Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Does a minimal latent world model know when drought breaks? A benchmark of drought- to-pluvial transitions over North America

Nicholas J. Sokol

nsokol@utk.edu

University of Tennessee

Research Article

Keywords: world models, drought, pluvial, compound events, forecast verification, SPEI

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11624732/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

We benchmark a minimal latent world model on drought-to-pluvial transitions over North America. Using SPEIbase v2.11 (SPEI-1, 1980 to 2023, 4,817 land cells), we evaluate 335,106 held-out drought origins under blocked cross-validation. The world model (WM1) is compared with persistence, climatology, a logistic model, gradient boosting, and a ConvLSTM on flip probability, footprint, withheld flips, and counterfactual ENSO forcing. The logistic model leads, with Brier skill of 0.015 to 0.022. The ConvLSTM also beats climatology but not the logistic model. WM1 scores minus 0.234 to minus 0.178 and is badly overconfident. Its median footprint is empty at lead 6. Among the forecast models WM1 times flips best, at 1.62 months on a common subset, but a climatological timing rule is better at 1.36 months, and a deterministic ablation shows that the timing is largely an ensemble-noise effect. The companion concurrent drought-pluvial event catalog (Sokol 2026) provides geographic context. The collapse is specific to this minimal latent model, and event-focused benchmarks expose failures that mean-state scores hide.

1 Introduction

Drought rarely ends quietly. Across much of North America, dry periods terminate in wet extremes, and the transition itself carries risk, because soils, reservoirs, flood defenses, and planting decisions that were tuned to scarcity are exposed to excess within a single season. Compound event research has established that such linked hazards are not well described by studying each hazard alone (Zscheischler et al. 2018), and the typology of compound events places events that follow one another in time, with the first conditioning the second, in their own class (Zscheischler et al. 2020). Concurrent drought and pluvial extremes are a spatial member of this family, in which a dry pole and a wet pole coexist in the same month. He and Sheffield (2020) documented a global lagged seesaw between droughts and pluvials over seven decades, and Swain et al. (2018) showed that precipitation whiplash in California is projected to rise through the twenty-first century. In their review of hydroclimate volatility, Swain et al. (2025) report that whiplash measured with the Standardized Precipitation Evapotranspiration Index (SPEI) has risen sharply since the middle of the twentieth century and they call explicitly for machine learning methods to study that volatility. Drought onset has its own rapid form, the flash drought (Otkin et al. 2018), and seasonal drought prediction has a mature literature of its own (Hao et al. 2018). The reverse transition, drought breaking into pluvial conditions, has so far been evaluated only with dynamical models. Cui et al. (2026) assessed the subseasonal forecast skill of dry-to-wet whiplash events over China, and Wang et al. (2026) evaluated ECMWF subseasonal reforecasts of the abrupt drought-to-flood transition in Henan Province in 2024, where the ensemble captured the transition at a one-week lead and skill dropped sharply at a two-week lead. No benchmark of learned models on these transitions exists. Zhao et al. (2025), reviewing 55 studies of drought and flood abrupt alternation, found a field concentrated on identification indices, with prediction by learned models absent.

A second literature has been building predictive models of the Earth system at remarkable speed. The benchmark tradition began with WeatherBench (Rasp et al. 2020) and continues with WeatherBench 2

(Rasp et al. 2024). GraphCast showed that a graph neural network can produce skillful medium-range forecasts (Lam et al. 2023), Pangu-Weather reached comparable skill with a three-dimensional transformer (Bi et al. 2023), and FourCastNet did so with a Fourier neural operator (Pathak et al. 2022). GenCast extended the frontier to probabilistic ensemble forecasts out to 15 days (Price et al. 2025). NeuralGCM couples learned components with a dynamical core and produces stable long rollouts (Kochkov et al. 2024), and Aurora shows that a single foundation model can be fine-tuned across several Earth system tasks (Bodnar et al. 2025). ClimateBench provides a complementary benchmark for emulating climate responses to forcing at annual scale (Watson-Parris et al. 2022). The headline scores in this literature are error and correlation measures on mean states, mostly at medium range. Extremes are covered only in part. Olivetti and Messori (2024) compared two learned models with a numerical model on hot, cold, and wind extremes and found that performance varies by region and lead time, and Ben-Bouallègue et al. (2024), in a first statistical assessment in an operational-like setting, also verified mean states. Sun et al. (2025) asked whether an artificial intelligence weather model can predict out-of-distribution Category 5 tropical cyclones that were removed from training, which is the closest design analogue to the withheld-flip test we run here.

The timescale mismatch is the core problem. Hydroclimatic regimes evolve over months, and the subseasonal to seasonal (S2S) range is where learned models are weakest. ChaosBench, a physics-based benchmark for S2S prediction, finds that weather-scale models collapse toward climatology at S2S range (Nathaniel et al. 2024). DroughtSet, a spatiotemporal benchmark for S2S drought prediction, predicts drought states but does not treat the transition into pluvial conditions as a task (Tan et al. 2024). We therefore do not know, from any published benchmark, whether learned models anticipate the flip from drought to pluvial conditions, whether their rollouts preserve the spatial footprint of each regime, or whether they can predict flips in regions and intensity ranges they have never seen.

A third literature supplies the architecture that our study puts on trial. World models, introduced by Ha and Schmidhuber (2018), learn a latent state-space representation of an environment, a transition model that can be rolled forward in imagination, and a decoder back to observations. Hafner et al. (2025) carried this design to a recurrent state-space model that masters more than 150 control tasks by training on imagined rollouts. LeCun (2022) argues, in a position paper, that predictive world models of this kind are a basis for machine intelligence. Reichstein et al. (2019) set an agenda for deep learning in Earth system science that calls for physically consistent evaluation of exactly this kind, and hydrology has its own evidence that recurrent networks learn long memory, from rainfall-runoff modeling across many catchments (Kratzert et al. 2018). No study in this tradition has been tested on a geophysical regime transition.

We close that gap with a benchmark of drought-to-pluvial transitions. The companion concurrent drought-pluvial event (CDPE) catalog (Sokol 2026) provides the geographic context, because it maps where dry and wet poles coexist across the continent. Our contributions are fourfold. First, we define regime states and transition targets on the same SPEI record that underlies the companion catalog, so the benchmark and the catalog describe the same phenomenon. Second, we score models with

transition-focused metrics, namely Brier skill and reliability for flip probability, the symmetric extremal dependence index for rare flips, fractions skill for regime footprints, and the continuous ranked probability score for the SPEI field. Third, we adapt the withheld-event design of Sun et al. (2025) to flips, withholding one region and the strongest droughts from training. Fourth, we add a counterfactual forcing test that separates a world model, which should respond to altered forcing in physically consistent ways, from a pattern-matching forecaster. Our research questions are the following. RQ1 (timing): can learned world models forecast the timing of drought to pluvial flips at 1 to 6 month leads better than persistence, climatology, and statistical baselines? RQ2 (footprint): do multi-step rollouts preserve the spatial extent and persistence of each regime, or do they smear toward climatology? RQ3 (generalization): can models predict flips in regions or intensity ranges withheld entirely from training? RQ4 (counterfactual consistency): when forcing is altered in the model, do rolled-out regime frequencies match observed teleconnection composites?

Our headline answer is negative, and we report it as such. The minimal latent world model we train is the worst probabilistic forecaster in the benchmark, beaten by climatology at every horizon, while a plain logistic regression leads the benchmark and a ConvLSTM is the only other model to beat climatology. The failure is specific rather than generic. A genuine spatiotemporal deep model does not collapse in the same way, so the collapse belongs to the minimal latent model and not to deep models in general. We test one minimal instance of the world model family, and our claims are scoped to that instance throughout. The benchmark separates two skills that mean-state evaluation conflates, and it isolates a specific failure, miscalibration with rollout collapse.

2 Data and event definitions

2.1 SPEI record

Our state variable is the one-month SPEI (SPEI-1) from SPEIbase, the multiscalar drought index of Vicente-Serrano et al. (2010), which combines precipitation with potential evapotranspiration and is therefore sensitive to the temperature-driven component of drying. The primary record is SPEIbase version 2.11, built on CRU TS 4.09, at 0.5 degree resolution. We use the window from January 1980 to December 2023, which gives 528 months. The domain is North America between 20 and 50 degrees north and 130 and 60 degrees west, a grid of 60 by 140 cells. We retain land cells with valid SPEI values, which gives 4,817 land cells. A sensitivity analysis replaces SPEI-1 with the three-month SPEI (SPEI-3) from SPEIbase version 2.10, built on CRU TS 4.08. We state that release difference wherever the sensitivity is reported, because the SPEI-3 file from version 2.11 was not in hand. Exact files, checksums, and releases for every input are recorded in the data provenance file that accompanies the code.

Two design choices follow from the companion studies in our CDPE program. First, we use SPEIbase rather than a reanalysis field, because the CDPE catalog of record (Sokol 2026) is detected on SPEIbase, and the North American Regional Reanalysis (Mesinger et al. 2006) and ERA5 (Hersbach et al. 2020) enter our wider program as further estimates of the same events rather than as the benchmark record.

Second, we end the window in 2023, the end of the SPEIbase record we hold for the sensitivity timescales. Our single forcing covariate is the Oceanic Niño Index (ONI), taken as a monthly series over the same window. Forcing enters the statistical models as a covariate and enters the world model as the exogenous input that plays the role that actions play in reinforcement learning.

2.2 Regime states and flip targets

We classify each land cell and month into one of three states from SPEI-1. A cell is in drought when SPEI is at most minus 1, in pluvial conditions when SPEI is at least plus 1, and in a normal state between those thresholds. The thresholds are the standard moderate-extreme cut points of the SPEI scale and they match the thresholds used in the companion catalog work (Sokol 2026), so the benchmark inherits a definition that has already been stress-tested rather than a new one tuned for this study.

The flip target defined in this section is what we call a drought-to-pluvial transition throughout the paper. An origin is a cell-month in drought. Over the full window we observe 415,302 drought origins. For each origin we record the first lead, from 1 to 6 months, at which the same cell reaches pluvial conditions. The flip target at horizon h is binary and equals 1 when the first pluvial month falls at or before h. Origins too close to the end of the record to be evaluated at horizon h are excluded from that horizon only. This definition makes the flip a property of the observed record alone. No model output enters the target, and the same target is scored for every model.

2.3 The companion CDPE catalog

The spatial counterpart of our temporal flip target is the CDPE catalog, which detects months in which a dry pole and a wet pole coexist over the domain. The catalog of record in our companion work (Sokol 2026) holds 52 qualifying months under the baseline detector settings on SPEIbase v2.11. An alignment exercise in the companion record (Sokol 2026), which we quote here rather than recompute, compared that catalog with the 46 months returned by the same detector on the earlier SPEIbase v2.10 release. The two sets share 32 months, which gives a Jaccard similarity of 0.485, far above the overlap expected by chance, and the difference between them is a release difference in the input fields rather than a window artifact. The 52 catalog months carry 104 poles. Of those poles, 61 fall inside the contiguous United States, 27 in Canada, and 16 in Mexico, so 58.7 percent of poles are anchored in the United States, and 15 of the 52 months have both poles inside the United States (Fig. 1). We use the catalog in two ways. It anchors the benchmark geographically, because the poles show where concurrent extremes actually occur, and it disciplines our claims, because a catalog that is sensitive to the SPEIbase release warns us that every score in this paper should be read as conditional on a stated data release. The catalog is context for this benchmark, and no catalog quantity is computed in this paper.

2.4 Blocked cross-validation

We evaluate with blocked cross-validation by year. The record is divided into five test blocks, 1988 to 1994, 1995 to 2001, 2002 to 2008, 2009 to 2015, and 2016 to 2023. For each block, models are trained on all remaining years with an embargo of 6 months on both sides of the block, so that no training origin

sits within the SPEI accumulation window of a test origin and no overlapping windows leak information across the split. The pooled test set holds 335,106 drought origins. Evaluated origins vary slightly by horizon, from 334,583 at h = 1 to 330,290 at h = 6, because origins near the end of the record cannot be evaluated at long horizons. Observed flips in the test set number 44,880 at h = 1, 84,142 at h = 2, 118,913 at h = 3, and 197,228 at h = 6, which correspond to base rates of 0.134, 0.252, 0.357, and 0.597. Flips are therefore rare at one month and common at six, and any useful model must be scored against that shifting base rate rather than against a fixed notion of accuracy. All randomness in the pipeline uses a single fixed seed, 20261103, and the full benchmark reruns end to end with one command.

3 Models

We compare seven forecasters, which span the range from no-skill references to the world model under test (Table 1). Every model is refitted inside every cross-validation fold, on training data only.

Table 1
Models compared in the benchmark. All models are refitted in each of the five blocked folds. Horizons are 1, 2, 3, and 6 months throughout.

Code	Model	Form	Output
B0	Persistence	Flip probability fixed at 0	Deterministic no-flip
B0s	Smoothed persistence	Flip probability fixed at 0.1 times the training base rate	Constant probability
B1	Climatology	Cell and calendar month flip frequency in training data, with a smoothing prior of 20 times the training base rate	Probability
B2	Logistic Markov model	Logistic regression on the nine covariates listed in the text, standardized	Probability
B3	Spatial gradient boosting	Histogram gradient boosting on the B2 covariates plus 3 by 3 and 5 by 5 neighborhood statistics	Probability
B4	ConvLSTM	Convolutional LSTM on the past 6 monthly SPEI-1 fields with land mask, calendar month, and ONI channels, hidden width 16	Probability
WM1	Minimal latent world model	Latent state-space model with a principal component encoder, a multilayer perceptron transition, and a linear decoder, rolled out as a 50 member ensemble	Ensemble probability

B0 is persistence expressed as a forecast of the flip event. Because the origin cell is in drought, persistence predicts that no flip occurs, so its flip probability is 0 at every horizon. B0s is a smoothed variant that assigns every origin a probability of 0.1 times the training base rate at that horizon. The pair separates the value of simply refusing to predict flips from the value of a small constant probability, and both are far weaker references than climatology, as Section 5 shows.

B1 is climatology, the reference against which all skill in this paper is measured. For each cell and calendar month, B1 is the flip frequency at that horizon in the training folds, shrunk toward the training base rate with a smoothing prior of 20 times the base rate. The shrinkage keeps sparse cell-months from producing degenerate probabilities of exactly 0 or 1.

B2 is a logistic regression that plays the role of a Markov regime model enriched with forcing. Its nine covariates are the current SPEI value, the three-month SPEI trend at the origin cell, the sine and cosine of calendar month, the current ONI value, the ONI value three months earlier, cell latitude, cell longitude, and the domain-mean SPEI of the origin month. Covariates are standardized and the model is fitted separately for each horizon. B2 is the strong statistical baseline that reviewers of any benchmark in this area should expect, because it uses the current regime state, its recent direction, the season, and the large-scale forcing, and nothing else.

B3 adds spatial context to B2. We use histogram gradient boosting with 150 iterations and a learning rate of 0.08. Its covariates are the nine B2 covariates plus, for 3 by 3 and 5 by 5 windows around the origin cell, the neighborhood mean SPEI, the neighborhood minimum SPEI, the neighborhood maximum SPEI, and the fraction of neighborhood land cells in drought. B3 is the flexible tabular baseline. The spatiotemporal deep learning baseline in the tradition of the ConvLSTM of Shi et al. (2015) is B4, which we describe next. Training for B3 is subsampled deterministically to at most 200,000 origins per fold and horizon, under the fixed seed, and B3 is evaluated on all test origins.

B4 is a ConvLSTM in the sense of Shi et al. (2015), implemented with a manually coded convolutional recurrent cell with 16 hidden channels and 3 by 3 kernels. For each origin month its inputs are the past 6 monthly SPEI-1 fields over the full domain, with missing values set to 0 and values clipped at plus or minus 5, plus four broadcast channels holding the land mask, the sine and cosine of calendar month, and the current ONI value. A 1 by 1 convolutional head maps the final hidden state to one logit per horizon, and the model is trained with binary cross-entropy on drought-origin cells only. B4 is trained on all training origins with no subsampling, for 10 epochs at a batch size of 16 origin months, with the Adam optimizer at a learning rate of 0.001 and a fold-specific seed derived from the fixed benchmark seed. B4 is a real spatiotemporal deep model, and it matters for scope. If B4 had collapsed alongside WM1, the negative result could have been read as a verdict on deep models in general. It did not, and Section 5 reports its scores alongside the other baselines.

WM1 is the minimal latent world model under test, and we call it that throughout. Formally, we learn an encoder q(z_t | x_t) that maps the observed SPEI field x_t to a latent state z_t, a forcing-conditioned transition p(z_{t + 1} | z_t, f_t) that advances the latent state under exogenous forcing f_t, and a decoder p(x_t | z_t) that maps the latent state back to the field. We implement the encoder as a principal component projection of the land-cell field onto 40 components fitted on training months, the transition as a multilayer perceptron with hidden layers of 128 and 64 units that takes the latent state together with the sine and cosine of calendar month and the current ONI value, held persistent, as its inputs, and the decoder as the inverse principal component transform. The transition is trained on consecutive training

months to predict the next latent state. Rollouts are probabilistic because, at each step, we add a residual drawn from the empirical distribution of training transition residuals, which gives an ensemble of 50 members per origin. Flip probabilities are ensemble fractions, footprints are ensemble medians, and counterfactual rollouts hold the ONI input fixed at a chosen value in place of the observed value. This is a deliberately modest member of the world model family of Ha and Schmidhuber (2018) and Hafner et al. (2025). It has the defining structure, a learned latent state, a forcing-conditioned transition that can be rolled forward in imagination, and a decoder, at a scale that trains on monthly fields without specialized hardware. Hydrological precedent for learned latent dynamics is strong, since recurrent networks have learned catchment memory across hundreds of basins (Kratzert et al. 2018), but no precedent covers regime transitions, which is the point of the test.

4 Tasks and metrics

We score four tasks, which map one to one onto our research questions.

T1, flip timing. For every test drought origin and every horizon in 1, 2, 3, and 6 months, each model issues a flip probability. We score probabilities with the Brier score and report the Brier skill score (BSS) against B1 climatology, so a positive value means the model beats the seasonal and spatial flip climatology and a negative value means it does worse. We draw reliability diagrams at h = 3, which bin origins by predicted probability and compare the mean prediction in each bin with the observed flip frequency. For the rare-event structure of the problem we also compute the symmetric extremal dependence index (SEDI) of Ferro and Stephenson (2011) on deterministic forecasts thresholded at a probability of 0.5. SEDI rewards models that discriminate rare flips without being rewarded for hedging, and it is the natural complement to the Brier score at h = 1, where the base rate is 0.134. Finally, we measure flip timing directly. For each model we take the origins where the model predicts a flip and a flip occurs, and we compute the mean absolute error, in months, between the predicted flip lead and the observed first pluvial lead. For B0 to B3 the predicted lead is the earliest horizon at which predicted probability reaches 0.5, and for WM1 it is the ensemble median flip lead. This measure is conditional by construction, and we interpret it only as a conditional timing measure among predicted flips, never as overall skill.

We attach uncertainty to the headline skill scores with a year-block bootstrap. We resample calendar years of test origins with replacement, keep each origin with its year, recompute every BSS against the resampled climatology Brier score, and repeat 1,000 times under the fixed seed. We report percentile 95 percent intervals for B2, B3, B4, and WM1 and for the B2 minus WM1 difference.

Because the conditional MAE compares models on different subsets, we also compute it on the common subset of origins where every model in B1, B2, B3, B4, and WM1 predicts a flip and a flip occurs. On the same subsets we score two naive timers that issue no probabilities at all. The first always predicts a flip at lead 1. The second is a climatology timer that predicts the median observed flip lead of training origins in the same calendar month, computed within each fold from origins that flipped within 6

months. Finally, we run a deterministic ablation of WM1 with the residual sampling switched off, so that each origin has a single conditional-mean trajectory. The deterministic flip probability at horizon h is 1 when that trajectory reaches pluvial conditions at or before h and 0 otherwise. The ablation separates what the learned conditional-mean dynamics carry from what the ensemble noise adds.

We also score the SPEI field itself with the continuous ranked probability score (CRPS), averaged over test origin months at leads 1, 3, and 6. WM1 contributes its 50 member ensemble. Persistence reduces to a mean absolute error, and the climatology ensemble is the set of training fields for the target calendar month. CRPS is our general check that the transition-focused scores are not an artifact of the binary target.

T2, footprint rollout. We roll each model forward 6 months from every test origin month and compare the predicted regime masks with the observed masks for drought and for pluvial conditions. Agreement is measured with the fractions skill score (FSS) of Roberts and Lean (2008) at neighborhood scales of 1, 3, 5, and 9 grid cells, pooled over origins. FSS is the right measure here because a footprint that is displaced by one cell should not be scored like a footprint that is absent, which is the failure we most want to detect. WM1 contributes its ensemble median field, B1 its climatological state probabilities thresholded at 0.5, and B0 the origin mask itself. Because a median field can hide member-level structure, we score T2 three ways for WM1, on the ensemble median field, on a probability field that thresholds the fraction of members in each state at 0.5, and as the mean of the per-member FSS values across the 50 members. We also score the deterministic ablation field at lead 6. We also compare regime persistence directly, by computing the observed three-by-three transition matrix between states at origin and at lead 6 and the corresponding matrix for each model, and reporting the Frobenius distance between model and observed matrices. Figure 2 previews the case view in a single origin month, pairing the observed flip outcome on the drought-origin cells with the WM1 and B4 flip probabilities on the same cells.

T3, withheld flips. We run two holdouts. In the spatial holdout, models are retrained with the Southern Plains box, from 25 to 37 degrees north and 106 to 93 degrees west, removed from training, and are tested only on origins inside that box. For B2 and B3 we exclude the box origins. For WM1 we replace the box cells in training fields with their training-period calendar month climatology, so that no information about flips in the box enters the encoder or the transition. In the intensity holdout, models are retrained without the strongest droughts. For B2 and B3 we exclude origins with SPEI at most minus 2, and for WM1 we winsorize training fields at SPEI equal to minus 2, so that no value below minus 2 enters training. The test sets are the withheld origins only. We report the BSS drop, which is the in-distribution BSS on the same origins minus the holdout BSS, with both scores referenced to the in-distribution B1 Brier score on those origins. A positive drop means the holdout hurt the model.

T4, counterfactual forcing. We hold the initial state of every test origin month fixed and run WM1 rollouts twice, once with the ONI input forced to plus 2.0 at every step and once with it forced to minus 2.0. At lead 3 we compute the domain fractions of land cells in drought and in pluvial conditions, averaged over

the 432 test origin months. The comparison targets are the observed composites over test years, namely the same area fractions in the 44 months with observed ONI at least plus 1 and in the 43 months with observed ONI at most minus 1. The composites are not lead-aligned to individual origins, and we report the comparison with that caveat attached. A world model that has learned the teleconnection should move its regime frequencies in the observed direction and by a comparable amount. A model that responds in the right direction but by far too little has learned the sign of the forcing and not its strength, a model whose warm-minus-cold contrast is present but whose regime fractions sit far below the observed levels has failed on levels rather than on the response, and a model that does not respond at all has learned a pattern matcher. T4 is designed to separate those cases.

5 Results

5.1 T1, flip timing: the logistic model leads and the ConvLSTM also beats climatology

Table 2 gives the headline scores. B2, the logistic Markov model, leads the benchmark, at 0.015 at h = 1, 0.021 at h = 2, 0.021 at h = 3, and 0.022 at h = 6. The gain is consistent rather than large. B2 has positive skill in every test block and at every horizon, which means the current SPEI value, its trend, the season, and ENSO forcing together carry real, if small, information about flips beyond climatology. The year-block bootstrap confirms that the gain is not a resampling artifact. The 95 percent interval for B2 runs from 0.009 to 0.022 at h = 1 and from 0.005 to 0.041 at h = 6, and it excludes zero at every horizon. The B2 minus WM1 gap is likewise significant at every horizon, at 0.249 with an interval of 0.210 to 0.289 at h = 1 and 0.223 with an interval of 0.133 to 0.325 at h = 6.

B4, the ConvLSTM, is the second model to beat climatology. Its BSS is 0.011 at h = 1, 0.012 at h = 2, 0.009 at h = 3, and 0.001 at h = 6, and it is positive in 17 of its 20 block and horizon cells. Its bootstrap intervals exclude zero at h = 1 and h = 2 and include zero at h = 3 and h = 6, so its skill is real at short leads and unresolved at long ones. B4 never beats B2, and its rare-event discrimination at h = 3, a SEDI of 0.029, sits below B2 at 0.085 and B3 at 0.053 and above WM1 at 0.014. A genuine spatiotemporal deep model therefore neither collapses like WM1 nor beats the logistic baseline. The negative result in this paper is specific to the minimal latent world model, and we scope it that way throughout.

Table 2 Task T1 flip timing. Brier skill score (BSS) against B1 climatology, pooled over the five test blocks, with Brier scores in the lower panel. Test drought origins total 335,106. Evaluated origins (n) and observed flips differ by horizon because origins near the end of the record cannot be evaluated at long horizons. SEDI is computed at h = 3 on forecasts thresholded at 0.5. Flip-month mean absolute error (MAE) is conditional on the model predicting a flip that occurs, with its own n in parentheses. B0 and B0s never reach the 0.5 threshold, so their MAE is undefined and their n is 0. The lowest panel gives year-block bootstrap 95 percent intervals for the BSS of the four probabilistic models, from 1,000 resamples of calendar years.

Model	BSS h = 1	BSS h = 2	BSS h = 3	BSS h = 6	SEDI h = 3	Flip-month MAE (n)
B0 persistence	-0.146	-0.328	-0.549	-1.488	0.000	NA (0)
B0s smoothed	-0.115	-0.262	-0.439	-1.195	0.000	NA (0)
B1 climatology	0.000	0.000	0.000	0.000	-0.000	2.89 (195,247)
B2 logistic	0.015	0.021	0.021	0.022	0.085	2.77 (180,742)
B3 gradient boosting	-0.036	-0.016	-0.016	-0.017	0.053	2.46 (165,414)
B4 ConvLSTM	0.011	0.012	0.009	0.001	0.029	2.84 (178,476)
WM1 world model	-0.234	-0.198	-0.178	-0.200	0.014	1.64 (198,572)

n and flips	h = 1	h = 2	h = 3	h = 6
Evaluated origins	334,583	333,554	333,379	330,290
Observed flips	44,880	84,142	118,913	197,228
Base rate	0.134	0.252	0.357	0.597

Brier score	h = 1	h = 2	h = 3	h = 6
B0	0.134	0.252	0.357	0.597
B0s	0.131	0.240	0.331	0.527
B1	0.117	0.190	0.230	0.240
B2	0.115	0.186	0.225	0.235
B3	0.121	0.193	0.234	0.244
B4	0.116	0.188	0.228	0.240
WM1	0.144	0.227	0.271	0.288

BSS 95 percent interval	h = 1	h = 2	h = 3	h = 6
B2	0.009 to 0.022	0.012 to 0.031	0.011 to 0.032	0.005 to 0.041
B3	minus 0.059 to minus 0.014	minus 0.045 to 0.011	minus 0.052 to 0.018	minus 0.060 to 0.022
B4	0.006 to 0.017	0.002 to 0.022	minus 0.004 to 0.020	minus 0.013 to 0.016
WM1	minus 0.272 to minus 0.195	minus 0.227 to minus 0.163	minus 0.220 to minus 0.131	minus 0.298 to minus 0.107

B3, the spatial gradient boosting model, falls below climatology at every horizon, with BSS from minus 0.036 at h = 1 to minus 0.017 at h = 6. Its neighborhood covariates do not rescue it. The extra flexibility fits structure in the training folds that does not transfer across year blocks, and at the deterministic threshold of 0.5 its rare-event discrimination is real but modest, with a SEDI at h = 3 of 0.053, below B2 at 0.085.

The two persistence references collapse as the horizon grows. By h = 6, where most droughts have flipped, refusing to predict flips gives BSS of minus 1.488 for B0 and minus 1.195 for B0s. Persistence is not a competitive reference for transition events, and a benchmark that reports skill against persistence alone would flatter every learned model in our set.

WM1 is the worst probabilistic model in the benchmark. Its BSS is minus 0.234 at h = 1, minus 0.198 at h = 2, minus 0.178 at h = 3, and minus 0.200 at h = 6, and it is negative in all five test blocks at every horizon (Figure 3, panel a). The reliability diagram at h = 3 shows why (Figure 3, panel b). Across most of its range, WM1 is underconfident at low probabilities and sharply overconfident at high ones. In its top bin, with a mean predicted probability of 0.963 over 3,345 origins, flips occur with a frequency of 0.355. The observed frequency stays near 0.35 to 0.40 across the whole upper half of the WM1 probability range, so its probabilities carry ranking information but almost no calibration. B2, by contrast, stays close to the diagonal, and in its top bin a mean prediction of 0.534 over 19,037 origins meets an observed frequency of 0.484. WM1 has the lowest SEDI of any discriminating model at h = 3, 0.014, below B3 at 0.053 and B2 at 0.085, so its deterministic forecasts at the 0.5 threshold also discriminate rare flips weakly.

The CRPS check on the SPEI field itself tells the same story in continuous form. WM1 scores 0.745 at lead 1, 0.590 at lead 3, and 0.579 at lead 6. That beats persistence, which scores 1.010, 1.081, and 1.093, by a wide margin, so the rollouts do track the field better than holding it fixed. It loses to the climatology ensemble at every lead, because climatology scores 0.577, 0.577, and 0.575. A model whose probabilistic field forecast cannot beat the distribution of past fields for the same calendar month has not learned usable transition dynamics, whatever its deterministic tracking shows.

One result cuts the other way, and it needs care. Conditional on predicting a flip that occurs, WM1 has the lowest flip-month error of the five forecast models. On its own subset its MAE is 1.64 months over 198,572 origins, and the ranking is not a subset artifact. On the common subset of 149,109 origins where every model predicts a flip and a flip occurs, WM1 scores 1.621 months against 2.428 for B3, 2.752 for B2, 2.856 for B4, and 2.902 for B1 (Table 3). Moving between the own and common subsets changes no model by more than 0.04 months.

Table 3 Flip-month timing. Mean absolute error in months on each model own predicted-flip subset and on the common subset of 149,109 origins where every forecast model predicts a flip and a flip occurs. The two naive timers issue no probabilities. The climatology timer predicts the training median flip lead for the calendar month. The deterministic WM1 row is the noise-free ablation, scored on the origins where it predicts a flip.

Model or timer	Own subset MAE (n)	Common subset MAE (n)
B1 climatology	2.89 (195,247)	2.90 (149,109)
B2 logistic	2.77 (180,742)	2.75 (149,109)
B3 gradient boosting	2.46 (165,414)	2.43 (149,109)
B4 ConvLSTM	2.84 (178,476)	2.86 (149,109)
WM1 world model	1.64 (198,572)	1.62 (149,109)
Always flip at lead 1	NA	2.09 (149,109)
Climatology timer	NA	1.36 (149,109)
WM1 deterministic	1.82 (13,487)	1.86 (10,760)

WM1 is not, however, the best flip timer in the room. A rule that always predicts a flip at lead 1 scores 2.089 months on the common subset, and WM1 beats it. A climatology timer that predicts the training median flip lead for the calendar month, which in practice predicts lead 3 in nearly every fold and month, scores 1.359 months on the common subset and 1.393 months on the WM1 subset. It beats WM1 on both. The honest statement is therefore narrow. Among the forecast models WM1 times flips best, and a training climatology timing rule times them better.

The deterministic ablation shows where the timing comes from, and it is not the learned dynamics. With the residual noise switched off, WM1 Brier skill falls to minus 0.389 at h = 1, minus 0.467 at h = 2, minus 0.616 at h = 3, and minus 1.421 at h = 6, far below the ensemble at every horizon. At h = 3 the deterministic trajectory crosses into pluvial conditions for only 6.5 percent of origins, and flips then occur 38.0 percent of the time against 35.5 percent when it does not cross, so the noise-free forecast is nearly non-discriminating and the ensemble overconfidence pattern disappears with the noise. The deterministic model predicts a flip for only 7.2 percent of the common-subset origins, and where it does, its timing error is 1.821 months on its own subset of 13,487 origins. The bulk of the ensemble probability

behavior, and of its 1.64 month timing, is therefore produced by residual noise pushing members across the plus 1 threshold rather than by learned conditional-mean dynamics. We read the timing result as an ensemble-noise effect, and we return to it in Section 6.

The SPEI-3 sensitivity, now run for every probabilistic model, points the same way. On SPEI-3 states, B2 scores 0.006 at h = 1, 0.014 at h = 3, and 0.020 at h = 6, B3 scores minus 0.537, minus 0.034, and minus 0.069, and WM1 scores minus 0.435, minus 0.179, and minus 0.161. The h = 3 value for B2 reproduces the fold-mean figure of 0.014 from the earlier B2-only sensitivity, on Brier scores of 0.1238 against 0.1255 (338,099 test origins). The h = 1 column needs its base rate attached. A three-month average almost never crosses from drought to pluvial in a single month, so the SPEI-3 flip base rate at h = 1 is only 0.0056, and at that base rate the Brier score punishes the overprediction of B3 and WM1 heavily and consistently across all five folds. We therefore read the SPEI-3 arm on its h = 3 and h = 6 columns, where the ordering matches the headline exactly. We note again that the sensitivity uses SPEIbase v2.10 while the headline uses v2.11.

The sustained-flip target tells the same story on the primary record. When a flip requires pluvial conditions in two consecutive months, which leaves 28,052 sustained flips by h = 3 and 55,357 by h = 6, B2 scores 0.010 at h = 3 and 0.011 at h = 6, B3 scores minus 0.038 and minus 0.043, and WM1 scores minus 0.065 and minus 0.071. The model ordering is unchanged, so the headline ranking is not an artifact of the single-month flip definition. The WM1 deficit shrinks in absolute terms only because the sustained target is rarer.

5.2 T2, footprint rollout: the world model footprint is empty

At lead 6, the WM1 ensemble median field produces a fractions skill score of exactly 0.000 for the drought mask and exactly 0.000 for the pluvial mask, at every neighborhood scale from 1 to 9 cells (Figure 4). The median field sits in the normal band so completely that thresholding it at the regime cut points recovers no footprint at all. Persistence, which simply keeps the origin mask, scores 0.232 for drought and 0.242 for pluvial at scale 5, rising to 0.260 and 0.271 at scale 9. Climatology thresholded at 0.5 scores 0.019 for drought and 0.000 for pluvial at scale 5. The ordering is damning for rollout-based use of the world model. A forecast that carries the origin footprint forward unchanged preserves more of the regime geometry at six months than the learned rollout does.

The zero is a property of the field summary, and we state its limits plainly. A probability field that thresholds the fraction of members in each state at 0.5 also scores 0.000 at every scale for both masks, so the median and the majority vote collapse in the same way. Scoring the members one at a time gives a different picture. The mean per-member FSS is 0.115, 0.138, 0.151, and 0.172 for drought and 0.140, 0.160, 0.173, and 0.196 for pluvial conditions at scales 1, 3, 5, and 9, with a spread across members of only about 0.007 to 0.009, so individual members carry real footprint signal and the value is stable across the ensemble. It still loses to persistence at every scale, because persistence scores 0.183 to 0.260 for drought and 0.200 to 0.271 for pluvial conditions over the same scales. The deterministic

ablation field scores about 0.001, so removing the noise does not rescue the footprint either. The corrected statement is that every central-tendency summary of the rollout has an empty footprint, member-level scoring recovers modest signal, and even that signal stays below doing nothing.

Figure 2 shows the case behind the pooled scores. All three panels use the 2,207 drought-origin cells of the origin month of May 1992 and nothing else. In the observed record, 1,936 of those cells flipped to pluvial conditions within 6 months, and the cells that did not flip sit mostly on the West Coast and in the far northeast. The WM1 flip probability at lead 6, the share of the 50 ensemble members that flip in each cell, averages 0.51 over the case cells, with a case Brier score of 0.250 against the observed outcome. The third panel of Figure 2 shows B4 for the same cells on the same scale. Its flip probability averages 0.68, with a case Brier score of 0.148, so in this case the ConvLSTM probability sits closer to the realized flip share of 0.88. The median field of the same WM1 ensemble is normal in every one of these cells, which is the case-level view of the empty footprint that Figure 4 reports pooled over all origin months. The case separates the two statements. As a probability, the world model is low but graded. As a footprint, its median is empty.

The transition matrices quantify the same shrinkage. In the observed record, a cell in drought at the origin is in drought at lead 6 with probability 0.182, in a normal state with probability 0.643, and in pluvial conditions with probability 0.175. The WM1 rollout compresses those probabilities toward the normal state, giving 0.088, 0.812, and 0.101 for the same row, and it overstates normal-state persistence as well, at 0.815 against an observed 0.661. The Frobenius distance between the model and observed transition matrices is 0.346 for WM1, against 0.034 for B1 climatology and 1.538 for the persistence identity. Climatology reproduces the average transition structure ten times more faithfully than the world model rollout, because the world model replaces real transition variability with a damped move to the center.

5.3 T3, withheld flips: the logistic model generalizes, the world model does not

Table 4 reports the holdout results. In the spatial holdout, B2 loses nothing. Its BSS drop is minus 0.002 at h = 1, minus 0.005 at h = 3, and minus 0.005 at h = 6 over 42,223 holdout origins at h = 1 and h = 3 and 41,490 at h = 6, and its holdout BSS stays positive at 0.042 at h = 6. A linear model in the current state, trend, season, and forcing transfers to an unseen region because the relationships it encodes are not regional. B3 also shows no degradation in the spatial holdout, with drops of minus 0.007, minus 0.058, and minus 0.024, though its holdout skill remains at or below climatology at h = 1.

WM1 fails the spatial holdout badly. Its drop is minus 0.070 at h = 1, but it rises to 0.300 at h = 3 and to 1.087 at h = 6, where its holdout BSS reaches minus 1.189 against an in-distribution value of minus 0.102 on the same origins. The latent representation has bound the dynamics of the Southern Plains to patterns learned from the rest of the continent, and when that region is removed from training the rollout for the region is not just weak but actively misleading at long leads.

The intensity holdout is milder for every model (Table 4). Over 20,625 holdout origins at h = 1 and h = 3 and 20,129 at h = 6, B2 drops by minus 0.002, minus 0.001, and 0.003, and B3 by minus 0.007, 0.003, and 0.022. WM1 is mixed, with a drop of 0.101 at h = 1 and negative drops of minus 0.040 at h = 3 and minus 0.066 at h = 6, which means its scores on the strongest droughts are no worse, and at longer horizons slightly better, than its in-distribution scores on the same origins. Withholding intensity is a weaker test than withholding geography for this model class, because the strongest droughts still occur inside regions whose ordinary dynamics were trained. The contrast between the two holdouts is itself a finding. What does not transfer is place.

Table 4 Task T3 withheld flips. Holdout Brier skill score (BSS) is referenced to the in-distribution B1 Brier score on the same origins. The drop is the in-distribution BSS minus the holdout BSS, so a positive drop means the holdout degraded the model. The spatial holdout removes the Southern Plains box (25 to 37 degrees north, 106 to 93 degrees west) from training. The intensity holdout removes origins, or field values, at SPEI of minus 2 or below from training.

Holdout	Model	h = 1 holdout BSS	h = 1 drop	h = 3 holdout BSS	h = 3 drop	h = 6 holdout BSS	h = 6 drop	n holdout (h = 1, 3 / h = 6)
Spatial	B2	0.015	-0.002	0.042	-0.005	0.042	-0.005	42,223 / 41,490
Spatial	B3	-0.033	-0.007	0.035	-0.058	0.000	-0.024	42,223 / 41,490
Spatial	WM1	-0.116	-0.070	-0.421	0.300	-1.189	1.087	42,223 / 41,490
Intensity	B2	0.030	-0.002	0.048	-0.001	0.049	0.003	20,625 / 20,129
Intensity	B3	-0.032	-0.007	0.025	0.003	0.006	0.022	20,625 / 20,129
Intensity	WM1	-0.334	0.101	-0.142	-0.040	-0.087	-0.066	20,625 / 20,129

5.4 T4, counterfactual forcing: right direction, wrong levels

The observed composites in test years are strong. In the 44 months with ONI of at least plus 1, the drought area fraction is 0.133 and the pluvial area fraction is 0.222. In the 43 months with ONI of at most minus 1, the drought fraction is 0.204 and the pluvial fraction is 0.157. Warm ENSO phases shift the continent toward pluvial conditions and cold phases shift it toward drought, in the observed record, by about 7 percentage points on each side.

WM1 responds to forced ONI in the observed direction (Table 5). With ONI forced to plus 2.0, its lead 3 fractions are 0.070 drought and 0.140 pluvial. With ONI forced to minus 2.0, they are 0.116 drought and

0.093 pluvial. Warm forcing lowers drought and raises pluvial output relative to cold forcing, as it should. The warm-minus-cold contrast is also substantial. Forced warm minus forced cold is minus 0.046 on drought and plus 0.047 on pluvial, against observed contrasts of minus 0.071 and plus 0.065, so the model produces about two thirds of the observed drought contrast and about seven tenths of the observed pluvial contrast. It does so under stronger forcing than the composites carry, because the forced runs hold the ONI input at plus or minus 2.0 while the composites collect months with observed ONI beyond plus or minus 1. The failure is in the levels, not in the contrast. Against the observed warm composite, the forced warm run is low by 0.062 on drought and by 0.082 on pluvial. Against the observed cold composite, the forced cold run is low by 0.088 on drought and by 0.064 on pluvial. Both forced runs sit far below both observed composites on both fractions, at roughly half to three fifths of the composite levels, because the rollout collapse of Section 5.2 drains probability mass out of both regime tails and into the normal state, whatever the forcing says. The model has learned the direction of the teleconnection and a fair share of its contrast, but it cannot hold the regime fractions at their observed levels. Under our design, that is a failed counterfactual test, and it is the same failure as the footprint test viewed from the forcing side.

Table 5 Task T4 counterfactual forcing at lead 3. Area fractions are shares of the 4,817 land cells in each regime state. Observed composites use test-year months grouped by observed ONI. Forced runs hold the ONI input to WM1 at plus 2.0 or minus 2.0 for all steps, from 432 test origin months. Differences are forced minus observed for the matching composite.

Series	Drought fraction	Pluvial fraction	Basis
Observed, ONI at least +1	0.133	0.222	44 months
Observed, ONI at most -1	0.204	0.157	43 months
WM1 forced ONI = +2.0	0.070	0.140	432 origins
WM1 forced ONI = -2.0	0.116	0.093	432 origins
Difference, +2.0 minus warm composite	-0.062	-0.082	
Difference, -2.0 minus cold composite	-0.088	-0.064	

6 Discussion

6.1 What failed, and what did not

Our benchmark was designed so that a negative result would still be diagnostic, and the result is negative in a specific way. The minimal latent world model fails on probability, on footprint, on regional generalization, and on counterfactual levels. Its conditional timing is the lowest among the forecast models, and Section 5.1 shows that even that is largely an ensemble-noise effect that a climatology timer beats. Those are separable skills, and the separation is the finding. A model can appear to time

regime change, because residual noise crossing a fixed threshold inherits the seasonal rhythm of the record, and still fail to attach correct probabilities to the events that matter, because its training objective never asked it to. The deterministic ablation is what lets us say this. The learned conditional mean on its own is nearly non-discriminating, so the timing belongs to the noise and the threshold, not to learned dynamics. One-step latent prediction error, which is what our transition was trained on, rewards a model for placing its next state near the conditional mean of the next field. At monthly scale the conditional mean of a regime field is a damped, central field, so the objective itself manufactures the collapse that T2 and T4 then detect. We state this as interpretation of a measured pattern, not as a new result. The pattern is measured. The collapse appears in the empty median footprint, in the transition matrix compressed toward the normal state, in CRPS values that lose to climatology while beating persistence, and in forced runs that undershoot both observed ENSO composites on both tails at once. Four different scores fail in the same direction, which is what a single mechanism, shrinkage toward the conditional mean under iterated rollout, would produce.

The comparison with the simple models disciplines any temptation to blame the task. Flips are predictable, a little. B2 beats climatology in every block and at every horizon on two SPEI timescales and two SPEIbase releases, with covariates that any forecaster would reach for first. The information is in the current state, its trend, the season, and ENSO. What the flexible models add on top of that information is mixed. B3 shows that adding flexible spatial features subtracts skill. B4 shows that a learned spatial model can beat climatology, though it never beats the logistic model. WM1 is the one that fails, and the contrast between B4 and WM1 is what scopes our claim. The failure belongs to the minimal latent world model, not to learned models as a class. The S2S literature predicts the direction of this result. Nathaniel et al. (2024) found that weather-scale models collapse toward climatology at S2S range on a physics-based benchmark, and our result extends that finding from general S2S fields to hydroclimatic regime transitions, with the added detail that the collapse is visible in the transition matrix and in the counterfactual response, not only in headline scores. The operational assessment of Ben-Bouallègue et al. (2024) and the extremes comparison of Olivetti and Messori (2024) both verified learned models on mean states or on single hazards, and both would have missed our failure mode, because a model can hold a respectable mean state while its regime tails are empty. Tan et al. (2024) built the drought-side benchmark, and their design stops at the drought state. Our benchmark is the other half of that design, and the two together suggest that state prediction and transition prediction are different tasks that need different training objectives.

6.2 The open question

Whether differently built learned models can pass this benchmark remains an open question, and the benchmark, its event definitions, and its per-origin predictions are released so that any such model can be scored on the same origins.

6.3 Limitations

We state five limits plainly. First, our world model is one modest architecture, a principal component encoder with a perceptron transition, and it is not a fine-tuned foundation model. Aurora (Bodnar et al. 2025), NeuralGCM (Kochkov et al. 2024), GenCast (Price et al. 2025), and the medium-range models of Lam et al. (2023) and Bi et al. (2023) solve harder problems at finer scales, and our result does not measure them. It measures what the defining world model structure achieves on this task when trained on this record, and it sets the bar that a larger model must clear on the same origins before claiming regime skill. The gray swan design of Sun et al. (2025) makes the same point for cyclones, that out-of-distribution claims need withheld-event tests rather than reputation. Second, WM1 is the only latent world model we test, and it is a single minimal instance of the family. The benchmark now contains a genuine ConvLSTM baseline, and B4 does not collapse, so the spatiotemporal baseline family is no longer represented by a substitute. What remains untested is the rest of the world model family itself, including recurrent state-space models at the scale of Hafner et al. (2025). Our conclusions are therefore scoped to minimal latent world models, and the title and abstract carry that scope deliberately. Third, our states come from one index and one threshold pair. The companion sensitivity work (Sokol 2026) shows that event counts move with definition, and while our headline ordering survives the SPEI-3 sensitivity and the sustained-flip target, a benchmark on streamflow states or on impact-based states could order models differently. Fourth, the catalog agreement we report is quoted from the companion alignment record and not recomputed here, and the release difference between SPEIbase v2.11 and v2.10 moves the catalog itself, so all of our numbers are conditional on stated releases. Fifth, reanalysis and gridded records such as NARR (Mesinger et al. 2006) and ERA5 (Hersbach et al. 2020) would give a second estimate of every score, and we have not run the benchmark on them. Observational uncertainty is therefore represented in our paper by a release contrast, not by a full second benchmark. The T4 comparison targets are composites over test-year months and are not lead-aligned to origins, so T4 measures response magnitude against a climatological yardstick, not event-level counterfactual accuracy.

7 Conclusions

We asked whether a minimal latent world model knows when drought breaks, and on observed North American data from 1980 to 2023 the answer is that it does not know the odds, and its apparent sense of timing is mostly noise. A forcing-conditioned minimal latent world model, rolled out as an ensemble, produced flip probabilities that lose to climatology at every horizon from 1 to 6 months, a lead 6 regime footprint that is empty under every central-tendency summary, a spatial generalization failure in the withheld Southern Plains, and counterfactual ENSO responses that move in the right direction, with a warm-minus-cold contrast at roughly two thirds of the observed contrast, while the regime fractions themselves sit at about half the composite levels. Two models beat climatology. A logistic regression on the current state, its trend, the season, and ENSO gained about 2 percent, and a ConvLSTM gained about 1 percent, so the collapse is specific to the minimal latent model and not a property of deep models or of the task. Among the forecast models the world model timed flips best, at 1.62 months on a common subset, but a climatological timing rule scored 1.36 months and a noise-free run of the same model is

nearly non-discriminating, so we report the timing as an ensemble-noise effect rather than learned skill. The failure is therefore one of probability and footprint in a minimal latent world model. We release the benchmark, the event definitions, and the per-origin predictions so that any future world model can be scored on the same 335,106 drought origins.

Declarations

Funding The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests The author has no relevant financial or non-financial interests to disclose.

Author contributions Nicholas J. Sokol conceived and designed the study, assembled the data, wrote the analysis code, performed the analysis, and wrote the manuscript.

Data availability SPEIbase is a third-party dataset available from its producers (Vicente-Serrano et al. 2010) and is not redistributed here. The Oceanic Niño Index series is a public product of the National Oceanic and Atmospheric Administration. The event definitions and the per-origin predictions generated in this study are available at github.com/Nsokol44/cdpe-worldbench.

Code availability The code that builds the benchmark is available at github.com/Nsokol44/cdpe-worldbench.

References

1. Ben-Bouallègue Z et al (2024) The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in an operational-like context. Bull Am Meteorol Soc 105:E864–E883
2. Bi K, Xie L, Zhang H, Chen X, Gu X, Tian Q (2023) Accurate medium-range global weather forecasting with 3D neural networks. Nature 619:533–538
3. Bodnar C, Bruinsma WP, Lucic A et al (2025) A foundation model for the Earth system. Nature 641:1180–1187
4. Cui B, Hsu PC, Lei L, Xie J, Jiang Q (2026) Dynamical controls on subseasonal forecast skill of dry-to-wet whiplash events over China. Atmospheric and Oceanic Science Letters 100856
5. Ferro CAT, Stephenson DB (2011) Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. Weather Forecast 26:699–713
6. Ha D, Schmidhuber J (2018) World models. arXiv:1803.10122
7. Hafner D, Pasukonis J, Ba J, Lillicrap T (2025) Mastering diverse control tasks through world models. Nature 640:647–653
8. Hao Z, Singh VP, Xia Y (2018) Seasonal drought prediction: Advances, challenges, and future prospects. Rev Geophys 56:108–141

9. He X, Sheffield J (2020) Lagged compound occurrence of droughts and pluvials globally over the past seven decades. Geophys Res Lett 47:e2020GL087924
10. Hersbach H, Bell B, Berrisford P et al (2020) The ERA5 global reanalysis. Q J R Meteorol Soc 146:1999–2049
11. Kochkov D, Yuval J, Langmore I et al (2024) Neural general circulation models for weather and climate. Nature 632:1060–1066
12. Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M (2018) Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol Earth Syst Sci 22:6005–6022
13. Lam R, Sanchez-Gonzalez A, Willson M et al (2023) Learning skillful medium-range global weather forecasting. Science 382:1416–1421
14. LeCun Y (2022) A path towards autonomous machine intelligence. OpenReview preprint, version 0.9.2
15. Mesinger F, DiMego G, Kalnay E et al (2006) North American Regional Reanalysis. Bull Am Meteorol Soc 87:343–360
16. Nathaniel J, Qu Y, Nguyen T, Yu S, Busecke J, Grover A, Gentine P (2024) ChaosBench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. NeurIPS 2024 Datasets and Benchmarks Track. arXiv:2402.00712
17. Olivetti L, Messori G (2024) Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather, and GraphCast. Geosci Model Dev 17:7915–7962
18. Otkin JA, Svoboda M, Hunt ED et al (2018) Flash droughts: A review and assessment of the challenges imposed by rapid-onset droughts in the United States. Bull Am Meteorol Soc 99:911–919
19. Pathak J, Subramanian S, Harrington P et al (2022) FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv:2202.11214
20. Price I, Sanchez-Gonzalez A, Alet F et al (2025) Probabilistic weather forecasting with machine learning. Nature 637:84–90
21. Rasp S, Dueben PD, Scher S, Weyn JA, Mouatadid S, Thuerey N (2020) WeatherBench: A benchmark data set for data-driven weather forecasting. J Adv Model Earth Syst 12:e2020MS002203
22. Rasp S, Hoyer S, Merose A et al (2024) WeatherBench 2: A benchmark for the next generation of data-driven global weather models. J Adv Model Earth Syst 16:e2023MS004019
23. Reichstein M, Camps-Valls G, Stevens B et al (2019) Deep learning and process understanding for data-driven Earth system science. Nature 566:195–204
24. Roberts NM, Lean HW (2008) Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Mon Weather Rev 136:78–97
25. Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-C (2015) Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Adv Neural Inf Process Syst 28

26. Sokol NJ (2026) A computational algorithm to identify concurrent drought-precipitation events. Manuscript submitted to Physical Geography
27. Sun YQ, Hassanzadeh P, Zand M, Chattopadhyay A, Weare J, Abbot DS (2025) Can AI weather models predict out-of-distribution gray swan tropical cyclones? Proceedings of the National Academy of Sciences 122:e2420914122
28. Swain DL, Langenbrunner B, Neelin JD, Hall A (2018) Increasing precipitation volatility in twenty-first-century California. Nat Clim Change 8:427–433
29. Swain DL, Prein AF, Abatzoglou JT et al (2025) Hydroclimate volatility on a warming Earth. Nat Reviews Earth Environ 6:35–50
30. Tan X, Zhao Q, Liu Y, Zhang X (2024) DroughtSet: Understanding drought through spatial-temporal learning. arXiv:2412.15075
31. Vicente-Serrano SM, Beguería S, López-Moreno JI (2010) A multiscalar drought index sensitive to global warming: The Standardized Precipitation Evapotranspiration Index. J Clim 23:1696–1718
32. Wang Y, Yuan X, Zhou S (2026) Subseasonal ensemble prediction of the 2024 abrupt drought-to-flood transition in Henan Province. China Water 18:635
33. Watson-Parris D, Rao Y, Olivié D et al (2022) ClimateBench v1.0: A benchmark for data-driven climate projections. J Adv Model Earth Syst 14:e2021MS002954
34. Zhao L, Aiyub K, Ata FM, Lam KC, Sun M (2025) A systematic review of methods for identifying drought-flood abrupt alternation: Advances and future directions. Front Environ Sci 13:1590613
35. Zscheischler J, Westra S, van den Hurk BJJM et al (2018) Future climate risk from compound events. Nat Clim Change 8:469–477
36. Zscheischler J, Martius O, Westra S et al (2020) A typology of compound weather and climate events. Nat Reviews Earth Environ 1:333–347

Figures

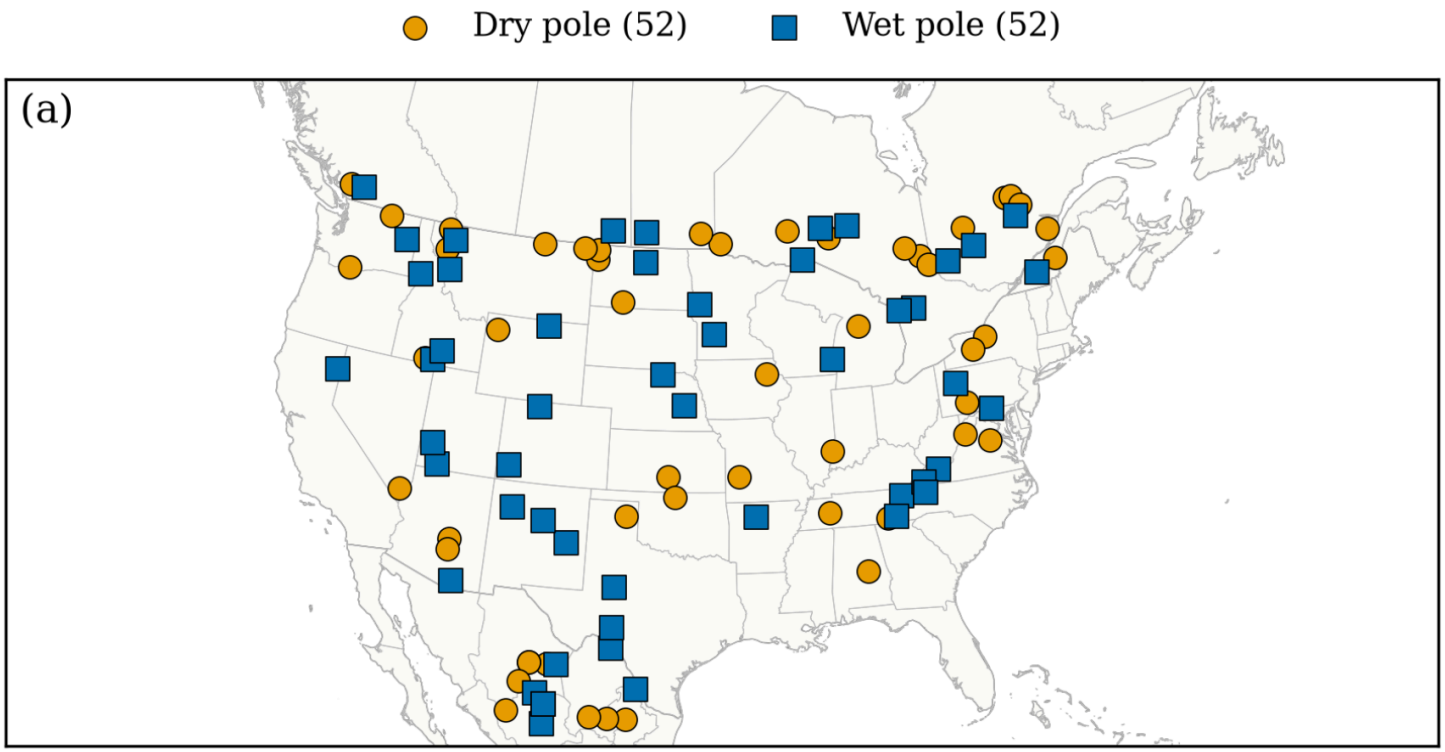

Figure 1

Poles of the companion concurrent drought-pluvial event catalog. The map shows the dry pole (gold circles) and the wet pole (blue squares) of each of the 52 catalog months detected on SPEIbase v2.11 at baseline settings over 1980 to 2023, for a total of 104 poles. Of the 104 poles, 61 fall in the contiguous United States, 27 in Canada, and 16 in Mexico. Read the map as the geography of concurrency itself. Most months pair a dry pole and a wet pole in different subregions of the continent, and 15 of the 52 months have both poles inside the contiguous United States. Projection is North America Albers. The catalog is the companion record of Sokol (2026), quoted here as motivating context and not computed in this paper

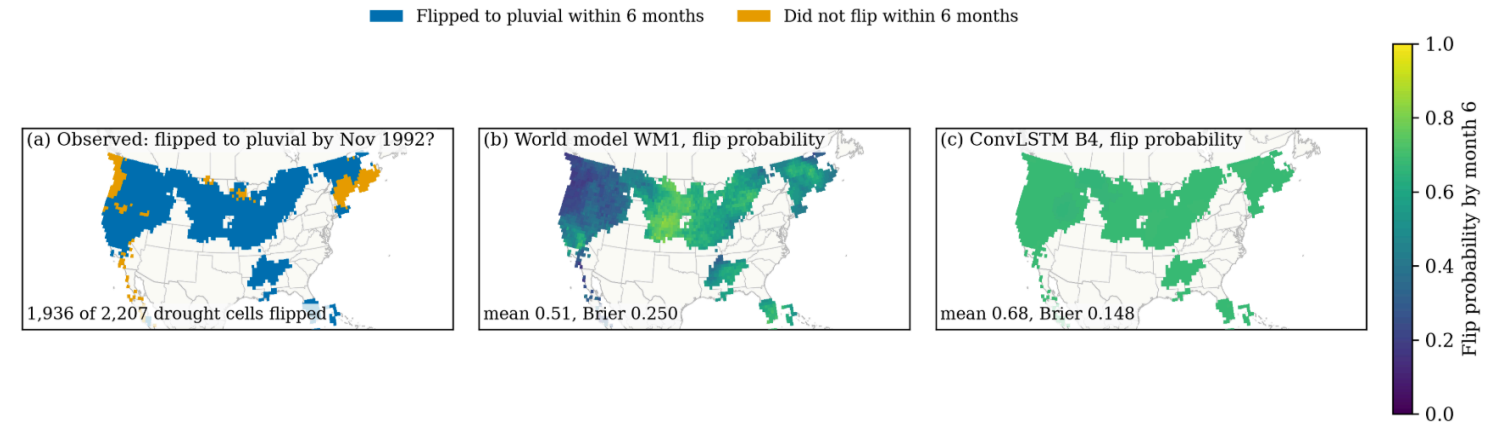

Figure 2

Flip outcome and flip probabilities at lead 6 for the origin month of May 1992, on one shared set of cells. All three panels show the 2,207 drought-origin cells of that origin month and nothing else. Panel (a) shows the observed outcome on those cells, with cells that flipped to pluvial conditions within 6 months in blue and cells that did not flip in gold. In total, 1,936 of the 2,207 drought cells flipped. Panel (b) shows the WM1 flip probability at lead 6 on the same cells, the share of the 50 ensemble members that flip in each cell, with a mean of 0.51 and a case Brier score of 0.250. Panel (c) shows the B4 ConvLSTM flip probability on the same cells, with a mean of 0.68 and a case Brier score of 0.148. Panels (b) and (c) share one probability scale from 0 to 1. Read the three panels as the same cells seen three ways. The observed outcome is mostly flip, the ConvLSTM probability is high and broad, and the world model probability averages near one half.

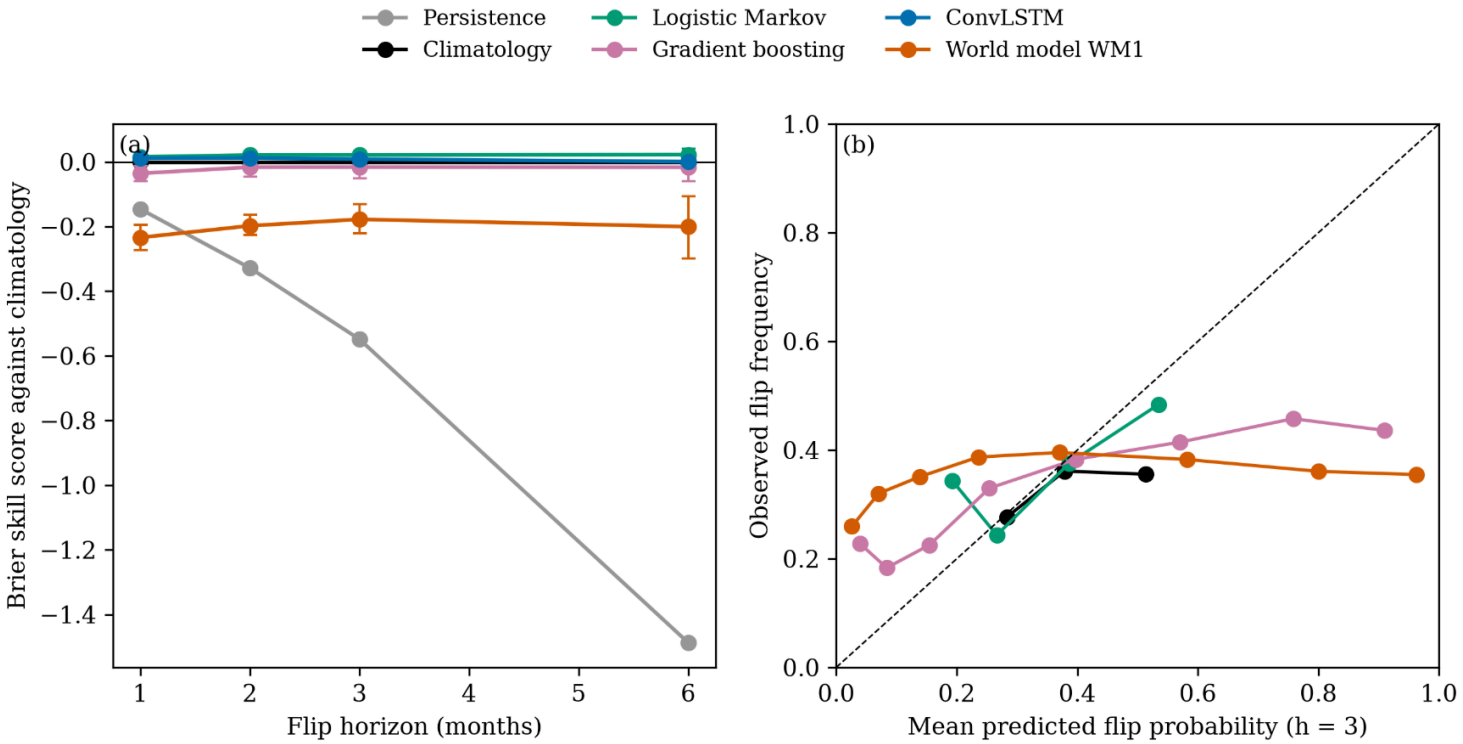

Figure 3

Flip probability skill and calibration. Panel (a) shows the Brier skill score against B1 climatology for persistence (B0), climatology (B1, zero by construction), the logistic Markov model (B2), gradient boosting (B3), the ConvLSTM (B4), and the minimal latent world model (WM1), pooled over the five test blocks at horizons of 1, 2, 3, and 6 months. Whiskers on B2, B3, B4, and WM1 are year-block bootstrap 95 percent intervals from 1,000 resamples of calendar years. Evaluated origins number 334,583 at h = 1 and 330,290 at h = 6, with observed flips of 44,880 and 197,228. Panel (b) shows reliability at h = 3, plotting the observed flip frequency against the mean predicted probability within probability bins, over 333,379 origins with 118,913 observed flips. The dashed diagonal is perfect calibration. Read the WM1 curve in panel (b) from left to right. Observed frequency stays near 0.35 to 0.40 while WM1 mean predictions rise to 0.96, which is the overconfidence that drives its negative skill in panel (a)

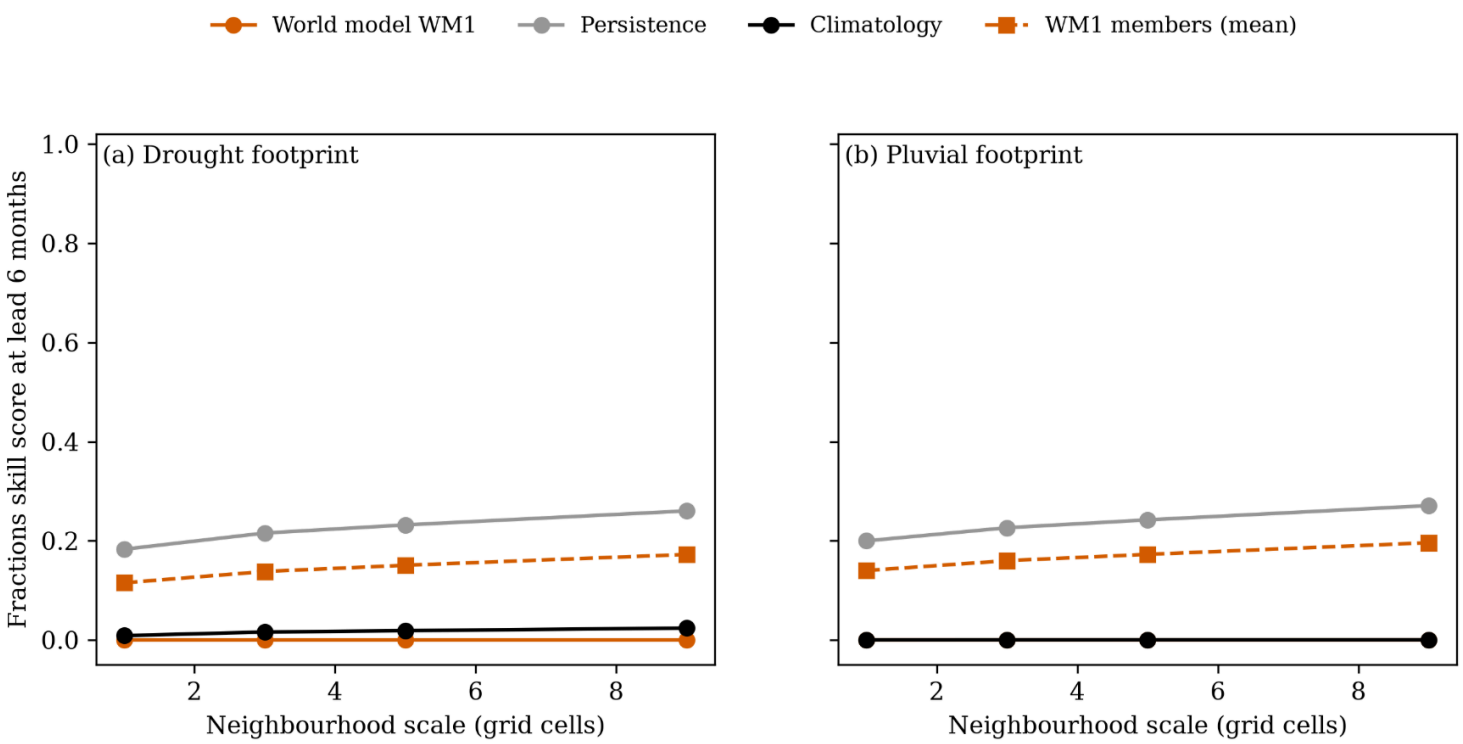

Figure 4

Fractions skill score (FSS) for regime footprints at lead 6. Panel (a) scores the drought mask and panel (b) scores the pluvial mask, each pooled over all test origin months, at neighborhood scales of 1, 3, 5, and 9 grid cells of 0.5 degrees. The world model (WM1, orange) is the ensemble median field, the dashed orange line is the mean per-member WM1 score, persistence (gray) is the origin mask carried forward, and climatology (black) is thresholded climatological state probability. One is a perfect score and zero is no skill. Read the orange line on the axis in both panels as an empty predicted footprint, and the gray line as the score a user gets by assuming that the regime map does not change at all