

Supplementary Information for “Measurement noise limits soil carbon forecasting and crediting across European monitoring points”

Md Rakib Hasan, Mst Anika Khatun Rupa and A. S. M. Mohiuddin

S1 The recurrent decoder

The main text treats the recurrent cell as an interchangeable component. This section gives the two cells and the convolution they read from, drawn in Figs. S1 to S4.

The fused sequence enters a gated recurrent layer whose cell is interchangeable. The long short-term memory cell (Hochreiter and Schmidhuber, 1997), drawn in Fig. S3, maintains a separate cell state c_t updated additively,

$$\begin{aligned} f_t &= \sigma(W_f[h_{t-1} \parallel u_t] + b_f), & i_t &= \sigma(W_i[h_{t-1} \parallel u_t] + b_i), \\ \tilde{c}_t &= \tanh(W_c[h_{t-1} \parallel u_t] + b_c), & c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \\ o_t &= \sigma(W_o[h_{t-1} \parallel u_t] + b_o), & h_t &= o_t \odot \tanh(c_t). \end{aligned} \quad (\text{S1})$$

The gated recurrent unit (Cho et al., 2014), drawn in Fig. S4, merges the forget and input roles into one update gate,

$$\begin{aligned} r_t &= \sigma(W_r[h_{t-1} \parallel u_t] + b_r), & u_t^g &= \sigma(W_g[h_{t-1} \parallel u_t] + b_g), \\ \tilde{h}_t &= \tanh(W_h[r_t \odot h_{t-1} \parallel u_t] + b_h), & h_t &= (1 - u_t^g) \odot h_{t-1} + u_t^g \odot \tilde{h}_t. \end{aligned} \quad (\text{S2})$$

Both cells solve the same problem and solve it the same way. A plain recurrent layer propagates gradient by repeated multiplication with one weight matrix (Fig. S2), so information decays or explodes geometrically with lag, whereas Eqs. (S1) and (S2) both route the state through an additive path whose scaling is learned per step. That difference makes a carbon input from three seasons ago recoverable at all.

The choice between them is an empirical question that the soil literature has not settled, and there is a specific reason it might matter here. A GRU carries three weight matrices where an LSTM carries four, so it holds roughly a quarter fewer parameters at a given hidden width, while performing comparably to the LSTM on the sequence tasks where the two were first compared (Chung et al., 2014). Whether that economy pays off on training sets as small as those available for soil carbon is untested, and holding the encoder, the fusion layer and the head identical while swapping only the cell makes it testable.

S2 Organic soils

The organic panel holds 202 observations from 128 points, classed by a previous carbon value above 120 g kg^{-1} , and is validated under three spatially blocked folds of the same block size as the mineral panel, which gives 12 blocks. No model beats persistence. The best variant, the hybrid without its learned encoder, reaches -0.017 , the hybrid with a GRU -0.034 , the full

hybrid -0.049 , the hybrid without its recurrent decoder -0.054 and gradient boosting without the previous value -0.058 , while withholding the previous observation or the static embedding lowers skill to -0.104 and -0.153 . Persistence error is 0.677 here against 0.427 on mineral soils, so there is more change to explain, but with 202 observations in three folds no model learns enough of it to improve on assuming no change. Random splitting gives -0.046 , close to the blocked -0.049 , and the whole spread of these results, 0.14 , rests on about eighty times fewer observations than the mineral panel. No bootstrap intervals were computed on this subset. The graph variants of Section S3 reach -0.035 with geographic and -0.027 with covariate-space neighbours, still below persistence.

S3 Graph extension

Two graph convolution layers were placed on the hybrid’s learned representation, mixing each observation with its eight nearest neighbours, defined geographically or by covariate similarity, with the configurations selected for the hybrid in each fold and three-member ensembles. Every point appears twice and the 2018 observation carries the 2015 target as its previous value, so edges between observations of the same point and edges from a later campaign to an earlier one were removed. Both observations of a point share a spatial block, and blocking alone does not prevent that leak. The graph model also trains on all observations at once rather than in batches, so a control with the same model and training but every edge removed separates the effect of the neighbours from that of the training.

On mineral soils the control reaches 0.101 against 0.111 for the hybrid, a difference of -0.010 (-0.022 to 0.005), so the full-batch training regime costs a little skill if anything. Geographic neighbours raise skill to 0.122 , a gain over the control of 0.021 (0.016 to 0.024) on the same folds and block bootstrap as the main comparison, which is 0.011 (0.001 to 0.023) above the hybrid and 0.022 (0.015 to 0.027) above boosting. Covariate-space neighbours reach 0.110 , from a single run without an interval.

Each observation’s neighbours carry their own previous carbon values and covariates, all known when the forecast is made, and giving the same information to the tabular models locates the gain. Over the same edges, each observation received the mean previous log carbon and the mean static properties of its neighbours. Over the five replicates of the main comparison these features raise the mean-reversion floor by 0.029 (0.011 to 0.041) and boosting by 0.010 (0.004 to 0.014) with fixed settings and by 0.012 (0.007 to 0.020) with nested selection. Boosting given them recovers about half of the graph’s lead, with the graph still 0.011 above it on the single assignment of folds, while boosting with neighbour features and the hybrid remain level at a difference of 0.007 (-0.010 to 0.021).

S4 Calibration and validation optimism

Prediction intervals are close to calibrated on mineral soils, with nominal 90 % intervals covering 0.896 to 0.924 of held-out observations. Random splitting raises the hybrid’s skill from 0.111 to 0.134 , an optimism of 2.6 % in RMSE, which confirms that the blocking is doing its work while showing that spatial leakage does not separate the methods here.

S7 Measurement noise in carbon-crediting terms

The crediting quantities of the main text are given here for both bulk-density cases. The central case uses the LUCAS fine-earth means of Chen et al. by land cover (1.16 , 1.00 and 1.02 g cm^{-3}

for cropland, grassland and other cover), and the upper case applies the measured cropland mean of Panagos et al. (1.26 g cm^{-3}) to every point. Stocks are for 0–20 cm without a coarse-fragment correction. Values in tonnes and currency scale linearly with bulk density and depth, whereas the shares, ratios and location counts do not depend on either: noise is 15 times (11 to 20) the six-year gain of the 4 per 1000 target in both cases, detection at 5 % with 80 % power needs 2,127 (1,503 to 2,982) locations, of which noise alone accounts for 2,013 (1,167 to 3,119), and the methodology’s equation (8) gives 193 (136 to 271) locations with a tolerable error of 0.7 times the gain and 95 (67 to 133) with a tolerable error equal to it. The probability that a project with no true change is credited, 0.30, and its expected credit as a share of the target gain, 0.25, follow from the 70 % confidence level alone and carry no sampling interval.

Table 1: Measurement noise in carbon-crediting terms under two bulk-density cases, mineral soils, per hectare, 0–20 cm, one six-year cycle. 95 % block-bootstrap intervals.

Quantity	Fine earth by land cover	1.26 g cm^{-3} everywhere
Median stock, t C ha^{-1}	34.6 (28.3 to 40.8)	38.6 (32.2 to 46.1)
Apparent change from noise, $\text{t CO}_2 \text{ ha}^{-1}$	46.8 (37.7 to 52.6)	52.1 (42.0 to 59.2)
Gain under 4 per 1000, six years, $\text{t CO}_2 \text{ ha}^{-1}$	3.07 (2.52 to 3.63)	3.43 (2.87 to 4.10)
Detectable change, 10 locations, $\text{t CO}_2 \text{ ha}^{-1}$	53.0 (46.2 to 58.3)	59.1 (51.5 to 65.8)
Detectable change, 100 locations, $\text{t CO}_2 \text{ ha}^{-1}$	14.8 (13.1 to 16.2)	16.5 (14.6 to 18.4)
No-change project, expected credit, $\text{t CO}_2 \text{ ha}^{-1}$	0.77 (0.63 to 0.91)	0.86 (0.72 to 1.03)
Reversion gain, lowest third, $\text{t CO}_2 \text{ ha}^{-1}$	10.4 (6.6 to 12.2)	11.5 (7.4 to 13.8)
Apparent change from noise, $\text{US\$ ha}^{-1}$	358 (288 to 403)	399 (322 to 454)
Apparent change from noise, EUR ha^{-1}	3,460 (2,786 to 3,891)	3,859 (3,108 to 4,384)
Reversion gain, $\text{US\$ ha}^{-1}$	80 (51 to 94)	88 (57 to 106)
Reversion gain, EUR ha^{-1}	769 (492 to 906)	855 (549 to 1,020)

S5 Hyperparameter selection

The nested search gives no sign that a different configuration would change the ordering. Across the twenty searches of the five spatial replicates, nine of the twelve configurations were selected at least once and none more than five times, and within a search the best and worst configurations differed by 0.014 to 0.099 on the inner held-out blocks. No width, learning rate or dropout is consistently better at this data density, as expected of a network with little further to fit. Boosting ran with fixed settings in the protocol comparison, and giving it the same nested selection lowers it across space rather than raising it (main text).

S6 Mean reversion, subgroups and spatial pattern

Part of what both methods predict needs no covariates at all (Table S5). On the single assignment of folds, a linear fit of the new value on the previous one reaches a skill of 0.042, and adding a term for the campaign raises it to 0.046, about two fifths of the 0.100 reached by boosting and of the 0.111 reached by the hybrid. The fit predicts that high values fall and low values rise, which is regression towards the mean, and the main text shows that a noisy repeated measurement produces it even where nothing has changed. Campaign-level shifts carry almost nothing, with a shift for each campaign scoring 0.006 and a shift for each campaign and land use 0.007, so the models are not learning an offset between surveys. Split by the previous concentration, both methods reach 0.19 to 0.20 in the lowest third, 0.05 to 0.06 in the highest and only about 0.02 in the middle third, where there is least room to revert.

No subgroup separates the two methods. The hybrid is ahead by 0.006 to 0.013 in every

land-use and concentration class, ahead by 0.004 to 0.056 in three of the four climate zones and behind by 0.008 in the temperate zone, and every interval includes zero. Skill is highest where there is most change to forecast, in the Mediterranean and arid zones where persistence errs most, and lowest in the continental and boreal zone, where it errs least. The arid zone holds 624 observations in five blocks, so its interval is wide.

The measured change itself differs by land use and by interval. On cropland the median change was -0.3 g kg^{-1} (-2.8%) between 2009 and 2015 and $+0.6 \text{ g kg}^{-1}$ ($+4.9\%$) between 2015 and 2018, with the share of points losing carbon falling from 54 to 39 %. On grassland the median change was $+0.9$ and $+1.1 \text{ g kg}^{-1}$, $+4.5$ and $+5.9\%$, with 43 and 41 % of points losing. A cropland loss over six years followed by a larger gain over three matches the negative correlation between successive changes of the main text better than a change in soil stocks.

In space the predictions do not reproduce regional differences in change (Fig. 3 of the main text). Averaged over the 354 cells of 100 km that hold at least ten observations, predicted and observed change correlate at only $r = 0.11$. The observed mean change in log carbon itself varies little with latitude, from 0.021 to 0.026 across bands south of 45°N , between 45 and 52°N and north of 52°N . Point by point the predictions are compressed, with a fitted slope of 0.24 against observed change, as expected when most of the observed change is noise that no forecast can follow.

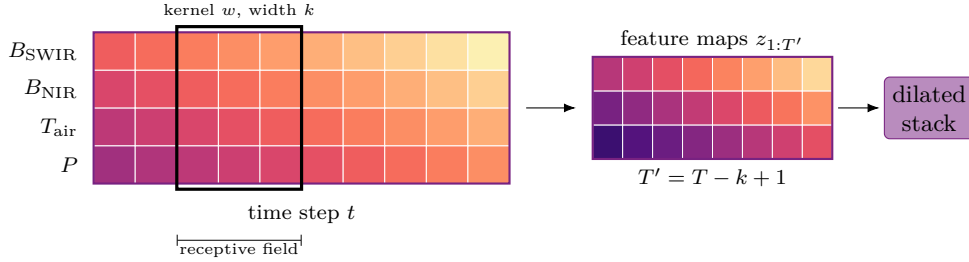


Figure S1: One-dimensional convolution applied across a multivariate covariate sequence. Each row is one covariate channel, drawn here as two reflectance bands and two climate drivers to show the general design, and each column is one time step. A kernel of width k spans every channel at once and slides along the time axis, so a single filter responds to a joint pattern such as a wet-season reflectance drop that coincides with a rainfall peak. Stacking dilated layers widens the receptive field geometrically while the parameter count grows only linearly, which is what allows a season of context to be summarised without a separate weight for every lag.

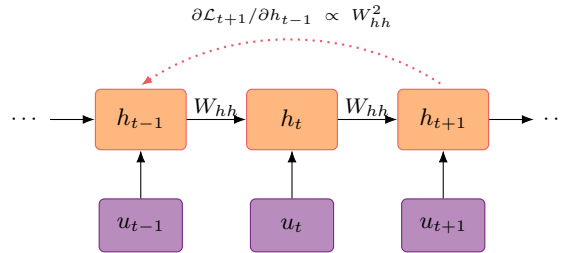


Figure S2: A recurrent layer unrolled over three steps, with u_t the fused input of Eq. (2) of the main text. The same weight matrix W_{hh} is reused at every step, so the hidden state h_t carries information forward indefinitely in principle. The dotted path shows why it does not in practice. A gradient reaching back n steps is multiplied by W_{hh} n times, so it decays geometrically whenever the leading eigenvalue of W_{hh} is below one and explodes when it is above. Gated cells (Figs. S3 and S4) replace this repeated multiplication with an additive path.

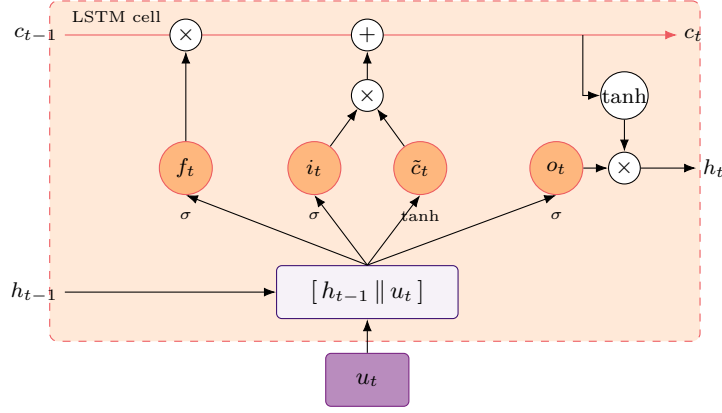


Figure S3: The long short-term memory cell (Hochreiter and Schmidhuber, 1997). The heavy horizontal line is the cell state, which is modified only by an elementwise multiplication and an addition, so a gradient travelling along it is scaled by the forget gate instead of by a weight matrix. Holding f_t near one therefore preserves information across many steps without the decay shown in Fig. S2. Each gate reads the same concatenation of the previous hidden state and the current input, with its activation shown beneath it.

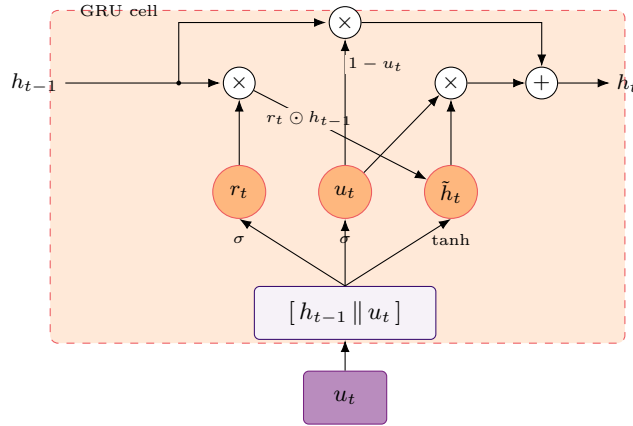


Figure S4: The gated recurrent unit (Cho et al., 2014). A single update gate u_t performs both jobs the LSTM splits between its forget and input gates, interpolating between the previous state and the candidate so the two weights sum to one by construction. The cell also carries no separate memory line, which removes one state vector and one gate. The result is roughly a quarter fewer parameters for a given hidden width, which may help where training data are scarce, a possibility ablation 3 of the main text tests directly.

Supplementary tables

Table S1: Skill against persistence under spatially blocked validation, mineral soils ($n = 16,496$), single assignment of blocks to folds. RMSE on log SOC; coverage of nominal 90 % intervals.

Model	Skill	RMSE	Coverage
Ablation 2, no learned encoder	0.117	0.377	0.917
Ablation 1, no recurrent decoder	0.115	0.377	0.909
Ablation 3, GRU in place of LSTM	0.115	0.378	0.924
Full model, LSTM	0.111	0.379	0.911
Gradient boosting, with previous value	0.100	0.384	–
Ablation 5, no previous observation	0.060	0.401	0.924
Gradient boosting, no previous value	0.060	0.401	–
Ablation 4, no static embedding	−0.293	0.552	0.896
Persistence	0.000	0.427	–
Full model under random k -fold (optimism reference)	0.134	0.370	0.921

Table S2: Contribution of the dynamic covariates, mineral soils, spatially blocked validation. CCC on log concentration.

Model	Skill	RMSE	CCC
Hybrid, real covariate series	0.111	0.379	0.809
Hybrid, series replaced by noise	0.081	0.392	0.786
Hybrid, series replaced by zeros	0.078	0.394	0.780
Gradient boosting, static properties and previous value only	0.064	0.400	0.781

Table S3: Forward temporal validation, mineral soils, training on 2015 targets and testing on 2018 targets. –, not computed.

Model	Skill	RMSE	CCC
<i>Same points</i>			
Full model, LSTM	0.143	0.342	0.846
Gradient boosting, with previous value	0.078	0.367	0.826
Gradient boosting, no previous value	0.055	–	–
Persistence	0.000	0.399	0.822
<i>New places</i>			
Full model, LSTM	0.139	–	–
Mean-reversion floor	0.056	–	–
Gradient boosting, with previous value	0.034	–	–

References

- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). 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)*, pages 1724–1734.
- Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Table S4: Stronger boosting baselines against the hybrid, mineral soils, with 95 % block-bootstrap intervals on the difference.

Model	Skill	Hybrid minus boosting
<i>Across space, spatial blocking, five replicates</i>		
Hybrid	0.113	
Boosting, summary statistics, fixed (protocol)	0.110	0.003 (−0.011 to 0.019)
Boosting, summary statistics, tuned	0.099	0.014 (−0.000 to 0.030)
Boosting, timing features, fixed	0.115	−0.003 (−0.021 to 0.026)
Boosting, timing features, tuned	0.102	0.010 (−0.009 to 0.029)
Mean-reversion floor	0.047	
<i>Later survey, new places, five replicates</i>		
Hybrid	0.139	
Boosting, summary statistics, fixed (protocol)	0.034	0.105 (0.021 to 0.190)
Boosting, summary statistics, tuned	0.119	0.020 (−0.003 to 0.042)
Boosting, timing features, fixed	0.049	0.090 (0.007 to 0.168)
Boosting, timing features, tuned	0.114	0.025 (−0.004 to 0.066)
Mean-reversion floor	0.056	
<i>Later survey, same points, single split</i>		
Hybrid	0.143	
Boosting, summary statistics, fixed (protocol)	0.078	0.064 (0.036 to 0.089)
Boosting, summary statistics, tuned	0.139	0.004 (−0.004 to 0.017)
Boosting, timing features, fixed	0.118	0.025 (0.013 to 0.036)
Boosting, timing features, tuned	0.142	0.001 (−0.011 to 0.022)

Table S5: Skill by subgroup, mineral soils, spatially blocked validation, with 95 % block-bootstrap intervals. Persistence RMSE in brackets.

	<i>n</i>	Hybrid	Boosting	Hybrid minus boosting
<i>Simple baselines, skill (95 % interval)</i>				
Shift for each campaign	16,496	0.006	0.003	0.008
Shift for each campaign and land use	16,496	0.007	0.003	0.010
Linear on previous value	16,496	0.042	0.026	0.052
Linear on previous value and campaign	16,496	0.046	0.029	0.056
<i>By land use</i>				
Cropland	11,271	0.127	0.115	0.012 (−0.011 to 0.028)
Grassland	4,495	0.064	0.055	0.009 (−0.003 to 0.023)
<i>By previous concentration</i>				
Lowest third, $\leq 12.0 \text{ g kg}^{-1}$	5,541	0.199	0.186	0.013 (−0.021 to 0.030)
Middle third	5,470	0.023	0.017	0.006 (−0.007 to 0.017)
Highest third, $> 20.1 \text{ g kg}^{-1}$	5,485	0.058	0.047	0.012 (−0.004 to 0.030)
<i>By climate zone (persistence RMSE)</i>				
Mediterranean, Cs (0.53)	4,298	0.149	0.128	0.021 (−0.010 to 0.028)
Temperate, Cf and Cw (0.38)	6,801	0.091	0.100	−0.008 (−0.017 to 0.006)
Continental and boreal, D (0.34)	4,773	0.046	0.042	0.004 (−0.006 to 0.016)
Arid and semi-arid, B (0.63)	624	0.160	0.103	0.056 (−0.168 to 0.085)

Table S6: Organic soils (previous SOC above 120 g kg^{-1} ; $n = 202$, three spatially blocked folds).

Model	Skill	RMSE	Coverage
No learned encoder	−0.017	0.689	0.856
GRU in place of LSTM	−0.034	0.700	0.876
Full model, LSTM	−0.049	0.710	0.886
No recurrent decoder	−0.054	0.714	0.896
Gradient boosting, no previous value	−0.058	0.716	–
No previous observation	−0.104	0.747	0.911
No static embedding	−0.153	0.781	0.901
Persistence	0.000	0.677	–