# Simulation-Scale Supervised Pretraining for Sealed Deterministic ENSO Skill

Scientific Report 4: Phase B, the PB-1 Transfer, and the First Hash-Gated Test

Jashua Luna*

25 July 2026

**Abstract**

Simulation results here are held-out climate-model evidence (`E2-SIM`), PB-1's selection is observed-validation evidence (`E3-VAL`), and the final result is the project's pristine sealed-test evidence (`E4a-TEST`). We test whether supervised forecast-task pretraining on CMIP6 historical simulations can overcome the observed transfer gap documented in Report 3. Eleven CMIP6 source models train a coefficient-native persistence-residual Earthformer on the exact 12-to-14-month forecast task (21,505 balanced windows); three whole source models (GFDL-ESM4, MCM-UA-1-0, MRI-ESM2-0) are held out for all checkpoint selection. Exact-continuation training to 30k steps (v1→v2→v3) shows an early held-out primary peak, late selector divergence, and broad late held-out overfit — grounds to stop simulation training. The predeclared PB-1 consultation then compares two transfers of the selected v3 backbone against the fixed V6 reference under the frozen Phase A protocol: the features-only zero-output transfer wins with the project's first resolvable observed primary improvement (Niño-3.4 correlation delta +0.489, 90% CI $[+0.256, +0.857]$; validation RMSE 0.5407, ACC +0.3579, Niño-3.4 correlation +0.6157). After Phases C and D were evaluated and rejected on validation, a written protocol froze PB-1 as the single final model and authorized exactly one hash-gated evaluation of the sealed 2016–2020 split. That evaluation returned pooled field RMSE 0.5406 versus persistence 0.6591 and zero anomaly 0.6200 (paired delta −0.1185, 90% CI $[-0.1561, -0.0858]$); 35/35 windows beating persistence and 14/14 leads beating zero; field ACC +0.2553; Niño-3.4 correlation +0.3432; and amplitude ratios 0.408/0.398/0.411 — confirmed pattern and phase skill with persistent amplitude damping.

## 1 Study Identity and Handoff

**Study ID** R4 / Phase B. **Incoming state:** the frozen Phase A transfer protocol and the V6 reference from Report 3. **Changed factor:** supervised forecast-task exposure at simulation scale. **Evidence tiers:** `E2-SIM` pretraining and budget study; `E3-VAL` predeclared transfer consultation; `E4a-TEST` final result. **Outgoing baseline:** immutable PB-1 (checkpoint SHA-256 `c6b49723...`) for Reports 5–8. **Timeline note:** reading order is thematic — the Phase C and Phase D studies of Reports 5–6 were executed *after* PB-1 selection but *before* this report's test section, and Phase E onward came after it. Phase B itself spans 2026-07-03 to 2026-07-05; the test read is 2026-07-06.

*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).

## 2 Introduction

Report 3 ended with a measured negative: masked reconstruction, even with regional auxiliaries, improved representation diagnostics without improving forecasts. The natural alternative inverts the pretext: train on the *target task itself* — 12 months in, 14 months out, persistence-residual, same loss — using the only source of abundant tropical-Pacific trajectories available: coupled climate simulations. The sim-to-real hypothesis is that model trajectories, despite simulation bias, teach forecastable ENSO evolution, *provided* checkpoint selection stays on held-out source models and the observed transfer is governed as tightly as Phase A prescribed. Precedent exists at the index level [1] and for medium-range fields [2]; the contribution here is the field-level, coefficient-native instance with explicit statistical governance and a sealed confirmatory endpoint.

## 3 Related Work

Climate-simulation pretraining entered ENSO deep learning with [1], which trained a CNN on CMIP historical simulations and SODA-era reanalysis before fine-tuning; [2] showed the same lever for gridded medium-range prediction; and transformer-based Earth system models routinely assume large training corpora [3]. Three practices distinguish this study. *Source-model holdout*: selection uses whole held-out models — generalization across model physics, not across windows of one model. *Exact continuation*: v1→v2→v3 resume model, optimizer, sampler position, and RNG state bit-exactly, making the 10k/20k/30k budget study a single controlled trajectory. *Residual initialization and sealed evaluation*: the zero-output transfer preserves exact persistence at step 0 on observations, and the endpoint is a one-shot, hash-gated test governed by a frozen written protocol [5]. The study is a task-alignment and evidence-governance experiment, not a comprehensive ENSO benchmark.

## 4 Common Forecast Coordinates

Observed: $24 \times 48$ tropical Pacific (30°S–30°N, 120°E–280°E); 12 context / 14 target months; train 1981–2010; validation 2011–2015 (35 windows); sealed test 2016–2020 (35 windows); cosine-weighted Niño-3.4 box, leads-3–14 summary; persistence and zero-anomaly baselines. CMIP uses analogous geometry but realization-local detrending, climatology, and normalization; simulation scores are therefore not absolute counterparts of ERA5 scores, and the CMIP selector's Niño-3.4 correlation (flattening windows and leads) is not numerically identical to the observed `D4` statistic.

## 5 CMIP6 Data and Source-Model Boundary

Discovery against the public Pangeo CMIP6 catalog (`CMIP/historical/Omon/tos/r1i1p1f1`) on 2026-07-03 returned 57 candidate datasets across 50 source models; metadata inspection rejected curvilinear and unstructured grids, incomplete 1850–2014 coverage, and unit problems, leaving **14 complete rectilinear stores** [4]. Each realization is processed independently: longitude normalization; selection of 30°S–30°N, 120°E–280°E with an interpolation halo; regrid to $24 \times 48$; Celsius conversion; cell-wise linear detrending fit over 1850–2014 with calendar-month fixed effects, applied *before* climatology; subtraction of a 1900–1999 realization-local monthly climatology; three-level DTCWT packing; and store-local per-channel normalization. Every store spans 1,980 months with coefficient geometry $[1980, 24, 48, 37]$, carries `allowed_use: pretrain_forecast` and

`simulation_calendar: true`, and is rejected by observed supervised loaders. Training balances windows by `source_id` (equal 1,955-window quotas), giving 21,505 train and 5,865 held-out windows.

### 5.1 Training and Held-Out Models

Eleven source models train: CESM2, CESM2-FV2, CESM2-WACCM, CESM2-WACCM-FV2, E3SM-1-0, E3SM-1-1, E3SM-1-1-ECA, GFDL-CM4, GISS-E2-1-G, GISS-E2-1-G-CC, GISS-E2-1-H. Three are held out whole: **GFDL-ESM4**, **MCM-UA-1-0**, and **MRI-ESM2-0** — MRI and MCM are family-unique in the compatible set, and GFDL supplies a third institution (its relation to GFDL-CM4 is the documented compromise required for a three-model holdout). Whole-source holdout matters because windows within one realization are highly dependent; aggregation weights models equally so no source dominates silently. All Phase B checkpoint decisions use these three models only, consuming zero observed-validation consultations (Table 1).

| Role | Source models (`historical` r1i1p1f1, one per source) | Windows |
|---|---|---|
| Training (11) | CESM2, CESM2-FV2, CESM2-WACCM, CESM2-WACCM-FV2; E3SM-1-0, E3SM-1-1, E3SM-1-1-ECA; GFDL-CM4; GISS-E2-1-G, GISS-E2-1-G-CC, GISS-E2-1-H | 21,505 |
| Held out (3) | GFDL-ESM4 (shares an institution with GFDL-CM4 — the documented compromise); MCM-UA-1-0 and MRI-ESM2-0 (family-unique) | 5,865 |

**Table 1:** The Phase B source-model boundary. Every store spans 1,980 months on the common grid with realization-local detrending, climatology, and normalization; training windows are balanced to equal 1,955-window quotas per source, and all checkpoint selection uses the three held-out models only.

## 6 Forecast-Pretraining Method

The model, target, and loss are exactly Report 2's supervised configuration — coefficient-native Earthformer $[12, 24, 48, 37] \to [14, 24, 48, 37]$, persistence-residual target, unweighted `residual_skill_v0` — with one difference declared in advance: *all* 166 optimizable tensors train in simulation (pretraining does not authorize observed full-backbone unfreezing; the observed transfer retains the frozen Phase A boundary). AdamW at $10^{-4}$, batch 1, balanced source sampling. Stage continuations resume `last.pt` exactly — model, optimizer, deterministic sampler position, and Python/NumPy/CPU/CUDA RNG state — so v1 (0–10k), v2 (10k–20k), and v3 (20k–30k) form one uninterrupted optimization trajectory with recorded stage hashes.

## 7 Simulation Experiments (`E2-SIM`)

### 7.1 Pilot and Expanded V1

A six-train/two-holdout real pilot (1,000 CUDA steps) validated ingestion, dataset, shape, and checkpoint paths (selected checkpoint held-out Niño-3.4 correlation 0.509). The expanded v1 trained 10,000 steps on the 11/3 split: selected step 5,000 with aggregate coefficient RMSE 1.1739, field RMSE 0.6579, field ACC 0.2730, Niño-3.4 correlation 0.5644 — against held-out persistence field

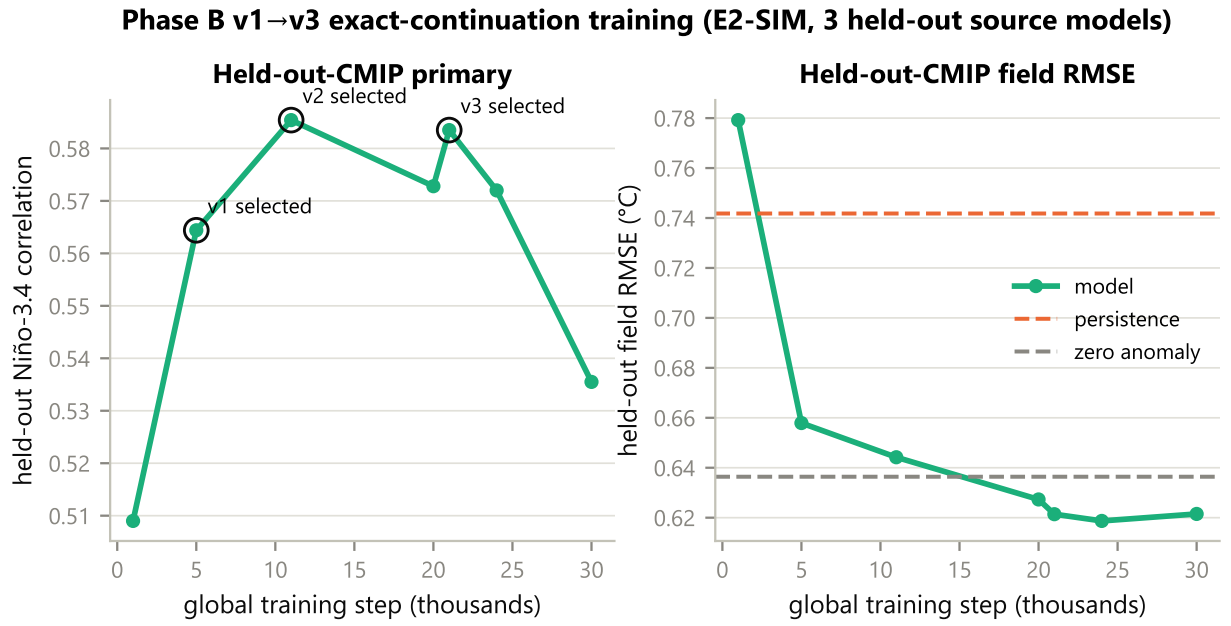


**Figure 1:** The v1→v3 exact-continuation budget study on the three held-out CMIP models (`E2-SIM`). Left: the held-out primary peaks early in each stage and declines late (30k endpoint 0.5355). Right: held-out field RMSE improves monotonically, crossing the zero-anomaly baseline in v3 — selector divergence that justifies stopping.

RMSE 0.7418 (correlation 0.3183) and zero-anomaly 0.6364. Variability ratios 0.828/0.771/0.742: simulation training preserves far more amplitude than any observed fine-tune of Reports 2–3.

### 7.2 V2 Exact Continuation to 20k

V2 selected stage step 1,000 (global 11,000): Niño-3.4 correlation 0.5854 (+0.0211 over v1), field RMSE 0.6442 (−0.0137), ACC 0.2965 (+0.0236), variability 0.829/0.838/0.816. The primary then oscillated, ending at 0.5728 at global 20,000 — an early-peak/late-regression warning, but not broad degradation (end-stage RMSE and ACC improved).

### 7.3 V3 Exact Continuation to 30k

V3 retained global step **21,000** (SHA-256 `2247c1f9...`): Niño-3.4 correlation 0.5835 (−0.0019 vs v2 — the primary has stopped improving), coefficient RMSE 1.1112, field RMSE 0.6214 (crossing below the zero baseline by 0.0149 for the first time), ACC 0.3015, but variability ratios worsened to 0.765/0.777/0.752. Selector divergence appeared: the composite-loss and field-RMSE selectors both chose global 24,000 (RMSE 0.6187, correlation 0.5720), while the global-30,000 endpoint's primary fell to 0.5355 — broad late held-out overfit relative to the early peak (Fig. 1). Per-model v3 primaries: GFDL-ESM4 0.574, MCM-UA-1-0 0.465, MRI-ESM2-0 0.668. Decision: stop simulation training; proceed to the predeclared observed consultation from the v3 selected state.

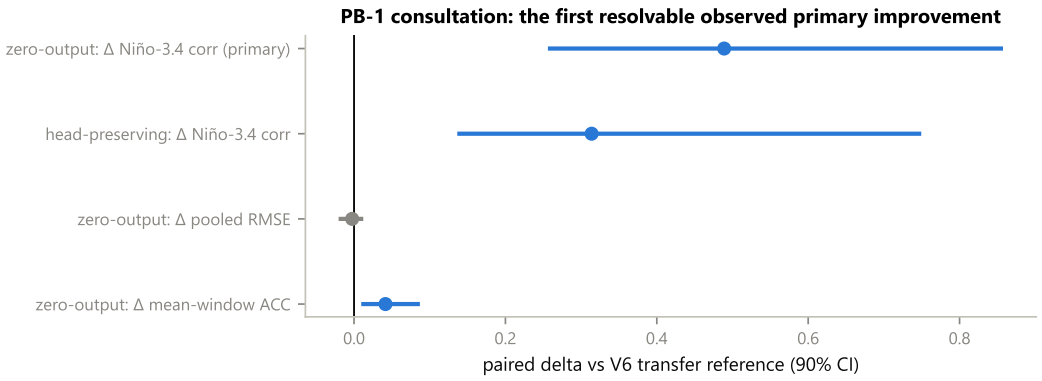


**Figure 2:** The predeclared PB-1 consultation against the fixed V6 reference (`E3-VAL`; 90% paired moving-block CIs). The zero-output transfer meets the full adoption rule: resolvable primary, unresolved RMSE, resolvably improved ACC.

# 8 PB-1 Observed Transfer Study (`E3-VAL`)

The consultation was predeclared in the ledger before any candidate ran: two transfers of v3's selected backbone against the fixed V6 pretrained reference, both under the frozen Phase A protocol (`decoder_last_block`, $10^{-4}/5 \times 10^{-6}$, epoch cadence, unweighted loss, guarded selection). The candidates differ only in residual-head handling: *head-preserving* (`none`: the pretrained residual head transfers, so no step-zero persistence equivalence, recorded as not applicable) and *features-only* (`zero_output`: the residual head is re-zeroed, exact persistence at step 0). Adoption rule: primary CI $> 0$ with delta $\geq +0.02$, no resolvable pooled-RMSE regression $\geq +0.005\,^{\circ}\mathrm{C}$, no resolvable mean-window-ACC regression $\leq -0.01$; branches for unresolved and resolvably-worse outcomes were declared with it. Bootstrap: paired moving blocks of 6, 10,000 resamples, 90%, seed 20260703.

## 8.1 Validation Outcome and Selection

Both 1,000-step CUDA fine-tunes completed; for the first time the selection metric did not saturate at the first epoch boundary (`zero_output` peaked at step 670; `none` improved through step 1,000). Selected `zero_output` validation: RMSE 0.540731, mean field ACC +0.357874, Niño-3.4 correlation +0.615723. Paired against V6: primary $+\mathbf{0.488977}$, CI $[+0.256068, +0.857422]$ — the first resolvable observed primary improvement in the project record; RMSE $-0.002497$ $[-0.020406, +0.012208]$ unresolved; mean-window ACC $+0.041523$ $[+0.009528, +0.087033]$ resolvably better (Fig. 2). The `none` candidate also passed the rule (primary $+0.313891$ $[+0.136393, +0.749443]$) but is dominated on primary and both guards. Declared diagnostics: the full-field amplitude ratio resolvably *decreased* ($-0.1009$) while the Niño-3.4 ratio resolvably increased ($+0.1090$); the 2015 case remained worse than persistence (0.7922 vs 0.7420, case-study-only). Decision: adopt the features-only transfer as **PB-1** (`best_skill.pt`, step 670, SHA-256 `c6b49723...`); the second Phase B consultation was not consumed.

# 9 Selection Exhaustion and Final-Test Gate

Between PB-1's adoption and the test read, two further phases ran to completion on validation and were rejected: Phase C (OHC300 input; Report 5) and Phase D (per-coefficient Gaussian heads; Report 6, first half). Both consumed their predeclared consultations; PB-1 remained the single

surviving model. On 2026-07-06 — before any 2016–2020 month was materialized — the final protocol was frozen in writing: exactly one evaluation of exactly one model (PB-1, named by full checkpoint SHA-256), plus persistence and zero-anomaly baselines (which consume no budget); the complete metric list with bootstrap settings; the rule that everything is reported favorable or not and that no test number ever selects, tunes, or gates anything; and a recorded honest expectation ("comparable modest skill"). The evaluator itself is fail-closed: it refuses to materialize test months unless the protocol file exists and its declared SHA-256 matches the checkpoint under evaluation.

# 10 First PB-1 Sealed-Test Result (`E4a-TEST`)

The single authorized evaluation ran on 2026-07-06 (35 windows):

| Metric (2016–2020, 35 windows) | PB-1 | Persistence | Zero anomaly |
|---|---|---|---|
| Pooled field RMSE (°C) | **0.5406** | 0.6591 | 0.6200 |
| Paired Δ vs persistence | −0.1185, 90% CI [−0.1561, −0.0858] (resolvable) | | |
| Windows beating persistence | **35 / 35** | — | — |
| Leads beating zero anomaly | **14 / 14** | — | — |
| Mean field ACC | **+0.2553** | +0.2292 | undefined |
| Niño-3.4 correlation, leads 3–14 | **+0.3432** | −0.015 | undefined |
| Amplitude ratios (full/eq/Niño-3.4) | 0.408/0.398/0.411 | ≈ 1 | 0 |

**Table 2:** The pristine sealed-test result (`E4a-TEST`). Persistence field ACC and Niño-3.4 correlation from the evaluation artifact's skill block; the CI attaches to RMSE versus persistence per the frozen protocol.

Leadwise structure appears in Fig. 3: PB-1 trails persistence only at lead 1 (where persistence is nearly perfect by construction), beats it at every later lead, and holds positive Niño-3.4 correlation through lead 14.

## 10.1 Interpretation

The positive field ACC and Niño-3.4 correlation are what make Table 2 more than a climatology-like win: Report 2 documented models that beat both baselines on RMSE with *zero* pattern correlation, and PB-1 is not that. Simultaneously, the amplitude ratios (≈ 0.41) record the persistent damping the deterministic frame imposes — the exact limitation the probabilistic program (Reports 6–8) subsequently attacks. Event-specific test behavior is reported descriptively only, and nothing in the test result was used to select or tune anything afterward.

# 11 Persistent Series Scorecard

The three evidence blocks are kept separate. `E2-SIM`: v1/v2/v3 selected rows (field RMSE 0.6579/0.6442/0.6214; Niño-3.4 correlation 0.5644/0.5854/0.5835; variability 0.83/0.83/0.76 full-field), with persistence 0.7418 and zero 0.6364 on the same holdout. `E3-VAL`: PB-1 selected validation row (RMSE 0.5407, ACC +0.3579, Niño-3.4 +0.6157) with the paired deltas of Fig. 2. `E4a-TEST`: Table 2, where the CI attaches to RMSE versus persistence only — other test CIs would require documented recomputation from the preserved per-window records. `X4`: v1 selected `10d8319b...`; v2 `3c822401...`; v3 `2247c1f9...` (plus guardrail `36550518.../a4034ad5...`); PB-1 `c6b49723...`; stage runtimes 4,944/4,767/5,464 s; exact 177-state reloads at every stage; `test_data_read: false` in every artifact before the authorized read.

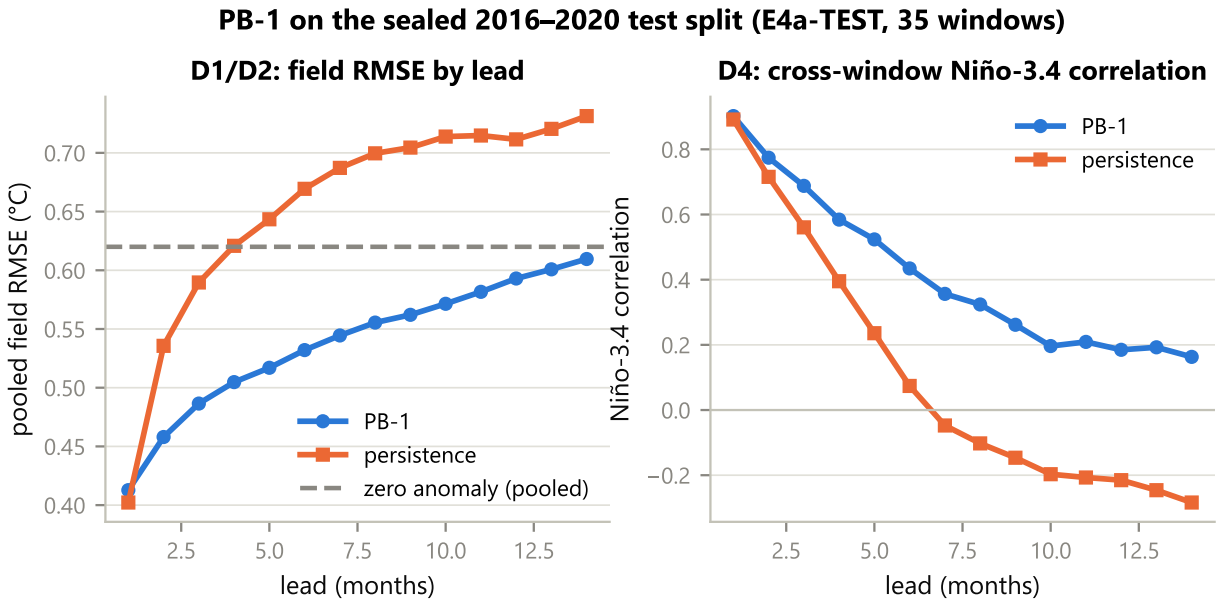


**Figure 3:** PB-1 sealed-test lead curves (`E4a-TEST`): pooled field RMSE by lead against persistence and the pooled zero-anomaly level (left); cross-window Niño-3.4 correlation by lead (right).

## 12 Discussion

*Why did supervised task alignment transfer where reconstruction did not?* The pretraining gradient now points at the downstream object itself — the conditional distribution of a 14-month future given a 12-month past — so the representation is organized around forecastable evolution rather than interpolation. Simulation scale (21,505 windows versus 335) provides the event diversity the observed record cannot.

*Why did features-only transfer dominate head preservation?* The pretrained residual head encodes CMIP-calibrated correction magnitudes in store-local normalized units; zeroing it lets observed fine-tuning rebuild the output map from exact persistence while keeping the pretrained features — initialization bias where it helps, none where it binds.

*Why did simulation selectors diverge late?* Long training overfits the held-out models unevenly across metrics: coefficient/field losses keep improving while the cross-window index correlation — the quantity closest to the observed primary — peaks early. Treating the primary's early peak as the selection signal, with RMSE/ACC as guards, is what the frozen policy prescribed, and the observed transfer result vindicates it.

*What does the sealed test establish?* That the core hypothesis — forecasting complete complex-wavelet coefficient maps with an Earthformer backbone yields real tropical-Pacific SST skill — holds on truly sealed data, with effect size comparable to its validation estimate (RMSE 0.5407 validation vs 0.5406 test against a harder persistence baseline).

## 13 Reproducibility and Protocol Audit

The full chain is machine-checkable: the 14-store manifest index (`cmip6_tos_phase_b_v1_manifest_index.json`, reused byte-for-byte by v2/v3); exact-resume states with recorded predecessor hashes; selector configurations and per-model metrics inside each checkpoint; the PB-1 transfer overlay with its

initialization report (177 loaded, zero skips); the frozen gate file naming the PB-1 SHA-256; the recorded attempted/executed test reads; the per-window test records; and the bootstrap artifact (`bootstrap_vs_persistence.json`). Two stale-metadata warnings of record: ancillary test files retain earlier diagnostic wording, and one YAML target-mode field disagrees with the executed checkpoint's persistence-residual contract — the claim rests on the executed checkpoint hash plus the `test_data_read` audit, not on either prose field.

## 14 Limitations

Climate-model dependence and shared CMIP biases; only three held-out source models (one sharing an institution with a training model); simulation metrics aggregated differently from observed metrics (flattened correlation; realization-local normalization); ERA5 as the observational target rather than independent in-situ products; a modest baseline suite; persistent amplitude damping; a single five-year test period; and no authorization to reuse the test result for later probabilistic selection. A separately predeclared bounded amendment was later exercised once by G2-long (Report 8); it did not alter PB-1's result and fully spent the split.

## 15 Conclusion and Transition

Supervised CMIP6 forecast-task pretraining produced the project's strongest pristine sealed result: resolvable held-out field skill with positive pattern and phase correlation, from a training recipe whose every selection decision is documented and whose test access was governed by a written one-shot gate. PB-1 is hereafter immutable — the SST-only deterministic mean reference for every subsequent report. Two questions leave this report open: whether subsurface predictors can add transferable information (Report 5, answered in the negative), and whether the damped amplitude can be cured probabilistically (Reports 6–8). Neither question carries test authority; Report 8 documents the separate second and final amended consultation.

## A Evidence Inventory

The paths below name the primary records in the project's internal research archive; the frozen protocol, per-window test records, and bootstrap artifact together establish the `E4a` claim.

```
docs/phase_b_cmip_forecast_pretraining_plan.md
docs/model_selection_policy.md
docs/final_protocol.md
outputs/manifests/cmip6_tos_phase_b_v1_manifest_index.json
outputs/runs/cmip6-tos-forecast-pretrain-v1-10k/
outputs/runs/cmip6-tos-forecast-pretrain-v2-20k/
outputs/runs/cmip6-tos-forecast-pretrain-v3-30k/
outputs/evaluations/
  phase_b/
  coeff-native-residual-skill-v0-cmipv3pretrained-1000step-validation/
  coeff-native-residual-skill-v0-cmipv3pretrained-zeroout-1000step-validation/
  final-sealed-test-pb1/
  validation_ledger.md
```

# References


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

[2] 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.

[3] Z. Gao, X. Shi, H. Wang, Y. Zhu, Y. Wang, M. Li, and D.-Y. Yeung, "Earthformer: Exploring space-time transformers for Earth system forecasting," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[4] V. Eyring et al., "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization," *Geoscientific Model Development*, 9:1937–1958, 2016.

[5] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, 102:359–378, 2007.