Cross-Domain Transfer of Upper-Ocean Heat-Content Predictors

Scientific Report 5: A Controlled OHC300 Negative Study (Phase C)

Jashua Luna*

25 July 2026

Abstract

Held-out-simulation outcomes here are `E2-SIM`; observed transfer outcomes are `E3-VAL`. We ask whether upper-300 m ocean heat content (OHC300) adds transferable information to the PB-1 SST pathway. A 74-channel coefficient-native model (37 SST + 37 OHC DTCWT channels in, 37 SST channels out) is pretrained on paired CMIP6 SST/OHC300 stores under the Phase B protocol and fine-tuned on observations with GODAS-derived OHC300, against equal-budget SST-only twins that zero the OHC input while keeping the identical pretraining lineage. Within simulation, OHC helps: at equal 30k budget the SST+OHC backbone reaches held-out CMIP Niño-3.4 correlation ≈ 0.610 versus 0.584 for the SST-only Phase B backbone. On observations it does not transfer — and actively hurts. The first consultation exposed a pretraining-budget confound (10k vs 30k) and, per the predeclared branch, triggered the fair-budget iteration; the second and final consultation at equal 30k budget isolated the input causally: using observed GODAS OHC resolvably degrades Niño-3.4 correlation (-0.126, 90% CI $[-0.209, -0.021]$), mean-window field ACC (-0.080, CI $[-0.140, -0.056]$), and pooled RMSE ($+0.016$, CI $[+0.012, +0.032]$) relative to zeroing it, while the SST-only twin remains comparable to PB-1 — so the lineage is sound and the observed OHC *input* is the harm. Phase C is closed as a rigorous negative: naive coefficient concatenation of a simulation-trained subsurface predictor does not survive the simulation-to-observation boundary. PB-1 is retained.

1 Study Identity and Handoff

Study ID R5 / Phase C. **Incoming baseline:** PB-1 (Report 4). **Changed factor:** add OHC300 context channels while retaining SST targets, the frozen transfer recipe, and (after the confound iteration) the pretraining budget. **Evidence tiers:** `E2-SIM` and `E3-VAL` only; no test access. **Terminal decision:** reject naive OHC concatenation; retain PB-1 for Report 6. **Timeline:** 2026-07-06, after PB-1 selection and before the E4a test read; both predeclared Phase C consultations were consumed.

2 Introduction

Subsurface heat content is the canonical slow variable of ENSO dynamics. In the recharge-oscillator description, equatorial warm-water volume builds before warm events and discharges through them [1], and observed warm-water volume leads Niño-3.4 SST by roughly two to three seasons [2] — long enough, in principle, to extend a 14-month forecast's usable horizon. If any additional predictor should help this system, it is this one.

*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).

The question, posed precisely, is not whether OHC carries ENSO signal — it does, in theory and in our simulations — but whether a predictor that improves *held-out climate-model* forecasting retains its benefit when the observed counterpart of that predictor comes from an ocean reanalysis (GODAS) with its own assimilation physics, climatology, and error structure. That is a simulation-to-observation transfer question, and Phase C is built as a controlled experiment on exactly that boundary.

3 Related Work

The recharge/discharge paradigm [1] and its observational confirmation via warm-water volume [2] motivate subsurface predictors for statistical and ML ENSO forecasting alike; operational analyses routinely monitor upper-ocean heat content as an ENSO precursor. Ocean reanalyses, however, disagree substantially on subsurface heat-content variability, especially before Argo [4]: the observed "ground truth" for OHC is itself a model product (GODAS assimilates profiles into an ocean model [3]), so a simulation-trained OHC pathway faces distribution shift on *both* sides of the transfer. CMIP6 models likewise differ from observations in thermocline sharpness, ENSO amplitude, and the phase relation between heat content and SST [5]. The literature thus supplies both the hypothesis and the specific risk this study is designed to isolate; multimodal fusion practice in ML (channel concatenation of co-registered fields) supplies the naive mechanism under test.

4 Data

Simulation and observed OHC products are described separately because their mismatch is central to the interpretation.

4.1 CMIP6 SST and OHC300

For each of the 14 Phase B source realizations, ocean potential temperature (`thetao`) is ingested, cropped to the domain *before* vertical reduction, and reduced by depth/thickness-weighted integration over the upper 300 m:

$$\mathrm{OHC300}(x,t) \;=\; \frac{\sum_{k:\, z_k \le 300\,\mathrm{m}} \theta(x, z_k, t)\, \Delta z_k}{\sum_{k:\, z_k \le 300\,\mathrm{m}} \Delta z_k}, \tag{1}$$

i.e. a thickness-weighted mean temperature of the upper 300 m (proportional to heat content up to constants). The result is paired 1:1 with the Phase B SST stores under identical policy: realization-local detrending, 1900–1999 climatology, 24×48 regrid, three-level DTCWT packing, store-local normalization, and the same 11-train/3-held-out source split.

4.2 Observed ERA5 and GODAS

Observed OHC300 is built from GODAS potential-temperature files [3] by the same reduction (1), timestamp-aligned with ERA5 SST, masked, and sealed to 1981–2015 (420 months): no 2016–2020 OHC material exists in the study. GODAS differs from CMIP OHC in assimilation constraints, mean thermocline structure, variance, and trend handling — the distribution shift under test.

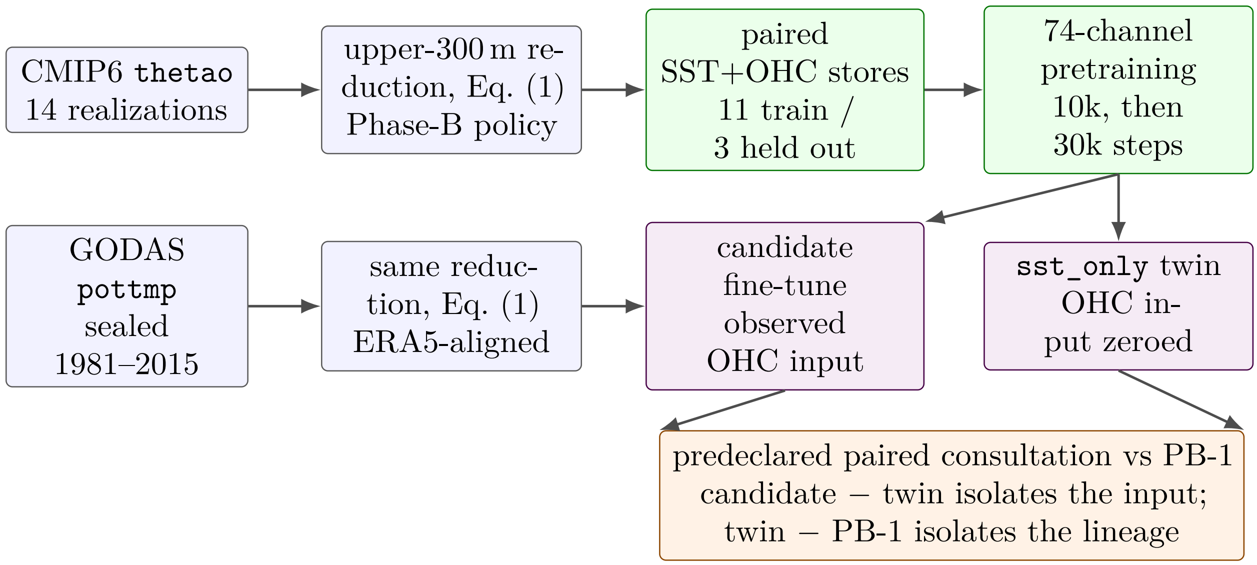

Figure 1: The controlled Phase C design. Simulation and observed OHC300 are produced by the same vertical reduction under the same store policy; the candidate and its `sst_only` twin share pretrained weights, recipe, and budget, differing only in whether the observed OHC channels are read or zeroed.

4.3 Common Forecast Coordinates

As in the series: 24×48 tropical Pacific, 12/14-month geometry, observed train 1981–2010, validation 2011–2015 (35 windows), 2016–2020 reserved and untouched, cosine-weighted Niño-3.4 with leads-3–14 summary, persistence and zero-anomaly baselines. Ensemble size: not applicable.

5 Model and Controlled Ablation

The candidate consumes 74 input channels — the 37 SST coefficients concatenated with the 37 OHC coefficients — through a fresh first convolution, with the coefficient-native backbone and the unchanged 37-channel SST output head; targets, loss, and the frozen transfer protocol are exactly PB-1's.

The decisive control is the `sst_only` twin: the *identical* OHC-lineage model — same pretrained weights, same fine-tune recipe, same budget — with the OHC input channels zeroed at transfer time. The twin separates two confounded explanations for any candidate-versus-PB-1 difference: pretraining lineage (different backbone history) versus input information (what observed OHC actually contributes). Candidate-minus-twin isolates the input; twin-minus-PB-1 isolates the lineage (Fig. 1).

6 Experimental Design and Gates

The predeclared consultation (ledger, 2026-07-06) fixes: reference PB-1; primary Niño-3.4 correlation leads 3–14; guards pooled RMSE and mean-window ACC; adoption thresholds identical to PB-1's (primary CI > 0, delta $\geq +0.02$, no resolvable RMSE $\geq +0.005\,^{\circ}$C or ACC ≤ -0.01 regression); paired moving-block bootstrap (block 6, 10,000 resamples, 90%, seed 20260703); and the two-consultation branch structure — if the candidate is unresolved without resolvable guard regression, one data/budget iteration is authorized as the second and final consultation. The first candidate carried a known asymmetry: its backbone had 10k pretraining steps versus PB-1's 30k (v1→v3), a budget confound the design anticipated.

7 Results

7.1 Within-Simulation OHC Signal (`E2-SIM`)

At equal 30k budget, the paired SST+OHC backbone improves held-out CMIP forecasting: Niño-3.4 correlation ≈ **0.610** versus 0.584 for the SST-only Phase B v3 backbone, with per-model and aggregate metrics improving. The flattened simulation correlation estimator is not numerically comparable to observed `D4`, but the within-simulation direction is unambiguous: the subsurface field carries usable predictive signal for this architecture when train and test share model physics.

7.2 First Observed Consultation: Budget Confound

The 10k OHC candidate lost to PB-1 on the primary — Niño-3.4 correlation 0.3985 vs 0.6157, delta −0.2172, CI $[-0.2781, -0.0576]$, resolvably worse — with unresolved guards (RMSE −0.0058; ACC −0.0304). But its equal-10k `sst_only` twin shows the OHC input was not the problem at that budget: within the 10k lineage, OHC resolvably *improved* pooled RMSE (−0.0063, CI $[-0.0084, -0.0008]$) and nudged the primary (+0.0505, unresolved $[-0.126, +0.118]$). The candidate lost because of three-fold less pretraining, not because of OHC. Per the predeclared branch, this triggered the one data-volume iteration: extend OHC pretraining to 30k and re-consult.

7.3 Second Observed Consultation: Equal-Budget Isolation

At matched 30k budget the picture inverts (Fig. 2). Against PB-1, the OHC candidate is not better (primary 0.4547 vs 0.6157; delta −0.161, CI $[-0.266, +0.023]$ unresolved) and is resolvably worse on both guards (ACC −0.122; RMSE +0.012). The causal isolation is candidate-versus-twin: feeding observed GODAS OHC, relative to zeroing it,

Paired delta (OHC − `sst_only`, 30k)	Point	90% CI
Niño-3.4 correlation, leads 3–14	−0.126	$[-0.209, -0.021]$
Mean-window field ACC	−0.0799	$[-0.1401, -0.0561]$
Pooled field RMSE (°C)	+0.0165	$[+0.0121, +0.0324]$

— all three resolvable, all three unfavorable. Meanwhile the twin itself is comparable to PB-1 (Niño-3.4 correlation 0.5808 vs 0.6157), confirming the OHC pretraining lineage is sound. The observed OHC *input* is what hurts.

8 Persistent Series Scorecard

`X4`: candidate/twin checkpoints `e4dba1fc.../a0865797...` (10k) and `358ac849.../c08e2e07...` (30k); GODAS store sealed 1981–2015; `test_data_read: false` throughout; 2015 case values (OHC 0.7786/0.7837 vs persistence 0.742) recorded as case-study-only.

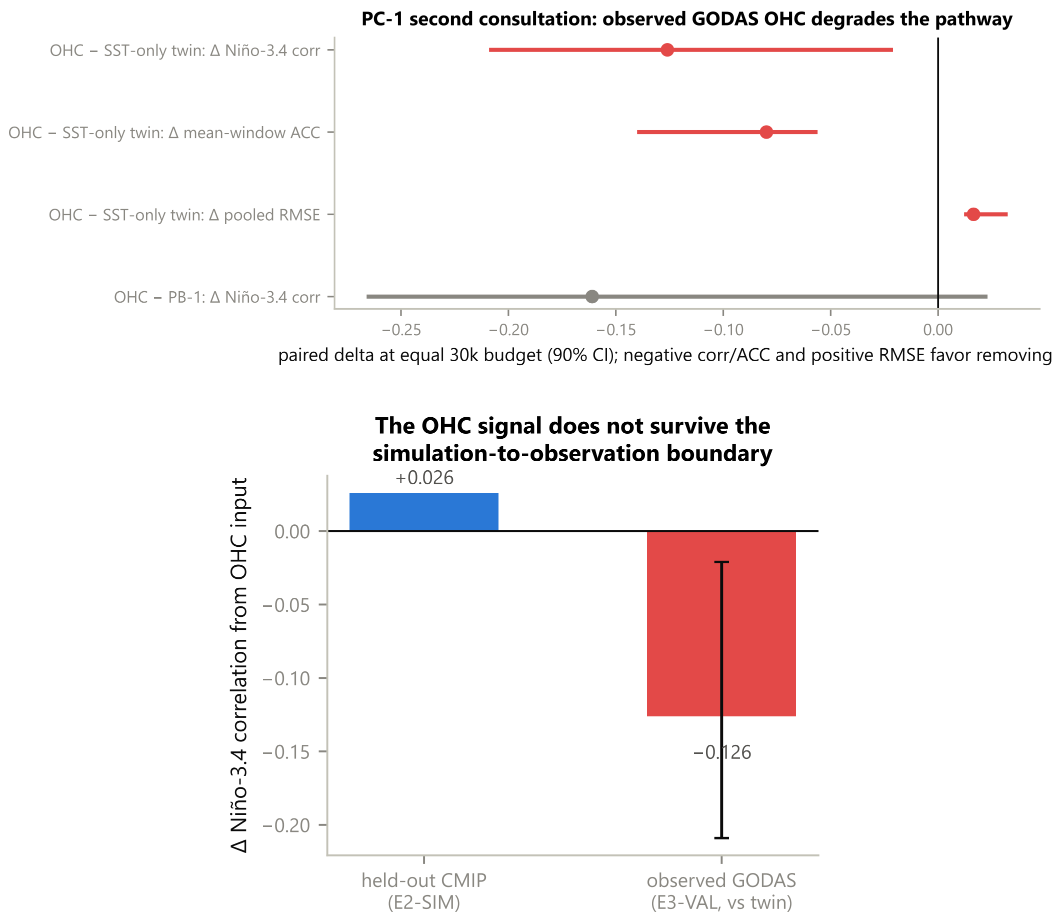

Figure 2: Top: the decisive second consultation (`E3-VAL`; 90% paired CIs). Every OHC-versus-twin delta is resolvable and unfavorable; the OHC-versus-PB-1 primary is unresolved with resolvably worse guards. Bottom: the sign of the OHC effect on Niño-3.4 correlation flips across the simulation-to-observation boundary — the study's one-figure summary.

9 Distribution-Shift Analysis

The demonstrated *outcome* is that observed OHC input harms this pathway; the *mechanism* — simulation-to-observation distribution mismatch — is inferred from converging circumstantial evidence rather than a single decisive diagnostic. The candidate's OHC→SST pathway was fit entirely on CMIP statistics: realization-local climatologies and trends, model thermocline structure, and model-specific OHC–SST phase relations. GODAS OHC differs in all of these — assimilation-constrained variance, different mean state and trend treatment, and product-specific noise — so the transferred pathway reads observed OHC through mis-calibrated weights and injects error rather than lead information. Consistent with this reading: the same pathway *helps* when train and evaluation share model physics (`E2-SIM`); the harm survives budget matching; and the twin construction rules out lineage effects. A planned next-pass addition is a direct comparison of coefficient-space distributions (per-channel variance profiles, regional means/variances, missingness, and conditional OHC–SST response) between the CMIP training pool and GODAS; the ingestion reports contain the required statistics.

Badge	Model	D1 RMSE	D3* ACC	D4 Niño-3.4	note
E2-SIM	SST-only v3 (held-out CMIP)	0.6214	0.302	0.584	Phase B selected
E2-SIM	SST+OHC 30k (held-out CMIP)	—	—	0.610	OHC helps in-sim
E3-VAL	PB-1 (reference)	0.5407	0.358	0.6157	frozen
E3-VAL	OHC candidate 10k	0.5349†	—	0.3985	budget-confounded
E3-VAL	OHC candidate 30k	+0.012 vs PB-1	−0.122 vs PB-1	0.4547	guards resolvably worse
E3-VAL	`sst_only` twin 30k	—	—	0.5808	lineage sound

Table 1: Scorecard with separated evidence blocks. *Delta rows use the bootstrap-compatible mean-window ACC; pooled ACC is a distinct estimator. †Implied by the recorded delta −0.0058 vs PB-1 (unresolved). The paired transfer-gap diagnostic is the report's headline: simulation benefit +0.026 versus observed input effect −0.126.

10 Discussion

The result argues against *naive coefficient concatenation*, not against subsurface predictors as such. The physics that motivated the study is intact — OHC300 demonstrably helps within simulation — and the failure is located at the product boundary. Constructive variants, each a properly separate future study: (i) bias-correcting observed OHC toward the simulation distribution (quantile mapping per channel) before the transferred pathway reads it; (ii) pretraining the OHC pathway on an observation-consistent ocean reanalysis ensemble rather than CMIP; (iii) collapsing OHC to low-dimensional summaries with robust cross-product meaning (warm-water volume, equatorial mean thermocline depth) instead of full coefficient maps; (iv) learned cross-product alignment (adapters fit on the overlap period); and (v) targeted low-mode conditioning, where OHC informs only the coherent basin-scale modes that Report 7 shows carry the index error. Each inherits this report's twin design as its control.

11 Limitations

One observed OHC product (GODAS) — reanalysis disagreement [4] means another product could transfer differently; 0–300 m mean-temperature aggregation discards vertical structure; no salinity/dynamic-height information; three held-out CMIP models; 35 overlapping validation windows; a two-stage adaptive design (the budget iteration was predeclared but still uses the validation split twice); and no test access. Within-simulation improvement cannot validate observational transfer — that asymmetry is the study's subject, not an oversight.

12 Conclusion and Transition to Report 6

OHC300 carries real predictive signal in simulations and fails, causally and resolvably, through the tested observation pathway. Phase C closes with both consultations spent, PB-1 frozen as the deterministic mean reference, and a clean negative that later phases repeatedly cite as the canonical distribution-shift precedent (it directly shapes the risk analysis of the CMIP-pretrained diffusion head in Report 6 and the calibration-transfer caveats of Report 8). The unresolved amplitude/uncertainty question passes to the probabilistic-head study of Report 6.

A Evidence Inventory

The paths below name the primary records in the project's internal research archive; the ingestion reports contain the per-channel statistics required for the coefficient-distribution comparison proposed in Section 9.

```
docs/phase_c_ohc_plan.md
outputs/manifests/cmip6_paired_sst_ohc300_phase_c_v1_manifest_index.json
outputs/reports/godas_ohc300_1981_2015_ingestion_report.json
outputs/runs/cmip6-ohc-forecast-pretrain-phase-c-v1-10k/
outputs/runs/cmip6-ohc-forecast-pretrain-phase-c-v2-30k/
outputs/evaluations/
  phase_c/
  coeff-native-residual-skill-v0-cmip-ohc-30k-pretrained-1000step-validation/
  coeff-native-residual-skill-v0-cmip-ohc-30k-pretrained-sstonly-1000step-validation/
tests/test_ohc_ingestion.py
tests/test_auxiliary_paired_dataset.py
```

References

[1] F.-F. Jin, "An equatorial ocean recharge paradigm for ENSO. Part I: Conceptual model," *Journal of the Atmospheric Sciences*, 54:811–829, 1997.

[2] C. S. Meinen and M. J. McPhaden, "Observations of warm water volume changes in the equatorial Pacific and their relationship to El Niño and La Niña," *Journal of Climate*, 13:3551–3559, 2000.

[3] D. Behringer and Y. Xue, "Evaluation of the global ocean data assimilation system at NCEP: The Pacific Ocean," in *Eighth Symposium on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface*, AMS, 2004.

[4] M. A. Balmaseda, K. E. Trenberth, and E. Källén, "Distinctive climate signals in reanalysis of global ocean heat content," *Geophysical Research Letters*, 40:1754–1759, 2013.

[5] 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.