Measure what the coarsening removes: a precondition for interpreting spatial-resolution comparisons in ecological models

Joshua Wang
Independent researcher, Surrey, British Columbia, Canada
Corresponding author: joshuawang2048@gmail.com

September 1, 2026

Preprint, not peer reviewed.

Key Points

- A resolution comparison can be unable to answer the question it asks, and that condition is measurable before any model is fitted.
- Coarsening removed 12% of predictor spatial variance over a municipality and 48% along a valley transect; only the second could be read.
- Below about 0.4 of spatial variance removed a null carries no information; above it, detection turns on the number of validation blocks.

Abstract

1. Ecological models increasingly use "scale-free" downscaled climate surfaces, assuming finer resolution predicts local response better. Even a comparison that varies resolution alone can fail to test it: over a small extent, coarsening barely changes the predictors. I provide a diagnostic detecting this before a model is fitted. Though the variance decomposition is classical, calibrating it against detection is not.

2. Two extents were chosen at opposite ends of that fraction, both in British Columbia: a low-relief municipality ~30 km across (153 corridor polygons × 4 summers) and a ~100 km valley transect spanning 4–1,920 m (300 forest stands × 4 summers). Model A used scale-free ClimateBC values at each polygon's location and elevation; Model B averaged *those same variables* within coarse cells. Both were random forests predicting a satellite water-stress index under spatially-blocked, polygon-grouped cross-validation.

3. At the municipal extent the comparison was unanswerable rather than null: coarsening to 4 km removed only 12.3% of the predictors' spatial variance, leaving them correlated at median $r = 0.998$. Over the transect it removed 48.4%, and the hypothesis was not supported at 25 km, at a power near one in four and at a cell size the label turns on. The coarse model was better on the point estimate (paired $\Delta\mathrm{RMSE} = +0.00093$, 95% CI $[-0.00061, +0.00257]$), but the interval spans zero at both cell sizes. Under blocking the direction favours the coarse model throughout; under random folds it reverses. R^2 was -0.138 for the fine model and $+0.003$ for the coarse. Terrain and stand structure together scored highest (CV $R^2 = +0.029$, positive in 80% of folds

against climate's 40%), and adding climate made it worse.

4. **Synthesis.** A resolution hypothesis can be untestable while looking like a null result. Before interpreting any such comparison, measure the fraction of predictor spatial variance the coarsening destroys, and treat a low fraction as disqualifying rather than a finding. The rule holds in one direction only, since below $f \approx 0.4$ a null carries no information while above it detectability turns on the number of validation blocks as much as on f.

Keywords: change of support; climate downscaling; ecological models; spatial resolution; spatially-blocked cross-validation; statistical power; water stress

1 Introduction

1.1 The resolution assumption

An ecologist asking how vegetation responds to local conditions needs climate values at the scale of the units being modelled, and the usual route to those values is a downscaling product. That chain carries an assumption worth testing.

Climate models produce output on grids. Global and regional models run at 100 km per cell and 10–50 km per cell respectively, so at 25 km an entire municipality can fall inside a single cell along with its neighbours, and every unit in it is given an identical climate value. The standard response is downscaling.

"Scale-free" downscaling products such as ClimateBC and ClimateNA [1] take a particularly aggressive approach. Instead of producing a finer grid, a combined interpolated climate surface with elevation-based lapse-rate adjustment is constructed so that a value can be queried at an arbitrary latitude, longitude and elevation. There is no cell to snap to. ClimateBC downscales 1961–1990 monthly normals unified at 4 km [2], and returns a point value rather than a cell average, so in principle every polygon in a study can receive its own climate. Returning a value at a location is not the same as the underlying field varying between locations that close together, and this paper measures that variation.

The assumption embedded in using such a product is that finer climate input produces better local ecological prediction. Environmental modelling has asked the corresponding question of its own inputs for some time, varying the resolution of a terrain surface and measuring what the model does differently [3], and propagating input resolution through to uncertainty in the output [4]. It is rarely tested by isolating resolution while holding the source data and algorithm strictly constant, and it nevertheless drives real decisions about which data to acquire and where to direct modelling effort.

1.2 Grounds for doubting the assumption

Organisms do not experience macroclimate. Canopy shading, cold-air drainage, slope aspect, soil depth and rooting access all decouple the conditions at a given stem from the regional atmosphere. Forest canopies buffer understorey temperatures by several degrees relative to the open air a macroclimate product represents [5, 6], and the microclimate literature argues that refining such a product need not improve how it represents the conditions organisms experience, because the processes responsible for the decoupling are not in the product at any resolution [7, 8]. Neither established response to that gap is a resolution increase: measuring near-surface conditions directly, as the SoilTemp network does [9], or modelling them mechanistically from terrain and canopy [10]. This study takes neither route, a limitation of it rather than a criticism of them, and it tests terrain and forest composition directly as competing predictor families against climate under the same validation (Supporting Information, Section S12); soil depth and rooting access were not (Section 4.1).

Spatial and temporal variation are different quantities. A predictor can vary strongly through time while barely varying across a study area, or the reverse. A model fitted to variation along one axis carries no guarantee of predicting variation along the other. The "space-for-time substitution" underpinning much of climate change ecology treats the two as interchangeable, and it is a documented failure mode: Perret et al. [11] tested it directly using a network of ponderosa pine tree-ring series and found that a species' response to spatial climatic variation does not predict its response to climate change through time (see also the commentary by Angert [12]).

Aggregation changes relationships. The Modifiable Areal Unit Problem establishes that statistical relationships between variables shift, and sometimes reverse, when the size and shape of the spatial units of aggregation change [13], in landscape ecology as much as in the human geography it was posed in [14]. A question about resolution is therefore inescapably a question about aggregation, which means its answer can depend on the extent over which it is asked, and Section 3 shows that happening.

1.3 This study

The assumption is tested here at two extents, against a satellite water-stress index (CDEI) computed per polygon per summer, which places a polygon against the driest state observed at its own greenness level. It is a relative feature-space index rather than a soil-moisture measurement, and that distinction is carried through everything that follows (Section 2.2).

In falsifiable form:

> H_1**.** Scale-free downscaled climate predictors, point-queried against a 4 km baseline, predict polygon-scale water stress better than coarse grid-based climate predictors.

Within each extent, the design isolates resolution and nothing else. Model A uses ClimateBC values sampled at each polygon's own location and elevation. Model B uses those same values, spatially averaged within coarse cells, with each polygon then taking its cell's value. Deriving B by degrading A rather than substituting a different coarse product is essential: otherwise the comparison would confound which dataset with what resolution, and only resolution is under test.

Whether that comparison can resolve a difference at all turns on something the design does not control, namely how much spatial variance the coarsening actually removes. That quantity depends on the study extent, it can be computed before any model is fitted, and a simulation reported below fixes what it is worth. It is the part of this work meant for reuse elsewhere; the two extents are where it is demonstrated.

2 Materials and Methods

2.1 Study extents

The experiment, code, variables, model and fold structure were identical at both extents, and only the geographic extent changed.

Extent 1: City of Surrey, British Columbia. The 153 polygons of the Green Infrastructure Network, obtained from the City of Surrey's open-data ArcGIS service [15]. They span 0.29–157.8 ha (median 5.8 ha) and are characteristically long and narrow instead of compact, which is what makes partial-pixel coverage a concern at 20–30 m imagery resolution (Limitations, item 7). Elevations run −1 to 116 m across roughly 30 km, low relief but not flat. Study area $\approx$ 408 km^2.

Extent 2: a south-of-Fraser elevation transect. Because Extent 1 proved structurally unable to test the hypothesis (Section 3.1), the identical experiment was repeated over a ~100 km west–east transect spanning 4–1,920 m elevation, on 300 treed stands from the BC Vegetation Resources Inventory (VRI)

[16]. Study area ≈ 2,841 km^2.

Whether the relief at either extent generates the spatial climate gradient the hypothesis needs is measured rather than assumed (Section 2.7). Layers, retrieval parameters, the stratified sample and the corridor-versus-polygon unit structure are in Supporting Information (Section S8).

All spatial operations were performed in EPSG:26910 (UTM zone 10N), and every layer is reprojected to it beforehand. The analysis unit is always the polygon, meaning a corridor polygon at Extent 1 and a stand at Extent 2. That matters because Model B averages predictors within coarse grid cells (Section 2.5): those cells are a device for degrading the predictors, and a polygon takes the value of the cell it falls in. No quantity is ever computed for a cell, and no cell is ever a row in the data.

2.2 Response variable: CDEI

The response is the Canopy Dryness–Edge Index (CDEI), computed per polygon per summer from Sentinel-2 and Landsat 8/9 composites for 2022–2025. It places each polygon against the driest state observed at its own greenness level, using summer NDVI, a shortwave canopy-water index and land-surface temperature. The bands, the compositing window, the fitting of the dry edge and the temperature normalisation are in Supporting Information (Section S9); nothing here depends on evidence reported elsewhere, and a companion paper [17] checks the index independently as an applied product.

Two properties of the index bear on what follows. It is a *relative* feature-space index and not a soil-moisture measurement, and it has never been validated against ground observation at either extent. And because RLST normalises each polygon against its own multi-summer mean, it divides out all between-polygon temperature variation, so between polygons CDEI is numerically almost identical to `dry_dist` alone (ρ = +0.99997).

2.3 Climate predictors

Climate came from the ClimateBC scale-free downscaling web API [1], queried once per polygon centroid with latitude, longitude and elevation, the last of these attached from the Copernicus GLO-30 DEM [18]. Five periods were retrieved per site: the 1961–1990 normal plus one annual period per summer (2022–2025). The endpoint and the acquisition constraints are in Supporting Information.

The full climate-only predictor set has 14 columns: seven seasonal summer variables, five anomalies against each site's own 1961–1990 normal, log(1+PPT_sm) and a Mahalanobis climate-novelty scalar, itemised in Supporting Information (Table S10).

Predictors are climate-only by design. Remote-sensing-derived quantities (NDVI, SWCI, LST) are excluded from the predictor set for the A-vs-B test, because CDEI is constructed from them and including them would be circular.

2.4 Panel assembly

Merging the zonal satellite table with the climate panel gives one row per polygon-year: 153 polygons × 4 summers = 612 rows over Surrey, 300 stands × 4 summers = 1,200 over the transect.

2.5 The resolution experiment (Model A vs Model B)

The two arms of the comparison are:

- **Model A**, ClimateBC values sampled at each polygon's own location and elevation.
- **Model B**, *those same values*, spatially averaged within coarse cells; each polygon then takes its cell's value.

Model B is therefore a *synthetic spatial degradation* of the downscaled product rather than a coarse climate model, since a regional climate model or reanalysis simulates atmospheric fluxes across its grid box whereas Model B only removes spatial detail from ClimateBC. The experiment therefore measures sensitivity to that detail, the content of H_1, not a head-to-head against any particular coarse product. The average is taken over the *sampled polygons* in a cell and so is area-blind (Supporting Information), and where a cell holds one polygon Model B equals Model A exactly, which affects 1 of 25 cells over Surrey and 2 of 10 over the transect.

Because the averaging is within (cell, year) rather than within cell alone (Section 2.7), the temporal signal survives untouched, so on a purely temporal target Models A and B are identical by construction and resolution can only be tested along the spatial axis.

Cell size: 4 km for Surrey (25 cells) and 25 km for the transect (10 cells); both were fixed before any model was fitted. The two extents are not independent: Surrey lies inside the transect's sampling frame, and the westernmost stands share its longitude band. The 25 km cell was chosen to approximate the grid resolution typical of the regional climate models the hypothesis set out to beat. Because the verdict proves sensitive to that choice, the transect experiment is also reported at 12 km (Supporting Information, Table S1).

Learner. Random forest [19], as implemented in scikit-learn [20], with 400 trees, a minimum leaf size of 3, `sqrt` features per split, no depth cap, and fixed seed 26910. With a few hundred polygons and a strongly grouped fold structure, an unconstrained forest memorises polygons instead of learning climate, so depth is governed by the leaf-size constraint rather than capped independently. The values were fixed a priori, with no hyperparameter search or nested tuning. Hyperparameters, seeds, feature list and folds are identical for A and B, and that was taken to mean the choice could not bias the *direction* of the contrast. It does not follow, because the arms differ in the cardinality of their predictors as well as in resolution: A carries one distinct vector per site and B roughly one per occupied cell, so only A has the freedom to overfit. Section 3.3 tests that directly. The choice also affects the *magnitude*, and materially, so the experiment was re-run across a grid of forest settings (Supporting Information, Table S3).

2.6 Validation design

Spatially-blocked, polygon-grouped cross-validation. Folds are k-means blocks ($k = 5$) over polygon coordinates, held out whole, because random splits would place a polygon's immediate neighbours in the training set and score spatial autocorrelation as skill [21]; Ploton et al. [22] show how large that inflation can be, reporting good random-split accuracy and near-zero spatially-validated accuracy for a forest biomass model built to reproduce published mapping practice. A further question this design does not answer is *where* a fitted model's predictions may be trusted, for which Meyer and Pebesma [23] propose an area-of-applicability estimate; none is computed here, and the Mahalanobis novelty scalar of Section 2.3 is a predictor, not a substitute for one. Blocking is applied to the polygon rather than the row, so a polygon cannot appear in both train and test. It is repeated 5 times with re-seeded k-means, so no result depends on a single partition. Folds with fewer than 5 test rows or fewer than 20 training rows are skipped.

Paired comparison. A and B are compared fold-by-fold on identical folds, so the fold-to-fold variation that dominates a small spatial CV cancels rather than drowning the contrast. The confidence interval is a bootstrap (10,000 resamples) instead of a t-interval, because a handful of folds is nowhere near normal. The resampling unit is the block *within* a repeat, with the endpoints averaged across repeats, because only the 5 blocks of one repeat partition the panel and pooling all 25 repeat-block differences would treat a re-partition as fresh sample, narrowing the interval by roughly $\sqrt{5}$. Every interval here is therefore a bootstrap over five units.

Two preconditions before any verdict. If either fails, the verdict is INCONCLUSIVE.

1. **Skill gate.** At least one model must beat predicting the mean, $R^2 > 0$. If neither clears that bar, their comparison carries no information about resolution. The threshold is exactly zero (`MIN_SKILL_R2 = 0.0`), a hard cut on a noisy quantity; where a model lands near it that is reported explicitly (Section 3.2).
2. **Contrast gate.** The upscale must destroy a material fraction of the predictors' spatial variance. If coarsening barely changes the predictors, A and B are nearly the same model, and a null difference between them measures their similarity rather than resolution. The gate is one-sided by construction, because the simulation of Section 3.4 establishes a floor at $f \approx 0.4$ below which a real effect is recovered at most one time in five, and finds no level of f at which the comparison becomes reliable. The implementation keeps a nominal cut at $f = 0.30$ (`MIN_CONTRAST_FRAC = 0.30`) for the verdict labels, though the evidence supports the floor rather than that particular number.

2.7 Measurement of the spatial variance removed

Write x_{ijt} for predictor j at polygon i in summer t, and $c(i)$ for the coarse cell holding polygon i, assigned by flooring the polygon's projected coordinates onto a square grid of side ℓ. Model B replaces every predictor with its cell mean, taken separately within each summer:

$$x^B_{ijt} = \frac{1}{|c(i)|} \sum_{i' \,:\, c(i')=c(i)} x_{i'jt}. \tag{1}$$

Averaging within cell *and* summer, rather than within cell alone, is what leaves the temporal signal untouched, so the two models differ in spatial detail only.

Both are then measured against the same spatial variation, taken about each summer's own mean $\bar{x}_{jt}$ so that between-summer variation never enters:

$$S^2_j = \frac{1}{N} \sum_{i,t} \left(x_{ijt} - \bar{x}_{jt}\right)^2, \qquad S^{2,B}_j = \frac{1}{N} \sum_{i,t} \left(x^B_{ijt} - \bar{x}_{jt}\right)^2, \tag{2}$$

over all N rows. The fraction removed for predictor j, and the figure reported throughout, are

$$f_j = 1 - \frac{S^{2,B}_j}{S^2_j}, \qquad f = \mathrm{median}_j\, f_j. \tag{3}$$

Because x^B is the unweighted mean of exactly the rows it replaces, the cross term in the expansion vanishes and the variance splits without residue:

$$S^2_j = S^{2,B}_j + \frac{1}{N} \sum_{i,t} \left(x_{ijt} - x^B_{ijt}\right)^2. \tag{4}$$

So f_j is the share of spatial variance lying within cells rather than between them, which is precisely what Model B discards and Model A keeps. On the Surrey panel the identity holds to 5×10^{-15}, the floating-point limit.

Computing it takes no model fit: assign each unit to a cell of side ℓ from its projected coordinates, apply Equations (1)–(3), and report the median across predictors. The implementation is `experiment.upscaling_diagnostics()`.

The diagnostic never reads the response variable, which is what allows it to be run before an experiment and what makes applying it after one defensible (Section 4.2).

2.8 Simulation design

So far the precondition rests on two observations, which fix a direction but no threshold. A simulation supplies one and needs no new data.

Synthetic panels of 300 sites × 4 summers over a 100 km square carry four Gaussian random-field climate predictors of correlation length L, an independent short-range local driver standing in for terrain and stand structure, and a response generated from the predictors at their *fine* locations, $y = \beta\bar{x}(s) + \gamma z(s) + \varepsilon$, so a genuine resolution advantage exists by construction. The weights are fixed *a priori* at equal values, putting climate at one third of response variance, on the grounds that a resolution question is only worth asking where climate matters materially. Crucially β is constant across the sweep, so varying L against a fixed 25 km cell moves the variance-removed fraction f *without moving the truth*, and any change in detection is power rather than effect size. Models A and B, the folds, the learner and the paired bootstrap interval are the same functions used throughout this paper, called unmodified. One hundred seeds at each of eight levels of L; a detection is a paired interval excluding zero in the generative direction. Falsification criteria were fixed in code before the run, and the full specification, including two design faults found in calibration, is in Supporting Information.

2.9 Disclosure of generative AI use

Generative AI tools were used in preparing this work. Claude (Anthropic), Opus 4.8 and then Opus 5, scaffolded code in the analysis pipeline and drafted technical prose, which the author then revised; Gemini 3.1 Pro (Google), accessed through NotebookLM, served as an adversarial reviewer of drafted sections. No tool is credited as a contributor, because a tool cannot take responsibility for what it produces. The author takes responsibility for all such code and text, has verified every analysis reported here, and has annotated AI-scaffolded code as such in the repository.

3 Results

The diagnostic is applied at both extents before either comparison is read. At the first it refuses the comparison; at the second it admits one, at 25 km.

3.1 Extent 1 (Surrey): model comparison and precondition diagnostics

On the full 612-row panel, Models A and B are statistically indistinguishable: paired $\Delta\mathrm{RMSE} = -0.00004$, 95% CI $[-0.00028, +0.00018]$, with A better in 48% of folds and neither model above the no-skill line ($R^2 = -0.042$ and -0.051). MAE gives the same answer, its interval spanning zero around a point estimate two orders of magnitude below the metric itself (Supporting Information, Table S7).

This is reported as INCONCLUSIVE and not as evidence against downscaling, because both preconditions of Section 2.6 fail, and Section 4.2 takes up what that distinction rests on.

Precondition 1 fails: no model has skill on any target. Every candidate response variable sits at or below the no-skill line under blocked CV (Supporting Information, Table S8), which rules out "the wrong index was chosen" as an explanation.

Precondition 2 fails: coarsening barely changes the predictors. The 4 km upscale removes only ~12% of the predictors' spatial variance, so A and B are very nearly the same model (Table 1). That cell is also the baseline's own grid, so what B discards here is mostly what the elevation adjustment adds beneath it.

Table 1. Surrey: meaningful resolution contrast requires cells so large that too few remain to spatially block Model B.

Coarse cell	Cells over Surrey	Median spatial variance removed
4 km	25	12.3%
8 km	12	24.1%
12 km	6	28.4%
20 km	4	39.6%
30 km	3	85.6%

Reaching meaningful contrast requires ~20 km cells, which leaves 4 across the whole municipality, too few to spatially block Model B.

Root cause: predictors and response vary along orthogonal axes. The variance ratios are in Table 2.

Table 2. Surrey: ClimateBC varies mostly in *time*; water stress varies mostly in *space*. Spatially-blocked CV asks the model to rank polygons it has never seen, the axis along which climate is nearly constant.

Predictor	Between-year sd	Between-corridor sd	Ratio
CMD_sm	77.53	7.63	**10.2× temporal**
PPT_sm	73.04	11.60	**6.3× temporal**
Tmin_sm	0.80	0.29	2.7× temporal
Rad_sm	0.19	0.07	2.5× temporal
DD18_sm	35.32	14.75	2.4× temporal
Eref_sm	9.16	3.90	2.3× temporal
Tmax_sm	0.67	0.29	2.3× temporal
Response	**Between-year sd**	**Between-corridor sd**	**Ratio**
ndvi_mean	0.0127	0.1194	**9.4× spatial**
CDEI	0.0031	0.0126	**4.1× spatial**
lst_mean	1.03	2.88	2.8× spatial

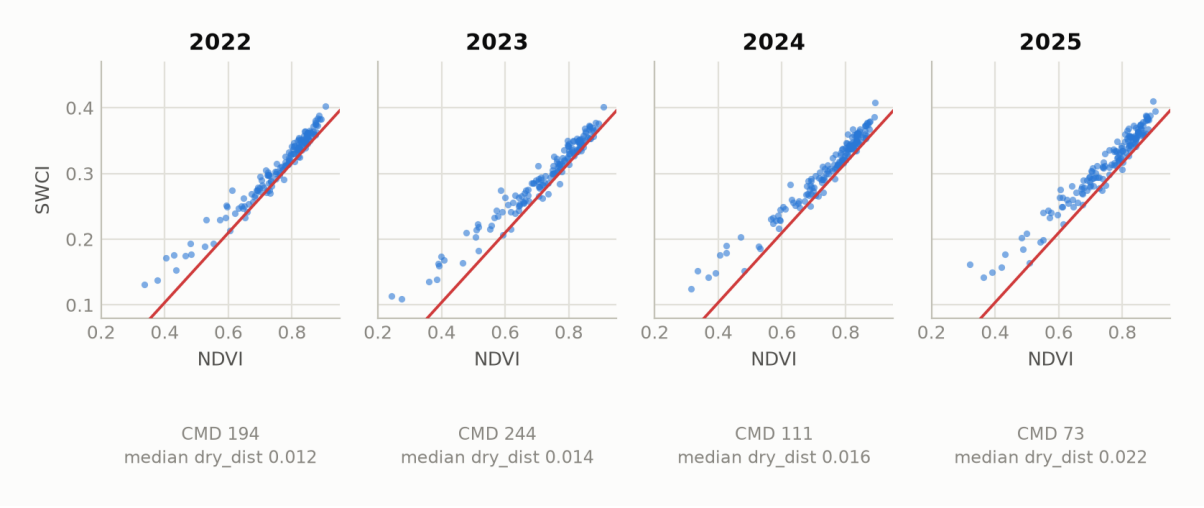

Figure 1. Surrey's 153 corridor polygons in the NDVI–SWCI plane, one panel per summer, with the fitted dry edge. As summer moisture swings from CMD 244 to 75 mm the whole cloud shifts (median distance from the edge 0.014, 0.012, 0.016, 0.022 for 2023, 2022, 2024, 2025, so the wettest summer sits furthest from the edge but the ordering is not monotone) while the spread *between* polygons stays essentially fixed, so climate moves the panel through time without ordering the polygons within it.

3.2 Extent 2 (transect): model comparison and sensitivity analyses

The identical experiment on 1,200 rows (300 stands × 4 summers) at a 25 km cell:

Table 3. Transect: the coarse model is better on the point estimate on 1,200 rows at a 25 km cell. Neither paired interval excludes zero.

Model	**RMSE**	**MAE**	R^2
A (scale-free)	0.02565	0.01583	−0.138
B (coarse 25 km)	0.02472	0.01560	+0.003
Paired difference	**Estimate**	**95% CI**	**A better**
ΔRMSE	+0.00093	[−0.00061, +0.00257]	40% of folds
ΔR^2	−0.14127	[−0.40788, +0.04426]	40% of folds

Model A's position on its own terms is not in doubt: at $R^2 = -0.138$ it does materially worse than predicting the mean.

This is **not** a result showing the coarse model is good. A model given fine-scale detail goes backwards here, while the coarse average is merely uninformative. The comparison is meaningful even though neither model is useful in absolute terms.

Precondition 2 now passes. Over the transect, coarsening removes 13.4% of the predictors' spatial variance at 4 km rising to 48.4% at 25 km, with the leading predictors losing 58–68% (Supporting Information, Table S1), so A and B are different models.

The skill gate passes only marginally. The gate requires $\max(R^2_A, R^2_B) > 0$. Model A is well below it at −0.138; Model B clears it at +0.003, which is to say it does almost exactly what predicting the mean would do.

Sensitivity analyses. Sensitivity is reported in full in Supporting Information: the coarse cell size (Table S1), the random seed (Table S2), the learner's hyperparameters (Table S3) and the canopy-density confound in the response (Table S4). Model B leads at every cell size, in 20 of 20 seeds and in eight of eight forest configurations, yet the paired interval spans zero everywhere except the least constrained forest, where Model A overfits hardest. Residualising the response on 24 independently measured stand attributes raises the point estimate ($\Delta\mathrm{RMSE} = +0.00113$ at 25 km, $+0.00146$ at 12 km, the latter excluding zero). Removing stand greenness as well inverts it (-0.00064 and -0.00046, the latter excluding zero for Model A), but greenness is a constitutive input to CDEI and neither model has skill on that response.

The fold scheme. Every configuration above holds the evaluation design fixed, whereas replacing the k-means spatial blocks with random folds over the same units, grouping otherwise unchanged, reverses the transect comparison, so that across seven seeds $\Delta\mathrm{RMSE}$ moves from a $+0.00085$ to $+0.00093$ band under blocking to -0.00040 to -0.00073 under random folds, non-overlapping bands, and the paired MAE interval, which spans zero under blocking, excludes zero for Model A in all seven (Supporting Information, Table S5). Wadoux et al. [24] argue that spatial cross-validation misestimates map accuracy, and Linnenbrink et al. [25] that random folds suit interpolation and blocked folds extrapolation, with most applications between; this comparison sits on that fault line. Blocking is retained because the question asked is whether downscaled climate transfers to unseen units, which is the extrapolation case, and because only the fine-resolution arm can exploit the neighbours random folds admit (Supporting Information, Section S2). The direction of the transect result is therefore a property of the evaluation design as much as of the data.

Why neither model has spatial skill: a different reason than at Surrey. What changed between the extents is the *predictor*, which over Surrey varied mostly in time and over the transect varies in space, so this is *not* an orthogonal-axes problem in the space-versus-time sense. The response is spatially dominant at both extents, varying 11× more across space than across time here (between-stand sd 0.0241 against 0.0022 between years) and 4.1× over Surrey (Table 2). There is a strong spatial signal to predict and the climate gradient does not align with it: across the 300 stands the Spearman correlation between stand-mean CDEI and the spatial gradient does not exceed $|\rho| = 0.15$ for any climate variable or for elevation.

3.3 A cardinality-matched control on the fine arm

Model A and Model B differ in the number of distinct predictor vectors they carry as well as in resolution, so a deficit for A can come from either. Model A′ separates them: it holds A's values and permutes them between sites *within* each coarse cell, drawing one permutation per cell and applying it to every summer, so each site keeps its own time series and moves to another location inside the cell it already occupied. Cardinality, the marginal distribution of every predictor, each site's temporal structure and every cell-by-year mean are left untouched, which is checked rather than assumed: Model B drifts by 2×10^{-16} and f is unchanged to six decimals, with 290 of 300 sites relocated. What A′ destroys is only the association between a predictor and the place it was measured.

Over 20 seeds A and A′ are indistinguishable. The paired difference is -0.00008 in RMSE, no interval excludes zero in any seed, and A′ is the better of the two in 45% of them, while Model B beats both in every seed. On R^2 A′ scores -0.100 against A's -0.138, so removing the within-cell detail does not hurt and may help slightly. The fine detail Model A carries is therefore not information this learner extracts, and A's deficit against B at this extent is not evidence that coarsening *helps*.

The control cannot separate detail that is noise from detail that is signal too weak for this learner at this sample size, which is the same detection limit the rest of the paper concerns. It agrees with the skill gate, which already found the removed variance uncorrelated with the response, so the two reach

the same conclusion by different routes.

3.4 Detection threshold in simulation

With the generative effect held constant, the detection rate rises with the fraction of predictor spatial variance the coarsening removes (Spearman $\rho = +0.833$ over the eight levels), while climate's share of response variance stays at 0.336 ± 0.001 throughout (Table 4).

F1 holds ($\rho = +0.833 \geq +0.8$), though as a trend rather than a level-by-level increase: the four lowest-f levels are statistically indistinguishable from one another. F2 holds decisively, with a minimum rate of 0.07 $[0.03, 0.14]$ against the 0.25 required, so there is a regime in which a genuine effect is unrecoverable and a null is therefore uninformative. **F3 fails at $k = 5$.** No level reaches 0.80; the maximum is 0.63 $[0.53, 0.72]$ at $f = 0.81$. That failure is not a property of f alone, since every level above was validated on five blocks and the interval is a bootstrap over those five. Sweeping the block count at fixed f and fixed n lifts detection at $f = 0.81$ from 0.68 to 0.84 as k goes from 5 to 20, reaching the bar, while at $f = 0.71$ it does not move (Supporting Information, Table S6). Detectability is set by the number of validation units as well as by f, so no value of f can be recommended on its own as licensing trust.

Below $f \approx 0.4$ detection is 0.12 $[0.09, 0.15]$ pooled over the runs rather than the tabulated levels, and in the band $0.30 \leq f < 0.45$, which passes a nominal 30% gate, it is 0.20 $[0.13, 0.29]$, so a study clearing 30% recovers a guaranteed effect about one time in five, while above the floor it is 0.45 $[0.40, 0.50]$. Neither figure moves when the sample rises from 300 sites to 2,400 at $k = 5$ (Section S14), which is what an interval built from five units would do.

What this does *not* establish matters as much, because the growth of the advantage *available* to Model A with f is an identity rather than a finding, B being A upscaled so that what B loses is exactly the within-cell residual whose variance f measures. What cannot be derived, and is the point of simulating, is where the detection threshold falls for a given sample size, fold structure and learner.

The calibration is conditional on the geometry it was measured in. The sweep above holds the study geometry fixed at Extent 2's, and re-running it across five geometries that vary the domain and the cell size together, with the correlation length swept in units of the cell so that f stays comparable, separates two claims that had travelled together. Detection still rises with f inside every geometry (Spearman $+0.84$ to $+0.99$), so each is a calibration curve in its own right. At matched f the five do not agree: near $f = 0.42$ detection runs from 0.20 $[0.08, 0.42]$ where the domain spans four coarse cells to 0.75 $[0.53, 0.89]$ where it spans ten, and those two intervals do not overlap. A control pair sharing a ratio but differing in absolute size returns the same f at all eight levels and a detection rate that agrees within its interval at all eight, so what moves the curve is how many coarse cells the extent contains rather than how large the extent is. The mechanism shows in the arms themselves, since Model A reaches positive CV R^2 only where the cells are many and the fine arm therefore has enough independent structure to learn from. A fraction licenses a power figure only alongside the geometry that produced it, which is a limit on the calibration rather than on the diagnostic, since f itself is a property of the panel and the cell (Supporting Information, Table S12).

Surrey's $f = 0.123$ sits where this design recovers a real effect between about one time in six and one in thirty, depending on the local driver's assumed correlation length (Supporting Information, Section S16), so the Surrey null is uninformative rather than negative. The transect's $f = 0.484$ sits just above the floor, in a regime with roughly one chance in four of resolving a real effect, which is what it did show: a consistent direction whose interval spans zero. No single null at moderate f should be read as evidence of no effect, this paper's transect included.

The transect null is not equivalence either, and that is testable on the data. The argument above is a simulation argument, and the same question can be put to the panel directly by two one-sided tests, which ask whether the paired difference lies inside a band too small to act on rather than whether it differs from zero. That band has to be derived rather than chosen, since setting it at the endpoint under test would guarantee its own verdict. Perturbing each corridor's CDEI and re-ranking the applied product shows a shift of 0.00100 leaves nineteen of the top twenty corridors in place, while 0.00257 displaces two and 0.00500 turns over a fifth, so 0.00100 in RMSE units, 4.0% of Model B's, is the largest band the ranking is indifferent to. The 90% interval is [−0.00043, +0.00243] and does not fall inside it, so the comparison establishes neither a difference worth acting on nor the absence of one. Two limits belong with the numbers, in that the simulated effect size is set by the weights above rather than calibrated to the observed one, making these rates against a constructed truth, and shortening L to raise f eases the blocked-CV extrapolation for *both* arms, so the marginal R^2 columns are not comparable down the table, whereas the paired difference is.

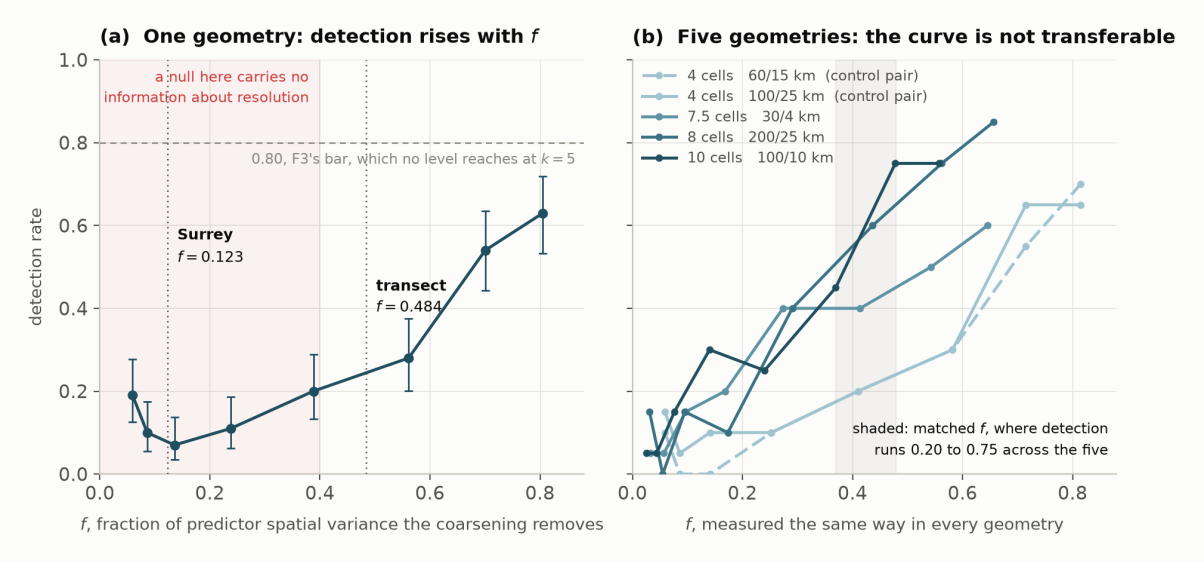

Figure 2. The precondition, calibrated and then limited. **(a)** At Extent 2's geometry, detection rises with f (Spearman $\rho = +0.833$ over the eight levels, 100 synthetic panels each, 95% Wilson intervals). Surrey and the transect are marked where they fall. The shaded region is the floor the paper defends; F3's bar of 0.80 is never reached at $k = 5$. **(b)** The same sweep in five geometries, 20 panels per level, coloured by how many coarse cells the domain spans. Every curve rises, so each is a calibration in its own right, and at matched f they run from 0.20 to 0.75. The control pair shares a ratio at 1.7× different absolute size, so the separation is the cell count rather than the extent.

Table 4. Detection rate against the fraction of predictor spatial variance removed, with the generative advantage held constant across rows. One hundred synthetic panels per level; a detection is a paired ΔRMSE interval excluding zero in the generative direction, reported as a proportion with a 95% Wilson interval. Surrey measures $f = 0.123$ and the transect $f = 0.484$.

f removed	Detection rate	95% CI	Mean ΔRMSE	R^2_A	R^2_B
0.060	0.19	[0.13, 0.28]	−0.021	−0.379	−0.409
0.086	0.10	[0.06, 0.17]	−0.005	−0.398	−0.396
0.136	0.07	[0.03, 0.14]	+0.003	−0.347	−0.331
0.238	0.11	[0.06, 0.19]	−0.005	−0.250	−0.251
0.388	0.20	[0.13, 0.29]	−0.041	−0.155	−0.208
0.561	0.28	[0.20, 0.37]	−0.072	−0.088	−0.179
0.700	0.54	[0.44, 0.63]	−0.101	−0.032	−0.157
0.805	0.63	[0.53, 0.72]	−0.116	+0.003	−0.139

3.5 Drivers of between-stand water stress

Under the same blocked CV, terrain and stand structure together are the only predictor family to reach CV R^2 above +0.02 (+0.029, positive in 80% of held-out folds against climate's 40%), and adding climate on top makes it worse (+0.029 → +0.020). By permutation importance on a held-out spatial block the leading drivers are all local, and no climate variable appears (Supporting Information, Section S7, Table S9).

Part of that may be a vegetation-type artifact: leading species and broadleaf/conifer class rank highest, which is consistent with CDEI partly encoding canopy *type* rather than water stress per se, and Table S4 bounds what that artifact can account for.

4 Discussion

4.1 Interpretation of the negative result

If finer climate data improved polygon-scale water-stress prediction, obtaining it would be a sensible use of constrained effort. It is not shown to improve that prediction here, at a power this paper puts near one in four (Section 3.4). The cost is not the data, which are CC-BY [2], but the acquisition and analysis the product commits an ecologist to, and the attention it draws from what does carry signal.

The microclimate literature offers a coherent explanation [7]: if the conditions a stand actually experiences are set by canopy structure, slope aspect, soil depth and rooting access, then refining the resolution of a macroclimate product may not resolve them, because the responsible processes are not represented in that macroclimate product at any resolution, and Section 3.5 is consistent with it.

That explanation has the same limit the model has: soil depth and rooting access are named as mechanism but were never measured, since the stand-structure features reach species, age, height and crown closure rather than the belowground properties (Section 4.3). The explanation is therefore consistent with these results without being tested by them, and the weakness of the explanatory model, best CV R^2 around +0.03, is what that distinction looks like quantitatively.

4.2 A transferable precondition for resolution comparisons

A resolution hypothesis can be untestable while looking like a null result, and that condition is measurable in advance.

Over Surrey, Models A and B are statistically indistinguishable, and that null is uninformative about resolution, which leaves the hypothesis unresolvable *at that extent* rather than untestable in principle. Coarsening removed only 12.3% of the predictors' spatial variance there, leaving the two predictor sets correlated at a median $r = 0.998$ across the 14 climate variables ($r = 0.922$ between polygons), so a null difference between models that near-identical says nothing about resolution. Over the transect the same operation removed 48.4% and pulled those correlations to $r = 0.847$ and $r = 0.707$. What separated an uninformative experiment from an informative one was not sample size, learner or index, but whether the extent was large enough that coarsening destroyed something.

The precondition others can apply directly is to measure, before running any resolution comparison, the fraction of predictor spatial variance the coarsening removes, and to treat a small fraction as disqualifying rather than as a finding. The simulation constrains that advice in one direction only (Section 3.4). A floor below which a null means nothing can be defended; a threshold above which a result may be believed cannot, because detectability also turns on the block count the comparison is validated on. I originally set the gate at 30% by judgement rather than derivation, and a study clearing exactly that bar recovers a guaranteed effect about one time in five, too weak to serve as a licence.

The contrast cutoff changes no verdict reported here, because wherever it fails the skill gate fails independently, so it could sit anywhere between roughly 13% and 48% without changing a label.

The quantity itself is not new, since aggregating a variable over blocks and asking how much of its variance survives is a case of the change-of-support problem of geostatistics [26]. The split between the variance lying within blocks and the variance lying between them is classical, and given a fitted variogram it can be obtained without coarsening anything and measuring it [27]. Equation (4) is that decomposition, estimated directly on the panel in hand instead of derived from a covariance model.

What the theory does not supply is the part this paper adds. That a coarsening removes 12% of spatial variance says nothing by itself about whether a comparison at that level can resolve a real difference, since that turns on sample size, fold structure and the learner, none of which the decomposition sees. Goodling et al. [28] reach a neighbouring conclusion from the evaluation side, that a spatial model scored at one scale has not been scored, and f is the input-side quantity that says which scales were ever distinguishable. Section 3.4 converts the fraction into a statement about power, and it is the reason to report f rather than treat support as a qualitative worry.

The same reasoning reaches the Modifiable Areal Unit Problem [13, 14]. If relationships change with the size and shape of the aggregation units, the extent does not merely affect the answer, it determines which questions are answerable, which is why more data inside Surrey would not have resolved H_1.

On one point of ordering, the precondition gate was written into the code *after* Surrey returned its null, because the null needed interpreting and a bare "not supported" would have misrepresented it. That ordering is a legitimate target for criticism, and disclosing it is the appropriate response. It is not, however, what the literature calls HARKing: a gate is an inclusion criterion, not a hypothesis presented as though it had been predicted in advance. It also concerns the experiment's *capacity to detect a difference*, which can be evaluated without reference to the outcome, since the diagnostic never sees the response variable. And it was applied unchanged to the transect, where it permitted a verdict contradicting the original hypothesis.

4.3 Limitations

1. **CDEI is a relative feature-space index, not a soil-moisture measurement**, and has not been validated against any ground observation. No public station network covers either extent at the polygon scale the analysis works at. Closing that gap needs paired in-situ measurements across the CDEI range in the same summers, a field campaign this study did not have and a companion paper [17] sets out.

2. **The explanatory model is weak in absolute terms** (best CV $R^2 \approx +0.03$ to $+0.04$) and sup-

ports a directional statement about *which family of drivers matters*, not prediction. The directional claim rests not on that magnitude, which would not carry it, but on the consistency and the separation, since terrain-plus-structure is positive in 20 of 25 held-out folds against climate's 10, over five independent partitions, and the families are separated by 0.167 in R^2 (Supporting Information, Table S9).

3. **Unmeasured drivers.** Soil depth, rooting depth, groundwater access, irrigation and management history are absent from the feature set and are plausible major contributors to the unexplained variance.

4. **A single downscaling product and a single learner.** The test covers ClimateBC with a random forest; other downscaling products or model families were not tested, though the diagnosis (orthogonality of gradient and response) is a property of the *fields*, not of the learner.

5. **One transect, one region.** The transect comparison rests on a single Fraser Valley elevation gradient, and whether it generalises to other terrain and climates is untested.

6. **The transect direction depends on the evaluation design.** Model B's advantage holds at every cell size, seed and forest setting under spatial blocking, and reverses under random folds (Section 3.2). The one blocked exception is the greenness-residualised response (Table S4), where the index has been stripped of one of the two axes it is defined on. Blocking matches the transfer question asked here, but a study whose target is the sampled units themselves would be scored the other way, and on this panel would reach the other answer.

7. **Analysis-unit mismatch between extents.** Surrey uses narrow urban corridor polygons; the transect uses VRI forest stands. The two are comparable in area, median 5.8 ha against 7.4 ha, but not in shape: effective width (twice area over perimeter) has a median of 39.7 m for Surrey corridors against 94.8 m for the stands, and the gap is roughly two-fold on either convention (49.4 m against 94.8 m unbuffered, 39.7 m against 87.0 m buffered). Stands were chosen because 39.7 m spans barely one or two 20–30 m pixels, so a Surrey corridor's zonal summary is substantially made of pixels it shares with whatever adjoins it, but the choice leaves the two extents unlike each other as ecological units.

8. **The skill gate threshold is a product of convenient construction.** The verdict rule requires $\max(R^2_A, R^2_B) > 0$, and no citation is available for it, because it is a heuristic introduced here in which $R^2 = 0$ is a natural but not a principled boundary. This paper's own results land on opposite sides of it at two defensible cell sizes while the paired difference is unchanged (Supporting Information, Table S2), and at 12 km it sets the label on its own. A rule stated in terms of the paired interval alone would be more robust, and is therefore the recommended form for reuse.

9. **The between-year variance estimates rest on four summers.** The orthogonality diagnosis that Surrey's null turns on compares each variable's between-year standard deviation with its between-polygon one, and four summers is a thin basis for the between-year half of every ratio. Supporting Information gives what those ratios would need to be disturbed, and why the qualitative separation survives the small n even though the individual figures should not be read as precise.

4.4 Conclusion

Scale-free downscaled climate data, ClimateBC's point-queried surface adjusting an interpolated 4 km field by elevation rather than producing a finer grid, was not shown to improve polygon-scale water-stress prediction. At a 25 km cell, the resolution of the regional models such downscaling is meant to beat, the coarse model led on the point estimate by a margin whose paired 95% confidence interval spans zero ($\Delta\mathrm{RMSE} = +0.00093$, CI $[-0.00061, +0.00257]$). Model B leads at every cell size, in 20 of 20 seeds and in eight of eight forest settings, while the paired interval excludes zero in one of those and

in no seed. Those configurations all held the fold scheme fixed, and substituting random folds reverses the sign (Section 3.2), so that consistency is a property of the design space explored and not of the data alone. Because Model B was built by spatially averaging the *same* ClimateBC variables Model A used, the result cannot be attributed to a poor downscaling or interpolation choice.

Past this pair of extents, the diagnostic generalises where the comparison does not: measuring the fraction of predictor spatial variance a coarsening removes costs no model fit and never touches the response, and it separates an answerable resolution question from an unanswerable one, though only in one direction, since below $f \approx 0.4$ a null carries no information while above it detectability turns on how many units the validation interval is built from as much as on f (Supporting Information, Table S6).

Data and code availability

The pipeline and the walkthrough notebooks are openly available at `https://github.com/concretemicrowave/surrey-biome-pipeline`, and the two processed analysis panels behind every result reported here are archived at `https://doi.org/10.5281/zenodo.22050486`. The package is `coarsegate` [29] (`https://github.com/concretemicrowave/coarsegate`), installable from PyPI. The diagnostic and its calibration are released as an installable package, so the precondition can be measured without reimplementing it. Raw and intermediate rasters are not redistributed. The environment is Python 3.12.13 with scikit-learn 1.9.0 [20], NumPy 2.5.1 and pandas 3.0.3, and all randomness is seeded (seed 26910). Acquisition and aggregation use GeoPandas, rasterio, xarray and exactextract, which returns exact pixel coverage fractions rather than a centroid rule.

Conflict of interest

None.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgements

The author designed the study, ran and verified every analysis reported here, and is responsible for all content and for any errors that remain.

References

[1] Tongli Wang, Andreas Hamann, Dave Spittlehouse, and Carlos Carroll. Locally downscaled and spatially customizable climate data for historical and future periods for north america. *PLOS ONE*, 11(6):e0156720, 2016. doi: 10.1371/journal.pone.0156720.

[2] Tongli Wang, Andreas Hamann, and Dave Spittlehouse. ClimateBC: scale-free climate data for British Columbia. Software, web API v7.30, 2026. URL `https://climatebc.ca/help/climateBC/help.htm`. Accessed 22 August 2026. Downscales 1961–1990 monthly normals,

unified at 4 km, to point locations by bilinear interpolation with a local elevation adjustment.

[3] Ping Yang, Daniel P. Ames, André Fonseca, Danny Anderson, Rupesh Shrestha, Nancy F. Glenn, and Yang Cao. What is the effect of LiDAR-derived DEM resolution on large-scale watershed model results? *Environmental Modelling & Software*, 58:48–57, 2014. doi: 10.1016/j.envsoft.2014.04.005.

[4] Mark Pogson and Pete Smith. Effect of spatial data resolution on uncertainty. *Environmental Modelling & Software*, 63:87–96, 2015. doi: 10.1016/j.envsoft.2014.09.021.

[5] Pieter De Frenne, Florian Zellweger, Francisco Rodríguez-Sánchez, Brett R. Scheffers, et al. Global buffering of temperatures under forest canopies. *Nature Ecology & Evolution*, 3(5):744–749, 2019. doi: 10.1038/s41559-019-0842-1.

[6] Florian Zellweger, David Coomes, Jonathan Lenoir, Leen Depauw, et al. Seasonal drivers of understorey temperature buffering in temperate deciduous forests across Europe. *Global Ecology and Biogeography*, 28(12):1774–1786, 2019. doi: 10.1111/geb.12991.

[7] David H. Klinges, J. Alex Baecher, Jonas J. Lembrechts, Ilya M. D. Maclean, Jonathan Lenoir, Caroline Greiser, Michael Ashcroft, Luke J. Evans, Michael R. Kearney, Juha Aalto, Isabel C. Barrio, Pieter De Frenne, Joannès Guillemot, Kristoffer Hylander, Tommaso Jucker, Martin Kopecký, Miska Luoto, Martin Macek, Ivan Nijs, Josef Urban, Liesbeth van den Brink, Pieter Vangansbeke, Jonathan Von Oppen, Jan Wild, Julia Boike, Rafaella Canessa, Marcelo Nosetto, Alexey Rubtsov, Jhonatan Sallo-Bravo, and Brett R. Scheffers. Proximal microclimate: Moving beyond spatiotemporal resolution improves ecological predictions. *Global Ecology and Biogeography*, 33(9):e13884, 2024. doi: 10.1111/geb.13884.

[8] Pieter De Frenne, Jonathan Lenoir, Miska Luoto, Brett R. Scheffers, et al. Forest microclimates and climate change: Importance, drivers and future research agenda. *Global Change Biology*, 27 (11):2279–2297, 2021. doi: 10.1111/gcb.15569.

[9] Jonas J. Lembrechts, Juha Aalto, Michael B. Ashcroft, Pieter De Frenne, Martin Kopecký, et al. SoilTemp: A global database of near-surface temperature. *Global Change Biology*, 26(11):6616–6629, 2020. doi: 10.1111/gcb.15123.

[10] Ilya M. D. Maclean, Jonathan R. Mosedale, and Jonathan J. Bennie. Microclima: An R package for modelling meso- and microclimate. *Methods in Ecology and Evolution*, 10(2):280–290, 2019. doi: 10.1111/2041-210X.13093.

[11] Daniel L. Perret, Margaret E. K. Evans, and Dov F. Sax. A species' response to spatial climatic variation does not predict its response to climate change. *Proceedings of the National Academy of Sciences*, 121(1):e2304404120, 2024. doi: 10.1073/pnas.2304404120.

[12] Amy L. Angert. The space-for-time gambit fails a robust test. *Proceedings of the National Academy of Sciences*, 121(4):e2320424121, 2024. doi: 10.1073/pnas.2320424121.

[13] Stan Openshaw. *The Modifiable Areal Unit Problem*. Number 38 in Concepts and Techniques in Modern Geography (CATMOG). Geo Books, Norwich, UK, 1984. ISBN 0-86094-134-5.

[14] Dennis E. Jelinski and Jianguo Wu. The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology*, 11(3):129–140, 1996. doi: 10.1007/BF02447512.

[15] City of Surrey. Green Infrastructure Network Corridors (OpenData MapServer, layer 1). City of Surrey Open Data, ArcGIS MapServer, 2026. URL https://gisservices.surrey.ca/arcgis/rest/services/OpenData/MapServer. Accessed 21 July 2026. Catalogue: https://opendata-surrey.hub.arcgis.com/. The corridor attributes are those published in Appendix J of the City's Biodiversity Conservation Strategy.

[16] British Columbia Ministry of Forests. Vegetation resources inventory (VRI): Forest vegetation composite rank 1 layer. BC Data Catalogue, 2024. https://catalogue.data.gov.bc.ca/dataset/vri-forest-vegetation-composite-rank-1-layer-r1.

[17] Joshua Wang. An exploratory satellite water-stress monitor for Surrey's green infrastructure corridors: construction, independent validation, and the limits of both. Manuscript in preparation, 2026.

[18] European Space Agency and Airbus. Copernicus DEM: Global and european digital elevation model, GLO-30. Copernicus Space Component Data Access, 2022. https://spacedata.copernicus.eu/collections/copernicus-digital-elevation-model.

[19] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

[22] Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11(1):4540, 2020. doi: 10.1038/s41467-020-18321-y.

[23] Hanna Meyer and Edzer Pebesma. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9):1620–1633, 2021. doi: 10.1111/2041-210X.13650.

[24] Alexandre M. J.-C. Wadoux, Gerard B. M. Heuvelink, Sytze de Bruin, and Dick J. Brus. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, 457:109692, 2021. doi: 10.1016/j.ecolmodel.2021.109692.

[25] Jan Linnenbrink, Jakub Nowosad, and Hanna Meyer. Moving beyond spatial and random cross-validation in environmental modelling: a call for prediction-domain adaptive evaluation, 2026. arXiv preprint.

[26] Carol A. Gotway and Linda J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648, 2002. doi: 10.1198/016214502760047140.

[27] Jean-Paul Chilès and Pierre Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 2 edition, 2012. doi: 10.1002/9781118136188.

[28] P. Goodling, K. Belitz, P. Stackelberg, and B. Fleming. A spatial machine learning model developed from noisy data requires multiscale performance evaluation: Predicting depth to bedrock in the Delaware river basin, USA. *Environmental Modelling & Software*, 179:106124, 2024. doi: 10.1016/j.envsoft.2024.106124.

[29] Joshua Wang. coarsegate: measure what a coarsening removes, 2026. URL https://doi.org/10.5281/zenodo.22132564.

Table S1. Cell-size sensitivity on the transect. Model A is the same model throughout ($R^2 = -0.138$); only the coarsening changes. At 4 km the contrast gate fails, so no verdict is issued. At 25 km the median predictor loses about half its spatial variance and the leading ones (Eref, Tmax, CMD) lose two-thirds, while 10 cells still leave enough geographic spread to spatially block Model B. The evaluation is the paired interval; the categorical labels the skill gate would attach are discussed in the text but deliberately not tabulated (main text, Limitations).

Cell	Cells	Var. removed	R^2_B	Paired ΔRMSE (95% CI)	Paired ΔR^2 (95% CI)	Excl. 0
4 km	101	13.4%	n/a	contrast gate fails, comparison not interpretable		
8 km	40	29.0%	−0.040	+0.00033 [−0.00124, +0.00237]	−0.097 [−0.399, +0.092]	no
12 km	23	34.4%	−0.003	+0.00095 [−0.00070, +0.00270]	−0.135 [−0.400, +0.051]	no
25 km	**10**	**48.4%**	+0.003	+0.00093 [−0.00061, +0.00257]	−0.141 [−0.408, +0.044]	no

Supporting Information

Measure what the coarsening removes: a precondition for interpreting spatial-resolution comparisons in ecological models

Joshua Wang

What is in this document. Part 1 reports results that support the main text: the four sensitivity analyses the result rests on, the Surrey skill screen, the between-year variance estimates, and the full driver comparison for stand water stress. Part 2 gives the methods detail behind them: the study extent and sampling, the construction of CDEI, the climate predictors and their acquisition, the experiment's implementation, the explanatory analysis, and the full specification of the precondition simulation.

Part 1. Additional results and sensitivity analyses

S1 Sensitivity analyses

These four analyses are reported here rather than in the main text because Methods in Ecology and Evolution excludes Supporting Information from the word limit. The prose and tables are the manuscript's own and are unchanged; nothing has been shortened in moving them. Section and table numbers referred to in the main text as Tables S1–S4 appear below in that order.

Sensitivity to the coarse cell size. Because the verdict label turns on three thousandths of R^2, the experiment was re-run at every cell size whose contrast is defensible, changing nothing else (Table S1). The pattern separates the two things that could be meant by "the result":

At 12 km the paired difference is indistinguishable from the 25 km one (ΔRMSE = +0.00095 against +0.00093), and neither interval excludes zero. The skill gate would nonetheless return INCONCLUSIVE there and NOT SUPPORTED at 25 km, because Model B lands three thousandths on the other side of a threshold at exactly zero. That is why this paper evaluates on the interval and not on the label. The point estimate and its direction hold at two independent cell sizes; the categorical label changes between them while the measured difference does not. At 8 km the difference shrinks, which is consistent with the mechanism: 8 km removes only 29% of the predictors' spatial variance, leaving little contrast to detect. So the direction is not an artifact of choosing 25 km, and the verdict word would

Table S2. Seed sensitivity on the transect. Each row aggregates independent re-runs of the complete experiment, re-seeding the k-means spatial blocking and the random forest together. The sign of R^2_B is deterministic given the cell size, and the paired difference excludes zero in none of the 40 runs. Values are mean ± sd across seeds.

Cell	**Seeds**	R^2_A	R^2_B	$R^2_B > 0$	**Paired** ΔRMSE	**CI excl. 0**
12 km	20	-0.1378 ± 0.0033	-0.0029 ± 0.0014	**0 / 20**	$+0.00094 \pm 0.00003$	0 / 20
25 km	20	-0.1378 ± 0.0033	$+0.0033 \pm 0.0005$	**20 / 20**	$+0.00093 \pm 0.00003$	0 / 20

Table S3. Hyperparameter sensitivity on the transect at a 25 km cell. Model B is better on the point estimate in all eight configurations, but the paired interval excludes zero in only one, the unconstrained single-sample leaf, which is also the configuration in which Model A overfits hardest. R^2_B is stable near zero throughout.

Trees	**Leaf**	**max_features**	R^2_A	R^2_B	**Paired** ΔRMSE **(95% CI)**	**Excl. 0**
400	1	sqrt	-0.249	$+0.001$	$+0.00222$ [$+0.00020$, $+0.00393$]	yes
400	**3**	**sqrt**	-0.138	$+0.003$	$+0.00093$ [-0.00061, $+0.00257$]	**no**
400	5	sqrt	-0.094	$+0.003$	$+0.00049$ [-0.00072, $+0.00203$]	no
400	10	sqrt	-0.060	$+0.004$	$+0.00023$ [-0.00086, $+0.00161$]	no
200	3	sqrt	-0.143	$+0.003$	$+0.00097$ [-0.00060, $+0.00263$]	no
800	3	sqrt	-0.136	$+0.003$	$+0.00090$ [-0.00065, $+0.00255$]	no
400	3	0.5	-0.191	$+0.003$	$+0.00146$ [-0.00060, $+0.00350$]	no
400	3	1.0	-0.236	$+0.002$	$+0.00195$ [-0.00050, $+0.00440$]	no

be a threshold artifact that the underlying measurement is not. The conclusion rests on the paired intervals.

Sensitivity to the random seed. A margin of three thousandths might simply be noise: an R^2 of $+0.003$ could plausibly be an artifact of which random partition the k-means blocking happened to draw, or of the forest's own randomness. This was tested directly rather than argued. The entire experiment was re-run across independent seeds, each re-seeding both the spatial blocking and the forest, changing nothing else (Table S2).

The measurement is stable, and the two claims that follow from it point in opposite directions. At 25 km, R^2_B is positive in every one of 20 seeds and its seed-to-seed standard deviation (0.0005) is roughly a sixth of its distance from zero, and at 12 km it is negative in every one of 20 seeds. The verdict label is a deterministic function of the cell size and not a lucky draw, a sharper statement than Table S1 alone can make, because it establishes that the 12 km / 25 km disagreement is a real property of the coarsening and not sampling variability. The paired difference is equally reproducible, with a seed-to-seed standard deviation two orders of magnitude below its own value, but reproducibly *indistinguishable from zero*: the interval excludes zero in **0 of 40 runs across both cell sizes**. The blocking seed is therefore not a source of doubt about either finding.

This establishes that $+0.003$ is a stable measurement, so the gate is passed reproducibly. It does not make $+0.003$ a *meaningful* amount of skill, and nothing in this paper claims it is. The weakness that remains is in the gate's design and not in the number: a threshold of exactly $R^2 = 0$ separating "informative comparison" from "uninformative" is a convenience, not a principled boundary, and this paper's own results land on opposite sides of it at two defensible cell sizes. That is a limitation of the method proposed here, and it is recorded as one in the main text's Limitations.

Sensitivity to the learner's hyperparameters. The forest's settings were fixed a priori rather than tuned (main text, Materials and Methods), so the experiment was re-run at 25 km across eight configurations, changing nothing else (Table S3). The direction is unanimous; the significance is essentially absent.

Three things follow. First, Model B is better than Model A on the point estimate in *every* configuration,

and R^2_B never leaves the interval $[+0.001, +0.004]$: the coarse model's position at the no-skill line is a stable property, not a tuning artifact. Second, the consistency of direction does not become statistical significance. The paired $\Delta\mathrm{RMSE}$ interval spans zero in seven of eight configurations, including the pre-registered one; it excludes zero only at `min_samples_leaf` = 1, the least constrained forest and the one in which Model A overfits hardest. The gap is therefore visible in the point estimate at every setting and resolvable at none of the sensible ones. Third, the pattern is mechanistically coherent: R^2_A climbs monotonically from -0.249 to -0.060 as the leaf size grows, while R^2_B does not move. Model A's deficit is substantially overfitting to fine-grained climate detail, and constraining the forest suppresses that overfitting, which is the same claim the paper makes about that detail being noise. But it means the *significance* of the gap depends on how hard the learner is allowed to chase it.

The coarse model is therefore never worse under any forest setting tested, and never distinguishably better under any of the reasonable ones. That is weaker than a bare categorical verdict would convey, and it is as far as the evidence goes. The Conclusion carries it in that form.

The narrowing gap might suggest the finding is a tuning artifact and not a property of the data. Three points argue against that. The gap narrows but does not invert or vanish: at `min_samples_leaf` = 10, Model A is still worse by 0.063 in R^2, and Model B remains pinned near zero throughout, so what is lost is significance and not sign. Heavy regularisation closes the gap by driving *both* models toward the mean predictor, which is the null model, instead of by rescuing Model A. And most importantly, the claim that the fine-scale climate gradient carries no usable signal for this response does not rest on the forest comparison at all: it is independently supported by the rank correlations below ($|\rho| \leq 0.15$ between stand-mean CDEI and every component of the spatial gradient) and by the explanatory analysis in the main text, where adding climate to a terrain-and-structure model makes it worse. Only one of those three lines of evidence involves a hyperparameter.

Sensitivity to the canopy-density confound in the response. CDEI is entangled with canopy density at both extents, measured below and independently reported for the Surrey application in a companion paper [1]. That bears on this comparison. Specifically, if CDEI is substantially a greenness measure, then "fine-scale climate does not predict CDEI" may be a statement about where the broadleaf stands are and not about water stress. The confound at Extent 2 had not previously been quantified. It is $\rho = -0.314$ between stand-mean CDEI and stand-mean NDVI ($n = 300$ stands), against Surrey's -0.347. The confound is the same size at both extents, so it is not answered by pointing at the transect's larger, better-behaved polygons. Crown closure gives $\rho = +0.180$ ($p = 0.002$), while stand height and age are indistinguishable from zero ($+0.005$ and $+0.045$), and land-cover class alone accounts for only 2.7% of between-stand CDEI variance. The confound runs along greenness, not along stand age, structure or broadleaf-versus-conifer type.

The test is then direct: partial canopy structure out of the response and re-run the identical comparison. Residualisation uses no climate information and is applied to Model A and Model B alike, so the paired difference remains a like-for-like contrast; it is fitted over the whole panel and not within each training fold, which leaks response structure into the folds but leaks the same structure into both arms. Residualisation is run at two strengths, bounding the question from either side (Table S4). *Structural* residualisation uses the 24 VRI stand attributes (crown closure, height, age, leading species, BEC zone and land-cover class), which come from the provincial forest inventory and not from the imagery CDEI is built from, and removes 14.9% of CDEI's variance. *Greenness* residualisation adds stand-mean NDVI and removes 73.2%.

Against the structural residual the point estimate grows rather than weakening: the paired $\Delta\mathrm{RMSE}$ rises from $+0.00093$ to $+0.00113$ at 25 km and from $+0.00095$ to $+0.00146$ at 12 km. The intervals are another matter. Neither arm at 25 km excludes zero, and only the 12 km structural arm does, by $+0.00009$. The argument this table can carry is therefore directional rather than statistical. Whatever the fine-scale climate gradient is failing to predict, it is not failing *because* the response is a forest-inventory artifact: strip the inventory out and the coarse model's edge grows rather than dissolving,

Table S4. The A-vs-B comparison against the original and residualised response. Removing canopy structure measured independently of the imagery leaves the point estimate intact and slightly raises it. Removing stand greenness inverts it, but does so in a regime where neither model has skill; see the text for why that inversion is not the counter-example it appears to be.

Cell	Response	R^2_A	R^2_B	Paired $\Delta\mathrm{RMSE}$ (95% CI)
25 km	CDEI (published)	−0.138	+0.003	+0.00093 [−0.00061, +0.00257]
25 km	less VRI structure	−0.176	−0.020	+0.00113 [−0.00028, +0.00266]
25 km	less structure and NDVI	−0.087	−0.229	−0.00064 [−0.00138, +0.00002]
12 km	CDEI (published)	−0.138	−0.003	+0.00095 [−0.00070, +0.00270]
12 km	less VRI structure	−0.176	+0.005	+0.00146 [+0.00009, +0.00283]
12 km	less structure and NDVI	−0.087	−0.179	−0.00046 [−0.00083, −0.00007]

which is the opposite of what a canopy-structure artifact would do. That answers the confound without establishing an effect.

The greenness residual inverts the sign, and at 12 km its interval excludes zero in favour of Model A. What stops it being a counter-example is that NDVI is not a confounder of CDEI in the way crown closure is; it is a *constitutive input*. CDEI is perpendicular distance in the NDVI–SWCI plane (main text, Materials and Methods), so regressing stand-mean NDVI out of it removes one of the two axes the index is defined on, and what remains is no longer a quantity that can be interpreted as water stress.

Both models are also worse than the mean on that response ($R^2_A = -0.087$, $R^2_B = -0.229$ at 25 km; -0.087 and -0.179 at 12 km), so the skill precondition stated in the main text fails there. That is offered as context and not as the reason, because on its own it does not discriminate: the 8 km and 12 km rows of Table S1 fail the same gate on the published response, and they are reported and counted towards the direction. Only the constitutive-input argument separates this row from those. In other words, the result survives de-confounding against independently measured canopy structure, and no test can separate CDEI from greenness altogether, because greenness is half of how CDEI is built. That second half is a limitation of the index (main text, Limitations), not a finding about climate resolution.

S2 The fold scheme: spatial blocks against random folds

Every sensitivity above varies the cell size, the seed, the learner or the response while holding the evaluation design fixed. This one varies the design. Random folds assign whole units to $k = 5$ groups at random instead of by k-means over coordinates; grouping, k, the number of repeats, the learner, the seeds and the paired repeat-level bootstrap are unchanged, so spatiality of the partition is the only difference. The spatial arm reproduces the published Surrey figures exactly (RMSE 0.01393, $R^2 = -0.042$, paired $\Delta\mathrm{RMSE} = -0.00004$), which fixes the harness.

Table S5. Paired ΔRMSE under the two fold schemes, seven seeds. Negative favours Model A. Under blocking the transect bands sit entirely above zero and under random folds entirely below it; the two do not overlap.

	Spatial blocks		Random folds	
Seed	**Transect**	**Surrey**	**Transect**	**Surrey**
26910	+0.00093	−0.00004	−0.00053	−0.00040
1	+0.00091	−0.00002	−0.00052	−0.00061
7	+0.00088	−0.00007	−0.00053	−0.00051
42	+0.00085	−0.00008	−0.00054	−0.00049
101	+0.00089	−0.00004	−0.00073	−0.00031
2024	+0.00090	−0.00008	−0.00040	−0.00055
555	+0.00086	+0.00003	−0.00051	−0.00059

On the transect the paired MAE interval spans zero under blocking and excludes zero in favour of Model A in seven seeds of seven under random folds.

The asymmetry has a structural cause. Model B's predictors are cell means, constant within a 25 km cell, so Model B cannot separate neighbouring stands at all and gains nothing when a stand's neighbours enter the training set (R^2 + 0.003 → −0.047); Model A retains the within-cell detail and gains (R^2 − 0.138 → −0.000). A model given only the projected coordinates is far from predictive under either scheme ($R^2 = -0.599$ blocked, -0.205 random), so this is not position predicting the response. It is consistent with fine predictor structure acting as a key to neighbours already seen, which is the inflation Ploton et al. [2] document, and which by construction only the fine-resolution arm can exploit. That reading is consistent with the evidence rather than established by it.

S3 The block count, and what F3's failure is a property of

The pre-registered F3 criterion (detection ≥ 0.80) fails in the main sweep. That sweep, and the sample-size sweep of Section S14, both hold the validation design fixed at $k = 5$ blocks. Since the paired interval is a bootstrap over the blocks of one repeat, raising the site count from 300 to 2,400 adds rows but not resampling units, which is a candidate explanation for detection saturating in n that f cannot supply.

This sweeps k instead, at the two highest-f levels, with the field, the site count, the learner, the cell size and the paired bootstrap unchanged (`scripts/k_sweep.py`, 50 panels per cell rather than the main sweep's 100, so the intervals are wider).

Table S6. Detection rate against block count at fixed f and fixed $n = 300$. Wilson 95% intervals on 50 panels per cell. At the highest f the block count moves detection substantially and the bar is reached; one level down it does nothing.

	$k = 5$	$k = 10$	$k = 20$
$f = 0.81$	0.68 [0.54, 0.79]	0.78 [0.65, 0.87]	**0.84** [0.71, 0.92]
$f = 0.71$	0.58 [0.44, 0.71]	0.56 [0.42, 0.69]	0.60 [0.46, 0.72]

Two things follow, and they cut against each other.

At $f = 0.81$ detection rises from 0.68 to 0.84 as k goes from 5 to 20, so the point estimate crosses the F3 bar that the main sweep reports as unreachable. The main text's $k = 5$ arm reproduces the main sweep at this level (0.68 [0.54, 0.79] here against 0.63 [0.53, 0.72] on 100 panels), so the difference is the block count and not the harness. **F3's failure is therefore not a property of f.** It

Table S8. Surrey: no candidate response variable is predictable from climate at any resolution.

Target	sd	Model A R^2	Model B R^2
CDEI	0.0138	−0.042	−0.051
`dry_dist`	0.0137	−0.043	−0.053
`ndvi_mean`	0.1216	−0.232	−0.154
`swci_mean`	0.0567	−0.280	−0.223
`lst_mean`	3.0962	−0.655	−0.578

is a property of f evaluated on five blocks. Because that cell carries the claim, it was re-run at the main sweep's 100 panels: the estimate is unchanged at 0.84 and the interval tightens to [0.76, 0.90] (`data/processed/k_sweep_confirm.csv`). It still spans 0.80, so the bar is reached and not cleared.

At $f = 0.71$, one level down, the block count does nothing at all: 0.58, 0.56, 0.60 across a fourfold increase in k. So a larger k does not rescue a comparison whose predictors were not sufficiently separated by the coarsening.

The honest summary is two-dimensional. Detectability turns on the fraction of predictor spatial variance the coarsening removes *and* on the number of units the validation interval is built from, and neither is sufficient alone. A study reporting f without its block count has reported half of what decides whether its comparison could have answered the question. Both caveats stated with the main sweep still apply: the interval at $k = 20$ spans 0.80 rather than excluding it, so the bar is reached and not cleared, and these remain rates against a constructed effect size.

S4 Surrey: the Extent 1 model comparison in full

Table S7. Surrey: Models A and B are statistically indistinguishable on the full 612-row panel. Paired differences are A − B over identical folds, with a 10,000-sample bootstrap interval.

Model	RMSE	MAE	R^2
A (scale-free)	0.01393	0.01057	−0.042
B (coarse 4 km)	0.01397	0.01039	−0.051
Paired difference	**Estimate**	**95% CI**	**A better**
ΔRMSE	−0.00004	[−0.00028, +0.00018]	48% of folds
ΔMAE	+0.00018	[−0.00017, +0.00045]	28% of folds

S5 Surrey: skill on every candidate response variable

Precondition 1 fails: no model has skill on any target. Every candidate response variable sits at or below the no-skill line under blocked CV, which rules out "the wrong index was chosen" as an explanation (Table S8).

S6 The between-year variance estimates

The between-year variance estimates rest on four summers. The orthogonality diagnosis of the main text's Results sets each variable's between-year standard deviation against its between-polygon

one and reports ratios up to 10.2× for the predictors and 9.4× the other way for the responses (main text). Those two terms are not estimated with equal confidence. The between-polygon term rests on 153 polygons; the between-year term rests on *four* summers, and those four are not a mild draw: their summer climatic moisture deficits are 194, 244, 111 and 73 mm, a spread of more than threefold. A four-summer window containing less interannual range would yield a smaller between-year standard deviation and therefore compress every ratio in the table. The figures are descriptive of 2022–2025 and not as properties of the ClimateBC field in general, and the same caveat applies to the response ratios, whose denominators come from the same four summers. One defence offered elsewhere does not apply here and should not be borrowed: that defence tests the *spatial* term only. What limits the damage is that the diagnosis turns on the direction of the inequality rather than its size. Spatially-blocked CV asks the model to rank corridors along the axis on which climate varies least, and that holds for any ratio comfortably above 1. The directional claim survives; the specific multipliers are a property of this record. Extending the climate panel past four summers would settle them, and is the same fix as item 4.

S7 Drivers of between-stand water stress: full comparison

The method for this analysis is in Section S12. The result quoted in the main text is the first line of the comparison below.

Table S9. Predictor families scored under the same blocked CV. Terrain + stand structure scores highest; stand structure alone and the all-in family also clear zero, and adding climate on top of terrain-plus-structure makes it *worse*.

Predictor family	**features**	**CV R^2**	**Folds with $R^2 > 0$**
climate (baseline)	14	−0.138	40%
terrain (slope/aspect/ruggedness/elevation)	5	−0.085	44%
stand structure (age/height/crown/species)	24	+0.004	68%
terrain + structure	**29**	+0.029	**80%**
terrain + structure + climate	43	+0.020	60%

Terrain + stand structure is the highest-scoring family (positive in 80% of folds; $R^2 = +0.044$ on pure stand-means). Adding climate on top makes it worse (+0.029 → +0.020), independent confirmation that the climate gradient is noise for this response. Moreover, by permutation importance on a held-out spatial block, the leading drivers are all local: leading species (bigleaf maple), stand area, aspect (eastness, then northness), and broadleaf class. No climate variable appears.

That the local variables ranking highest are compositional ones is what raises the vegetation-type reading noted in the main text. Broadleaf and conifer stands occupy different regions of the NDVI–SWCI–LST space in which CDEI is constructed, so a leading-species split can order stands by canopy type rather than by water stress. Table S4 bounds how much of the A-versus-B contrast that could account for.

Random forests are often described as robust to uninformative features, so a drop from +0.029 to +0.020 on adding 14 of them looks like more than indifference. The robustness intuition does not hold for this configuration. With `max_features="sqrt"` the forest samples only a subset of features as split candidates at each node, so adding 14 climate columns to 29 terrain-and-structure columns means roughly a third of the candidates at every node are now draws from an uninformative set, and each node has a correspondingly smaller chance of seeing the informative feature it needs. Feature dilution under `sqrt` sampling is the expected mechanism, and no appeal to collinearity between climate and structure is required to explain a degradation of this size. The consequence for the interpretation is

mild: the drop confirms the climate block is uninformative, but its magnitude reflects the learner's sampling rule as well as the data.

Part 2. Methods and materials detail

S8 Study extent: sampling detail

Extent 1. The corridor polygons are the layer "Green Infrastructure Network Corridors" (layer id 1) of the City of Surrey open-data ArcGIS MapServer [3]. Their mean area is 9.1 ha against a median of 5.8 ha. The terrain is low-relief but not flat: rolling till uplands (the Newton and Sunnyside uplands, Panorama Ridge) separated by the lowland valleys of the Serpentine and Nicomekl rivers.

The 153 polygons carry 144 distinct Green Infrastructure Network corridor ids, because seven corridors are digitised in several pieces. The polygon is the modelling unit throughout the main text, so every Surrey n reported there is $n = 153$; the corridor is the reporting unit of the applied product [1]and is not used in the analyses reported here.

Extent 2. Stands were retrieved from the BC Data Catalogue WFS (`openmaps.gov.bc.ca/geo/pub/ows`) over a bounding box of $(-123.15, 49.00, -121.40, 49.20)$ in WGS84. The 49.20°N cap is deliberate: without it the box reaches the North Shore mountains and the sample stops being a south-of-Fraser transect. Forest stands were used rather than corridor polygons because a Surrey corridor's median effective width spans barely one or two 20–30 m pixels (main text, Limitations).

Of 12,168 candidate treed stands, 300 were sampled, stratified by BEC (Biogeoclimatic Ecosystem Classification) zone [4] × easting quintile with allocation proportional to $\sqrt{n}$ rather than to stratum size n: BEC zone proxies elevation before any DEM call, easting quintiles cover the transect end to end, and $\sqrt{n}$ allocation stops the abundant lowland strata from swamping the sparse montane strata that carry the gradient. Protected-area polygons were evaluated as analysis units and rejected (11–19 polygons spanning 3–61,594 ha, too few and too heterogeneous); they were retained only as a validation subset.

S9 Response variable: CDEI construction

Summer (1 June–31 August) cloud-masked composites are built for each year 2022–2025: NDVI and a shortwave canopy-water index SWCI = (B11 − B12)/(B11 + B12) from Sentinel-2, and land-surface temperature from Landsat 8/9. SWCI is a SWIR–SWIR ratio and is therefore *not* Gao's NIR–SWIR NDWI [5], which it is easily mistaken for. Rasters are reprojected to EPSG:26910 and summarised zonally per polygon. In the NDVI–SWCI plane the polygons occupy a wedge whose lower envelope is the driest state observed at each greenness level. That construction is an analogue of the surface-temperature/vegetation-index feature space rather than an instance of it, and the difference is worth stating because the lineage is sometimes cited as though it transferred directly. In that space the dry edge is the *upper envelope*, the maximum surface temperature at a given greenness [6]; Moran et al. [7] give the trapezoid form of it against surface–air temperature and a soil-adjusted vegetation index, and Carlson and Ripley [8] supply the NDVI-to-fractional-cover relation the greenness axis rests on rather than any dry edge. Substituting a *shortwave water index* for surface temperature, and taking the lower envelope instead of the upper, is this paper's own step and is not established by those sources. A *dry edge* is fitted once per panel, as a line through the 5th percentile of SWCI within each of 12 equal-frequency NDVI bins. A polygon's `dry_dist` is its perpendicular distance above that edge; CDEI divides this by RLST, the polygon's land-surface temperature relative to its own multi-summer mean. A low CDEI places a polygon close to the driest state observed for its greenness.

Table S10. The seven seasonal summer climate variables forming the core of the predictor set.

Variable	Meaning
`Tmax_sm`	mean summer maximum temperature
`Tmin_sm`	mean summer minimum temperature
`PPT_sm`	summer precipitation
`Rad_sm`	summer solar radiation
`Eref_sm`	summer reference evaporation
`CMD_sm`	summer climatic moisture deficit
`DD18_sm`	summer degree-days above 18 °C

S10 Climate predictors: variable table and acquisition

The remaining seven columns of the predictor set are anomalies against each site's own 1961–1990 normal for `Tmax_sm`, `Tmin_sm`, `PPT_sm`, `Eref_sm` and `CMD_sm`; $\log(1 + \texttt{PPT\_sm})$; and the Mahalanobis climate-novelty scalar.

The novelty term is scaled by the covariance *of the departures from normal*, not by the spatial covariance of the normals themselves. Over a low-relief 30 km area the normals barely vary between corridors, so their spatial covariance is near-singular, and scaling by it would rank *unusual locations* rather than *unusual climate*. As defined, the scalar has a median of 1.87 (95th percentile 3.21).

Acquisition. The endpoint used was `LatLonEl` on the host `api6.climatebc.ca`; elevation was attached from the Copernicus GLO-30 DEM [9] via a Planetary Computer STAC search. The ClimateBC web API enforces an undocumented limit of roughly 50 calls per hour per IP with no batch endpoint, so acquisition was paced, checkpointed and resumable against a permanent on-disk cache. Both extents completed in full, 765 of 765 calls for Surrey and 1,500 of 1,500 for the transect, giving a complete 5-period panel for all 153 corridor polygons and all 300 stands.

S11 Experiment implementation detail

The coarse grid is not floating: cells are anchored to the analysis CRS origin. A polygon's cell index is $\lfloor x/c \rfloor, \lfloor y/c \rfloor$ on its EPSG:26910 easting and northing for cell size c, so cell boundaries fall at exact integer multiples of c in UTM zone 10N metres and the partition is reproducible from the coordinates alone. This matters because a grid of the same cell size shifted by half a cell would group different polygons together and yield different coarse means; anchoring removes that degree of freedom rather than leaving it to chance. Averaging is over the polygons falling in a cell in a given year, so it is area-blind by design, which is what a coarse grid cell actually delivers.

Forward temporal holdout. A second, stricter test was specified, training on the earlier summers and testing on the most recent so the model has never seen the test *year*. It was not completed, and no result from it is reported anywhere in this paper.

S12 Explanatory analysis: method

A follow-up analysis, reported in the main text, asked what *does* organise between-stand water stress, scored under the same blocked CV. Three predictor families were compared, and their skill is reported in the main text. *Climate* is the 14 variables of the main text's climate predictor set. *Terrain* is 5

variables derived from the Copernicus GLO-30 DEM [9]: elevation, slope, ruggedness, and aspect decomposed into northness and eastness. *Stand structure* is 24 variables from the VRI: PROJ_AGE_1, PROJ_HEIGHT_1, CROWN_CLOSURE, SPECIES_PCT_1 and stand area, plus one-hot encodings of leading species, BEC zone and BCLCS land-cover class. The encoding retains *every* level observed in the 300 sampled stands, each as its own indicator column: no level is dropped as a reference category, and rare levels are neither pooled into an "other" class nor thresholded on a minimum count. That fixes the family at 24 variables, 5 numeric plus 19 indicators. Anyone reproducing this should expect the count to change if a different stand sample yields a different set of leading species.

NDVI, SWCI and LST are excluded as predictors here: CDEI is built from them, so including them would be circular. Aspect is decomposed into northness and eastness because aspect in degrees is circular and cannot be used as a raw numeric feature. Relative importance is assessed by permutation importance (20 repeats) on a held-out spatial block, not on training data.

On comparing families of unequal size. The families contain 5, 14 and 24 features, so the largest could win simply by having the most ways to split. Families are ranked by *held-out* R^2 under the same blocked CV, not by any in-sample importance measure, and held-out skill does not reward extra features: an uninformative predictor costs generalisation instead of buying it, so feature count alone cannot manufacture a ranking. The permutation ranking of *individual* features is the weaker instrument, because leading species, BEC zone and land-cover class enter as one-hot dummies, so a categorical variable's contribution is divided across its levels while a continuous variable's is not. Individual importances are comparable within an encoding but not across encodings, and are read here only as an indication of which *kinds* of predictor matter.

CDEI is used here as an unvalidated index, and no conclusion in this paper rests on a check reported elsewhere: the resolution comparison asks only whether two predictor sets differ against the same response, which does not require that response to be calibrated. Two independent checks on the index are carried out in a companion paper [1], and the limits that remain are in the main text's Limitations. The software environment and seeding are under Data and code availability.

S13 Precondition simulation: full specification

Fields are generated on a 512×512 lattice over the 100 km square by Gaussian-smoothing white noise, the filter width setting the correlation length L, and sampled at 300 site locations where they are standardised. The eight sweep levels are $L = 40, 28, 20, 14, 10, 7, 5$ and 3.5 km against a fixed 25 km cell. Weights are $\beta = \gamma = \sigma_\varepsilon = 1$ with the climate composite standardised, the local driver has $L = 3$ km, and summers carry a common offset in the predictors but no term in the response, so the four summers are replicate reads of one spatial truth.

Two design faults were found in calibration and are recorded here so they are not reintroduced. Correlation lengths at or above the domain width drove the smoother to a constant, and renormalising then amplified floating-point noise into a spurious field, which made f *non-monotone* in L; L is therefore capped well inside the domain and the field reflected rather than wrapped, so it is not forced periodic. And redrawing the field independently each summer made the response's between-summer variance dominate its between-site variance, so both models predicted the summer level almost perfectly (CV $R^2 \approx 1.00$) and the spatial contrast, the only thing the precondition concerns, was swamped.

The three criteria were fixed in the script's docstring before the run: that detection rise with f at Spearman $\rho \geq +0.8$; that a low-f regime exist with detection ≤ 0.25, so a real effect is genuinely unrecoverable there; and that a high-f regime exist with detection ≥ 0.80. Observed over one hundred panels per level: $+0.833$, 0.07 and 0.63. The first two are met, the second decisively. **The third is not met.** Detection never reaches 0.80 at any level of f tested; its maximum is 0.63 $[0.53, 0.72]$ where four fifths of the predictors' spatial variance has been destroyed. The interval excludes 0.80, so this

is a property of the design rather than a sampling artifact: an earlier run at twenty panels per level returned 0.65 $[0.43,\ 0.82]$, which could not be separated from the criterion, and raising the panel count fivefold resolved the question against it.

The consequence is stated in the main text and repeated here because it is the most important limitation of the proposed diagnostic. The precondition is one-sided. It identifies extents at which a resolution comparison cannot answer the question put to it, which is what F2 establishes and what the Surrey case needed. It does not identify extents at which such a comparison can be trusted, because within this design no such extent exists. Had detection been flat in f, the diagnostic would not have been doing the work claimed for it and belonged out of the paper rather than in it weakly; had detection been high at $f \approx 0.12$, Surrey's null would have been informative after all and the paper's framing wrong. Both of those outcomes were avoided; the failure of F3 is a third outcome, and it bounds what may be recommended rather than what may be concluded about Surrey.

S14 Sample size and the failure of F3

F3 was evaluated at 300 sites, matching Extent 2. A detection rate is a power calculation and power grows with the sample, so the failure could have been a property of that one panel size rather than of the diagnostic. If it were, the recommendation would have to change: a floor that only binds small studies is a different claim from one that binds every study.

The four highest levels of L were therefore crossed with 300, 600, 1,200 and 2,400 sites, one hundred panels per cell, everything else unchanged and called from the same functions. β is fixed here as it is in the main sweep, so the generative truth does not move and the climate share of response variance stays at 0.334 ± 0.024 throughout. Only the site count and f vary.

Table S11 gives the result. No cell reaches 0.80. The highest is 0.78 $[0.69,\ 0.85]$ at 600 sites, and detection then stops responding: 0.78, 0.77 and 0.75 across a fourfold increase from there. The one gain is from 300 to 600 sites at high f, 0.63 against 0.78 with intervals that barely overlap. Beyond that the sample is not what limits the comparison.

Below the floor it makes no difference whatever. At $f \approx 0.40$ detection is 0.20, 0.20, 0.20 and 0.22 over an eightfold range of n. A study whose coarsening destroys little of its predictors' spatial variance therefore cannot recover the comparison by measuring more sites, and that is worth more to a reader than the floor on its own: it says which remedy will not work.

Two bounds on the above. At $n \geq 600$ the intervals reach 0.85, 0.84 and 0.82, so each includes 0.80; no configuration reached the bar, and the bar is not shown to be out of reach. And these are rates against a constructed effect size, as in the main sweep, so they bound this design rather than any particular study.

Table S11. Detection rate against site count at the four highest levels of f, with the generative advantage held constant across every cell. One hundred synthetic panels per cell; a detection is a paired ΔRMSE interval excluding zero in the generative direction, reported with a 95% Wilson interval.

Sites	f **removed**	**Detection rate**	**95% CI**	**Mean** ΔRMSE
300	0.388	0.20	[0.13, 0.29]	−0.0406
300	0.561	0.28	[0.20, 0.37]	−0.0716
300	0.700	0.54	[0.44, 0.63]	−0.1006
300	0.805	0.63	[0.53, 0.72]	−0.1160
600	0.397	0.20	[0.13, 0.29]	−0.0461
600	0.574	0.38	[0.29, 0.48]	−0.0827
600	0.716	0.59	[0.49, 0.68]	−0.1124
600	0.824	0.78	[0.69, 0.85]	−0.1352
1,200	0.403	0.20	[0.13, 0.29]	−0.0417
1,200	0.584	0.31	[0.23, 0.41]	−0.0880
1,200	0.728	0.52	[0.42, 0.62]	−0.1180
1,200	0.836	0.77	[0.68, 0.84]	−0.1406
2,400	0.402	0.22	[0.15, 0.31]	−0.0377
2,400	0.584	0.36	[0.27, 0.46]	−0.0810
2,400	0.730	0.50	[0.40, 0.60]	−0.1148
2,400	0.840	0.75	[0.66, 0.82]	−0.1416

S15 The geometry sweep: what the calibration is conditional on

The section below transfers a rate to one other geometry and finds the transfer depends on an assumption. This sweep asks the more general question directly, by varying the domain and the coarse cell together over five geometries and sweeping the correlation length in units of the cell so that f stays comparable across them. Twenty synthetic panels per level, eight levels per geometry, everything else the main simulation's own machinery called unmodified. The local driver is held at 3 km in *metres* rather than rescaled with the cell, which is the opposite choice to the one made below, and the control pair is what tests whether that choice matters.

Table S12. The five geometries. `transect` reproduces the main sweep's design exactly. `mid` is the control: it shares `transect`'s domain-to-cell ratio at 0.6× the absolute size, and returns an identical f at every level.

Geometry	Domain (km)	Cell (km)	Ratio	f range	Spearman ρ
transect	100	25	4.0	0.059–0.815	+0.843
mid	60	15	4.0	0.059–0.815	+0.892
surrey	30	4	7.5	0.033–0.645	+0.988
wide	200	25	8.0	0.031–0.657	+0.862
fine-cell	100	10	10.0	0.025–0.558	+0.964

Detection rises with f inside all five, so every geometry is a calibration curve on its own terms. Across them at matched f it does not agree, and the disagreement is ordered by how many coarse cells the domain contains rather than by its size.

Table S13. Detection near $f = 0.42$, ordered by domain-to-cell ratio. The extreme pair's intervals do not overlap. R^2 is the mean over 20 panels; Model A turns positive only where the cells are many.

Geometry	Ratio	f	Detection	95% CI	R^2 A	R^2 B
mid	4.0	0.410	0.20	[0.08, 0.42]	−0.185	−0.218
transect	4.0	0.410	0.20	[0.08, 0.42]	−0.149	−0.197
surrey	7.5	0.414	0.40	[0.22, 0.61]	−0.084	−0.168
wide	8.0	0.437	0.60	[0.39, 0.78]	+0.036	−0.037
fine-cell	10.0	0.369	0.45	[0.26, 0.66]	+0.032	−0.022
fine-cell	10.0	0.478	0.75	[0.53, 0.89]	+0.058	−0.018

The control pair settles what the sweep would otherwise leave open. `transect` and `mid` share a ratio and differ in absolute size by 1.7×, and they place the local driver at 0.12 and 0.20 of a cell respectively, so the fixed 3 km driver is not held constant relative to the geometry. Their f is identical at all eight levels, to four decimals, and their detection agrees within its interval at all eight, differing at four of them by one or two panels in twenty (0.70 against 0.65, 0.55 against 0.65, 0.00 against 0.10, 0.00 against 0.05), every pair of Wilson intervals overlapping. That is agreement at the resolution twenty panels per level can supply rather than exact agreement, and it rules out both absolute extent and the driver's cell-relative scale as the cause, leaving the cell count.

Three criteria were fixed before the run. Detection rising with f within each geometry holds in all five. A low-f floor at or below 0.25 holds in four, and breaches in `fine-cell` at one level on 6 successes out of 20 against a bar of 5, whose Wilson interval [0.15, 0.52] contains the bar, so it is not read as a failure. Agreement at matched f fails, which is the result above. Analysis code: `scripts/geometry_sweep.py`.

S16 Transferring a detection rate to another geometry

The main sweep is run at one geometry: a 100 km square carrying 300 sites at a 25 km cell, matched to Extent 2. The main text reads a rate off it for Extent 1 as well, and Extent 1 is 153 polygons over 408 km^2 at a 4 km cell. This section reports what happens when the sweep is re-run at Surrey's own geometry, because the transfer turns out to depend on an assumption the simulation has to make and the main sweep never had to state.

The assumption. The local driver standing in for terrain and stand structure has a fixed correlation length of 3 km, chosen so that it survives coarsening: against the 25 km cell it leaves 84% of its variance within cells. Against a 4 km cell the same 3 km field leaves only 9% within cells, which turns the driver into a coarse field and inverts the role it was given. Preserving the design's own ratio gives $3000 \times 4/25 = 480$ m.

Surrey's geometry. A 20 km square carries 400 km^2 and exactly 25 occupied cells at 4 km, matching the main text's Surrey cell-size table; $n = 153$. One hundred panels per level, everything else unchanged.

Table S14. Detection at Surrey's geometry (20 km square, 4 km cell, $n = 153$, local driver 480 m). Surrey measures $f = 0.123$.

f removed	Detection rate	95% CI
0.079	0.07	[0.03, 0.14]
0.138	0.16	[0.10, 0.24]
0.240	0.26	[0.18, 0.35]
0.331	0.32	[0.24, 0.42]
0.532	0.48	[0.38, 0.58]

All three pre-registered criteria return the same verdicts here. F1 holds more cleanly than in the main sweep (Spearman $\rho = +1.000$ against $+0.833$, without the non-monotone low-f region), F2 holds with a minimum of 0.07, and F3 fails, its maximum reaching 0.48.

What the assumption is worth. Holding the geometry and f fixed and varying only the local driver's correlation length:

Table S15. Sensitivity of the detection rate to the local driver's correlation length, at $f = 0.138$ and Surrey's geometry, against the same test at Extent 2's geometry and $f = 0.468$. One hundred panels per cell.

Geometry	Local driver	Within-cell share	Detection rate
Surrey, 4 km cell	480 m	0.55	0.16 [0.10, 0.24]
Surrey, 4 km cell	3000 m	0.09	0.03 [0.01, 0.08]
Extent 2, 25 km cell	1000 m	0.95	0.34 [0.25, 0.44]
Extent 2, 25 km cell	3000 m	0.86	0.24 [0.17, 0.33]
Extent 2, 25 km cell	6000 m	0.64	0.19 [0.13, 0.28]

At Surrey's geometry the two arms are 5.3× apart and their intervals do not overlap, so no single rate is identified there without fixing that scale. At Extent 2's geometry the spread across a wider range of the same parameter is 1.8× with every interval overlapping, and the published 3 km setting returns 0.24, which is what the main text's detection table interpolates at the transect's $f = 0.484$. The asymmetry has a cause rather than being noise: at a 25 km cell every scale tested keeps the driver majority within-cell, so its role holds throughout, while at a 4 km cell the unscaled driver stops being sub-cell at all.

Two consequences. The transect figure quoted in the main text is stable across this assumption. The Surrey figure is not, which is why the main text gives a range there, and why a detection rate should be reported against a stated geometry rather than against f alone.

References

[1] Joshua Wang. An exploratory satellite water-stress monitor for Surrey's green infrastructure corridors: construction, independent validation, and the limits of both. Manuscript in preparation, 2026.

[2] Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11(1):4540, 2020. doi: 10.1038/s41467-020-18321-y.

[3] City of Surrey. Green Infrastructure Network Corridors (OpenData MapServer, layer 1). City of Surrey Open Data, ArcGIS MapServer, 2026. URL https://gisservices.surrey.ca/arcgis/rest/services/OpenData/MapServer. Accessed 21 July 2026. Catalogue: https://opendata-surrey.hub.arcgis.com/. The corridor attributes are those published in Appendix J of the City's Biodiversity Conservation Strategy.

[4] British Columbia Ministry of Forests. Biogeoclimatic ecosystem classification (BEC) zone/subzone/variant/phase map. BC Data Catalogue, 2024. https://catalogue.data.gov.bc.ca/dataset/bec-map.

[5] Bo-Cai Gao. NDWI — a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3):257–266, 1996. doi: 10.1016/S0034-4257(96)00067-3.

[6] Inge Sandholt, Kjeld Rasmussen, and Jens Andersen. A simple interpretation of the surface temperature/vegetation index space for assessment of surface moisture status. *Remote Sensing of Environment*, 79(2–3):213–224, 2002. doi: 10.1016/S0034-4257(01)00274-7.

[7] M. S. Moran, T. R. Clarke, Y. Inoue, and A. Vidal. Estimating crop water deficit using the relation between surface-air temperature and spectral vegetation index. *Remote Sensing of Environment*, 49(3):246–263, 1994. doi: 10.1016/0034-4257(94)90020-5.

[8] Toby N. Carlson and David A. Ripley. On the relation between NDVI, fractional vegetation cover, and leaf area index. *Remote Sensing of Environment*, 62(3):241–252, 1997. doi: 10.1016/S0034-4257(97)00104-1.

[9] European Space Agency and Airbus. Copernicus DEM: Global and european digital elevation model, GLO-30. Copernicus Space Component Data Access, 2022. https://spacedata.copernicus.eu/collections/copernicus-digital-elevation-model.