Deterministic Complex-Wavelet SST Forecasting under Limited Observations

Scientific Report 2: Architecture, Objective, and Persistence-Residual Ablations

Jashua Luna*

25 July 2026

Abstract

All outcomes in this report are repeatedly consulted observed-validation evidence (`E3-VAL`); none is a held-out skill claim. We ask whether direct coefficient-native Earthformer adaptation can forecast tropical-Pacific SST evolution from 335 observed training windows (1981–2010), evaluated on 35 overlapping validation windows (2011–2015). Nine controlled families are studied: the one-channel bridge versus native construction; four audited trainability boundaries; off/weak/medium/calibrated amplitude losses; skill-aware direct objectives (skill-v0, skill-v1); persistence-residual targets with zero-output initialization; a longer residual confirmation; and a target-aware anti-cancellation penalty. The scientific finding is uniform in direction: low RMSE was repeatedly achieved by near-climatological damping (field standard-deviation ratios as low as 0.02), attempts to restore amplitude traded pattern and Niño-3.4 correlation away, and persistence-residual learning — while producing genuinely structured corrections — spent most of its energy canceling persistence (residual/persistence cosine ≈ -0.79 to -0.83; up to 94% opposing energy). The report closes each family under predeclared stop rules and concludes that the deterministic objective/data regime, not architectural access, is the binding constraint — motivating the representation-learning and pretraining program of Report 3.

1 Study Identity and Handoff

Study ID R2. **Incoming baseline:** the validated system of Report 1. **Changed factors:** coefficient-native architecture, trainability boundary, deterministic objective, and target parameterization. **Evidence tier:** `E3-VAL`, repeatedly consulted, prior to the later bootstrap governance — these comparisons are exploratory validation evidence. **Terminal decision:** predeclared stops for the amplitude-loss, skill-curriculum, plain-residual, and anti-cancellation families. **Outgoing finding:** the deterministic objective/data regime is the bottleneck; hand forward coefficient-native persistence-residual modeling, the enriched scorecard, and the case for more representation experience. **Timeline:** 2026-06-30 through 2026-07-02; test never accessed.

2 Introduction

RMSE against persistence alone is an unsafe skill criterion for anomaly forecasting. Persistence carries full-amplitude anomalies that decorrelate with lead, so a forecast that simply damps toward

*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).

climatology (zero anomaly) scores a *lower* RMSE without tracking any evolution. A nearly uniform field can therefore "beat persistence" while possessing no pattern or phase skill. This report's first result is a concrete instance, and the report exists to characterize the phenomenon and its attempted cures under controlled changes.

We test five hypotheses, each mapped to a family below: (H1) native coefficient access (all 37 channels through the official backbone) improves expressivity over the one-channel bridge; (H2) widening the decoder-side trainability boundary restores anomaly amplitude; (H3) explicit amplitude penalties can restore variance without destroying accuracy; (H4) skill-aware objectives (correlation and index terms) can buy phase skill at acceptable amplitude; (H5) persistence-residual parameterization, with exact persistence at initialization, lets the model learn corrections rather than reconstruct the field.

3 Data, Baselines, and Fixed Protocol

The ERA5 setup of Report 1 is fixed throughout: 12-month context, 14-month target, 335 training windows, 35 validation windows, train-only climatology and normalization, persistence and zero-anomaly baselines. Persistence pooled validation RMSE is 0.6558 °C; zero anomaly is 0.5940 °C — note that zero anomaly is the *stronger* baseline, a fact invisible until it was added to the scorecard mid-study (§5). These comparisons preceded the paired-bootstrap governance introduced in Report 3; point deltas here are read qualitatively.

3.1 Common Forecast Coordinates

Domain 30°S–30°N, 120°E–280°E; 24×48 grid; 12 context and 14 target months; train 1981–2010; validation 2011–2015 (35 windows); test 2016–2020 untouched; cosine-weighted Niño-3.4 box (5°S–5°N, 190°E–240°E) with the leads-3–14 correlation summary; persistence and zero-anomaly baselines. Ensemble size: not applicable.

4 Model Variants

4.1 One-Channel Bridge and Coefficient-Native Earthformer

The bridge mode wraps the released one-channel ICAR backbone with project-owned $37 \to 1$ and $1 \to 37$ projections; it validated the pipeline but forces all coefficient structure through one channel. The native mode constructs the official model at $[12, 24, 48, 37] \to [14, 24, 48, 37]$ with shape-compatible checkpoint reuse (173/177 tensors) and fresh input-convolution/output-head boundary tensors (Report 1). All families below are native unless stated.

4.2 Trainability Boundaries

Four audited policies define the trainable set: `output_head_only` (the four fresh boundary tensors); `output_head_and_final_norm` (+ final decoder normalization; 6 tensors / 23,909 parameters); `decoder_last_block` (+ output-side decoder stage 0; 53 tensors / 223,621 parameters); and `decoder_output_side` (+ upsampling and final decoder; 61 tensors / 371,397 parameters). Fresh and reused groups take separate learning rates (10^{-4} / 5×10^{-6} unless noted); encoder tensors are never trainable; every run's audit records trainable names, frozen counts, and a maximum frozen gradient of exactly zero.

4.3 Direct and Persistence-Residual Targets

Direct runs predict future coefficients $z_{1:14}$ outright. Residual runs predict departures from coefficient persistence p (the final context month repeated over 14 leads):

$$r_{1:14} \;=\; z^{\text{true}}_{1:14} - p, \qquad \hat{z}_{1:14} \;=\; p + \hat{r}_{1:14}, \tag{1}$$

with the native residual head (`dec_final_proj`) zeroed after checkpoint reuse, so the step-zero forecast is *exactly* persistence (verified: zero residual norm and zero coefficient/map difference). Future coefficients never enter the baseline, so (1) has no temporal leakage.

5 Evaluation Reform inside the Study

The study began with RMSE-versus-persistence and amplitude diagnostics; it ended with the enriched scorecard that the whole series now uses. Mid-study, the evaluation added: the zero-anomaly baseline; field ACC by lead (per-window spatial mean and pooled space–time, cosine-weighted); Niño-3.4 correlation and RMSE by lead across windows (undefined correlations reported as undefined, never zero-filled); and the three amplitude ratios (`D6`). A retrospective audit (2026-07-01) re-evaluated every prior checkpoint under the new metrics, reproducing each original RMSE exactly and revealing that the RMSE "winners" had essentially no pattern skill: the fresh-layer and amp035 runs beat both baselines on RMSE with mean field ACC near or below zero (-0.049, -0.003), while the amplitude-preserving bridge-residual run had the strongest correlations (field ACC 0.29, Niño-3.4 0.24) and the worst RMSE (0.665). No checkpoint both retained amplitude and forecast the 2015 El Niño. The audit changed interpretation only; it altered no checkpoint.

6 Experiment Matrix and Gates

7 Results

Results are grouped causally; RMSE is always paired with correlation and amplitude so a damped forecast cannot masquerade as a successful one (Fig. 1).

7.1 Fresh Layers and Decoder Adaptation (H1, H2)

Training only the four fresh native boundary tensors for 500 steps reached pooled RMSE 0.5636 versus persistence 0.6558 (30/35 windows better), but with full/equatorial/Niño-3.4 amplitude ratios of **0.0218**/0.0172/0.0131: the model learned a lower-error, near-climatological field, not evolving anomaly structure. Representative maps are nearly uniform and miss the developing 2015 warming entirely. Widening to `decoder_output_side` (61 tensors) raised the ratios to 0.3813/0.4388/0.3107 — amplitude *can* be restored by adaptation depth — but pooled RMSE worsened to 0.5857, lead 1 fell behind persistence, and the 2015 event remained severely damped. H1 holds in the narrow sense (native access is workable and efficient); H2 buys amplitude only at accuracy cost.

7.2 A–E and amp035 Amplitude Ablations (H3)

Five 500-step ablations crossed two boundaries (`output_head_and_final_norm`, `decoder_last_block`) with amplitude presets off/weak/medium, under min-std guards. The amplitude loss adds two-sided standard-deviation matching and bias terms at preset weights. Outcomes: A/B/C/D retained RMSE near 0.560 but stayed collapsed (minimum regional ratio < 0.022); medium-weight E lifted

Run (500–5000 steps)	Trainable set	Objective change	Selector	Guards / stop rule
fresh-layers v0	4 boundary tensors	none (coef+field MSE)	val. loss	none (baseline family)
decoder-unfreeze v0	61 t. / 371k p.	+ amplitude terms	guarded amp.	amplitude-guarded checkpoint
A / B (head+norm)	6 t. / 24k p.	amp. off / weak	loss, RMSE	min-std guards 0.10/0.20
C / D / E (last block)	53 t. / 224k p.	amp. off / weak / medium	loss, RMSE	same
amp035	53 t. / 224k p.	std wt. 0.035, bias 0.0075	best RMSE	0.10 guard; fallback rule
skill-v0	53 t. / 224k p.	+ ACC, Niño-3.4, variance, bias	Niño-3.4 corr 3–14	positive-ACC + amplitude
skill-v1 curriculum	53 t. / 224k p.	hinge floors + curriculum	guarded skill	floors + pos. corr + RMSE ceiling; stop on corr collapse
residual-skill v0 (1k)	53 t. / 224k p.	residual target (1)	Niño-3.4 corr 3–14	pos. ACC, pos. corr, RMSE $\leq$ pers.
residual confirmation (5k cap)	53 t. / 224k p.	unchanged	frozen criteria	early stop (fired at 1675)
anticancel v0 (2k)	53 t. / 224k p.	+ excess-cancellation term	unchanged	joint stop rule

Table 1: The controlled experiment matrix. Every run is CUDA, bounded, and validation-only; every audit reports zero encoder trainables, zero bridge tensors, and zero maximum frozen gradient.

the ratios to 0.0543/0.0379/0.0344 at RMSE 0.5640. The final calibration (amp035: std weights 0.035, bias 0.0075) reached ratios 0.0770/0.0460/0.0448 at best-RMSE 0.5608 (31/35 windows beat persistence), rising to 0.0867/0.0516/0.0488 by step 500 while RMSE deteriorated — and the 0.10 guard *never passed*, so evaluation used the predeclared `best_rmse` fallback. No candidate cleared the regional amplitude guard while preserving the intended skill. H3 is closed for this family: simple amplitude-weight tuning cannot fix the collapse.

7.3 Skill-Aware Direct Objectives (H4)

skill-v0 makes the scorecard gradient-bearing:

$$\mathcal{L}_{\text{skill}} \;=\; \lambda_c\,\mathcal{L}_{\text{coef}} + \lambda_f\,\mathcal{L}_{\text{field}} + \lambda_a\,(1-\text{ACC}) + \lambda_n\,\mathcal{L}_{Niño-3.4} + \lambda_v\,\mathcal{L}_{\text{var}} + \lambda_b\,\mathcal{L}_{\text{bias}}, \tag{2}$$

with correlation losses epsilon-guarded so a constant prediction receives finite ACC loss 1. Selected by Niño-3.4 correlation, skill-v0 reached RMSE 0.5528 (zero anomaly 0.5940), mean field ACC +0.138, and Niño-3.4 correlation +0.365 — the family's best phase result — at amplitude ratios of only 0.041/0.046/0.033. skill-v1 replaced the symmetric variance term with one-sided hinge floors,

$$\mathcal{L}_{\text{amp}} \;=\; \text{ReLU}\Big(\phi - \tfrac{\sigma(\hat{y})}{\sigma(y)}\Big)^2, \tag{3}$$

raised by a step curriculum together with the correlation weights. The guarded checkpoint passed its amplitude floors (0.1345/0.1336/0.1023; RMSE 0.5665 below zero anomaly) but correlation collapsed: ACC +0.024, Niño-3.4 +0.117, and the 2015 final/peak prediction was +0.065 °C versus +2.64 °C observed. The pair is a clean accuracy–amplitude exchange, not a monotone improvement; the declared stop rule rejected further amplitude-weight escalation.

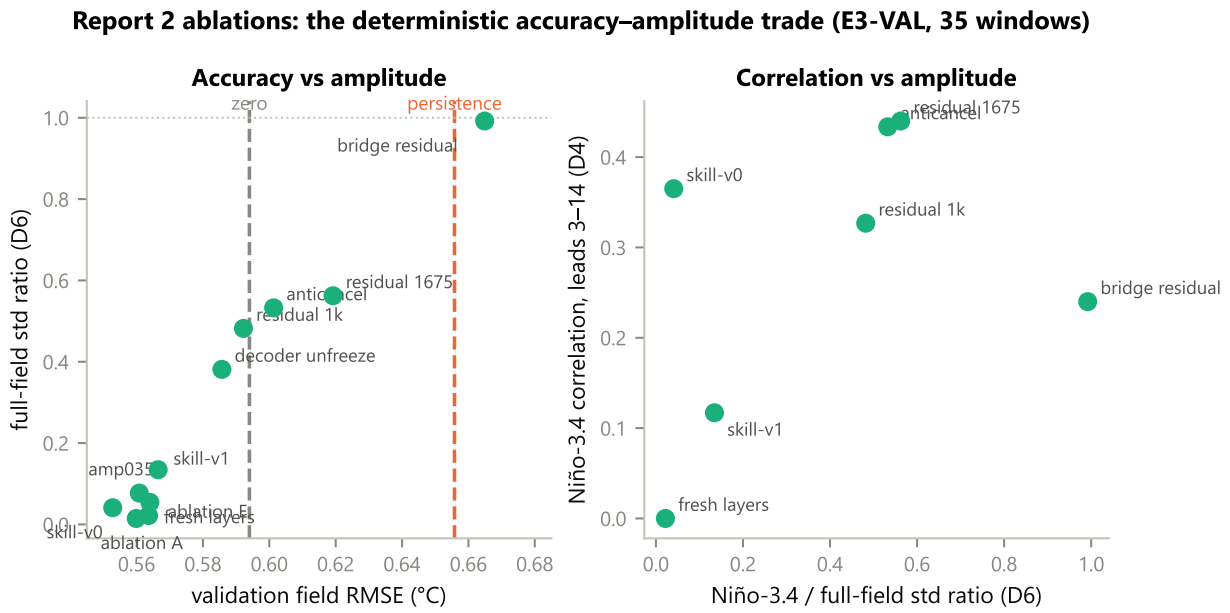

Figure 1: The deterministic trade, all decision-relevant runs (`E3-VAL`). Left: validation field RMSE versus full-field amplitude ratio (`D6`); dashed rules mark the zero-anomaly and persistence baselines. The low-RMSE cluster is nearly amplitude-free; amplitude-bearing runs pay RMSE. Right: Niño-3.4 correlation versus amplitude for runs where the correlation is defined — no run reaches high correlation at realistic amplitude.

7.4 Persistence-Residual Learning (H5)

The 1,000-step residual run (Eq. 1; step-zero persistence verified exact) selected step 335 — exactly the first epoch boundary — and produced RMSE 0.5922 (24/35 windows beat persistence; zero anomaly 0.5940), field ACC +0.1345, Niño-3.4 correlation +0.3268. Its residuals were genuinely structured: predicted residual standard deviation 0.4204 °C (0.646× the target residual), residual field ACC +0.4758. But the learned correction largely canceled persistence — final amplitude ratios 0.482/0.456/0.343 — and the 2015 final Niño-3.4 prediction was +0.352 °C versus +2.64 °C, worse than persistence on that window.

7.5 Longer Training and Cancellation Audit

The frozen-criteria confirmation stopped at step 1,675 (= 5 × 335) under validation-loss early stopping. Selected metrics: Niño-3.4 correlation 0.3268 → **0.4399**, but RMSE 0.5922 → 0.6192 (now behind zero anomaly), field ACC 0.1345 → 0.0758, 20/35 windows. The cancellation module (`X3`) makes the mechanism explicit: residual/persistence correlation −0.8278, cosine −0.7862, 93.96% of predicted residual energy opposing persistence, final variability 0.5623× persistence. The 2015 final/peak forecast improved only to +0.450/ + 0.622 °C. The predeclared pivot rule fired: stop plain residual-skill tuning.

7.6 Target-Aware Anti-Cancellation

Cancellation is not inherently wrong — decay toward climatology *requires* a residual opposing persistence. The anti-cancellation term therefore penalizes only cancellation in excess of what the

target itself requires. Per sample and lead, with cosine-latitude-weighted projections,

$$\beta \;=\; \frac{\langle r,\, p\rangle}{\langle p,\, p\rangle + \epsilon}, \qquad a \;=\; \mathrm{ReLU}(-\beta), \qquad \mathcal{L}_{\mathrm{xc}} \;=\; \Big\langle \mathrm{ReLU}\big(a_{\mathrm{pred}} - a_{\mathrm{tgt}} - \tau\big)^2 \Big\rangle, \tag{4}$$

with weight 0.10 and margin $\tau = 0.05$; zero-persistence slices are marked inactive. The bounded run (selected step 1,340) gave RMSE 0.6013 (still behind zero anomaly), ACC +0.1210, Niño-3.4 correlation +0.4335. Excess cancellation fell (0.0418 → 0.0361 versus the plain step-1675 checkpoint), but cosine anti-alignment *worsened* (−0.7862 → −0.8080), final/persistence variability fell (0.5623 → 0.5326), and 2015 stayed missed (+0.405 °C). The joint stop rule closed the loss family: the penalty moved its own target without moving amplitude or accuracy.

8 Persistent Series Scorecard

Run	`D1` RMSE	vs pers.	vs zero	`D3` ACC	`D4` Niño-3.4	`D6` full/eq/Niño-3.4
persistence	0.6558	—	+0.0618	(by lead)	(by lead)	≈ 1
zero anomaly	0.5940	−0.0618	—	undefined	undefined	0
fresh layers	0.5636	−0.0922	−0.0304	−0.049	≈ 0	0.022/0.017/0.013
decoder unfreeze	0.5857	−0.0701	−0.0083	n/r	n/r	0.381/0.439/0.311
amp035 (best RMSE)	0.5608	−0.0950	−0.0332	−0.003	≈ 0	0.077/0.046/0.045
skill-v0	0.5528	−0.1030	−0.0412	+0.138	+0.365	0.041/0.046/0.033
skill-v1 (guarded)	0.5665	−0.0893	−0.0275	+0.024	+0.117	0.135/0.134/0.102
residual 1k (step 335)	0.5922	−0.0636	−0.0018	+0.135	+0.327	0.482/0.456/0.343
residual 1675	0.6192	−0.0366	+0.0252	+0.076	+0.440	—/—/0.562*
anticancel (step 1340)	0.6013	−0.0545	+0.0073	+0.121	+0.434	—/—/0.533*

Table 2: Persistent scorecard, `E3-VAL`, exploratory (pre-bootstrap). "n/r" = not recorded under the enriched scorecard at run time; *final-to-persistence variability from the cancellation module (`X3`), not the `D6` target-ratio estimator. 2015 case values are kept out of this table by rule.

The `X3` module for the residual family: cosine −0.786 (1k) / −0.828 correlation (1675) / −0.808 (anticancel); opposing energy 93.96–96%; excess cancellation 0.0418 → 0.0361 under (4). `X4` provenance: all runs 53 tensors / 223,621 parameters trainable (except decoder-unfreeze 61/371,397 and A/B 6/23,909), 113 tensors / 1,193,780 frozen at the last-block boundary, epoch-boundary selection, CUDA, `test_data_read: false`.

9 Synthesis of Ablations

Falsified: (i) that low RMSE indicates skill — the audit shows the RMSE winners are near-climatological (H1's success is expressivity, not skill); (ii) that decoder-side capacity restores useful amplitude (H2: amplitude without accuracy); (iii) that deterministic amplitude pressure at any tested weight preserves correlation (H3, H4: a clean exchange in both directions); (iv) that penalizing excess cancellation redirects the residual solution (the penalty's own target moved; the mechanism did not).

Supported: persistence-residual parameterization learns genuinely structured corrections (residual ACC +0.48) and yields the family's best index correlations — but under this objective the optimal correction is predominantly anti-persistence damping.

Unresolved by design: whether the constraint is information (335 windows), representation experience, or the deterministic frame itself. The record supports a data/representation intervention next — masked multi-product pretraining (Report 3) and supervised simulation pretraining

(Report 4) — rather than further loss-weight search; the deterministic-frame question is taken up by the probabilistic program (Reports 6–8), whose launch document explicitly builds on this report's stop rules.

10 Threats to Validity

The 35 validation windows overlap heavily and were consulted repeatedly across families; metric expansion occurred mid-study (the audit mitigates but cannot undo selection under the earlier metric); baselines are sparse (persistence, zero); runs are short and bounded (335-step epochs; several selections at the first epoch boundary, later shown by Phase A to be partly a cadence-of-observation artifact); no multiple-comparison control existed before the Report-3 governance; and the 2015 event dominates qualitative judgment — it is reported as a labeled case study and never used as a selector. None of these results accesses the test split.

11 Conclusion and Transition to Report 3

The diagnosed failure mode is now precise: under squared-error-dominated deterministic objectives on 335 windows, the model's optimal move is climatological damping; amplitude can be purchased only with correlation; and residual learning converts the same pressure into persistence cancellation. Frozen families — amplitude-weight tuning, skill-v1 curricula, plain and anti-cancellation residual variants — are closed and must not be re-searched. Report 3 inherits the coefficient-native persistence-residual architecture, the enriched scorecard, and the hypothesis that the model needs more *representation experience* than 360 observed months can provide, testing masked multi-product SST pretraining and, when that fails to transfer, handing the task-alignment hypothesis to Report 4.

A Evidence Inventory

The paths below name the primary records in the project's internal research archive; per-run metrics, guards, and selection records are stored with each evaluation directory.

```
docs/experiment_log.md
outputs/evaluations/
  skill_metric_audit.md
  coeff-native-partial-unfreeze-ablation-summary.md
  coeff-native-fresh-layers-v0-500step-validation/
  coeff-native-decoder-unfreeze-v0-500step-validation/
  coeff-native-decoder-lastblock-amp035-v0-500step-validation/
  coeff-native-skill-v0-500step-validation/
  coeff-native-skill-v1-curriculum-1000step-validation/
  coeff-native-residual-skill-v0-1000step-validation/
  coeff-native-residual-skill-v0-5000step-validation/
  coeff-native-residual-anticancel-v0-2000step-validation/
```