Projection-Space Calibration and Coherent Lowpass Field Lifting

Scientific Report 7: The Phase F Ladder (F0–F3b, F1b)

Jashua Luna*

25 July 2026

Abstract

Held-out conditional-variance tests here are `E2-SIM`; all calibrated observed products are `E3-VAL`. We test whether explicit calibration of the Niño-3.4 projection can solve the regional under-dispersion left by the spatial diffusion ensemble PE-2c (Report 6). Six rungs run under a leave-one-context-start-year-out (LOYO) protocol: F0 (band decomposition of the index error and the ensemble spread), F1a (flat EMOS/ECC-Q index calibration), F2 (a CMIP model-analog trajectory baseline), F3 (state-conditional variance learned from CMIP persistence errors), F3b (extended conditioning features), and F1b (coherent lowpass field lifting). The central results are that average reliability is readily corrected — flat EMOS lifts Niño-3.4 spread–skill $0.401 \to 0.964$ with near-nominal coverage and exact mean identity — and that calibrated index spread can be materialized in sharp fields without disturbing the mean or highpasses (F1b: member std ratio $0.652 \to 0.787$, into the sharpness band, at exact index fidelity and mean-field invariance to $1.5 \times 10^{-7}\,^{\circ}\mathrm{C}$), but that available state features on a three-year validation slice do not deliver sufficient *flow-dependent* width: conditional variance is decisively learnable in simulation (held-out log-score +0.185, rank correlation 0.431) yet transfers to only 0.160 observed flow-dependence against a 0.30 target. F0 explains the shape of the problem: 89% of PB-1's index-error variance lives in the two coarsest bands, and the ensemble spread is already 81% lowpass — so the deficit is amplitude and case-dependence of the coherent mode, not band mis-allocation. Every Phase F CRPS value is affected by an archived lead/member-axis scorer defect and must be canonically recomputed before its magnitude is retained; the qualitative gate directions are unaffected. PB-1 remains the adopted deterministic model at this phase.

1 Study Identity and Handoff

Study ID R7 / Phase F. **Incoming baseline:** PE-2c and the regional calibration wall from Report 6. **Changed factor:** make the Niño-3.4 projection explicit through calibration, analogs, conditional variance, and lowpass lifting — no retraining of the generator. **Evidence tier:** repeatedly consulted validation with LOYO fitting; no test access. **Outgoing diagnosis:** temporal and flow-aware structure must move into the generator (Report 8); the calibration and lifting machinery built here is reused downstream unchanged. **Timeline:** 2026-07-10; budgets waived, every rung ledger-predeclared.

2 Introduction

Report 6 established that a spatially structured generative ensemble around the frozen PB-1 mean still under-disperses the Niño-3.4 index, across head family, loss, sampler, and a 64-fold data increase

*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 wall is architectural in the sense that the generator does not place enough variance in the modes that survive a box average. Phase F attacks the projection directly. The Niño-3.4 index is one linear functional of the 37-channel coefficient tensor; nothing in the generator gives gradient or sampling pressure to that direction specifically. Three questions organize the ladder:

1. Where does PB-1's index error live, by scale, and where does the current ensemble spread live? (F0)
2. Do the existing member ranks contain calibratable information — i.e. does inflation plus rank-preserving remapping reach nominal reliability? (F1)
3. Can climate simulations identify *state-dependent* coherent-mode variance that transfers to observations? (F2, F3)

A discipline governs all fitting: 35 overlapping validation windows overfit easily, so every fitted parameter is fit and scored under LOYO by context-start year (2011/2012/2013), with gate metrics pooled over held-out folds only. Table 1 maps each rung to the question it answers.

Rung	Question	Outcome
F0	Where do index error and spread live, by band?	89% of error variance in the two coarsest bands; spread already 81% lowpass — the deficit is amplitude and case-dependence, not routing.
F1a	Does flat inflation reach reliability?	Spread–skill 0.964 with near-nominal coverage and exact mean identity; archived CRPS worsens.
F2	Does the CMIP archive hold conditional uncertainty?	Raw spread–skill 0.738, the project's best uncalibrated value — but no flow dependence (0.087).
F3	Does CMIP-fit conditional variance transfer?	Decisively learnable in simulation (+0.185 log-score, rank corr. 0.431); observed flow dependence only 0.160 against the 0.30 target.
F3b	Do added SST precursors help?	Stage-1 gate failed against the base; observations correctly never consulted.
F1b	Can calibrated index spread materialize in sharp fields?	Exact index fidelity and mean invariance; member std ratio 0.652 → 0.787, into the sharpness band.

Table 1: The Phase F ladder: one question per rung, each predeclared in the ledger before its result was inspected.

3 Focused Literature Review

Phase F is a disciplined application of ensemble post-processing to one aggregated functional. Ensemble model output statistics (EMOS / non-homogeneous Gaussian regression) correct bias and dispersion with few parameters [1]; ensemble copula coupling (ECC) and its dependence-preserving variants reorder calibrated marginals by the raw ensemble ranks, retaining lead-to-lead structure [2]. Analog ensembles provide state-conditioned members by retrieval [3]. Conditional heteroscedastic models predict $\sigma(\text{state})$ [4], and conformal methods offer distribution-free coverage as a fallback [5]. The specific warning the external review raised — that direct multivariate seasonal calibration

can look excellent in-sample and vanish in cross-validation for lack of data relative to parameters — is exactly why the design is deliberately low-parameter and LOYO-blocked; operational seasonal systems variance-correct Niño-3.4 indices with restrained statistical models rather than flexible neural calibrators, treating the regional index as a special calibration target rather than a quantity that emerges correct from a field ensemble. Each rung below tests one recommendation from that review.

4 Validation and Metric Protocol

Three LOYO folds hold out context-start years 2011, 2012, and 2013 in turn (fold sizes 12/12/11 windows); leads 3–14 define the index headline scope. Fair CRPS is scored per window and lead over the member axis. A recorded defect governs all CRPS reading in this report: the archived projection scorer reshaped [window, member, lead] arrays without moving the member axis last, mixing leads and members; spread, coverage, and deterministic metrics are unaffected, but every Phase F CRPS value must be canonically recomputed before its magnitude is retained. These values are also not interchangeable with the all-lead legacy probabilistic evaluator's CRPS. Where a CRPS number appears below it is the archived value, labeled as such.

4.1 Common Forecast Coordinates

24×48 tropical Pacific; 12/14-month geometry; train 1981–2010; validation 2011–2015 (35 windows); no 2016–2020 access; cosine-weighted Niño-3.4 box; persistence and zero-anomaly baselines; 32 members. Calibration headlines restrict scoring to leads 3–14.

5 F0: Bandwise Error and Spread Decomposition

Band-limited inverse DTCWT decomposes the PB-1 error field (truth minus mean head) and the ensemble correction fields into lowpass and highpass levels 1–3, and the Niño-3.4 index of each band-limited component is taken. Writing the index error variance as a sum of band contributions,

$$\mathrm{Var}\big(N[\,\mathrm{err}\,]\big) \;=\; \sum_{b\in\{\mathrm{lp},1,2,3\}} \mathrm{Var}\big(N[\,\mathrm{err}_b\,]\big) + \text{cross terms}, \tag{1}$$

the measured shares are **66% lowpass, 23% level 3, 11% level 2, 0.1% level 1** — 89% in the two coarsest bands, verifying the scale-routing premise (Fig. 1). The ensemble's index *spread* is already 81% lowpass. So the failure is not band mis-allocation at the index level: the ensemble puts its index spread in roughly the right band. The deficit is the *amplitude* and the *case-dependence* of that coherent lowpass mode. F0 carries no gate; it scopes every design downstream.

6 F1a: Flat Projection-Space EMOS/ECC-Q

The predictive model for the index deviation at lead ℓ is Gaussian with a lead-smooth error floor,

$$Y_\ell - A(\mu_{\mathrm{PB1},\ell}) \;\sim\; \mathcal{N}\big(0,\ \sigma^2(\ell)\big), \qquad \sigma^2(\ell) = \exp\big(P_d(\ell)\big) + b^2\, s^2_{\mathrm{raw}}(\ell), \tag{2}$$

with P_d a low-degree polynomial in lead (≤ 3 effective parameters) and b an optional dependence on raw ensemble spread ($b = 0$ first, $b > 0$ second). Members are reconstructed by rank-preserving Gaussian-quantile remapping (ECC-Q, Eq. 9 of the series charter): each member keeps its per-lead

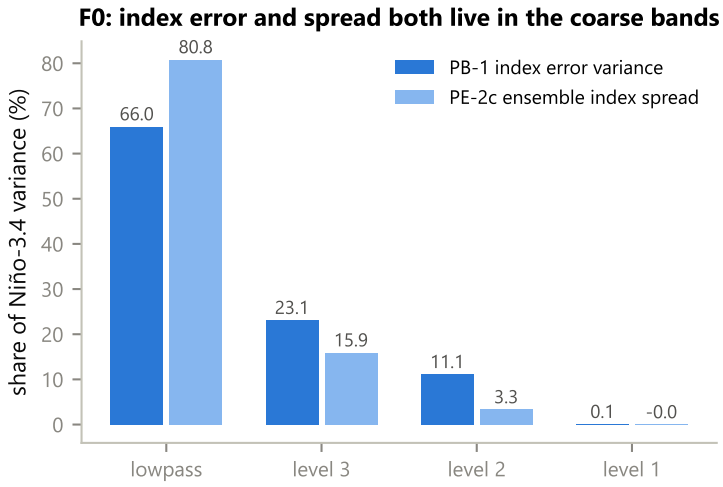

Figure 1: F0 decomposition (`E3-VAL`, 35 windows). PB-1's Niño-3.4-index error variance and the PE-2c ensemble's index spread both concentrate in the lowpass and level-3 bands — the coherent scales that survive the box average. The problem is amplitude and case-dependence there, not band routing.

rank, its deviation becoming $\sigma(\ell)\,\Phi^{-1}((k-\frac{1}{2})/M)$, recentered exactly so the index ensemble mean equals PB-1's.

At $b=0$: spread–skill **0.401** → **0.964** (in band), coverage 56/77/81% (all within 10 points of nominal), exact mean identity — but archived fair CRPS *worsens*, 0.245 → 0.270 (+10%) for PE-2c. Three of four gates pass; the CRPS gate fails, pending canonical recomputation. Crucially, $b=1$ is nearly indistinguishable from $b=0$ in the preserved diagnostics, and the correlation between |index error| and raw spread across windows is only 0.05–0.13: the raw ensemble spread is a flow-independent constant ribbon (0.198/0.200/0.198 by fold) while fold RMSE swings 0.246 → 0.758 through the 2015-event fold (Fig. 3). Climatological inflation ($\sim 2.5\times$) therefore fixes *average* reliability but over-widens quiet folds and still under-covers the event fold — exactly the small-case-dependence effect the external review warned about. Per the predeclared interpretation rule, the ensemble lacks flow-dependent coherent-mode uncertainty, so F3 becomes mandatory rather than optional.

7 F2: CMIP Model-Analog Trajectory Baseline

Each observed initialization is represented by a compact standardized state (12-month Niño-3.4-3.4/-3/-4 context trajectories); candidates are season-matched on context-end calendar month across all 14 CMIP stores; per-model top-k=3 nearest neighbors (Euclidean, per-model ENSO-amplitude standardized, $\geq$ 12-month separation) give a 42-member, model-balanced ensemble; analog future 14-lead index deviations are recentered across the ensemble, rescaled by the observed context σ, and added to the PB-1 index trajectory (mean preserved exactly). The predeclared signals: raw analog spread–skill > 0.65 would show the CMIP archive contains usable conditional uncertainty; corr(|error|, spread) materially above 0.05–0.13 would show analogs supply the missing flow dependence.

Outcome. Signal 1 **met**: raw spread–skill 0.738 (the best uncalibrated value in the project; respectable raw coverage 51/76/82%) — the archive does contain conditional uncertainty. Signal 2 **not met**: corr(|error|, spread) = 0.087, no better than the diffusion ensembles — forced k-per-model

retrieval makes ensemble width reflect the model pool's climatological diversity, near-constant across windows, not state-conditional uncertainty. LOYO-calibrated ($b = 0$): spread–skill 0.972, 3/4 gates, archived CRPS again fails (0.294 vs raw 0.281); $b = 1$ over-disperses (1.206). The synthesis after F1a+F2: average reliability is cheaply solvable but the CRPS gate fails *identically* in every ensemble because the singular remaining deficit is flow dependence.

8 F3: Conditional Variance from CMIP

The design consequence of F0–F2 is to model $\sigma(\text{state}, \ell)$ explicitly, with CMIP informing amplitude and observations supplying only a tiny adaptation. A conditional log-variance model is fit by model-balanced OLS on ∼380k rows of CMIP persistence-error pairs (persistence error being the observable CMIP proxy for a persistence-residual forecast's error):

$$\log e^2 \;=\; \beta_0 + \beta_1\,\ell + \beta_2\,\ell^2 + \beta_3\,|s_0| + \beta_4\,|s_0|\,\ell + \beta_5\,\tau_6 + \beta_6\,|\tau_6| + \beta_7\,\sin\tfrac{2\pi m}{12} + \beta_8\,\cos\tfrac{2\pi m}{12}, \quad (3)$$

with $|s_0|$ the per-model-standardized context-end Niño-3.4 amplitude, τ_6 the 6-month context tendency, and m the target calendar month (spring-barrier seasonality). Observed adaptation is one scale c per LOYO fold, $c^2 = \langle e^2/\sigma^2_{\text{CMIP}}\rangle_{\text{fit}}$; members are reconstructed with ECC-Q driven by $c\,\sigma_{\text{CMIP}}(\text{state}, \ell)$; the PB-1 mean is fixed exactly.

8.1 Held-Out-Simulation Falsification (`E2-SIM`)

Stage 1 gates observed use on CMIP generalization. On the three held-out models the conditional model beats the lead-only baseline by +0.185 Gaussian log-score with rank correlation 0.431 between predicted σ and |error|. The learned physics is sensible: $\hat\beta_{|s_0|} = +0.25$ and $\hat\beta_{|s_0|\times\ell} = +1.28$ (larger initial ENSO state → larger error, growing with lead), with active seasonal terms. State-conditioning of persistence-forecast error is learnable from CMIP; proceed to observations.

8.2 Observed Conditional Calibration (`E3-VAL`)

Stage 2, three of five gates: spread–skill **1.089** (pass), coverage 60/83/90% (pass; 90% exactly nominal), mean identity exact (pass); archived fair CRPS 0.2645 vs raw 0.2450 (fail, though better than flat F1a's 0.2700 — conditioning recovers ∼20% of the flat cost); flow dependence **0.160** (fail vs the 0.30 target, though 2–3× every prior ensemble's 0.05–0.13). The per-fold observation scales are $0.90/1.02/0.80 \approx 1$: the CMIP-standardized persistence-error magnitude transfers to PB-1's observed errors nearly one-to-one, validating the proxy.

The named ceiling. Predicted σ varies 0.12–1.83 across windows (33% variation) yet flow dependence tops out at 0.160 — in significant part because the 2013-start contexts preceding the 2014–15 event ramp were near-*neutral* states (the famously unforecastable 2014 "failed El Niño" transition), so |initial state| genuinely under-predicts those errors on this three-year slice. Higher flow dependence may simply be unidentifiable here without richer precursors — a new predeclared variant, not a tweak. This is the best-balanced probabilistic index product to date, not adopted under the strict gates.

9 F3b: Extended SST Precursors

Five added SST-index precursors (zonal gradient nino4 − nino3, absolute gradient, gradient tendency, context volatility, and interactions — all computable identically on CMIP and observations, no

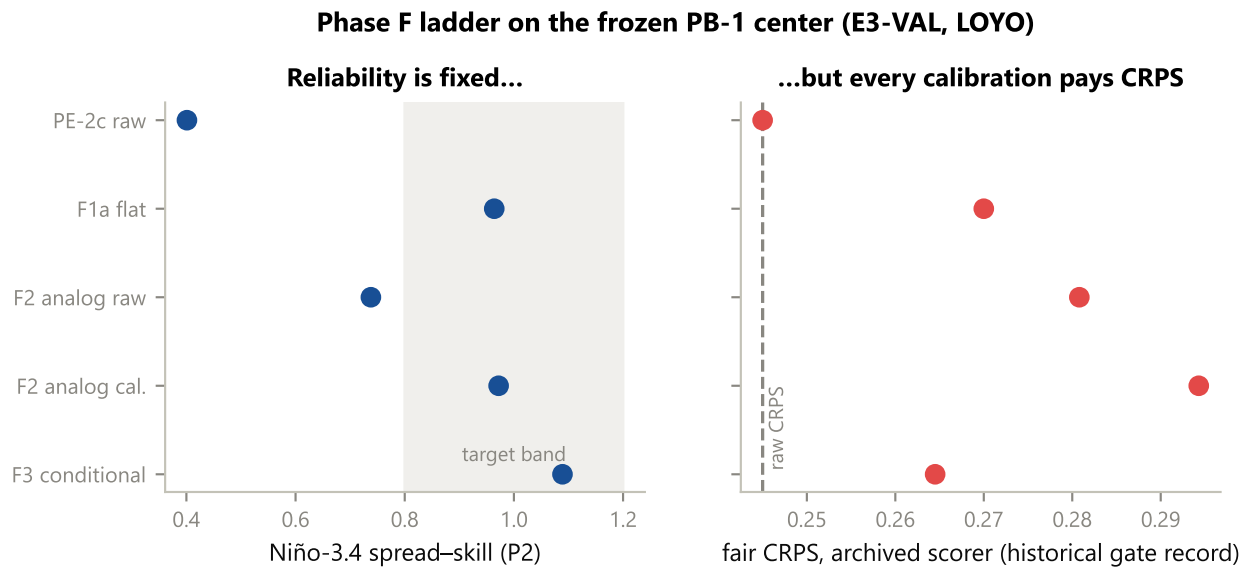

Figure 2: The Phase F ladder on the frozen PB-1 center (`E3-VAL`, LOYO). Left: every calibration reaches the spread–skill band. Right: every calibration pays archived fair CRPS relative to raw — the flow-dependence deficit made visible as a scoring penalty (values pending canonical recomputation).

subsurface data) were tested at stage 1 against the F3 base. The gate is a clean vs-base comparison on held-out CMIP. Result: the full 14-feature log-score is 1.0495 versus the 9-feature base's 1.0364 (i.e. *worse*) with rank correlation 0.428 vs 0.431 — the added surface features fit within-model idiosyncrasies rather than transferable physics. Stage 2 (observations) was correctly never run per the predeclared rule. The F3 nine-feature σ stands; genuinely new information (subsurface OHC/WWV) would be the next candidate precursor class, carrying the Report 5 observed-transfer caveat.

10 F1b: Coherent Lowpass Field Lifting

The calibrated index change is lifted back into member fields through a leadwise basin pattern B_ℓ — a ridge regression ($\lambda = 0.1\,\langle q^2 \rangle$) of fit-fold PB-1 error fields on their Niño-3.4 residuals, band-limited to the DTCWT lowpass band and normalized to unit Niño-3.4 average $A(B_\ell) = 1$ — applied to the zero-member-mean deviation difference:

$$r'_m \;=\; r_m + B_\ell\,[\,(q'_m - \bar{q}') - (q_m - \bar{q})\,], \qquad \textstyle\sum_m [(q'_m - \bar{q}') - (q_m - \bar{q})] = 0. \tag{4}$$

Because the lifted quantity has zero member mean, the ensemble-mean field is invariant by construction; because B_ℓ is lowpass and unit-normalized, member indices reproduce the calibrated values exactly and highpasses are preserved. Results: index fidelity max error 0.00000 °C; ensemble-mean field invariant to 1.5×10^{-7} °C; highpass band-energy changes 0.3–2.0% ($< 5\%$ gate; lowpass +19.4%, where the calibrated spread intentionally lives); member std ratio (`P4b`) 0.652 → **0.787**. Three of four gates pass. The declared-failed gate — member std change < 0.05 — is triggered precisely because the calibrated $\sim 2.5\times$ index spread *materializes* in the fields (the lift's purpose), landing inside the sharpness target $[0.7, 1.3]$; its small-perturbation premise is incompatible with conditional-σ inflation, so the "failure" is an improvement on the sharpness metric, recorded without changing the gate. An implementation note of record: the first version lifted $q' - q$ directly and

shifted the mean field up to 0.43 °C because the raw ensemble-mean index drifts from PB-1's; the corrected v2 lifts only the deviation difference, restoring exact mean invariance and leaving the small raw-mean offset in the mean (where PE-2c's raw ensemble mean is the better point forecast).

11 Persistent Series Scorecard

Rung	Badge	P2	P3 50/80/90	CRPS*	flow-dep.	gates
PE-2c raw	E3	0.401	33/54/64	0.2450	0.054	—
F1a flat (b=0)	E3	0.964	56/77/81	0.2700	—	3/4 (CRPS)
F2 analog raw	E3	0.738	51/76/82	0.2808	0.087	signal 1 met
F2 analog cal. (b=0)	E3	0.972	—	0.2943	—	3/4 (CRPS)
F3 conditional	E3	1.089	60/83/90	0.2645	0.160	3/5
F3b (14-feature)	E2-SIM	stage-1 gate failed vs base (log-score 1.0495 > 1.0364)				
F1b lifted fields	E3	index fidelity exact; mean inv. 1.5×10^{-7}; P4b 0.652 → 0.787; 3/4				

Table 2: Phase F scorecard (E3-VAL / E2-SIM). *archived fair CRPS, leads 3–14, affected by the lead/member-axis defect — provenance column, not P1; every value requires canonical per-window/per-lead recomputation before its magnitude is retained.

Calibration diagnostics of record: F3 per-fold scales 0.90/1.02/0.80; mean-identity errors at machine precision throughout; predicted-σ variation 0.12–1.83; sigma–|error| correlation 0.160; F1b lifting integrity as above. X4: no new checkpoint (inference-only post-processing of PE-2c 8ad14ac0...); artifacts phase_f_f0_band_decomposition.json, phase_f_f1a_pe2c_b0.json, phase_f_f2_analog_baseline.json, phase_f_f3_conditional_calibration.json, phase_f_f3b_conditional phase_f_f1b_lifted_members.json; LOYO folds throughout; test_data_read: false.

12 Synthesis and Decision

F1 fixes average reliability; F2 provides broader raw spread but without state dependence; F3 transfers conditional physics from CMIP but remains too weak on the observed slice (0.160 vs 0.30); F3b adds no transferable signal; and F1b successfully constructs the field product. The through-line is one deficit, measured five ways: *flow dependence*. Any flow-independent width that fixes average coverage necessarily over-widens quiet periods and under-covers events, so the CRPS gate fails identically across F1a, F2, and F3. Every CRPS-dependent gate must be rechecked after canonical recomputation; that audit is distinct from the historical decision, under which no F variant cleared every strict gate and PB-1 remained the adopted deterministic model at this phase.

13 Discussion

The report separates two calibration goals that the field often conflates: *average reliability* (fixed cheaply, by any inflation) and *resolution / flow dependence* (the hard part, requiring spread that tracks the state). CRPS penalizes the second failure even when the first is solved, because mean error enters CRPS and event windows dominate it. Event-fold influence is severe here: one three-year slice contains essentially one event ramp, grown from neutral states, so the very windows that most need wide spread are the ones the available precursors least predict. The constructive implication — and the explicit handoff to Report 8 — is that a generator modeling *joint lead structure* may improve both the conditional mean and the coherent low-mode uncertainty *before* calibration, changing

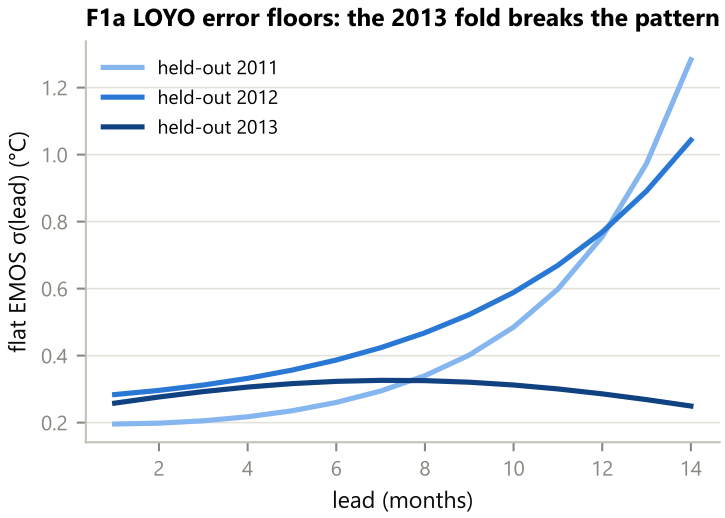

Figure 3: F1a flat error floors $\sigma(\ell)$ by LOYO fold. The 2013 fold — which holds out the 2014–15 event ramp — has a qualitatively different, non-monotone shape, the concrete signature of the flow-dependence and event-fold problem that defeats flow-independent calibration.

the terms of the trade. The Phase-F-era "$\sim$0.22 ceiling" on flow dependence is therefore feature-, center-, seed-, and estimator-specific, not a universal constant; Report 8's canonical G2-long raw flow dependence (+0.354) is measured on a different center and reported separately.

14 Limitations

Only three LOYO folds; repeated validation use (budgets waived); strong-event leverage (the 2015 ramp dominates the event fold); simple linear σ features; analog dependence among CMIP models; raw-spread weakness (F1a b=0 $\approx$ b=1); calibration-estimator uncertainty; and no test access within Phase F. Report 8 later tests the validation-fit flat calibration unchanged as one component of the amended G2-long read. The archived lead/member-axis CRPS defect touches every CRPS number in this report; a frozen corrected-score artifact is required before any magnitude is quoted as `P1`.

15 Conclusion and Transition to Report 8

Phase F hands forward the explicit Niño-3.4 calibration and lifting machinery (EMOS/ECC-Q; the CMIP-fit conditional σ; the lowpass unit-index lifting pattern) — all reused unchanged in Report 8 — and the sharpened diagnosis that flow-independent spread is the remaining wall. The next hypothesis follows directly: a joint *sequence* generator with direct temporal and coarse-band objective pressure may learn better mean trajectories and state-dependent corrections, moving flow dependence into the generator where post-hoc calibration could not manufacture it.

A Evidence Inventory

The paths below name the primary records in the project's internal research archive; each rung's JSON artifact carries its fitted parameters, gate outcomes, and LOYO fold assignments.

`docs/ensemble_calibration_research_brief.md`
`docs/Fixing Under-Dispertion in a Wavelet Diffusion ENSO Forecaster.md`
`docs/phase_f_calibration_plan.md`
`outputs/evaluations/phase_f_f0_band_decomposition.json`
`outputs/evaluations/phase_f_f1a_pe2c_b0.json`
`outputs/evaluations/phase_f_f2_analog_baseline.json`
`outputs/evaluations/phase_f_f3_conditional_calibration.json`
`outputs/evaluations/phase_f_f3b_conditional_calibration.json`
`outputs/evaluations/phase_f_f1b_lifted_members.json`
`src/cwe_enso/projection_calibration.py`
`tests/test_projection_calibration.py`
`tests/test_conditional_variance.py`

References

[1] T. Gneiting, A. E. Raftery, A. H. Westveld, and T. Goldman, "Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation," *Monthly Weather Review*, 133:1098–1118, 2005.

[2] R. Schefzik, T. L. Thorarinsdottir, and T. Gneiting, "Uncertainty quantification in complex simulation models using ensemble copula coupling," *Statistical Science*, 28(4):616–640, 2013.

[3] L. Delle Monache, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, "Probabilistic weather prediction with an analog ensemble," *Monthly Weather Review*, 141:3498–3516, 2013.

[4] D. A. Nix and A. S. Weigend, "Estimating the mean and variance of the target probability distribution," in *IEEE International Conference on Neural Networks*, 1994.

[5] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, Springer, 2005.