BIOAg Project Final Report

Report Type:
Final

Title:
Scalable assessment of soil organic carbon for carbon incentive programs

Principal Investigator(s) and Cooperator(s):
PI
Kirti Rajagopalan (Department of Biological Systems Engineering)

Co-PI(s)
Jana Doppa (School of Electrical, Electronic and Computer Science)
Deirdre Griffin-LaHue (Department of Crop and Soil Sciences)
Dani Gelardi (WA Department of Agriculture)
Jordan Jobe (School of Electrical, Electronic and Computer Science)

Cooperator(s)
Georgine Yorgey (Center of Sustaining Agriculture and Natural Resources)

Abstract:
Incentive programs aimed at promoting climate mitigation and soil health often rely on practice-based payments and coarse estimates of benefits, rather than directly quantifying soil carbon accrual. Transitioning toward outcome-based incentives—aligned with principles of true cost accounting—requires scalable, reliable methods for measuring soil organic carbon (SOC) and associated uncertainties. In this project, an interdisciplinary team focused on dryland agricultural systems in the Pacific Northwest to address this challenge. We developed and evaluated a machine learning framework for predicting SOC in the 0–30 cm soil profile by integrating multispectral satellite imagery with environmental and management covariates. Given the small-data context, we further implemented uncertainty quantification methods to generate instance-level prediction intervals and to guide efficient data collection. Results demonstrate that SOC can be predicted with meaningful accuracy using remotely sensed and ancillary data, and that uncertainty-informed sampling strategies can substantially reduce data requirements while maintaining model performance. These findings provide critical preliminary evidence supporting scalable SOC monitoring approaches and directly inform ongoing efforts to develop outcome-based carbon incentive programs.

Project Description:
There is growing global interest among government agencies, non-profit organizations, and industry groups in providing direct financial incentives to farmers for increasing soil organic carbon (SOC) storage on agricultural lands. Soils represent the largest terrestrial carbon pool, and even a 10% increase in this pool has the potential to offset approximately 30 years of anthropogenic greenhouse gas emissions. Despite the emergence of carbon markets and a substantial $3.1 billion investment in the USDA Climate-Smart Commodities Program in 2022, a critical gap persists between scientific understanding and practical implementation in agricultural carbon farming. Current programs often incentivize practice adoption rather than measurable outcomes and lack robust approaches to quantify actual SOC accrual

and associated uncertainties. This challenge is fundamentally one of measurement, monitoring, reporting, and verification (MMRV).

Addressing this gap requires advances at the intersection of agricultural and data sciences. We assembled an interdisciplinary team spanning multiple institutions, with support from growers, commodity groups, conservation districts, and regional and national stakeholders and submitted a proposal to the USDA NIFA Data Science for Food and Agricultural Systems program, using dryland systems in the Pacific Northwest (PNW) as a case study. While the proposal was well received, reviewers raised key concerns regarding (1) the feasibility of using satellite imagery to estimate SOC beyond the near-surface layer, particularly within the top 30 cm, and (2) the challenges of training robust machine learning (ML) models in small-data contexts.

The primary purpose of this seed project was to directly address these reviewer concerns and strengthen a competitive resubmission for extramural funding. Specifically, we aimed to develop and evaluate a prototype ML framework for predicting SOC stocks in the top 30 cm of soil by integrating multispectral satellite observations with environmental and management covariates, using data from the Washington Soil Health Initiative. The project was designed to test two key hypotheses: (a) that SOC in the 0–30 cm soil profile can be accurately predicted using satellite-derived and ancillary data, and (b) that incorporating uncertainty quantification into ML models can guide targeted data collection strategies in resource-constrained settings. By resolving these uncertainties, this work provides critical preliminary evidence to support scalable SOC monitoring approaches and positions the team for a stronger, more competitive extramural proposal.

Outputs:

Overview of Work Completed

This project successfully developed a prototype ML framework for predicting surface SOC using multispectral satellite imagery and environmental covariates. The model achieved reasonable predictive accuracy, directly addressing prior reviewer concerns regarding the feasibility of estimating SOC from remotely sensed data. These results provide preliminary evidence that SOC exhibits sufficient predictability from satellite-derived signals, strengthening the technical foundation for future extramural proposals.

In addition, we implemented a prototype uncertainty quantification framework based on conformal prediction, enabling statistically rigorous (“provable”) and instance-specific uncertainty estimates for ML-based remote sensing applications. This represents a significant advancement, as most existing remote sensing models do not provide uncertainty estimates at the individual prediction level. Incorporating this capability is critical for stakeholder trust and for enabling decision-making in carbon MMRV systems.

Building on this, we evaluated an intelligent data collection framework that leverages model uncertainty to guide new sampling efforts. Using active learning experiments, we demonstrated that model performance comparable to that achieved with the full training dataset (based on conventional convenience sampling) could be attained with approximately half the number of samples when data collection is guided by uncertainty. This highlights the potential for substantial cost savings and improved efficiency in resource-constrained data collection contexts.

Finally, we developed educational materials to support training of undergraduate and high school students in the use of Google Earth Engine for agricultural applications. These materials are designed to lower barriers to entry for students—particularly those in agricultural disciplines—who may not traditionally engage with data science and remote sensing tools.

Methods, Results, and Discussion

Methods

***Data and Study Region*:** The SOC prediction model was developed using 513 observations of soil organic carbon (0–30 cm depth) from dryland agricultural systems in eastern Washington, collected through the Washington Soil Health Initiative.

***Environmental, Management, and Remote Sensing Covariates*:** A comprehensive set of environmental and management variables was assembled. These included daily weather data (4 km resolution, gridMET, 1979–present), soil properties (30 m), and topographic variables derived from a 30 m Digital Elevation Model. Management information—including crop type, rotation—was obtained from WSDA and USDA Crop Data Layers. Tillage information was added based on model predictions from prior work. Multispectral satellite data were compiled from Sentinel-1 (radar), Sentinel-2 (optical), and Landsat (optical and thermal), including both raw spectral bands and a suite of vegetation, soil moisture, and tillage-related indices identified in prior literature.

***Data Integration and Unit of Analysis*:** All datasets were harmonized to a common unit of analysis defined by field polygons. Variables at finer spatial resolution (e.g., soils, terrain, satellite data) were aggregated using summary statistics such as mean and distributional measures, while coarser-resolution data (e.g., weather) were assigned uniformly to fields.

***Feature Engineering and Selection*:** The dataset was randomly partitioned into 80% training and 20% independent test sets. Within the training data, SOC outliers were removed using an interquartile range criterion. Feature engineering steps were applied to capture nonlinearities and interactions, including square-root and squared transformations, pairwise interactions among the most SOC-correlated predictors, and sample-level summary statistics (mean, standard deviation, minimum, maximum, and range).

To reduce dimensionality and collinearity, a multi-stage feature selection process was implemented. Features were first ranked by correlation with SOC, followed by removal of highly collinear variables. Mutual information regression was then used to retain the most informative predictors, and Random Forest-based importance scores were used to identify the final set of features. These features were transformed using a Yeo-Johnson power transformation and standardized using z-score normalization.

***Machine Learning Models and Evaluation*:** Multiple tree-based machine learning models—including Random Forest, Gradient Boosting, XGBoost, LightGBM, and Extra Trees—were trained to establish predictive performance. To explicitly account for spatial structure, additional models included inverse-distance-weighted k-nearest neighbors (KNN) and a geographically adjusted Gaussian mixture model. Model performance was evaluated using R^2, RMSE, and MAE on both training and independent test datasets to assess generalization.

***Uncertainty Quantification*:** Prediction uncertainty was quantified using a split conformal prediction framework. The training data were further divided into model-training and calibration subsets. A hybrid ensemble model combining spatial KNN and Extra Trees was used as the base predictor. Prediction intervals were derived from calibration residuals, yielding statistically valid, distribution-free, instance-level uncertainty estimates under the exchangeability assumption.

***Intelligent Data Sampling*:** To optimize data collection, an active learning framework was implemented. A Query-by-Committee (QBC) approach, consisting of an ensemble of models with varying random seeds, was used to quantify uncertainty based on prediction disagreement. High-uncertainty samples were identified and further refined using a Determinantal Point Process (DPP), which promotes diversity

across both geographic and feature space. This hybrid uncertainty–diversity strategy enabled efficient selection of new samples, maximizing information gain under limited data collection budgets.

Results

Comparative Model Performance: Ten machine learning models were evaluated for their ability to predict SOC, with performance assessed using R^2, RMSE, and MAE on both training and independent test datasets. Overall, results indicate strong predictive capability, particularly for hybrid models that integrate spatial and feature-based learning. The Ensemble-KNN-k5-ET-80-19 model achieved the best overall performance, with a test R^2 of 0.724, RMSE of 3.26 g/kg, and MAE of 2.36 g/kg, demonstrating superior generalization relative to all other approaches. Notably, ensemble variants that combined spatial KNN with tree-based models consistently outperformed standalone machine learning methods, with test R^2 values ranging from 0.713 to 0.724 (Figure 1). This highlights the importance of incorporating spatial structure alongside high-dimensional feature information for SOC prediction. Among standalone models, Extra Trees achieved the strongest performance (test R^2 = 0.661), followed closely by Random Forest (R^2 = 0.659) and Gradient Boosting (R^2 = 0.650), although Gradient Boosting exhibited comparatively higher RMSE and MAE. In contrast, the spatially driven GAGMM models showed substantially weaker performance (test R^2 ≈ 0.406), indicating that spatial information alone is insufficient for accurate SOC estimation without integration of environmental and remote sensing covariates.

Training performance was consistently high across models, with R^2 values exceeding 0.93 for all tree-based methods and approaching near-perfect fits (>0.99) for ensemble models (Figure 1). This suggests the presence of overfitting in the current prototype, particularly for more complex ensembles. However, this is expected in a small-data context and can be addressed in future work through expanded datasets, improved regularization, and more robust validation strategies.

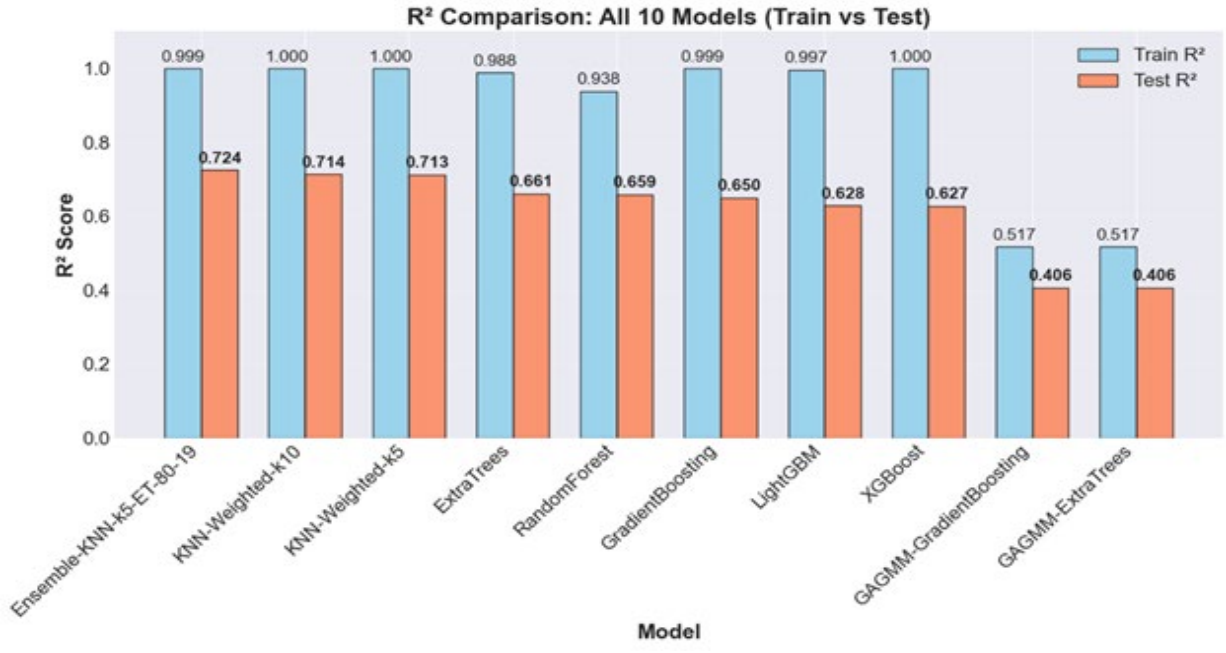

Figure 1. The R^2 score of the models for training and test data.

For further analysis we use only the best-performing Ensemble-KNN-k5-ET-80-19 model having an RMSE of 3.259 g/kg and MAE of 2.361 g/kg on the test set, with an R^2 of 0.7236. The scatter plot of predicted

versus actual SOC values revealed a strong agreement along the 1:1 line for most samples, with most predictions falling within ±5 g/kg of actual values (Figure 2). Error magnitude, represented by absolute error in the color bar, shows that larger discrepancies occurred primarily at higher SOC concentrations (>20 g/kg), where the model tends to underpredict.

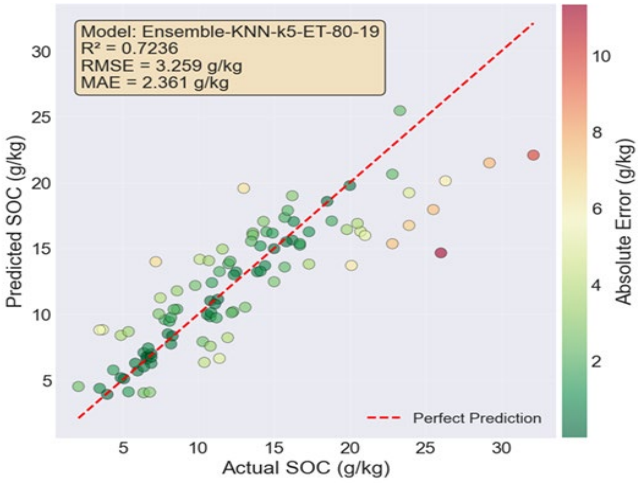

Figure 2. Scatter plot of predicted vs actual SOC for the top performing model plotted using test data. The color bar indicates absolute error.

Active Learning Performance: The QBC-DPP demonstrates rapid model improvement as labeled samples increased (Figure 3). Starting from an initial training set of only 20 labeled samples, the R^2 score improved from approximately 0.38 to 0.72 after incorporating 200 labeled points, with diminishing returns observed beyond 300 samples. RMSE also decreased sharply from approximately 4.8 g/kg to 3.3 g/kg within the first 150 labeled samples, then stabilized around 3.25 g/kg. Similarly, MAE declined from 3.6 g/kg to approximately 2.3 g/kg, stabilizing after 300 labeled samples. These results indicate that the intelligent sampling strategy achieved near-optimal performance using only about half of the available training data, suggesting significant potential for cost reduction in the field data collection.

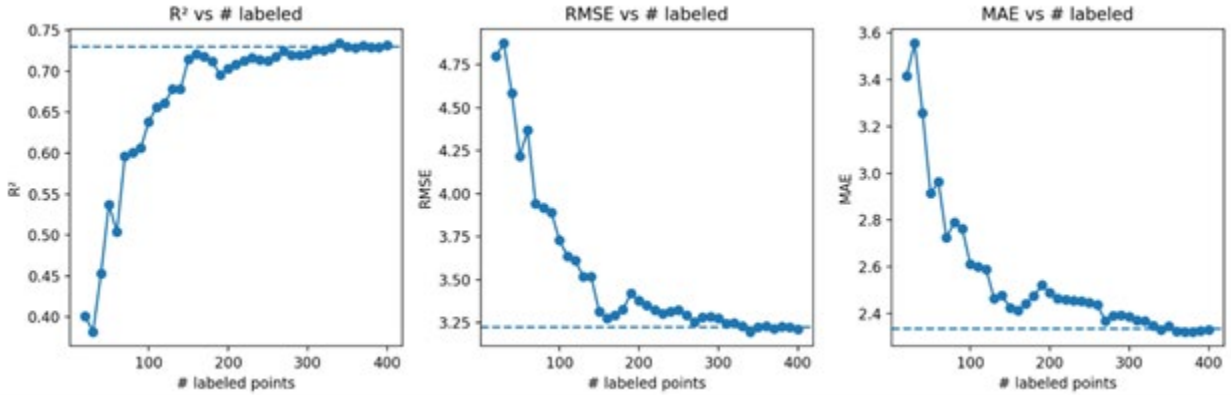

Figure 3. The active learning curve for intelligent data sampling using QBC-DPP method, plotted for R^2 score, RMSE and MAE.

Discussion

Environmental and Management Controls on SOC: Correlation analyses and feature engineering (results not shown here to keep the report short) provided insight into the environmental and management drivers of SOC in eastern Washington. Positive correlations with clay content and negative correlations with sand content are consistent with established pedological understanding that fine-textured soils protect organic matter through aggregation and mineral–organic associations. The SOC also showed a moderate positive association with mean annual precipitation and a weak negative relationship with mean annual temperature, supporting the expectation that cooler, wetter conditions enhance carbon accumulation through increased biomass inputs and reduced decomposition rates.

Crop- and county-level patterns further reinforce the role of management. Perennial or residue-rich systems such as winter wheat and grass/pasture exhibited higher SOC relative to fallow or intensively managed annual systems. These findings align with broader evidence that continuous ground cover, residue retention, and reduced tillage contribute to improved soil health and carbon storage.

The strong seasonal signal in the Normalized Difference Tillage Index (NDTI) highlights the importance of management timing and residue dynamics. Higher summer NDTI values—and their stronger association with SOC—suggest that fields maintaining residue or crop cover tend to retain more carbon. Weaker relationships in winter and spring likely reflect reduced optical sensitivity under snow cover and variable soil moisture conditions, underscoring the importance of multi-season and multi-sensor (including radar/SAR) observations for robust SOC estimation.

Interpretation of Model Performance*:* Across model classes, ensemble approaches that combine spatial and feature-based learning performed best. In particular, the hybrid KNN–Extra Trees framework achieved strong predictive accuracy (test $R^2 \approx 0.72$, RMSE ≈ 3.26 g/kg, MAE ≈ 2.36 g/kg). While there remains room for improvement, these results demonstrate that SOC variability in eastern Washington can be effectively captured using a combination of remote sensing, environmental covariates, and spatial structure. The improved performance of ensemble models relative to standalone approaches highlights the value of explicitly incorporating spatial autocorrelation alongside high-dimensional predictor information. This is particularly important in heterogeneous agricultural landscapes, where both local spatial context and broader environmental gradients influence SOC patterns.

Value of Active Learning for Soil Sampling*:* Active learning experiments demonstrate that intelligently selected samples can substantially reduce the data requirements for accurate SOC prediction. Near-optimal model performance was achieved with approximately 250–300 labeled samples, compared to the full dataset of over 400 samples used in conventional training. This indicates that random or convenience sampling is likely inefficient in this context. The observed pattern of rapid performance gains during early sampling rounds followed by diminishing returns is consistent with effective uncertainty-based sampling strategies. These findings have clear operational implications: targeted sampling guided by model uncertainty can significantly reduce labor, time, and cost associated with soil data collection while maintaining predictive accuracy. This is particularly important for large-scale SOC monitoring efforts, where field sampling remains a major bottleneck.

Publications, Handouts, Other Text & Web Products:

Presentations:

Dubey N., Norouzi Kandelati, A., Gharsallaoui MA., Doppa JR. & Rajagopalan, K. *Uncertainty-Aware Surface Soil Organic Carbon Mapping in Agricultural Lands.* American Geophysical Union Fall Meeting, Washington DC, Dec 14-20, 2025.

Norouzi Kandelati, A., Gharsallaoui MA., Dubey N., Doppa JR. & Rajagopalan, K. *Uncertainty-Aware Tillage Mapping in Data-limited Agroecosystems.* American Geophysical Union Fall Meeting, Washington DC, Dec 14-20, 2025.

Journal papers:

Norouzi Kandelati, A., Gharsallaoui MA., Dubey N., Doppa JR. & Rajagopalan, K. Uncertainty-Aware Remote Sensing Predictions in Data-limited Agroecosystems. Remote Sensing of Environment. Ready for submission (Expected submission date: April 2026).

Dubey N., Norouzi Kandelati, A., Gharsallaoui MA., Doppa JR. & Rajagopalan, K. Uncertainty-aware surface soil organic carbon mapping in agricultural lands. Computers and Electronics in Agriculture. In preparation (Expected submission date: June 2026).

Models:

The ML models and uncertainty quantification wrappers produced as part of this work are available as open-source software on GitHub for public use. The URL links to open-source code will be available in the journal publications.

Outreach & Education Activities:

Graduate student Amin Norouzi Kandelati delivered two hands-on Google Earth Engine (GEE) training workshops for undergraduate students at Heritage University (10 participants) and Wenatchee Valley College (10 participants) in Fall 2025 in collaboration with Jessica Black from Heritage University and Sai Ramaswami from Wenatchee Valley College. The workshop successfully demystified cloud-based Earth Systems monitoring and ML applications for a diverse group of students, many of whom were first generation college students with little to no prior experience in coding or GEE. By moving away from intimidating technical jargon and focusing on hands-on discovery, the hands-on session transformed GEE from a complex tool into an accessible resource for problem-solving. Students moved from simply viewing maps to actively interacting with satellite data, gaining the confidence to ask—and answer—environmental questions using imagery. This addressed a common perception among students in agricultural domains that they lack the skills to engage with AI and machine learning applications, and provided the 20 participants with confidence to pursue these technologies in their programs and careers. The training material developed as part of this effort has been curated for reuse and broader dissemination as part of other projects.

Impacts:

Short-Term:

- Two graduate students from agricultural and data science backgrounds gained hands-on experience in interdisciplinary research and collaboration.
- A postdoctoral researcher received training in grant development, and contributed to extramural proposals.
- The feasibility of predicting SOC in the top 30 cm of soil using satellite imagery and environmental covariates was established.
- A machine learning pipeline incorporating uncertainty quantification was developed, demonstrating the potential for intelligent, uncertainty-guided data collection that achieves comparable model performance with reduced sampling effort.
- The project generated critical preliminary results that directly address prior reviewer concerns, strengthening the team's competitiveness for extramural funding submissions. Results from this work were incorporated into extramural proposal submissions currently under review.

- Undergraduate and high school students from Wenatchee Valley College and Heritage University were introduced to AI and machine learning applications in agriculture through hands-on workshops, increasing awareness and confidence in engaging with these tools.

Intermediate-Term:

- Too early to report

Long-Term:

- Too early to report

Additional Funding Applied for/Secured:

Rajagopalan, Yorgey et al. *AI-Enabled Tools to Overcome Barriers to Agricultural Natural Climate Solutions.* Allen Family Philanthropy. September 2026-August 2029. $1.425M. Pending review.

Rajagopalan et al. *From Uncertain Data to Trustworthy AI-Driven Agricultural Decisions through Provable Uncertainty Quantification.* USDA NIFA Data Sciences for Food and Agricultural Sciences. September 2026-August 2029. $650K. Pending review.

Graduate Students Funded:

Md. Amine Gharsallaoui, PhD candidate, School of Electrical Engineering and Computer Science, Voiland College of Engineering and Architecture

Amin Norouzi Kandelati, PhD candidate, Department of Biological Systems Engineering, College of Agricultural, Human, and Natural Resources Sciences

Recommendations for Future Research:

While this project demonstrates the feasibility of SOC prediction using integrated remote sensing and environmental data, several key areas warrant further investigation to advance toward operational deployment.

Addressing Overfitting: The high training performance observed across models ($R^2 > 0.93$ and near-perfect fits for ensemble approaches) indicates overfitting in the current prototype setting. Rather than relying solely on expanding the dataset, future work should prioritize more robust model development and evaluation strategies. This includes assessing the degree of extrapolation by analyzing the alignment between training and test data distributions (e.g., using dimensionality reduction techniques such as t-SNE to evaluate coverage of feature space). In addition, repeated train–test splits and ensemble evaluation frameworks can be used to assess the sensitivity of model performance to data partitioning. Voting or aggregation across models trained on different splits can help determine whether predictive performance is stable or dependent on specific training subsets. Together, these approaches will provide a more rigorous understanding of model generalization and reliability in small-data, high-dimensional settings.

Multitask Learning for Joint Prediction: An important new direction will be the development of multitask learning frameworks that jointly predict SOC alongside key management variables such as crop type, rotation, and tillage practices. Because the drivers and covariates influencing SOC and management practices are closely related, joint learning has the potential to improve predictive performance for both tasks. This is particularly valuable given that SOC and management ground truth data are often collected at different locations and resolutions; multitask approaches can help leverage shared information across tasks.

Uncertainty Quantification under Noisy Data: The conformal prediction framework implemented in this study provides valid uncertainty estimates under standard assumptions; however, both machine learning models and uncertainty quantification methods typically assume that input features and target variables are measured without error. In practice, SOC measurements, management records, and remote sensing data all contain varying degrees of noise and uncertainty. Future research should focus on developing ML and uncertainty quantification frameworks that explicitly account for noisy inputs and labels, enabling more realistic and reliable uncertainty estimates for real-world MMRV applications.

Scaling Active Learning for Operational Use: The promising results from active learning highlight the potential for more efficient soil sampling strategies. Future work should evaluate these approaches in real-world field campaigns, including integration with logistical constraints, cost considerations, and stakeholder priorities. Extending this framework to larger spatial domains will be essential for translating these gains into operational SOC monitoring programs.