PHYSICS-INFORMED MACHINE LEARNING UNDER CLIMATE DOMAIN SHIFT: PDE-FREE PHYSICS REGULARISATION FOR CLOUD PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study out-of-distribution generalisation in geophysical prediction and propose CC-PINN, a physics-informed multi-layer perceptron (MLP) that encodes the Clausius–Clapeyron thermodynamic relation as a gradient-based regularisation term. Unlike prior PINNs, CC-PINN requires no explicit governing-equation. CC-PINN introduces a lightweight constraint on humidity-temperature consistency without altering network architecture. Trained on atmospheric reanalysis data (temperature, pressure, relative humidity, specific humidity, vertical velocity) using modest computational resources, CC-PINN matches a capacity-matched MLP in-distribution and improves out-of-distribution performance. CC-PINN achieves a 12.3% reduction in global area-weighted RMSE over a capacity-matched MLP baseline. Under a stringent covariate-shift test - training only on the polar latitudes - CC-PINN reduces tropical area-weighted root mean squared error (RMSE) by 22.6% relative to the baseline, while maintaining in-distribution parity. Ablations show the performance gains are substantially attenuated when the physics term is removed, highlighting the role of targeted domain knowledge inclusion in improving extrapolation. These findings suggest that compact, domain-motivated regularisation can deliver robust generalisation in scientific ML tasks.

1 INTRODUCTION

Out-of-distribution (OOD) generalisation remains a fundamental challenge in machine learning, where models, often trained under one data regime are routinely deployed in others, (Koh et al., 2021). In climate modelling, these shifts are not rare anomalies but the norm: evolving climate states can produce input distributions that differ significantly from historical training data, (Beucler et al., 2024). Standard neural networks excel at in-distribution prediction but frequently overfit to spurious correlations, resulting in degraded performance when faced with new regimes, (Arjovsky et al., 2020).

Cloud fraction prediction is a critical example of this problem. Global circulation models (GCMs) cannot explicitly resolve the small-scale processes that govern cloud formation, so they rely on parameterisations—empirical or semi-empirical approximations that introduce uncertainty into climate projections (Stephens, 2005; Bony et al., 2006; Smith, 1990a; Tiedtke, 1993). Recent work has explored replacing or augmenting these schemes with machine-learning surrogates trained on reanalysis or high-resolution simulation data (Rasp et al., 2018; Brenowitz & Bretherton, 2019). While such models can achieve low error in familiar conditions, they often fail to maintain accuracy when extrapolating to different climatic zones or unseen meteorological states (Dueben & Bauer, 2018; Beucler et al., 2024).

We address this gap by introducing CC-PINN, a physics-informed MLP that incorporates the Clausius–Clapeyron (CC) thermodynamic relation directly into the loss function as a gradient-based regularisation term (Raissi et al., 2019). The CC relation governs the dependence of saturation vapour pressure on temperature, a fundamental driver of cloud formation (Wallace & Hobbs, 2006). The constraint nudges predictions to vary with temperature and humidity in the CC-consistent *direction* (qualitative coupling), rather than enforcing a strictly units-exact magnitude. By encoding this as

a soft constraint, our approach enforces physically consistent humidity–temperature relationships without restricting the network architecture or requiring explicit PDE supervision.

Using ERA5 pressure-level data (Hersbach et al., 2020), we evaluate two protocols: (i) *Global in-distribution* (random global split), and (ii) *Polar→Tropics OOD*, where models train and validate on polar latitudes only ($|\phi| \geq 66.33°$) and are tested in the tropics ($|\phi| \leq 23.5°$); midlatitudes are reported as a secondary OOD band (Stocker et al., 2013). We report *area-weighted* RMSE with cosine-latitude weights, a standard metric for model evaluation (Gleckler et al., 2008), and aggregate results over twenty five random seeds, i.e., different random initialisations of network weights and shuffles of training data Our contributions are:

- A minimal-intrusion physics regulariser based on the CC relation, applicable to any feed-forward architecture without modifying the forward pass.
- A systematic OOD benchmark for cloud fraction prediction using ERA5 reanalysis data, in which all tropical and mid-latitude data (-66.33° to 66.33°) is excluded from training to induce a strong covariate shift.
- Compared with a capacity-matched MLP (i.e., a network with the same architecture and number of parameters as CC-PINN), CC-PINN preserves global in-distribution parity (it performs equally well on test data drawn from the training spatial distribution) and reduces area-weighted RMSE by $\sim$12.3% globally and $\sim$22.6% in the tropics under Polar→Tropics OOD.
- Ablation studies demonstrating that the OOD gains are no longer present when the physics term is removed, highlighting the role of the inductive bias (i.e., a built-in modeling assumption guiding the learning). This indicates that improvement stems from the CC-guided bias rather than capacity or sampling artefacts.

By framing cloud fraction prediction as a case study in physics-guided OOD generalisation, our work contributes to the broader ML discourse on targeted inductive biases. These results indicate that small, well-chosen physics constraints can materially improve generalisation in complex, data-limited scientific tasks, with implications for other domains facing similar challenges.

2 Related Work

Physics-informed neural networks. Physics-informed neural networks (PINNs) embed physical knowledge during training, classically by penalising residuals of governing equations (Raissi et al., 2019). Alongside full PDE-residual supervision, lighter-weight constraints have been used to stabilise learning and improve robustness in scientific settings without modifying model architecture (Beucler et al., 2021; Karpatne et al., 2017). We follow this latter line: our method adds a *Clausius–Clapeyron (CC)-guided* gradient constraint to a capacity-matched MLP, providing a soft inductive bias rather than enforcing a units-exact equality.

Machine-learning components in climate models. A growing literature replaces or augments parameterisations in GCMs/ESMs with ML surrogates (Rasp et al., 2018; Brenowitz & Bretherton, 2019; Yuval & O'Gorman, 2020). These studies report strong in-distribution skill for subgrid processes (e.g., moist convection, boundary-layer turbulence), yet robustness can degrade under regime shifts or altered forcings (Dueben & Bauer, 2018; Beucler et al., 2024). This motivates physics-guided inductive biases that help retain fidelity when transferring across climatic regimes.

Cloud-fraction modelling. Cloud fraction is a long-standing source of uncertainty in climate projections (Stephens, 2005). Cloud prediction in climate models is critical not only for estimating cloud cover itself, but also because clouds have strong influences on climate through shortwave (SW) and longwave (LW) radiation feedbacks (Ramanathan et al., 1989; Bony et al., 2015a; Forster et al., 2021). Classical schemes in GCMs (Sundqvist et al., 1989; Smith, 1990b) are physically motivated but require empirical tuning and can exhibit regime-dependent biases, notably in the tropics and marine stratocumulus (Hourdin et al., 2017; Nam et al., 2012). Data-driven surrogates that learn cloud fraction from reanalysis or high-resolution simulations (Krasnopolsky et al., 2013; Yuval & O'Gorman, 2021) reduce heuristic assumptions but, without explicit thermodynamic constraints,

may reproduce non-physical behaviour or overfit dominant correlations in the training distribution, (Dueben & Bauer, 2018).

OOD under climate change: parallels to our set-up. Assessing whether a learned cloud scheme will remain reliable under warming typically requires testing out of the training distribution. In climate terms, forced warming shifts the joint distribution of temperature and humidity, with CC-implied increases in saturation vapour pressure and moisture availability (Held & Soden, 2006). Our Polar→Tropics protocol is not a future-scenario emulator, but it is a purposeful analogue: by training only on polar latitudes and evaluating in the tropics, we expose the model to thermodynamic states (higher T, higher q, different RH structure) that lie outside the training envelope - precisely the kind of regime shift that challenges cloud schemes in warmer climates (Bony et al., 2015b). This thermodynamics-first OOD test isolates the humidity–temperature coupling that CC highlights, while holding other confounders fixed. As such, it complements hybrid/online evaluations and provides a controlled proxy for climate-change robustness (Koh et al., 2021).

Thermodynamic constraints and positioning. Recent work incorporates conservation and thermodynamic relationships to encourage physical consistency in atmospheric ML (Beucler et al., 2021; Yuval & O'Gorman, 2021). To our knowledge, however, the CC relation has not been used as a *differentiable, CC-guided gradient constraint* specifically for cloud-fraction emulation. Our contribution is to introduce such a minimal-intrusion constraint—leaving architecture and capacity unchanged—and to evaluate robustness under a challenging regime transfer. To avoid common evaluation artefacts on latitude–longitude grids, details on non-overlapping (leakage-resistant) splits and area-weighted metrics are given with the experimental protocol.

3 METHOD

3.1 PROBLEM FORMULATION

Let $x_i = (T_i, RH_i, q_i, \omega_i, p_i) \in \mathbb{R}^5$ denote ERA5 predictors (outlined in Table 1) at sample i and let $c_i \in [0, 1]$ be cloud fraction on pressure levels (Hersbach et al., 2020). We learn $f_\theta : \mathbb{R}^5 \to [0, 1]$ with prediction $\hat{c}_i = f_\theta(x_i)$. Our design goals are: (i) *simplicity and fairness* via a capacity-matched MLP baseline. Fairness meaning the only systematic difference between models is the cc-slope regulariser; all other factors constant or symetrically tuned. And (ii) *robustness under thermodynamic regime shift*. The dataset and metrics are detailed in section 4.

3.2 MODELS

Baseline MLP. A fully connected network with three hidden layers of 11 ReLU units and a sigmoid output producing $\hat{c} \in [0, 1]$. We chose 3×11 architecture via preliminary hyper-parameter tuning; other sizes like 10 or 12 neurons yielded similar validation performance. Full details of the MLP architecture and training in Appendix Tables A1, A2.

CC-PINN. Identical architecture; the only change is an additional *gradient supervision* term that aligns the model's temperature sensitivity with the Clausius–Clapeyron (CC) slope.

3.3 CC-SLOPE MATCHING (GRADIENT SUPERVISION)

We use the standard CC relation (Wallace & Hobbs, 2006) for saturation vapour pressure $e_s(T)$:

$$\frac{de_s}{dT} \;=\; \frac{L_v\, e_s}{R_v\, T^2}, \tag{1}$$

and define a per-sample residual that matches the network's partial derivative with respect to temperature (holding the other inputs fixed) to the CC slope:

$$r \;=\; \frac{\partial \hat{c}}{\partial T}\Big|_{RH,\,q,\,\omega,\,p} \;-\; \frac{de_s}{dT}, \qquad L_{\text{phys}} \;=\; r^{\,2}. \tag{2}$$

We compute $\left(\partial\hat{c}/\partial T\right)\big|_{RH,q,\omega,p}$ via automatic differentiation with RH, q, ω, and p treated as constants at their sample values (i.e., no graph path from T into those tensors). This yields the intended *partial* derivative.

Although de_s/dT (Pa K^{-1}) and $\partial\hat{c}/\partial T$ (K^{-1}) carry different units, this only sets the overall scale of the penalty: any fixed unit conversion (e.g., dividing de_s/dT by a reference pressure e_0) is a constant rescaling that α (a tunable weight controlling the strength of the physics term) absorbs. If $s \equiv de_s/dT$ and we replace s by κs for any fixed $\kappa > 0$, the term becomes $\alpha\,(g - \kappa s)^2 = \alpha\,\kappa^2\,(g/\kappa - s)^2$ with $g \equiv \partial\hat{c}/\partial T$. Thus the absolute scale of the CC slope only sets the effective weight of the constraint; tuning α compensates for κ.[1]

As inputs are min–max normalised, $X' = (X - X_{\min})/(X_{\max} - X_{\min})$, we convert network Jacobians to physical units before equation 2.

3.4 Training objective

We minimise a normalised, area-weighted objective over minibatch $\mathcal{B}$:

$$L \;=\; \frac{1}{\sum_{i\in\mathcal{B}} w_i} \sum_{i\in\mathcal{B}} w_i \Big((c_i - \hat{c}_i)^2 \;+\; \alpha\, L_{\text{phys},i} \Big), \qquad w_i = \cos\phi_i, \tag{3}$$

where ϕ_i is latitude (in radians). Using the same weighting at training and evaluation so the optimisation objective matches the evaluation objective, avoiding train–eval metric mismatch ('metric drift').

3.5 Optimisation

We use Adam (Kingma & Ba, 2015) with early stopping on validation error. The learning rate, physics weight α, dropout rate, and batch size are tuned by Bayesian optimisation (Akiba et al., 2019) on the train-validation split, see Appendix Table A2. All other hyperparameters are fixed and shared across models, seen in Appendix Table A1. We run twenty five seeds and report mean±SEM.

Note on scale invariance. Constant rescaling of the CC target (e.g., using $d\tilde{e}_s/dT = (1/e_0)\,de_s/dT$) only changes the effective weight; the tuned α compensates for this scaling.

4 Data and Experimental Protocol

4.1 ERA5 variables and units

We use ERA5 reanalysis on pressure levels at $0.25^\circ \times 0.25^\circ$ resolution. Inputs are air temperature T (K), relative humidity RH (%), specific humidity q (kg kg^{-1}), pressure vertical velocity ω (Pa s^{-1}), and the pressure level index p (hPa). The target $c \in [0, 1]$ is ERA5 cloud fraction on pressure levels.

Table 1: ERA5 predictors and target used in this study. Features are min–max normalised using *training-split only* statistics; target remains dimensionless (fraction).

Symbol	ERA5 short name	Unit	Level	Notes
T	`t`	K	pressure	Air temperature
RH	`r`	%	pressure	Relative humidity (0–100)
q	`q`	kg kg^{-1}	pressure	Specific humidity
ω	`w`	Pa s^{-1}	pressure	Pressure vertical velocity
p	`level`	hPa	pressure	Pressure level identifier[2]
c (target)	`cc`	[0,1]	pressure	Cloud fraction (fraction)

[1] This argument is at the loss level; optimiser details can change step sizes, but cross-validated α reliably absorbs constant rescalings of s.

[2] Pressure levels used in this work: {1000, 975, 950, 925, 900, 875, 850, 825, 800, 775, 750, 700, 650, 600, 550, 500, 450, 400, 350, 300, 250, 225, 200, 175, 150, 125, 100, 70, 50, 30, 20, 10, 7, 5, 3, 2, 1} hPa.

4.2 TEMPORAL SET-UP

To enforce temporal disjointness, we use a single timestamp for training/validation and a distinct timestamp for testing (UTC):

$$\text{Train/val: } \mathbf{2024\text{-}08\text{-}01\ 14{:}00} \quad \text{and} \quad \text{Test: } \mathbf{2024\text{-}12\text{-}12\ 09{:}00}.$$

This induces both diurnal and seasonal shifts (boreal summer $\rightarrow$ boreal winter) without mixing times across splits.

4.3 LATITUDE BANDS AND REGIONS OF ASSESSMENT

We report both *global* and *band-wise* scores to expose regime dependence. Unless stated otherwise, band edges are the canonical

$$\textbf{Tropics: } |\phi| \le 23.5^\circ, \qquad \textbf{Midlatitude: } 23.5^\circ < |\phi| < 66.33^\circ, \qquad \textbf{Polar: } |\phi| \ge 66.33^\circ,$$

with ϕ the geographic latitude. Band-wise metrics restrict the evaluation set $\mathcal{S}$ to the band and *renormalise* cosine weights within that band (Eq. 5), yielding $\mathrm{RMSE}_w^{(\text{band})}$. This breakdown highlights thermodynamic contrasts (e.g., higher T/q structure in the Tropics) and complements the global score.

4.4 PREPROCESSING AND NORMALISATION

Per-feature min–max scalers are fit on the *training split only* and then frozen:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \tag{4}$$

where $(X_{\min}, X_{\max})$ are computed on training data from 2024-08-01 14:00. We pass RH internally as a fraction in $[0, 1]$ (not %) for consistent gradient units.

4.5 SAMPLING AND SPLITS (LEAKAGE-ROBUST)

Each sample is a unique spatio–temporal–pressure coordinate (ϕ, λ, p, t) (latitude, longitude, pressure level, time). Exact duplicate coordinates are removed prior to splitting. We split *by coordinates* so no (ϕ, λ, p, t) appears in more than one split:

- Train/validation (temporal slice): all coordinates with $t =$ 2024-08-01 14:00 are grouped and partitioned 80/20 for train/val.
- Test (temporal hold-out): all coordinates with $t =$ 2024-12-12 09:00 form the test set and are never used for tuning, early stopping, or normaliser fitting.

When sub-sampling for efficiency, we sample without replacement *within* each temporal slice to preserve the grouped constraint. For band-wise reporting, we also aggregate test errors over latitude bands (Tropics, Midlatitude, Polar) without altering the temporal split.

4.6 MODEL SELECTION AND REPORTING

Model selection (early stopping; Bayesian optimisation of learning rate and α) is performed on the validation split from 2024-08-01 14:00 only. No December keys are used for tuning or normalisation. All results are reported as mean$\pm$standard error of the mean (SEM) over twenty five random seeds.

4.7 METRICS

Because a latitude–longitude grid over-represents high latitudes, we report *area-weighted* RMSE with cosine-latitude weights. Let $\mathcal{B}$ denote the evaluation set (global or a latitude band) and ϕ_i the latitude of sample i in radians. Define weights:

$$w_i \;=\; \cos(\phi_i), \tag{5}$$

and area-weighted RMSE

$$\mathrm{RMSE}_w = \sqrt{\sum_{i\in\mathcal{B}} w_i \left(c_i - \hat{c}_i\right)^2}. \tag{6}$$

We use the same weighting scheme in the training loss to align optimisation and evaluation.

5 RESULTS

Here We evaluate two protocols: (i) **Global train** (Temporal split: train/val at 2024-08-01 14:00; test at 2024-12-12 09:00) and (ii) **Polar train** (Spatial OOD variant: same temporal split but train/val restricted to $|\phi| \geq 66.33°$). Metrics are *area-weighted* RMSE and standard error of the mean (mean±SEM over 25 seeds) with cosine-latitude weights.

In Fig.1, we report area-weighted RMSE (lower is better) by **test region** for each **model/training regime**. This presentation makes both central tendency and variability across random initialisations/data shuffles explicit, and allows direct comparison of in-distribution and out-of-distribution evaluations.

5.1 GLOBAL TRAIN (TEMPORAL SPLIT): PARITY WITH THE BASELINE

On the December test timestamp, the global trained CC-PINN and the baseline are statistically comparable. Globally, CC-PINN achieves a predictive error of 0.1010 ± 0.0002 vs $0.1033\pm0.0.0005$ for the baseline. Band-wise: Tropics 0.0873 ± 0.0002 vs 0.0883 ± 0.0004, Midlatitude 0.0992 ± 0.0002 vs 0.1017 ± 0.0005, and Polar 0.1528 ± 0.0006 vs 0.1566 ± 0.0012). Overall, the CC term maintains in-distribution parity, and demonstrates minor improvements in accuracy and spread under the temporal shift.

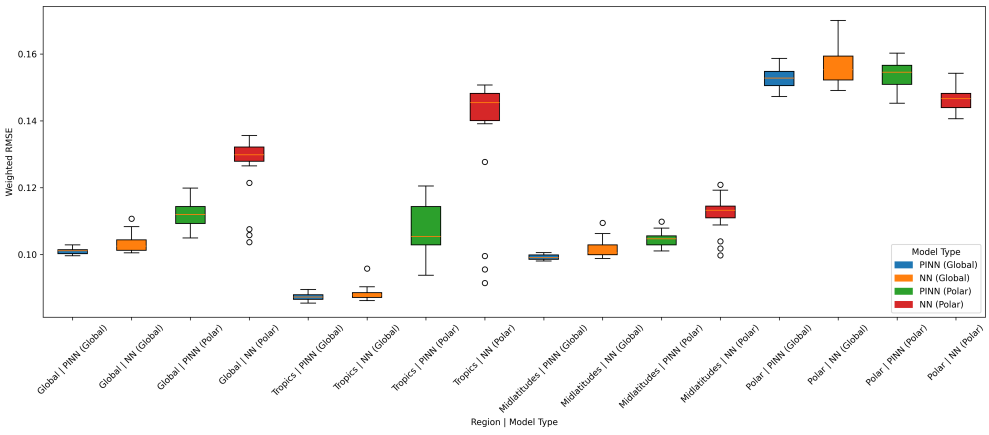

Figure 1: Area-weighted RMSE (lower is better) by **test region** for each **model/training regime**. Notation: **`TestRegion | Model(TrainRegion)`**; e.g., `Tropics | PINN(Global)` denotes a PINN trained on Global and evaluated on the Tropics. Small circles are seed-wise results ($n = 25$). Boxes show the median (line) and interquartile range (box); whiskers use $1.5 \times$ IQR.

5.2 POLAR TRAIN (SPATIAL OOD): CC-PINN IMPROVES TRANSFER

When trained only on Polar latitudes and evaluated globally in December, CC-PINN improves the global score from 0.1275 ± 0.0018 to 0.1119 ± 0.0008 (∼**12.3%** relative reduction). The gains concentrate where the thermodynamic shift is largest:

- **Tropics:** $0.1391\pm0.0034 \rightarrow 0.1076\pm0.0015$ (∼**22.6%** lower RMSE).
- **Midlatitude:** $0.1125\pm0.0010 \rightarrow 0.1045\pm0.0004$ (∼**7.1%** lower).

- **Polar:** a modest degradation $0.1468 \pm 0.0008 \rightarrow 0.1535 \pm 0.0007$ ($\sim$4.6% higher), consistent with a trade-off that prioritises correct temperature sensitivity in warmer/moister regimes.

These results indicate that *aligning $\partial \hat{c}/\partial T$ with the CC slope* meaningfully improves extrapolation to thermodynamically distinct states.

5.3 Significance of improvements

Polar–train global reduction (0.1275→0.1119) is significant by a two-sample Welch t-test, $t(32.62) = -8.17$ (two-sided $p = 2.17 \times 10^{-9}$); the PINN mean RMSE is lower than the NN, with a large effect (Cohen's $d \approx 2.31$).

5.4 Gradient alignment with CC

We verify that gradient supervision changes the model's temperature sensitivity in the intended direction. Define a tolerance-aware sign function

$$\mathrm{sign}_\tau(x) = \begin{cases} -1, & x \le -\tau, \\ 0, & |x| < \tau, \\ 1, & x \ge \tau, \end{cases}$$

and let $g_i = \frac{\partial \hat{c}}{\partial T}\big|_i$ and $s_i = \frac{de_s}{dT}\big|_i$. We report the directional agreement (area-weighted)

$$\mathrm{DA}_w = \frac{\sum_{i \in \mathcal{S}} w_i \, \mathbf{1}\big[\mathrm{sign}_{\tau_g}(g_i) = \mathrm{sign}_{\tau_s}(s_i)\big]}{\sum_{i \in \mathcal{S}} w_i},$$

with $w_i = \cos \phi_i$ and small tolerances τ_g, τ_s to treat near-zero pairs as agreement.

and the Spearman correlation ρ between $\partial \hat{c}/\partial T$ and de_s/dT across samples. Under **Polar train**, CC-PINN improves DA and ρ in the Tropics and Midlatitude (Table 2), indicating the loss steers sensitivities toward CC-consistent behaviour.

Table 2: Clausius–Clapeyron alignment under **Polar train**: area-weighted directional agreement (DA_w) and Spearman $\rho(g, s)$ (mean $\pm$ SEM over 25 seeds. Higher is better.

Region	PINN DA_w	PINN ρ	NN DA_w	NN ρ
Tropics	0.554 ± 0.056	0.191 ± 0.035	0.260 ± 0.057	0.034 ± 0.043
Midlatitude	0.647 ± 0.053	0.135 ± 0.038	0.385 ± 0.055	-0.008 ± 0.043
Polar	0.613 ± 0.050	0.125 ± 0.038	0.405 ± 0.050	-0.017 ± 0.039
Global	0.613 ± 0.051	0.154 ± 0.035	0.359 ± 0.053	0.013 ± 0.042

Table 3: Area-weighted RMSE (mean $\pm$ SEM over 25 seeds.). Bands: Tropics $|\phi| \le 23.5°$, Midlatitude $23.5° < |\phi| < 66.33°$, Polar $|\phi| \ge 66.33°$. Lower is better.

	Global train		Polar train	
Region	**PINN**	**NN**	**PINN**	**NN**
Tropics	0.0873 ± 0.0002	0.0883 ± 0.0004	0.1076 ± 0.0015	0.1391 ± 0.0034
Midlatitude	0.0992 ± 0.0002	0.1017 ± 0.0005	0.1045 ± 0.0004	0.1125 ± 0.0010
Polar	0.1528 ± 0.0006	0.1566 ± 0.0012	0.1535 ± 0.0007	0.1468 ± 0.0008
Global	0.1010 ± 0.0002	0.1033 ± 0.0005	0.1119 ± 0.0008	0.1275 ± 0.0018

5.5 Where do gains arise? Stratification by temperature

We stratify the December test set into equal-area temperature bins and report area-weighted RMSE within each bin. Contrary to the simplest "warmer ⇒ larger gains" expectation, the largest relative reductions occur in the coldest bins, while improvements are modest in the warmest bin. This pattern is consistent with the dual role of our CC-slope term: (i) in cold regimes where $de_s/dT \approx 0$,

Table 4: Temperature–stratified errors using **equal-weight** bins (bin centres shown, in K). RMSE is computed within each bin; $\Delta\% = 100\,(\mathrm{RMSE_{NN}} - \mathrm{RMSE_{PINN}})/\mathrm{RMSE_{NN}}$ (higher is better).

Bin centre (K)	CC-PINN RMSE	NN RMSE	$\Delta\%$
202.69	0.1042	0.1339	22.2
228.85	0.1039	0.1359	23.6
250.52	0.1278	0.1319	3.1
271.41	0.1179	0.1247	5.4
297.53	0.1030	0.1082	4.8

it damps spurious temperature sensitivity by nudging $\partial\hat{c}/\partial T \to 0$, yielding sizable error reductions; (ii) in warm regimes — often dry, the baseline already attains low RMSE and cloud fraction is more dynamics-limited than thermodynamically limited, so aligning the temperature sensitivity offers limited additional benefit. Importantly, the warmest bin's *absolute* errors are already small, so relative gains naturally appear muted even when absolute gaps are comparable across bins. Overall, the stratified view reinforces that CC-slope supervision improves robustness where T-sensitivity is most error-prone (cold/mid bins) while preserving parity in warm states. Full results are shown in Table 4

Removing the CC-slope term (α=0) erases OOD gains (Polar trained Tropics and Midlatitude), confirming that improvements derive from the gradient supervision rather than capacity or sampling.

5.6 Training stability

CC-PINN reduces variance across seeds in the OOD setting (e.g., Tropics SEM 0.0015 vs 0.0034; Global SEM 0.0008 vs 0.0018 under Polar train), suggesting the CC term regularises optimisation by discouraging spurious temperature responses.

6 Discussion

Key finding. A single, lightweight CC–slope supervision (Eq. 2) materially improves extrapolation to thermodynamically distinct states. Under **Polar train**, CC-PINN reduces *global* RMSE from 0.1275 ± 0.0018 to 0.1119 ± 0.0008 ($\sim$12.3%), with the largest gain in the *Tropics* ($0.1391\pm0.0034 \to 0.1076\pm0.0015$; $\sim$22.6%). Under **Global train** the two models are statistically comparable, indicating the constraint does not harm in-distribution performance.

Warm-bin trade-off. Our temperature-stratified results show the largest gains in cold bins and near parity in the warmest bin, which aligns with CC-slope supervision acting chiefly as a guardrail against spurious T-sensitivity in regimes where de_s/dT is small, and offering limited headroom where baseline errors are already low and clouds are dynamics-dominated.

6.1 Implications for OOD generalisation

These results support the broader claim that *task-relevant* inductive biases can curb spurious correlations that otherwise limit transfer. Here, aligning $\partial\hat{c}/\partial T$ to the CC slope produces gains concentrated in warm/moist regimes—the very states where CC effects amplify humidity availability—while leaving the temporal split essentially unchanged. This specificity matters: the constraint targets the mechanism most likely to shift under warming, rather than imposing a broad architectural prior.

6.2 Relevance to climate modelling

For climate applications, the constraint is *deployable*: it leaves the forward pass untouched, adds negligible overhead, and can be switched on/off via a single weight α. As such, it is compatible with hybrid deployments (residual correction or emulator settings). That said, our present evaluation is *offline* and uncoupled; online stability and conservation in a prognostic model remain to be tested.

A practical next step is single-column online coupling (e.g., SCM or aquaplanet), where the CC-guided sensitivity can be stress-tested under feedbacks with radiation and dynamics (Brenowitz & Bretherton, 2019; Rasp et al., 2018).

6.3 Methodological strengths

(i) *Capacity fairness:* identical depth/width across models isolates the effect of the physics term. (ii) *Leakage-robust evaluation:* grouped splits over (ϕ, λ, p, t) with train-only normalisation, and area-weighted metrics. (iii) *Two axes of shift:* a temporal split (August→December) and a spatial OOD variant (Polar train), showing parity in one and gains in the other. (iv) *Compute efficiency:* all runs are desktop-trainable, supporting reproducibility.

6.4 Limitations and threats to validity

- **Target fidelity.** ERA5 cloud fraction is model-derived; learned biases may reflect the reanalysis operator. Observation-based targets (e.g., CloudSat/Calipso or satellite retrievals (Stephens et al., 2002; Winker et al., 2010)) would strengthen external validity.
- **Temporal coverage.** Results use two timestamps (14:00 Aug 1 and 09:00 Dec 12, 2024). Broader diurnal/seasonal sampling would test robustness to synoptic diversity.
- **Metric narrowness.** We focus on RMSE; calibration and error asymmetry (e.g., under/over-clouding) are not assessed.

6.5 Broader impact

Physics-guided losses like the CC term provide a simple bridge between domain knowledge and ML robustness. Because they are transparent, tunable, and architecture-agnostic, they are accessible to groups without large compute budgets and can foster more trustworthy ML components in climate workflows. Care is still required: improvements here do not guarantee safe behaviour in coupled models or extremes; rigorous online testing and calibration are prerequisites for operational use.

Take-away. A minimal, interpretable constraint on temperature sensitivity—implemented as gradient supervision—delivers measurable OOD gains where they matter most, without architectural changes. This pattern (soft, mechanism-targeted physics guidance) is a pragmatic path for advancing robust scientific ML.

7 Conclusion

We introduced CC-PINN, which embeds the Clausius–Clapeyron (CC) law as a gradient-based supervision term on temperature sensitivity. With no architectural changes and negligible overhead, this *CC-slope* constraint preserves parity under a temporal split while improving extrapolation under a spatial OOD stress test: under **Polar train**, global RMSE drops from 0.1275 ± 0.0018 to 0.1119 ± 0.0008 ($\sim$12.3%), with the largest gain in the **Tropics** ($0.1391 \pm 0.0034 \rightarrow 0.1076 \pm 0.0015$; $\sim$22.6%). A modest Polar-band trade-off is observed, consistent with prioritising correct warm-state sensitivity.

Takeaways. (1) A small, task-relevant physics prior can cut OOD error without degrading in-distribution performance, with benefits concentrated in regimes where CC effects are strongest. (2) Gains are robust across seeds and capacity controls; they stem from the gradient supervision rather than model size or sampling, and are insensitive to constant rescalings of the CC target (absorbed by α). (3) The approach is lightweight, architecture-agnostic, and compute-efficient, making it practical for broader scientific ML use.

Future work. Extend to multi-constraint objectives (e.g., simple moisture/energy closures or monotonicity in RH), evaluate against observation-based cloud products across more times/years and report calibration, and test online in single-column/GCM settings to assess stability, conservation, and long-horizon forecast skill.

Reproducibility Statement.

We specify the dataset, variables, and units in Sec. 4.1; exact timestamps and latitude-band definitions in Secs. 4.2–4.3; preprocessing and normalisation in Sec. 4.4; leakage-robust sampling/splits and model-selection protocol in Secs. 4.5–4.6; and the area-weighted RMSE metric (cosine-latitude weights) in Sec. 4.7 with formulas in Eqs. (5)–(6). Model architectures, losses, and optimisation are detailed in Sec. 3 (CC slope Eq. (1), gradient supervision Eq. (2), training objective Eq. (3); optimiser/early stopping in Secs. 3.4–3.5), with fixed settings and tuned hyperparameters in Appendix Tables A1–A2. We report mean ± SEM over 25 seeds, use a temporal hold-out test set unseen by tuning, and provide significance tests (Sec. 5.3), gradient-alignment diagnostics (Sec. 5.4), temperature-stratified analyses (Sec. 5.5), and stability notes (Sec. 5.6). ERA5 acquisition is reproducible via the Copernicus Climate Data Store using the variable names, levels, and exact times listed in Secs. 4.1–4.2.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019. doi: 10.1145/3292500.3330701.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Tom Beucler, Michael S. 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.

Tom Beucler, Pierre Gentine, Janni Yuval, Ankitesh Gupta, Liran Peng, Jerry Lin, Sungduk Yu, Stephan Rasp, Fiaz Ahmed, Paul A. O'Gorman, J. David Neelin, Nicholas J. Lutsko, and Michael Pritchard. Climate-invariant machine learning. *Science Advances*, 10(6):eadj7250, 2024. doi: 10.1126/sciadv.adj7250.

Sandrine Bony, Robert Colman, Vladimir M. Kattsov, Richard P. Allan, Christopher S. Bretherton, Jean-Louis Dufresne, Alex Hall, Stéphane Hallegatte, Melinda M. Holland, William Ingram, David A. Randall, Brian J. Soden, George Tselioudis, and Mark J. Webb. How well do we understand and evaluate climate change feedback processes? *Journal of Climate*, 19(15):3445–3482, 2006. doi: 10.1175/JCLI3819.1.

Sandrine Bony, Bjorn Stevens, Dargan M. W. Frierson, Christian Jakob, Masa Kageyama, Robert Pincus, Theodore G. Shepherd, Steven C. Sherwood, A. Pier Siebesma, Adam H. Sobel, Masahiro Watanabe, and Mark J. Webb. Clouds, circulation and climate sensitivity. *Nature Geoscience*, 8: 261–268, 2015a. doi: 10.1038/ngeo2398.

Sandrine Bony, Bjorn Stevens, Dargan M. W. Frierson, Christian Jakob, Masa Kageyama, Robert Pincus, Theodore G. Shepherd, Steven C. Sherwood, A. Pier Siebesma, Adam H. Sobel, Masahiro Watanabe, and Mark J. Webb. Clouds, circulation and climate sensitivity. *Nature Geoscience*, 8: 261–268, 2015b. doi: 10.1038/ngeo2398.

Noah D. Brenowitz and Christopher S. Bretherton. Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, 11 (8):2728–2744, 2019. doi: 10.1029/2019MS001711.

Peter D. Dueben and Peter Bauer. Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, 11(10):3999–4009, 2018. doi: 10.5194/gmd-11-3999-2018.

P. Forster, T. Storelvmo, K. Armour, W. Collins, J.-L. Dufresne, D. Frame, D. J. Lunt, T. Mauritsen, M. D. Palmer, M. Watanabe, M. Wild, and H. Zhang. The earth's energy budget, climate feedbacks, and climate sensitivity. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.), *Climate Change*

2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the IPCC, pp. 923–1054. Cambridge University Press, Cambridge, UK and New York, NY, USA, 2021. doi: 10.1017/9781009157896.009.

Peter J. Gleckler, Karl E. Taylor, and Charles Doutriaux. Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, 113(D6):D06104, 2008. doi: 10.1029/2007JD008972.

Isaac M. Held and Brian J. Soden. Robust responses of the hydrological cycle to global warming. *Journal of Climate*, 19(21):5686–5699, 2006. doi: 10.1175/JCLI3990.1.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, and et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Frédéric Hourdin, Thorsten Mauritsen, Andrew Gettelman, Jean-Christophe Golaz, V. Balaji, and et al. The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, 98(3):589–602, 2017. doi: 10.1175/BAMS-D-15-00135.1.

Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop R. Ganguly, Shashi Shekhar, Nagiza F. Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017. doi: 10.1109/TKDE.2017.2720168.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, and et al. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 5637–5664, 2021.

Vladimir M. Krasnopolsky, Michael S. Fox-Rabinovitz, and Alexander A. Belochitski. Using ensemble of neural networks to learn stochastic convection parameterizations for climate and numerical weather prediction models from data simulated by a cloud resolving model. *Advances in Artificial Neural Systems*, 2013:485913, 2013. doi: 10.1155/2013/485913.

Cécile Nam, Jean-Louis Dufresne, Hélène Chepfer, and Sandrine Bony. The 'too few, too bright' tropical low-cloud problem in cmip5 models. *Geophysical Research Letters*, 39:L21801, 2012. doi: 10.1029/2012GL053421.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

V. Ramanathan, R. D. Cess, E. F. Harrison, P. Minnis, B. R. Barkstrom, E. Ahmad, and D. Hartmann. Cloud-radiative forcing and climate: Results from the earth radiation budget experiment. *Science*, 243(4887):57–63, 1989. doi: 10.1126/science.243.4887.57.

Stephan Rasp, Michael S. Pritchard, and Pierre Gentine. Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39):9684–9689, 2018. doi: 10.1073/pnas.1810286115.

R. N. B. Smith. A scheme for predicting layer clouds and their water content in a general circulation model. *Quarterly Journal of the Royal Meteorological Society*, 116(492):435–460, 1990a. doi: 10.1002/qj.49711649210.

R. N. B. Smith. A scheme for predicting layer clouds and their water content in a general circulation model. *Quarterly Journal of the Royal Meteorological Society*, 116(492):435–460, 1990b. doi: 10.1002/qj.49711649210.

Graeme L. Stephens. Cloud feedbacks in the climate system: A critical review. *Journal of Climate*, 18(2):237–273, 2005. doi: 10.1175/JCLI-3243.1.

Graeme L. Stephens, Deborah G. Vane, Robert J. Boain, and et al. The CloudSat mission and the A-train. *Bulletin of the American Meteorological Society*, 83(12):1771–1790, 2002. doi: 10.1175/BAMS-83-12-1771.

T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley (eds.). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2013. doi: 10.1017/CBO9781107415324.

H. Sundqvist, E. Berge, and J. E. Kristjansson. Condensation and cloud parameterization studies with a mesoscale numerical weather prediction model. *Monthly Weather Review*, 117(8):1641–1657, 1989. doi: 10.1175/1520-0493(1989)117⟨1641:CACPSW⟩2.0.CO;2.

Michael Tiedtke. Representation of clouds in large-scale models. *Monthly Weather Review*, 121 (11):3040–3061, 1993. doi: 10.1175/1520-0493(1993)121⟨3040:ROCILS⟩2.0.CO;2.

John M. Wallace and Peter V. Hobbs. *Atmospheric Science: An Introductory Survey*. Academic Press, 2nd edition, 2006.

David M. Winker, Mark A. Vaughan, Ali Omar, Yongxiang Hu, Kenneth A. Powell, Zhaoyan Liu, William H. Hunt, and Stuart A. Young. Overview of the CALIPSO mission and CALIOP data processing algorithms. *Journal of Atmospheric and Oceanic Technology*, 27(3):231–249, 2010. doi: 10.1175/2009JTECHA1281.1.

Janni Yuval and Paul A. O'Gorman. Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Journal of Advances in Modeling Earth Systems*, 12(12):e2020MS002268, 2020. doi: 10.1029/2020MS002268.

Janni Yuval and Paul A. O'Gorman. Learning cloud and precipitation physics for climate modeling: Representing uncertainties with stochasticity. *Journal of Advances in Modeling Earth Systems*, 13(2):e2020MS002386, 2021. doi: 10.1029/2020MS002386.

A Appendix

Declaration of Generative AI and AI-assisted technologies in the writing process.

We used ChatGPT (GPT-4 and GPT-5) in a limited editorial role: (1) code refactoring suggestions for non-novel boilerplate (renaming variables, reformatting functions, extracting helpers); (2) light text polishing for grammar/clarity; and (3) LaTeX formatting guidance (environments, floats, and bibliography style). We did *not* use AI tools to propose or validate methods, derive results, tune hyperparameters, generate figures/tables, create datasets/labels, or perform literature reviews without manual verification. All AI-assisted changes were inspected and, where appropriate, rewritten by the authors; the authors accept full responsibility for the content.

Table A1: Fixed training settings shared by all runs.

Item	Value
Architecture	3×11 ReLU; sigmoid output
Optimiser	Adam
Epochs (max)	1000 (early stopping; patience 20 epochs)
Seeds	25 (report mean±SEM)
Normalisation	Per-feature min–max (train split only; frozen for val/test)
Area weighting	$w_i = \cos(\phi_i)$ (radians)

Table A2: Tuned hyperparameters by model and training protocol. Values selected on validation and then fixed for 25 seeds; all other settings are shared (see Table A1).

	Global train		**Polar train**	
Hyperparameter	**PINN**	**NN**	**PINN**	**NN**
Learning rate η	1.461×10^{-3}	5.285×10^{-3}	5.659×10^{-4}	1.858×10^{-3}
Batch size B	16	64	32	16
Physics weight α	1.015×10^{-4}	—	1.010×10^{-4}	—
Dropout rate	0.0112	0.0118	0.00770	0.0206