

How should machine-learning bias correction of satellite solar irradiance be built and validated? A methodological protocol from a tropical monsoon station network

Eko Riyanto

ekoriyanto25@students.undip.ac.id

Diponegoro University

Iwan Setiawan

Diponegoro University

Uky Yudatama

Universitas Muhammadiyah Magelang

Research Article

Keywords: bias correction, machine learning, validation design, spatial cross-validation, satellite irradiance, tropical meteorology

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10706667/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Machine-learning (ML) bias correction is increasingly used to reconcile satellite-derived global horizontal irradiance (GHI) with ground observations, yet the methodological choices behind these models—quality screening, predictor selection, model family, and above all the validation design—receive far less scrutiny than the algorithms themselves. We propose and demonstrate a four-stage protocol (network quality control, physically informed predictor construction, multi-family model comparison, and dual validation) using 5,075 daily observations from five climatological stations in a tropical monsoon network in Central Java, Indonesia (2022–2024), paired with NASA POWER satellite estimates. Uncorrected satellite GHI shows near-zero network-median bias but a station-level range from -3.2% to $+21.7\%$, the largest offset traceable to a single under-reading sunshine recorder. Under random cross-validation and a temporal 2024 hold-out, gradient boosting is the best of five model families, cutting pooled RMSE from 3.25 to $2.55 \text{ MJ m}^{-2} \text{ d}^{-1}$ (-21%) and removing seasonal bias structure. Under leave-one-station-out validation the ranking inverts: every ML model performs worse than the uncorrected satellite product, and the most flexible learners degrade most. A sensitivity experiment shows the correction at the anomalous station succeeds only when that station's own history is in the training set, and permutation importance identifies surface pressure—an elevation fingerprint—as a leading predictor, indicating that the models memorise station identity rather than a transferable atmospheric relationship. The protocol converts a seemingly contradictory pair of results into a coherent, deployment-specific answer and generalises to any setting where predictions are made from clustered observational networks.

Introduction

Reliable surface solar irradiance data underpin photovoltaic siting, agricultural modelling, and hydrological analysis, yet ground pyranometer networks remain sparse across most of the tropics (Besharat, Dehghan, and Faghih 2013; Rumbayan, Abudureyimu, and Nagasaka 2012). Satellite-derived and reanalysis products such as NASA POWER fill this gap with continuous spatial coverage, but their estimates carry systematic errors that vary with cloud regime, aerosol loading, and terrain (White et al. 2011; Urraca et al. 2018; Bright 2019). The standard remedy is site adaptation or bias correction: a statistical model maps the satellite estimate, together with auxiliary atmospheric variables, onto the ground reference (Polo et al. 2016, 2020). Machine-learning (ML) regressors now dominate this literature because they capture non-linear, regime-dependent error structure that linear methods miss (Voyant et al. 2017; Yagli, Yang, and Srinivasan 2019).

A methodological problem hides inside this apparently settled workflow. Published comparisons of bias-correction models overwhelmingly report accuracy from random k-fold cross-validation or a single temporal hold-out, and then recommend the winning algorithm for use across a region. Both steps embed an assumption that is rarely stated and even more rarely tested: that the held-out data are exchangeable with the deployment conditions. When observations come from a small number of fixed stations, random cross-validation places records from every station in both training and test folds, so the reported skill answers only the question “how well can the model correct stations it has already seen?”

The question practitioners usually care about—“how well will it correct a location without a ground instrument?”—requires holding out entire stations, a design known in spatial statistics as leave-one-station-out (LOSO) or spatial cross-validation (Roberts et al. 2017; Meyer and Pebesma 2021). Assessments in ecology and environmental mapping show that ignoring this distinction can overstate predictive skill by large margins and even reverse model rankings (Ploton et al. 2020; Meyer and Pebesma 2022), and the climate-downscaling literature has long warned that bias correction can add value under one evaluation design and destroy it under another (Maraun 2016; Maraun et al. 2017). Solar-energy validation guidelines acknowledge multiple performance indicators (Gueymard 2014) but give little guidance on the resampling design itself, and time-series work shows that even the temporal dimension of cross-validation is routinely mishandled (Bergmeir and Benítez 2012).

The tropics sharpen the problem. Monsoon cloud regimes generate large day-to-day irradiance variability; ground references in national networks are often sunshine-duration recorders rather than pyranometers, so instrument artefacts propagate into the “truth”; and station density is so low that several stations may fall inside a single satellite grid cell, leaving the satellite product with no information to distinguish between them. Under these conditions the choice of validation design is not a technical detail. It decides the answer.

The stakes are practical. Indonesia, an equatorial archipelago with among the most homogeneous solar climates on earth, plans a large expansion of photovoltaic capacity, and resource assessment there leans almost entirely on satellite and reanalysis products precisely because the ground network is thin (Rumbayan, Abudureyimu, and Nagasaka 2012). Agricultural and hydrological models consume the same products. If a bias-correction model validated one way improves those inputs and validated another way corrupts them, then the choice of validation design silently propagates into siting decisions, yield estimates, and reservoir operation. A protocol that makes the design choice explicit, and ties each reported number to the deployment claim it supports, is therefore useful well beyond the five stations analysed here.

This paper treats the construction and validation of an ML bias-correction model as a methodological object in its own right. We propose a four-stage protocol—(i) network-level quality control, including coordinate verification and instrument-consistency screening; (ii) physically informed predictor construction; (iii) comparison across model families of increasing flexibility; and (iv) dual validation that reports temporal-transfer and spatial-transfer skill side by side—and we demonstrate it on a complete empirical case: five climatological stations in Central Java, Indonesia, over 2022–2024, paired with NASA POWER daily estimates. Three questions structure the analysis. First, which model family corrects satellite GHI best when the deployment target is the observed network itself? Second, does the answer survive when the validation design is switched from random and temporal resampling to LOSO? Third, what mechanism explains any divergence, and what does it imply for how such models should be reported?

The contribution is methodological rather than algorithmic. We show that the same data, predictors, and models yield a 21% RMSE improvement or a net degradation relative to the uncorrected satellite product depending solely on the resampling design, that the ranking of model families inverts between designs, and that diagnostic tools—a training-set ablation at an anomalous station and permutation importance—can distinguish genuine atmospheric learning from memorisation of station identity. The protocol turns an apparent contradiction into a pair of honest, deployment-specific claims, and it applies wherever predictions are built from clustered observational networks, from air-quality sensors to hydrological gauges.

Background

Bias correction of satellite irradiance

Site adaptation methods range from simple linear mappings and quantile matching to non-parametric ML regressors (Polo et al. 2016). Benchmark studies generally find that flexible learners—random forests (Breiman 2001), gradient boosting (Friedman 2001), and stacked ensembles (Wolpert 1992)—outperform linear corrections on held-out data drawn from the same stations (Polo et al. 2020; Yagli, Yang, and Srinivasan 2019). NASA POWER, the product corrected here, provides global daily radiation and meteorology at coarse grid resolution and has been validated extensively against mid-latitude networks (White et al. 2011), with fewer assessments in tropical monsoon climates.

Validation design for clustered data

Cross-validation estimates generalisation error only under exchangeability between folds (Stone 1974). Clustered data violate this assumption twice: serial dependence links consecutive days at one station, and spatial dependence links stations to themselves across folds. Remedies include blocked temporal designs (Bergmeir and Benítez 2012) and spatial or grouped resampling (Roberts et al. 2017). Meyer and Pebesma (2021) formalise the related notion of an area of applicability, and Ploton et al. (2020) and Meyer and Pebesma (2022) document large-scale mapping studies whose reported skill collapsed under spatially independent evaluation. The bias-correction literature in climate science reaches a parallel conclusion from a different direction: corrections calibrated on one set of conditions can degrade physical consistency when transferred (Maraun 2016; Maraun et al. 2017). What is missing, and what this paper supplies, is a worked protocol that places these designs side by side for satellite-irradiance correction in a data-sparse tropical network and traces the divergence to its mechanism.

Seen from the standpoint of general quantitative methodology, the issue is one of external validity. A cross-validation estimate is an inference from a sample of evaluation conditions to a population of deployment conditions, and its credibility rests on how the folds are drawn relative to that population (Stone 1974). Random folds treat station-days as the sampling unit and therefore license conclusions only about the station-days regime; station-blocked folds treat stations as the sampling unit and license

conclusions about new stations. Conflating the two is the resampling analogue of generalising from a convenience sample to a population it was never drawn from. The methodological literature on clustered and hierarchical data has made this point for inferential statistics; the present paper makes it operational for predictive model selection, where the consequence is not an understated standard error but a reversed substantive recommendation.

Ground references in tropical networks

Many national networks in the tropics record sunshine duration with Campbell–Stokes recorders rather than measuring irradiance directly; daily GHI is then estimated through the Angström–Prescott relation (Ångström 1924; Prescott 1940; Allen et al. 1998). The chain from a burn trace on a card to an irradiance value introduces instrument-specific error that standard radiation quality-control procedures (Journée and Bertrand 2011) are not designed to catch, because each individual value can look plausible while the station as a whole drifts. Network-level screening—comparing each station’s distribution against its neighbours—is therefore a prerequisite, not an optional extra.

Data and study area

Station network and reference irradiance

The ground reference comes from five climatological stations operated by the provincial water-resources agency of Central Java, Indonesia: DANGI, JRAGUNG, JUWERO, KEDUNGOMBO, and SEMPUR, spanning 7.01° – 7.41° S and 110.14° – 110.85° E. The five were selected from a 28-station network by requiring at least 90% completeness in all predictor variables over 2022–2024. Daily reference GHI ($\text{MJ m}^{-2} \text{d}^{-1}$) was derived from Campbell–Stokes sunshine-duration records through the Angström–Prescott relation (Allen et al. 1998), the standard approach where pyranometers are unavailable. The final dataset contains 5,075 station-days (932–1,064 per station) with no missing values in any variable used below.

The climate is tropical monsoon: a wet season centred on December–February (DJF), a dry season centred on June–August (JJA), and transition months between. Mean reference GHI ranges from $14.9 \text{ MJ m}^{-2} \text{d}^{-1}$ at SEMPUR to $18.9 \text{ MJ m}^{-2} \text{d}^{-1}$ at KEDUNGOMBO.

Satellite data and pairing

Satellite estimates were obtained from the NASA POWER project for each station’s coordinates: daily all-sky and clear-sky GHI, direct and diffuse components, cloud amount, cloud optical depth, aerosol optical depth, surface albedo, air temperature (mean, maximum, minimum), relative and specific humidity, surface pressure, wind speed at 2 m and 10 m, precipitation, top-of-atmosphere irradiance, and daylight length. Because POWER’s radiation fields inherit the coarse (1°) resolution of the underlying CERES products, all five stations fall within a single radiation grid cell; the satellite series at the five sites are

therefore nearly identical (mean $18.1\text{--}18.2\text{ MJ m}^{-2}\text{ d}^{-1}$ at every station), a structural fact that becomes central in Section 5.

Stage 1 of the protocol: network quality control

Three screening steps preceded any modelling. First, station coordinates were verified against independent sources before satellite retrieval; in an earlier iteration of this dataset a transcription error displaced one station's coordinates by roughly 53 km, silently degrading the pairing until the check caught it. Second, physically implausible values (e.g., temperatures outside instrument range, negative irradiance) were screened; none remained in the final extract. Third, and most consequentially, each station's reference series was compared against the network distribution. SEMPUs's sunshine-duration record sits far below the network mean, and its derived GHI is correspondingly low, producing an apparent satellite overestimation of + 21.7% at that station while the neighbouring stations in the same satellite pixel show biases between -3.2% and $+4.6\%$ (Fig. 1). Because the satellite series is common to all five sites, the anomaly is attributable to the ground instrument, not the satellite. We retain SEMPUs in the analysis deliberately: an under-reading reference is exactly the kind of contamination a bias-correction workflow encounters in practice, and Section 5.4 uses it as a diagnostic instrument.

Methods

Stage 2: predictor construction

The predictor set contains 24 variables available wherever the satellite product is available: the satellite GHI itself, clear-sky GHI, clear-sky index, two clearness indices, direct and diffuse irradiance, surface albedo, cloud amount, cloud optical depth, aerosol optical depth, three temperature variables, two humidity variables, surface pressure, two wind speeds, precipitation, extraterrestrial irradiance, daylight length, and sine–cosine transforms of day of year. Ground-measured variables were excluded from the predictors so that the fitted models remain deployable at locations without instruments. The target is daily reference GHI; predicting the reference directly (rather than the bias) makes the uncorrected satellite estimate a natural baseline against which every model must justify itself.

Stage 3: model families

Five regressors of increasing flexibility were compared under identical predictors: ordinary least squares; ridge regression with generalised cross-validated penalty (Hoerl and Kennard 1970); a random forest (Breiman 2001); gradient-boosted trees (Friedman 2001); and a stacked ensemble combining the three preceding learners through a linear meta-model with out-of-fold base predictions (Wolpert 1992). The five span the flexibility spectrum deliberately. The two linear models can represent only a global affine recalibration of the predictors, the classical model-output-statistics correction; the tree ensembles can represent regime-dependent corrections that differ between clear and cloudy days, wet and dry seasons, or—the possibility at the heart of this paper—between stations, whenever any predictor combination

fingerprints a station; the stacked ensemble tests whether combining the families adds skill beyond its best member (Wolpert 1992; Antonanzas et al. 2016). All models were implemented in scikit-learn (Pedregosa et al. 2011) with fixed random seeds. Hyperparameters were held at moderate, regularised settings (100–200 trees or boosting iterations, minimum leaf sizes of five, penalised meta-learner) rather than tuned per design, so that differences between validation designs reflect the designs themselves rather than differential tuning effort; Section 6 discusses why per-design tuning could sharpen but not reverse the findings.

Stage 4: dual validation

Each model was evaluated under three resampling designs chosen to answer three distinct deployment questions.

1. **Random 5-fold cross-validation.** Days are assigned to folds at random, ignoring station and time. This is the design most often reported in the bias-correction literature; it answers “how well does the model interpolate within the observed station-days?”
2. **Temporal block hold-out.** Models train on 2022–2023 and predict all of 2024. This answers the operational question “how well will the model correct future data at the stations it was calibrated on?”
3. **Leave-one-station-out (LOSO).** Each station is held out in turn; the model trains on the other four and predicts the held-out station’s full record. This answers the spatial-transfer question “how well will the model correct a location it has never seen?”

Performance is reported with four complementary indicators, pooled over all held-out predictions in each design, following standard solar-validation practice (Gueymard 2014). For n held-out days with predictions $\hat{y}_i$ (predicted) and reference values y_i (reference),

RMSE = square root of the mean of (predicted – reference)² over all n held-out days; MAE = mean of |predicted – reference| over all n held-out days

MBE = mean of (predicted – reference) over all n held-out days; $R^2 = 1 - (\text{sum of squared prediction errors} / \text{sum of squared deviations of the reference from its mean})$

RMSE penalises large errors and is the headline dispersion measure; MAE is robust to the heavy-tailed error days that monsoon convection produces; MBE isolates systematic offset, the quantity bias correction exists to remove; and R^2 situates each model against the trivial climatological predictor. Reporting the four jointly matters because the designs separate them: a model can shrink dispersion while retaining offset, or remove offset at the price of dispersion, and the tables in Section 5 contain both patterns. Pooling across held-out predictions (rather than averaging per-fold scores) weights every day equally and avoids the small-fold instability that per-station averaging would introduce with only five stations.

Two diagnostics probe the mechanism behind any disagreement between designs: a training-set ablation in which SEMPUPU is removed from the temporal-design training data while evaluation is unchanged, and permutation importance (Fisher, Rudin, and Dominici 2019) of the gradient-boosting model on the 2024 hold-out.

Results

Uncorrected satellite performance

Table 1 summarises the uncorrected product. Pooled over all station-days, satellite GHI has RMSE 3.26 MJ m⁻² d⁻¹, MBE +0.69 MJ m⁻² d⁻¹, and R² = 0.37. The pooled bias is dominated by SEMPUPU; the network median bias is close to zero. Correlations between daily satellite and reference series range from 0.52 (DANGI) to 0.77 (SEMPUPU), modest values that reflect both the coarse satellite pixel and the indirect sunshine-based reference.

Uncorrected NASA POWER daily GHI against the ground reference by station, 2022–2024. MBE and RMSE in MJ m⁻² d⁻¹

Station	n	Reference mean	MBE	Relative bias (%)	RMSE
DANGI	1,064	18.06	+0.04	+0.2	3.92
JRAGUNG	1,026	18.29	-0.10	-0.5	2.65
JUWERO	1,027	17.42	+0.79	+4.6	2.87
KEDUNGOMBO	932	18.86	-0.61	-3.2	2.76
SEMPUPU	1,026	14.90	+3.23	+21.7	3.81

Model comparison under the three designs

Table 2 and Fig. 2 present the central result. Under random cross-validation, every ML model beats the uncorrected satellite, and flexibility pays: gradient boosting attains RMSE 2.59 against 3.26 for the raw product, with the random forest close behind and the linear models trailing. The temporal design tells the same story with almost identical numbers—gradient boosting reaches RMSE 2.55 on the unseen year 2024, a 21% reduction, with MBE shrinking from +0.75 to +0.36 MJ m⁻² d⁻¹. Agreement between the two designs indicates that the error structure is stable from year to year at these stations.

LOSO reverses the verdict. Every ML model now performs worse than the uncorrected satellite product (RMSE 3.45–3.68 versus 3.26), and the ordering of model families inverts: the linear and ridge corrections, weakest under the first two designs, are now the least bad, while gradient boosting—the winner under temporal validation—is the worst performer of all. Flexibility, an asset when the deployment station is in the training set, becomes a liability when it is not.

Two secondary patterns in Table 2 reinforce the interpretation. The linear models degrade least under LOSO because a global affine map has no capacity to encode station identity; what it learns is, by construction, the network-average relationship, which transfers imperfectly but does not import another station’s idiosyncrasies. And the stacking ensemble, whose linear meta-learner is fitted on out-of-fold base predictions, sits between the linear and tree families in every design while achieving the smallest absolute MBE under the temporal hold-out ($-0.01 \text{ MJ m}^{-2} \text{ d}^{-1}$); its internal cross-fitting acts as a mild guard against the memorisation that the raw tree ensembles indulge, at some cost in dispersion. Neither pattern rescues spatial transfer, but both are diagnostic: when model ordering correlates with capacity to encode group identity, group leakage—not atmospheric signal—is doing the work.

Pooled performance of five model families under three validation designs. Best value per design in bold; $\text{MJ m}^{-2} \text{ d}^{-1}$ for RMSE, MAE, MBE

Design	Model	RMSE	MAE	MBE	R ²
Random 5-fold CV	Uncorrected satellite	3.26	2.60	+0.69	0.37
	Linear	2.78	2.15	0.00	0.54
	Ridge	2.78	2.15	0.00	0.54
	Random forest	2.61	1.96	-0.01	0.60
	Gradient boosting	2.59	1.92	-0.01	0.61
	Stacking ensemble	2.69	2.11	-0.59	0.57
Temporal (2024 hold-out)	Uncorrected satellite	3.25	2.59	+0.75	0.36
	Linear	2.83	2.17	+0.63	0.51
	Ridge	2.83	2.17	+0.63	0.51
	Random forest	2.63	1.99	+0.39	0.58
	Gradient boosting	2.55	1.92	+0.36	0.60
	Stacking ensemble	2.63	2.08	-0.01	0.58
Leave-one-station-out	Uncorrected satellite	3.26	2.60	+0.69	0.37
	Linear	3.45	2.81	+0.04	0.30
	Ridge	3.45	2.81	+0.04	0.30
	Random forest	3.58	2.86	+0.41	0.25
	Gradient boosting	3.68	2.95	+0.42	0.20
	Stacking ensemble	3.50	2.92	-0.11	0.28

Station-level and seasonal structure under temporal validation

For the deployment scenario the temporal design represents—continuing correction at instrumented stations—the gradient-boosting model improves or holds every station (Table 3). The largest gain occurs precisely at SEMPU, where RMSE falls from 4.20 to 2.02 and MBE from +3.69 to +0.58 MJ m⁻² d⁻¹: with SEMPU’s history in the training data, the model learns that station’s instrument offset and reproduces it. Seasonally, the raw product’s bias in the 2024 hold-out is largest in the dry season (JJA MBE +0.96); the correction compresses bias across all seasons and cuts dry-season RMSE from 3.15 to 1.99, with the wet season improving least (Fig. 3), consistent with the harder cloud-detection problem under monsoon convection.

Station-level performance in the 2024 hold-out year: uncorrected satellite versus the gradient-boosting correction trained on 2022–2023. MJ m⁻² d⁻¹

		Uncorrected		Corrected	
Station	n	MBE	RMSE	MBE	RMSE
DANGI	357	+0.29	3.63	+0.79	3.03
JRAGUNG	333	-0.20	2.49	+0.42	2.42
JUWERO	348	+0.54	2.84	-0.03	2.56
KEDUNGOMBO	305	-0.70	2.68	+0.01	2.61
SEMPU	340	+3.69	4.20	+0.58	2.02

Mechanism: why the designs disagree

Two diagnostics locate the source of the divergence. The ablation experiment (Table 4) retrains the gradient-boosting model on 2022–2023 with SEMPU excluded and evaluates on 2024 as before. Performance at the other four stations barely moves (RMSE 2.67 versus 2.72). At SEMPU it collapses: RMSE rises from 2.02 to 4.78 and MBE from +0.58 to +3.90 MJ m⁻² d⁻¹—worse than the uncorrected satellite at that station. The successful correction of SEMPU under temporal validation was therefore not a correction of satellite error at all; it was a learned reproduction of that station’s instrument offset, available only because the station corrected itself through its own training history. LOSO makes every station face the SEMPU condition, which is why all learned corrections fail under it.

Permutation importance on the 2024 hold-out points the same way. Beyond the satellite GHI itself (importance 0.23), the second-ranked predictor is surface pressure (0.22), ahead of every radiative variable. Daily pressure variation is small; between-station pressure differences, driven by elevation, are persistent. A model that leans on pressure is triangulating *which station* an observation belongs to and

applying a station-specific offset—efficient memorisation, not atmospheric physics. The structural precondition is visible in the raw data: because all five stations share one satellite radiation pixel, the satellite predictors cannot distinguish stations whose reference climatologies differ by up to $4 \text{ MJ m}^{-2} \text{ d}^{-1}$, so between-station variance in the target vastly exceeds the between-station variance in the predictors, and any flexible learner fills the gap with whatever station fingerprints the auxiliary variables provide.

Training-set ablation under the temporal design (gradient boosting, evaluated on 2024). Removing SEMPUR from training leaves the other stations unaffected but destroys the correction at SEMPUR, showing that the model learns station identity rather than transferable satellite-error structure. $\text{MJ m}^{-2} \text{ d}^{-1}$

Training set	Evaluated at	RMSE	MBE
All five stations	SEMPUR	2.02	+0.58
All five stations	Other four stations	2.67	+0.31
Excluding SEMPUR	SEMPUR	4.78	+3.90
Excluding SEMPUR	Other four stations	2.72	+0.30

Discussion

One dataset, two honest answers

Read separately, the temporal and LOSO results appear contradictory: the same gradient-boosting model is simultaneously the best correction available and worse than doing nothing. The protocol resolves the contradiction by attaching each number to its deployment question. For continuing operation at instrumented stations—gap-filling, retrospective harmonisation, extending a record forward—the temporal result governs, and the recommendation is clear: a regularised gradient-boosting correction, delivering a 21% RMSE reduction and near-zero seasonal bias. For mapping GHI at ungauged locations in this network’s region, the LOSO result governs, and the honest recommendation is equally clear: use the uncorrected satellite product, because every learned correction transfers negatively. A paper reporting only the first design would recommend deploying a model that degrades accuracy at exactly the locations bias correction is meant to serve; a paper reporting only the second would discard a correction that works well for the purpose most networks actually use it for.

Methodological implications

Three general lessons follow. First, validation design is a substantive modelling choice that must be matched to the deployment claim, echoing findings from spatial ecology (Roberts et al. 2017; Ploton et al. 2020; Meyer and Pebesma 2022) and time-series evaluation (Bergmeir and Benítez 2012), and it belongs in the methods section with the same prominence as the learning algorithm. Second, model-

family rankings are design-dependent: the common practice of selecting the most flexible learner by random cross-validation and then asserting regional applicability combines the design that most favours flexibility with the claim that flexibility least supports. Third, mechanism diagnostics are cheap and decisive. The SEMPu ablation required one additional model fit; permutation importance required none. Together they converted a bare performance gap into an interpretable statement—the model memorises station identity through elevation fingerprints—that tells a reader precisely when the correction can and cannot be trusted. We suggest that studies correcting clustered environmental data report, at minimum, one within-network and one between-network design, plus an identity diagnostic of this kind.

For an operating agency the translation is direct. A provincial meteorological or water-resources office that wants to extend, gap-fill, or harmonise records at its existing stations should calibrate a regularised boosting correction on the full multi-station history and revalidate annually with a rolling temporal hold-out; the 21% dispersion gain and the near-elimination of the anomalous station's offset are available today. The same office asked to supply irradiance for an uninstrumented district should hand over the uncorrected satellite series with its published uncertainty, and treat any pressure from users for a “corrected” value as a request it cannot honestly fulfil until the network spans enough satellite pixels—or gains a pyranometer at the target site—for spatial transfer to be testable. The protocol's value to such an office is that both recommendations come from one analysis, with the evidence for each attached.

A reporting checklist for corrections of clustered data

The protocol condenses into five reporting items that a reader, reviewer, or meta-analyst can check in any study correcting satellite or reanalysis products against a station network. (1) *Network structure*: state the number of stations, their spatial arrangement relative to the product's grid resolution, and how many stations share a grid cell; this single fact determines whether spatial transfer is even identifiable from the data. (2) *Reference provenance*: state whether the ground truth is measured directly or derived (e.g., from sunshine duration), and report a network-consistency screen of each station against its neighbours before attributing any discrepancy to the satellite. (3) *Design–claim matching*: for every accuracy number, name the resampling design and the deployment question it answers; a claim of regional applicability requires a station-blocked design, full stop. (4) *Baseline discipline*: report the uncorrected product under the same designs, since a correction that loses to its own input is worse than no model. (5) *Identity diagnostics*: report at least one probe—an ablation, a grouped-importance analysis, or an inspection of predictors that proxy station identity—showing whether the model's skill survives removal of station-specific information. None of these items requires new data or expensive computation; all five could be verified from this paper's tables alone, which is the standard we suggest the genre adopt.

The checklist also clarifies what this study does *not* claim. It does not claim that ML corrections of satellite irradiance are generally futile; with a network spanning many satellite pixels and direct pyranometer references, the between-station signal that transfer requires would exist, and flexible learners might well transfer. It claims that whether they transfer is an empirical question whose answer

is invisible to the designs most studies report, and it supplies the minimal design pair that makes the answer visible.

The quality-control stage carries its own methodological weight. SEMPUs + 21.7% “satellite bias” would, in a single-station study, have been reported as a property of NASA POWER in tropical conditions and corrected as such. Network-level screening reattributed it to the ground instrument, using the shared satellite pixel as a natural control. Where the reference itself is derived from sunshine recorders, as across much of the tropics, distinguishing satellite error from reference error is a precondition for the phrase “bias correction” to mean what it says.

Limitations

The network is small—five stations inside one radiation grid cell—which is simultaneously the study’s limitation and its point: this is the data regime in which most tropical correction studies operate, and the regime in which validation-design effects are strongest. The LOSO conclusion is conditional on that structure; a network spanning many satellite pixels would give the predictors genuine between-station signal and could support transferable corrections. The reference GHI is Angström–Prescott-derived rather than pyranometer-measured, so part of the residual error budget belongs to the reference; the protocol’s screening stage mitigates but cannot eliminate this. Hyperparameters were deliberately held fixed across designs; tuning within each design would sharpen absolute numbers but could not reverse a gap of the size observed. Finally, three years of data limit the temporal design to a single hold-out year, though the close agreement between random and temporal designs suggests stable inter-annual error structure.

Conclusion

We proposed a four-stage protocol for building and validating machine-learning bias corrections of satellite irradiance in data-sparse networks and demonstrated it end-to-end on a tropical monsoon network in Central Java. The protocol’s dual-validation core produced two answers that a single-design study would have collapsed into one: gradient boosting cuts RMSE by 21% and removes seasonal bias for deployment at calibrated stations, while no learned correction—and the flexible ones least of all—should be transferred to unseen locations within this network’s satellite pixel, where the uncorrected product remains the best available estimate. Diagnostics traced the divergence to memorisation of station identity, visible both in a one-fit ablation and in the prominence of an elevation fingerprint among the predictors. The broader claim is simple and portable: in clustered observational settings, the validation design is part of the model, and reporting it in matched pairs—one design per deployment claim—costs little and prevents the two symmetric errors of deploying corrections that degrade accuracy and discarding corrections that work.

Declarations

Acknowledgements

The authors thank the Central Java Provincial Public Works Agency (PUSDATARU) for access to the climatological station records, and the NASA POWER project for open provision of satellite data.

Declarations

Funding

The authors would like to thank Universitas Diponegoro for providing financial support for the Article Processing Charge (APC) of this publication.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Data availability

The dataset supporting this study's findings is openly available in Mendeley Data at <https://data.mendeley.com/datasets/43kx3kppgh/1> (DOI: 10.17632/43kx3kppgh.1). Ground observations originate from the provincial climatological network of Central Java; satellite data are openly available from the NASA POWER project (<https://power.larc.nasa.gov>).

Code availability

Analysis code (Python 3.12, scikit-learn 1.8) reproducing all tables and figures will be archived with the dataset.

Author contributions

E.R. conceived the study, collected and curated the station and satellite data, developed the analysis code, performed all statistical and machine-learning analyses, prepared Figures 1–3 and Tables 1–4, and wrote the main manuscript text. I.S. and U.Y. contributed to the study conception and methodological design, supervised the research, and critically revised the manuscript for important intellectual content. All authors reviewed and approved the final manuscript.

Supplementary information

The complete station–satellite dataset (5,075 daily records, 31 variables) is provided as Online Resource 1.

References

1. Allen, Richard G., Luis S. Pereira, Dirk Raes, and Martin Smith. 1998. "Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements." *FAO Irrigation and Drainage Paper* 56.
2. Ångström, Anders. 1924. "Solar and Terrestrial Radiation." *Q. J. R. Meteorol. Soc.* 50: 121–26. <https://doi.org/10.1002/qj.49705021008>.
3. Antonanzas, Javier, Natalia Osorio, Rodrigo Escobar, Ruben Urraca, Francisco J. Martínez-de-Pisón, and Fernando Antonanzas-Torres. 2016. "Review of Photovoltaic Power Forecasting." *Sol. Energy* 136: 78–111. <https://doi.org/10.1016/j.solener.2016.06.069>.
4. Bergmeir, Christoph, and José M. Benítez. 2012. "On the Use of Cross-Validation for Time Series Predictor Evaluation." *Inf. Sci.* 191: 192–213. <https://doi.org/10.1016/j.ins.2011.12.028>.
5. Besharat, Fariba, Ali A. Dehghan, and Ahmad R. Faghih. 2013. "Empirical Models for Estimating Global Solar Radiation: A Review and Case Study." *Renew. Sustain. Energy Rev.* 21: 798–821. <https://doi.org/10.1016/j.rser.2012.12.043>.
6. Breiman, Leo. 2001. "Random Forests." *Mach. Learn.* 45: 5–32. <https://doi.org/10.1023/A:1010933404324>.
7. Bright, Jamie M. 2019. "Solcast: Validation of a Satellite-Derived Solar Irradiance Dataset." *Sol. Energy* 189: 435–49. <https://doi.org/10.1016/j.solener.2019.07.086>.
8. Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. 2019. "All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." *J. Mach. Learn. Res.* 20 (177): 1–81.
9. Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Ann. Stat.* 29 (5): 1189–1232. <https://doi.org/10.1214/aos/1013203451>.
10. Gueymard, Christian A. 2014. "A Review of Validation Methodologies and Statistical Performance Indicators for Modeled Solar Radiation Data: Towards a Better Bankability of Solar Projects." *Renew. Sustain. Energy Rev.* 39: 1024–34. <https://doi.org/10.1016/j.rser.2014.07.117>.
11. Hoerl, Arthur E., and Robert W. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12 (1): 55–67. <https://doi.org/10.1080/00401706.1970.10488634>.
12. Journée, Michel, and Cédric Bertrand. 2011. "Quality Control of Solar Radiation Data Within the RMIB Solar Measurements Network." *Sol. Energy* 85 (1): 72–86. <https://doi.org/10.1016/j.solener.2010.10.021>.
13. Maraun, Douglas. 2016. "Bias Correcting Climate Change Simulations—a Critical Review." *Curr. Clim. Change Rep.* 2: 211–20. <https://doi.org/10.1007/s40641-016-0050-x>.

14. Maraun, Douglas, Theodore G. Shepherd, Martin Widmann, Giuseppe Zappa, Daniel Walton, José M. Gutiérrez, Stefan Hagemann, et al. 2017. "Towards Process-Informed Bias Correction of Climate Change Simulations." *Nat. Clim. Change* 7: 764–73. <https://doi.org/10.1038/nclimate3418>.
15. Meyer, Hanna, and Edzer Pebesma. 2021. "Predicting into Unknown Space? Estimating the Area of Applicability of Spatial Prediction Models." *Methods Ecol. Evol.* 12 (9): 1620–33. <https://doi.org/10.1111/2041-210X.13650>.
16. ———. 2022. "Machine Learning-Based Global Maps of Ecological Variables and the Challenge of Assessing Them." *Nat. Commun.* 13: 2208. <https://doi.org/10.1038/s41467-022-29838-9>.
17. Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *J. Mach. Learn. Res.* 12: 2825–30.
18. Ploton, Pierre, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, et al. 2020. "Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models." *Nat. Commun.* 11: 4540. <https://doi.org/10.1038/s41467-020-18321-y>.
19. Polo, Jesús, Carlos Fernández-Peruchena, Vasileios Salamalikis, Luis Mazorra-Aguiar, Mathilde Turpin, Luis Martín-Pomares, Andreas Kazantzidis, Philippe Blanc, and Jan Remund. 2020. "Benchmarking on Improvement and Site-Adaptation Techniques for Modeled Solar Radiation Datasets." *Sol. Energy* 201: 469–79. <https://doi.org/10.1016/j.solener.2020.03.040>.
20. Polo, Jesús, Stefan Wilbert, José A. Ruiz-Arias, Richard Meyer, Christian Gueymard, Marcel Sári, Luis Martín, et al. 2016. "Preliminary Survey on Site-Adaptation Techniques for Satellite-Derived and Reanalysis Solar Radiation Datasets." *Sol. Energy* 132: 25–37. <https://doi.org/10.1016/j.solener.2016.03.001>.
21. Prescott, James A. 1940. "Evaporation from a Water Surface in Relation to Solar Radiation." *Trans. R. Soc. S. Aust.* 64: 114–18.
22. Roberts, David R., Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Aroita, Severin Hauenstein, et al. 2017. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40 (8): 913–29. <https://doi.org/10.1111/ecog.02881>.
23. Rumbayan, Meita, Asifujang Abudureyimu, and Ken Nagasaka. 2012. "Mapping of Solar Energy Potential in Indonesia Using Artificial Neural Network and Geographical Information System." *Renew. Sustain. Energy Rev.* 16 (3): 1437–49. <https://doi.org/10.1016/j.rser.2011.11.024>.
24. Stone, Mervyn. 1974. "Cross-Validatory Choice and Assessment of Statistical Predictions." *J. R. Stat. Soc. Ser. B* 36 (2): 111–47. <https://doi.org/10.1111/j.2517-6161.1974.tb00994.x>.
25. Urraca, Ruben, Thomas Huld, Ana Gracia-Amillo, Francisco J. Martínez-de-Pisón, Frank Kaspar, and Andres Sanz-Garcia. 2018. "Evaluation of Global Horizontal Irradiance Estimates from ERA5 and COSMO-REA6 Reanalyses Using Ground and Satellite-Based Data." *Sol. Energy* 164: 339–54. <https://doi.org/10.1016/j.solener.2018.02.059>.

26. Voyant, Cyril, Gilles Notton, Soteris Kalogirou, Marie-Laure Nivet, Christophe Paoli, Fabrice Motte, and Alexis Fouilloy. 2017. "Machine Learning Methods for Solar Radiation Forecasting: A Review." *Renew. Energy* 105: 569–82. <https://doi.org/10.1016/j.renene.2016.12.095>.
27. White, Jeffrey W., Gerrit Hoogenboom, Paul W. Wilkens, Paul W. Stackhouse, and James M. Hoell. 2011. "Evaluation of Satellite-Based, Modeled-Derived Daily Solar Radiation Data for the Continental United States." *Agron. J.* 103 (4): 1242–51. <https://doi.org/10.2134/agronj2011.0038>.
28. Wolpert, David H. 1992. "Stacked Generalization." *Neural Netw.* 5 (2): 241–59. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).
29. Yagli, Gokhan M., Dazhi Yang, and Dipti Srinivasan. 2019. "Automatic Hourly Solar Forecasting Using Machine Learning Models." *Renew. Sustain. Energy Rev.* 105: 487–98. <https://doi.org/10.1016/j.rser.2019.02.006>.

Figures

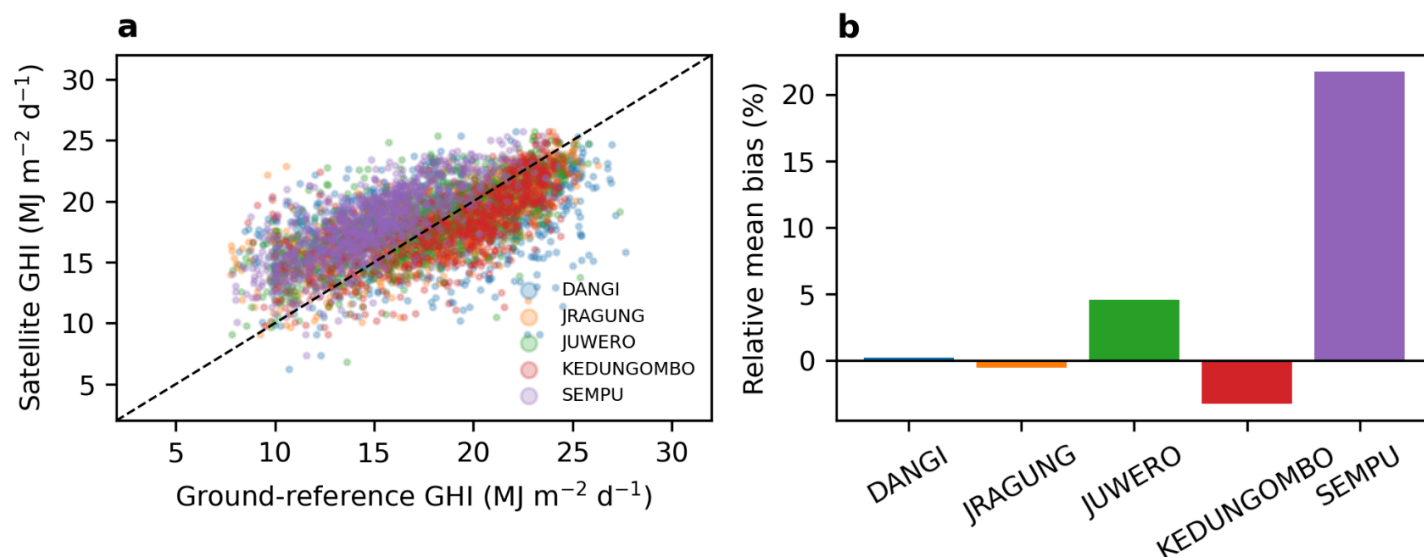


Figure 1

Uncorrected satellite performance across the network, 2022–2024. **a** Daily satellite versus ground-reference GHI by station; the dashed line is the 1:1 relation. **b** Relative mean bias by station. Four stations lie within $\pm 5\%$; SEMPUR shows $+21.7\%$, an offset traceable to its sunshine recorder rather than to the satellite product, since all five stations share one satellite radiation grid cell

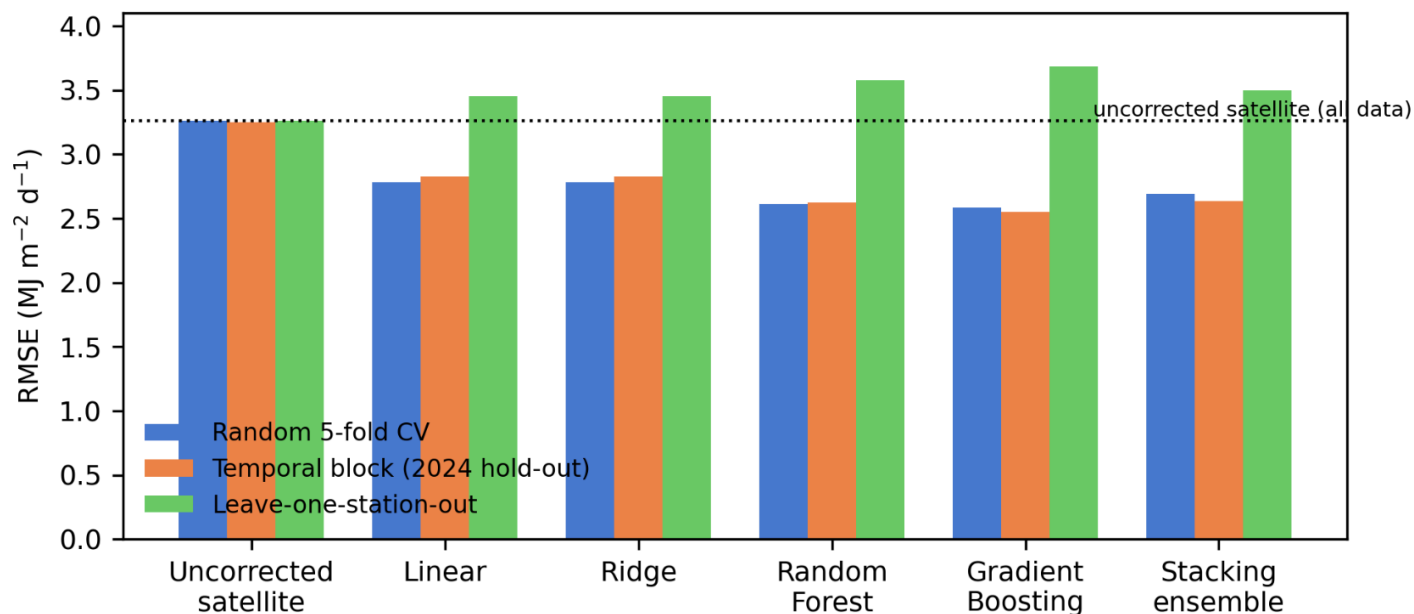


Figure 2

RMSE of each model family under the three validation designs. Random cross-validation and the temporal hold-out agree closely and reward flexibility; leave-one-station-out validation inverts the ranking and places every learned correction above the uncorrected-satellite benchmark (dotted line)

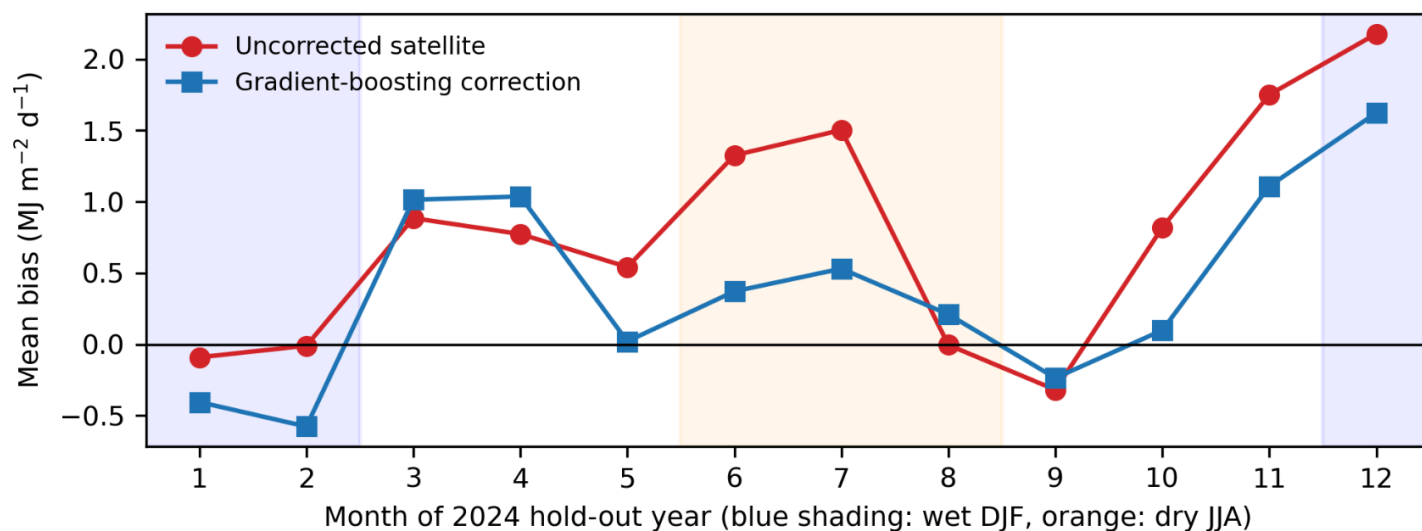


Figure 3

Monthly mean bias in the 2024 hold-out year for the uncorrected satellite product and the gradient-boosting correction. Shaded bands mark the wet (DJF, blue) and dry (JJA, orange) seasons. The correction removes the seasonal bias structure that the raw product carries

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [resultslosostation.csv](#)
- [resultsmain.csv](#)
- [resultssempu.csv](#)
- [resultsseasonal.csv](#)
- [resultstemporalstation.csv](#)
- [featureimportance.csv](#)