PHYSICS AS REPRESENTATION, NOT REGULARIZATION: MECHANISM AND MEASUREMENT IN PROBABILISTIC SOLAR FORECASTING UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Physics-informed machine learning is motivated by two promises: better predictions in-distribution, and better behaviour outside it. We test both in probabilistic solar forecasting by separating two mechanisms for injecting the same physical information into one shared backbone. *Representation* exposes physically derived quantities (clear-sky irradiance, solar geometry, air mass) to the predictor as inputs; *regularization* places physics-based penalties in the training objective, where they shape weights and then vanish at inference. Using 42 inverter systems on five campuses, 19 quantile levels and a parameter-matched physics-free control, we find the two mechanisms are not interchangeable. Representation improves CRPS by 2.21% (95% CI $[-4.13, -0.45]$, five seeds, campus-clustered bootstrap), improves all 42 inverters, cuts implausible night-time generation by a factor of 50, and replicates in a gradient-boosting model of a different family (-1.03%, $[-1.49, -0.46]$). Regularization alone yields -0.61% ($[-2.76, +0.35]$), indistinguishable from zero. Under leave-one-campus-out evaluation, however, transfer is *heterogeneous rather than uniform*: the representation advantage persists on two campuses, is absent on a third, and reverses on a single-inverter fourth. A single fold, the usual practice, could have supported any of these conclusions. We further show that the large cross-population degradation such a protocol reports ($+66\%$ on one fold) is a capacity artefact; measured on identical rows it is at most about 9%. With five campuses the smallest attainable pooled p-value is 0.0625, so we report intervals rather than verdicts. The practical conclusion is that how physics enters a forecaster, and how transfer is measured, change the reported answer more than the physics itself.

1 INTRODUCTION

Physics-informed machine learning (Raissi et al., 2019; Karniadakis et al., 2021) promises models that need less data and behave better outside their training distribution. In applied forecasting the first promise is routinely checked; the second is asserted far more often than it is measured. Solar power forecasting is a clean setting in which to measure it, because part of the signal is fixed by astronomy and atmospheric optics, clear-sky irradiance, solar zenith angle, air mass, all computable exactly from site coordinates and a clock, while the residual (clouds, soiling, curtailment, array geometry) is site-specific and has no reason to transfer.

Two design choices are usually conflated in this literature. Physical information can be supplied to the model as an *input representation*, in which case it is available whenever the model predicts; or it can be imposed as a *regularizer* on the training objective, in which case it shapes the weights and is absent at inference. Papers report that "adding physics helps" without separating these, and typically evaluate on data drawn from the same sites used for training.

We separate the two mechanisms on one backbone (Figure 1), hold the training budget fixed, add a parameter-matched physics-free control, and then ask whether any in-distribution benefit survives a change of site. Three findings follow.

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.

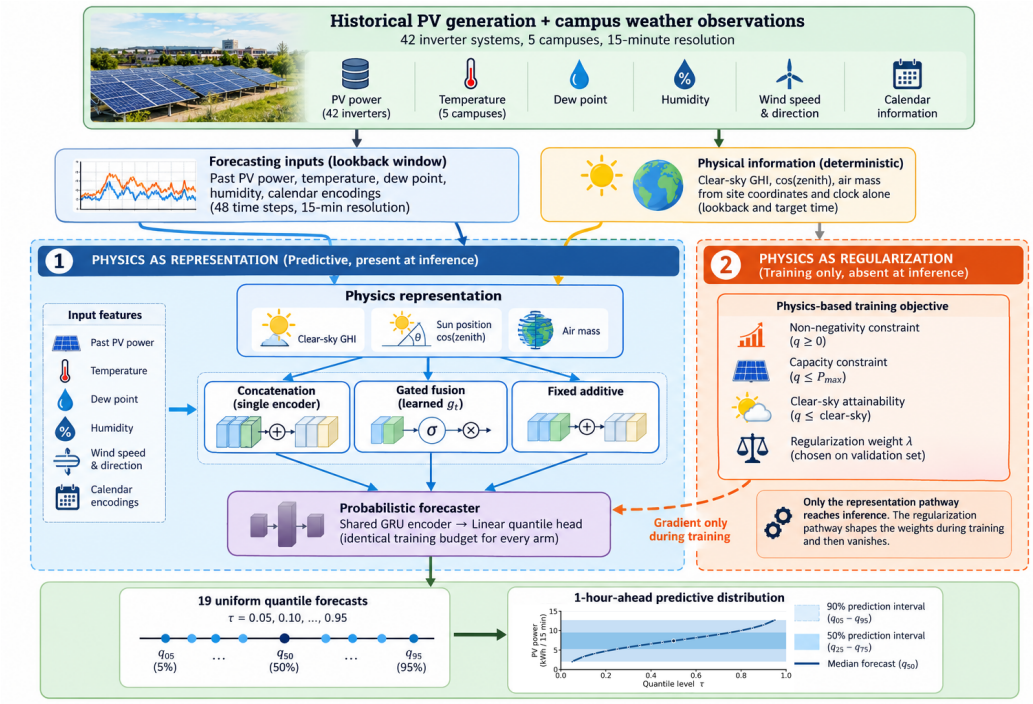


Figure 1: **Two mechanisms, one backbone.** The same physical information, clear-sky GHI, cos(zenith) and air mass, computed from site coordinates and clock time alone, enters the forecaster either as an input *representation* (pathway 1; concatenation, learned gate, or fixed additive; present whenever the model predicts) or as a *regularizer* on the training objective (pathway 2; non-negativity, capacity and clear-sky-attainability penalties, which shape the weights during training and are absent at inference). Every arm shares one GRU encoder, one linear 19-quantile head and one training budget. The control arms are the same backbone with the physics pathway removed: *no physics* (d=96) and the primary comparator, *parameter-matched no physics* (d=141), which carries the same parameter count as the gated representation arm. The output panel shows the predicted quantile function for one target step; values are illustrative.

1. **Mechanism matters.** Representation improves probabilistic accuracy and physical plausibility; regularization alone does not. The representation effect replicates in a gradient-boosting model that shares none of the neural architecture.
2. **Transfer is heterogeneous, not uniform.** Across five leave-one-campus-out (LOCO) folds the representation advantage persists on two campuses, vanishes on one, and reverses on another. Any single fold would have yielded a confident and different headline.
3. **The measurement protocol carries much of the reported effect.** The $+66\%$ degradation our own preliminary single fold produced is an artefact of comparing populations of different inverter capacity. Evaluated on identical rows, the cost of an unseen campus is at most about 9%.

We also report the study's negative results in full: with five campuses no pooled campus-level test can reach $p < 0.05$; the headline effect is significant at five seeds but not at three; and the best forecaster we measured is not the neural model at all, but quantile LightGBM.

2 Related Work

Physics-informed neural networks (Raissi et al., 2019) inject governing equations as soft constraints in a loss, and recent surveys organise the wider field explicitly by *how* physical knowledge is integrated rather than by what knowledge is used (Karniadakis et al., 2021; Meng et al., 2025). That taxonomy is the premise of this paper; what is missing is a measurement of whether the choice of integration point changes the outcome. In PV forecasting specifically, a recent review of hybrid

physics/ML models (de Oliveira Santos et al., 2024) documents the same label covering clear-sky normalisation, irradiance covariates, capacity clipping and penalty terms, frequently combined in one model and reported as a single intervention; physics-informed ensembles for PV follow the same pattern (Naghapushanam et al., 2026). What is rarely done is the controlled separation, identical backbone, identical budget, parameter-matched control, that isolates which of those choices produced the reported gain.

Probabilistic evaluation follows standard practice: quantile regression trained with the pinball loss (Koenker & Bassett, 1978) and assessed with CRPS (Gneiting & Raftery, 2007), a strictly proper score; recurrent quantile architectures of the kind we use remain a current baseline for probabilistic PV forecasting (Ait Mouloud et al., 2025). Cross-site generalisation is an active topic, transfer from a single site to new geographies (Zhang et al., 2025) and domain-alignment methods that try to separate site-invariant from site-specific structure (Ge & Ji, 2026), but held-out-site evaluations are still usually reported on one fold, with per-site statistics that treat co-located systems as independent. Our analysis departs from that: the campus, not the inverter, is the unit of inference, and we use a nested block bootstrap (Efron & Tibshirani, 1986) rather than a Diebold–Mariano test (Diebold & Mariano, 1995), whose HAC assumptions our measured autocorrelation does not support.

3 Experimental setting

Data. UNISOLAR (Wimalaratne et al., 2022): 42 inverter systems on five campuses of one university, 15-min PV generation from 2020-01-01 to 2022-04-23, co-located weather (temperature, humidity, dew point, wind) and site metadata. Splits are shared by every arm: train to 2021-06-30, validation 2021-07-01 to 2021-10-31, test 2021-11-01 to 2022-04-23. Targets are normalised per site by that site's train-period 99th percentile (scale only, no centring). The daytime evaluation mask uses *modelled* clear-sky GHI at the target step > 50 W/m^2, so it never reads an observation; it leaves 320,662 test rows.

Task. One target step at a 1-hour lead (4×15 min ahead) from a 48-step (12 h) lookback, predicted as 19 uniform quantiles $\tau \in \{0.05, \ldots, 0.95\}$ trained with the pinball loss (Koenker & Bassett, 1978). Quantile monotonicity is restored post hoc by rearrangement; raw violation rates are measured before that step.

Physical information. Clear-sky GHI, $\cos(\text{zenith})$ and air mass, computed with `pvlib` (Holmgren et al., 2018; Ineichen & Perez, 2002) from site coordinates and clock time alone, over the lookback and at the target step. No observed irradiance is used anywhere.

Arms. All share a 1-layer GRU encoder (Cho et al., 2014) and a linear quantile head, with an identical training budget (AdamW, lr 10^{-3}, cosine schedule, batch 512, 600 steps/epoch, max 25 epochs, patience 5). *No physics* (d=96, 68,371 parameters); *parameter-matched no physics* (d=141, 144,826 parameters, matched to the gated arm within 0.56%), the primary control; *physics representation* (concatenation, 68,947); *gated physics representation* (144,019); *physics regularization* (non-negativity, capacity and clear-sky-attainability penalties, λ selected from $\{0.01, 0.1, 1.0\}$ on validation, $\lambda = 0.1$). A quantile-LightGBM baseline (Ke et al., 2017) with and without the same physics features receives an identical protocol (same cache, splits, lookback, target, quantiles, normalisation, daytime mask and post-processing).

Statistics. Estimands are row-weighted means of *per-row paired* CRPS differences on identical rows. Uncertainty comes from a nested block bootstrap resampling seeds $\rightarrow$ campuses $\rightarrow$ inverters within campus $\rightarrow$ calendar-day blocks ($B = 2000$, percentile CIs). The clustering unit is the campus: within-campus inverter correlation of daily paired differences is $r = 0.76$ against 0.14 between campuses, so 42 inverters carry roughly five independent units. Inverter-level Wilcoxon tests are reported as descriptive only. Pooled campus-level statements use a sign-flip permutation over the five folds, whose minimum attainable two-sided p is 0.0625; Holm correction (Holm, 1979) is applied within families.

Seeds. Neural in-distribution arms use five seeds (42–46); LOCO folds and LightGBM use three (42–44). Five seeds was declared in the compute budget before results were seen, but the headline contrast moved from $p = 0.087$ to $p = 0.015$ when seeds 45–46 were added, so we report both.

Table 1: In-distribution results on 320,662 daytime test rows. Neural arms: five seeds (42–46); LightGBM: three seeds. CRPS is reported as mean ± s.d. over seeds. The parameter-matched control is the primary comparator for the representation arms.

Arm	Capacity	CRPS ↓	RMSE ↓	R^2 ↑	PICP_{90}
LightGBM, physics representation	48 feat.	$\mathbf{1.2286 \pm 0.0005}$	$\mathbf{3.8295}$	$\mathbf{0.9072}$	0.8615
LightGBM, no physics	37 feat.	1.2405 ± 0.0007	3.8489	0.9063	0.8767
Neural, physics representation (concat)	68,947	1.2625 ± 0.0157	3.8798	0.9048	$\mathbf{0.8984}$
Neural, gated physics representation	144,019	1.2669 ± 0.0127	3.8714	0.9052	0.8847
Neural, physics regularization (λ=0.1)	68,371	1.2800 ± 0.0205	3.8871	0.9044	0.8865
Neural, no physics (d=96)	68,371	1.2876 ± 0.0240	3.8847	0.9045	0.8863
Neural, parameter-matched no physics	144,826	1.2909 ± 0.0178	3.9127	0.9031	0.8778

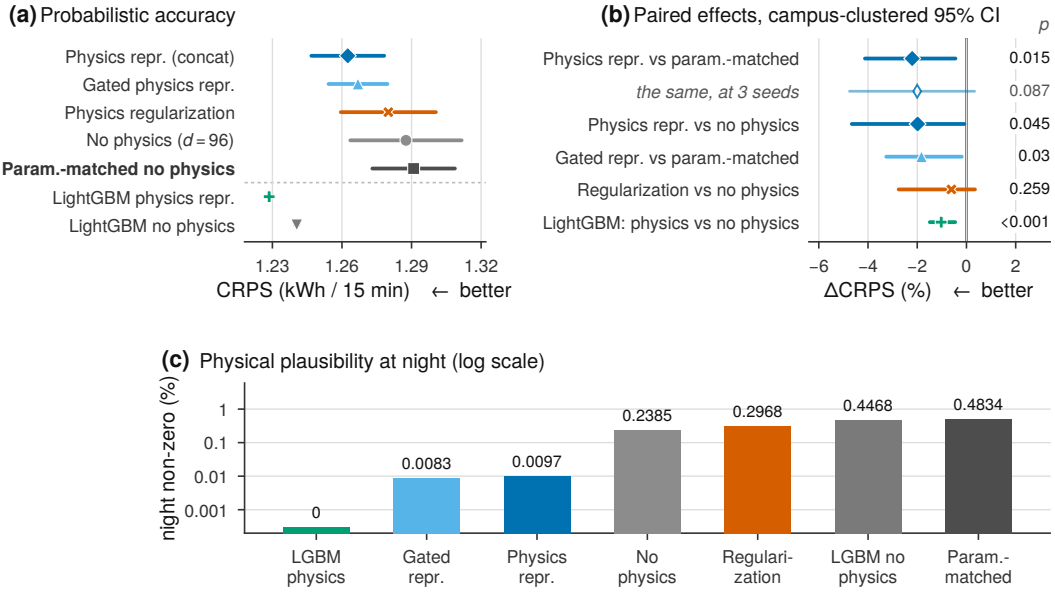


Figure 2: **In-distribution comparison.** **(a)** CRPS per arm; marker = mean over seeds, bar = ±1 s.d. **(b)** Paired effects on identical rows with campus-clustered 95% bootstrap CIs; filled markers denote intervals excluding zero, the open marker is the same headline contrast computed at three seeds. **(c)** Median rate of non-zero generation predicted at night, log scale, physics as representation suppresses it by roughly 50×; physics as regularization does not.

4 In-distribution: the mechanism decides

Table 1 and Figure 2 give the in-distribution picture.

Representation helps. Against the parameter-matched control, physics representation lowers CRPS by 2.21% (95% CI $[-4.13\%, -0.45\%]$, $p = 0.015$), and improves *all 42 of 42* inverters. Against the unmatched d=96 control the effect is -1.98% ($[-4.66, -0.08], p = 0.045$). The gated variant behaves similarly (-1.83%, $[-3.26, -0.19]$, $p = 0.030$).

Regularization does not. Physics penalties alone give -0.61% ($[-2.76, +0.35], p = 0.259$): the point estimate is not zero, but it is roughly four times smaller than the representation effect and its interval comfortably contains zero. Physics used only to shape weights during training, and absent when the model predicts, buys very little.

The effect replicates in a different model family. Quantile LightGBM given exactly the same physics features improves on LightGBM without them by -1.03% ($[-1.49, -0.46], p < 0.001$), on 37 of 42 inverters, and reaches its optimum with less than half the trees (1,868 vs 4,010). The benefit of physics-as-representation is therefore not a property of one GRU.

Table 2: Leave-one-campus-out, per fold. Held-out ΔCRPS of physics representation (concatenation) against two controls, with campus-clustered 95% bootstrap CIs; negative favours physics. The gated arm's held-out interval contains zero on every fold (p from 0.613 to 0.995).

Held-out campus Inverters / rows	1 27 / 212,831	2 5 / 37,074	3 8 / 55,272	4 1 / 7,493	5 1 / 7,992
vs no physics (%)	-2.70	-3.34	$+0.01$	-2.24	$+1.46$
95% CI	$[-4.51, -1.18]$	$[-6.03, -1.12]$	$[-3.34, +4.09]$	$[-7.80, +2.14]$	$[+0.23, +2.65]$
verdict	persists	persists	absent	indeterminate	reversed
vs param.-matched (%)	-2.87	-6.30	-3.37	$+2.90$	-1.07
95% CI	$[-4.38, -1.70]$	$[-8.11, -4.09]$	$[-8.01, -0.22]$	$[-2.47, +8.41]$	$[-4.75, +3.24]$

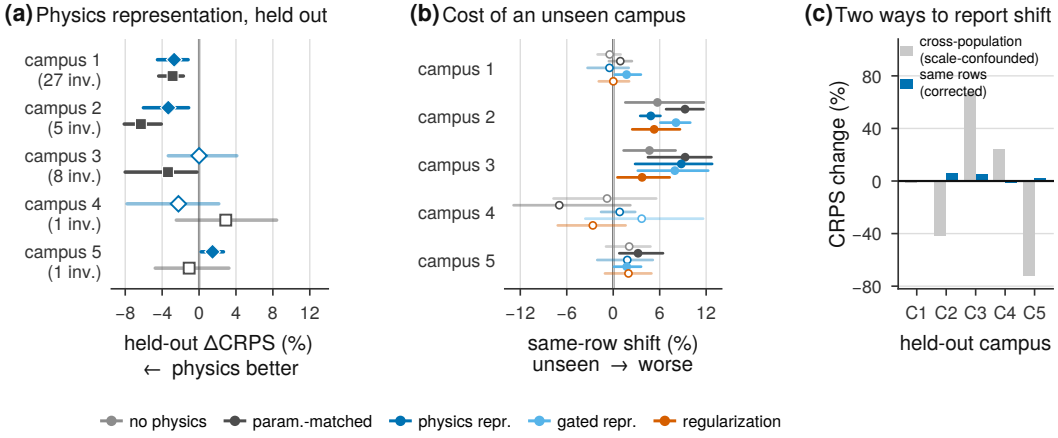


Figure 3: **Leave-one-campus-out across all five folds.** **(a)** Held-out effect of physics representation against the no-physics control (blue diamonds) and the parameter-matched control (dark squares); filled markers denote 95% CIs excluding zero. **(b)** Cost of an unseen campus measured on identical rows, per arm. **(c)** The same five folds measured two ways: the cross-population comparison a single-fold protocol reports (grey; ranges from -72% to $+66\%$) against the same-row shift for the same arm (blue; -0.8% to $+5.7\%$). The grey bars are a capacity artefact, not a measure of shift.

Physical plausibility improves sharply. The median rate at which a model predicts non-zero generation at night falls from 0.483% (parameter-matched control) and 0.239% (d=96 control) to 0.010% (concat) and 0.008% (gated), a factor of roughly 50 (Figure 2c). Physics regularization, whose penalties target exactly this behaviour, achieves 0.297%: no better than the controls. This is the sharpest expression of the mechanism result. Making "the sun is down" visible to the predictor works; penalising the model for forgetting it does not.

Two results that cut against us. First, calibration is imperfect everywhere: no arm reaches nominal 90% coverage, and LightGBM's coverage gets *worse* with physics (0.8767 → 0.8615) while its CRPS improves, sharper intervals not matched by a coverage gain. No post-hoc conformal calibration is applied anywhere in this study. Second, the best forecaster we measured is not the neural model: LightGBM without physics beats the best neural arm by 1.24%, though not distinguishably so ($[-3.73, +2.60], p = 0.657$), and beats the neural parameter-matched control by 3.22% ($[-4.81, -0.25], p = 0.038$). The study is a controlled comparison of how physics enters a forecaster, not a leaderboard entry, but the ranking should be stated plainly.

5 Cross-campus transfer: heterogeneous, not uniform

Each LOCO fold removes one campus entirely from training and validation and evaluates on its rows in the same calendar window as the in-distribution test split, so the only difference is whether the campus was seen.

The representation advantage does not transfer uniformly. Table 2 and Figure 3a show the concatenation arm beating the no-physics control with an interval excluding zero on campuses 1 and 2 (-2.70% and -3.34%), showing no effect on campus 3 ($+0.01\%$), being indeterminate on campus 4, and being *worse* on campus 5 ($+1.46\%$, $[+0.23, +2.65]$). Against the parameter-matched control the effect persists on three folds. Pooled over the five folds the mean effect is -0.018 kWh/15 min with sign-flip $p = 0.250$, and no five-fold design can do better than $p = 0.0625$.

The gated arm is different and cleaner: its in-distribution advantage (-1.83%) does not persist on *any* held-out campus, homogeneously across folds. Two representations with indistinguishable in-distribution scores therefore behave differently under shift, which in-distribution evaluation cannot reveal.

A single fold would have misled us. Our own preliminary work used only campus 3, and reported that the benefit vanishes under shift. With all five folds that conclusion is one of three possible single-fold headlines: "persists" (campus 1 or 2), "absent" (campus 3), "reverses" (campus 5). The folds are not exchangeable, campus 1 holds 27 of 42 inverters and its fold therefore trains on only 15; campuses 4 and 5 are single inverters with 44% and 74% missing air-temperature records, so the spread is partly structural, which is exactly why one fold cannot stand for the protocol.

Most of the reported "degradation" is a capacity artefact. Comparing a model's CRPS on seen campuses against its CRPS on the held-out campus mixes the effect of shift with the fact that the two populations contain inverters of different size. That comparison gives $+66\%$ on campus 3, $+24\%$ on campus 4, and -72% on campus 5 (Figure 3c, grey), a quantity whose sign is set by which campus happens to be held out. Scoring the *same model on the same rows*, trained with and without that campus, gives the actual cost of not having seen it: -0.46% (campus 1, CI includes zero), $+5.71\%$ (campus 2), $+4.68\%$ (campus 3), -0.80% (campus 4), $+2.06\%$ (campus 5) for the no-physics arm, the estimates whose intervals exclude zero span $+1.7\%$ to $+9.3\%$ (Figure 3b). Capacity-normalised CRPS tracks these within about one percentage point. The corrected shift is real, campus-dependent, and an order of magnitude smaller than the headline figure a naive protocol produces.

6 What we could not establish

We list these as findings, not caveats, because they bound what any study of this size can claim.

Five campuses is five units. Inverters on one campus share a weather record ($r = 0.76$). Inverter-level tests on this data give absurdly small p-values, the headline contrast yields Wilcoxon $p \approx 5 \times 10^{-11}$ across 42 inverters, against a campus-clustered bootstrap p of 0.090 at the same three seeds. The difference is roughly nine orders of magnitude and is entirely an artefact of treating clustered units as independent. We believe this inflation is widespread in multi-site forecasting papers.

The headline is seed-sensitive. At three seeds the representation-vs-matched contrast is -2.00% ($[-4.75, +0.33], p = 0.087$); at five it is -2.21% ($[-4.13, -0.45], p = 0.015$). The point estimate barely moved and the interval narrowed, as adding seeds should; but a contrast that crosses 0.05 on the addition of seeds invites an optional-stopping objection, so both are reported and neither is presented alone.

We dropped a heterogeneity statistic we had computed. A DerSimonian–Laird analysis of the five folds gave $Q = 10.9$, $p \approx 0.03$, $I^2 = 0.63$. An audit showed 95.7% of the meta-analytic weight fell on a *single-inverter* campus, because absolute CRPS scales with capacity and so does its bootstrap SE; removing that campus collapses Q to 0.65 ($p = 0.885$). I^2 at $k = 5$ had a bootstrap interval of $[0.00, 0.89]$ (Higgins & Thompson, 2002), and the five folds are not independent, any two share three of five training campuses. We therefore report the per-fold intervals directly and no pooled heterogeneity statistic.

Other limits. λ and the penalty thresholds were selected once on an in-distribution validation pool containing pre-test rows from every campus, so the regularization arm's LOCO folds inherit a cross-campus selection (not test-period leakage). Held-out sites use their own train-period capacity scale, a "nameplate known" deployment assumption. One dataset, one university, one climate; the parameter match is within 0.56%, not exact; and no conformal calibration is applied to any arm.

7 Conclusion

Physical information helped this forecaster when it was supplied as a representation the model can read at prediction time, and did not help when it was supplied only as a penalty during training, a distinction visible in probabilistic accuracy, in all 42 inverters, in a 50-fold reduction of night-time implausibility, and replicated in a gradient-boosting model that shares none of the architecture. Whether that benefit survives a move to an unseen site turned out not to have a single answer: it persisted on two campuses, disappeared on a third, and reversed on a fourth. On top of that, the large degradation numbers a conventional single-fold protocol reports are mostly a capacity artefact, and the per-site significance such papers report is inflated by roughly nine orders of magnitude by ignoring site clustering.

The uncomfortable implication is that in this literature the evaluation design which mechanism, which control, which fold, which clustering unit moves the reported conclusion more than the physics does. We would rather report that than a cleaner story the evidence does not support.

AI use statement

We used generative AI tools to aid and polish writing, and to draft sections of this paper from the authors' results and analysis outputs. We did not use generative AI tools for research ideation, experimental design, or the interpretation of results, and proof generation is not applicable to this work. All AI-assisted text was reviewed and revised by the authors, and every number reported in this paper is regenerated from the stored prediction files by the analysis scripts released with the submission. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

Ethics statement

This work uses a public dataset of aggregate photovoltaic generation and co-located weather observations from institutional installations (Wimalaratne et al., 2022); it involves no human subjects, no personal or identifiable data, and no sensitive attributes. We see no dual-use concern. The one ethical risk we consider material is inferential: probabilistic forecasts of this kind inform grid dispatch and procurement decisions, and both the inflated per-site significance and the scale-confounded shift metric documented in Section 6 would, if propagated uncorrected, overstate how reliable such a model is at a site it has never seen. We therefore report cluster-aware intervals, the per-fold heterogeneity, and the statistics we discarded, rather than a single pooled verdict. The authors declare no competing interests.

Reproducibility statement

Every number in this paper is regenerated from the stored per-row prediction files by the analysis scripts described in Section 3; none is transcribed by hand. The dataset is public (Wimalaratne et al., 2022). Splits, masks, feature definitions, penalty thresholds, the λ grid, optimiser settings and software versions are given in Section 3 and in full in the supplementary material, together with a run manifest covering all 106 training runs (experiment id, arm, seed, fold, config hash, timings), per-seed and per-inverter scores, the complete 65-row statistical analysis, and the heterogeneity audit summarised in Section 6. A bit-exact reproduction of one registered run was verified before any result reported here was produced.

References

Louiza Ait Mouloud, Aissa Kheldoun, Samira Oussidhoum, Hisham Alharbi, Saud Alotaibi, Thabet Alzahrani, and Takele Ferede Agajie. Seasonal quantile forecasting of solar photovoltaic power using Q-CNN-GRU. *Scientific Reports*, 15:27270, 2025. doi: 10.1038/s41598-025-12797-8.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*, pp. 1724–1734, 2014.

Leticia de Oliveira Santos, Tarek AlSkaif, Giovanni Cordeiro Barroso, and Paulo Cesar Marques de Carvalho. Photovoltaic power estimation and forecast models integrating physics and machine learning: A review on hybrid techniques. *Solar Energy*, 284:113044, 2024. doi: 10.1016/j.solener.2024.113044.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

Yantong Ge and Xunsheng Ji. CausalTransPV: Causal invariant representation learning for cross-site photovoltaic power forecasting via selective domain alignment. *Energies*, 19(6):1410, 2026. doi: 10.3390/en19061410.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Julian P. T. Higgins and Simon G. Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558, 2002.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

William F. Holmgren, Clifford W. Hansen, and Mark A. Mikofski. pvlib python: A python package for modeling solar energy systems. *Journal of Open Source Software*, 3(29):884, 2018.

Pierre Ineichen and Richard Perez. A new airmass independent formulation for the Linke turbidity coefficient. *Solar Energy*, 73(3):151–157, 2002.

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.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Chuizheng Meng, Sam Griesemer, Defu Cao, Sungyong Seo, and Yan Liu. When physics meets machine learning: a survey of physics-informed machine learning. *Machine Learning for Computational Science and Engineering*, 1(1):20, 2025. doi: 10.1007/s44379-025-00016-0.

Manimaran Naghapushanam, Baskaran Jeevarathinam, and C. Sankari. Physics-informed voting ensemble for solar power generation forecasting: integrating domain knowledge with machine learning. *Energy Informatics*, 9(1):3, 2026. doi: 10.1186/s42162-025-00604-7.

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.

Sachith Wimalaratne, Dilantha Haputhanthri, Supun Kahawala, Gihan Gamage, Damminda Alahakoon, and Andrew Jennings. UNISOLAR: An open dataset of photovoltaic solar energy generation in a large multi-campus university setting. In *2022 15th International Conference on Human System Interaction (HSI)*, pp. 1–5, 2022. doi: 10.1109/HSI55341.2022.9869474.

Liwenbo Zhang, Robin Wilson, Mark Sumner, and Yupeng Wu. Transfer learning in very-short-term solar forecasting: Bridging single site data to diverse geographical applications. *Applied Energy*, 377:124353, 2025. doi: 10.1016/j.apenergy.2024.124353.