Self-Supervised Complex-Wavelet Learning and Statistically Governed Transfer

Scientific Report 3: The V0–V7 Pretraining Ladder, the Transfer Gap, and Phase A

Jashua Luna*

25 July 2026

Abstract

Pretext outcomes in this report are training-only evidence (`E1-TRAIN`); transfer outcomes are observed-validation evidence (`E3-VAL`). We study whether masked coefficient reconstruction over ERA5 and three historical SST reconstruction products (ERSSTv5, HadISST1, COBE-SST2) improves observed forecast transfer for the coefficient-native persistence-residual Earthformer of Report 2. The V0–V7 ladder improves pretext reconstruction monotonically — the four-store V6 improves all seven tracked holdout diagnostics for every store, and the regional-auxiliary V7 further improves field, equatorial, and Niño-3.4 reconstruction, index MSE, and tendency correlation — yet the gains do not become forecast skill: under an identical supervised fine-tune, V7 *worsens* all three headline transfer metrics relative to V6 (Niño-3.4 correlation $0.1267 \rightarrow 0.0380$). A coefficient-energy audit shows why fine-scale reconstruction dominates the pretext gradient (highpasses carry 7.98% of isolated Niño-3.4-relevant variance but 98.19% of normalized coefficient loss). Motivated by small unresolvable deltas, the study also installs the project's statistical governance — an append-only consultation ledger, practical minimum effects, and a paired moving-block bootstrap — and runs Phase A, two predeclared consultation batches that freeze a conservative transfer protocol (`decoder_last_block`, learning rates $10^{-4}/5 \times 10^{-6}$, epoch-cadence validation, unweighted `residual_skill_v0`, guarded best-skill selection). Central finding: representation reconstruction improved consistently, but pretext gains did not reliably transfer; the project redirects to supervised forecast-task simulation pretraining (Report 4).

1 Study Identity and Handoff

Study ID R3. **Incoming state:** the deterministic bottleneck of Report 2. **Changed factors:** self-supervised data/task scale (V0–V7), then the transfer protocol itself (Phase A). **Evidence tiers:** `E1-TRAIN` pretext metrics; `E3-VAL` transfer comparisons — early ones exploratory, Phase A under the new bootstrap governance. **Outgoing protocol (frozen):** `decoder_last_block`; fresh/reused learning rates $10^{-4}/5 \times 10^{-6}$; validation once per completed epoch; unweighted `residual_skill_v0`; guarded `best_skill.pt` selection with Niño-3.4 correlation (leads 3–14) primary. **Timeline:** 2026-07-01 through 2026-07-03; test never accessed.

2 Introduction

Report 2 located the deterministic bottleneck in the objective/data regime: 335 supervised windows cannot teach a 1.4M-parameter decoder boundary what evolving tropical-Pacific anomalies look

*Deep Wave Research. Correspondence: `jluna@deepwave.ca`. Part of the Complex-Wavelet Earthformer ENSO report series; metric identifiers and evidence tiers are defined in the series charter (document 00).

like. Masked spatiotemporal coefficient learning is the natural response — reconstruction tasks manufacture supervision from unlabeled sequences, and historical SST reconstructions extend the record by more than a century. Two questions structure the study:

1. Do distinct historical SST reconstructions improve multiscale representation quality, measured by masked-coefficient and physical-space reconstruction diagnostics on fixed holdouts?
2. Does that quality transfer to phase/amplitude forecasting under a controlled supervised fine-tune?

The answers turn out to be "yes, consistently" and "no, not reliably" — and the divergence between them is the report's central scientific object.

3 Related Work and Data-Dependence Caveat

Masked autoencoding is the dominant self-supervised recipe for vision and video [1, 2]: hide most of the input, reconstruct it, and fine-tune the encoder. Its transfer success is typically reported for *semantic* downstream tasks; whether reconstruction quality transfers to *dynamical extrapolation* — forecasting — is far less established, and our results bear directly on that question. In climate ML, pretraining has helped when the pretext matched the downstream task: CMIP-pretrained CNNs for ENSO indices [3] and climate-simulation-pretrained ResNets for medium-range fields [4] both pretrain on (approximately) the forecast task itself. Our ladder isolates the complementary case: same architecture, same data modality, but a reconstruction pretext.

A caveat governs all multi-product claims: ERSSTv5 [5], HadISST1 [6], and COBE-SST2 [7] are distinct *reconstructions* built from substantially shared observations (ICOADS-era ship and buoy records). They diversify reconstruction methodology, gridding, and noise structure — not the set of ENSO events. Store counts below are therefore never read as independent sample sizes.

4 Common Forecast Coordinates

For every transfer comparison: tropical Pacific 30°S–30°N, 120°E–280°E on the 24×48 grid; 12-month context, 14-month target; observed train 1981–2010; observed validation 2011–2015 (35 windows); 2016–2020 reserved and untouched; cosine-weighted Niño-3.4 box with the leads-3–14 correlation summary; persistence and zero-anomaly baselines. Ensemble size: not applicable. Pretext stores extend *earlier* than 1981 only (next section).

5 Pretraining Data and Leakage Boundaries

The pretext corpus is: ERA5 1981–2010 training months (never validation/test); ERSSTv5 1854-01–1980-12 (1,524 months, 1951–1980 climatology); HadISST1 1870-01–1980-12 (1,332 months, same climatology interval); COBE-SST2 1850-01–1980-12 (1,572 months). Every external store is declared `pretrain_only`, is rejected by supervised loaders, and fails closed on any 2011–2020 overlap. Each store carries source-local per-channel normalization: ERSST's 2° sampling has a different highpass variance profile from ERA5, and applying ERA5 statistics produced an immediately rejected coefficient-scale diagnostic. Manifests record spans, month counts, coefficient/normalization hashes, sealed-period overlap flags, `allowed_use`, and `test_data_read` (Table 1).

Store	Span	Months	Role
ERA5	1981-01–2010-12	360	supervised source; training months only
ERSSTv5	1854-01–1980-12	1,524	pretrain-only; 1951–1980 climatology
HadISST1	1870-01–1980-12	1,332	pretrain-only; 1951–1980 climatology
COBE-SST2	1850-01–1980-12	1,572	pretrain-only

Table 1: The pretext corpus. Every external store is declared `pretrain_only`, carries source-local per-channel normalization, and fails closed on any 2011–2020 overlap; the three reconstructions share observational ancestry and are never counted as independent samples.

5.1 Store Sampling and Comparability

Sampling modes are proportional, exact-quota balanced, and weighted (largest-remainder); realized per-store batch counts are recorded in checkpoints (e.g. V6: 732/756/732/780 for ERA5/ERSST/HadISST/COBE). Because stores are normalized separately, pooled *normalized* losses are explicitly marked not strictly comparable across stores; cross-store discussion prioritizes inverse-normalized physical-space RMSE (`X2`) — field, equatorial band, and Niño-3.4 — which live in common SST-anomaly units.

6 Pretext Tasks and Checkpoint Contract

The pretext model is the same 177-state coefficient-native Earthformer used for forecasting, run at 12-month geometry. Corruption operators are: element masks, whole-month temporal masks, and space–time block masks, plus two-month middle/future gaps in later rungs. With mask set $\mathcal{M}$ and normalized coefficients z, the base objective is hidden-only reconstruction,

$$\mathcal{L}_{\mathrm{X1}} \;=\; \frac{1}{|\mathcal{M}|} \sum_{(t,c,h,w)\in\mathcal{M}} \big(\hat{z}_{t,c,h,w} - z_{t,c,h,w}\big)^2, \tag{1}$$

with validity-masked cells excluded. Physical diagnostics (`X2`) inverse-normalize and inverse-DTCWT merged predictions to score reconstruction RMSE for the field, equatorial band, and Niño-3.4. Train/holdout sequence partitions are deterministic; a timestamp audit immediately before array materialization rejects any month after 2010-12. Checkpoints round-trip all 177 states exactly; supervised initialization loads 172 tensors with exactly five declared 12-to-14-month temporal-geometry skips, and residual persistence equivalence after zero-output initialization is verified to be exact.

6.1 V0–V6 Reconstruction Ladder

Each rung changes one factor and exact-loads its predecessor:

Rung	Changed factor	Steps	Selected checkpoint
V1	ERA5-only masked pretraining	2,000	best holdout loss
V2	stronger tasks (block space–time + month gaps)	3,000	step 3,000
V3	+ ERSSTv5 store	1,000[†]	best holdout loss
V4	exact balanced ERA5/ERSST quotas	3,000	step 3,000
V5	+ HadISST1 (three-store balance)	3,000	step 2,750
V6	+ COBE-SST2 (four-store balance)	3,000	step 3,000
V7	+ regional auxiliary tasks (§6.2)	3,000	step 2,500

[†]reduced bound; the 186-sequence holdout made the projected runtime excessive.

6.2 V7 Regional Phase/Amplitude Auxiliary Task

V7 keeps the masked objective and adds three physical-space regional terms computed only for hidden target months: predicted values at masked locations are merged with visible targets, inverse-normalized with the store's own statistics, inverse-DTCWT'd, and reduced to cosine-weighted indices. With hidden-month index N_t, equatorial index E_t, and monthly tendency $\Delta N_t = N_t - N_{t-1}$,

$$\mathcal{L}_{\mathrm{V7}} \;=\; \mathcal{L}_{\mathrm{X1}} + 0.25\,(\hat{N}_t - N_t)^2 + 0.10\,(\hat{E}_t - E_t)^2 + 0.25\,(\widehat{\Delta N}_t - \Delta N_t)^2. \tag{2}$$

The targets remain self-supervised functions of the unlabeled SST sequence; no forecast target, validation month, or supervised optimization enters. The state schema is unchanged, so V7 checkpoints remain compatible with the supervised initialization surgery.

7 Pretraining Results (`E1-TRAIN`)

The ladder improves nearly monotonically on its fixed holdouts. Selected milestones (pooled unless stated): V1, masked MSE 1.0138 → 0.7352, field reconstruction RMSE 0.1310 → 0.0958; V2, hidden MSE 0.7774 → 0.7510 with full/equatorial/Niño-3.4 RMSE 0.1216/0.1680/0.1627 → 0.1174/0.1578/0.1544; V4 (balanced two-store), pooled hidden 0.7553 → 0.7375 with both stores improving individually; V5 (three-store), pooled hidden 0.6822 → 0.6686, with ERA5 and HadISST improving on all seven tracked diagnostics but small ERSST physical/regional regressions — the one non-monotone rung; V6 (four-store), pooled hidden 0.6990 → 0.6693, field 0.1952 → 0.1887, Niño-3.4 reconstruction 0.2648 → 0.2444, with *all seven diagnostics improving for every store* and the V5 ERSST regression not persisting. V7 then improves everything its auxiliary targets: field 0.1887 → 0.1796, equatorial 0.2278 → 0.2040, Niño-3.4 reconstruction 0.2444 → 0.2063, index MSE 0.0460 → 0.0294, tendency MSE 0.0378 → 0.0271, tendency correlation 0.8016 → 0.8524 — all ten tracked diagnostics, every store, no cross-source regression. These are representation diagnostics, not forecast skill.

8 Observed Transfer Experiments (`E3-VAL`)

Three pretrained initializations receive an *identical* bounded supervised fine-tune (coefficient-native, persistence-residual, `decoder_last_block`, unweighted `residual_skill_v0`, 1,000 CUDA steps, guarded selection; all selected step 335):

Initialization	`D1` RMSE	`D3` ACC	`D4` Niño-3.4 corr	`D6` full/eq/Niño-3.4	2015 final (°C)
V2-pretrained	0.5478	+0.2699	+0.0422	0.464/0.462/0.346	+0.252
V6-pretrained	**0.5432**	**+0.2876**	**+0.1267**	0.465/0.441/0.320	+0.221
V7-pretrained	0.5450	+0.2768	+0.0380	0.465/0.438/0.319	+0.204
non-pretrained residual (R2)	0.5922	+0.1345	+0.3268	0.482/0.456/0.343	+0.352

Table 2: Identical supervised fine-tunes from three pretraining rungs (`E3-VAL`, exploratory; 2015 values are case-study-only, observed final index +2.640 °C). Pretraining improves field RMSE and ACC over the non-pretrained run but *costs* index correlation; and V7 — the best pretext checkpoint by every diagnostic — is worse than V6 on all three headline metrics.

The decisive contrast is V6 versus V7 (Fig. 1): the auxiliary tasks improved exactly the pretext quantities that look forecast-relevant (hidden-month index and tendency reconstruction)

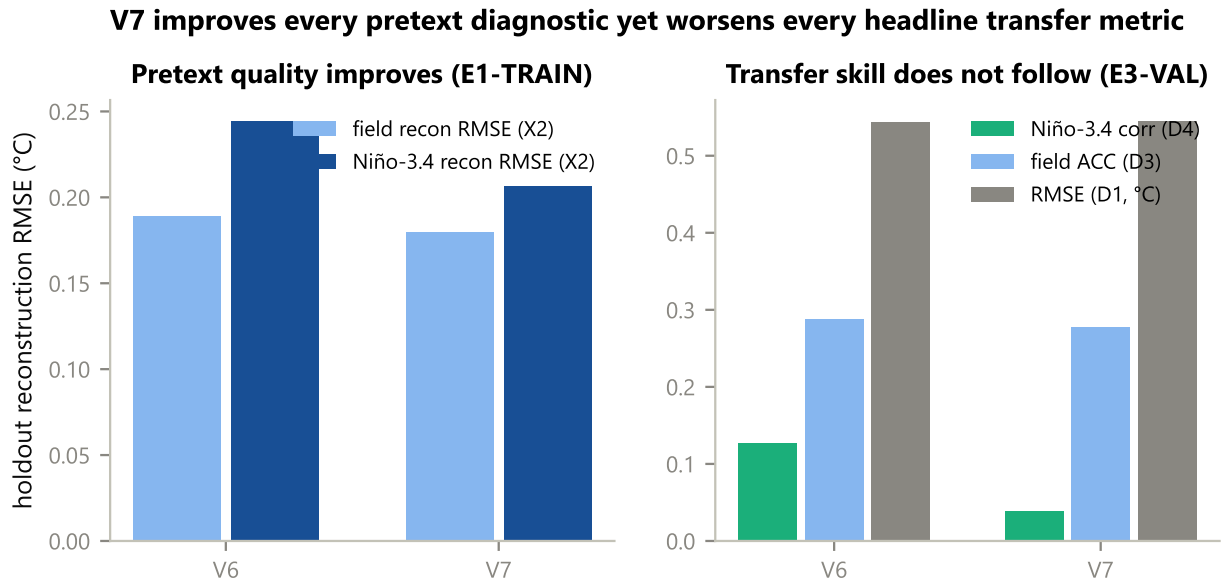

Figure 1: The reconstruction-to-forecast transfer gap. Left: V6→V7 improves holdout reconstruction (`X2`; lower is better). Right: the same step worsens every headline transfer metric (`D1`, `D3`, `D4`).

and still *worsened* transfer — Niño-3.4 correlation 0.1267 → 0.0380, ACC 0.2876 → 0.2768, RMSE 0.5432 → 0.5450. Cancellation eased marginally under V7 (cosine −0.851 → −0.837; opposing energy 96.2 → 95.6%) while amplitude stayed damped and 2015 worsened. Reconstruction quality, even regionally targeted, is not the transfer bottleneck.

9 Measurement and Decision-Governance Reform

Table 2's deltas are small — 0.002 °C of RMSE, a few points of correlation — and the study's early rungs were being compared without uncertainty. This forced the governance reform that the rest of the project runs on:

- **Append-only consultation ledger.** Every observed-validation family comparison is entered *before* results are inspected, declaring candidate, fixed reference, primary metric, guards, practical minimum effects, bootstrap settings, decision branches, checkpoint-selection source, and the consultations remaining in the phase (budgeted at two per phase for Phases A–D).
- **Paired moving-block bootstrap.** For metrics recomputed from the same ordered 35 windows in both runs: blocks of 6 consecutive windows, 10,000 resamples, 90% percentile intervals, seed 20260703, identical resampled indices applied to both runs. Aggregate summaries are never expanded into synthetic samples; a missing per-window artifact is reported as missing.
- **Decision language.** *Resolvable improvement* requires the CI to exclude zero favorably *and* the predeclared minimum effect; an interval containing zero is *unresolved*, never a win or loss. Mixed primary/guard outcomes follow the predeclared branch.
- **Selection separation.** Checkpoint selection moves off 2011–2015 for Phases B–D (held-out simulations or a frozen training-tail holdout); 2011–2015 is reserved for predeclared family-level comparisons.

- **Case-study boundary.** The 2015 event may be reported and rendered, but its numbers are never selectors, family criteria, stop rules, or minimum effects.

A retrospective decision-resolvability audit re-examined the ladder's earlier choices under paired bootstrap where artifacts permitted, distinguishing decisions the data actually resolved from those made on unresolved deltas — the direct motivation for Phase A's predeclared design.

10 Phase A Transfer-Protocol Study

Phase A selects a *reusable transfer protocol* for Phases B–C — not a forecast model — through two predeclared consultation batches under the §9 machinery. Primary: Niño-3.4 correlation, leads 3–14; guards: pooled RMSE and mean-window field ACC; adoption thresholds: primary CI > 0 with delta $\geq +0.02$, no resolvable RMSE regression $\geq +0.005\,^{\circ}\mathrm{C}$, no resolvable ACC regression ≤ -0.01 (Fig. 2).

10.1 A1: Full-Backbone Low-Rate Transfer

Candidates make all 166 optimizable parameters trainable (4 fresh at 10^{-4}; 162 reused, including 87 encoder parameters, at 10^{-6} or 10^{-5}) against the fixed V7 `decoder_last_block` reference. Outcomes: lr 10^{-6} (selected step 1675) — RMSE 0.5756, primary delta $+0.219$ $[-0.077, +0.797]$ unresolved, RMSE delta $+0.0306$ $[+0.0174, +0.0430]$ *resolvably worse*; lr 10^{-5} (step 670) — primary -0.030 $[-0.388, +0.708]$ unresolved, RMSE $+0.0323$ $[+0.0165, +0.0415]$ resolvably worse. Neither meets the rule; the smaller boundary is retained and the conditional V6 replication branch is not triggered.

10.2 A2: Validation Cadence and Headroom

Report 2's repeated step-335 selections (the first epoch boundary) suggested the fine-tune saturates in one epoch. A2 tests every-84-step validation against epoch cadence on V6. The finer cadence *does* find a pre-335 within-run peak (selected step 84: RMSE 0.5458, ACC $+0.2897$, Niño-3.4 $+0.2330$), refining saturation as partly an observation-cadence artifact — but at the family level the primary is unresolved ($+0.106$ $[-0.141, +0.218]$) and the ACC guard *resolvably regresses* (-0.0260 $[-0.0483, -0.0156]$). Epoch cadence is retained.

10.3 A3: Coefficient Energy and Scale Weighting

A training-only audit reconstructs physical maps from isolated coefficient groups (zeroing all others; isolated shares do not allocate cross-group DTCWT covariance). Normalized isolated Niño-3.4-relevant variance shares for lowpass/highpass-1/2/3 are 92.02/0.00/0.42/7.56%, while mean normalized coefficient SSE shares across selected V6/V7 checkpoints are 1.81/76.64/17.70/3.84%: the highpasses carry $\mathbf{7.98}\%$ of isolated index-relevant variance and $\mathbf{98.19}\%$ of the loss. This triggered the one predeclared scale-weighted run (V7, weighted coefficient MSE): RMSE improved resolvably (-0.0037 $[-0.0053, -0.0004]$) but the primary did not resolve ($+0.023$ $[-0.080, +0.127]$), so the unweighted loss remains frozen. The audit's significance outlives Phase A: the same loss-allocation imbalance, remeasured on correction temporal structure, becomes the causal lever of Phase G (Report 8).

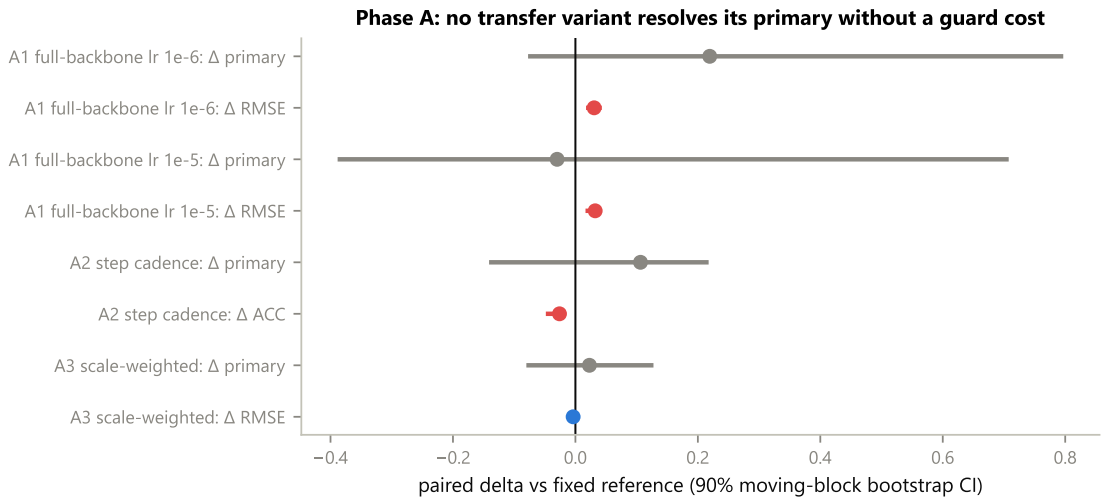

Figure 2: Phase A paired deltas with 90% moving-block bootstrap CIs (`E3-VAL`, predeclared). Blue: resolvable in the favorable direction; red: resolvable regression; gray: unresolved. No candidate resolves its primary without a resolvable guard cost, so the conservative protocol is frozen.

Element	Frozen value
Trainable boundary	`decoder_last_block` (53 tensors / 223,621 parameters)
Learning rates	10^{-4} fresh / 5×10^{-6} reused
Validation cadence	once per completed epoch
Objective	unweighted `residual_skill_v0`
Selection	guarded `best_skill.pt`; primary Niño-3.4 corr. leads 3–14

Table 3: The frozen Phase A transfer protocol, adopted after A1–A3 and reused verbatim by every observed transfer in Reports 4–5. Each element is the conservative branch of its predeclared comparison: no candidate change resolved its primary without a resolvable guard cost.

11 Persistent Series Scorecard

Pretraining (`E1-TRAIN`): `X1` masked-coefficient and `X2` physical/regional reconstruction diagnostics by store, as summarized above; per-store values live in the run summaries and cross-source diagnostic reports. Transfers (`E3-VAL`): Table 2 carries `D1`–`D6`; Phase A adds paired deltas and CIs (Fig. 2). `X4` provenance: V6 checkpoint SHA-256 `5a81ab7a...`, V7 `16b4a801...`; every transfer audit records 172 loaded / 5 skipped tensors, the 53/223,621-parameter trainable boundary, 113/1,193,780 frozen, exact residual-persistence equivalence, and `test_data_read: false`.

12 Discussion

Why does reconstruction fail to transfer while its diagnostics improve? Four mechanisms are consistent with the record. (i) *Objective mismatch*: masked reconstruction learns a conditional *interpolant* of hidden months given visible ones; forecasting requires a conditional *extrapolant* 14 months forward. The interpolant's sufficient statistics need not contain the extrapolant's. (ii) *Fine-scale gradient dominance*: with per-channel normalization, 98% of pretext gradient serves bands carrying 8% of index-relevant variance (A3) — representation capacity is spent where the downstream task needs it least. (iii) *Normalization and source heterogeneity*: store-local statistics

preserve each product's variance profile, but the supervised task lives in ERA5's; cross-store averaging can pull shared features toward reconstruction-generic solutions. (iv) *Short fine-tuning*: 335 windows, one effective epoch of adaptation — but A1/A2 show that opening the backbone or refining cadence does not release the constraint. V7 sharpens the puzzle: a *correct local task* (hidden-month index and tendency) still fails, because reconstructing a hidden month among visible neighbors is far easier than predicting it from the past alone. The decisive next intervention is supervised *forecast-task* pretraining at simulation scale — same architecture, same residual target, abundant trajectories — which Report 4 executes.

13 Limitations

Shared observational ancestry across SST products; distinct per-store normalization limiting pooled-loss comparability; unequal effective event counts across stores; a single observed validation slice (35 overlapping windows) consulted repeatedly; adaptive rung-to-rung choices in the ladder (V0–V6 predate the ledger's bootstrap discipline); no random-seed replication of pretraining runs; and bootstrap governance covering transfer comparisons only — pretext metric uncertainty is not quantified. The 2016–2020 split is untouched throughout.

14 Conclusion and Transition to Report 4

The frozen Phase A protocol (Table 3) — `decoder_last_block`, $10^{-4}/5 \times 10^{-6}$, epoch cadence, unweighted `residual_skill_v0`, guarded best-skill selection — is this study's durable product, along with a negative lesson now measured three ways: pretext reconstruction quality, even with regional phase/amplitude auxiliaries, does not become observed forecast skill in this regime. Report 4 inherits the coefficient-native checkpoint bridge, the persistent scorecard, the held-out-selection principle, and the explicit hypothesis that training on the *actual* 12-to-14-month forecast task across CMIP6 models will transfer where masked reconstruction did not.

A Evidence Inventory

The paths below name the primary records in the project's internal research archive; per-store reconstruction diagnostics and the Phase A ledger entries are stored with the corresponding run and evaluation directories.

```
docs/pretraining.md
docs/next_phase_plan.md
docs/phase_a_transfer_protocol_plan.md
docs/model_selection_policy.md
outputs/evaluations/
  decision_resolvability_audit.md
  phase_a/phase_a_summary.md
  coeff-native-residual-skill-v0-v6pretrained-1000step-validation/
  coeff-native-residual-skill-v0-v7auxpretrained-1000step-validation/
outputs/reports/coefficient_energy_audit.md
outputs/runs/
  coeff-native-external-mixed-pretrain-v4-3000step/
  coeff-native-external-mixed-pretrain-v5-3000step/
```

```
coeff-native-external-mixed-pretrain-v6-3000step/
coeff-native-external-aux-pretrain-v7-3000step/
```

References

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[2] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[3] Y.-G. Ham, J.-H. Kim, and J.-J. Luo, "Deep learning for multi-year ENSO forecasts," *Nature*, 573:568–572, 2019.

[4] S. Rasp and N. Thuerey, "Data-driven medium-range weather prediction with a ResNet pretrained on climate simulations: A new model for WeatherBench," *Journal of Advances in Modeling Earth Systems*, 13:e2020MS002405, 2021.

[5] B. Huang et al., "Extended Reconstructed Sea Surface Temperature, version 5 (ERSSTv5)," *Journal of Climate*, 30:8179–8205, 2017.

[6] N. A. Rayner et al., "Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century," *Journal of Geophysical Research*, 108(D14):4407, 2003.

[7] S. Hirahara, M. Ishii, and Y. Fukuda, "Centennial-scale sea surface temperature analysis and its uncertainty," *Journal of Climate*, 27:57–75, 2014.