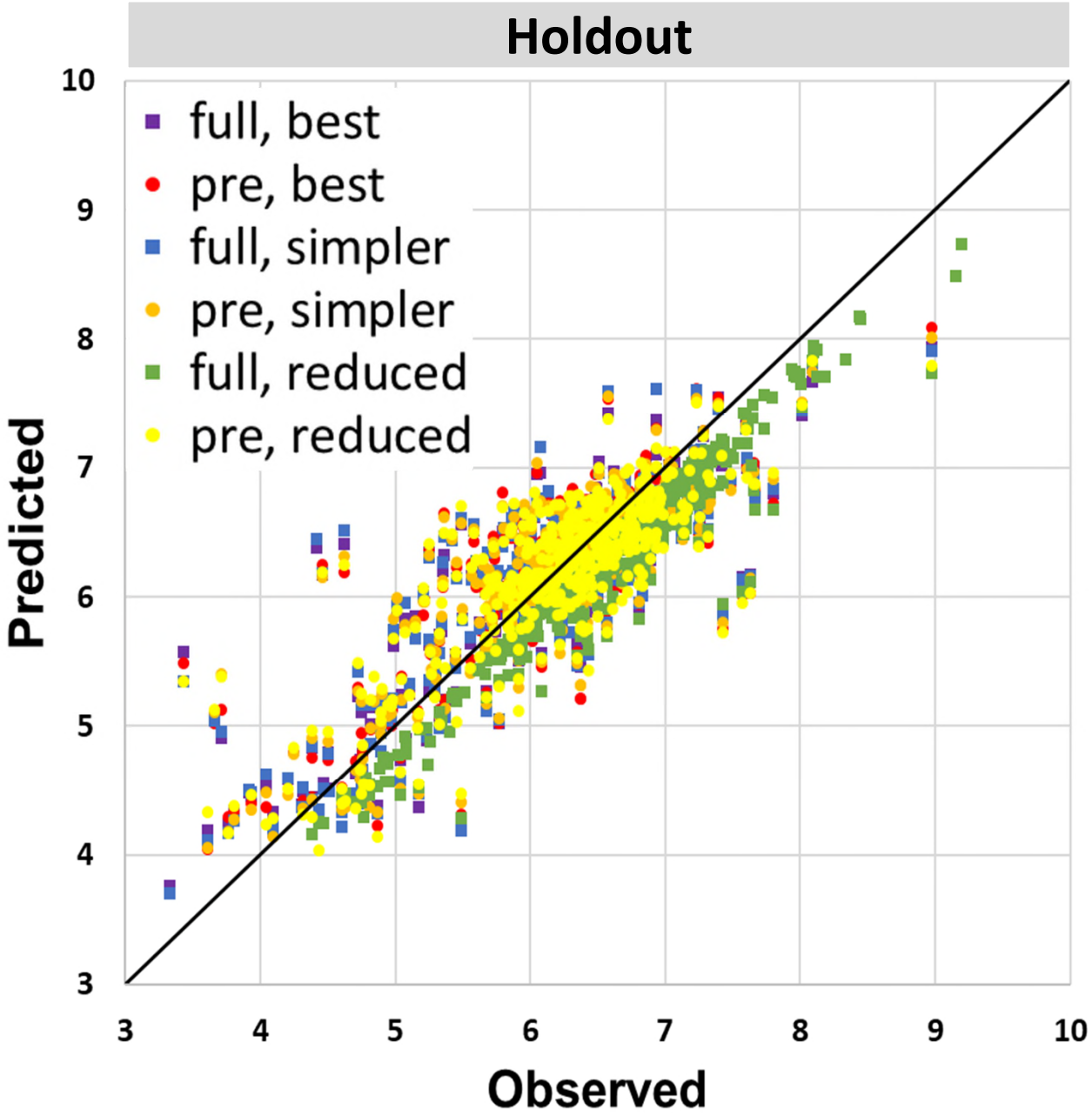




Specific Conductance Models

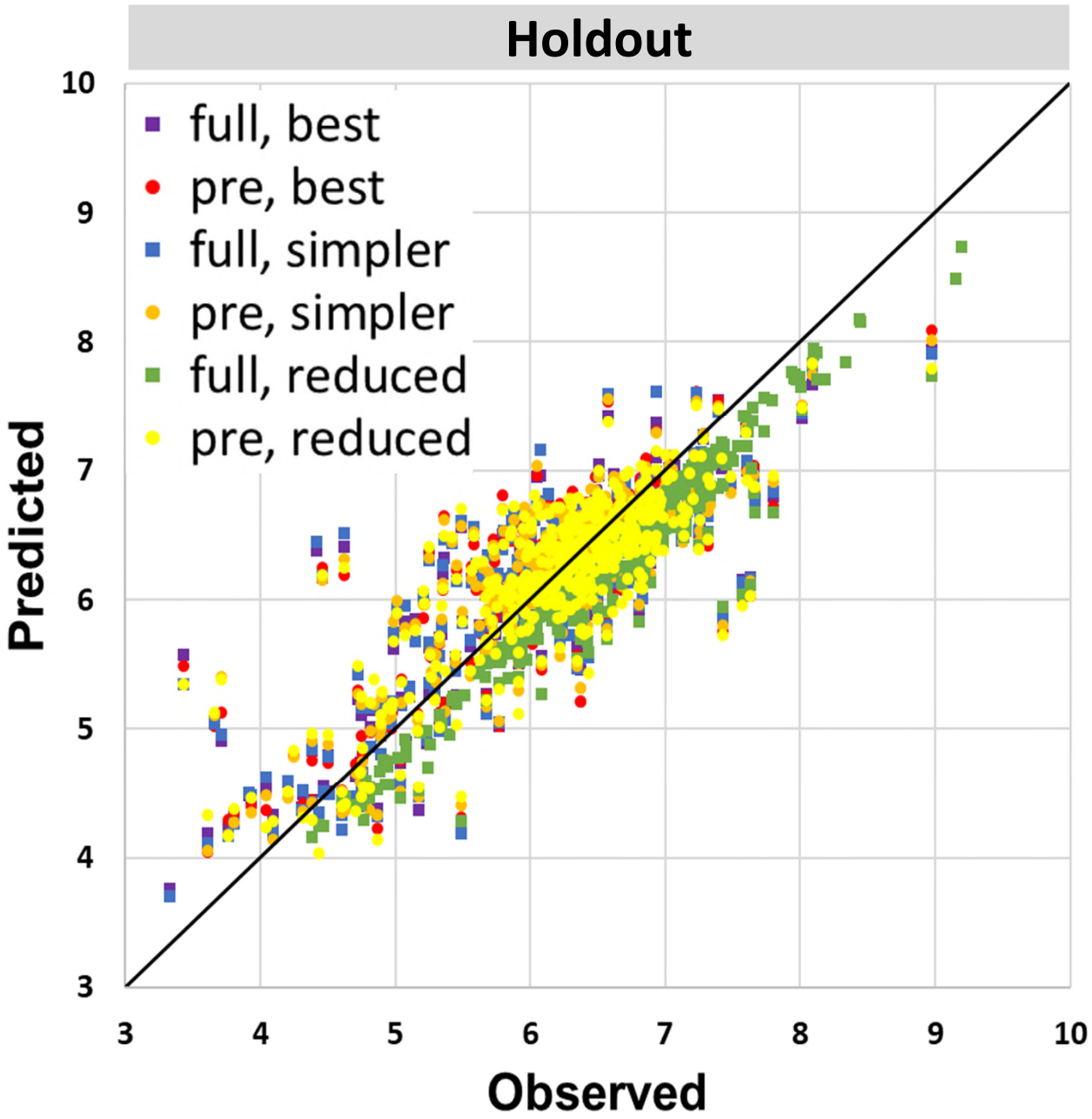
Pre-processing = remove linear correlated variables

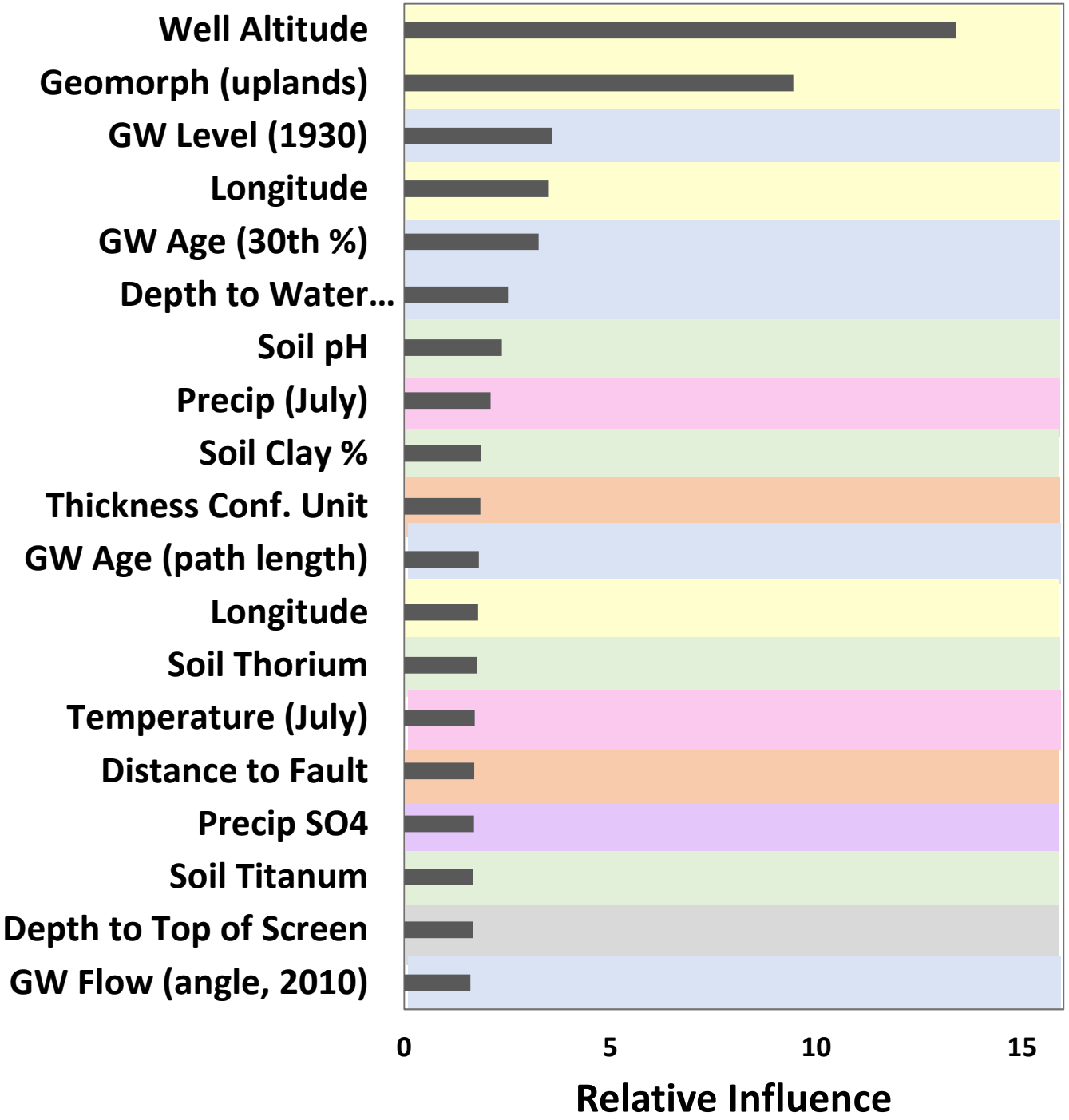




Specific Conductance Models

model	r^2	no. vars	int. depth	min. obs.	lrn. rate	no. trees
Full, best	0.71	431	16	8	0.004	4500
Pre-processed, best	0.68	225	18	8	0.006	4000
Full, simpler	0.69	431	6	8	0.014	3000
Pre-processed, simpler	0.67	225	6	10	0.010	5000
Full, reduced	0.69	56	6	8	0.014	3000
Pre-processed, reduced	0.67	50	6	10	0.010	5000





Relative Influence of Explanatory Variables

Position in System

GW Flow Model / Age

Soil Chemistry / Physical Properties

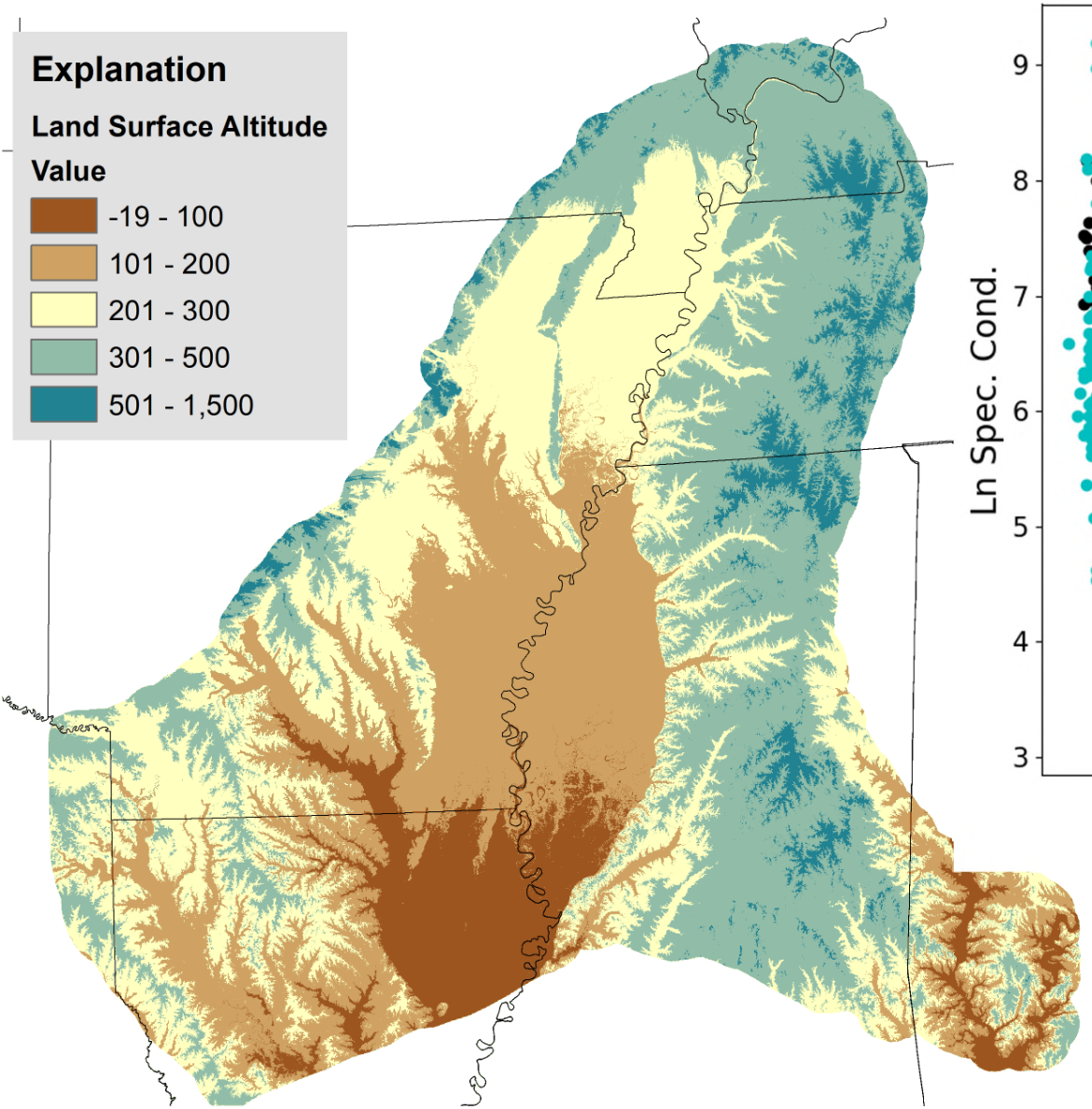
Climate

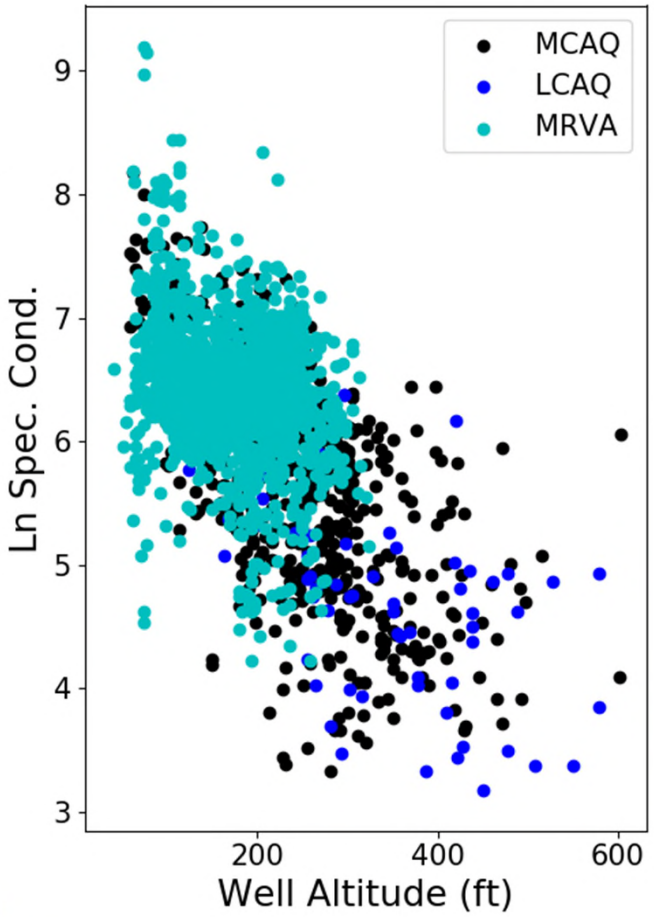
Hydrogeology

Precip. Chemistry

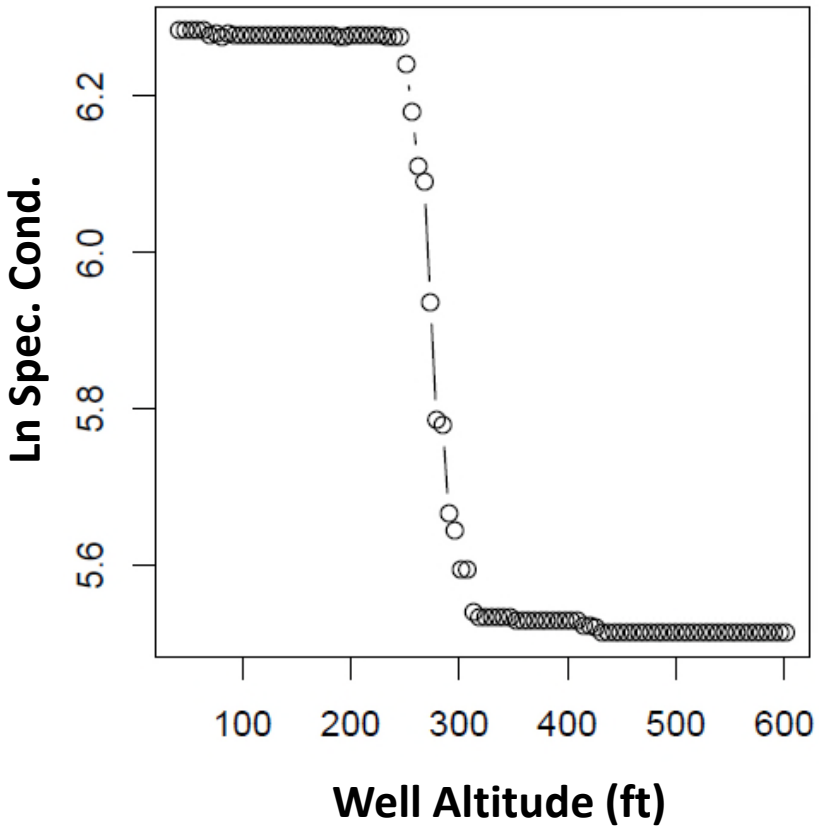
Well Geometry

Well altitude (DEM)

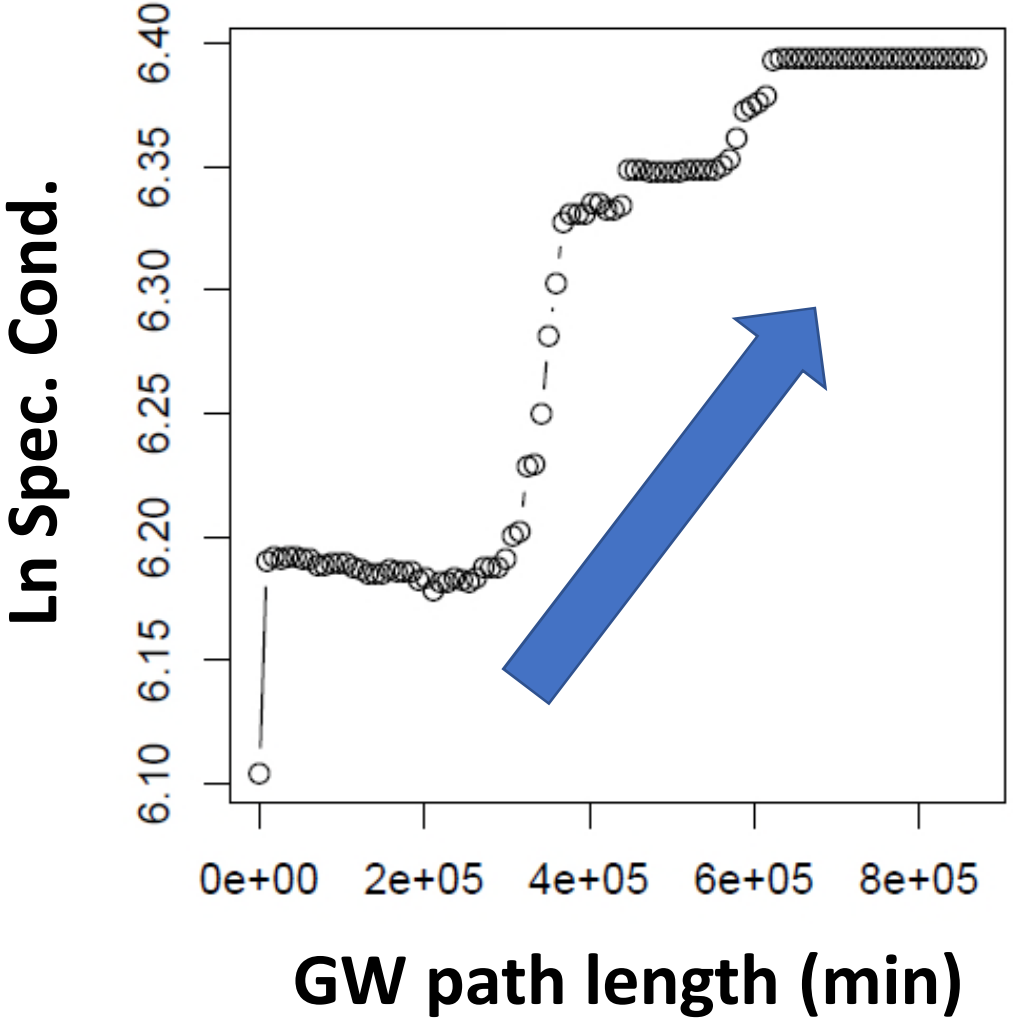


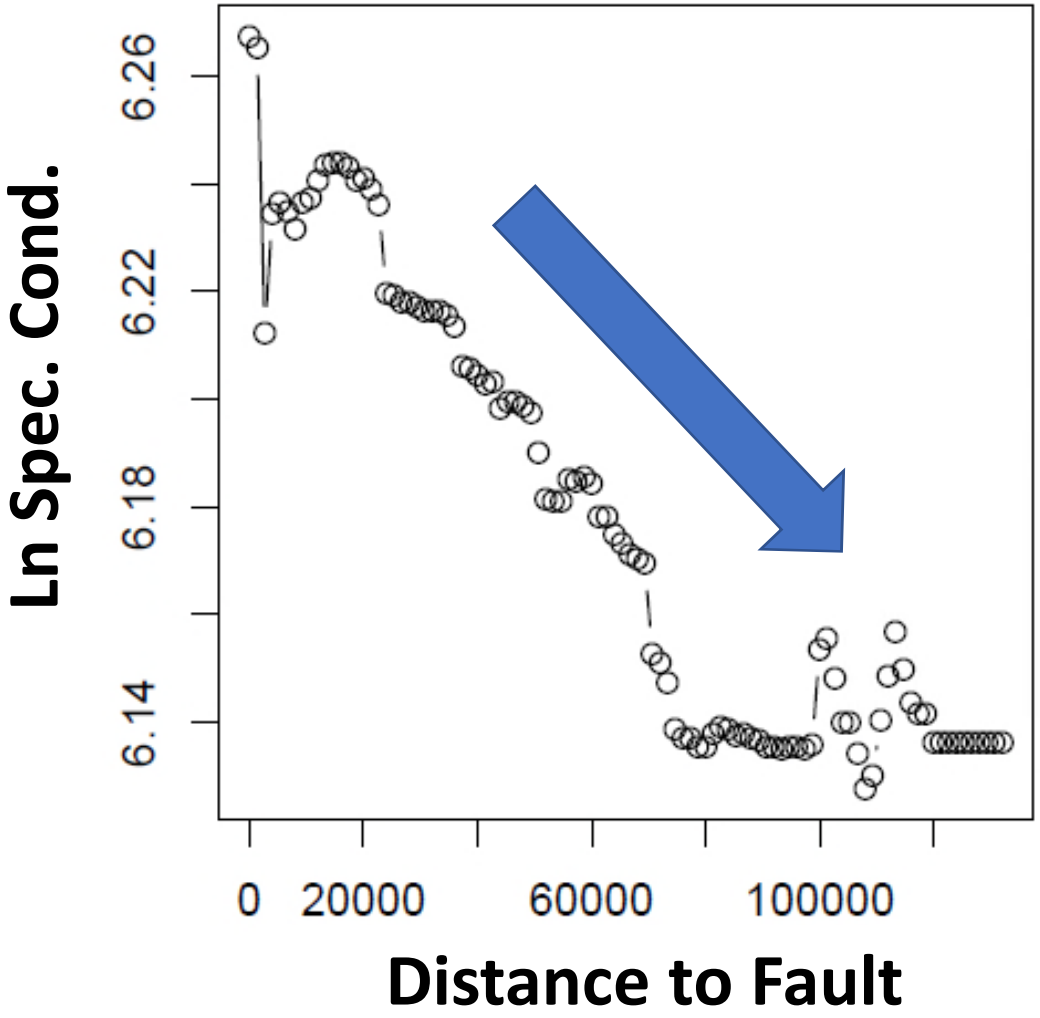


partial dependence plots



Hydrogeology

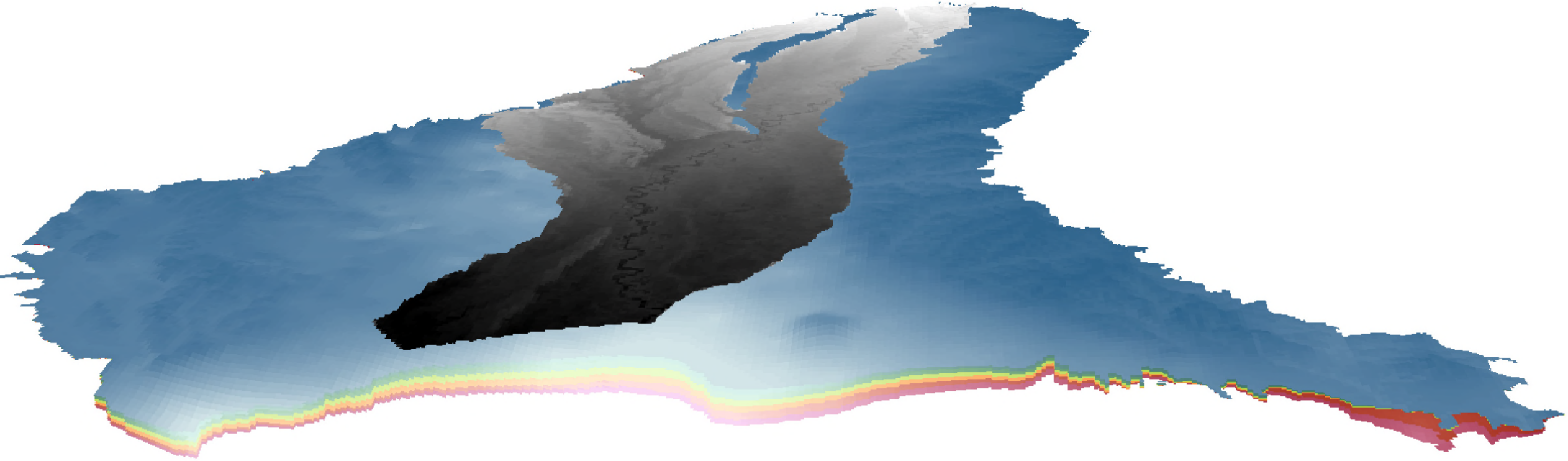




Comparing Explanatory Variables

Specific Cond.	Chloride	pH
Well Altitude	Longitude	Altitude Top of Screen
Geomorph (uplands)	GW (min path length)	GW Age (minimum)
GW Level (1930)	Soil Kaolinite	Longitude
Longitude	GW Level (1930)	GW Age (30th %)
GW Age (30th %)	Thickness Conf. Unit	Geomorph (uplands)
Depth to Water (change)	Well Altitude	Soil Chemistry (Ti)
Soil pH	Normalized Position	Altitude Bottom of Screen
Precip (July)	GW Age (90th %)	Distance to Fault
Soil Clay %	Soil (% fine)	Normalized Position
Thickness Conf. Unit	GW (max path length)	Latitude

Next steps… 3D prediction



QUESTIONS?

Kathy Knierim
kknierim@usgs.gov

USGS
science for a changing world