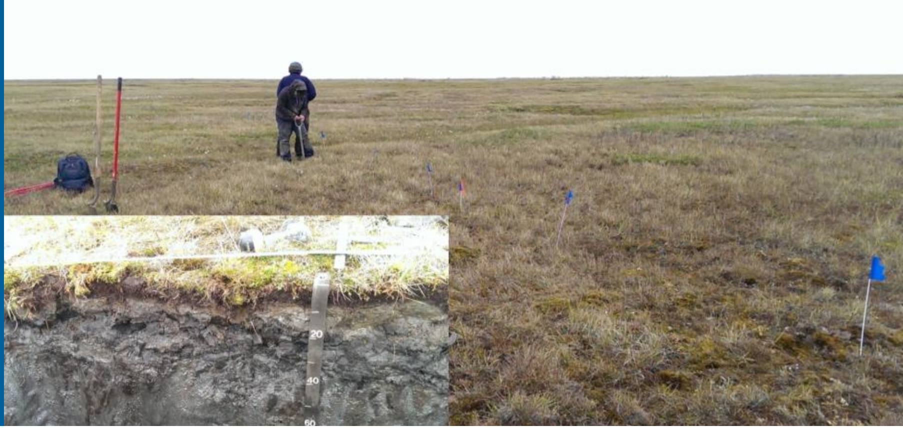
MACHINE LEARNING TO INVESTIGATE SOIL ORGANIC CARBON STORAGE AND DYNAMICS

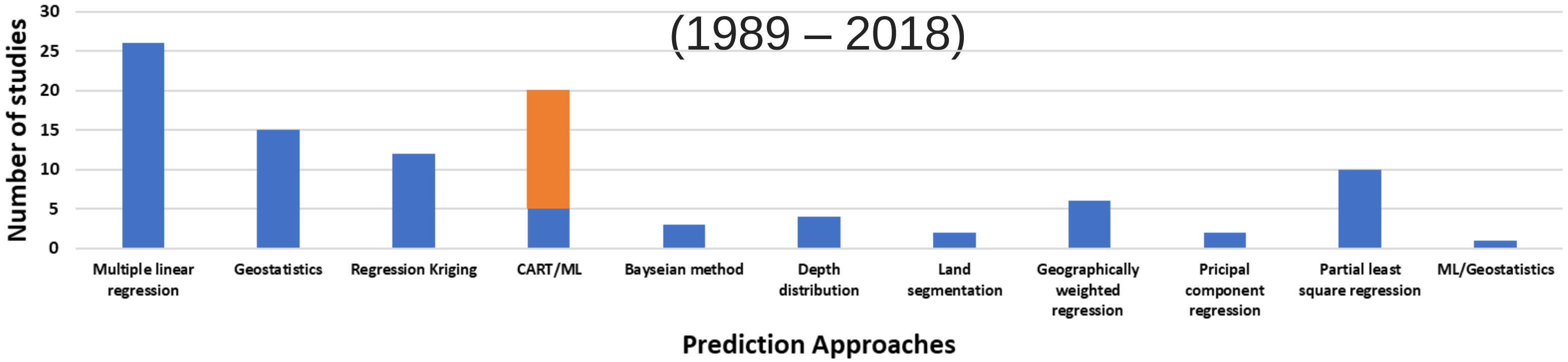


UMAKANT MISHRA

Environmental Science Division

03/20/2020

SPATIAL PREDICTION APPROACHES OF SOC STOCKS



Terrain attributes, climatic data, soil reflectance, land use

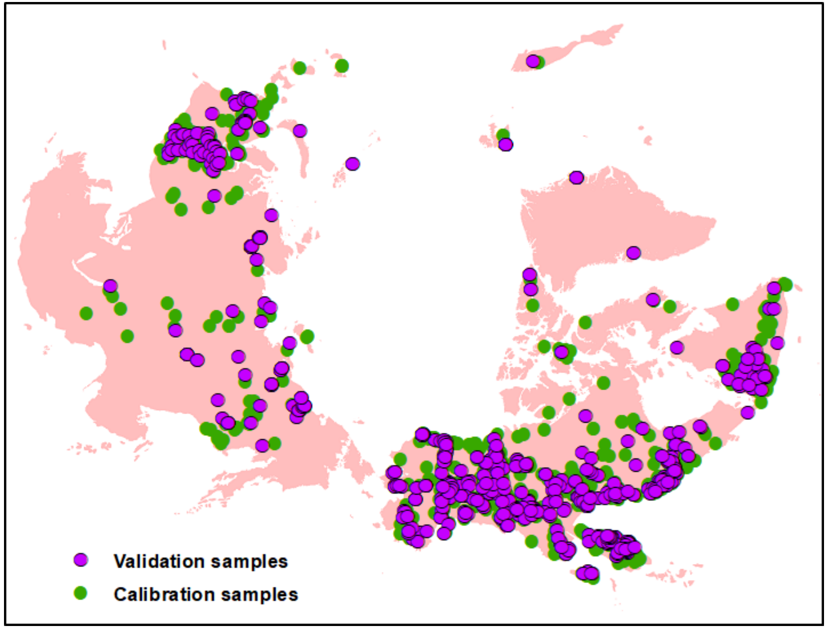
- Various approaches of differing mathematical complexities are being applied for spatial prediction of SOC stocks.
- Regression kriging, which has been reported to produce highest prediction accuracy, is a hybrid approach which combines correlation between SOC and environmental controllers with spatial autocorrelation between soil observations.
- Recently, number of studies using ML has increased.

(Updated from Mishra & Lal, 2011)

COMPARING REGRESSION KRIGING WITH MACHINE LEARNING APPROACHES

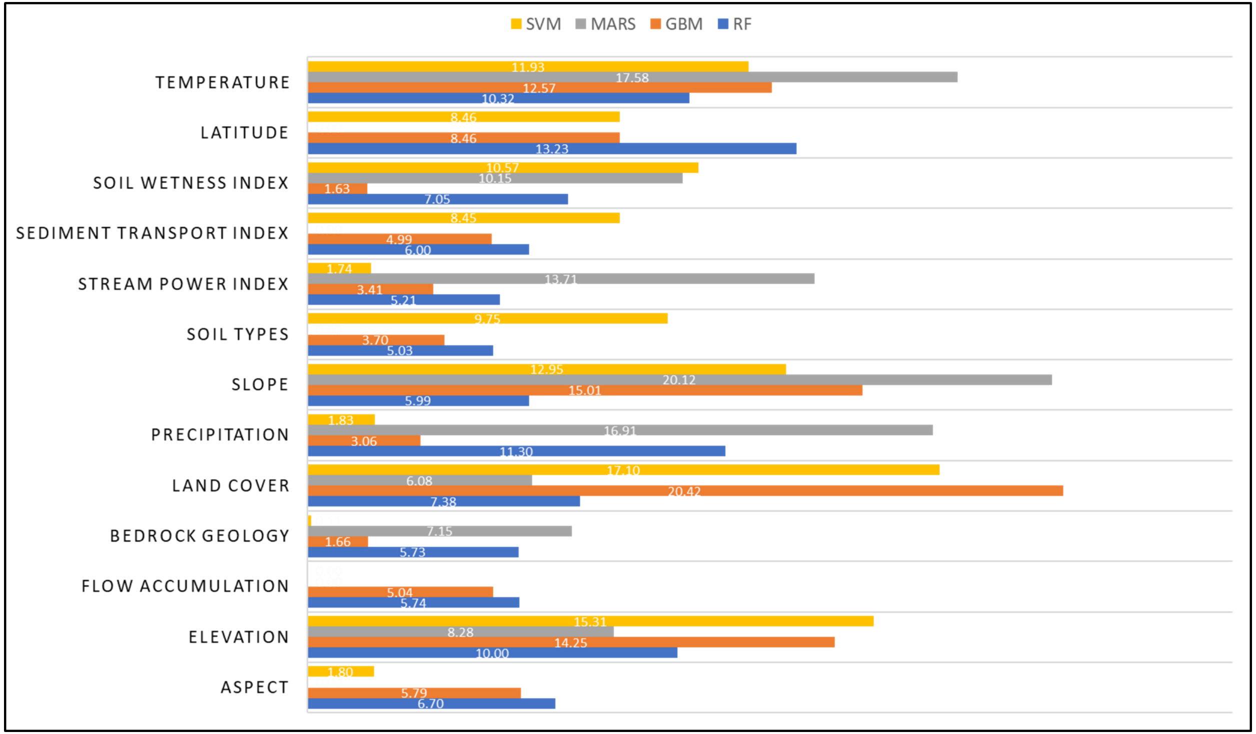
- We compared four machine learning approaches (gradient boosting machine [GBM], multinarrative adaptive regression spline [MARS], random forest (RF), and support vector machine [SVM]) with regression kriging to predict the spatial heterogeneity of surface (0-30 cm) SOC stocks.

- We used 2374 surface soil samples and a variety of environmental covariates to predict the spatial heterogeneity of SOC stocks at 250-m spatial resolution across the northern circumpolar permafrost region.



VARIABLE IMPORTANCE IN DIFFERENT SPATIAL PREDICTION APPROACHES

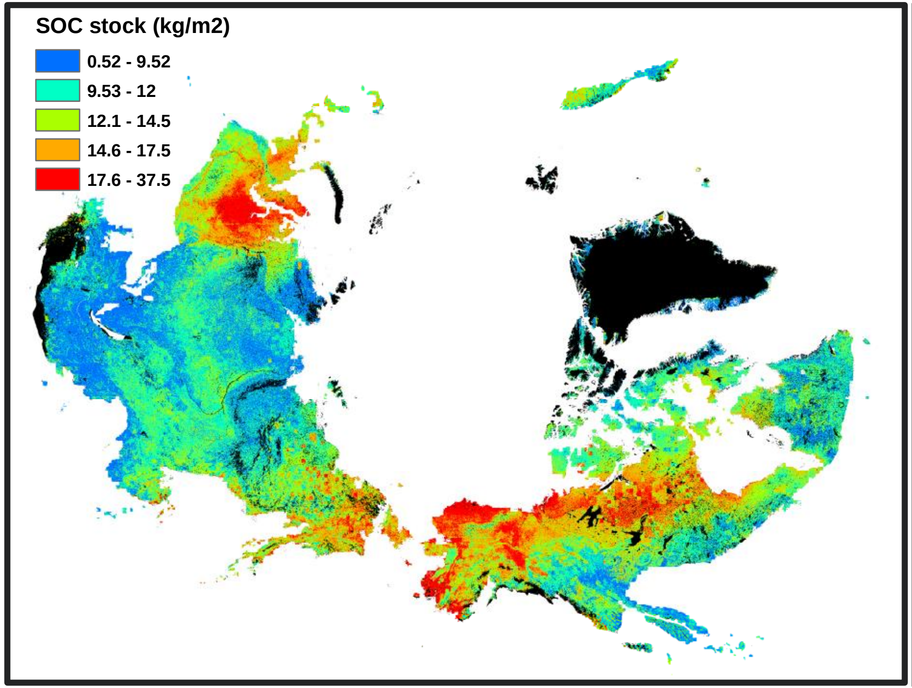
SVM = support vector machine, MARS = multinarrative adaptive regression spline, GBM = Gradient Boosting Machine, RF = random forest

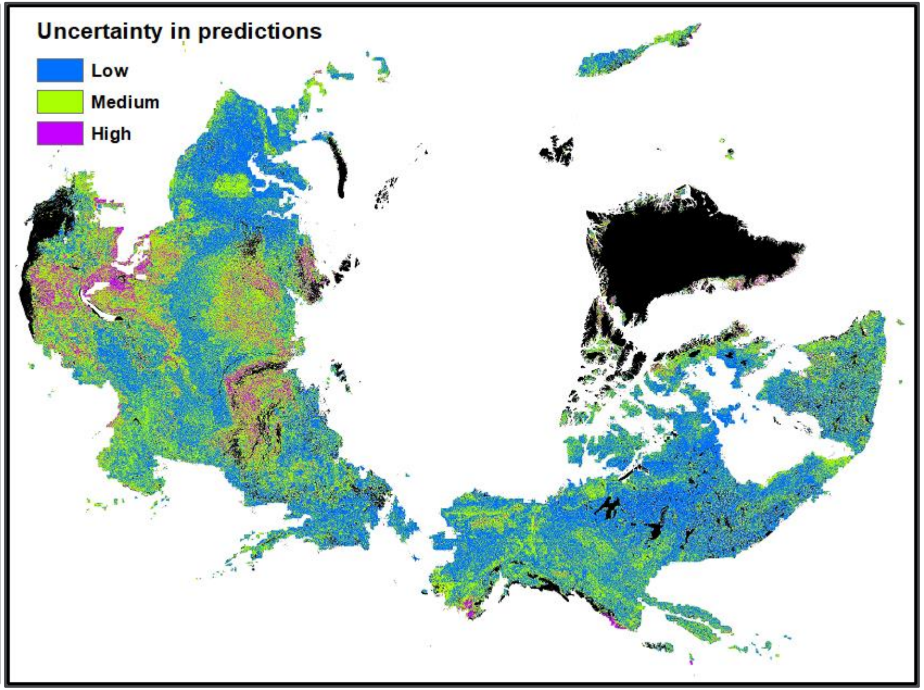


Regression kriging
Temperature
Latitude
Bedrock geology
Land cover types
Soil types
Sediment transport index
Stream power index

(Mishra et al., under review)

ENSEMBLE MACHINE LEARNING APPROACH BETTER PREDICTS SOIL ORGANIC CARBON STOCKS





Median predictions from
4 Machine Learning approaches

Low = <20%
Medium = 20-49%
High = >50% uncertainty
in predicted SOC stocks

PREDICTION ACCURACY OF DIFFERENT SPATIAL PREDICTION APPROACHES (N = 714 SITES)

- Regression kriging approach produced lower prediction errors in comparison to MARS and SVM, and comparable prediction accuracy with GBM and RF techniques.

- The ensemble median prediction of SOC stocks obtained from all four machine learning techniques showed highest prediction accuracy.

Prediction approaches	Validation Indices				
	r	RMSE ($kg\ m^{-2}$)	MEE ($kg\ m^{-2}$)	SDE ($kg\ m^{-2}$)	RPD
Gradient boosting machine	0.57	8	0.3	5	1.2
Multivariate adaptive regression spline	0.38	9	0.2	4	1.1
Random forest	0.60	8	0.1	5.6	1.2
Support vector machine	0.50	8.6	2	4.4	1.1
Multiple linear regression	0.31	9.5	2.64	4	1.0
Regression Kriging	0.58	8	0.65	6.6	1.2
Ensemble machine learning	0.63	7.5	0.4	4.2	1.8

KEY FINDINGS OF COMPARING REGRESSION KRIGING WITH MACHINE LEARNING APPROACHES

- Different prediction techniques inferred different importance and used different number of environmental predictors for SOC stocks.

- Regression kriging approach produced lower prediction errors in comparison to MARS and SVM, and comparable prediction accuracy with GBM and RF techniques.

- The ensemble median prediction of SOC stocks obtained from all four machine learning techniques showed highest prediction accuracy.

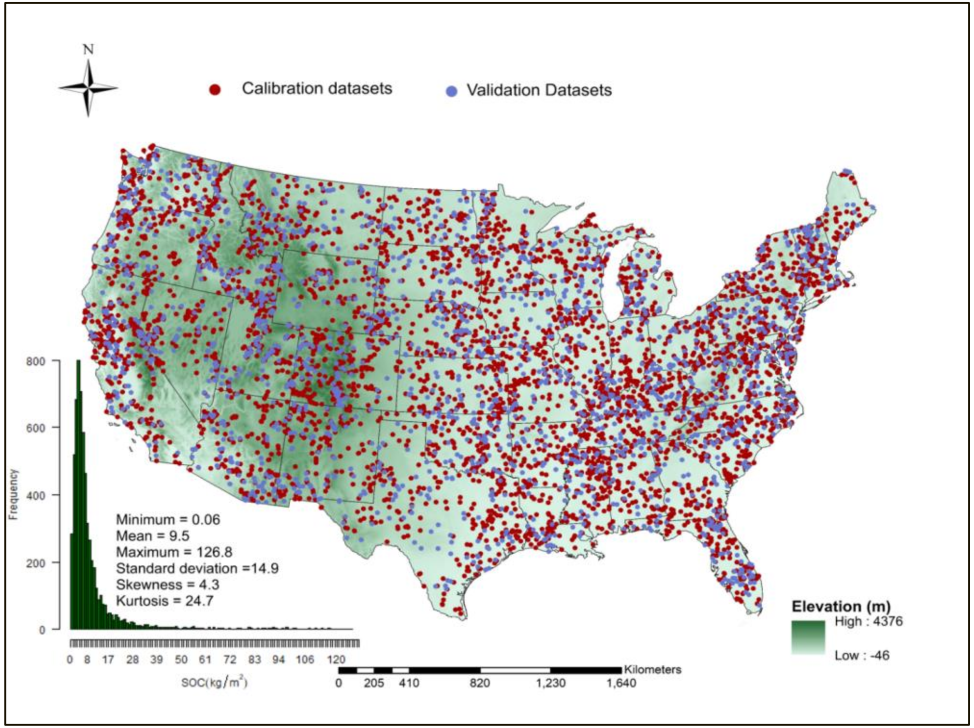
PREDICTING DECADAL SOC CHANGE: COMPARISON OF MACHINE LEARNING MODELS WITH CMIP6 MODEL PROJECTIONS

PREDICTING DECADAL SOC CHANGE: COMPARISON OF MACHINE LEARNING MODELS WITH CMIP6 MODEL PROJECTIONS

- Recent results are suggesting ensemble mean predictions of ML techniques are providing more realistic results for both baseline and SOC change predictions.

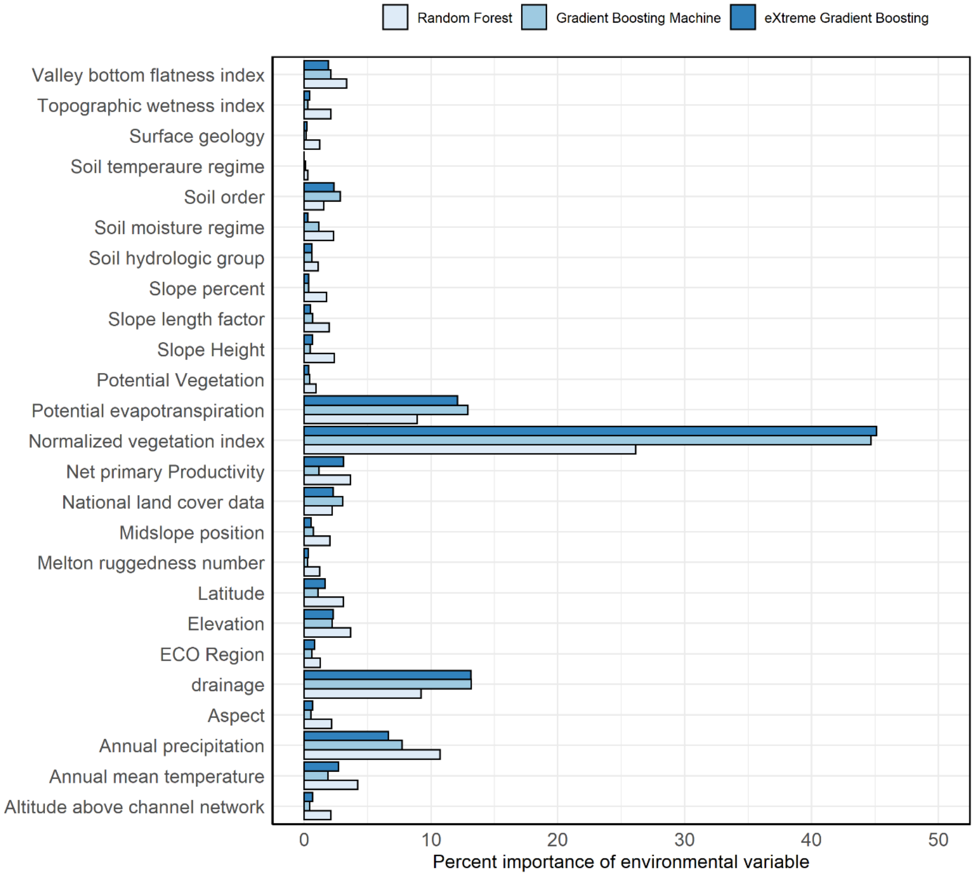
- We compared ensemble ML predictions (RF, GBM, and XGB) of baseline and decadal SOC change with results of recently available CMIP6 ESM projections.

- 100 m spatial resolution for SSP2 4.5 w m^{-2} and SSP5 8.5 w m^{-2} scenarios.

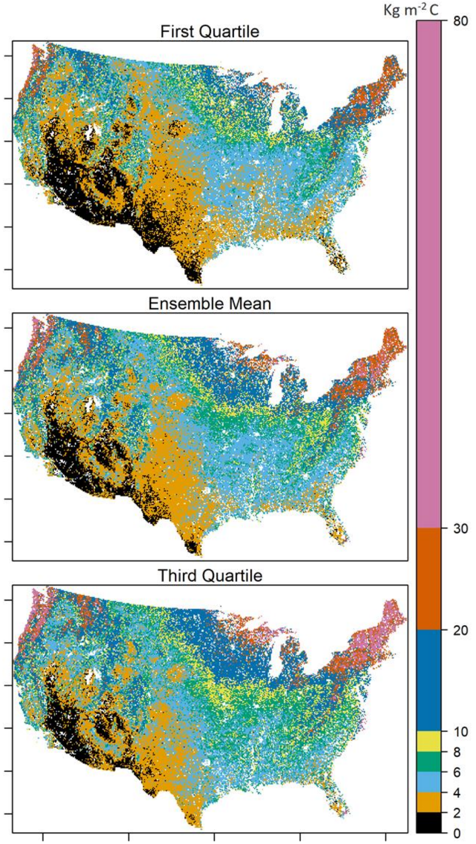


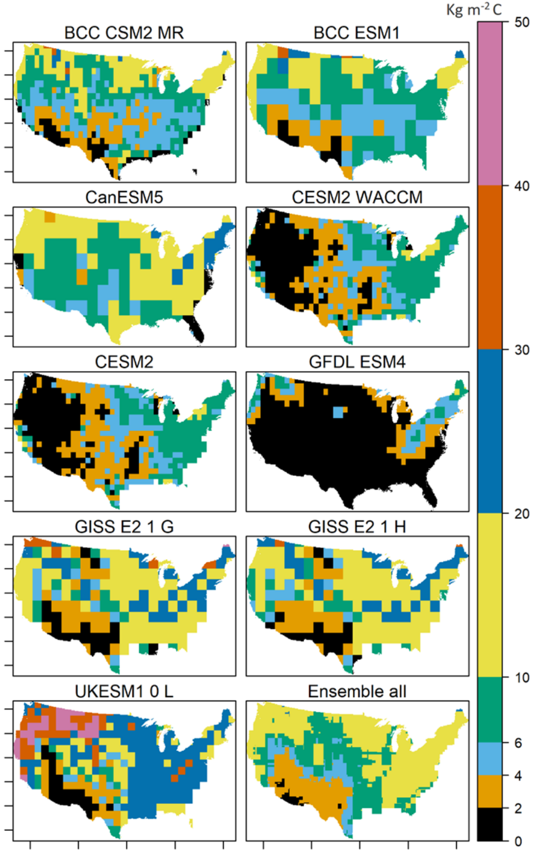
IMPORTANT ENVIRONMENTAL CONTROLLERS OF CONTINENTAL US SURFACE SOIL ORGANIC CARBON STOCKS

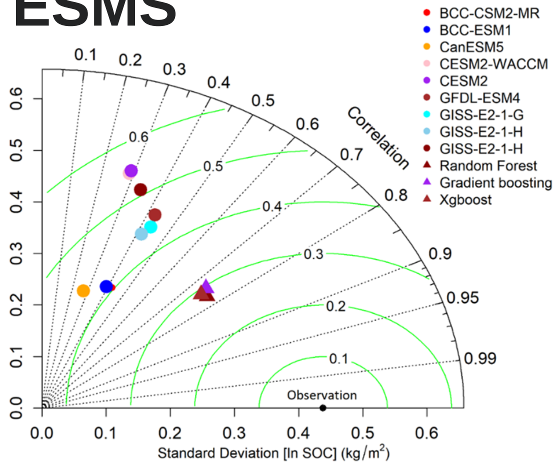
- Out of 32 environmental factors we evaluated different ML approaches used 25 environmental factors.
- Normalized Difference Vegetation Index, potential evapotranspiration, drainage condition and annual precipitation were most important predictors of surface SOC stocks.
- Other important environmental controllers of SOC stocks were temperature, elevation, and soil order.

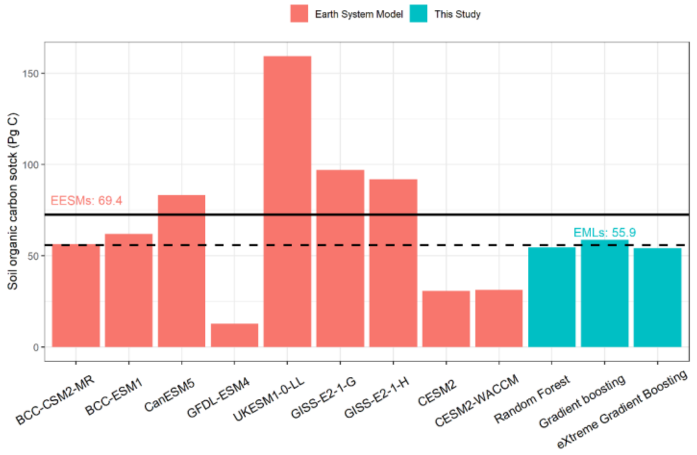


BASELINE CONTINENTAL US SURFACE SOC STOCKS: ML PREDICTIONS IN COMPARISON TO CMIP6 ESMS



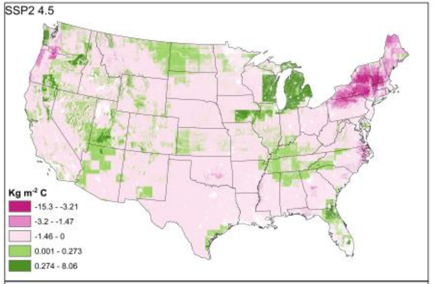


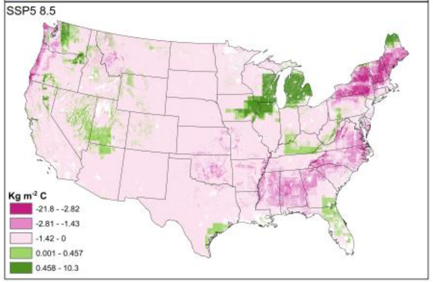




PROJECTED SPATIAL PATTERNS OF SURFACE SOC CHANGE (PG C) IN CONTINENTAL US BY 2100

Ensemble ML predictions

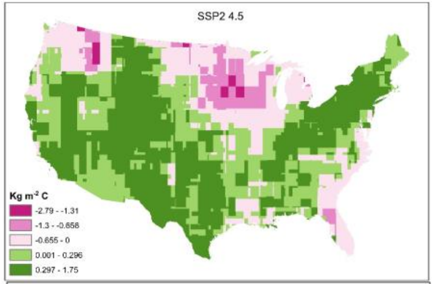


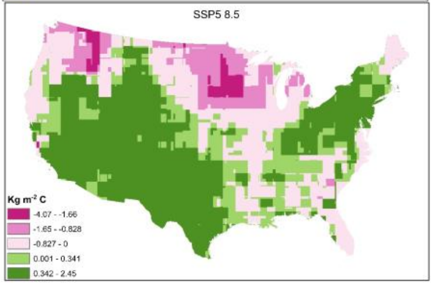


Land cover types	ML		ESM	
	SSP2 4.5 w m^{-2}	SSP5 8.5 w m^{-2}	SSP2 4.5 w m^{-2}	SSP5 8.5 w m^{-2}
Forest	- 0.97	- 1.53	2.9×10^{-3}	-4×10^{-4}
Croplands	- 0.21	- 0.53	-1.1x10-3	-3.6x10-3
Wetlands	-6.8×10^{-2}	- 0.10	-2×10^{-4}	-6×10^{-4}
Other (Pasture + herbaceous)	- 0.56	- 1.28	8.3×10^{-3}	8.8×10^{-3}

Negative sign show SOC loss and positive sign show SOC sequestration

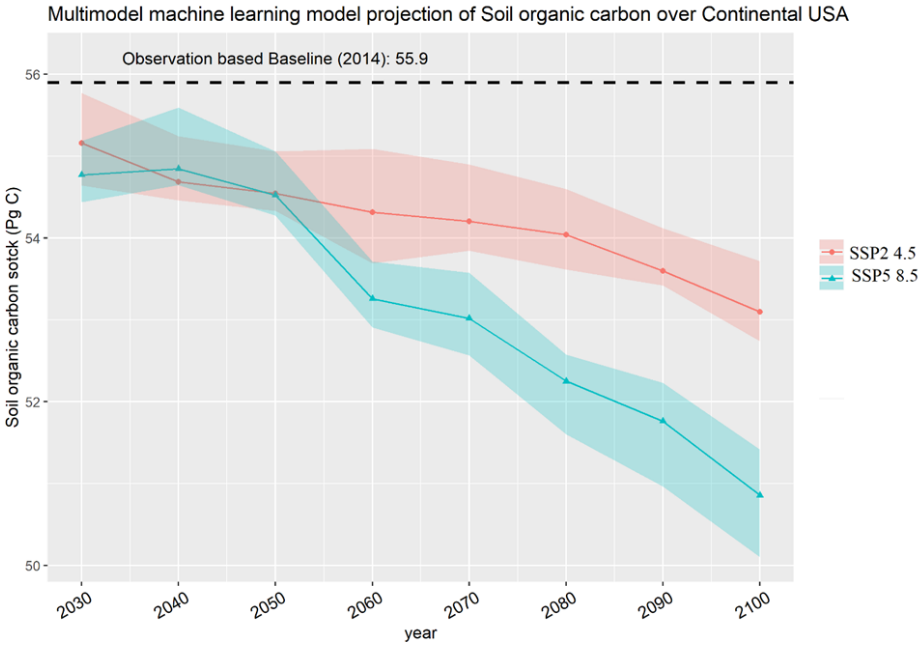
Ensemble ESM predictions

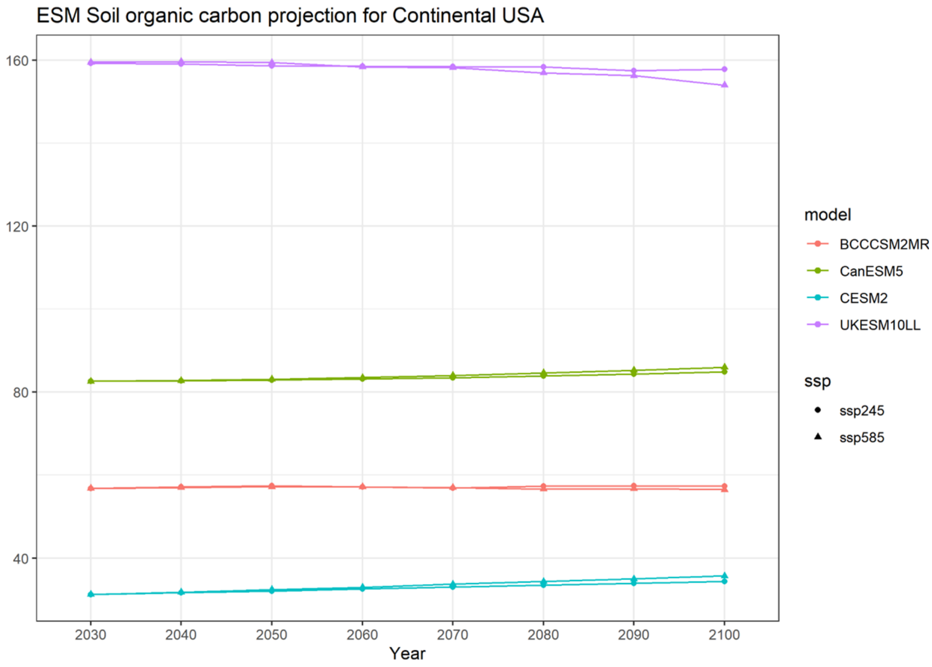




- ML approaches are showing SOC loss under both scenarios, with higher SOC losses under higher emissions.
- ESMs are showing mixed results of SOC change.
- Both types of models are consistently showing SOC loss from croplands and wetlands.

PROJECTED DECADAL SOC CHANGES IN US SURFACE SOILS





- ML approaches are not in agreement with ESMs in predicting decadal and total changes in continental US surface SOC stocks.
- ESM predictions differ in orders of magnitude and show different sign of change.

KEY FINDINGS OF SOIL ORGANIC CARBON CHANGE STUDY

- Baseline representation of continental US surface SOC stocks in CMIP6 ESMs are not consistent with observations. This disagreement could be due to absence of important environmental predictors in current ESMs.

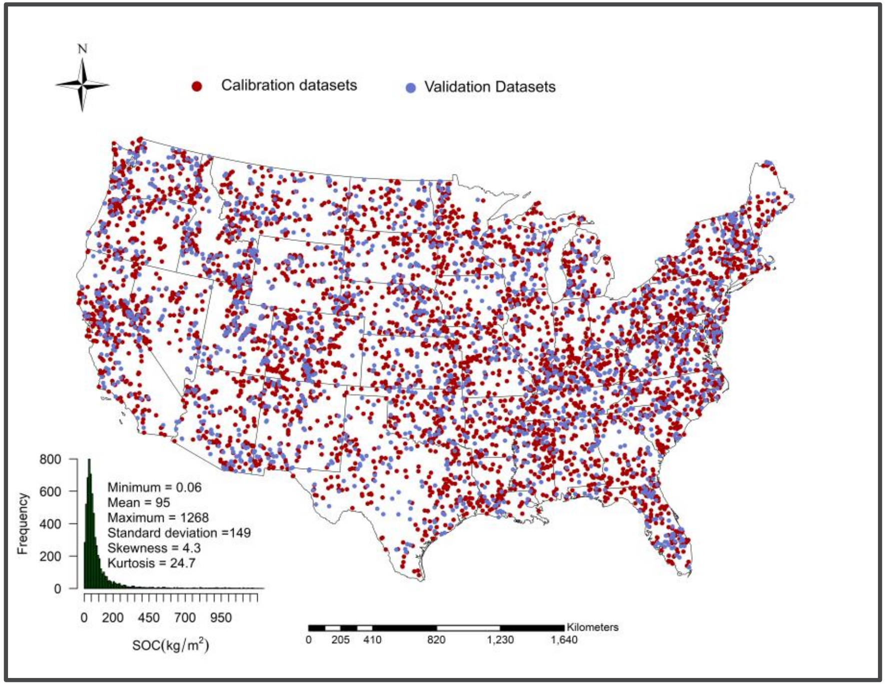
- Ensemble ML approach predicts SOC loss under both moderate (2.1 Pg C) and high emission scenarios (3.9 Pg C). In contrast, ESMs predict both SOC sequestration and loss over continental US.

- Ensemble ML approach predicts larger changes in SOC stocks in comparison to ESMs, but both ML and ESMs are consistently predicting SOC loss from croplands and wetlands.

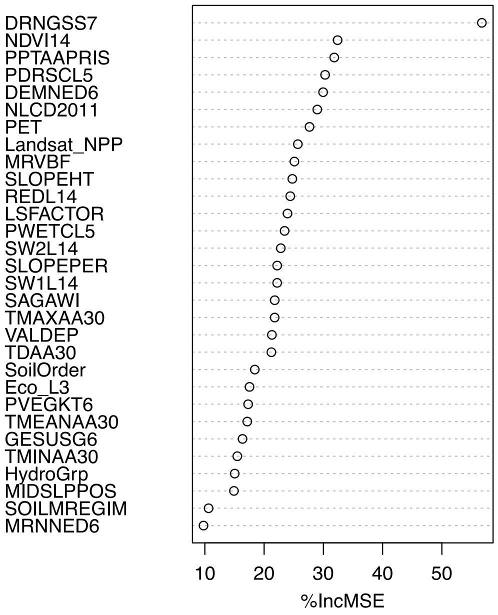
DERIVING FUNCTIONAL RELATIONSHIPS OF ENVIRONMENTAL CONTROLLERS OF SOC STOCKS

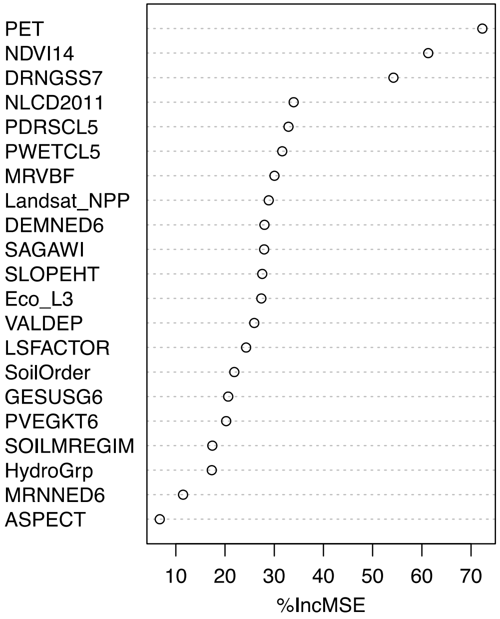
FUNCTIONAL RELATIONSHIPS BETWEEN ENVIRONMENTAL PREDICTORS AND SOIL ORGANIC CARBON STOCKS

- We need better model benchmarks which could reduce the disagreement between SOC observations and their model representations.
- We used ~6300 recently available SOC stock observations and 32 environmental covariates representing different soil-forming factors.
- We combined Random Forest with generalized additive models to develop functional relationships of important environmental controllers.



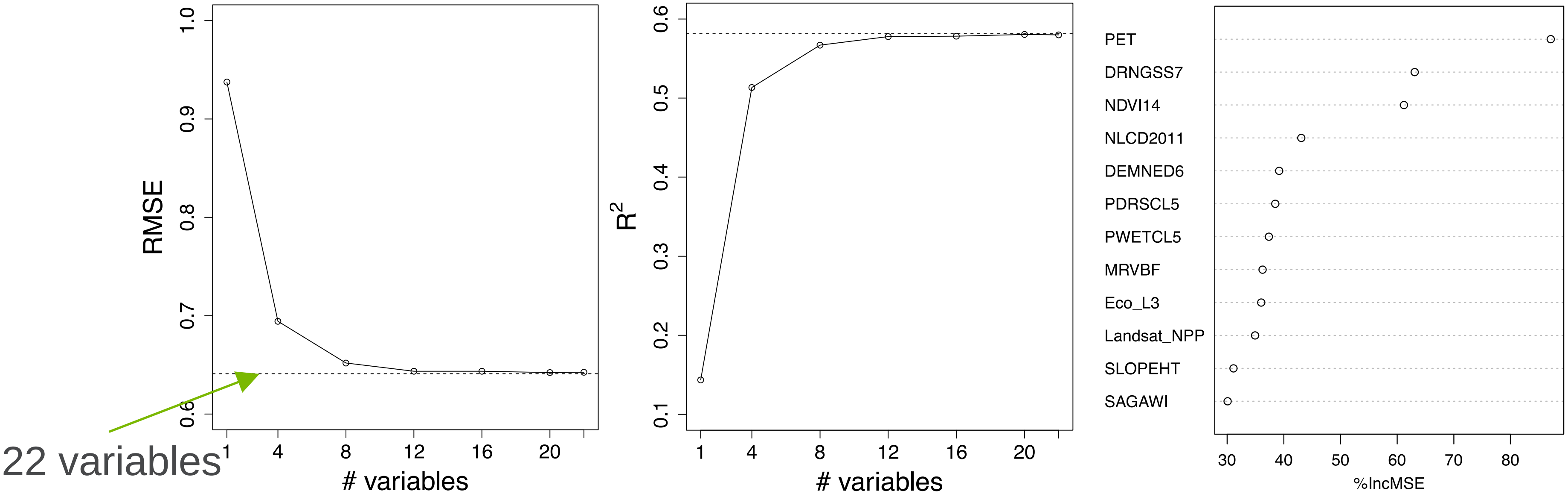
IMPORTANT ENVIRONMENTAL CONTROLLERS OF CONTINENTAL US SURFACE SOIL ORGANIC CARBON STOCKS





- First, we used all 32 environmental factors in random forest to predict SOC stocks.
- We removed correlated variables (r=0.7) and identified 22 environmental factors.

RANDOM FOREST: NUMBER OF VARIABLES VS PREDICTION ACCURACY



- With additional number of variables prediction accuracy increased, but after 12 variables improvement in prediction accuracy was minimal.

GENERALIZED ADDITIVE MODELS

Find polynomial functions to fit the target variable

$$E[Y] = \sum_{i=1}^{N} f_i(x_i) + C$$

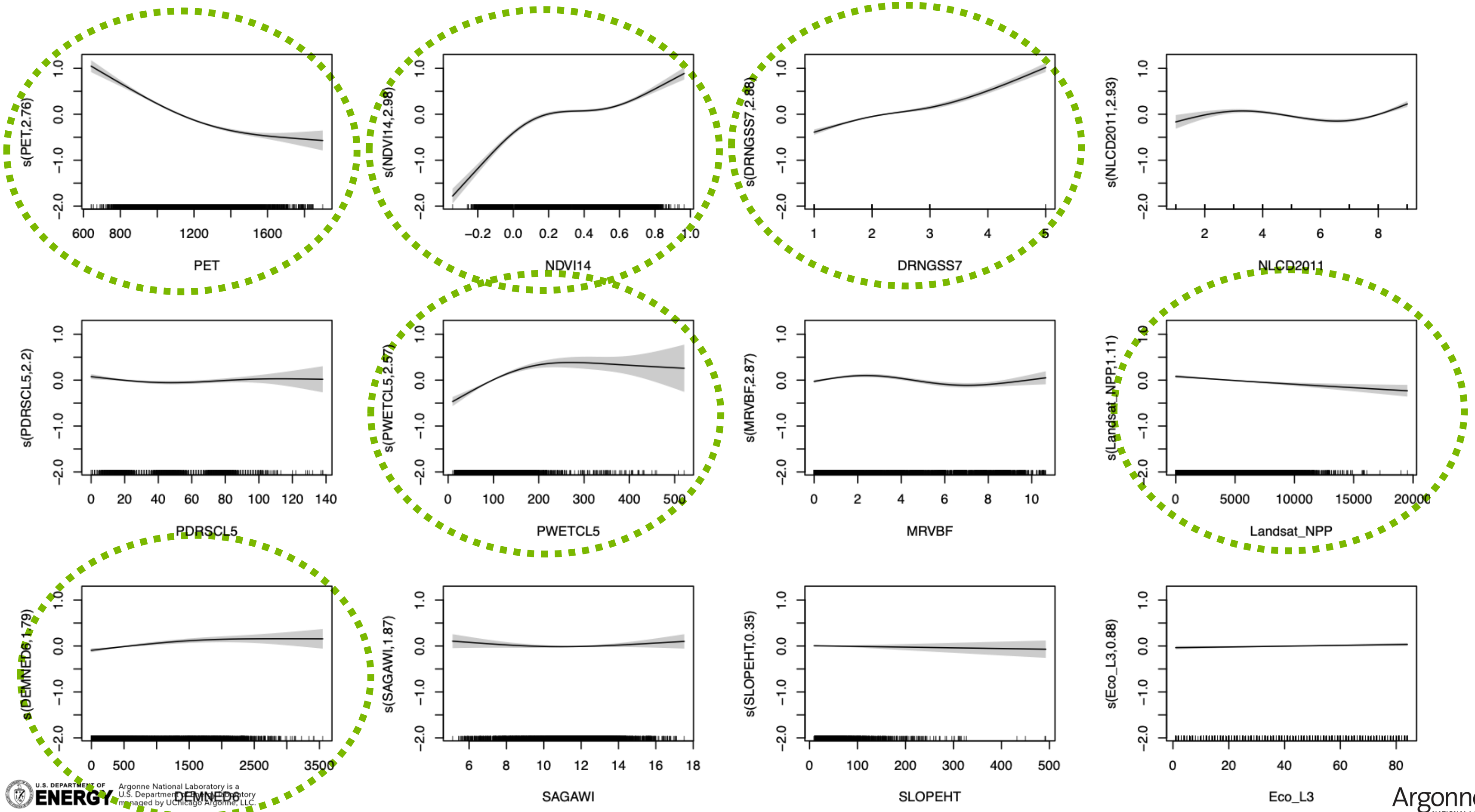
$f_i(x)$ is usually a spline

Only 12 variables identified by the random forest
are used.

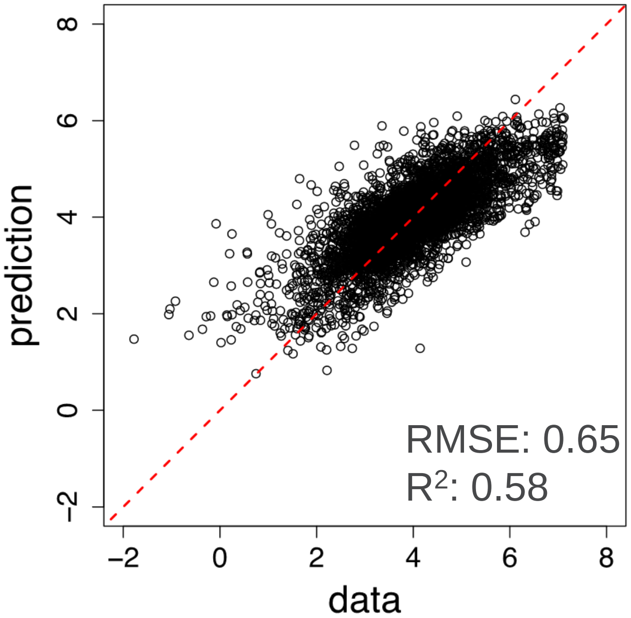
- We kept 11 variables at median value and then changed a test variable from minimum to maximum, and plotted test variable vs SOC stock.

- Fitted a non-linear function that captured the response surface.

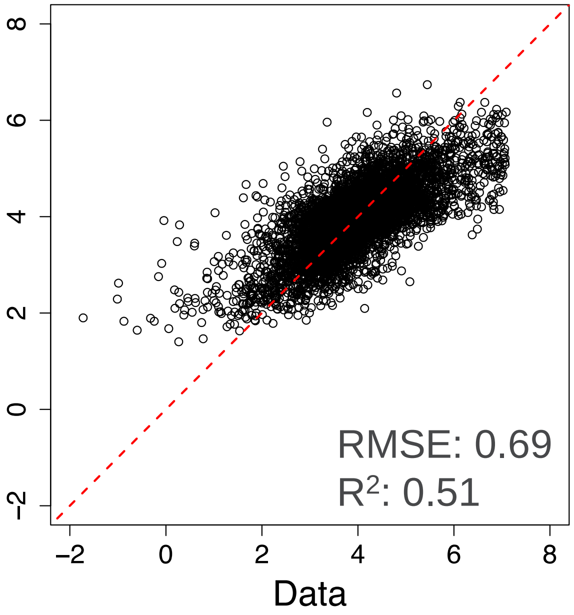
RESPONSE SURFACES OF 12 ENVIRONMENTAL FACTORS



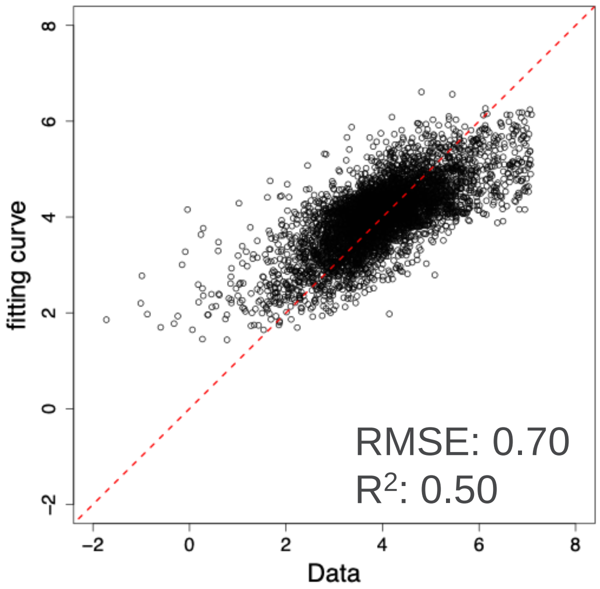
PREDICTION ACCURACY USING FUNCTIONAL RELATIONS OF 6 ENVIRONMENTAL PREDICTORS IN COMPARISON TO RANDOM FOREST



Random forest using all 32 variables



Random forest using 22 variables



Using functional relations of 6 variables

KEY FINDINGS FROM DEVELOPING FUNCTIONAL RELATIONSHIPS BETWEEN ENVIRONMENTAL FACTORS AND SOC STOCKS

- Using random forest we can identify important environmental predictors of SOC stocks.

- Response surface of environmental factors on SOC stocks can be derived using generalized additive models.

- Derived non-linear response surfaces produced similar prediction accuracy as of the random forest in predicting surface SOC stocks of continental USA.

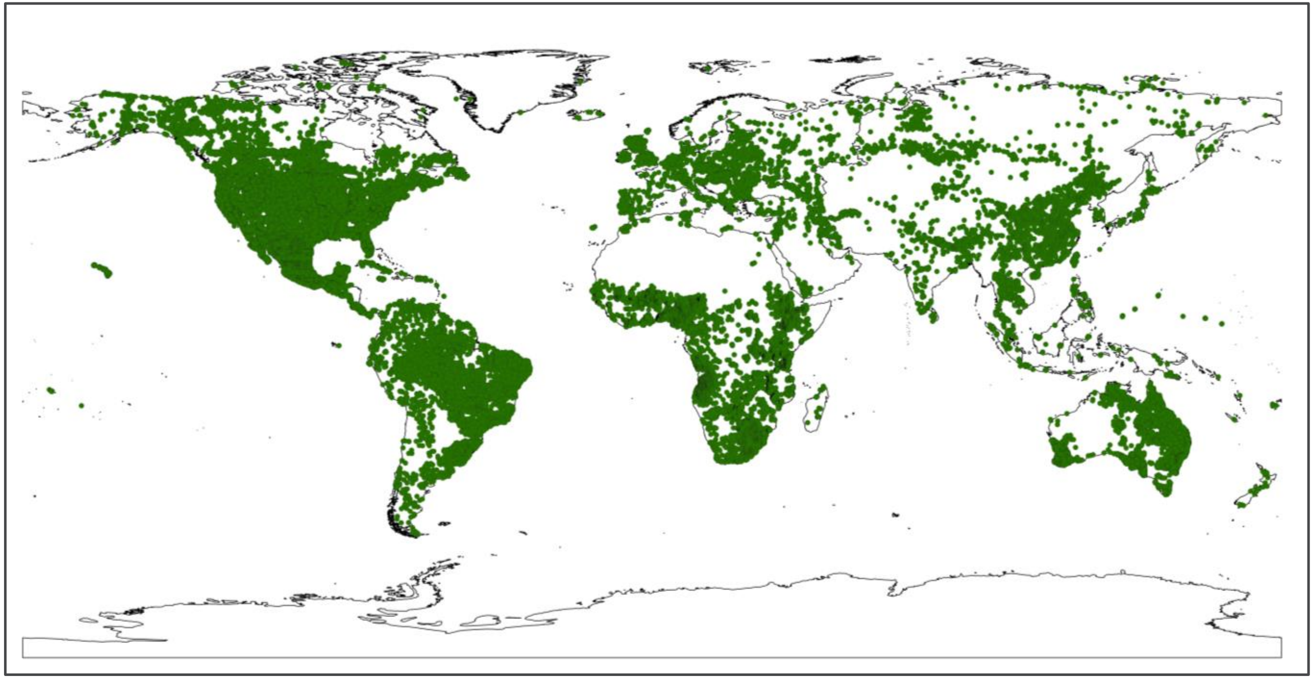
SUMMARY

- THE ENSEMBLE MEDIAN PREDICTION PROVIDES GREATER SPATIAL DETAILS AND PRODUCES HIGHER PREDICTION ACCURACY, AND THUS CAN BE A BETTER CHOICE FOR PREDICTING SPATIAL HETEROGENEITY OF SOIL PROPERTIES.

- ENSEMBLE MACHINE LEARNING APPROACH PREDICTS MORE REALISTIC DECADAL CHANGES IN SOIL ORGANIC CARBON STOCKS OF CONTINENTAL US IN COMPARISON TO 4 CMIP6 ESMS.

- BY COMBINING MACHINE LEARNING WITH GENERALIZED ADDITIVE MODELING FUNCTIONAL RELATIONSHIPS BETWEEN ENVIRONMENTAL FACTORS AND SOC STOCKS CAN BE DEVELOPED, WHICH MAY SERVE AS POTENTIAL LAND MODEL BENCHMARKS.

LARGE DATASETS FOR GLOBAL STUDIES



We have acquired ~114,000 soil profile data and 30 environmental covariates from various sources, and plan to conduct SOC storage and dynamics studies at global scale.

Acknowledgements

U.S. DEPARTMENT OF ENERGY | Office of Science

THANK YOU FOR YOUR TIME AND ATTENTION!