

Measurement noise limits soil carbon forecasting and crediting across European monitoring points

Md Rakib Hasan

`rakibhridoy@fermium.systems`

University of Dhaka

Mst Anika Khatun Rupa

University of Dhaka

A. S. M. Mohiuddin

University of Dhaka

Article

Keywords:

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11493951/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Measurement noise limits soil carbon forecasting and crediting across European monitoring points

Md Rakib Hasan^{1,2,*}, Mst Anika Khatun Rupa¹ and A. S. M. Mohiuddin¹

¹Department of Soil, Water and Environment, University of Dhaka, Dhaka 1000, Bangladesh

²Fermium Systems, Dhaka 1207, Bangladesh

*Corresponding author: rakibhridoy@fermium.systems

Abstract

Soil organic carbon is now measured to monitor soil health under the European Soil Monitoring Law and to pay for its gains through EU carbon farming certification. Here, three surveys of 8,275 European points show that measurement noise makes up at least 94 % of the measured six-year change. The noise in one remeasured sample is 15 times the six-year gain of the 4 per 1000 target, detecting that gain needs about 2,100 resampled locations, the EU methodology's 70 % confidence rule credits three in ten projects whose carbon did not change, and enrolling fields on a low baseline credits a reversion gain six times the target. A deep network and gradient boosting forecast change with the same skill, 0.11, below a noise ceiling of 0.28. Crediting should therefore rest on measurement error estimated from repeated samples, a mean-reversion benchmark and enrolment rules that ignore single low readings.

1 Introduction

The carbon held in soils exceeds that of the atmosphere and terrestrial vegetation together, and its change under management is still one of the hardest terms of the terrestrial carbon budget to predict¹⁻³. Decomposition kinetics, aggregate protection and mineral association are characterised in detail⁴⁻⁶, so the obstacle lies less in process understanding than in the size of the signal. Annual change in soil organic carbon (SOC) is usually a fraction of a percent of the standing stock, with resurveyed Chinese croplands averaging 0.22 % per year⁷, while measured concentration can differ by more than that between sampling points a few metres apart⁸. Detecting change therefore requires dense resampling or intervals of a decade or more^{3;8}. On the European LUCAS survey, predicted change at single sites rarely exceeds its own uncertainty and becomes detectable only when averaged over larger areas⁹, as space-time models of Hungarian monitoring data also find¹⁰. Measurement, reporting and verification frameworks for mitigation targets such as the four-per-mille initiative depend on models to interpolate between sparse resamplings¹¹⁻¹⁴.

The question has also become regulatory. The European Soil Monitoring Law, Directive (EU) 2025/2360, lists soil organic carbon among the descriptors of soil health in its Annex I, sets a six-year cycle of soil measurements, and points to the LUCAS soil survey as a basis for that monitoring¹⁵. Member states will therefore receive a carbon value at each sampled point every six years, and models will be asked to fill the space and time between them. How much of the change measured at a point over six years is real, and how much a model could forecast, are questions the directive does not answer and that bear directly on how its measurements can be used. The same measurements are also becoming the basis of payment, since Regulation (EU) 2024/3012 establishes voluntary certification of carbon farming and its methodologies credit a gain in soil organic carbon measured by resampling fixed locations or verified against such

resampling^{16;17}, so the error of a remeasured value decides how much of a credited gain is real.

Two modelling traditions have addressed the gap, and their strengths and limitations are complementary. Process models such as RothC and DayCent carry the mechanism explicitly but need a dated management history and calibration data that exist at a handful of long-term stations^{18;19}, whereas machine learning learns from whatever covariates exist without any guarantee that a fitted relationship holds beyond the conditions that produced it^{20;21}. Digital soil mapping has largely converged on random forests and gradient boosting^{22–25}, which predict spatial pattern well and say much less about change, and a national space-time application in Argentina found prediction uncertainty to be substantial²⁶. Estimates of change itself have come from a quantile additive model fitted to the revisited LUCAS points²⁷ and from a spatiotemporal model of bare-soil reflectance on German monitoring sites²⁸, and most implementations treat each observation as an independent row, so a covariate series enters as summary statistics and its ordering is discarded.

Deep learning entered this space with a specific promise, which is that a network learns the representation from the data instead of requiring covariates to be engineered in advance, and that matters where the informative structure lies in the shape of a signal^{29;30}. The promise is settled for soil spectroscopy, where convolutional networks have outperformed partial least squares on the largest spectral libraries^{31;32}, and less settled for mapping, where gains over conventional learners each rest on a single country and the Chilean gain depends on spatial context that blocked validation could remove^{33–37}. Recurrent networks have outperformed calibrated conceptual models for streamflow, including in basins withheld from training^{38;39}, and have extended satellite soil moisture records^{40;41}, but their use for soil carbon remains rare. Convolutional-recurrent hybrids have outperformed XGBoost and random forests in forecasting river water quality⁴², although training and test data there came from the same four stations. For soil carbon, Zhang et al.⁴³ paired a convolutional network over terrain layers with an LSTM over ten years of MODIS phenology, and later convolutional-recurrent and dual-channel networks have mapped soil carbon and organic matter from multi-temporal imagery^{44;45}, but all of these predict concentration at one time rather than change between surveys.

Whichever method is used, the claim of skill depends on the validation design and on the reference it is measured against. Spatial and temporal autocorrelation inflate apparent skill under random cross-validation^{21;46}, severely enough that a forest biomass model which appeared to explain more than half the variation under random splitting showed quasi-null skill under spatial validation⁴⁷. In the Ganges Delta, holding out whole locations lowered the R^2 of gradient boosting for SOC from an inflated 0.91 to 0.64, deeper neural networks gave no reliable improvement at that sample size, and models fitted in one of fifteen deltas failed in almost every other⁴⁸. Repeated measurement adds a second trap, since a value measured with error tends to be followed by one closer to the mean, so that apparent change depends on the starting level even where nothing has changed^{49;50}.

This study tests a convolutional-recurrent hybrid for forecasting SOC change against the tabular standard, on 16,496 mineral-soil observations from 8,275 European points sampled three times between 2009 and 2018. Like the model of Zhang et al.⁴³, the hybrid couples a convolutional network with an LSTM, but here the convolution encodes the monthly covariate series, the recurrent decoder reads the encoded sequence, and a static embedding carries the soil properties. Three campaigns give each point two intervals, close to what a six-year monitoring cycle yields over a decade, so the test asks whether a sequence model helps at the density of measurement that monitoring will provide rather than at its best case. The study asks first whether the hybrid beats gradient boosting given the same inputs under spatially blocked validation, and then, through five ablations, which of its components carry any skill it has, since a network that matches a tree ensemble may do so through its sequence machinery or in spite of it. Three visits to each point also allow the measurement error of a single LUCAS value to be estimated. That

estimate gives a ceiling on the skill any model can reach, the share of measured change that is noise and the error that a carbon credit inherits when it is paid on a remeasured value, and it allows a test of whether change depends on the starting level once reversion towards the mean is removed. Skill throughout is one minus the ratio of a model's root-mean-square error to that of persistence, the forecast of no change, so that zero matches persistence and positive values beat it.

2 Results

2.1 Hybrid and gradient boosting compared

Pooled over five replicates, the hybrid reaches a skill of 0.113 and gradient boosting 0.110, and the difference of 0.003 has a 95 % interval of -0.011 to 0.019 (Table 1, Fig. 1). The hybrid leads in three of the five replicates and in 63 % of bootstrap draws, so the interval includes zero and under spatial blocking the hybrid does not earn its added complexity. The 90 % interval, -0.009 to 0.015 , lies inside the margin of ± 0.02 , so the two methods are equivalent at that margin, and any advantage for the hybrid is at most 0.019. The single assignment of Supplementary Table S1 gives a slightly larger gap, 0.011 with an interval of -0.006 to 0.025 . With blocks of 390 km both methods gain slightly, the hybrid to 0.118 and boosting to 0.117, and the difference of 0.001 (-0.013 to 0.022) again has a 90 % interval inside the margin, so the comparison does not depend on the block size. Withholding the previous value from boosting lowers it to 0.060, so the hybrid's lead over that variant, 0.051 (0.028 to 0.075), measures the value of the previous measurement rather than of the architecture.

Both methods clear the mean-reversion floor by a clear margin. Over the same five replicates the linear fit on the previous value and the campaign reaches 0.047. The hybrid exceeds it by 0.066 (0.039 to 0.089) and boosting by 0.063 (0.037 to 0.084), so reversion towards the mean supplies about two fifths of either model's skill and each method recovers structure beyond it.

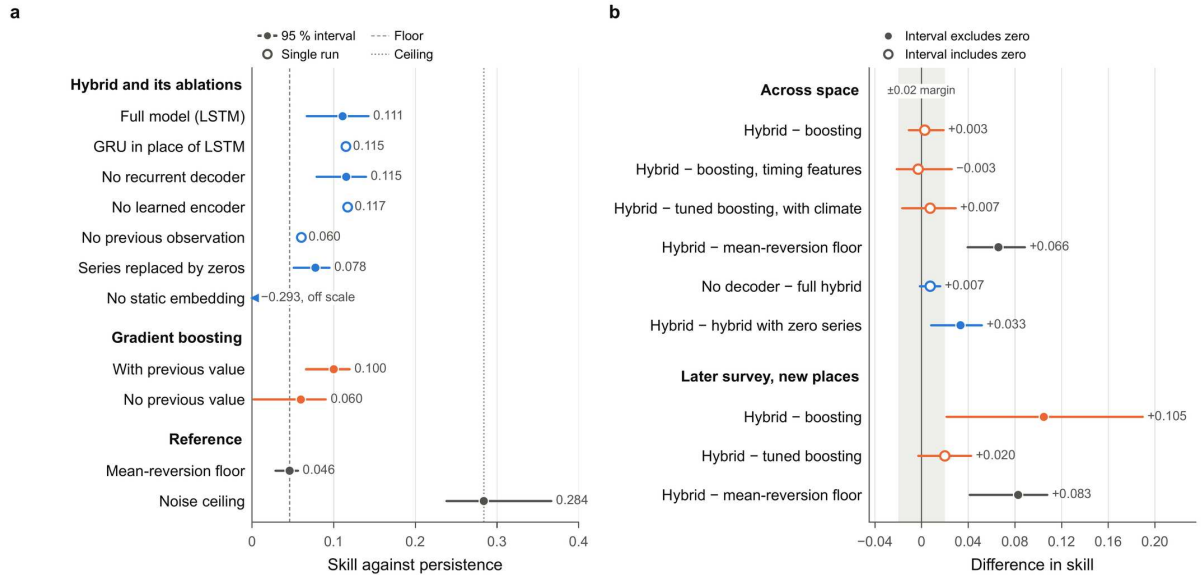


Figure 1: Skill against persistence, mineral soils, spatially blocked validation. (a) Skill of each model on the single assignment of blocks to folds; hollow markers, single runs without an interval; dashed line, mean-reversion floor; dotted line, noise ceiling. (b) Paired differences in skill across space and at a later survey at new places; filled markers, interval excludes zero; shaded band, equivalence margin of ± 0.02 . Lines, 95 % block-bootstrap intervals.

2.2 Signal, noise and the ceiling on skill

Measurement noise accounts for most of the change measured at a LUCAS point (Table 2). On the 8,221 points whose first two campaigns were mineral, the change in log carbon between 2009 and 2015 correlates with the change between 2015 and 2018 at -0.54 (-0.60 to -0.43). That correlation puts the error variance of a single measurement at 0.093 (0.054 to 0.145) in squared log units, about half of the mean squared change of 0.182 that persistence misses. Noise therefore makes up at least 94% (76 to 105%) of the mean squared change over the six years from 2009 to 2015 and 120% (96 to 134%) of that over the three years from 2015 to 2018, so the three-year change cannot be distinguished from pure noise. The implied standard deviation of true change over six years is 0.105 in log units, about 11% , with an interval from zero to 0.19 , and over three years it is zero (0 to 0.07).

No model can therefore exceed a skill of 0.28 (0.24 to 0.37) on the forecast targets, or 0.33 for the 2015 targets and 0.23 for the 2018 targets (Eq. 1, Fig. 1). The hybrid achieves 44% of the attainable reduction in mean squared error, boosting 43% and the mean-reversion floor 19% , which leaves more than half of the attainable reduction for a better model or covariate. The ceiling is lowest where the correlation between successive changes is most negative, on cropland, on the most clayey soils and in the middle third of starting carbon, and highest on grassland, on sandy soils and on the most carbon-rich mineral soils. A change of land-cover class between 2015 and 2018 leaves the ceiling where it is, 0.26 against 0.28 where the class did not change.

The same signature appears in the two control properties. Successive changes in pH correlate at -0.44 and in log total nitrogen at -0.41 , and over three years their noise shares, 1.04 (0.97 to 1.11) and 0.95 (0.81 to 1.07), are as close to one as that of carbon. Total nitrogen is measured on the same samples and changes with organic matter, so its agreement with carbon points to sampling and analytical variability common to every property rather than to a feature of the carbon method alone.

Change appears to depend on the starting level only because the starting value carries error. Regressing the change from 2015 to 2018 on the 2015 value gives a slope of -0.22 (-0.29 to -0.15), the pattern usually read as faster loss from carbon-rich soils. The slope on the mean of the 2015 and 2018 values shrinks to -0.06 (-0.10 to -0.006), and instrumenting the 2015 value with the 2009 value, which shares none of its error, gives -0.02 (-0.05 to 0.01). On cropland the instrumented slope is -0.004 (-0.060 to 0.030) and on grassland -0.036 (-0.088 to 0.007), so no dependence of true change on the level is detectable in either.

2.3 Measurement noise and carbon credits

Translated into stock, the error of a single LUCAS value far exceeds the gains that carbon farming is paid for (Table 3). At a typical point, holding 35 t C ha^{-1} in the top 20 cm, one remeasured composite sample shows an apparent change of 37% (27 to 48%) of the stock even where nothing has changed, about $47\text{ t CO}_2\text{ ha}^{-1}$, which is 15 times (11 to 20) the gain of $3.1\text{ t CO}_2\text{ ha}^{-1}$ that the 4 per 1000 target implies over six years. At the voluntary-market price for agriculture credits that noise is worth about US\$360 per hectare against US\$24 for the target gain, and at the EU allowance price about EUR 3,460 against EUR 227.

Detecting the target gain against this noise needs about 2,100 ($1,500$ to $3,000$) resampled locations in a conventional one-sided test at 5% with 80% power, and noise alone, without true heterogeneity, needs about 2,000. Ten locations detect only a change of 42% (34 to 51%) and a hundred one of 12% (10 to 14%). The sample-size rule of the carbon farming methodology, at a one-sided confidence of 70% , asks for about 190 locations, and at that design a project whose true change is zero is credited in 30% of certification periods and receives on average a quarter of the target gain, $0.77\text{ t CO}_2\text{ ha}^{-1}$.

Selection on the baseline adds a larger and systematic credit. Points in the lowest third by their 2015 value gained 14 % (8 to 21 %) more over the next interval than points in the lowest third by their 2009 value, about 10 t CO₂ ha⁻¹ or six times the target gain, and that difference is reversion towards the mean rather than sequestration, since the instrumented dependence of change on the level is close to zero. The methodology’s model test cannot tell reversion from sequestration either, because its criterion that the R^2 of predicted on observed change exceed zero is met by the mean-reversion floor, at 0.088 (0.085 to 0.094 across replicates), which forecasts no change beyond reversion, as it is by the hybrid and boosting at 0.21.

2.4 Ablations and the covariate series

The ablations show where the skill comes from, and the answer runs against the design argument of Sect. 4.5. Removing the recurrent decoder raises skill from 0.111 to 0.115 on the single assignment. Pooled over three replicates the gain of 0.007 has an interval of -0.002 to 0.016 , and its 90 % interval lies inside the margin of ± 0.02 . Replacing the learned encoder with yearly summary statistics gives 0.117, and swapping the LSTM for a GRU with every upstream component held identical gives 0.115. Every variant that removes or simplifies a sequence component therefore scores at least as well as the full model, so neither the convolutional encoder nor the recurrent decoder is doing measurable work.

Removing the static embedding is the one change that matters, dropping skill to -0.293 , far below persistence. That embedding carries the soil properties and the previous carbon value, and withholding the previous value alone lowers skill to 0.060. The loss of 0.051 arises even though log total nitrogen measured at the same campaign correlates with the previous log carbon at $r = 0.91$, so a close proxy for the level does not replace the measured level itself.

The covariate series carry about a third of either model’s skill (Supplementary Table S2). Replacing every series with zeros, with the architecture unchanged, costs the hybrid 0.033 with an interval of 0.008 to 0.052, and replacing them with random noise costs 0.030. Gradient boosting given only the static properties and the previous value reaches 0.064, against 0.100 when the summary statistics of the series are added. On this single assignment the previous measurement therefore supplies about 0.046 through reversion, the other soil properties about 0.018 more, and the satellite series 0.033 to the hybrid and 0.036 to boosting. A network that learns from the full monthly trajectory extracts no more from the series than a tree ensemble given their means, extremes and slopes.

Because the mix of Landsat sensors differs between the two target campaigns (Sect. 4.4), part of that gain could be a proxy for the campaign rather than information about the soil. Given the campaign as an explicit input on the same folds, boosting gains 0.003 (-0.004 to 0.008) from the campaign alone and 0.037 (0.029 to 0.048) from the series beyond it, so across space the series carry information about carbon change that the campaign does not, and both models receive them, which leaves the comparison unaffected. The nested search gives no sign that a different configuration of the hybrid would change the ordering, since nine of its twelve configurations were selected at least once across twenty searches and none was consistently better (Supplementary Section S5).

2.5 Forward temporal validation

Forward temporal validation asks whether a model generalises to a later survey rather than to new places (Supplementary Table S3), training on the 2015 targets alone and testing on the 2018 targets.

At the same points the hybrid reaches 0.143 against 0.078 for boosting with its protocol settings, a lead of 0.064 with an interval of 0.036 to 0.089. That design lets every test point

contribute its earlier observation to training, and withholding the test points in space as well leaves the hybrid almost where it was, at 0.139 over five replicates, while protocol boosting falls to 0.034, below the mean-reversion floor of 0.056. The hybrid’s lead over protocol boosting at new places is 0.105 (0.021 to 0.190), and its margin over the floor is 0.083 (0.041 to 0.108), so at a later date and at places it has not seen the hybrid forecasts change clearly beyond reversion towards the mean.

Protocol boosting loses that skill through the covariate series. On the single assignment of folds, boosting given only the static properties and the previous value reaches 0.077 at new places and adding the series lowers it to 0.036, whereas under spatial blocking the same series raise it by 0.036, so a relation learned from the series in 2015 does not carry to 2018. A shift in the Landsat sensor mix between the two windows is one candidate, although the comparison does not isolate it.

2.6 Stronger boosting baselines

Refitting the protocol baseline with timing features, with nested selection or with both, on the same folds and replicates, leaves the spatial comparison where it was, with the hybrid between 0.003 below and 0.014 above every variant and every interval reaching zero (Supplementary Table S4). At a later survey selection matters, raising boosting at new places from 0.034 to 0.119 with summary statistics, so the hybrid’s lead there shrinks from 0.105 to 0.020 (−0.003 to 0.042) and at the same points vanishes, and most of its advantage in temporal extrapolation came from boosting’s fixed settings.

2.7 Observed climate covariates

Adding ERA5-Land temperature, precipitation and soil water (Fig. 2) changes neither method across space, moving the hybrid by 0.001 (−0.007 to 0.015) and tuned boosting by 0.008 (−0.002 to 0.016) over the five replicates, and with climate included the hybrid leads tuned boosting by 0.007 (−0.017 to 0.029). A lead of 0.015 on a single assignment of folds sits near the top of the range across replicates, −0.006 to 0.021, so observed climate does not separate the methods.

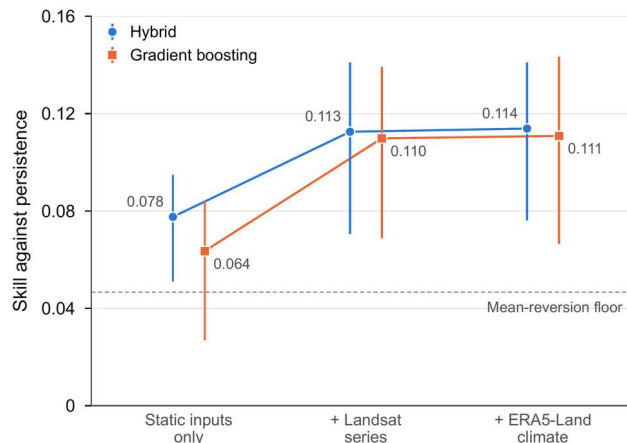


Figure 2: Skill against persistence as covariate series are added to the static inputs, mineral soils, spatially blocked validation. The Landsat series raise both models by a similar amount, and climate leaves both nearly unchanged. The Landsat and climate steps are pooled over five replicates of the fold assignment and the static-only step is the single assignment. Whiskers, 95 % block-bootstrap intervals; dashed line, mean-reversion floor.

2.8 Mean reversion, subgroups and spatial pattern

A linear fit of the new value on the previous one reaches a skill of 0.042 on the single assignment of folds, and shifts for each campaign add almost nothing, so the models are not learning an offset between surveys (Supplementary Table S5, Supplementary Section S6). No subgroup separates the two methods, since every interval across land use, starting concentration and climate zone includes zero, and both reach 0.19 to 0.20 in the lowest third of previous concentration against about 0.02 in the middle third, where there is least room to revert. In space the predictions do not reproduce regional differences in change (Fig. 3). Averaged over the 354 cells of 100 km that hold at least ten observations, predicted and observed change correlate at only $r = 0.11$, and point by point the predictions are compressed to a slope of 0.24 against observed change, as expected when most of the observed change is noise that no forecast can follow.

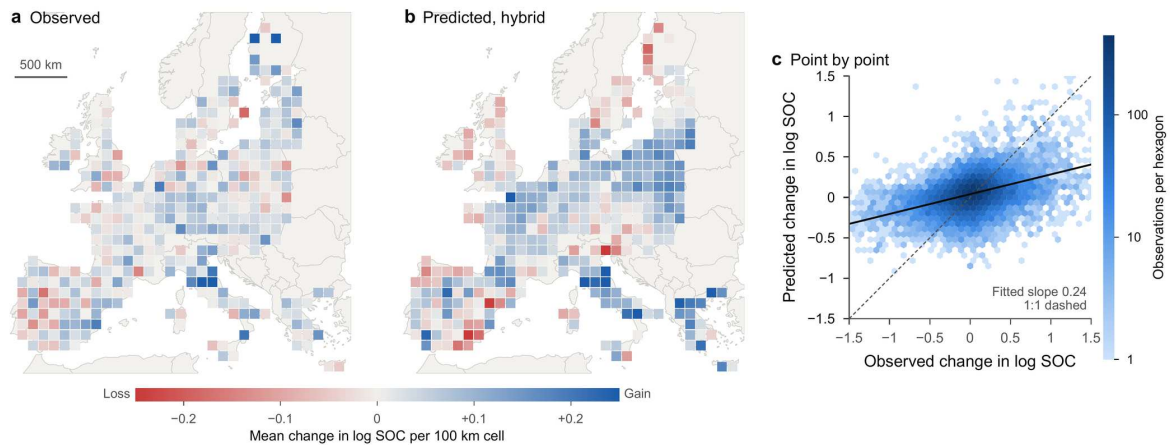


Figure 3: Change in log SOC, mineral soils. (a) Observed and (b) hybrid-predicted mean change per 100 km cell; red, loss; blue, gain. (c) Predicted against observed change, with the fitted line (solid) and the 1:1 line (dashed). Country outlines, Natural Earth.

2.9 Organic soils

On the organic soils, 202 observations from 128 points under three blocked folds, no model beats persistence (Supplementary Table S6). The best variant, the hybrid without its learned encoder, reaches -0.017 and the full hybrid -0.049 . Persistence errs more here than on mineral soils, 0.677 against 0.427 , so there is more change to explain, but 202 observations are too few to learn it, and no bootstrap intervals were computed on this subset (Supplementary Section S2).

3 Discussion

3.1 Sequence components across space

The architecture assumes that the trajectory of a covariate series carries information about carbon change beyond its summary statistics. The Landsat series do carry about a third of either model's skill, but the hybrid extracts 0.033 of it and boosting, given means, extremes and slopes, 0.036, and timing features add only 0.006 more, so the information lies in the level and spread of the series rather than in the timing that a convolution and a recurrent decoder are built to read. Benchmarks across 45 tabular datasets of around 10,000 samples found tree ensembles still ahead of deep networks⁵¹, and a panel of 16,496 rows with two intervals per point falls in that regime.

Across space, skill comes from three sources of roughly known size. Reversion towards the

mean supplies about 0.046, the other soil properties about 0.018, consistent with the positive association of SOC with total nitrogen and clay reported for deltaic soils⁵², and the satellite series about 0.035, and both methods assemble the same total from them. Measured against the noise ceiling of 0.28, both capture under half of the attainable reduction in error, so the remaining headroom is real, but neither the sequence machinery nor tuned trees claim any of it. A graph extension mixing each observation with its geographic neighbours added 0.022 beyond boosting, and boosting given the mean previous carbon of the same neighbours recovers about half of it (Supplementary Section S3), so part of what is missing is local information about each point’s surroundings rather than a better reading of its own history.

3.2 Temporal extrapolation

The one setting in which the hybrid leads is temporal extrapolation, with training on the 2015 targets and testing on 2018. Its large lead over protocol boosting, 0.105 at new places, arises because boosting with fixed settings learns a relation between the series and carbon in 2015 that does not hold in 2018, and against boosting with nested selection the lead shrinks to 0.020 with an interval that just reaches zero. The hybrid’s distinctive property is therefore robustness rather than reach, since it holds its skill at a later survey without the selection boosting needs, and its margin over reversion towards the mean there, 0.083 (0.041 to 0.108), is the clearest evidence in this study of change forecast beyond reversion. The ablations were run under spatial blocking only, so which component carries that robustness remains open.

3.3 Measurement noise, the Soil Monitoring Law and carbon credits

The noise analysis has a direct reading for the Soil Monitoring Law. At a single LUCAS point, noise makes up at least 94 % of the mean squared change in log carbon over six years, the interval of the directive’s measurement cycle, and the three-year change cannot be told from noise at all. The implied true change over six years has a standard deviation of about 11 % with an interval reaching zero. A change measured at one point between two cycles is therefore mostly a statement about sampling and laboratory variability, and the directive’s carbon descriptor can be read as a trend only when aggregated over many points or when each value is made more precise. Two features of the result argue that the noise is in the measurement rather than in carbon itself. Total nitrogen and pH, measured on the same samples, show the same negative correlation between successive changes, and a change of land-cover class, which should alter true carbon, leaves the ceiling unchanged.

Lower measurement noise would raise the ceiling, shrink the share of skill that reversion towards the mean can supply, and leave more of the measured change for covariates to explain. Composites of more cores around each point, replicate analyses that separate laboratory from field error, and marked or precisely relocated sites each address a different part of the error variance estimated here, and none of them needs a better model. The finding also bears on how model skill should be reported. A forecast given the previous measurement draws about two fifths of its skill against persistence from reversion towards the mean, and a ceiling of 0.28 bounds the rest, so skill reported without a mean-reversion floor and a noise ceiling cannot be read as skill at forecasting change. The same floor and ceiling belong to an evaluation protocol set out for forecasting SOC change in Bangladesh rice land, where repeated measurements are sparser still⁵³.

For carbon crediting the same error sets a threshold that a project must clear before a credited gain can be told from noise. The carbon farming methodology under Regulation (EU) 2024/3012 credits the mean measured change less a deduction at a one-sided confidence of 70 %¹⁷, a level that credits three in ten projects whose true change is zero and, at the sample size the

methodology prescribes, pays them on average a quarter of the target gain. The error variance estimated here makes the scale concrete, since detecting a gain at the 4 per 1000 rate within one six-year cycle with conventional confidence would need about 2,100 resampled locations of LUCAS precision, eleven times the 190 that the methodology’s own rule asks for. Two provisions of the methodology deserve scrutiny in that light. Its measurement error for mineral soils is taken from the laboratories, whereas the variation between repeated samples of a site is several times the analytical error⁵⁴, and the sampling error it adds may come from repeated sampling of some locations, which would capture the error of remeasurement, or from local information on carbon variability, which need not. Its model test, an R^2 on observed change above zero, is passed by a model that forecasts reversion towards the mean and nothing else, so a model can be validated while crediting the noise of the baseline measurement. Crediting rules that pay on remeasured change therefore need a sample size set from an error variance estimated on repeated samples, a validation benchmark of mean reversion rather than of zero, and an enrolment rule that does not select fields on a single low measurement.

The level analysis carries a second implication. Regressing measured change on the starting value gives the familiar pattern of faster loss from carbon-rich soils, and that pattern underlies claims that high-carbon soils are most at risk. Here it vanishes once the starting value is instrumented by an earlier, independent measurement, so on LUCAS mineral soils over this period a dependence of change on the level is an artefact of measurement error rather than a property of the soils. The same artefact turns into credit when fields are enrolled on a low baseline value, since the gain of 14 % that reversion alone then produces over the next measurement is six times the six-year gain of the 4 per 1000 target.

3.4 Relation to earlier work

Zhang et al.⁴³ found that adding ten years of MODIS phenology to a convolutional-recurrent network produced a more accurate map of soil carbon, and the result here is compatible with theirs, since the dynamic inputs again contribute, about a third of the skill, while the recurrent decoder that processes them contributes nothing across space. The gains reported for deep learning in digital soil mapping^{33;35} concern mapping and not change, and the graph extension of Supplementary Section S3 agrees with the spatial-context argument in direction, since geographic neighbours added 0.021 on mineral soils.

Studies of change on the same survey point the same way. Tian et al.⁹ found predicted change at single LUCAS sites to be poorly resolved and detectable only after spatial averaging, and the noise share estimated here gives the reason from the measurement side. De Rosa et al.²⁷ estimated change across the revisited LUCAS points with a quantile additive model, whereas on the restricted panel of this study the hybrid’s cell averages do not reproduce regional contrasts. This study adds the comparison those studies did not make, between a sequence model and a tabular one given identical inputs, together with an estimate of the measurement error of a single LUCAS value and the bound it places on any forecast.

Reversion towards the mean has long been recognised in soil monitoring as a source of apparent change^{49;50}, and in 51 Canadian forest plots measured three times, uncorrelated changes over the two intervals led Buchkowski et al.⁵⁵ to read the apparent gain over the first as a statistical artefact. The LUCAS panel shows the signature far more strongly, with a correlation of -0.54 over 8,221 points, and the present analysis turns it into estimates of the error variance, the noise share, the attainable skill, the level dependence that remains once the starting error is removed and the error that a carbon credit inherits.

3.5 Limitations

The comparison covers one family of sequence models, a convolutional encoder with a recurrent decoder in LSTM and GRU forms, and attention-based or tabular deep networks were not tested. The result also rests on one region and one survey programme, so it shows what the sequence components contribute on LUCAS and not what they contribute for soil carbon everywhere. Covariates are Landsat at 30 m and ERA5-Land at about 9 km, and a denser or finer record might carry structure these do not. Landsat reflectance was not adjusted across sensors, and although the series carry information beyond the campaign across space (Sect. 2.4), their failure to transfer under protocol boosting at a later survey may partly reflect the sensor shift. Management, which process models treat as the main driver of change on cropland, is not recorded for the LUCAS grid, so no process model such as RothC or DayCent could be run as a further reference^{18;19}.

The crediting figures convert concentration to stock with literature bulk densities and the 0–20 cm LUCAS depth, whereas the methodology requires 0–30 cm and a measured or validated fine-earth bulk density, and values in tonnes and currency scale linearly with both, unlike the ratios and location counts. They also assume that the error of a LUCAS composite applies to a project’s resampling. Projects that relocate points precisely, composite more cores and analyse both rounds in one laboratory batch would carry less error, and the location counts scale with the error variance, whereas an error shared by every location of a project, such as a laboratory offset between rounds, is not reduced by more locations and is not part of the estimate here.

The noise estimates rest on two assumptions, that measurement errors are independent between visits and that true changes over the two intervals are not negatively correlated. A real reversal of change at many points, such as a management cycle that adds carbon and then removes it, would make part of the negative correlation true signal and lower the noise share. Errors that differ between campaigns, such as a change of laboratory protocol, would move the estimates in either direction, and the three-year share above one suggests that the error variances are not exactly equal. LUCAS publishes no replicate measurements, so neither assumption can be checked directly. The variogram range of 780 km leaves 18 blocks, although halving the block size to give 56 blocks leaves the comparison unchanged.

The shape of the panel is itself a limit, since an architecture built for long sequences works here with two intervals per point, and with two target campaigns the covariate windows for 2015 and 2018 overlap by 24 months. The temporal results rest on one pair of campaigns and organic soils on 202 observations, and the equivalence margin of ± 0.02 was set before the corrected analysis, although the 95 % interval for the hybrid-minus-boosting difference includes zero whatever margin is chosen.

3.6 Conclusions

A convolutional encoder coupled to a gated recurrent decoder forecasts change in soil organic carbon across space as well as gradient boosting on summary statistics given the same inputs, and no better. On 16,496 mineral-soil observations from 8,275 European points measured three times, the hybrid reaches a skill of 0.113 and boosting 0.110 under spatially blocked validation with hyperparameters chosen inside each training fold and pooled over five replicates, and the difference of 0.003 (-0.011 to 0.019) is equivalent to zero within 0.02. Removing the recurrent decoder or the learned encoder does not lower skill, and the Landsat series, which supply about a third of the skill, are read as well by boosting as by the network.

Measurement noise sets the limits of the problem. Successive changes at a point correlate at -0.54 , so noise makes up at least 94 % of the mean squared change over six years and all of it over three, a pattern shared by pH and total nitrogen. No model can exceed a skill of about 0.28, and the apparent faster loss of carbon from carbon-rich soils disappears once the starting

value is instrumented by an earlier measurement. Change measured at a single point every six years, the cycle of the European Soil Monitoring Law, is therefore readable as a trend only in aggregate or with more precise measurement, and on this panel a lower measurement error would raise the attainable skill more than any change of model.

In carbon-crediting terms one remeasured composite sample carries noise of about $47 \text{ t CO}_2 \text{ ha}^{-1}$, fifteen times the gain of the 4 per 1000 target over a six-year cycle, so a credited gain at that rate needs about 2,100 resampled locations to be detected. A methodology that credits at 70 % confidence pays three in ten projects whose true carbon did not change, enrolment on a low baseline credits a reversion gain six times the target gain, and a model that forecasts only reversion passes its validation test. The hybrid's one advantage appears at a later survey at unseen places, where it exceeds the mean-reversion floor by 0.083 (0.041 to 0.108) and tuned boosting by 0.020 with an interval reaching zero.

4 Methods

4.1 The LUCAS soil panel

The Land Use and Coverage Area frame Survey (LUCAS) samples topsoil across the European Union on a common grid, returning to the same identified points in successive campaigns⁵⁶. Three campaigns carry organic carbon at a shared set of points, 2009, 2015 and 2018, which is the structure a sequence model needs. Matching the campaigns by point identifier gives 27,333 points with a usable carbon value in at least one campaign. Restricting to cropland and grassland, resolved per point because the 2009 release does not record land cover, leaves 15,201, and requiring a carbon value in all three campaigns leaves 8,349 points. Targets are the 2015 and 2018 campaigns, each paired with its predecessor, which gives 16,698 target observations. Each release is read by its organic-carbon column alone, and a check in the code refuses any column that is not a spelling of organic carbon, since the 2015 and 2018 files also carry electrical conductivity.

Carbon concentration is strongly right-skewed, with a mean of 22.7 g kg^{-1} against a median of 15.6 in 2009, and a skewness of 8.5 that falls to 0.59 under a log transform, so the target is log concentration throughout. Stock is not reported because bulk density is not available across the panel. An observation is classed as organic when its previous carbon value exceeds 120 g kg^{-1} , a rule that uses only what is known when the forecast is made, whereas classing on every campaign would select on the outcome by dropping points whose carbon rose past the threshold. Organic observations are 202 from 128 points, 1.2 % of the panel, and they carry a large share of the spread of change, since restricting to the points below the threshold in both earlier campaigns cuts the standard deviation of the 2009 to 2018 difference from 17.6 to 10.5 g kg^{-1} . Mineral and organic soils are therefore analysed separately, and the main results are for the 16,496 mineral-soil observations from 8,275 points, 8,247 of them targeting 2015 and 8,249 targeting 2018.

4.2 Signal and noise in measured change

Successive changes at a point show how much of the measured change a covariate could hope to explain. A point that rises and then falls more often than chance allows is the signature of reversion towards the mean rather than of a trend⁴⁹, and LUCAS returns to a point rather than to the identical spot, so sampling and laboratory variability alone would produce it. LUCAS sites are not marked, so a repeated sample represents the surrounding area, and the variation between repeated samples is several times that between repeated laboratory analyses of one sample⁵⁴. A model that learns only to pull high values down and low values up will therefore beat persistence without forecasting any real change, and persistence cannot separate those two

outcomes. A linear fit of the new log value on the previous one and the campaign, fitted on the training folds, is reported beside it as a mean-reversion floor (Sect. 4.7).

Three visits to a point separate measurement error from true change in the way that three waves of a test-retest design separate unreliability from instability⁵⁷. Suppose each measured log value is the true value plus an error that is independent between visits and of every input. The changes Δ_1 (2009 to 2015) and Δ_2 (2015 to 2018) at a point then share the error of the middle measurement with opposite signs, so that their covariance equals the covariance of the true changes minus the error variance σ_e^2 . Provided the true changes at a point are not negatively correlated, σ_e^2 is at least the negative of the observed covariance. The error of the new measurement cannot be forecast, so no model reaches a mean squared error below σ_e^2 , and its skill against persistence is bounded by

$$S_{\max} = 1 - \sqrt{\sigma_e^2 / \overline{\Delta^2}}, \quad (1)$$

where $\overline{\Delta^2}$ is the mean squared change between consecutive measurements, the error of persistence. Using the lower bound for σ_e^2 makes $S_{\max}$ an upper bound under an equal error variance at every campaign. A model of skill S then achieves the share $(1 - (1 - S)^2) / (1 - \sigma_e^2 / \overline{\Delta^2})$ of the attainable reduction in mean squared error.

The same error variance gives three further quantities, computed on the 8,221 points whose first two campaigns were both mineral. Both ends of an interval carry an error, so the share of the mean squared change over an interval that is noise is $2\sigma_e^2 / \overline{\Delta^2}$, a lower bound for the same reason as above, and it is reported for the six-year and the three-year interval. A share above one is possible only if the error variance differs between campaigns or the true changes are negatively correlated, and is read as change indistinguishable from noise. Second, regressing change on the starting value confounds a real dependence on the level with reversion towards the mean, because the starting error enters both sides. The 2009 value carries none of the errors of 2015 and 2018, so it serves as an instrument for the 2015 value in a regression of Δ_2 on it, giving the slope of true change on the true starting level if that change is linear in the level. That slope is reported beside the naive slope on the 2015 value and the slope on the mean of the 2015 and 2018 values that Lark et al.⁴⁹ proposed. Third, the ceiling is recomputed within strata defined by values that carry none of the later errors, namely land cover in 2015, a change of land-cover class between 2015 and 2018, and terciles of clay and of 2009 carbon. The identity is also applied to pH in CaCl_2 and to log total nitrogen, two properties measured on the same samples and expected to change slowly, as controls. Particle size, which should not change at all, was measured only in 2009 and cannot serve. All intervals in this section resample the spatial blocks of the main analysis.

4.3 Measurement noise in carbon-crediting terms

The error variance is translated into the terms of the carbon farming methodologies under Regulation (EU) 2024/3012¹⁷, per hectare over one six-year cycle. Stock is concentration multiplied by fine-earth bulk density and by the 0–20 cm depth that LUCAS samples, with the LUCAS fine-earth means of 1.16, 1.00 and 1.02 g cm⁻³ for cropland, grassland and other cover⁵⁸ and, as an upper case applied to every point, the measured cropland mean of 1.26 g cm⁻³⁵⁹, without a correction for coarse fragments. A typical point is the median stock. The reference gain is the 4 per 1000 target¹² accumulated over six years, $1.004^6 - 1 = 2.4\%$. Carbon is converted to carbon dioxide by 44/12 and valued at the 2024 average transaction price of agriculture credits in the voluntary market, US\$7.66 per tonne⁶⁰, and at the 2025 average price of EU emission allowances, EUR 74 per tonne⁶¹, an upper reference, since certified carbon farming units are not traded in the EU emissions trading system.

Five quantities follow. The apparent change from noise at one remeasured location with

no true change is $\exp(\varepsilon) - 1$ with $\varepsilon \sim N(0, 2\sigma_e^2)$, reported as its mean absolute value. The number of resampled locations needed to detect the reference gain uses the standard deviation of the measured six-year change, which holds both noise and true heterogeneity, in a one-sided test at 5 % with 80 % power, and in equation (8) of the methodology, which sets the sample size at a one-sided confidence of 70 % ($z = 0.52$) with a tolerable error equal to the effect size multiplied by the confidence level. The methodology credits the mean measured change less an uncertainty deduction of t_{70} standard errors, or 8 % if larger (its equation 12), so a project whose true change is zero is credited whenever the noise in its mean exceeds 0.52 standard errors, and the probability and expected size of that credit follow from the normal distribution. Fields are often enrolled because their baseline value is low, and the gain that reversion alone then produces is the difference between the mean 2015–2018 change in the lowest third of points selected by their 2015 value, which shares the error of that change, and in the lowest third selected by their 2009 value, which does not. Finally, the methodology accepts a model whose R^2 on observed change exceeds zero (its equation 7), and that R^2 is computed for the mean-reversion floor and both models on the five spatial replicates. Intervals resample the spatial blocks.

4.4 Covariates

Dynamic covariates come from Landsat 5, 7, 8 and 9 Collection-2 surface reflectance⁶², masked for cloud and shadow and composited to monthly medians at 30 m through Google Earth Engine⁶³. One sensor family carries every campaign, which avoids switching to Sentinel-2 partway through the panel, although the mix of Landsat sensors shifts between windows. The windows before 2009 draw on TM and ETM+, those before 2015 add OLI from 2013, and those before 2018 draw mostly on ETM+ and OLI, whose reflectance and NDVI differ slightly⁶⁴. No cross-sensor adjustment was applied, so the series could carry the identity of the campaign, and Sect. 2.4 tests whether they do. Six reflectance bands are carried with NDVI for vegetation and NDWI and MNDWI for the water regime, and a tenth channel records which months were observed, so the model can distinguish an interpolated month from a measured one. Across the panel 75 % of months carry a clear observation.

Each observation carries the 60 months ending in December of the year before its survey, since LUCAS does not publish per-point sampling dates and a window reaching into the survey year would risk conditioning a forecast on weather that followed the measurement. Static covariates are clay, sand, silt, pH and total nitrogen measured at the previous campaign, together with the previous carbon value, so everything an observation carries was knowable when the forecast would have been made. LUCAS measures texture only on a point’s first visit, so the clay, sand and silt measured then are carried forward to later campaigns, since texture does not change on these timescales. Cation exchange capacity is recorded in 2009 but not in 2015 and is excluded.

For a robustness check, monthly 2 m air temperature, total precipitation and top-layer volumetric soil water from ERA5-Land⁶⁵ were extracted at every point over the same windows and appended as three further dynamic channels. Coastal points that fall on masked ocean cells took the mean of land cells within 20 km.

4.5 Model architecture

The hybrid separates within-season variation from between-year change, and dynamic from static information, so that neither is flattened into one feature vector (Fig. 4a). The dynamic sequence $X \in \mathbb{R}^{C \times T}$, here $C = 10$ channels over $T = 60$ months, passes through a stack of one-dimensional dilated convolutions, with layer l computing

$$z_t^{(l)} = \phi \left(\sum_{j=0}^{k-1} W_j^{(l)} z_{t-d_{lj}}^{(l-1)} + b^{(l)} \right), \quad (2)$$

where k is the kernel width, $d_l = 2^{l-1}$ the dilation and ϕ a rectified linear activation⁶⁶. With $k = 3$ and three layers, each of 32 or 64 filters as selected in each fold (Sect. 4.6), the receptive field is 15 months, long enough to detect a seasonal trajectory, such as a drop in NDVI coinciding with a rise in NDWI, wherever it falls in the window. A fixed set of monthly summary statistics cannot provide that shape-detecting invariance, which is the usual argument for convolution over engineered features^{29;43}.

The static vector s passes through a dense embedding of width 16 and is broadcast across all T' encoder output steps,

$$u_t = \text{ReLU}(W_u [z_t \parallel \text{Emb}(s)] + b_u), \quad t = 1, \dots, T', \quad (3)$$

so that the same wet-season signal can be read differently on a clay soil than on a sandy one, the interaction that process models encode as a moisture modifier on decomposition^{4;18}. The fused sequence enters a gated recurrent layer whose cell is interchangeable between a long short-term memory (LSTM) cell⁶⁷ and a gated recurrent unit (GRU)⁶⁸, with everything upstream held identical, so the two cells are compared under controlled conditions. The cell equations are given in Supplementary Section S1 (Supplementary Figs. S1 to S4).

The head emits a mean $\hat{\mu}$ and a predictive standard deviation $\hat{\sigma}$, trained under a Gaussian negative log-likelihood so the model learns its own uncertainty⁶⁹. Training uses Adam⁷⁰ with early stopping on whole spatial blocks held out from the training fold, so no stopping observation has a neighbour, or a second observation of its own point, on the fitting side. Each reported network is a three-member deep ensemble.

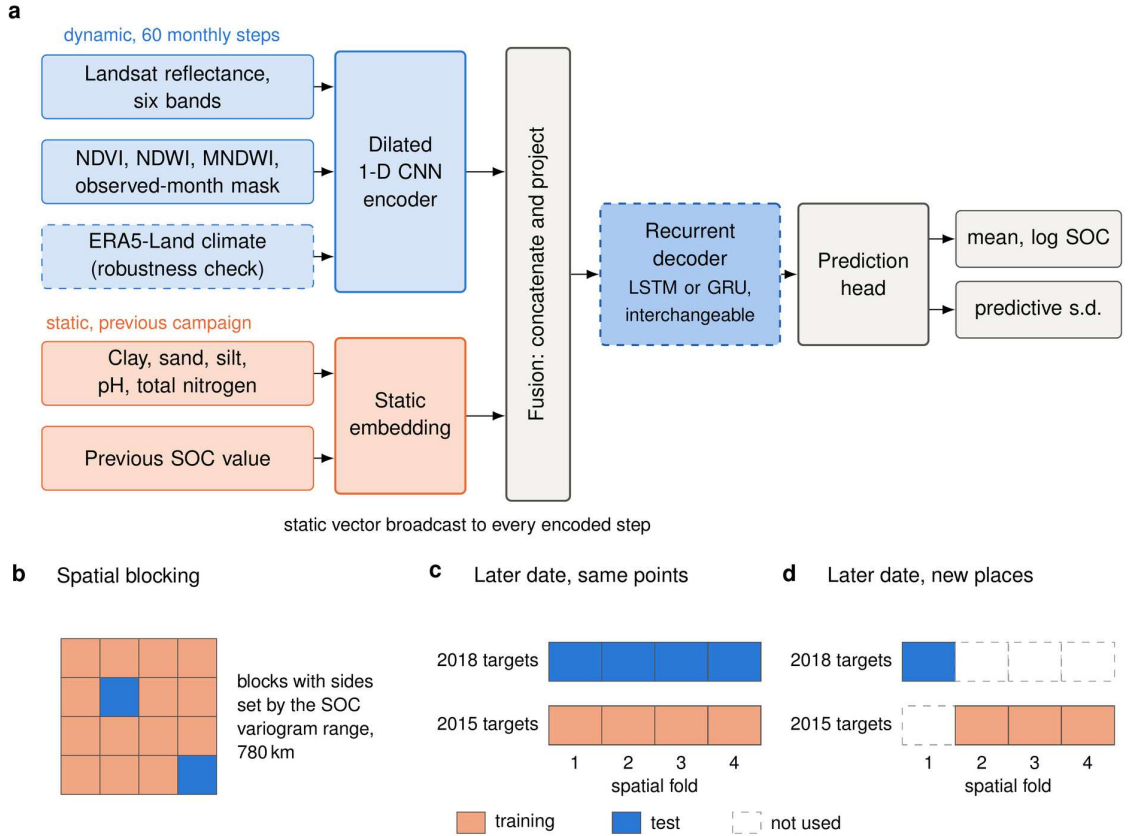


Figure 4: Model and validation designs. (a) Convolutional-recurrent hybrid; dashed parts are optional (ERA5-Land climate) or interchangeable (LSTM or GRU decoder). (b) Spatial blocking. (c) Forward validation at the same points. (d) Forward validation at new places.

4.6 Evaluation design

Every result is set against three references, the first being a persistence forecast that predicts no change from the last observation and so sets the floor, because SOC is so persistent that a model reproducing the last value scores well on concentration while saying nothing about change. Skill is therefore reported as $1 - \text{RMSE}_{\text{model}}/\text{RMSE}_{\text{persistence}}$, so that zero matches persistence and a negative value does worse than assuming no change, with RMSE on log concentration and the concordance correlation coefficient alongside. Persistence errs by exactly the observed change, so this is also the skill with which a model predicts the change itself. Gradient boosting²³ on temporal summary statistics, given the same inputs as the architecture, represents the working standard of the field, and this is the comparison that decides whether the architecture earns its complexity. For each dynamic channel it receives the mean, standard deviation, minimum, maximum and linear slope, together with the static properties and the previous carbon value, and it runs with fixed settings (400 trees of depth 4, learning rate 0.05, rows and columns subsampled at 0.8). Those five statistics discard seasonality and timing, and the fixed settings forgo the selection the hybrid receives, so both choices favour the hybrid. Three stronger variants remove them (Sect. 2.6). Timing features give each channel, beyond the five statistics, the mean of each calendar month, the mean of each year, the amplitude and peak month of the mean seasonal cycle, the last year’s mean against the four before it and the mean month-to-month change, 270 features in all. Nested selection repeats the hybrid’s protocol, so that inside every training fold a grid of twelve settings of tree depth (3, 5, 8), minimum child weight (1, 20) and column subsampling (0.5, 0.9) is fitted with early stopping on whole blocks and scored on a held-out quarter of the blocks, and the winner is refitted on the fold. Extending the grid to depths of 10 and 12, which were then chosen in one of the 66 searches, moved no skill or difference by more than 0.001. The third reference is the mean-reversion floor of Sect. 4.2.

Validation withholds contiguous spatial blocks rather than random points (Fig. 4b–d), with block size set from the range of the SOC variogram so that a held-out point has no training neighbour inside its autocorrelation distance^{46;47}. The fitted range of 780 km gives 18 blocks, with a median of 496 observations each, assigned to four folds (Fig. 5). Forward temporal validation trains on the 2015 targets and tests on the 2018 targets. Random k -fold validation is run only as an optimism reference, and calibration is checked by the share of held-out observations inside a nominal 90 % prediction interval (Supplementary Section S4). Organic soils use the same block size, which gives 12 blocks in three folds, because their variogram, on 202 observations, does not level off to a sill.

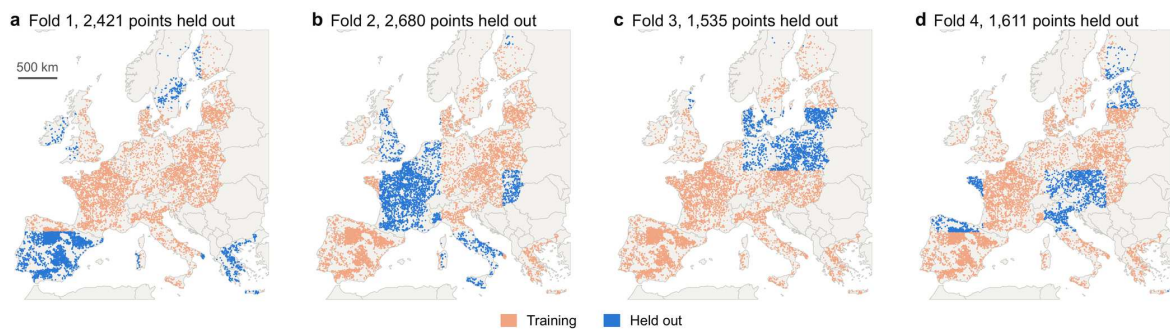


Figure 5: Mineral-soil LUCAS points ($n = 8,275$) and the four spatially blocked folds, with (a)–(d) holding out folds 1 to 4. Blue, held out; orange, training.

Five ablations, each a subtraction from the full model, isolate where any skill originates. They remove the recurrent decoder, replace the convolutional encoder with yearly summary statistics, swap the LSTM for a GRU, remove the static embedding, and withhold the previous carbon value. Width, learning rate and dropout are selected by a nested search, so that no test

observation influences the configuration. Inside each training fold a quarter of the blocks is held out, twelve configurations varying width between 32 and 128 units, learning rate and dropout are fitted on the remaining blocks and scored on those held out, and the best is used for that fold. Every ablation of a fold runs at the configuration selected for the full model, so a difference between ablations reflects the component removed and not its hyperparameters.

The hybrid is judged to earn its added complexity only if it outperforms gradient boosting given the same inputs under spatially blocked validation, which requires the 95 % interval on the hybrid’s advantage to lie above zero. Whether the two methods can instead be called equivalent is tested separately.

4.7 Comparing models

Differences in skill are small, so each is reported with a 95 % interval from a spatial block bootstrap. Whole validation blocks are resampled with replacement 2,000 times and every model is scored on the same resampled set, which keeps the pairing between models and the spatial correlation within each replicate. Resampling single observations would treat neighbouring points as independent and give intervals that are too narrow.

A single run depends on the network seeds and on which blocks fall in which fold, so the central comparison is repeated over five replicates, each with its own seeds and its own random assignment of blocks to folds. The bootstrap draws from all five, so the intervals carry both sources of variation. Equivalence is judged by whether the 90 % interval on the difference lies inside ± 0.02 skill, a two one-sided test. The margin was set during an earlier analysis of the panel and is kept unchanged here. On the corrected panel it is wider than the whole spread that reassigning blocks to folds produces, with the hybrid’s skill ranging from 0.108 to 0.116 over the five replicates and its difference from boosting from -0.004 to 0.011 . The mean-reversion floor of Sect. 4.2 is fitted in every replicate. To test whether the comparison depends on the block size, it is repeated over three replicates with blocks of 390 km, half the variogram range, which gives 56 blocks in place of 18, with every hyperparameter search redone on the new folds. A second temporal design trains on the 2015 targets outside a fold and tests on the 2018 targets inside it, so that no test point also contributes a training observation.

Skill is also broken down by subgroup on the single assignment of folds (Supplementary Table S5). Land use is the cover recorded at the target campaign, with the 730 observations under covers other than cropland and grassland left out. Climate zones follow the Köppen-Geiger rules of Peel et al.⁷¹, applied to each point’s mean monthly ERA5-Land temperature and precipitation over its covariate windows, and no point falls in the polar zone. Persistence RMSE is reported for each zone as a measure of how much change there is to forecast, and the subgroup intervals are not adjusted for the nine comparisons.

Data availability

No new measurements were made. The LUCAS topsoil campaigns are available from the European Soil Data Centre on request and are not redistributed here, since their licence grants a personal, non-transferable right of use. Covariates were derived from Landsat Collection-2 surface reflectance through Google Earth Engine⁶³, and climate from ERA5-Land⁶⁵.

Code availability

Analysis code, the covariate extraction and the scripts that rebuild the panel from ESDAC files are archived at Zenodo (version 3.2.0, doi:10.5281/zenodo.23583865) and developed at

<https://github.com/rakibhhridoy/soc-sequence-test>, so a reader with their own ESDAC access can reproduce every table. Archived versions before 3.0.0 read electrical conductivity as organic carbon for the 2015 and 2018 campaigns and should not be used.

References

- [1] Batjes, N. H. Total carbon and nitrogen in the soils of the world. *European Journal of Soil Science* **47**, 151–163 (1996).
- [2] Post, W. M., Emanuel, W. R., Zinke, P. J. & Stangenberger, A. G. Soil carbon pools and world life zones. *Nature* **298**, 156–159 (1982).
- [3] Stockmann, U. *et al.* The knowns, known unknowns and unknowns of sequestration of soil organic carbon. *Agriculture, Ecosystems & Environment* **164**, 80–99 (2013).
- [4] Parton, W. J., Schimel, D. S., Cole, C. V. & Ojima, D. S. Analysis of factors controlling soil organic matter levels in Great Plains grasslands. *Soil Science Society of America Journal* **51**, 1173–1179 (1987).
- [5] Six, J., Conant, R. T., Paul, E. A. & Paustian, K. Stabilization mechanisms of soil organic matter: Implications for C-saturation of soils. *Plant and Soil* **241**, 155–176 (2002).
- [6] Wiesmeier, M. *et al.* Soil organic carbon storage as a key function of soils: A review of drivers and indicators at various scales. *Geoderma* **333**, 149–162 (2019).
- [7] Yan, X., Cai, Z., Wang, S. & Smith, P. Direct measurement of soil organic carbon content change in the croplands of China. *Global Change Biology* **17**, 1487–1496 (2011).
- [8] Smith, P. *et al.* How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. *Global Change Biology* **26**, 219–241 (2020).
- [9] Tian, X., de Bruin, S., Schneider, F., Herold, M. & de Beurs, K. Modeling soil organic carbon changes using signal-to-noise analysis: A case study using European soil survey datasets. *Global Change Biology* **32**, e70813 (2026).
- [10] Szatmári, G. *et al.* Space-time modelling of soil organic carbon stock change at multiple scales: Case study from Hungary. *Geoderma* **451**, 117067 (2024).
- [11] Lal, R. Soil carbon sequestration impacts on global climate change and food security. *Science* **304**, 1623–1627 (2004).
- [12] Minasny, B. *et al.* Soil carbon 4 per mille. *Geoderma* **292**, 59–86 (2017).
- [13] Paustian, K. *et al.* Climate-smart soils. *Nature* **532**, 49–57 (2016).
- [14] Bossio, D. A. *et al.* The role of soil carbon in natural climate solutions. *Nature Sustainability* **3**, 391–398 (2020).
- [15] European Parliament and Council of the European Union. Directive (EU) 2025/2360 of the European Parliament and of the Council of 12 November 2025 on soil monitoring and resilience (Soil Monitoring Law). Official Journal of the European Union, L series, 26 November 2025 (2025). URL <https://eur-lex.europa.eu/eli/dir/2025/2360/oj/eng>.

- [16] European Parliament and Council of the European Union. Regulation (EU) 2024/3012 of the European Parliament and of the Council of 27 November 2024 establishing a Union certification framework for permanent carbon removals, carbon farming and carbon storage in products. Official Journal of the European Union, L series, 6 December 2024 (2024). URL <https://eur-lex.europa.eu/eli/reg/2024/3012/oj/eng>.
- [17] European Commission. Annex to the Commission Delegated Regulation supplementing Regulation (EU) 2024/3012 by establishing the certification methodologies for carbon farming activities. C(2026) 4666 final, Brussels, 9 July 2026 (2026).
- [18] Coleman, K. & Jenkinson, D. S. RothC-26.3: A model for the turnover of carbon in soil. In *Evaluation of Soil Organic Matter Models*, 237–246 (Springer, 1996).
- [19] Parton, W. J., Hartman, M., Ojima, D. & Schimel, D. DAYCENT and its land surface submodel: description and testing. *Global and Planetary Change* **19**, 35–48 (1998).
- [20] Wadoux, A. M. J.-C., Minasny, B. & McBratney, A. B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews* **210**, 103359 (2020).
- [21] Meyer, H. & Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution* **12**, 1620–1633 (2021).
- [22] Breiman, L. Random forests. *Machine Learning* **45**, 5–32 (2001).
- [23] Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).
- [24] Hengl, T. *et al.* SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE* **12**, e0169748 (2017).
- [25] Poggio, L. *et al.* SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL* **7**, 217–240 (2021).
- [26] Heuvelink, G. B. M. *et al.* Machine learning in space and time for modelling soil organic carbon change. *European Journal of Soil Science* **72**, 1607–1623 (2021).
- [27] De Rosa, D. *et al.* Soil organic carbon stocks in European croplands and grasslands: How much have we lost in the past decade? *Global Change Biology* **30**, e16992 (2024).
- [28] Broeg, T., Don, A., Wiesmeier, M., Scholten, T. & Erasmi, S. Spatiotemporal monitoring of cropland soil organic carbon changes from space. *Global Change Biology* **30**, e17608 (2024).
- [29] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324 (1998).
- [30] Reichstein, M. *et al.* Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204 (2019).
- [31] Ng, W. *et al.* Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma* **352**, 251–267 (2019).
- [32] Tsakiridis, N. L., Keramaris, K. D., Theocharis, J. B. & Zalidis, G. C. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. *Geoderma* **367**, 114208 (2020).

- [33] Padarian, J., Minasny, B. & McBratney, A. B. Using deep learning for digital soil mapping. *SOIL* **5**, 79–89 (2019).
- [34] Emadi, M. *et al.* Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. *Remote Sensing* **12**, 2234 (2020).
- [35] Wadoux, A. M. J.-C. Using deep learning for multivariate mapping of soil with quantified uncertainty. *Geoderma* **351**, 59–70 (2019).
- [36] Odebiri, O., Mutanga, O. & Odindi, J. Deep learning-based national scale soil organic carbon mapping with Sentinel-3 data. *Geoderma* **411**, 115695 (2022).
- [37] Odebiri, O., Odindi, J. & Mutanga, O. Basic and deep learning models in remote sensing of soil organic carbon estimation: A brief review. *International Journal of Applied Earth Observation and Geoinformation* **102**, 102389 (2021).
- [38] Kratzert, F., Klotz, D., Brenner, C., Schulz, K. & Herrnegger, M. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences* **22**, 6005–6022 (2018).
- [39] Kratzert, F. *et al.* Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research* **55**, 11344–11354 (2019).
- [40] Fang, K., Shen, C., Kifer, D. & Yang, X. Prolongation of SMAP to spatiotemporally seamless coverage of continental U.S. using a deep learning neural network. *Geophysical Research Letters* **44**, 11030–11039 (2017).
- [41] Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research* **54**, 8558–8593 (2018).
- [42] Hasan, M. R., Rahman, A., Jolly, Y. N., Zubyer, S. & Mohiuddin, A. S. M. Comparative analysis of water quality forecasting for the urban Turag river using enhanced and hybrid deep learning architectures. *Next Research* **2**, 100597 (2025).
- [43] Zhang, L. *et al.* A CNN-LSTM model for soil organic carbon content prediction with long time series of MODIS-based phenological variables. *Remote Sensing* **14**, 4441 (2022).
- [44] Dong, Z. *et al.* Prediction of soil organic carbon content in complex vegetation areas based on CNN-LSTM model. *Land* **13**, 915 (2024).
- [45] Geng, J. *et al.* 2C-Net: A novel spatiotemporal dual-channel network for soil organic matter prediction using multi-temporal remote sensing and environmental covariates. *Remote Sensing* **17**, 3358 (2025).
- [46] Roberts, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017).
- [47] Ploton, P. *et al.* Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications* **11**, 4540 (2020).
- [48] Hasan, M. R. *et al.* Geochemical predictability quantification and spatial transferability of soil organic carbon in Ganges Delta. *EGUsphere [preprint]* (2026). Doi:10.5194/egusphere-2026-4799.
- [49] Lark, R. M., Bellamy, P. H. & Kirk, G. J. D. Baseline values and change in the soil, and implications for monitoring. *European Journal of Soil Science* **57**, 916–921 (2006).

- [50] Bradford, M. A. *et al.* Testing the feasibility of quantifying change in agricultural soil carbon stocks through empirical sampling. *Geoderma* **440**, 116719 (2023).
- [51] Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems 35*, 507–520 (2022).
- [52] Uddin, M. J., Hasan, M. R., Arabi, F. Z., Akter, S. & Ali, A. H. M. Z. Spatial variability of soil and carbon sequestration in the Ganges moribund delta of Bangladesh. *Dhaka University Journal of Biological Sciences* **35**, 37–46 (2026).
- [53] Hasan, M. R. Process models, gradient boosting and deep sequence learning for soil organic carbon change in Bangladesh rice land (2026). Preprint, Zenodo, doi:10.5281/zenodo.23511068.
- [54] Hiederer, R. Data evaluation of LUCAS soil survey laboratory data: Survey 2009, 2012 and 2015. JRC Technical Report EUR 30092 EN, Publications Office of the European Union, Luxembourg (2020).
- [55] Buchkowski, R. W., Polussa, A. & Bradford, M. A. Designing national forest inventories for accurate estimation of soil carbon change. *Global Change Biology* **32**, e70868 (2026).
- [56] Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A. & Fernández-Ugalde, O. LUCAS Soil, the largest expandable soil dataset for Europe: a review. *European Journal of Soil Science* **69**, 140–153 (2018).
- [57] Heise, D. R. Separating reliability and stability in test-retest correlation. *American Sociological Review* **34**, 93–101 (1969).
- [58] Chen, S. *et al.* European topsoil bulk density and organic carbon stock database (0–20 cm) using machine-learning-based pedotransfer functions. *Earth System Science Data* **16**, 2367–2383 (2024).
- [59] Panagos, P. *et al.* Soil bulk density assessment in Europe. *Agriculture, Ecosystems & Environment* **364**, 108907 (2024).
- [60] Ecosystem Marketplace. State of the voluntary carbon market 2025: Meeting the moment, renewing trust in carbon finance. Tech. Rep., Forest Trends Association, Washington, DC (2025).
- [61] European Securities and Markets Authority. EU carbon markets 2026. ESMA Market Report ESMA50-481369926-34622, ESMA, Paris (2026).
- [62] Wulder, M. A. *et al.* Fifty years of Landsat science and impacts. *Remote Sensing of Environment* **280**, 113195 (2022).
- [63] Gorelick, N. *et al.* Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* **202**, 18–27 (2017).
- [64] Roy, D. P. *et al.* Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. *Remote Sensing of Environment* **185**, 57–70 (2016).
- [65] Muñoz-Sabater, J. *et al.* ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data* **13**, 4349–4383 (2021).
- [66] Bai, S., Kolter, J. Z. & Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling (2018). ArXiv:1803.01271.

- [67] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Computation* **9**, 1735–1780 (1997).
- [68] Cho, K. *et al.* Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734 (2014).
- [69] Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 6402–6413 (2017). ArXiv:1612.01474.
- [70] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization (2015). International Conference on Learning Representations (ICLR); arXiv:1412.6980.
- [71] Peel, M. C., Finlayson, B. L. & McMahon, T. A. Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences* **11**, 1633–1644 (2007).

Funding

No funding was received for this work.

Author contributions

MRH conceived the study, developed the methodology and software, performed the analysis, and wrote the manuscript. MAKR contributed to validation and to reviewing and editing the manuscript. ASMM provided supervision and reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Tables

Table 1: Paired differences in skill with 95 % block-bootstrap intervals. Bracketed numbers give the replicates pooled, other rows use the single assignment; share > 0 is the share of bootstrap draws with a positive difference.

Comparison	Difference	95 % interval	Share > 0
<i>Spatial blocking</i>			
Full model minus gradient boosting, with previous value (5)	0.003	−0.011 to 0.019	0.63
Full model minus mean-reversion floor (5)	0.066	0.039 to 0.089	1.00
No recurrent decoder minus full model (3)	0.007	−0.002 to 0.016	0.95
Full model minus gradient boosting, no previous value	0.051	0.028 to 0.075	1.00
Full model minus series replaced by zeros	0.033	0.008 to 0.052	1.00
<i>Spatial blocking, with ERA5-Land climate</i>			
Full model minus gradient boosting, with previous value (5)	0.003	−0.021 to 0.054	0.38
Full model minus gradient boosting, tuned (5)	0.007	−0.017 to 0.029	0.76
<i>Later date, same points</i>			
Full model minus gradient boosting, with previous value	0.064	0.036 to 0.089	1.00
<i>Later date, new places</i>			
Full model minus gradient boosting, with previous value (5)	0.105	0.021 to 0.190	1.00
Full model minus mean-reversion floor (5)	0.083	0.041 to 0.108	1.00

Table 2: Signal and noise in repeated LUCAS measurements, mineral points with three campaigns, with 95 % block-bootstrap intervals. Noise share is the lower bound $2\sigma_e^2/\Delta^2$ on the share of mean squared change that is measurement error; ceiling from Eq. (1) on the changes at these points, which gives 0.27 where all 16,496 forecast targets give 0.28 (Sect. 2.2).

	Points	$r(\Delta_1, \Delta_2)$	Noise share, 6 years	Ceiling
Log SOC, all points	8,221	−0.54	0.94 (0.76 to 1.05)	0.27 (0.23 to 0.35)
<i>Strata</i>				
Cropland in 2015	5,719	−0.59	1.01 (0.78 to 1.13)	0.24 (0.20 to 0.33)
Grassland in 2015	2,167	−0.39	0.72 (0.60 to 0.84)	0.38 (0.34 to 0.43)
Same land-cover class in 2018	6,848	−0.54	0.93 (0.73 to 1.05)	0.28 (0.23 to 0.35)
Other land-cover class in 2018	1,373	−0.55	1.04 (0.89 to 1.13)	0.26 (0.24 to 0.32)
Clay, lowest third	2,735	−0.41	0.70 (0.56 to 0.78)	0.37 (0.33 to 0.44)
Clay, highest third	2,775	−0.63	1.12 (0.85 to 1.25)	0.21 (0.17 to 0.30)
2009 SOC, lowest third	2,733	−0.55	0.95 (0.74 to 1.07)	0.27 (0.25 to 0.36)
2009 SOC, middle third	2,723	−0.64	1.21 (0.85 to 1.41)	0.21 (0.15 to 0.34)
2009 SOC, highest third	2,765	−0.48	0.72 (0.61 to 0.84)	0.34 (0.29 to 0.38)
<i>Control properties, same samples</i>				
pH (CaCl ₂)	8,221	−0.44	0.73 (0.70 to 0.76)	0.35 (0.33 to 0.36)
Log total nitrogen	8,221	−0.41	0.64 (0.56 to 0.75)	0.38 (0.34 to 0.42)

Table 3: Measurement noise in carbon-crediting terms, mineral soils, per hectare and 0–20 cm, at the median stock of 35 t C ha⁻¹ with fine-earth bulk density. Values per hectare at US\$7.66 (voluntary agriculture credits, 2024) and EUR 74 (EU allowances, 2025) per t CO₂; 95 % block-bootstrap intervals; R^2 ranges are across replicates. The upper bulk-density case is in the Supplementary Information.

Quantity	Estimate	t CO ₂ ha ⁻¹	US\$ / EUR ha ⁻¹
<i>One six-year cycle, no true change</i>			
Apparent change from noise, one location	37 % (27 to 48)	47 (38 to 53)	358 / 3,460
Gain under the 4 per 1000 target	2.4 %	3.1 (2.5 to 3.6)	24 / 227
<i>Detectable change, 5 % one-sided test, 80 % power</i>			
One location	202 % (153 to 270)	256 (209 to 294)	1,960 / 18,933
Ten locations	42 % (34 to 51)	53 (46 to 58)	406 / 3,922
A hundred locations	12 % (10 to 14)	15 (13 to 16)	113 / 1,096
Locations to detect the 4 per 1000 gain	2,127 (1,503 to 2,982)	–	–
<i>Carbon farming methodology, 70 % one-sided confidence</i>			
Locations required (its equation 8)	193 (136 to 271)	–	–
No-change project, probability of credit	0.30	–	–
No-change project, expected credit	25 % of gain	0.77 (0.63 to 0.91)	6 / 57
<i>Enrolment on a low baseline, lowest third</i>			
Reversion gain over the next interval	14 % (8 to 21)	10.4 (6.6 to 12.2)	80 / 769
<i>Model test, R^2 on change (its equation 7)</i>			
Mean-reversion floor	0.088 (0.085 to 0.094)	–	–
Gradient boosting	0.205 (0.188 to 0.214)	–	–
Hybrid	0.210 (0.202 to 0.216)	–	–

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.pdf](#)