GEOSTATE: PRETRAINED GEOSPATIAL EMBEDDINGS ARE FORECASTABLE WHERE THE LAND CHANGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Geospatial foundation models are evaluated almost exclusively as static feature extractors. We ask whether their representations also behave as *predictive states*: can next year's embedding of a location be forecast from its recent history? We study annual 64-dimensional AlphaEarth Foundations embeddings over five Indonesian regions. We forecast the residual $Z_{t+1} - Z_t$ with extrapolation, linear, MLP, GRU and Transformer forecasters under matched budgets and five seeds. We evaluate them under temporal and geographic holdouts with spatial-block bootstrap intervals. On the primary 2024 test, aggregate gains over persistence are modest and no architecture clearly wins. Stratifying by change explains why. Learned forecasters lose slightly to persistence at stable locations but gain substantially at dynamic ones, including when change is identified from information available at forecast time. The apparent geographic attenuation disappears once differences in dynamism are accounted for. A non-foundation control shows that aggregate error can mislead. PCA, autoencoder and quarterly representations of Sentinel-2 are easier to forecast overall, but their skill is not consistently concentrated at land-cover change and brings little downstream benefit. AlphaEarth instead shows a strong, model-consistent association between forecast skill and observed land-cover change. Its predictable component is state-conditional and anti-persistent, consistent with mean reversion, and it varies across years. Aggregate forecast error alone is therefore a poor test of whether a representation is a useful predictive state.

1 INTRODUCTION

Pretrained geospatial foundation models (Cong et al., 2022; Jakubik et al., 2023; Tseng et al., 2023; Brown et al., 2025) produce representations that transfer well to land-cover mapping, biophysical regression and retrieval. In almost all of this work the embedding is *descriptive*: a summary of a location's current state, consumed by a task head. Whether the same embedding can serve as a *predictive* state, one whose trajectory is itself forecastable, is much less studied, yet it matters for any application that must anticipate rather than describe land change.

We study this question for AlphaEarth Foundations (Brown et al., 2025), whose public Satellite Embedding dataset provides a global, annual, 10 m, 64-dimensional embedding field. Given $Z_{t-k:t}$ at a location, we forecast the residual $\Delta Z_t = Z_{t+1} - Z_t$ and reconstruct $\hat{Z}_{t+1} = Z_t + \widehat{\Delta Z}_t$, so persistence is the zero function and every forecaster must beat it. We call this setting **GeoState**.

The answer turns out to depend on *where* one looks. Averaged over all locations, learned forecasters barely beat persistence, and no architecture clearly dominates. Most locations are nearly stable from one year to the next, and at those locations persistence is close to optimal. On the primary 2024 test the gains concentrate at locations that are actually changing, where every learned forecaster we tried beats persistence by a wide margin, in held-out regions as well as in-domain. The effect recurs in every held-out year we test, but it is not temporally universal: in-domain, it fails in 2023.

Contributions. (1) A task and evaluation protocol for forecasting a fixed, independently pretrained geospatial representation. It separates latent accuracy, downstream semantic retention, and change-stratified skill, and it reports seed variance and spatially blocked intervals (§3). (2) A controlled benchmark on AlphaEarth with extrapolation, linear, MLP, GRU and Transformer forecasters under

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

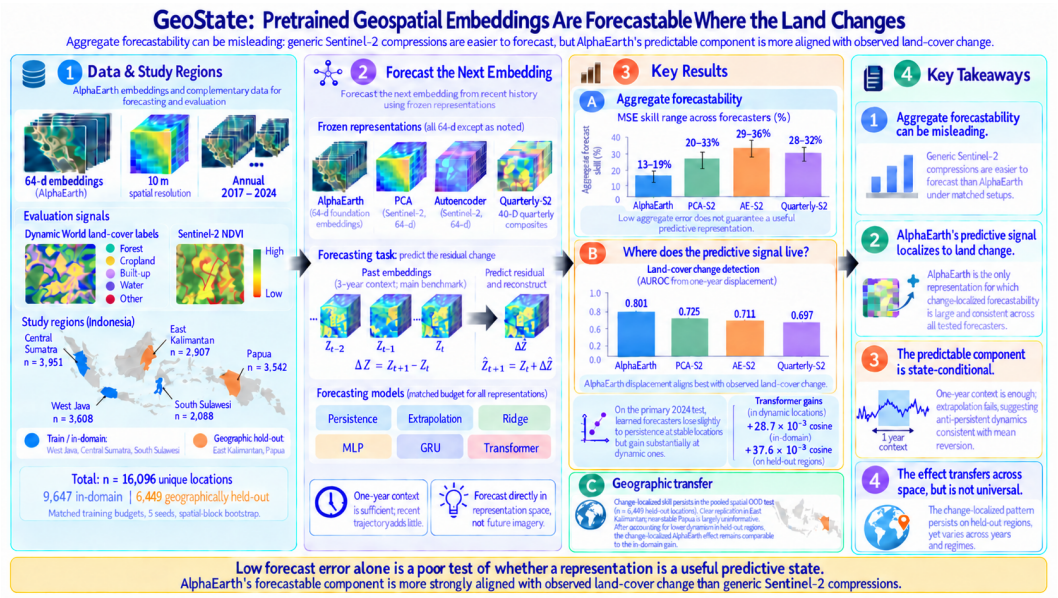


Figure 1: Graphical overview. (1) Data and study regions. (2) GeoState forecasts the next annual embedding from recent history, $\hat{Z}_{t+1} = Z_t + \widehat{\Delta Z}_t$, for AlphaEarth and for generic Sentinel-2 representations under matched budgets. (3) Key results: aggregate forecastability, where the predictive signal lies, and geographic transfer. (4) Takeaways, developed in Section 6.

one training budget (§5), built on an audited dataset (§4). (3) The finding that forecastability is *change-localized* and *model-agnostic*. It survives three stratifications, including one available at forecast time, and it transfers to unseen regions once differences in dynamism are controlled for (§6.2–6.3). (4) Evidence about *what* is predictable: a state-conditional, anti-persistent component consistent with mean reversion, rather than a trajectory trend, which varies from year to year (§6.4, §6.6). (5) A non-foundation control (PCA, autoencoder and quarterly-composite representations of raw Sentinel-2) on a matched common subset, with identical target years, forecasters, training protocol and seeds. Generic representations are *more* forecastable by aggregate error. AlphaEarth is the only representation whose forecastable component is concentrated at observed land-cover change, strongly and for every forecaster tested (§6.5).

Scope of the claim. The control shows that AlphaEarth's change-localized forecastability is not reproduced by generic compressions of same-year Sentinel-2 reflectance, in-domain. It does not isolate *which* property of AlphaEarth is responsible, whether self-supervised pretraining, multi-sensor assimilation or temporal gap-filling. Sentinel-2 was available only for the training regions, so the control is not repeated under geographic shift (§7).

2 Related Work

Geospatial foundation models and their evaluation. Self-supervised Earth-observation models such as SatMAE (Cong et al., 2022), Prithvi (Jakubik et al., 2023) and Presto (Tseng et al., 2023) pretrain on multispectral, and often multi-temporal, imagery. They are evaluated by fine-tuning or probing on downstream tasks such as classification, segmentation and regression. AlphaEarth Foundations (Brown et al., 2025) instead releases a precomputed annual embedding field that assimilates multiple sensors and is designed for mapping from sparse labels. In all of these, time is an *input* that improves the description of the present. To our knowledge, prior work has not directly evaluated whether the pretrained representation itself is forecastable as a target.

Analyses of AlphaEarth embeddings. Recent work shows that individual AlphaEarth dimensions align with physical land-surface variables and that these relationships are stable across years (mean inter-year correlation 0.963 over 2017–2023; Rahman, 2026), and that the dimensions form

a specialist-to-generalist hierarchy in which a few dimensions suffice for land-cover classification (Benavides-Martinez et al., 2026). Temporal consistency of the embedding–variable mapping is a precondition for our question but a different one. We ask whether the embedding *trajectory* at a location can be predicted, and where.

AlphaEarth as context for forecasting. Aalaila et al. (2026) use AlphaEarth embeddings as static spatial covariates in a log-Gaussian Cox process that forecasts emergency-service demand. There the embedding is an exogenous predictor of another variable. In GeoState the embedding is the forecast target.

Forecasting Earth observations. EarthNet2021 (Requena-Mesa et al., 2020) frames localized climate-impact forecasting as weather-conditioned prediction of future Sentinel-2 imagery, and EO-WM (Luo et al., 2026) forecasts multispectral observations with a weather-conditioned video diffusion transformer. Both predict observations in pixel space. We forecast in the space of a frozen, independently pretrained representation, which removes reconstruction of imagery and lets a single forecast serve any downstream probe.

Earth-system foundation models. Aurora (Bodnar et al., 2025) learns an encoder–processor–decoder whose latent state is rolled forward to forecast atmospheric and ocean variables. Its latent space is learned jointly *for* forecasting. GeoState asks the complementary question: whether a representation pretrained *without* any forecasting objective already carries forecastable dynamics.

Sequence models. Our recurrent and attention-based forecasters follow Cho et al. (2014) and Vaswani et al. (2017). With three annual tokens, we treat an MLP on the flattened context as the natural control for whether sequence architectures add anything.

3 Task and Evaluation Protocol

Task. Let $Z_t \in \mathbb{R}^{64}$ be the AlphaEarth embedding of a location in year t (unit-norm by construction). A forecaster F_θ maps a context $Z_{t-k+1:t}$ to $\widehat{\Delta Z}_t$, and $\hat{Z}_{t+1} = Z_t + \widehat{\Delta Z}_t$. We use $k = 3$ unless stated otherwise, and ablate k and the residual parameterization (§6.4).

Forecasters. *Extrapolation*: persistence ($\hat{Z}_{t+1} = Z_t$); constant velocity ($Z_t + (Z_t - Z_{t-1})$); mean velocity ($Z_t + \frac{1}{2}[(Z_t - Z_{t-1}) + (Z_{t-1} - Z_{t-2})]$). *Learned*: ridge regression on the flattened context (regularization tuned on validation); a two-layer MLP on the flattened context; a two-layer GRU; and a two-layer pre-norm Transformer encoder that reads the last token. Neural models minimize $\alpha\,\mathrm{MSE}(\widehat{\Delta Z}_t, \Delta Z_t) + (1-\alpha)(1 - \cos(\hat{Z}_{t+1}, Z_{t+1}))$ with $\alpha = 0.8$.

Three levels of evaluation. (i) *Latent*: RMSE, mean cosine similarity, and *MSE skill* $1 - \mathrm{MSE}_{\text{model}}/\mathrm{MSE}_{\text{persistence}}$, a scale-free quantity (though not, on its own, a measure of meaningful forecastability; §6.5). (ii) *Semantic retention*: a land-cover classifier (Dynamic World labels; macro-F1) and an NDVI regressor (R^2) are trained once on *true* training embeddings and applied unchanged to forecasts. Retention is $\mathrm{Score}(\hat{Z}_{t+1})/\mathrm{Score}(Z_{t+1})$. Both targets derive from Sentinel-2, which AlphaEarth also ingests, so these probes show that forecasts preserve remote-sensing semantics, *not* that they match independently measured outcomes. (iii) *Change-stratified skill*, described next.

Stratification. We report the per-location cosine gain over persistence within strata defined in three ways. *Oracle*: terciles of the realized change $\|Z_{t+1} - Z_t\|_2$. This is a diagnostic that conditions on the future. Persistence is near-optimal by construction in its stable stratum, and the question is how much of the dynamic stratum's change a model recovers. *Observable*: terciles of past change $\|Z_t - Z_{t-1}\|_2$, which is known when the forecast is made. *Label change*: whether the Dynamic World label changes between t and $t+1$, which is defined without the embedding metric. All thresholds are computed once on the *training* split and applied unchanged to every test split (thresholds 0.192 and 0.299 for realized change), so that a stratum denotes the same absolute amount of change in every region.

Matched-change analysis. Held-out regions may simply change less. We therefore (a) plot gain against realized change in training-decile bins, and (b) importance-reweight held-out locations so that their distribution of realized change matches the in-domain test split, then recompute the aggregate gain.

Uncertainty. All neural results use five training seeds (mean ± s.d.). Pixels are spatially autocorrelated, so intervals come from a paired bootstrap that, in each of 2,000 replicates, resamples 0.25° (≈28 km) spatial blocks *and* independently draws a training seed for each model. The interval therefore reflects both spatial dependence and initialization and minibatch-order variance. Appendix Table 8 shows that conclusions are unchanged from point resampling up to 0.5° blocks. Because the held-out set contains only two regions, we also report each region separately.

4 DATA

Sources and regions. We sample 10 m locations from the AlphaEarth Satellite Embedding dataset (2017–2024) in five Indonesian areas of interest. West Java, Central Sumatra and South Sulawesi are the training regions; East Kalimantan and Papua are held out entirely. Each location carries annual embeddings, the annual modal Dynamic World label (Brown et al., 2022), and a June–September median Sentinel-2 (Drusch et al., 2012) NDVI. Missing labels and NDVI never cause a location to be dropped; only embedding validity does. This avoids the cloud-driven selection bias that arises when a location is required to have valid auxiliary variables in every year.

Data audit. Each region was exported as four independent random chunks (distinct seeds, 1,000 candidates each). An earlier loader additionally concatenated a 3,594-row legacy single-file export of West Java. Its seed coincides with chunk 0, so 898 of these rows duplicated chunk-0 locations. We discard the entire legacy export, keep only the four current chunks, and assert that every (region, location) pair is unique. This reduces West Java from 7,202 rows to 3,608 locations, and the full dataset from 19,690 rows to 16,096 locations (Appendix Table 4). One South Sulawesi chunk failed at the Earth Engine task level ("computation timed out"). A whole-chunk failure is independent of which pixels would have been sampled, so South Sulawesi contributes three of its four random chunks. All results in this paper are computed on the audited data.

Splits. Training targets are 2020–2022, the validation target is 2023, and the test target is 2024, all within the training regions. The in-domain test thus reuses training *locations* at an unseen *year* ($n = 9{,}647$). The geographic test uses 2024 targets in the held-out regions ($n = 6{,}449$), whose locations never appear in training. Random pixel-level splits are never used.

5 EXPERIMENTAL SETUP

All neural forecasters share one budget: AdamW, batch size 128, weight decay 10^{-4}, at most 60 epochs, early stopping with patience 8 on validation loss, and a learning rate selected per model from $\{10^{-4}, 3\cdot10^{-4}, 10^{-3}\}$ on validation loss (seed 0), after which each model is retrained with five seeds. The MLP, GRU and Transformer have 131,648, 198,592 and 430,656 parameters respectively (Appendix Table 5). Ridge's regularization strength is selected on validation ($\alpha = 0.0001$). Probes are gradient-boosted trees trained on true 2020–2022 training embeddings. The entire study runs on one 6 GB consumer GPU.

6 RESULTS

6.1 AGGREGATE METRICS SHOW MODEST GAINS AND NO CLEAR WINNER

Table 1 gives the conventional view. Extrapolating the recent trajectory is harmful: constant velocity has an MSE skill of -301.5% in-domain and mean velocity -109.2%, so year-to-year displacements are strongly anti-persistent. The learned forecasters improve on persistence, but modestly. In-domain, the MLP ($+15.0\%$) and ridge ($+14.7\%$) lead, ahead of the Transformer ($+6.0\%$) and GRU ($+1.3\%$), and the neural models vary noticeably across seeds. In the held-out regions all skills fall to a few percent. The Transformer has the highest held-out cosine, but its margin over ridge is small ($+1.7[+0.4, +2.7]\times10^{-3}$) and its margin over the MLP is not distinguishable from zero ($+1.1[-0.4, +2.5]\times10^{-3}$; Table 3). Semantic retention does not separate the models either, and it is not even monotone in latent accuracy. Held-out land-cover retention is highest for the mean-velocity extrapolator (0.845), whose MSE skill is strongly negative, and persistence (0.813) beats every learned model. Latent accuracy and probe retention therefore measure different things and have to be reported separately. Read alone, this table would suggest that AlphaEarth embeddings

Table 1: Aggregate results on 2024 targets (mean±s.d. over five seeds for neural models; extrapolation and ridge are deterministic). Skill is MSE skill over persistence in percent. LC: land-cover probe; retention = forecast score / score on the true embedding. Best per column and split in bold.

Model	RMSE ↓	Cosine ↑	Skill (%) ↑	LC F1 ↑	LC retention ↑	NDVI R^2 ↑
In-domain (2024), n = 9,647						
Persistence	0.0394	0.9504	0.0	0.478	0.919	0.928
Const. velocity	0.0789	0.8540	-301.5	0.413	0.794	0.843
Mean velocity	0.0570	0.9089	-109.2	0.451	0.866	0.894
Ridge	0.0364	0.9568	14.7	**0.514**	**0.988**	0.946
MLP	**0.0363**±0.0005	**0.9589**±0.0009	**15.0**±2.3	0.508±0.010	0.976±0.020	**0.947**±0.000
GRU	0.0391±0.0013	0.9561±0.0020	1.3±6.6	0.511±0.015	0.982±0.030	0.945±0.002
Transformer	0.0382±0.0010	0.9578±0.0016	6.0±4.8	0.507±0.007	0.974±0.013	0.946±0.001
Geographic OOD (2024), n = 6,449						
Persistence	0.0295	0.9722	0.0	0.337	0.813	0.865
Const. velocity	0.0511	0.9297	-201.1	0.307	0.741	0.839
Mean velocity	0.0394	0.9544	-78.6	**0.350**	**0.845**	0.854
Ridge	**0.0290**	0.9728	**3.5**	0.312	0.753	**0.876**
MLP	0.0291±0.0002	0.9734±0.0004	2.8±1.6	0.315±0.009	0.760±0.023	0.876±0.001
GRU	0.0304±0.0019	0.9734±0.0014	-6.4±13.4	0.334±0.009	0.807±0.021	0.874±0.003
Transformer	0.0290±0.0005	**0.9745**±0.0007	3.1±3.3	0.331±0.007	0.799±0.016	0.876±0.001

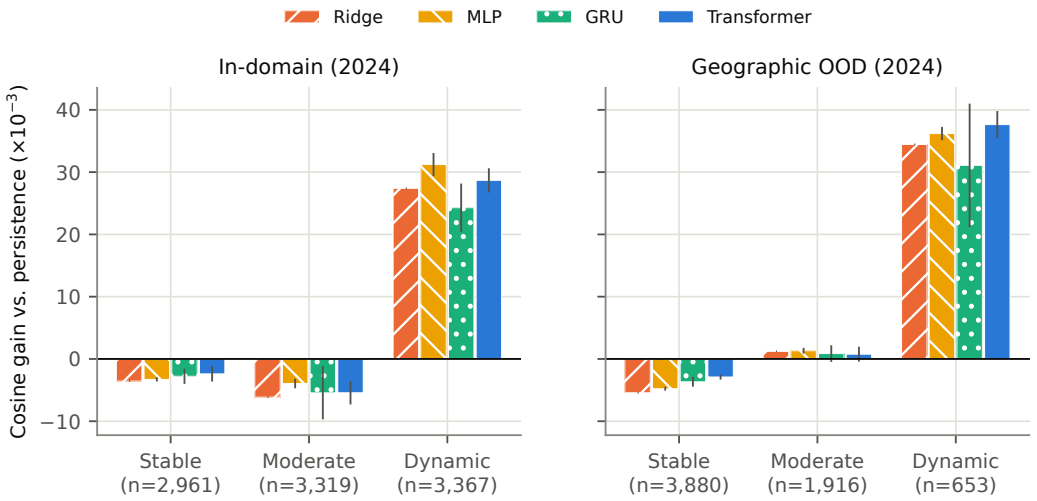


Figure 2: Cosine gain over persistence by realized-change stratum (training-derived thresholds applied to both test splits; bars show mean±s.d. over seeds). Every learned forecaster pays a small cost at stable locations and gains substantially at dynamic ones, in-domain and on held-out regions. Numbers in Table 2.

are barely forecastable and that the choice of architecture is immaterial. The next sections show that the first reading is wrong and the second is right.

6.2 Forecastability is change-localized and model-agnostic

Figure 2 and Table 2 show where the aggregate gains come from. At stable locations every learned forecaster is slightly *worse* than persistence, for example -2.4×10^{-3} for the Transformer in-domain, because it predicts small changes that do not occur. At dynamic locations every one of them is much better: ridge $+27.4$, MLP $+31.2$, GRU $+24.3$ and Transformer $+28.7\times10^{-3}$ in-domain, with spatial-block intervals far from zero (Table 3). The effect belongs to the represen-

Table 2: Cosine gain over persistence ($\times 10^{-3}$) by stratum, for every forecaster. **Top**: oracle strata by realized change. **Middle**: observable strata by past change $\|Z_t - Z_{t-1}\|$, known at forecast time. **Bottom**: Dynamic World label changed vs. unchanged between t and $t+1$. All thresholds come from the training split. Best per row in bold.

Split	Stratum	n	Const. vel.	Mean vel.	Ridge	MLP	GRU	Transformer
Oracle: realized change $\|Z_{t+1} - Z_t\|$								
ID	Stable	2,961	-33.5	-11.6	-3.6	-3.3$\pm$0.3	-2.8$\pm$1.2	**-2.4$\pm$1.2**
ID	Moderate	3,319	-84.6	-33.3	-6.1	**-3.9$\pm$0.8**	-5.4$\pm$4.3	-5.4$\pm$1.9
ID	Dynamic	3,367	-163.5	-75.9	+27.4	**+31.2$\pm$1.9**	+24.3$\pm$3.9	+28.7$\pm$1.9
OOD	Stable	3,880	-24.4	-9.2	-5.4	-4.8$\pm$0.3	-3.7$\pm$0.8	**-2.9$\pm$0.4**
OOD	Moderate	1,916	-57.3	-23.1	+1.2	**+1.3$\pm$0.5**	+0.9$\pm$1.3	+0.8$\pm$1.2
OOD	Dynamic	653	-107.2	-53.8	+34.4	+36.2$\pm$1.1	+31.1$\pm$9.9	**+37.6$\pm$2.2**
Observable: past change $\|Z_t - Z_{t-1}\|$								
ID	Stable	2,087	-25.5	-9.3	-2.4	-2.4$\pm$0.3	-1.9$\pm$0.7	**-1.7$\pm$1.2**
ID	Moderate	2,638	-53.6	-20.9	-2.8	**-1.6$\pm$0.8**	-1.8$\pm$1.7	-1.7$\pm$1.6
ID	Dynamic	4,922	-149.5	-66.2	+14.9	**+18.6$\pm$1.3**	+13.0$\pm$3.2	+16.1$\pm$2.1
OOD	Stable	3,776	-21.9	-9.1	-5.0	-4.5$\pm$0.3	-3.5$\pm$0.8	**-2.7$\pm$0.4**
OOD	Moderate	1,967	-55.4	-21.5	+0.8	**+1.1$\pm$0.4**	+0.7$\pm$1.2	+0.9$\pm$1.0
OOD	Dynamic	706	-117.4	-54.5	+29.9	+32.0$\pm$1.2	+27.9$\pm$8.7	**+32.8$\pm$2.3**
Dynamic World label change $t \to t+1$								
ID	Label stable	7,785	-91.4	-38.0	+3.5	**+5.1$\pm$0.6**	+2.6$\pm$2.6	+3.8$\pm$1.5
ID	Label changed	760	-170.2	-83.5	+34.3	**+40.3$\pm$3.1**	+33.2$\pm$6.8	+39.4$\pm$2.2
OOD	Label stable	4,439	-41.7	-17.1	-0.4	-0.1$\pm$0.4	-0.0$\pm$1.1	**+0.6$\pm$0.8**
OOD	Label changed	192	-73.7	-41.6	+33.6	+36.2$\pm$1.7	+32.9$\pm$12.8	**+41.5$\pm$2.0**

Table 3: Paired seed-and-spatial-block bootstrap (0.25° blocks, 2,000 replicates): mean difference in cosine similarity ($\times 10^{-3}$) with 95% interval. "Dyn.": oracle dynamic stratum (training thresholds). Bold: interval excludes zero.

	In-domain 2024		Geographic OOD 2024	
Comparison	All	Dyn.	All	Dyn.
Ridge − Persistence	**+6.4 [+4.5, +8.4]**	**+27.4 [+23.7, +31.4]**	+0.6 [-0.8, +1.9]	**+34.4 [+27.2, +42.2]**
MLP − Persistence	**+8.5 [+6.0, +11.1]**	**+31.2 [+26.1, +36.7]**	+1.2 [-0.3, +2.7]	**+36.2 [+28.8, +45.0]**
GRU − Persistence	**+5.8 [+2.1, +8.7]**	**+24.3 [+18.3, +32.3]**	+1.2 [-1.0, +3.9]	**+31.1 [+12.4, +45.6]**
Transformer − Persistence	**+7.4 [+4.6, +11.0]**	**+28.7 [+23.4, +34.5]**	**+2.3 [+0.5, +4.1]**	**+37.6 [+28.2, +48.0]**
Transformer − Ridge	+1.0 [-0.8, +3.9]	+1.3 [-2.0, +5.0]	**+1.7 [+0.4, +2.7]**	+3.2 [-0.8, +8.2]
Transformer − GRU	+1.7 [-1.7, +6.5]	+4.4 [-2.4, +11.3]	+1.1 [-2.1, +3.8]	+6.5 [-4.5, +28.4]
Transformer − MLP	-1.1 [-3.6, +2.7]	-2.5 [-6.8, +2.8]	+1.1 [-0.4, +2.5]	+1.4 [-2.5, +6.6]
GRU − Ridge	-0.6 [-3.4, +1.7]	-3.1 [-8.3, +2.5]	+0.6 [-1.6, +2.9]	-3.4 [-23.6, +6.7]

tation, not to an architecture. Within the dynamic stratum, the Transformer is indistinguishable from ridge ($+1.3[-2.0, +5.0]\times 10^{-3}$ in-domain, $+3.2[-0.8, +8.2]$ held out) and from the MLP ($-2.5[-6.8, +2.8]$). The earlier draft of this work attributed the effect to the Transformer; that attribution does not survive a matched comparison.

Two checks address the concern that oracle strata favor persistence by construction. First, stratifying by *past* change, which is available when the forecast is made, preserves the pattern: at locations that moved most in the previous year, the Transformer gains $+16.1\times 10^{-3}$ in-domain and $+32.8$ held out. The high-gain locations can therefore be identified in advance. Second, at locations whose Dynamic World label changes, which is defined without any embedding distance, gains are $+39.4$ (in-domain, $n = 760$) and $+41.5\times 10^{-3}$ (held out, $n = 192$), versus $+3.8$ and $+0.6$ where the label is unchanged.

6.3 THE EFFECT TRANSFERS; ITS APPARENT ATTENUATION WAS A MATTER OF COMPOSITION

With thresholds fixed from training data, the held-out regions are much less dynamic: only 10% of held-out locations fall in the dynamic stratum, against 35% in-domain (Appendix Figure 4). This explains why aggregate held-out gains are small, and it also explains the result reported in an earlier draft, where the dynamic-stratum gain appeared roughly halved under geographic shift. That draft defined strata by terciles computed *within* each split, so the held-out "dynamic" tercile contained much less absolute change (Appendix Table 6). At equal absolute change the gains do not shrink. The held-out dynamic-stratum gain ($+37.6\times10^{-3}$ for the Transformer) is at least as large as in-domain, and the gain–change curves overlap (Appendix Figure 8). Reweighting held-out locations to the in-domain change distribution raises the Transformer's aggregate held-out gain from $+2.3$ to $+10.7\times10^{-3}$, matching or exceeding its in-domain $+7.4$ (ridge: $+0.6 \rightarrow +9.4$ vs. $+6.4$; Appendix Table 7).

The two held-out regions behave very differently, and Appendix Table 10 treats each as its own replication. East Kalimantan, the more dynamic of the two, shows clear gains for every model over all locations (Transformer $+8.3[+4.5, +12.0]\times10^{-3}$) and in its dynamic stratum ($+39.6[+29.9, +50.8]$). Papua is overwhelmingly stable: only 43 of its locations reach the dynamic threshold, and *every* learned forecaster is worse than persistence over all Papuan locations (Transformer $-2.6[-3.3, -2.0]$, ridge $-5.8[-6.3, -5.3]\times10^{-3}$). The Transformer's small aggregate held-out lead in Table 1 comes from a smaller penalty on Papua's stable locations, not from larger gains where change occurs. With only two held-out domains we treat geographic transfer as supported for a changing frontier region and uninformative for a near-static one.

6.4 WHAT IS PREDICTABLE: THE CURRENT STATE, NOT THE TRAJECTORY

Three results point to the same mechanism (Appendix Table 11). *Context*: a Transformer that sees only the current year ($k = 1$) matches or beats one that sees three (in-domain skill $+16.6\%$ vs. $+4.0\%$ and dynamic gain $+32.7$ vs. $+28.1\times10^{-3}$; held-out skill $+4.9\%$ vs. $+1.9\%$, with a slightly smaller held-out dynamic gain, $+34.3$ vs. $+37.2$). The multi-year trajectory adds little or nothing. *Extrapolation*: velocity baselines are the worst models in every stratum (Table 2), so recent displacements do not continue. *Residual form*: predicting Z_{t+1} directly instead of ΔZ_t is competitive in-domain (skill $+14.2\%$) but collapses on held-out regions (skill -25.2%), so anchoring on Z_t is what makes the forecaster transfer. Together, these results indicate that the forecastable part of ΔZ_t is a *state-conditional, anti-persistent* component, consistent with mean reversion. Given where a location currently sits in embedding space, part of its next displacement is predictable, plausibly as a partial return of an anomalous state towards what is typical for such states, although we do not measure that conditional equilibrium directly. The same finding explains why a linear model and an MLP are as good as sequence models: with the current state as the informative input, three annual tokens leave nothing for attention or recurrence to exploit. Our default loss is also not optimal for squared error: dropping the cosine term raises in-domain skill to $+17.3\%$ with an unchanged dynamic-stratum gain.

6.5 IS IT THE REPRESENTATION? A NON-FOUNDATION CONTROL

Is AlphaEarth's forecastable structure a property of the foundation representation, or would any compressed representation of the same imagery show it? We built three generic representations from raw Sentinel-2 surface reflectance (10 bands, monthly composites): PCA to 64 dimensions (retaining 99.85% of the variance), a 64-dimensional autoencoder, and a denoised 40-dimensional set of per-band quarterly medians of observed months. All are fitted on training-region imagery from 2019–2022 only and rescaled by a single scalar to unit mean norm. In our sampled regions, 81–99% of Sentinel-2 month-band cells are missing in 2017–2018, so the control starts in 2019. Every representation, AlphaEarth included, uses 2019–2024 with a one-year context (k=1, which §6.4 shows is sufficient) and the same target years. The Sentinel-2 exports cover the three training regions, so the control is in-domain, on $8{,}218$ common locations, with the benchmark otherwise unchanged (Appendix G).

Raw MSE skill cannot rank representations by forecastability. Year-to-year noise makes persistence worse and rewards any model that shrinks towards typical states. We therefore ask *where*

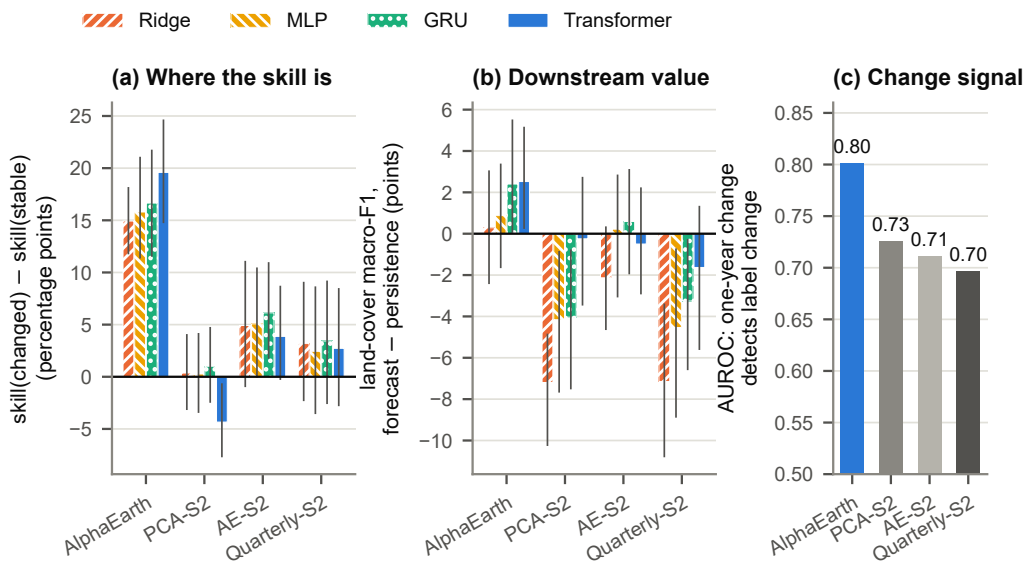


Figure 3: AlphaEarth vs. generic Sentinel-2 representations (in-domain test 2024, k=1, common point set; 95% seed-and-block intervals). **(a)** MSE skill at locations whose Dynamic World label changed minus skill where it did not. AlphaEarth's skill is strongly and consistently concentrated at observed label change; for generic representations the effect is small or absent. **(b)** Land-cover macro-F1 of the probe on the forecast minus on persistence: generic forecasts do not help, and PCA-S2 and Quarterly-S2 forecasts mostly hurt. **(c)** How well one-year change of each representation detects an observed label change.

the skill arises, using Dynamic World label change, which is the same for every representation, and whether forecasts carry downstream value. By aggregate skill, the generic representations are *more* forecastable: $+19.9$ to $+33.3\%$ for PCA, $+29.1$ to $+36.1\%$ for the autoencoder and $+28.1$ to $+31.8\%$ for quarterly composites, against $+13.2$ to $+19.1\%$ for AlphaEarth (Appendix Table 13). Where that skill arises differs sharply (Figure 3a). In AlphaEarth, skill at label-changed locations exceeds skill at stable ones by $+14.9$ to $+19.6$ percentage points for every forecaster, with intervals well clear of zero. For PCA the difference is nil or negative (ridge $+0.4$, interval $[-3.2, +4.1]$; Transformer -4.3). The autoencoder shows a small effect ($+4$ to $+6$, significant for one of four models) and quarterly composites none. AlphaEarth is therefore the only representation for which change-localized forecastability is large and consistent across all tested forecasters. In the 2023 validation year the generic representations' skill is significantly concentrated at *stable* locations (PCA ridge -9.8, quarterly ridge -14.3), while AlphaEarth's remains at changing ones ($+11.2$ to $+15.2$; Appendix Table 15).

Downstream value follows the same split (Figure 3b). In 2024, no AlphaEarth forecaster reduces next-year land-cover F1 relative to persistence, and the GRU and Transformer improve it (Transformer $+2.5$, interval $[+0.2, +5.2]$). Most PCA and quarterly forecasters reduce it (ridge -7.2 and -7.1), and autoencoder forecasts leave it unchanged. In the 2023 validation year every representation's forecasts lower F1. The losses are significant for 9 of 12 generic forecasters and for none of AlphaEarth's. The representations themselves differ in the same way: one-year change detects a label change with AUROC 0.801 for AlphaEarth, against 0.725, 0.711 and 0.697 (Figure 3c). Forecast change also anticipates next year's label change better in AlphaEarth, modestly in 2024 (0.764 vs. at most 0.745) and markedly in 2023 (0.833 vs. at most 0.711). A two-year context gives the same picture (Appendix Table 16).

The generic representations' extra skill is therefore largely the reversal of year-to-year fluctuation, from cloud gaps, compositing and phenology, that persistence carries forward and a shrinking model removes. It is spread evenly over the map and carries little land-cover information. In AlphaEarth the

predictable component is smaller but concentrates where observed land cover changes. Aggregate forecast error on its own conflates the two.

6.6 Temporal robustness: year dependence and horizon

A single test year is a weak basis for a temporal claim, so Appendix Table 12 repeats the one-step evaluation for every target year in the held-out regions. There, the dynamic-stratum gain is positive in all five years and for every learned forecaster (Transformer: $+38.6$ in 2020, $+34.4$ in 2021, $+33.1$ in 2022, $+37.6\times10^{-3}$ in 2024), but in 2023 it drops to $+13.9$ and aggregate skill falls to roughly zero or below (ridge -0.8%, Transformer $+0.1\%$). In-domain, 2023 is a clear failure even though it is the validation year: every learned forecaster is worse than persistence in aggregate (Transformer -9.0%) *and* in the dynamic stratum (-3.1×10^{-3}). 2023 coincided with El Niño and widespread drought across Indonesia. This may contribute to the failure, although we do not test that attribution. Forecastability is a property of a representation *and a period*.

Recursive rollouts (Appendix Figure 6 and Table 9) show the expected compounding. On held-out regions every learned forecaster beats persistence for the first horizons (Transformer 0.9725 vs. 0.9604 at h=1), but persistence leads by h=4–5 (0.9505 vs. 0.9334). The in-domain rollout begins in 2023 and never recovers from it.

7 Discussion and Limitations

What the evidence supports. Annual AlphaEarth embeddings contain a forecastable component with three properties. It is concentrated at changing locations, and those locations can be identified in advance from recent change. Linear and nonlinear forecasters recover it equally well. It transfers to an unseen frontier region at matched change. Generic compressions of raw Sentinel-2 do not share this property to a comparable degree, even though they score higher on aggregate forecast error. The effect is not temporally universal, since it fails in-domain in 2023 (see also Appendix Figure 7). Two earlier claims do not survive these controls: that the Transformer specifically captures change dynamics, and that the effect attenuates under geographic shift.

What the control does not show. The generic representations use Sentinel-2 optical data alone, with our own compositing and gap interpolation, whereas AlphaEarth assimilates several sensors and fills gaps through its own temporal modelling. The control shows that change-localized forecastability is not a generic consequence of compressing the same-year optical record. It cannot attribute that property to self-supervised pretraining rather than to multi-sensor fusion or temporal regularization. It is also in-domain only.

Other limitations. (i) The semantic probes and the label-change strata rely on Dynamic World and NDVI, both derived from Sentinel-2. An evaluation against an independent outcome, Landsat-based Hansen forest loss (Hansen et al., 2013), is implemented but not yet complete. (ii) Oracle strata condition on the future. We therefore treat the observable-change strata as the operational result. (iii) Two held-out regions are two domain replications, not thousands, and they differ sharply. (iv) Results depend on the target year (§6.6). (v) The scale is a pilot: about 16,000 locations in one country, with deterministic forecasts.

8 Conclusion

Forecasting a frozen geospatial foundation embedding is possible, but not everywhere and not always. On an audited Indonesian benchmark with matched baselines, five seeds and spatially blocked intervals, every learned forecaster beats persistence at changing locations in the primary 2024 evaluation, while losing slightly at stable ones. The effect persists across all geographically held-out years but is not temporally universal, as the in-domain 2023 failure shows. It is visible from information available at forecast time, holds on an unseen frontier region once dynamism is matched, and comes from a state-conditional, anti-persistent component that simple models capture as well as Transformers. Generic Sentinel-2 compressions look more forecastable by aggregate error, but not at change and not usefully downstream. To evaluate embeddings as predictive states, ask where forecast skill arises and whether it carries semantic value, not only how large it is.

AI USE STATEMENT

Generative AI (Claude, Anthropic) assisted with writing and debugging the data pipeline, model, training and evaluation code, and the experiment notebook, under the author's direction and review.

ETHICS STATEMENT

The study uses only public satellite-derived data products and involves no human subjects or personal data. Land-change forecasting can support conservation and planning. It could also be misused to target land for conversion, and forecasts in regions where a model has not been validated, such as near-static landscapes like Papua in our data, should not be used for decisions.

REPRODUCIBILITY STATEMENT

The supplementary material contains the complete pipeline: Earth Engine export scripts with fixed seeds, the audited dataset preparation with its uniqueness assertion, the benchmark library, and a single notebook that regenerates every table, figure and in-text number in this paper from the raw exports in about one hour on a 6 GB GPU. Seeds, the learning-rate grid, and all selection criteria are fixed in the notebook's first cell.

REFERENCES

Yahya Aalaila, Mouad Elhamdi, Gerrit Großmann, Daniel Jenson, Elizaveta Semenova, and Sebastian Vollmer. When context compensates for sparse event history: AlphaEarth for spatio-temporal point-process forecasting. *arXiv preprint arXiv:2607.01082*, 2026.

Ivan Felipe Benavides-Martinez, Justin Guthrie, Jhon Edwin Arias, Yeison Alberto Garces-Gomez, Angela Ines Guzman-Alvis, Cristiam Victoriano Portilla-Cabrera, Somnath Mondal, Andrew J. Allyn, and Auroop R. Ganguly. What on earth is AlphaEarth? hierarchical structure and functional interpretability for global land cover. *arXiv preprint arXiv:2603.16911*, 2026.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, et al. A foundation model for the Earth system. *Nature*, 641:1180–1187, 2025.

Christopher F. Brown, Steven P. Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J. Pasquarella, Robert Haertel, Simon Ilyushchenko, Kurt Schwehr, Mikaela Weisse, Fred Stolle, Craig Hanson, Oliver Guinan, Rebecca Moore, and Alexander M. Tait. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9:251, 2022.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. AlphaEarth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 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)*, pp. 1724–1734, 2014.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems*, 2022.

Matthias Drusch, Umberto Del Bello, Sabine Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120:25–36, 2012.

Matthew C. Hansen, Peter V. Potapov, Rebecca Moore, Matt Hancher, Svetlana A. Turubanova, Alexandra Tyukavina, David Thau, Stephen V. Stehman, Scott J. Goetz, Thomas R. Loveland, Anil Kommareddy, Alexey Egorov, Louise Chini, Christopher O. Justice, and John R. G. Townshend. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160): 850–853, 2013.

Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

Junwei Luo, Shuai Yuan, Zhenya Yang, Yansheng Li, Zhe Liu, and Hengshuang Zhao. EO-WM: A physically informed world model for probabilistic earth observation forecasting. *arXiv preprint arXiv:2606.27277*, 2026.

Mashrekur Rahman. Physically interpretable AlphaEarth foundation model embeddings enable LLM-based land surface intelligence. *arXiv preprint arXiv:2602.10354*, 2026.

Christian Requena-Mesa, Vitus Benson, Joachim Denzler, Jakob Runge, and Markus Reichstein. EarthNet2021: A novel large-scale dataset and challenge for forecasting localized climate impacts. In *Tackling Climate Change with Machine Learning workshop at NeurIPS*, 2020.

Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

A Data audit

Table 4: Export provenance per region. v1: rows produced by the earlier loader, which concatenated every file; v2: unique locations used here.

Region	Role	Chunks	Missing	v1 rows	v2 locations
Central Sumatra	train	4/4	–	3,951	3,951
East Kalimantan	OOD	4/4	–	2,907	2,907
Papua	OOD	4/4	–	3,542	3,542
South Sulawesi	train	3/4	1	2,088	2,088
West Java	train	4/4	–	7,202	3,608

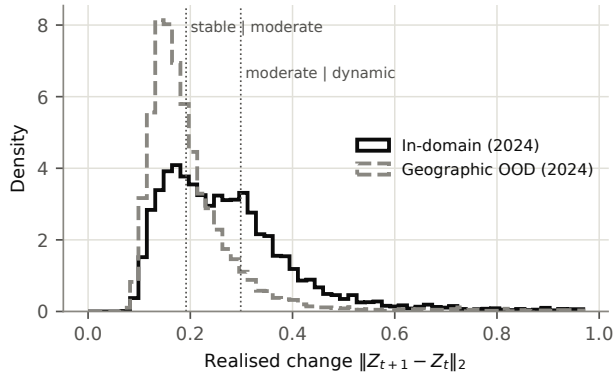


Figure 4: Distribution of realized one-year change in the 2024 test splits, with the training-derived tercile thresholds. Held-out regions are markedly less dynamic.

B TRAINING DETAILS

Table 5: Neural forecasters: parameter count, validation-selected learning rate, and mean best/total epochs over five seeds (max 60, patience 8).

Model	Parameters	LR	Best epoch	Epochs run
MLP	131,648	1e-03	13.4	21.4
GRU	198,592	1e-03	13.4	21.4
Transformer	430,656	1e-03	10.6	18.6

All models share: hidden width 128 (GRU, Transformer; MLP width 256), dropout 0.1, two layers, 4 attention heads (Transformer), gradient clipping at 1.0, fp32. Ridge is fit in closed form on the flattened context; its regularization strength is chosen from $\{10^{-4}, \ldots, 10^{2}\}$ on validation RMSE.

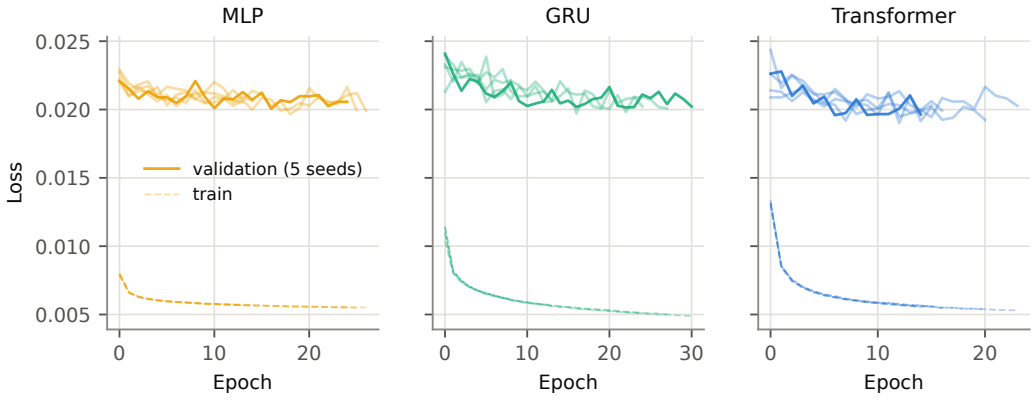


Figure 5: Training (dashed) and validation (solid) loss for all five seeds.

C ADDITIONAL STRATIFIED AND STATISTICAL RESULTS

Table 6: Oracle strata defined by terciles computed *within* each split (the earlier draft's definition). Held-out "dynamic" then denotes less absolute change, which produces an apparent attenuation of the gain. Compare Table 2.

Split	Stratum	n	Const. vel.	Mean vel.	Ridge	MLP	GRU	Transformer
ID	Stable	3,216	-35.2	-12.3	-3.6	-3.3±0.3	-2.8±1.3	**-2.4±1.2**
ID	Moderate	3,215	-89.0	-35.3	-6.1	**-3.8±0.8**	-5.5±4.5	-5.5±1.9
ID	Dynamic	3,216	-165.1	-77.0	+28.7	**+32.6±1.9**	+25.5±4.1	+30.1±1.9
OOD	Stable	2,150	-19.4	-7.0	-6.1	-5.3±0.3	-4.1±0.8	**-3.1±0.4**
OOD	Moderate	2,149	-32.7	-12.9	-4.2	-3.8±0.3	-3.0±0.8	**-2.5±0.6**
OOD	Dynamic	2,150	-75.6	-33.7	+12.1	**+12.7±0.5**	+10.6±3.6	+12.6±1.4

Table 7: Aggregate cosine gain over persistence ($\times 10^{-3}$): in-domain, held-out, and held-out after importance reweighting to the in-domain distribution of realized change.

Model	In-domain	OOD (raw)	OOD (reweighted)
Ridge	+6.4	+0.6	+9.4
MLP	+8.5±0.9	+1.2±0.4	+10.2±0.5
GRU	+5.8±2.0	+1.2±1.4	+8.6±3.2
Transformer	+7.4±1.6	+2.3±0.7	+10.7±1.1

Table 8: Sensitivity of the Transformer–persistence 95% interval ($\times 10^{-3}$) to the spatial block size of the bootstrap ("point" = classic resampling of locations). At $0.25°$ the in-domain and held-out tests contain 460 and 426 blocks.

Split	Scope	Point	$0.05°$	$0.1°$	$0.25°$	$0.5°$
ID	all	[+5.4, +10.4]	[+5.3, +10.6]	[+5.2, +10.7]	[+4.6, +11.0]	[+3.9, +11.6]
ID	dynamic	[+25.0, +33.0]	[+25.0, +33.1]	[+24.3, +33.3]	[+23.4, +34.5]	[+21.8, +36.7]
OOD	all	[+0.9, +3.5]	[+0.8, +3.5]	[+0.7, +3.7]	[+0.5, +4.1]	[+0.1, +4.5]
OOD	dynamic	[+30.9, +44.6]	[+31.1, +45.0]	[+30.3, +46.1]	[+28.2, +48.0]	[+28.6, +47.0]

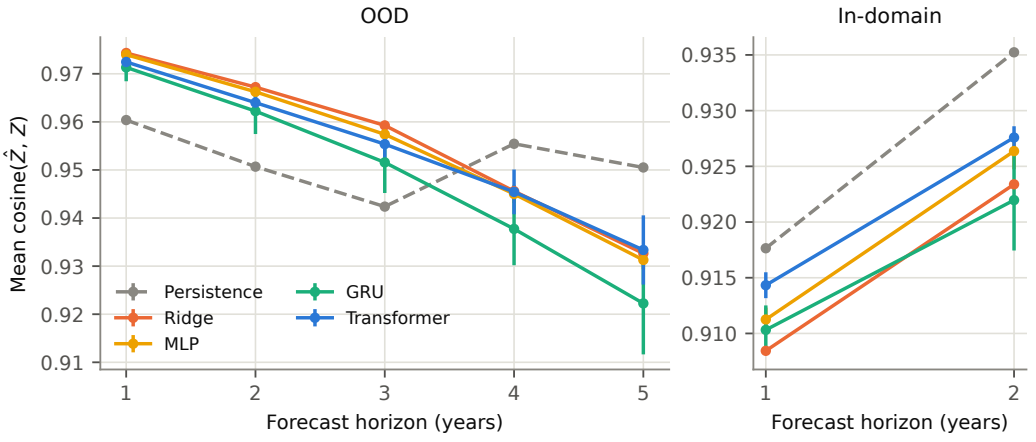


Figure 6: Recursive multi-horizon rollout (mean±s.d. over seeds). Left: held-out regions, context 2017–19, horizons 2020–24. Right: in-domain, context 2020–22, horizons 2023–24. Learned forecasters lead persistence for the first held-out horizons and lose it as errors compound; the in-domain rollout starts in 2023, the year in which all learned models underperform.

Table 9: Recursive rollout: mean cosine similarity by horizon (mean±s.d. over seeds).

Group	h (year)	Persistence	Const. vel.	Mean vel.	Ridge	MLP	GRU	Transformer
OOD	1 (2020)	0.9604	0.9083	0.9315	**0.9743**	0.9740$\pm$0.0005	0.9713$\pm$0.0029	0.9725$\pm$0.0008
OOD	2 (2021)	0.9507	0.8152	0.8652	**0.9672**	0.9663$\pm$0.0008	0.9622$\pm$0.0048	0.9640$\pm$0.0022
OOD	3 (2022)	0.9424	0.7245	0.8063	**0.9593**	0.9574$\pm$0.0016	0.9516$\pm$0.0064	0.9554$\pm$0.0037
OOD	4 (2023)	**0.9554**	0.6797	0.7852	0.9456	0.9450$\pm$0.0019	0.9377$\pm$0.0076	0.9454$\pm$0.0047
OOD	5 (2024)	**0.9505**	0.6106	0.7280	0.9326	0.9313$\pm$0.0019	0.9222$\pm$0.0106	0.9334$\pm$0.0072
In-domain	1 (2023)	**0.9176**	0.8676	0.8831	0.9084	0.9112$\pm$0.0011	0.9103$\pm$0.0022	0.9143$\pm$0.0012
In-domain	2 (2024)	**0.9352**	0.8135	0.8537	0.9234	0.9264$\pm$0.0016	0.9220$\pm$0.0045	0.9276$\pm$0.0010

D LAND-COVER TRANSITIONS

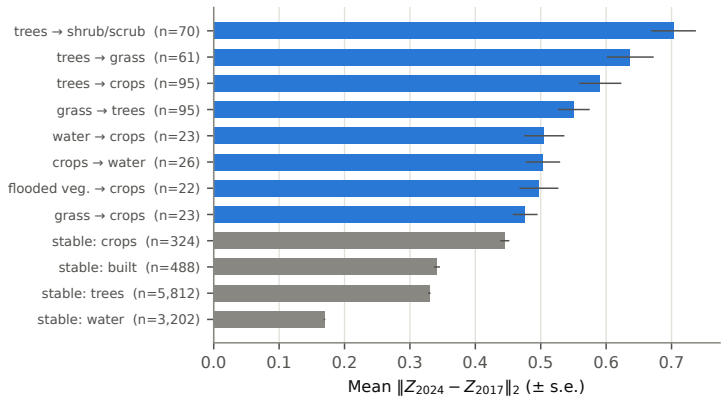


Figure 7: Mean embedding displacement 2017–2024 (±s.e.) for the most displaced observed Dynamic World transitions (blue) and the four largest stable classes (gray).

E GEOGRAPHIC TRANSFER: MATCHED CHANGE AND PER-REGION RESULTS

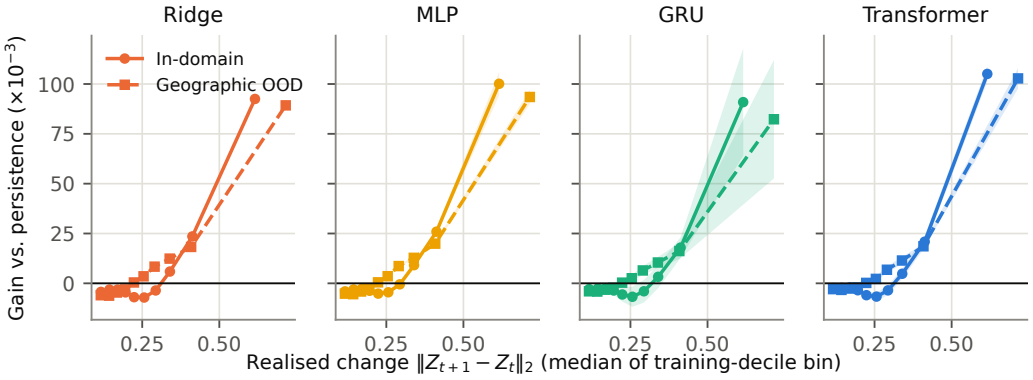


Figure 8: Gain over persistence as a function of realized change (training-decile bins), in-domain (solid) vs. held-out regions (dashed); bands: ±s.d. over seeds. At matched change, held-out gains track in-domain gains for every forecaster.

Table 10: Per held-out region (and per training region, 2024): cosine gain over persistence ($\times 10^{-3}$) with 95% seed-and-block intervals, over all locations and within the oracle dynamic stratum. n = all / dynamic. Bold: interval excludes zero.

		All locations			Dynamic stratum		
Region	n	Ridge	GRU	Transformer	Ridge	GRU	Transformer
Central Sumatra	3,951 / 1,024	**+2.7 [+0.5, +5.2]**	**+4.8 [+1.9, +8.5]**	**+5.5 [+1.6, +10.3]**	**+31.1 [+23.2, +39.9]**	**+35.3 [+12.6, +52.1]**	**+40.3 [+29.2, +52.3]**
South Sulawesi	2,088 / 703	**+12.0 [+6.6, +18.1]**	**+11.4 [+4.8, +17.3]**	**+12.4 [+6.8, +18.8]**	**+39.8 [+29.4, +49.7]**	**+36.2 [+24.4, +48.3]**	**+38.3 [+27.9, +49.4]**
West Java	3,608 / 1,640	**+7.1 [+4.6, +9.5]**	+3.5 [-3.0, +11.1]	**+6.6 [+2.2, +10.5]**	**+19.7 [+16.9, +22.8]**	**+12.3 [+3.1, +20.9]**	**+17.3 [+11.0, +22.8]**
East Kalimantan	2,907 / 610	**+8.3 [+6.1, +11.0]**	**+7.0 [+2.9, +11.8]**	**+8.3 [+4.5, +12.0]**	**+36.8 [+29.0, +45.4]**	**+32.8 [+13.4, +47.7]**	**+39.6 [+29.9, +50.8]**
Papua	3,542 / 43	**-5.8 [-6.3, -5.3]**	**-3.6 [-5.0, -2.1]**	**-2.6 [-3.3, -2.0]**	+0.8 [-7.7, +9.8]	+6.9 [-2.6, +17.8]	**+9.7 [+2.9, +19.0]**

F ABLATIONS AND PER-YEAR REPLICATION

Table 11: Transformer ablations (three seeds each; learning rate 0.001). Skill: MSE skill (%); Dyn.: cosine gain in the oracle dynamic stratum ($\times 10^{-3}$); LC ret.: land-cover retention.

	In-domain 2024				Geographic OOD 2024			
Variant	Cosine	Skill	Dyn.	LC ret.	Cosine	Skill	Dyn.	LC ret.
Default (residual, MSE+cos, k=3)	0.9572$\pm$0.0010	+4.0$\pm$3.0	+28.1$\pm$1.6	0.976$\pm$0.016	0.9744$\pm$0.0008	+1.9$\pm$2.8	+37.2$\pm$2.7	0.797$\pm$0.009
Direct target	0.9571$\pm$0.0005	+14.2$\pm$0.9	+29.5$\pm$1.9	0.892$\pm$0.013	0.9656$\pm$0.0014	-25.2$\pm$5.1	+34.5$\pm$3.8	0.726$\pm$0.025
MSE-only loss	0.9584$\pm$0.0012	+17.3$\pm$2.2	+27.5$\pm$2.1	0.936$\pm$0.014	0.9739$\pm$0.0014	+7.4$\pm$4.8	+36.9$\pm$2.3	0.708$\pm$0.033
Context k=2	0.9551$\pm$0.0015	-0.9$\pm$2.3	+24.4$\pm$2.0	0.966$\pm$0.021	0.9746$\pm$0.0005	+3.1$\pm$2.8	+36.7$\pm$2.1	0.801$\pm$0.025
Context k=1	0.9615$\pm$0.0004	+16.6$\pm$2.1	+32.7$\pm$1.2	0.965$\pm$0.016	0.9745$\pm$0.0007	+4.9$\pm$2.2	+34.3$\pm$2.3	0.768$\pm$0.012

Table 12: One-step forecasts per target year. Held-out (OOD) locations are never trained on, so every year is a replication (2020–22 overlap in-domain training *years*). ID (val) is the validation year, which is biased *towards* learned models. Cells: MSE skill (%) and dynamic-stratum cosine gain ($\times 10^{-3}$). n_{dyn}: dynamic-stratum size.

			Ridge		MLP		GRU		Transformer	
Split	Year	n_{dyn}	Skill	Dyn.	Skill	Dyn.	Skill	Dyn.	Skill	Dyn.
OOD	2020	1,644	+35.9	+42.0	+33.2$\pm$1.5	+40.7$\pm$1.0	+18.4$\pm$14.2	+35.6$\pm$7.0	+25.5$\pm$2.4	+38.6$\pm$2.0
OOD	2021	862	+15.4	+36.3	+15.4$\pm$2.1	+37.7$\pm$1.4	+6.6$\pm$18.5	+33.7$\pm$12.9	+13.6$\pm$3.4	+34.4$\pm$2.4
OOD	2022	372	-2.4	+28.7	-0.5$\pm$2.6	+31.2$\pm$2.1	-8.7$\pm$14.0	+29.1$\pm$11.2	+2.1$\pm$2.5	+33.1$\pm$2.3
OOD	2023	989	-0.8	+8.0	-1.5$\pm$2.7	+11.3$\pm$1.4	-11.8$\pm$12.4	+9.1$\pm$4.6	+0.1$\pm$1.9	+13.9$\pm$2.1
OOD	2024	653	+3.5	+34.4	+2.8$\pm$1.6	+36.2$\pm$1.1	-6.4$\pm$13.4	+31.1$\pm$9.9	+3.1$\pm$3.3	+37.6$\pm$2.2
ID (val)	2023	5,521	-9.8	-11.6	-10.9$\pm$2.0	-6.7$\pm$1.1	-17.3$\pm$8.4	-8.7$\pm$3.8	-9.0$\pm$2.1	-3.1$\pm$1.8
ID (test)	2024	3,367	+14.7	+27.4	+15.0$\pm$2.3	+31.2$\pm$1.9	+1.3$\pm$6.6	+24.3$\pm$3.9	+6.0$\pm$4.8	+28.7$\pm$1.9

G Representation control: data and additional results

Sentinel-2 inputs. For each location and year we sample monthly median composites of ten Sentinel-2 Level-2A bands (B2–B8A, B11, B12; scene-classification cloud mask) at the AlphaEarth points, giving 120 features per year. The exports used the earlier point list, so duplicated West Java rows are dropped and only audited locations are kept. South Sulawesi contributes 684 locations, from the one chunk exported before the export list was finalized. In 2017–2018, 81–99% of month-band cells are empty over our regions. From 2019 on, 9–28% are empty and are filled by within-year linear interpolation across months per band; locations with fewer than three observed months in any year are dropped (28 of 8,246). The quarterly representation uses per-band medians of *observed* months only. PCA, the autoencoder (120–256–64, 60 epochs) and the quarterly standardization are fitted on 2019–2022 training-region imagery. Learning rates are selected per representation on validation, exactly as for AlphaEarth.

Table 13: Representation control (in-domain test 2024, k=1, n = 8,218 locations). Skill: MSE skill vs. persistence (%) overall and at Dynamic World label-stable and label-changed locations. Contrast: changed − stable (percentage points). ΔLC-F1: land-cover macro-F1 on the forecast minus on persistence (×100). Brackets: 95% seed-and-block intervals; bold excludes zero. Anticip.: AUROC of forecast change $\|\hat{Z}_{t+1} - Z_t\|$ for an upcoming label change.

		MSE skill (%)					
Representation	Model	All	Stable	Changed	Contrast	ΔLC-F1	Anticip.
AlphaEarth	Ridge	+17.0	+13.7	+28.7	**+14.9 [+11.7, +18.2]**	+0.3 [-2.4, +3.1]	0.757
	MLP	+18.7±1.7	+15.2	+31.0	**+15.8 [+11.3, +21.1]**	+0.9 [-1.7, +3.4]	0.764±0.005
	GRU	+19.1±1.6	+14.9	+31.6	**+16.6 [+11.5, +21.8]**	**+2.4 [+0.0, +5.5]**	0.760±0.010
	Transformer	+13.2±1.9	+8.4	+28.0	**+19.6 [+14.7, +24.7]**	**+2.5 [+0.2, +5.2]**	0.757±0.004
PCA-S2	Ridge	+33.3	+32.2	+32.6	+0.4 [-3.2, +4.1]	**-7.2 [-10.3, -4.8]**	0.745
	MLP	+26.2±1.5	+25.5	+25.8	+0.3 [-3.5, +4.2]	**-4.2 [-7.7, -0.8]**	0.709±0.002
	GRU	+24.1±1.8	+23.5	+24.5	+1.0 [-2.5, +4.8]	**-4.0 [-7.5, -0.8]**	0.694±0.009
	Transformer	+19.9±1.0	+19.7	+15.4	**-4.3 [-7.7, -0.6]**	-0.2 [-3.5, +2.7]	0.727±0.004
AE-S2	Ridge	+36.1	+34.0	+38.9	+4.9 [-1.0, +11.1]	-2.1 [-4.7, +0.4]	0.737
	MLP	+33.7±0.3	+31.9	+36.9	+5.1 [-0.1, +10.5]	+0.2 [-3.1, +2.9]	0.726±0.001
	GRU	+32.5±0.5	+30.3	+36.4	**+6.1 [+1.7, +11.0]**	+0.6 [-2.0, +3.1]	0.719±0.004
	Transformer	+29.1±0.4	+27.5	+31.3	+3.9 [-0.3, +8.7]	-0.5 [-2.9, +2.2]	0.716±0.002
Quarterly-S2	Ridge	+31.8	+28.5	+31.7	+3.2 [-2.3, +9.1]	**-7.1 [-10.8, -3.4]**	0.700
	MLP	+30.0±1.5	+27.2	+29.7	+2.4 [-3.6, +8.7]	**-4.5 [-8.9, -0.7]**	0.685±0.008
	GRU	+29.7±0.8	+26.7	+30.1	+3.4 [-2.6, +9.2]	**-3.3 [-6.6, -0.3]**	0.681±0.010
	Transformer	+28.1±0.8	+25.3	+28.0	+2.7 [-2.8, +8.5]	-1.6 [-5.6, +1.3]	0.690±0.008

Table 14: Representation-level diagnostics on the common point set. AUROC: one-year change $\|Z_{t+1} - Z_t\|$ as a detector of Dynamic World label change (all target years 2020–2024). Ratio: median change at label-changed over label-stable locations. Probe columns: land-cover macro-F1 on the true 2024 embedding (oracle) and on the 2023 embedding (persistence), and NDVI R^2 (oracle).

Representation	Dim.	AUROC	Ratio	LC F1 (oracle)	LC F1 (persist.)	NDVI R^2
AlphaEarth	64	0.801	1.75	0.509	0.465	0.958
PCA-S2	64	0.725	1.42	0.430	0.441	0.977
AE-S2	64	0.711	1.38	0.389	0.385	0.974
Quarterly-S2	40	0.697	1.39	0.439	0.444	0.984

Table 15: Representation control on the 2023 validation year (k=1). Columns as in Table 13.

		MSE skill (%)					
Representation	Model	All	Stable	Changed	Contrast	ΔLC-F1	Anticip.
AlphaEarth	Ridge	-1.8	-4.9	+6.3	**+11.2 [+7.7, +15.0]**	-4.4 [-7.7, +4.2]	0.833
	MLP	-1.8±1.2	-5.1	+6.9	**+11.9 [+7.6, +17.0]**	-3.5 [-6.5, +4.9]	0.824±0.003
	GRU	-2.5±1.3	-6.2	+7.8	**+14.0 [+7.3, +20.3]**	-2.4 [-5.6, +6.0]	0.805±0.008
	Transformer	-5.1±1.5	-8.9	+6.2	**+15.2 [+10.4, +20.4]**	-2.6 [-5.4, +5.7]	0.802±0.005
PCA-S2	Ridge	+35.1	+35.1	+25.3	**-9.8 [-12.7, -6.9]**	**-11.8 [-14.9, -4.2]**	0.705
	MLP	+23.6±2.2	+23.0	+17.6	**-5.4 [-9.4, -1.6]**	**-9.4 [-13.1, -0.8]**	0.670±0.003
	GRU	+20.6±2.6	+20.3	+16.0	-4.3 [-8.3, +0.0]	**-9.1 [-12.2, -1.3]**	0.655±0.006
	Transformer	+15.9±0.9	+16.7	+6.7	**-9.9 [-13.7, -6.1]**	-5.2 [-7.9, +2.5]	0.711±0.001
AE-S2	Ridge	+40.9	+39.9	+27.9	**-12.0 [-16.2, -7.7]**	**-7.8 [-11.1, -0.4]**	0.686
	MLP	+37.2±0.4	+36.5	+27.0	**-9.5 [-13.4, -5.5]**	**-8.6 [-11.7, -1.4]**	0.682±0.001
	GRU	+35.0±0.6	+35.3	+27.8	**-7.5 [-11.1, -3.8]**	**-8.3 [-11.3, -0.8]**	0.677±0.002
	Transformer	+31.0±0.4	+31.5	+24.3	**-7.2 [-10.7, -3.7]**	**-9.2 [-11.8, -2.5]**	0.674±0.001
Quarterly-S2	Ridge	+34.7	+35.3	+21.0	**-14.3 [-19.4, -9.4]**	**-8.7 [-11.4, -1.3]**	0.633
	MLP	+32.7±1.5	+33.3	+18.8	**-14.5 [-19.9, -9.2]**	**-8.3 [-11.3, -0.5]**	0.625±0.009
	GRU	+31.5±0.6	+32.6	+19.4	**-13.2 [-17.9, -8.2]**	-7.5 [-10.7, +0.3]	0.629±0.010
	Transformer	+28.4±1.3	+29.6	+17.1	**-12.5 [-17.0, -7.9]**	-6.9 [-10.0, +0.7]	0.666±0.008

Table 16: Representation control with a two-year context (k=2; training targets 2021–2022; three seeds; test 2024). Columns as in Table 13.

		MSE skill (%)				
Representation	Model	All	Stable	Changed	Contrast	ΔLC-F1
AlphaEarth	Ridge	+9.0	+3.3	+24.1	**+20.8 [+16.8, +24.9]**	+1.5 [-1.4, +4.5]
	MLP	+12.7$_{\pm 2.7}$	+7.4	+27.7	**+20.3 [+15.5, +25.4]**	+1.5 [-2.4, +5.8]
	GRU	+9.5$_{\pm 1.9}$	+2.8	+25.4	**+22.6 [+17.3, +29.2]**	+0.4 [-2.2, +2.7]
	Transformer	+5.8$_{\pm 20.2}$	+2.3	+16.9	**+14.6 [+10.7, +18.4]**	-0.2 [-2.5, +1.5]
PCA-S2	Ridge	+32.9	+31.3	+32.3	+1.0 [-3.0, +5.1]	-2.0 [-5.3, +0.8]
	MLP	+20.9$_{\pm 0.2}$	+19.2	+21.8	+2.6 [-1.9, +7.4]	+2.2 [-1.6, +5.6]
	GRU	+18.5$_{\pm 1.4}$	+16.9	+16.7	-0.3 [-4.2, +3.8]	+1.6 [-1.4, +4.4]
	Transformer	+11.9$_{\pm 2.1}$	+11.3	+5.7	**-5.6 [-10.1, -1.1]**	+0.9 [-2.7, +3.7]
AE-S2	Ridge	+35.2	+33.1	+37.3	+4.3 [-1.9, +10.6]	-1.1 [-4.1, +1.8]
	MLP	+31.5$_{\pm 0.5}$	+29.5	+33.6	+4.1 [-1.4, +10.0]	+1.0 [-1.7, +3.8]
	GRU	+29.5$_{\pm 1.4}$	+27.3	+32.0	**+4.6 [+0.1, +9.6]**	+0.2 [-2.9, +3.2]
	Transformer	+26.0$_{\pm 1.8}$	+24.3	+26.2	+1.9 [-2.4, +6.9]	-0.4 [-3.9, +2.8]
Quarterly-S2	Ridge	+30.7	+27.1	+32.0	+4.9 [-1.8, +11.6]	-2.2 [-4.2, +0.1]
	MLP	+26.4$_{\pm 0.8}$	+23.2	+28.6	+5.4 [-1.4, +12.5]	-0.7 [-3.3, +2.8]
	GRU	+24.3$_{\pm 1.8}$	+21.5	+24.5	+3.0 [-4.2, +10.0]	+1.3 [-1.0, +3.9]
	Transformer	+22.4$_{\pm 1.0}$	+19.9	+21.3	+1.4 [-4.8, +7.6]	+1.0 [-2.1, +4.2]