Transfer and uncertainty for soil spectroscopy: a parsimonious approach using linear mixed models and soil surveys

Eric Potash[1,2*], Colleen Partida[3], Jose Safanelli[3], Elizabeth Forbes[4], Alex Polussa[5,6], Tiffany Runge[4], Matthew Sheffer[4], Jonathan Sanderman[3]

[1] Agroecosystem Sustainability Center, Institute for Sustainability, Energy, and Environment, University of Illinois Urbana-Champaign, Urbana, IL, USA
[2] Department of Natural Resources and Environmental Sciences, College of Agricultural, Consumer and Environmental Sciences, University of Illinois Urbana-Champaign, Urbana, IL, USA
[3] Woodwell Climate Research Center, 149 Woods Hole Rd., Falmouth, 02540 MA, USA
[4] Hudson Carbon, 357 County Route 12, Hudson NY 12534
[5] The Forest School, Yale School of the Environment, 360 Prospect St., Yale University, New Haven, CT 06511, USA
[6] Yale Center for Natural Carbon Capture, Yale University, New Haven, CT 06511, USA

* Corresponding author: epotash@illinois.edu

Abstract

Two challenges that limit the adoption of soil spectroscopy are (1) model transfer to settings with limited data, and (2) uncertainty quantification, especially with correlated errors. Linear mixed models (LMMs) could address these challenges but, despite their success in other fields, LMMs have not to our knowledge been applied in soil spectroscopy. Compared to machine learning transfer algorithms, LMMs are more straightforward and interpretable. LMMs also leverage auxiliary variables, in our case: field, soil depth, and external soil survey (SSURGO) estimates of pH and texture. But instead of using auxiliary variables to rigidly stratify a dataset into groups and fit separate models, LMMs partially pool the entire dataset to balance local and global modeling. In this study, we predict soil carbon from field moist VisNIR spectra of 459 samples of three soil layers (0-15, 15-30, 30-60 cm) in 155 cores across 18 agricultural fields in Northeastern USA (data and code provided). Using cross validation, we find that LMMs succeed in model transfer between fields, outperforming existing approaches such as memory-based learning and Cubist. Moreover, as probability models, LMMs automatically provide excellent uncertainty quantification, including for aggregated predictions where existing methods fail due to correlated errors. We conclude that LMMs show promise as a parsimonious approach for model transfer and uncertainty quantification in soil spectroscopy.

Keywords: spectroscopy, model transfer, uncertainty quantification, soil carbon, linear mixed models, multilevel modeling, hierarchical modeling

1 Introduction

It is well established that the infrared spectrum of a soil can be used to predict various soil properties including carbon concentration. While spectroscopy promises to reduce the processing and analysis costs for soil carbon and other properties, challenges remain that limit the uptake of this technology, including: (1) transfer, i.e. the performance of spectroscopy prediction models in soils with limited training data, and (2) uncertainty, i.e. the quantification of uncertainty in model predictions.

1.1 Model transfer

One way to improve spectroscopy model performance in a target soil (e.g. an agricultural field) is to collect additional training data in the target soil or similar soils. However, this adds costs for sampling, spectroscopic scans, and lab measurements. Alternatively, the spectroscopy model itself can be enhanced to improve performance on the target soil without new training data, and therefore without added cost. The performance of a model in different settings is called *model transfer* (Rossel et al., 2024).

A standard (baseline) model for soil spectroscopy is partial least squares regression (PLSR). This model specifies a single linear equation to predict a soil property from a spectrum. However, the underlying relationship between the spectrum and the target soil property may vary across different soils. Thus the performance of a global model like PLSR can be hindered by the presence of training data from soils that are substantially different from the target soil.

One approach to model transfer, then, is to select a subset of available data for training, excluding dissimilar soils. Similarity can be estimated from the spectra themselves (e.g. memory-based learning, MBL) or by subsetting the data according to mapped soil types (e.g. model stratification). When selecting a subset, there is often a practical trade-off between its size (more training data is better) and its relevance (more relevant data is scarce) (Moura-Bueno et al., 2020).

Another approach is to use a model that, unlike PLSR, allows the relationship between the spectrum and the predicted soil property to vary based on auxiliary information or the spectrum itself. Examples of such models include tree-based methods such as Cubist and quantile regression forest (QRF). Because such models use all available training data, they potentially avoid the trade-off between training data size and relevance encountered by models that subset the data.

1.2 Uncertainty quantification

Uncertainty quantification helps us understand whether our spectroscopy predictions are accurate and precise enough to support a given task, such as mapping or mean-area estimation, and subsequent decision-making. Despite its importance, uncertainty quantification is a relatively under-explored area of soil spectroscopy (Wadoux and Ramirez-Lopez, 2025). For example, most software implementations of PLSR do not quantify uncertainty.

We identified three main existing approaches to uncertainty quantification in the soil spectroscopy literature. The first is a set of techniques specific to PLSR (Zhang and

Garcia-Munoz, 2009). The second is quantile regression forest (QRF), which leverages the ensemble nature of random forest to quantify uncertainty (Meinshausen and Ridgeway, 2006;Wadoux and Ramirez-Lopez, 2025). The third is conformal prediction (Anastasios and Bates, 2023;Sanderman et al., 2025; Safanelli et al., 2025), a general framework for adding uncertainty quantification to any prediction model by estimating the "nonconformity" of new data with respect to the trained model.

Beyond making predictions for individual observations (soil samples), an important task is making aggregated predictions, e.g. estimating mean soil carbon concentration across a field, farm, or other geography. This is relevant for national inventories as well as market and non-market carbon projects (Potash et al., 2025a). While the “point prediction” of the mean is simply the mean of the point predictions of each individual observation, quantifying the uncertainty of this prediction is complicated because of correlated errors (Wadoux and Heuvelink, 2025).

1.3 This study

In this study, we apply linear mixed models (LMMs) to soil spectroscopy. LMMs, also known as hierarchical or multilevel regression models, are widely applied in ecology (Bolker et al., 2009) and agronomy (Gbur et al., 2020). However, they have received less attention in soil science due in part to the recent appeal of machine learning (ML) for modeling complex relationships with limited human input (Schmidinger et al., 2025).

An LMM is a regression model in which some of the coefficients (intercept and slopes) vary across groups of observations according to a probability distribution. For example, in this study, we allow the coefficients of a spectroscopy regression model to vary across agricultural fields and soil depth layers. Moreover, LMMs can model this variation in coefficients by leveraging readily available auxiliary information. In this study, the LMM coefficients vary with field-level soil survey estimates of soil pH and texture.

LMMs have several important features that emerge from their definition. These features support their success in a wide range of applications, and make them good candidates for both the model transfer and uncertainty quantification challenges of soil spectroscopy. In particular, the so-called “partial pooling” of observations across groups is a potential solution to the above trade-off between relevance and size of training datasets encountered by subsetting approaches to model transfer (section 1.2). Additionally, LMMs use auxiliary information to make localized predictions in unobserved groups, e.g. fields with no training measurements. Finally, LMMs can address uncertainty quantification because, as Bayesian probability models, they automatically quantify the uncertainty of their predictions using posterior distributions that include both structural (parameter) uncertainty and residual uncertainty.

The purpose of this study is to investigate the potential of LMMs for soil spectroscopy. Specifically, we wish to assess LMM performance in comparison to existing approaches for the simultaneous challenges of (1) model transfer and (2) uncertainty quantification. We examine soil carbon concentration in a cluster of agricultural fields in Northeastern USA. This study is especially relevant to agricultural carbon inventories and markets, where spectroscopy has the potential to reduce the costs of measurement to make carbon projects more economical (Potash et al., 2025a).

2 Materials and methods

2.1 Data

The study was conducted in the Hudson Valley region of Northeastern U.S.A. (42 °N, 73 °W), at 18 fields located within 30 km of each other (figure 1). The temperate climate has a mean annual temperature of 9 °C and precipitation of 11 cm. The crops were a mix of corn-soy and hay and varied in size (3 - 12 ha) and elevation (60 - 280 m above sea level) (c.f. Polussa et al., 2026). For each field, we estimated soil pH and texture to 60cm depth using area-weighted averages from USDA NRCS Soil Survey Geographic Database (SSURGO); pH ranged from 5 to 6.5, and texture was predominantly silt loam (figure 2).

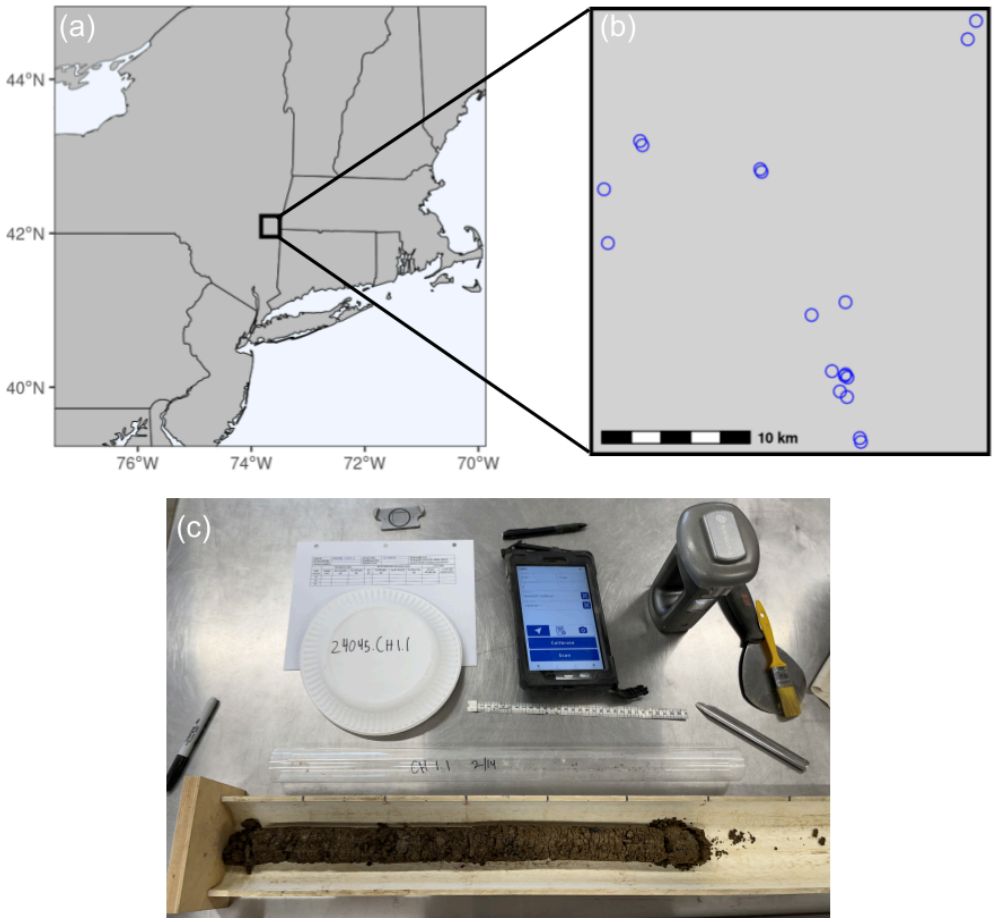

Figure 1: Study design. (a) Study region of Northeastern USA. (b) Locations of 18 fields sampled for study. (c) Handheld spectrometer scanning moist soil samples.

In winter 2023-2024, we used spatial coverage sampling (Walvoort et al., 2010) to select 4 to 15 locations per field (1 to 1.7 samples ha^{-1}). At each sampling location, intact soil cores (4 cm diameter) were collected to a target depth of 60 cm using a truck-mounted hydraulic drill

corer, or gas-powered handheld slide hammer where vehicular access was constrained. Cores were collected in plastic liners that were capped and refrigerated.

In the lab, field moist cores were extracted into a trough (split 10 cm diameter PVC pipe) and separated into depth layers of 0-15 cm, 15-30 cm, and 30-60 cm. Each layer was manually broken up and large rocks and roots removed. Spectroscopy scans were collected for each sample (depth layer) using a Neospectra Scanner Handheld NIR Analyzer (Si-Ware, Egypt; now BUCHI Proxiscout) to measure reflectances at 257 wavelengths between 1350 and 2550 nm. Six replicate scans were performed on each sample and averaged. Standard normal variate (SNV) transformation was applied to the averaged spectra. For the LMM, spectra were downsampled prior to SNV transformation (section 2.2.2).

The same samples were then air-dried and passed through a 2 mm sieve. Carbon concentration (%) was measured by dry combustion with an elemental analyzer (Vario MAX cube, Elementar Analysensysteme GmbH, Germany). According to their classification, these soils have negligible inorganic carbon, so total carbon concentrations are likely to reflect organic carbon concentrations. Carbon concentrations were modeled on the log1p scale (Safanelli et al., 2025).

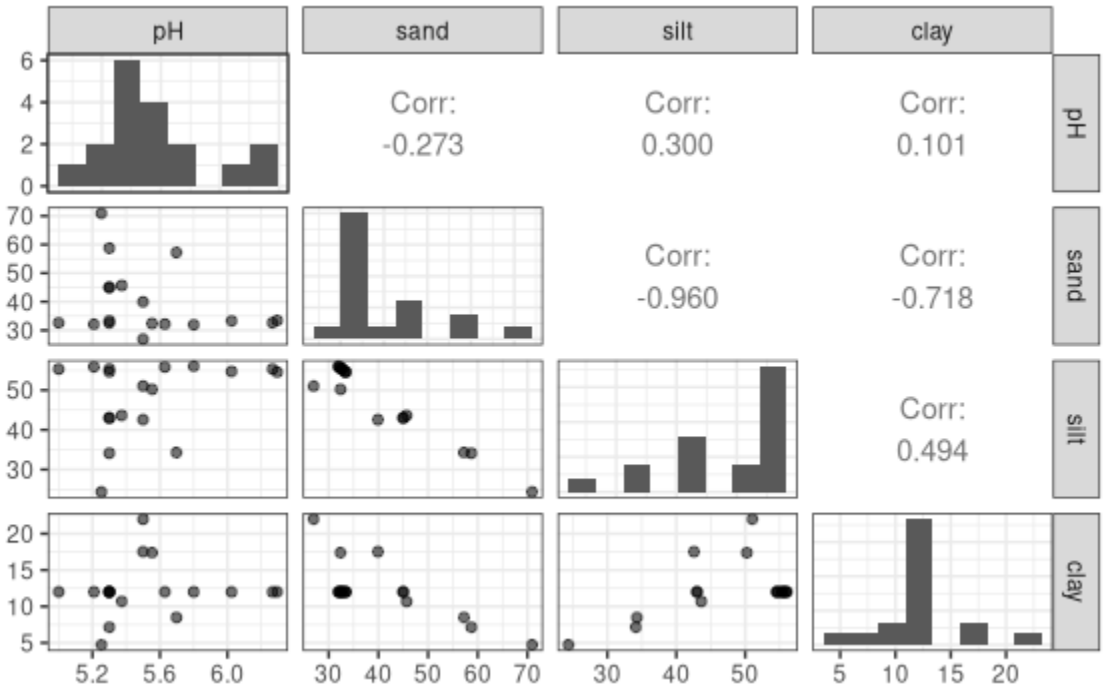

Figure 2: Average soil pH and texture properties estimated by the national soil survey (SSURGO) of the 18 study fields.

2.2 Models

2.2.1 Existing models

An important feature of spectroscopy modeling is the relatively large number of predictors. Traditional regression typically does not work well when the number of predictors (here n=257) is large relative to the number of observations (here n≤459). As a result, dimensionality reduction and regularization are common in spectroscopy modeling.

Partial least squares regression (**PLSR**) is a classic model for spectroscopy (Geladi and Kowalski, 1986) that combines regression with dimensionality reduction that accounts for the covariance between predictors and outcomes (Barra et al., 2021). We adopted the uncertainty quantification method of Faber and Kowalski (1996), which performed best in the evaluation of Zhang and Garcia-Munoz (2009). PLSR provides a global linear regression, though this may be transferred to a given prediction set (e.g. an agricultural field). One approach to transfer is to **stratify** the training data based on auxiliary information. In this study we stratified on soil texture (loam, sandy loam, or silt loam). However, stratification rigidly partitions the training data, creating a trade-off between training set relevance and training set size (section 1.1).

Another approach to PLSR model transfer is memory-based learning (**MBL**). MBL searches for a subset of samples (neighbors) from the training set before fitting a prediction model (weighted average PLSR) for each sample to be predicted (Ramirez-Lopez et al., 2013). We implemented MBL using the `resemble` R package. The number of neighbors and number of PLSR components was selected by leave-nearest-neighbor-out cross-validation (`validation_type = "NNv"`). The dissimilarity matrix was included as a predictor in the model (`diss_usage = "predictors"`). Spiking was used to ensure that local observations, when available, were included as neighbors.

An alternative to dimensionality reduction is regularization, a feature of ensemble methods. **Cubist** is a tree ensemble with linear regressions at its leaf nodes. Random forest is another tree ensemble where a leaf node prediction is simply the average of the training observations in that node. Both Cubist and random forest inherently support model transfer because they are non-linear models with predictions that are localized to different regions of the spectroscopy input space. Cubist is a standard model in soil spectroscopy following McBratney and Minasny (2009). Random forest has been advanced for soil spectroscopy because of quantile random forest (**QRF**), a variant that quantifies uncertainty (Wadoux and Ramirez-Lopez, 2025). Sanderman et al. (2025) quantified uncertainty for Cubist using conformal prediction.

2.2.2 Linear mixed model (LMM)

We refer the reader to textbooks such as Gelman and Hill (2007) for a general introduction to LMMs and to section 1.3 for their pertinence to model transfer and uncertainty quantification. In R notation our LMM formula was

```
log1p(Y) = 1 + Z + X + Z:X + (1 + X|field + layer + field:layer)
```

where `Y` is soil carbon concentration, `X` is shorthand for `K` reflectances `X_1 + … + X_K`, and `Z` is shorthand for L group-level predictors `Z_1 + … + Z_L`. For computational reasons, we used cubic splines to interpolate the original 257 wavelengths (followed by SNV) into `K=64`. We used four group-level predictors at the field level: pH, sand, silt, and clay (all estimated from SSURGO), because they summarize soil chemical and physical properties and are known to be strongly predicted by the visNIR spectrum (Safanelli et al., 2025). Note that the terms `Z` and `Z:X` induce the regression intercept and slopes, respectively, to vary as functions of the group-level predictors.

We fit the LMM in a Bayesian framework using Markov Chain Monte Carlo (Gelman et al., 2013). We used weakly informative priors, including a hierarchical regularization prior to handle the large number of spectroscopy predictors (see supplement). We used the probabilistic programming language Stan (Carpenter et al., 2017) via the `brms` R package (Burkner, 2017). Our convergence diagnostics are described in the supplement. Data and R source code are provided.

2.3 Cross validation scenarios

We used cross validation (CV) to evaluate transfer and uncertainty quantification under two scenarios: when no local observations (i.e. soil carbon measurements) are available and when some local observations are available. These scenarios emulate mode transfer to a target field with substantial amounts of training data from similar, nearby fields but little or no training data from the target field.

Under "**without local observations**" CV, training included all samples outside the target field (17 fields; 399 to 447 samples), and testing included all samples in the target field (12 to 45 samples). Under "**with local observations**" CV, training additionally included 3 random cores (3 depths each, 9 samples total) in the target field (18 fields; 408 to 456 samples), and testing consisted of the remaining samples in the target field (3 to 36 samples). "Without local observations" is akin to a "global" model and "with local observations" is akin to a "spiked" model (Ng et al., 2022). In both scenarios, auxiliary data (pH and texture) was used by the LMM for all fields.

For each CV fold, we fit the models (section 2.2) on the training set to predict carbon concentration. We made point predictions for each sample in the test set. For models with uncertainty quantification (all except MBL), we generated 95% intervals for each of the test samples. Finally, we generated 95% intervals for the predicted mean carbon concentration across each soil layer in each test set. Note that the LMM jointly models uncertainty across samples. Other models produced intervals using traditional standard error formulas which assume that predictions are unbiased (supplement).

We evaluated point predictions using standard metrics such as RMSE, resulting in one value for each of the 18 folds in each CV scenario. Prediction intervals were evaluated by their empirical coverage rate (proportion that contained the measured value, ideally 95%). Because of the small size of the test sets, these were pooled into a single coverage rate under each scenario. Finally, we evaluated the interval width as well as coverage using Winkler scores (supplement).

3 Results

3.1 Soil carbon distributions

Among the 459 soil samples, mean carbon was 1.2% with a standard deviation of 0.8%. We used a random-effects model to decompose the carbon distribution into 5 sources of variation (figure 3c): layer, field, core, field × layer, and sample, i.e. residual. The largest source of carbon variability was the soil layer (figure 3a).

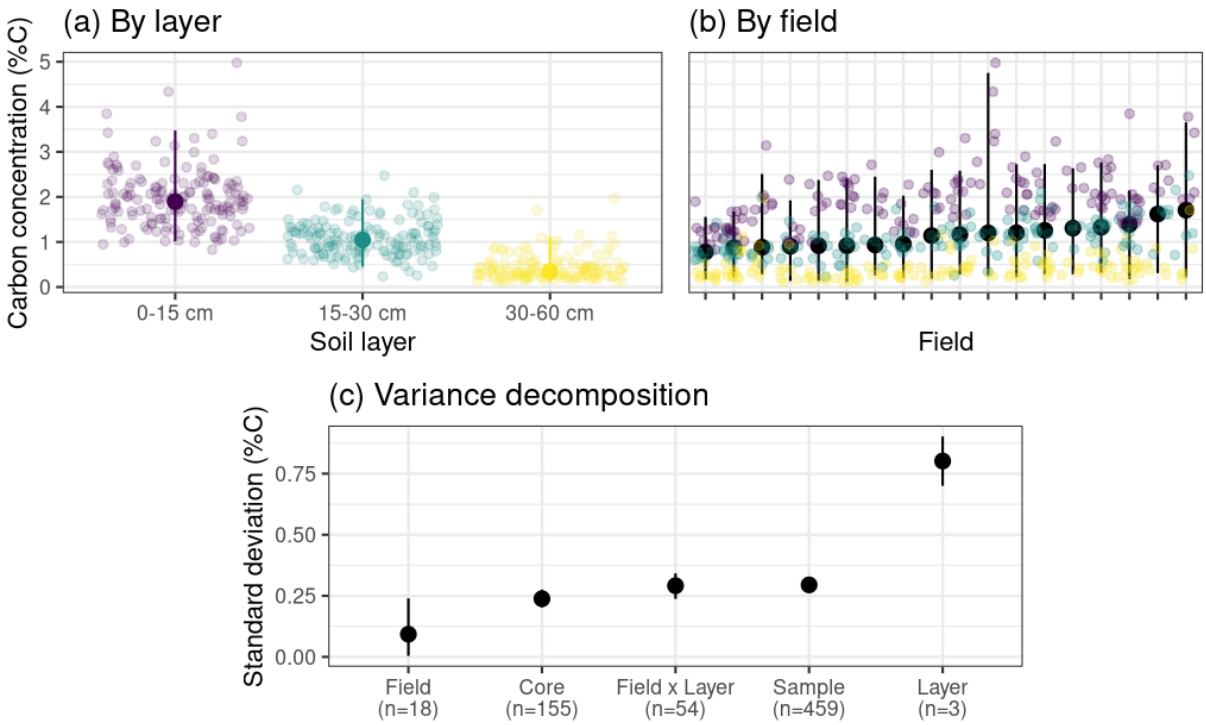

Figure 3: Soil carbon statistics. (a) Distributions of raw measurements by soil layer. (b) Distributions of raw measurements by field (n=18), sorted by median (dot) with 95% interval (line). (c) Variance decomposition from a random effects model.

3.2 Model transfer

Without local observations, mean RMSE (across 18 folds, one for each field) under PLSR was 0.49%, RPD was 2.1, absolute bias was 0.16% and R^2 was 0.79. With local observations, these metrics improved to 0.38%, 2.4, 0.14%, and 0.81, respectively.

When comparing models, we observed the following patterns across metrics and CV scenarios (figure 4). First, both the stratified (PLSR) and QRF models substantially under-performed the PLSR baseline. Second, MBL tended to improve on PLSR though the benefit was greater without local observations. Third, Cubist significantly improved on PLSR,

especially without local observations. Finally, the LMM performed best, especially with local observations.

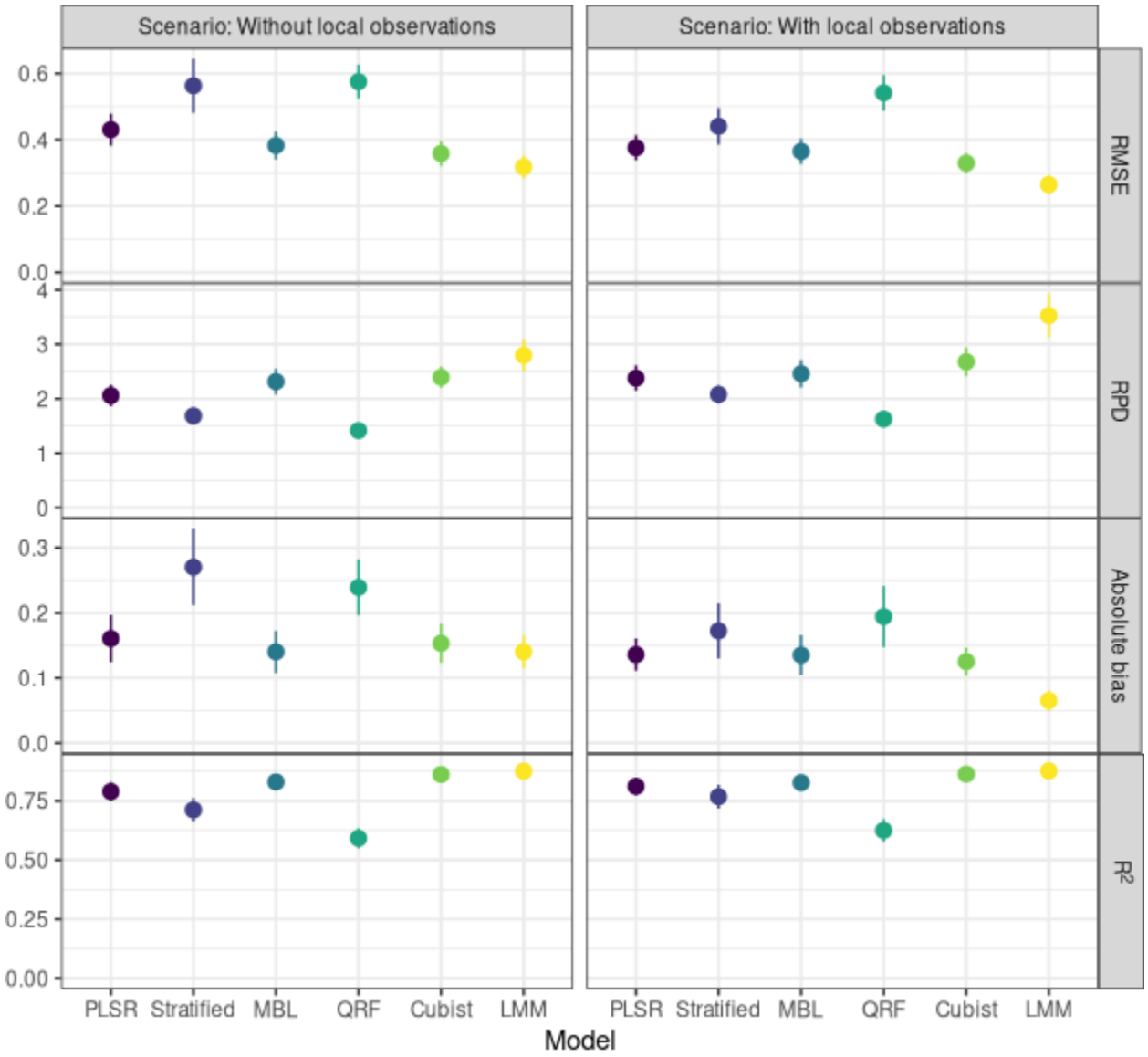

Figure 4: Model transfer performance. Each model is evaluated under two transfer scenarios (section 2.3): without local observations, and with local observations. Each model's performance in each scenario consists of 18 cross validation folds which are visualized by a point (mean) and bar (standard error).

3.3 Uncertainty quantification

We generated 95% prediction intervals for each model under both transfer scenarios (without local observations, with local observations) at both the sample (n=459) and field-layer (n=56) levels. All models except MBL had individual sample-level uncertainty quantification. All models including MBL generated 95% intervals for field-layer means, whether directly (LMM) or through a standard error formula (other models).

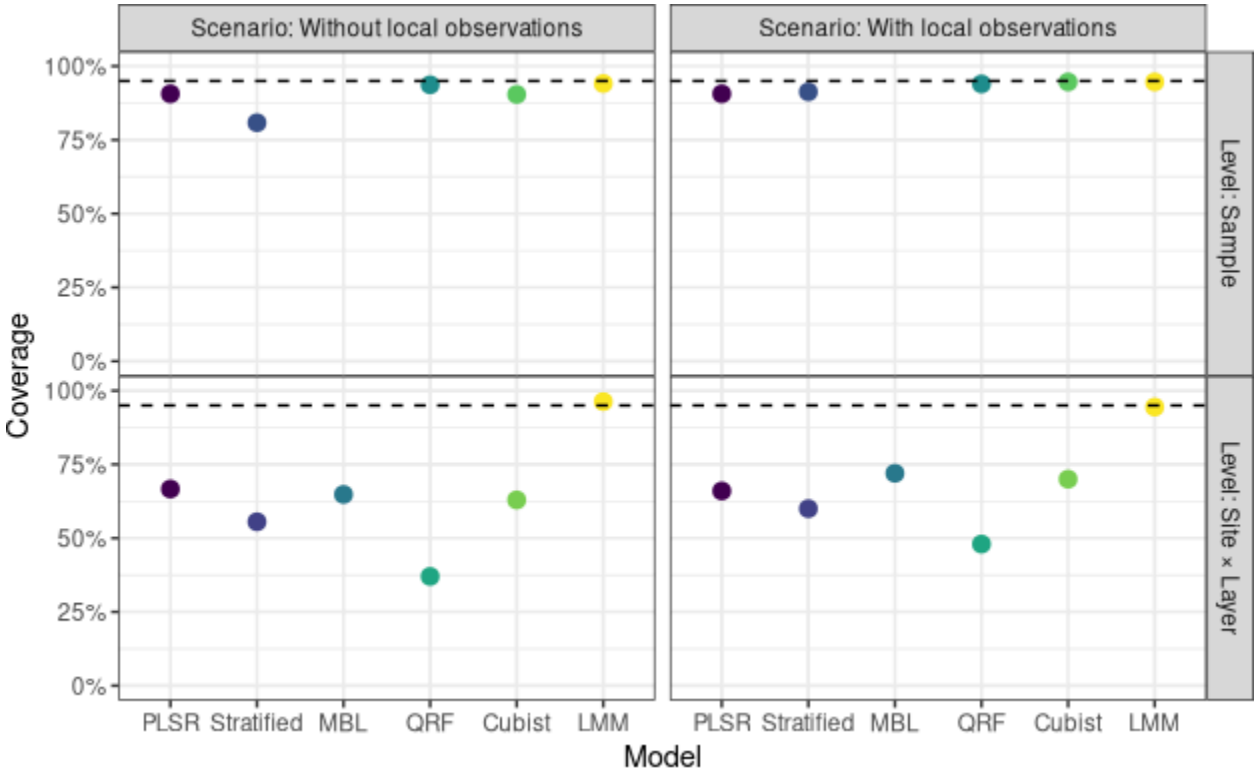

Figure 5: Uncertainty quantification at (top row) sample and (bottom row) field-layer scale, i.e. one prediction for each soil layer (0-15 cm, 15-30 cm, 30-60 cm) in each field. Points show empirical coverage rate of 95% intervals compared to the target coverage rate of 95% (dashed line).

Our first evaluation of these intervals was to compare their empirical coverage to the nominal 95% (figure 5). At the sample level, all models had close to nominal coverage except the stratified model without local observations. At the field-layer level and under both transfer scenarios, all models except LMM significantly underestimated their prediction uncertainty, with coverage rates between 25 and 75%. We note that the extent of this undercoverage is a function of model bias that is not accounted for in standard errors (supplement). Meanwhile, the LMM correctly produced intervals with approximately 95% coverage. The Winkler score evaluates both interval coverage and width and also shows that the LMM performs best, followed by Cubist (figure S1).

3.4 Linear mixed model interpretation

The coefficients of the visNIR reflectances (figure 7a) align with prior studies of soil carbon spectroscopy, specifically the importance of the 2000-2500 nm band. We also see influences of the surveyed soil properties on the regression intercept (figure 7b). The coefficient of pH is moderately large, negative, and precisely estimated. The coefficients of the soil texture attributes are more uncertain, which is not surprising since the attributes are highly correlated (figure 2). The soil layer coefficients (figure 7c) align with the fact that deeper soil layers have

lower carbon concentrations (figure 3). We also investigated the varying slopes of the LMM and found the variations were small (figure S2) but contributed meaningfully to prediction performance (figure S3).

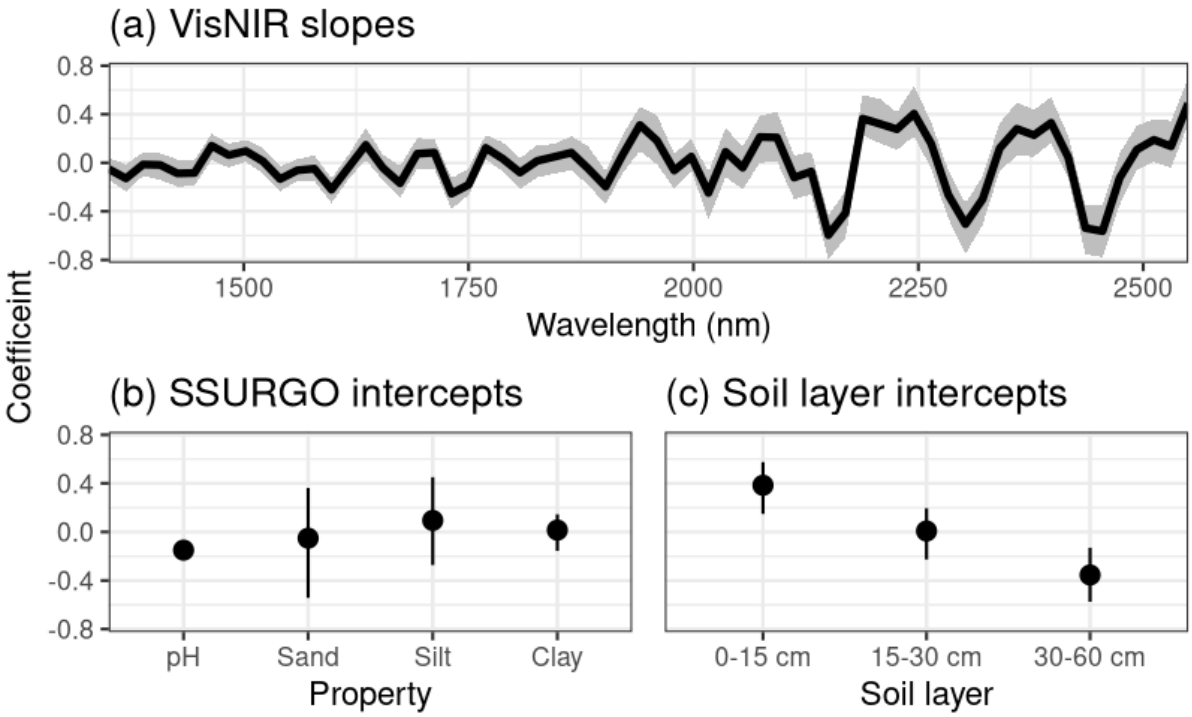

Figure 6: Linear mixed model (LMM) coefficients of (a) visNIR pseudo-reflectances, (b) group-level soil survey (SSURGO) predictors, and (c) soil layer intercepts. Note that outcome (on log1p scale), reflectances, soil properties have been standardized. All panels show Bayesian point estimates (posterior median) and uncertainties (posterior 50% interval).

4 Discussion

4.1 Main findings

In this study we evaluated the transfer and uncertainty quantification performance of a variety of spectroscopy models in the context of a regional cluster of 18 agricultural fields. We compared existing models to a new approach: a linear mixed model (LMM) that leverages auxiliary information (field, soil depth, and surveyed soil properties). In summary, we found that the LMM outperformed five existing approaches in both prediction performance and uncertainty quantification.

The prediction performance of the existing models varied; stratified PLSR and QRF consistently performed worse than the PLSR baseline, while MBL and Cubist performed better.

However, the LMM consistently performed best, reducing RMSE compared to PLSR by 19% and 24% on average without local observations and with local observations, respectively.

All models except MBL quantified uncertainty well for sample-level predictions. Specifically, they generated 95% intervals for these sample-level predictions with close to 95% coverage. However, when quantifying uncertainty for field-layer mean carbon concentration, existing models did not account for bias and severely underestimated the uncertainty of their field-layer predictions. The LMM was the only model accounting for correlated errors and so was the only model to correctly quantify uncertainty at the field-layer level (figure 5).

4.2 Considerations and future work

4.2.1 Setting

The relatively small geographic scale (30km) and homogeneous soils in our study make this a relatively favorable setting for a global PLSR model and therefore a relatively challenging setting for model transfer to improve on the PLSR baseline. Despite this, the LMM was a substantial improvement, suggesting the potential for even greater LMM transfer benefits in more heterogeneous soils (e.g. larger geographic scale).

4.2.2 Dataset size

In general, LMMs benefit from a large number of groups and samples per group in their training data (Gelman and Hill, 2007). We had 18 fields and 54 field-layers, with at least 3 samples per group. In the extreme cases of just one or two groups, or just one observation in every group, the LMM will be unable to estimate a group-level structure. However, the LMM should “fail gracefully” in these situations and reduce to a global regression. Future work should explore the sensitivity of LMM performance to the dimensions of training data.

4.2.3 Group-level predictors

For the local regression for a particular group (e.g. field), the LMM relies on: (i) observations in that group, if any are available in the training data; and (ii) group-level predictors in conjunction with observations from other groups. Without local observations, it is not surprising that the group-level predictors were crucial for LMM performance. Conversely, with local observations, i.e. including 9 samples from the target field in training, group-level predictors did not benefit LMM performance (figure S4).

Without local observations, the availability and quality of group-level predictors will have a large effect on LMM performance. We used SSURGO estimates of texture and pH, which are widely available in the USA, though their quality and relevance for spectroscopy prediction may vary. More precise estimates of these variables could improve LMM performance. Future work could validate these predictors in other regions of the USA and for outcomes besides carbon concentration, as well as in other countries (from other sources) and consider other predictors altogether. However, we emphasize that these variables were only important in our study without observations.

4.2.4 Modeling extensions

One benefit of the LMM is that it can be extended using well-developed Bayesian regression techniques. For example, a multi-outcome model (also known as multi-output or multi-task), in which multiple soil properties are modeled jointly. Specifically, a joint model of bulk density and carbon concentration would produce a model of carbon stocks, while accounting for potential residual correlation. Another example is a Gaussian process (kriging) component to account for residual spatial autocorrelation (e.g. Potash et al., 2025b). The random effects structure of the LMM could be expanded to include: finer geography (map units instead of fields), different years, different projects, and different sampling and measurement techniques. It is straightforward to implement these extensions (e.g. using the `brms` R package). However, the LMM in this study already requires considerable computation (~15 minutes on a budget laptop), though a simpler LMM (random intercepts but not slopes) performed almost as well (figure S3) in seconds.

ML models can be extended to include the LMM features of (i) the grouped structure of fields and soil layers and (ii) auxiliary group-level predictors. Including these variables in Cubist improved performance, though the LMM still performed best (figure S5; Moura-Bueno et al., 2021). To handle correlated errors, conformal prediction could be extended to include a group structure, though to our knowledge this is not standard. More generally, while many extensions to ML models are possible, they are less straightforward than for LMMs (above).

4.2.5 Modeling philosophy

While LMM coefficients vary across fields and soil layers, this variation is parameterized by linear relationships with group-level predictors and normally distributed residuals. Compared to less parametric ML, this parameterization may seem overly constrained. However, the LMM performed best in our study and is successful in many other applications. Moreover, the LMM is often easier to interpret (figure 6) than black-box ML.

The LMM is fit in-sample with Bayesian (MCMC) inference. This is unlike ML models which rely on internal cross validation to optimize hyperparameters or obtain residuals for conformal prediction. Cross validation is problematic if the training data is so small that each observation contributes significantly to model fit. It is also a computational burden for models with a large number of hyperparameters.

5 Conclusion

Two challenges that limit the adoption of soil spectroscopy are model transfer and uncertainty quantification. In this study, we collected a novel dataset of traditional soil carbon measurements as well as NIR spectra in a cluster of 18 agricultural fields in Northeastern USA. We evaluated transfer and uncertainty for a variety of existing approaches as well as a new approach that we proposed using a linear mixed model (LMM) that leverages pre-existing soil survey data.

We found that our LMM outperformed existing approaches, especially in (i) prediction performance when local observations were available to train the models and (ii) uncertainty

quantification of aggregated predictions. This strong performance, combined with a proven track record in many other applications, makes LMMs a promising approach for soil spectroscopy.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The authors acknowledge financial support from: NASA Acres Consortium (Potash); Environmental Defense Fund (Partida, Polussa, Sanderman) with awards from King Philanthropies and Arcadia, a charitable fund of Lisbet Rausing and Peter Baldwin; The Yale Center for Natural Carbon Capture (Polussa); and Lawrence Livermore National Laboratory (LLNL) LDRD Program (24-SI-002) (Safanelli).

Data availability

The data and code that support the findings of this study will be made available at the time of publication.

References

Angelopoulos, Anastasios N., and Stephen Bates. "Conformal prediction: A gentle introduction." *Foundations and Trends in Machine Learning* 16.4 (2023): 494-591.

Barra, Issam, et al. "Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–A review." *TrAC Trends in Analytical Chemistry* 135 (2021): 116166.

Bolker, Benjamin M., et al. "Generalized linear mixed models: a practical guide for ecology and evolution." *Trends in ecology & evolution* 24.3 (2009): 127-135.

Faber, Klaas, and Bruce R. Kowalski. "Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler." *Chemometrics and Intelligent Laboratory Systems* 34.2 (1996): 283-292.

Geladi, Paul, and Bruce R. Kowalski. "Partial least-squares regression: a tutorial." *Analytica chimica acta* 185 (1986): 1-17.

Gelman, Andrew, and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2007.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gbur, Edward E., et al. *Analysis of generalized linear mixed models in the agricultural and natural resources sciences*. John Wiley & Sons, 2020.

Meinshausen, Nicolai, and Greg Ridgeway. "Quantile regression forests." *Journal of machine learning research* 7.6 (2006).

Moura-Bueno, Jean Michel, et al. "When does stratification of a subtropical soil spectral library improve predictions of soil organic carbon content?." *Science of the Total Environment* 737 (2020): 139895.

Moura-Bueno, Jean Michel, et al. "Environmental covariates improve the spectral predictions of organic carbon in subtropical soils in southern Brazil." *Geoderma* 393 (2021): 114981.

Ng, Wartini, et al. "To spike or to localize? Strategies to improve the prediction of local soil properties using regional spectral library." *Geoderma* 406 (2022): 115501.

Polussa, Alexander, et al. "Toward causal designs to quantify management-driven soil carbon change at multi-field scales." *agriRxiv* 2026 (2026): 20260167252.

Potash, Eric, et al. "Measure-and-remeasure as an economically feasible approach to crediting soil organic carbon at scale." *Environmental Research Letters* 20.2 (2025a): 024025.

Potash, Eric, et al. "Think outside the plots: Perimeter measurements and spatial modeling mitigate confounding in a 145‐year experiment." *Agricultural & Environmental Letters* 10.1 (2025b): e70020.

Principato, Guillaume, et al. "Conformal prediction for hierarchical data." *arXiv preprint arXiv:2411.13479* (2024).

Ramirez-Lopez, Leonardo, et al. "The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets." *Geoderma* 195 (2013): 268-279.

Safanelli, José L., et al. "Open Soil Spectral Library (OSSL): Building reproducible soil calibration models through open development and community engagement." *PloS one* 20.1 (2025): e0296545.

Sanderman, Jonathan, et al. "Application of a Handheld Near Infrared Spectrophotometer to Farm‐Scale Soil Carbon Monitoring." *European Journal of Soil Science* 76.1 (2025): e70053.

Schmidinger, Jonas, et al. "LimeSoDa: A dataset collection for benchmarking of machine learning regressors in digital soil mapping." *Geoderma* 459 (2025): 117337.

Somarathna, P. D. S. N., Budiman Minasny, and Brendan P. Malone. "More data or a better model? Figuring out what matters most for the spatial prediction of soil carbon." Soil Science Society of America Journal 81.6 (2017): 1413-1426.

Rossel, Raphael A. Viscarra, et al. "An imperative for soil spectroscopic modelling is to think global but fit local with transfer learning." *Earth-Science Reviews* 254 (2024): 104797.

Wadoux, Alexandre MJ-C., and Leonardo Ramirez-Lopez. "Uncertainty of predictions in absorption spectroscopy: Modelling with quantile regression forest." *Chemometrics and Intelligent Laboratory Systems* 265 (2025): 105473.

Wadoux, Alexandre MJ-C., and Gerard BM Heuvelink. "Scientists yet to consider spatial correlation in assessing uncertainty of spatial averages and totals." *International Journal of Applied Earth Observation and Geoinformation* 139 (2025): 104472.

Walvoort, Dennis JJ, D. J. Brus, and J. J. De Gruijter. "An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means." *Computers & geosciences* 36.10 (2010): 1261-1267.

Winkler, Robert L. "A decision-theoretic approach to interval estimation." *Journal of the American Statistical Association* 67.337 (1972): 187-191.

Zhang, Lin, and Salvador Garcia-Munoz. "A comparison of different methods to estimate prediction uncertainty using Partial Least Squares (PLS): a practitioner's perspective." *Chemometrics and intelligent laboratory systems* 97.2 (2009): 152-158.

Supplementary information for Transfer and uncertainty for soil spectroscopy: a parsimonious approach using linear mixed models and soil surveys

Eric Potash, Colleen Partida, Jose Safanelli, Elizabeth Forbes, Alex Polussa, Tiffany Runge, Matthew Sheffer, Jonathan Sanderman

Corresponding author email: epotash@illinois.edu

This PDF file includes:

Supplementary text

Uncertainty quantification of averages

Suppose we have *n* soil samples with true soil properties (e.g. carbon concentrations) $Y_1,\ldots,Y_n$ and corresponding spectroscopy model predictions $\widehat{Y_1},\ldots,\widehat{Y_n}$. Then we can predict the mean soil property $\overline{Y}$ using the mean prediction $\overline{\widehat{Y}}$.
It is standard to generate a prediction distribution by assuming a normal distribution with standard deviation

$$SE_{\widehat{Y}} = sd(\widehat{Y})/\sqrt{n}.$$

For example, the 95% interval would be $\overline{\widehat{Y}} \pm 1.96 \cdot SE_{\widehat{Y}}$. This approach, based on the central limit theorem, implicitly assumes that the prediction errors

$$\epsilon_i = \widehat{Y_i} - Y_i$$

are random samples from a distribution. Moreover, it assumes that the errors have mean zero, i.e. the prediction model is unbiased. When the errors are biased, the width of the interval is systematically underestimated by the above formula.

Linear mixed model priors

The specification of our linear mixed model (LMM) can be found in the accompanying R source function `fit_lmer.brms()`. Here we summarize the choices of prior distributions which were chosen to be weakly informative (Gelman et al., 2013).

As described in the manuscript, the R formula for our linear mixed model (LMM) is:

```
log1p(Y) = 1 + Z + X + Z:X + (1 + X|field + layer + field:layer)
```

Where:

- Y is carbon concentration
- Z are 4 group-level predictors (SSURGO 0-60 cm field-level estimates of pH, sand, silt, and clay)
- X are 64 visNIR predictors.

Note that Y, Z, X, Z:X are standardized before prior specification and model fitting.

We start with standard Normal(0, 2.5) prior on the intercept and exponential(1) prior on the residual. The group-level predictor slopes Z receive Normal(0, 1) priors.

To handle the relatively large number of spectral slopes (K=64) for the spectral predictors X_1,..., X_K receive a hierarchical shrinkage prior with standard deviation σ_{β_X}:

$$\beta_{X_k} \sim Normal(0,\ \sigma_{\beta_X})$$
$$\sigma_{\beta_X} \sim Normal(0,\ 1)$$

The interactions between the group-level predictors and spectral predictors (L*K = 256) receive a similar prior with a distinct standard deviation $\sigma_{\beta_{ZX}}$:

$$\beta_{Z_l X_k} \sim Normal(0,\ \sigma_{\beta_{ZX}})$$
$$\sigma_{\beta_{ZX}} \sim Normal(0,\ 1)$$

Next there are six standard deviation parameters, one for each of the three levels (field, layer, and field-layer) times two types (intercepts, slopes). Each of these receives a standard normal prior.

Model diagnostics

Models were fit using Stan via the brms interface (see provided R code). We generated four chains with 200 iterations each, saving the last 100 to produce 400 samples from the joint posterior distribution of each model. Convergence was assessed using the criteria $\hat{R}$ < 1.05, n_{eff}> 100, and n_{eff}/n > 10% where $\hat{R}$ is the Gelman-Rubin convergence statistic, n_{eff} is the effective sample size, and n_{eff}/n is the effective sample size ratio (Gelman et al., 2013).

Winkler scores

The Winkler score is defined by

$$\text{Winkler}(l, u, y, \alpha) = (u - l) + \frac{2}{\alpha} \begin{cases} l - y, & \text{if } y < l \\ 0, & \text{if } l \le y \le u \\ y - u, & \text{if } y > u \end{cases}$$

where l and u are the lower and upper bounds of the prediction interval, y is the true value, and $1 - \alpha = 0.95$ is our desired coverage (Winkler, 1972).

Supplementary figures

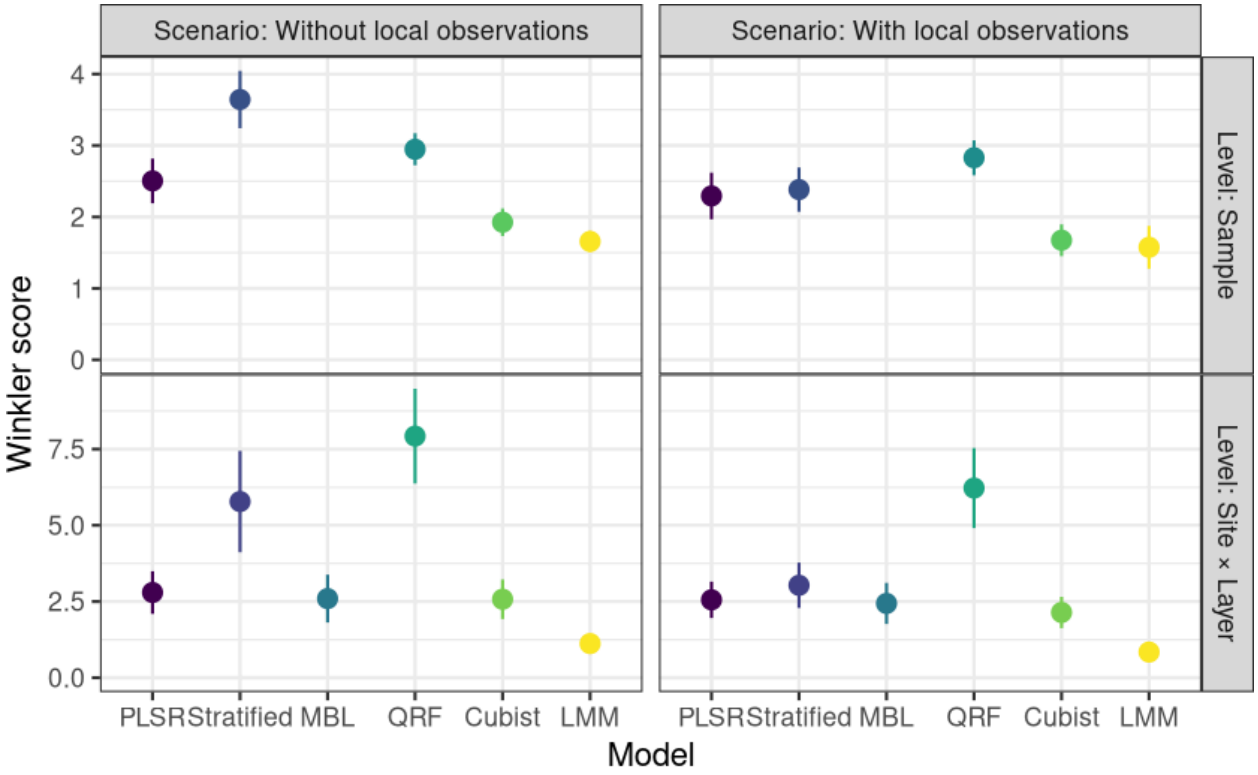

Figure S1: Winkler scores (lower is better) at (top row) sample and (bottom row) field-layer scale. Points and lines are mean and standard error, respectively.

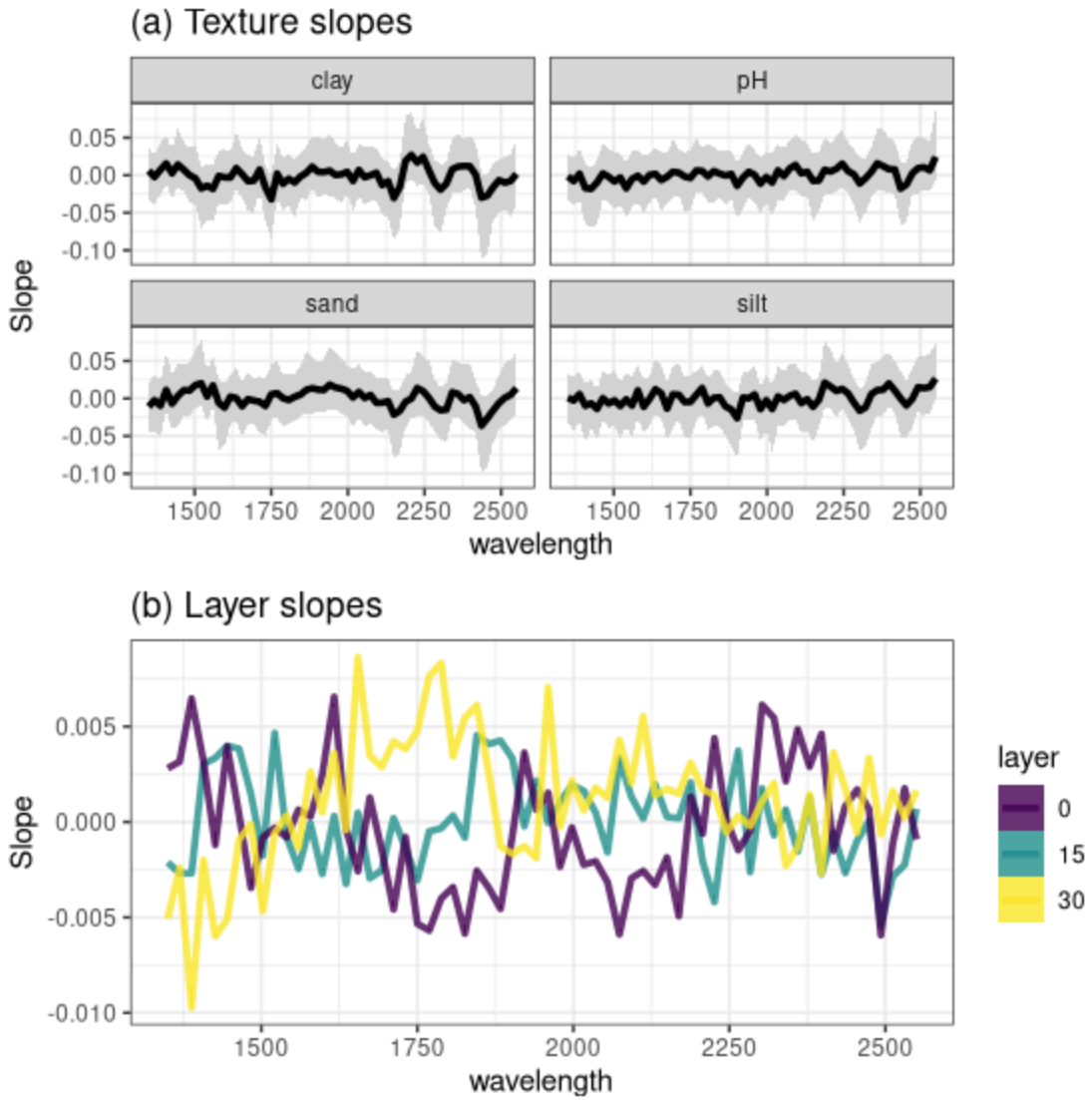

Figure S2: Variation in slopes by (a) field-level SSURGO predictors and (b) soil layer.

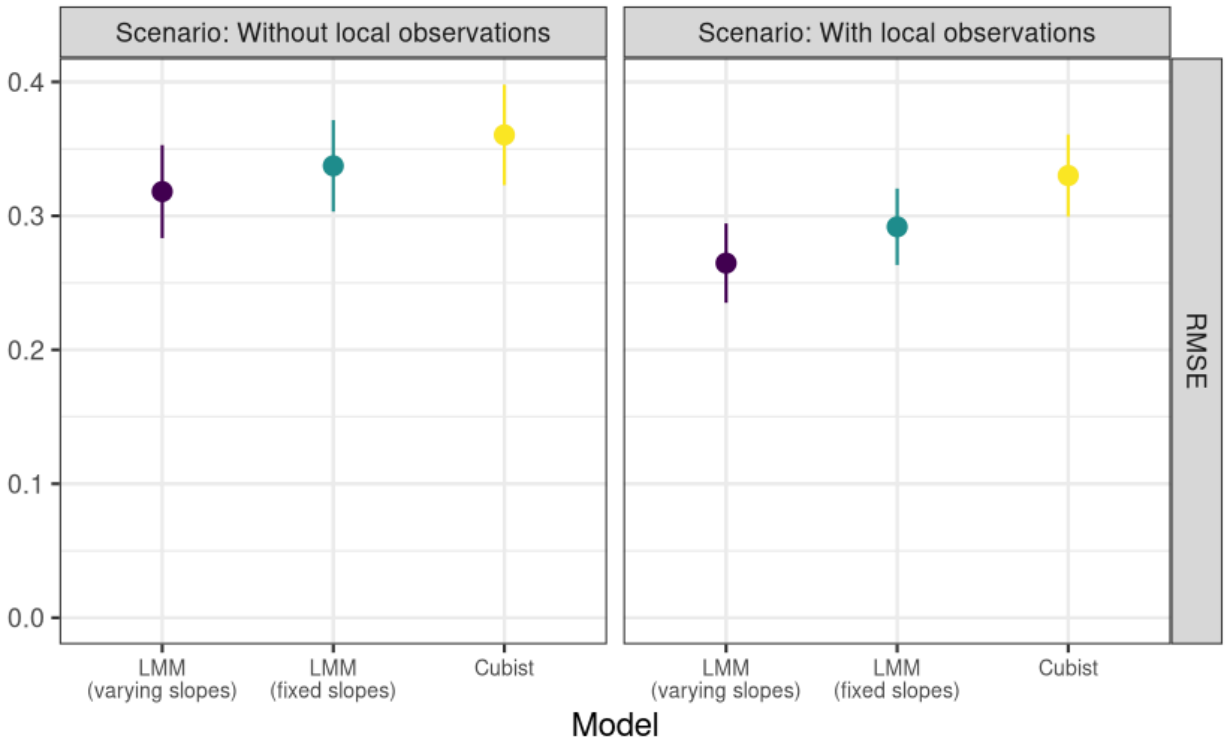

Figure S3: Influence of varying slopes on LMM performance. Fixed slope (varying intercept) LMM does not perform as well as varying slope LMM, but performs better than Cubist. Note that the fixed slope model can be fit about 100 times faster (9 seconds vs 900 seconds).

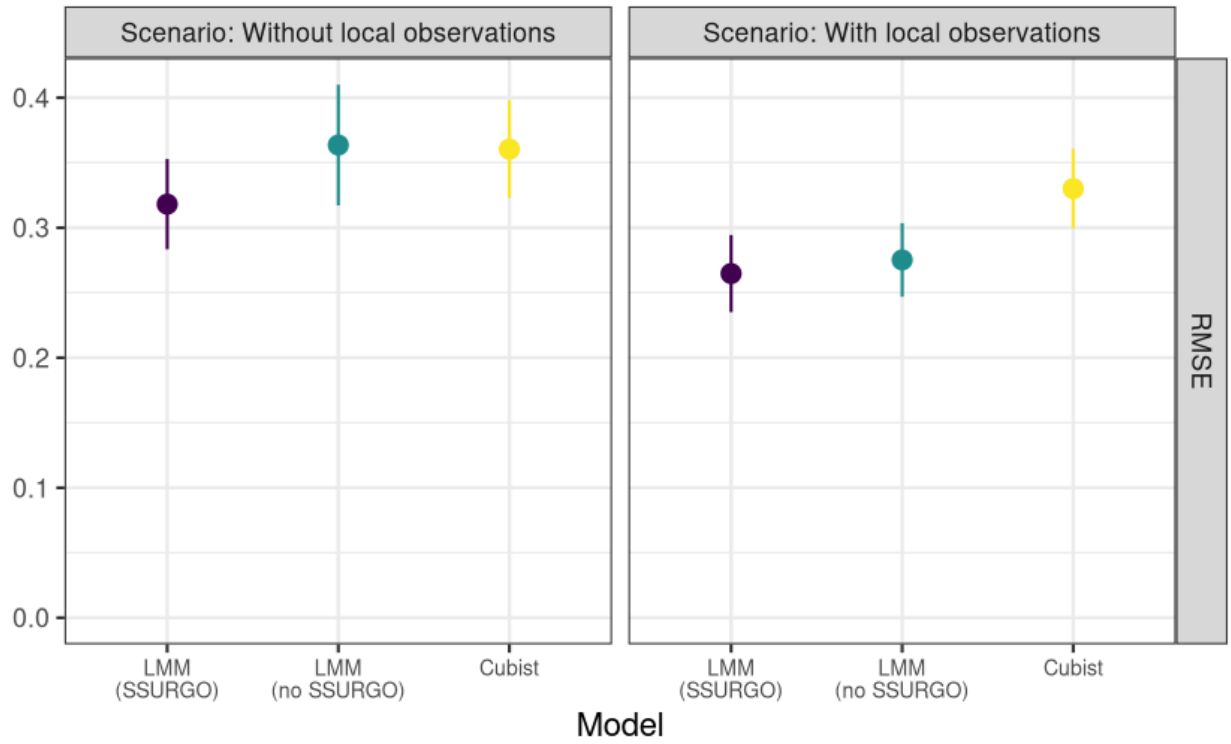

Figure S4: Influence of SSURGO group-level predictors on LMM performance. SSURGO is important for performance in no local observations scenario but not important in local observations scenario.

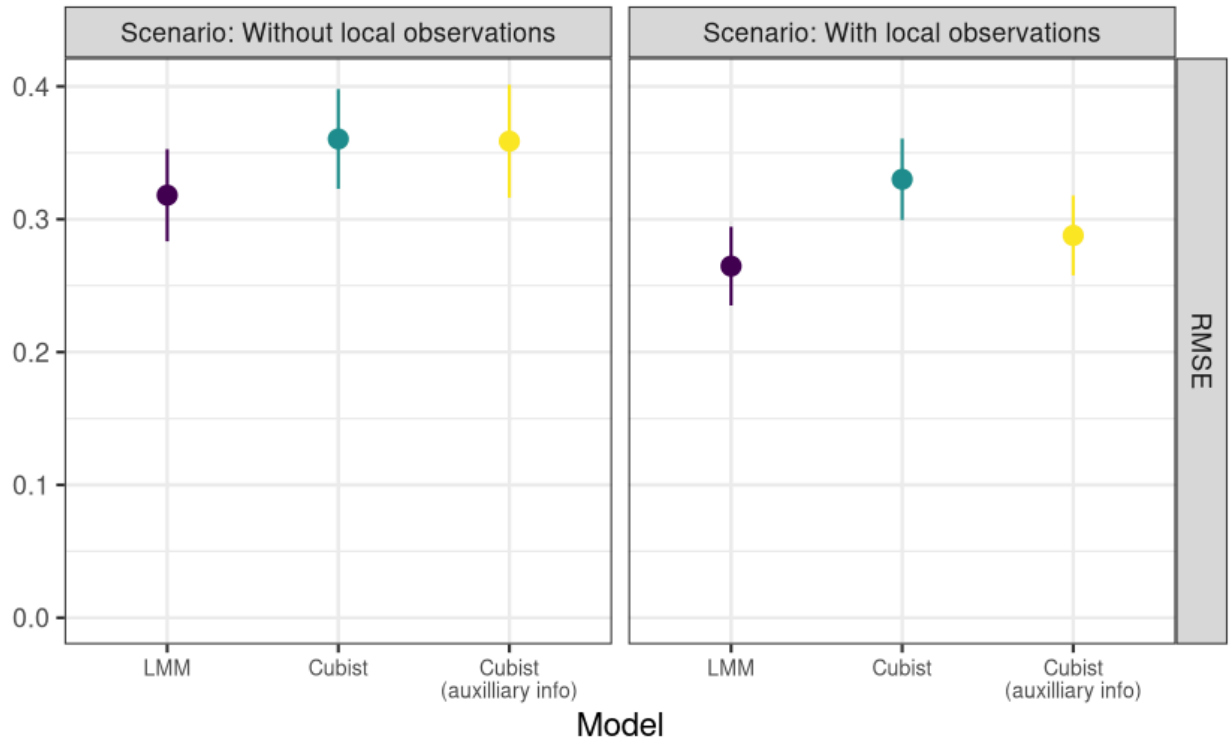

Figure S5: Influence of auxiliary information on Cubist performance. Adding field ID, soil layer, and SSURGO texture and pH predictors to Cubist improves performance in the local observations scenario but not in the no local observations scenario. In both cases, LMM performs best.