EVALUATING PREDICTOR-CORRECTOR DEPENDENCE IN HARD-CONSTRAINED CLIMATE EMULATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Climate emulators often enforce physical constraints by correcting each state predicted by a neural network, but evaluating only the corrected forecast conflates predictor skill with the effect of the correction. We introduce an evaluation framework that separates these contributions by comparing corrected forecasts, pre-correction predictions along corrected trajectories, and autonomous rollouts with correction disabled. We further cross correction during training and inference to distinguish its immediate effect from how it shapes the learned predictor, and use matched nonphysical controls to test whether improvements are specific to the physical correction. In ACE2, correction substantially improves physical consistency. In a three-seed ACE2 pilot, models trained with correction produce less accurate uncorrected precipitation forecasts at short horizons than models trained without correction, but more accurate forecasts at 10 and 30 days. This contradicts our preregistered hypothesis that the deficit would grow with forecast horizon. In a separate nine-seed experiment, directly supervising the uncorrected output reverses its precipitation forecast deficit at 72 hours while maintaining the accuracy of corrected forecasts at 24 hours. In ConCerNet's heat-equation system, projection-trained models have higher uncorrected trajectory-based vector-field RMSE than models trained without projection, while restoring the physical projection recovers accuracy far more effectively than a matched nonphysical control. Together, these results show that hard corrections can reshape the predictors trained alongside them, creating dependencies that are not visible from corrected forecasts alone. Hard-constrained emulators should therefore be evaluated by separately measuring predictor accuracy, correction effects, and their interaction during training. Code available at `https://anonymous.4open.science/r/ICLR-Corrector`.

1 INTRODUCTION

Learned weather and climate models increasingly combine a neural predictor with a deterministic operation that enforces conservation, positivity, diagnostic closure, or prescribed boundary conditions (Beucler et al., 2021; Watt-Meyer et al., 2025). These operations can improve a forecast while concealing errors in the state predicted by the network (Chapman et al., 2026). Evaluation of the corrected output alone cannot distinguish an accurate learned transition from an inaccurate transition whose errors are subsequently removed.

Training may improve the predictor itself; the corrector may repair a largely unchanged prediction; or the predictor may adapt to a correction available at every step (Um et al., 2020; Chapman et al., 2026).

Our proposed evaluation uses three complementary comparisons. Matched controls test whether the physical correction contributes more than an equal computation budget or a nonphysical perturbation. We then evaluate network outputs immediately before correction along corrected trajectories. Finally, we run models trained with and without correction in both inference modes, keeping each model's learned weights fixed. Disabling correction throughout forecasting makes the predictor repeatedly receive its own uncorrected outputs, exposing the feedback-dependent input distributions studied in sequential prediction (Ross et al., 2011). Figure 1 outlines evaluation (a,b), findings (c), and raw-output supervision (d).

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.

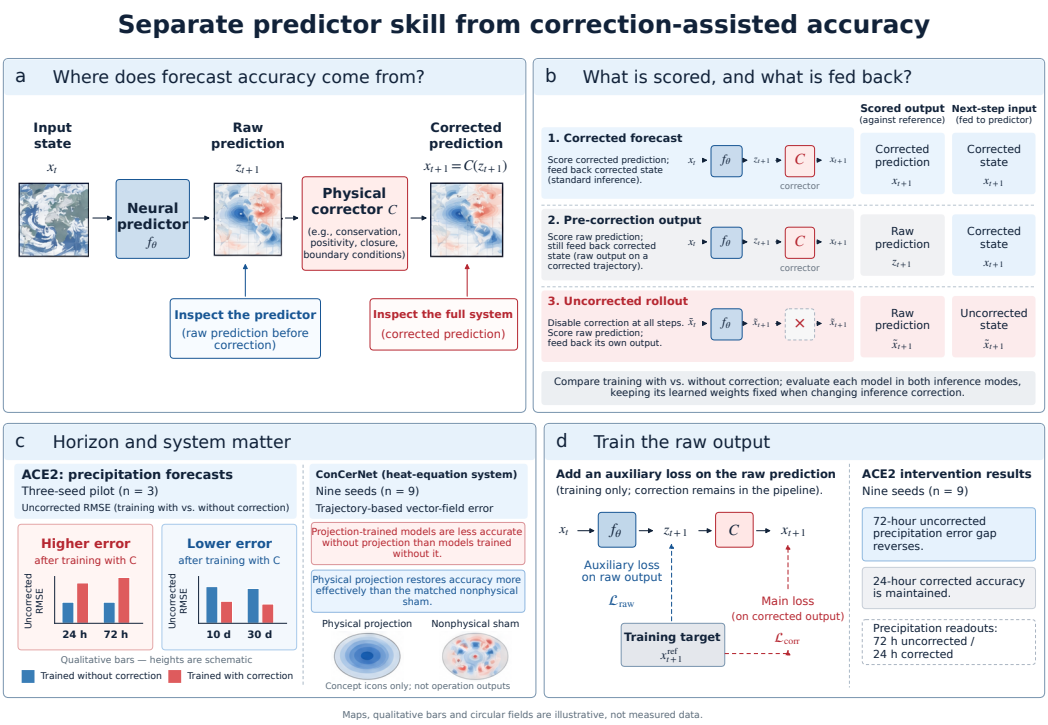


Figure 1: **Separating predictor skill from correction-assisted accuracy.** (a) The neural predictor produces z_{t+1} and the physical corrector C produces x_{t+1}; other inputs are suppressed. (b) The scored output and recurrent input define three complementary readouts. Each trained model retains its weights across inference modes. (c) Schematic bars summarize the ACE2 pilot's horizon-dependent training effect; blue/red denote training without/with correction. ConCerNet compares uncorrected vector-field error and physical versus sham projection. Quantitative results appear in Figures 2 and 4. (d) Dashed paths indicate training supervision. Table 2 reports the nine-seed intervention's 72-hour uncorrected and 24-hour corrected precipitation outcomes. The state maps, circular motifs, and bar heights are illustrative. ACE2 state RMSE and ConCerNet trajectory-based vector-field RMSE measure different quantities.

ACE2 (Watt-Meyer et al., 2025) provides a revealing case study. Its corrector greatly improves conservation and positivity, but the precipitation advantage depends on forecast horizon. With correction disabled throughout forecasting, models trained with correction have higher precipitation RMSE than models trained without it at 24 and 72 h, but lower RMSE at 10 and 30 days. Direct supervision of the uncorrected output reduces the short-horizon deficit. ConCerNet's heat-equation system (Zhang et al., 2023) exhibits a different pattern: projection-trained models rely strongly on the physical projection to recover predictive accuracy.

Our contribution is an experimental method for attributing forecast skill to a learned predictor, its corrector, and their interaction during training. We evaluate it through the ACE2 case study, a nine-seed ConCerNet experiment in a separate constrained dynamical system, and within-family comparisons of total-energy correction in ACE2S-SHiELD+.

2 Related Work

Neural atmospheric models. GraphCast and Pangu-Weather learn weather forecasts (Lam et al., 2023; Bi et al., 2023); NeuralGCM couples learned components to differentiable dynamics (Kochkov et al., 2024), and ACE2 targets weather-to-climate rollouts (Watt-Meyer et al., 2025).

Physics-guided and closed-loop learning. Physical priors (Karniadakis et al., 2021) include analytic atmospheric constraints (Beucler et al., 2021) and hard-constrained turbulence architectures (Mohan

et al., 2023). Learned solver corrections (Um et al., 2020; Dresdner et al., 2022) and feedback-induced distribution shifts (Ross et al., 2011) motivate evaluating how learned components behave when their outputs become future inputs.

Corrector dependence and negative controls. Chapman et al. (2026) show that precipitation rescaling can hide deterioration in raw network outputs and study remedies using raw supervision and an imbalance penalty. We additionally remove correction from autoregressive feedback and cross training with inference conditions. Spatial shams provide target-independent negative controls (Lipsitch et al., 2010) by disrupting physical alignment.

Constraint-enforcing operations. Optimization layers (Amos & Kolter, 2017), completion-and-correction (Donti et al., 2021), homeomorphic projections (Liang et al., 2023), and constrained generation (Cheng et al., 2025) incorporate feasibility into learned predictions. ConCerNet preserves invariants through vector-field projection (Zhang et al., 2023), providing a setting for our crossed training–inference evaluation.

3 METHOD

3.1 PREDICTOR–CORRECTOR FORMULATION

We first describe the state-correction formulation used for ACE2 (Watt-Meyer et al., 2025). Let x_t denote the model state and b_t the prescribed boundary or forcing information. The predictor produces

$$z_{t+1} = f_\theta(x_t, b_t), \tag{1}$$

and the physical corrector produces the next state,

$$x_{t+1} = C(z_{t+1}, x_t, b_t). \tag{2}$$

We evaluate prediction error against the reference state x^{ref}_{t+1} in three settings, distinguished by the scored output and the next-step input in Figure 1b.

Corrected rollout. We evaluate x_{t+1} and feed it into the next prediction step. This measures the forecast accuracy of the predictor–corrector system.

Pre-correction prediction along a corrected rollout. Along the same corrected trajectory, we evaluate z_{t+1} while feeding the corrected state x_{t+1} into the next step. This measures the accuracy of the network output conditional on a history of corrected states.

Uncorrected rollout. We disable correction throughout forecasting and recursively feed the network's outputs back into the predictor:

$$\tilde{x}_{t+1} = f_\theta(\tilde{x}_t, b_t), \qquad \tilde{x}_0 = x_0. \tag{3}$$

We evaluate $\tilde{x}_{t+1}$ against x^{ref}_{t+1}. This measures forecast accuracy without correction, including the effects of earlier predictions on subsequent inputs. For the same predictor, initial state, and prescribed forcing, $\tilde{x}_1 = z_1$. Later predictions can differ because the two rollouts feed back different states.

3.2 MATCHED-CONTROL EVALUATION IN ACE2

We compare the official physical correction with four controls: no correction, longitude shift, latitude-band permutation, and spectrum-preserving phase randomization. All models receive 2400 adaptation updates, with the common update budget selected using validation data. Each model uses the same correction condition during adaptation and forecasting, so these comparisons include both adaptation and inference effects.

Evaluation uses nine seeds, four months from 1940, forecast horizons of 24 and 72 hours, and three variables: near-surface temperature, 500 hPa height, and precipitation. The evaluation months are held out from local adaptation but fall within the released model's pretraining period (Watt-Meyer et al., 2025).

The three spatial shams disrupt the spatial alignment of the correction. Spatial phase randomization follows the surrogate-data principle of preserving spectral properties while disrupting other structure (Theiler et al., 1992). We evaluate predictive accuracy together with conservation, positivity, correction magnitude, and decorrelation.

3.3 Training–Inference Evaluation in ACE2

We independently vary correction during adaptation and forecasting. Models adapted with and without correction are each evaluated with correction enabled and disabled, yielding four combinations. Within each adapted model, learned weights, normalization, initial states, prescribed forcing, and evaluation settings remain fixed across inference conditions.

This pilot uses three seeds, four months held out from adaptation, and forecast horizons of 24 hours, 72 hours, 10 days, and 30 days.

For a given variable and forecast horizon, let $R_{a,i,s}$ denote the RMSE for seed s, averaged over the four evaluation months. The indices a and i indicate correction during adaptation and inference, respectively, with 1 denoting enabled and 0 denoting disabled. The relative benefit of inference correction is

$$I_{a,s} = 100\frac{R_{a,0,s} - R_{a,1,s}}{R_{a,0,s}}. \tag{4}$$

We evaluate this effect separately for predictors adapted with and without correction. The adaptation comparison with inference correction disabled is

$$A_s = 100\frac{R_{0,0,s} - R_{1,0,s}}{R_{0,0,s}}. \tag{5}$$

Positive A_s indicates lower uncorrected RMSE after adaptation with correction. We also report $I_{1,s} - I_{0,s}$ in percentage points. This contrast measures the difference between proportional inference benefits, each normalized by the corresponding predictor's uncorrected RMSE.

Our preregistered hypothesis predicts that adaptation with correction reduces uncorrected forecast accuracy relative to adaptation without correction, with the deficit increasing across forecast horizons.

3.4 Auxiliary Supervision of Pre-Correction Predictions

Motivated by the raw-supervision remedy studied by Chapman et al. (2026), we supervise pre-correction precipitation while retaining the official corrector during adaptation. The per-step objective is

$$\mathcal{L}_{\text{step}} = \mathcal{L}_{\text{main}} + \lambda\ \text{MSE}\big(N_p(z_{t+1,p}), N_p(x^{\text{ref}}_{t+1,p})\big)\,, \tag{6}$$

where p denotes surface precipitation and N_p is its normalization in the main loss. The auxiliary term averages squared errors over batch and spatial entries, supervises only precipitation, and leaves the main loss unchanged.

We select λ from $\{0.005, 0.02, 0.1\}$ using March 1940 validation forecasts averaged over seeds 7, 31, and 71. Eligible candidates have corrected 72-hour precipitation RMSE no more than 1% above the original model. We select the candidate with the lowest uncorrected 72-hour RMSE, breaking ties toward the smaller weight. This selects $\lambda = 0.005$ before held-out evaluation. All models receive 2400 adaptation updates. We evaluate the selected model with correction enabled and disabled; Appendix B.1 gives the evaluation design.

3.5 Evaluation in a Heat-Equation System

We apply the crossed training–inference design to ConCerNet's 101-dimensional heat-equation system (Zhang et al., 2023). Our experiment uses a known analytic invariant in place of learning one. The network predicts a vector field, and physical projection subtracts the mean of its components from each component. The projected vector field therefore sums to zero, enforcing conservation of the sum of state components at the vector-field level.

Networks trained with and without projection start from identical weights and use the same training data and minibatch order. Each network is evaluated with and without physical projection. An additional sham projection is matched in norm and constructed independently of evaluation targets.

The prediction metric is trajectory-based vector-field RMSE. At a forecast horizon of one time unit, we evaluate the model's vector field at the forecast state and the reference vector field at the reference state, then compute their RMSE. This metric combines differences in the vector fields and their evaluation states; state RMSE is a separate quantity.

3.6 Statistical Analysis

Aggregation before comparison. In the matched-control experiment, RMSE is averaged over seven paired forecast starts within each seed and month. Relative effects are then computed within month and averaged equally over months and seeds. The training–inference pilot first averages RMSE over four months within seed, then computes paired relative effects. Thus, the former averages month-specific relative effects, whereas the latter compares month-averaged errors. Appendix A.3 gives the estimands.

ACE2 matched-control evaluation. Training seed is the independent unit (Bouthillier et al., 2021); months and starts are nested paired observations. Across 24 confirmatory comparisons (four controls, two horizons, three variables), we report seed-level means, hierarchical bootstrap 95% intervals (Efron, 1979), and exact one-sided sign-flip tests (Ernst, 2004), with Holm correction (Holm, 1979). Long-horizon analyses are descriptive.

ACE2 training-inference evaluation. For the three-seed pilot, Figure 2 shows mean paired effects and descriptive 95% Student-t intervals, $\bar{e} \pm t_{0.975,2} s_e/\sqrt{3}$, where s_e is the sample standard deviation across seeds. Table 4 reports the same means and observed seed ranges, not confidence intervals. With three seeds, normality cannot be assessed reliably, and the minimum one-sided exact sign-test p-value is 0.125; these analyses remain descriptive.

ACE2 direct-supervision evaluation. The direct-supervision analysis uses nine adaptation seeds. We use bootstrap resampling (Efron, 1979), with adaptation seeds as clusters: 20,000 resamples preserve paired model comparisons within each resampled seed.

ConCerNet evaluation. We evaluate nine independent seeds at 0.1, 0.5, 1.0, and 2.0 model-time units using 32 held-out initial states. The confirmatory horizon is 1.0; other horizons are descriptive. Intervals use 20,000 paired seed-bootstrap resamples. Lower confidence bounds are the lower endpoints of two-sided 95% percentile intervals. Table 5 lists invariant-drift, prediction-error, physical–sham, and seed-consistency criteria.

4 Results

4.1 Practical Gains in Temperature

Across the 24 matched-regime comparisons, 23 favor the official correction and remain significant after Holm correction (Holm, 1979). Eight mean effects meet the practical-importance thresholds, all for temperature. The 72-hour precipitation comparison against no correction is inconclusive. Table 1 reports all effects and their uncertainty; Figure 5 in Appendix C.1 provides the corresponding interval plot.

4.2 Physical Consistency and Long-Horizon Error

A separate three-seed pilot evaluates ACE2's conservation and nonnegativity corrector (Watt-Meyer et al., 2025) using seven October 1940 forecast starts (Appendix B.3). Mean maximum dry-air drift was 0.0625 Pa with official correction during both adaptation and forecasting, compared with 1332 Pa without correction at either stage. Mean water-budget residual magnitude was 1.15×10^{-12} versus 7.93×10^{-5} kg m^{-2} s^{-1}, and the frequency of negative precipitation was 0% versus 33.1%.

In the longer matched-regime evaluations, precipitation RMSE was higher with correction at both stages than without correction at either stage: by 2.46% over the 10-day rollout window and 1.75% over the 30-day window. These effects are based on RMSE averaged over each window. No divergence was observed. Figure 6 reports the variable-specific effects.

4.3 Horizon-Dependent Reversal of the Uncorrected Gap

With correction disabled throughout forecasting, predictors adapted with correction have higher precipitation RMSE at short horizons and lower RMSE at longer horizons than predictors adapted

Table 1: **ACE2 matched-control forecast improvements.** Relative RMSE reduction (%) with official correction versus each control; correction conditions differ during both adaptation and inference.

	2 m temperature		500 hPa height		Precipitation	
Control	24 h	72 h	24 h	72 h	24 h	72 h
No correction	**0.98**† [0.90, 1.06]	**0.68**† [0.44, 0.92]	4.80† [3.97, 5.66]	0.87† [0.48, 1.27]	2.62† [2.26, 2.98]	0.00 [-0.53, 0.54]
Longitude shift	**0.95**† [0.89, 1.01]	**0.83**† [0.64, 1.02]	0.92† [0.78, 1.06]	0.69† [0.50, 0.89]	3.53† [3.15, 3.90]	1.79† [1.34, 2.23]
Latitude-band permutation	**0.95**† [0.89, 1.01]	**0.83**† [0.63, 1.03]	0.92† [0.79, 1.06]	0.70† [0.50, 0.89]	3.52† [3.13, 3.87]	1.80† [1.36, 2.24]
Phase randomization	**0.95**† [0.89, 1.01]	**0.85**† [0.65, 1.04]	0.97† [0.84, 1.11]	0.74† [0.54, 0.93]	3.60† [3.21, 3.98]	1.86† [1.40, 2.31]

Positive values favor official correction. Brackets: pointwise hierarchical bootstrap 95% intervals over nine seeds, four months per seed, and seven paired starts per month. †: Holm-adjusted one-sided sign-flip $p = 0.047$ across 24 tests; the unmarked comparison has $p = 0.514$. **Bold**: the unrounded mean meets the practical threshold (temperature: 0.5%; height and precipitation: 5%). The 72-hour precipitation mean versus no correction is slightly negative before rounding to 0.00%.

without correction (Figure 2; Table 4). Mean seed-level relative increases are 59.7% at 24 hours and 7.13% at 72 hours; the 72-hour increases range from 7.04% to 7.17% across the three seeds. At 10 and 30 days, the comparison reverses in all three seeds, with mean relative RMSE reductions of 13.48% and 15.69%. This reversal contradicts our preregistered hypothesis that the uncorrected deficit grows with forecast horizon.

The pattern also varies across variables. At 24 hours, adaptation with correction increases uncorrected 500 hPa height RMSE by 93.72% on average, while enabling inference correction for that predictor reduces its height RMSE by 51.58%. At 10 days, the 13.48% reduction in uncorrected precipitation RMSE contrasts with increases of 11.51% in temperature RMSE and 9.17% in height RMSE. The improvement at 10 days is therefore specific to precipitation among the three evaluated variables.

4.4 Direct Supervision Reduces the Short-Horizon Gap

Direct supervision reverses the uncorrected 72-hour precipitation RMSE gap in all nine seeds and lowers corrected 24-hour RMSE on average (Table 2). The mean uncorrected gap changes from +4.92% to -4.43% relative to the same-seed predictor adapted without correction. This experiment evaluates June and September 1940, whereas the training–inference pilot averages four months; the baseline gaps therefore refer to different evaluation sets. Figure 7 shows the paired seed-level changes.

Table 2: **Direct-supervision effects on precipitation forecasts** ($n = 9$)**.**

Comparison	Mean	95% interval
A. Uncorrected forecasts: 72-hour precipitation		
Original model: gap (%)	$+4.92$	$[+4.87, +4.97]$
Direct supervision: gap (%)	-4.43	$[-4.47, -4.39]$
Paired gap change: supervised − original (pp)	-9.35	$[-9.39, -9.32]$
B. Corrected forecasts: 24-hour precipitation		
Direct supervision: RMSE change (%)	-0.67	$[-0.73, -0.59]$

References: A, same-seed model adapted without correction; B, original correction-adapted model. RMSE averages June and September 1940 within seed. Intervals: 20,000 paired seed-bootstrap resamples; gap changes are paired within seed (pp: percentage points). All 9/9 seeds reverse the gap and improve corrected accuracy. The corrected upper bound (-0.59%) is below the prespecified $+1\%$ allowable increase.

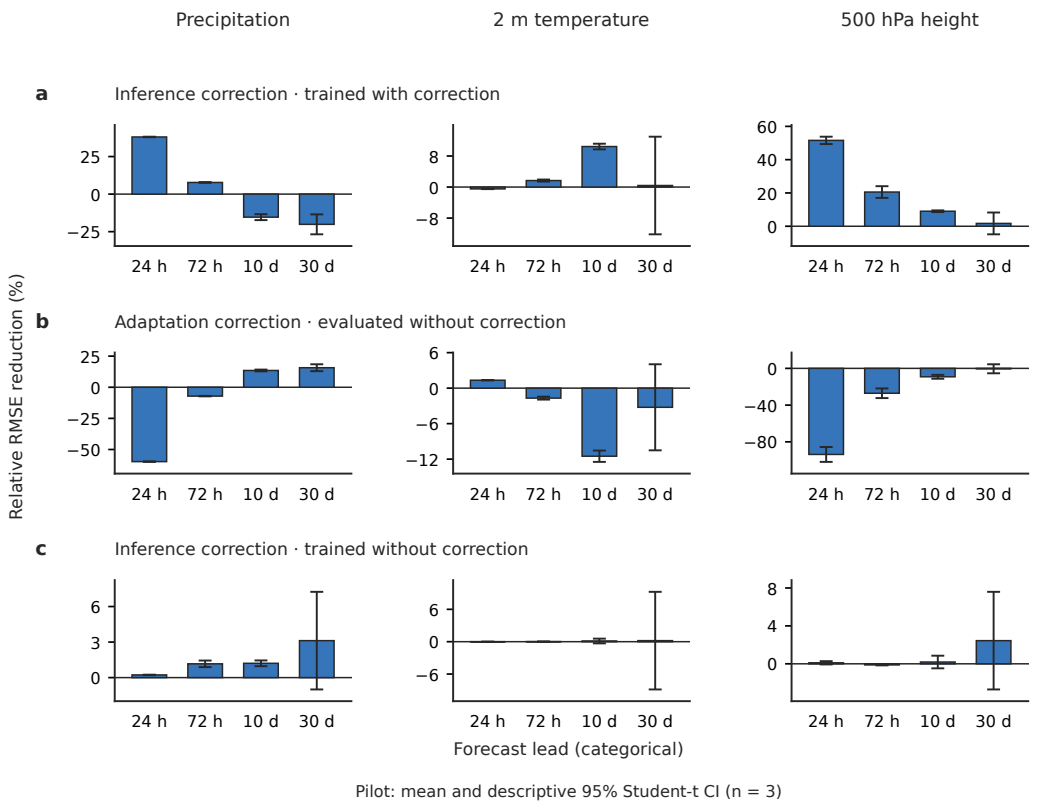


Figure 2: **ACE2 correction effects in the three-seed pilot.** Rows show the effect of enabling inference correction after adaptation with correction (a), adaptation with versus without correction under uncorrected inference (b), and enabling inference correction after adaptation without correction (c). Bars show mean paired relative RMSE reductions; capped error bars show descriptive 95% Student-t intervals ($n = 3$, two degrees of freedom). Positive values favor inference correction in a/c and adaptation with correction in b. Within each seed, RMSE is averaged over four months before computing relative effects. Forecast leads are categorical and vertical scales differ across panels.

4.5 Within-Family Energy-Correction Evaluation

At fixed ACE2S-SHiELD+ checkpoints (Clark et al., 2026), enabling energy correction reduces 24-hour 500 hPa height RMSE by 4.29% and 32.00% in the two checkpoint families trained with correction, rs0 and rs1 (Figure 3). For checkpoints trained without energy correction, the corresponding 24-hour height effects are below 0.1%, while enabling correction increases 30-day precipitation RMSE by 6.68% and 9.85%. Each effect averages four initialization dates. These descriptive comparisons show variation across variables, horizons, and checkpoints.

These fixed-checkpoint comparisons hold learned weights constant when toggling inference correction. A separate comparison evaluates checkpoints trained with and without energy correction, with correction disabled at inference (Figure 8). The released rs1 checkpoints used 213408 training batches with energy correction and 328320 without it. That between-checkpoint comparison therefore includes both training-budget and correction-condition differences; the fixed-checkpoint results above isolate the inference change.

4.6 Projection Dependence in ConCerNet

In our analytic-invariant adaptation of ConCerNet (Zhang et al., 2023), the mean seed-level ratio of uncorrected vector-field error for projection-trained models to that for models trained without projection is 26.87 at one model-time unit. Applying physical projection reduces the projection-

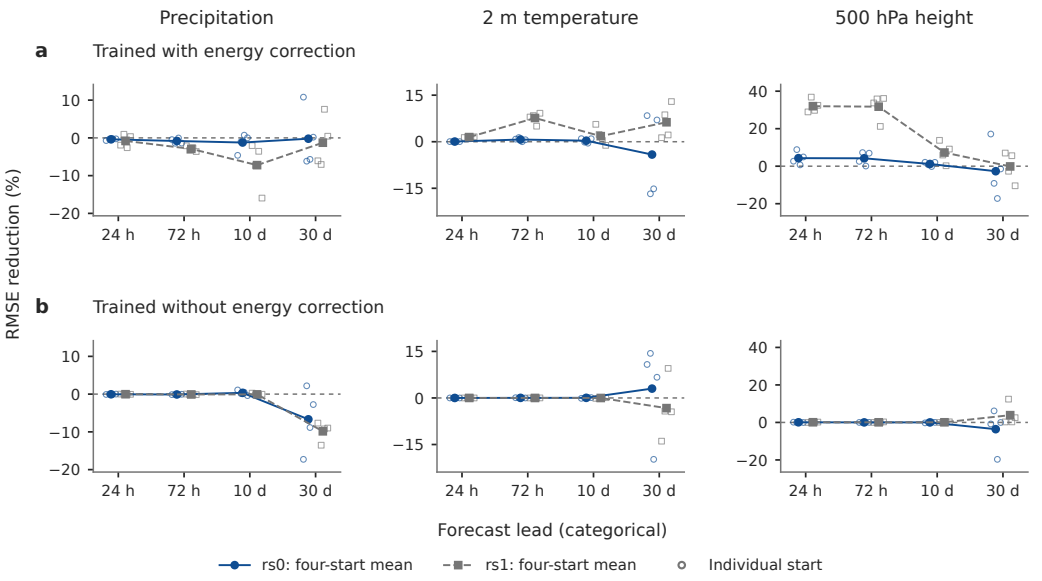


Figure 3: **ACE2S energy-correction effects at fixed checkpoints.** Rows enable inference correction after adaptation with (a) or without (b) it; positive RMSE reductions favor correction. Blue circles/solid lines denote rs0; gray squares/dashed lines denote rs1. Hollow marks show January, April, July and October 1979 starts, ordered left to right within each seed–lead group; filled marks show their mean. Forecast starts are nested within checkpoint families. Leads are categorical; variable-specific scales match across rows.

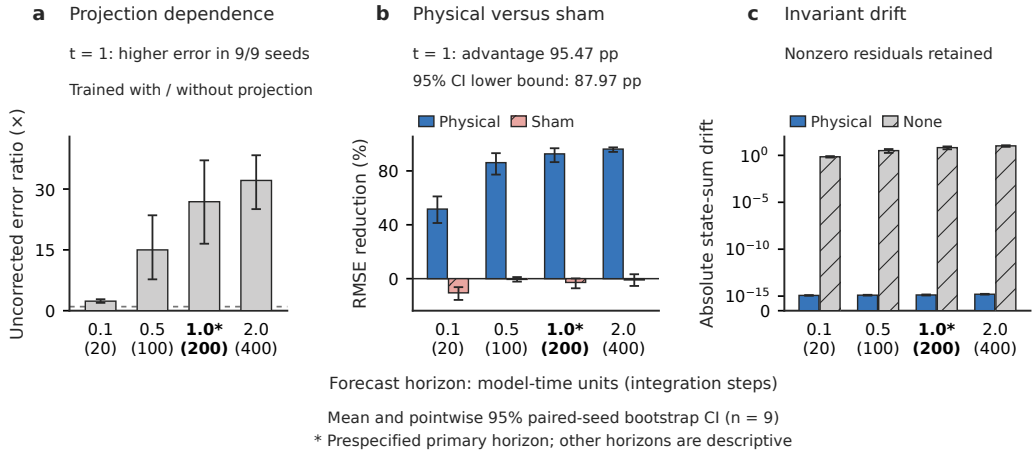


Figure 4: **ConCerNet projection dependence across forecast horizons.** The horizontal axis is elapsed time from initialization in model-time units, not hours or days; parentheses give integration steps ($\Delta t = 0.005$). Groups are categorical, not linearly spaced in time. The starred $t = 1$ (200 steps) is the prespecified primary horizon; others are descriptive. Bars: nine-seed means; error bars: pointwise 95% CIs from 20,000 paired seed-bootstrap resamples; 32 initial conditions per seed. (a) Uncorrected RMSE ratio, trained with/without projection; dashed line: equal error. (b) Physical (blue) and sham (pink, hatched) RMSE reductions versus no projection. Annotation: paired physical–sham difference. (c) Physical (blue) and no-projection (gray, hatched) state-sum drift. Panels b–c use projection-trained models; errors are trajectory-based vector-field RMSE, not state RMSE. Drift bars start at zero; the axis is linear below 10^{-16} and logarithmic above. Statistics use untransformed values; nonzero residuals are retained.

trained models' error by 92.65% on average, whereas the norm-matched sham increases it by approximately 2.82%. Both changes use the same model's error without inference projection as the reference.

Figure 4 extends these comparisons across four forecast times with pointwise 95% seed-bootstrap intervals. At the primary time $t = 1$, the physical–sham error-reduction difference is 95.47 percentage points. Table 5 reports the prespecified criteria and lower confidence bounds using the excess error ratio (total ratio minus one).

5 Discussion

In the ACE2 pilot, the effect of correction varies with adaptation history and forecast horizon. At longer horizons, the predictor adapted with correction has lower uncorrected precipitation RMSE than the predictor adapted without it, yet enabling its corrector increases RMSE. For the predictor adapted without correction, enabling correction reduces mean precipitation RMSE at every evaluated horizon. These results motivate evaluating coupled trajectories, as in prognostic parameterization tests (Brenowitz & Bretherton, 2018) and solver-in-the-loop learning (Um et al., 2020).

The variable-specific patterns in Table 4 suggest useful diagnostics. Following ACE2's separate assessments of bias and variability (Watt-Meyer et al., 2025), we propose examining precipitation bias, variability, and spatial agreement alongside errors in other fields. Tracking correction magnitudes and water-budget residuals over forecast lead could help assess whether changes in the applied correction coincide with the reversal.

The direct-supervision result links the short-horizon deficit to the adaptation objective: adding a loss on pre-correction outputs improves autonomous precipitation forecasts while retaining corrected forecast accuracy (Table 2). This complements the raw-amplitude stabilization studied by Chapman et al. (2026) with an evaluation of recursive forecasts. The reported intervention criteria focus on uncorrected 72-hour and corrected 24-hour precipitation error.

ConCerNet shows that projection dependence also occurs in a constrained dynamical system with an analytic invariant (Zhang et al., 2023). Its physical–sham comparison identifies a benefit specific to the physical projection under the tested control (Figure 4). Together with ACE2, this result motivates testing predictor–corrector dependence within each system. The two experiments use different error metrics, so their effect magnitudes are not directly comparable.

6 Limitations

The climate experiments adapt released ACE2-family checkpoints with prescribed sea-surface temperature and sea ice (Watt-Meyer et al., 2025). The 1940 evaluation months fall within checkpoint pretraining, so the results characterize adaptation and correction effects within that period. The horizon-dependent reversal is based on a three-seed pilot; the longer-horizon matched-regime analyses are descriptive. ConCerNet broadens the study to a heat-equation system with a different error metric, while climate-model coverage remains within the ACE2 family. The direct-supervision conclusions reported here concern 72-hour uncorrected and 24-hour corrected precipitation forecasts. Other physical correction schemes and broader temporal generalization require further evaluation.

7 Conclusion

Separating corrected forecasts, pre-correction outputs, and uncorrected rollouts reveals how physical correction interacts with the learned predictor. Across the tested ACE2 and ConCerNet settings, this interaction depends on training conditions, forecast horizon, and system. Direct supervision improves the short-horizon ACE2 predictor while preserving corrected accuracy. Evaluating these components separately therefore provides a clearer account of forecast skill in models that use physical correction.

AI USE STATEMENT

We used generative AI tools to assist with coding, experiment management, manuscript editing, and draft figure illustrations. The authors reviewed the resulting code, experimental settings, text, and figures and take responsibility for the final manuscript and accompanying artifacts.

REFERENCES

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *Proceedings of the International Conference on Machine Learning*, 2017.

Tom Beucler, Michael Pritchard, Stephan Rasp, Jordan Ott, Pierre Baldi, and Pierre Gentine. Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters*, 126(9):098302, 2021. doi: 10.1103/PhysRevLett.126.098302.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619:533–538, 2023.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021. URL https://proceedings.mlsys.org/paper_files/paper/2021/hash/0184b0cd3cfb185989f858a1d9f5c1eb-Abstract.html.

N. D. Brenowitz and C. S. Bretherton. Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, 45(12):6289–6298, 2018. doi: 10.1029/2018gl078510.

William E. Chapman, John Schreck, and Yingkai Sha. Hard conservation correctors can hide a degrading model when training autoregressive emulators. arXiv:2607.18416, 2026.

Chaoran Cheng, Boran Han, Danielle C. Maddix, Abdul Fatir Ansari, Andrew Stuart, Michael W. Mahoney, and Yuyang Wang. Gradient-free generation for hard-constrained systems. In *International Conference on Learning Representations*, 2025.

Spencer K. Clark, Troy Arcomano, James P. C. Duncan, Brian Henn, Anna Kwa, Jeremy McGibbon, W. Andre Perkins, Elynn Wu, Lucas M. Harris, Oliver Watt-Meyer, and Christopher S. Bretherton. Disentangling the effects of sea surface temperature and CO_2 in global machine learned weather-climate emulators. arXiv:2606.07928, 2026.

Priya L. Donti, David Rolnick, and J. Zico Kolter. DC3: A learning method for optimization with hard constraints. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2104.12225.

Gideon Dresdner, Dmitrii Kochkov, Peter Norgaard, Leonardo Zepeda-Núñez, Jamie A. Smith, Michael P. Brenner, and Stephan Hoyer. Learning to correct spectral methods for simulating turbulent flows. arXiv:2207.00556, 2022.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Michael D. Ernst. Permutation methods: A basis for exact inference. *Statistical Science*, 19(4): 676–685, 2004. doi: 10.1214/088342304000000396.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. doi: 10.1038/s42254-021-00314-5.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Christian Norgaard, Jamie Smith, Garrett Mooers, Milan Klower, James Lottes, Stephan Rasp, Peter Dueben, Thomas Kluyver, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632:1060–1066, 2024.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421, 2023.

Enze Liang, Minghua Chen, and Steven H. Low. Low complexity homeomorphic projection to ensure neural-network solution feasibility for optimization over (non-)convex set. In *Proceedings of the International Conference on Machine Learning*, 2023.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010.

Arvind T. Mohan, Nicholas Lubbers, Misha Chertkov, and Daniel Livescu. Embedding hard physical constraints in neural network coarse-graining of three-dimensional turbulence. *Physical Review Fluids*, 8(1):014604, 2023. doi: 10.1103/PhysRevFluids.8.014604.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020. doi: 10.1029/2020ms002203.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

James Theiler, Stephen Eubank, André Longtin, Bryan Galdrikian, and J. Doyne Farmer. Testing for nonlinearity in time series: The method of surrogate data. *Physica D: Nonlinear Phenomena*, 58 (1–4):77–94, 1992. doi: 10.1016/0167-2789(92)90102-S.

Kiwon Um, Robert Brand, Yun (Raymond) Fei, Philipp Holl, and Nils Thuerey. Solver-in-the-loop: Learning from differentiable physics to interact with iterative PDE-solvers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6111–6122, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/43e4e6a6f341e00671e123714de019a8-Abstract.html.

Oliver Watt-Meyer et al. ACE2: Accurately learning subseasonal to decadal atmospheric variability and forced responses. *npj Climate and Atmospheric Science*, 8:205, 2025. doi: 10.1038/s41612-025-01090-0.

Wang Zhang, Tsui-Wei Weng, Subhro Das, Alexandre Megretski, Luca Daniel, and Lam M. Nguyen. ConCerNet: A contrastive learning based framework for automated conservation law discovery and trustworthy dynamical system prediction. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41694–41714, 2023. URL https://proceedings.mlr.press/v202/zhang23ao.html.

A Evaluation Definitions

A.1 Prediction and Correction

We index the corrector in Section 3.1 by condition d, writing C_d. With correction disabled, $C_0(z_{t+1}, x_t, b_t) = z_{t+1}$. The three readouts and their feedback paths are defined in the main text and Figure 1b. Raw outputs are reconstructed by subtracting the recorded correction from the corrected output at each step of the corresponding rollout.

A.2 Evaluation Contrasts

Table 3 summarizes the four training–inference contrasts. Training includes local adaptation of released checkpoints.

Table 3: Training–inference evaluation contrasts. Correction is enabled or disabled at every forecast step.

Contrast	**Held fixed**	**Comparison and interpretation**
Inference correction after training with correction	Learned weights of the predictor trained with correction.	Enable versus disable correction during forecasting; measures its effect on the same predictor.
Training comparison without inference correction	Correction disabled throughout both forecasts.	Compare predictors trained with versus without correction; measures the difference between their uncorrected forecasts.
Inference correction after training without correction	Learned weights of the predictor trained without correction.	Enable versus disable correction during forecasting; measures the effect of adding correction without retraining.
Difference between inference effects	Weights within each inference pair; metric, horizon, effect scale, and sign convention.	Subtract the third-row inference effect from the first-row effect; measures the difference in relative correction benefits between the two predictors.

We use the seed-level RMSE $R_{a,i,s}$ and effects $I_{a,s}$ and A_s defined in Equations 4–5. The difference between inference effects is

$$J_s = I_{1,s} - I_{0,s}, \tag{7}$$

reported in percentage points. Each inference effect is normalized by its own predictor's uncorrected RMSE. Thus, J_s describes a difference between proportional improvements. We compute the contrast within each seed before summarizing its mean and range.

A.3 Aggregation of Prediction Errors

The following expressions formalize the aggregation procedures described in the main text.

For the matched-control experiment, let $\overline{R}_{d,s,m}$ denote mean RMSE over seven paired forecast starts for correction condition d, seed s, and month m. The improvement relative to control d is

$$G_{d,s,m} = 100\frac{\overline{R}_{d,s,m} - \overline{R}_{\text{official},s,m}}{\overline{R}_{d,s,m}}. \tag{8}$$

These relative effects are averaged equally over four months and then over nine training seeds.

The training–inference effects use Equations 4–5.

The long-horizon matched-regime experiment uses window-averaged error. For each seed, month, and correction condition, spatial RMSE uses cosine-latitude weighting, as in WeatherBench (Rasp et al., 2020), and is averaged over all saved six-hour forecast leads and initialization dates within the rollout window. The relative change from the matched no-correction model is

$$H_{d,s,m} = 100\left(\frac{R_{d,s,m}}{R_{0,s,m}} - 1\right). \tag{9}$$

Positive $H_{d,s,m}$ indicates higher RMSE with correction condition d.

B Experimental and Statistical Details

B.1 Direct Supervision of Uncorrected Outputs

Equation 6 defines the auxiliary objective; its normalization and validation-only weight selection are described in the main text. Here we give the prespecified initiation rule and held-out evaluation design.

The preregistered joint hypothesis requires a 72-hour uncorrected precipitation deficit in all three pilot seeds, growth of the deficit across horizons, and a mean 24-hour precipitation RMSE reduction of at least 0.5% when correction is restored. The direct-supervision experiment is initiated when the first and third criteria hold.

The direct-supervision evaluation uses nine adaptation seeds (7, 11, 23, 31, 43, 59, 71, 89, and 101) and two evaluation months, June and September 1940. Forecasts are evaluated at 24, 72, 240, and 720 hours. The supervised model is run with and without correction and compared with three matched baselines: the correction-trained model with correction enabled or disabled, and the model trained and evaluated without correction. The reported intervention criteria use the 72-hour uncorrected precipitation gap and the 24-hour corrected precipitation RMSE change.

B.2 Statistical Analysis

To account for training randomness (Bouthillier et al., 2021), training seed is the independent unit. Months and forecast starts are paired observations nested within seed.

For the matched-control experiment, we use bootstrap resampling (Efron, 1979) to report seed-level effects and pointwise 95% hierarchical percentile-bootstrap intervals, using 10,000 resamples of seeds, nested months, and paired starts. Exact one-sided sign-flip tests (Ernst, 2004) are Holm-adjusted (Holm, 1979) across the prespecified family of 24 comparisons. Long-horizon analyses are descriptive secondary outcomes.

Practical importance is assessed by comparing each unrounded mean relative RMSE improvement with a threshold of 0.5% for temperature and 5% for height and precipitation. The threshold classification is based on the mean effect; bootstrap intervals describe uncertainty around that estimate.

For the three-seed training–inference pilot, Figure 2 reports descriptive Student-t intervals and Table 4 reports seed means and ranges. The direct-supervision experiment uses 20,000 paired cluster-bootstrap resamples of its nine adaptation seeds.

B.3 Physical-Consistency Pilot

The physical-consistency comparison uses seeds 7, 11, and 23 and seven forecast starts, 1, 5, 9, 13, 17, 21, and 25 October 1940, with 72-hour rollouts. For each seed, dry-air drift is the maximum absolute change in area-weighted global-mean dry-air surface pressure over forecast starts and saved leads, relative to initialization. We then average these maxima across seeds. Water-budget residuals are area-weighted absolute means, averaged over starts, saved leads, and seeds. The negative-precipitation fraction is area-weighted and averaged over starts, the 24- and 72-hour endpoints, and seeds.

C Supplementary ACE2 Results

C.1 Matched-Control Comparisons

Figure 5 displays all 24 matched-control effects with intervals; numerical results appear in Table 1.

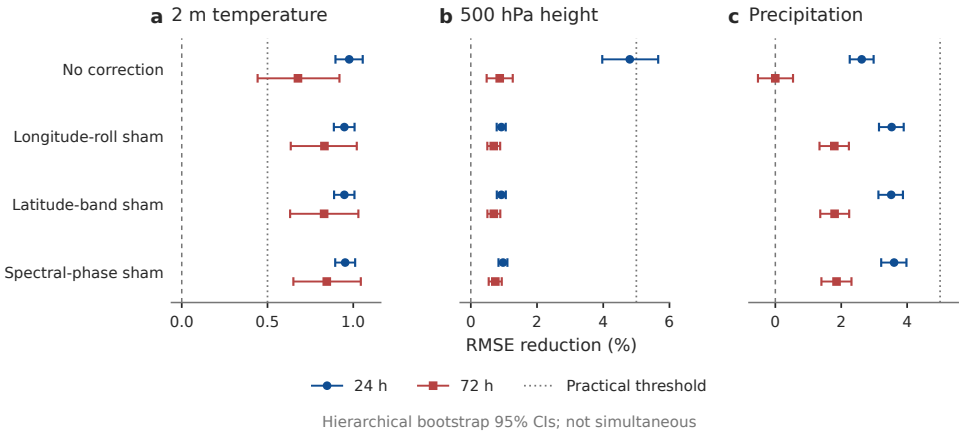


Figure 5: **All fixed matched-condition comparisons.** Points show the official correction's relative RMSE reduction against each control; blue circles identify 24 h and red squares identify 72 h. Intervals are pointwise hierarchical-bootstrap 95% confidence intervals. The dotted line is the variable-specific practical threshold and the dashed line denotes equal error. All 24 prespecified comparisons are shown. There are nine independent training seeds, with months and initial conditions nested within seeds. Separate numerical axes are retained for each variable.

C.2 Long-Horizon Comparisons

The long-horizon matched-regime evaluation uses February, May, August, and November 1940 and training seeds 7, 31, and 71. Ten-day forecasts use four initialization dates per month; 30-day forecasts use one. Figure 6 reports window-averaged RMSE changes relative to the matched no-correction model.

C.3 Training–Inference Comparisons

C.4 Direct-Supervision Seed Comparisons

Figure 7 shows the individual paired-seed changes summarized in Table 2.

D Additional Systems and Controls

D.1 Within-Family Energy-Correction Evaluation

The ACE2S-SHiELD+ experiment evaluates total-energy correction using the model and released checkpoints of Clark et al. (2026), from two seed families, rs0 and rs1. Each checkpoint is evaluated from four forecast starts: 1 January, 1 April, 1 July, and 1 October 1979.

The fixed-checkpoint and unequal-budget between-checkpoint comparisons are distinguished in the main text. Both between-checkpoint forecasts disable energy correction at inference.

Figure 3 in the main text shows the fixed-checkpoint inference comparisons. Figure 8 below reports the separate, budget-confounded between-checkpoint comparison.

D.2 ConCerNet Evaluation Criteria

The following seed-level criteria apply to the nine-seed ConCerNet experiment at its primary horizon, one model-time unit.

Let $E_{T,0}$ and $E_{U,0}$ denote errors without inference projection for models trained with and without projection, respectively. Let $E_{T,P}$ and $E_{T,S}$ denote errors for projection-trained models evaluated

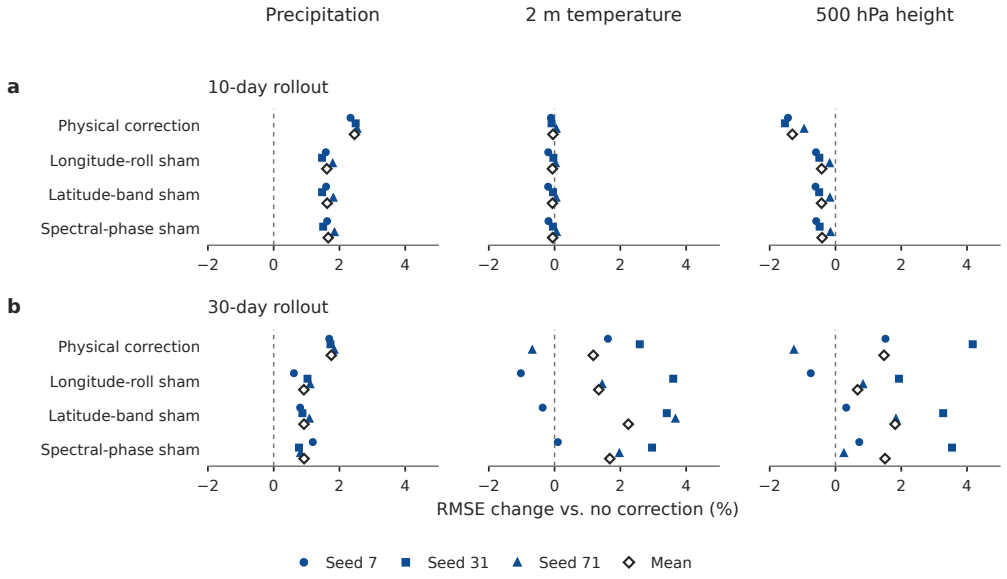


Figure 6: **Long-horizon RMSE changes relative to no correction.** Rows show 10- and 30-day rollouts and columns show forecast variables. Blue circles, squares and triangles identify training seeds 7, 31 and 71; open black diamonds show their mean. The dashed line denotes no change. Negative values indicate lower RMSE; this sign convention differs from the RMSE-reduction figures. No confidence intervals are drawn.

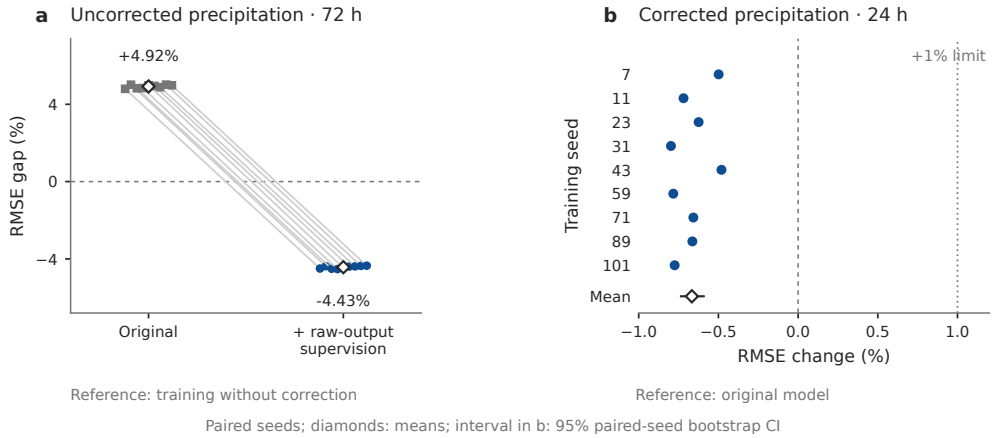


Figure 7: **Direct supervision of uncorrected predictions.** (a) Paired 72-hour precipitation RMSE gaps relative to training without correction. Gray squares and blue circles show models before and after raw-output supervision; lines connect the same seed and open diamonds mark means. (b) Corrected 24-hour precipitation RMSE changes relative to the original correction-trained model. Circles show nine seeds; the mean diamond and horizontal interval show the mean and its 95% paired-seed bootstrap interval. Dashed lines mark zero; the dotted +1% line is the prespecified allowable increase. The reference model differs between panels a and b.

Table 4: Descriptive effects of correction during adaptation and inference. Each entry reports the mean on the first line and [minimum, maximum] on the second line across three training seeds. Blocks 1–3 report relative RMSE reductions (%); positive values indicate lower RMSE than the specified uncorrected reference. Block 4 reports differences between relative reductions in percentage points.

1. Retaining correction at inference (%)

Adapted with correction; reference: the same predictor run without correction.

Variable	24 h	72 h	10 d	30 d
2 m temperature	-0.47 $[-0.50, -0.46]$	$+1.67$ $[1.55, 1.76]$	$+10.44$ $[10.11, 10.64]$	$+0.40$ $[-4.54, 5.58]$
500 hPa height	$+51.58$ $[50.61, 52.37]$	$+20.55$ $[19.11, 21.94]$	$+9.03$ $[8.80, 9.17]$	$+1.70$ $[-1.03, 4.23]$
Precipitation	$+38.04$ $[37.96, 38.10]$	$+7.75$ $[7.64, 7.83]$	-15.35 $[-16.00, -14.47]$	-20.13 $[-22.45, -17.21]$

2. Adaptation with correction (%)

Both run without correction; reference: the predictor adapted without correction.

Variable	24 h	72 h	10 d	30 d
2 m temperature	$+1.34$ $[1.33, 1.36]$	-1.69 $[-1.80, -1.61]$	-11.51 $[-11.94, -11.22]$	-3.24 $[-5.39, 0.10]$
500 hPa height	-93.72 $[-96.76, -90.27]$	-27.10 $[-29.21, -24.98]$	-9.17 $[-10.13, -8.50]$	-0.44 $[-2.72, 0.80]$
Precipitation	-59.73 $[-59.82, -59.65]$	-7.13 $[-7.17, -7.04]$	$+13.48$ $[13.30, 13.81]$	$+15.69$ $[14.78, 16.93]$

3. Adding correction at inference (%)

Adapted without correction; reference: the same predictor run without correction.

Variable	24 h	72 h	10 d	30 d
2 m temperature	-0.01 $[-0.01, -0.01]$	$+0.01$ $[0.01, 0.02]$	$+0.13$ $[0.01, 0.33]$	$+0.19$ $[-3.76, 3.41]$
500 hPa height	$+0.11$ $[0.03, 0.16]$	-0.12 $[-0.13, -0.10]$	$+0.19$ $[-0.01, 0.50]$	$+2.44$ $[1.15, 4.84]$
Precipitation	$+0.23$ $[0.22, 0.24]$	$+1.16$ $[1.04, 1.25]$	$+1.21$ $[1.12, 1.31]$	$+3.12$ $[1.96, 5.02]$

4. Difference in correction effects (percentage points)

Block 1 minus block 3, paired within each seed; positive means a larger effect in block 1.

Variable	24 h	72 h	10 d	30 d
2 m temperature	-0.47 $[-0.49, -0.45]$	$+1.66$ $[1.54, 1.74]$	$+10.32$ $[10.07, 10.57]$	$+0.20$ $[-5.47, 3.92]$
500 hPa height	$+51.47$ $[50.59, 52.24]$	$+20.66$ $[19.24, 22.03]$	$+8.84$ $[8.30, 9.18]$	-0.74 $[-5.87, 2.88]$
Precipitation	$+37.82$ $[37.74, 37.86]$	$+6.59$ $[6.58, 6.60]$	-16.55 $[-17.12, -15.78]$	-23.25 $[-24.41, -22.23]$

Within each seed, RMSE is averaged over the same four evaluation months before relative effects are calculated. Each inference effect uses its own predictor's uncorrected RMSE as the denominator. Block 4 subtracts the two relative inference effects within each seed before computing the mean and range.

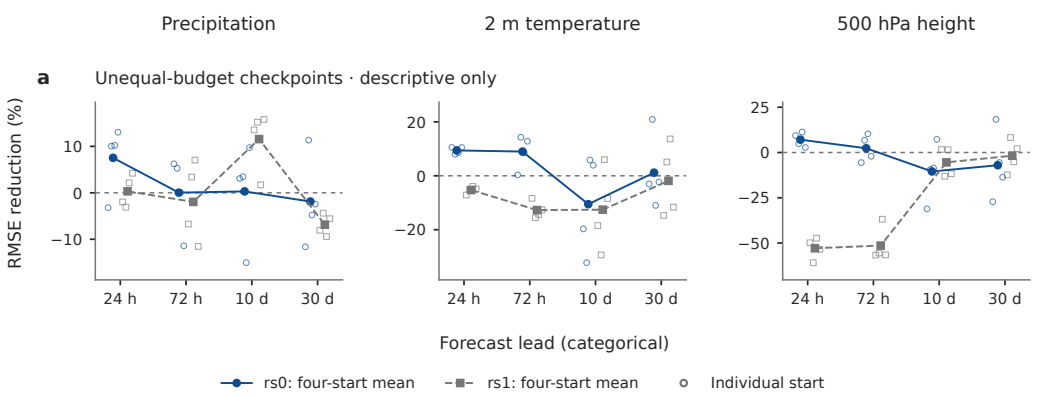


Figure 8: **Descriptive comparison of unequal-budget ACE2S checkpoints.** Both checkpoint families are evaluated without the inference energy corrector. Positive RMSE reduction favors the checkpoint adapted with correction. Hollow marks show initial-condition-specific effects and connected marks show each seed's four-start mean. Circles with solid blue lines identify rs0; squares with dashed gray lines identify rs1. The four starts are January, April, July and October 1979, ordered left to right within each seed–lead group. Lead times are categorical. Unequal adaptation budgets confound the between-checkpoint contrast; it does not isolate a causal effect of correction during adaptation.

with physical and sham projection. The excess uncorrected error ratio is

$$Q_s = \frac{E_{T,0} - E_{U,0}}{E_{U,0}}. \tag{10}$$

Adding one gives the total error ratio reported in the main text.

The physical-projection error reduction is

$$P_s = 100\frac{E_{T,0} - E_{T,P}}{E_{T,0}}, \tag{11}$$

and its advantage over the sham projection is

$$S_s = 100\frac{E_{T,S} - E_{T,P}}{E_{T,0}}, \tag{12}$$

reported in percentage points. These quantities are computed within each seed.

Invariant drift is the mean absolute change in the conserved state sum $g(x) = \sum_j x_j$ from its initial value. Its reduction under physical projection is

$$100\frac{D_{T,0} - D_{T,P}}{D_{T,0}}. \tag{13}$$

Table 5 reports the decision criteria and results.

Table 5: ConCerNet decision criteria and results at one simulation-time unit.

Comparison	Unit	Mean/ count	Lower 95% CI endpoint	Criterion
Excess uncorrected vector-field error: projection-trained vs. trained without projection	ratio	25.874	15.516	$\bar{x} > 0,\ L > 0$
Vector-field error reduction: physical projection vs. no projection	%	92.65	86.53	$\bar{x} > 0,\ L > 0$
Error-reduction difference: physical vs. sham projection	pp	95.47	87.97	$\bar{x} > 0,\ L > 0$
Invariant-drift reduction: physical vs. no projection	%	100.00	—	$\bar{x} \geq 90$
Seeds with nonnegative excess uncorrected error	seeds	9/9	—	$\geq 8/9$

Values are seed means ($\bar{x}$), except the final count. L is the lower endpoint of a two-sided 95% percentile bootstrap interval for the seed mean, based on 20,000 resamples of nine seeds. The dash denotes a criterion that does not use an interval. All prediction errors are trajectory-based vector-field RMSE. The excess ratio equals the total error ratio minus one. The physical–sham contrast is measured in percentage points. The invariant-drift reduction of 100.00% is rounded.