Supplementary Material for:
Geochemical predictability quantification and spatial transferability of soil organic carbon in Ganges Delta

Md Rakib Hasan[1], Fazla Zawadul Arabi[1], Mst Anika Khatun Rupa[1], Shoumik Zubyer[1], Mohammad Fazle Alam Rabbi[1], and Md Jashim Uddin*[1]

[1]*Department of Soil, Water and Environment, University of Dhaka, Dhaka-1000, Bangladesh*

Supplementary Tables

S1. Full feature list

Table S1: Complete feature list (43 features). OM is excluded from all models due to data leakage ($r = 0.996$ with SOC).

Category	Features	Count
Soil properties	EC, CEC, TN, pH, BD, Moisture, Clay, Sand, Silt	9
Spatial/topographic	Latitude, Longitude, Elevation, Slope, Aspect	5
Interaction terms	EC×CEC, Clay×Moisture, pH×EC, EC×Moisture, CEC×pH, Clay×CEC, TN×EC, TN×CEC, Moisture×BD	9
Ratio features	TN/EC, EC/CEC, Sand/Clay, Clay/Silt, Moisture/BD, TN/Clay, EC/Moisture, CEC/Clay	8
Polynomial terms	EC^2, EC^3, TN^2, CEC^2, $Moisture^2$, $Clay^2$, pH^2, BD^2	8
Log transforms	log(EC+1), log(CEC+1), log(TN+1), log(Moisture+1)	4
Total		**43**

S2. Descriptive statistics by zone

Table S2 summarizes the mean and standard deviation of key soil properties across the four geomorphological zones. Tidal Active zones carried the highest SOC, EC, CEC, moisture and clay content, consistent with mangrove-influenced settings where salinity-suppressed decomposition promotes carbon accumulation. The Mature Delta showed the lowest SOC and the highest bulk density, consistent with long-cultivated, well-drained floodplain soils. Values are reproduced by `zone_descriptives.py` in the code repository.

*Corresponding author: juswe@du.ac.bd

Table S2: Descriptive statistics of key soil properties by geomorphological zone (mean ± SD).

Property	Tidal Active ($n = 25$)	Active Delta ($n = 25$)	Mature Delta ($n = 25$)	Moribund Delta ($n = 25$)
SOC (t C ha^{-1})	127.3 ± 78.7	28.5 ± 15.2	8.4 ± 8.6	16.2 ± 10.5
EC (dS m^{-1})	4.72 ± 1.58	0.39 ± 0.39	0.12 ± 0.04	0.13 ± 0.04
CEC ($cmol_c$ kg^{-1})	28.4 ± 2.0	19.2 ± 4.3	4.2 ± 3.2	2.9 ± 2.4
TN (%)	0.10 ± 0.02	0.14 ± 0.06	0.09 ± 0.08	0.04 ± 0.03
pH	8.43 ± 0.26	7.82 ± 0.52	8.32 ± 0.38	8.23 ± 0.25
BD (g cm^{-3})	1.43 ± 0.11	1.45 ± 0.07	1.62 ± 0.12	1.59 ± 0.09
Moisture (%)	37.8 ± 4.2	30.0 ± 5.6	27.5 ± 6.9	27.6 ± 5.3
Clay (%)	41.9 ± 7.7	30.3 ± 15.5	27.2 ± 8.2	27.4 ± 7.6

S3. Hyperparameter configurations

Table S3 lists the hyperparameter settings for all models. The gradient-boosting regressor was run at scikit-learn defaults, without hyperparameter search. This is deliberate: at $n = 100$ with 20 spatial groups, a search over folds would itself risk the optimism that this study sets out to quantify, and the reported leak-free $R^2 = 0.64$ is therefore obtained by an untuned model. Deep-learning architecture parameters were set manually from the Methods description, and all deep models used early stopping with a patience of 30 epochs.

Table S3: Hyperparameter configurations. Gradient boosting is untuned; the deep-learning settings describe the reference re-implementation (`dl_reconstruction.py`), since the original training code was not preserved.

Model	Parameter	Value	Selection
GBR	n_estimators	100	scikit-learn default
GBR	max_depth	3	scikit-learn default
GBR	learning_rate	0.1	scikit-learn default
GBR	subsample	1.0	scikit-learn default
GNN	Hidden dim	16 per head (64 concatenated)	Manual
GNN	Layers	3	Manual
GNN	Attention heads	4	Manual
GNN	Dropout	0.3	Manual
GNN	Learning rate	0.001	Scheduler
MoE	Experts	4	Manual
MoE	Top-k	2	Manual
MoE	Expert hidden	[64, 32]	Manual
MoE	Load balance λ	0.01	Manual
ViT	Tokens	3 + CLS	Architecture
ViT	Layers	3	Manual
ViT	Heads	4	Manual
ViT	Embed dim	64	Manual
All DL	Epochs	200	Early stopping
All DL	Batch size	32	Fixed

S4. Analytical methods

Table S4 details the laboratory analytical methods used for each soil physicochemical property. All analyses followed standard protocols; quality assurance included duplicate analysis (every 10th sample) and certified reference material (ISE 921, Wageningen), with analytical precision for SOC of ±5% CV.

Table S4: Analytical methods for soil physicochemical properties.

Property	Method	Reference
SOC	Walkley–Black wet oxidation (CF = 1.32)	Walkley and Black (1934)
Organic matter	Walkley–Black and Van Bemmelen	Walkley and Black (1934)
Bulk density	Core method (100 cm^3 cylinders)	—
pH	1:2.5 soil:water, glass electrode (Hanna HI 2210)	—
EC	1:5 soil:water extract (Hanna HI 8633)	—
CEC	NH_4OAc at pH 7.0	Chapman (1965)
TN	Micro-Kjeldahl	Bremner (1996)
Particle size	Hydrometer method	Bouyoucos (1962)
Moisture	Gravimetric (105°C, 24 h)	—
Available P	Olsen ($NaHCO_3$ extraction)	—
Exchangeable K	NH_4OAc extraction, AAS	—

S5. Ablation study

Table S5 presents the gradient-boosting feature-group ablation under leak-free location-grouped cross-validation, isolating each component's contribution to predictive skill. The H1-informed engineered features add the largest reproducible gain ($\Delta R^2 = +0.10$ over the base feature set and +0.11 over a soil-only model), whereas the spatial/topographic covariates and EC alone are progressively weaker in isolation. These values are reproduced by `engineered_ablation.py` and `ml_spatial_cv.py` in the data release and match the ablation panel of main-text Fig. 4d.

Table S5: Feature-group ablation under leak-free location-grouped 5-fold cross-validation (pooled out-of-fold R^2; gradient boosting; OM excluded). "Base" is the 14 soil and spatial features; "+ engineered" adds the H1-informed engineered features. Reproduced by `engineered_ablation.py` and `ml_spatial_cv.py`.

Condition	Location-grouped R^2	Δ vs base
Base + engineered features	0.74	+0.10
Base features (14)	0.64	—
Soil-only (9 features)	0.62	−0.02
Spatial-only (5 features)	0.39	−0.25
EC-only baseline (1 feature)	0.11	−0.53

S6. Cross-delta transfer panel

Table S6 lists the 14 external deltas used in the transferability analysis (H4), primarily WoSIS 2023 (Batjes et al., 2024), with the BIS Nederland national soil survey (4TU.ResearchData, 2020; CC BY 4.0) augmenting Rhine-Meuse (504 WoSIS profiles + 2,148 BIS profiles after SOM→SOC conversion via the van Bemmelen factor of 0.58, rescaling of percentages to g kg^{-1}, RD-New→WGS84 coordinate transformation, and removal of BIS profiles within ~200 m of an existing WoSIS profile to avoid double-counting). Within-delta skill is positive where sample size is adequate, yet cross-delta transfer of these models is near-uniformly negative (Fig. 5, main text): of 210 train-on-one/test-on-another delta pairs only 6% exceed $R^2 = 0$ (median -0.55), and Leave-One-Delta-Out CV is negative for 11 of 14 deltas. The tropical Asian monsoon deltas most analogous to the Ganges (Mekong, Irrawaddy, Godavari, Krishna) hold WoSIS organic-carbon values but lack co-measured texture and pH, and so could not enter the feature-based transfer test.

Table S6: Cross-delta transfer panel (H4). Bounding box, complete-case top 1 m profiles, and within-delta 5-fold CV R^2 (pH + sand/silt/clay predictors). All profiles from WoSIS 2023 (Batjes et al., 2024) except where noted.

Delta	Lat range	Lon range	n	Within-delta R^2
Mississippi	28–33 N	93.5–88 W	2043	−0.24
Rhine-Meuse[†]	50.5–53.5 N	3–7.5 E	2652	0.19
Niger	4–6.5 N	4.5–7.5 E	327	0.25
Yellow River	34–40 N	116–122 E	315	−0.23
Limpopo	24.5–25.5 S	33–34.2 E	268	0.37
Chao Phraya	13–15 N	100–101 E	204	0.10
Paraná–Plata	32–35 S	59–57 W	179	0.62
Sacramento	37.5–38.5 N	122.3–121.3 W	178	0.84
Indus	23–25.5 N	66.5–69 E	167	0.73
Mackenzie	67.5–70.5 N	137–132 W	77	−0.10
Danube	44.3–45.6 N	28–30.2 E	68	0.79
Amazon	2.5 S–1.5 N	52–48 W	62	−0.32
Senegal	15.5–16.5 N	16.7–15.8 W	57	0.99
Pearl	21.8–23.5 N	112.5–114.2 E	55	0.56

[†]Rhine-Meuse: 504 WoSIS + 2,148 BIS Nederland profiles (4TU.ResearchData, CC BY 4.0). Folding BIS in raises the Rhine-Meuse within-delta R^2 from +0.15 (WoSIS only) to +0.19, while leaving the cross-delta medians and percent-positive essentially unchanged (within-median +0.25, cross-median −0.55, 6.2% of 210 pairs $R^2 > 0$). The conclusion that cross-delta transfer fails is therefore not a data-volume artifact (see `h4_bis_robustness.py` in the archive).

S6b. Expanded H4 analyses: distance-decay, feature universality, few-shot recovery

These three diagnostics jointly establish that the cross-delta transfer failure (Fig. 5) is not the simple consequence of a smooth ecological gradient between deltas (panel a), nor of which predictors are relevant in each delta (panel b), but of the irreducibly delta-specific quantitative mapping between predictors and SOC, a mapping that pooled global data cannot substitute for, even when augmented with local calibration (panel c).

S7. Summary of all 25 sub-hypotheses

Table S7 provides the comprehensive summary of all 25 sub-hypotheses across the four research tracks, including raw p-values, Benjamini–Hochberg FDR-corrected q-values, and outcome status. Under Bonferroni correction ($\alpha_{\text{adj}} = 0.002$), 7 of 10 geochemical sub-hypotheses survived; no climate remote-sensing proxy yielded a robust association at the location level ($n = 20$), so the H2 assessment rests on the process model (H2.6; see main text §H2). H4 results are based on 14 external

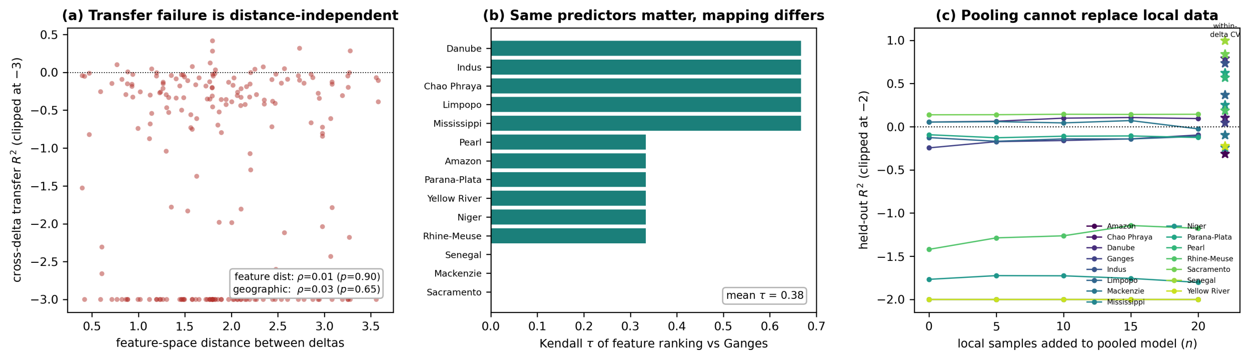

Figure S1: Expanded H4 analyses on the 15-delta panel. (a) Cross-delta transfer R^2 is uncorrelated with geographic, feature-space, or latitudinal distance between deltas (all Spearman $|\rho| \leq 0.11$, $p \geq 0.10$): transfer fails uniformly, not as a smooth function of environmental separation. (b) Per-delta GBR feature-importance rankings vs Ganges: all 14 external deltas show non-negative Kendall τ (mean +0.38); pH is the top predictor in 6 of 14 (matching Ganges). The same predictors are broadly relevant everywhere. (c) Few-shot recovery curves for every delta: pooled-other-deltas model plus n local samples vs held-out local R^2 (lines), with the within-delta CV ceiling marked by a star at right. Across all 15 deltas, the pooled+local model never reaches 50% of the within-delta ceiling by $n = 20$, and is positive in only 2 deltas; in deltas with strong within-delta skill (e.g. Senegal, Sacramento, Danube) pooling is actively harmful. Reproduced by `h4_expanded_analysis.py` in the archive.

deltas (6,652 harmonized soil profiles from WoSIS, with BIS Nederland augmenting Rhine-Meuse) plus the Ganges; see Table S6.

S8. Literature comparison

Table S8 compares the authoritative leak-free gradient-boosting performance reported here (location-grouped $R^2 \approx 0.64$) against published SOC prediction studies. Despite a substantially smaller sample size ($n = 100$), the model achieves competitive accuracy, underscoring the value of domain-informed feature engineering over brute-force data scaling. This comparison should be interpreted with caution given differences in spatial extent, target variables, and validation strategies across studies.

Supplementary Figures

S9. Sampling location coordinates

Table S9 lists the geographic coordinates and elevation of all 20 sampling locations, organized by geomorphological zone. Each location was sampled at five depth intervals (0–100 cm), yielding the 100 samples used in this study.

S10. Within-zone correlations

Table S10 reports soil property–SOC correlations computed separately within each geomorphological zone (Reviewer-requested per-zone analysis). Because the zones are distinct biogeochem-

Table S7: Summary of all 25 sub-hypotheses. Bonferroni threshold: $\alpha_{\text{adj}} = 0.05/25 = 0.002$. H4 results from 14 external deltas (6,652 harmonized soil profiles from WoSIS, with BIS Nederland augmenting Rhine-Meuse) plus the Ganges; see Table S6. †Bulk density is correlated against SOC concentration to avoid algebraic dependency with stock. ‡Random k-fold values are optimistic in this depth-replicated design; leak-free location-grouped CV gives $R^2 \approx 0.64$ (see main text, §H3).

ID	Hypothesis	Statistic	Raw p	BH q	Status
H1: Geochemical Controls					
H1.1	Clay mineralogy	$r = 0.323$	0.001	0.003	Significant (weak effect)
H1.2	Anoxia/Moisture	$r = 0.552$	<0.0001	<0.001	Confirmed
H1.3	pH nonlinearity	$R^2 = 0.108$	0.031	0.04	Weak
H1.4	Nutrient (TN)	$r = 0.045$	0.657	0.657	Not confirmed
H1.5	CEC aggregation	$r = 0.636$	<0.0001	<0.001	Confirmed
H1.6	Salinity (EC)	$r = 0.664$	<0.0001	<0.001	Confirmed
H1.7	Bulk density†	$r = -0.511$	<0.0001	<0.001	Confirmed
H1.8	Clay×Moisture	$\Delta R^2 = 0.068$	<0.05	0.042	Confirmed (BH)
H1.9	CEC×pH	$\Delta R^2 = 0.095$	<0.0001	<0.001	Confirmed
H1.10	EC×Moisture	$\Delta R^2 = 0.028$	0.021	0.035	Confirmed (BH)
H2: Climate Dynamics					
H2.1	Flooding (NDWI)	$r = 0.339$	0.143	0.179	Not confirmed
H2.2	Salinity (EC proxy)	$r = 0.271$	0.248	0.283	Not confirmed
H2.3	Monsoon intensity	$r = 0.401$	0.080	0.111	Not confirmed
H2.4	Vegetation change	$r = 0.252$	0.284	0.304	Not confirmed
H2.5	Temperature ($T_{\max}$)	proxy not robust	—	—	Not confirmed
H2.6	Management	modeled	—	—	Confirmed (modeled)
H3: ML/DL Prediction					
H3.1	ML $R^2 \geq 0.80$	random CV 0.91; grouped 0.64‡	—	—	Partial (optimistic)
H3.2	DL > ML	GBR ≥ all DL (leak-free)	—	—	Not supported
H3.3	Spatial generalization	LOLO $R^2 = 0.867$; grouped 0.64‡	—	—	Partial (LZO fails)
H3.4	Ensemble > individuals	ensemble ≈ GBR (no gain)	—	—	Not supported
H4: Cross-Delta Transfer (14 deltas + Ganges; WoSIS + BIS Nederland)					
H4.1	Zero-shot transfer skill	6% of 210 pairs $R^2 > 0$; median -0.55	—	—	Not transferable
H4.2	Forward Ganges → deltas	all 14 zero-shot $R^2 < 0$	—	—	Not transferable
H4.3	Reverse pooled → Ganges	$R^2 = -0.25$; few-shot stays <0 to n=20	—	—	Not transferable
H4.4	LODO (hold-out delta)	negative for 11/14 deltas	—	—	Not transferable
H4.5	Climate portability (G-RothC)	4.6–7.2× mgmt > climate (all deltas)	—	—	Confirmed

Table S8: Comparison of SOC prediction performance with published studies.

Study	Region	n	Best model	R^2
Were et al. (2015)	Kenya (tropical)	321	RF	0.73
Wadoux et al. (2021)	Global (review)	varies	Ensemble	0.60–0.85
Padarian et al. (2019)	Chile	2000+	CNN	0.68
Hengl et al. (2017)	Global (SoilGrids)	150k+	RF	0.56
Zhou et al. (2020)	China (subtropical)	689	XGBoost	0.78
Keskin et al. (2024)	Florida (subtropical)	284	RF+FE	0.81
This study (leak-free CV)	Ganges Delta	100	GBR	0.64

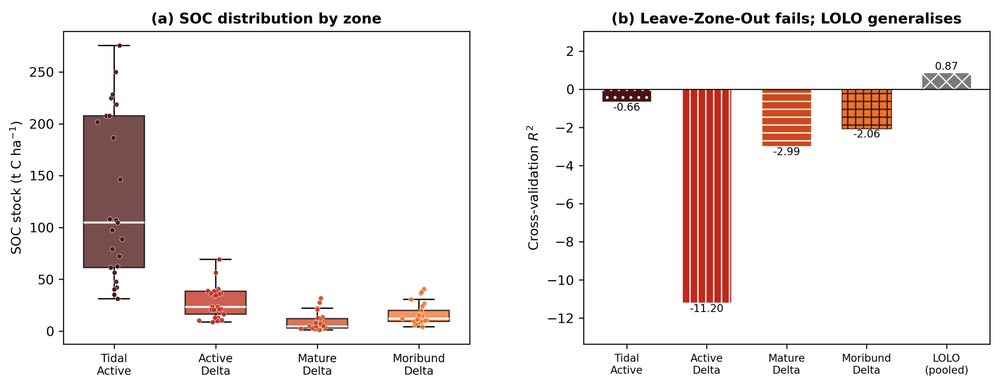

Figure S2: H3 spatial structure: (a) SOC distribution by geomorphological zone; (b) spatial cross-validation: LOLO passes ($R^2 = 0.867$) but LZO fails for all zones ($R^2 < 0$).

Table S9: Sampling locations with coordinates and zone assignments.

ID	Zone	Latitude (°N)	Longitude (°E)	Elevation (m)
L01	Tidal Active	21.82	89.18	2.1
L02	Tidal Active	21.95	89.32	1.8
L03	Tidal Active	22.08	89.05	2.5
L04	Tidal Active	21.75	89.45	1.5
L05	Tidal Active	22.15	89.22	3.2
L06	Active Delta	22.85	89.55	5.8
L07	Active Delta	23.02	89.72	6.2
L08	Active Delta	22.68	89.88	4.9
L09	Active Delta	23.15	89.42	7.1
L10	Active Delta	22.92	89.65	5.5
L11	Mature Delta	23.45	89.15	8.4
L12	Mature Delta	23.62	89.32	9.1
L13	Mature Delta	23.78	88.95	10.2
L14	Mature Delta	23.55	89.48	8.8
L15	Mature Delta	23.88	89.08	11.5
L16	Moribund Delta	24.12	88.75	14.2
L17	Moribund Delta	24.28	88.92	15.8
L18	Moribund Delta	24.05	88.58	13.5
L19	Moribund Delta	24.35	88.82	16.4
L20	Moribund Delta	24.18	88.65	14.9

ical settings, pooled correlations can mask, and even reverse, within-zone relationships (e.g. TN: $r = -0.79$ in Tidal Active versus +0.73 in Active Delta).

Table S10: Within-zone soil property–SOC correlations (Pearson r vs SOC concentration), computed separately for each geomorphological zone alongside the pooled value. $^*p < 0.05$; $n = 100$ pooled, $n = 25$ per zone. Reproduced by `per_zone_correlations.py`.

Predictor	Pooled	Tidal Active	Active Delta	Mature Delta	Moribund Delta
EC	+0.66*	−0.20	+0.34	+0.65*	+0.55*
CEC	+0.63*	−0.19	−0.14	+0.62*	+0.30
Moisture	+0.56*	+0.54*	−0.23	−0.11	+0.29
Clay	+0.33*	−0.30	−0.24	+0.47*	+0.30
pH	+0.25*	+0.61*	−0.29	−0.67*	−0.37
TN	+0.04	−0.79*	+0.73*	+0.35	+0.27

S11. Robustness to delta delineation: geomorphological polygons versus bounding boxes

The main-text panel delineates each external delta by a rectangular bounding box (Table S6), which is simple and reproducible but can admit non-deltaic upland soils. As a robustness check, the entire external panel was re-extracted using the published active-delta polygons of Nienhuis et al. (2020) (100 largest global deltas), buffered outward by 0.5° to capture the immediate delta margin, and re-ran the full transfer analysis. Two of the fourteen deltas required special handling: Senegal is not among the 100 mapped deltas, so a 0.5°-buffered version of its bounding box was retained; and the Yellow River polygon retained only $n = 9$ complete-case profiles, below the threshold for within-delta cross-validation. The polygon delineation is far stricter, retaining 1,449 of the original 6,652 profiles and a substantially more balanced panel (Table S11). The cross-delta conclusion is unchanged and marginally stronger: cross-delta median $R^2 = -0.66$ (versus −0.55), with 5.2% of 210 pairs positive (versus 6%), reverse pooled-to-Ganges $R^2 = -0.74$ (versus −0.25), and Leave-One-Delta-Out CV negative for 13 of 14 deltas (versus 11). The within-delta median skill attenuates from +0.25 to +0.05, an expected consequence of the smaller per-delta sample sizes and the greater internal homogeneity of the active delta plain relative to the wider bounding-box region. The headline result, that SOC–property models do not transfer between deltas in either direction, is therefore not an artifact of the box delineation. The check is reproduced by `h4_polygon_robustness.py` and the delta polygons (`delta_polygons.geojson`) in the data release.

Table S11: Robustness of the H4 transfer result to delta delineation: rectangular bounding box (main text) versus 0.5°-buffered active-delta polygon (Nienhuis et al., 2020). n is the number of complete-case top 1 m profiles; within-R^2 is the within-delta 5-fold CV (pH + sand/silt/clay). Summary rows give the cross-delta statistics. †Senegal is absent from the 100-delta polygon set; a buffered bounding box was used. ‡Yellow River retained $n < 10$ under the polygon, below the within-delta CV threshold.

	n		within-delta R^2	
Delta	Box	Polygon	Box	Polygon
Rhine-Meuse	2,652	595	+0.19	+0.35
Mississippi	2,043	130	−0.24	−0.26
Niger	327	29	+0.25	+0.06
Yellow River	315	9	−0.23	—‡
Limpopo	268	192	+0.37	+0.16
Chao Phraya	204	117	+0.10	−0.10
Paraná–Plata	179	18	+0.62	+0.59
Sacramento	178	222	+0.84	+0.36
Indus	167	17	+0.73	+0.34
Mackenzie	77	26	−0.10	−0.06
Danube	68	18	+0.79	−0.85
Amazon	62	34	−0.32	+0.08
Senegal†	57	22	+0.99	−0.36
Pearl	55	20	+0.57	−0.44
Total external n	**6,652**	**1,449**		
Within-delta median R^2			+0.25	+0.05
Cross-delta median R^2			−0.55	−0.66
Cross-delta % pairs > 0			6%	5.2%
Reverse pooled→Ganges R^2			−0.25	−0.74
LODO negative (of 14)			11	13