LEARNING EDITABLE PREDICTIVE STATES FOR OCEAN WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurately modeling the coupled evolution of the ocean carbon–physics system and its accompanying events is a long-standing challenge in climate science. Conventional data-driven methods often rely on the regression capacity of neural networks to map past surface fields to future ones without a state into which a physical change could be written, or model individual events in isolation, neglecting the information that sparse event records carry about the continuous fields. In this paper, we propose the State-Perturbation Joint-Embedding Predictive Architecture (SP-JEPA), a world model that represents the environment by a compact state and advances this state autonomously. Central to our method is a basis-anchored state: each frame combines the coefficients of a fixed spatial basis with a learned residual representation, so that a perturbation written into the coefficients changes the decoded field exactly by the prescribed pattern, and a single predictor propagates both the natural and the perturbed futures without reading future observations. Historical fields and source-grounded event observations jointly initialize this state, from which future fields and events are read out. We evaluate natural prediction on three decades of global monthly surface carbon–physics fields with six event families, and response prediction in a controlled transport–exchange environment that provides the true future before and after each perturbation. On the validation splits, SP-JEPA reaches a normalized root mean squared error of 0.131, level with the best field forecaster and 18% below the best compared latent world model, and fusing historical events lowers event-prediction error by 9–11% against absent or mismatched event inputs without increasing field error. On mechanisms unseen during training, its rollout reproduces the mean response of each pulse and reaches a response skill of 0.575 and a latent counterfactual skill of 0.206, whereas compared latent world models score below zero.

1 INTRODUCTION

The ocean absorbs roughly a quarter of anthropogenic CO_2 emissions through coupled physical, chemical, and biological processes (Gruber et al., 2019). Monthly surface fields describe part of this evolution, while storms, upwelling, the El Niño–Southern Oscillation (ENSO), and marine heatwaves reveal processes at different temporal and spatial scales (Feely et al., 2002; Mignot et al., 2022). A model of this system should predict its natural evolution from both kinds of observation and also predict how a specified physical change would alter that evolution.

Field reconstruction and forecasting estimate ocean variables from gridded histories (Landschützer et al., 2014; Li et al., 2021; Nguyen et al., 2023). Event records are often analyzed separately, although they can supplement processes poorly resolved by monthly fields or summarize patterns within them. Joint-embedding predictive architectures (JEPA) make a learned predictive state explicit (Assran et al., 2023; Maes et al., 2026), but its coordinates alone do not define how to apply a prescribed field change. Accurate natural forecasting therefore leaves open whether the same state can predict a perturbation's consequences (Vakalis, 2026).

We propose **SP-JEPA**, which represents each state frame by the coefficients of a fixed spatial basis and a learned residual representation, adds historical event observations to the residual, and advances natural and perturbed states with one predictor that outputs coefficient increments and new residual vectors, from which future fields and events are read out.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The *Global Ocean Carbon–Physics dataset* combines five monthly surface fields from 1993–2023 with six event families and tests natural forecasting. Because this record contains only one realized future, we evaluate responses separately in the *Paired Transport–Exchange environment*, which supplies paired futures under sixteen localized pulses. The two settings share the state, predictor, and edit, differ in the basis and readouts, and are trained separately; only the ocean model fuses events.

Our contributions are:

1) **A predictive state with exact physical edits.** Each state frame combines the coefficients of a fixed spatial basis with a learned residual representation, and a complementary readout prevents the residual from altering the coefficients. A perturbation written into the coefficients therefore changes the decoded field by exactly the prescribed pattern, and the same predictor advances the natural and the edited state. The coefficients also carry the seasonal cycle of the rollout: removing them raises the full-scale ocean field error by 27%.

2) **Historical events as state observations.** Six source-grounded event families enter the initial state as standardized anomalies alongside the field encoding, while future events serve only as prediction targets. Compared with models that receive no event history or the event history of other years, the observed records lower the error of the three event families with learned readouts by 9–11% without increasing field error, so the gain comes from the observed values rather than from the added input pathway.

3) **Capture–propagation evaluation of responses.** In an environment with paired true futures, we separate the response error introduced by encoding and readout from the error added by the rollout, which gives a lower bound on response skill. This decomposition shows that the fixed basis keeps a pulse in the state and that response supervision, rather than paired training data alone, lets the rollout propagate it. On mechanisms unseen during training, SP-JEPA reproduces the mean response of each pulse, whereas the compared latent world models lose most of the response already at encoding.

2 Related Work

Three strands of work motivate our design and evaluation; Appendix E extends this review.

Field reconstruction and forecasting. Statistical and neural reconstructions estimate surface pCO_2 and air–sea flux from sparse observations (Landschützer et al., 2014; 2016), supporting carbon-budget assessments (Le Quéré et al., 2018), and thermal and non-thermal decompositions attribute the resulting variability (Takahashi et al., 2002). Forecasting instead targets future fields given the observed history: neural operators, transformers, and climate foundation models learn this mapping over multiple variables and scales (Li et al., 2021; Holzschuh et al., 2025; Nguyen et al., 2023; Bodnar et al., 2025) and provide our natural-prediction comparisons. We additionally evaluate a compact recurrent state whose coordinates specify how a physical perturbation enters the rollout.

Event processes and observations. Studies of cyclones (Bates et al., 1998), marine heatwaves (Mignot et al., 2022), ENSO (Feely et al., 2002; Liao et al., 2020), and upwelling (Feely et al., 2008) relate specific processes to carbon exchange. Because their source records differ in duration and spatial support, we define source-derived monthly summaries with known-flags and use them as observations and prediction targets, bringing process information into the initial state while preserving each family's measurement scale.

Latent dynamics and event-aware representations. World models learn transitions for prediction and control (LeCun, 2022; Hafner et al., 2025), and JEPA predicts representations of future observations (Assran et al., 2023; Bardes et al., 2024); LeWorldModel provides our shared-encoder and regularization recipe (Maes et al., 2026). EAWM derives event boundaries from observations to shape policy representations (Peng et al., 2026), whereas our source-defined events are an additional observation of the initial environment. Geophysical time-series and sensing models extend latent prediction to physical systems (Nie et al., 2026; Deng et al., 2026), but action conditioning alone does not define an exact correspondence between a spatial perturbation and a latent edit. We establish that correspondence with fixed basis coordinates and compare the predicted response with paired outcomes, evaluating intervention fidelity separately from factual prediction (Vakalis, 2026).

3 Problem Setup

At time t, observations contain a standardized multichannel field $x_t \in \mathbb{R}^N$, a validity mask m_t, and a calendar encoding c_t. Ocean observations add 13 event values e_t and six known-flags κ_t (Appendix A.3). From the last L observations, H_t, the model constructs a state window S_t and predicts K future fields and events. The rollout is **autonomous**: each step uses only the preceding state window and known calendars, and no true future observation or re-encoded field prediction enters the recurrence. Field readouts use predicted frames, event readouts the updated window.

For response prediction, an action $u \in \mathcal{U}$ adds a prescribed pattern δ_u to the current field while keeping the dynamics and earlier history fixed. Its response is

$$r^*_{t+k}(u) = x^u_{t+k} - x_{t+k}, \qquad k = 1, \dots, K, \tag{1}$$

where x_{t+k} and x^u_{t+k} are the natural and perturbed futures. We compare this difference with two model rollouts from the same encoded history (Figure 1b). The paired environment supplies both true futures; the ocean record supports only the natural forecasting task.

4 Method

Figures 1 and 2 show how the state is formed, edited, and advanced; Appendix C gives pseudocode.

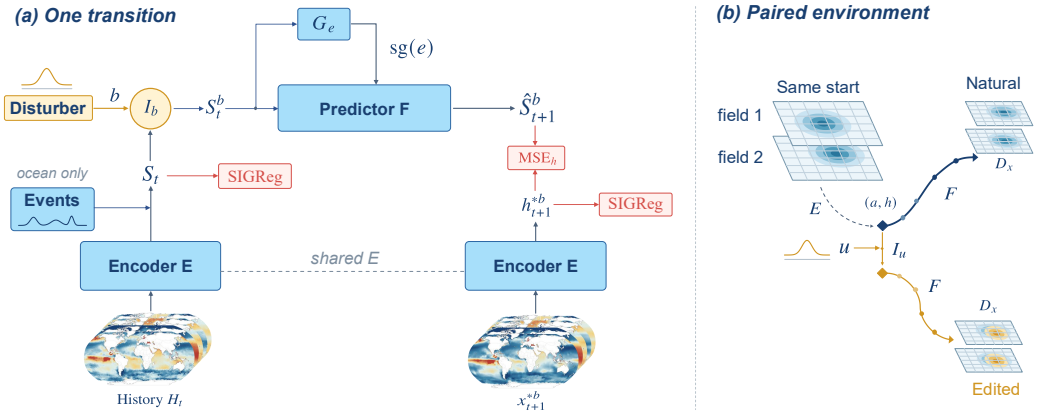

Figure 1: One prediction step and a paired query. (a) Historical fields and events form S_t; the disturber edits its last coefficients, and F_θ predicts the next frame, whose residual representation JEPA matches to the shared encoding of the true future. In the paired environment and the first ocean stage, $\mathrm{sg}(G_e(S_t))$ also conditions F_θ. (b) Natural and edited branches share F_θ; stacked grids are the two simulated fields, and contours are schematic.

4.1 Basis-Anchored State

Each frame $z_t = (a_t, h_t)$ contains $r = 32$ basis coefficients and a $d = 128$ learned representation. Let $A \in \mathbb{R}^{N \times r}$ be a fixed spatial basis and W a diagonal matrix of grid weights, with $A^\top W A$ invertible. We project the field onto A and encode the remaining field:

$$C_A = (A^\top W A)^{-1} A^\top W, \quad Q = I - A C_A, \qquad a_t = C_A x_t, \quad h^x_t = E_\theta(Q x_t, m_t, c_t), \tag{2}$$

where E_θ is a patch transformer. A learned linear map d_θ supplies the residual field readout,

$$D_x(a, h) = A a + Q\, d_\theta(h). \tag{3}$$

The identities $C_A A = I$ and $C_A Q = 0$ give $C_A D_x(a, h) = a$: **the residual readout cannot alter the basis coefficients**. For any $b \in \mathbb{R}^r$,

$$D_x(a + b, h) - D_x(a, h) = A b, \tag{4}$$

so coefficient changes have a fixed physical meaning. The ocean basis contains scaled leading singular vectors of standardized training fields, with cell areas only in W; the paired basis has 32 periodic Gaussian templates covering both fields and all action centers, with $W = I$ (Appendix C.1).

4.2 Event-Observation Fusion

Storm tracks, upwelling indices, and carbon-flux records observe processes beyond the five monthly fields, and temperature-derived indices summarize patterns within them; we use these six event families as observations of the historical state. The fusion layer subtracts training-period monthly means $\bar{e}_{\mathrm{mo}(\tau)}$, divides by the anomaly scale σ_e, and appends the known-flags:

$$\xi_\tau = \Big[\, \omega_\tau \odot \big(e_\tau - \bar{e}_{\mathrm{mo}(\tau)}\big) \oslash \sigma_e,\ \ \kappa_\tau \,\Big] \in \mathbb{R}^{19}, \qquad h_\tau = h^x_\tau + B_e \xi_\tau + b_e, \tag{5}$$

where ω_τ masks unknown values and the cyclone upper bound, which is confirmed only retrospectively. The zero-initialized fusion changes h but leaves a and Eq. (4) intact. It enters only the initial ocean history, future events serve as targets, and the paired environment encodes fields alone.

4.3 Increment Predictor and Autonomous Rollout

The predictor F_θ, which plays the role of the state-transition function, processes $S_t = [z_{t-L+1}, \ldots, z_t]$ with a causal transformer conditioned on the next calendar. Its two output heads predict a coefficient increment and a new residual representation:

$$\big(\Delta a_{t+1},\, \hat{h}_{t+1}\big) = F_\theta\big(S_t, c_{t+1}\big), \qquad \hat{a}_{t+1} = a_t + \Delta a_{t+1}. \tag{6}$$

The zero-initialized coefficient head starts from persistence. Each step drops the oldest frame and appends $(\hat{a}, \hat{h})$, so later predictions use the preceding predicted states. In the paired environment and the first ocean training stage, F_θ is also conditioned on $\mathrm{sg}(G_e(S_t))$, the detached event readout of its current window (Section 4.4); this path is disabled in the final ocean model.

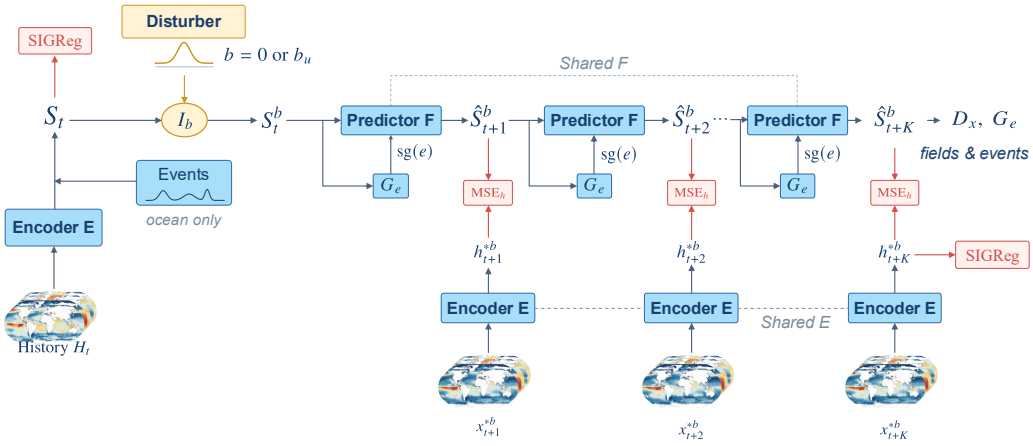

Figure 2: Autonomous rollout after one initial edit. Each step appends a predicted frame to the window; true futures enter only the training targets. Readouts act at each lead, and SIGReg regularizes the context and target encodings of natural windows ($b = 0$). $\mathrm{sg}(G_e)$ acts as in Figure 1.

4.4 Field and Event Readouts

Fields are read from each predicted frame, $\hat{x}_{t+k} = D_x(\hat{a}_{t+k}, \hat{h}_{t+k})$. The event readout G_e reads events from the updated window, which retains the consecutive months needed for flux reversals and the trailing Niño 3.4 index. The final ocean G_e adds a learned residual to a training-period seasonal baseline,

$$\ell_{t+k} = \bar{\ell}_{\mathrm{mo}(t+k)} + \chi \odot r_\phi\big(\mathrm{sg}(\hat{S}_{t+k})\big), \tag{7}$$

where sg stops gradients and $\bar{\ell}$ contains logit-transformed monthly means or log class frequencies. The gate χ, selected on the validation split, retains learned residuals for warm area, carbon-flux reversal, and carbon-sink area and keeps the seasonal baselines elsewhere; training supervises all six families with $\chi = \mathbf{1}$. The paired readout predicts, without a seasonal baseline, the fraction of each field above its training 90th percentile.

4.5 Basis-Coordinate Disturber and Counterfactual Rollout

Each supported action has $\delta_u = Ab_u$. The **disturber** changes only the last frame's coefficients, $S_t^b = I_b(S_t) = \big[z_{t-L+1}, \ldots, z_{t-1}, (a_t + b, h_t)\big]$, with $b = 0$ for the natural branch and $b = b_u$ for the edited branch; write $I_u = I_{b_u}$. Since $C_A(x_t + Ab_u) = a_t + b_u$ and $Q(x_t + Ab_u) = Qx_t$, editing is equivalent to perturbing the last observed field before encoding, with masks, calendars, and recorded events fixed. The compared models receive the same pulse in this form.

The two windows are advanced with the same F_θ and calendars, giving

$$\hat{r}_{t+k}(u) = D_x\big(\hat{z}^u_{t+k}\big) - D_x\big(\hat{z}_{t+k}\big), \qquad k = 1, \ldots, K. \tag{8}$$

The edit preserves earlier history, so the predictor must respond to the changed current frame. Ocean coefficient edits serve as sensitivity probes; responses are scored only in the paired environment.

4.6 Training Objectives

The JEPA objective matches a predicted residual representation to the shared encoder's representation of the corresponding true future x^{*b}_{t+k}:

$$\mathcal{L}_{\text{JEPA}} = \frac{1}{Kd} \sum_{k=1}^{K} \big\| \hat{h}^b_{t+k} - h^{*b}_{t+k} \big\|_2^2, \qquad h^{*b}_{t+k} = E_\theta\big(Q x^{*b}_{t+k}, m_{t+k}, c_{t+k}\big), \tag{9}$$

Target gradients remain active, following LeWorldModel (Maes et al., 2026). We combine this objective with SIGReg to discourage collapse and an area-weighted field loss to constrain observable predictions: $\mathcal{L}_D = \mathcal{L}_{\text{JEPA}} + 0.09\mathcal{L}_{\text{SIG}} + \mathcal{L}_{\text{field}}$ (reductions: Appendix C.2).

Ocean training first learns from fields with auxiliary event supervision and predicted-event conditioning. The second stage enables historical event fusion and the seasonal readout and disables predicted-event conditioning; starting fusion from a trained first-stage model lets models that differ only in the fused input continue one checkpoint (Section 5.2). The data loss trains the encoder, fusion, predictor, and field readout; the event loss directly trains only the event readout. The first stage and paired training project conflicting event gradients away from the data gradient (Appendix C.2).

After natural-trajectory pretraining, paired training supervises both branches and their signed field and event responses,

$$\mathcal{L}_{\text{resp}} = \frac{\mathbb{E}\big\| q(\hat{r}_x, \hat{r}_e) - q(r^*_x, r^*_e) \big\|^2}{V_{\text{train}}}, \tag{10}$$

where q fixes the field and event scales and V_{train} is training response energy (Appendix C.2). With probability 0.5, it also replaces the earlier history with another history from the same mechanism, retaining the current frame and its futures. The simulator's Markov dynamics preserve the target under this replacement, while the changed history discourages continuation of the old trajectory.

4.7 Capture–Propagation Evaluation of Responses

We compare the true response r^* with an encoded response r^E, obtained by encoding and decoding both true futures, and the rollout response $r^M = \hat{r}$. Both model responses use the same frozen field readout. For true-response energy $V = \mathbb{E}\|r^*\|^2 > 0$, the **capture error** $C = \mathbb{E}\|r^E - r^*\|^2/V$ measures error introduced by encoding and readout, and the **propagation error** $P = \mathbb{E}\|r^M - r^E\|^2/V$ the gap between rollout and encoded responses. We define the **response skill** (RS) and the **latent counterfactual skill** (LCS) as

$$\text{RS} = 1 - \frac{\mathbb{E}\|r^M - r^*\|^2}{V}, \qquad \text{LCS} = 1 - \big(\sqrt{C} + \sqrt{P}\big)^2. \tag{11}$$

The triangle inequality gives $\text{LCS} \le \text{RS}$, so positive LCS certifies response error below the zero-response baseline through both comparisons. The norm averages squared standardized field values over cells and channels; expectations average roots, actions, and leads within each mechanism, then mechanisms equally. Appendix C.5 gives the bound and evaluation checks.

Table 1: Main results on the validation splits. Ocean dataset: 35 origins, three seeds (nRMSE seed standard deviation ≤ 0.004; Table 10); nRMSE and ACC average over leads and channels, and P90 is the 90th percentile of per-origin nRMSE. Paired environment: four mechanisms, six seeds; field MSE averages both branches, and field forecasters were not trained on it. The action-mean table returns each action's mean training response and holds no state. **Bold**: best; underline: second best.

	Ocean dataset				Paired environment		
Method	nRMSE ↓	ACC ↑	P90 ↓	nRMSE by lead ↓	Field MSE ↓	RS ↑	LCS ↑
LeWM	0.173	0.185	0.196	0.166/0.174/0.179	0.0113	−0.16	−1.20
TC-LeWM	0.171	0.227	0.195	0.163/0.172/0.179	0.0071	−0.42	−1.80
EAWM-inspired	0.185	0.175	0.211	0.177/0.186/0.192	0.0111	−0.14	−1.22
DreamerV3 RSSM	0.160	0.253	0.178	0.160/0.160/0.161	0.0445	−5.40	−13.08
SP-JEPA (ours)	**0.131**	0.374	**0.146**	0.128/**0.132/0.134**	**0.0070**	0.575	**0.206**
FNO	0.132	0.481	0.151	**0.120**/0.134/0.141	–	–	–
ConvLSTM	0.137	0.440	0.149	0.127/0.139/0.144	–	–	–
PDE-Transformer	0.142	0.355	0.156	0.132/0.143/0.149	–	–	–
U-Net	0.147	0.386	0.161	0.142/0.148/0.152	–	–	–
NeuralOM	0.149	0.460	0.159	0.132/0.152/0.161	–	–	–
ClimaX	0.155	0.359	0.168	0.147/0.156/0.162	–	–	–
Persistence	0.155	**0.505**	0.174	0.125/0.161/0.175	–	–	–
Action-mean table	–	–	–	–	–	**0.64**	–

5 Experiments

We compare natural ocean prediction (Section 5.2), identify the designs behind its accuracy (Section 5.3), and test what the state represents and how it responds to a physical edit (Section 5.4).

5.1 Experimental Setup

The ocean dataset supplies 12-month histories of five 90×180 fields with three monthly targets, and the paired environment supplies six frames of two 16×16 fields, six future steps, and the solver's future after each of sixteen Gaussian pulses (Appendices A and B). All numbers come from the validation splits, which also select checkpoints; the paired validation split consists of four mechanisms unseen in training. On the ocean dataset, nRMSE is the root of the area-weighted mean squared error of standardized fields, and ACC is the pattern correlation of anomalies from the training climatology; on the paired environment, RS and LCS (Section 4.7) score predicted responses against the solver's. The compared latent world models are LeWM (Maes et al., 2026), TC-LeWM (Liu et al., 2026), an EAWM-inspired LeWM (Peng et al., 2026), and DreamerV3 RSSM (Hafner et al., 2025). The field forecasters FNO (Li et al., 2021), ConvLSTM (Shi et al., 2015), PDE-Transformer (Holzschuh et al., 2025), U-Net (Ronneberger et al., 2015), NeuralOM (Gao et al., 2025), and ClimaX (Nguyen et al., 2023) map the observed window directly to future fields, and seasonal-anomaly persistence is a reference without learning. On the paired environment, each latent world model trains with its own objective on the same trajectories as SP-JEPA and receives a pulse as δ_u added to its last observed field, the observation-space form of the disturber edit (Section 4.5). Appendix D.1 details budgets, metrics, and tests.

5.2 Natural Prediction on the Ocean Dataset

SP-JEPA reaches an nRMSE of 0.131, 18% below the best latent world model (DreamerV3 RSSM, 0.160) and **level with the best field forecaster** (FNO, 0.132; Table 1). With batches of eight, it sees more training windows than the compared methods, but a 5,000-update model trained on fewer windows than any of them still reaches 0.133 (Appendix D.1). FNO, persistence, and ConvLSTM lead at the first month, but **their errors grow faster**: from month one to three, the nRMSE rises from 0.128 to 0.134 for SP-JEPA, from 0.120 to 0.141 for FNO, and from 0.125 to 0.175 for persistence.

The ACC ranks the methods differently: SP-JEPA leads the latent world models (0.374) but trails persistence (0.505) and four of the six field forecasters, led by FNO (0.481). The two metrics weigh

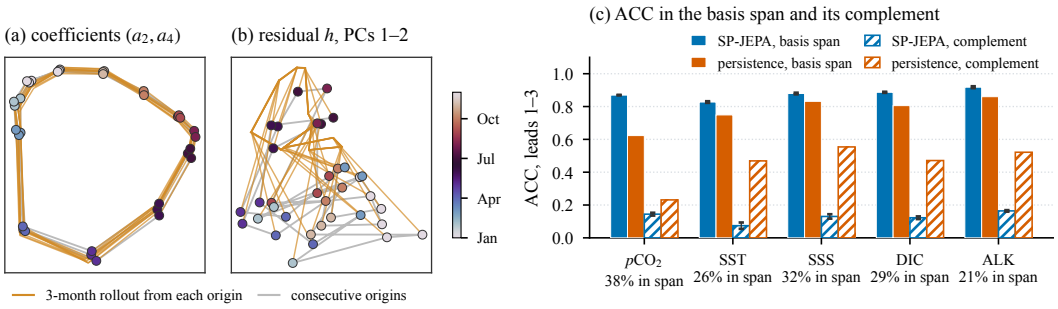

Figure 3: Where the state holds its skill (ocean validation origins). (a) Coefficients a_2 and a_4 of each origin, colored by month, with its three-month rollout; (b) the same for the two leading principal components of h. (c) ACC within the basis span (solid) and its complement (hatched); labels give the span's share of anomaly variance about each map's mean, and error bars the seed standard deviation.

the span of the fixed basis differently. The span captures most of the spatially uniform part of the anomalies, above all the excess pCO_2 of the validation years over the training climatology, and holds 89% of the standardized anomaly energy on which the nRMSE is computed; once the spatial mean of each map is removed, as in the ACC, it holds 21–38% of the anomaly variance (Figure 3c). Within the span, SP-JEPA is more skillful than persistence in all five channels (ACC 0.83–0.92 against 0.62–0.86); in the complement, its ACC falls to 0.07–0.16, against 0.23–0.55. **The rollout thus forecasts what the basis coefficients represent.** This part dominates the nRMSE, where SP-JEPA has the lowest pCO_2 error of all methods, but holds a minority of the variance that the ACC measures (per-channel results in Appendix D.2).

Event fusion passes the recent record of each event family into the state. Three models that continue the same first-stage checkpoint differ only in the fused input (Table 2, ‡ rows): observed events lower the event error by 8.8% against no fusion and by 10.7% against event histories from training years with the same calendar months (one-sided 95% bounds: 1.8% and 3.5%), while the field error falls by 0.6%. Because the second control keeps the fusion pathway and its input statistics, **its gain comes from the observed values**. Most of the gain falls on the warm area (21%; Appendix D.3), whose mean rises from 0.12 of the ocean area in the training months to 0.26 in the validation months. The fused record carries this level, whereas the event readout otherwise infers it from the state around a seasonal anchor fitted to the training years.

5.3 Ablations

The largest effects in Table 2 arise between modules. Without the basis coefficients, the predictor must advance h alone, and the field MSE rises by 51% at development scale (5,000 updates), 27% at full scale, and 72% on the paired environment; Section 5.4 shows that the coefficients carry the seasonal cycle. Without attention across frames, the paired error rises by 62%, because only the history reveals the advection and exchange rates of a mechanism. Without history replacement, the natural paired error falls by 33%, but the rollout stops following the current frame: after this frame is swapped with that of another initial state, the sixth-step prediction reaches a skill of 0.05 toward the new future, against 0.87 with replacement (Appendix D.8). Since an edit changes the current frame and not the history, such a model lets an edit fade within a few steps. Without the JEPA term, the ocean error rises by 4.9% at development scale in all seven seeds; at full scale, the validation-best checkpoints differ by only 0.7% [0.0%, 1.6%], because two of three runs without the term reach their best at update 1,000 and degrade afterwards (Appendix D.6). The term also keeps the encoded state decodable: after the first ocean stage, the encoded current frame decodes with an nRMSE of 0.125 with the term and 0.434 without it, and on the paired environment, dropping the term lowers the natural error by 30% but raises the decoding error of the current frame eightfold.

Within modules, screening arms indicate that the coefficient update matters more than the sequence network: a randomly initialized increment head and absolute outputs raise the error by 2.8% and a further 4.1%, whereas a GRU changes it by 0.7%. Teacher forcing leaves the ocean months unchanged, but on the paired environment its error ratio grows from 0.84 at the first step to 1.35 at

Table 2: Ablations by where the change acts. Entries are ratios of validation field MSE (and event error) to the full model on the same schedule unless a row names another reference; the paired column uses natural-trajectory models. Pre-specified ocean arms (development scale) use seven seeds (†: Holm-significant), others three to five; ‡: continued from the full-scale model; –: not applicable or not run (all arms: Appendix D.4).

Change	Ocean field [95% CI]	Seeds worse	Ocean event	Paired, natural
SP-JEPA (full model)	1	–	1	1
Within a module				
Trainable basis, same initialization	0.989 [0.973, 1.005]	1/7	–	0.44
Teacher forcing instead of open loop	1.002 [0.992, 1.011]	5/7	–	1.26
Increment head, random initialization	1.028 [1.009, 1.059]	3/3	–	–
Absolute coefficients (vs. row above)	1.041 [1.018, 1.068]	3/3	–	–
GRU instead of causal transformer	1.007 [1.000, 1.015]	3/3	–	–
Between modules				
Remove basis coefficients	1.510 [1.370, 1.669]†	7/7	–	1.72
Remove JEPA term	1.049 [1.036, 1.063]†	7/7	–	0.70
No attention across frames	0.999 [0.991, 1.008]	1/3	–	1.62
No history replacement	–	–	–	0.67
No historical event fusion‡	1.006 [1.001, 1.013]	3/3	1.097	–
Properties				
Residual width $d = 64$	1.020 [1.004, 1.037]	3/3	–	–
Residual width $d = 256$	1.016 [1.004, 1.031]	3/3	–	–
Other-year event history‡	1.007 [1.001, 1.013]	3/3	1.120	–
Fusion from update 0 (vs. 3,000)	1.007 [0.998, 1.015]	4/5	0.81	–

the sixth (Appendix D.5), so open-loop training pays off with the horizon. The fixed paired basis contains every action template, which makes each action's edit exact by Eq. (4); a trainable basis lowers the paired error by 56% but gives up this guarantee, since training moves its columns. Among properties, widths of 64 and 256 raise the error by about 2%, and fusing events from the first update lowers event error by 19% at a field ratio of 1.007, so the event gain does not require two stages.

5.4 State Interpretation and Counterfactual Response

The basis coefficients, which carry the largest ablation effect, keep the seasonal cycle explicit (Figure 3a,b). Because the basis is computed from uncentered monthly fields, an annual harmonic explains 98% and 86% of the variance of a_2 and a_4 across the validation origins, and the learned predictor advances each rollout along the orbit that the two coefficients trace. The same harmonic explains only 41% and 51% of the two leading principal components of h. Edits of the coefficients change the ocean rollout almost linearly and reproducibly across seeds (Appendix D.7), but the ocean products contain no paired outcomes against which such responses could be scored. We therefore score responses on the paired environment.

On the four validation mechanisms, SP-JEPA reaches an RS of 0.575 and an LCS of 0.206 (interval above zero for five of six seeds), whereas every compared latent world model scores below zero (Table 1). In Figure 4a, SP-JEPA reproduces how the pulse decays in its own field and reappears in the second, whereas LeWM predicts a weak, nearly uniform change. Over all queries, SP-JEPA's responses correlate with the truth at 0.76 with 84% of the true amplitude, against -0.02 to 0.19 for the compared models (Figure 4b).

Decomposing the true response shows what this score measures (Figure 4d). Of the response energy, 72% is the mean response of each action, 28% the deviation of each mechanism from it, and 0.2% that of each initial state. A stateless table of each action's mean training response therefore reaches an RS of 0.64 (Table 1). SP-JEPA's rollout reproduces the action-mean component (correlation 0.92) but, like every model, not the mechanism component: mechanism 4 advects the true response by about 3.5 grid cells in six steps, whereas SP-JEPA keeps it at the pulse (Figure 4a). The environment thus shows that **SP-JEPA propagates the average consequence of an action through its own state**; mechanism-dependent responses remain open, and state-dependent ones cannot be tested here.

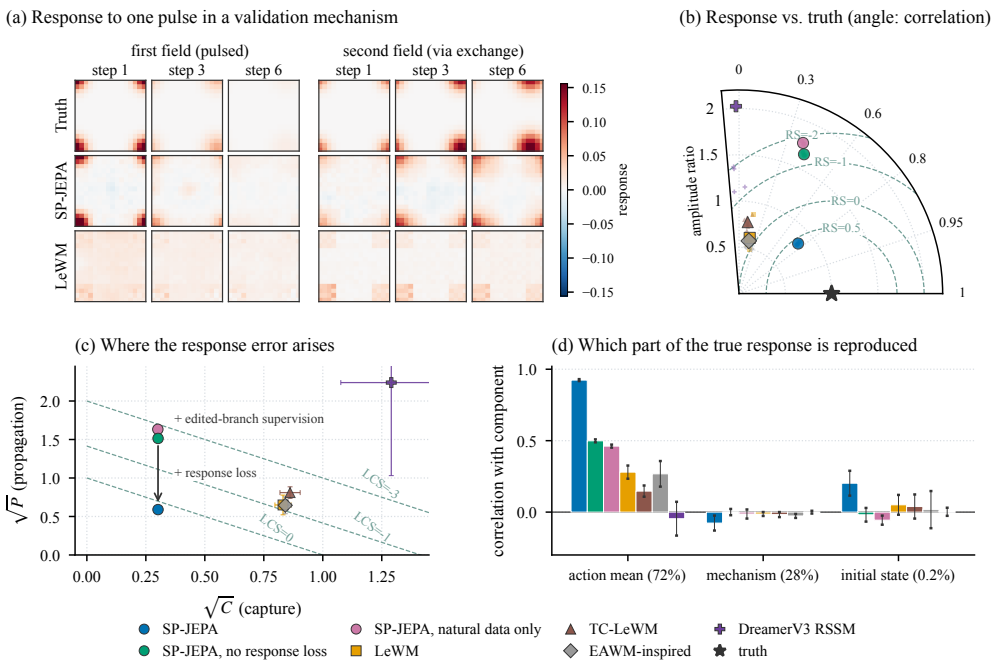

Figure 4: Responses on the four validation mechanisms (six seeds). (a) True and predicted responses to one pulse in mechanism 4, in the pulsed field and the field reached through exchange (all mechanisms: Appendix D.8). (b) Taylor diagram: angle is correlation with the true response, radius the amplitude ratio, dashed arcs RS; small markers are seeds. (c) Capture and propagation errors with LCS isolines; arrows add edited-branch supervision, then the response loss. (d) Correlation with three components of the true response (energy shares in parentheses). Bars: seed standard deviation.

The compared models miss even the action-mean component (correlation at most 0.28), and the capture and propagation errors show why (Figure 4c). Because $(\sqrt{C} + \sqrt{P})^2 \geq C$, LCS cannot exceed $1 - C$. The compared models lose most of the response when encoding the perturbed futures ($\sqrt{C} \geq 0.83$), which caps their LCS at 0.31. The paired basis of SP-JEPA contains every action template (Section 4.1), so its 32 templates hold 91% of the response energy, and $\sqrt{C}$ stays at 0.30 under every supervision, which caps its LCS at 0.91. **Propagation then depends on the response loss.** Trained on both branches without it, like the compared models, SP-JEPA correlates with the true response at 0.42 but overshoots its amplitude 1.7-fold ($\sqrt{P} = 1.51$), since the response, about 3×10^{-4} in mean energy per cell, contributes little to a two-branch field MSE of 0.0055. Supervising the response lowers $\sqrt{P}$ to 0.59 and raises the two-branch field MSE to 0.0070 (Appendix D.8).

6 Limitations and Conclusion

SP-JEPA combines a compact predictive state with an exact interface for spatial perturbations. Fixed basis coordinates account for much of its natural forecasting skill, historical event observations improve aggregate event prediction, and paired response supervision enables the state to propagate the average effect of a pulse. The main remaining difficulty is predicting how that response changes with the dynamical mechanism. Because the tested simulator varies little within a mechanism and the ocean products provide no paired intervention outcomes, these results establish short-horizon forecasting and finite-pulse response prediction; broader state-dependent responses and sequential planning require environments that expose those distinctions. For ocean forecasting, the principal accuracy gap lies in anomalies outside the fixed basis span, and fused historical frames occupy a different region of the residual space than the frames predicted by F_θ (Appendix D.6), so the event record informs the initial state without yet becoming part of the predicted state. A richer residual representation with event information in its prediction targets is a direct way to address both.

AI USE STATEMENT

In this work, we used generative AI tools for helping construct the conceptual framework, by suggesting mathematical methods that reduce discrepancies in some gradient computations; for designing and giving feedback on experiments, namely the visualization of results and the design of the inter-module ablations; for implementing methods, where AI supplemented boundary-case code and missing conditions in an author-led implementation; and for assisting with translation. We have not used generative AI tools for generating synthetic datasets, formulating mathematical claims, providing critical ingredients for proofs, assisting in writing proofs, proposing or refining hypotheses, cleaning or reformatting datasets, or interpreting results, and supporting qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for writing and editing code, summarizing and analyzing literature, identifying relevant literature, and editing the paper for readability, particularly the transitions between paragraphs. We have reviewed all AI-assisted work: we examined each suggested method and ablation design before adopting it, ran AI-assisted code and inspected its outputs, and checked every AI-identified reference against its original source. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendices A and B specify data sources, preprocessing, chronological splits, event definitions, and the paired simulator. Appendix C provides the architecture, objectives, training schedules, hyperparameters, and rollout algorithm; Appendix D reports the complete ablations and supporting diagnostics. The anonymous code package provides training and evaluation scripts and configuration files: `https://anonymous.4open.science/r/SP-JEPA-5DB1`.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. doi: 10.48550/arXiv.2506.09985.

Guillaume Astruc, Nicolas Gonthier, Clément Mallet, and Loic Landrieu. AnySat: One Earth observation model for many resolutions, scales, and modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19530–19540, 2025. doi: 10.1109/CVPR52734.2025.01819.

Marco Bagatella, Matteo Pirotta, Ahmed Touati, Alessandro Lazaric, and Andrea Tirinzoni. TD-JEPA: Latent-predictive representations for zero-shot reinforcement learning. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=SzXDuBN8M1`.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=QaCCuDfBk2`.

Nicholas R. Bates, Anthony H. Knap, and Anthony F. Michaels. Contribution of hurricanes to local and global estimates of air–sea exchange of CO_2. *Nature*, 395(6697):58–61, 1998. doi: 10.1038/25703.

Ce Bian, Zijie Zhao, Neil J. Holbrook, Peter G. Strutton, and Lixin Wu. Marine heatwaves shift ocean net primary productivity from the tropics toward the poles. *Nature Communications*, 17(1): 4624, 2026. doi: 10.1038/s41467-026-71238-w.

Mariana B. Bif, Colleen T. E. Kellogg, Yibin Huang, Julia Anstett, Sachia Traving, M. Angelica Peña, Steven J. Hallam, and Kenneth S. Johnson. Marine heatwaves modulate food webs and carbon transport processes. *Nature Communications*, 16(1):8535, 2025. doi: 10.1038/s41467-025-63605-w.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the Earth system. *Nature*, 641(8065):1180–1187, 2025. doi: 10.1038/s41586-025-09005-y.

Copernicus Climate Change Service. ORAS5 global ocean reanalysis monthly data from 1958 to present. Copernicus Climate Data Store, 2021. URL https://doi.org/10.24381/cds.67e8eeb7. Surface salinity is read from the ORAS5 reanalysis produced by ECMWF; accessed 19 September 2026.

Copernicus Marine Service. Global ocean surface carbon product MULTIOBS_GLO_BGC_CARBON_SURFACE_MYNRT_015_008. Copernicus Marine Service, 2024. URL https://doi.org/10.48670/moi-00047. Monthly Level-4 reconstructed surface ocean carbon fields; accessed 19 September 2026.

Damien Couespel, Jerry Tjiputra, Klaus Johannsen, Pradeebane Vaittinada Ayar, and Bjørnar Jensen. Machine learning reveals regime shifts in future ocean carbon dioxide fluxes inter-annual variability. *Communications Earth & Environment*, 5(1):99, 2024. doi: 10.1038/s43247-024-01257-2.

Nathaniel Cresswell-Clay, Bowen Liu, Dale R. Durran, Zihui Liu, Zachary I. Espinosa, Raul A. Moreno, and Matthias Karlbauer. A deep learning Earth system model for efficient simulation of the observed climate. *AGU Advances*, 6(4):e2025AV001706, 2025. doi: 10.1029/2025AV001706.

Huayu Deng, Jinghui Zhong, Xiangming Zhu, Yunbo Wang, and Xiaokang Yang. LASER: Learning active sensing for continuum field reconstruction. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2604.19355.

Amanda R. Fay and Galen A. McKinley. Global trends in surface ocean pCO_2 from in situ data. *Global Biogeochemical Cycles*, 27(2):541–557, 2013. doi: 10.1002/gbc.20051.

Richard A. Feely, Jacqueline Boutin, Catherine E. Cosca, Yves Dandonneau, Jacqueline Etcheto, Hisayuki Y. Inoue, Masao Ishii, Corinne Le Quéré, Denis J. Mackey, Michael McPhaden, Nicolas Metzl, Alain Poisson, and Rik Wanninkhof. Seasonal and interannual variability of CO_2 in the equatorial Pacific. *Deep Sea Research Part II: Topical Studies in Oceanography*, 49(13-14): 2443–2469, 2002. doi: 10.1016/S0967-0645(02)00044-9.

Richard A. Feely, Christopher L. Sabine, J. Martin Hernandez-Ayon, Debby Ianson, and Burke Hales. Evidence for upwelling of corrosive "acidified" water onto the continental shelf. *Science*, 320(5882):1490–1492, 2008. doi: 10.1126/science.1155676.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C. Lisaius, Markus Immitzer, Toby Jackson, James Ball, David A. Coomes, Anil Madhavapeddy, Andrew Blake, and Srinivasan Keshav. TESSERA: Temporal embeddings of surface spectra for Earth representation and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34818–34831, 2026.

Jennifer Gahtan, Kenneth R. Knapp, Carl J. Schreck, III, Howard J. Diamond, James P. Kossin, and Michael C. Kruk. International best track archive for climate stewardship (IBTrACS) project, version 4r01. NOAA National Centers for Environmental Information, 2024. URL https://doi.org/10.25921/82ty-9e16. Global tropical-cyclone best-track dataset; accessed 19 September 2026.

Yuan Gao, Hao Wu, Fan Xu, Yanfei Xiang, Ruijian Gou, Ruiqi Shu, Qingsong Wen, Xian Wu, Kun Wang, and Xiaomeng Huang. NeuralOM: Neural ocean model for subseasonal-to-seasonal simulation. *arXiv preprint arXiv:2505.21020*, 2025. doi: 10.48550/arXiv.2505.21020.

Nicolas Gruber, Peter Landschützer, and Nicole S. Lovenduski. The variable Southern Ocean carbon sink. *Annual Review of Marine Science*, 11(1):159–186, 2019. doi: 10.1146/annurev-marine-121916-063407.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Alexander Hayward. Marine heat waves and cold spells – multiple analysis / definitions (MHW-MAD): A multi-definition global marine heatwave database from satellite sea surface temperature data. Zenodo, 2026. URL https://doi.org/10.5281/zenodo.21835325. Multi-definition global marine-heatwave catalog from satellite sea surface temperature; accessed 19 September 2026.

Benjamin Holzschuh, Qiang Liu, Georg Kohl, and Nils Thuerey. PDE-transformer: Efficient and versatile transformers for physics simulations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23562–23602, 2025.

Zhentao Hu, Moritz Mathis, Hongmei Li, Tatiana Ilyina, Yoana G. Voynova, Vlad A. Macovei, Feng Zhou, Qicheng Meng, Corinna Schrum, and Wenyan Zhang. Polar and subpolar shelf seas dominate enhanced coastal CO_2 uptake during marine heatwaves. *Nature Communications*, 17 (1):9346, 2026. doi: 10.1038/s41467-026-77165-0.

Michael G. Jacox, Christopher A. Edwards, Elliott L. Hazen, and Steven J. Bograd. Coastal upwelling revisited: Ekman, Bakun, and improved upwelling indices for the U.S. west coast. *Journal of Geophysical Research: Oceans*, 123(10):7332–7350, 2018. doi: 10.1029/2018JC014187.

Armand Kassaï Koupaï, Lise Le Boudec, Louis Serrano, and Patrick Gallinari. ENMA: Token-wise autoregression for continuous neural PDE operators. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=3CYXSMFv55.

Kenneth R. Knapp, Michael C. Kruk, David H. Levinson, Howard J. Diamond, and Charles J. Neumann. The international best track archive for climate stewardship (IBTrACS): Unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91(3):363–376, 2010. doi: 10.1175/2009BAMS2755.1.

Peter Landschützer, Nicolas Gruber, Dorothee C. E. Bakker, Ute Schuster, Shin-ichiro Nakaoka, Matthew R. Payne, Tobias P. Sasse, and Jiye Zeng. A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink. *Biogeosciences*, 10(11): 7793–7815, 2013. doi: 10.5194/bg-10-7793-2013.

Peter Landschützer, Nicolas Gruber, Dorothee C. E. Bakker, and Ute Schuster. Recent variability of the global ocean carbon sink. *Global Biogeochemical Cycles*, 28(9):927–949, 2014. doi: 10.1002/2014GB004853.

Peter Landschützer, Nicolas Gruber, and Dorothee C. E. Bakker. Decadal variations and trends of the global ocean carbon sink. *Global Biogeochemical Cycles*, 30(10):1396–1417, 2016. doi: 10.1002/2015GB005359.

Corinne Le Quéré, Robbie M. Andrew, Pierre Friedlingstein, Stephen Sitch, Judith Hauck, Julia Pongratz, Penelope A. Pickers, Jan Ivar Korsbakken, Glen P. Peters, Josep G. Canadell, Almut Arneth, Vivek K. Arora, Leticia Barbero, Ana Bastos, Laurent Bopp, Frédéric Chevallier, Louise P. Chini, Philippe Ciais, Scott C. Doney, Thanos Gkritzalis, Daniel S. Goll, Ian Harris, Vanessa Haverd, Forrest M. Hoffman, Mario Hoppema, Richard A. Houghton, George Hurtt, Tatiana Ilyina, Atul K. Jain, Truls Johannessen, Chris D. Jones, Etsushi Kato, Ralph F. Keeling, Kees Klein Goldewijk, Peter Landschützer, Nathalie Lefèvre, Sebastian Lienert, Zhu Liu, Danica Lombardozzi, Nicolas Metzl, David R. Munro, Julia E. M. S. Nabel, Shin-ichiro Nakaoka, Craig Neill, Are Olsen, Tsuneo Ono, Prabir Patra, Anna Peregon, Wouter Peters,

Philippe Peylin, Benjamin Pfeil, Denis Pierrot, Benjamin Poulter, Gregor Rehder, Laure Resplandy, Eddy Robertson, Matthias Rocher, Christian Rödenbeck, Ute Schuster, Jörg Schwinger, Roland Séférian, Ingunn Skjelvan, Tobias Steinhoff, Adrienne Sutton, Pieter P. Tans, Hanqin Tian, Bronte Tilbrook, Francesco N. Tubiello, Ingrid T. van der Laan-Luijkx, Guido R. van der Werf, Nicolas Viovy, Anthony P. Walker, Andrew J. Wiltshire, Rebecca Wright, Sönke Zaehle, and Bo Zheng. Global carbon budget 2018. *Earth System Science Data*, 10(4):2141–2194, 2018. doi: 10.5194/essd-10-2141-2018.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview preprint, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf.

Zongyi Li, Nikola B. Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Enhui Liao, Laure Resplandy, Junjie Liu, and Kevin W. Bowman. Amplification of the ocean carbon sink during El Niños: Role of poleward Ekman transport and influence on atmospheric CO_2. *Global Biogeochemical Cycles*, 34(9):e2020GB006574, 2020. doi: 10.1029/2020GB006574.

Chang Liu, Fei Suo, Yanzhou Jin, Zeyu Ping, Yusuke Iwasawa, Yutaka Matsuo, and Yaonan Zhu. Temporally centered SIGReg improves LeWorldModel representations for robot policy learning. *arXiv preprint arXiv:2607.26924*, 2026. doi: 10.48550/arXiv.2607.26924.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026. doi: 10.48550/arXiv.2603.19312.

Alexandre Mignot, Karina von Schuckmann, Peter Landschützer, Florent Gasparin, Simon van Gennip, Coralie Perruche, Julien Lamouroux, and Tristan Amm. Decrease in air–sea CO_2 fluxes caused by persistent marine heatwaves. *Nature Communications*, 13(1):4300, 2022. doi: 10.1038/s41467-022-31983-0.

Jens Daniel Müller, Nicolas Gruber, Aline Schneuwly, Dorothee C. E. Bakker, Marion Gehlen, Luke Gregor, Judith Hauck, Peter Landschützer, and Galen A. McKinley. Unexpected decline in the ocean carbon sink under record-high sea surface temperatures in 2023. *Nature Climate Change*, 15(9):978–985, 2025. doi: 10.1038/s41558-025-02380-4.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25904–25938, 2023.

Weizhi Nie, Weichao Liu, Honglin Guo, and Yuting Su. Phys-JEPA: Physics-informed latent world models for multivariate time-series forecasting. *arXiv preprint arXiv:2606.16076*, 2026. doi: 10.48550/arXiv.2606.16076.

NOAA Climate Prediction Center. Cold and warm episodes by season (ERSSTv5): Oceanic Niño index. NOAA CPC historical ONI record, 2024. URL https://www.cpc.ncep.noaa.gov/products/analysis_monitoring/enso/oni/v5/. Centered three-month ERSSTv5 Niño 3.4 index; accessed 19 September 2026.

NOAA National Centers for Environmental Information. Optimum interpolation sea surface temperature (OISST). NOAA product page, 2021. URL https://www.ncei.noaa.gov/products/optimum-interpolation-sst. Monthly SST files from the OISST product family; accessed 19 September 2026.

NOAA Southwest Fisheries Science Center and Michael G. Jacox. Coastal upwelling transport index (CUTI) daily data. NOAA ERDDAP, 2018. URL https://oceanview.pfeg.noaa.gov/erddap/info/erdCUTIdaily/index.html. Seventeen one-degree latitude bands from 31°N to 47°N; accessed 19 September 2026.

Zhao-Han Peng, Shaohui Li, Zhi Li, Shulan Ruan, Yu Liu, and You He. From observations to events: Event-aware world model for reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=OWkkFaq1IZ.

Rajiv Ranjan, Udaiveer Singh, Shashank Tamaskar, and Dharmendra Saraswat. SPEAR NeXT causal latent forecasting across multiple horizons for spectral temporal Earth representation learning. *arXiv preprint arXiv:2609.16871*, 2026. doi: 10.48550/arXiv.2609.16871.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

Salva Rühling Cachay, Miika Aittala, Karsten Kreis, Noah Brenowitz, Arash Vahdat, Morteza Mardani, and Rose Yu. Elucidated rolling diffusion models for probabilistic forecasting of complex dynamics. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=NnRQOPzL9P.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, pp. 802–810, 2015.

Eduardo Soares, Emilio Vital Brazil, Victor Shirasuna, Breno W. S. R. de Carvalho, and Cristiano Malossi. Towards a foundation model for partial differential equations across physics domains. *arXiv preprint arXiv:2511.21861*, 2025. doi: 10.48550/arXiv.2511.21861.

Mary Margaret V. Stoll, Curtis A. Deutsch, Hana Jurikova, James W. B. Rae, Hartmut Frenzel, Anne M. Gothmann, Simone R. Alin, and Alexander C. Gagnon. A century of change in the California Current: Upwelling system amplifies acidification. *Nature Communications*, 16(1): 9661, 2025. doi: 10.1038/s41467-025-63207-6.

Taro Takahashi, Stewart C. Sutherland, Colm Sweeney, Alain Poisson, Nicolas Metzl, Bronte Tilbrook, Nicolas Bates, Rik Wanninkhof, Richard A. Feely, Christopher Sabine, Jon Olafsson, and Yukihiro Nojiri. Global sea–air CO_2 flux based on climatological surface ocean pCO_2, and seasonal biological and temperature effects. *Deep Sea Research Part II: Topical Studies in Oceanography*, 49(9-10):1601–1622, 2002. doi: 10.1016/S0967-0645(02)00003-6.

Donna Vakalis. The intervention gap in latent world models. *arXiv preprint arXiv:2608.29998*, 2026. doi: 10.48550/arXiv.2608.29998.

Sifan Wang, Zehao Dou, Siming Shan, Tong-Rui Liu, and Lu Lu. FunDiff: Diffusion models over function spaces for physics-informed generative modeling. *Nature Communications*, 17(1):5749, 2026. doi: 10.1038/s41467-026-72292-0.

Hexin Ye, Zhanhong Ma, Jianfang Fei, Kevin E. Trenberth, Toste Tanhua, and Lijing Cheng. Reduction of tropical cyclone-induced ocean carbon outgassing since 1993. *Nature Geoscience*, 19 (6):630–638, 2026. doi: 10.1038/s41561-026-01985-4.

Varvara E. Zemskova, Tai-Long He, Zirui Wan, and Nicolas Grisouard. A deep-learning estimate of the decadal trends in the Southern Ocean carbon storage. *Nature Communications*, 13(1):4056, 2022. doi: 10.1038/s41467-022-31560-5.

A GLOBAL OCEAN CARBON–PHYSICS DATASET

The Global Ocean Carbon–Physics dataset contains one monthly ocean record, aligned fields, validity masks, and six source-derived event families. This section gives the construction used in the experiments.

A.1 STATE, SOURCES, AND SPATIAL SUPPORT

For month t, the observed field is

$$o_t = (X_t, M_t), \qquad X_t \in \mathbb{R}^{5\times 90\times 180}, \tag{12}$$

where M_t is the five-channel common-valid mask. The channels are ordered as surface partial pressure of CO_2 (pCO_2), sea-surface temperature (SST), sea-surface salinity (SSS), dissolved inorganic carbon (DIC), and total alkalinity (ALK). The release covers January 1993–December 2023 (372 months) on a regular 2° grid with latitude centers $-89^\circ, \ldots, 89^\circ$ and longitude centers $-179^\circ, \ldots, 179^\circ$, so the stored state has shape `[372,5,90,180]`. Table 3 lists the source product of each channel.

Table 3: Field channels and source products. L4 denotes a Level-4 reconstructed product. The common mask and 2° aggregation are applied after source alignment.

Channel	Unit	Product used	Status in this protocol
pCO_2	μatm	CMEMS surface carbon fields	L4 reconstructed
SST	$^\circ$C	NOAA OISST monthly SST	observational analysis
SSS	PSU	ECMWF ORAS5, surface level	reanalysis
DIC	μmol kg^{-1}	CMEMS surface carbon fields	L4 reconstructed
ALK	μmol kg^{-1}	CMEMS surface carbon fields	L4 reconstructed

The carbon fields are from the Copernicus Marine product `MULTIOBS_GLO_BGC_CARBON_SURFACE_MYNRT_015_008` (DOI 10.48670/moi-00047), a monthly Level-4 reconstruction from in-situ observations and auxiliary variables (Copernicus Marine Service, 2024). The SST source is the NOAA OISST monthly product (NOAA National Centers for Environmental Information, 2021), mapped by exact calendar-month and coordinate indices. SSS is the surface value of the ECMWF ORAS5 global ocean reanalysis (DOI 10.24381/cds.67e8eeb7) (Copernicus Climate Change Service, 2021). Because the carbon channels are reconstructions, the forecasting task predicts reconstructed fields and is retrospective rather than operational.

The products are first aligned on a common 713×1440 working grid of approximately 0.25° and then aggregated to the target grid with exact spherical cell areas. Let $q_{t,i} = \prod_{c'=1}^{5} m_{t,c',i}$ denote joint support at native cell i. For target cell j and channel c, the pooled value is

$$\bar{X}_{t,c,j} = \frac{\sum_{i\in j} a_i q_{t,i} X_{t,c,i}}{\sum_{i\in j} a_i q_{t,i}}, \qquad a_i = R^2(\sin\phi_{i,+} - \sin\phi_{i,-})(\lambda_{i,+} - \lambda_{i,-}). \tag{13}$$

A target cell is valid only when $\sum_{i\in j} a_i q_{t,i}/A_j \geq 50\%$, where A_j is the full spherical area of the target cell. Area weighting corrects for the unequal area of equal-degree cells, and joint support ensures that all five channels describe the same water. Invalid entries are stored as zeros together with $M_t = 0$ and are excluded from aggregation, losses, and metrics.

The common support contains 9,182 cells, from approximately 75°S to 79°N, and the field mask is constant across all 372 months (Figure 5).

A.2 MONTHLY TASK AND FOLD CONSTRUCTION

A forecasting sample at origin t is

$$(X_{t-11:t}, M_{t-11:t}, c_{t-11:t}, e_{t-11:t}, \kappa_{t-11:t}) \longrightarrow (X_{t+1:t+3}, M_{t+1:t+3}, e_{t+1:t+3}), \tag{14}$$

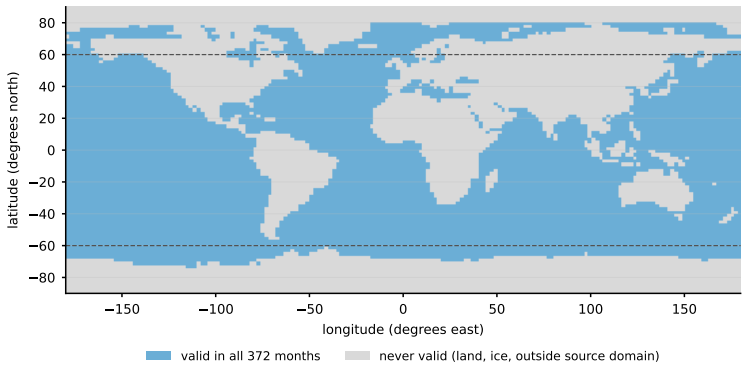

Figure 5: Common support of the five fields. Colored cells meet the 50%-area rule in all 372 months; dashed lines mark $\pm 60°$ latitude.

Table 4: Global origins per chronological fold (train/validation/test). Reported experiments use Main rolling. Earlier folds and purged variants are included in the data release.

Fold	Rolling	Purged	Field-month intervals (train / validation / test)
Main	283/35/36	283/23/24	1993-01–2017-09 / 2017-10–2020-10 / 2020-11–2023-12
F1	178/34/34	178/22/22	1993-01–2008-12 / 2009-01–2011-12 / 2012-01–2014-12
F2	238/34/34	238/22/22	1993-01–2013-12 / 2014-01–2016-12 / 2017-01–2019-12

where c encodes the calendar, e contains 13 event values, and κ contains six known-flags. The model receives future calendars but uses future fields and events only as targets. Temporal availability follows the month represented by each product, not its historical publication date.

Each fold computes per-channel normalization, climatologies, and event thresholds from its training months only. The stored standardization uses the unweighted mean and population standard deviation of valid coarse-grid values,

$$\tilde{X}_{t,c,j} = (X_{t,c,j} - \mu_{c,\mathrm{train}})/\sigma_{c,\mathrm{train}}, \tag{15}$$

while losses and metrics weight cells by spherical area.

Each origin is one complete global sequence. The Main rolling split assigns origins by their three target months and allows histories to extend into the preceding interval. Target months do not cross split boundaries. Table 4 also lists the available earlier folds and purged variants, which require the history to remain within each interval.

Adjacent origins share months, so the ocean uncertainty estimates resample contiguous blocks of origins.

A.3 Event Families Used by the Model

Five event families use lower and upper bounds, and Niño 3.4 uses three class indicators, giving 13 values (Table 5). Bounds describe source identification: the lower bound includes confirmed contributions, and the upper bound also includes unresolved contributions. They coincide under complete support and are not forecast confidence intervals.

Construction details. The SST, carbon, and Niño supports admit cells that are valid in at least 95% of the training months, and area fractions use the full spherical area of the admitted support as denominator. Missing values inside the support widen the interval, whereas cells outside it never contribute. With $w_j = A_j S_j$ for admitted cells $S_j = 1$ and $q_{t,j}$ indicating that cell j is observed, a binary indicator $a_{t,j}$ is summarized by

$$L_t = \frac{\sum_j w_j q_{t,j} a_{t,j}}{\sum_j w_j}, \qquad U_t = L_t + \frac{\sum_j w_j (1 - q_{t,j})}{\sum_j w_j}. \tag{16}$$

Table 5: The six event families. Thresholds, climatologies, and admitted supports are computed from the training months of each fold. IBTrACS: International Best Track Archive for Climate Stewardship; CUTI: Coastal Upwelling Transport Index; UTC: Coordinated Universal Time.

Family	Source	Monthly definition	Output
Tropical-cyclone load	IBTrACS v04r01	$n/(1+n)$, where n is the monthly mean of distinct storm IDs per UTC day; the lower bound counts tropical-storm records with USA wind $\geq$ 34 kt, and the upper bound adds records whose nature or wind is unresolved	bounds in $[0, 1)$
SST warm-area fraction	NOAA OISST (state channel)	Area fraction of admitted cells whose monthly SST exceeds the cell's training-period same-month 90th percentile	bounds
Coastal-upwelling activity	NOAA CUTI	Monthly mean number of active 1° latitude bands between 31 and 47°N, divided by 17	bounds
Carbon-flux reversal	CMEMS $fgco_2$	Area fraction of admitted cells whose air–sea flux changes sign between consecutive months	bounds
Carbon-sink area	CMEMS $fgco_2$	Area fraction of admitted cells with ocean uptake ($fgco_2 > 0$)	bounds
Niño 3.4 state	NOAA OISST (state channel)	Mean SST anomaly over 5°S–5°N, 170–120°W in the current and two preceding months; warm above $+0.5^\circ$C, cold below -0.5°C	3 classes

For CUTI, a known monthly active-band count n_t gives equal bounds $L_t = U_t = n_t/17$; unknown months retain $[0, 1]$ and a false known-flag. Positive $fgco_2$ denotes ocean uptake, and reversal is unknown in the first month. The Niño 3.4 proxy averages the current and two preceding monthly anomalies and is reported separately from official ENSO episodes.

Role in the model. Six known-flags mark families with identified values and gate event losses and metrics. The fusion rule in Section 4.2 excludes the cyclone upper bound because it uses retrospective confirmation.

Official catalogs that are not used. The model uses monthly warm-area and trailing Niño 3.4 indices. The available daily marine-heatwave catalog uses a 1993–2022 baseline, and official ENSO episode labels use subsequent months for confirmation. The selected proxies instead use past monthly values and training-period statistics.

A.4 Source Event Catalogs

Cyclone observations use NOAA IBTrACS v04r01 (DOI 10.25921/82ty-9e16) (Knapp et al., 2010; Gahtan et al., 2024). Upwelling observations use CUTI over seventeen latitude bands from 31°N to 47°N (Jacox et al., 2018; NOAA Southwest Fisheries Science Center & Jacox, 2018). Carbon-event definitions use the CMEMS flux product and its positive-uptake convention (Copernicus Marine Service, 2024). The release also retains the DMI marine-heatwave catalog (Hayward, 2026) and historical ONI (NOAA Climate Prediction Center, 2024) for source inspection; these two catalogs are not model inputs. Counts from different sources are kept separate because storms, sampled heatwave components, and active latitude bands have different units and spatial support.

A.5 Event Case Studies

Two cases illustrate the different temporal resolutions of the sources and fields. Heatmaps show each channel's seasonal anomaly divided by the training-period anomaly standard deviation at the same location, clipped to $[-3, 3]$; gray cells are invalid. The captions state the case selection rules.

Haiyan intensified from 34 to 170 kt in 3.5 days (Figure 6). Its track gives timing and intensity that the monthly cell cannot resolve. The 1997–1998 El Niño evolves over several months: regional SST rises while pCO_2 and DIC fall (Figure 7), a pattern consistent with earlier Pacific carbon analyses (Feely et al., 2002). These examples describe source resolution and co-variation; the fusion experiment tests the predictive value of the monthly event summaries.

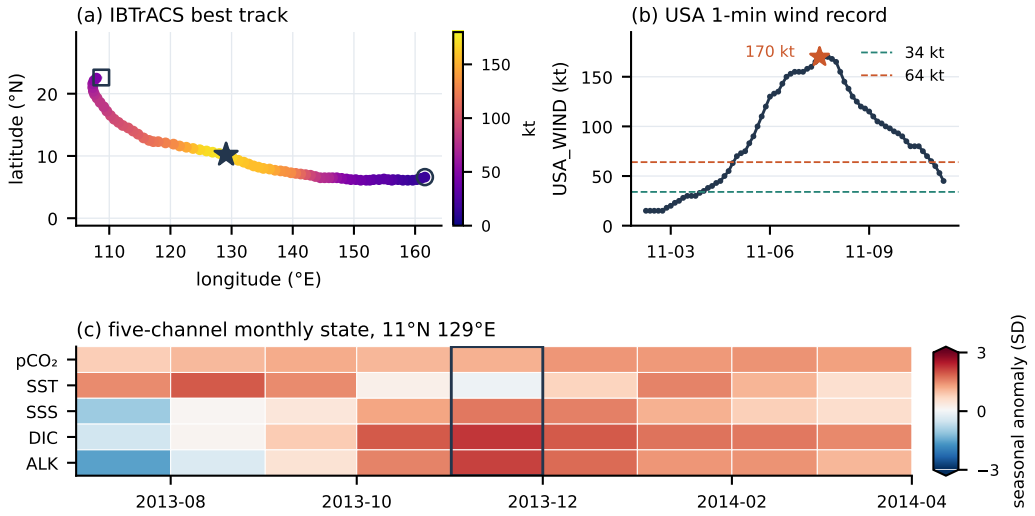

Figure 6: Typhoon Haiyan (2013), selected as the unique IBTrACS record named HAIYAN in season 2013. (a) Best-track positions colored by one-minute sustained wind; the star marks the 170-kt peak at 10.2°N, 129.1°E on 2013-11-07 12:00 UTC. (b) USA one-minute wind with the 34- and 64-kt thresholds. (c) Five-channel monthly state in the fixed 2° cell containing the peak (11°N, 129°E); the box marks November 2013.

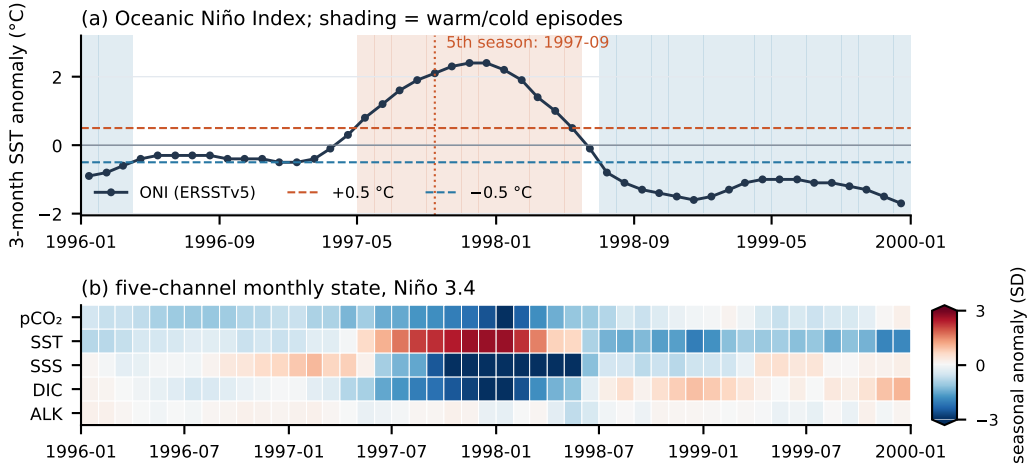

Figure 7: The 1997–1998 El Niño over the continuous window 1996–1999. (a) Historical ONI with the ± 0.5°C thresholds; shading marks retrospectively identified warm and cold episodes. (b) Five-channel monthly state averaged over the Niño 3.4 region.

B PAIRED TRANSPORT–EXCHANGE ENVIRONMENT

The Paired Transport–Exchange environment generates natural and perturbed futures under a known two-field dynamical system.

B.1 GOVERNING EQUATIONS AND SOLVER

Two positive fields x_1, x_2 live on the periodic domain $[0, 2\pi)^2$ and obey

$$\partial_t x_1 + \mathbf{v} \cdot \nabla x_1 = \nu_1 \Delta x_1 - \alpha x_1 + \beta x_2 - \gamma x_1 x_2, \tag{17}$$

$$\partial_t x_2 + \mathbf{v} \cdot \nabla x_2 = \nu_2 \Delta x_2 + \alpha x_1 - \beta x_2 + \gamma x_1 x_2, \tag{18}$$

with $\gamma = 0.2$. A mechanism fixes the spatially constant velocity $\mathbf{v}$, diffusivities ν_1, ν_2, and exchange rates α, β. Periodic boundaries and cancellation of the exchange terms conserve the integral of $x_1 + x_2$ within each branch. A pulse changes the initial integral, after which each branch conserves its own total.

The double-precision solver uses Fourier collocation on a 16×16 grid and Strang splitting with step $1/16$. Each step applies half a linear transport–exchange step, the exact local nonlinear exchange, and the second linear half-step. The linear operator uses a two-by-two matrix exponential per Fourier mode and an advection phase; first derivatives set the even-grid Nyquist mode to zero, while diffusion retains it. Records are one simulation-time unit apart. The same solver generates both branches independently.

B.2 MECHANISMS, INITIAL STATES, AND ACTIONS

Velocity components are drawn independently from $\mathcal{U}[-0.35, 0.35]$, diffusivities from $\mathcal{U}[0.015, 0.060]$, and exchange rates from $\mathcal{U}[0.020, 0.120]$. Each initial field is a constant, drawn from $\mathcal{U}[0.90, 1.20]$ for x_1 or $\mathcal{U}[0.70, 1.00]$ for x_2, plus four cosine modes. Mode amplitudes are uniform on $[-0.06, 0.06]$, phases on $[0, 2\pi)$, and integer wavenumbers on $\{0, 1, 2, 3\}^2$, with $(0, 0)$ replaced by $(1, 0)$. Separate deterministic NumPy random streams are keyed by profile, split, mechanism, and initial-state IDs in the released simulator. Splits use disjoint mechanisms (Table 6); each sample has six history frames and six future steps.

An action adds one periodic Gaussian pulse to x_1 at the initial time. The sixteen actions combine four centers $(0, 0)$, $(\pi, 0)$, $(0, \pi)$, (π, π), two signs, and two peak amplitudes, 0.05 and 0.10, all with width 0.6. An action changes only the initial state, not the equations or the mechanism parameters. Every initial state receives the identical pulse, so differences between responses arise from the state and the mechanism.

The basis in Appendix C.1 contains each action template and templates on the second field to represent transferred responses. The two event labels are fractions of grid points above each field's training-period 90th percentile. They are recomputed from each branch's true field; numerical error tolerances determine their lower and upper bounds.

B.3 NUMERICAL CHECKS AND RESPONSE STRUCTURE

Table 7 reports numerical checks over four mechanisms, four initial states per mechanism, and all sixteen actions, in standardized evaluation units. The numerical-error energy is 1.4×10^{-7} of the within-mechanism response-difference energy in these checks.

Action means account for 71.7% of response energy, between-mechanism deviations for 28.1%, and within-mechanism deviations for 0.15%, with similar shares after refinement. These scales motivate the action-mean lookup control and the component-wise evaluation in Figure 4.

Table 6: Splits of the Paired Transport–Exchange environment. Mechanisms are disjoint across splits.

Split	Mechanisms × initial states	Initial states	Use
Train	32×16	512	training and training-period statistics
Selection	4×16	64	checkpoint selection and development evaluation
Calibration	4×16	64	readout calibration and scales
Test	16×16	256	final evaluation after freezing

Table 7: Numerical checks of the paired-environment solver (maximum over checked cases).

Check	Result
Equation residual at step 10^{-3} / 10^{-4}	2.1×10^{-4} / 2.1×10^{-5}
Single Fourier mode vs. analytic solution	7×10^{-12}
Spatial refinement (16^2 vs. 64^2): natural / response	2×10^{-7} / 3.7×10^{-6}
Temporal refinement ($1/16$ vs. $1/128$): natural anomaly / response	1.2×10^{-3} / 2.4×10^{-5}
Conservation within a branch / branch difference vs. pulse integral	1×10^{-14} / 2.9×10^{-12}
Linear path ($\gamma = 0$) vs. exact linear operator	1×10^{-13}
Numerical-error energy / within-mechanism response-difference energy	1.4×10^{-7}
Minimum field value at record times	0.14

C Implementation Details

This section specifies the architecture, objectives, and training procedure. Field normalization, ocean basis construction, event scales, and seasonal baselines use training data only. The model stores these quantities with its weights.

C.1 Architecture

Basis. Stack the standardized current fields at the n training origins into a matrix X, with zeros at cells outside the support that remains valid throughout training. For the uncentered, unweighted decomposition $X = U\Sigma V^\top$, the ocean basis is $A = V_{:,1:32}\Sigma_{1:32,1:32}/\sqrt{n}$. Cell areas enter W in the subsequent weighted projection. The paired basis has sixteen periodic Gaussian templates per field, centered on $\{0, \pi/2, \pi, 3\pi/2\}^2$ with spatial width 0.6. Each template is divided by its field's training standard deviation, so the action coefficient retains the pulse amplitude in native simulator units.

Encoder and field readout. The encoder E_θ is a four-layer, four-head transformer over non-overlapping patches of the residual field Qx_t, its validity mask, and the grid coordinates, with 6×6 patches on the ocean grid and 4×4 patches on the paired grid. A summary token carrying the calendar embedding is mapped by a two-layer projector to $h_t^x \in \mathbb{R}^{128}$. The residual branch d_θ of the field readout is a linear map from $\mathbb{R}^{128}$ to the field.

Predictor. Each frame becomes a token $h_\tau + W_a a_\tau + W_c c_\tau + p_\tau$, with a learned position embedding p_τ. Four causal transformer blocks use four heads and feature-wise linear modulation (FiLM), whose scale and shift parameters are computed from the next calendar. In the first ocean stage and paired setting, a second FiLM contribution comes from the detached event readout $\mathrm{sg}(G_e)$ of the current window. The coefficient head is zero-initialized; the residual head predicts $\hat{h}$ directly.

Event readout. The network has one hidden layer of 128 units. In the second ocean stage, its output layer starts at zero. Each bounded family uses two sigmoid outputs sorted into lower and upper bounds; Niño 3.4 uses a three-class softmax. Equation (7) defines the ocean inference gate. The first ocean stage uses an unanchored head, and the paired head predicts four bounds for its two events.

C.2 OBJECTIVES

The SIGReg term compares random one-dimensional projections of the residual vectors with a standard Gaussian through the empirical characteristic function,

$$\mathcal{L}_{\mathrm{SIG}} = \frac{B}{M(L+K)} \sum_{\tau=t-L+1}^{t+K} \sum_{j=1}^{M} \int_{-3}^{3} e^{-s^2/2} \left| \frac{1}{B} \sum_{i=1}^{B} e^{\mathrm{i}s\langle v_j, h_{i,\tau}\rangle} - e^{-s^2/2} \right|^2 \mathrm{d}s, \tag{19}$$

with $M = 1024$ random unit directions and $B = 8$ independent natural samples. Context vectors include event fusion when enabled; future vectors are field-only target encodings. We evaluate the even integrand on 17 points in $[0, 3]$ with trapezoidal weights and double the result. The field term averages over valid cells, channels, and leads,

$$\mathcal{L}_{\mathrm{field}} = \frac{1}{Kn_v} \sum_{k=1}^{K} \sum_{v=1}^{n_v} \frac{\sum_j w_j m_{t+k,v,j} \left(\hat{x}_{t+k,v,j} - x_{t+k,v,j}\right)^2}{\sum_j w_j m_{t+k,v,j}}, \tag{20}$$

where n_v is the number of channels and w_j the area of cell j (uniform on the paired grid).

The event loss uses squared error for known bounds and cross-entropy for known Niño 3.4 classes. The second ocean stage uses $\mathcal{L}_G = 0.02[\mathcal{L}_{\mathrm{event}} + 0.01\overline{r_\phi^2}]$. In the first stage and paired setting, data and event gradients are computed separately for each shared module and combined as

$$g = g_D + g_G - \frac{\min(0, \langle g_G, g_D\rangle)}{\|g_D\|^2} g_D. \tag{21}$$

When $g_D = 0$, the event gradient is retained. The projection removes the conflicting component at the current parameters. All parameter gradients subsequently share a global norm clip.

For paired response training, q assigns half the squared norm to fields and half to events. The field part is divided by $\sqrt{2n_v HW}$; the event part is divided by its training scale and $\sqrt{2n_e}$, with unknown components set to zero. Thus the loss averages standardized field-response and event-response errors. Its denominator $V_{\mathrm{train}} = \mathbb{E}\|q(r_x^*, r_e^*)\|^2$ is estimated from the first 32 training pairs. A single response weight is calibrated once so that its gradient norm on F_θ is 25% of the field-loss gradient norm.

C.3 TRAINING SCHEDULES

Ocean dataset. Primary ablations use 3,000 first-stage and 2,000 second-stage updates and compare fixed endpoints. Screening ablations use the first-stage endpoint; the 30,000-update checks follow the main schedule. The main model trains the first stage for up to 28,000 updates, validates every 500 updates, and selects the lowest field-error checkpoint as the parent. The second stage restarts AdamW and the sampler, trains for 2,000 updates, validates every 100 updates, and selects its lowest field-error checkpoint. The three main seeds select parents at updates 26,500, 27,500, and 24,500 and second-stage checkpoints at 1,400, 1,500, and 1,200, respectively. Primary ablation seeds are 17, 29, 43, 61, 73, 89, and 101; three-seed runs use the first three.

Paired environment. Training uses 1,000 natural updates followed by 5,000 paired updates. Uniform selection of 288 candidate pairs yields 576 perturbed futures, all available from the start of paired training. Each pair contains two roots from the same mechanism, matched by natural-future distance, with one shared action. A batch uses eight distinct mechanisms. SIGReg uses a separate batch of eight natural roots. History replacement exchanges the first five frames within a pair, or uses a same-mechanism donor during natural training, with probability 0.5. Mechanism IDs are used only for sampling. Checkpoints minimize two-branch field error on the selection mechanisms, evaluated every 250 updates. Six-seed comparisons use 17, 29, 43, 61, 73, and 89.

Optimization. Both settings use AdamW with weight decay 0.01 and a global gradient-norm clip of 1, with learning rate 3×10^{-4} on the ocean dataset and 5×10^{-5} on the paired environment. A batch contains eight windows, or eight pairs of initial states in the paired stage.

C.4 PSEUDOCODE

Algorithm 1 specifies the forward computation. Here $F_\theta(S, c)$ includes the event readout, stop-gradient, and FiLM conditioning internally when that pathway is enabled; its condition is a function of S. Training applies the objectives above to the predictions and corresponding true futures. Table 8 summarizes hyperparameters.

Algorithm 1 SP-JEPA: state initialization and autonomous rollout

Require: history H_t; calendars $c_{t+1:t+K}$; optional action u; fixed A, C_A, Q; model setting
for $\tau = t - L + 1, \dots, t$ **do** ▷ basis-anchored state
  $a_\tau \leftarrow C_A x_\tau$; $h_\tau \leftarrow E_\theta(Q x_\tau, m_\tau, c_\tau)$
  if historical event fusion is enabled **then**
    $h_\tau \leftarrow h_\tau + B_e \xi_\tau + b_e$
  end if
end for
$S^0 \leftarrow [(a_\tau, h_\tau)]_{\tau=t-L+1}^{t}$; $\mathcal{B} \leftarrow \{0\}$
if an action u is supplied **then**
  $S^u \leftarrow I_u(S^0)$; $\mathcal{B} \leftarrow \{0, u\}$
end if
for $k = 1, \dots, K$ **do** ▷ autonomous rollout with a shared F_θ
  for each branch $\beta \in \mathcal{B}$ **do**
    condition F_θ on $\mathrm{sg}(G_e(S^\beta))$ if enabled
    $(\Delta a, \hat{h}) \leftarrow F_\theta(S^\beta, c_{t+k})$; $\hat{a} \leftarrow a_{\mathrm{last}}(S^\beta) + \Delta a$
    $S^\beta \leftarrow \mathrm{shift}\big(S^\beta, (\hat{a}, \hat{h})\big)$ ▷ drop oldest frame, append prediction
    $\hat{x}^\beta_{t+k} \leftarrow A\hat{a} + Q d_\theta(\hat{h})$
    $\hat{e}^\beta_{t+k} \leftarrow G_e(S^\beta, c_{t+k})$ ▷ readout for the current setting and stage
  end for
  if u is supplied **then**
    $\hat{r}_{t+k} \leftarrow \hat{x}^u_{t+k} - \hat{x}^0_{t+k}$
  end if
end for
return field and event forecasts, and response forecasts when u is supplied

Table 8: Hyperparameters of SP-JEPA.

Setting	Ocean dataset	Paired environment
History L / horizon K	12 / 3 months	6 / 6 steps
Grid / channels n_v	90×180 / 5	16×16 periodic / 2
Basis A ($r = 32$)	leading singular vectors of training fields	periodic Gaussian templates, width 0.6
Residual width d	128	128
Encoder	4 layers, 4 heads, 6×6 patches	4 layers, 4 heads, 4×4 patches
Predictor	4 causal layers, 4 heads, FiLM	4 causal layers, 4 heads, FiLM
Event readout	seasonal anchor, 128 hidden units	two bounded events, 128 hidden units
Schedule	3,000 + 2,000 updates (ablations); 28,000 + 2,000 (main comparison)	1,000 natural + 5,000 paired updates
Optimizer	AdamW, lr 3×10^{-4}, wd 0.01	AdamW, lr 5×10^{-5}, wd 0.01
Batch	8 windows	8 pairs from 8 mechanisms
λ_{SIG} / SIGReg directions	0.09 / 1024	0.09 / 1024
Event loss weight	0.02 (residual penalty 0.01)	0.02
Response weight	—	gradient on F_θ at 25% of field gradient
History replacement	—	probability 0.5
Checkpoint	end of each stage (ablations); lowest validation field error per stage (main comparison)	lowest two-branch field error, every 250 updates
Gradient clip	global norm 1	global norm 1

C.5 Response metrics and readout checks

For each true future, $\mathrm{enc}(x) = (C_A x, E_\theta(Qx, m, c))$. The encoded response is $r^E = D_x(\mathrm{enc}(x^u)) - D_x(\mathrm{enc}(x))$; the rollout response uses the same readout with predicted states. Let $\|v\|_w = (\mathbb{E}\|v\|^2)^{1/2}$ under the common evaluation weights. Then

$$\|r^M - r^*\|_w \le \|r^M - r^E\|_w + \|r^E - r^*\|_w = \sqrt{PV} + \sqrt{CV},$$

which proves LCS $\le$ RS. Capture includes encoder and readout error; propagation compares rolled-out and encoded futures rather than isolating a single module. Both scores equal one for exact responses and zero when both model responses vanish.

Before response evaluation, independent calibration mechanisms compare decoded current fields with the training-mean predictor and encoded future responses with the zero-response predictor, separately for each field. All six reported checkpoints have lower errors than both corresponding baselines. Confidence intervals use 2,000 mechanism bootstrap samples. The first-field threshold event has zero response on the calibration set, so the reported LCS evaluates fields. Branch-wise field errors accompany the response metrics because subtraction can cancel errors shared by both branches.

D Additional Results

This section gives the evaluation protocol of Section 5, per-channel results and forecast maps on the ocean dataset, the complete event-fusion and ablation results, and the diagnostics cited in the main text. All numbers are computed on the validation splits.

D.1 Evaluation protocol

Ocean dataset. The chronological split has 283 training, 35 validation, and 36 test origins. The validation origins run from September 2017 to July 2020, so their targets cover October 2017 to October 2020. Fields are standardized per channel with training means and standard deviations, and the monthly climatology used for anomalies, ACC, and persistence is the mean of the training fields for each calendar month. All methods in Table 1 are evaluated by one script with the same area weights, validity masks, and climatology. The compared methods train with batches of two windows for 30,000 updates, are validated every 1,000 updates, and are selected by validation field error, with seeds 17, 29, and 43. SP-JEPA follows the schedule of Appendix C.3 with batches of eight windows and sees 205,600–232,000 windows up to its selected checkpoints. Table 9 separates this budget from the comparison of Table 1. The development-scale model, trained for 3,000 first-stage and 2,000 second-stage updates and taken at its final update without validation-based selection, sees 40,000 windows, fewer than any compared method, and still reaches a lower nRMSE than every latent world model and every field forecaster except FNO. Continuing the second stage from the final first-stage checkpoint instead of the validation-best one, and reporting the final update, changes the field MSE of the main model by $+0.7\%$.

Table 9: Training budget on the ocean dataset. Windows count the training windows seen up to the reported checkpoint. DreamerV3 RSSM and FNO are the strongest latent world model and field forecaster of Table 1 and use their validation-best checkpoints, whereas the 5k model uses its final update.

Model	Updates	Batch	Windows	nRMSE ↓	ACC ↑
DreamerV3 RSSM	30,000	2	60,000	0.160	0.253
FNO	30,000	2	60,000	0.132	0.481
SP-JEPA, 5k updates (7 seeds)	5,000	8	40,000	0.133	0.355
SP-JEPA, main model	25,700–29,000	8	205,600–232,000	0.131	0.374

Paired environment. Every model trains for 6,000 updates with learning rate 5×10^{-5} on the same 576 perturbed futures: 1,000 updates on natural trajectories, followed by 5,000 updates on pairs of natural and edited trajectories. Each compared latent world model applies its own loss to

both branches with equal weight and receives neither a response loss nor event labels. The compared models are selected every 500 updates, and SP-JEPA every 250, by the two-branch field MSE on the four validation mechanisms. A learning rate of 3×10^{-4} lowers the RS of LeWM (six seeds), TC-LeWM, and the EAWM-inspired model (one seed each) and raises that of DreamerV3 RSSM to -1.04 (one seed); every compared model stays below zero. The encoded response of a compared model is obtained by encoding the full causal window of each branch and decoding its last frame with the model's own decoder; DreamerV3 uses the modes of its categorical posterior and prior, so its evaluation is deterministic. RS, LCS, $\sqrt{C}$, and $\sqrt{P}$ are computed by one routine for all models, with confidence intervals from 2,000 mechanism bootstrap samples.

Metrics and tests. The nRMSE is the square root of the area-weighted MSE of standardized fields, averaged over leads and channels before the root; the field-MSE ratios of the ablations use the same MSE. The ACC is the area-weighted pattern correlation of centered anomalies for each origin, lead, and channel, averaged afterwards. Event error is the squared error of the predicted bounds for the three families that the gate χ keeps (warm area, carbon-flux reversal, and carbon sink), and event ratios are geometric means of the per-family ratios. Error ratios are geometric means of per-seed ratios. Confidence intervals come from a paired two-level bootstrap with 4,000 samples, which resamples seeds with replacement and 12-month circular blocks of the 35 consecutive origins, with the same indices for both arms. The four pre-specified ablations use a one-sided exact sign-flip test on the log ratios of seven seeds, whose smallest attainable p-value is $1/128$, with Holm correction at a family-wise level of 0.05. An arm has a substantive effect when it is significant, its ratio is at least 1.02, and the lower bound of its interval exceeds 1. Arms with three seeds serve as screening and receive no significance test.

D.2 Per-channel results and forecast maps on the ocean dataset

Table 10: Table 1 by channel: nRMSE with its seed standard deviation, RMSE of the standardized fields, and ACC, averaged over leads (three seeds; persistence is deterministic). **Bold**: best; underline: second best.

		RMSE of standardized fields ↓					ACC ↑				
Method	nRMSE ↓	pCO_2	SST	SSS	DIC	ALK	pCO_2	SST	SSS	DIC	ALK
LeWM	0.173 ± 0.002	0.304	0.078	0.130	0.131	0.132	0.274	0.062	0.241	0.216	0.130
TC-LeWM	0.171 ± 0.004	0.303	0.076	0.130	0.127	0.128	0.254	0.111	0.284	0.264	0.222
EAWM-inspired	0.185 ± 0.004	0.324	0.084	0.136	0.140	0.144	0.322	0.037	0.232	0.203	0.082
DreamerV3 RSSM	0.160 ± 0.000	0.263	0.073	0.139	0.130	0.133	0.437	0.077	0.268	0.246	0.240
SP-JEPA (ours)	**0.131** ± 0.000	**0.208**	0.058	0.125	0.107	0.113	0.522	0.276	0.330	0.400	0.342
FNO	0.132 ± 0.001	0.223	0.062	**0.110**	**0.103**	**0.107**	**0.558**	0.314	0.539	0.511	0.485
ConvLSTM	0.137 ± 0.001	0.227	0.069	0.115	0.110	0.112	0.507	0.248	0.507	0.473	0.466
PDE-Transformer	0.142 ± 0.002	0.230	0.067	0.124	0.114	0.121	0.494	0.127	0.403	0.393	0.356
U-Net	0.147 ± 0.001	0.242	0.078	0.124	0.118	0.120	0.460	0.183	0.443	0.428	0.418
NeuralOM	0.149 ± 0.001	0.260	0.066	0.115	0.112	0.114	0.392	0.387	0.536	0.487	0.499
ClimaX	0.155 ± 0.004	0.250	0.072	0.141	0.124	0.130	0.462	0.207	0.350	0.396	0.378
Persistence	0.155	0.283	**0.056**	0.113	0.109	0.110	0.325	**0.516**	**0.581**	**0.540**	**0.565**

Table 10 splits Table 1 by channel. The pCO_2 anomalies of the validation years carry a large uniform offset, so this channel contributes 50–67% of the mean squared error of every method. SP-JEPA has the lowest pCO_2 error of all methods and the second-lowest SST and DIC errors, whereas FNO is more accurate in SSS, DIC, and alkalinity, and persistence has the highest ACC in the four channels other than pCO_2. The maps below show where these differences arise.

Figure 8 shows a single forecast in physical units. The origin is the one with the median lead-3 error. Because the pCO_2 of the validation years exceeds the training climatology by 24 μatm on average, the maps show anomalies with their global mean removed. The true anomalies contain the cold equatorial Pacific of the La Niña conditions of winter 2017/18 and, over the same region, a positive pCO_2 anomaly, consistent with stronger upwelling of carbon-rich water. The basis-span part of the truth keeps these large-scale structures and drops the smaller ones. The forecast of SP-JEPA resembles this span part more closely than the full truth: its correlation with the span part is 0.87 for pCO_2 and 0.68 for SST, against 0.56 and 0.26 with the full truth. Persistence keeps the small-scale structure of September, some of which has decayed by December, and reaches an ACC

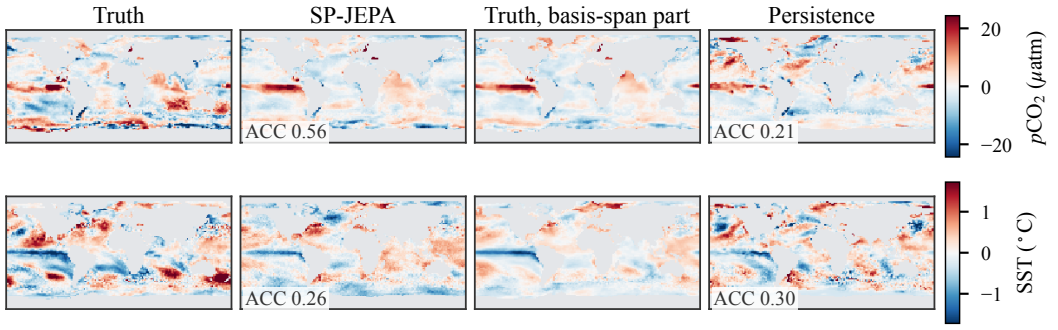

Figure 8: Anomalies of pCO_2 (top) and SST (bottom) for December 2017, forecast from September 2017 (lead 3), the validation origin with the median lead-3 error of SP-JEPA (seed 17). Anomalies are taken from the training climatology, and the global mean of each map is removed. The third column is the part of the true anomaly in the span of the fixed basis. Labels give the ACC of each forecast.

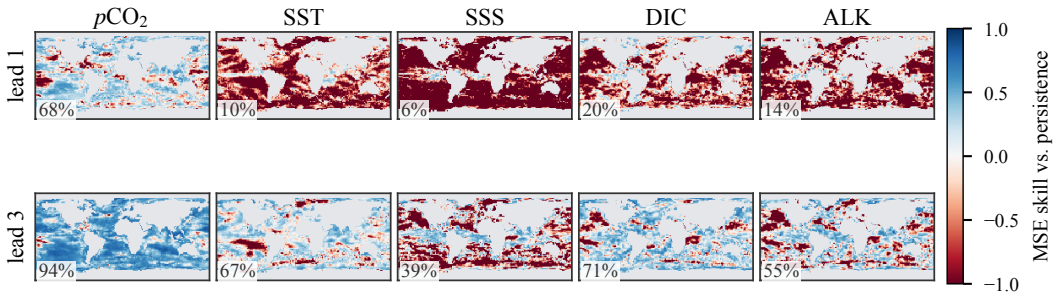

Figure 9: MSE skill of SP-JEPA relative to seasonal-anomaly persistence in each grid cell over the 35 validation origins, $1 - \mathrm{MSE}_{\text{SP-JEPA}}/\mathrm{MSE}_{\text{persistence}}$, at leads 1 and 3 (SP-JEPA averaged over three seeds). Blue marks cells where SP-JEPA is more accurate; labels give the share of the ocean area where it is.

of 0.21 for pCO_2 and 0.30 for SST. The single case thus shows the division of Figure 3c in space: SP-JEPA forecasts the part of the anomaly that the basis represents, and persistence retains the part that it does not.

Figure 9 extends this comparison to all origins and all channels. At the first month, persistence is more accurate over most of the ocean for every channel except pCO_2, where SP-JEPA is more accurate over 68% of the area. At the third month, the share rises to 94% for pCO_2, 71% for DIC, 67% for SST, and 55% for alkalinity. Salinity remains the exception at 39%, and it also has the smallest share at the first month (6%). The spatial pattern agrees with the lead structure of Table 1: persistence wins where the last observation is still informative, and the advantage of the rollout grows as that information decays.

D.3 Event-observation fusion

Table 11: Event-observation fusion at full scale. Three models continue the same first-stage checkpoint, selected within 28,000 updates, for 2,000 updates (three seeds). Ratios divide the error of SP-JEPA by that of the control, so values below 1 favor fusion; parentheses give one-sided 95% upper bounds.

Control	Field ratio [95% CI]	Event ratio	Warm area	Carbon-flux reversal	Carbon sink
No fusion	0.994 [0.987, 0.999]	0.912 (0.982)	0.787	0.914	1.054
Training-year events	0.993 [0.987, 0.999]	0.893 (0.965)	0.786	0.909	0.996

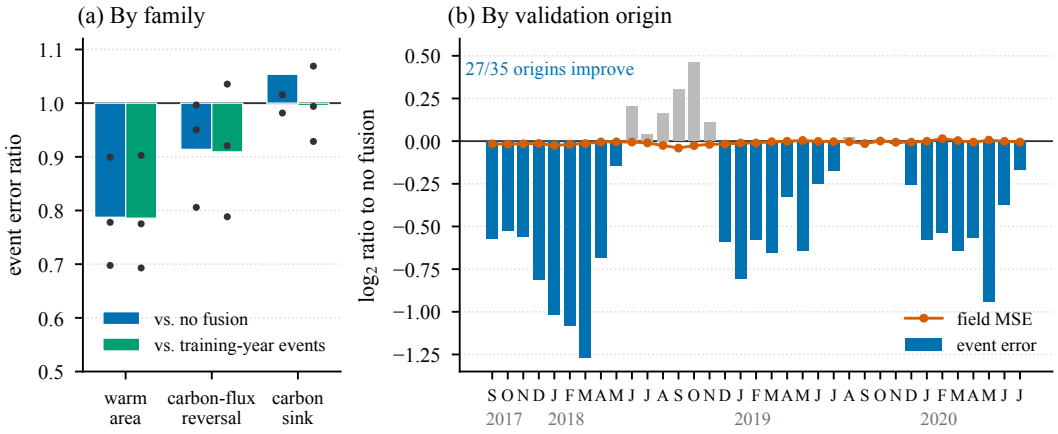

Figure 10: Event-observation fusion at full scale. (a) Error ratio of each event family to the two controls; dots are seeds. (b) Log ratio of event error (bars) and field MSE (line) to the model without fusion for each validation origin, geometric mean over three seeds; letters give the calendar month of the origin.

The second control replaces the event history of each window by that of a training window ending in the same calendar month and keeps only the entries known in both. Table 11 and Figure 10a separate the event gain by family. Warm area improves by 21% against both controls and carbon-flux reversal by 9%, whereas the carbon sink changes by $+5\%$ against no fusion and by less than 1% against training-year events. The mean warm-area fraction is 0.12 over the training months and 0.26 over the validation target months, so the observed record carries a level that the training-year histories and the seasonal anchor of the readout do not. Figure 10b shows that the gain is spread over the validation period: 27 of the 35 origins improve, and six of the eight losses fall on the consecutive origins from June to November 2018. The field error of every origin changes by between -2.8% and $+1.0\%$.

The remaining event arms change the field error little. Replacing additive fusion by a gated fusion $h \odot (1 + \tanh(B_e \xi + b_e))$ changes the field error by -0.7% and the event error by $+6\%$. The three first-stage event mechanisms, which keep the event loss out of the encoder and predictor, remove the

event conditioning of F_θ, or add the event gradient without projection, each change the field error by less than 0.3%. The data and event gradients are nearly orthogonal throughout the first stage, with cosines between -0.03 and 0.07, so the projection of Eq. (21) rarely removes anything.

D.4 All ablation arms

Table 12: All ocean ablation arms, grouped as in Table 2; each ratio compares an arm with the full model on the same schedule. Pre-specified arms (†: significant after Holm correction) use the development schedule of 3,000 and 2,000 updates with seven seeds; screening arms end after 3,000 first-stage updates with three seeds; event arms continue the same first-stage parents with five seeds, except the fusion-timing arm, which trains for 5,000 updates without an optimizer restart and is compared with delayed fusion. Intervals resample seeds and 12-month blocks of origins.

Change	Field ratio [95% CI]	Seeds worse	Event ratio
SP-JEPA (full model)	1	–	1
Within a module			
Trainable basis, same initialization	0.989 [0.973, 1.005]	1/7	–
Teacher forcing instead of open loop	1.002 [0.992, 1.011]	5/7	–
Two-layer MLP residual readout	0.983 [0.976, 0.990]	0/3	–
GRU instead of causal transformer	1.007 [1.000, 1.015]	3/3	–
Linear attention	1.000 [0.990, 1.007]	2/3	–
Increment head, random initialization	1.028 [1.009, 1.059]	3/3	–
Absolute coefficients (vs. row above)	1.041 [1.018, 1.068]	3/3	–
Summed instead of projected event gradient	0.998 [0.992, 1.002]	1/3	–
Gated instead of additive fusion	0.993 [0.986, 0.999]	0/5	1.06
No seasonal anchor in the event readout	1.000 [1.000, 1.000]	–	3.39
Between modules			
Remove basis coefficients	1.510 [1.370, 1.669]†	7/7	–
Remove JEPA term	1.049 [1.036, 1.063]†	7/7	–
No direct field gradient to predicted h	0.995 [0.990, 1.000]	1/3	–
Fixed coefficients, $\Delta a = 0$	7.32 [6.74, 7.90]	3/3	–
No attention across frames	0.999 [0.991, 1.008]	1/3	–
No calendar input	1.020 [0.993, 1.050]	3/3	–
Event loss kept out of E_θ and F_θ	0.999 [0.996, 1.001]	2/3	–
No event conditioning of F_θ	0.997 [0.990, 1.001]	1/3	–
Properties			
Residual width $d = 64$	1.020 [1.004, 1.037]	3/3	–
Residual width $d = 256$	1.016 [1.004, 1.031]	3/3	–
Fusion from update 0 (vs. 3,000)	1.007 [0.998, 1.015]	4/5	0.81

Table 13: Paired-environment arms trained from scratch on natural trajectories for 6,000 updates (three seeds). The swap probe gives the skill of the sixth-step prediction toward the future of a replaced current frame (Figure 17).

Arm	Natural error ratio	Seeds worse	Swap probe, step 6
SP-JEPA (full model)	1	–	0.87
Trainable basis, same initialization	0.44	0/3	0.94
Teacher forcing instead of open loop	1.26	3/3	0.82
Remove basis coefficients	1.72	3/3	0.86
Remove JEPA term	0.70	0/3	0.90
Fixed coefficients, $\Delta a = 0$	26.2	3/3	-1.13
No attention across frames	1.62	3/3	0.86
No history replacement	0.67	0/3	0.05

Tables 12 and 13 complete Table 2. Two pre-specified arms were also trained at full scale with the main protocol (three seeds): removing the basis coefficients raises the field error by 27.4% [20.3%, 35.1%], and removing the JEPA term by 0.7% [0.0%, 1.6%] at the validation-best checkpoints and by 6.4% at the final update (Appendix D.6).

Among the arms that Table 2 omits, a two-layer residual readout lowers the error by 1.7% in the screening runs; the main model keeps the linear readout of Eq. (3). Cutting the direct gradient of the field loss to the predicted residual representation and removing the calendar input give intervals that include no change, although the calendar arm moves the event families in opposite directions, raising the carbon-flux-reversal error 2.25-fold and lowering the carbon-sink error to 0.79. Freezing the coefficients, $\Delta a = 0$, raises the error 7.3-fold on the ocean dataset and 26-fold on the paired environment, so the evolution of the coefficients carries most of the forecast.

The event arms qualify the event readout and the training schedule. Removing the seasonal anchor raises the error of the carbon-flux-reversal and carbon-sink families 47-fold and 2.2-fold, whereas the warm-area error falls to 0.38 of its anchored value; the ratio of 3.39 in Table 12 is the geometric mean of these effects. Fusing events from the first update instead of after 3,000 updates changes the field error by $+0.7\%$ with a one-sided 95% bound of $+1.4\%$, below the pre-specified non-inferiority margin of 2%, and lowers the event error by 19% (one-sided bound: at least 6%), with the warm-area error falling to 0.66 of its value under delayed fusion. The main model keeps the two-stage schedule, under which the three models of the fusion test in Section 5.2 continue one shared parent.

D.5 Horizon and history

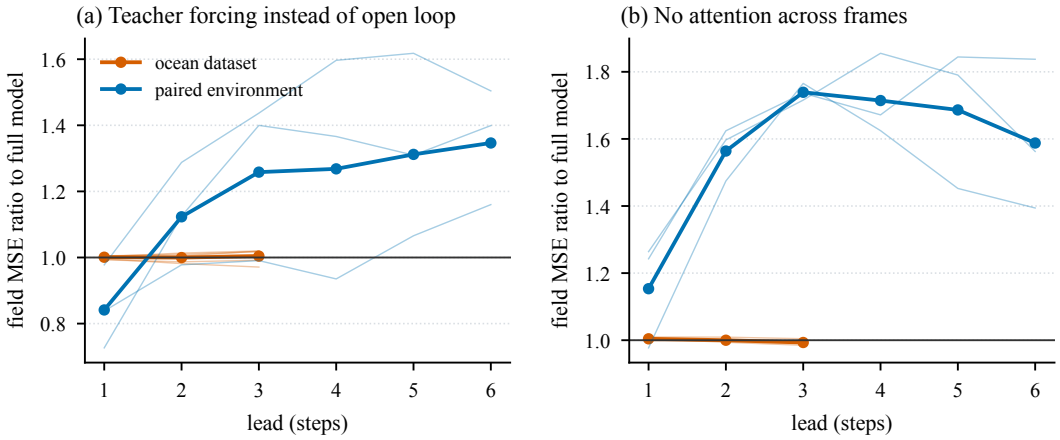

Figure 11: Field MSE ratio to the full model by lead for teacher forcing (a) and for removing attention across frames (b), on the ocean dataset (development scale) and the paired environment (natural training). Thin lines are seeds and thick lines their geometric mean.

Figure 11 resolves two ablations of Section 5.3 by lead. On the paired environment, teacher forcing is more accurate than open-loop training at the first step (ratio 0.84) and less accurate at every later step, up to 1.35 at the sixth, the growth of exposure bias with the horizon; over the three ocean leads, the two schedules do not differ. Removing attention across frames raises the paired error from the first step (1.15) to a maximum of 1.74 at the third and leaves the ocean error unchanged at every lead. Given the current field and the calendar, the history of the ocean dataset thus adds little information over three months.

D.6 Mechanism of the JEPA term

Figure 12a shows the full-scale behavior behind the small gap at validation-best checkpoints: without the JEPA term, two of three seeds reach their best validation error at update 1,000 and degrade afterwards, whereas the full model improves until updates 24,500–27,500, and by the final update of the second stage the field MSE without the term is 6.4% higher. Two mechanisms accompany this difference. With the term, the effective rank of the encoded and predicted residual representations stays at about eight during the first stage, whereas without it the predicted representation first expands to a rank of 27 and is still about 16 at update 3,000 (Figure 12b). With the term, the gradients that reach the predictor and the field readout are also 5.5 and 7.8 times larger, against 1.6 times for the encoder (Figure 12c), although the term does not act on the field readout; over the same updates, the training field loss is 50% higher with the term (0.0076 against 0.0050).

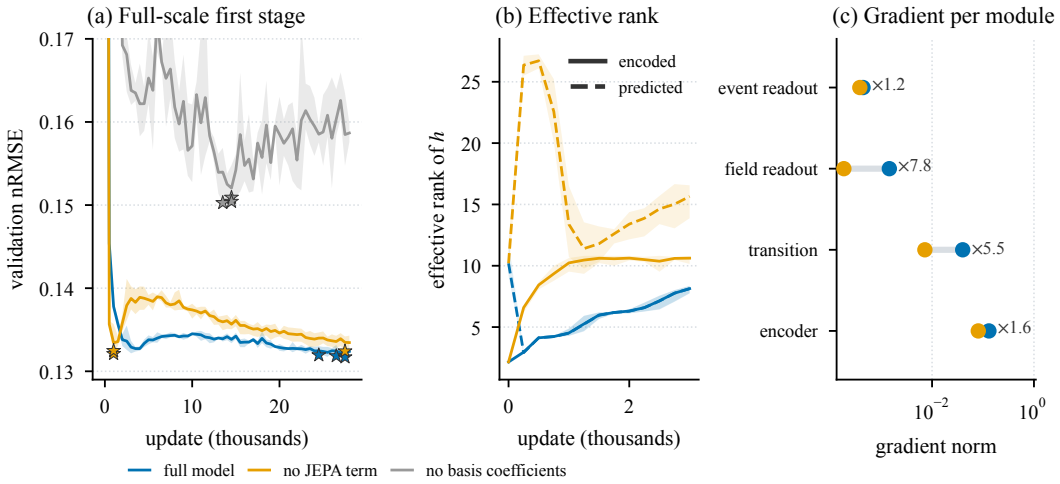

Figure 12: Effect of the JEPA term on the ocean dataset. (a) Validation nRMSE during the full-scale first stage; lines are seed means, bands the range over three seeds, and stars the validation-best update of each seed. (b) Effective rank of the encoded and predicted residual representations during the development-scale first stage (seven seeds; bands give the range; colors as in a); the two curves of the full model coincide after about 250 updates. (c) Gradient norm per module, averaged over updates 500–3,000; labels give the ratio of the full model to the model without the JEPA term.

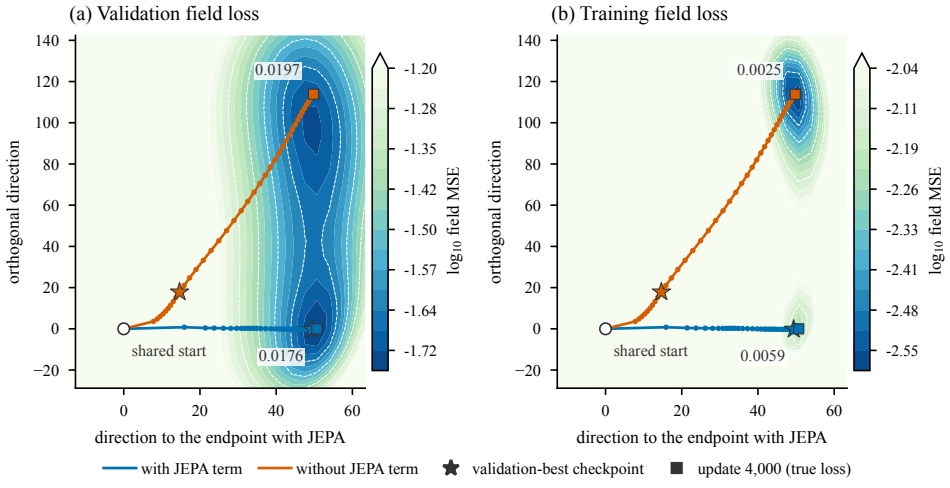

Figure 13: Validation (a) and training (b) field loss on the plane through a shared initialization and the final weights of two 4,000-update runs that differ only in the JEPA term (seed 29). Lines trace the checkpoints saved every 100 updates, projected onto the plane (up to 48 units off it); stars mark the validation-best checkpoint, and squares the final checkpoint, labeled with its true loss.

Two short runs from the same initialization and data order reproduce the full-scale behavior (Figure 13). Without the JEPA term, the weights travel 124 units from the start instead of 51, the final training field error falls to 0.0025 instead of 0.0059, and the final validation error rises to 0.0197 instead of 0.0176. The run without the term thus reaches a lower training loss that generalizes worse.

The decoding of the encoded current frame, which Section 5.3 uses to show that the JEPA term keeps the encoded state decodable, also depends on event fusion: after the second stage, the main model decodes this frame with an nRMSE of 0.43–0.50, against 0.124 without fusion and 0.17–0.24 with training-year events, while its forecast error is that of Table 1.

D.7 Basis vectors and edits on the ocean dataset

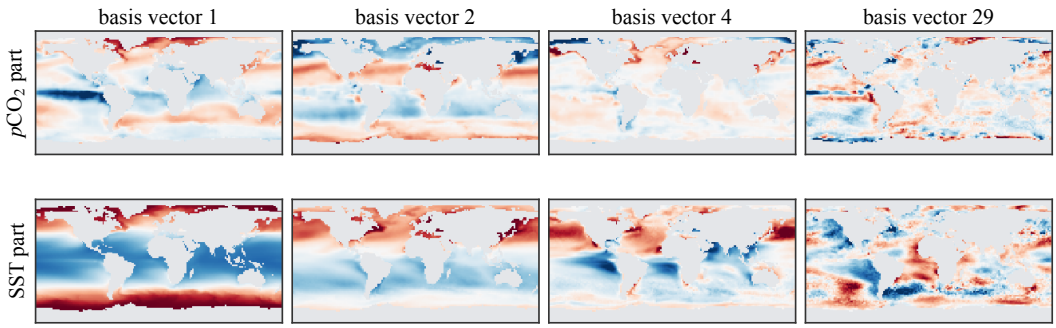

Figure 14: pCO_2 and SST parts of four vectors of the fixed ocean basis: the first, second, and fourth vectors, and vector 29, whose coefficient has the largest training anomaly variance. Signs and scales are arbitrary.

Figure 14 shows what the leading coefficients represent. Because the basis is computed from uncentered fields, the first vector resembles the mean state, with the meridional temperature gradient in its SST part. The second vector contrasts the two hemispheres, so its coefficient follows the annual cycle, peaking in August; the coefficient of the fourth vector peaks about three months later, and the two together trace the orbit of Figure 3a. Vector 29, whose coefficient has the largest anomaly variance in training, varies on smaller scales across the Pacific and the Southern Ocean.

Figure 15 examines edits on the ocean dataset. Each of the 32 coefficients of every validation origin is edited by plus and minus half the mean standard deviation of the training coefficient anomalies, and the rollout is compared with the unedited one. The responses are close to linear, since the responses to positive and negative edits, divided by the edit, differ by 3.6% of their mean magnitude, and reproducible, since the mean response matrices of the three seeds correlate at 0.88 or more. The LIM supplies a data-estimated linear response for comparison. On the sixteen coefficients with the largest anomaly variance, the SP-JEPA response correlates with it at 0.91, 0.80, and 0.71 at leads 1, 2, and 3, while a response equal to the edit itself correlates at 0.94, 0.83, and 0.65, so SP-JEPA agrees with the LIM less than this trivial response at the first two leads and more at the third. The cross-coefficient terms correlate only at 0.25–0.35, and the edits of SP-JEPA decay faster than those of the LIM at every lead: the mean response of an edited coefficient to its own edit is 0.61 against 0.86 at the first month and 0.29 against 0.54 at the third. Whether these magnitudes are correct cannot be determined from the ocean record, which contains no paired outcomes.

D.8 Responses on the paired environment

Table 14 lists the capture and propagation errors behind Table 1 and Figure 4c. The three SP-JEPA rows share a warm-up on natural trajectories and then differ only in supervision: natural trajectories alone, both branches without the response loss, and both branches with it. Their capture error is identical to two decimals, whereas the propagation error falls from 1.63 to 1.51 with edited-branch supervision and to 0.59 with the response loss. The LCS of the six SP-JEPA seeds ranges from 0.189 to 0.221, and its mechanism-bootstrap interval excludes zero for five of them.

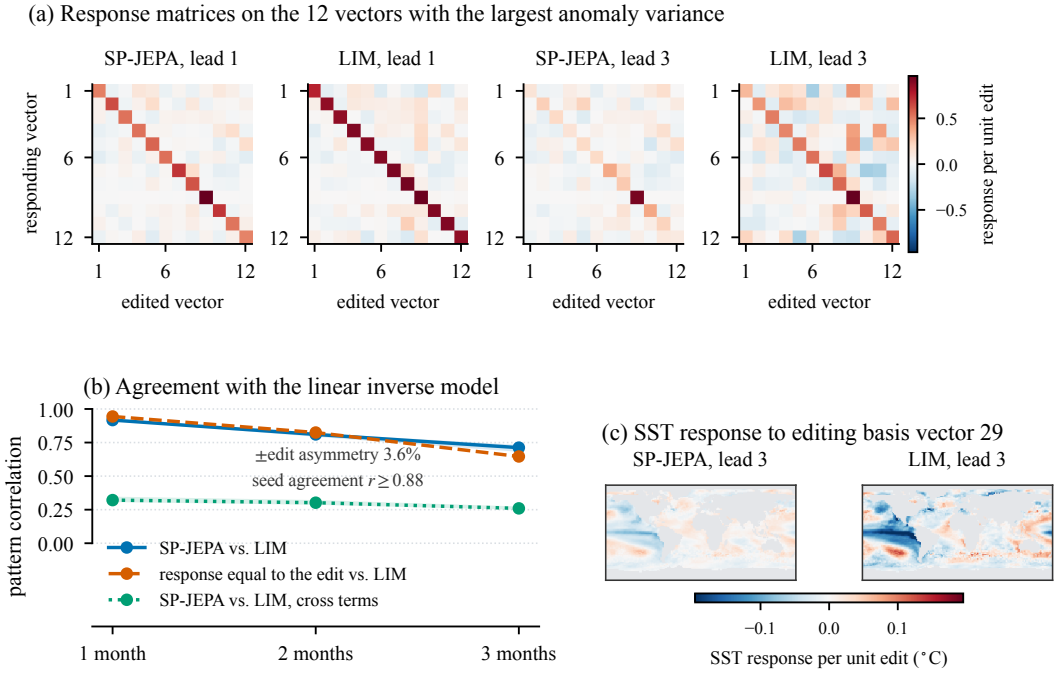

Figure 15: Edits of the basis coefficients on the ocean dataset (three seeds). (a) Mean response of each coefficient to a unit edit of each other coefficient, for SP-JEPA and for a linear inverse model (LIM) fitted to the training coefficient anomalies, on the twelve coefficients with the largest anomaly variance. (b) Pattern correlation of the SP-JEPA response with the LIM on the sixteen coefficients with the largest training anomaly variance, compared with that of a response equal to the edit itself; bands give the seed range. (c) SST pattern of the lead-3 response to a unit edit of basis vector 29.

Table 14: Complete response results on the four validation mechanisms of the paired environment (six seeds). Field MSE averages the natural and edited branches; RS_1 and RS_6 are the RS at the first and sixth step. Top: each model is trained with its own objective on both branches. Middle: SP-JEPA with less supervision after a shared warm-up. The action-mean table predicts the mean training response of each action and holds no state. **Bold** and underline: best and second best in the top block.

Model	Field MSE ↓	RS ↑	RS_1 ↑	RS_6 ↑	LCS ↑	$\sqrt{C}$ ↓	$\sqrt{P}$ ↓
LeWM	0.0113	−0.16	0.09	−0.47	−1.20	0.83	0.65
TC-LeWM	0.0071	−0.42	0.08	−0.94	−1.80	0.86	0.81
EAWM-inspired	0.0111	−0.14	0.10	−0.43	−1.22	0.84	0.65
DreamerV3 RSSM	0.0445	−5.40	−1.13	−9.32	−13.08	1.29	2.24
SP-JEPA (ours)	**0.0070**	**0.575**	**0.88**	**0.27**	**0.206**	**0.30**	**0.59**
SP-JEPA, natural data only	0.0054	−1.75	0.44	−3.26	−2.74	0.30	1.63
SP-JEPA, no response loss	0.0055	−1.38	0.48	−2.48	−2.30	0.30	1.51
Action-mean table	–	0.64	–	–	–	–	–

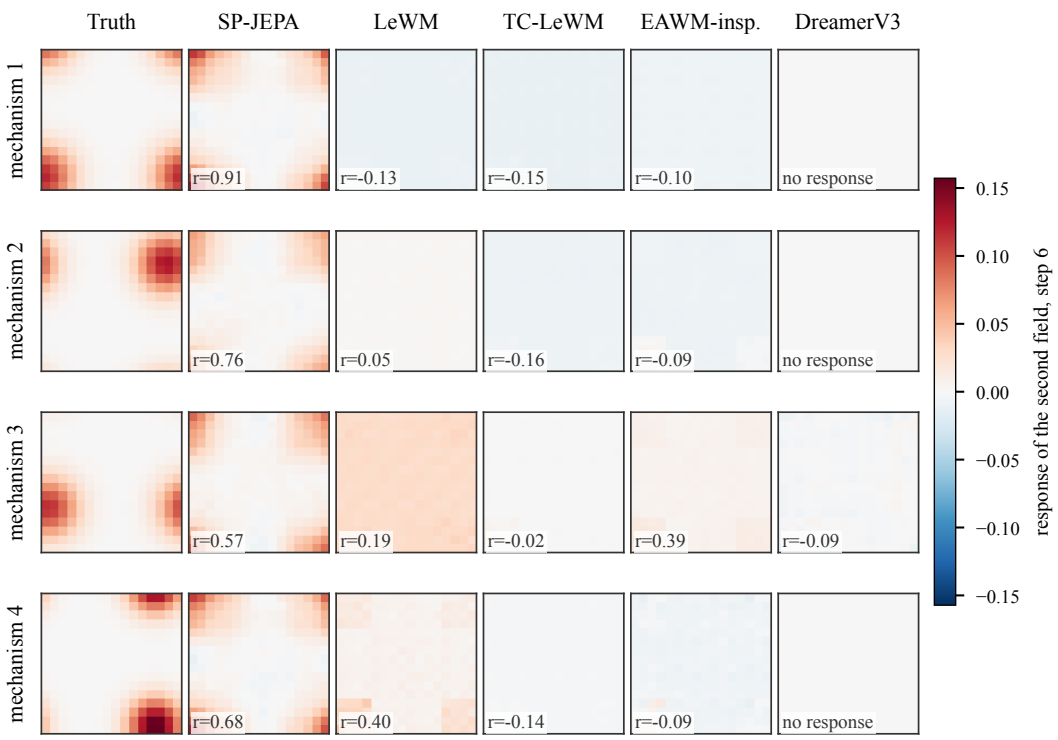

Figure 16: Response of the second field at step 6 to the same positive pulse from the same initial-state index in each of the four validation mechanisms. Labels give the correlation with the true response over all six steps and both fields; "no response" marks a model whose predicted response is constant.

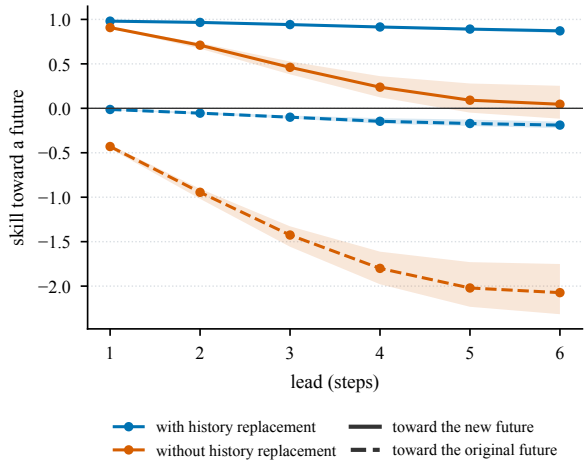

Figure 17: Swap probe on the paired environment for models trained on natural trajectories with and without history replacement (three seeds; bands give the range). The current frame of an initial state is replaced by that of another initial state from the same mechanism. The skill toward a future is $1 - \mathrm{MSE}(\hat{x}, x')/\mathrm{MSE}(x^{\mathrm{new}}, x^{\mathrm{orig}})$, where $\hat{x}$ is the prediction after the swap, x^{new} and x^{orig} are the true futures of the new and the original initial state, and x' is x^{new} (solid) or x^{orig} (dashed). The skill is 1 when the prediction matches that future and negative when the prediction is farther from it than the two futures are from each other.

Figure 16 extends Figure 4a to all validation mechanisms and all compared models. In each mechanism, the true response is carried away from the pulse in a direction set by the advection of that mechanism. SP-JEPA predicts the decay of the pulse and its transfer to the second field in every mechanism, with RS of 0.82, 0.54, 0.40, and 0.56 in the four mechanisms (six seeds), but it does not predict the direction of transport, which belongs to the mechanism component of Figure 4d. The compared latent world models predict weak and nearly uniform responses in every mechanism, and DreamerV3 RSSM predicts none in three of them. The skill of SP-JEPA declines with the horizon, from an RS of 0.88 at the first step to 0.27 at the sixth (Table 14), whereas LeWM, TC-LeWM, and the EAWM-inspired model are near zero at the first step and negative from the second or third, and DreamerV3 RSSM is negative throughout.

Figure 17 underlies the history-replacement result of Section 5.3. With history replacement, the prediction reaches a skill of 0.98 toward the new future at the first step and 0.87 at the sixth, and its skill toward the original future stays between -0.2 and 0. Without replacement, the skill toward the new future falls from 0.91 at the first step to 0.05 at the sixth, and the predictions move away from both futures, so an edit written into the current frame fades within a few steps.

E Extended Related Work

The following studies provide further context for the observations and modeling choices in Section 2.

Reconstruction and forecasting. Ocean carbon reconstructions combine sparse measurements with environmental predictors to estimate global surface fields (Landschützer et al., 2013; Fay & McKinley, 2013). These estimates support carbon-budget assessments (Le Quéré et al., 2018; Müller et al., 2025) and analyses of trends and regimes (Zemskova et al., 2022; Couespel et al., 2024). Forecasting has a different target: future fields conditional on the available history. Diffusion forecasters and Earth-system foundation models extend this task across variables and scales (Kassaï Koupaï et al., 2025; Rühling Cachay et al., 2025; Soares et al., 2025; Wang et al., 2026; Cresswell-Clay et al., 2025). Temporal Earth-observation models also learn representations across spatial resolutions and modalities (Feng et al., 2026; Ranjan et al., 2026; Astruc et al., 2025).

Events and carbon variability. Cyclone studies quantify changes in air–sea exchange during storms (Bates et al., 1998; Ye et al., 2026). Marine-heatwave studies examine carbon uptake and biological responses across open-ocean and coastal settings (Mignot et al., 2022; Hu et al., 2026; Bif et al., 2025; Bian et al., 2026). ENSO analyses connect Pacific carbon variability to basin-scale conditions (Feely et al., 2002; Liao et al., 2020), while upwelling studies relate carbon-rich subsurface water to surface chemistry (Feely et al., 2008; Stoll et al., 2025). These process studies motivate the source selection; the model uses the monthly definitions in Table 5.

Predictive representations. Video JEPA models extend representation prediction to temporal observations (Assran et al., 2025). TD-JEPA studies latent-predictive representations for zero-shot reinforcement learning (Bagatella et al., 2026), and temporally centered SIGReg adapts LeWorldModel's regularization for robot policy learning (Liu et al., 2026). SP-JEPA adopts a shared-encoder JEPA objective and separately evaluates the physical meaning of state edits. Our EAWM-inspired comparison transfers event-tendency supervision to the same prediction backbone; it is an adaptation rather than a reproduction of EAWM's complete reinforcement-learning system.