Enhancing the Reliability of Portable Soil Probes Through Machine Learning Optimization

A.N. Cambouris[1], E. Lord[2], M. Duchemin[1]

[1]Quebec Research and Development Centre (QRDC), Agriculture and Agri-Food Canada (AAFC), 2560 Hochelaga Blvd., Quebec City, QC, G1V 2J3, Canada.

[2]St-Jean-sur-Richelieu Research and Development Centre, AAFC, 430 Gouin Boulevard, Saint-Jean-sur-Richelieu, QC, J3B 3E6, Canada.

A paper from the Proceedings of the
17th International Conference on Precision Agriculture and 11th Brazilian Congress on Precision and Digital Agriculture
13-15 July 2026
Porto Alegre, Brazil

Abstract.

Precision agriculture requires high-resolution soil information, yet traditional laboratory analyses limit sampling intensity due to high labor and analytical costs. Portable spectrographic probes offer rapid in-field soil characterization, but their accuracy often falls short of laboratory standards. This study evaluated whether machine learning (ML) models can improve the reliability of probe-based measurements by calibrating them against laboratory reference values.

A dataset of 392 soil samples from seven fields with contrasting textures in Eastern Canada was collected using a Dutch auger and analyzed in the laboratory for water pH (pHw), buffer pH (pHb), and total carbon (excepted for TotC, 249 samples; 5 fields). The ChrysaLabs probe was inserted into the same sampling holes to obtain in situ measurements. Five or three fields were used for model training for pH and TotC, respectively, while two fields served as an independent test set. Calibration augmentation experiments were conducted by progressively adding 5–30% of samples from the test fields into the training dataset. Descriptive statistics revealed substantial discrepancies between laboratory and probe datasets, confirming the need for ML-based correction.

A baseline Laplacian kernel Support Vector Machine (kSVM) regression model was developed and compared with two automated machine learning workflows using Support Vector Machine models implemented through the scikit-learn (sklearn) AutoML and Optuna AutoML platforms. Baseline performance (R^2_{kSVM}) reached 0.369 for pHw, 0.461 for pHb, and 0.027 for TotC.

For pHw, sklearn AutoML reached R^2 = 0.590 with no improved with calibration augmentation, whereas Optuna AutoML improved from 0.477 to 0.500 (+5%) with 10% calibration augmentation. For pHb, sklearn AutoML achieved R^2 = 0.741 and improved slightly to 0.746 with 15% calibration augmentation, while Optuna AutoML increased from R^2 = 0.720 to 0.730 with 15% augmentation. TotC predictions showed the largest relative improvement: sklearn AutoML reached R^2 = 0.252 (+36%), and Optuna AutoML R^2 increased from 0.187 to 0.316 (+69%) with 15% augmentation.

These results demonstrate that automated hyperparameter optimization substantially enhances the correspondence between probe-based soil measurements and laboratory reference data.

Optuna AutoMl was the most responsive to additional calibration samples, highlighting the value of field-specific calibration for improving the reliability of portable soil sensing tools.

Keywords. *Portable soil spectroscopy; Machine learning calibration; Support Vector Regression (SVR); Soil pH and carbon prediction; AutoML optimization.*

Introduction

Precision agriculture requires high-resolution soil data, yet traditional laboratory analyses limit sampling because of high labor and analysis costs. This limitation is well documented, as laboratory soil analyses remain costly, time-consuming, and logistically restrictive, often preventing the dense sampling required for site-specific management (Adamchuk et al., 2004; Viscarra Rossel & McBratney, 2008; Reyna Bowen et al., 2025; Albuquerque et al., 2024). Portable spectrographic tools such as the ChrysaLabs probe, Veris® VIS-NIR sensors, SoilCares/NutriScan, and AgroCares handheld spectrometers allow rapid in-field soil characterization with multiple measurements, offering a promising alternative for increasing sampling density. However, their accuracy often falls short of standard laboratory protocols, as portable proximal sensors can exhibit biases, reduced precision, and sensitivity to field conditions (Wetterlind et al., 2013; Ge et al., 2019). Recent studies have demonstrated that machine learning algorithms can substantially improve the accuracy of portable VIS-NIR soil sensors by correcting systematic biases and enhancing prediction performance relative to laboratory standards (Ng et al., 2020; Padarian et al., 2019). Machine Learning (ML) -based calibration models, particularly those using nonlinear or deep learning approaches, have been shown to outperform traditional chemometric methods and significantly strengthen the reliability of *in situ* spectroscopic measurements (Ng et al., 2020). In recent years, automated machine learning (AutoML) frameworks have emerged as effective tools for optimizing model selection and hyperparameters, reducing the need for manual tuning while improving predictive performance in soil spectroscopy workflows (Ramirez-Lopez & Schmidt, 2020). AutoML approaches have also demonstrated enhanced robustness and transferability of soil spectral models across sites and instruments, making them particularly suitable for calibrating portable spectroscopic probes such as the ChrysaLabs system (Hengl et al., 2021). These limitations highlight the need for calibration strategies capable of aligning probe-based measurements with laboratory reference values. Consequently, in this study, our hypothesis was that ML calibration models would enhance the reliability of the ChrysaLabs portable spectroscopic soil probe by aligning its *in situ* predictions more closely with laboratory reference values. In addition, we hypothesized that enriching the training dataset with 0, 5, 10, 15, 20, or 30% of field specific samples would allow the model to better learn site specific spectral–property relationships, leading to improved estimation performance.

Methodology

Fields studied, soil sampling and analysis

Seven agricultural fields (referred to as field #1 to 7) located in Eastern Canada were selected for this study (Fig. 1). These sites represent a diversity of soil textures and cropping systems, with textures ranging from sandy loam to clay. In total, 392 soil samples (pH; 7 fields) or 249 (total carbon; 5 fields) were collected using two approaches, i.e. conventional laboratory analysis and a ChrysaLabs probe (ChrysaLabs, 2023) to determine pH extracted with water (pHw), pH buffer pH (pHb), and total carbon (TotC).

The traditional sampling approach consisted of using a Dutch auger (2.5-cm diameter) to collect a composite soil sample (0–20 cm depth) composed of four soil cores taken within a 1.5-m radius around each georeferenced point. Following the manufacturer's recommendations, the ChrysaLabs probe was then inserted into the same four auger holes to obtain *in situ* measurements.

Soil pHw, pHb, and TotC were obtained using both the conventional laboratory method (hereafter Lab) and the ChrysaLabs probe measurements (hereafter Probe). The pHw was determined using a 1:2 soil-to-water ratio (Hendershot et al., 2008). The pHb was measured using the Shoemaker–McLean–Pratt (SMP) single-buffer method (Ziadi & Tran, 2008) and validated using soil reference materials. TotC was quantified via dry combustion using a LECO CN-2000 analyzer (LECO Corporation, St. Joseph, MI, USA). The soil pHw, pHb, and the TotC determination were done in the QRDC (AAFC), according to the reference method book of Carter & Gregorich, (2008).

Reflectance measurements collected by the probe were transmitted directly to the company's web portal, where machine learning algorithms were applied to estimate soil properties. The processed probe data were subsequently downloaded from the company website and integrated into the experimental database. Descriptive statistics (mean, minimum, maximum, SD, CV) revealed substantial differences between the laboratory and the probe dataset, highlighting the need for ML optimization to improve probe-based predictions.

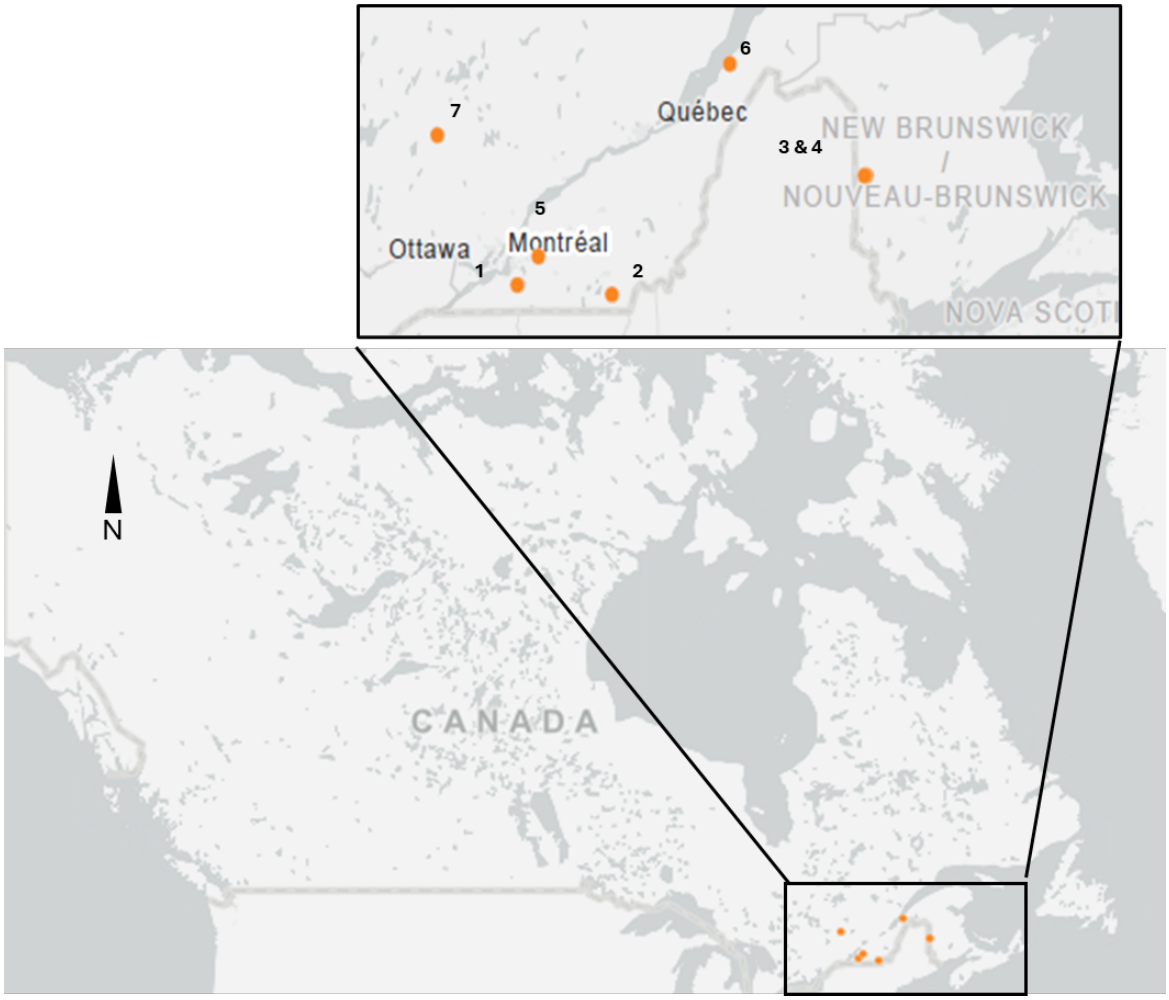

Fig. 1 Canada map with the seven fields located in Eastern Canada. Source: ESRI Canada, ArcGis Online.

Data preprocessing and outlier handling

Data from the different fields were used based on the laboratory data availability. For the pH models, data from fields #1 to 5 were used for training, while data from fields #6 and #7 were reserved for testing, as these fields exhibited a broader range of variability. A different number of fields (# 3, 6, and 7) was used to train the models for TotC and tested on data from fields #2 and #4 since data was not available for the other fields. This partitioning strategy ensured that the

proposed methodology could be applied to previously unseen data, while mitigating the risk of model overfitting.

Prior to the model fitting, an outlier filter was applied to the target variable (Lab value). Sample for which the target value fell outside the interval [μ − 1.5 σ, μ + 1.5 σ], where μ and σ denote the sample average and standard deviation of the target, were removed. This reduced the influence of extreme observations during the hyperparameter optimization procedure. Filtering was only performed on the training partitions. Consequently, dataset train size was 392 (Fields #1 to 5) and 123 (Fields # 3, 6, 7) for the pH and TotC, respectively; and dataset test size was therefore 70 (Fields # 6 and 7) and 126 (Fields # 2 and 4) for the pH, and TotC, respectively.

Methodological ML approaches

A baseline Laplacian kernel Support Vector Machine (kSVM) regression model was developed and compared with two automated machine learning workflows using Support Vector Machine models implemented through the scikit-learn and Optuna AutoML platforms.

As a baseline approach, a Laplacian kernel regression model was determined as the best SVM regression model empirically. After this initial selection, the kSVM was trained using R (4.5.1) with manually tuned parameters using the ε-SVR formulation (type = "eps-svr"). For this task, the ksvm() function from the kernlab package, using a Laplacian kernel (kernel = "laplacedot"), was used. For each replicate (N=5), the kSVM model was trained using the mlr library to search a discrete parameter grid: ε ∈ {0.015, 0.025, …, 0.15} (15 values), C ∈ {10, 20, …, 100} (10 values), σ ∈ {0.25, 0.35, …, 2.5} (23 values). A 5-fold cross-validation scheme was applied, and a grid-search control with resolution = 5 was used to identify the combination minimizing the mean squared error (MSE) on the validation folds (80% training/20% validation of the training set). The model iteration yielding the highest R^2 on the validation set was retained as the best model for that replicate and tested on the test set.

The AutoML platforms Auto-Sklearn (v1.4.2) and Optuna (v4.6.0) were selected based their availability and performance (Akiba et al. 2019, Feurer et al. 2022). Specifically, they allow to use different kernel Support Vector Regression with linear and Radial Basis Function (RBF) kernels. A Laplacian kernel was also implemented to compare the AutoML results to the R version. For all AutoML simulations, five replicates (N=5) were carried out using unique random seed.

For Sklearn, three Support Vector Regression models were compared: a linear kernel SVR, a Radial Basis Function (RBF) kernel SVR, and a Laplacian kernel SVR, all with a regularization parameter C = 1.0, with the RBF and Laplacian kernels using automatic bandwidth scaling (gamma = 'scale') and the Laplacian model additionally specifying an insensitive-loss margin of ε = 0.1. Other default parameters were used for the different models.

For similar to the Sklearn, the Optuna platform with the Tree-structured Parzen Estimator (TPE) algorithm sampler was used and set to maximize R^2. A total of 50 trials per family were executed. Search spaces were defined per algorithm using the default parameters. For example, Random Forest search was tuned over n estimators between 50-500, max depth between 3 and 20, min samples split ranging from 2 to 20 and min samples leaf of 1 to 10. The same Laplacian kernel SVR was used with the Optuna platform.

Test set augmentation experiment

To quantify how rapidly model performance on a new field could be improved by incorporating a small number of samples from that field into the training set, a controlled augmentation experiment was performed. For each algorithm an augmentation fraction (p) of 0, 5, 10, 15, 20, or 30% of stratified random sample was drawn from the test set with a fixed seed, concatenated with the original training set, and used to refit the model. Performance was then evaluated on the complementary (1 − p) fraction of the test set, so that no sample contributed to both training and evaluation. The same sampling was used across algorithms to ensure that all models faced the identical augmented training and test partitions at each level of p.

Performance evaluation, statistical significance and reporting

Model performance was quantified using on the test partition using the coefficient of determination (R^2), all implemented in scikit-learn's metrics module. Fig.s were created using the python matplotlib (v3.7.5) library. Statistical differences were evaluated using the Wilcoxon signed-rank test from the scipy.stats module (v1.11.4).

Results and Discussion

Comparison data Lab versus data Probe per field for pHw, pHb, and TotC

When comparing the distribution of soil pH (pHw and pHb) and TotC across fields, consistent differences were observed between the ChrysaLabs probe and the laboratory reference method (Fig. 2). For both pHw and pHb, the mean values varied noticeably between the two approaches, with the probe sometimes overestimating and sometimes underestimating pH depending on the field. A similar pattern was observed for TotC, where the probe produced lower mean values in some fields but approached or exceeded laboratory values in others. Despite these field-specific fluctuations in mean estimates, a consistent trend emerged in the range and spread of the data: for every soil property and in every field, the laboratory method displayed a wider distribution, with larger interquartile ranges and more pronounced extremes. This broader spread indicates that the laboratory measurements were more sensitive to within-field variability, capturing a greater diversity of soil conditions than the probe. In contrast, the probe distributions were systematically narrower, suggesting a reduced ability to detect fine-scale differences among samples. Overall, the field-level comparison reinforces the earlier dataset-level findings: while the probe provides values within plausible ranges, its tendency to either overestimate or underestimate depending on the field, combined with its consistently narrower data range, highlights the higher sensitivity and discriminative capacity of the laboratory reference method for all three soil properties.

Comparison data Lab versus data Probe for all dataset

Across the three soil properties evaluated, clear differences were observed between the ChrysaLabs probe and the laboratory reference method. For pHw, the probe produced a higher mean value (6.44) than the laboratory measurement (6.11), indicating that the probe overestimated soil pH relative to the reference method. A similar pattern was observed for pHb, where the probe again reported a higher mean (7.00) compared with the laboratory value (6.75). In contrast, for total carbon (TotC), the probe yielded a substantially lower mean (1.29%) than the laboratory analysis (2.17%), demonstrating that the probe underestimated TotC in this dataset. When comparing variability, the coefficient of variation (CV) was consistently higher for the laboratory method across all three properties: pHw (10.45% vs. 8.93%), pHb (5.38% vs. 3.99%), and TotC (36.11% vs. 28.55%). This pattern suggests that the laboratory method exhibited greater sensitivity to differences among samples, capturing a wider range of variability than the probe. Overall, these results show that while the probe provided values within the expected range for each property, it tended to overestimate pH and underestimate TotC, and its lower CV values indicate a reduced capacity to detect fine-scale variability compared with the laboratory reference method.

Because of the variability observed between laboratory and probe measurements, we empirically selected Fields 6 and 7 as the test dataset for both pH properties. Together, these two fields captured a broad portion of the pH variability present in the full dataset. Across Fields 6 and 7, pHw values ranged from 4.72 to 7.61, while pHb values ranged from 5.86 to 7.36, which are close to the minimum and maximum values reported in the overall dataset (Table 1). Selecting test fields that span the full observed range ensures that model performance is evaluated under representative low and high pH conditions, rather than being biased toward a narrower subset of the distribution.

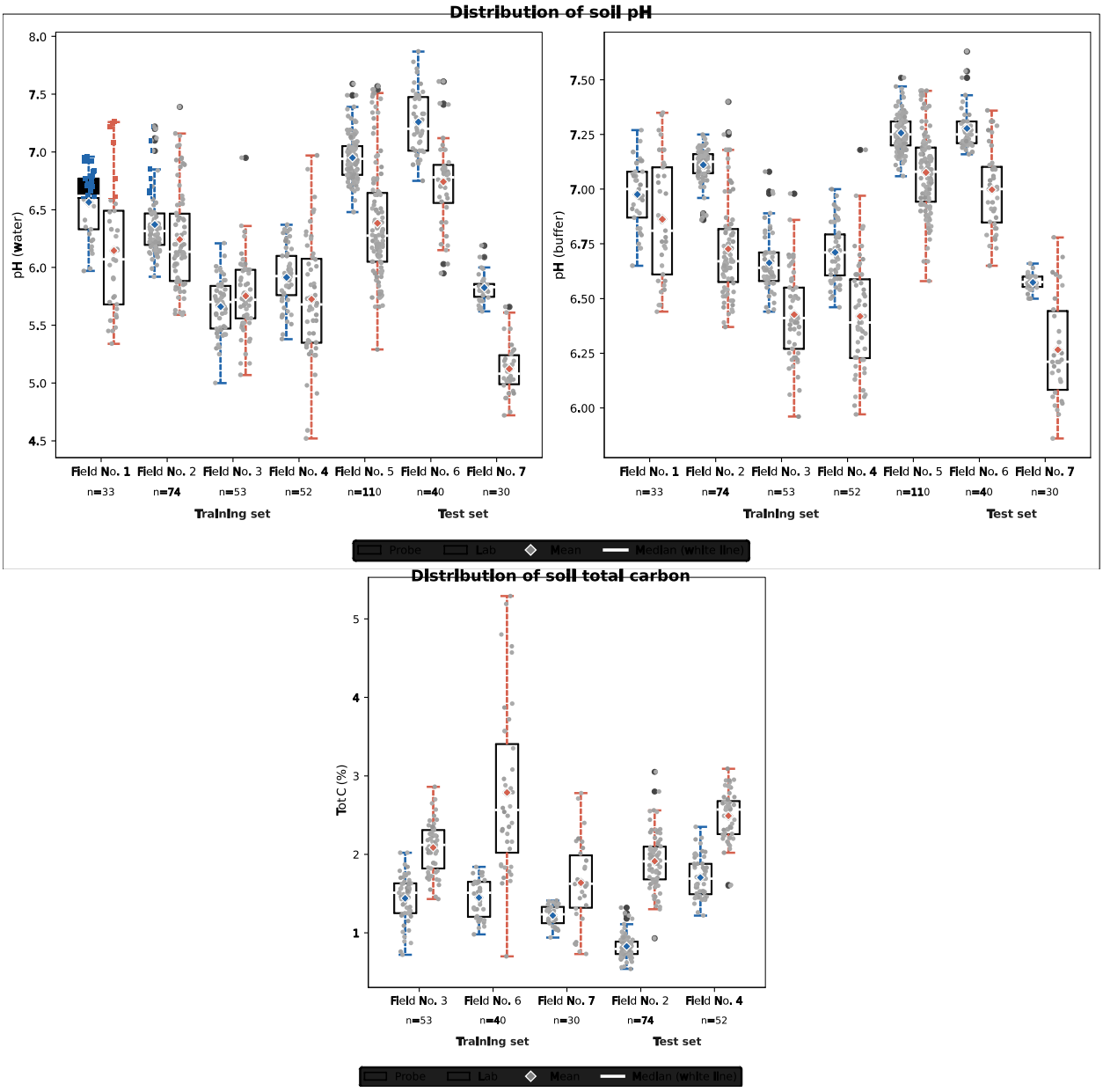

Fig. 2 Distribution of soil pH (water), pH (buffer), and total carbon and dataset split information per field and laboratory analysis (Lab) and the ChrysaLabs probe (Probe).

Table 1. Descriptive statistics of the final dataset for the traditional laboratory analysis (Lab) and the ChrysaLabs (Probe) approaches for the three soil properties (pH n=392, TotC n=249).

Soil property	Approach	Descriptive statistics							
		mean	std	min	25%	50%	75%	Max	CV (%)
pHw	Probe	6.44	0.59	5.00	5.93	6.39	6.93	7.87	8.93
	Lab	6.11	0.62	4.52	5.69	6.09	6.49	7.83	10.45
pHb	Probe	7.00	0.28	6.44	6.71	7.11	7.22	7.63	3.99
	Lab	6.75	0.37	5.86	6.49	6.46	7.04	7.45	5.38
TotC	Probe	1.29	0.40	0.54	0.91	1.30	1.61	2.35	28.55
	Lab	2.17	0.66	0.70	1.77	2.13	2.49	5.29	36.11

For TotC, the distribution was narrower and the number of available samples smaller, requiring a different field-selection strategy for the training and test sets. Fields # 3, 6 and 7 were retained in the training set because it exhibited the widest range of TotC values in the laboratory analysis (Fig. 2), making it suitable for model calibration. For the test set, Fields # 2 and 4 were selected because their TotC values spanned a range from 0.93 to 3.09%, covering much of the variability observed in the dataset while providing independent field conditions for model evaluation. This approach ensures that both the training and test sets contain meaningful variability, which is essential for assessing the model's ability to generalize beyond the calibration field.

AutoML

Our first investigation was to test if we could achieve better correlation between the probe and lab results. Our hypothesis was that the systematic bias of the probe could be corrected using a ML model that could learn some properties of the data.

Our first attempt was to apply a simple regression model, but this approach was unable to capture the variability in the data (data not shown). We then investigated SVM models, which are semi-supervised trained models able to model non-linear relations using different kernel functions such as RBF or the Laplacian kernel (Chen et al. 2025). We first attempted to use human optimized intuition (kSVM) to select the best generalizable models, before moving to AutoML methods using the sklearn and Optuna platforms which are dedicated to the selection of the best models (Table 2).

General Model performance

For pHw, predictive performance was moderate across all configurations, with baseline R^2 values ranging from 0.369 for the human-optimized SVR Laplacian model to 0.590 for the sklearn SVR RBF model (Table 2). Across both the human and AutoML platforms, the RBF kernel consistently produced the strongest pHw predictions, and the performance of sklearn and Optuna models was generally comparable (R^2= 0.472 to 0.590). For pHb, model performance was substantially higher. Support vector regression (SVR) models achieved the strongest results overall, with baseline R^2 values ranging from 0.461 (human-optimized SVR Laplacian) to 0.741 (sklearn SVR Linear), the latter representing the highest R^2 obtained across all soil properties and configurations. Within pHb, the linear kernel performed best for both platforms, with Optuna achieving an R^2 of 0.720, confirming the strong predictability of this property relative to pHw and TotC. In contrast, TotC was far more difficult to predict. Baseline R^2 values ranged from near zero for the human-optimized SVR Laplacian model (0.027) to a maximum of only 0.275 for the sklearn SVR Laplacian model. Unlike the pH properties, the Laplacian kernel under sklearn outperformed both the linear and RBF kernels for TotC, highlighting a different response of this property to kernel choice.

These results collectively support the application of SVM regression for correcting probe-based pH measurements. Similar findings were reported by Shin et al. (2025), who used related models to denoise VIS-NIR spectral data for predicting soil properties such as organic carbon. The weak performance observed for TotC in our study likely reflects the limited size and narrow distribution of the dataset, which also constrained the use of more advanced machine-learning approaches such as neural networks (Shin et al. 2025). The differential predictability among soil properties is consistent with previous studies and reflects the inherent challenges of *in situ* soil measurement (Viscarra Rossel et al., 2006; Stenberg et al., 2010).

Data augmentation

With good predictive performance analysis using the AutoML platforms for the pH, we next tried to verify if calibration of the model with sample augmentation (addition of some known lab data for the test fields) would improve the correlation from the test fields for the three soil properties. Data augmentation with additional point is a recognized strategy for improving the local-scale accuracy of VIS-NIR calibrations by enriching the calibration set with a few samples from the

target site (Wetterlind & Stenberg, 2010; Guerrero et al. 2014).

Table 2. Machine Learning results for the different soil properties including the inclusion of new sampling field points

Soil property	Platform	Model	Baseline R^2	Best Addition (% add. samples)	Best R^2	Improvement
pHw	Human-optimized	SVR Laplacian	0.369	-	-	-
	sklearn	SVR Laplacian	0.493	0	0.493	0%
	sklearn	SVR Linear	0.493	5	0.501	2%
	sklearn	SVR RBF	0.590	0	0.590	0%
	Optuna	SVR Laplacian	0.443	20	0.490	10%
	Optuna	SVR Linear	0.460	5	0.472	2%
	Optuna	SVR RBF	0.477	10	0.500	5%
pHb	Human-optimized	SVR Laplacian	0.461	-	-	-
	sklearn	SVR Laplacian	0.630	15	0.631	0%
	sklearn	SVR Linear	0.741	15	0.746	1%
	sklearn	SVR RBF	0.687	10	0.700	2%
	Optuna	SVR Laplacian	0.666	10	0.672	1%
	Optuna	SVR Linear	0.720	15	0.730	1%
	Optuna	SVR RBF	0.690	10	0.693	0%
TotC	Human-optimized	SVR Laplacian	0.027	-	-	-
	sklearn	SVR Laplacian	0.275	0	0.275	0%
	sklearn	SVR Linear	0.185	20	0.252	36%
	sklearn	SVR RBF	0.220	20	0.240	8%
	Optuna	SVR Laplacian	0.113	20	0.289	155%
	Optuna	SVR Linear	0.211	30	0.309	46%
	Optuna	SVR rbf	0.187	15	0.316	69%

For all models, addition of some calibrated points from the test field resulted in small improvement for pH (1-10%) to moderate improvements for TotC (8-66%) (Table 2 and Fig. 3). More specifically, for pHb, the best sklearn model (SVR Linear) improved by +1% to (R^2 = 0.746) with 15% additional test samples, while the best Optuna model (SVR Linear) also showed an improvement of +1% to (R^2 = 0.730) with 15% additional test samples. For pHw, the best sklearn model (SVR RBF) showed no improvement following sample augmentation, whereas the best Optuna model (SVR Laplacian) demonstrated substantial improvement from (R^2 = 0.443) to (R^2 = 0.490) with 20% additional test samples (+10%). Predictions of TotC exhibited the highest improvement with the best sklearn model (SVR Laplacian) achieving (R^2 = 0.275) at baseline and the best Optuna model (SVR RBF) improving (+41%) from (R^2 = 0.187) to (R^2 = 0.316) with 15% additional test samples.

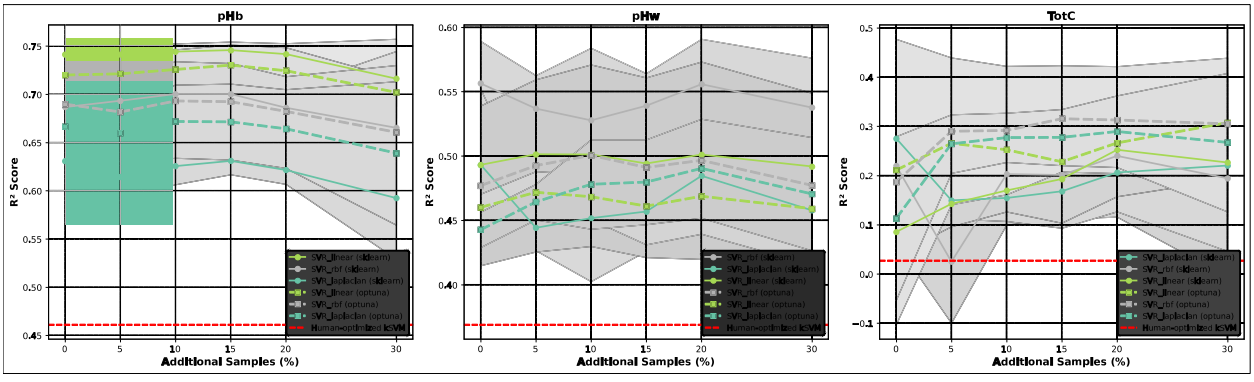

Fig. 3 Effect of sample augmentation on the SVM models performance (R^2) against the baseline. Shaded area represents the standard deviation (SD) of five independent retraining of the models (N=5)

To evaluate whether model calibration could be improved by incorporating a small number of known laboratory samples from the test fields, we applied a data-augmentation strategy in which 0–30% of the test-set samples were progressively added to the training dataset. This approach is widely recognized for improving local-scale VIS–NIR calibration performance by enriching the training set with representative samples from the target site (Wetterlind & Stenberg, 2010; Guerrero et al., 2014). For pHw, augmentation produced limited gains. The best sklearn model (SVR RBF) showed no improvement beyond its baseline performance (R^2 = 0.590), while the best Optuna model (SVR Laplacian) improved from R^2 = 0.443 to 0.490 with 20% added samples (+10%) (Table 2 and Fig. 3). These modest gains suggest that the initial calibration already captured sufficient variability for pHw, allowing the models to generalize reasonably well to new fields. For pHb, augmentation effects were similarly small. The best sklearn model (SVR Linear) increased slightly from R^2 = 0.741 to 0.746 with 15% added samples (+1%), while the best Optuna model (SVR Linear) improved from R^2 = 0.720 to 0.730 with the same augmentation level (+1%). These minimal improvements indicate that pHb was already highly predictable from the original calibration set, consistent with its strong baseline performance. In contrast, TotC showed the largest response to data augmentation. The best sklearn model (SVR Laplacian) remained unchanged at R^2 = 0.275, but several Optuna models exhibited substantial gains. The strongest improvement was observed for the Optuna SVR RBF model, which increased from R^2 = 0.187 to 0.316 with 15% added samples (+69%). These results align with previous findings that augmentation is most beneficial when baseline accuracy is low and when the calibration dataset is limited in size or variability (Guerrero et al., 2014). In our case, the narrow TotC distribution and smaller sample size constrained initial model performance, making TotC particularly responsive to augmentation.

Overall, these results show that while AutoML platforms are effective for calibrating probe-based pH measurements, their performance for TotC is strongly dependent on dataset size and representativeness. Although AutoML offers advantages in automated hyperparameter optimization, it lacks domain-specific knowledge and may generate models that do not generalize well without careful validation (Karmaker et al., 2021). By validating models across independent fields, we demonstrated that data augmentation can meaningfully improve TotC predictions, whereas pH properties require little additional information to achieve stable performance. Future work could explore spatially informed augmentation strategies and more complex ML algorithms available within AutoML frameworks, such as Random Forests (Hengl et al., 2021), to further enhance model robustness.

Agronomic impacts of the study

The agronomic appeal of the ChrysaLabs probe lies in its ability to dramatically increase sampling density and spatial coverage at relatively low operational cost, enabling finer-scale soil mapping than is feasible with conventional laboratory analyses alone. In principle, such high-resolution information can support more precise lime and fertilizer recommendations, particularly for pH management and soil carbon monitoring in precision agriculture (Adamchuk et al., 2004; Viscarra Rossel & McBratney, 2008; Ge et al., 2019). However, our results show that uncalibrated probe measurements systematically overestimate soil pH and underestimate total carbon relative to laboratory reference values, with narrower within-field variability and reduced sensitivity to local soil differences. These biases, if used directly for agronomic decisions, could lead to under- or over-liming and misestimation of soil organic carbon stocks, thereby affecting nutrient availability, crop response, and long-term soil health (Wetterlind et al., 2013; Skjemstad & Baldock, 2008).

The observed discrepancies between probe and laboratory data, together with the strong field-specific effects on model performance, highlight the need for site-specific calibration strategies. Our augmentation experiments confirm that adding a relatively small fraction of local laboratory samples (5–30%) to the training set substantially improves the alignment between probe and lab measurements, especially for TotC, consistent with previous work on spiked VIS–NIR calibrations and local spectral enrichment (Wetterlind & Stenberg, 2010; Guerrero et al., 2014). In this context,

ML, particularly SVM-based and AutoML-optimized models, acts as a bridge between rapid *in situ* sensing and robust agronomic interpretation, correcting systematic biases and enhancing predictive reliability (Padarian et al., 2019; Ng et al., 2020; Ramirez-Lopez & Schmidt, 2020; Hengl et al., 2021). Taken together, these findings suggest that portable probes like ChrysaLabs can be agronomically valuable tools when embedded in a workflow that couples targeted laboratory calibration with ML-based corrections, rather than being used as standalone replacements for reference analyses.

Conclusion

Overall, these results show that AutoML-based hyperparameter optimization substantially improves the correspondence between probe-based soil measurements and laboratory reference data, with Optuna models being the most responsive to additional calibration samples. The study also demonstrates that reliable pH correction models can be strengthened when a small number of field-specific laboratory samples are available, supporting the use of portable probes in precision agriculture.

The systematic biases observed in uncalibrated probe measurements, combined with their reduced sensitivity to within-field variability, highlight the importance of integrating portable sensing tools with targeted laboratory calibration. Our augmentation experiments confirm that adding even 5–30% of local samples can meaningfully enhance model performance, particularly for TotC.

Taken together, these findings indicate that portable spectroscopic probes such as the ChrysaLabs system can provide agronomically useful, high-resolution soil information when embedded within an ML-assisted calibration workflow. Future work should explore spatially informed augmentation strategies and additional ML algorithms available within AutoML platforms to further improve model robustness.

Acknowledgement

This work was funded by Agriculture and Agri-Food Canada (J-003174). We thank Andrée-Dominique Baillargeon, Camille Lambert-Beaudet for their technical support in laboratory and field work.

References

Adamchuk, V. I., Hummel, J. W., Morgan, M. T., & Upadhyaya, S. K. (2004). On-the-go soil sensors for precision agriculture. *Computers and Electronics in Agriculture*, doi:10.1016/j.compag.2004.03.002

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631), doi: 10.1145/3292500.333070

Albuquerque, J. R. de P., Makara, C. N., Ferreira, V. G., Brazaca, L. C., & Carrilho, E. (2024). Low-cost precision agriculture for sustainable farming using paper-based analytical devices. *RSC Advances*, doi:10.1039/d4ra02310b

Barbudo, R., Ventura, S., & Romero, J. R. (2023). Eight years of AutoML: categorisation, review and trends. *Knowledge and Information Systems*, 65(12), 5097-5149. doi:10.1007/s10115-023-01935-1

Carter, M. R., Gregorich, E. G. (2008). Soil sampling and methods of analysis. CRC press.

Chen, H., Xie, X., & Li, D. (2025). Semi-supervised learning with deep laplacian support vector

machine. Pattern Analysis and Applications, 28(1), 14. doi:10.1007/s10044-024-01395-5

Chrysalabs (2023). Optical probe and method for real-time and in situ measurements of soil properties. Retrieved December 20, 2024, from https:// www.chrysalabs.com

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., & Hutter, F. (2022). Auto-sklearn 2.0: Hands-free automl via meta-learning. Journal of Machine Learning Research, 23(261), 1-61.

Ge, Y., Morgan, C. L. S., & Grunwald, S. (2019). Soil sensing technologies for precision agriculture. *Soil Science Society of America Journal*, doi:10.2136/sssaj2018.09.0339

Guerrero, C., Stenberg, B., Wetterlind, J., Viscarra Rossel, R. A., Maestre, F. T., Mouazen, A. M., et al. (2014). Assessment of soil organic carbon at local scale with spiked NIR calibrations: effects of selection and extra-weighting on the spiking subset. European Journal of Soil Science, 65(2), 248-263. doi:10.1111/ejss.12129

Hendershot, W., Lalande, H., Duquette, M. (2008). Soil reaction and exchangeable acidity. Soil Sampling and Methods of Analysis, 16. P. 173-173.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2021). Random forest as a generic framework for predictive modeling of soil properties with automated hyperparameter tuning. *Soil*, doi:10.5194/soil-7-1-2021

Karmaker, S. K., Hassan, M. M., Smith, M. J., Xu, L., Zhai, C., & Veeramachaneni, K. (2021). Automl to date and beyond: Challenges and opportunities. Acm computing surveys (csur), 54(8), 1-36. doi:10.1145/3470918

Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., & McBratney, A. B. (2020). Convolutional neural network for soil spectral modelling. European Journal of Soil Science, 71(6), 1048–1066. https://doi.org/10.1111/ejss.12938

Padarian, J., Minasny, B., & McBratney, A. B. (2019). Using deep learning for digital soil mapping. SOIL, 5(1), 79–89. doi:10.5194/soil-5-79-2019

Ramirez-Lopez, L., & Schmidt, K. (2020). Deep learning and automated machine learning for soil spectroscopy: State-of-the-art and future directions. *Geoderma*, doi:10.1016/j.geoderma.2020.114251

Reyna Bowen, J. L., Vera Montenegro, L., & Delgado Moreira, M. I. (2025). Optimizing soil analysis in precision agriculture: Evaluating alternative methods for SOC prediction. *Journal of Ecological Engineering*, doi:10.12911/22998993/195882

Shin, S. K., Lee, S. J., & Park, J. H. (2025). Prediction of soil properties using vis-nir spectroscopy combined with machine learning: A review. *Sensors*, *25*(16), 5045. doi.org/10.3390/s25165045

Skjemstad, J. O., Baldock, J. A., (2008). Total and organic carbon. Soil sampling and methods of analysis, 2008, p. 225-237. In M. R. Carter & E. G. Gregorich (Eds.), *Soil Sampling and Methods of Analysis* (2nd ed., pp. 225–237). CRC Press, Boca Raton, FL.

Viscarra Rossel, R. A., & McBratney, A. B. (2008). Diffuse reflectance spectroscopy as a tool for soil analysis. *Geoderma*, doi:10.1016/j.geoderma.2007.09.004

Wetterlind, J., Stenberg, B., & Söderström, M. (2013). The use of near infrared spectroscopy to improve soil mapping at the farm scale. *Precision Agriculture*, doi:10.1007/s11119-013-9313-9

Ziadi, N., Tran, T.S. (2008). Lime requirement. Soil Sampling and Methods of Analysis, 12. P. 129-134. In M. R. Carter & E. G. Gregorich (Eds.), *Soil Sampling and Methods of Analysis* (2nd ed., pp. 225–237). CRC Press, Boca Raton, FL.