Band-Weighted Temporal Sequence Diffusion for Probabilistic ENSO Trajectories

Scientific Report 8: Phase G, the Adopted Stack, and the Second and Final Sealed Test

Jashua Luna*

25 July 2026

Abstract

Held-out head-pretraining results here are `E2-SIM`; all observed trajectory, calibration, lifting, and adoption results are `E3-VAL`; the terminal sealed-test consultation is `E4b-TEST`. We study whether temporally joint diffusion can replace independently denoised monthly textures with coherent ENSO trajectories. An AR(1) sampling control (G0) fixes month-to-month jitter (index temporal-difference ratio 1.279 → 0.977) but provably not flow dependence; a joint-sequence architecture (G1) matches the folded head's denoising loss yet learns *no* temporal coherence (member increment lag-1 autocorrelation stuck at −0.45 at every sampler setting). A predeclared correction-structure audit convicts the objective, not the architecture: the true corrections are temporally correlated in every band (lowpass lag-1 0.645), but per-cell denoising MSE spends 84.4% of its weight on the finest band — which carries 0.1% of index-error variance — and only 1.1% on the lowpass. Re-weighting the loss equally across bands (G2′), the sole changed factor, causally unlocks coherence (sampled lowpass lag-1 0.019 → 0.308) and the project's best conditional mean; doubling CMIP pretraining to 16k steps (G2-long) compounds both (0.392; ensemble-mean Niño-3.4 RMSE 0.4748 versus PB-1's 0.5256). Flat calibration had failed its CRPS gate on every PB-1-centered ensemble; recentering on the learned mean makes it CRPS-neutral — the mean, not the calibration, was the bottleneck. The adopted stack (G2-long + flat LOYO EMOS/ECC-Q + coherent lifting) is the project's best validation product (spread–skill 0.958, coverage 59/77/82%, fair CRPS 0.2396 → 0.2463 canonical), adopted by user direction with the strict CRPS-not-worse-than-raw gate waived (+0.0095 archived, +0.0067 corrected). Its single authorized `E4b-TEST` read returned a mixed result: field RMSE at PB-1 (0.5390 vs 0.5406), Niño-3.4 correlation marginally higher (+0.3532 vs +0.3432), but field ACC (+0.2229) and Niño-3.4 RMSE (0.6912) worse than PB-1 and worse than zero anomaly (0.5821); validation-fit calibration improved raw test reliability (spread–skill 0.263 → 0.641; fair CRPS 0.4369 → 0.3890 canonical) without fully calibrating it. The validation advantages did not fully transfer; the 2016–2020 split is now fully spent.

1 Study Identity and Final Status

Study ID R8 / Phase G. **Incoming state:** the flow-independent calibrated ensemble and reusable machinery of Report 7. **Changed factors:** leadwise noise correlation, joint temporal architecture, wavelet-band objective weighting, CMIP pretraining length, calibration center, and lifting. **Evidence tiers:** `E2-SIM` pretraining; adaptively reused `E3-VAL` development; one post-adoption `E4b-TEST` reported-not-gated consultation. **Terminal result:** G2-long is near PB-1 only in field RMSE,

*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).

worse on field ACC and regional-index RMSE, and accompanied by partially calibrated uncertainty; no further 2016–2020 read is authorized by any document. **Timeline:** Phase G development 2026-07-11; adoption and the E4b read 2026-07-12.

2 Introduction

Report 6's PE-2c diffusion head folds the lead axis into the batch axis: each of the 14 monthly corrections is denoised independently, with independent noise, conditioned only on that month's mean correction field (`uncertainty.py::denoising_loss`). Two consequences, both confirmed: members are not trajectories (temporally independent textures, visible as month-to-month jitter in member Hovmöllers), and the spread carries no flow dependence (Report 7: corr(|error|, spread) 0.05–0.16). Phase F proved the missing conditional-variance information is learnable from CMIP; Phase G moves temporal structure and state-awareness *into the generator*. Four questions separate the causal factors: does imposed continuity suffice (G0); does temporal architecture capacity suffice (G1); does the objective supply the needed gradient pressure (G2′); and does the learned head, once coherent, calibrate (G3′, G2-long)? Figure 6 summarizes the resulting ladder on the trajectory metrics.

3 Focused Literature Review

The relevant evidence comes from probabilistic weather forecasting, video diffusion, and spatiotemporal downscaling. Two lessons recur. First, *temporal coherence is created by architecture and conditioning, not post-hoc noise fixes*: CorrDiff-style residual diffusion restores fine spatial detail but does not guarantee time coherence [1], while methods that model time either add a lead/trajectory axis or make the model continuous in time [2]. Second, *correlated noise is a control, not a cure*: continuous ensemble forecasting shows correlated driving noise mainly changes trajectory autocorrelation while leaving per-lead skill governed by the underlying model [2], and ERDM treats temporally correlated noise as orthogonal to the forecasting architecture. For the joint head itself, video-diffusion practice factorizes space and time rather than collapsing time into channels [3, 4], and EDM-style formulations [5] underlie the strongest weather diffusion systems [6]. On scoring, the variogram score is preferred over the energy score as the primary multivariate temporal score because it is more sensitive to mis-specified dependence [7]. We test architecture and objective as *separate* causal factors — a discipline the ledger enforces and the results reward.

4 Temporal Metrics and Gates

Beyond the carried `D1`–`D6`, `P1`–`P3`, `P4a`/`P4b`, `P5a`/`P5b`, and `P6`, this report adds the trajectory extension `T1`–`T5` (series charter): the temporal-difference ratio `T1` (target $[0.8, 1.2]$), increment autocorrelation `T2` (versus observed $+0.161$ lag-1), the variogram and energy scores `T3`, the high-frequency power fraction `T4`, and the center-specific flow-dependence diagnostics `T5a` (learned center, canonical) and `T5b` (frozen-PB-1-error, historical). Carried gates: frozen-mean identity, member sharpness, per-band energy, spread–skill $[0.8, 1.2]$, coverage ± 10 points, fair CRPS $\leq$ raw, and the aspirational flow-dependence > 0.3; lifecycle metrics are gated on held-out CMIP only (the observed slice holds essentially one event ramp).

4.1 CRPS Estimator Audit Boundary

Canonical `P1` is fair CRPS evaluated per window/lead over the member axis. The shared Phase F/G projection scorer instead reshaped [window, member, lead] directly, mixing leads and members; its recorded CRPS values are historical decision evidence, not canonical `P1`. This report carries verified corrected values only for the final G2-long validation and E4b artifacts; every earlier rung's CRPS is the archived value, labeled as such, pending recomputation.

4.2 Common Forecast Coordinates

24×48 tropical Pacific; 12/14-month geometry; train 1981–2010; validation 2011–2015 (35 windows). Phase G development had *no* test access; only after adoption did the written amendment authorize one E4b read. Cosine-weighted Niño-3.4 box, leads-3–14 calibration scope; persistence and zero-anomaly baselines; the canonical 32-member ensemble with recorded seed, η, ρ, and sampler steps.

5 G0: AR(1) Correlated-Noise Control

The independent per-month DDIM noise — both the initial latent and the η=1 per-step noise — is replaced by a stationary AR(1) process across lead with unit marginal variance,

$$z_1 = e_1, \qquad z_t = \rho\, z_{t-1} + \sqrt{1-\rho^2}\, e_t, \qquad \rho = e^{-1/\tau}, \tag{1}$$

a sampling knob on PE-2c (no retraining). Sweeping $\rho \in \{0, 0.607, 0.779, 0.882, 1.0\}$ paired by window seed: index `T1` moves 1.279 (too jittery) $\rightarrow$ 0.977 (near-perfect at τ_2, ρ=0.607) $\rightarrow$ 0.168 (frozen, at common noise); the τ_2 guards pass (spread–skill 0.399; archived CRPS +0.4%). But flow dependence at the operating point moves only +0.080 vs control ($<$ the predeclared 0.10 threshold); it reaches +0.255 only at common noise, which collapses the ensemble (`T1` 0.168, guards fail). No ρ delivers real flow dependence without destroying the ensemble — a textbook confirmation that the architectural time axis is required, not a noise prior. A structural limit also appears: observed increments are positively autocorrelated (+0.161: ENSO has momentum), and AR(1) noise around a *frozen* mean can only reduce the differencing-induced negative autocorrelation, reaching +0.161 only by overshooting to common noise (+0.568). τ_2 is selected as the presentation sampler and G1 baseline; nothing is adopted as a skill claim.

6 G1: Joint Sequence Architecture

The new head (replacing the folded denoiser; Fig. 1) is a 1.80M-parameter 2-D residual U-Net over 24×48 (scales $24{\times}48 \rightarrow 12{\times}24 \rightarrow 6{\times}12 \rightarrow 3{\times}6$) with a depthwise temporal convolution (k=3) at every scale and temporal self-attention with relative-lead bias at the two coarsest scales, adaptive-norm diffusion-time embedding, and an explicit lead embedding in every block. It denoises the whole $[B, 14, 37, 24, 48]$ correction sequence at one per-sequence noise level. The formulation is *identical* to PE-2c (ε-prediction, cosine schedule, 50-step DDIM η=1, x_0 clamp 4.0), so G1 isolates the time-axis and full-trajectory-conditioning effect. Conditioning v1 (free): the full 14-month mean trajectory via temporal mixing, lead embedding, target-month encoding, and a compact context state. Training: 8,000-step CMIP pretrain (held-out selection) $\rightarrow$ 1,000-step observed head-only fine-tune.

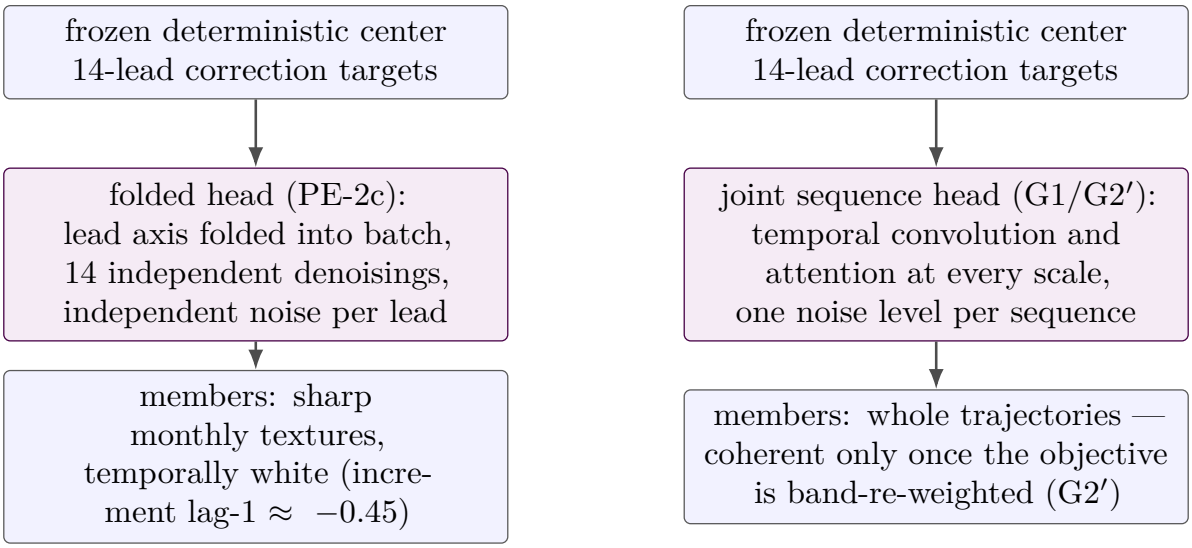

Figure 1: Folded versus joint denoising around the same frozen center. The architectural change alone (right, G1) supplies temporal *capacity* but leaves samples temporally white; the equal-band objective re-weighting of G2′ is what converts that capacity into learned temporal coherence.

6.1 Architectural Negative

The head fits the correction distribution as well as the folded head (held-out denoising 0.401, observed 0.394, both ≈ PE-2c's 0.40) and learns *no* temporal coherence. Across the η sweep: member `T1` 1.305–1.765 (worse than the folded 1.279 — *more* jittery); lag-1 increment autocorrelation stuck at -0.45 at every η (the MA(1)≈ -0.5 signature of temporally independent textures), never approaching observed $+0.161$; flow dependence sign-flipping around zero ($+0.08/-0.28/+0.05/-0.08$). The sweep *refutes* the hypothesis that η=1 was masking learned coherence: even at η=0 (the setting most favorable to it), members are temporally independent. η is a spread↔jitter knob (raising it lifts both spread–skill 0.369 → 0.521 and jitter), not a coherence knob. The one real gain is spread–skill; the mean-bypass RMSE (0.067 °C) exceeds its gate. *Diagnosis:* per-cell denoising MSE supplies temporal *capacity* but no temporal *pressure* — the temporal analogue of Phase E.

7 Correction-Structure Audit

A predeclared measurement resolves the G1 negative precisely (Fig. 2). Decomposing the true corrections (truth minus frozen PB-1) and G1's sampled corrections (η=0) by band:

band	true corr. lag-1	G1 sampled lag-1	share of denoising-MSE weight
lowpass	**0.645**	0.019	**1.1%**
level 3	0.487	−0.077	2.3%
level 2	0.377	0.013	12.1%
level 1 (finest)	0.267	0.034	**84.4%**

The true index-error trajectory is extremely persistent (level lag-1 +0.816; increment lag-1 +0.190 ≈ observed). Three facts convict the objective: (i) the true corrections carry rich temporal structure in every band, strongest where the index lives; (ii) per-channel normalization makes every lifted cell equal weight, and the finest band owns ∼ 84% of the valid cells, so the objective spends ∼ 84% of its gradient on the band with the *least* temporal structure and (per F0) 0.1% of index-error

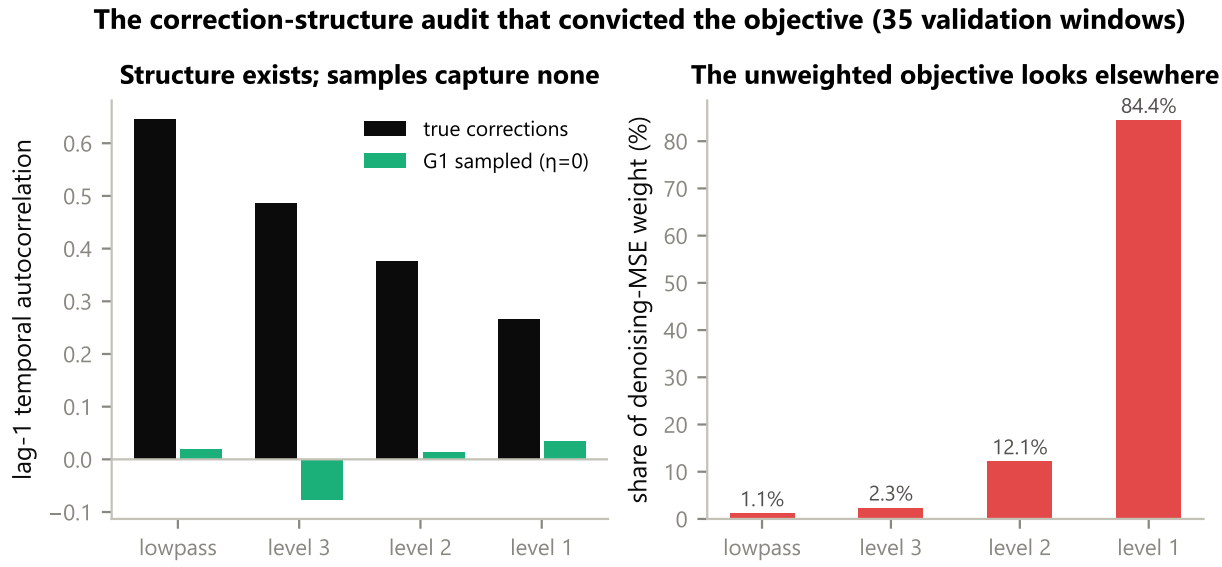

Figure 2: The audit that convicted the objective (`E3-VAL`, 35 windows). Left: true corrections are temporally correlated in every band; G1 samples are temporally white everywhere. Right: the unweighted per-cell objective spends 84.4% of its gradient on the finest, least-structured band.

variance, while the index-carrying lowpass gets $\sim 1\%$; (iii) G1's samples captured essentially none of the structure in any band. This also retroactively illuminates Phase E: the fine-scale over-cook and the calibration wall's indifference to 64× data are what an objective dominated by unpredictable fine-band texture looks like.

8 G1b: Correlated Sampling on the Sequence Head

AR(1) sampling (1) on the trained G1 checkpoint gives the best variogram measured to date (1.503 at τ_2) and higher spread–skill (0.532), but a $\sim 14\%$ CRPS regression under the archived scorer (0.2808 vs G0 τ_2's 0.2463). By the predeclared CRPS-priority rule, G0 τ_2 is retained as the presentation product; the trajectory diagnostics rule out simply stacking imposed noise correlation on G1 as the main solution. (Magnitudes await canonical rescoring; the split verdict does not.)

9 G2′: Equal-Band Denoising Objective

Changing only the wavelet-band weights — architecture, data, steps, and sampler held fixed — from cell-count weighting to *equal-band* (each band the same total weight, mean-normalized over valid cells),

$$\mathcal{L}_{\text{eqband}} \;=\; \mathbb{E}_{t,\varepsilon}\Big[\textstyle\sum_b \omega_b \,\|\varepsilon - \varepsilon_\theta(\cdot)\|_b^2\Big], \qquad \omega_{\text{lp}}{:}\omega_1 \approx 23{:}1 \text{ per cell}, \tag{2}$$

passes the structure gate causally: sampled lowpass lag-1 autocorrelation 0.019 → **0.308** (≥ 0.3; true 0.645) from the objective change alone. Further wins: the *best point forecast in the project* (ensemble-mean Niño-3.4 RMSE 0.4827 vs PE-2c-ensemble 0.4916 vs PB-1 0.5256 — the 0.339 °C ensemble-mean shift is learned predictable lowpass error signal, not harmful bias, so the mean-bypass gate "failure" is recorded as inapplicable in this regime), and the best fair CRPS in the project at η=0.5 (0.2435 < G0 τ_2's 0.2463, archived, pending recomputation). Remaining deficits are

one phenomenon — insufficient stochastic variation around a now-smooth conditional mean: `T1` 0.67–0.74 (over-smooth), spread–skill 0.25–0.32, coverage 22/35/42%, variogram 1.93. Only the lowpass crossed the learning threshold (fine bands still ≈ 0); and the 8k CMIP pretrain was still improving, an unexploited lever.

10 G3′: Calibration Center Experiment

Applying the Phase F machinery (flat F1a and conditional F3) to the G2′ η=0.5 ensemble in two centering variants isolates the role of the mean: G3′-A recenters on frozen PB-1 (the exact F-phase protocol); G3′-B recenters on G2′'s own better ensemble mean.

variant	mean RMSE	`P2`	`P3` 50/80/90	fair CRPS*	vs raw
G2′ raw (η=0.5)	0.489	0.313	—	0.2472	—
G3′-A F3 (PB-1 center)	0.526	1.089	60/83/90	0.2739	+0.027
G3′-B F1a-flat (ens. center)	**0.489**	**0.972**	**59/76/82**	**0.2496**	**+0.0024**
G3′-B F3-cond (ens. center)	0.489	1.258	68/88/92	0.2558	+0.009

Table 1: The center experiment (`E3-VAL`, LOYO; only the recentering target differs between A and B). *archived scorer, pending canonical recomputation.

The decisive contrast is A vs B, which differ *only* in the recentering target: recenter on PB-1 and flat calibration reproduces the familiar +11% CRPS wall; recenter on G2′'s own better mean and it becomes CRPS-neutral (+1%) while reaching spread–skill 0.972 and near-nominal coverage. This cleanly proves the hypothesis: **G2′'s learned mean was the CRPS bottleneck** — fix the mean and flat calibration fixes reliability without paying CRPS. The conditional variant over-disperses (1.258) because the CMIP σ is scaled to PB-1-magnitude errors and the LOYO per-fold scales (0.97/1.04/0.91) do not shrink enough for G2′'s smaller errors — but it delivers the best flow dependence in the project (0.223, up from 0.160).

11 G3″ and F1b

A partial-conditioning blend (compressing the conditional σ's window-variation toward its per-lead geometric mean, then applying the conditional per-fold scale) was built to convert the 0.223 flow dependence into an in-range spread. The finding is a rigorous negative *bounded by the data*: flow dependence never exceeds 0.223 at any blend strength, so the > 0.3 gate is unreachable on this three-year slice — the named F3 ceiling (the 2013 neutral-onset windows) is a property of the slice, and corr(|G2′ error|, σ) = 0.298 $\approx$ PB-1's 0.318 confirms the better mean did not invert the state-dependence. (An honest implementation note: a first blend design gave the held-out event fold the *smallest* spread and negative flow dependence; the bug was found and fixed during the run, and the corrected tool reproduces the full conditional at strength 1, unit-tested.) G3′-B F1a-flat remains the best index product. Coherent lowpass F1b lifting (Report 7) — lifting the zero-member-mean deviation difference through a lowpass, unit-Niño-3.4-average pattern B_ℓ,

$$r'_m = r_m + B_\ell[(q'_m - \bar{q}') - (q_m - \bar{q})], \qquad A(B_\ell) = 1, \tag{3}$$

— then materializes the calibrated spread in fields: exact index fidelity, invariant mean, highpass changes 2.0/0.5/1.1%, member std 0.587 $\rightarrow$ 0.822 (into the sharpness band).

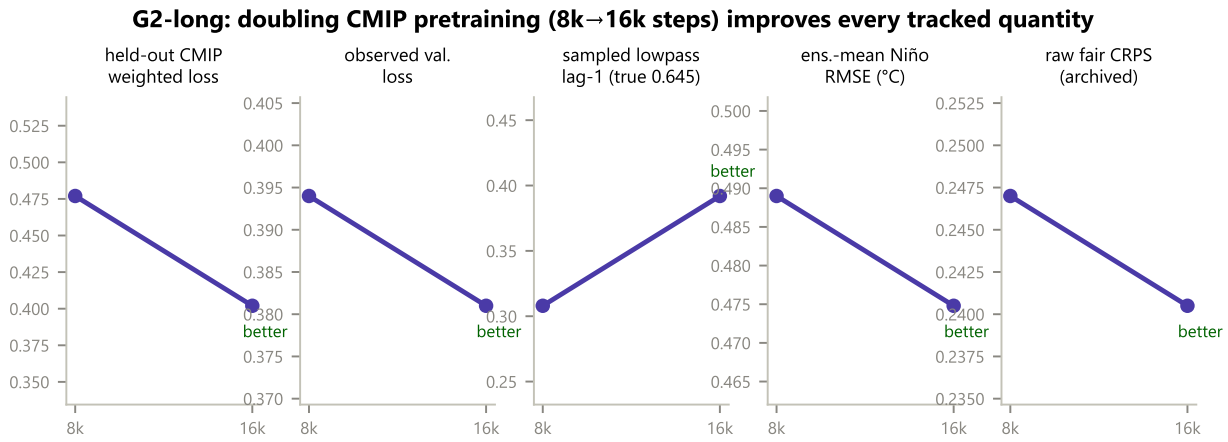

Figure 3: Doubling CMIP pretraining (`E2-SIM` → `E3-VAL`) improves every tracked quantity: more temporal structure across all bands, a better mean, lower CRPS, and the first positive raw flow dependence. The optimal sampler shifts from η=0.5 to η=1.0.

12 G2-long: Pretraining-Length Ablation

Changing only the CMIP pretraining from 8,000 to 16,000 steps (Fig. 3) is the single biggest Phase G improvement, confirming the 8k base had not converged:

metric	G2′ (8k)	G2-long (16k)
pretrain held-out weighted loss	0.477	**0.402**
observed validation loss	0.394	**0.381**
sampled lowpass lag-1 (true 0.645)	0.308	**0.392**
best raw fair CRPS (at η)*	0.247 (0.5)	**0.2405 (1.0)**
ensemble-mean Niño-3.4 RMSE (PB-1 = 0.5256)	0.489	**0.4748**
raw flow dependence `T5a`	−0.08	+0.152

More pretraining learned more temporal structure across *all* bands (fine-band lag-1 0.049/0.061/0.065 vs ≈ 0.01–0.03), a better mean, and the first positive raw flow dependence. Metric-provenance note: the raw canonical `T5a` on the final seed-42 artifact is +0.354; the separately reported +0.152 is the historical sequence diagnostic `T5b` (frozen PB-1 error, different per-window seeds) and is not `T5a` for the learned-center product. Archived raw fair CRPS 0.2405; canonical audit 0.2396.

13 Final Adopted Validation-Basis Product

The adopted stack: G2-long sequence diffusion checkpoint (CMIP pretrain `e2dde94a...`, 16k steps, equal-band; 1,000 observed head-only steps around frozen PB-1 `c6b49723...` on v3 backbone `2247c1f9...`), sampled at η=1, ρ=0; flat LOYO EMOS/ECC-Q recentered on the learned ensemble mean; coherent lowpass F1b lifting. Calibrated metrics: spread–skill 0.958, coverage 59/77/82%, fair CRPS 0.2396 → 0.2463 (canonical audit; archived 0.2405 → 0.2500), exact index fidelity, mean-field shift $\approx 2.4 \times 10^{-7}$ °C, highpass changes (`P5b`) 2.3/0.5/1.0%, and field-lifting ratio (`P4b`) 0.608 → 0.800. Flat inflation flattens canonical `T5a` from +0.354 raw to −0.492 calibrated — the reliability↔flow-dependence tension; $b > 0$ raw-spread blending was swept and does not help. `P4b`/`P5b` are not compared with PE-style `P4a`/`P5a`.

13.1 Adoption, Waiver, and Evidence Boundary

Per the latest append-only ledger, user direction formally adopted the stack on 2026-07-12 despite missing the strict CRPS-not-worse-than-raw gate under both the archived scorer (+0.0095) and the corrected audit (+0.0067) — a Monte-Carlo-noise margin, but a miss, waived exactly as budgets were earlier waived. Disclosed with the adoption: flat inflation changes canonical `T5a` from +0.354 raw to −0.492 calibrated; validation was adaptively reused; 2016–2020 was already known project-wide from PB-1; and adoption preceded all G2-long test access.

14 Final-Protocol Amendment and Pre-Read Expectation

A written amendment — exercising the bounded provision predeclared in the Phase E plan ("if and only if a generative head is adopted on validation, one additional single-shot test of that one model plus PB-1 and the two baselines") — authorized exactly one report-only evaluation of the checkpoint SHA-256 `16a865bc...fc381c1`. Frozen: η=1, ρ=0, 32 members, seed 42, the learned ensemble center, and the flat $\sigma(\ell)$ fit on all 2011–2015 windows and applied unchanged to test. The recorded pre-read expectation was a modest point-skill improvement and reasonable reliability. E4b is genuine held-out but lower-weight, reported-not-gated evidence: G2-long never saw a test month during development (verifiable), but far more researcher degrees of freedom were exercised on validation than for PB-1. Declared scope: raw G2-long fields/center and calibrated Niño-3.4 trajectories only; F1b lifted fields and `T1`–`T5` were not tested.

15 Second and Final Sealed-Test Result (`E4b-TEST`)

15.1 Pipeline Sanity and Deterministic Ensemble Center

The frozen `mean_head` first reproduces PB-1's sealed-test numbers to the reported digit (RMSE 0.5406, ACC +0.2553, Niño-3.4 correlation +0.3432, 35/35), confirming identical test data and pipeline. For the sampled ensemble center (Table 2, Fig. 4):

Metric (2016–2020, 35 windows)	G2-long ens. mean	PB-1 (frozen)	baseline
pooled field RMSE (°C)	**0.5390**	0.5406	pers. 0.6591
windows beating persistence	34 / 35 (97.1%)	35/35	—
mean field ACC	+0.2229	**+0.2553**	pers. +0.2292
Niño-3.4 corr. leads 3–14	**+0.3532**	+0.3432	pers. −0.015
Niño-3.4 RMSE leads 3–14 (°C)	0.6912	**0.5714**	zero 0.5821; pers. 0.7774

Table 2: The E4b deterministic center (`E4b-TEST`). G2-long ties PB-1 on field RMSE, is marginally better on Niño-3.4 phase correlation, and worse on field ACC and Niño-3.4 amplitude error — worse than both PB-1 and zero anomaly on the index.

The result splits at the index level: all 14 full-field leads beat zero, but G2-long beats zero on only 5/14 Niño-3.4-RMSE leads. The headline paired interval (center minus persistence, Niño-3.4 RMSE, block 6, 10,000 resamples, seed 20260703) is −0.086, 90% CI [−0.432, +0.090]: *unresolved* (35 event-dominated windows give a wide index CI; the field-RMSE win at 97% of windows is the stronger persistence comparison). No paired interval versus PB-1 was computed, so the point estimates are called *mixed* — not statistically equivalent or improved. Amplitude stays damped (std ratio 0.40).

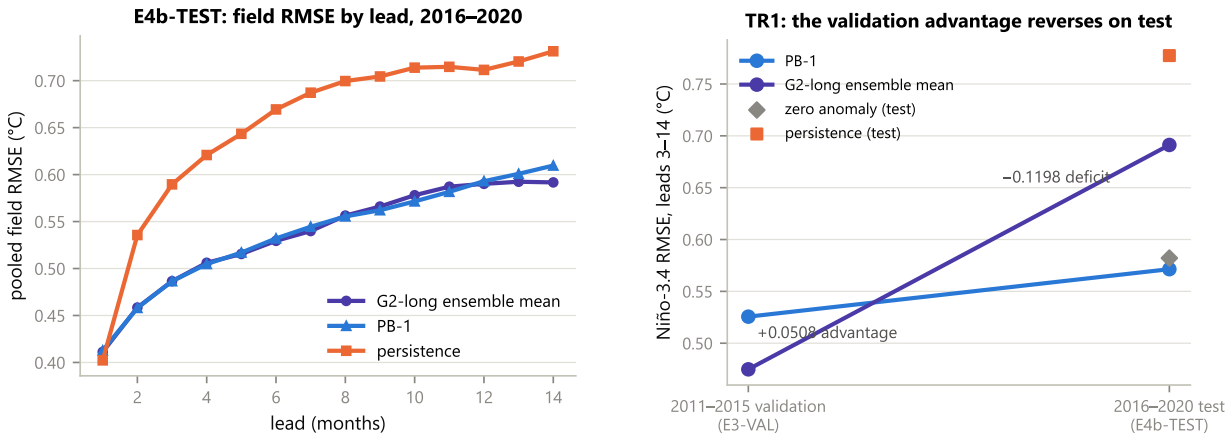

Figure 4: Left: E4b field RMSE by lead — G2-long's ensemble mean tracks PB-1 and beats persistence at all leads ≥ 2. Right: the transport reversal (`TR1`): the validation-period Niño-3.4-RMSE advantage over PB-1 (+0.0508 °C) becomes a −0.1198 °C deficit on test (contrast −0.1706 °C).

15.2 Held-Out Calibration and Metric Audit

Applying the validation-fit flat σ unchanged: raw-to-calibrated spread–skill 0.263 → 0.641; coverage 20/38/45% → 43/61/68% (Fig. 5). Canonical per-window/per-lead fair CRPS 0.4369 → 0.3890; the archived axis-mixed pair 0.4450 → 0.4107 is a provenance note only. The qualitative conclusion — calibration improves CRPS but the ensemble remains under-dispersed because 2016–2020 errors exceed the 2011–2015 scale σ was tuned on — is unchanged between scorers. Raw member sharpness (`P4a`) is 0.579/0.552/0.463 (full/equatorial/Niño-3.4), below the prior sharpness gates, and reported as such rather than as an unqualified "sharp" claim.

15.3 Validation-to-Test Transport and Final Status

The reversal is explicit (`TR1`, Fig. 4): G2-long's Niño-3.4-RMSE advantage over PB-1 moves from +0.0508 °C on validation to −0.1198 °C on test, a transport contrast of −0.1706 °C under a convention where positive favors G2-long. The near-nominal validation calibration (`TR2`) becomes only partial on test. This second and final read fully spends 2016–2020; later calculations from the frozen records are audit-only and cannot create new confirmatory evidence.

16 Persistent Series Scorecard

Raw and calibrated `D1`–`D6`; `P1`–`P3`, `P4a`/`P4b`, `P5a`/`P5b`, `P6`; and `T1`–`T4` plus `T5a`/`T5b` for PE-2c, G0, G1, G2′, and G2-long under one canonical artifact (validation values above). A separate `E4b` block covers only the tested D and P metrics (Tables 2, and the raw/calibrated 0.263/0.641 P2, 20/38/45 → 43/61/68 P3, 0.4369 → 0.3890 P1); F1b `P4b`/`P5b` and all T metrics stay at E3. `TR1` = −0.1706 °C; `TR2` the raw-to-calibrated reliability changes above. `X4`: adopted checkpoint `16a865bc...`; CMIP pretrain `e2dde94a...` (16k); frozen mean `c6b49723...`; v3 backbone `2247c1f9...`; sampler η=1, ρ=0, 32 members, seed 42; authorizing amendment and materialization audit establish E4b status (a stale "validation-only" status string in the probabilistic summary despite `split=test` is superseded by the amendment and audit). Every value labels CRPS estimator version, lead scope, center identity, seed, η, ρ, member count, post-processing, and calibration fit/evaluation mode; lifecycle metrics on held-out CMIP are kept separate from descriptive observed cases.

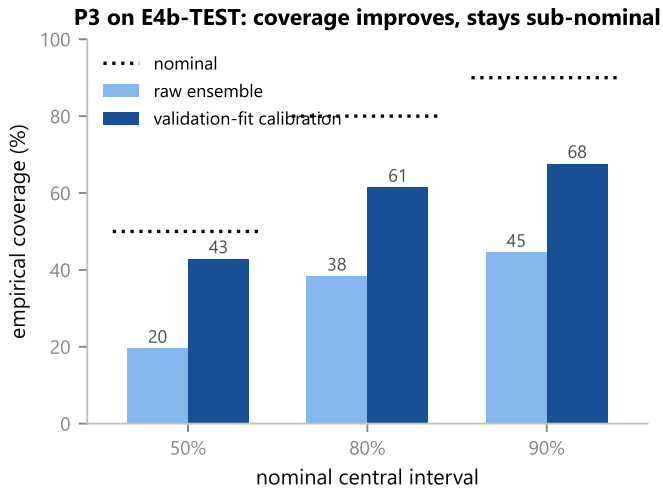

Figure 5: E4b interval coverage (`P3`), raw and under validation-fit calibration; dotted rules mark nominal. Calibration improves every interval but none reaches nominal — partial transfer.

17 Discussion

Mechanism. Architecture supplies capacity; equal-band weighting supplies coarse-scale gradient pressure; longer simulation training consolidates temporal structure and mean skill. The G1/G2′ single-variable contrast is an unusually clean demonstration that a sequence model can fit the marginal denoising task perfectly while learning none of the temporal structure the loss does not reward, and that re-weighting the same objective across scales causally unlocks it.

The reliability–flow-dependence tension and its sensitivity. The raw canonical `T5a` is positive (+0.354), but flat calibration makes it negative (−0.492) while improving average spread reliability — inflation that fixes the mean level of the spread flattens its state variation. The magnitude is center- and seed-sensitive (the historical +0.152 used a different diagnostic on different seeds), which is itself part of the report's caution.

The E3-to-E4b reversal. A mean advantage and near-nominal calibration developed under heavy adaptive reuse of 35 overlapping windows dominated by one event cycle reverted on 35 held-out windows with different events. The robust result is field RMSE at PB-1; the validation-specific results are the regional-mean and calibration advantages. The held-out evidence is *mixed* — neither equivalence nor superiority — which is what a single honest sealed read of a heavily consulted product should be expected to deliver.

18 Limitations and Required Future Evidence

Adaptive validation reuse and waived budgets through Phases E–G; only three LOYO folds; seed sensitivity (one fixed 32-member realization per artifact); no strict-gate pass before the adoption waiver; one lower-weight probabilistic E4b read; simple flat calibration; a residual increment-autocorrelation gap (members never reach observed +0.161); and no operational ONI/Brier verification. The archived CRPS axis-order defect touched every earlier Phase F/G gate record; canonical rescoring is complete for the final G2-long artifacts, pending for earlier rungs. Thin posthoc provenance; no test

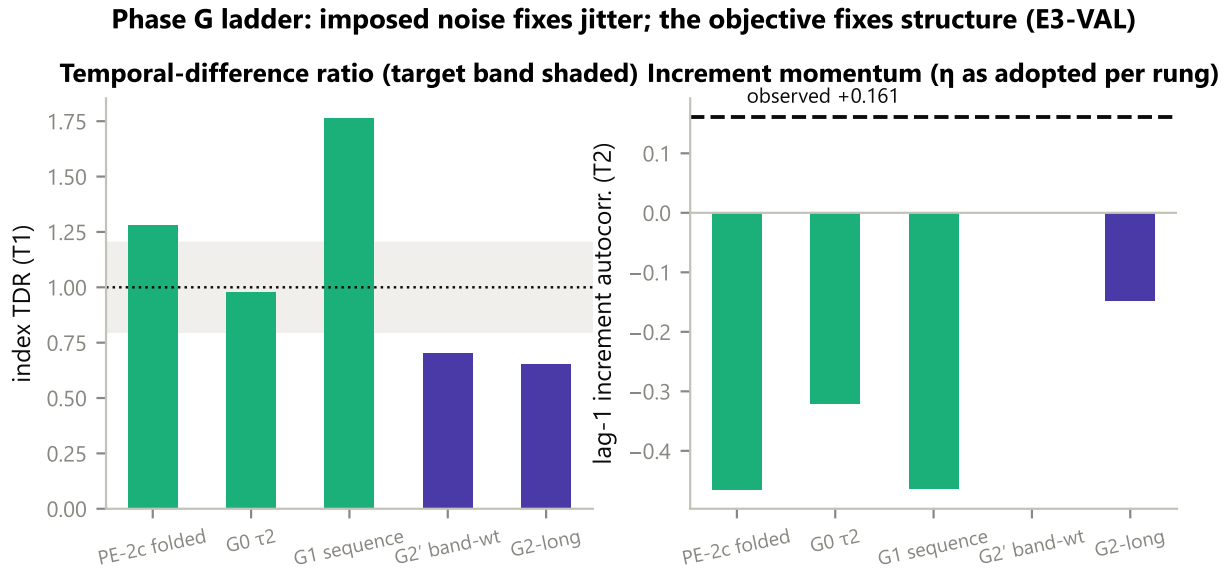

Figure 6: The Phase G ladder on temporal metrics (`E3-VAL`). Imposed noise (G0) fixes `T1` without touching increment momentum; the sequence architecture alone (G1) worsens jitter; the band-weighted objective (G2′/G2-long) is the only rung that moves `T2` toward observed while approaching the `T1` band.

of F1b/`T1`–`T5`; and no G2-long-versus-PB-1 equivalence interval. A new independent period/product or a prospective evaluation is required; the protocol-spent 2016–2020 split cannot support another held-out claim. The runtime gate verifies the declared checkpoint/document but maintains no global one-shot counter — finality rests on the amendment and the append-only ledger.

19 Conclusion

The temporal-head arc reached a real destination: a coherent probabilistic research product and a causal explanation for the temporal-objective breakthrough (a temporal architecture supplies capacity, but the per-cell objective must be re-weighted to give the coarse bands gradient pressure — and that single change unlocked learned coherence and a better mean, which in turn made post-hoc calibration CRPS-competitive for the first time; longer pretraining compounds it). The final claim is scoped exactly: the system was adopted for project use on validation evidence, and its later E4b result showed mixed point estimates plus partially calibrated uncertainty rather than a resolvable improvement over PB-1; raw physical members remained below the declared `P4a` amplitude thresholds. PB-1's first pristine `E4a-TEST` result stands as the project's strongest confirmatory evidence, and the shared test period is fully spent.

A Evidence Inventory

The paths below name the primary records in the project's internal research archive; the authorizing amendment, materialization audit, and per-window test records together establish the `E4b` claim, and the canonical rescoring artifacts carry the corrected CRPS values.

```
docs/temporal_generative_head_research_brief.md
docs/Temporarally Coherent Flow-Aware Diffusion ... Forecasting.md
```

```
docs/phase_g_temporal_head_plan.md
docs/final_protocol_amendment_g2long_sealed_test.md
configs/evaluate_coeff_native_g2long_sequence_bandweighted_test.yaml
outputs/evaluations/
  validation_ledger.md
  phase_g0_correlated_noise.json
  phase_g1_sequence_evaluation.json
  phase_g_correction_structure.json
  phase_g2_correction_structure.json
  phase_g2_sequence_evaluation.json
  phase_g3b_calibration.json
  phase_g3pp_blended_current.json
  phase_g2long_sequence_evaluation.json
  g2long_eta10_recentered.json
  phase_g2long_eta10_calibration.json
  phase_g2long_f1b_lifted.json
  g2long_sealed_test_posthoc.json
  coeff-native-g2long-sequence-bandweighted-eta10-SEALED-TEST/
outputs/forecasts/adopted_g2long_gallery/
src/cwe_enso/projection_calibration.py
tests/test_trajectory_metrics.py
tests/test_sequence_diffusion.py
tests/test_final_protocol_gate.py
```

References

[1] M. Mardani et al., "Residual corrective diffusion modeling for km-scale atmospheric downscaling," arXiv:2309.15214, 2023.

[2] M. Andrae, T. Landelius, J. Oskarsson, and F. Lindsten, "Continuous ensemble weather forecasting with diffusion models," arXiv:2410.05431, 2024.

[3] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[4] A. Blattmann et al., "Stable Video Diffusion: Scaling latent video diffusion models to large datasets," arXiv:2311.15127, 2023.

[5] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[6] I. Price et al., "Probabilistic weather forecasting with machine learning," *Nature*, 637:84–90, 2025.

[7] M. Scheuerer and T. M. Hamill, "Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities," *Monthly Weather Review*, 143:1321–1334, 2015.