

A Held-Out Site Is a Sample of One: What Random, Spatial and Site-Disjoint Splits Each Estimate, Why Leakage That Differs Between Configurations Reorders a Comparison, and How Many Sites a General Claim Needs

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. A September 2026 study of farm-machinery traversability reports that a forward-facing near-infrared camera appears to beat a colour camera on a paddy-boundary class under its corpus's own sequence-level split, that the advantage disappears once recording sites are held apart, that the near-infrared configuration gained more from having seen a site than colour did, and that two leads the second camera held at one held-out site changed sign when the held-out site was rotated. Each of those findings touches a literature that disagrees with itself: ecologists and remote-sensing researchers dispute whether spatial cross-validation corrects an optimistic bias or introduces a pessimistic one, clinical machine learning had the same dispute over subject-wise splitting, and machine-learning benchmarks report both that rankings survive distribution shift and that they do not. In this analysis the author reviews and synthesises published results to show that these disputes concern three different questions an evaluation can be asked - how accurate a model is at a stated target (level), which of two configurations is better there (order), and whether an advantage holds across the population of sites a deployment will meet (generality) - and that each needs a different design. The level dispute resolves once the prediction target is specified, and the published positions on both sides agree on that resolution more than their titles suggest. The order question has a failure mode the level dispute does not cover: when leakage inflates configurations unequally, a split that shares sites can reverse a comparison, and six published results in four fields show it doing so or coming close: reversals in Kang (2026), Zech et al. (2018), Liang et al. (2017) and, on small margins, Meyer et al. (2019) and Schratz et al. (2018), and a several-fold shrinkage in Milioto et al. (2018). The generality question is a small-sample problem in sites, not in images: one held-out site is one observation. The paper consolidates 24 published results in a table, derives the condition under which leakage preserves an ordering, states a procedure that assigns each evaluation the claim it licenses, and specifies five studies that would settle what remains open.

Keywords: spatial cross-validation; data leakage; site-disjoint evaluation; model comparison; domain generalization; agricultural robotics; evaluation methodology; clustered inference

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

A deployed machine meets places it has never seen. A farm robot drives onto a field absent from its training set; a pneumonia classifier reads films from a hospital that contributed none of its training images; a camera-trap classifier sees backgrounds it never trained on; a forest-biomass map is read at locations far from any inventory plot. The evaluations that justify such systems are drawn from data in which nearby samples resemble one another, and the question of how to split those data into training and test portions has produced a literature that is large, partly settled and, on some points, openly in dispute.

The dispute has a familiar first half. Roberts et al. (2017) reviewed cross-validation for data with temporal, spatial, hierarchical or phylogenetic structure and concluded that random splitting regularly understates predictive error and that blocked splitting is nearly always the more appropriate choice when the goal is prediction to new data or new predictor space. Ploton et al. (2020) reproduced the method of large-scale biomass mapping

on an inventory of 11.8 million trees in central Africa and found that a random 10-fold cross-validation reported an R^2 of 0.53 where a spatial 44-fold cross-validation reported 0.14, and that a random forest given only geographic coordinates performed about as well as the one given 36 remote-sensing and environmental predictors. Meyer et al. (2019), Kattenborn et al. (2022) and Karasiak et al. (2022) reported the same direction in land-cover classification, drone-image segmentation and Sentinel-2 forest mapping. The recommendation that followed is now widespread: block the split in space.

The second half is less familiar outside geostatistics. Wadoux et al. (2021) argued, under the title "Spatial cross-validation is not the right way to evaluate map accuracy", that the case against standard cross-validation rests on a misconception: unbiased map accuracy is obtained by probability sampling and design-based inference, and in their own biomass experiment standard cross-validation had smaller bias than spatial cross-validation. De Bruin et al. (2022) followed with an experiment over five sampling designs in which conventional random cross-validation suited regularly spread samples, weighted cross-validation suited

moderately clustered ones, and blocked spatial cross-validation came closest only for the most strongly clustered design, where much of the map was extrapolation, and was still biased there. Linnenbrink et al. (2024) describe the appropriateness of random and spatial cross-validation for map accuracy as "currently the subject of controversy". Clinical machine learning had the same argument in 2017: Saeb et al. (2017) showed that record-wise cross-validation can grossly overstate diagnostic accuracy when records from one subject appear on both sides of a split, and Little et al. (2017), in a published exchange including the original authors, argued that subject-wise cross-validation can itself be inconsistent and that the real problem is confounding, which no choice of split fixes.

The anchor for this paper is a study that sits at the junction of these arguments. Kang (2026) evaluated colour, near-infrared, a capacity-matched luminance control and their fusion for segmenting traversable ground ahead of agricultural machinery, on a public Korean corpus that pairs the two bands. Under the corpus's own sequence-level split, near-infrared appeared to segment a paddy-boundary class far better than colour; under site-disjoint protocols it did not, and neither did fusion or near-infrared beat colour consistently on any of five traversability classes. The study adds two observations that the level dispute in ecology does not address. The leakage was sensor-dependent - near-infrared gained more from having seen a site than colour or the luminance control did - so holding sites apart did not lower every configuration alike but removed one configuration's lead. And the two effects that did favour the second camera at the single site the main protocol held out both changed sign when a different site was held out.

This paper argues that the published disagreements become tractable once three questions are separated. The level question asks how accurate a fitted model is at a stated prediction target. The order question asks which of two configurations - two sensors, two models, two feature sets - is better at that target. The generality question asks whether an ordering holds across the population of sites a deployment will meet, rather than at the sites a study happened to hold out. The dispute between Ploton et al. and Wadoux et al. is a dispute about the level question, and Section 4 argues that it resolves once the target is written down, in a way that later work from both camps (Mila et al., 2022; de Bruin et al., 2022) endorses. The order question has a separate failure mode, which Section 5 derives: a split that shares sites inflates each configuration by an amount that depends on how well it can exploit what the shared sites give away, and when those amounts differ by more than the true difference, the comparison reverses. The generality question, treated in Section 6, is a small-sample problem whose unit is the site, and a study with three sites has three observations however many frames

it scores.

In this analysis the author reviews and synthesises published results to show that leakage which differs between configurations can reorder a comparison and that one held-out site licenses no general claim. The contributions are as follows. (1) A decomposition of what an evaluation estimates into level, order and generality, with the split each one requires (Section 3). (2) A reading of the spatial cross-validation and subject-wise cross-validation disputes as disagreements about the target of the level question, with the published criterion that settles them (Section 4). (3) A derivation of the condition under which leakage preserves or reverses an ordering, and a collection of published cases in which it reversed one or came close (Section 5). (4) An account of why one held-out site licenses no general claim, and what site-level inference requires (Section 6). (5) A list of the routes by which a split labelled disjoint still shares information (Section 7). (6) Table 1, which consolidates 24 published results; Figure 1, which plots the subset with two endpoints; and Algorithm 1, a procedure that assigns an evaluation the claim it licenses (Sections 8 and 9). Section 10 applies the procedure to the multispectral-sensor literature the anchor addresses, Section 11 states the strongest case against this reading, and Section 12 specifies the studies that would settle the open part.

Where a value in this analysis is the difference of two published values, the text marks it "by subtraction".

2. What the anchor study reports

Kang (2026) works from the AI Hub Precision Agriculture Farm Machinery Autonomous Driving corpus, which supplies forward-facing colour and near-infrared imagery from machinery driving through rice paddies and dry fields, with a separate pixel-level annotation for each modality. The study recovers three location clusters ("super-sites") from the recorded GPS tracks by single-linkage clustering at an 11 m threshold. The corpus's distributed partition is drawn over whole recording sequences, so that no sequence is on both sides, but separate sequences revisit the same plots; by the study's count its training and validation sequences share 270 GPS-footprint overlaps and every super-site contributes frames to both partitions.

The design has four features that matter for what follows. First, the luminance control is the greyscale of the colour image, presented through the same single-map input as near-infrared, so that a near-infrared gain over it is attributable to the band rather than to the change from three chromatic channels to one intensity map. Second, the main leakage-controlled comparisons are scored on the pixel intersection of the two annotations, which removes any advantage a configuration gains from matching quirks of its own annotation. Third, the leakage inflation is measured as a

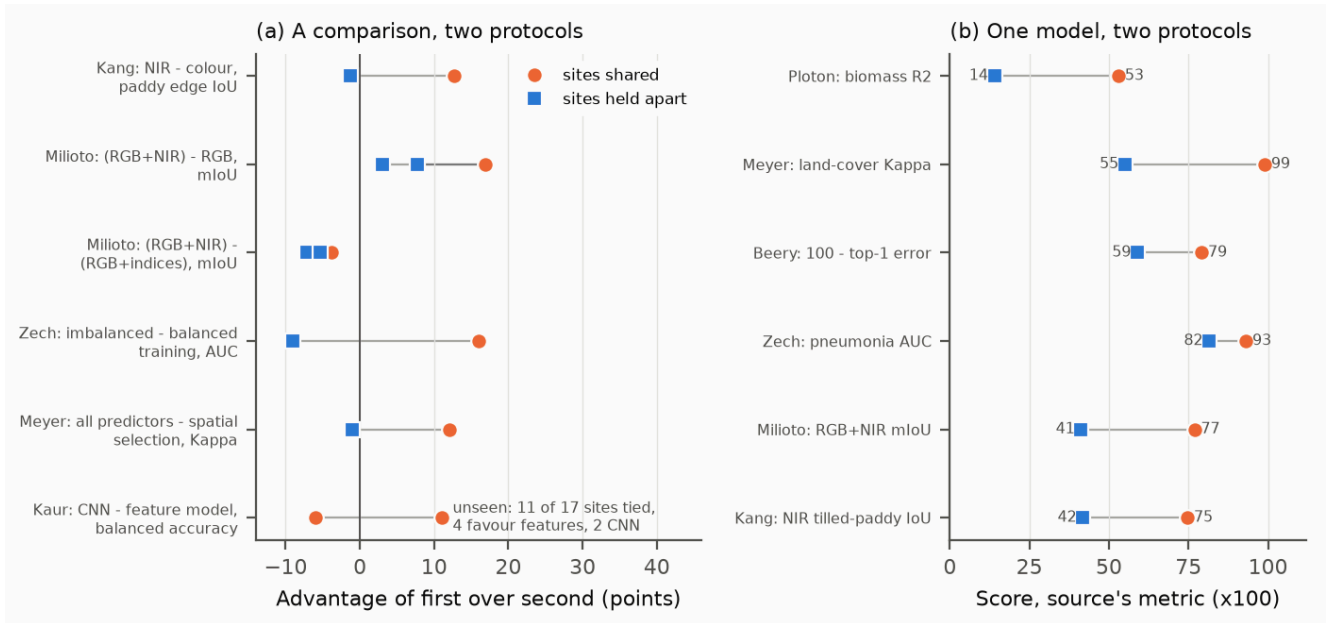


Figure 1. Published values under a protocol that shares sites between training and test (circles) and one that holds them apart (squares). (a) The advantage of one configuration over another, in points (0-1 metrics x100): Kang (2026), NIR minus colour on the paddy edge; Milioto et al. (2018), RGB+NIR minus RGB and minus RGB with derived indices, held-out Bonn images against Zurich and Stuttgart, by subtraction from their Table III; Zech et al. (2018), MSH Severe minus Balanced training cohort, internal against external AUC; Meyer et al. (2019), all predictors minus spatially selected predictors, Kappa (" >0.99 " plotted as 0.99); Kaur et al. (2026), CNN minus feature model on seen sites under each of two rating sets, with the unseen-site outcome as a per-site tally (tied = balanced-accuracy difference under 0.1). (b) One model's score under each protocol: Ploton et al. (2020) biomass R2; Meyer et al. (2019) land-cover Kappa; Beery et al. (2018) 100 minus image-level top-1 error; Zech et al. (2018) joint MSH-NIH model AUC, external at Indiana; Milioto et al. (2018) RGB+NIR mIoU at Bonn and Zurich; Kang (2026) NIR tilled-paddy IoU, whose two protocols score different frames.

paired quantity: the per-frame difference between a model trained under the sequence-level split and one trained under a site-disjoint protocol, over frames both score. Fourth, confidence intervals resample whole driving sequences rather than frames, and each criterion that pools folds of a site rotation is accompanied by a one-sample t-test on the per-site means, at two degrees of freedom.

The results that bear on this paper are these. Under the site-blocked protocol, near-infrared beats colour on none of the five target classes. On the paddy edge, near-infrared leads colour by +0.127 IoU on the paired statistic under the sequence-level split and trails it by -0.013 (95% interval [-0.028, -0.000]) under the site-blocked split. On a fixed set of frames from the held-out super-site, a model that trained on other frames of that super-site scores 0.172 on the near-infrared paddy edge against 0.016 for a model that never saw it. The paired gain from having seen the site is +0.067 for near-infrared, +0.027 for the luminance control and -0.008 for colour; the between-configuration contrasts (near-infrared over colour +0.076, near-infrared over luminance +0.040, luminance over colour +0.035) all have intervals excluding zero, but those intervals resample the same four driving sequences inside one super-site, a limit the study states in so many words. Near-infrared is again the most inflated configuration at the second held-out super-site on which the class is evaluable and under two further model configurations.

The rotation results are the second half. Fusion's site-blocked margin over colour on the tilled paddy surface is +0.032 at the site the main protocol held out, +0.010 at the second and -0.075 at the third, for a site-level mean of -0.011 ($p = 0.77$). Near-infrared's margin over the luminance control, pooled over frames from three folds, is +0.080 (95% interval [+0.028, +0.137], 7,317 frames), which satisfies the study's frame-level criterion; its per-fold values are +0.140, -0.027 and +0.073, and the site-level test gives $p = 0.33$. The study notes that a pooled figure weights each fold by its frame count, so one large site can carry a pooled pass alone, and that at two degrees of freedom the 0.05 critical value is 4.30 standard errors. It also reports that on near-infrared's margin over the luminance control the frame-level interval excludes zero while the sequence-level interval on the same data spans it.

The study is careful about its own scope. It evaluates daylight, clear-weather imagery from one province, initialises every encoder from visible-band weights, and states that a gain smaller than the site-to-site spread across its three sites is not detectable on this corpus. One of the three folds of its unrestricted rotation is not a test of site generalisation at all: it holds out the super-site containing 66.4% of all dry-field frames, the farm-road control drops to 0.248 IoU there, and the study reads the fold as out-of-domain transfer. The paddy-restricted rotation, which supplies the sign changes above, has no such fold: its lowest fold mean on the tilled paddy

surface is 0.392. Its conclusion is accordingly narrow: for daytime paddy and dry-field driving on this corpus the second camera is unproven, and a location-leaking evaluation can reorder the sensors a procurement decision depends on.

3. Three questions an evaluation can be asked

The disputes reviewed below are easier to read once the estimand is fixed. Let a configuration c be a complete pipeline - input bands, architecture, training recipe. Let s index sites, where a site is whatever unit a deployment meets as new: a field, a hospital, a camera location, a forest region, a subject. Let D be the training data, drawn from a set of training sites, and let T be the deployment target.

Level. The level question asks for the expected loss of a model fitted to D , evaluated at T . Two targets dominate the literature. The interpolation target is the set of locations inside the sampled region, at their actual distances from training data; this is what a map is read at, and it is the target in Wadoux et al. (2021) and de Bruin et al. (2022). The new-site target is a site that contributed nothing to D ; this is what a robot, a clinic deploying a purchased model, or a camera trap at a new location meets, and it is the target in Kang (2026), Beery et al. (2018) and Zech et al. (2018). A level estimate is a number with an interval, and its correctness depends on whether the test portion stands in the same relation to the training portion as T stands to D .

Order. The order question asks for the sign, and ideally the size, of the difference between two configurations at a target, $\Delta(A, B; T) = L(B; T) - L(A; T)$ in loss, or equivalently the score of A minus the score of B . Order is what a procurement or design decision consumes: whether to fit the second camera, whether to adopt the newer architecture, whether a feature-selection step helps. An order estimate inherits every bias of the two level estimates it is built from, but it also cancels any bias they share. That is its advantage and, Section 5 argues, its specific vulnerability.

Generality. The generality question asks about the distribution of $\Delta(A, B; s)$ over the population of sites a deployment will meet: its mean, its spread, and the share of sites at which A is better. A claim that "near-infrared adds traversability information that colour lacks", or that "domain-adapted training improves external performance", is a generality claim. It is a claim about a population whose unit is the site, and its sample size is the number of sites, not the number of frames, pixels or patients.

The three questions are nested in their demands. A correct level estimate at a new-site target requires that the test site be new; a correct order estimate additionally requires that whatever the split gives away does not favour one configuration over the other; a generality claim additionally requires enough sites to estimate a

distribution. A study can answer the first and fail the third. The anchor study's site-blocked protocol answers the level and order questions at one site; its rotation is an attempt at the third with three sites; and its explicit refusal to read a pooled-frame pass as general is the step most published sensor comparisons do not take.

The distinction is not new in pieces. Varoquaux, in Little et al. (2017), framed the choice between subject-wise and record-wise splitting as the choice between an expectation marginal over the confound and one conditional on it, which is the level question with two targets. Bates et al. (2022) showed, for the linear model fitted by least squares, that cross-validation estimates the average error of models fitted to other training sets from the same population rather than the error of the model in hand, which separates a property of the pipeline from a property of a fitted model. Gorman and Bedrick (2019) and Sogaard et al. (2021) argued about whether system rankings in language technology survive a change of split, which is the order question. What has not been done, as far as the author can find, is to set the three side by side and ask which published evaluation design answers which.

4. The level question and the spatial cross-validation dispute

4.1 The case that random splits are optimistic

The case is empirical and large. Ploton et al. (2020) report, beyond the 0.53-versus-0.14 contrast, a buffered leave-one-out analysis on 1,000 test observations: with no exclusion buffer the R^2 was about 0.50, at about 50 km from the nearest training pixel it fell to about 0.15, and beyond 100 km it was virtually null, while a random forest given only pixel coordinates tracked the full model's performance across buffer radii. Their reading is that most of the model's apparent skill was spatial interpolation of the response rather than information carried by the predictors. Meyer et al. (2019) found that a land-use classifier and a leaf-area-index regression evaluated by random cross-validation reported Kappa above 0.99 and R^2 of 0.87, against a Kappa of 0.55 and a larger RMSE (1.25 against 0.96) under spatial cross-validation. Kattenborn et al. (2022), segmenting tree species in drone imagery across multiple acquisitions, found that randomly sampled hold-outs overestimated performance by up to 28% relative to spatially blocked hold-outs, more so at small training sizes, and that autocorrelation among observations was significantly higher within an acquisition than between acquisitions. Karasiak et al. (2022), mapping two forest classes from Sentinel-2 imagery, report that spatial leave-one-out gives "performance metrics" that "are consistent with the real quality of the resulting map contrary to traditional non-spatial cross-validation that overestimates accuracy". Schratz et al. (2018), on forest-disease presence, report non-spatial estimates of AUROC 24 to 39% higher than

their spatial counterparts. Valavi et al. (2019) and Mahoney et al. (2023) supply software and simulation evidence for blocked designs, the latter finding that spatial cross-validation with exclusion buffers generally improved on random V-fold estimates. Le Rest et al. (2014) showed that spatial leave-one-out also matters for variable selection, where ignoring autocorrelation favours overly complex models.

The pattern extends beyond geography wherever samples cluster within a unit. In brain MRI, Yagis et al. (2021) found that splitting 3D volumes at the 2D-slice level raised average slice-level test accuracy by 29 to 55% across four datasets, and that on a randomly relabelled dataset a slice-level split still produced about 96% accuracy against about 50% for a subject-level split. Tampu et al. (2022) report inflation of 0.07 to 0.43 in Matthews correlation (5 to 30% in accuracy) from improper splitting of optical coherence tomography volumes. In hyperspectral image classification, Liang et al. (2017) found that an SVM given only pixel coordinates and no spectral information exceeded 93.8% overall accuracy on every benchmark scene with 5% of pixels used for training under random pixel sampling, against roughly 75.5 to 93.2% for the same SVM on spectra; Nalepa et al. (2018) and El-Hajri et al. (2026) report the same family of problem, the latter quantifying a 30.56-point overall-accuracy gap between random pixel splits and a spatially disjoint split on one benchmark scene. Kapoor and Narayanan (2023) count leakage errors in 17 fields affecting 329 papers and give a taxonomy of eight types of leakage.

4.2 The case that spatial splits are pessimistic

The counter-case is theoretical first and empirical second. Wadoux et al. (2021) argue that the accuracy of a map over a region is a population quantity with a design-based estimator: with a probability sample of validation locations, design-based inference gives unbiased estimates of map accuracy without any adjustment for spatial autocorrelation. On the author's reading of that framing, spatial proximity between training and validation points is not a defect but the condition under which a map is used, since a map is read everywhere, including next to its training data. Their abstract states the conclusion without hedging: spatial cross-validation methods "have no theoretical underpinning and should not be used for assessing map accuracy". In their biomass experiment, standard cross-validation was less biased than spatial cross-validation; where the data were heavily clustered and sampling density varied widely, standard cross-validation fell short, but spatial cross-validation fell shorter.

De Bruin et al. (2022) sharpen the point by locating the problem in the sampling design rather than in autocorrelation. They treat the optimism of conventional cross-validation on clustered samples as a selection bias, which the author reads as the over-representation of

densely sampled areas where prediction is easy, and they propose to correct it by weighting with inverse sampling intensity or by a model-based estimator that lets residual variance depend on sampling intensity. Over 100 realisations of five designs, those two methods were less biased than both conventional and blocked spatial cross-validation in every design except the one with the strongest clustering. Their abstract describes the belief that proximity between calibration and validation points invalidates classical validation as a "widespread misconception". Roberts et al. (2017) themselves warned that blocking can induce extrapolation by restricting the predictor combinations available for training, and so overestimate interpolation error.

The clinical and time-series literatures have the same counter-current. Little, in Little et al. (2017), constructed a model in which observations are independent but clustered in value by subject, showed that subject-wise splitting then makes training and test sets differently distributed, and argued that it can fail where record-wise splitting does not; he also noted that leave-one-subject-out estimates are high-variance when subjects are few and heterogeneous. Bergmeir et al. (2018) showed, with theoretical arguments and experiments, that standard K-fold cross-validation can be used for purely autoregressive models whose errors are uncorrelated, which covers the common case of a flexible model nesting a more appropriate one, and found that it performed favourably against out-of-sample evaluation in a simulation and a real-world example. Cerqueira et al. (2020) found cross-validation usable on stationary series and out-of-sample methods more accurate when non-stationarity was present.

4.3 What the dispute is about

Read side by side, the two cases are less opposed than their titles. Every author in Section 4.1 who states a target states a new-location or new-subject target: prediction "to new data or predictor space" (Roberts et al., 2017), new acquisitions (Kattenborn et al., 2022), new subjects (Saeb et al., 2017). Every author in Section 4.2 states an interpolation target: the map over the sampled region (Wadoux et al., 2021; de Bruin et al., 2022), or the next observation of a series whose structure the model captures (Bergmeir et al., 2018). Each side's empirical finding is a finding about its own target.

The criterion that adjudicates is stated formally by Rabinowicz and Rosset (2020). Their Theorem 2.1 gives a condition for cross-validation to be unbiased under correlated data: the partition must preserve the distributional relation between the prediction set and the training set. The condition does not require the observations to be independent or exchangeable. When the relation is not preserved, they propose a bias-corrected estimator, which Yuval and Rosset (2025) extend to arbitrary learners including deep networks and

to clustered and spatial dependence. Saeb et al. (2017) state the same criterion in clinical language - the relation between training and validation sets should mimic that between the training set and the data expected in clinical use - and Varoquaux's contribution to Little et al. (2017) states it a third way, as a choice between marginal and conditional expectations over the confounding variable.

The mapping-validation literature has since built the criterion into its methods. Mila et al. (2022) proposed nearest-neighbour distance matching leave-one-out cross-validation, which chooses the training data excluded around each test point so that the distribution of test-to-training distances during cross-validation matches the distribution of prediction-to-training distances over the target area; in their simulations, random leave-one-out was adequate for interpolation with randomly distributed samples, buffered leave-one-out was adequate for prediction into new areas but overestimated interpolation error, and the matched method was reliable in every scenario. Linnenbrink et al. (2024) gave a k-fold version that scales to large samples and reported that the quality of the distance match predicted the reliability of the estimate. Meyer and Pebesma (2021) proposed the area of applicability, the region in predictor space within which the cross-validation error holds, and Meyer and Pebesma (2022) applied the concern to global maps of ecological variables. Brenning (2021) proposed prediction-error profiles as a function of prediction distance, which report the level at every target distance at once. Rolf (2023) sets out the evaluation challenges specific to geospatial models trained on global or remotely sensed data, and treats the correct way to measure their performance as a matter still under debate.

This paper's reading is therefore that the level dispute is settled in principle and open only in application. A level estimate is correct when the split reproduces the target's relation to the training data; random splitting reproduces an interpolation target when samples are regular, blocked splitting reproduces a new-site target, and distance-matched splitting reproduces either when the target can be written as a distribution of distances. The remaining disagreement is about which target a given study meant, which is a question its authors can answer and frequently do not. The anchor study answers it explicitly: Kang (2026) cites Wadoux et al. (2021), agrees that blocked validation is pessimistic for map accuracy over a sampled region, and states that the regime Wadoux et al. identify as pessimistic is the regime a deployed machine meets.

Two qualifications keep this from being too tidy. First, the design-based route Wadoux et al. recommend requires a probability sample from the target population. For a new-site target, the population is the set of sites a deployment will meet, and a probability sample of it means randomly selected sites - which no corpus in this paper's evidence base supplies. Second, the distance-

matching methods are defined by geographic distance, but the leakage the anchor study documents runs through a channel that geographic distance indexes only roughly: the visual identity of a site. A frame 761 m from the nearest training frame - the minimum separation at the anchor's site-blocked partition - can still be recognisably the same farm. Where the target is "a site never seen", the relevant distance is whether the site was seen, not how far away the nearest seen frame is.

5. The order question: when leakage differs between configurations

5.1 The condition

Let $S_P(c)$ be the score of configuration c under protocol P , measured on a fixed set of target frames. Write $S_{\text{shared}}(c) = S_{\text{disjoint}}(c) + I(c)$, where $I(c)$ is the inflation configuration c receives from a protocol that shares sites between training and test, measured as a paired difference over the same frames, which is how Kang (2026) measures it. Then for two configurations A and B ,

$$\Delta_{\text{shared}}(A, B) = \Delta_{\text{disjoint}}(A, B) + [I(A) - I(B)].$$

The ordering under the shared protocol equals the ordering under the disjoint protocol whenever the inflation difference $I(A) - I(B)$ is smaller in magnitude than Δ_{disjoint} , or has the same sign. It reverses when the inflation difference has the opposite sign and a larger magnitude. This is an identity, not an empirical result, and it has three consequences.

First, a large inflation does not by itself threaten an ordering. If every configuration gains the same amount from seeing a site, the comparison under a leaking split is unbiased even though every level estimate is wrong. This is why the large level inflations in Section 4.1 do not, on their own, imply that published rankings are wrong, and why the "accuracy on the line" findings discussed in Section 11 are consistent with them.

Second, what threatens an ordering is a difference in the capacity to exploit what the shared sites give away. A configuration that can recognise a site - from a spectral signature, a background, a scanner artefact, a coordinate, a prevalence - and that has a site-specific shortcut to the label, gains more than one that cannot. The size of $I(A) - I(B)$ is a property of the configurations and the corpus jointly, not of the protocol alone.

Third, the inflation difference is measurable. A study that trains each configuration under both protocols and scores the shared frames obtains $I(c)$ for each configuration directly, as the anchor study does. Most studies train under one protocol only, and the inflation difference is then not merely unreported but unknowable from the paper.

5.2 Published cases

Agricultural sensing. In Kang (2026) the paired inflation on the paddy edge is +0.067 for near-infrared, +0.027 for the luminance control and -0.008 for colour; near-infrared's +0.127 lead under the sequence-level split becomes a -0.013 deficit under the site-blocked one, each protocol scored on its own validation frames. Because the +0.127 lead and the +0.076 inflation contrast between near-infrared and colour are measured on different frame sets, the identity in Section 5.1 cannot be applied by subtracting one from the other. The same direction appears, without a reversal, in Milioto et al. (2018), who trained crop-weed segmentation networks on 70% of a dataset recorded in Bonn and tested on a held-out 15% of Bonn images and on the whole of two datasets from Zurich and Stuttgart. By subtraction from their Table III, the network given RGB plus near-infrared beats the RGB-only network by 16.94 mIoU points on held-out Bonn images, by 2.98 at Zurich and by 7.73 at Stuttgart. The near-infrared advantage shrinks by between half and four-fifths at new sites. Their RGB-plus-derived-indices network beats RGB plus near-infrared at all three test sets (80.8 against 76.92, 48.36 against 41.23, 61.12 against 55.82), so that ordering is preserved; the shrinking advantage is the near-infrared one.

Ecology and remote sensing. Meyer et al. (2019) compared a model using all predictors with one whose predictors were chosen by forward feature selection under spatial cross-validation. Under random cross-validation the all-predictors model scored Kappa above 0.99 and the spatially selected one 0.87; under spatial cross-validation they scored 0.55 and 0.56. A random-split comparison would have rejected the selection step that the spatial comparison marginally favours. Ploton et al. (2020) supply the limiting case: a model with no ecological predictors at all, only coordinates, matched the full model across buffer radii, so the predictors' contribution, which is the comparison an ecologist wants, is close to zero under every protocol, while a random split assigns the full model an R^2 of 0.53. Liang et al. (2017) supply the hyperspectral version: under random pixel sampling a coordinates-only SVM beat a spectral SVM on every scene, and spectral-spatial methods that smooth over a pixel's neighbourhood gained accuracy with filter size, up to 92.4% on Indian Pines and 98.0% on Pavia University, because, as the authors argue, larger filters enlarge the overlap between training and test pixels; under their controlled sampling, accuracy improved marginally with small filters and then fell. Spatial-context methods were systematically flattered relative to spectral ones, which is a differential inflation by construction. In the arXiv version of Schratz et al. (2018), the gap between non-spatial and spatial AUROC estimates is 0.167 for a GAM and 0.213 for a random forest; the spatial estimates are 0.708 and 0.699, and the preprint's Table 2 reports the non-spatial ones as

0.874 and 0.912, and the two models trade places between protocols on a margin that is small under either. The journal version of the study (Schratz et al., 2019) reports Brier scores, under which the random forest ranks first, so this reversal is specific to the preprint's metric and should be read as an illustration of the identity rather than as a robust result.

Medical imaging. Zech et al. (2018) trained pneumonia classifiers on chest radiographs from two hospital systems with very different pneumonia prevalence and tested internally (on held-out films from the same systems) and externally (at a third system). A ranking of films by hospital system alone, using no image content, achieved AUC 0.861 on the joint internal test set. When the authors engineered training cohorts that differed in prevalence between sites, every imbalanced cohort beat the balanced one internally and lost to it externally; for the cohort labelled MSH Severe the internal AUCs were 0.899 against 0.739 and the external ones 0.641 against 0.732. On the authors' account, CNNs could detect which hospital system a film came from, and a model trained where site and prevalence were entangled gained on internal test data that shared hospitals with training and paid for it at a hospital it had not seen; the ordering reversed. Glocker et al. (2019) show the upstream mechanism in brain MRI: after state-of-the-art preprocessing, a classifier could still identify which of two studies a scan came from with very high accuracy, and harmonisation did not remove the scanner signal. DeGrave et al. (2021) report the same failure for COVID-19 radiograph classifiers, which relied on confounding factors and failed at new hospitals. Kaur et al. (2026) compared a convolutional quality-assessment network with a hand-crafted-feature model on 1,098 brain MRI volumes from 17 sites: on new images from seen sites the network had higher balanced accuracy under one set of quality ratings (0.86 against 0.75) and lower under the other (0.76 against 0.82); under leave-one-site-out evaluation the two were within 0.1 of each other on 11 sites, the feature model was ahead on 4 and the network on 2, and neither exceeded 0.76.

Wildlife and benchmarks. Beery et al. (2018) designed a camera-trap benchmark with cis-location and trans-location test sets and found image-level top-1 error of 20.83% at training locations against 41.08% at new ones, a 97% increase, with the gap persisting under bounding-box cropping and sequence information. Their paper is a level result, but it points at the configuration-dependence directly: generalisation to new locations was poor especially for classification systems, and, the author suggests, a method that leans on background plausibly gains more from a shared location than one that does not. Gulrajani and Lopez-Paz (2020) found that, with all conditions equal, no domain-generalisation algorithm beat empirical risk minimisation by more than one point under any model-selection criterion, and that selection by a training-domain validation set outperformed leave-

one-domain-out cross-validation; the authors attribute the difficulty of fair comparison to inconsistent datasets, architectures and model-selection criteria, which makes their finding an order result about protocols. Kapoor and Narayanan (2023) report the starkest order failure outside spatial data: in civil war prediction, every paper they examined that claimed complex machine-learning models outperform logistic regression failed to reproduce because of leakage, and the complex models were not substantively better than the decades-old baseline.

5.3 What the cases share

In the agricultural, ecological, remote-sensing and radiographic reversals above, one configuration had a channel to the label through the identity of the site that the other lacked or used less: a spectral band from which networks learned more site-specific structure (Kang, 2026), coordinates or neighbourhood smoothing (Liang et al., 2017), predictors that encode location (Meyer et al., 2019), a learned site-prevalence prior (Zech et al., 2018). This is the shortcut mechanism Geirhos et al. (2020) describe, applied to comparisons: a decision rule that works on the benchmark because of how the benchmark was built. The anchor study's contribution here is to separate the channel from three alternatives without training anything - frame brightness alone does not identify the site better than the majority baseline on near-infrared, the two annotations identify the site about equally, and cross-camera misregistration erodes both comparison sets alike - and so to place the excess in the band as presented to a network initialised on visible-band images. It does not establish why the band carries more site information; Section 13 lists that as unknown.

The cases also show the order question's peculiar asymmetry. A level estimate under a leaking split is wrong in a known direction - optimistic for a new-site target. An order estimate under a leaking split can be wrong in either direction, because the inflation difference can have either sign. A study that reports only that its ranking was "robust to the split" has to say which split, since the robust ranking may be the leaking one.

6. The generality question: one held-out site is one observation

6.1 The unit of inference

A site-disjoint split fixes the level and order questions at the held-out site. It does not tell a reader whether the ordering holds at another site, and the evidence base contains several direct demonstrations that it may not. In Kang (2026), fusion over colour changed sign across the three paddy-restricted folds, and near-infrared over the luminance control did too, while a frame-pooled test passed. Kaur et al. (2026), with 17 sites, show the per-site picture a three-site study cannot: a mix of ties and wins for each side. Milioto et al. (2018), with two new sites, show the near-infrared advantage varying by a

factor of 2.6 between them (2.98 against 7.73 mIoU points). Sa et al. (2018) report the extra channels raising crop-weed AUC on WeedMap from [0.607, 0.681, 0.576] for an RGB-only baseline to [0.839, 0.863, 0.782] with nine input channels, evaluated on whole orthomosaic maps held out from training - one held-out map per camera. That is a site-disjoint result and, by the argument here, a result about one site.

The statistical point is standard outside machine learning and rarely applied inside it. When observations are correlated within clusters, the effective sample size for a cluster-level claim is closer to the number of clusters than to the number of observations. Cameron, Gelbach and Miller (2008) show that standard cluster-robust tests over-reject with few clusters (five to thirty), and Cameron and Miller (2015) review the consequences, including that default standard errors can greatly overstate precision. Ibragimov and Muller (2010) give the construction the anchor study uses: estimate the parameter separately within each of q groups and run an ordinary t-test on the q estimates, which remains valid, if conservative, under heterogeneous variances as long as the group estimates are approximately independent. With $q = 3$ the test has two degrees of freedom and a 0.05 critical value of 4.30, so it can detect only effects large relative to their site-to-site spread. That is not a defect of the test. It is the correct statement of what three sites can show.

The same logic governs intervals. Field and Welsh (2007) show that the cluster bootstrap, which resamples whole clusters, gives consistent variance estimates under both the random-effect and the transformation models of clustered data. The anchor study's observation that a frame-level interval excluded zero while a sequence-level interval on the same data spanned it is the practical form of that result: autocorrelation that inflates a random-split score also narrows a frame-level interval, by the same route. Varoquaux (2017) makes the sample-size version of the point for neuroimaging: with 100 samples, cross-validation error bars are about plus or minus 10%, and the standard error across folds strongly underestimates them. Bates et al. (2022) show that the usual cross-validation intervals can have coverage far below nominal because fold accuracies are correlated. Nadeau and Bengio (2003) showed that variance estimators based only on a cross-validation experiment must be biased, and that ignoring training-set variability can lead to the wrong conclusion that one algorithm is significantly better; Dietterich (1998) had shown that a paired t-test over random train-test splits has high Type I error. Bouthillier et al. (2021) found that data sampling, initialisation and hyperparameter choice each move benchmark results markedly.

6.2 Why site variation is not noise to be averaged away

A generality claim treats sites as draws from a population. Two literatures warn that the population may not be one population. D'Amour et al. (2020) show that pipelines are underspecified: many predictors with equivalent held-out performance in the training domain behave very differently in deployment, so the configuration that wins at one held-out site may be one of several equivalent ones that would lose at another. Dehghani et al. (2021) show that the relative performance of algorithms can be altered by the choice of benchmark task alone, and Gorman and Bedrick (2019) replicated nine part-of-speech taggers' results on a standard split but could not reliably reproduce some of their rankings under randomly generated splits. Sogaard et al. (2021) answer Gorman and Bedrick from the other side: random splits are also optimistic, even adversarial splits underestimate the error on new in-domain samples, and benchmarks should include multiple independent test sets. Koh et al. (2021) built WILDS around distribution shifts of exactly this kind - across hospitals, camera traps, locations and times - and report out-of-distribution performance substantially below in-distribution performance under standard training, with the gap persisting under methods designed to close it.

The direction of the argument is not one-sided. Futoma et al. (2020) argue that an emphasis on broad generalisability can overlook settings in which a model is useful where it was built, and that usefulness at the bedside rather than portability should be the goal. In the present vocabulary, they argue that the generality question is not always the one a deployment asks. A farm that operates on its own fields every season may care about the order question at its own sites, which a temporally held-out evaluation at those sites answers better than any site-disjoint one. The requirement this paper defends is narrower: a study that claims generality should have evidence at the level of sites, and a study that does not should claim less.

6.3 How many sites

The evidence base does not support a universal number, and the author does not propose one. It does support a rule of thumb about what can be said with a given count. With one held-out site, a study can report the level and order at that site and nothing about their spread. With two or three, it can report whether the sign was stable, as Milioto et al. (2018) and Kang (2026) do, and a site-level test can detect only effects several times their between-site standard deviation. With a dozen or more, as in the 17-site evaluation of Kaur et al. (2026), it can report a distribution and a share of sites favouring each configuration. A claim that a sensor or model "generalises" across sites is a claim about that distribution, and the number of sites bounds how much of it the study saw.

7. Disjoint in name only

A split labelled disjoint can still share the information that matters. The evidence base names at least five routes.

Grouping at the wrong level. The anchor corpus's partition is disjoint in sequences but not in locations; Kang (2026) counts 270 GPS-footprint overlaps between its partitions. Separating recordings is not the same as separating sites when recordings revisit sites.

Spatial windows. Rachamalla et al. (2026) report, across 24 hyperspectral scenes, that assigning patches to partitions by their centre pixel leaves between 6.7% and 81.6% (median 23.5%) of validation and test patches sharing pixels with a training window, so that a disjoint split of one benchmark scene "would be disjoint in name only" without a guard band. Nalepa et al. (2018) make the same point about the benchmark partitions then in use.

Shared acquisition conditions. Kattenborn et al. (2022) found autocorrelation significantly higher within an image acquisition than between acquisitions, so a split that separates tiles but not acquisitions retains a shared signature; Glocker et al. (2019) found scanner identity recoverable after harmonisation. A split that separates sites but not seasons, sensors or operators can share any of those.

The annotation. Kang (2026) scores its leakage-controlled comparisons on the intersection of two annotations because a configuration trained and scored on its own modality's labels can gain from matching that annotation's quirks; its supplementary measurement shows both annotations identify the site far above the majority baseline. Labels drawn by a site-specific team, protocol or tool are a site signature.

Pretraining and selection. A test site that is new to the fine-tuning data may not be new to the pretraining data, and a test site that is used to choose hyperparameters is no longer held out. Saeb et al., in Little et al. (2017), identify the implicit reuse of test data by the developer as the most important remaining problem with subject-wise validation. Gulrajani and Lopez-Paz (2020) treat model selection as part of the algorithm for the same reason and require that oracle-selected results be disclosed as such. Kaufman et al. (2012) define leakage as information about the target that would not legitimately be available at prediction time and propose a learn-predict separation as the general remedy; Elangovan et al. (2021) quantify train-test overlap in public NLP datasets and its effect on what an evaluation measures. Roberts et al. (2021) reviewed 2,212 studies of machine learning for COVID-19 imaging, retained 61 after quality screening, and judged none of the models they identified to be of potential clinical use, because of methodological flaws, underlying biases or both.

The list matters for the order question because each route can be configuration-dependent. A configuration that uses pretraining weights matched to its input inherits more from pretraining than one that does not - the anchor study notes that colour is the only configuration whose input matches its visible-band initialisation. A configuration that reads a modality with a stronger site signature gains more from shared acquisition conditions. A residual channel of either kind turns a nominally disjoint split back into a shared one for one configuration only.

8. Consolidated evidence

Table 1 collects the published values on which the argument rests, one row per result, with the protocol pair and what moved between protocols stated so that rows of different kinds are not read as one series. Figure 1 plots the subset that has a value under both protocols.

Three things are visible in Figure 1 that the table states less directly. Panel (b) shows that level inflation is the rule across fields and is large - often tens of points - whenever the target is new sites. Panel (a) shows that order is a different matter: in three of the six comparisons the shared-site protocol and the disjoint one disagree in sign; in one (Kaur et al.) a seen-site advantage whose sign depends on the rating set gives way to mostly ties at unseen sites; in one (Milioto et al.'s RGB+NIR over RGB) the protocols agree in sign and disagree in size by a factor of two to six; and in one (RGB+NIR against RGB with derived indices) they agree. The figure is not a sample of the literature. It is a collection of the cases in which both endpoints were published, and such cases are published disproportionately by authors who suspected a problem. The rate at which comparisons reverse under leakage is unknown, and Section 13 says so.

9. A procedure for reading an evaluation

Algorithm 1 states, as a decision procedure, what Sections 3 to 7 argue: which claim a published evaluation licenses, given what it reports about its target, its split, its configurations and its sites.

The procedure is conservative in one direction on purpose. Line 15 refuses an ordering from a leaking split whenever one configuration has a plausible channel to site identity that the other lacks, even if the study's numbers look clean, because the inflation difference is the quantity the study did not measure. Line 23 refuses a general claim from a frame-pooled pass, because a pooled figure weights each site by its frame count; the anchor study's own rotation shows a pooled interval excluding zero over a per-site sign change.

The procedure is not conservative about level. A random split with regular sampling and an interpolation target receives a valid level estimate at line 9, as Wadoux et al. (2021), de Bruin et al. (2022) and Mila et al. (2022)

argue it should. The procedure does not treat spatial blocking as a virtue in itself; it treats correspondence to the target as the virtue, which is the Rabinowicz and Rosset (2020) criterion restated as a rule.

10. The multispectral sensor question through the procedure

The anchor study's motivating question - whether a forward-facing near-infrared camera repays its cost on a farm machine - is a generality question about an order. Read through Algorithm 1, the published evidence on it sorts as follows.

The near-infrared gains Kang (2026) tests come from two benchmarks. By the anchor study's account, the urban RGB-NIR dataset of Choe et al. (2018) draws its annotated key frames from 50 driving videos and splits them by scene category, and the forested multispectral benchmark of Valada et al. (2017) describes no split. Under Algorithm 1, both are "shared: unknown or true", and both orders are provisional at best; Kang (2026) also notes that the forested benchmark itself reports colour alone above near-infrared alone, with a gain for the band in fusion. GOOSE (Mortimer et al., 2024), which pairs forward-facing colour with near-infrared in off-road scenes, has a fixed split that, again by the anchor's account, is balanced across seasons, weather and classes rather than locations.

The agricultural nadir-view literature supplies site-disjoint results. Milioto et al. (2018) train on data from one location and test on two others, and report the near-infrared advantage over RGB shrinking at the new locations, with an RGB-derived-index input beating RGB plus near-infrared at every test set: an order result at two new sites, with a stable sign for the indices and a site-dependent size for the band. Sa et al. (2018) hold out a whole orthomosaic map per camera and report large AUC gains from multispectral channels: an order result at one site per camera. Sa et al. (2017) generate ground truth from NDVI on plots that contain only crop or only weed and report an F1 of about 0.8; under Algorithm 1 the comparison between input-channel sets there is trained on labels that are themselves a near-infrared product, while the test images were labelled by hand, a residual route of the annotation kind. Chebrolu et al. (2017) supply a widely used sugar-beet dataset recorded by a robot carrying a four-channel multispectral camera over three months on one farm, which by construction cannot answer a new-farm question on its own.

The procedure's verdict is that the generality of a near-infrared advantage for agricultural segmentation is not established in either direction by the published record. The site-disjoint evidence is positive over RGB alone in nadir crop-weed segmentation at one or two sites (Milioto et al., 2018; Sa et al., 2018), though not over RGB-derived indices in the one study that tested them (Milioto et al., 2018), and null in forward-facing

Table 1. Published results bearing on level, order and generality, by source. "Shared" is a protocol that lets training and test share sites, subjects or acquisitions; "disjoint" is one that does not. Values are in each source's own metric; differences marked "by subtraction" are computed from values the source reports.

#	Source	Comparison or quantity	Shared	Disjoint	Moved
1	Kang (2026)	NIR - colour, paddy edge IoU	+0.127	-0.013	Order
2	Kang (2026)	Paired inflation, paddy edge: NIR / lum. / colour	+0.067 / +0.027 / -0.008	-	Differs by config.
3	Kang (2026)	NIR - lum., tilled paddy, per held-out site	-	+0.140, -0.027, +0.073 (frame-pooled +0.080, 7,317 frames); site-level $p = 0.33$	Sign by site
4	Kang (2026)	Fusion - colour, tilled paddy, per held-out site	-	+0.032, +0.010, -0.075; $p = 0.77$	Sign by site
5	Milioto et al. (2018)	(RGB + NIR) - RGB, mIoU (by subtraction)	+16.94	+2.98; +7.73	Size
6	Milioto et al. (2018)	(RGB + NIR) - (RGB + indices), mIoU (by subtraction)	-3.88	-7.13; -5.30	None
7	Sa et al. (2018)	RGB vs 9-channel AUC, one held-out map	-	[0.61, 0.68, 0.58] vs [0.84, 0.86, 0.78]	One site
8	Zech et al. (2018)	Imbalanced vs balanced training, AUC	0.899 vs 0.739	0.641 vs 0.732	Order
9	Zech et al. (2018)	Joint model AUC; hospital-only ranking	0.931; 0.861	0.815	Level
10	Meyer et al. (2019)	All predictors vs spatial selection, Kappa	>0.99 vs 0.87	0.55 vs 0.56	Order
11	Ploton et al. (2020)	Biomass R2, random vs spatial CV	0.53	0.14	Level
12	Liang et al. (2017)	Coordinates-only vs spectral SVM, OA	>93.8 vs 75.5-93.2	-	Level (shared only); the order change is in filter size (Section 5.2)
13	Schratz et al. (2018)	GAM vs RF AUROC	0.874 vs 0.912	0.708 vs 0.699	Order (small)
14	Kaur et al. (2026)	CNN vs feature model, balanced accuracy	0.86 vs 0.75; 0.76 vs 0.82 (two rating sets)	17 sites: 11 tied; features ahead on 4, CNN on 2	Mostly ties (seen-site sign depends on rating set)
15	Beery et al. (2018)	Image-level top-1 error	20.83%	41.08%	Level
16	Kattenborn et al. (2022)	Random vs blocked hold-out	up to 28% higher	-	Level
17	Karasiak et al. (2022)	Non-spatial vs spatial leave-one-out CV, accuracy	overestimates accuracy	consistent with real map quality	Level
18	Yagis et al. (2021)	Slice vs subject split; random labels	+29 to +55%; 96%	50% on random labels	Level
19	Tampu et al. (2022)	Improper vs proper split, MCC	+0.07 to +0.43	-	Level
20	El-Hajri et al. (2026)	Random pixel vs disjoint split, OA	+30.56 pts	-	Level
21	Wadoux et al. (2021)	Standard vs spatial CV, map accuracy	smaller bias	larger bias	Level, opposite sign
22	de Bruin et al. (2022)	Five designs, map accuracy	random CV suits regular samples	blocked closest only when strongly clustered	Level, by design
23	Rachamalla et al. (2026)	"Disjoint" test patches overlapping training	-	6.7-81.6%, median 23.5%	Protocol
24	Gulrajani and Lopez-Paz (2020)	Training-domain vs leave-one-domain-out selection	-	former better; no method > ERM by 1 pt	Selection

traversability at the three held-out sites of a paddy-restricted rotation, with one of the three folds of the unrestricted rotation effectively out of domain (Kang, 2026). Those differ in viewpoint, task, crop and sensor, so they are not one comparison. What they share is that every site count is in single digits, and Section 6 argues that single-digit site counts can show a sign is unstable but cannot show it is stable. Sevastopoulos and Konstantopoulos (2022) survey traversability estimation and describe the field's dependence on large labelled

datasets that are expensive to collect; a multi-site benchmark for the sensor question is a case of that general cost.

11. The strongest case against this reading

The strongest objection is that the order problem this paper highlights is rare, and that the evidence base is a collection of exceptions. Two published results support the objection directly. Miller et al. (2021) found strong correlations between in-distribution and out-of-

Algorithm 1. Assigning an evaluation the claim it licenses

```

input: study E comparing configurations A and B, reported target T, split P, held-out sites H
1 if T is not stated: T <- inferred from the use the paper names; flag "target inferred"
2 unit <- the unit a deployment meets as new (field, hospital, camera, subject, region)
3 if P lets training and test share any instance of the unit from line 2 (e.g. two sequences recorded at the same
site):
4   shared <- true; record the shared channel(s)           # Section 7 routes
5 else if P separates units but a residual route exists (windows, annotation, selection):
6   shared <- partial; record the route
7 else shared <- false
8 # level
9 if T is interpolation and samples are regular: LEVEL ok under random or distance-matched P
10 else if T is new units and shared != false: LEVEL optimistic; claim "upper bound only"
11 else if T is new units and shared == false: LEVEL ok at H only
12 # order
13 if shared != false:
14   if I(A), I(B) and D_shared all scored on the same frames: D_disj <- D_shared - (I(A) - I(B)); ORDER at H =
sign(D_disj)
15   else if A has a site-identity channel B lacks: ORDER unsafe; claim "order unknown"
16   else ORDER provisional; claim "order assumes equal inflation"
17 else ORDER ok at H
18 # generality
19 n <- number of distinct held-out units in H
20 if n == 1: GENERAL none; claim "at one site"
21 else if per-unit D reported:
22   report sign count over n; site-level test on n per-unit means (df = n - 1)
23   GENERAL ok only if the site-level test rejects and no fold is out-of-domain
24 else GENERAL none; claim "pooled over n sites; per-site sign unknown"
25 intervals: require resampling at the unit level; frame-level intervals -> flag "too narrow"
26 return (LEVEL, ORDER, GENERAL, flags)

```

distribution performance across many models and shifts, including satellite and wildlife benchmarks from WILDS, holding across architectures, hyperparameters, training-set sizes and durations. Recht et al. (2019) built new test sets for CIFAR-10 and ImageNet, found accuracy drops of 3 to 15 and 11 to 14 points, and found that gains on the original test sets translated into larger gains on the new ones, so rankings largely survived. If rankings survive shift in general, the objection runs, differential inflation is a curiosity.

The reply has three parts. First, the identity in Section 5.1 predicts exactly this pattern: when configurations differ in capacity but not in their access to site identity, inflation is roughly shared and order is preserved. The models compared by Miller et al. (2021) and Recht et al. (2019) were, as far as their descriptions go, trained on the same inputs within each comparison, so none had a channel to the test distribution's identity that the others lacked. The reversals in Section 5.2 are comparisons between configurations that differ in what they can see - a band, coordinates, a neighbourhood, a site-prevalence prior. Sensor comparisons are comparisons of that second kind by definition. Second, Teney et al. (2023) found inverse correlations between in- and out-of-distribution performance on real datasets and explained how earlier studies could miss them through biased selection of models; so "rankings survive shift" is a regularity with known exceptions, not a law. Third, Miller et al. themselves report cases in which the correlation is weaker, and the paper's own scope is shifts between test sets, not leakage between training and test.

A second objection is that the anchor study's own result is fragile: its key between-configuration contrasts

resample four driving sequences inside one super-site, the evaluable rotation has two sites for the paddy edge, and its main protocol is a single held-out site. The objection is correct, and the anchor study states it. It is also the paper's point: a result that removes a sensor's lead under site holdout is itself a result at one or two sites. The anchor study is evidence that the order can reverse on this corpus, not evidence that near-infrared fails in general; Section 10 reads it that way.

A third objection comes from the level dispute. If Wadoux et al. (2021) are right that spatial cross-validation is biased for map accuracy, a reader might infer that site-disjoint evaluation is biased for sensor comparisons too. The inference does not follow. Wadoux et al.'s argument is about an interpolation target; for a new-site target, the bias they identify becomes the correct answer, as the anchor study notes. And for the order question the relevant bias is not in either level estimate but in their difference, which Wadoux et al. do not address.

12. Studies that would settle the open part

Study 1: paired inflation as a reported quantity. For any comparison of configurations that differ in input, train each under both a shared-site and a site-disjoint protocol and report $I(c)$ for each, paired over common frames, with unit-level intervals. Prediction under this paper's reading: the inflation difference is near zero for configurations that share inputs and differ in architecture, and non-zero for configurations that differ in input modality. Either outcome is informative; a near-zero inflation difference for modality comparisons would weaken Section 5's central claim.

Study 2: a many-site sensor benchmark. Record paired colour and near-infrared imagery for forward-facing traversability at a few dozen farms, with farms selected at random from a defined region, and report the per-farm difference for each configuration. The design-based estimator Wadoux et al. (2021) recommend then applies at the level of farms. This is the study that answers the anchor's procurement question in general; no published corpus supports it.

Study 3: an identity probe as a predictor of inflation. For each configuration, measure how well a frozen representation identifies the site (a linear probe on held-out sequences), and test whether probe accuracy predicts $I(c)$ across configurations and corpora. Kang (2026) measures site identifiability from raw frame statistics and from annotations; the representation-level version is the step that would connect the inflation to Geirhos et al.'s shortcut account and would let a study predict the inflation difference without training twice.

Study 4: pretraining matched to the modality. Repeat a sensor comparison with encoders pretrained on near-infrared imagery as well as visible-band imagery. Kang (2026) names this as the experiment that would separate a band effect from a band's reliance on site cues under a mismatched initialisation.

Study 5: a census of reversals. Collect published studies that report both a shared-site and a disjoint evaluation of at least two configurations, and count how often the sign changes, stratified by whether the configurations differ in input. This would replace Figure 1's collection of cases with a rate. It is a literature study, and it requires only that authors have reported both protocols - which, on the evidence of this paper's search, most have not.

13. What is not known

The rate at which a comparison reverses between a shared-site and a site-disjoint protocol is not known; the published cases are selected. Why near-infrared carries more site information than luminance on the anchor corpus is not known; the anchor study excludes three mechanisms and does not identify the one that operates. Whether a near-infrared advantage holds across forward-facing agricultural sites is not known in either direction; no published corpus has enough sites to say. Whether distance-matched cross-validation, designed for geographic distance, transfers to targets defined by unit identity rather than distance has not been tested. How many sites a sensor comparison needs depends on a between-site variance that no study in the evidence base estimates for that task.

14. Limitations

This analysis rests on 24 consolidated results, chosen because they report values under two protocols or bear directly on the protocol question; it is not a systematic

review, and Section 8 states the selection effect this creates. Several sources are 2026 preprints (Kang; Kaur et al.; El-Hajri et al.; Rachamalla et al.) that have not been peer reviewed. The Schratz et al. reversal is specific to the preprint version's metric and is used as an illustration only. Differences computed by subtraction inherit whatever variance the source's numbers carry, which most sources do not report at the unit level. The decomposition into level, order and generality is a stipulation, and other decompositions - for example, by the decision a result feeds - are possible. Algorithm 1 has not been applied to a sample of papers by independent readers, and its rules on lines 15 and 23 will refuse some claims that are in fact correct.

15. Conclusion

The arguments over spatial cross-validation and subject-wise splitting are, for the most part, arguments about which target an accuracy figure describes, and the published positions on both sides converge on the same criterion: split so that the test portion stands to the training portion as the deployment target stands to the training data. That settles the level question in principle. It does not settle the order question, because a split that shares sites inflates configurations by amounts that depend on how much of the site each can see, and when those amounts differ by more than the true difference, the comparison reverses; the anchor study and five other published comparisons in agriculture, ecology, remote sensing and medical imaging (Zech et al., 2018; Meyer et al., 2019; Liang et al., 2017; Schratz et al., 2018, as reversals; Milioto et al., 2018, as a several-fold shrinkage) show it reversing or shrinking several-fold. Nor does it settle the generality question, because a held-out site, however many frames it contains, is one observation of a population of sites. A sensor comparison that holds sites apart, measures the inflation it removed, and reports its result site by site has answered the question it can answer. One that pools frames from one or three sites into a single significant number has answered a smaller question than its conclusion states.

References

- Bates, S., Hastie, T., & Tibshirani, R. (2022). Cross-validation: what does it estimate and how well does it do it? arXiv:2104.00673. doi:10.48550/arXiv.2104.00673.
- Beery, S., Van Horn, G., & Perona, P. (2018). Recognition in Terra Incognita. ECCV 2018. arXiv:1807.04975. doi:10.48550/arXiv.1807.04975.
- Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis, 120, 70-83. doi:10.1016/j.csda.2017.11.003.
- Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Sepah, N., Raff, E., Madan, K., Voleti, V., Kahou, S. E., Michalski, V., et al. (2021). Accounting for Variance in Machine Learning Benchmarks. arXiv:2103.03098. doi:10.48550/arXiv.2103.03098.
- Brenning, A. (2021). Spatial machine-learning model diagnostics: a model-agnostic distance-based approach. arXiv:2111.08478.

- doi:10.48550/arXiv.2111.08478.
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics*, 90(3), 414-427. doi:10.1162/rest.90.3.414.
- Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2), 317-372. doi:10.3368/jhr.50.2.317.
- Cerqueira, V., Torgo, L., & Mozetic, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109, 1997-2028. arXiv:1905.11744. doi:10.48550/arXiv.1905.11744.
- Chebolu, N., Lottes, P., Schaefer, A., Winterhalter, W., Burgard, W., & Stachniss, C. (2017). Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *The International Journal of Robotics Research*, 36(10), 1045-1052. doi:10.1177/0278364917720510.
- Choe, G., Kim, S.-H., Im, S., Lee, J.-Y., Narasimhan, S. G., & Kweon, I. S. (2018). RANUS: RGB and NIR Urban Scene Dataset for Deep Scene Parsing. *IEEE Robotics and Automation Letters*, 3(3), 1808-1815. doi:10.1109/LRA.2018.2801390.
- D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlby, N., et al. (2020). Underspecification Presents Challenges for Credibility in Modern Machine Learning. arXiv:2011.03395. doi:10.48550/arXiv.2011.03395.
- de Bruin, S., Brus, D. J., Heuvelink, G. B. M., van Ebbenhorst Tengbergen, T., & Wadoux, A. M. J.-C. (2022). Dealing with clustered samples for assessing map accuracy by cross-validation. *Ecological Informatics*, 69, 101665. doi:10.1016/j.ecoinf.2022.101665.
- DeGrave, A. J., Janizek, J. D., & Lee, S.-I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3, 610-619. doi:10.1038/s42256-021-00338-7.
- Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlby, N., Diaz, F., Metzler, D., & Vinyals, O. (2021). The Benchmark Lottery. arXiv:2107.07002. doi:10.48550/arXiv.2107.07002.
- Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7), 1895-1923. doi:10.1162/089976698300017197.
- El-Hajri, I., Karrakchou, O., & Mousist, A. (2026). Adaptive Band Selection for Hyperspectral Classification with Spatially Disjoint Evaluation. arXiv:2606.06684. doi:10.48550/arXiv.2606.06684.
- Elangovan, A., He, J., & Verspoor, K. (2021). Memorization vs. Generalization: Quantifying Data Leakage in NLP Performance Evaluation. *EACL 2021*. arXiv:2102.01818. doi:10.48550/arXiv.2102.01818.
- Field, C. A., & Welsh, A. H. (2007). Bootstrapping Clustered Data. *Journal of the Royal Statistical Society Series B*, 69(3), 369-390. doi:10.1111/j.1467-9868.2007.00593.x.
- Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9), e489-e492. doi:10.1016/S2589-7500(20)30186-2.
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2, 665-673. arXiv:2004.07780. doi:10.48550/arXiv.2004.07780.
- Glocker, B., Robinson, R., Castro, D. C., Dou, Q., & Konukoglu, E. (2019). Machine Learning with Multi-Site Imaging Data: An Empirical Study on the Impact of Scanner Effects. *Medical Imaging meets NeurIPS Workshop 2019*. arXiv:1910.04597. doi:10.48550/arXiv.1910.04597.
- Gorman, K., & Bedrick, S. (2019). We need to talk about standard splits. *Proceedings of ACL 2019*. doi:10.18653/v1/P19-1267.
- Gulrajani, I., & Lopez-Paz, D. (2020). In Search of Lost Domain Generalization. arXiv:2007.01434. doi:10.48550/arXiv.2007.01434.
- Ibragimov, R., & Muller, U. K. (2010). t-Statistic Based Correlation and Heterogeneity Robust Inference. *Journal of Business & Economic Statistics*, 28(4), 453-468. doi:10.1198/jbes.2009.08046.
- Kang, S. (2026). Forward-Facing Near-Infrared Adds Little to Colour for Farm-Machinery Traversability: A Site-Disjoint Evaluation of Sensor-Dependent Spatial Leakage. arXiv:2609.13265. doi:10.48550/arXiv.2609.13265.
- Kapoor, S., & Narayanan, A. (2023). Leakage and the Reproducibility Crisis in ML-based Science. *Patterns*, 4(9). arXiv:2207.07048. doi:10.48550/arXiv.2207.07048.
- Karasiak, N., Dejoux, J.-F., Monteil, C., & Sheeren, D. (2022). Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*, 111, 2715-2740. doi:10.1007/s10994-021-05972-1.
- Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., & Dormann, C. F. (2022). Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 5, 100018. doi:10.1016/j.jphoto.2022.100018.
- Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), 1-21. doi:10.1145/2382577.2382579.
- Kaur, P., Thornton, J. S., Barkhof, F., Yousry, T. A., Vos, S. B., & Zhang, H. (2026). Quality assessment of brain structural MR images: Comparing generalization of deep learning versus hand-crafted feature-based machine learning methods to new sites. arXiv:2603.18050. doi:10.48550/arXiv.2603.18050.
- Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., et al. (2021). WILDS: A Benchmark of in-the-Wild Distribution Shifts. arXiv:2012.07421. doi:10.48550/arXiv.2012.07421.
- Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., & Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, 23(7), 811-820. doi:10.1111/geb.12161.
- Liang, J., Zhou, J., Qian, Y., Wen, L., Bai, X., & Gao, Y. (2017). On the Sampling Strategy for Evaluation of Spectral-spatial Methods in Hyperspectral Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2), 862-880. arXiv:1605.05829. doi:10.48550/arXiv.1605.05829.
- Linnenbrink, J., Mila, C., Ludwig, M., & Meyer, H. (2024). kNNDM CV: k-fold nearest-neighbour distance matching cross-validation for map accuracy estimation. *Geoscientific Model Development*, 17, 5897-5912. doi:10.5194/gmd-17-5897-2024.
- Little, M. A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). Using and understanding cross-validation strategies. *Perspectives on Saeb et al. GigaScience*, 6(5). doi:10.1093/gigascience/gix020.
- Mahoney, M. J., Johnson, L. K., Silge, J., Frick, H., Kuhn, M., & Beier, C. M. (2023). Assessing the performance of spatial cross-validation approaches for models of spatially structured data. arXiv:2303.07334. doi:10.48550/arXiv.2303.07334.
- Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9), 1620-1633. doi:10.1111/2041-210X.13650.
- Meyer, H., & Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13. doi:10.1038/s41467-022-29838-9.
- Meyer, H., Reudenbach, C., Wollauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. *Ecological Modelling*, 411, 108815.

- arXiv:1908.07805. doi:10.48550/arXiv.1908.07805.
- Mila, C., Mateu, J., Pebesma, E., & Meyer, H. (2022). Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation. *Methods in Ecology and Evolution*, 13(6), 1304-1316. doi:10.1111/2041-210X.13851.
- Milioto, A., Lottes, P., & Stachniss, C. (2018). Real-time Semantic Segmentation of Crop and Weed for Precision Agriculture Robots Leveraging Background Knowledge in CNNs. ICRA 2018. arXiv:1709.06764. doi:10.48550/arXiv.1709.06764.
- Miller, J., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., & Schmidt, L. (2021). Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization. arXiv:2107.04649. doi:10.48550/arXiv.2107.04649.
- Mortimer, P., Hagmanns, R., Granero, M., Luettel, T., Petereit, J., & Wuensche, H.-J. (2024). The GOOSE Dataset for Perception in Unstructured Environments. ICRA 2024. arXiv:2310.16788. doi:10.48550/arXiv.2310.16788.
- Nadeau, C., & Bengio, Y. (2003). Inference for the Generalization Error. *Machine Learning*, 52, 239-281. doi:10.1023/A:1024068626366.
- Nalepa, J., Myller, M., & Kawulok, M. (2018). Validating Hyperspectral Image Segmentation. arXiv:1811.03707. doi:10.48550/arXiv.1811.03707.
- Ploton, P., Mortier, F., Rejou-Mechain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., & Pelissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11, 4540. doi:10.1038/s41467-020-18321-y.
- Rabinowicz, A., & Rosset, S. (2020). Cross-Validation for Correlated Data. arXiv:1904.02438. doi:10.48550/arXiv.1904.02438.
- Rachamalla, T., Das, A., Kaushik, S., & Roy, S. K. (2026). A PyTorch Library for Hyperspectral Image Models: Technical Report. arXiv:2609.39871. doi:10.48550/arXiv.2609.39871.
- Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet Classifiers Generalize to ImageNet? arXiv:1902.10811. doi:10.48550/arXiv.1902.10811.
- Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schroder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913-929. doi:10.1111/ecog.02881.
- Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3, 199-217. arXiv:2008.06388. doi:10.48550/arXiv.2008.06388.
- Rolf, E. (2023). Evaluation Challenges for Geospatial ML. ICLR 2023 Workshop on Machine Learning for Remote Sensing. arXiv:2303.18087. doi:10.48550/arXiv.2303.18087.
- Sa, I., Chen, Z., Popovic, M., Khanna, R., Liebis, F., Nieto, J., & Siegwart, R. (2017). weedNet: Dense Semantic Weed Classification Using Multispectral Images and MAV for Smart Farming. arXiv:1709.03329. doi:10.48550/arXiv.1709.03329.
- Sa, I., Popovic, M., Khanna, R., Chen, Z., Lottes, P., Liebis, F., Nieto, J., Stachniss, C., Walter, A., & Siegwart, R. (2018). WeedMap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming. *Remote Sensing*. arXiv:1808.00100. doi:10.48550/arXiv.1808.00100.
- Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). The need to approximate the use-case in clinical machine learning. *GigaScience*, 6(5). doi:10.1093/gigascience/gix019.
- Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2018). Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data. arXiv:1803.11266. doi:10.48550/arXiv.1803.11266.
- Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, 406, 109-120. doi:10.1016/j.ecolmodel.2019.06.002.
- Sevastopoulos, C., & Konstantopoulos, S. (2022). A Survey of Traversability Estimation for Mobile Robots. arXiv:2204.10883. doi:10.48550/arXiv.2204.10883.
- Sogaard, A., Ebert, S., Bastings, J., & Filippova, K. (2021). We Need to Talk About Random Splits. EACL 2021. arXiv:2005.00636. doi:10.48550/arXiv.2005.00636.
- Tampu, I. E., Eklund, A., & Haj-Hosseini, N. (2022). Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Scientific Data*, 9. doi:10.1038/s41597-022-01618-6.
- Teney, D., Lin, Y., Oh, S. J., & Abbasnejad, E. (2023). ID and OOD Performance Are Sometimes Inversely Correlated on Real-world Datasets. arXiv:2209.00613. doi:10.48550/arXiv.2209.00613.
- Valada, A., Oliveira, G. L., Brox, T., & Burgard, W. (2017). Deep Multispectral Semantic Scene Understanding of Forested Environments Using Multimodal Fusion. *International Symposium on Experimental Robotics (ISER 2016)*, Springer Proceedings in Advanced Robotics, 465-477. doi:10.1007/978-3-319-50115-4_41.
- Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guilleria-Arroita, G. (2019). blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2), 225-232. doi:10.1111/2041-210X.13107.
- Varoquaux, G. (2017). Cross-validation failure: small sample sizes lead to large error bars. arXiv:1706.07581. doi:10.48550/arXiv.1706.07581.
- Wadoux, A. M. J.-C., Heuvelink, G. B. M., de Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, 457, 109692. doi:10.1016/j.ecolmodel.2021.109692.
- Yagis, E., Atnafu, S. W., Garcia Seco de Herrera, A., Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L., & Diciotti, S. (2021). Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific Reports*, 11. doi:10.1038/s41598-021-01681-w.
- Yuval, O., & Rosset, S. (2025). Cross Validation for Correlated Data in Regression and Classification Models, with Applications to Deep Learning. arXiv:2502.14808. doi:10.48550/arXiv.2502.14808.
- Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Confounding variables can degrade generalization performance of radiological deep learning models. *Published in PLoS Medicine*. arXiv:1807.00431. doi:10.48550/arXiv.1807.00431.