

Validation of the pvSpline photovoltaic simulation engine against 49 grid-connected systems in three climates: a no-calibration protocol

pvSpline
contact@pvspline.fr

Draft of October 19, 2026

Abstract

Simulation software underpins the design and financing of photovoltaic (PV) plants, yet published accuracy evidence for commercial tools is scarce. We validate the pvSpline PV simulation engine against measured AC production from 49 grid-connected systems in four public datasets, spanning hot-desert, temperate-oceanic and temperate-continental climates at 1 to 60 min resolution (53 million records, about 470 system-years). No model parameter is calibrated against the scored series and every site parameter is provenance-tracked; two nameplate back-fits are disclosed. Median capacity-normalized RMSE is 7.1 % fleet-wide and 5.8 % on the 31 fixed-tilt desert systems, whose median energy deviation is -2.2% ; energy deviation is within $\pm 15\%$ for 36 of 49 systems. The error tail is dominated by tracker failures, meter outages and two metadata errors, diagnosed, corrected and reported per site. Per-site attribution puts the defensible sub-model envelope at a median 0.8 points on the fixed-tilt fleet, against 4.5 points for a weather-input swap and a 7.1-point residual: data quality and inputs dominate, sub-model choice does not. Substituting reanalysis for on-site weather raises the desert RMSE from 5.8 % to 9.9 %. A same-input pvlib-python re-implementation predicts a median 8 points more energy on clean desert records, a gap the loss terms it omits close arithmetically; NREL PVWatts and end-to-end PVGIS, with heavier flat losses, sit closer on the temperate rooftops, where our uniform climate priors prove too light; and PVGIS reaches a fleet-median absolute energy deviation of 10.9 % against 9.6 % for pvSpline.

Keywords: PV performance modeling; simulation validation; energy yield; measured PV data; error attribution; weather input

1 Introduction

Yield simulation is the quantitative backbone of photovoltaic project development: it drives system sizing, bankability assessments, performance guarantees, and the arbitration of under-performance claims. The de-facto industry standard, PVsyst [1, 2], and alternatives such as NREL’s System Advisor Model [3] and PVWatts [4] (the v5 manual; the current engine is v8), and the open-source pvlib-python [5, 6] implement broadly similar physics chains, but differ in sub-model choices, default loss parameters, and implementation detail. Cross-tool comparisons against measured plants have been conducted at Sandia [7] and NREL [8], and community benchmarking is coordinated through the PV Performance Modeling Collaborative [9] — most recently as *blind* intercomparisons, in which participants predict withheld measured output from shared inputs [10], the closest published analogue of this paper’s no-calibration discipline, applied at inter-modeler rather than single-tool scope; individual sub-models have been validated extensively in isolation (e.g. transposition [11], separation models [12]). Whole-tool validations, however, are typically conducted by the tool’s own authors on proprietary datasets and seldom disclose site-parameter provenance.

Table 1: This validation against prior whole-tool validations, on the two study-design axes above. “Provenance” is whether each site parameter is traceable to a recorded source. Freeman et al. compared four tools against nine instrumented systems; the last row characterizes the common tool-author study (e.g. Sandia [7], and vendor validations) rather than a single work.

Study	Systems	Provenance	Public data
This work (pvSpline)	49	tracked ^a	yes
PVPMC blind intercomparison [10]	6	blind ^b	released after
Freeman et al. 2014 [8]	9	not stated	not fully
Tool-author studies (typical)	few–tens	seldom	seldom

^aper-field provenance recorded in the machine-readable configurations, with the disclosed nameplate back-fits of Section 3.2 the only exceptions. ^bthe blind protocol enforces no-calibration by construction (measured output withheld from participants); the organizers, not the participants, control provenance.

This paper presents an accuracy validation of pvSpline, a C++/Qt desktop PV design and simulation application whose engine implements a PVsyst-aligned model chain and directly imports PVsyst component libraries (.PAN/.OND module and inverter files). Our study design targets two weaknesses common in prior tool validations:

1. **Parameter-provenance discipline.** Every site parameter is traceable to manufacturer nameplate metadata, dataset documentation, or a published literature prior (climate-class soiling rates, technology-class degradation rates). No parameter was tuned against the measured power series used for scoring, and the machine-readable site configurations record the provenance of each value (Section 3.2).
2. **Climate and system diversity.** The validation fleet spans 49 scored systems across four independent public datasets in three Köppen climate classes (BWh, Cfb, Cfb/Dfb), fixed-tilt and tracked mounting, residential and research-facility scale, and native time resolutions from 1 min to 60 min (Section 3).

Table 1 situates the study against prior whole-tool validations along these two axes. Cross-tool comparisons at Sandia [7] and NREL [8] and the community benchmarking coordinated through the PV Performance Modeling Collaborative [9] established the measured-vs-modeled methodology, but — being tool inter-comparisons or author-run studies — they typically draw on proprietary or partner-instrumented systems and seldom record per-parameter provenance. The distinction this paper claims is not accuracy alone but the combination of a public fleet and a no-calibration provenance discipline.

The contributions are: (i) a documented, citable description of the pvSpline irradiance-to-AC model chain (Section 2); (ii) a no-calibration validation protocol with IEC 61724-1-aligned metrics [13] (Section 4); (iii) fleet-wide accuracy results including sub-model ablations and weather-input error separation (Section 5); and (iv) same-input comparisons against pvlib-python and NREL’s PVWatts together with an end-to-end comparison against the PVGIS screening service (Section 5.8).

2 The pvSpline simulation engine

pvSpline converts a per-step weather record (global horizontal irradiance or measured plane-of-array irradiance, ambient temperature, wind speed, rainfall, humidity) and a static site configuration (geometry, nameplate, thermal and electrical parameters) into predicted AC power through an ordered chain: solar geometry → decomposition → transposition → optical losses (including the soiling accumulator, applied to the plane-of-array irradiance) → spectral correction → cell temperature → DC conversion → DC losses and degradation → inverter — the

Table 2: Sub-models of the pvSpline chain in execution order. Options marked (d) are the defaults used for the headline results.

Stage	Model	Reference
Solar position	NREL SPA	[15]
Decomposition GHI→DNI/DHI	Erbs (d); Engerer2	[16, 17]
Clear-sky reference (Engerer2)	Ineichen–Perez	[18]
Transposition → POA	Hay–Davies (d); Perez	[19, 20]
Incidence-angle modifier	Martin–Ruiz (beam)	[21, 22]
Air mass	Kasten–Young	[23]
Spectral correction	Lee–Panchula (+PWAT)	[24, 25]
Cell temperature	Faiman	[26, 27]
DC power	linear- γ (d); De Soto; PVsyst	[1, 2, 28]
DC losses	quadratic ohmic; flat mismatch	PVsyst convention
Inverter	tabulated η -curve + AC clip	EN 50530 η_{euro} ; [29]
Soiling	Kimber + exponential wash	[30, 31]
Degradation	linear LID + annual rate	[32]

canonical structure of hourly PV performance models [14]. Table 2 summarizes the sub-models; the remainder of this section specifies each stage. The validation harness calls the same physics functions as the production solver — solar position, decomposition and transposition, incidence-angle and spectral corrections, cell temperature, the DC models, and the inverter efficiency curve — but assembles them in its own per-record driver over the measured weather. It does not run the product’s simulation pipeline, which by default feeds those functions a PVGIS typical year and applies its own default loss stack, so the results in Section 5 score the shared model chain rather than every product default.

2.1 Solar geometry

Solar zenith and azimuth are computed with the NREL Solar Position Algorithm [15] (stated accuracy $\pm 0.0003^\circ$ over the years -2000 to 6000), including nutation, aberration, topocentric parallax, and atmospheric refraction. The angle of incidence on the collector follows the standard spherical relation. For tracked systems, the per-step effective orientation is derived from the sun vector using either the astronomic (AOI-minimizing) rotation or an irradiance-maximizing search, bounded by the configured rotation limit and, where a ground-coverage ratio is available, by backtracking; a wind-stow override supersedes both whenever measured wind speed exceeds the site’s stow threshold.

2.2 Decomposition

When no measured diffuse component exists, GHI is split using the per-step clearness index $k_t = \text{GHI}/I_{0h}$, where I_{0h} is extraterrestrial horizontal irradiance with Spencer’s eccentricity correction [33]. (The Erbs coefficients are fit to hourly data; following common practice they are applied at each dataset’s native resolution, 1–60 min.) The default is the Erbs piecewise polynomial [16]; the alternative is the Engerer2 logistic separation model [17] (paper predictor conventions: $\Delta k_{tc} = k_{tc} - k_t$, AST as decimal hour-of-day, and the $\beta_5 K_{de}$ cloud-enhancement term; the published 1-minute coefficient set — Engerer 2015 publishes no hourly set), whose clear-sky reference uses the Ineichen–Perez closed form [18] driven by site altitude and Linke turbidity, with the Spencer eccentricity factor and pressure-corrected air mass. A temporal self-consistency guard clamps input GHI to the physical ceiling I_{0h} — hourly-mean GHI paired with instantaneous low-sun geometry otherwise produces unbounded DNI — and every clamped row is counted and reported in the per-site diagnostics rather than silently corrected.

2.3 Transposition

POA irradiance is assembled from beam, sky-diffuse, and ground-reflected components (albedo 0.20 throughout this study). The default sky-diffuse model is Hay–Davies [19]; the alternative is the Perez model [20] with the published all-sites composite coefficient set. Both models rank among the better-performing transposition (not decomposition) models in empirical inter-comparisons, with hourly relative RMS differences of roughly 6–9% against measured tilted irradiance [11], and the difference between the two concentrated in clear-sky circumsolar and horizon-brightening conditions — precisely the axis probed by the transposition ablation of Section 4.3. Sites with a plane-of-array pyranometer bypass decomposition and transposition entirely; for tracked systems, a fixed reference sensor is not representative of the moving plane, and the GHI-derived path is always used.

2.4 Optical and spectral losses

Beam irradiance is attenuated by the Martín–Ruiz incidence-angle modifier [21] with angular characteristic $a_r = 0.157$ (consistent with the 0.16–0.17 range Martín & Ruiz fit for uncoated c-Si front glass); diffuse and ground components use fixed hemispheric factors of 0.95 and 0.85 — the sky factor matches the Martín–Ruiz diffuse extension’s sky integral [21] (≈ 0.953 – 0.956 over 20 – 35° tilt); the ground factor is a fixed engineering constant (the corresponding Martín–Ruiz ground integral is strongly tilt-dependent, 0.67 – 0.83 over the same range, reaching 0.85 only near 38 – 40°). On the measured-POA path, where no component split exists, a clear-sky-weighted blended factor of 0.96 is applied. A spectral mismatch factor — the First Solar / Lee–Panchula closed form [24] with the published per-technology coefficients, evaluated at the reference water column when no humidity data exists (PWAT otherwise from the Gueymard closed form [25]) and normalized to unity at the ASTM G-173 reference — multiplies the irradiance seen by the electrical model; cell temperature deliberately uses the uncorrected POA.

2.5 Cell temperature

Cell temperature uses the Faïman model [26], standardized in IEC 61853-2 [27], with PVsyst’s thermal-balance convention: $T_{\text{cell}} = T_{\text{amb}} + \alpha(1 - \eta) G_{\text{POA}} / (U_c + U_v v)$ with $\alpha = 0.9$, $\eta = 0.15$, and U_c keyed to mounting class ($29 \text{ W m}^{-2} \text{ K}^{-1}$ free-standing — PVsyst’s mounting default, defined against this absorbed flux) and the wind term dropping out when no wind series exists.

2.6 DC conversion and DC losses

Three module models are selectable per site, with automatic downgrade on parameter-extraction failure (PVsyst $\rightarrow$ De Soto $\rightarrow$ linear): the linear- γ model $P_{\text{dc}} = P_{\text{STC}} (G_{\text{POA}}/1000) (1 + \gamma(T_{\text{cell}} - 25))$; the De Soto five-parameter single-diode model [28] with parameters extracted from datasheet STC values by Newton iteration (including De Soto’s linear bandgap temperature dependence $E_g(T) = E_{g,\text{ref}}(1 - 0.0002677(T - T_{\text{ref}}))$, applied consistently in both the extractor and the operating-point translation; note De Soto (2006) validated the model on mono-Si, poly-Si, thin-film-Si, and triple-junction a-Si — applying it to CdTe and HIT, as this study’s fleet does, extrapolates beyond that validated technology set); and the PVsyst variant [1, 2] with irradiance-dependent shunt resistance $R_{sh}(G) = R_{sh,\text{base}} + (R_{sh,0} - R_{sh,\text{base}}) e^{-R_{sh,\text{exp}} G/G_{\text{ref}}}$, where $R_{sh,\text{base}} = \max((R_{sh,\text{ref}} - R_{sh,0} e^{-R_{sh,\text{exp}}})/(1 - e^{-R_{sh,\text{exp}}}), 0)$ so that $R_{sh}(G_{\text{ref}}) = R_{sh,\text{ref}}$ exactly. For the single-diode variants, per-module maximum-power output is scaled to the array by the ratio of system to module nameplate. Two PVsyst-aligned derates are applied before the inverter: a load-aware quadratic ohmic wiring loss (fractional loss $L_{\text{ohmic}} \cdot P_{\text{dc}}/P_{\text{STC}}$) and a flat mismatch fraction.

2.7 Inverter

DC power passes through a tabulated efficiency curve — a single-variable load-fraction curve of the family analyzed in [29], using a typical six-point shape normalized to 96 % *European* weighted efficiency (EN 50530 η_{euro} weights — not the CEC protocol’s [34]) for every site in this study (per-site nameplate rescaling is exercised only by the AC-electrical companion checks of Section 4.4). The load fraction is normalized to the inverter DC nameplate, derived from the AC rating assuming a typical 0.96 DC-side over-rating where the rating is documented, and to the array’s STC nameplate otherwise. AC output is clipped at the inverter AC rating where documented; sites without a known inverter rating are not clipped. Clipping is essential for residential systems with 1.3–2× DC/AC oversizing, whose annual energy is otherwise systematically under-predicted.

2.8 Time-varying losses: soiling and degradation

Soiling follows a daily accumulator combining Kimber-style dry deposition [30] with a continuous exponential rain wash (this work, replacing the all-or-nothing rain-threshold cleaning of both Kimber [30] and the HSU model of Coello & Boyle [31] with a smooth decay):

$$L_{d+1} = L_d e^{-r_d/\tau} + \rho f_{\text{tilt}}, \quad (1)$$

with day- d rain total r_d , wash constant τ , deposition rate ρ , and a linear tilt factor f_{tilt} (1 at $\leq 5^\circ$, 0 at 90° — an engineering construction of this work; Kimber’s published model takes no tilt input); the loss is capped at 25 % and applied *optically* to the plane-of-array irradiance entering the module (upstream of the thermal, spectral, and IV models and the inverter clip). r_d is the per-source-day rainfall *total*, summed over per-row increments (cumulative-daily channels such as DKASC’s are differenced at load time), and the accumulator opens at its periodic steady state — one recurrence pass over the record’s own rain history — rather than at a pristine panel. Degradation uses the two-parameter linear form $f(t) = 1 - \text{LID}/100 - (r_{\text{annual}}/100) \max(0, t_{\text{years}})$ with fractional years t_{years} measured from a July-1 mid-year commissioning anchor (only the install year is documented), applied per row on the DC side upstream of the inverter, with technology-class rates from the Jordan–Kurtz compendium [32]. An optional climate-aware mode Arrhenius-scales the annual rate by the site’s 98th-percentile cell temperature, with an effective activation energy of 0.30 eV spanning the encapsulant- and interconnect-driven mechanism bands (moisture-permeation kinetics $\approx 0.35\text{--}0.38$ eV [35]; hot-climate failure modes surveyed in [36] — neither publishes a power-degradation E_a , so 0.30 eV is an engineering assumption, with the T_{98} percentile metric following IEC TS 63126); it is disabled for the headline results.

2.9 Scope

Production features not exercised by this validation — horizon and row-to-row shading, bifacial rear-side irradiance, bypass-diode partial-shading IV, snow coupling, transient (lumped-RC) cell temperature, and module-specific IAM profiles from .PAN files — are omitted conservatively: the validated sites are unshaded, monofacial reference systems. A GPU-raytraced 3D extension, in which the application’s Unreal Engine site viewer bakes dimensionless near-shading, bifacial rear-gain, and substring-mismatch factor tables from full 3D site geometry while the engine described here retains all radiometry, is under development and will be validated separately against reference raytracing and shaded instrumented sites.

Table 3: Validation datasets. “POA” indicates an on-site plane-of-array sensor used directly (bypassing decomposition/transposition). “Span” is the per-system length of the scored (inner-joined) record. In total, 53.0 million measured–predicted pairs covering roughly 470 system-years were scored.

Dataset	Location (Köppen)	Systems	Res.	Span	Weather	POA
DKASC [37]	Alice Springs, AU (BWh)	34 ^c	5 min	7–17 yr	on-site	yes ^a
StoreNet [38]	Dingle, IE (Cfb)	10	1 min	1 yr	on-site GHI	no
OPSD CoSSMic [39]	Konstanz, DE (Cfb)	1	60 min	2 yr	reanalysis ^b	no
Trivedi [40]	Dingle, IE (Cfb)	4	30 min	2.5–2.7 yr	reanalysis ^b	no

^afixed-tilt systems only; tracked systems use the GHI-derived path.

^bOpen-Meteo/ERA5-family archive co-located with the site.

^cof 39 configured meters; five excluded for measurement corruption (Table 4).

3 Validation datasets and site configuration

3.1 Datasets

Four independent public datasets are used (Table 3 and Figure 1). **DKASC** — the Desert Knowledge Australia Solar Centre [37] publishes 5-minute AC power for tens of co-located systems of diverse module technologies (mono-Si, poly-Si, CdTe, CIGS, amorphous, and one CPV concentrator) and mounting (fixed, single- and dual-axis tracked), with an on-site weather station providing GHI, DHI, tilted irradiance, temperature, wind, humidity, and rainfall.¹ **StoreNet** [38] provides 1-minute production for ten residential rooftop systems in Dingle, Ireland, with a shared local weather station (GHI). **OPSD CoSSMic** [39] provides 60-minute cumulative energy for a rooftop system in Konstanz, Germany (the CoSSMic `industrial1` building), differenced to mean power. **Trivedi** [40] provides 30-minute PV energy for four residential “ambassador” households, also on the Dingle peninsula (the label *Trivedi* is the harness’s historical nickname for this dataset, retained here so tables match the archived reports; the data descriptor is by Avgoloupis et al. [40], a follow-up to the StoreNet release). The four households share an identical technology package — a 2.1 kWp rooftop array, a 5 kWh DC-coupled hybrid battery, an EV, and a heat pump — with PV metered as gross generation by a dedicated meter upstream of the battery; the descriptor publishes no tilt or azimuth, and meter outages are zero-filled in the released files. Both gaps matter in Section 5. Köppen climate classes follow the Peel et al. world map [41]. Where no on-site weather exists, hourly meteorology is drawn from the Open-Meteo archive [42], which serves the ERA5 reanalysis family [43]; the production application additionally supports PVGIS [44] as a prospective-weather source. This heterogeneity is exploited in Section 5.7 to separate weather-input error from model error.

3.2 Parameter provenance

Site parameters follow a strict provenance discipline, recorded per field in the machine-readable site configurations:

- *Geometry and nameplate* (tilt, azimuth, kWp, inverter rating, module technology, install year): manufacturer/operator metadata from the dataset portals and documentation.
- *Soiling* (Eq. 1): Köppen-class literature priors (identical for all systems within a climate class); no per-site fitting against the validated series. The BWh (desert) class carries the Kimber-family deposition/wash prior; the Cfb/Dfb (European) classes carry an explicit,

¹Desert Knowledge Australia, the Australian Government, the Northern Territory Government and the project managers, Ekistica do not endorse, and accept no legal liability whatsoever arising from, or connected to, the outcomes and conclusions associated with the use of data from the Desert Knowledge Australia Solar Centre.

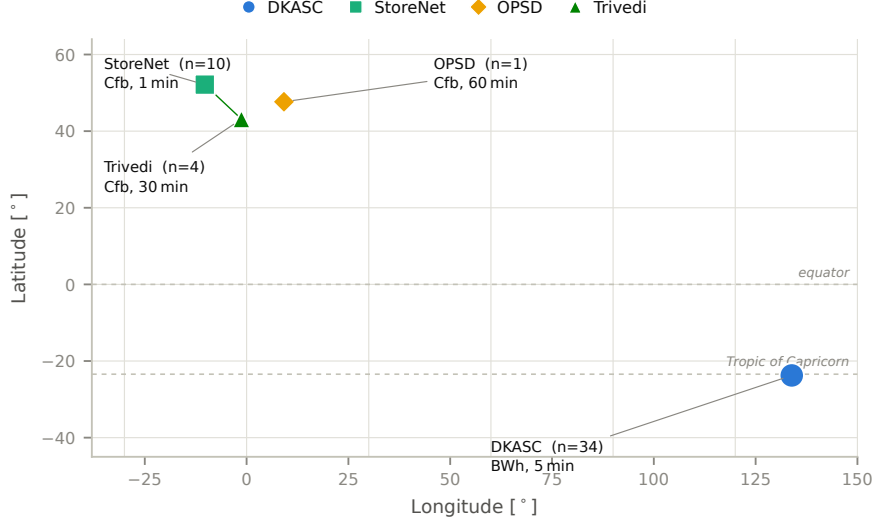


Figure 1: Geographic and climatic spread of the validation fleet: the four datasets by site location, marker area proportional to scored-system count and colour keyed to the dataset palette used throughout. The desert fleet (DKASC, Alice Springs) sits on the Tropic of Capricorn; the European fleets are temperate mid-latitude, with StoreNet and Trivedi co-located at Dingle (the Trivedi marker offset for legibility). A longitude/latitude locator, generated from the site configurations by `doc/paper/scripts/make_figures.py --fleetmap`.

provenance-tagged *zero* deposition prior — frequent rain keeps the equilibrium soiling level near zero, and no rainfall channel is loaded for those datasets — a posture Section 5 prices via the performance-ratio gap, and which any nonzero literature prior would revise only as a disclosed sensitivity leg, never as a silent headline change.

- *Degradation*: technology-class median rates from the Jordan–Kurtz compendium [32] on the DKASC fleet, whose portal documents commissioning years. The European strata (StoreNet, OPSD, Trivedi) carry an explicit, provenance-tagged *zero* for both aging and LID — no commissioning date is armed in their configurations, and the engine disables the derate without an install year — so the system-years those records accumulated before and during the measurement window contribute unbooked early fade. The posture is disclosed because it acts in the over-prediction direction on exactly the strata Section 5 reports over-predicted; arming descriptor-provenanced commissioning dates with the same class priors is future work to be stated before the sweep that adopts it, like any prior refinement under this protocol.
- *Thermal* (U_c): IEC 61853-2 mounting-class defaults.
- *Electrical derates*: the chain’s load-aware ohmic and flat mismatch derates (Section 2) are configured to *zero* on every validated site — no as-built wiring or string-layout metadata exists for any of them, and inventing PVsyst-typical values would be exactly the unprovenanced tuning this protocol forbids. The zero posture is disclosed here because it is load-bearing for the baseline gap accounting of Section 5.8 and the year-zero level analysis of Section 5.5: real wiring and mismatch losses exist on these plants but appear in the residual, not in the model.

No parameter of any kind was calibrated against the measured power series used for scoring, with one disclosed exception: for two DKASC meters (M19_C, M4_C.II) the nameplate is not recoverable from the portal, and their kWp was back-fit as the 99.5th percentile of the measured

Table 4: Excluded sites (all DKASC; no StoreNet, OPSD, or Trivedi system was excluded). The first four are configuration-level skips (`skip` with a recorded `skipReason`); the last was quarantined automatically at run time by the capacity-sanity gate.

Site	Mechanism	Reason
M2_C	manual skip	Per-row corruption of <code>Active_Power</code> (spikes to 154 MW on a ≈ 5 kWp meter); sister phase M2_A on the same meter is clean.
M3_A, M3_B, M3_C	manual skip	<code>Active_Power</code> implausibly large across the whole M3 cluster (300–500 kW per phase on few-kWp systems); suspected unit or signal mismatch, pending raw-data investigation.
M2_B	auto-quarantine	Measured peak 25.9 kW on a 2.16 kWp nameplate ($\approx 12\,000$ W kWp ⁻¹ , above the 5 000 W kWp ⁻¹ gate); a single-row sensor spike.

power series divided by the fleet’s peak-to-nameplate anchor (1.286 kW kWp⁻¹, established on the portal-documented systems). This anchors only the nameplate scale of those two sites — no loss or physics parameter — to a robust peak statistic of the validated series; both sites are flagged in the machine-readable configurations and in the per-site results (Appendix A), and removing them moves the DKASC median nRMSE_{cap} by less than 0.3 percentage points (6.3 % $\rightarrow$ 6.1 %). The *initial* headline run (2026-07-12) additionally carried a nameplate back-fit *not* covered by that disclosure, and it is this study’s cautionary tale: the four Trivedi households ran at 3.6–4.9 kWp derived from their observed measured peaks, values a post-hoc cross-check against the data descriptor shows to be $\times 1.71$ –2.33 the true, uniform 2.1 kWp — the observed peaks turn out to be meter register artifacts, not production (Section 5). The configurations now carry the descriptor nameplate with per-field provenance and produced the headline run reported here; Trivedi tilt and azimuth remain disclosed assumptions (south-facing, 35°), as the descriptor publishes neither.

3.3 Data-quality gates and exclusions

Two mechanisms guard against corrupt measurements dominating the metrics, and both are disclosed rather than silent. (i) A capacity-sanity gate auto-quarantines any site whose measured peak exceeds 5 000 W per kWp of nameplate ($\approx 3.9\times$ the fleet’s clear-sky peak ratio), indicating schema or unit corruption. (ii) For datasets whose meters record outages as zero-or-trickle readings rather than missing rows (DKASC, Trivedi), a per-site *system-offline filter* excludes daylight rows whose measured power is below 10 % of predicted while the model predicts at least 1 W; every excluded row is counted, the per-site dropped fraction is reported alongside the metrics (Appendix B), and the excluded rows are exported to the per-site report for audit. On the 31 fixed-tilt DKASC systems the filter removes a median 5.2 % of the pre-filter inner join (up to 30.8 % on M19_C); these fractions become load-bearing in Section 5.8, where external tools score the same records *without* the filter. The row-wise gate is blind between its threshold and full output by design — a partial outage (one string of three, one inverter of two) sits exactly there, as the M11_3 case study of Section 5 shows — so the analysis tooling additionally runs a persistence-based *detector*: the daily measured-to-predicted energy ratio is weather-neutral (predictions are driven by the same weather record, so overcast days move both series together), and a run of ≥ 14 days below a 0.6 ratio is reported as a candidate outage epoch with its median ratio and its contribution to ΔE , cross-checked against the site’s declared `availabilityEpochs`. Detection never masks anything by itself — masking authority stays with the human-confirmed, portal-provenanced config entries — it exists to surface suspected undocumented *step* epochs (dead record heads, unrecorded failure onsets — the finds of Section 5.5) for exactly that confirmation; progressive drifts that never sustain the ratio threshold are, by the same division of labor, the longitudinal fit’s territory. Table 4 lists every excluded site with its reason.

4 Methodology

4.1 Prediction protocol

Each system is simulated independently from its weather record and static configuration; the engine never sees measured power (the outage filter of Section 3.3 compares measured against predicted rows, but only to *exclude* rows from scoring — it feeds nothing back into the model). Predictions are scored against measurements on the *inner join* of the two series by timestamp: rows present in only one series are excluded from both the error statistics and the energy integrals, so data-coverage gaps cannot masquerade as model bias. Negative measured values (e.g. inverter standby draw) are retained to keep the bias estimate honest. Peak statistics are computed on each full series. Time stamps follow the start-of-interval convention throughout; dataset-native conventions are converted at load time.

4.2 Metrics

For the joined set $\{(m_i, p_i)\}_{i=1}^n$ of measured and predicted power, we report the IEC 61724-1-aligned [13] set:

$$\text{MBE} = \frac{1}{n} \sum_i (p_i - m_i), \quad \text{MAE} = \frac{1}{n} \sum_i |p_i - m_i|, \quad (2)$$

$$\text{RMSE} = \left(\frac{1}{n} \sum_i (p_i - m_i)^2 \right)^{1/2}, \quad (3)$$

$$\text{nRMSE}_{\text{mean}} = \text{RMSE} / \bar{m}, \quad \text{nRMSE}_{\text{cap}} = \text{RMSE} / P_{\text{STC}}, \quad (4)$$

$$R^2 = 1 - \frac{\sum_i (p_i - m_i)^2}{\sum_i (m_i - \bar{m})^2}, \quad \Delta E = \frac{\sum_i p_i \Delta t_i - \sum_i m_i \Delta t_i}{\sum_i m_i \Delta t_i}, \quad (5)$$

where Δt_i is the record interval, so ΔE compares inner-joined *energy* integrals. Positive MBE and ΔE indicate over-prediction. MAE, RMSE, $\text{nRMSE}_{\text{mean}}$, MBE, and R^2 are additionally reported in a daylight-only variant, masked by solar elevation from the SPA; $\text{nRMSE}_{\text{cap}}$ and ΔE are all-hours quantities. Daylight-only variants avoid the optimistic dilution of error statistics by trivially-correct night hours; capacity-normalized nRMSE enables comparison across system sizes.

Interval estimates and tests reported alongside the medians use two conventions throughout. 95 % confidence intervals on stratum medians are site-level percentile bootstraps (20 000 resamples, fixed seed); because sites within a dataset share a weather record, these quantify site-resampling variability rather than fully independent-site sampling. Paired sub-model comparisons (Section 4.3) are tested with the two-sided Wilcoxon signed-rank on per-site deltas — exact null distribution up to $n = 20$, normal approximation with tie and continuity corrections above. Both are regenerated from the archived run summaries by the manuscript tooling (Section 7).

4.3 Model configurations

The headline configuration uses the chain defaults of Table 2. Four sensitivity axes are evaluated on the identical fleet:

- *Transposition.* The harness sweeps its transposition models in a single run by construction — Hay–Davies and Perez [11], each reported separately per site. A third slot, Perez driven from raw horizontal PVGIS irradiance, exists for the production PVGIS pipeline; on this fleet, where every site already drives transposition from GHI, it reproduces the Perez results identically and serves as a consistency check of the sweep plumbing (Section 5.2). The headline numbers pin Hay–Davies.

- *Decomposition.* Erbs vs. Engerer2 — separation-model choice is a known first-order error source for GHI-only pipelines [12].
- *DC model.* Linear- γ vs. De Soto vs. PVSyst single-diode [2, 28], on the subset of sites with full datasheet parameters.
- *Temporal resolution.* Hourly records are re-shaped into energy-conserving sub-hourly records along the clear-sky envelope (per-record means preserved exactly, so hourly energy is unchanged while intra-hour peaks reappear), quantifying the resolution error that hourly averaging hides — inverter-curve nonlinearity, and clipping where an AC rating is documented. Native sub-hourly datasets (DKASC 5 min, StoreNet 1 min) pass through unchanged.

4.4 AC-electrical companion checks

Beyond power, the harness validates AC-side electrical quantities on the subset of DKASC meters that report per-phase current, line voltage, apparent power, and signed power factor. Predicted DC power is passed through the inverter efficiency curve (rescaled to the site’s nameplate European efficiency where documented) and compared against the measured AC quantities, yielding a realized-efficiency check that isolates the inverter stage from the irradiance chain. A voltage-drop check computes the IEC code-book drop for a synthetic compliant reference run (10 m, 6 mm² Cu at the measured load and ambient temperature) and anchors $|V_{\text{measured}} - V_{\text{nominal}}|$ against it; this scores the wiring model against a compliant baseline, not the site’s unknown as-built cabling, and is reported as a plausibility bound rather than a point prediction.

4.5 Baselines

Three external baselines are scored — two *same-input* tools (pvlib-python and NREL’s PVWatts engineering model), which isolate implementation differences by holding the weather and site parameters fixed, and one *end-to-end prospective* service (PVGIS), which is deliberately fed nothing but the site geometry. Every external leg scores on its own inner join against the same measured series and applies no system-offline filter (Section 3.3) — a deliberate asymmetry that Section 5.8 exploits diagnostically. All three are reported in Section 5.8.

PVSyst. PVSyst is not among the scored baselines. The `pvsyst-compare` subcommand builds a procedural PVSyst $\leftrightarrow$ pvSpline parity suite — matched project pairs sweeping a feature matrix over hand-built PVSyst template archetypes, since PVSyst’s CLI can run but not create projects — but every pair sits at a single location (Paris) on its typical-year weather, so the suite compares the two tools with each other rather than either with measured output. It cannot score this fleet and is outside the scope of this paper.

pvlib-python. A baseline script (`validation/scripts/pvlib_baseline.py`; requires pvlib $\geq$ 0.13, developed against 0.15.2) re-implements the chain of Section 2 entirely with pvlib-python [5, 6] on the *same* machine-readable site configurations and weather records: NREL SPA solar position, the same GHI physical-ceiling guard, Erbs decomposition, Hay–Davies transposition (albedo 0.20), Martin–Ruiz beam IAM ($a_r = 0.157$) with the fixed 0.95/0.85 diffuse/ground factors, First Solar (Lee–Panchula) spectral correction with Gueymard precipitable water, Faiman cell temperature with the site’s U_c , linear- γ DC conversion, the same LID-plus-linear degradation arithmetic, and a 96 %-nominal inverter efficiency curve with AC clipping. Three documented scope limits remain: tracked sites are excluded (no drop-in pvlib twin for the tracker rotation/stow logic); pvSpline’s PVSyst-convention DC derates and Kimber–Coello soiling accumulator have no exact pvlib equivalent and are omitted, which biases the baseline high by roughly the omitted loss terms (on this fleet the DC derates are configured to zero — Section 3.2 — so

the omission reduces to the soiling accumulator); and the baseline always drives transposition from GHI, whereas pvSpline uses the measured plane-of-array channel on the DKASC fixed-tilt systems whenever the tilted-pyranometer reading is present for that row (Table 3). That channel is not continuous, however: across the fixed-tilt fleet a median of 31 % of scored weather rows fall back to the same GHI→POA Hay–Davies path the baseline uses (per-site GHI→POA fallback fractions are in the archived warnings), so the DKASC pair is a partial-measured-POA path against a fully GHI-driven re-implementation — roughly 69 % measured-POA, not two identical irradiance paths and not a pure measured-versus-derived contrast either. This makes the near-uniform clean-record gap below (attributed to the omitted loss terms) if anything a conservative read, since a third of the rows already share the irradiance path. All three limits are noted in the script and its output header, which also pins the pvlib and pandas versions of each run.

NREL PVWatts. A second same-input leg (`validation/scripts/pysam_baseline.py`; requires `nrel-pysam`) hands the *same* weather and site geometry to NREL’s PVWatts v8 engineering model [4] through PySAM, the Python binding to the open-source SAM Simulation Core (SSC) [3]. PVWatts is license-free and among the most widely *used* reference yield models, with published validations available — though in the NREL cross-tool study [8] the then-current PVWatts v1 was one of the two outliers (average −11.9 %); the modern v5+ engine run here substantially reworked those losses. It supplies an independent, license-free second opinion alongside the pvlib re-implementation. Where the pvlib leg re-implements the pvSpline chain stage by stage, this leg deliberately runs PVWatts’s *own* transposition (Perez), NOCT-based cell temperature, DC model, single-parameter system loss, and inverter, isolating the end-to-end system model rather than any single stage. Two design choices preserve the same-input contract. First, PVWatts transposes from beam and diffuse, not global horizontal (a GHI-only resource table is rejected), so this leg splits GHI with the *same* Erbs decomposition as the pvlib leg and hands PVWatts the resulting DNI/DHI — the only thing that changes between the two same-input legs is the POA→AC system model. Second, PVWatts is an hourly engineering model: the leg aggregates the native record to hourly and scores energy-first, exactly as the PVGIS leg does, so its nRMSE is hourly and, like PVGIS, is not compared against the native-resolution runs (a discipline the comparison generator enforces). Three scope limits mirror the pvlib leg and PVWatts’s design: tracked sites are excluded (fixed open-rack array only); the explicit ohmic/mismatch/soiling submodels collapse into PVWatts’s single system-loss parameter, left at its 14 % default (a screening-grade choice, as with PVGIS); and PVWatts sets the temperature coefficient from a discrete module class rather than the per-site γ , so the configured γ is advisory here. Two implementation notes: SSC requires a contiguous single-year 8760-hour grid, so each calendar year is run separately on a full-year grid (leap days dropped, absent hours zero-filled and never scored), and daylight-saving zones are handled by running the leg on a fixed standard-time offset derived from the site `tzId`, so no wall-clock hour is duplicated or skipped across a transition. The leg has run over the same 41 systems as the pvlib leg (31 fixed-tilt DKASC systems and 10 StoreNet houses; driven by `run_baselines.py` alongside the pvlib and PVGIS legs; summaries in `validation/reports/pysam_baseline/`); its column appears in Table 14 and is read in Section 5.8.

PVGIS (end-to-end prospective). Unlike the two same-input baselines, PVGIS [44] cannot be fed the validation weather: it is queried as a complete prospective stack — its own irradiance database driving its own PV model with a flat user-level system-loss parameter. It therefore answers a different, practically central question: *what would a free screening study have predicted for these exact systems?* A third script (`validation/scripts/pvgis_baseline.py`) queries the v5.3 hourly API (`seriescalc` with PV calculation) once per site directly from the machine-readable site configurations — geometry converted to PVGIS’s equator-facing azimuth

convention, dual-axis-tracked sites requested as such, module technology mapped to the nearest `pvtechchoice` — caches the responses, and scores the returned hourly AC power against hourly-aggregated measurements (UTC, start-of-interval; hours with under half their expected sub-hourly rows are dropped). Four protocol disclosures: (i) the radiation database is auto-assigned by PVGIS and recorded per site — SARA3 satellite data for the European sites, but ERA5 reanalysis at Alice Springs, which lies outside satellite coverage, so the DKASC leg is an ERA5-driven comparison; (ii) the system-loss parameter stays at the PVGIS default of 14% — that is what a screening user gets, and tuning it would defeat the framing; (iii) PVGIS output is hourly, so its nRMSE and R^2 are not comparable to the native-resolution runs — energy deviation is the cross-tool metric, a discipline enforced by the comparison generator (`validation/scripts/compare_baselines.py`); (iv) PVGIS applies its own DEM-derived terrain-horizon shading by default (`usehorizon` on), which neither `pvSpline` nor the `pvl` twin models on this fleet — an asymmetry that favors PVGIS wherever local terrain actually shades a site (the Dingle peninsula sites are the candidates), and that the fetch script’s `--horizon off` switch is designed to isolate; the paper-data driver runs that A/B leg as its own step (`pvgishorizon`, summaries kept apart in `pvgis_baseline_nohorizon/` and never folded into the comparison tables). The A/B has now run on the full data: switching PVGIS’s terrain horizon off shifts its per-site ΔE by at most 0.5 pp (median +0.3 pp, PVGIS predicting slightly *more* without terrain shading, as expected), on the Irish coastal sites included. Terrain-horizon modeling therefore explains essentially none of PVGIS’s temperate-site agreement — the asymmetry is real but immaterial on this fleet, and the DEM-horizon hypothesis for the remaining Cfb gap can be retired.

5 Results

All numbers in this section come from the canonical full-dataset transposition-sweep run of `pvSpline validate`, archived at `validation/reports/20260730-021242/` (the stamp pinned in `doc/paper/canonical_runs.json`, which the statistics and table generators read) and run with both metadata corrections below (Trivedi nameplate, M11.3 identity). Retained re-execution demonstrations bracket it at two strengths. Within a single session and binary, the fleet was re-scored end-to-end twice on 2026-07-22 (20260722-155212 and 20260722-165442). Across rebuilds, the canonical run is the fifth execution of one sweep: 20260723-013216, 20260723-201917, 20260723-233610 and 20260725-052241 precede it, and the five span four build commits (69ed856, 3075551b, 18bc374, 8d5215b). The merges between the first three leave the simulation chain untouched: the interpolation kernels are unchanged, the only `src/calculation` diff is an additive chart-geometry helper, and the rest lands in the job-registry, catalogue and `pvsyst-compare` layers, with the validation harness itself extended additively (no line removed). The interval to the fourth does land engine code — Fresnel anti-reflective IAM support surfaced for the comparison suite of Section 4.5, sub-hourly far-horizon shading, and virtual-battery credit-expiry accounting — but none of these paths is exercised by this fleet’s configurations (no site config carries an IAM or horizon profile, and the scored channel is PV production upstream of any household storage), and the bit-identical fifth execution is the end-to-end demonstration. All five agree on every per-site metric exactly — 49 of 49 sites, every reported statistic — as do the same-session pair, and automated claims of the linter in Section 7 enforce both agreements against whichever sweep is pinned. Relative to the 2026-07-22 pair, the canonical engine gained the 2026-07-22 accuracy and data-quality follow-ups (the precipitable-water spectral term armed from the DKASC humidity channel, a row-level plausibility gate and changepoint outage detection in the filter stack, and a robust 30-minute peak statistic with QCRAD-lite weather QC in the diagnostics layer); these are deliberate, disclosed changes: they moved per-site energy deviation by a median of 0.2 pp (at most 0.7 pp, on M6_A), a bound the linter also enforces. The earlier transition — the yield-physics

audit and its two accuracy follow-up passes after the superseded 2026-07-17 canonical run (20260717-050926, retained alongside, as are the 2026-07-14 and 2026-07-13 sweeps that reproduced it) — moved per-site energy deviation by a median of 2.7 pp (at most 10.6 pp, on Trivedi id03); after each transition, every literal in this draft was re-derived from the new archives. The metadata-correction chain below is unaffected: its basis, the initial 2026-07-12 sweep at commit 8db89c8, is retained at `validation/reports/20260712-193827/` as the basis of the two case studies below. 49 systems are scored under each of the three transposition slots (147 site-model rows). The generated report (`summary.md` plus one JSON per transposition model) is archived in the source tree; per-site values are reproduced in Appendix A. Unless stated otherwise, numbers refer to the chain defaults (Hay-Davies, Erbs, linear- γ). Baseline numbers (Section 5.8) come from the archived runs of the baseline scripts of Section 4.5 (pvlib 0.15.2, pandas 3.0.3; PVGIS v5.3 API), whose per-site summaries are archived in `validation/reports/pvlib.baseline/` and `validation/reports/pvgis.baseline/` and reproduced in Appendix B. The ablation, electrical, and weather-substitution numbers of Sections 5.2–5.7 come from dedicated harness runs archived under `validation/reports/ablation_*/` and alongside the sweep; each family’s canonical stamp is pinned in `canonical_runs.json`, regenerated with the canonical sweep (ablation_engerer2/20260730-025641, ablation_desoto/20260730-032435, ablation_substep/20260730-030946, ablation_weathersub/20260730-031016, and the satellite leg ablation_weathersat/20260730-033919), with every earlier stamp — the 2026-07-17 family regenerated with the superseded canonical sweep, the 2026-07-14 and 2026-07-13 runs (row-identical to it at reported precision except the pre-armed Engerer2 variant, disclosed in Section 5.2), and the two disclosed no-op runs — retained alongside. Dataset identifiers, the citable data DOIs the originals carry, and the checksum manifest that pins the corpus byte-for-byte are given in Section 7.

5.1 Fleet-wide accuracy

Table 5 summarizes accuracy by dataset over the 53.4 million scored records, Figures 2 and 3 show the per-dataset distribution and the per-site picture, and Figure 4 the underlying measured-vs-predicted density for one representative system per dataset. Fleet-median capacity-normalized RMSE is 7.1 % (IQR 5.5–18.0 %; 95 % CI 6.0–9.7 %), fleet-median R^2 is 0.940, and fleet-median energy deviation is +1.0 % (95 % CI –2.2 to +8.1 %). Whole-record energy deviation $|\Delta E|$ is within 10 % for 28 of 49 systems, within 15 % for 36, and within 25 % for 43. Note that the error statistics are computed over *all* hours of the inner join at native resolution; night rows, which both series predict as (near-)zero, inflate R^2 relative to the daylight-only variant recorded in the per-site harness reports, so the R^2 column should be read as a fleet-internal ranking rather than an absolute goodness-of-fit claim. Figure 5 illustrates the ordered per-stage loss cascade the engine applies, from plane-of-array irradiance to delivered AC energy, for a representative grid-connected fixed-tilt configuration; the system-side derates in its lower panel are the structure the same-input pvlib baseline omits — though on the validated fleet the ohmic and mismatch entries are configured to zero (Section 3.2), so there the omission reduces to soiling (Section 5.8).

Fixed-tilt desert systems. The 31 fixed-tilt DKASC systems — 7–17 years of 5-minute data each, on-site weather, five module technologies — form the cleanest stratum: median $\text{nRMSE}_{\text{cap}}$ 5.8 % (95 % CI 5.5–6.3 %), median R^2 0.959, and $|\Delta E| \leq 15$ % for 28 of 31 systems. The energy bias leans negative (20 of 31 systems; median –2.2 %, 95 % CI –8.8 to +0.3 % — the interval includes zero). The longitudinal decomposition of Section 5.5 locates the conservatism in the *year-zero* loss stack — the uniformly applied climate-class soiling and quality terms (Section 3.2) — while the technology-class degradation priors turn out to under-book the fleet’s real fade, so the deviation rises across each record, crossing from negative to positive, rather than holding steady.

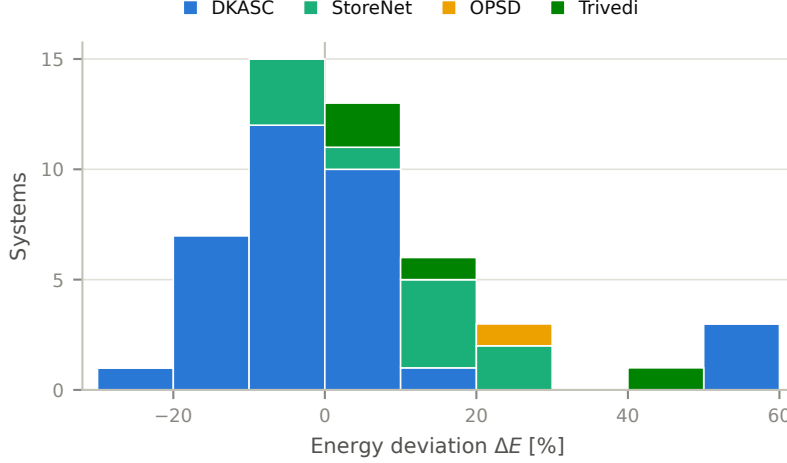


Figure 2: Distribution of whole-record energy deviation ΔE by dataset (10-point bins, stacked; chain defaults, Hay–Davies slot of the archived canonical 20260730–021242 sweep). Generated from the archived run summary by `doc/paper/scripts/make_figures.py`.

Table 5: Fleet-wide accuracy by dataset (chain defaults: Hay–Davies, Erbs, linear- γ). Medians with interquartile ranges in brackets; all-hours inner join at native resolution. The three dual-axis-tracked DKASC aggregate meters are broken out from the 31 fixed-tilt DKASC systems.

Dataset	n_{sites}	med. nRMSE _{cap} [%]	med. R^2	med. ΔE [%]
DKASC (all)	34	5.9 [5.3, 8.2]	0.958	−1.4 [−9.7, +4.7]
fixed-tilt	31	5.8 [5.3, 6.7]	0.959	−2.2 [−10.7, +3.5]
tracked ^a	3	20.1 [17.4, 21.0]	0.384	+51.9 [+32.5, +54.1]
StoreNet	10	19.4 [19.0, 20.0]	0.102	+15.2 [+2.1, +19.4]
OPSD	1	15.3	0.546	+21.8
Trivedi	4	10.1 [9.6, 11.1]	0.656	+13.9 [+8.8, +24.2]
Fleet	49	7.1 [5.5, 18.0]	0.940	+1.0 [−8.1, +12.1]

^adual-axis-tracked systems behind combined multi-phase meters (M9_AC, M9_B, M10_BC).

Tracked aggregate meters. The three dual-axis-tracked systems are measured behind combined multi-phase meters and carry the fleet’s largest errors (ΔE +13.0 % to +56.3 %). The portal cross-check of Section 6 adds structural explanations: M10_BC is a SolFocus *CPV concentrator* (triple-junction cells, purely DNI-driven) that the chain can only model as a flat-plate array, and the portal’s maintenance log records recurring tracker-drive and inverter failures across the tracked fleet — availability losses an always-tracking simulation deliberately does not model. A per-era decomposition of the M9 meters makes the availability attribution quantitative: while its DEGER trackers actually tracked (2013–2016), M9_AC scored −1 to +14 % annual energy deviation at 2 240–2 260 kWh kWp^{−1}, three of the four years within $\pm 4\%$ — the two-axis prediction is close to unbiased when the hardware works — and its +56 % whole-record figure is its portal-documented 2019–2021 tracker/inverter failure epoch; M9_B runs the same arc with a 2023–2025 failure epoch, and since April 2022 five of the six Site-1 trackers are parked fixed (north, 10°). The stratum extreme, M11.3, turned out not to belong to the stratum at all — see the case study below. We report these sites individually rather than exclude them post hoc.

A second metadata case study: M11.3 (mistaken plant identity). As archived, M11.3 scores ΔE +122.8 % with R^2 −0.97 as a 31.5 kWp dual-axis-tracked system. A meter-level inves-

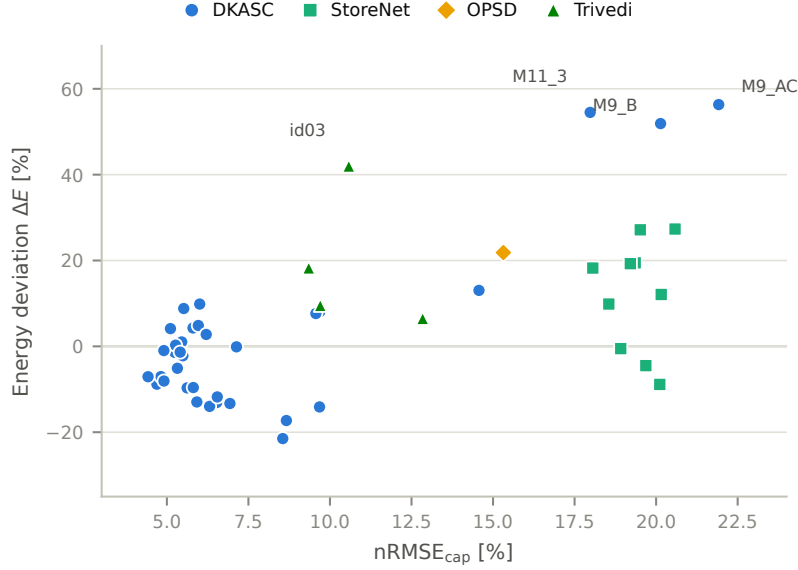


Figure 3: Per-site $\text{nRMSE}_{\text{cap}}$ vs. whole-record ΔE (chain defaults). The labeled DKASC systems are the tracked aggregate meters and the mis-joined M11.3 (both discussed in the text); id03 is the Trivedi extreme.

tigation (per-era energy decomposition of the per-site exports, then a portal cross-check of the data-export machinery itself) found the harness’s scraper had joined this meter’s data export to the wrong portal page: export 78 is meter M11 on *Site 2* — the eco-Kinetics 26.52 kWp mono-Si array on a single ADES dual-axis tracker (commissioned August 2010) — while the scraped “Dual Large Scale Trackers” page (31.5 kWp, Trina) belongs to export 102, the retired combined channel of the Site-1 trackers that M9_AC/M9_B already cover per-phase. The archived M11.3 error is therefore a stack of metadata-versus-plant mismatches: nameplate $\times 1.19$ (31.5 vs. 26.52 kWp); a tracking model applied to an array whose tracker was *disabled in September 2012* and fixed at tilt 20° /north ever since ($\approx \times 1.3$ – 1.4 in this climate); two of three inverters dead from ≈ 2016 until December 2021 (the 10–50 % -of-prediction daylight band that dominates those years is exactly the offline filter’s blind zone); a zero-filled pre-commissioning record head (2008–2010, filter-absorbed); and a $\approx 2\times$ metering-scale error over the first seven weeks of operation — the real origin of the “53 kW peak” on what is a 26.52 kWp array. The record’s one full tracking year closes the loop on the engine: 2011 scored $1936 \text{ kWh kWp}^{-1}$ under the wrong 31.5 kWp nameplate, which is $2302 \text{ kWh kWp}^{-1}$ under the true one — matching the portal’s own 61 049 kWh annual figure and an accurate tracked prediction. The configurations are corrected in-tree (identity, nameplate, fixed mounting, provenance URL, and a defect note), which reclassifies M11.3 into the fixed-tilt stratum, and the corrected re-run behaves as the diagnosis predicts: ΔE falls from +122.8 to +54.5 %, R^2 turns from -0.97 to $+0.31$, $\text{nRMSE}_{\text{cap}}$ from 26.2 to 18.0 %, and the site lands in the outage-affected fixed class — the residual is the disclosed, unmodeled availability history (two of three inverters dead 2016–2021), which both external baselines corroborate by missing *larger* on the same unfiltered record (pvlib +86.8 %, PVGIS +81.4 %, Section 5.8). The capacity anchor of Section 6 dates the same history from the meter side, model-free: per-year PVUSA-style realized capacity holds at 520–584 W kWp $^{-1}$ through 2013–2015, collapses to 122–283 across 2016–2021, and recovers to 584–620 from 2022 — the availability epoch in capacity units, the plant’s rating intact on either side. The protocol lesson complements the Trivedi one: the shallow portal cross-check of Section 6 had “confirmed” the 31.5 kWp nameplate by consulting the same mis-joined page — a cross-check against the error’s own source cannot catch it; only the meter-level data forced the contradiction.

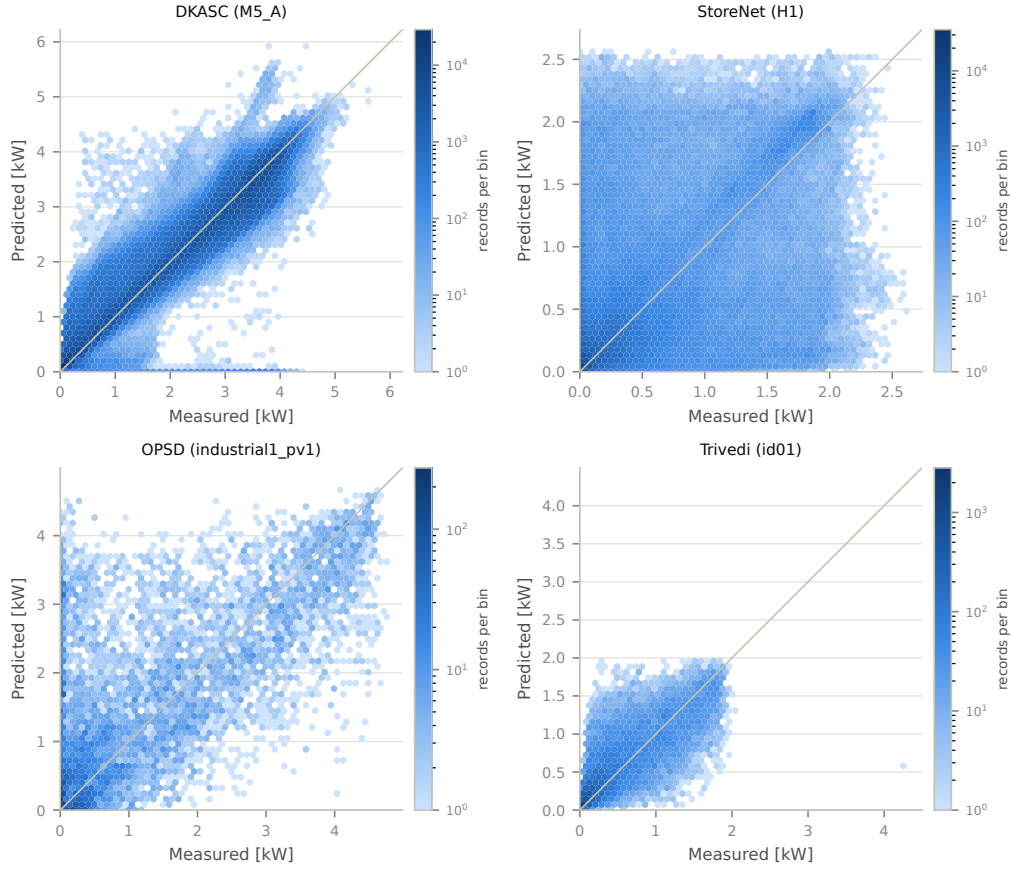


Figure 4: Measured vs. predicted AC power for one representative system per dataset (chain defaults, Hay–Davies slot of the archived sweep): hexagonal-bin density with logarithmic count shading and the identity line for reference. Scored daylight rows of the inner join; rows excluded by the offline filter of Section 3.3 are omitted. Generated from the sweep’s per-site exports by `doc/paper/scripts/make_figures.py --scatter` (the exports are not committed — any full run regenerates them).

Residential 1-minute systems. The ten StoreNet rooftops score $\text{nRMSE}_{\text{cap}}$ 18.1–20.6 % with low pointwise R^2 (median 0.102): at 1-minute resolution, a single shared GHI sensor cannot resolve per-roof cloud transients, so pointwise error is dominated by timing noise rather than systematic bias. The energy integral is the more meaningful quantity at this resolution: ΔE spans -8.9% to $+27.3\%$ (median $+15.2\%$) with a small median bias of $+31\text{ W}$ on 1.7–2.2 kWp systems. The baseline comparisons of Section 5.8 localize this median over-prediction to the inputs: the same-input pvlib twin reproduces it to within 5 points per house (so it is not the implementation), and PVGIS, with its heavier flat loss, removes most of it (so it is not the meters). The irradiance input is not the culprit either: substituting satellite irradiance for the shared sensor leaves the over-prediction in place (Section 5.7), which points to the loss priors and per-house metadata. The performance-ratio layer (Section 5, Table 6) sharpens the per-house reading: measured PR spans 0.72 to 1.02 across ten houses sharing one sensor and one assumed geometry, while the modeled PR is 0.95 on every house (shared inputs), so the stratum’s spread is dominated by per-house nameplate, geometry, or local-irradiance metadata rather than physics — H4/H5 sit at the low measured-PR pole (the $+27\%$ over-predictions), while H7/H10 measure PR near or above unity, marking their configured 1.7 kWp as understated or their local irradiance as above the shared station’s.

Reanalysis-driven sites. The single OPSD system over-predicts by $+21.8\%$ ($R^2 = 0.55$ at 60 min). Its -26% peak-anchor excursion is *not* evidence of a realized-capacity deficit: the predicted peak sits at the same level (4.74 kW modeled vs. 4.75 kW measured), so both series fall equally short of a 5-minute desert-derived anchor, as any hourly Cfb record must. The over-prediction instead localizes to near-zero-measured daylight epochs: a diagnostic 10%-of-predicted offline filter — which the headline deliberately does *not* apply to this dataset (Section 3.3) — removes 18.1 % of daylight rows and collapses ΔE from $+21.8$ to $+3.0\%$ (already $+6.3\%$ at a 5 % gate; Table 13), and the clear-sky bin carries the bias (Table 10, k_t 0.8–1.0). Snow cover — a declared scope omission (Section 2) that a Konstanz winter actually exercises — and/or meter outages zero-filled in the released record are the candidate causes, and the persistence detector of Section 3.3 now bounds the question: it segments 3 undeclared low-ratio epochs, every one confined to October–March (October 2015 to March 2016, plus two spanning winter 2016/17), jointly worth $+21.2\text{ pp}$ — nearly the site’s entire $+21.8\%$ over-prediction, so outside these winter windows the record is approximately unbiased ($+0.6\text{ pp}$ net). Their median daily ratios of 0.25–0.54 — persistent *partial* production, not the ≈ 0 of a dead meter — lean toward physical winter suppression (snow cover, with possible winter inverter trips) rather than zero-filled telemetry alone. A snowfall overlay — the site’s Open-Meteo daily snowfall series, committed beside the analysis artifacts (`snow.OPSD.industrial1_pv1.csv`) and crossed with the detected epochs — now splits that verdict per epoch. The two deep-winter epochs are as snowy as their winters: snowfall fell on 26 and 18 % of their days against whole-winter shares of 23 and 13 %, and between them they carry 96 % of the two winters’ cumulative snowfall (66 and 22 cm), so persistent snow cover remains a viable driver of their partial ratios. The short October–November 2016 epoch, by contrast, is effectively snow-free (0.4 cm over its span, 8 % of days), which hardens the telemetry-or-inverter reading for at least that one — the causes are mixed, not single. The snow scope exclusion stands (Section 2). A Marion-model sensitivity leg has run on this site (`validation/reports/ablation_snowmarion/20260730-034133/`) but cannot price that snow: its cover model starts only at an hourly snowfall of 1 cm, which the reanalysis record rarely reaches (49 covered daylight rows across 216 snowfall days, 0.05 % of energy), so it moves the site’s energy deviation by only -0.08 pp and neither prices nor rules out snow inside the detected epochs. Either way the miss is a bounded, seasonal, single-site effect, not a chain bias.

The Trivedi stratum: a nameplate-metadata case study. The headline numbers above (Table 5) show the four Trivedi households as an ordinary residential stratum: ΔE +6.5 to +42.0 % (median +13.9 %), $\text{nRMSE}_{\text{cap}}$ 9.3–12.8 %, R^2 0.52–0.76. That is the *corrected* state of what was, in the initial 2026-07-12 run (retained at `validation/reports/20260712-193827/`), the fleet’s heaviest outlier stratum: ΔE +69 to +163 % with R^2 of 0.00 to –3.4, despite predicted *peaks* 23–24 % *below* measured peaks. A post-hoc root-cause investigation traced that signature to the harness’s own site metadata rather than to the engine, the weather input, or (primarily) the meters. The initial run configured these systems at 3.6–4.9 kWp, back-derived from their observed 30-minute peaks because no nameplate had been located; the data descriptor, consulted after the fact, specifies an *identical 2.1 kWp array* for all four households (Section 3), so predictions were inflated $\times 1.71$ – 2.33 before any physics ran. The peaks that seeded the wrong nameplate are themselves artifacts, provable from the initial run’s summaries alone: each house’s best 30-minute slot reaches 1.29 – $1.31 \times$ the model’s best slot while its best *hour* reaches only 0.75 – $0.77 \times$ the model’s — a 1.74 – 1.79 ratio between two aggregations of the same series, possible only if the top slots are isolated single-slot spikes. The raw files indeed contain one or two register catch-up values per house, up to 2.95 kWh in one half-hour: a 5.9 kW “peak” on a 2.1 kWp array. The measurement channel is otherwise sound in kind — gross PV generation from a dedicated meter upstream of the battery — but its outages are zero-filled rather than flagged, including a 197-day run (id02) and two ≈ 63 -day runs at a mid-2021 metering-platform changeover during which one house demonstrably kept exporting to the grid; this is what the offline filter’s drop fraction — 14–21 % of the joined record in the initial run, 11.2–19.0 % after correction, as the 10 %-of-predicted gate re-equilibrates at the smaller predictions — was absorbing (Section 3.3). Because the linear- γ DC model scales predictions with nameplate, the correction was predictable algebraically (+69.4, +86.0, +163.1, +101.2 % $\rightarrow$ –1.2, +5.6, +20.1, –13.8 %), and the first corrected re-run landed within a few points of that, the gap being the filter re-equilibration just described; the post-audit canonical engine (Section 5) adds a further +7 to +11-point upward shift on this GHI-only stratum, to +9.5, +18.3, +42.0, +6.5 %. The largest residual (id03, +42.0 %) is consistent with the missing geometry metadata — its morning-shifted production centroid and reproducibly low clean-year yield (≈ 640 kWh kWp $^{-1}$ in both 2019 and 2020, vs. 847–896 for the healthiest house-years) suggest an east-shifted or shaded array that the assumed south/35° plane over-predicts. The clear-day orientation screen of Section 6 turns that centroid into a number: fitted purely from the *shape* of its twenty clearest days, id03’s best plane is 15° tilt at azimuth 130° — rotated toward the southeast and flattened, 27.5° from the assumed south/35° plane — with id04 deviating 48.7° under an east-northeast 25° fit, id02’s 42.0° deviation *dropped* by the screen’s identifiability guard (its fit recovers only 0.46 of the configured plane’s shape residual — the weakly identified azimuth of a 10° fitted tilt), and id01, the household that agrees best with the assumed geometry (+9.5 %), fitting the configured plane back to within 5.7°, unflagged — the control behaving exactly where the metadata is least in doubt. The id03 residual is also input-invariant: under the satellite source swap of Section 5.7 it stays at +36.1 % — consistent with geometry, not weather. Per the no-calibration protocol the fits are diagnostics, not corrections; a fitted-geometry sensitivity leg pricing the missing metadata stays future work. The corrected nameplate is in-tree with per-field provenance. Two protocol lessons stand. First, peak-derived nameplate — which the provenance discipline permits as a disclosed last resort (Section 3.2) — is unsafe on spike-contaminated channels; the cross-aggregation peak ratio above is a cheap guard. The performance-ratio layer of Table 6 adds the mirror-image lesson: the two sites whose nameplate *is* the disclosed percentile back-fit score measured PR of 1.02 (M19_C) and 0.95 (M4_C_II) — levels at or above the top of the plausible band, marking the back-fit as having *under-sized* both arrays (their –21.5 and –14.1 % energy deviations are the fleet’s rank-1 and rank-3 negative extremes), plausibly because a percentile peak taken over a long degraded, soiled record sits below the clean-peak anchor ratio. Peak-derived nameplate therefore fails in both directions — spike-inflated high (Trivedi), degradation-deflated low (the

Table 6: Measured and modeled performance ratio by stratum (IEC 61724-1; scored rows). The reference yield is the engine’s *effective* (post-IAM) in-plane irradiance, so absolute PR reads $\approx 4\%$ above a bare-pyranometer PR; the measured-vs-modeled contrast is unaffected (shared denominator). ΔPR is the median of per-site differences, so rows need not equal the difference of the two medians. $\text{PR}_{\text{meas}}^T$ is the IEC 61724-1 temperature-corrected measured PR (each row’s reference yield scaled by $1 + \gamma(T_{\text{cell}} - 25)$ with the modeled Faiman cell temperature, from the `t_cell_c` export column); ‘—’ = stratum configurations carry no module γ .

Stratum	med. PR_{meas}	med. PR_{pred}	med. ΔPR	med. $\text{PR}_{\text{meas}}^T$
DKASC fixed-tilt	0.781	0.755	-0.016	0.852
DKASC tracked	0.489	0.745	+0.256	0.547
StoreNet	0.803	0.947	+0.144	—
OPSD	0.777	0.954	+0.177	—
Trivedi	0.846	0.964	+0.118	—

back-fits) — and a measured-PR flag ($\text{PR} > 0.95$) joins the cross-aggregation ratio in the harness as the corresponding guard. The capacity anchor of Section 6 closes the loop from the standards side: the two back-fit sites are exactly the two whose PVUSA-style realized capacity per configured kWp sits far *above* the fleet (M19_C +32.8%, M4_C_II +30.0% against the 753 W kWp^{-1} median) — an understated denominator reads as excess capacity in every year of the record. Second, the failure was diagnosable entirely from published per-site artifacts — peak ratios across aggregations, drop fractions, and two independent external tools missing in the same direction (Section 5.8) — which is the argument for publishing them.

Performance ratio. For readers who think in the language of plant operations, we report the IEC 61724-1 performance ratio (PR): the ratio of final yield $Y_f = E_{\text{AC}}/P_{\text{STC}}$ to reference yield $Y_r = H_{\text{POA}}/G_{\text{STC}}$, computed per site over the scored rows as $\text{PR} = \sum_i P_i \Delta t_i / (P_{\text{STC}} \sum_i G_{\text{POA},i} \Delta t_i / G_{\text{STC}})$. Because measured and modeled PR share the same plane-of-array reference yield (the chain’s in-plane irradiance) and the same nameplate, their difference is exactly the whole-record energy deviation reported above; their *level* is the new information — where the fleet sits against the 0.75–0.85 band typical of well-run systems. Table 6 reports the levels. The healthy fixed-tilt desert stratum measures a median PR of 0.78 against a modeled 0.75 — the -2.2% energy bias in PR form. The tracked aggregate meters measure 0.47–0.49 on the two M9 meters (0.66 on the CPV concentrator): a whole-record PR of about one half quantifies their documented availability history directly, independent of any model. On the European rooftops the modeled PR sits far above the measured one — StoreNet median 0.95 modeled vs. 0.80 measured (median per-site gap +0.14; OPSD +0.18, Trivedi +0.12) — which prices the missing Cfb loss structure explicitly: these sites book *zero* soiling (an explicit, provenance-tagged zero prior, Section 3.2; Appendix C shows the accumulator inactive), no shading, and no snow. The temperature-corrected PR of IEC 61724-1 (the $\text{PR}_{\text{meas}}^T$ column of Table 6) now fills from the `t_cell_c` export column, and it completes the desert story: pricing out the thermal derate lifts the fixed-tilt stratum’s measured PR from 0.78 to 0.852 — squarely healthy once the one loss an Alice Springs summer makes unavoidable is removed — while the tracked meters’ corrected 0.547 remains far below any plausible plant state, confirming that their deficit is availability, not temperature. The column stays ‘—’ on the European and Indian strata, whose dataset metadata documents no module temperature coefficient to correct with.

5.2 Sub-model sensitivity

Transposition. The Hay–Davies/Perez choice is second-order on this fleet. Paired per-site differences (Perez minus Hay–Davies) in $\text{nRMSE}_{\text{cap}}$ have a fleet median of -0.002 percentage

points (Perez marginally lower on 27 of 49 sites; signed-rank $p = 0.203$, i.e. indistinguishable) and never exceed 0.38 points at any site; the kWp-weighted fleet nRMSE_{cap} is 11.22 % under either model. Perez systematically raises predicted energy — median $\Delta(\Delta E)$ is +0.15 points on DKASC, +1.6 on StoreNet, +1.7 on OPSD, and +2.8 on Trivedi (maximum +3.5); the direction is statistically unambiguous ($p < 10^{-8}$) even though its magnitude stays second-order — consistent with its larger circumsolar/horizon-brightening terms, and on the already-over-predicting Irish and German GHI-only sites this uniformly worsens agreement, while on DKASC it is neutral. The default (Hay–Davies) is therefore retained. The third sweep slot (Perez-from-raw-GHI, used by the production PVGIS pipeline) reproduced the Perez metrics identically on every site and metric, as expected when all sites already drive from GHI — a consistency check of the sweep plumbing rather than an independent result.

Decomposition. The Erbs→Engerer2 swap (dedicated run at the same commit, archived in `validation/reports/ablation_engerer2/`) engages only where the chain must split GHI itself: measured diffuse short-circuits decomposition, so DKASC — whose station reports DHI — moves by less than 0.01 pp on every metric (a structural null, not a finding; Figure 6), and measured-POA rows never reach the decomposition stage at all. On the three GHI-only datasets the choice is second-order and, on this fleet, mildly helpful: Engerer2 lowers predicted energy on every site (exact signed-rank $p < 10^{-4}$, $n = 15$) — StoreNet median ΔE +15.2 → +13.1 % (per-site shift −1.6 to −2.2 pp), OPSD +21.8 → +20.3 %, Trivedi +13.9 → +10.8 % (per-site shift −2.6 to −3.4 pp) — trimming the over-prediction of the European and Indian sites without reordering it, while pointwise nRMSE_{cap} barely moves (medians −0.4, −0.1, and −0.2 pp respectively). An earlier draft of this ablation reported the opposite — a uniform +9 to +16 pp shift away from the measurements — which the 2026-07 implementation audit traced to defects in the Engerer2 port itself (an inverted Δk_{tc} sign, apparent solar time fed as hours-from-noon, and a missing $\beta_5 K_{de}$ term); that reading was anomalous against Engerer2’s literature standing (best-overall of 140 models in the Gueymard–Ruiz–Arias global separation review), the anomaly disappeared with the fix, and the corrected run reported here — which also gained the Bright & Engerer 2019 hourly re-parameterisation [45] for hourly rows alongside the 1-minute southeastern-Australia set for sub-hourly ones — behaves in line with that standing (the pre-audit runs are retained alongside, disclosed rather than replaced). Two caveats still temper the reading. First, both coefficient sets are Engerer’s published fits, not a European re-fit, and the two European (Cfb/Dfb) climates lie outside their calibration domain; a climate-local separation model might behave differently. Second, the clear-sky reference is site-armed rather than default — DEM-derived altitudes (12–558 m, Open-Meteo/Copernicus GLO-90) and climatological annual-mean Linke turbidities with per-field provenance — so the result is conditional on those inputs. Erbs is retained as the shipped default: the gain is second-order on this fleet, but the corrected result removes this draft’s earlier objection to it — notably, recent PVsyst releases default to an Engerer2 separation (v8.1 additionally re-fits the coefficients worldwide), a choice this GHI-only subset now mildly supports rather than contradicts.

DC model. With datasheet parameter blocks populated for 27 of the 30 sites with known module models — manufacturer datasheets and the CEC module database, per-field source provenance recorded in the machine-readable configurations; the remaining three (the Sungrid SG-280 pair M8_B/M8_C, and the eco-Kinetics module of the reclassified M11_3) are defunct-supplier models with no recoverable datasheet and stay on linear- γ — the linear- γ → De Soto swap ran with zero silent downgrades: every armed site that reaches scoring moved (26 of the 27 armed configurations; the 27th, M2_B, is the auto-quarantined spike meter of Table 4) (Figure 7). (An earlier attempt without datasheet blocks reproduced the default results bit-for-bit, confirming the documented downgrade path end-to-end; it remains archived.) For the crystalline bulk the choice is second-order: the 31-site fixed-tilt medians move from −2.2 to

Table 7: Whole-record energy deviation by module technology on the 31 fixed-tilt DKASC systems (chain defaults). A descriptive breakdown of where the fleet’s negative bias sits, not a causal attribution: technology is confounded with install era and the per-class soiling/degradation priors. Medians over each class; Installed is the class’s range of installation years.

Technology	n	Installed	med. nRMSE _{cap} [%]	med. ΔE [%]
poly-Si	18	2008–2016	5.7	−7.1
mono-Si	6	2008–2013	5.9	+2.3
CdTe	2	2008–2010	5.4	+1.9
CIGS	1	2017	8.7	−17.3
a-Si / HIT [§]	2	2008–2010	5.9	−6.6
unlabeled [‡]	2	—	9.1	−17.8

[§]the Kaneka amorphous-Si and Sanyo HIT heterojunction systems. [‡]the two sites whose nameplate is the disclosed peak back-fit (Section 3.2), so their deviation partly reflects that scale. The mono-Si median is robust to M11.3’s disclosed outage residual (Section 5). Desert Knowledge Australia, the Australian Government, the Northern Territory Government and the project managers, Ekistica do not endorse, and accept no legal liability whatsoever arising from, or connected to, the outcomes and conclusions associated with the use of data from the Desert Knowledge Australia Solar Centre.

−1.0 % ΔE with nRMSE_{cap} essentially unchanged (5.8 → 5.8 %), the poly bulk drifts up by +0.1 to +2.8 pp of energy — the deep-negative poly sites gain 1–3 pp toward zero with pointwise error improving — and the newer poly systems trade up to −0.8 pp of pointwise error (M16_A_II) against +1–3 pp of energy. The first-order effect lands where the literature expects it: all five thin-film/HIT systems gain +1.7 to +7.3 pp of low-irradiance energy, improving $|\Delta E|$ on three of the five — M15_A (CdTe) crosses zero from −5.1 to +2.1 %, the HIT system trims −11.8 to −8.7 % — while the Kaneka a-Si overshoots (−1.4 → +5.9 %) alongside the already-overpredicting M6_A (+8.8 → +11.6 %). The energy shift is statistically detectable across the armed set as a whole (signed-rank $p < 10^{-3}$, $n = 26$) but stays second-order outside the thin-film/HIT group: the first-order magnitudes remain technology-specific, not fleet-wide. Linear- γ is retained as the headline default; De Soto reads as a worthwhile refinement specifically for the high- V_{oc} thin-film technologies. That the single-diode gain concentrates on the non-standard technologies rather than the crystalline bulk mirrors PVsyst’s own reassessment of the model: its recent work reports that the highest-fill-factor contemporary cells strain the standard single-diode parameterization [46] and proposes a generalized high-performance module model in response [47] — the same class of non-standard cell technologies our thin-film and HIT results probe empirically. The third arm of the DC-model axis defined in Section 4.3 — the PVsyst-variant single-diode parameterization — has not been run, so the contrast reported here is the two-way linear- γ versus De Soto one; the configuration generator already emits that arm, which makes closing it a run rather than a code change.

Bias by module technology. Breaking the fixed-tilt energy bias down by module technology (Table 7) localizes where it lives: the poly-Si bulk — 18 systems, mostly BP and Q-Cells modules — carries a median −7.1 % deviation, while the mono-Si systems sit at +2.3 % and the two CdTe systems at +1.9 %. The reading is descriptive rather than causal: technology is confounded with install era and with the uniform per-class soiling and degradation priors, so it does not say that any technology is modeled worse. The longitudinal degradation check below (Section 5.5) bears on the degradation half of that confound.

Temporal resolution. On the fleet’s single hourly dataset (OPSD), synthesizing energy-conserving 5-minute records along the clear-sky envelope (Section 4.3) moves whole-record energy deviation from +21.8 % to +21.3 % (−0.5 pp) and leaves nRMSE_{cap} essentially unchanged

Table 8: Median capacity-normalized RMSE by stratum as the scored series are aggregated from native resolution to monthly means (energy-conserving).

Stratum (med. $\text{nRMSE}_{\text{cap}}$ [%])	native	15 min	hourly	daily	monthly
DKASC fixed-tilt	5.8	5.7	5.4	2.9	2.6
DKASC tracked	20.1	20.2	19.9	12.7	12.0
StoreNet	19.4	17.7	16.2	8.3	2.8
OPSD	15.3	15.3	15.3	6.5	4.7
Trivedi	10.1	10.1	9.7	3.7	1.5

Per-window MBE is in `pr_aggregation.json`; on this fleet it is sign-stable across windows (no site flips sign native→monthly).

(15.3 vs. 15.4 %; both series reshaped, metrics at 5 min) — down from the -3.8 pp the pre-audit engine showed here, most of which the audit’s timestamp-geometry fixes absorbed: the hourly path now evaluates solar position and transposition at the interval centre (and the end-labeled OPSD register deltas and Open-Meteo radiation means are re-keyed to a common start-of-interval basis), geometry the 5-minute synthesis was already sampling, so the residual gap is the genuine sub-hourly effect rather than a half-interval geometry handicap. One attribution caveat: this site has no documented inverter AC rating, so no clipping engages (the predicted peak in fact rises from 4.74 to 4.81 kW as intra-hour peaks reappear); the -0.5 pp therefore measures the non-clipping resolution effects — evaluating the inverter’s convex low-load efficiency region and the low-sun transposition geometry per 5-minute step rather than at hourly means — and is a lower bound on the resolution error for clipping-limited systems ($1.3\text{--}2\times$ DC/AC oversizing, Section 2), of which this stratum has none. The same intra-hour averaging error — the mismatch between an hourly-mean irradiance and the convex inverter-efficiency and clipping response it drives — has motivated a sustained line of work in PVsyst, which keeps the hourly simulation and adds a statistical sub-hourly clipping correction fitted to the lost energy [48–50]; our energy-conserving synthesis instead reconstructs the sub-hourly series directly, so the effect is resolved within the same physics chain rather than by a separate correction term.

5.3 Error versus temporal aggregation

The temporal-resolution result above reshapes coarse records finer; the complementary question is how the native-resolution error behaves as it is aggregated coarser. Section 5 argues that at 1-minute resolution the StoreNet pointwise error is dominated by timing noise (a shared sensor cannot resolve per-roof cloud transients) rather than systematic bias, so the energy integral is the meaningful quantity. Table 8 makes that quantitative: recomputing $\text{nRMSE}_{\text{cap}}$ after resampling each site’s measured and predicted power (energy-conservingly) to 15-minute, hourly, daily, and monthly means. If the error were bias, it would survive aggregation; if it were timing noise, it would average away. Table 8 and Figure 9 confirm the split. StoreNet’s median $\text{nRMSE}_{\text{cap}}$ collapses from 19.4 % at native 1-minute resolution to 2.8 % at monthly means — equal to the desert fleet’s 2.6 % — with the knee between hourly and daily, exactly where per-roof cloud-transient timing averages out: for monthly-energy purposes the engine is as accurate on Irish rooftops with one shared sensor as on desert research systems. What survives aggregation is bias: the tracked meters floor at 12.0 % (their availability-driven over-prediction), and per-site MBE is sign-stable across windows (0 sign flips native→monthly, magnitude within ± 40 % on 45/46 of sites with $|\text{MBE}| > 5$ W — consistency with the noise reading rather than proof, since energy-conserving resampling nearly preserves the mean by construction). One disclosure: the StoreNet monthly statistic rests on twelve monthly points per house (a one-year record).

Table 9: Irradiance-stage accuracy against the DKASC 20° reference sensors (daylight rows, elevation > 5°; full station record). rMBE and rRMSE are normalized to the mean measured irradiance of each channel, so the decomposition row — normalized to the small diffuse mean — reads large in relative terms. Generated from the archived check by `make_irradiance_table.py`.

Comparison	n [10 ⁶]	rMBE [%]	rRMSE [%]	R^2
Decomposition: Erbs DHI vs. measured DHI	0.81	+20.9	57.7	0.593
Transposition (GHI-driven) vs. tilted	0.51	+0.8	15.0	0.933
Transposition (measured DHI) vs. tilted	0.51	+1.0	14.9	0.935

5.4 Irradiance-stage validation

The headline metrics score AC power, which convolves the irradiance chain (decomposition and transposition) with everything downstream (module, inverter, losses). DKASC lets us separate the two: beside GHI, its weather station logs a measured diffuse-horizontal channel and a measured global-tilted channel at a fixed 20° equator-facing reference plane, so the chain’s *intermediate* irradiance products can be scored directly against co-located reference sensors, independent of any PV system. We report three comparisons on the daylight rows (solar elevation > 5°) of the station record: (i) *decomposition* — Erbs-predicted DHI against the measured diffuse channel; (ii) *transposition*, *GHI-driven* — the full prospective chain (Erbs then Hay–Davies) predicting global-tilted irradiance against the tilted sensor; and (iii) *transposition alone* — Hay–Davies fed the *measured* diffuse split, which removes decomposition error and isolates the transposition model. The decomposition and transposition mathematics are exactly those of the production chain (Section 2); the reference implementation used here is cross-checked against pvlib-python to < 1 W m^{−2}, and the harness-native version that scores the shipped C++ engine on the same sensors is prepared (Section 7).

Table 9 reports the result over the full station record, and it splits exactly along the expected line [11], with one instructive twist. Decomposition is the weak stage in isolation: Erbs-predicted DHI runs +20.9 % above the measured diffuse channel (rRMSE 57.7 %, R^2 0.59) — the known clear-sky behaviour of hourly separation models applied to a high- k_t desert record. Transposition is strong: fed the measured diffuse split, Hay–Davies tracks the tilted sensor at +1.0 % rMBE (rRMSE 14.9 %). The twist is that the full prospective path (Erbs then Hay–Davies) is *just as good* on global tilted irradiance — +0.8 % rMBE, rRMSE 15.0 %, R^2 0.93 — because on a 20° plane the decomposition error largely cancels in the beam-plus-diffuse recombination: over-predicted diffuse is compensated by the correspondingly under-predicted beam. Two readings follow. First, the irradiance chain delivers plane-of-array energy within about one percent on this station, so the fixed-tilt fleet’s −2.2 % energy bias is set *downstream* of the irradiance stage — consistent with the year-zero loss-stack attribution of Section 5.5. Second, the large decomposition-stage error is latent rather than harmless: it cancels only where beam and diffuse recombine on a near-horizontal plane, and it is exactly the term that surfaces at steeper tilts, in beam-sensitive geometries, and on the GHI-only European sites where the decomposition ablation of Section 5.2 is first-order.

5.5 Residual structure, degradation trend, and filter sensitivity

Three further analyses read the per-site row-level exports the harness writes for every scored system, localizing the residual and testing its robustness. All are generated by `doc/paper/scripts/make_residual_analyses.py`; the exports are regenerated by any full run (they are not committed).

Residual structure. Binning the per-row normalized bias (predicted minus measured, as a percentage of P_{STC}) by clearness index k_t and by solar elevation separates candidate error

Table 10: Median per-row bias [% of P_{STC}] by clearness-index bin and stratum (scored daylight rows, chain defaults).

Stratum (med. bias [% P_{STC}])	k_t 0.0–0.2	k_t 0.2–0.4	k_t 0.4–0.6	k_t 0.6–0.8	k_t 0.8–1.0
DKASC fixed-tilt	+0.0	+0.7	+1.3	-1.2	-3.2
DKASC tracked	-0.3	+1.4	+14.5	+24.8	+22.7
StoreNet	-1.7	+1.1	+10.3	+23.6	+17.8
OPSD	-2.2	+2.4	+4.3	+4.3	+40.2
Trivedi	+0.0	+2.4	+7.6	+13.0	+10.7

Italic = fewer than 500 scored rows in the bin.

Table 11: Median per-row bias [% of P_{STC}] by solar-elevation bin and stratum (scored daylight rows, chain defaults).

Stratum (med. bias [% P_{STC}])	0–15°	15–30°	30–45°	45–90°
DKASC fixed-tilt	+0.9	+0.6	-1.7	-3.1
DKASC tracked	+4.9	+27.6	+21.4	+13.2
StoreNet	+0.1	+2.1	+4.5	+3.6
OPSD	+1.5	+6.1	+2.0	-2.2
Trivedi	+0.4	+3.0	+5.6	+4.8

sources: a monotone k_t trend implicates the decomposition/transposition path, an elevation trend the low-sun geometry, and a flat profile points downstream to the thermal or electrical model. Tables 10 and 11 report both binnings. The k_t profile separates the strata: DKASC fixed-tilt is nearly flat and mildly negative at high clearness (-0.9% of P_{STC} at k_t 0.6–0.8) — a multiplicative deficit at high power, consistent with the year-zero loss-stack conservatism established below, not a decomposition signature — while every GHI-only stratum rises steeply with k_t , peaking at $+13$ – 24% in the k_t 0.6–0.8 bin (OPSD sits lower there, its excess concentrated in a thin $k_t > 0.8$ tail), and the tracked meters mirror that shape for the availability reason already established. The elevation binning adds a finer desert signature: over-prediction at low sun ($+0.9\%$ at 0–15°, the transposition/IAM credit region) against under-prediction at high sun (-3.1% at 45–90°), while the European strata instead peak at mid or high elevation — the shape of a multiplicative loss or geometry deficit rather than of low-sun horizon blocking (with the caveat that a bias expressed in % of P_{STC} scales mechanically with power level, so the two binnings must be read jointly). A modeled- T_{cell} -binned residual and a month×hour heatmap remain future analysis work; the `t_cell_c` export column they need now ships in the per-site exports, where it already feeds the temperature-corrected PR column of Table 6.

Longitudinal degradation. The 7–17-year DKASC records let us test the degradation sub-model directly, a check most whole-tool validations omit: for each site we fit the slope of annual energy deviation against array age and compare it to the Jordan–Kurtz technology-class rate the configuration asserts (Table 12). A slope near zero means the prior tracks the array’s real fade; a positive slope means the measured array degrades faster than modeled, so predictions drift high across the record. Only clean-epoch years (offline-filtered) enter the fit.

The fixed-tilt result is two-sided. The median trend is $+0.75 \text{ pp yr}^{-1}$ (30 of 31 sites positive; $+0.61$ after excluding the fault-tail sites below) against booked class rates of 0.4 – $0.6 \text{ \% yr}^{-1}$: at face value the measured fleet fades two to three times as fast as the temperate-median Jordan–Kurtz priors book, consistent with the hot-climate end of the published rate distributions [32, 51].

That reading now carries a quantified instrument-side bound (`make_sensor_drift.py`, committed as `sensor_drift.json`). Every DKASC prediction rides one shared station GHI channel, so a secular trend in that channel aliases one-for-one into the fitted slope, and no witness inside

the harness could catch it. The audit tracks, per record-year, the 98th-percentile envelope of the daily clearness index (daily Σ GHI over ΣI_{0h}) — the clearest days’ atmosphere is climatologically far more stable than a drifting thermopile — and fits a Theil–Sen trend, robust to a single recalibration step. The station’s clear-day envelope *rises* by a median $+0.54\% \text{ yr}^{-1}$ across the 34 DKASC site-records (30 beyond the ± 0.3 flag; per-record slopes -0.02 to $+0.60$, the spread of differing record spans over the one shared channel; the one-year StoreNet record supports no trend): the same sign as, and roughly two-thirds the magnitude of, the $+0.75 \text{ pp yr}^{-1}$ deviation trend. If the envelope rise is instrument-side — rising sensitivity, a cleaning-regime change, or an over-corrected recalibration history — the genuinely under-booked fade shrinks to $\approx +0.2 \text{ pp yr}^{-1}$ and the two-to-three-times reading above overstates; if it is a real clear-sky brightening trend, predictions track a true input and the fade reading stands. A reanalysis cross-trend separates the two (`make_drift_witness.py`, committed as `drift_witness.json`): the identical envelope statistic run on Open-Meteo archive reanalysis GHI at the station coordinates is *flat* over the same 17 years ($-0.03\% \text{ yr}^{-1}$), while the station-to-reanalysis envelope ratio trends $+0.56\% \text{ yr}^{-1}$ — essentially the full station slope of $+0.57\% \text{ yr}^{-1}$. The independent sky view sees a stationary atmosphere, which places the station envelope’s rise on the instrument side to the witness’s resolution (reanalysis is a model, and assimilation-stream inhomogeneities are its noise floor; trend agreement is evidence, not proof). Read together, the audit and its witness move the default reading of the $+0.75$ slope: most of it is channel drift, and the genuinely under-booked fade is nearer $+0.2$ than the face-value figure. Per the detection-only protocol the bound is disclosed, not applied — no series is corrected — and the *level* conclusions (the first-two-year -4.6% and the year-zero loss-stack attribution) are slope-free and unaffected.

Read jointly with the whole-record bias, this splits the “conservative priors” attribution into a level term and a slope term: the median deviation over each site’s *first* two years is -4.6% (the year-zero loss stack over-derates), it then narrows along the positive trend, and the median over each site’s *final* two years is $+3.3\%$ ($n = 31$ sites with ≥ 6 scored years) — the two errors cancel near mid-record, which flatters the whole-record -2.2% , and a 25-year prospective run with these priors would under-book lifetime fade. Re-centering either term against this fleet is the calibration the protocol forbids; per-technology hot-climate rates are future work. The fit hygiene visible in the table — a year-coverage gate (≥ 300 measured days per fit year) and masking of the documented availability epochs carried in the site configurations — earned its keep against the pathologies an earlier draft of this fit had flagged: M16_C and M4_B, whose commissioning-era head years poisoned their raw fits (M16_C’s 2011 stub year scores $+758\%$), fit cleanly once the coverage gate drops those years, and M9_B fits once its documented failure epochs are masked (the two qualifying tracked meters carry a $+2.96 \text{ pp yr}^{-1}$ median — faster than the fixed fleet, as mechanically expected of trackers). Exactly one guard rejection survives the hygiene: M9_AC at $+14.8 \text{ pp yr}^{-1}$ *after* its declared 2019–2021 epoch is masked — consistent with the persistence detector’s independent finding (Section 5.5) that this meter’s collapse began around September 2018, months before the portal-documented window, i.e. an under-declared data pathology rather than aging.

Two accepted fixed-tilt slopes are large enough to read as progressive faults rather than fade: M8_C walks monotonically from $+4$ to $+49\%$ annual deviation over 2019–2025 and M18_C from -3 to $+34\%$ over 2017–2025 — accelerating arcs no linear degradation produces. A portal cross-check (July 2026) finds no documented outage *window* for either, so both stay unmasked; but for M18_C the portal’s maintenance feed does record a hot-spot inspection finding (two modules, May 2017) immediately preceding its acceleration — consistent with progressive cell damage on a micro-inverter array — while M8_C, a defunct-supplier array (Section 5.2), remains wholly undocumented. The capacity anchor of Section 6 corroborates both reads from the meter side: across the fleet exactly these two sites trip its capacity-decline flag, their PVUSA-style realized capacity falling -4.0 (M8_C) and $-3.6\% \text{ yr}^{-1}$ (M18_C). The persistence detector of Section 3.3, run on the full data, draws the boundary between the two tools exactly where these meters sit:

Table 12: Median annual energy-deviation trend [pp yr⁻¹] against the asserted degradation prior, by stratum — per-year fits with documented availability epochs masked and partial years coverage-gated (footnotes in the table).

Stratum	med. ΔE trend [pp yr ⁻¹]	n ($n_{\uparrow}$)	med. prior [% yr ⁻¹]
DKASC fixed-tilt	+0.75	31 (30)	0.60
DKASC tracked	+2.96	2 (2)	0.40
StoreNet	—	—	0.00
OPSD	—	—	0.00
Trivedi	—	—	0.00

Trend needs ≥ 5 fit-eligible yr (each covering ≥ 300 d, outside documented availability epochs) and $|\text{trend}| \leq 5$ pp yr⁻¹; ‘—’ = none qualify. $n_{\uparrow}$ = sites with a positive trend.
Guard-rejected: DKA/M9_AC +14.8.

neither accelerating arc trips it — a drift that passes through months at daily ratios of 0.7–0.9 never sustains 14 days below 0.6, so progressive faults stay the longitudinal fit’s catch — yet on the same two meters it surfaces what annual fits cannot resolve. M18_C’s record opens with an *undeclared dead head*: daily ratio ≈ 0 from February 2010 to August 2013, worth +29.6 pp of whole-record deviation on its own — which demystifies this site’s outsized 11.1 % offline-drop fraction in Table 17 (the filter was silently absorbing a dead 3.5-year head, not scattered noise). M8_C shows only a 123-day dead commissioning head and two short 2025 dips (+2.9 pp combined), confirming its post-2019 walk is continuous drift, not hidden outage mass.

The detector’s control behavior across the fleet supports trusting those two reads. On M11_3 — the reclassified aggregate whose three failure windows are portal-documented — all three detected epochs overlap declared windows (including the 2016–2021 two-inverters-dead epoch at daily ratio 0.24) and nothing undeclared surfaces. On M9_AC it *re-dates* the portal story: the collapse onset lands at 22 September 2018, three months before the declared 2019–2021 window opens, preceded through 2018 by two sub-0.6-ratio dips the maintenance feed never recorded — prodromal tracker trouble, and the reason its degradation fit stays guard-rejected even after masking. On the Trivedi houses it segments the known zero-filled outage mass into 6–7 undeclared epochs per house worth +21 to +69 pp each — the mass the offline filter was built to absorb — and on four StoreNet houses it finds one or two short winter epochs apiece (+6 to +17 pp), consistent with the Cfb winter pathologies discussed below. At the other extreme, the CIGS system M15_C.II carries a ≈ -12 % *year-one* level offset (2016: -12.4 %) while its heavy 1.44 % yr⁻¹ thin-film class prior *over-books* the array’s real fade by ≈ 0.8 pp yr⁻¹ at face value (fitted trend -0.83 ; ≈ 1.4 under the channel-drift adjustment of the bound above, so the over-booking conclusion only strengthens): its whole-record -17.3 % is an initial capacity/quality offset compounded by an over-heavy aging prior, not under-modeled degradation.

Offline-filter sensitivity. The system-offline filter (Section 3.3) is load-bearing on the zero-inflated meters, so we verify the headline is not an artifact of its 10 %-of-predicted, ≥ 1 W thresholds by sweeping the gate over 5/10/20 % and the floor over 1/10 W and recomputing the stratum energy deviations (Table 13, with each stratum’s headline configuration in bold: the harness filters only the zero-inflated datasets, so the StoreNet and OPSD headline column is *none* and their filtered cells are diagnostics, not sensitivities of the headline). On the strata the filter is built for, the dependence is bounded and monotone — DKASC fixed-tilt moves from -1.8 to -3.0 % across the full gate range, Trivedi from $+16.5$ to $+9.6$ % — so the 10 %/1 W default is an operating point, not a tuned one. Applied where it does *not* belong, the same gate is heavily load-bearing: StoreNet would swing from $+7.3$ to -9.2 % across gates, because at 1-minute resolution a measured-versus-predicted gate eats legitimate per-roof cloud transients — the quantitative case for restricting the filter to zero-inflated meters. The OPSD column of

Table 13: Median stratum energy deviation [%] with no offline filter (‘none’) and as the gate/floor sweep. Bold marks each stratum’s headline configuration — the harness applies the filter only to the zero-inflated DKASC and Trivedi meters (Section 3.3). All cells, including ‘none’, recompute on daylight rows of the per-site exports. The final ‘GHI-gate’ column is a *prediction-independent* variant (the record’s GHI channel $\geq 150 \text{ W m}^{-2}$ while the meter reads under 2 % of nameplate — no reference to the model’s own prediction; on the reanalysis-driven strata that channel is the model’s input rather than a measurement). On the stratum the filter is load-bearing for, it corroborates the model-relative gate without presupposing the model (DKASC fixed-tilt -1.6 vs. the headline -2.2); on Trivedi it recovers most of the outage mass ($+63.6 \rightarrow +25.5$, vs. $+14.1$ model-relative — zero-filled epochs whose daylight rows sit under the irradiance threshold escape it), and it barely moves the tracked meters, whose sub-prediction *partial* production only a model reference can flag.

Stratum (med. ΔE [%])	none	5%/1W	5%/10W	10%/1W	10%/10W	20%/1W	20%/10W	GHI-gate
DKASC fixed-tilt	+5.6	-1.8	-1.8	-2.2	-2.2	-3.0	-3.0	-1.6
DKASC tracked	+55.8	+53.0	+53.0	+51.9	+51.9	+47.9	+47.9	+55.6
StoreNet	+17.5	+7.3	+7.3	+1.2	+1.2	-9.2	-9.2	+11.0
OPSD	+22.2	+6.3	+6.3	+3.0	+3.0	-1.9	-1.9	+11.4
Trivedi	+63.6	+16.5	+16.5	+14.1	+14.1	+9.6	+9.6	+25.5

Bold = the stratum’s headline configuration (the harness filters only DKASC and Trivedi; ‘none’ = no offline filter).

the same sweep is what Section 5 repurposes diagnostically (the site’s whole over-prediction sits below the gate). One convention to note: every cell, *none* included, recomputes on daylight rows of the per-site exports, so the fixed-tilt 10 %/1 W cell reads -2.2 , coinciding at this precision with the all-hours headline. The *none* column is the filter’s own justification, read jointly with the epoch accounting above: unfiltered, the fixed-tilt median swings from -2.2 to $+5.6$ % — a $+7.8$ pp shift carried almost entirely by the dead heads and failure windows the persistence detector itemizes per site — and Trivedi swings from $+14.1$ to $+63.6$ %, the per-house $+21$ to $+69$ pp epoch masses landing in the median. The filter is thus not shaving noise at the margin; it is excluding discrete, now-individually-attributed outage mass, and scoring *through* it (as the external tools of Section 5.8 do) is the single largest distortion available in this benchmark.

5.6 AC-electrical checks

The electrical companion checks (Section 4.4) ran with the sweep; their summary is archived alongside it. Two of the three legs turn out to be unexercisable on this fleet: the per-phase DKASC meter files behind the scored sites carry no line-voltage, apparent-power, or power-factor channels (those exist only on the site-level master-meter rollups, which Section 3 excludes from scoring as mixed-orientation aggregates), so the voltage-drop plausibility bound and the power-factor check report no measured anchor on any site and remain open pending data with per-meter electrical channels. The realized-efficiency leg did run everywhere. At the 50 %-load bin the chain’s curve point is 0.967, and the measured-AC-to-predicted-DC ratio bundles the inverter with every real loss downstream of the predicted DC point, plus the site’s prediction bias at that operating slice; since the chain books soiling optically at the plane-of-array stage and degradation on DC (Section 2.8), the predicted-DC reference already carries both derates and a healthy site should sit near the bare curve point. It is a consistency check, not an isolated inverter measurement. Read that way it behaves as expected (Figure 8): on the eleven fixed-tilt systems with whole-record $|\Delta E| \leq 5$ % the ratio clusters at 0.95–1.02 (median 0.98), bracketing the curve point as the derate-inclusive reference predicts, while the tracked aggregate meters collapse far below it (0.60–0.62 on the M9 meters; the reclassified M11.3 rose from 0.30 to 0.61 under its corrected fixed-array prediction). The corrected Trivedi households now sit at

0.58–0.80 — id01/id02 (0.80/0.74) at the residential band’s edge, id03 (0.58) low in proportion to its geometry residual; in the initial inflated-nameplate run the same ratio read 0.24–0.46, the denominator error made visible — reproducing the outlier structure of Section 5 from an independent electrical angle rather than revealing inverter-stage error.

5.7 Weather-input error separation

For the datasets with on-site weather (DKASC, StoreNet), a dedicated run substitutes hourly Open-Meteo/ERA5-family reanalysis for the station record: GHI, ambient temperature, and precipitable water are replaced, the measured plane-of-array and diffuse channels are dropped (the chain runs GHI→POA on every row, as any prospective study must), and station wind and rainfall are deliberately retained so tracker wind-stow and the soiling accumulator behave identically to the headline run. The measured delta is therefore the full instrumented-to-prospective input gap — irradiance source and irradiance path together — which is exactly what a yield study without site instrumentation faces. The two datasets already driven by reanalysis (OPSD, Trivedi) reproduce their headline results bit-for-bit, an end-to-end control of the substitution plumbing. (A first attempt at this run was itself a silent no-op — the DKASC and StoreNet loaders ignored the substitution flag — and remains archived alongside the valid run; the loader gap is fixed in-tree and the run reported here, the canonical 20260730-031016 stamp of Section 5, postdates the fix, the two metadata corrections, and the yield-physics audit.)

On the 31 fixed-tilt desert systems the weather input costs +4.1 pp of pointwise error: median $\text{nRMSE}_{\text{cap}}$ rises from 5.8 % to 9.9 % (per-site shift -0.4 to $+5.6$ pp; Figure 10), median R^2 falls from 0.959 to 0.876, and median energy deviation deepens from -2.2 % to -8.9 % — a difference of medians of -6.7 pp, while the median *per-site* shift, the statistic quoted in the abstract, is -4.5 pp (signed-rank $p < 10^{-5}$) — and $|\Delta E|$ grows on 24 of 31 systems: the substituted inputs deliver less plane-of-array energy than the station path, stacking on loss priors that were already conservative. StoreNet separates along the other axis: pointwise error does not worsen (median $\text{nRMSE}_{\text{cap}}$ eases from 19.4 % to 17.8 %; at 1-minute resolution it is bound by per-roof representativeness, not by the GHI source), while energy over-prediction grows ($+15.2 \rightarrow +22.5$ %, every house shifting $+5.8$ to $+8.1$ pp). The tracked aggregate meters appear to improve ($+51.9 \rightarrow +34.2$ % median ΔE), but that is lower reanalysis irradiance coincidentally offsetting outage-driven over-prediction, not a better model. Read against Section 5.8, the split carries a practical lesson: on the 16 clean-record desert systems the same reanalysis family lands at -5.7 % median ΔE under our explicit literature priors but $+0.3$ % under PVGIS’s flat 14 % screening loss — prospective inputs shift the calibration burden onto the loss model, and priors that read conservative under station weather read doubly so under reanalysis.

The satellite corner of this input-attribution triangle has now run for the stratum whose loader consumes an hourly weather file natively: the Trivedi family re-ran with hourly PVGIS-SARAH3 satellite GHI through the same chain (dumped by `pvgis.baseline.py --dump-weather` in the loader-contract weather-file shape; the `site_configs.weathersat` family pins the reanalysis fallback off, and the archived stamp is pinned in `canonical_runs.json`). A pure source swap with the model held fixed moves every house the same way: energy deviation shifts down by -6.0 to -4.0 pp (median -4.6), taking the stratum median from $+13.9$ to $+9.3$ %, while pointwise $\text{nRMSE}_{\text{cap}}$ barely moves (median 10.1 vs. 10.1 %). Roughly a third of the Trivedi over-prediction is therefore carried by the reanalysis irradiance itself, shown here as a source swap inside our own chain, while the id03 geometry residual survives the swap essentially intact ($+36.1$ %), staying attributed to metadata rather than input. The same swap has also run on the StoreNet and OPSD records, through a loader-level substitution override (`weatherSubstituteFile`: GHI and ambient temperature replaced after the native load, measured plane-of-array channels cleared, wind and precipitation kept native; stamps 20260730-033923 and 20260730-034123), and the two answer differently. On StoreNet, where the swap replaces the shared on-site sensor, the over-prediction stays: every house moves up

Table 14: Baseline comparison: median whole-record energy deviation ΔE [%] by stratum and tool, with median $|\Delta E|$ in parentheses. Each tool is scored on its own inner join of the same measured series — pvSpline and pvlib at native resolution (pvSpline with the offline filter of Section 3.3, pvlib without), PVWatts and PVGIS on hourly aggregates without the filter. The DKASC fixed-tilt stratum is split at its median per-site offline-drop fraction ($\approx 4\%$ of the pre-filter join, Appendix B). The pvlib and PVWatts baselines cover fixed-tilt DKASC and StoreNet only (native weather loaders); pvlib’s StoreNet numbers are the re-run after the disclosed units-defect fix (see text); PVWatts consumes the same Erbs-split irradiance with its single 14% system-loss bucket.

Stratum	n	pvSpline	pvlib	PVWatts	PVGIS
DKASC fixed-tilt, clean record	16	−1.8 (4.5)	+5.2 (6.8)	−2.3 (5.6)	+0.3 (5.4)
DKASC fixed-tilt, outage-affected	15	−7.1 (9.6)	+27.8 (27.8)	+19.1 (19.1)	+25.7 (25.7)
DKASC fixed-tilt, all	31	−2.2 (8.1)	+9.0 (9.3)	+3.7 (8.1)	+5.6 (10.1)
DKASC tracked ^a	3	+51.9 (51.9)	—	—	+48.6 (48.6)
StoreNet	10	+15.2 (15.2)	+10.7 (13.0)	−4.7 (6.3)	+0.3 (7.5)
OPSD	1	+21.8 (21.8)	—	—	−1.5 (1.5)
Trivedi	4	+13.9 (13.9)	—	—	+33.5 (33.5)
Fleet	49	+1.0 (9.6)	—	—	+5.6 (10.9)

^arequested from PVGIS as two-axis-tracked; out of scope for the pvlib and PVWatts baselines.

by 0.7 to 1.0 pp, taking the median from +15.2 to +16.0%, while pointwise $\text{nRMSE}_{\text{cap}}$ falls on every house (median 19.4 to 17.8%). Driven by independent satellite irradiance, the chain over-predicts just as it does on the station record, so the StoreNet over-prediction is not the shared sensor reading high; it stays with the loss priors and per-house metadata. On OPSD, where the swap replaces reanalysis irradiance, the energy deviation falls from +21.8 to +15.9%, the same direction as on Trivedi. DKASC has no distinct satellite leg to run — Alice Springs lies outside the SARA3/Meteosat field of view, and PVGIS’s ERA5 fallback there reproduces the reanalysis posture already covered by the substitution leg above.

5.8 Comparison with pvlib-python, PVWatts, and PVGIS

Table 14 compares whole-record energy deviation — the cross-resolution metric of Section 4.5 — across tools by stratum, and Figures 11 and 12 show the per-site picture behind those medians; Appendix B lists every site. Two protocol reminders frame the numbers. First, each tool is scored on its own inner join: the external baselines apply no system-offline filter, so on systems with documented outage windows (Section 3.3) they score intervals that the pvSpline harness excludes and discloses. Second, the PVGIS leg is end-to-end — its own irradiance database, its default 14% system loss — so its agreement measures the whole screening stack, not the model chain. The second same-input reference — NREL’s PVWatts model (Section 4.5) — covers the same scope as the pvlib leg (31 fixed-tilt DKASC systems and 10 StoreNet houses); its column appears in Table 14 and its reading follows the pvlib and PVGIS paragraphs below.

pvlib-python (same input). On the 31-site fixed-tilt DKASC subset the pvlib baseline predicts more energy than pvSpline on 30 of the 31 systems (the one exception, M2_A, sits −1.2 pp below), with a median paired ΔE difference of +9.0 percentage points (pp). Splitting the subset at its median offline-drop fraction is diagnostic (16 clean-record / 15 outage-affected, the split of Table 14). On the 16 clean-record systems the gap is a tight, near-uniform +7.7 pp (range +4.5 to +9.8) — the direction and rough magnitude of the terms the baseline deliberately omits. On this fleet those terms reduce to the BWh-prior soiling accumulator: the chain’s ohmic and mismatch derates are configured to *zero* on every validated site (Section 3.2), so the gap decomposes as soiling plus the GHI-driven irradiance path. The accounting closes arithmetically: the

harness emits the energy-weighted mean *applied* derates per site (`soiling_mean_applied_pct` in the summary JSON), and on the 16 clean-record systems the median gap of 7.7 pp reconciles as 7.4 pp of applied soiling plus a 0.3 pp path-and-residual remainder — consistent with the irradiance-stage check of Section 5.4, which independently prices the GHI-driven path at about +0.8 % of plane-of-array energy on the rows where pvSpline uses the tilted sensor (per-site reconciliations are in `attribution.json`; M15-A closes to within 0.5 pp). The applied-soiling level is also of the same order as the −4.6 % year-zero under-prediction of Section 5.5: the same loss stack read two independent ways. Pointwise accuracy is essentially shared (median paired $\text{nRMSE}_{\text{cap}}$ difference +0.56 pp). On the 15 outage-affected systems the gap explodes (+ −1.2 to +122.9 pp, median +30.4) and rises with the harness’s per-site offline-drop fraction (Spearman rank correlation 0.74, 95 % CI 0.41–0.92; Figure 12): pvlib is scoring the meter-outage windows that the harness excludes and discloses. The decisive corroboration is that pvlib’s per-site energy deviations agree with PVGIS’s — computed from completely independent weather — at $r = 0.97$ (95 % CI 0.96–1.00) across the 31 sites: where the two baselines depart from pvSpline, they depart *together* and toward the same values, the signature of measurement defects rather than implementation divergence. On StoreNet, where both tools drive transposition from the same shared GHI record (no plane-of-array channel exists, Table 3), the two implementations track each other house by house (Figure 11, StoreNet row): pvlib lands 3.6–5.0 pp below pvSpline on every one of the ten houses (median paired ΔE difference −4.5 pp; paired $\text{nRMSE}_{\text{cap}}$ within 0.8 pp), reproducing house by house — including the +27 % extremes on H4 and H5 — the over-prediction pattern reported for pvSpline in Section 5. The Cfb over-prediction is therefore not an implementation artifact; the PVGIS leg below narrows it further. One audit note: the first archived StoreNet run of this baseline was invalid — it read the shared weather station’s `soltot` column as W m^{-2} when it is J cm^{-2} per 1-minute window (Met Éireann convention), feeding the chain irradiance $\approx 167\times$ too small and producing −100 % energy deviation on all ten houses. The loader defect was disclosed and fixed, the leg re-run, and the superseded summary retained in the repository history.

PVWatts (same input, screening loss). The PVWatts leg completes a 2×2 the other tools only sample: same-input/screening-loss, where pvlib is same-input/no-derates and PVGIS is independent-input/screening-loss. Its column reads accordingly. Against pvlib, whose irradiance it shares, it lands lower in every stratum (clean −2.3 vs. +5.2; outage +19.1 vs. +27.8; StoreNet −4.7 vs. +10.7) — the direction the flat 14 % bucket dictates, with the stratum-to-stratum spacing modulated by PVWatts’s simpler hourly system model (NOCT thermal, no spectral term, its own inverter curve). pvSpline in turn sits below *both* on the desert strata: the explicit prior stack books more year-zero loss than even a 14 % screening bucket — the conservative-priors finding of the longitudinal analysis (Section 5.5) in a third, cross-tool form. On the clean half a generic derate is competitive with the literature-prior chain for whole-record energy ($|\Delta E|$ median 5.6 vs. our 4.5) — the lesson of the PVGIS clean column, now with input differences removed. On the outage half it scores the excluded windows like every unfiltered tool (+19.1), and on StoreNet its flat loss overshoots into compensation (−4.7 against our +15.2), reproducing from the *same* shared sensor the pattern PVGIS shows from independent satellite irradiance — a third tool corroborating that the Cfb over-prediction is a level effect any sufficiently large uniform derate can mask, i.e. input-side priors, not chain physics. Across its 41-site scope PVWatts edges the harness on median $|\Delta E|$ (7.99 vs. 8.81 %) while being the closer tool on only 17 of 41 sites — flat-loss screening wins medians on mixed-quality fleets by splitting the difference between clean and faulty records, while the explicit chain is more often nearer the truth per site.

PVGIS (end-to-end screening). Fleet-wide, the free screening stack is a stronger reference than its flat-loss framing suggests: median $|\Delta E|$ of 10.9 % against pvSpline’s 9.6 %, closer to

the measurement on 18 of 49 sites. The structure is more informative than the near-tie. On the 16 clean-record desert systems (the ERA5 leg), PVGIS reaches a median $|\Delta E|$ of 5.4 % against pvSpline’s 4.5 %: for whole-record energy on well-behaved fixed-tilt systems, prospective reanalysis weather with a default loss concedes almost nothing to on-site weather with literature priors — a useful calibration of what the extra modeling buys, which on clean sites is sub-hourly fidelity, loss attribution, and the per-site diagnostics this section rests on, rather than annual energy. On the temperate-European rooftops (the SARA3 leg) the comparison lands against us: PVGIS is closer on 7 of 10 StoreNet houses (median $|\Delta E|$ 7.5 % vs. 15.2 %) and on the OPSD system (−1.5 % vs. +21.8 %). Because the SARA3-driven run brackets the StoreNet measurements (per-site predicted-to-measured peak ratios 0.86–1.02), the systematic Cfb over-prediction — which the same-input twin above pins to the input side — reads as uniform climate-class priors too light for these rooftops, and/or per-roof losses (local shading, orientation) that a single shared irradiance input cannot see, rather than a sensor that reads high: our own satellite substitution (Section 5.7) leaves it in place. It is in any case not an artifact of the measured data. (A third candidate, local horizon shading that the chain does not model but PVGIS’s default DEM horizon partially does, is now excluded by the horizon A/B of Section 4.5: turning PVGIS’s terrain horizon off moves its agreement by at most 0.5 pp, so terrain shading is not what separates the tools here.) On the tracked aggregate meters, the screening stack misses *in the same direction* on every system (median +48.6 % vs. +51.9 %; PVGIS closer on 1 of 3) — what one expects when the dominant error source is the measurement record itself rather than any tool’s physics. The Trivedi stratum played the same corroborating role in the nameplate diagnosis of Section 5: at the inflated nameplate both tools missed together (+164.2 % median for PVGIS vs. +93.6 %, PVGIS larger because it also scores the zero-filled outage windows the pvSpline offline filter drops). At the corrected nameplate the re-fetched PVGIS leg lands at +10.4 to +55.4 % (median +33.5 %) against pvSpline’s +6.5 to +42.0 %, PVGIS closer on 0 of 4; the per-house excess of +4 to +36 points is consistent with exactly those unfiltered outage windows, which PVGIS’s inner join retains. (The first post-correction PVGIS run silently reused responses cached at the old nameplate; the cache is now keyed to the full request parameters, and the re-fetch replaced it.)

6 Discussion and threats to validity

Published whole-tool validations on high-quality hourly data report annual-energy deviations of roughly ± 3 –8 % (Freeman et al. [8], four quantitative tools against nine instrumented systems, “within ± 8 %” excluding identified outliers; the ± 3 % lower edge traces to the companion SAM-only NREL study TP-6A20-60204 on fixed-tilt systems). Independent uncertainty budgets place the compounded spread of long-term yield predictions — weather source, model, and interannual variability together — at several percent in its own right [52], so agreement within the single-digit-percent band approaches the noise floor of the exercise. Our closest-comparable stratum — the 31 fixed-tilt DKASC systems with on-site weather — lands at a median ΔE of −2.2 % (IQR −10.7 to +3.5 %) under a strictly no-calibration protocol with uniform literature priors, over records four to sixteen times longer than a single year and at 5-minute rather than hourly resolution. We consider that band consistent with the published state of the art, with two caveats. First, the deviation is predominantly one-sided, and the longitudinal decomposition (Section 5.5) splits its attribution: the year-zero loss stack (soiling, LID, quality) is conservative — first-two-year deviations median −4.6 % — while the degradation priors are the opposite, booking roughly a third to a half of the fade the fleet measures (+0.75 pp yr^{−1} median drift, of which $\approx +0.54$ % yr^{−1} reads as clear-envelope drift of the shared irradiance channel rather than array fade — reanalysis-witnessed, Section 5.5), so records end above zero deviation and the whole-record figure flatters the long horizon. Re-centering either term would likely tighten the band — we deliberately did not, as that is precisely the calibration this protocol forbids.

Second, the fleet tail is dominated by metering, metadata, and weather-input defects rather than chain physics: the tracked aggregate meters remain the surviving tail — a reading corroborated externally, since the unfiltered external baselines, which score the outage windows the harness excludes, miss on exactly those systems, together, and in the same direction (Section 5.8) — while the Trivedi stratum, initially the heaviest tail member, collapsed into the residential band once its nameplate metadata was corrected against the data descriptor (Section 5). This is the practically relevant failure mode for unattended batch studies and the reason the harness emits per-site anchor checks and input-quality warnings instead of a single fleet average.

The model-versus-data reading can be made quantitative from the runs this paper already contains. For every site, the committed transposition sweep and the Engerer2/De Soto ablations span a *defensible model-choice envelope* — the set of whole-record energy deviations reachable by swapping sub-models the literature would accept — and the committed diagnostics (offline-drop fraction, measured PR, declared availability epochs, and the pvlb/PVGIS triangulation of Section 5.8) carry the data-quality signals. Table 15 classifies each site by the strongest committed signal. The headline is the envelope itself: on the 31 fixed-tilt desert systems the median model-choice span is 0.8 pp of energy deviation, while the median distance from the measurements that any defensible choice can reach ($\min |\Delta E|$ over the envelope) is 7.1 pp — an order of magnitude apart, so sub-model selection is essentially irrelevant to the desert residual. The input axis sits between the two: swapping the weather source alone (the reanalysis substitution of Section 5.7) moves the same sites by a median $|\Delta(\Delta E)|$ of 4.5 pp — nine times the model envelope — fixing the hierarchy *data quality* > *input* $\gg$ *sub-model choice* for this fleet. What the evidence does attribute: 8 sites agree within the measurement floor, 11 carry a data-side signature (nameplate metadata, documented epochs, or outage-heavy records where the independent baselines miss together), only 1 are genuinely rescued by a model choice (the thin-film systems that the De Soto ablation moves onto the measurements — 1 sites fleet-wide), and the remaining 11 carry a systematic level residual no defensible model choice explains on a clean record — the year-zero loss-prior stack of Section 5.5. The technology cross-tab localizes that class: 8 of the 11 are poly-Si — matching the -7.1% poly median of Table 7 — while the mono-Si systems dominate the *agrees* class (3 of 8). The GHI-only strata invert the envelope picture without changing the conclusion: StoreNet’s envelope is wide (median span 3.7 pp, dominated by the decomposition choice) but one-sided — swapping models moves predictions *away* from the measurements, so the reachable minimum stays at 13.1 pp and the over-prediction reads as input and prior level, not model form. The classification is deliberately coarse — first matching rule wins, so a site with several defects (e.g. id03’s zero-filled outages *and* missing geometry) is counted once under the strongest committed signal — and M4_C.II illustrates the rules’ honesty: its measured PR of 0.947 sits just under the 0.95 metadata gate, so it files under input/prior even though Section 5 argues its back-fit nameplate is the likelier cause.

Known threats to validity, and their treatment:

- *Metadata quality.* Nameplate metadata from public portals can be wrong; the capacity-sanity gate catches gross corruption, but subtle errors (e.g. an undocumented string outage) deflate apparent accuracy. The Trivedi diagnosis of Section 5 is this study’s concrete instance on the harness’s own side: a plausible-looking peak-derived nameplate, wrong by $\times 1.71$ – 2.33 , dominated a whole stratum’s headline metrics until cross-checked against the data descriptor (and rehabilitated the stratum once corrected). We report per-site results so outliers are visible rather than averaged away. In the initial run the peak-to-nameplate anchor check flagged five scored DKASC sites on the high side (M7_A +55 %, M8_B +53 %, M8_C +50 %, M10_BC +34 %, M11_3 +31 %). A cross-check against the DKASC portal (July 2026) confirmed the configured nameplate of four — M7_A 5.805 kW ($27 \times$ SPR-215), M8_B/M8_C 5.04 kW ($18 \times$ SG-280), M10_BC 16.8 kW — so those excursions are measurement artifacts, not configuration errors. The fifth “confirmation” proved circular: the check read M11_3’s 31.5 kW off the same mis-joined portal page the scraper

Table 15: Per-site error attribution by the strongest committed signal (first matching rule): *agrees* = $|\Delta E| \leq 3\%$, the measurement floor; *metadata* = measured PR > 0.95 (nameplate understated); *avail-doc* = declared availability epochs; *avail-inf* = offline-filter drop > 10% or both baselines agreeing ≥ 10 pp above pvSpline; *model* = some defensible sub-model choice (Hay–Davies/Perez $\times$ Erbs/Engerer2 $\times$ linear/De Soto) lands within the floor; *input/prior* = the systematic remainder. “Span” is the median width of the model-choice envelope [pp]; “min $|\Delta E|$ ” the median closest approach any defensible choice achieves; “[input swap]” the median per-site shift under the reanalysis weather substitution (an input change, kept outside the model envelope; identically zero on the strata already driven by reanalysis). Generated from the committed run summaries by `make_attribution.py`; per-site classes in `attribution.json`.

Stratum	agrees	metadata	avail-doc	avail-inf	model	input/prior	med. span	med. min $ \Delta E $	med. [input swap]
DKASC fixed-tilt	8	1	1	9	1	11	0.8	7.1	4.5
DKASC tracked	0	0	3	0	0	0	1.4	51.1	17.7
StoreNet	1	2	0	0	0	7	3.7	13.1	7.3
OPSD	0	0	0	0	0	1	3.3	20.3	0.0
Trivedi	0	0	0	4	0	0	5.9	10.8	0.0

had used, and the meter-level investigation of Section 5 later overturned it (wrong plant entirely; true nameplate 26.52 kW, fixed mounting); at the corrected nameplate M11_3 re-flags harder, at +45%, now attributable to the commissioning metering spike below rather than to the nameplate. The canonical run anchors on a robust 30-minute peak statistic (adopted after this diagnosis), which corroborates the spike reading: the three single-row-spike sites fall from +50–+55 to +30–+32 once isolated rows stop carrying the peak, while M11_3’s seven-week metering epoch survives the robust statistic. The check is two-sided, and the corrected run’s complete flag set is in Appendix C: most flagged desert systems are on the *low* side, corroborating the conservative-prior reading of the fleet’s negative bias rather than signalling further metadata error. A third identity anchor joined the peak and PR checks with the 2026-07-22 follow-ups: a PVUSA/ASTM-E2848-style regression of measured power on measured in-plane irradiance over clear high-irradiance rows ($P = G(a + bG)$, evaluated at 1000 W m^{-2}), fitted per site-year on measured-sensor rows only (`make_capacity_anchor.py`, committed as `capacity_anchor.json`; 31 of the 34 DKASC meters reach a fit, fleet median 753 W kWp^{-1} ; the European strata carry no measured plane-of-array channel and deliberately no-op rather than laundering model output through a model-independent label). Its census partitions the known suspects cleanly: the only three sites beyond $\pm 20\%$ of the fleet median are the three identity/nameplate cases — M11_3 at -28.5% (mistaken plant; its per-year series dates the 2016–2021 collapse, Section 5) and the two disclosed back-fits M19_C/M4_C.II at $+32.8/+30.0\%$ (under-sized denominators) — and the only two sites whose realized capacity declines beyond $2\% \text{ yr}^{-1}$ are M8_C (-4.0) and M18_C (-3.6), the two progressive-fault arcs of Section 5.5, confirmed from the capacity side. The clear-day orientation screen (`make_geometry_check.py`, `geometry_check.json`) is the same family’s screening-grade member: a shape-only tilt/azimuth grid fit over each site’s twenty clearest days, level-normalized so it is orthogonal to nameplate scale, with an identifiability guard — a deviation flags only when the fitted plane also at least halves the configured plane’s shape residual, so argmin wander over a flat objective cannot flag a correct configuration (the guard’s selftest clips a correctly-configured synthetic plant until the bare threshold false-flags it, a case the guard drops while keeping a true 22° mismatch flagged). On the fleet the guard cuts the census from the bare threshold’s 43 to 31 of 49 sites, and its verdicts sort cleanly. It clears every noise-floor case: no StoreNet or OPSD site survives (their 1-minute and hourly shapes cannot materially beat the configured plane; StoreNet median best-fit SSE 0.51), and id02’s 42° deviation drops as unidentifiable (improvement 0.46). It

keeps the validated Trivedi verdicts: id03 and id04 flag identifiably while the id01 control fits the configured plane back to within 5.7° (Section 5). What survives beyond them is 29 desert sites whose fits nominally beat portal-documented geometries by a median 0.88 of the configured plane’s shape residual at deviations of $12\text{--}35^\circ$. Given those documented geometries, the measured plane-of-array channel they agree with (Section 5.4), and the stratum’s headline accuracy at the configured planes, we read the surviving desert flags not as metadata error but as the shape objective absorbing systematics it does not model — the thermal and spectral shape terms a pure transposition candidate lacks — into plane compensation. The screen therefore stays ranked evidence for human investigation, not a census; its validated verdicts on this fleet are the Trivedi ones. For the three fixed-tilt sites the excursion is a documented single-row sensor spike that corrupts only the peak statistic (their bulk metrics are unremarkable: $\text{nRMSE}_{\text{cap}}$ 5.1–7.3 %, $|\Delta E| \leq 3.3\%$). For the trackers, M10_BC is a SolFocus CPV concentrator whose clear-sky, two-axis DNI response can legitimately exceed the flat-plate peak anchor; M11_3’s 53 kW “peak” is a seven-week $\approx 2\times$ metering-scale error at commissioning, undocumented on the portal. The portal’s maintenance history also records real availability losses the simulation does not model — intermittent midday AC dropouts on M7_A through 2022 (inverters replaced December 2022) and recurring tracker-drive failures on M10_BC, including one unit offline from 2021 to April 2022 — which act in the direction of the observed over-prediction on the tracked stratum.

- *Zero-inflated meters.* The Trivedi outage filter (Section 3.3) risks excluding genuine zero-production rows; the excluded rows are published per site for audit, and the unfiltered stratum energy deviations are reported in the ‘none’ column of Table 13 (populated from the canonical run’s per-site exports). The dataset’s released files zero-fill outages up to 197 days long (Section 5), so the filter is load-bearing there; the +69 to +163 % deviation that survived it in the initial run was attributed primarily to that run’s spike-derived nameplate rather than to residual under-recording — confirmed by the corrected canonical run’s +7 to +42 % — and the unfiltered external baselines missed larger still at the same inflated nameplate (+94 to +240 % for PVGIS, Section 5.8) exactly because the zero-filled windows stay in their joins.
- *Shared implementation between harness and product.* The harness reuses the production physics modules, but its per-record chain assembly (loading, joining, soiling accumulation, derates) is validation-specific and unit-tested separately. The product’s own simulation pipeline, with its typical-year weather and default loss stack, is not itself scored here, so the results carry over to the product only as far as those two wrappers agree.
- *Selection of datasets.* All four datasets are public and were selected before results were known; no dataset was dropped after scoring.
- *Uncertainty of the measurements themselves.* Pyranometer and revenue-meter uncertainties (typically 2–5 % per the NREL solar-resource best-practices handbook [53], and 0.5–2 % for revenue-class meters per their accuracy-class standards [54, 55]) bound the achievable agreement; sub- % deviations should not be over-interpreted.
- *Fleet composition.* The validated fleet is diverse in module technology and resolution but bounded in two ways the results should be read against. Mounting: it contains fixed-tilt and *dual*-axis-tracked systems only — no single-axis tracker, the dominant utility-scale configuration, is scored, so the tracker findings speak to two-axis geometry alone. Climate: three Köppen classes (BWh, Cfb, Cfb/Dfb) leave humid-subtropical, tropical, snow-dominated, and high-altitude regimes untested, and the hourly, continental-leaning stratum is a single system (OPSD). Extending the fleet to single-axis trackers and to those

climates (candidate public, sub-hourly, provenance-documented datasets exist) is the most useful next step for external validity, kept separate here so no dataset is added after seeing results (previous item).

- *Statistical dependence between sites.* Systems within a dataset share a weather record (and, at DKASC, a location), so the 49 sites are not independent samples of “PV systems in general”; the bootstrap intervals of Section 4 quantify site-resampling variability within this fleet, not sampling from a broader population, and the per-dataset stratification of every results table is the honest unit of generalization. A design-effect estimate makes the dependence quantitative (one-way-ANOVA intraclass correlation over the four dataset clusters, Kish cluster size for the unequal groups): whole-record energy deviation clusters weakly (ICC 0.13, effective sample size ≈ 11 of 49), but pointwise $\text{nRMSE}_{\text{cap}}$ is largely *cluster-determined* (ICC 0.76, effective sample size ≈ 2.5) — unsurprising, since native resolution and the shared-sensor geometry set that metric’s scale per dataset — so fleet-level pointwise medians should be read as summaries of roughly two to three independent observations, and the energy metric as the one that generalizes. The StoreNet stratum shows this most sharply: the houses pair into near-identical per-site metrics (H2/H3, H4/H5, H7/H10), because all ten are driven by one weather station under a single assumed 35° /south geometry, so their *predictions* nearly coincide. The *measured* records are nonetheless distinct meters, not re-exports of a shared feed — their independent household consumption and battery channels differ on every overlapping row, and most pairs differ in measured production as well — so the near-degeneracy is a prediction-side artifact of shared inputs, and the stratum carries fewer effective degrees of freedom for *pointwise* error than its ten meters suggest, while remaining ten genuine measured series for the energy comparison.

7 Reproducibility

The validation harness is part of the pvSpline source tree, which is proprietary and not publicly distributed. Within it, the headline sweep is a single invocation over the public datasets,

```
pvSpline validate --data-root <datasets> --out validation/reports/ --diag-csv
```

and `validation/scripts/generate_paper_data.py` chains it with the ablations, the baseline legs, and the figure, table, and claim-lint regeneration. Site configurations (including per-field provenance annotations), dataset loaders, metric implementations, and the report generator are in `validation/` and `src/validation/`. The headline results in this draft come from the canonical sweep of 2026-07-30, committed with its `summary.md` and per-model JSON reports at `validation/reports/20260730-021242/` (pinned in `canonical_runs.json`) — run on the engine carrying the 2026-07-22 accuracy and data-quality follow-ups and both metadata corrections (Trivedi nameplate + M11.3 identity). Determinism is demonstrated at two strengths. Within one session and binary, the retained 2026-07-22 canonical pair (`validation/reports/20260722-165442/` and the same-session re-execution `validation/reports/20260722-155212/`) is bit-identical per site. Across rebuilds, the canonical sweep is the fifth execution of one run: it reproduces `validation/reports/20260723-013216/`, `20260723-201917/`, `20260723-233610/` and `20260725-052241/` (all retained) bit-identically, the five spanning four build commits over seven days. The linter enforces both identities — against whichever sweep is pinned, not against a fixed stamp — together with the disclosed median 0.2pp/max 0.7pp follow-ups shift from the 2026-07-22 canonical (Section 5). The superseded 2026-07-17 canonical sweep is retained at `validation/reports/20260717-050926/`, together with the 2026-07-14 sweep that reproduced it to within 3×10^{-4} pp (`validation/reports/20260714-195130/`) and the first fully-corrected sweep at

`validation/reports/20260713-084406/`; the intermediate Trivedi-only correction is retained at `validation/reports/20260713-060315/`, and the initial sweep — commit `8db89c8`, clean tree, 2026-07-12; its archived summaries embed the full hash — at `validation/reports/20260712-193827/` as the basis of the two case studies; within that 2026-07-12...17 lineage each metadata correction changed only the corrected sites' rows, every other row being bit-identical across those archives. Every generated report embeds the producing binary's commit hash and a dirty-tree flag.

The reproduction is tractable on commodity hardware. The headline sweep scores 53 million measured–predicted records across 49 systems under each of three transposition models (147 site–model passes) in a single invocation, single-threaded through the shared physics implementations; the external baselines add one `pvl` pass per instrumented dataset and one cached PVGIS query per site. The full-data driver run behind this draft bears the budget out: the 147-pass canonical sweep completed in ≈ 32 minutes, and the end-to-end pipeline — unit tests, the sweep, the four ablation families, the `pvl`/PVWatts/PVGIS and PVGIS-horizon legs, irradiance-stage scoring, and the figure, table, and claim-lint regeneration — in ≈ 1.6 hours on a single desktop machine. Both are wall-clock timings taken on the 2026-07-17 engine, and they are quoted as an order-of-magnitude budget for a rerun rather than as a benchmark: the harness records neither the host specification nor the sweep's peak resident set, so neither is stated here, and a re-timing carrying both accompanies the submission run.

One provenance wart is disclosed rather than hidden: the embedded hash is captured at build-configure time, so the 2026-07-13 summaries record the binary's last-configure commit rather than the config-only commits that corrected the Trivedi nameplate; the repository history of the Trivedi site-configuration file is the authoritative provenance for that change. The superseded 2026-07-17 sweep did not inherit the wart: its summaries embed the clean-tree commit the producing binary was configured and built at (`d5314af`, dirty flag clear), so its provenance is exact. The 2026-07-22 canonical pair re-opened a variant of it: its summaries embed commit `602ecb5` with the dirty-tree flag *set* — the producing working tree carried the paper-pipeline driver fix merged immediately afterwards (PR #1548) when the binary was configured. The families that follow repeat the pattern four times: the first 2026-07-23 sweep and ablations embed commit `69ed856` (the merged follow-ups implementation), the 2026-07-23 re-executions plus the satellite leg embed `3075551b` (the merged paper-content PR), the 2026-07-25 family embeds `18bc374` (the merged PVsyst template-patch PR), and the canonical 2026-07-30 family embeds `8d5215b` (a web-front-end package bump, three days before the run), each with the dirty-tree flag set — though the bit-identical reproduction across all five of those sweeps bounds what any of the dirt could have touched to nothing that reaches a scored number. What those trees actually carried is not recoverable after the fact — an uncommitted working state leaves no record — so we report the flags as the binaries stamped them rather than dropping them, and let the cross-build reproduction bound their reach. The submission archive removes the ambiguity at the source instead of arguing it away: a clean rebuild at a tagged commit, re-run end to end.

The statistics layer of this manuscript is itself generated: the confidence intervals, signed-rank p-values, and correlation intervals quoted in Section 5 are computed from the archived summaries by `doc/paper/scripts/make_numbers.py` into a macro file the manuscript inputs, and the same script's `--check` mode re-derives the hand-written literals (headline medians, counts, split sizes) from the archives and fails on any mismatch — the guard that caught this draft's own pre-correction residue. Further generators produce paper tables from the archives and from a full run's per-site exports: `make_diagnostics.py` (Appendix C) and `make_tech.bias.py` (Table 7) read the committed summaries directly; `make_pr_aggregation.py` (the performance ratio of Section 5 and the aggregation curve of Section 5.3) and `make_residual_analyses.py` (the residual, degradation, and filter-sensitivity tables of Section 5.5) read the per-site row-level exports, each carrying a `--selftest` or placeholder mode so the manuscript compiles before those exports

exist. Both restrict themselves to the sweep summary’s scored-site list — a per-site CSV on disk is not proof the site survived the skip flags or the capacity quarantine, so the summary is the allow-list and quarantined meters cannot enter the stratum medians — the degradation fits gate out partial calendar years and mask the documented availability epochs recorded in the site configurations, the residual generator also emits the persistence-detected candidate outage epochs of Section 3.3, and both generators rebuild the committed tables and JSONs from their committed outputs (`--rebuild-from-json`) when the row exports are absent. A further generator, `make_attribution.py`, derives the error-attribution matrix of Table 15 purely from the committed sweep, ablation, and baseline summaries. Three instrument-layer generators added with the 2026-07-22 follow-ups read the same per-site exports and commit their verdicts beside the other analysis artifacts — `make_sensor_drift.py` (`sensor_drift.json`, the clear-envelope irradiance-drift bound of Section 5.5), `make_capacity_anchor.py` (`capacity_anchor.json`, the PVUSA-style capacity anchor of Section 6), and `make_geometry_check.py` (`geometry_check.json`, the clear-day orientation screen) — each carrying a `--selftest` that recovers an injected defect (a $-0.7\% \text{ yr}^{-1}$ sensor drift, an understated nameplate and a capacity decline, a wrong-by- 22° orientation). A fourth, `make_drift_witness.py` (`drift_witness.json`, the reanalysis cross-trend of Section 5.5), needs only the committed station envelopes plus an Open-Meteo archive fetch, so it reruns anywhere with network access; its selftest pins the cross-trend algebra (a drifting station over a flat sky moves the ratio by the drift, co-rising envelopes leave it flat). The snowfall overlay of Section 5 rests on a committed sidecar (`snow_OPSPD_industrial1_pv1.csv`, an Open-Meteo daily export for the Konstanz site fetched 2026-07-23, stored in mm) that `make_numbers.py` crosses with the detector’s committed epochs; a full run picks the same sidecar up and stamps `snow_day_share` on each detected epoch.

The `pvlib-python`, `PVWatts`, and `PVGIS` baselines live in `validation/scripts/` as `pvlib_baseline.py`, `pysam_baseline.py`, and `pvgis_baseline.py` (shared loaders and metrics in `baseline.common.py`, kept in lockstep with the C++ harness loaders); all three pin their library versions and protocol parameters in every output they write, and `validation/scripts/compare_baselines.py` folds their summaries into the comparison tables of Section 5.8. A single wrapper, `validation/scripts/run_baselines.py`, executes the three baseline legs, the irradiance-stage check of Section 5.4 (`irradiance_stage.py`, whose ported decomposition and transposition are self-tested against `pvlib` via `--selftest`), and the comparison in one command (Section 4.5). The baseline summaries behind Section 5.8 (`pvlib` 0.15.2, `pandas` 3.0.3; `PySAM` 7.1.1; `PVGIS` v5.3 with the radiation database and request window recorded per site) are archived under `validation/reports/pvlib_baseline/`, `validation/reports/pysam_baseline/`, and `validation/reports/pvgis_baseline/` with the comparison generator’s output alongside; the StoreNet `pvlib` summary is the post-fix re-run, and the superseded defective first run disclosed in Section 5.8 is retained in the repository history. One archival step remains, stated here rather than left implied: a release tag pinning the exact revision behind the submission run; the runs in this draft are pinned by commit hash alone. The datasets need no deposit of their own. The OPSD and Trivedi originals carry citable data DOIs (10.25832/household_data and 10.6084/m9.figshare.30294478, CC BY 4.0), as does StoreNet (10.6084/m9.figshare.c.6829134, CC0); DKASC is cited by portal and access date, since its terms cap redistribution of unaltered data at 5,000 cells without the owner’s consent. Byte-level identity with a later download is checkable rather than assumed: `validation/scripts/dataset_manifest.py` writes and verifies a sorted SHA-256 manifest of the scored data root that plain `sha256sum -c` also reads.

8 Conclusion

We validated `pvspline`, a `PVsyst`-compatible desktop PV simulation engine, against 49 grid-connected systems from four public datasets — 53.4 million measured–predicted records span-

ning roughly 470 system-years across hot-desert and temperate European climates at 1–60 minute native resolution — under a protocol in which every site parameter is provenance-tracked and none is calibrated against the scored series (disclosed nameplate back-fits excepted, Section 3.2). On the cleanest stratum, 31 fixed-tilt desert systems with on-site weather and 5-minute metering, the engine achieves a median capacity-normalized RMSE of 5.8%, median R^2 of 0.959, and median whole-record energy deviation of -2.2% — within the $\pm 3\text{--}8\%$ band reported for established hourly tools on high-quality data — and fleet-wide, whole-record energy deviation is within $\pm 15\%$ for 36 of 49 systems. The residual structure is informative: the mild negative bias on the desert fleet decomposes longitudinally into a conservative year-zero loss stack and degradation priors that book only a third to a half of the measured fade (median drift $+0.75\text{ pp yr}^{-1}$, about two-thirds of which a reanalysis-witnessed envelope audit attributes to irradiance-sensor drift, so records end above zero deviation), while the heavy tail is traceable per site to tracked aggregate meters with documented failure epochs, one dataset whose back-fit nameplate proved $1.7\text{--}2.3\times$ too large on top of zero-filled meter outages — diagnosed post hoc against its data descriptor, corrected in-tree, and thereby collapsed from $+69\ldots+163\%$ to $+7\ldots+42\%$ energy deviation — and one meter whose data export had been joined to the wrong portal page (corrected: $+123 \rightarrow +55\%$, the residual being its documented availability epochs) — input-quality failure modes that a validation protocol must surface rather than average away.

The transposition ablation shows the Hay–Davies/Perez choice to be second-order on this fleet (fleet-median $\text{nRMSE}_{\text{cap}}$ shift < 0.01 percentage points; Perez adds $+0.1$ to $+2.8$ points of energy depending on climate), supporting Hay–Davies as a sound default in a GHI-driven pipeline. The external baselines sharpen the residual-structure reading from both sides: a same-input pvlib-python re-implementation sits a near-uniform $+8$ points of energy deviation above pvSpline exactly where records are clean — the loss terms it omits — and departs wildly exactly where the offline filter flags meter outages, agreeing there with the independent-weather PVGIS run ($r = 0.97$) rather than with pvSpline, while on the shared-GHI Irish rooftops it reproduces pvSpline’s per-house deviations within 5 points; PVGIS itself matches pvSpline on clean-desert annual energy and beats it on the clean-metered European rooftops, quantifying both what the full chain buys (sub-hourly fidelity and per-site diagnostics more than clean-site annual energy) and where the uniform priors need work (Cfb rooftop losses). The decomposition ablation is inert wherever measured diffuse exists, and on the GHI-only sites the audit-corrected Engerer2 port trims a second-order 1.6 to 3.4 points off the over-prediction — leaving the Erbs default defensible rather than contradicted — while restoring intra-hour structure on the hourly dataset recovers -0.5 points of its over-prediction with no clipping in play. The weather substitution separates the input share of prospective error: $+4.1$ points of desert $\text{nRMSE}_{\text{cap}}$ and a median per-site energy shift of -4.5 points under reanalysis inputs, with StoreNet’s pointwise error not degrading — and, set against the PVGIS result, it shows that prospective inputs shift the calibration burden onto the loss model. The IV-model ablation is second-order for the crystalline fleet and a technology-specific refinement for thin film ($+1.7$ to $+7.3$ points of low-irradiance energy on those systems). The PVsyst-variant single-diode run remains for the submission version.

Data and code availability

All four measurement datasets are public. DKASC data and per-system metadata are downloadable from the DKA Solar Centre portal [37], whose terms cap redistribution of unaltered data at 5,000 cells without the owner’s consent, so the portal and access date are cited rather than a snapshot. Desert Knowledge Australia, the Australian Government, the Northern Territory Government and the project managers, Ekistica do not endorse, and accept no legal liability whatsoever arising from, or connected to, the outcomes and conclusions associated with the use of data from the Desert Knowledge Australia Solar Centre. StoreNet was released

by the IERC/StoreNet consortium with the Trivedi et al. descriptor [38] and is on figshare (10.6084/m9.figshare.c.6829134, CC0); the OPSD CoSSMic household data carry a citable data DOI (10.25832/household_data, CC-BY) [39]; the Dingle ambassador-household data are on figshare (10.6084/m9.figshare.30294478, CC BY 4.0) [40]; reanalysis weather is served by the freely accessible Open-Meteo archive [42]. The pvSpline engine and its validation harness (loaders, site configurations with per-field provenance, metrics, report generator, baseline scripts, and figure/statistics generators) are proprietary and not publicly distributed; per-site results are tabulated in Appendices A–C.

Acknowledgements

This study exists because its measurement data is public. We thank the DKA Solar Centre (Alice Springs) for the DKASC archive and its per-system metadata and maintenance history; the International Energy Research Centre and the StoreNet consortium, and Trivedi et al. [38], for the Dingle trial data; the Open Power System Data and CoSSMic projects [39]; Avgoloupis et al. [40] for the Dingle ambassador-household data descriptor; and the Open-Meteo project [42] for open reanalysis access. This work received no external funding.

Competing interests

The author is the developer of pvSpline. This conflict is inherent to a tool-author validation and is mitigated by the paper’s design: public datasets, provenance-tracked parameters with a no-calibration protocol, and per-site reporting.

A Per-site results

Table 16 lists every scored system under the chain defaults (Hay–Davies transposition). Values are computed on the all-hours inner join at native resolution; n is the joined row count. The per-site input-quality diagnostics the harness records — GHI→POA fallback coverage, DNI-ceiling clamps, soiling accruals, and anchor-check excursions — are tabulated in Appendix C (offline-filter drop fractions are in Appendix B); the Perez sweep and the full raw report are archived at validation/reports/20260730-021242/.

Table 16: Per-site accuracy, chain defaults (Hay–Davies slot of the archived canonical 20260730–021242 sweep). DKASC rows: Desert Knowledge Australia, the Australian Government, the Northern Territory Government and the project managers, Ekistica do not endorse, and accept no legal liability whatsoever arising from, or connected to, the outcomes and conclusions associated with the use of data from the Desert Knowledge Australia Solar Centre.

Dataset	Site	kWp	n [10^6]	nRMSE _{cap} [%]	R^2	ΔE [%]
DKASC	M10.BC [†]	16.80	1.37	14.6	0.833	+13.0
	M11.3	26.52	1.65	18.0	0.309	+54.5
	M15.A	5.40	1.56	5.3	0.971	−5.1
	M15.B	5.50	1.19	6.5	0.958	−13.1
	M15.C.II	5.20	0.918	8.7	0.937	−17.3
	M16.A.II	6.05	0.935	5.6	0.968	−9.7
	M16.C	5.50	1.40	5.8	0.961	−9.6
	M17.A	5.28	1.44	6.9	0.950	−13.3
	M17.B	5.50	1.28	5.9	0.961	−13.0
	M17.C	5.20	1.28	4.7	0.973	−8.8
	M18.B.II	5.30	0.884	4.8	0.975	−7.1
	M18.C	5.00	1.38	5.8	0.952	+4.3
	M19.A	5.50	1.30	4.4	0.976	−7.1
	M19.C [‡]	4.73	1.08	8.6	0.930	−21.5
	M1.A	1.98	1.70	5.1	0.964	+4.1
	M1.B	1.98	1.69	6.0	0.942	+9.9
	M1.C	1.98	1.67	9.7	0.863	+8.1
	M20.A	5.61	1.38	6.3	0.958	−14.0
	M20.B	6.00	1.19	4.9	0.974	−8.1
	M2.A	1.98	1.63	9.6	0.869	+7.6
	M4.A	6.00	1.67	5.2	0.966	−1.4
	M4.B	6.30	1.61	6.5	0.956	−11.8
	M4.C.II [‡]	4.69	0.751	9.7	0.932	−14.1
	M5.A	4.95	1.69	5.5	0.962	+1.0
	M5.B	5.10	1.67	6.0	0.954	+4.9
	M5.C	4.95	1.67	5.5	0.964	−2.2
	M6.A	6.96	1.56	5.5	0.959	+8.8
	M6.B	5.25	1.70	5.3	0.967	+0.3
	M7.A	5.80	1.67	7.1	0.940	−0.1
	M8.A	4.92	1.51	5.4	0.964	−1.3
	M8.B	5.04	1.56	4.9	0.971	−1.0
	M8.C	5.04	1.58	6.2	0.946	+2.8
	M9.AC [†]	23.40	1.19	21.9	0.271	+56.3
	M9.B [†]	10.50	1.18	20.1	0.384	+51.9
StoreNet	H1	2.20	0.525	18.6	0.212	+9.9
	H2	2.20	0.526	19.4	0.011	+19.5
	H3	2.20	0.526	19.2	0.050	+19.3
	H4	2.10	0.524	20.6	−0.172	+27.3
	H5	2.20	0.525	19.5	−0.167	+27.2
	H7	1.70	0.525	19.7	0.300	−4.5
	H10	1.70	0.525	20.1	0.339	−8.9
	H11	2.00	0.524	18.1	0.135	+18.2
	H13	2.00	0.523	20.2	0.068	+12.1
	H17	2.00	0.525	18.9	0.293	−0.5
OPSD	industrial1_pv1	5.00	0.017	15.3	0.546	+21.8
Trivedi	id01	2.10	0.047	9.7	0.760	+9.5
	id02	2.10	0.044	9.3	0.705	+18.3
	id03	2.10	0.045	10.6	0.517	+42.0
	id04	2.10	0.048	12.8	0.607	+6.5

[†]dual-axis-tracked system behind a combined multi-phase meter.

[‡]kWp back-fit from the measured 99.5th-percentile peak — the disclosed provenance exception of Section 3.2.

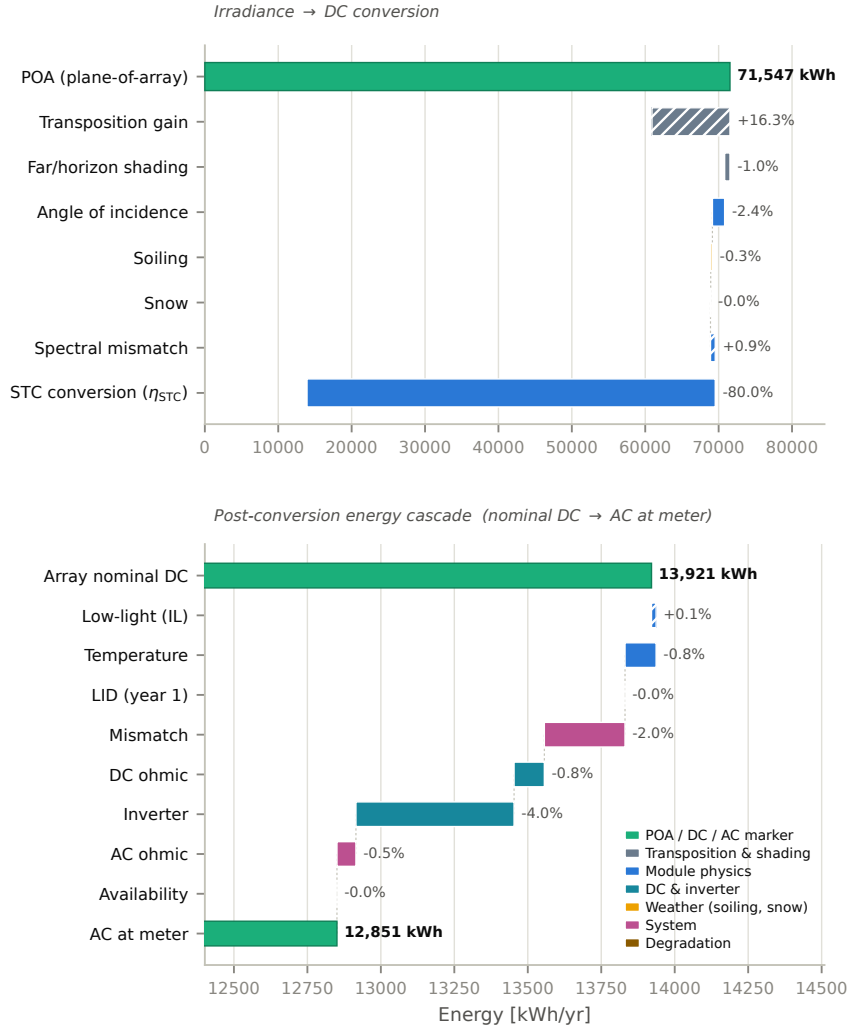


Figure 5: Per-stage loss cascade for a representative grid-connected fixed-tilt configuration, from plane-of-array irradiance to delivered AC energy, as computed by pvSpline’s unified loss model. *Top:* the irradiance-to-DC conversion — a transposition credit and far-horizon shading step the resource to the received plane-of-array irradiance, the optical steps (angle of incidence, soiling, snow, spectral) take it to what the cells absorb, then the module STC conversion (η_{STC}) collapses it to the array’s nominal DC energy. *Bottom:* the post-conversion cascade (nominal DC to AC at meter), zoomed so the system-side derates are legible; each bar is the energy removed at that stage (hatched: added), labeled with its fraction of the energy entering it. Each step is the signed net of its hourly term; on this deck the spectral and irradiance-level steps are small credits. The deck’s loss inputs mirror the PVsyst template of the A/B protocol of Section 4.5 rather than pvSpline’s standalone defaults — mismatch 2%, availability and first-year light-induced degradation configured to zero (their slots render at -0.0%), no module-quality derate — and its angle-of-incidence step evaluates the module’s Fresnel anti-reflective-coating IAM profile. The soiling term of the upper panel and the mismatch and ohmic entries of the lower panel are the structure the same-input pvlib baseline partially omits (Section 5.8; on the validated fleet the ohmic/mismatch entries are configured to zero, Section 3.2). Numbers are the ordered `LossModel` entries of the grid-fixed deck of the `pvsyst-compare` parity suite of Section 4.5 (Paris, 30° south), run with Hay–Davies transposition, which keeps PVGIS’s own plane-of-array irradiance ([validation/reports/paper_losswaterfall/](#)); rendered by `make_figures.py --losswaterfall`. The figure illustrates the loss *structure*, not a fleet accuracy result.

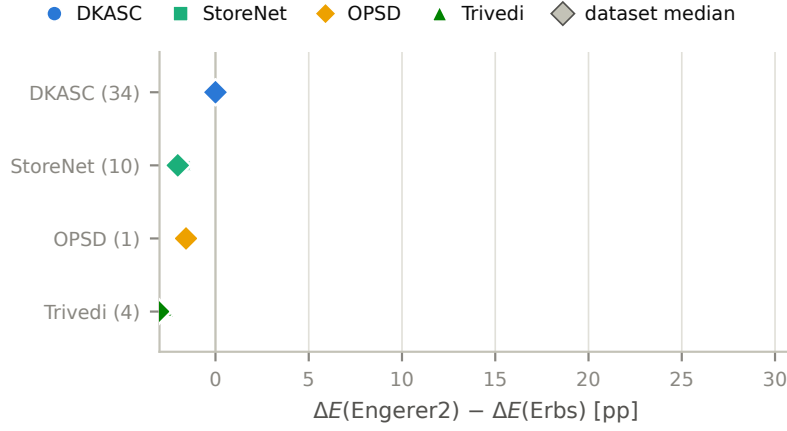


Figure 6: Per-site shift in whole-record energy deviation when the decomposition model swaps Erbs→Engerer2 (Section 5.2). The 34 DKASC marks coincide at zero — measured diffuse short-circuits decomposition — while the GHI-only datasets shift -1.6 to -3.4 pp. Generated from the archived run summaries by `doc/paper/scripts/make_figures.py --ablations` (as are Figures 8 and 10).

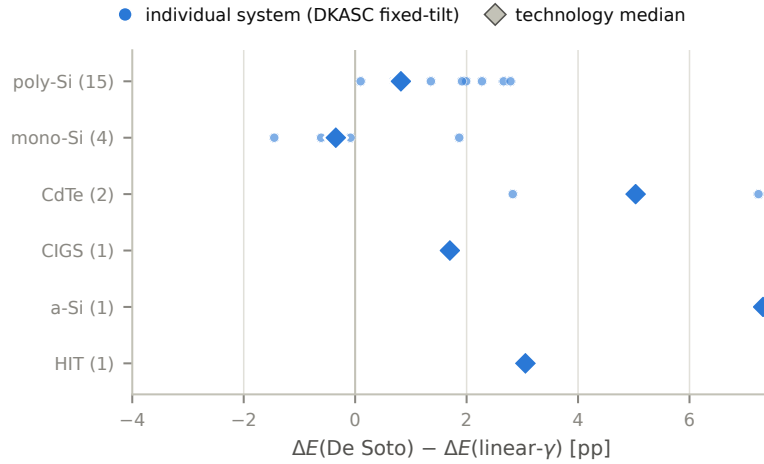


Figure 7: Per-site shift in whole-record energy deviation when the DC model swaps linear- $\gamma \rightarrow$ De Soto single-diode, by module technology (fixed-tilt DKASC systems with populated datasheet blocks). The crystalline bulk pins near zero while every thin-film and HIT system gains low-irradiance energy.

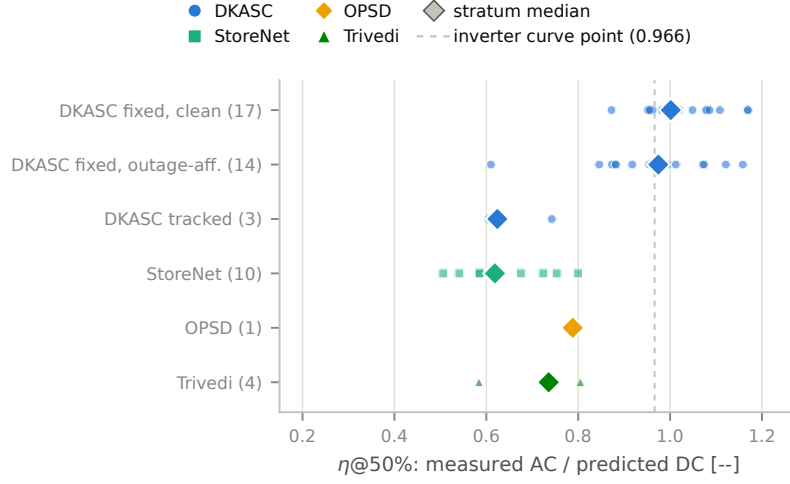


Figure 8: Realized-efficiency consistency check (Section 5.6): per-site measured-AC-to-predicted-DC ratio at the 50%-load bin, by stratum. Clean fixed-tilt systems cluster at the inverter curve point (the predicted-DC reference already carries the booked soiling and degradation derates, Section 2.8); the tracked aggregate meters collapse far below it, while the corrected Trivedi households rejoin the residential band.

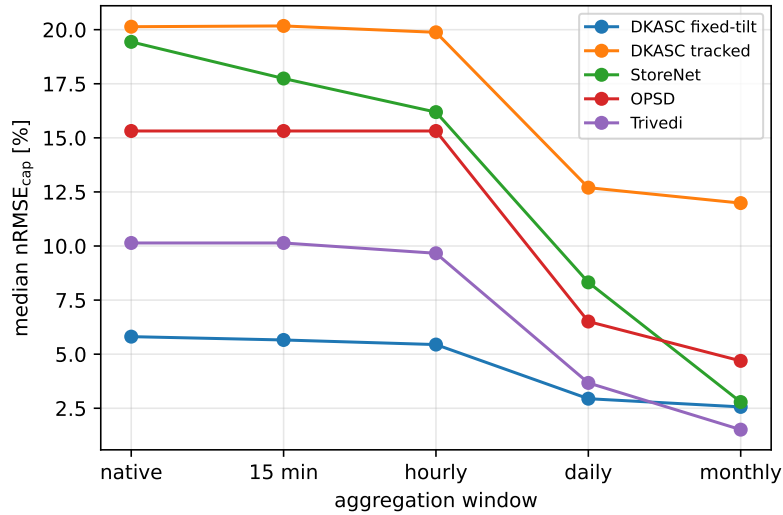


Figure 9: Error-versus-aggregation curve: median $nRMSE_{cap}$ per stratum against the resampling window (energy-conserving means). Generated from a full run's per-site exports by `make_pr_aggregation.py --figure`.

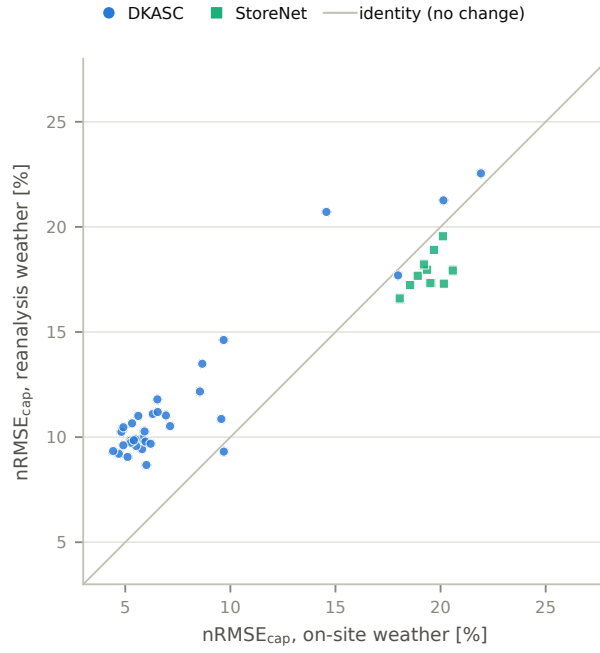


Figure 10: Weather substitution (Section 5.7): per-site $nRMSE_{cap}$ under on-site weather vs. the reanalysis substitution. The fixed-tilt desert fleet lifts off the identity line while StoreNet sits at or just below it; OPSD and Trivedi already run on reanalysis and reproduce identically (omitted).

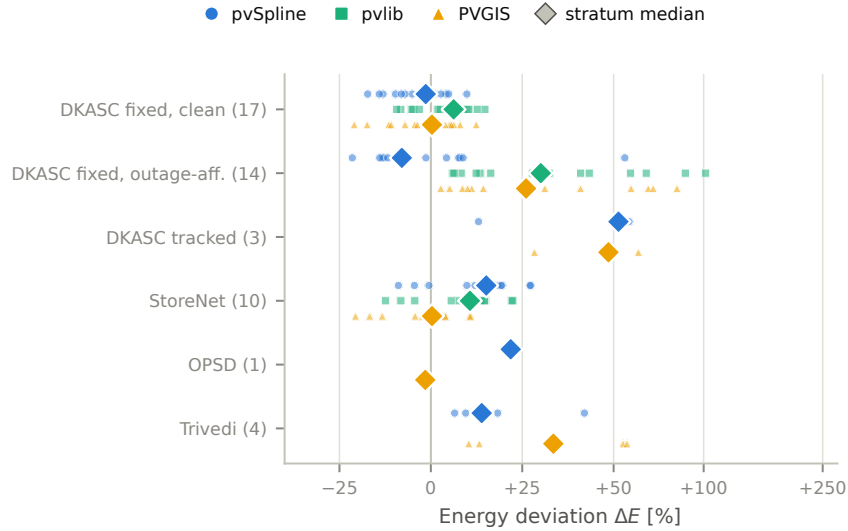


Figure 11: Per-site whole-record energy deviation by stratum and tool: small marks are individual systems, diamonds the stratum medians of Table 14. The axis is linear within $\pm 50\%$ and logarithmic beyond; pvlib covers fixed-tilt DKASC and StoreNet only; the PVWatts leg (same coverage) is tabulated in Table 14 and omitted here for legibility. Generated from the archived run summaries by `doc/paper/scripts/make_figures.py --baselines` (as is Figure 12).

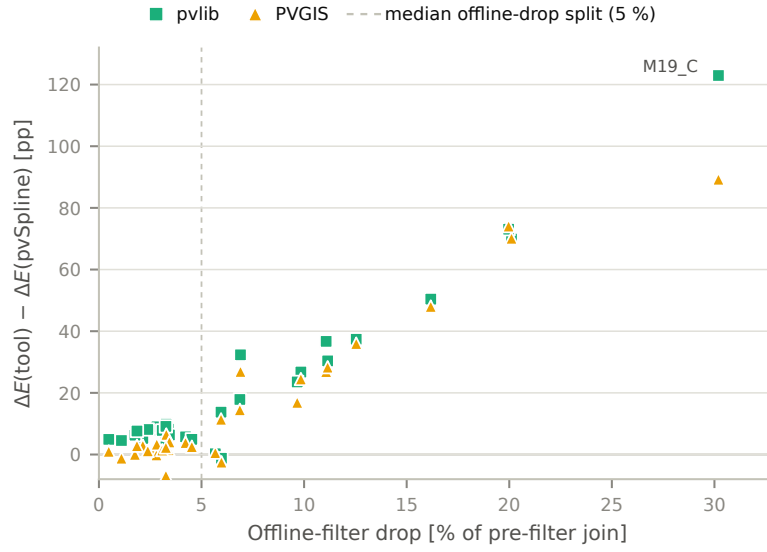


Figure 12: Paired per-site ΔE difference from pvSpline vs. the offline-filter drop fraction on the 31 fixed-tilt DKASC systems: on clean records the same-input pvlib baseline sits a near-uniform +8 pp above pvSpline (dominated by the omitted BWh soiling accumulator — the configured DC derates are zero on this fleet, Section 3.2) while end-to-end PVGIS scatters around zero, and both baselines rise together with the outage fraction that they score and the harness excludes.

B Per-site baseline comparison

Table 17 lists whole-record energy deviation per site for pvSpline (chain defaults, native resolution, system-offline filter of Section 3.3 applied), the same-input pvlib-python baseline (native resolution, no offline filter), and the end-to-end PVGIS baseline (hourly aggregates, no offline filter), together with the fraction of the pre-filter inner join removed by the offline filter (“drop”). The PVGIS radiation database is auto-assigned and recorded per site: ERA5 for DKASC (Alice Springs lies outside SARA3 coverage), SARA3 for the European sites. Raw per-site metrics for every leg are archived under `validation/reports/`; the superseded first StoreNet pvlib run (Section 5.8) is retained in the repository history.

Table 17: Per-site whole-record energy deviation ΔE [%] by tool, and the pvSpline offline-filter drop fraction [% of the pre-filter join]. “—” = out of the pvlib baseline’s scope (tracked systems; datasets without a native weather loader). Per-site PVWatts deviations (same scope as pvlib) are in the committed comparison artifact `validation/reports/baseline_comparison.md`; Table 14 carries its stratum medians.

Dataset	Site	pvSpline	pvlib	PVGIS	drop [%]
DKASC	M10_BC [†]	+13.0	—	+28.3	12.1
	M11_3	+54.5	+86.8	+81.4	6.9
	M15_A	−5.1	+2.7	−4.4	2.7
	M15_B	−13.1	−4.9	−11.5	3.4
	M15_C_II	−17.3	−8.2	−17.4	2.8
	M16_A_II	−9.7	−5.1	−10.9	1.1
	M16_C	−9.6	+27.8	+26.4	12.5
	M17_A	−13.3	+13.5	+11.2	9.8
	M17_B	−13.0	+57.0	+57.2	20.1
	M17_C	−8.8	+64.3	+65.3	20.0
	M18_B_II	−7.1	+1.9	−3.7	2.8
	M18_C	+4.3	+41.0	+31.2	11.1
	M19_A	−7.1	+43.3	+40.9	16.2
	M19_C [‡]	−21.5	+101.5	+67.8	30.2
	M1_A	+4.1	+9.9	+8.0	4.2
	M1_B	+9.9	+14.8	+12.4	4.5
	M1_C	+8.1	+8.3	+8.7	5.7
	M20_A	−14.0	+16.4	+14.3	11.1
	M20_B	−8.1	−3.2	−7.0	0.5
	M2_A	+7.6	+6.4	+5.2	6.0
	M4_A	−1.4	+8.4	+5.1	3.3
	M4_B	−11.8	+6.1	+2.7	6.9
	M4_C_II [‡]	−14.1	−9.3	−21.0	3.3
	M5_A	+1.0	+6.2	+3.9	2.1
	M5_B	+4.9	+12.7	+6.3	3.1
	M5_C	−2.2	+4.2	+1.9	3.4
	M6_A	+8.8	+32.4	+25.7	9.7
	M6_B	+0.3	+6.5	+0.3	1.8
	M7_A	−0.1	+9.0	+2.2	3.3
	M8_A	−1.3	+12.4	+10.1	6.0
	M8_B	−1.0	+7.1	+0.2	2.4
	M8_C	+2.8	+10.4	+5.6	1.9
	M9_AC [†]	+56.3	—	+60.5	3.1
	M9_B [†]	+51.9	—	+48.6	3.2
StoreNet	H1	+9.9	+5.6	−4.3	0.0
	H2	+19.5	+14.8	+4.1	0.0
	H3	+19.3	+14.6	+3.9	0.0
	H4	+27.3	+22.4	+11.0	0.0
	H5	+27.2	+22.2	+10.7	0.0
	H7	−4.5	−8.2	−16.7	0.0
	H10	−8.9	−12.4	−20.6	0.0
	H11	+18.2	+13.7	+3.0	0.0
	H13	+12.1	+7.7	−2.3	0.0
	H17	−0.5	−4.4	−13.3	0.0
OPSD	industrial1_pv1	+21.8	—	−1.5	0.0
Trivedi	id01	+9.5	—	+13.3	11.8
	id02	+18.3	—	+53.8	19.0
	id03	+42.0	—	+55.4	14.4
	id04	+6.5	—	+10.4	11.2

[†]dual-axis-tracked system behind a combined multi-phase meter.

C Per-site input-quality diagnostics

Table 18 reproduces, per site, the input-quality counters the harness records during the sweep and writes into each run’s per-site warnings — the quantities Section 2 and Section 6 refer to but do not tabulate in the body. They are diagnostic, not accuracy metrics: a large GHI→POA share simply means the site leaned on the transposition chain rather than a tilted sensor, a high DNI-clamp rate flags a record whose GHI/geometry pairing often reached the physical ceiling, and the anchor column reports the peak-to-nameplate excursion of the robust 30-minute peak for every site; excursions beyond $\pm 15\%$ trip the per-site warning flag. The excursion is genuinely two-sided: besides the five high-side DKASC sites discussed in Section 6 and the four high-side Trivedi households (the corrected-nameplate residual of Section 5), the majority of the flagged desert systems sit on the *low* side (-17 to -26%), the same direction as the fleet’s mild negative energy bias and consistent with the deliberately conservative uniform priors — not with metadata error. The table is generated from the archived summaries by `doc/paper/scripts/make_diagnostics.py`.

Table 18: Per-site input-quality diagnostics recorded by the harness during the chain-defaults sweep (Hay–Davies slot of the archived 20260730–021242 run). “GHI→POA” is the share of a site’s weather rows transposed from GHI rather than read from a tilted-plane sensor (100% where no POA channel exists, and for tracked systems whose fixed reference sensor is discarded); the complement used the measured plane-of-array reading. “DNI clamp” is the fraction of rows whose input GHI hit the extraterrestrial ceiling (Section 2). Soiling columns are the Kimber–Coello accumulator’s dry-day and wash-day counts and the cumulative fractional loss washed off over the record (“—” where the soiling prior is inactive). “Anchor” is the peak-to-nameplate excursion of the robust 30-minute peak statistic, reported for every site since the harness emits it first-class; excursions beyond $\pm 15\%$ trip the per-site warning flag of Section 6 (“—” = not recorded). Offline-filter drop fractions are in Appendix B.

Dataset	Site	GHI→POA [%]	DNI clamp [% rows]	Soiling dry / wash [d]	washed [pts]	Anchor [%]
DKASC	M10_BC	97	0.32	4 776 / 833	11.0	+32
	M11_3	35	0.14	5 408 / 963	12.6	+45
	M15_A	30	0.10	4 948 / 898	11.5	−13
	M15_B	12	0.04	3 850 / 659	8.9	−14
	M15_C.II	10	0.03	2 987 / 534	7.1	−13
	M16_A.II	10	0.03	2 988 / 533	7.1	−22
	M16_C	30	0.10	4 949 / 897	11.5	−17
	M17_A	30	0.10	4 949 / 897	11.5	−14
	M17_B	30	0.10	4 949 / 897	11.5	−18
	M17_C	30	0.10	4 949 / 897	11.5	−23
	M18_B.II	11	0.03	2 879 / 512	6.8	−26
	M18_C	30	0.10	4 825 / 853	11.2	−25
	M19_A	30	0.10	4 825 / 853	11.2	−23
	M19_C	30	0.10	4 825 / 853	11.2	−8
	M1_A	35	0.14	5 408 / 963	12.6	−22
	M1_B	35	0.14	5 407 / 964	14.8	−18
	M1_C	35	0.14	5 408 / 963	12.6	−25
	M20_A	30	0.10	4 824 / 854	11.2	−17
	M20_B	12	0.04	3 789 / 616	8.6	−22
	M2_A	36	0.15	5 283 / 920	12.3	−20
	M4_A	36	0.15	5 282 / 921	12.3	−22
	M4_B	36	0.15	5 270 / 920	12.2	−18
	M4_C.II	12	0.04	2 445 / 420	5.8	−8
	M5_A	36	0.15	5 269 / 921	12.2	−22
	M5_B	36	0.15	5 270 / 920	12.2	−24
	M5_C	36	0.15	5 270 / 920	12.2	−21
	M6_A	36	0.15	5 270 / 920	12.2	−24
	M6_B	36	0.15	5 270 / 920	12.2	−21
	M7_A	36	0.15	5 270 / 920	12.2	+32
	M8_A	31	0.10	4 868 / 862	11.3	−22
	M8_B	31	0.10	4 868 / 862	11.3	+32
	M8_C	31	0.10	4 868 / 862	11.3	+30
	M9_AC	97	0.31	3 796 / 616	8.6	−23
	M9_B	97	0.31	3 776 / 616	8.6	−22
StoreNet	H1	100	2.34	0 / 0	0.0	−18
	H10	100	2.34	0 / 0	0.0	−24
	H11	100	2.34	0 / 0	0.0	−25
	H13	100	2.34	0 / 0	0.0	−19
	H17	100	2.34	0 / 0	0.0	−17
	H2	100	2.34	0 / 0	0.0	−27
	H3	100	2.34	0 / 0	0.0	−25
	H4	100	2.34	0 / 0	0.0	−24
	H5	100	2.34	0 / 0	0.0	−27
	H7	100	2.34	0 / 0	0.0	−24
OPSD	industrial1_pv1	100	1.61	0 / 0	0.0	−26
Trivedi	id01	100	0.99	0 / 0	0.0	+59
	id02	100	0.99	0 / 0	0.0	+63
	id03	100	0.97	0 / 0	0.0	+104
	id04	100	0.99	0 / 0	0.0	+118

References

- [1] André Mermoud and Thibault Lejeune. Performance assessment of a simulation model for PV modules of any available technology. In *Proc. 25th European Photovoltaic Solar Energy Conference*, Valencia, Spain, 2010. doi: 10.4229/25thEUPVSEC2010-4BV.1.114.
- [2] Kenneth J. Sauer, Thomas Roessler, and Clifford W. Hansen. Modeling the irradiance and temperature dependence of photovoltaic modules in PVsyst. *IEEE Journal of Photovoltaics*, 5(1):152–158, 2015. doi: 10.1109/jphotov.2014.2364133.
- [3] Nate Blair, Nicholas DiOrio, Janine Freeman, Paul Gilman, Steven Janzou, Ty Neises, and Michael Wagner. System Advisor Model (SAM) general description. Technical Report NREL/TP-6A20-70414, National Renewable Energy Laboratory, 2018.
- [4] Aron P. Dobos. PVWatts version 5 manual. Technical Report NREL/TP-6A20-62641, National Renewable Energy Laboratory, 2014.
- [5] 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. doi: 10.21105/joss.00884.
- [6] Kevin S. Anderson, Clifford W. Hansen, William F. Holmgren, Adam R. Jensen, Mark A. Mikofski, and Anton Driesse. pvlib python: 2023 project update. *Journal of Open Source Software*, 8(92):5994, 2023. doi: 10.21105/joss.05994.
- [7] Christopher P. Cameron, William E. Boyson, and Daniel M. Riley. Comparison of PV system performance-model predictions with measured PV system performance. In *Proc. 33rd IEEE Photovoltaic Specialists Conference*, 2008. doi: 10.1109/pvsc.2008.4922865.
- [8] Janine Freeman, Jonathan Whitmore, Nate Blair, and Aron P. Dobos. Validation of multiple tools for flat plate photovoltaic modeling against measured data. Technical Report NREL/TP-6A20-61497, National Renewable Energy Laboratory, 2014.
- [9] Joshua S. Stein. The photovoltaic performance modeling collaborative (PVPMC). In *Proc. 38th IEEE Photovoltaic Specialists Conference*, pages 3048–3052, 2012. doi: 10.1109/pvsc.2012.6318225.
- [10] Marios Theristis, Nicholas Riedel-Lyngskær, Joshua S. Stein, Lelia Deville, Leonardo Micheli, Anton Driesse, William B. Hobbs, et al. Blind photovoltaic modeling intercomparison: A multidimensional data analysis and lessons learned. *Progress in Photovoltaics: Research and Applications*, 31(11), 2023. doi: 10.1002/pip.3729.
- [11] P. G. Loutzenhiser, H. Manz, C. Felsmann, P. A. Strachan, T. Frank, and G. M. Maxwell. Empirical validation of models to compute solar irradiance on inclined surfaces for building energy simulation. *Solar Energy*, 81(2):254–267, 2007. doi: 10.1016/j.solener.2006.03.009.
- [12] Christian A. Gueymard and José A. Ruiz-Arias. Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. *Solar Energy*, 128:1–30, 2016. doi: 10.1016/j.solener.2015.10.010.
- [13] IEC. IEC 61724-1:2021 photovoltaic system performance — part 1: Monitoring. International Electrotechnical Commission, Geneva, 2021.
- [14] John A. Duffie and William A. Beckman. *Solar Engineering of Thermal Processes*. Wiley, Hoboken, NJ, 4th edition, 2013. doi: 10.1002/9781118671603.

- [15] Ibrahim Reda and Afshin Andreas. Solar position algorithm for solar radiation applications. *Solar Energy*, 76(5):577–589, 2004. doi: 10.1016/j.solener.2003.12.003.
- [16] D. G. Erbs, S. A. Klein, and J. A. Duffie. Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation. *Solar Energy*, 28(4):293–302, 1982. doi: 10.1016/0038-092x(82)90302-4.
- [17] N. A. Engerer. Minute resolution estimates of the diffuse fraction of global irradiance for southeastern Australia. *Solar Energy*, 116:215–237, 2015. doi: 10.1016/j.solener.2015.04.012.
- [18] Pierre Ineichen and Richard Perez. A new airmass independent formulation for the Linke turbidity coefficient. *Solar Energy*, 73(3):151–157, 2002. doi: 10.1016/s0038-092x(02)00045-2.
- [19] J. E. Hay and J. A. Davies. Calculation of the solar radiation incident on an inclined surface. In *Proc. First Canadian Solar Radiation Data Workshop*, pages 59–72, 1980.
- [20] Richard Perez, Pierre Ineichen, Robert Seals, Joseph Michalsky, and Ronald Stewart. Modeling daylight availability and irradiance components from direct and global irradiance. *Solar Energy*, 44(5):271–289, 1990. doi: 10.1016/0038-092x(90)90055-h.
- [21] N. Martín and J. M. Ruiz. Calculation of the PV modules angular losses under field conditions by means of an analytical model. *Solar Energy Materials and Solar Cells*, 70(1): 25–38, 2001. doi: 10.1016/s0927-0248(00)00408-6.
- [22] David L. King, William E. Boyson, and Jay A. Kratochvil. Photovoltaic array performance model. Technical Report SAND2004-3535, Sandia National Laboratories, 2004.
- [23] Fritz Kasten and Andrew T. Young. Revised optical air mass tables and approximation formula. *Applied Optics*, 28(22):4735–4738, 1989. doi: 10.1364/ao.28.004735.
- [24] Mitchell Lee and Alex Panchula. Spectral correction for photovoltaic module performance based on air mass and precipitable water. In *Proc. IEEE 43rd Photovoltaic Specialists Conference*, pages 1351–1356, 2016. doi: 10.1109/pvsc.2016.7749836.
- [25] Christian Gueymard. Analysis of monthly average atmospheric precipitable water and turbidity in Canada and Northern United States. *Solar Energy*, 53(1):57–71, 1994. doi: 10.1016/s0038-092x(94)90606-8.
- [26] David Faiman. Assessing the outdoor operating temperature of photovoltaic modules. *Progress in Photovoltaics: Research and Applications*, 16(4):307–315, 2008. doi: 10.1002/pip.813.
- [27] IEC. IEC 61853-2:2016 photovoltaic (PV) module performance testing and energy rating — part 2: Spectral responsivity, incidence angle and module operating temperature measurements. International Electrotechnical Commission, Geneva, 2016.
- [28] W. De Soto, S. A. Klein, and W. A. Beckman. Improvement and validation of a model for photovoltaic array performance. *Solar Energy*, 80(1):78–88, 2006. doi: 10.1016/j.solener.2005.06.010.
- [29] Anton Driesse, Praveen Jain, and Steve Harrison. Beyond the curves: modeling the electrical efficiency of photovoltaic inverters. In *Proc. 33rd IEEE Photovoltaic Specialists Conference*, 2008. doi: 10.1109/pvsc.2008.4922827.

- [30] Adrianne Kimber, Lawrence Mitchell, Sara Nogradi, and Howard Wenger. The effect of soiling on large grid-connected photovoltaic systems in California and the Southwest region of the United States. In *Proc. IEEE 4th World Conference on Photovoltaic Energy Conversion*, pages 2391–2395, 2006. doi: 10.1109/wcpec.2006.279690.
- [31] Michael Coello and Lin Boyle. Simple model for predicting time series soiling of photovoltaic panels. *IEEE Journal of Photovoltaics*, 9(5):1382–1387, 2019. doi: 10.1109/jphotov.2019.2919628.
- [32] Dirk C. Jordan, Sarah R. Kurtz, Kaitlyn VanSant, and Jeff Newmiller. Compendium of photovoltaic degradation rates. *Progress in Photovoltaics: Research and Applications*, 24(7):978–989, 2016. doi: 10.1002/pip.2744.
- [33] J. W. Spencer. Fourier series representation of the position of the sun. *Search*, 2(5):172, 1971.
- [34] Ward Bower, Chuck Whitaker, William Erdman, Michael Behnke, and Mark Fitzgerald. Performance test protocol for evaluating inverters used in grid-connected photovoltaic systems. Technical report, California Energy Commission, 2005.
- [35] Philip Hülsmann and Karl-Anders Weiss. Simulation of water ingress into PV-modules: IEC-testing versus outdoor exposure. *Solar Energy*, 115:347–353, 2015. doi: 10.1016/j.solener.2015.03.007.
- [36] Dirk C. Jordan, Timothy J. Silverman, John H. Wohlgemuth, Sarah R. Kurtz, and Kaitlyn T. VanSant. Photovoltaic failure and degradation modes. *Progress in Photovoltaics: Research and Applications*, 25(4):318–326, 2017. doi: 10.1002/pip.2866.
- [37] Desert Knowledge Australia Centre. Download Data. Alice Springs. <http://dkasolarcentre.com.au/historical-data/download>, 2026. Date accessed: July 2026; byte-level contents pinned by the study’s SHA-256 dataset manifest.
- [38] R. Trivedi, M. Bahloul, A. Saif, S. Patra, and S. Khadem. Comprehensive dataset on electrical load profiles for energy community in Ireland. *Scientific Data*, 11:621, 2024. doi: 10.1038/s41597-024-03454-2.
- [39] Open Power System Data. Data package: Household data (CoSSMic). https://data.open-power-system-data.org/household_data/, 2020. Version 2020-04-15 (second release of all CoSSMic households).
- [40] D. Avgoloupis, V. Michalakopoulos, E. Sarvas, V. Marinakis, M. Couto, A. Coccia, and M. Jimenez. Residential energy dataset with electric vehicles, photovoltaic generation and tariff variability in Ireland. *Scientific Data*, 13:834, 2026. doi: 10.1038/s41597-026-07186-3. Data: <https://doi.org/10.6084/m9.figshare.30294478> (CC BY 4.0).
- [41] M. C. Peel, B. L. Finlayson, and T. A. McMahon. Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*, 11:1633–1644, 2007. doi: 10.5194/hess-11-1633-2007.
- [42] Patrick Zippenfenig. Open-Meteo.com weather API. Zenodo, 2023.
- [43] Hans Hersbach, Bill Bell, Paul Berrisford, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

- [44] Thomas Huld, Richard Müller, and Attilio Gambardella. A new solar radiation database for estimating PV performance in Europe and Africa. *Solar Energy*, 86(6):1803–1815, 2012. doi: 10.1016/j.solener.2012.03.006.
- [45] Jamie M. Bright and Nicholas A. Engerer. Engerer2: Global re-parameterisation, update, and validation of an irradiance separation model at different temporal resolutions. *Journal of Renewable and Sustainable Energy*, 11(3):033701, 2019. doi: 10.1063/1.5097014.
- [46] Agnès Bridel-Bertomeu, Michele Olios, André Mermoud, and Bruno Wittmer. Limits of the single diode model in view of its application to the latest PV cell technologies. In *Proc. 40th European Photovoltaic Solar Energy Conference*, Lisbon, Portugal, 2023.
- [47] Luca Antognini, Michele Olios, Auriane Canesse, Robin Vincent, André Mermoud, and Bruno Wittmer. A general approach to model high-performance PV modules for accurate energy yield simulations. In *Proc. 42nd European Photovoltaic Solar Energy Conference*, Bilbao, Spain, 2025.
- [48] Adrien Villos, Bruno Wittmer, André Mermoud, Michele Olios, and Agnès Bridel-Bertomeu. A model correcting the effect of sub-hourly irradiance fluctuations on overload clipping losses in hourly simulations. In *Proc. 8th World Conference on Photovoltaic Energy Conversion (WCPEC-8)*, Milan, Italy, 2022. doi: 10.4229/WCPEC-82022-4EO.2.2.
- [49] Agnès Bridel-Bertomeu, Michele Olios, André Mermoud, and Bruno Wittmer. Accounting for sub-hourly irradiance fluctuations in hourly performance simulations. In *Proc. 40th European Photovoltaic Solar Energy Conference*, Lisbon, Portugal, 2023.
- [50] Michele Olios, Bruno Wittmer, André Mermoud, Agnès Bridel-Bertomeu, and Robin Vincent. Implementation of a sub-hourly clipping correction in PVsyst. In *Proc. 41st European Photovoltaic Solar Energy Conference*, Vienna, Austria, 2024.
- [51] Dirk C. Jordan and Sarah R. Kurtz. Photovoltaic degradation rates — an analytical review. *Progress in Photovoltaics: Research and Applications*, 21(1):12–29, 2013. doi: 10.1002/pip.1182.
- [52] Didier Thevenard, Anton Driesse, Sophie Pelland, Dave Turcotte, and Yves Poissant. Uncertainty in long-term photovoltaic yield predictions. Technical Report 2010-122 (RP-TEC), CanmetENERGY, Natural Resources Canada, Varennes, Canada, 2010.
- [53] Manajit Sengupta, Aron Habte, Stefan Wilbert, Christian Gueymard, and Jan Remund. Best practices handbook for the collection and use of solar resource data for solar energy applications. Technical Report NREL/TP-5D00-77635, National Renewable Energy Laboratory, 2021.
- [54] ANSI C12.20-2015: Electricity meters — 0.1, 0.2, and 0.5 accuracy classes. American National Standards Institute / NEMA, Rosslyn, VA, 2015.
- [55] IEC 62053-22:2020 electricity metering equipment — particular requirements — part 22: Static meters for ac active energy (classes 0,1s, 0,2s and 0,5s). International Electrotechnical Commission, Geneva, 2020.