LEARNING TO REPAIR PHYSICAL VIOLATIONS IN FROZEN WEATHER FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Weather foundation models can achieve strong predictive accuracy while producing states that violate physically meaningful relationships. We study whether such violations can be repaired *after* forecasting, without modifying the forecasting model. ConserveFM learns an additive correction conditioned on a frozen forecast, the preceding atmospheric state, and forecast lead, and optimizes predictive fidelity together with moisture, hydrostatic, advection, and spectral consistency. We evaluate post-hoc repair on ERA5 under 60 matched stress conditions spanning three forecast leads, five structured corruption families, and four severities. An accuracy-oriented variant achieves the best stress NRMSE of 0.3008 and ACC of 0.9503. A physics-oriented variant reduces the aggregate physical-violation ratio from 0.852 to 0.333, while increasing stress NRMSE to 0.3159. Direct state prediction occupies an intermediate Pareto-optimal regime. These results show that predictive accuracy and physical consistency are distinct objectives: stronger physical repair can be obtained without retraining the backbone, but not without a measurable predictive cost.

1 INTRODUCTION

Large pretrained models are increasingly used for atmospheric and spatiotemporal forecasting, where they provide transferable representations and strong predictive performance across environmental tasks Nguyen et al. (2025); Shan et al. (2026); Song et al. (2026); Liao et al. (2025); Yuan et al. (2026). Yet forecast accuracy alone does not ensure that a predicted state respects relationships among moisture, pressure, temperature, wind, and spatial scales. A forecast can be close to the reference state under a statistical metric while remaining physically inconsistent, and enforcing physical structure too aggressively can in turn degrade an otherwise useful forecast.

Most physics-informed approaches address this problem during training by adding physical losses, architectural constraints, or conservation terms. This strategy is less attractive when the forecaster is a large pretrained model that is expensive to retrain, externally supplied, or already optimized for predictive skill. It also entangles two separate questions: whether the forecaster predicts accurately, and whether its output can be modified toward greater physical consistency.

We instead study *post-hoc physical repair*. Given a frozen forecast, ConserveFM learns a separate correction conditioned on the forecast, the preceding atmospheric state, and forecast lead. The forecasting backbone is never updated. This formulation makes physical intervention modular and allows its predictive cost to be measured directly.

A second motivation is that physical consistency is inherently multi-objective. Prior work has shown that physical priors may improve or degrade predictive performance depending on the imposed relation and its weighting Soltani et al. (2024); He et al. (2023); Jiang et al. (2025); Yang et al. (2025). We therefore do not assume that the most physically consistent repair must also minimize forecast error. Instead, we explicitly evaluate the resulting accuracy–physics operating points. Our experiments use a controlled synthetic stress benchmark to isolate specific failure modes; they are not intended to model the full spectrum of intrinsic or autoregressive forecast drift.

Our contributions are:

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.

- We formulate physical consistency correction as a learned post-hoc repair problem for frozen weather forecasting models.
- We combine predictive reconstruction with moisture, hydrostatic, advection, and spectral consistency and evaluate repair under 60 matched lead–corruption–severity stress conditions.
- We show that the strongest predictive and strongest physics-oriented repairs are distinct Pareto-optimal solutions: substantially stronger physical consistency is achievable, but at a measurable predictive cost.

2 Related Work

Pretrained and foundation models have been applied to tropical-cyclone, irradiance, precipitation, and broader spatiotemporal forecasting Nguyen et al. (2025); Shan et al. (2026); Song et al. (2026); Liao et al. (2025). These studies primarily optimize forecasting or transfer performance rather than correction of an already generated forecast.

A complementary literature incorporates physical information directly into learned models. Physical priors have been used in meteorological prediction Pang et al. (2025); Gijón et al. (2026); Katiba et al. (2026); Gjylapi et al. (2026), while conservation-aware learning has been studied through flux continuity, governing-equation residuals, and structured reconstruction Jagtap et al. (2020); Zheng et al. (2020); Deng & Han (2026). Such constraints can improve physical behavior, but their effectiveness depends strongly on formulation and optimization.

Of particular relevance are approaches that distinguish soft from hard physical constraints. Projection layers can impose selected relations by construction Chen et al. (2024), while output projection can map arbitrary predictions onto predefined physical manifolds Valente et al. (2025). These methods provide strong guarantees when a tractable feasible set is known. ConserveFM instead learns the correction itself, allowing the intervention to depend on forecast state, history, and lead time.

Finally, several studies report conflicts between predictive fit and physical validity Soltani et al. (2024); He et al. (2023); Jiang et al. (2025), and optimization conflicts between data and physics terms are a known source of instability Yang et al. (2025); Huang & Agarwal (2023). ConserveFM builds on this observation by treating forecast repair explicitly as a multi-objective problem rather than assuming that physical regularization and predictive accuracy share a common optimum.

3 Method

Figure 1 provides an overview of the ConserveFM framework, from frozen-backbone forecasting to physics-aware post-hoc correction and the final accuracy- and physics-oriented operating modes.

3.1 Post-Hoc Forecast Repair

Let x_t denote the most recent atmospheric state, $\hat{x}_{t+\tau}$ a forecast produced by a frozen backbone at lead τ, and $x_{t+\tau}$ the corresponding reference state. ConserveFM learns

$$f_\theta(\hat{x}_{t+\tau}, x_t, \tau) \to \Delta x_{t+\tau},$$

and returns

$$\hat{x}^*_{t+\tau} = \hat{x}_{t+\tau} + \Delta x_{t+\tau}.$$

Only f_θ is trained; the forecasting backbone remains fixed.

The repair model receives normalized versions of the raw forecast and previous state together with lead-time conditioning. Its primary output is the atmospheric correction. Optional localization and corruption-type heads are used in the broader ablation study but are not required at inference.

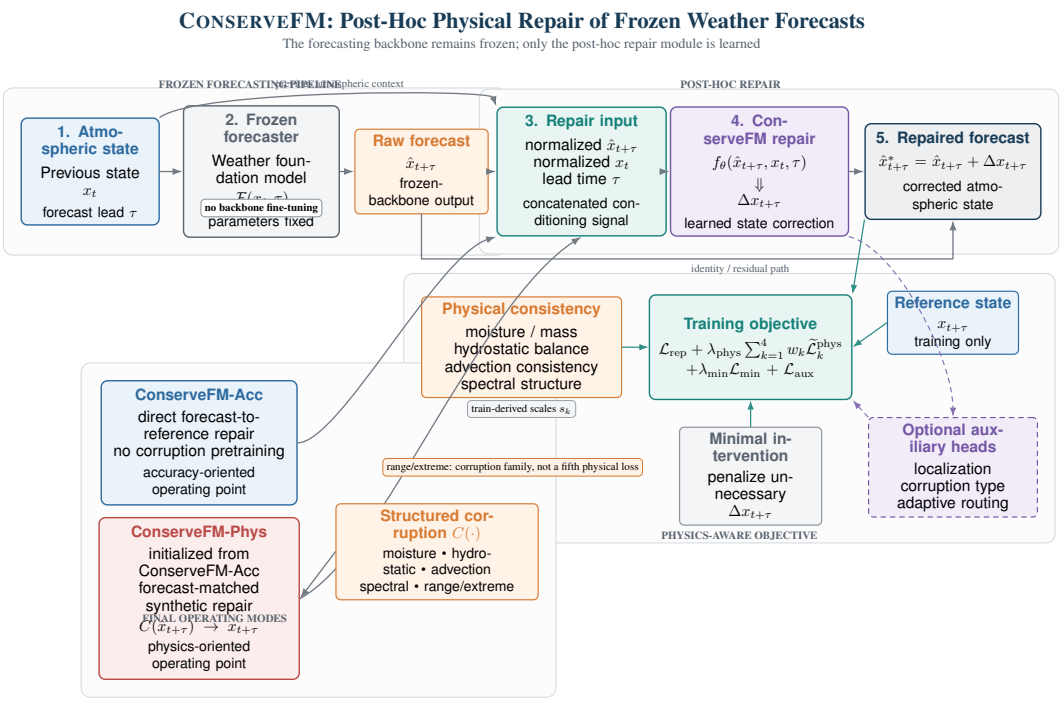


Figure 1: ConserveFM post-hoc repair framework. A frozen weather backbone first produces the raw forecast $\hat{x}_{t+\tau}$. ConserveFM conditions on this forecast, the previous atmospheric state x_t, and forecast lead τ to predict an additive correction $\Delta x_{t+\tau}$. Training combines forecast reconstruction, four normalized physical-consistency terms, and a minimal-intervention penalty, with optional auxiliary heads. ConserveFM-Acc emphasizes predictive accuracy, whereas ConserveFM-Phys adds forecast-matched synthetic corruption repair to move the operating point toward stronger physical consistency.

3.2 PHYSICS-AWARE OBJECTIVE

Training combines reconstruction, physical consistency, and minimal intervention:

$$\mathcal{L} = \mathcal{L}_{\text{rep}} + \lambda_{\text{phys}} \sum_{k=1}^{4} w_k \tilde{\mathcal{L}}_k^{\text{phys}} + \lambda_{\text{min}} \mathcal{L}_{\text{min}} + \mathcal{L}_{\text{aux}}.$$

The four physical terms measure complementary properties: (i) moisture or mass consistency, (ii) hydrostatic consistency across pressure levels, (iii) endpoint advection consistency with the preceding state and wind field, and (iv) spatial spectral structure. These terms are diagnostic residuals used to compare repaired states; they are not claimed to constitute a complete discretization of the atmospheric primitive equations. In particular, the advection term at 24–72 h is interpreted as an endpoint consistency surrogate rather than an exact trajectory integration.

Because the raw residuals have different numerical scales, we normalize each term with a reference scale s_k estimated from the training split and kept fixed during evaluation:

$$\tilde{\mathcal{L}}_k^{\text{phys}} = \frac{\log\left(1 + \mathcal{L}_k^{\text{phys}} / s_k\right)}{\log 2}.$$

The minimal-change term discourages unnecessary intervention when the frozen forecast is already suitable.

We additionally use five structured corruption families for training and stress evaluation: moisture leakage, hydrostatic inconsistency, advection inconsistency, spectral artifacts, and range/extreme perturbations. The last is a corruption family rather than a fifth physical constraint. The same corruption families are used for ConserveFM-Phys training and stress testing, so the stress benchmark

measures held-out-state performance within these families rather than out-of-family corruption generalization.

3.3 Accuracy- and Physics-Oriented Repair

We retain two representative ConserveFM operating modes.

ConserveFM-Acc. The accuracy-oriented model learns directly from frozen-backbone forecasts and reference states without corruption pretraining. It is selected for predictive robustness in the broad screening experiment.

ConserveFM-Phys. Starting from ConserveFM-Acc, we perform an additional forecast-matched repair stage in which a structured perturbation is applied to the frozen forecast:

$$\widetilde{x}_{t+\tau} = C(\hat{x}_{t+\tau}), \qquad \Delta\widetilde{x}_{t+\tau} = f_\theta(\widetilde{x}_{t+\tau}, x_t, \tau),$$

with repaired output

$$\widetilde{x}^*_{t+\tau} = \widetilde{x}_{t+\tau} + \Delta\widetilde{x}_{t+\tau},$$

trained against $x_{t+\tau}$. Thus ConserveFM-Phys retains the additive residual parameterization; it does not become a direct-state predictor. Applying C to the forecast rather than the reference state aligns synthetic training with corrupted forecast inputs and shifts the solution toward stronger physical consistency.

Internal comparison variants. *Direct state* receives the same conditioning variables but predicts the repaired atmospheric state directly, without the additive identity path. *Original full* denotes the initial joint formulation using the complete corruption/physics objective together with the auxiliary localization, corruption-type, and adaptive-routing components. These are controlled internal ablations rather than external post-processing baselines.

4 Experimental Setup

Data. We use ERA5 on a 1.5° global grid at 6-hour resolution. Each state contains 71 channels comprising surface variables and pressure-level geopotential, temperature, specific humidity, and horizontal winds over 13 pressure levels. We use 1979–2018 for training, 2019–2020 for validation, and 2021–2022 for testing. Statistics required for normalization are estimated from the training split only.

Backbone and screening. ClimaX is the primary frozen forecasting backbone. A broad preliminary study covers 20 repair variants, including alternative parameterizations, physical constraint ablations, corruption ablations, weighting strategies, and auxiliary-head ablations. The preliminary screen uses seeds 42, 43, and 44 only for model selection and stability analysis. Its physical-space RMSE mixes channels expressed in different units and is therefore treated strictly as a screening diagnostic, not as a calibrated atmospheric error metric.

Stress benchmark. We evaluate leads

$$\tau \in \{6, 24, 72\}\ \mathrm{h}$$

under five corruption families and four severity levels

$$\alpha \in \{0.25, 0.5, 1.0, 2.0\},$$

giving $3 \times 5 \times 4 = 60$ matched stress conditions per method. All methods receive the same corrupted examples. The test years are disjoint from the training period, but the corruption families themselves are intentionally shared with the additional ConserveFM-Phys training stage; conclusions from this benchmark therefore concern robustness to these specified perturbation families, not arbitrary operational distribution shift.

Table 1: Main predictive and physical-consistency results. NRMSE is lower-is-better and ACC is higher-is-better. Physical columns report repaired-to-corrupted violation ratios; values below one indicate improvement. The uncorrected corrupted input has ratio 1 by definition. Clean and stress NRMSE use different aggregation protocols and should not be compared as a paired before/after measurement.

Method	Clean NRMSE ↓	Stress NRMSE ↓	Stress ACC ↑	Moist. ↓	Hydro. ↓	Adv. ↓	Spec. ↓	Physics GM ↓
ConserveFM-Acc	**0.3070**	**0.3008**	**0.9503**	0.494	0.827	1.394	0.925	0.852
Direct state	0.3130	0.3071	0.9487	0.730	**0.215**	1.001	**0.422**	0.507
ConserveFM-Phys	0.3193	0.3159	0.9458	**0.263**	0.464	**0.168**	0.598	**0.333**
Original full	0.3164	0.3339	0.9401	0.635	0.982	1.361	0.997	0.959
No repair (ratio reference)	–	–	–	1.000	1.000	1.000	1.000	1.000

Metrics. Predictive performance is measured in standardized channel space with cosine-latitude area weighting using normalized RMSE (NRMSE) and anomaly correlation (ACC). Physical consistency is measured independently for the four constraints. For constraint k, we report

$$r_k = \frac{\mathcal{V}_k(\hat{x}^*)}{\mathcal{V}_k(\widetilde{x})},$$

where $r_k < 1$ indicates reduced violation. The uncorrected corrupted input is therefore the physical reference $r_k = 1$ by construction. We summarize the four ratios by their geometric mean, denoted Physics GM.

Clean and stress NRMSE are not a paired before/after quantity. Clean metrics aggregate unperturbed test forecasts, whereas stress NRMSE is an equal-weight macro-average over the 60 condition-level evaluations. Their different aggregation can therefore yield a stress macro-average numerically below the clean aggregate; only comparisons between methods under the same protocol are interpreted.

The final normalized comparison uses fully evaluated seed-42 models. Uncertainty across the stress benchmark is quantified with a paired bootstrap over the 60 matched conditions. These intervals measure condition-level variation and are not estimates of training-seed uncertainty.

5 Results

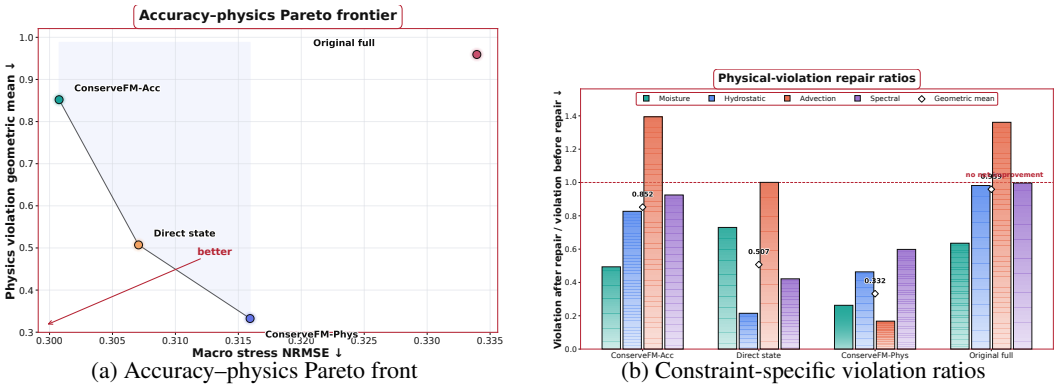
5.1 Accuracy–Physics Trade-off

Table 1 summarizes the main comparison. ConserveFM-Acc achieves the strongest predictive performance, with clean NRMSE of 0.3070, stress NRMSE of 0.3008, and stress ACC of 0.9503. ConserveFM-Phys instead provides the strongest aggregate physical consistency, with a Physics GM of 0.333. Direct state prediction occupies an intermediate operating point, whereas the original full formulation is dominated.

ConserveFM-Phys reduces moisture, hydrostatic, advection, and spectral violations by approximately 74%, 54%, 83%, and 40%, respectively. Its aggregate physical-violation ratio decreases to 0.333, compared with 0.852 for ConserveFM-Acc, while stress NRMSE increases from 0.3008 to 0.3159, corresponding to an approximately 5% predictive cost. Direct state is stronger on the hydrostatic (0.215) and spectral (0.422) diagnostics, whereas ConserveFM-Phys gains most strongly on moisture and advection; the aggregate GM therefore should not be read as uniform superiority on every constraint.

Notably, ConserveFM-Acc increases the advection violation ratio to 1.394 despite achieving the lowest forecast error. Predictive accuracy and physical consistency therefore do not identify the same solution. As shown in Fig. 2, ConserveFM-Acc, Direct state, and ConserveFM-Phys form three distinct non-dominated operating points.

A paired bootstrap over the 60 matched stress conditions gives an NRMSE advantage for ConserveFM-Acc over Direct state of 0.00631 (95% CI $[0.00385, 0.00868]$) and over ConserveFM-Phys of 0.01517 (95% CI $[0.01320, 0.01726]$). These intervals quantify variation across stress conditions rather than uncertainty across independently trained models.



(a) Accuracy–physics Pareto front (b) Constraint-specific violation ratios

Figure 2: Accuracy and physical consistency yield different operating points. ConserveFM-Acc minimizes predictive error, whereas ConserveFM-Phys provides the strongest aggregate physical repair.

5.2 Stress Severity and Forecast Horizon

The predictive ordering is not driven by a single corruption family. ConserveFM-Acc obtains the lowest NRMSE for moisture, hydrostatic, advection, and spectral perturbations, while Direct state is marginally better for the range/extreme corruption.

Performance degrades with forecast horizon. ConserveFM-Acc increases from 0.2035 NRMSE at $+6$ h to 0.4044 at $+72$ h. At the strongest corruption severity, its NRMSE is 0.3121, compared with 0.3163 for Direct state, 0.3282 for ConserveFM-Phys, and 0.4066 for the original full formulation. Additional spatial diagnostics are provided in Appendix A; they are descriptive complements to the aggregate stress metrics rather than additional headline evidence.

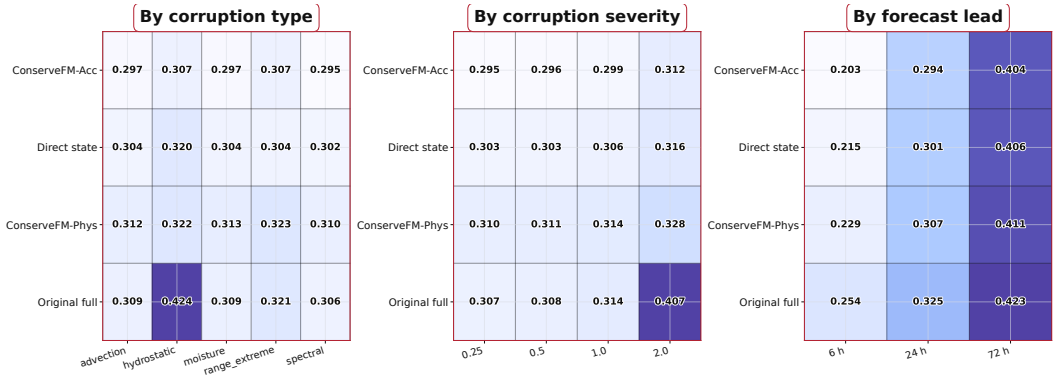


Figure 3: Stress robustness decomposed by corruption family, severity, and forecast lead. Longer horizons and stronger perturbations increase predictive error, while the relative ordering remains broadly stable.

5.3 Optimization Stability and Negative Results

The preliminary three-seed screen reveals substantial instability in several joint physics formulations. Direct state has the lowest preliminary mean stress RMSE (144.998 ± 9.164), while the no-corruption-pretraining variant that provides the basis for ConserveFM-Acc is second but markedly more stable (166.150 ± 2.240). The original full formulation reaches 428.878 ± 269.561. These values are mixed-unit physical-space screening statistics and are not directly comparable with the standardized NRMSE used in Table 1.

The auxiliary tasks provide an additional negative result. Corruption-type accuracy remains approximately 0.20–0.21, close to five-class chance performance, and localization IoU remains low. The

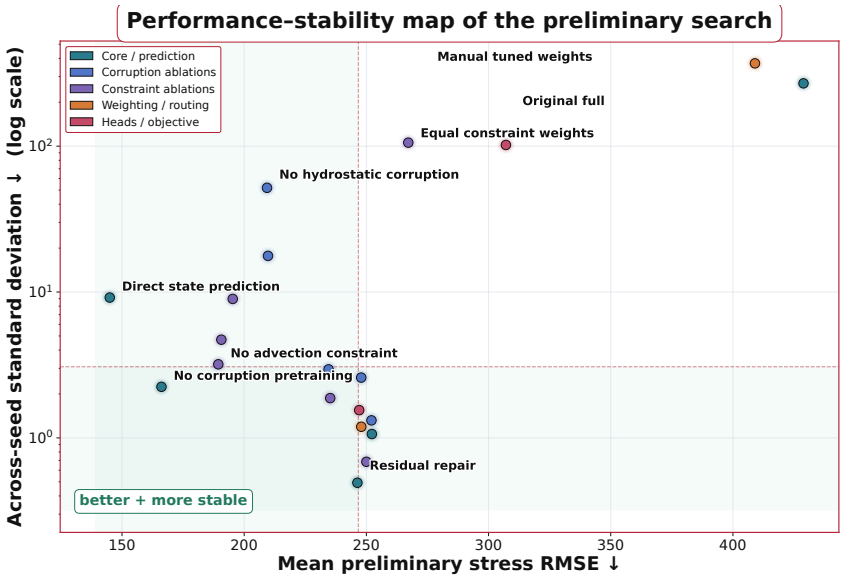


Figure 4: Preliminary performance–stability analysis across training seeds. The original joint formulation exhibits substantially larger variability than the strongest predictive variants.

main conclusions therefore rely on repaired-state predictive accuracy and directly measured physical violations rather than on auxiliary corruption diagnosis.

Overall, the results identify a consistent multi-objective structure: ConserveFM-Acc favors predictive robustness, ConserveFM-Phys favors physical consistency, and Direct state occupies an intermediate Pareto-optimal regime.

6 Discussion and Limitations

The main empirical finding is that physical repair is genuinely multi-objective. ConserveFM-Acc is preferable under predictive error, whereas ConserveFM-Phys substantially improves all four physical diagnostics. The accuracy-oriented solution can even worsen an individual physical quantity, as illustrated by its advection ratio above one. Similar conflicts between data fidelity and physical constraints have been reported in other physics-informed systems Soltani et al. (2024); He et al. (2023); Jiang et al. (2025).

The instability of the original joint formulation also highlights the importance of scale and optimization. Physical residuals differ substantially in magnitude, so an adaptive weighting mechanism may optimize numerical scale rather than physical relevance. This is consistent with previously reported conflicts between data and physics objectives Yang et al. (2025). Train-derived normalization improves interpretability of the joint objective, but does not eliminate the underlying multi-objective nature of the problem.

ConserveFM uses soft physical objectives and therefore does not guarantee exact constraint satisfaction. Hard projection methods can provide such guarantees for tractable constraint sets Chen et al. (2024); Valente et al. (2025). A natural extension is to combine learned state-dependent repair with hard projection for constraints that admit differentiable feasible sets.

The evaluation has several important limitations. First, the corruption benchmark is synthetic and uses the same corruption families as the additional ConserveFM-Phys training stage. It therefore measures recovery on held-out atmospheric states within specified perturbation families, not general recovery from unknown operational failure modes. Second, the headline normalized comparison is fully evaluated for one training seed; the paired bootstrap quantifies variation across stress conditions, not training stochasticity. Third, Table 1 compares internal repair formulations and uses the corrupted input only as the physical-ratio reference; it does not provide a fully standardized

predictive row for raw, uncorrected ClimaX or an external analytical-projection baseline. We therefore claim an accuracy–physics trade-off among the evaluated repair strategies rather than absolute superiority over all post-processing alternatives.

Finally, the present evaluation is open-loop: repaired states are assessed at fixed 6, 24, and 72 h leads and are not recursively fed back into the frozen backbone. This distinction matters especially because ConserveFM-Acc increases the advection violation ratio. Closed-loop rollout stability, multi-seed final evaluation, empirically derived forecast failures, and direct comparison with tractable hard projections are important next steps.

7 Conclusion

ConserveFM treats physical consistency as a post-hoc repair problem rather than a property that must be learned by the forecasting backbone itself. On 60 matched ERA5 stress conditions, ConserveFM-Acc achieves the strongest predictive robustness (NRMSE 0.3008, ACC 0.9503), while ConserveFM-Phys reduces the aggregate physical-violation ratio to 0.333 at a stress NRMSE of 0.3159. The resulting Pareto structure shows that predictive accuracy and physical consistency are distinct objectives under the evaluated stress protocol. Post-hoc repair can substantially change the physical-consistency profile of a frozen weather forecast without retraining its backbone, but stronger physical enforcement has a measurable and constraint-dependent predictive cost.

AI use statement

Generative AI tools were used only for language editing, including grammar, clarity, and conciseness. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final manuscript.

Reproducibility statement

Code and experimental materials supporting this study, including dataset manifests, inference and evaluation scripts, statistical analysis code, generated outputs, and figure-generation materials, are provided here `https://github.com/cataug/ConserveFM`.

References

Hao Chen, Gonzalo E. Constante Flores, and Can Li. Physics-informed neural networks with hard linear equality constraints. *Computers and Chemical Engineering*, 189, 2024. doi: 10.1016/j.compchemeng.2024.108764.

Jie Deng and Lijia Han. Modified pinn approach integrating conservation laws for efficient multi-stage training of coupled nonlinear systems. *Chinese Physics B*, 35(4), 2026. doi: 10.1088/1674-1056/ae030b.

Alfonso Gijón, Ainhoa Pujana-Goitia, Eugenio Perea, Miguel Molina-Solana, and Juan Gómez-Romero. Prediction of wind turbines power with physics-informed neural networks and evidential uncertainty quantification. *Engineering Applications of Artificial Intelligence*, 164, 2026. doi: 10.1016/j.engappai.2025.113331.

Dezdemona Gjylapi, Alketa Hyso, and Astrit Denaj. Hybrid physics-informed temporal attention model with external physical constraints for pv power forecasting. *Baltic Journal of Modern Computing*, 14(2):356 – 375, 2026. doi: 10.22364/bjmc.2026.14.2.05.

GaoYuan He, YongXiang Zhao, and ChuLiang Yan. A physics-informed generative adversarial network framework for multiaxial fatigue life prediction. *Fatigue and Fracture of Engineering Materials and Structures*, 46(10):4036 – 4052, 2023. doi: 10.1111/ffe.14123.

Archie J. Huang and Shaurya Agarwal. On the limitations of physics-informed deep learning: Illustrations using first-order hyperbolic conservation law-based traffic flow models. *IEEE Open Journal of Intelligent Transportation Systems*, 4:279 – 293, 2023. doi: 10.1109/OJITS.2023.3268026.

Ameya D. Jagtap, Ehsan Kharazmi, and George Em Karniadakis. Conservative physics-informed neural networks on discrete domains for conservation laws: Applications to forward and inverse problems. *Computer Methods in Applied Mechanics and Engineering*, 365, 2020. doi: 10.1016/j.cma.2020.113028.

Zixin Jiang, Xuezheng Wang, and Bing Dong. Physics-informed modularized neural network for advanced building control by deep reinforcement learning. *Advances in Applied Energy*, 19, 2025. doi: 10.1016/j.adapen.2025.100237.

Imad Katiba, Hanae Belmajdoub, and Khalid Minaoui. Deep reinforcement control for precipitation prediction and cloud-seeding decisions in morocco and the iberian peninsula. *IEEE Transactions on Geoscience and Remote Sensing*, 64, 2026. doi: 10.1109/TGRS.2026.3698664.

Wenlong Liao, Shouxiang Wang, Dechang Yang, Zhe Yang, Jiannong Fang, Christian Rehtanz, and Fernando Porté-Agel. Timegpt in load forecasting: A large time series model perspective. *Applied Energy*, 379, 2025. doi: 10.1016/j.apenergy.2024.124973.

Duc Long Nguyen, Duc Tien Du, Xuan Manh Nguyen, Ngoc Tu Nguyen, Khanh Hung Mai, Dinh Quan Dang, Gia Nam Hoang, Anh Duc Tran, Thanh Hung Nguyen, Diep N. Nguyen, Phi Le Nguyen, and Van Khiem Mai. Enhancing tropical cyclone intensity forecasting over the bien dong sea with foundation model and prompt tuning. *Engineering Applications of Artificial Intelligence*, 162, 2025. doi: 10.1016/j.engappai.2025.112567.

Bin Pang, Hongqing Zhang, and Ziyang Hao. Kapnet: A physics-informed neural network with physical constraints modeled by kolmogorov-arnold network for wind speed prediction. *AIP Advances*, 15(11), 2025. doi: 10.1063/5.0292276.

Shuo Shan, Weijing Dou, Victor Sreeram, Chenxi Li, Kai Wang, Kanjian Zhang, and Haikun Wei. Solarfm: Spatio-temporal solar irradiance forecasting based on time series foundation model. *Applied Energy*, 410, 2026. doi: 10.1016/j.apenergy.2026.127532.

Maryam Soltani, Ghasem Akbari, and Nader Montazerin. Trade-off between reconstruction accuracy and physical validity in modeling turbomachinery particle image velocimetry data by physics-informed convolutional neural networks. *Physics of Fluids*, 36(8), 2024. doi: 10.1063/5.0218499.

MinJi Song, YeonJi Chung, Ju-Young Shin, and Myoung-Jin Um. A comparative study on monthly precipitation forecasting using time-series foundation models in south korea. *Engineering Applications of Artificial Intelligence*, 179, 2026. doi: 10.1016/j.engappai.2026.115143.

Matilde Valente, Tiago C. Dias, Vasco Guerra, and Rodrigo Ventura. Physics-consistent machine learning with output projection onto physical manifolds. *Communications Physics*, 8(1), 2025. doi: 10.1038/s42005-025-02329-1.

Tengmao Yang, Zhihao Qian, Nianzhi Hang, and Moubin Liu. S-pinn: Stabilized physics-informed neural networks for alleviating barriers between multi-level co-optimization. *Computer Methods in Applied Mechanics and Engineering*, 447, 2025. doi: 10.1016/j.cma.2025.118348.

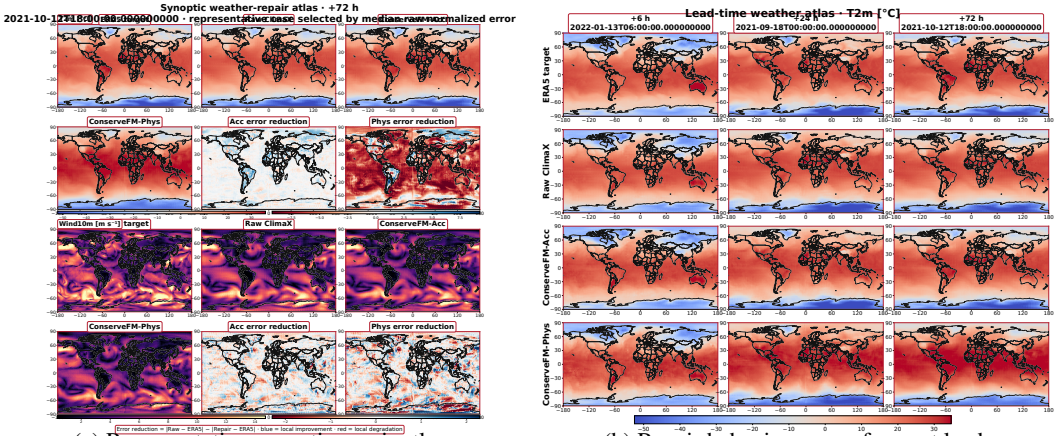
Yuan Yuan, Jingtao Ding, Zhongpu Qiu, Jingfang Fan, and Yong Li. Learning the coupled dynamics of global climate modes. *Nature Machine Intelligence*, 8(6):930 – 941, 2026. doi: 10.1038/s42256-026-01245-5.

Qiang Zheng, Lingzao Zeng, and George Em Karniadakis. Physics-informed semantic inpainting: Application to geostatistical modeling. *Journal of Computational Physics*, 419, 2020. doi: 10.1016/j.jcp.2020.109676.

A Additional Geographic Repair Atlases

This section provides additional spatial diagnostics of ConserveFM on sampled held-out test forecasts. The atlases complement the aggregate results by showing where the repair module modifies

the forecast, where these modifications reduce local standardized prediction error, and how this behavior changes with forecast horizon. The maps are descriptive diagnostics rather than additional stress-benchmark metrics. Spatial error maps use the implementation's common per-grid-cell aggregation over standardized atmospheric channels, so they should be interpreted as aggregate intervention/error patterns rather than as single-variable physical residuals.



(a) Representative synoptic repair atlas (b) Repair behavior across forecast leads

Figure 5: Additional qualitative repair atlases. (a) Representative ERA5 target, raw forecast, and repaired atmospheric fields with corresponding local error changes. (b) Comparison of repair behavior at $+6$, $+24$, and $+72$ h. The lead-time panels are independently selected representative test cases and are not a single autoregressive trajectory.

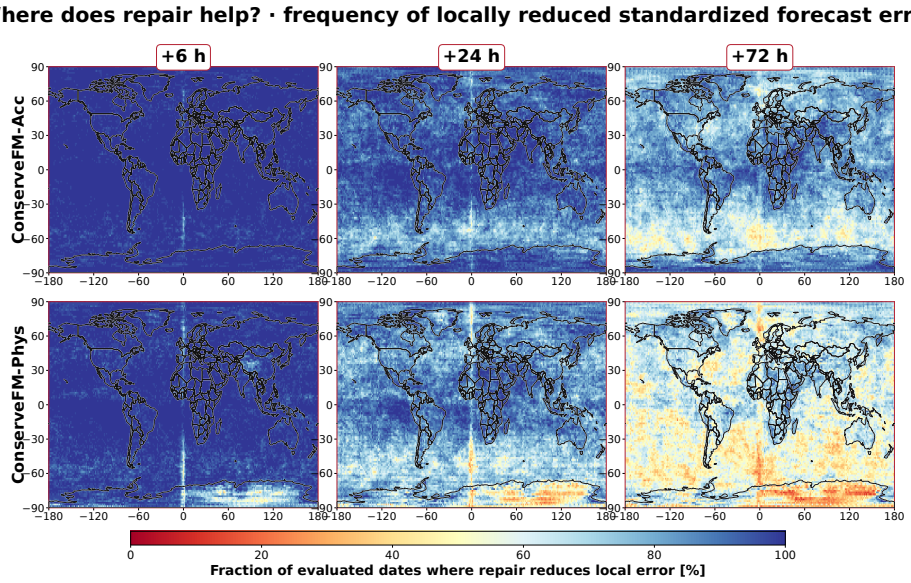


Figure 6: **Geographic frequency of beneficial repair.** For each grid cell, the atlas reports the fraction of sampled held-out test cases in which the repaired forecast has lower local standardized error than the raw forecast, shown across repair variants and forecast leads.

These atlases show that the learned corrections are spatially structured and lead-dependent rather than uniform global offsets. Importantly, regions with large interventions do not necessarily coincide with regions of the largest forecast-error reduction.

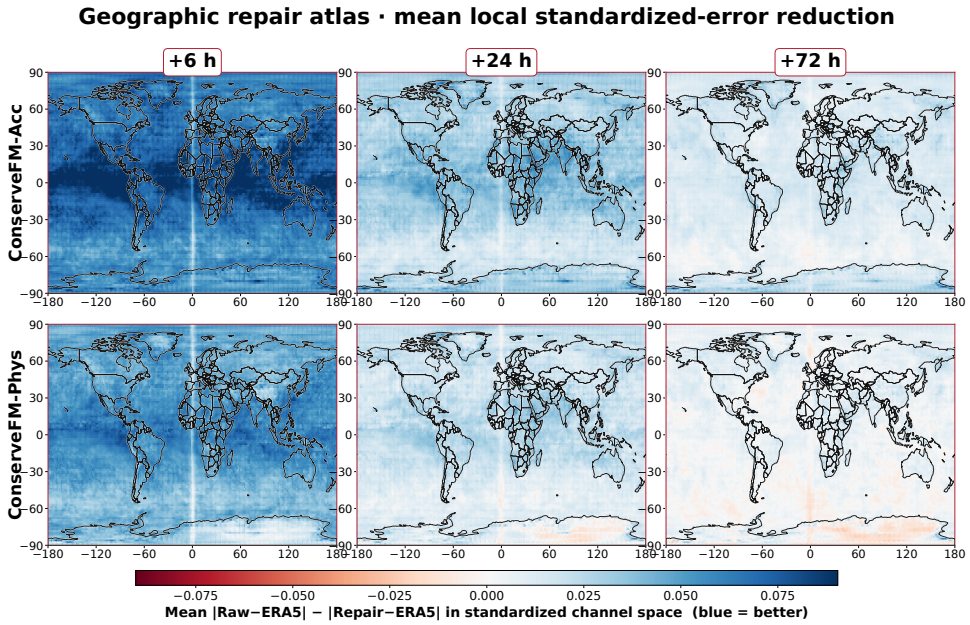


Figure 7: **Mean geographic error reduction.** Positive values indicate locations where post-hoc repair decreases the mean local standardized forecast error relative to the frozen backbone. This figure should be read jointly with Fig. 8, since large interventions need not imply large error reductions.

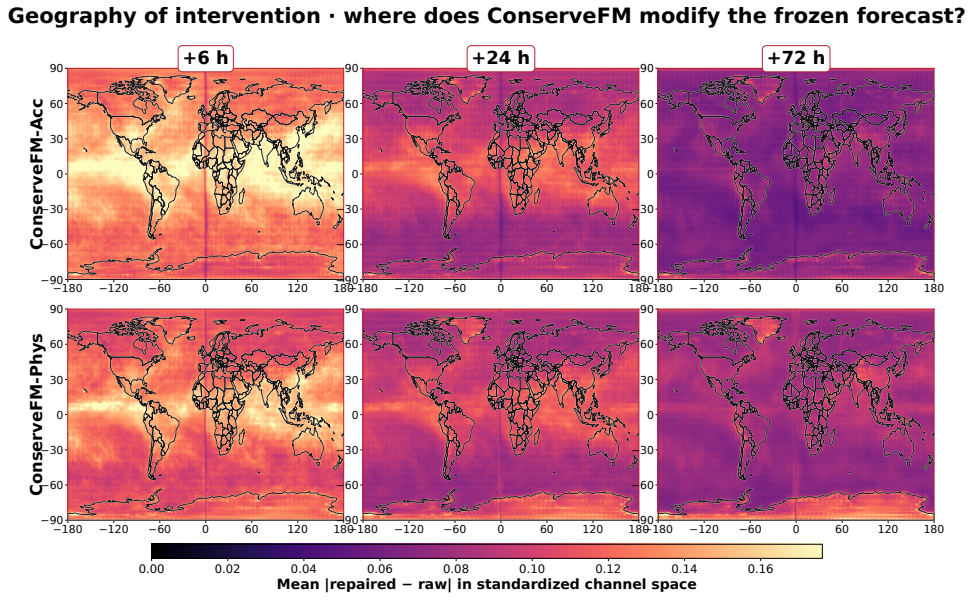


Figure 8: **Geographic magnitude of intervention.** The atlas shows the mean absolute difference between repaired and raw forecasts, separating where ConserveFM acts from whether the resulting correction improves forecast accuracy.