

A Composite Field-Evaluation Framework for Digital Agriculture Maps

Phillip Lanza

Cornell University

David G. Rossiter

ISRIC-World Soil Information

Diana S. Sinton

Cornell University

Louis Longchamps

11928@cornell.edu

Cornell University

Research Article

Keywords: digital agriculture, field evaluation, crop scouting, spatial decision support, observer bias

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11283264/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. Phillip Lanza is co-founder and Chief Technology Officer of Overfield Labs, Inc., holds equity in the company, and has a commercial interest in applications related to STADS. Cornell University owns intellectual property related to STADS, which is managed by the Cornell Center for Technology Licensing. Phillip Lanza and Louis Longchamps are inventors of this intellectual property and may be eligible to receive a share of licensing revenue. The remaining authors declare no relevant financial or non-financial interests.

Abstract

Purpose: Evaluating maps from digital agriculture systems requires limiting biased observations while providing enough spatial guidance to inspect mapped features accurately. This study developed a composite framework combining output-blinded and system-output-guided scouting as complementary evidence streams.

Methods: The framework combines output-blinded paired-point scouting with masked agronomic review and high-precision, system-output-guided scouting with the output visible. The blinded stream tests whether observations support feature presence and selected attributes at mapped locations while limiting expectation bias; the guided stream assesses the same correspondence with greater spatial precision. Agronomic reviewers use prespecified rules to convert subjective scouting observations into standardized determinations of feature presence and attribute agreement. The framework was demonstrated using 2025 Spatiotemporal Anomaly Detection System (STADS) crop-anomaly maps from New York.

Results: The composite evaluation comprised separate output-blinded and system-output-guided evidence sets. Of 99 output-blinded point pairs scouted, 55 were retained after excluding observations that could not be confidently matched to original locations; 53 had at least two valid reviewer classifications for primary analysis, and 18 (34.0%) supported both anomaly presence and the predicted direction. Of 85 system-output-guided real-time kinematic global navigation satellite system (RTK-GNSS) pairs originally planned, 57 were retained after exclusions and 53 were analyzable; 37 (69.8%) supported both feature presence and the evaluated attribute. Framework implementation identified practical requirements for record linkage, masking, sampling, positioning, and observation timing.

Conclusion: Combining complementary output-blinded and system-output-guided evidence provides a transferable and transparent approach for evaluating mapped agricultural outputs when complete field-scale ground truth is impractical.

Key Points

- Several spatial digital agriculture tools map subfield features such as management zones and anomalous crop responses. Before these maps are used for interpretation or management, they should be evaluated through targeted field observations conducted by scouts to determine how well they represent actual field conditions.
- Field scouting also introduces uncertainty. An evaluation protocol could reduce observer bias while ensuring that scouts inspect the intended mapped locations reliably.
- We developed a composite field-evaluation framework that combines blind paired-point scouting with high-precision, system-output-guided scouting so that the strengths of each approach address limitations of the other.

- Output-blinded scouting tests whether crop differences associated with mapped features can be observed without scouts seeing the system output, minimizing expectation bias from prior knowledge of the mapped features.
- Output-guided high-precision real-time kinematic global navigation satellite system (RTK-GNSS) scouting allows evaluators to view the system output and directly compare conditions inside and outside mapped features, improving spatial interpretation but increasing the potential for observer bias.
- Multiple agronomic reviewers separately interpreted descriptive blind-scouting observations using prespecified criteria to classify mapped-feature presence and selected output attributes.

Introduction

Digital agriculture tools produce maps that identify areas within fields for further investigation or targeted management. Examples include crop-condition alerts, anomaly maps, and subfield management zones. Before relying on these maps, farmers and agronomists need evidence that the mapped data that these tools generate correspond to actual and relevant field conditions. Obtaining this evidence is difficult because crop conditions vary across fields and change over time. Some differences are visible only at particular growth stages or under certain environmental conditions, and comprehensive observations across entire fields are rarely available (Pierce & Nowak, 1999; Plant, 2001; Maestrini & Basso, 2021; Trevisan et al., 2021).

Field evaluation must also account for uncertainty in the observations used to assess a map. Scouts must reach the intended locations, observe relevant crop differences, and record enough information to compare their observations with the mapped prediction. Their judgments may be influenced by what they expect to find. A useful evaluation protocol must therefore address both where observations are made and how they are interpreted. Evaluation approaches for mapped agricultural outputs have included field phenotyping (Araus & Cairns, 2014), crop scouting (Shaw & Kelley, 2005), proximal sensing (Tagarakis & Ketterings, 2017), remote sensing (Weiss et al., 2020), yield-monitor data (Cho et al., 2021; Maestrini & Basso, 2021), and integrated agronomic datasets (Bullock et al., 2019; Leroux et al., 2019; Trevisan et al., 2021).

These approaches provide different kinds of evidence. Field observations help interpret crop conditions but require substantial labor and involve observer judgment (Bock et al., 2020, 2022). Sensor measurements and remote-sensing products can cover larger areas, but interpreting a mapped pattern may still require information about pests, diseases, nutrient status, management, or environmental conditions (Pancorbo et al., 2021; Weiss et al., 2020). The choice of evaluation evidence should therefore reflect what the map represents and which aspects can be checked in the field.

Showing scouts a map helps them locate the predicted feature and compare it with surrounding conditions. However, knowing what the map predicts may influence where scouts look and how they interpret uncertain observations (Nickerson, 1998; Hróbjartsson et al., 2012). Withholding the map

reduces this source of expectation bias but prevents scouts from using the mapped boundaries and predicted pattern to guide their inspection. The two approaches therefore answer related but different questions: can scouts observe the predicted difference without seeing the map, and do field observations support the mapped feature when the map guides inspection?

This study developed a field-evaluation framework that combines these two approaches. In blind paired-point scouting, scouts visit one location inside a mapped areal feature and a nearby comparison location outside it. They do not see the feature map and are not told which point is inside the areal feature. Agronomic reviewers then assess the scouting records without seeing the map or knowing which observation represents the mapped feature. In contrast, in output-guided scouting, evaluators view the map and use high-precision positioning to inspect conditions inside and outside the areal feature. Each of these two approaches assesses whether scouted observations confirm the presence of the mapped feature and, where relevant, an additional characteristic, such as whether crop condition is better or worse than in the comparison area. Results from the two approaches are reported separately and interpreted together.

This two-approach framework was applied to crop-anomaly maps produced by the Spatiotemporal Anomaly Detection System (STADS). STADS identifies areas within fields where crop condition differs from a modeled reference response under local spatial and temporal conditions. Field evaluation examined whether these areas showed observable crop differences and whether the nature of those differences agreed with the map. A mapped anomaly was treated as a prediction to be evaluated rather than as an established field condition. The evaluation did not seek to determine the cause of each anomaly or reproduce the methods used to generate the maps.

The objectives were to develop a field-evaluation protocol combining blind and output-guided scouting and to demonstrate its application using STADS anomaly maps. The case study also identified practical requirements for collecting, linking, and interpreting field observations. Because the two scouting approaches evaluated different anomaly samples, the study assessed their complementary contributions rather than testing which approach performed better.

Methods

Composite field-evaluation framework

The two-approach framework addresses two questions. First, can scouts observe differences associated with a mapped feature without seeing the map? Second, when scouts use the map to guide inspection, do field observations confirm the feature and its predicted characteristics? Blind paired-point scouting addresses the first question by withholding the feature map and concealing which point is inside the feature. Output-guided scouting addresses the second by combining the visible map with high-precision positioning. The approaches provide complementary evidence and are analyzed separately

because observations are collected under different conditions. Figure 1 outlines the framework, and Table 1 summarizes the design requirements and expected strengths and limitations of each approach.

Table 1
Comparison of the two evidence streams in the composite field-evaluation framework

Design element	Blind paired-point stream	Output-guided stream
Primary evaluation target	Independent field observability	Targeted spatial correspondence and operational interpretability
Output visibility	Mapped output and target membership concealed	Mapped output and real-time position visible
Observer task	Compare paired locations using a standardized blind record	Inspect the mapped feature and adjacent reference area directly
Navigation	Positioning adequate to reach concealed paired points	Positioning precision matched to output resolution and boundary scale
Expected strength	Reduced expectation bias	Greater spatial control and direct map interpretation
Expected limitation	Less control over exact placement and feature extent	Susceptibility to expectation and confirmation bias
Required integrity controls	Randomized labels, persistent identifiers, complete coordinates, masked review	Persistent identifiers, achieved-accuracy records, direct georeferencing, structured comparisons
Interpretation	Evidence that the mapped difference is independently observable	Evidence of correspondence under an output-informed workflow

Output-blinded paired-point evidence stream

Blind scouting compares two point locations: a target inside a mapped areal feature and a nearby reference outside it. The two locations should have similar crop, management, and environmental conditions so that unrelated differences are less likely to affect the comparison. Each point should also be far enough from the mapped boundary to reduce uncertainty about whether the observation falls inside or outside the feature. When these conditions are met, the two point locations serve as an “evaluable pair.” Scouts visit both point locations without having seen the areal feature map or knowing which of the two point locations is the target. The primary outcome is the proportion of paired observations for which the field evidence supports a difference at the target location relative to the nearby reference location.

Blind scouting can use high-precision positioning, including real-time kinematic global navigation satellite system (RTK-GNSS) devices, to guide a scout to a location while not revealing the mapped areal feature. However, the cost of equipping every scout with such a receiver may be prohibitive for routine scouting operations. The protocol therefore accommodates consumer-grade GNSS devices, provided

that their positioning performance is adequate for the size of the mapped feature and the planned observation locations. When positioning is less accurate or precise, points should be placed farther from mapped boundaries to reduce the risk of observing the wrong area. To be evaluated reliably, small or narrow features may require higher-precision equipment.

When scouting records contain text descriptions rather than directly comparable quantitative measurements, multiple agronomic reviewers assess each pair separately using predefined criteria. Prespecified rules define the minimum number of valid reviews and the agreement required to classify each pair. These classifications indicate whether the observations support the mapped areal feature and any additional characteristic being evaluated (Longchamps et al., 2014, 2016). Categories allow observations from different crops and field conditions to be assessed consistently; numerical measurements can be used when they can be compared directly with the mapped prediction. The protocol should specify what information reviewers can see, which records can be evaluated, how classifications are combined, and how missing responses are handled. Figure 2a illustrates the blind scouting approach.

Output-guided high-precision evidence stream

In output-guided scouting, evaluators view the mapped feature and their real-time position while comparing conditions inside the feature with a nearby reference area outside it. The reference area should have similar crop, management, and environmental conditions. Evaluators assess whether the observed characteristics match with conditions within the mapped areal feature and its predicted characteristics. Because seeing the map may influence their judgments, the results describe agreement under guided inspection and remain subject to expectation bias.

Positioning must be accurate enough to distinguish the target and reference locations relative to the mapped boundary. Requirements depend on the feature's size and shape, the distance of each point from its boundary, and uncertainty in both the mapped boundary and the recorded position. For maps derived from raster data, pixel size is relevant but does not alone determine the required positioning accuracy. Figure 2b illustrates the output-guided approach.

Composite synthesis

Results from each scouting approach are first assessed using its predefined rules to determine whether field observations support the presence of the mapped feature. If the map predicts an additional characteristic, such as whether the crop is experiencing high weed pressure relative to the reference area, agreement with that prediction is assessed separately.

These assessments are combined into three categories. Full correspondence means that observations support both the feature's presence and the additional predicted characteristic. Presence-only correspondence means that observations support the feature's presence, but the additional

characteristic is inconsistent with the prediction or cannot be determined. No correspondence means that observations do not support the feature's presence. Any records missing information needed for a particular assessment are excluded from that assessment but remain eligible for others.

When the two scouting approaches use different samples or observation procedures, their percentages are reported separately. The results provide complementary evidence about the presence and characteristics of mapped features but do not establish whether one scouting approach performs better than the other.

STADS case study design

Evaluated outputs and study area

STADS was used as the case-study system to demonstrate the two-approach evaluation framework during normal cropping operations in New York in 2025. STADS produces mapped polygon features representing crop anomalies, which are here defined as subfield areas differing from a modeled reference crop response beyond a defined threshold. These features may also contain attributes describing characteristics of the detected difference. This study evaluated whether those mapped features and selected attributes corresponded to field observations, while the proprietary methods used to generate them were outside the scope of the study. Non-confidential attributes used in the evaluation included crop, anomaly area and perimeter, detection date, and a signed severity score. The score-to-direction mapping was used to evaluate directional agreement and the composite correspondence endpoint; calculation of the proprietary score itself was outside the scope of the study.

Eight monitored regions, each approximately 400 km², represented a range of crop-production conditions across New York. Five regions were allocated to output-blinded paired-point scouting by external crop consultants. Three separate regions were allocated to system-output-guided scouting by the research team using real-time kinematic global navigation satellite system (RTK-GNSS) positioning. Because the two evidence streams evaluated different regions and different anomaly samples, the study was not designed to estimate the effect of blinding, positioning technology, or output visibility on evaluation results.

Sampling and paired-location generation

The lead author selected candidate field anomalies in QGIS 3.42.0 (QGIS Development Team, 2009) while viewing polygon geometry, crop, region, detection date, apparent size and severity, field boundaries, and nearby comparison areas. Selection sought diversity across crops, regions, dates, sizes, and severities, subject to field access and the availability of a suitable non-anomalous comparison area. No fixed sampling fractions or quotas were used. Accordingly, the resulting sample was a diversity-oriented convenience sample and was not statistically representative of all anomalies identified via STADS.

Preference was given to polygons larger than 9 m², equivalent to one 3 m by 3 m pixel, to reduce sensitivity to source-imagery and navigation error. One point was placed inside each selected polygon and one nearby point was placed outside it. Where feasible, points were positioned approximately 10 m inside and 10 m outside the boundary. For polygons smaller than 3m x 3m, the internal point was placed near the centroid. External points were placed outside the selected anomaly and approximately 6 m or more from the other mapped anomalies.

Output-blinded field protocol

Ninety-nine blind point pairs were provided to an independent crop consulting firm. Each pair contained one point within a field anomaly identified via STADS and one nearby point outside it. Points were transferred to the firm's system within 7 days of detection, although consultants visited them according to operational schedules and were not required to scout within 7 days.

Experienced crop consultants used their standard procedures and consumer-grade GNSS to reach the assigned points. They were not shown the anomaly map layer and were not told which point was inside the mapped anomaly. Structured records included scouting notes, crop information, plant stage when available, and georeferenced photographs for many of the point locations. Georeferenced photographs provided a convenient means of recording location coordinates for scouting observations. Device-level information and achieved positional accuracy were not retained. De-identified observations are provided in Online Resource S1.

Output-guided RTK-GNSS field protocol

The research team also evaluated STADS outputs while viewing paired points, mapped anomalies, and real-time field position. Eighty-five point pairs were originally planned. The lead author conducted the observations with assistance from a trained field intern for a subset of locations. An Arrow Gold GNSS receiver (Eos Positioning Systems, Inc., Terrebonne, QC, Canada) connected to New York State Department of Transportation real-time corrections (NYSNet) was used for navigation. The receiver could provide centimeter-level positions when it achieved a fixed RTK solution, but complete achieved-accuracy and solution-status records were not retained for every observation. Observations, coordinates, and photographs were recorded in ArcGIS Survey123 (Environmental Systems Research Institute, 2024). The data collection form and de-identified observations are provided in Online Resource S2.

At each point pair, scouts compared crop conditions inside the mapped anomaly with the external comparison area. Structured fields described crop height, health, growth stage, appearance, and visible pest, disease, weed, nutrient, management, or other conditions. Scouts classified the observed crop response as better (positive deviance), worse (negative deviance), the same, or indeterminate relative to the comparison area. The protocol evaluated anomaly presence and direction rather than causal diagnosis or correspondence with a numerical severity score. Representative field examples are provided in Online Resource S7.

Observation linkage and quality control

Output-blinded records

All 99 planned output-blinded point pairs were scouted, but 44 were excluded because the returned observations could not be linked reliably to their original target-reference pair and corresponding STADS feature. Because linkage depended on positional information from georeferenced photographs, missing photographs left insufficient location data to determine where the returned observations had been made and therefore to associate them reliably with their prescribed scouting points. The resulting exclusions therefore represent record-linkage failures rather than field non-confirmations. These linkage failures occurred at the pair level: 55 pairs were fully reconstructed, whereas none was only partially reconstructed. Among the 110 observations in the 55 retained pairs, 96 were linked directly using spatial evidence and 14 required supporting metadata or point-label information. For direct spatial matches, the median distance between the planned point and matched observation coordinate was 3.2 m (interquartile range 1.6-6.0 m). These distances represent record-matching distances rather than measurements of device positional accuracy.

Linkage success was not evenly distributed across the original sample. Thirty-six of 41 corn-source pairs (87.8%) were retained compared with 19 of 58 soybean-source pairs (32.8%), and retention also varied among sampling periods, farms, and imagery dates. Linkage showed little association with mapped anomaly direction or planned target-reference separation. However, uneven retention means the analyzed subset may not fully represent the original sample, so linkage-related selection bias cannot be excluded. Detailed reconstruction procedures, audit results, and retained-versus-excluded comparisons are provided in Online Resource S3.

Output-guided records

Output-guided records contained coordinates and original point-pair identifiers. Exported records were reviewed against source points, anomaly polygons, field boundaries, and crop metadata. Point-in-polygon overlay verified inside-outside assignments. No completed pair was excluded because of uncertain source linkage.

Of the 85 originally planned pairs, 27 were not scouted: 12 were associated with an incorrect crop because farmer-reported crop information used during planning was outdated or inaccurate for several fields, 9 were in active mining area with restricted access, 1 was inaccessible because of a flooded road, and 5 were missed. One replacement pair was added, yielding 59 completed pairs. Two completed non-target pairs, one clover and one grass location, were excluded. Fifty-seven pairs remained for analysis. The RTK-GNSS record-reconciliation manifest and associated audit materials are provided in Online Resource S8.

Masked reviewer classification of output-blinded observations

The 55 retained output-blinded pairs were evaluated separately by five agronomic reviewers using identical workbooks. The five-member panel included a plant-science undergraduate, a master's-trained agronomist, an assistant professor of digital agriculture, a PhD candidate in digital agronomy, and a PhD student in digital agronomy. Agronomic experience ranged from moderate to extensive. Reviewers could not see each other's classifications or the STADS output. Inter-reviewer agreement was evaluated using multiple complementary measures, including Krippendorff's alpha, Fleiss' kappa, Gwet's AC1, pairwise Cohen's kappa, and raw agreement. Two reviewers were study authors, so classifications were treated as structured agronomic interpretations rather than fully external ground truth, and reviewer-subset sensitivity analyses were used to assess the influence of reviewer composition. Detailed methods and results of these evaluations are provided in Online Resource S3.

Each workbook presented the crop, paired scouting observations, and plant stage when available. Photographs were not included because they had been collected primarily for georeferencing and did not consistently document the described crop condition. Proprietary STADS attributes, anomaly direction and severity, precise coordinates, scout identity, and farm and field identifiers were withheld from reviewer-facing records. During processing, the within-anomaly observation was placed in the primary field and the outside-anomaly observation in the comparison field, but reviewers were not informed of this assignment. This ordering was not randomized and was considered a case-study limitation. A blank reviewer workbook is provided in Online Resource S4.

Reviewers classified anomaly presence as Confirmed, Possibly confirmed, Not confirmed, Indeterminate, or Excluded, and each classification was also assigned a High, Medium, or Low confidence rating. Interpretations of each classification are shown in Table 2. Reviewers also independently classified the direction of the primary observation as Positive, Negative, or Indeterminate relative to the paired observation and recorded separate confidence. Free-text rationale fields documented each judgment. The full rubric is provided in Online Resource S5. De-identified, completed reviewer workbooks and associated processing scripts are provided in Online Resource S6.

Table 2
Blind anomaly-presence classifications and their interpretations

Classification	Indication
Confirmed	Clear agronomically meaningful difference in crop response as compared to associated reference area
Possibly confirmed	Subtle, incomplete, or ambiguous evidence
Not confirmed	No meaningful difference
Indeterminate	Insufficient or conflicting evidence
Excluded	Invalid crop condition for comparison

Endpoints and statistical analysis

Anomaly-presence consensus of output-blinded observation classifications

The point pair was the unit of analysis for output-blinded anomaly presence. A reviewer classification was valid for this endpoint when it was Confirmed, Possibly confirmed, or Not confirmed. Indeterminate and Excluded classifications were omitted. A pair was evaluable when at least two reviewers provided valid classifications.

The Moderate rule was the primary presence definition. A pair supported anomaly presence when at least 75% of reviewers selected Confirmed or Possibly confirmed, with the required count rounded upward. The effective requirements were 2 of 2, 3 of 3, 3 of 4, or 4 of 5 valid reviewers. Strict, Liberal, confidence-weighted, and complete-case alternatives were secondary sensitivity analyses and are described with their results in Online Resource S3.

Direction and composite correspondence of classifications

For the point-pair composite endpoint, Positive and Negative classifications were treated as valid direction responses. Indeterminate, Excluded, and missing responses were omitted. The unique modal response was assigned as the point-pair direction. Tied modes were treated as indeterminate. A direction match required a modal Positive or Negative direction equal to the STADS-derived direction.

Full correspondence required anomaly-presence confirmation under the Moderate rule and a direction match. A presence-confirmed pair with a mismatched, tied, or unavailable reviewer direction was classified as presence-only correspondence. A pair that did not satisfy the Moderate presence rule was classified as no correspondence regardless of its direction result. Composite evaluability required an evaluable presence determination and an available binary STADS direction; it did not require an evaluable reviewer-direction mode. The RTK-GNSS stream used the same three categories, with one field-evaluation record per anomaly.

Estimation and secondary analyses

Confidence intervals for proportion estimates were calculated using a field-cluster percentile bootstrap with replacement to account for observations from the same field not being fully independent. Rather than resampling individual observations, entire fields were resampled. For the blind stream, clusters were defined using the original field identifiers. For the RTK-GNSS-guided stream, clusters were defined using field-boundary identifiers assigned spatially. In each of 10,000 bootstrap replicates, fields were sampled with replacement and all observations from each selected field were included together. The statistic was then recalculated for each replicate. The 95% confidence interval was defined by the 2.5th and 97.5th percentiles of the resulting bootstrap distribution.

Secondary analyses examined crop-specific outcomes, alternative presence rules, reviewer agreement, reviewer confidence, complete cases, and the reviewer subset without identified competing interests.

Their methods and results are provided in Online Resource S3 because they assess robustness of the case-study classifications rather than define the composite protocol.

Data handling and reproducibility

Python 3.11.15 (Van Rossum & Drake, 2009) scripts reconstructed blind-record linkages when georeferenced photographs and supporting metadata provided sufficient evidence to associate returned observations confidently with their original target-reference pairs. Records lacking sufficient evidence were flagged as linkage failures. The scripts also standardized categorical responses, constructed point-pair-level consensus records, calculated endpoints, and generated reported summaries and figures. Internal datasets preserved source identifiers and audit fields, while public materials use anonymized identifiers and omit precise coordinates, farm and field identities, scout identity, confidential location information, and proprietary STADS attributes. Online Resources S1-S8 contain the de-identified blind scouting observations, RTK-GNSS Survey123 form and observations, blind-record reconstruction and secondary-analysis materials, blank reviewer workbook, classification rubric, completed reviewer workbooks and processing scripts, de-identified field examples, and RTK-GNSS reconciliation materials, respectively.

Results

Application of the composite protocol to STADS

Record flow and evaluability

Data produced by the two evaluation approaches had different record flows because they used different data-collection and quality-control procedures. In the output-blinded stream, all 99 blind pairs were scouted, but 44 could not be linked reliably to the intended source pair and anomaly. Fifty-five were retained and 53 met the primary presence-evaluability requirement. In the output-guided stream, 27 of 85 originally planned pairs were not scouted, one replacement was added, and two completed non-target pairs were excluded. Fifty-seven were retained and 53 were evaluable for the primary endpoint. Table 3 summarizes the record flow and primary results.

Table 3. Record flow and primary outcomes for the STADS case study

Record stage or endpoint	Blind paired-point method	RTK-GNSS-guided method
Originally planned pairs	99	85
Additional replacement pairs	0	1
Not scouted	0	27
Completed pairs excluded before analysis	44 linkage failures	2 non-target pairs
Retained pairs	55	57
Evaluable for primary presence endpoint	53	53
Presence supported under stream-specific primary rule	25 of 53	44 of 53
Full correspondence	18 of 53	37 of 53
Presence-only correspondence	7 of 53	7 of 53
No correspondence	28 of 53	9 of 53

Blind paired-point results

Under the Moderate rule, 25 of 53 evaluable pairs supported anomaly presence, corresponding to 47.2% of the blind sample (95% field-cluster bootstrap confidence interval [CI] 32.0% to 61.1%). Three evaluable pairs had two valid presence classifications, five had three, twelve had four, and thirty-three had five. Alternative prespecified rules produced presence-support estimates from 37.7% to 60.4% and are reported in Online Resource S3.

Of the 25 pairs supporting presence, 18 had a modal reviewer direction matching the STADS-derived direction and were classified as full correspondence. Five had a mismatched modal direction and two had tied positive and negative classifications; these seven were classified as presence-only correspondence. The remaining 28 evaluable pairs did not satisfy the Moderate presence rule and were classified as no correspondence. All 18 full-correspondence pairs had four or five valid direction classifications, and each matching modal direction represented a majority of valid direction responses.

Reviewer-level direction, agreement, confidence, and competing-interest sensitivity analyses are reported in Online Resource S3. They were not used to redefine the primary point-pair composite endpoint.

Output-guided RTK-GNSS results

Field observations confirmed 44 of 53 evaluable output-guided anomalies, corresponding to 83.0% of the guided sample (95% field-cluster bootstrap CI 76.8% to 89.8%). Among direction-evaluable records, observed direction agreed with STADS in 37 of 43 cases (86.0%; 95% CI 76.1% to 94.4%).

Thirty-seven of 53 records met both the presence and direction criteria and were classified as full correspondence. Seven were classified as presence-only correspondence, including six direction mismatches and one indeterminate direction. Nine did not confirm anomaly presence and were classified as no correspondence.

Composite case-study outcomes

Full correspondence occurred in 18 of 53 blind evaluations (34.0%) and 37 of 53 output-guided evaluations (69.8%). Presence-only correspondence occurred in 7 of 53 evaluations (13.2%) in each stream. No correspondence occurred in 28 of 53 blind evaluations (52.8%) and 9 of 53 output-guided evaluations (17.0%). Figure. 3 shows the composite distributions.

The streams evaluated separate anomaly samples in different regions and differed in evaluators, positioning, output visibility, observation timing, and record linkage. Their percentages therefore do not estimate the causal effect of the field protocol or RTK guidance.

Discussion

Methodological contribution

The composite framework separates two questions that are commonly conflated in evaluations of maps produced by digital agriculture tools. Blind paired-point scouting asks whether a mapped feature corresponds to a difference that can be documented without knowledge of the output. Output-guided scouting asks whether the feature can be located and interpreted when the output is used to direct field investigation. Neither approach alone provides complete validation. The blind method limits expectation bias but offers less spatial control. The guided method provides direct spatial interpretation but permits the output to influence observation. Preserving separate sampling frames, endpoints, and denominators makes the resulting evidence easier to interpret.

The composite categories distinguish different types of agreement. A binary confirmation rate cannot distinguish an output that identified a meaningful location but assigned the wrong direction from one that had no field-observable difference. Full, presence-only, and no correspondence separate these outcomes while retaining indeterminate direction as a conservative failure to meet the joint endpoint. The categories can be adapted to other mapped attributes, including class, severity, boundary correspondence, timing, or persistence, provided the attribute and field evidence are defined before data collection.

Multiple reviewers can convert qualitative observations into a reproducible categorical endpoint when direct measurements are unavailable, an approach that has precedent in agronomic image and field assessment in which independent expert classifications were integrated using predefined decision rules (Longchamps et al., 2014, 2016). The reviewer workflow should not, however, be described as producing

ground truth. It produces an explicit consensus from masked agronomic interpretations. Its credibility depends on the classification rubric, reviewer independence, minimum valid responses, treatment of missing ratings, aggregation threshold, and separation of presence from direction. Detailed agreement coefficients can help diagnose reviewer variation, but they are secondary to a clearly defined primary consensus rule.

Lessons from implementation and recommended protocol

Applying the protocol showed that output-blinded evaluation requires reliable traceability between planned scouting points, returned observations, and mapped features to maintain data provenance and support reliable record linkage. Forty-four of the 99 scouted pairs were excluded because observations returned through the field-scouting workflow lacked sufficient georeferenced photographs to reconstruct their association with the original target-reference points and corresponding anomaly features confidently. This high exclusion rate indicates a weakness in the data-capture and record-provenance workflow rather than evidence for or against the STADS anomaly outputs themselves. It reduced the effective sample size and, because linkage success was not evenly distributed across the original sample, may also have affected the representativeness of the retained records. Future deployments should assign immutable, machine-readable identifiers to every mapped feature, pair, point, observation, photograph, attachment, reviewer record, and analysis record. These identifiers should be preserved unchanged through external data systems, and automated checks should reject submissions containing missing, duplicate, or inconsistent identifiers.

Field forms should also capture the primary evaluation outcome directly. An output-blinded scout can report whether an A-B difference is present, whether A or B has the better crop condition, and the confidence of that judgment without being told which point is the mapped target. Narrative observations and structured agronomic fields remain useful, but direct pair-level responses reduce dependence on retrospective interpretation. If photographs will support later review, the protocol should specify framing, scale, orientation, and required views rather than rely on photographs primarily for reconstructing observation locations. Every scouting observation should include recoverable coordinates so it can be linked to its prescribed point and mapped feature. Coordinates may be recorded directly by the field form or extracted from required georeferenced photographs. Photograph metadata can therefore provide a low-burden method of collecting location information even when the images themselves are not used for later review, provided coordinates can be extracted reliably and are sufficiently accurate for the evaluation scale. The coordinate-capture method should minimize field burden while providing sufficient positional accuracy for the spatial scale of the evaluation.

Masking must continue through field collection and reviewer assessment. Random A-B labels concealed anomaly membership from scouts in this case study, but the within-anomaly observation was consistently placed in the primary reviewer field. Although reviewers were not told this ordering and their rationale did not show clear evidence that they inferred it, future workbooks should randomize the ordering of target and comparison observations in the reviewer workbook. When the same feature is

evaluated in both approaches, the output-blinded visit should occur first so that previous exposure to the mapped output cannot influence the masked assessment.

Sampling decisions determine the population to which findings apply. The purposive STADS sample provided geographic and agronomic diversity but cannot support population-wide estimates of system performance. Future applications should retain the complete candidate-feature frame, document selection and replacement, and use random sampling within prespecified strata when feasible. Null pairs in which neither location corresponds to a mapped feature are needed to estimate how often observers or reviewers identify differences in the absence of a predicted feature. A partially overlapping set of output-blinded and output-guided observations is required when direct comparison of protocols is an objective.

Positioning quality and observation timing should be recorded as protocol variables. Navigation precision should reflect feature size and output resolution, and field records should retain the device, achieved accuracy, correction status, and distance from the intended point or boundary. Describing a receiver as RTK-capable does not establish centimeter-level accuracy for every record. Similarly, points should be scouted as close as practicable to the imagery date, and the detection-to-observation interval should be preserved for every visit. These records allow investigators to distinguish disagreement with the digital output from changes in the field or inadequate observation placement.

Reviewer and analysis rules should be prespecified. Future studies should define a fixed reviewer panel or a minimum number of valid ratings that does not inadvertently change the effective consensus threshold. The STADS Moderate rule required unanimity with two or three valid ratings, three of four, or four of five. A future protocol could instead require a fixed number of valid reviewers and use the same decision threshold for every pair. Direction aggregation should likewise specify the minimum number of valid responses, required majority, and treatment of ties in classifications. Alternative rules and reviewer subsets are best used as sensitivity analyses rather than competing primary endpoints.

Finally, each evidence stream should retain its own deployment manifest, record-flow account, exclusion criteria, primary evaluation target, denominator, and uncertainty estimate. Composite synthesis should occur only after these stream-specific results are constructed. Pooling the two streams into one accuracy estimate would obscure the different observation conditions and imply comparability that the design may not support.

Interpretation of the STADS case study

The blind approach provided evidence that some STADS anomalies corresponded to differences observable through routine masked scouting. Under the primary rule, 25 of 53 (47.2%) evaluable pairs supported anomaly presence and 18 (34.0%) met the joint presence-and-direction criterion. These results should be interpreted as conditional on the retained, purposively selected sample. The

substantial linkage loss also means that the blind estimate characterizes records that could be reconstructed, not all 99 scouted pairs.

The output-guided method showed greater correspondence within its separate anomaly sample, with 44 of 53 (83.0%) observations supporting anomaly presence and 37 of 53 (69.8%) observations meeting the joint presence-and-direction criterion. Viewing the anomaly and real-time position allowed the evaluator to compare mapped and reference areas directly. The higher percentages could reflect spatial control, targeted observation, knowledge of the expected feature, differences in temporal lag, different evaluators or regions, or differences between the selected anomalies. The study cannot separate these explanations because the same anomalies were not evaluated under controlled blind and guided conditions.

Limitations and transferability

The study did not estimate false-negative rates because comprehensive field observations were unavailable for identifying anomalies that were present but not detected by STADS. The absence of null pairs prevented estimation of the background rate at which the blind evaluation would identify differences where no mapped feature was predicted. Purposive sampling may have favored larger, more visible, accessible, or severe anomalies. Incorrect boundaries and crop records, restricted access, and missed locations further shaped the guided sample. The guided approach relied primarily on one evaluator and did not retain records suitable for a separate masked reassessment. The study also did not assess numerical severity, anomaly cause, scouting efficiency, management response, or decision value.

These limitations restrict claims about STADS performance but do not prevent evaluation of the framework's design. The protocol can be adapted to crop-injury maps, pest and disease risk products, management zones, yield-stability layers, variety-performance predictions, and other maps produced by digital agriculture tools. The mapped feature, comparison unit, field-observable endpoint, and additional attribute must be defined for each application. The framework is most useful when comprehensive ground truth is impractical and both observer independence and precise spatial interpretation matter.

Conclusion

This study developed a composite field-evaluation framework that combines output-blinded scouting with output-guided high-precision scouting of paired points. The two approaches answer different evaluation questions: whether a mapped difference is observable without seeing the output, and whether field conditions correspond to the mapped feature during guided inspection. Prespecified reviewer and synthesis rules convert qualitative observations into reproducible classifications. The STADS case study demonstrated the framework and identified requirements for persistent identifiers, complete field metadata, randomized masking, documented sampling, spatial and temporal fidelity, and prespecified

reviewer aggregation. Future evaluations should incorporate these requirements before field deployment and use overlapping samples only when direct protocol comparison is intended.

Declarations

Funding

This work was supported by the Hatch Act of 1887, project accession no. 7003680 from the U.S. Department of Agriculture's National Institute of Food and Agriculture. Additional support was provided by the 2023 Research Innovation Fund of the Cornell Institute for Digital Agriculture.

Competing Interests

Phillip Lanza is co-founder and Chief Technology Officer of Overfield Labs, Inc., holds equity in the company, and has a commercial interest in applications related to STADS. Cornell University owns intellectual property related to STADS, which is managed by the Cornell Center for Technology Licensing. Phillip Lanza and Louis Longchamps are inventors of this intellectual property and may be eligible to receive a share of licensing revenue. The remaining authors declare no relevant financial or non-financial interests.

Ethics approval

In accordance with Cornell University Institutional Review Board Policy 1 and the Cornell IRB decision tree for determining whether an activity constitutes human participant research, the investigators determined that this study did not constitute human participant research. Information collected for the study concerned crop and field conditions and evaluation of the digital agriculture method rather than the individuals conducting scouting or evaluation. Accordingly, submission to the Cornell University IRB and IRB approval were not required.

Consent to participate

The study did not constitute human participant research under Cornell University IRB policy; therefore, IRB-regulated informed consent to participate was not required. Crop consultants and research personnel contributed scouting and evaluation information concerning crop and field conditions in their professional roles and were not studied as research participants.

Consent for publication

The participating crop consultants consented to the publication and sharing of their de-identified scouting reports, with precise location data and personally identifiable information removed.

Data and code availability

The supplementary materials listed in Table 4 are provided in this paper’s Online Resources repository, version 1.2.0, archived in Zenodo at <https://doi.org/10.5281/zenodo.23107259>.

Table 4. Online resource sections and their contents

Online Resource	Contents
S1	De-identified blind paired-point scouting observations
S2	RTK-GNSS Survey123 data-collection form and de-identified scouting observations
S3	Blind-record reconstruction and linkage rules, linkage audit, detailed secondary statistical methods, and sensitivity-analysis results
S4	Blank reviewer-evaluation workbook
S5	Reviewer classification rubric
S6	De-identified completed reviewer workbooks and associated Python processing and statistical-analysis scripts
S7	Representative de-identified field examples
S8	De-identified RTK-GNSS record-reconciliation and correction-audit materials

Original scouting photographs (aside from example photos provided in Online Resource S7), precise coordinates, anomaly geometries, confidential farm and field identifiers, personally identifiable scout information, and proprietary STADS attributes and anomaly-generation methods are not publicly available. The reported statistical evaluation can be reproduced from the shared materials, but the original anomaly-generation process and complete spatial reconstruction cannot be reproduced.

Use of generative artificial intelligence

During manuscript preparation and development of the supplementary materials, the authors used OpenAI’s ChatGPT to assist with identifying inconsistencies among reported statistical results; editing and reorganizing text; generating initial drafts of selected Python scripts used for data processing, statistical analysis, record reconciliation, and audit generation; and preparing the de-identified reconciliation and audit materials provided in Online Resource S8. ChatGPT was not used to generate or

modify primary data, conduct field evaluations, determine reviewer classifications, or independently make statistical or scientific interpretations. All AI-assisted code, documentation, and outputs were reviewed by the authors before use. The authors examined the underlying logic, parameters, assumptions, and source-data mappings; tested and modified the scripts as necessary; and verified the resulting record counts, analytical outputs, and statistical summaries against the underlying data and independent checks. The authors take full responsibility for the validity, accuracy, interpretation, and content of the analyses, manuscript, and supplementary materials.

Author contributions - CRediT

Phillip Lanza: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing - original draft, Writing - review & editing, Visualization, Project Administration. **David G.**

Rossiter: Conceptualization, Methodology, Writing - review & editing. **Diana S. Sinton:** Writing - review & editing. **Louis Longchamps:** Conceptualization, Methodology, Resources, Supervision, Project Administration, Funding Acquisition, Writing - review & editing.

References

1. Araus, J. L., & Cairns, J. E. (2014). Field high-throughput phenotyping: The new crop breeding frontier. *Trends in Plant Science*, 19(1), 52–61. <https://doi.org/10.1016/j.tplants.2013.09.008>
2. Bock, C. H., Barbedo, J. G. A., Ponte, D., Bohnenkamp, E. M., D., & Mahlein, A. K. (2020). From visual estimates to fully automated sensor-based measurements of plant disease severity: Status and challenges for improving accuracy. *Phytopathology Research*, 2(1), 9. <https://doi.org/10.1186/s42483-020-00049-8>
3. Bock, C. H., Chiang, K. S., & Del Ponte, E. M. (2022). Plant disease severity estimated visually: A century of research, best practices, and opportunities for improving methods and practices to maximize accuracy. *Tropical Plant Pathology*, 47(1), 25–42. <https://doi.org/10.1007/s40858-021-00439-z>
4. Bullock, D. S., Boerngen, M., Tao, H., Maxwell, B., Luck, J. D., Shiratsuchi, L., Puntel, L., & Martin, N. F. (2019). The Data-Intensive Farm Management Project: Changing Agronomic Research Through On-Farm Precision Experimentation. *Agronomy Journal*, 111(6), 2736–2746. <https://doi.org/10.2134/agronj2019.03.0165>
5. Cho, J. B., Guinness, J., Kharel, T. P., Sunoj, S., Kharel, D., Oware, E. K., van Aardt, J., & Ketterings, Q. M. (2021). Spatial estimation methods for mapping corn silage and grain yield monitor data. *Precision Agriculture*, 22(5), 1501–1520. <https://doi.org/10.1007/s11119-021-09793-z>
6. Environmental Systems Research Institute (2024). *Survey123 for ArcGIS* (Version 3.20.x) [Computer software]. <https://www.esri.com/en-us/arcgis/products/arcgis-survey123/overview>
7. Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., Ravaud, P., & Brorson, S. (2012). Observer bias in randomised clinical trials with binary outcomes: Systematic

- review of trials with both blinded and non-blinded outcome assessors. *Bmj*, 344, e1119.
<https://doi.org/10.1136/bmj.e1119>
8. Leroux, C., Jones, H., Pichon, L., Taylor, J., & Tisseyre, B. (2019). Automatic harmonization of heterogeneous agronomic and environmental spatial data. *Precision Agriculture*, 20(6), 1211–1230.
<https://doi.org/10.1007/s11119-019-09650-0>
 9. Longchamps, L., Panneton, B., Reich, R., Simard, M. J., & Leroux, G. D. (2016). Spatial Pattern of Weeds Based on Multispecies Infestation Maps Created by Imagery. *Weed Science*, 64(3), 474–485.
<https://doi.org/10.1614/WS-D-15-00178.1>
 10. Longchamps, L., Panneton, B., Simard, M. J., & Leroux, G. D. (2014). An Imagery-Based Weed Cover Threshold Established Using Expert Knowledge. *Weed Science*, 62(1), 177–185.
<https://doi.org/10.1614/WS-D-13-00050.1>
 11. Maestrini, B., & Basso, B. (2021). Subfield crop yields and temporal stability in thousands of US Midwest fields. *Precision Agriculture*, 22(6), 1749–1767. <https://doi.org/10.1007/s11119-021-09810-1>
 12. Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2(2), 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>
 13. Pancorbo, J. L., Camino, C., Alonso-Ayuso, M., Raya-Sereno, M. D., Gonzalez-Fernandez, I., Gabriel, J. L., Zarco-Tejada, P. J., & Quemada, M. (2021). Simultaneous assessment of nitrogen and water status in winter wheat using hyperspectral and thermal sensors. *European Journal of Agronomy*, 127, 126287. <https://doi.org/10.1016/j.eja.2021.126287>
 14. Pierce, F. J., & Nowak, P. (1999). Aspects of Precision Agriculture. In D. L. Sparks (Ed.), *Advances in Agronomy* (Vol. 67, pp. 1–85). Academic Press. [https://doi.org/10.1016/S0065-2113\(08\)60513-1](https://doi.org/10.1016/S0065-2113(08)60513-1)
 15. Plant, R. E. (2001). Site-specific management: The application of information technology to crop production. *Computers and Electronics in Agriculture*, 30(1), 9–29. [https://doi.org/10.1016/S0168-1699\(00\)00152-6](https://doi.org/10.1016/S0168-1699(00)00152-6)
 16. QGIS Development Team. (2009). *QGIS Geographic Information System*. Open Source Geospatial Foundation. <http://qgis.org>
 17. Shaw, D. R., & Kelley, F. S. (2005). Evaluating Remote Sensing for Determining and Classifying Soybean Anomalies. *Precision Agriculture*, 6(5), 421–429. <https://doi.org/10.1007/s11119-005-3681-9>
 18. Tagarakis, A. C., & Ketterings, Q. M. (2017). In-Season Estimation of Corn Yield Potential Using Proximal Sensing. *Agronomy Journal*, 109(4), 1323–1330.
<https://doi.org/10.2134/agronj2016.12.0732>
 19. Trevisan, R. G., Bullock, D. S., & Martin, N. F. (2021). Spatial variability of crop responses to agronomic inputs in on-farm precision experimentation. *Precision Agriculture*, 22(2), 342–363.
<https://doi.org/10.1007/s11119-020-09720-8>
 20. Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Figures

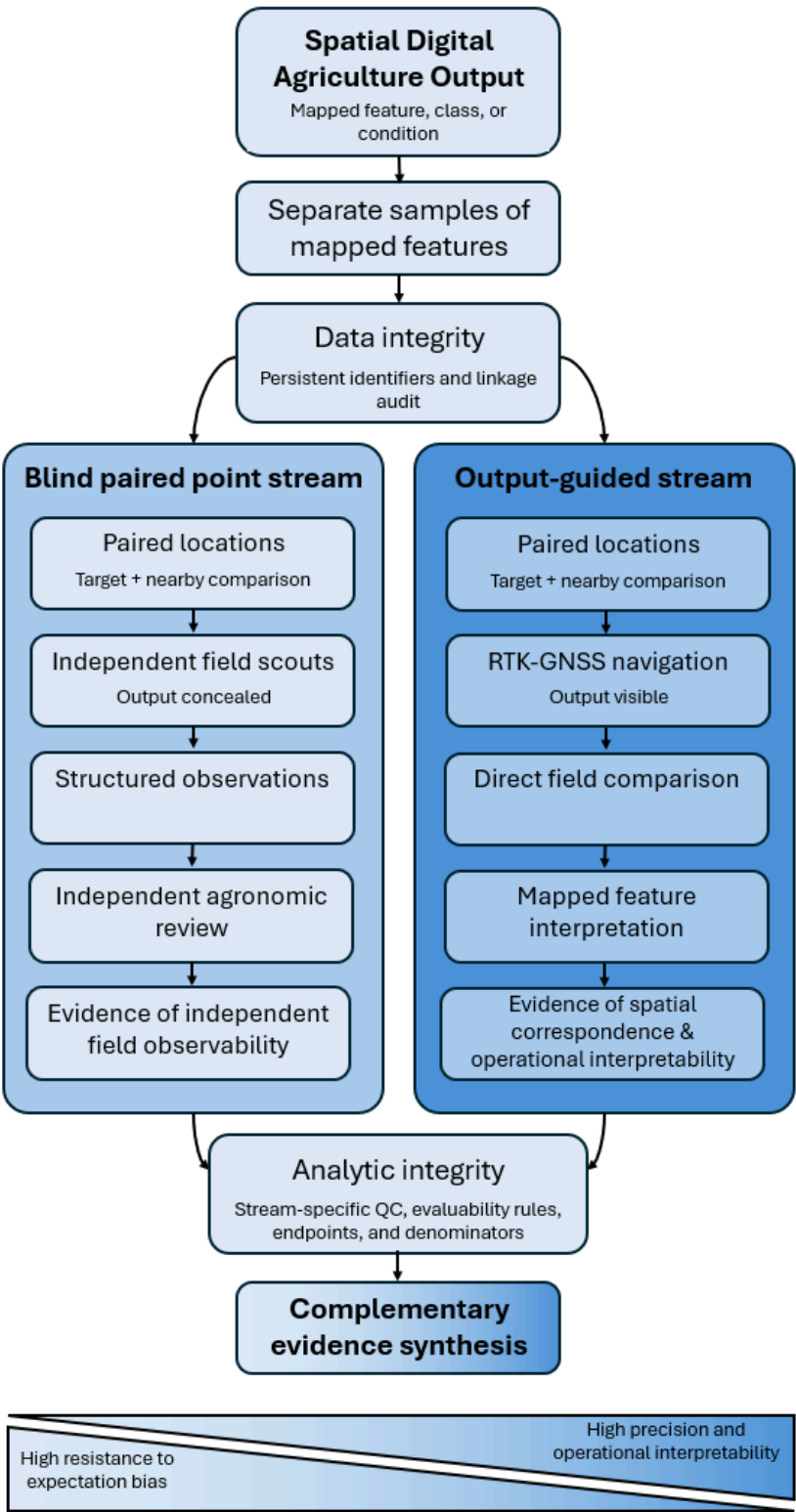


Figure 1

Composite scouting and evaluation protocol for spatial digital agriculture outputs. Prespecified stream-specific samples of mapped features are evaluated through two complementary streams. Blind scouts assess paired target and comparison locations without seeing the mapped output, after which structured observations can undergo masked agronomic review. Output-guided evaluators use navigation equipment capable of high positional accuracy with the mapped feature visible. Data integrity and analytic integrity apply across both streams, and evidence is synthesized without pooling unlike observations into one accuracy estimate.

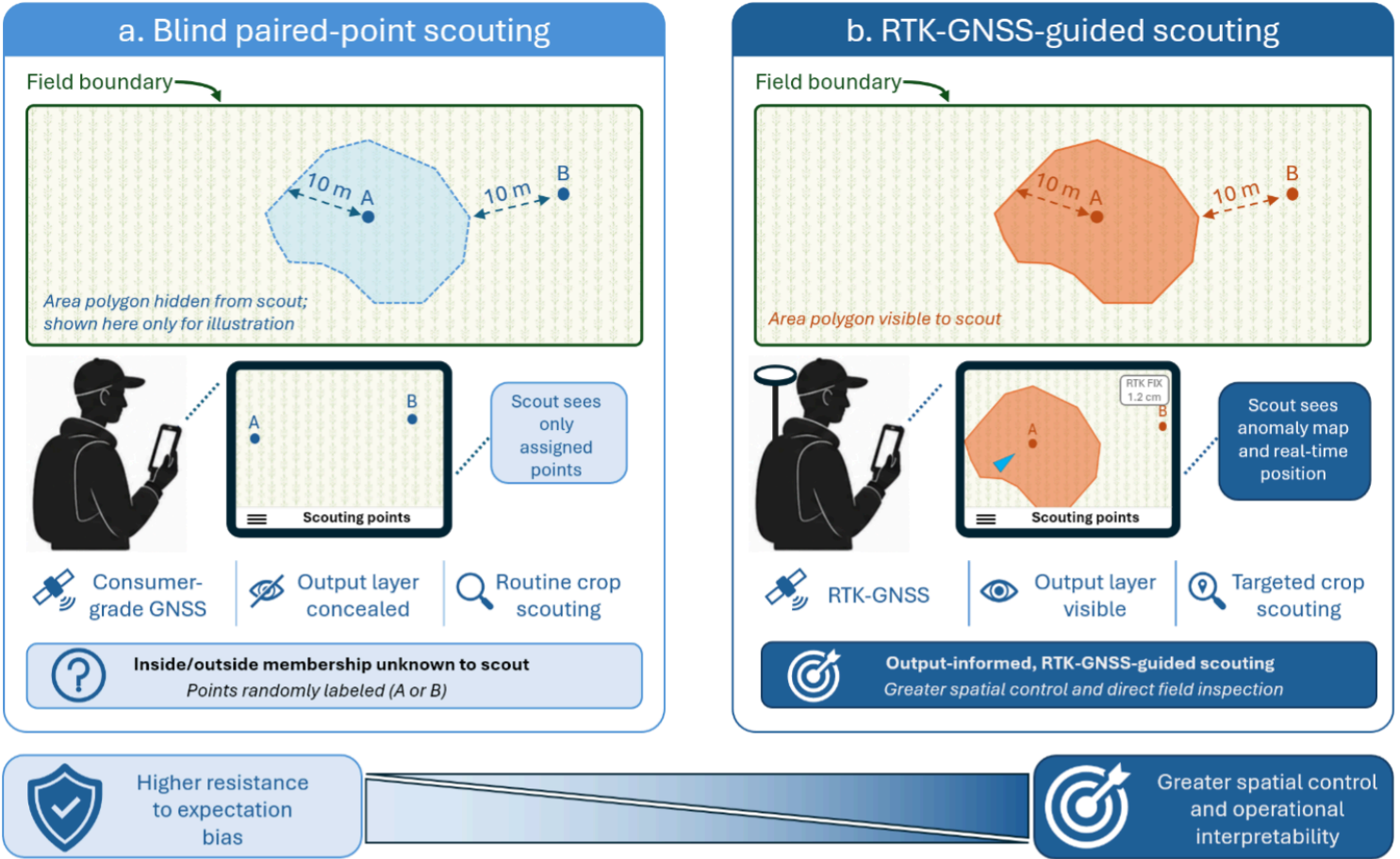


Figure 2

Field implementation of the blind paired-point and output-guided RTK-GNSS protocols. Each pair contained one location inside a mapped output feature and a nearby comparison location outside the feature. Where feasible, points were placed approximately 10 m inside and 10 m outside the feature boundary. Output-blinded consultants used randomized A and B labels without seeing the anomaly. Output-guided scouts viewed the anomaly and their real-time position. The diagram is schematic and not to scale.

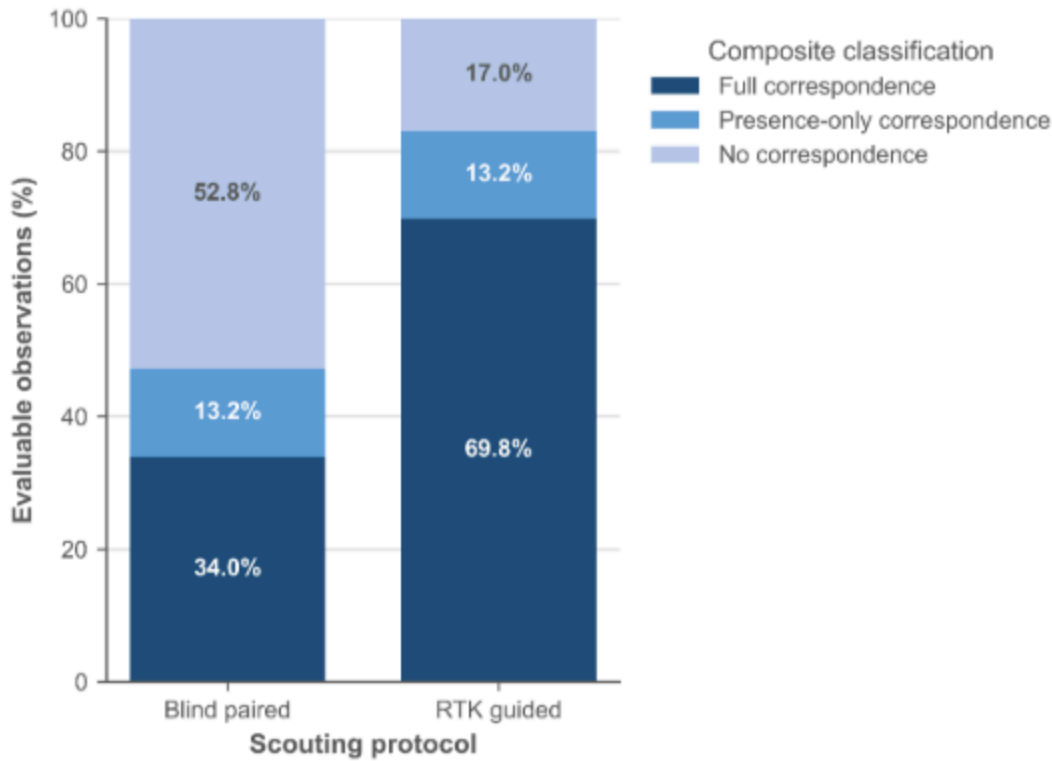


Figure 3

Composite classification of STADS field-evaluation outcomes by evidence stream. Full correspondence required support for anomaly presence and agreement with the STADS-derived direction. Presence-only correspondence indicated confirmed presence with a mismatched or indeterminate direction. No correspondence indicated that field observations did not confirm presence. Each stream contained 53 evaluable records. Because the streams evaluated different samples under different conditions, the percentages are descriptive.