Balancing Robustness and Accuracy in Rooftop PV Estimation: Benchmarking the French TSO's Approach and Physics-Based Models

Gabriel Kasmi, Augustin Touron, Laurent Dubus, Yves-Marie Saint-Drenan, Philippe Blanc

Balancing Robustness and Accuracy in Rooftop PV Estimation: Benchmarking the French TSO's Approach and Physics-Based Models

Gabriel Kasmi[a,b], Augustin Touron[b], Laurent Dubus[b,c], Yves-Marie Saint-Drenan[a], Philippe Blanc[a]

[a] *Mines Paris, Université PSL, Centre Observation Impacts Energie (O.I.E.), Sophia Antipolis, 06904, France*
[b] *Direction de la Recherche et du Développement, RTE France, Paris La Défense, 92073, France*
[c] *WEMC (World Energy & Meteorology Council), Norwich, NR4 7TJ, United Kingdom*

Abstract

The deployment of photovoltaic (PV) systems accelerates worldwide and reached a global installed capacity reaching 1,858 GW_p in 2024, France accounting for 25 GW_p. However, PV grid integration is hindered by limited rooftop PV observability, as transmission system operators (TSOs) lack direct measurements of small-scale systems. This study provides the first national-scale benchmarking of the French TSO's statistical method against a physics-based approach using remote-sensing PV registries and weather data. Using ground-truth measurements from rooftop PV systems across France (2020-2023), we evaluate accuracy, robustness, and operational trade-offs of both methods. Results show comparable accuracy (aggregated errors of approximately 4%), with the TSO method demonstrating superior robustness and the physics-based approach offering greater adaptability. The TSO method performs better than expected under current conditions but faces challenges from evolving PV systems, as self-consumption installations now represent 60% of new capacity. Findings highlight the critical importance of accurate PV registries for grid reliability, with implications extending beyond France to other regions with rooftop PV integration challenges.

Keywords: Rooftop PV, Grid integration, PV estimation, TSO methods, Observability, Remote sensing

1. Introduction

The global installed capacity of photovoltaic (PV) systems has been growing exponentially to support the decarbonization of electricity systems (Haegel et al., 2017). In 2024, worldwide PV capacity reached 1,858 GW_p, with France accounting for 25 GW_p as of March 2025. Between 2023 and 2024, this represented growth rates of nearly 20% globally and 25% in France (International Renewable Energy Agency (IRENA), 2025; RTE France, 2025).

A key challenge accompanying this rapid expansion is the integration of rooftop PV electricity into power grids. Transmission system operators (TSOs) often lack direct measurements of rooftop PV production, particularly for installations below 36 kW_p—hereafter referred to as "rooftop PV" or "small-scale distributed PV systems"—which account for nearly 25% of France's total installed PV capacity (approximately 5 GW_p as of May 2025). This lack of accurate measurements, known as limited rooftop PV observability (Kasmi, 2024), can create significant operational risks for TSOs. As highlighted by RTE, the French transmission system operator, and the International Energy Agency (IEA): *"With current models and data, the solar PV contingency would dramatically increase during sunshine hours (especially at sunrise, when forecasts are less accurate), particularly at short-term horizons if generation forecasts are not improved"* (RTE France and IEA, 2021).

Improving rooftop PV observability is therefore a prerequisite for safe grid integration. This raises a central question: *How accurate are the estimation methods currently employed by TSOs for estimating aggregated rooftop PV production, and can alternative approaches improve observability at national scale?*

Grid integration of rooftop PV requires accurate knowledge of installed capacity at regional or national scales and the ability to reliably estimate and forecast rooftop PV power production (Kasmi, 2024; Huxley et al., 2022). Significant uncertainties persist in both installed capacity assessment and power estimation, despite recent methodological improvements (Pan et al., 2022b; Walch et al., 2021).

Regarding knowledge of underlying PV systems, recent advances in remote sensing and deep learning have enabled the creation of detailed rooftop PV registries. Many regions, particularly in Europe and the United States, have been mapped using airborne and satellite imagery (Yu et al., 2018; Kasmi et al., 2022; Arnaudo et al., 2023). These registries provide key system characteristics—location, installed capacity, tilt, and azimuth angles—that

can be combined with weather data and physical conversion models to estimate PV power production (Dobos, 2014).

Kasmi et al. (2024) recently demonstrated that such a physics-based approach, relying on the PVWatts model, can produce accurate estimates of individual rooftop PV power production. However, no comparison or evaluation of the TSO's approach at the country scale has been proposed.

The present study builds on the framework of Kasmi et al. (2024) to propose the first evaluation and benchmark of the French TSO's method for estimating rooftop PV power production, thereby contributing to the assessment of how accurate and reliable current methods are. We evaluate the TSO's method using a curated dataset of ground-truth power measurements from rooftop PV systems across France over the 2020–2023 period. Building on Kasmi et al. (2024), we benchmark the TSO's approach against a method that estimates individual power production of rooftop PV systems using sparse system characterization combined with solar irradiance and temperature data, all retrievable at the scale of a country or a TSO's control area.

Our central results show that the TSO's estimation method performs adequately under three specific conditions: (i) sufficient density of measured PV systems to avoid interpolation errors, (ii) accurate knowledge of installed capacity (the primary source of uncertainty), and (iii) minimal self-consumption or behavioral changes in distributed PV usage.

When condition (i) is not met, the physics-based method of Kasmi et al. (2024) offers a viable alternative, performing comparably to the TSO's approach without sensitivity to interpolation errors. Condition (ii) shifts uncertainty from estimation methods to fleet characterization, highlighting the importance of accurate PV registries. Condition (iii) is already changing, as self-consumption systems now represent 60% of rooftop PV systems, raising immediate questions about how evolving consumption patterns will affect estimation accuracy. On this latter point, accurate individual estimates of rooftop PV power production can serve as inputs to more sophisticated models that account for consumption patterns and behavioral changes.

Contributions. This work offers the following contributions: First, we present the first national-scale evaluation of the French TSO's rooftop PV power estimation methodology against ground-truth data for small-scale PV systems. Second, we quantify the accuracy of the TSO's method and compare it with a physics-based approach leveraging remote-sensing-derived rooftop PV reg-

istries and weather data. Third, we demonstrate that the physics-based method of Kasmi et al. (2024) slightly outperforms the TSO's approach at the individual system level, albeit with lower robustness to errors in individual system characterization. Finally, we discuss the implications of these findings for rooftop PV observability and grid integration strategies beyond the French use case.

2. Background and related works

2.1. Background: PV observability in France

Definition and overview. Kasmi (2024) defines observability as "*the ability of the TSO to accurately estimate a power unit's real-time and future production.*" This definition encompasses three distinct levels of data availability that determine how the TSO, in our case RTE, monitors and estimates power production.

At the highest level, RTE has direct access to real-time measurements of power production. This applies to systems connected to the transmission network, including conventional power plants, approximately 80% of wind power plants, and a limited number of PV plants (24 systems representing 869 MW_p as of May 2025). When real-time data is unavailable, RTE relies on ex-post measurements—actual production data typically available 48 hours after generation occurs. These delayed measurements serve as ground truth for calibrating power estimation models that provide real-time production estimates for similar systems. At the lowest level are systems with no measurements at all, termed "not observable." For these installations, RTE cannot assess the accuracy of estimation and forecasting methods, creating significant operational uncertainty.

Therefore, in practice, observability translates to a simple operational criterion: a system is observable if and only if RTE has access to either real-time or ex-post measurements of its actual production. Table 1 illustrates how PV system observability in France varies dramatically by installation size. Until recently, most unobserved systems were small installations below 36 kW_p (Kasmi et al., 2024). However, the situation has deteriorated significantly: by May 2025, approximately 50% of systems between 250–1000 kW_p and 35% of systems above 1000 kW_p remain unobserved. This trend likely reflects the challenge of deploying metering infrastructure at scale—in periods of rapid market growth, new installations are commissioned faster than monitoring capabilities can be implemented.

Table 1: PV installation observability by power class (May 2025). Percentages in italics. TN: transmission network; DN: distribution network. **Bold values** highlight shares of unobserved PV capacity and installations. Source: RTE.

Power class	Observed		Not observed	
	Installed capacity [MW_p]	Number of installations [-]	Installed capacity [MW_p]	Number of installations kW_p¿kW_p [-]
≤ 36	54.4	2,254	4816.6	1,002,962
(%)	*1.1*	*0.2*	*98.9*	*99.8*
36-250	7707.8	60,278	801.6	6,325
(%)	*90.6*	*90.5*	*9.4*	*9.5*
250-1000	575.3	1,045	658.9	1,119
(%)	*46.6*	*48.3*	*53.4*	*51.7*
≥ 1000 (DN)	9998.3	1,950	5553.6	1,651
(%)	*64.3*	*54.2*	*35.7*	*45.8*
≥ 1000 (TN)	869	24	0	0
(%)	*100*	*100*	*0*	*0*
Total	19162.8	65,527	11,830.7	1,012,057
(%)	*75.3*	*6.1*	***38.2***	***93.9***

Operational and economic implications of poor observability. The implications of limited observability are substantial and multifaceted. Observability underpins all TSO operational tasks, from real-time grid management to forecasting, as RTE's models require historical production data for training and validation. For unobservable systems, the TSO must rely on statistical estimation methods (detailed in Section 4.2.3); however, without ground-truth measurements, the accuracy of these methods cannot be assessed. This lack of validation introduces significant uncertainty into operational decision-making.

The operational implications extend beyond technical accuracy to measurable economic consequences. While TSOs can provide real-time estimates of unobserved PV production, the significant uncertainties associated with these estimates can lead to suboptimal operational decisions and inefficient resource allocation. The cost-causation principle established by Milligan et al. (2011)—that "*those individuals who cause costs to the system should pay for those costs*"—becomes difficult to implement when estimation uncertainties are large. High uncertainty in individual system production estimates prevents precise attribution of grid impacts to specific Balance Responsible Parties (BRPs), forcing the TSO to absorb imbalance costs that cannot be

accurately traced to their source. As the "balancing authority of last resort," the TSO ultimately bears the financial burden of system imbalances when uncertainties prevent proper cost attribution, leading to inefficient risk allocation and socializing costs that should be borne by individual market participants.

This deteriorating situation has grown more acute over time: the non-observed installed capacity increased from 3,771 to 11,831 MW_p between September 2023 and May 2025, representing growth from 22% to 38% of total installed capacity (Kasmi et al., 2024). As highlighted in RTE France and IEA (2021), this trend raises concerns at the national level, with warnings that continued poor observability could significantly increase costs associated with grid imbalances and redispatch measures.

2.2. Regional TSO approaches to PV estimation

To address observability challenges, transmission system operators have developed operational estimation methods that can be broadly classified into two categories: statistical approaches that exploit spatial correlations between systems, and physics-based approaches that combine meteorological data with PV system models. These TSO methods, which form the basis of our evaluation, prioritize operational reliability and computational efficiency over methodological sophistication.

Statistical upscaling methods. The upscaling approach (Schierenbeck et al., 2010; Lorenz et al., 2011) represents the traditional method for PV production estimation in the absence of telemetry. This technique involves estimating a capacity factor—the ratio between actual power output and rated capacity—based on the production of nearby metered reference plants, then applying this capacity factor to unmetered plants for which only installed capacity is known.

A key limitation is that reference plants may not share the same technical characteristics as target plants, leading to representativeness errors (Saint-Drenan et al., 2016). These errors become particularly problematic when meteorological conditions vary spatially or when system configurations differ significantly between reference and target installations. For distributed rooftop PV systems, this representativeness challenge is amplified by the heterogeneous nature of installations, which vary significantly in technology, orientation, shading conditions, and maintenance practices compared to the utility-scale reference plants typically used for upscaling.

Weather-based physics approaches. Physics-based regional methods combine meteorological data with physical PV models to estimate regional production without requiring detailed plant characteristics. This approach has gained attention in the TSO community due to its ability to leverage numerical weather prediction (NWP) models and satellite-based irradiance data.

Saint-Drenan et al. (2015) proposed a probabilistic approach based on physical PV system models and solar irradiance data, requiring only basic parameters—location, tilt and azimuth angles, and nameplate capacity. Building on this framework, Saint-Drenan et al. (2018) developed a comprehensive approach that couples physical PV models with statistical distributions of plant configurations, achieving relative RMSE of 3.8-4.2% and biases of -2.4 to 0.1% when validated against TSO data from France and Germany at three-hourly resolution.

The weather-based approach leverages satellite-based remote sensing for irradiance data (5-15 minute resolution) and NWP models for meteorological variables at hourly resolution (Qu et al., 2017).

Critical limitations of regional approaches. Despite strong aggregate performance metrics (3.8-5% RMSE), regional TSO methods reveal fundamental limitations when applied to distributed rooftop PV systems that actually contribute to, rather than resolve, observability challenges.

Validation methodologies in this field typically compare regional methods against TSO aggregate data, effectively using the TSO as ground truth. However, as demonstrated in Table 1, TSOs themselves rely on estimation methods for 38% of installed capacity, meaning the "ground truth" used for validation is itself uncertain. This study addresses this validation challenge by using independent ground truth measurements, avoiding the circular dependency on TSO-estimated data for validation purposes.

The scaling problem is particularly acute for rooftop PV systems. Unlike utility-scale installations where individual plant characteristics can be surveyed and monitored, rooftop PV presents a "many-small-systems" challenge where individual system diversity cannot be captured by traditional regional statistical approaches. Weather-based methods partially address this through physical modeling, and probabilistic approaches like those developed by Saint-Drenan (2016) attempt to account for uncertainty in system configurations through statistical distributions of plant parameters. However, the reliance on assumed distributions of "typical system configurations" may still not fully reflect the actual heterogeneity of system characteristics in real

deployments, particularly as rooftop PV adoption expands into more diverse installation contexts.

2.3. Global challenges in distributed PV estimation

While TSOs have developed regional operational approaches, the broader research community has pursued alternative methodologies to address similar observability challenges worldwide, primarily focusing on distributed behind-the-meter systems. This research has followed a parallel methodological dichotomy: statistical disaggregation approaches that exploit consumption patterns, and physics-based methods that combine system characteristics with meteorological data.

Scale and methodological diversity of the invisible PV problem. The observability challenge extends globally, affecting grid operators across diverse markets and regulatory frameworks. Shaker et al. (2015) systematically characterized this "invisible solar" phenomenon, documenting over 246,000 unmonitored small-scale distributed PV projects totaling 2,374 MW_p in California alone by 2015. Australian projections suggest distributed energy resources could provide 13-22% of total electricity consumption by 2040, predominantly through unmonitored behind-the-meter systems (Pan et al., 2022a).

Researchers have developed numerous approaches for estimating behind-the-meter PV generation, broadly classified into data-driven and model-based methods (Wu et al., 2022). Data-driven approaches subdivide by supervision requirements: supervised methods utilize historical measurements from representative sites, semi-supervised approaches leverage partial labeling, while unsupervised methods attempt to extract PV generation using only net load measurements and exogenous variables.

Recent advances include federated learning for privacy preservation (Lin et al., 2022), deep learning architectures incorporating spatiotemporal correlations (Pan et al., 2022a), and neural network approaches achieving mean absolute errors below 6.2% under full observability and 8.6% under partial observability scenarios (Cheung et al., 2019).

Fundamental constraints of disaggregation approaches. Despite methodological sophistication, invisible PV research faces critical limitations that constrain practical deployment. Most approaches focus on aggregated estimation at feeder or regional levels, providing limited insight into individual system performance needed for detailed grid planning. These methods increasingly

struggle with evolving distributed energy systems, particularly solar-plus-storage deployments that smooth out fluctuations typically exploited for disaggregation (Chen and Ardakanian, 2022).

Their reliance on statistical patterns makes them vulnerable to changes in consumption behavior and technology adoption that can invalidate historical relationships. Most methods require substantial training data or representative proxies, limiting applicability in data-sparse regions or for emerging system configurations. Validation challenges are compounded by varying temporal scopes, geographic scales, and performance metrics across studies, preventing standardized performance comparisons.

Individual system-level approaches have begun emerging to address these constraints. Studies have analyzed systems ranging from 15 Swiss homes (Walch et al., 2021) to 740 Australian systems (de Hoog et al., 2021), but validation reveals significant accuracy challenges. Walch et al. (2021) demonstrated that while PV area detection can be accurate, hourly production estimation suffers from systematic overestimation, with mean annual errors of 16% and higher errors in winter conditions.

2.4. Physics-based individual system estimation

The limitations of both aggregate invisible PV methods and traditional TSO approaches have motivated research into individual system-level estimation techniques that combine operational practicality with methodological rigor.

Remote sensing for system characterization. Tracking individual rooftop PV systems poses significant challenges for system operators due to their decentralized nature. While some countries maintain public registries, uncertainties regarding data accuracy persist (Huxley et al., 2022). Remote sensing approaches using aerial or satellite imagery have emerged as promising methods for reducing uncertainty in rooftop PV capacity estimation, with applications demonstrated across multiple countries (Yu et al., 2018; Mayer et al., 2022; Kasmi et al., 2022).

Advanced approaches construct technical registries containing system characteristics such as location, capacity, tilt, and azimuth angles that can be combined with weather data and physical conversion models to estimate individual PV power production.

Validation challenges and research gaps. Kasmi et al. (2024) demonstrated that individual rooftop PV production can be accurately estimated using minimal technical parameters derived from remote sensing combined with weather data and physical conversion models, achieving approximately 10% pRMSE for individual system estimation. This physics-based approach addresses several limitations of existing methods by providing system-level estimates, relying on physical principles rather than statistical patterns, and requiring only geometric and location data obtainable through remote sensing.

However, this work did not benchmark physics-based individual estimation methods against statistical upscaling approaches currently used by TSOs, leaving critical questions about their relative accuracy and operational viability unanswered. The literature reveals a consistent methodological pattern where statistical approaches compete with physics-based approaches across spatial scales, but direct performance comparisons are complicated by varying validation scales, temporal resolutions, and geographic contexts.

Research positioning and contribution. To our knowledge, this study provides the first systematic comparison between TSO statistical methods and physics-based individual estimation using ground-truth power measurements for rooftop PV systems at national scale. This comparison is unprecedented in its geographical coverage (entire France), temporal granularity (hourly resolution), and integration of remote-sensing-derived technical data.

Our research addresses three critical gaps: First, the lack of head-to-head comparison between statistical upscaling (the current TSO standard) and physics-based individual estimation under identical validation conditions. Second, the absence of national-scale validation studies that account for the full heterogeneity of system types, locations, and performance characteristics present in real deployments. Third, the limited understanding of how estimation method performance scales from regional aggregation to individual system accuracy—a critical consideration as grid management becomes increasingly distributed.

By bridging operational TSO practices with emerging research in individual rooftop PV estimation and breaking the circular validation problem through independent ground-truth data, this study aims to inform evidence-based decisions about observability investment priorities and method selection for different operational contexts.

3. Data

This study builds upon the dataset introduced in Kasmi et al. (2024), extending it with operational TSO measurements to enable national-scale benchmarking of PV estimation methods. The integration of production measurements from the French Transmission System Operator (TSO), RTE, with independent ground truth measurements provides a comprehensive validation of operational TSO estimation methods against external data sources at national scale.

3.1. Ground Truth Data: Individual Rooftop PV Systems

3.1.1. Dataset characteristics

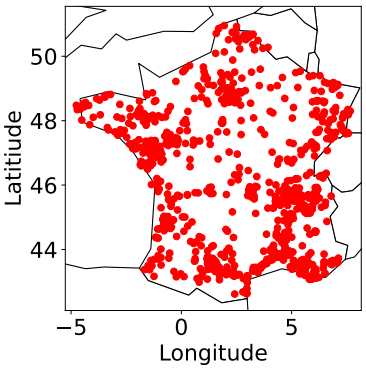
High-resolution power output data from individual rooftop PV installations across mainland France were obtained from the nonprofit association *Asso BDPV*. The installations form a subset of the BDAPPV dataset (Kasmi et al., 2023), with power values recorded at 30-minute intervals for grid-connected systems.

The systems are distributed across mainland France (Figure 1a) with diverse technical characteristics (Figure 2), providing comprehensive coverage of French rooftop PV deployment patterns. The temporal coverage spans multiple seasons and weather conditions, capturing operational variability essential for robust estimation method validation.

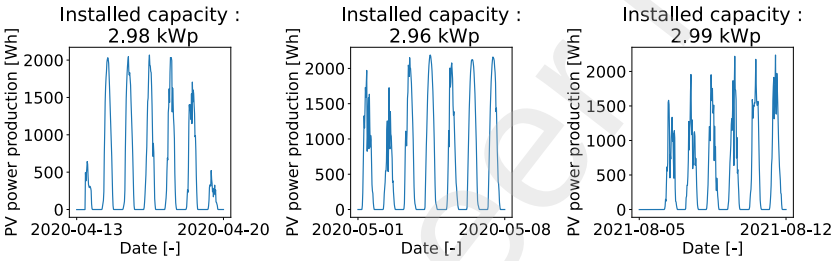
Regarding self-consumption, all systems in our dataset represent total PV production rather than net grid injection. The measurement infrastructure connects directly to the PV inverter output, capturing gross generation before any local consumption. This configuration was verified through manual inspection of power production curves, which consistently align with physics-based simulations that model gross generation without consumption offsets. Therefore, our ground truth data represents the actual PV production that estimation methods aim to predict, eliminating potential bias from unknown consumption patterns.

3.1.2. System representativeness

The 906 systems provide national geographic coverage with characteristics representative of French residential PV deployment. System capacities range from 1.2 to 9.8 kWp, with an average of 3.5 kW_p, consistent with residential rooftop installations that dominate the unobserved PV population identified in Section 2.1.



(a) Spatial distribution of metered PV systems from BDPV.



(b) Individual 30-minute power output time series from representative installations.

Figure 1: Ground truth data overview: (**a**) spatial distribution of rooftop PV systems with power measurements, and (**b**) sample power output time series.

While the volunteer-contributed nature of our dataset may introduce selection bias toward more engaged system owners, this does not necessarily indicate better-maintained systems. Rather, our sample likely represents "solar enthusiasts" willing to share performance data, which may include both high and low-performing installations. Importantly, our systems exhibit representative characteristics in terms of geographic distribution, capacity range (1.2-9.8 kW_p, average 3.5 kW_p), and technical configurations when compared to the broader French residential PV population. The 255-system subset with both production and remote-sensing data represents a random subsample, further supporting the validity of our findings for typical rooftop installations.

3.2. System Technical Parameters

Physics-based estimation requires three fundamental system parameters: installed capacity, tilt angle, and azimuth angle. Two complementary data

sources provide these parameters for comprehensive method evaluation.

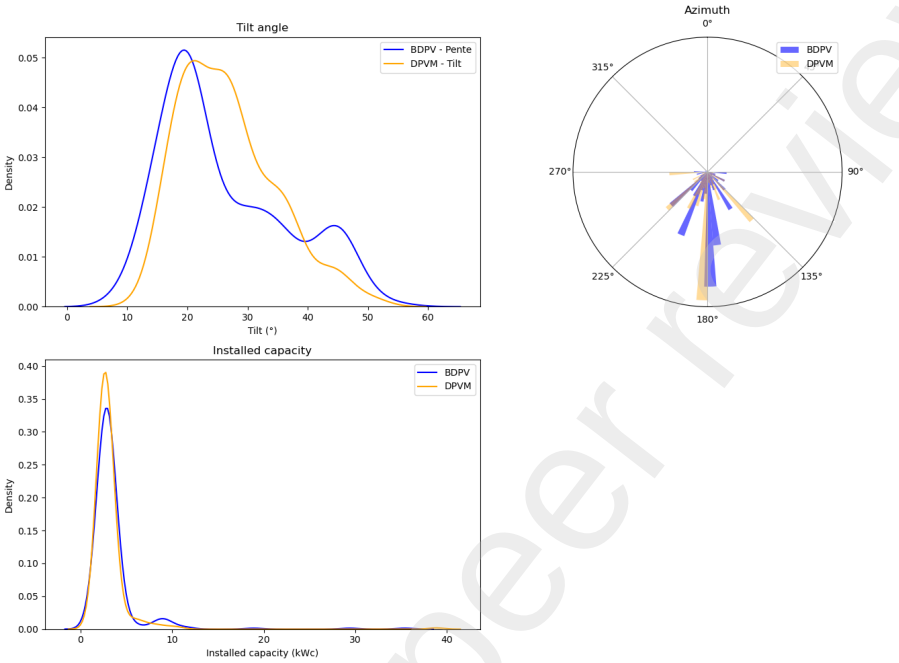


Figure 2: System parameter distributions: ground truth (BDPV, blue) versus remote sensing estimates (DPVM, orange). Differences quantify the accuracy of automated parameter extraction for operational applications.

Ground truth technical registry (BDPV). Self-reported system parameters from BDPV provide ground truth values for installed capacity, tilt, and azimuth angles. These owner-reported specifications represent the most accurate available characterization of actual system configurations for validation purposes.

Remote sensing registry (DPVM). The DeepPVMapper (DPVM) registry Kasmi et al. (2022) provides automated parameter extraction from aerial imagery for over 400,000 installations. System capacity estimates derive from detected panel areas using linear power density assumptions, while tilt and azimuth angles are estimated using the Python library `PyPVRoof` (Trémenbert et al., 2023).

The intersection of 1,485 DPVM installations with 906 validated BDPV systems yields 255 systems with both ground truth power measurements and

remotely-derived parameters. These systems maintain broad geographic distribution and preserve variability in orientation and capacity characteristics required for national benchmarking.

Figure 2 compares parameter distributions between ground truth (BDPV) and remote sensing estimates (DPVM). DPVM systematically underestimates tilt angles and shows discretization effects in azimuth estimates, reflecting automated extraction limitations. The average capacity is 3.5 kW_p, representative of French residential installations.

3.3. Meteorological Data

Physics-based PV estimation requires high-quality meteorological inputs synchronized with power measurements. Two data sources provide comprehensive environmental coverage.

Solar irradiance. Solar irradiance data from the Copernicus Atmospheric Monitoring Service (CAMS) (Qu et al., 2017) provide 15-minute historical values at 3 km resolution across France. The dataset includes global, direct, and diffuse horizontal irradiance components derived from MSG SEVIRI satellite imagery and CAMS aerosol forecasts. Values were aggregated to 30-minute intervals and spatially interpolated to PV system locations using bilinear interpolation.

Temperature. Ambient air temperature at 2 m height from ERA5 reanalysis (Hersbach et al., 2020) provides hourly values at 31 km resolution. Temperature data were linearly interpolated to 30-minute resolution and spatially matched to maintain synchronization with irradiance and power measurements.

3.4. TSO Operational Data

Integration of operational measurements from RTE enables independent validation of TSO estimation methods using comprehensive ground truth data, addressing limitations of previous validation approaches that rely on TSO aggregate data.

3.4.1. Observable fleet characteristics

As of May 2025, RTE's registry contains 1,107,584 grid-connected PV systems, with 6.1% classified as observable (Table 1). From this observable fleet, 10-minute power measurements for 68,000 systems covering 2020-2023 provide operational data for upscaling method evaluation. Measurements

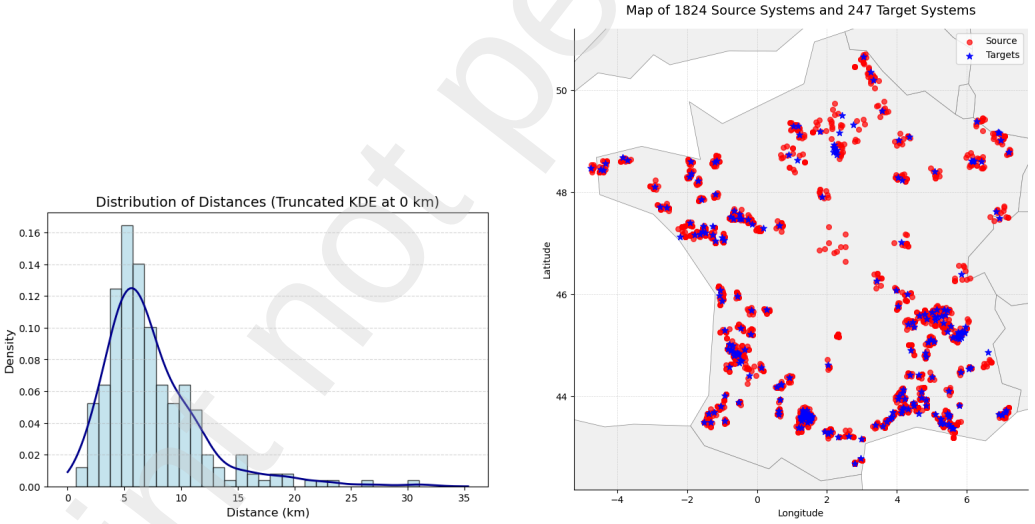
were converted to 30-minute energy values through integration and resampled to match ground truth temporal resolution.

3.4.2. Reference system selection

Implementation of RTE's spatial neighbor selection algorithm for 255 target systems identified 1,824 unique reference systems. These systems exhibit mean capacity of 283 kW_p (80 times larger than target rooftop systems) with average spatial separation of 7.44 km (median: 6.39 km, maximum: 30.97 km).

Spatial distances fall within single CAMS pixels (11 km × 8 km), where solar irradiance remains constant (Saint-Drenan et al., 2016), validating meteorological consistency assumptions. However, the substantial capacity difference (283 kW_p vs. 3.5 kW_p) demonstrates the scale mismatch inherent in current TSO upscaling methodology.

Figure 3 shows distance distributions and geographic locations, demonstrating national validation coverage.



(a) Distance distribution between RTE reference and target systems.

(b) Geographic distribution of reference (red) and target (blue) systems.

Figure 3: TSO upscaling validation setup: spatial relationships between reference and target systems enable national-scale evaluation of operational estimation methods.

3.4.3. Validation methodology

The integration of operational TSO data with ground truth measurements enables independent assessment of estimation method accuracy. TSO performance is evaluated using RTE reference data and operational algorithms, while physics-based methods use meteorological inputs and remote sensing parameters. Both methods are validated against identical ground truth targets, enabling direct performance comparison.

This approach provides independent validation of TSO estimation accuracy using ground truth data, addressing limitations of previous studies that validate TSO methods against TSO aggregate data. We provide the first comprehensive validation for the French context, which has received limited academic attention despite France's position as a major European PV market.

4. Methods

4.1. PV Power Estimation Methods

This section presents two methodological approaches for estimating individual rooftop PV power output: statistical methods that extrapolate from nearby metered systems, and physics-based methods that simulate PV production using environmental and system characteristics. The comparison enables assessment of operational accuracy and scalability for grid applications.

Statistical approaches include the classic upscaling method (Schierenbeck et al., 2010; Lorenz et al., 2011) and RTE's operational implementation. Physics-based models utilize simplified physical relationships and are particularly suited for applications with detailed system metadata from PV registries Kasmi et al. (2024).

4.2. Statistical Upscaling Methods

4.2.1. Upscaling framework

The upscaling approach infers normalized power output of an unmonitored system k at time t, denoted $p^{(k)}_{norm,t}$, using a weighted combination of normalized outputs from n nearby metered systems:

$$p^{(k)}_{norm,t} = \sum_{i=1}^{n} w_i p^{(i)}_{norm,t} \quad \text{where} \quad \sum_{i=1}^{n} w_i = 1 \tag{1}$$

Weights w_i are derived from spatial proximity, exploiting meteorological correlation between geographically close systems.

4.2.2. Distance-weighted upscaling

The baseline upscaling implementation assigns weights inversely proportional to geographic distance:

$$p_{norm,t}^{(k)} = \sum_{i=1}^{n} \frac{d(k,i)^{-p}}{\sum_{i=1}^{n} d(k,i)^{-p}} p_{norm,t}^{(i)}, \quad p > 0 \tag{2}$$

where $d(k,i)$ is the geographic distance between systems k and i. This approach, denoted **Upscaling**, provides continuous weighting based on spatial correlation.

4.2.3. RTE operational method

RTE implements median-based upscaling for operational PV estimation. The method selects the $n = 10$ nearest metered systems and computes the median normalized power:

$$p_{norm,t}^{(k)} = \underset{i \in \{1,\ldots,n\}}{\text{median}} \left\{ p_{norm,t}^{(i)} \;\middle|\; w(i,k) = 1 \right\} \tag{3}$$

All neighbors receive equal weighting. Absolute power is recovered using installed capacity:

$$p_{PV,t}^{(k)} = p_{inst}^{(k)} \times p_{norm,t}^{(k)} \tag{4}$$

This method, denoted **RTE**, prioritizes operational robustness over spatial optimization.

It should be noted that RTE operationally uses monthly energy aggregation from distribution system operators rather than installed capacity for power curve fitting. For consistency with physics-based methods and to enable direct comparison, we implemented the TSO method using installed capacity. Appendix B.2 demonstrates that this implementation choice does not substantially affect performance metrics, though it represents a limitation in directly replicating operational conditions. This methodological adaptation allows for controlled comparison while maintaining the essential characteristics of the TSO approach.

4.3. Physics-Based Estimation

4.3.1. PVWatts model implementation

The physics-based approach uses the PVWatts model (Dobos, 2014), selected for low data requirements compatible with remote sensing applications. DC power output is estimated as:

$$p_{DC,t} = \frac{POA_{eff}(\theta, \phi)}{G_{stc}} \cdot P_{PV} \cdot (1 + \gamma_{pdc}(T_{module,t} - T_{stc})) \tag{5}$$

where POA_{eff} is effective plane-of-array irradiance, $T_{module,t}$ is module temperature, P_{PV} is installed capacity, $\gamma_{pdc} = -0.002$ K^{-1} is the temperature coefficient, and $G_{stc} = 1000$ Wm^{-2} is standard irradiance.

AC power conversion uses load-dependent inverter efficiency:

$$p_{AC,t} = p_{DC,t} \times \eta_{inv}(p_{DC,t}) \tag{6}$$

where $\eta_{inv}(p_{DC,t})$ represents the inverter efficiency as a function of DC power loading, following the default parameterization provided by Dobos (2014).

4.3.2. System parameterization strategies

The physics-based model, denoted **PVLib**, is evaluated with multiple parameterization strategies for tilt angle θ and azimuth ϕ:

- **BDPV**: Ground truth parameters from system owner reports
- **DPVM**: Remote sensing estimates from DeepPVMapper Kasmi et al. (2022)
- **Constant**: Fixed assumptions ($\theta = 30°, \phi = 180°$)
- **Naive**: Horizontal installation assumption ($\theta = 0°$)
- **Oracle**: Parameters optimized against ground truth power measurements (performance upper bound)

These parameterizations assess sensitivity to system characteristic accuracy and provide operational deployment scenarios. When referenced without specification, **PVLib** refers to the **BDPV** parameterization.

4.3.3. Method comparison example

Figure 4 illustrates power output estimation from statistical and physics-based methods for a representative rooftop system. The physics-based approach captures diurnal patterns through solar irradiance modeling, while statistical methods reflect averaged behavior from reference systems.

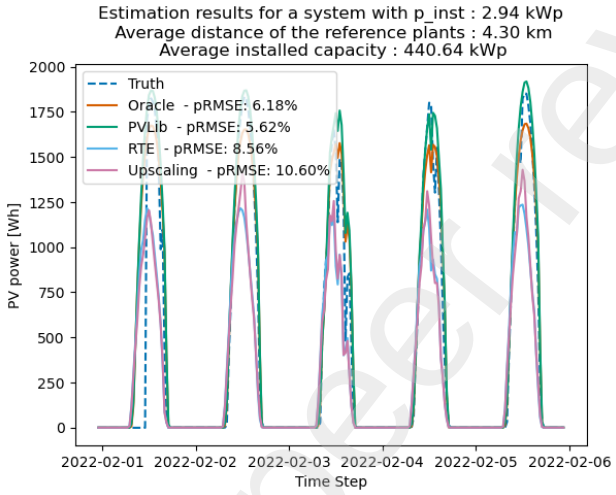


Figure 4: Power output estimation comparison for a rooftop PV system: RTE (statistical) versus physics-based methods demonstrate different temporal characteristics and estimation accuracy.

4.4. Performance Evaluation Framework

The evaluation quantifies estimation accuracy and assesses operational scalability for grid applications through comprehensive error analysis and aggregation effects.

4.4.1. Performance metrics

Root mean squared error (RMSE) serves as the primary accuracy metric:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2} \tag{7}$$

where $\hat{y}_t$ and y_t are estimated and measured power outputs. Percentage RMSE (pRMSE) normalizes by installed capacity $p_{PV,j}$:

$$pRMSE_j = \frac{RMSE_j}{p_{PV,j}} \times 100 \tag{8}$$

Mean bias error (MBE) quantifies systematic estimation bias:

$$MBE = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t) \tag{9}$$

Positive MBE indicates overestimation; negative values indicate underestimation.

4.4.2. Operational analysis framework

Temporal error characteristics. Error analysis examines patterns across time-of-day intervals (30-minute resolution, 48 daily intervals) and seasonal conditions (summer: June–September; winter: December–March). This temporal decomposition identifies operational conditions where estimation methods perform optimally or exhibit systematic errors.

System aggregation effects. Accuracy improvement through system aggregation is quantified using:

$$RMSE_{agg} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sigma_0^2(t)}{n} + \mu_0^2(t)\right)} \tag{10}$$

from Saint-Drenan et al. (2016), where $\sigma_0(t)$ is error standard deviation and $\mu_0(t)$ is mean bias across n systems. This analysis quantifies operational benefits of regional versus individual system estimation.

Error source decomposition. Following Saint-Drenan (2016), estimation errors are categorized as interpolation errors and representativeness errors. Interpolation errors correspond to spatial meteorological mismatches between reference and target systems. Representativeness errors, on the other hand, arise from system characteristic differences (rooftop versus utility-scale configurations).

4.4.3. Implementation validation

Data preprocessing excludes nighttime periods (zero production) to focus on operational generation periods. Timestamp alignment resolves timezone inconsistencies across data sources. RTE method implementation validation

ensures consistency with operational procedures. Statistical method error profiles are verified against established literature benchmarks.

The evaluation framework provides comprehensive assessment of method accuracy, operational scalability, and deployment requirements for grid applications.

5. Results and Analysis

5.1. Comparative Performance of TSO and Physics-Based Methods

The evaluation demonstrates that physics-based models achieve marginally superior accuracy compared to RTE's operational method when system parameters are precisely known, but exhibit significantly greater sensitivity to parameter uncertainty. RTE maintains consistent performance across diverse system configurations, indicating superior operational robustness for grid applications. The analysis provides quantitative benchmarks for TSO operational methods and identifies practical trade-offs for large-scale deployment.

Table 2: Performance comparison of TSO operational method (**RTE**), baseline **Upscaling**, physics-based method (**PVLib**) with different parameterizations, and theoretical optimum (**Oracle**). **BDPV**: ground truth parameters; **DPVM**: remote sensing parameters; **Constant**: fixed assumptions (30° tilt, 180° azimuth); **Naive**: horizontal assumption. Best performance **bold**, second-best underlined.

			Error (pRMSE) [%] (↓)				**Bias (MBE) [W] (→ 0)**			
			Mean	Min	Max	Median	Mean	Min	Max	Median
Physical		Oracle	7.23	4.36	23.57	6.54	32.11	-646.1	544.35	24.59
	PVLib	BDPV	**7.74**	**3.65**	26.41	**6.98**	141.33	-632.05	683.89	117.54
		DPVM	10.31	4.02	49.94	8.25	**41.29**	-2075.66	1494.28	**24.81**
		Constant	9.30	4.47	39.75	8.39	-78.75	-2578.18	1161.28	-73.83
		Naive	11.17	5.07	54.25	9.21	105.29	-1991.58	1623.48	99.28
Statistical		RTE	7.98	5.02	**16.65**	7.68	51.34	**-417.33**	**558.35**	49.0
		Upscaling	8.32	4.64	17.99	7.79	56.64	-570.22	778.73	50.35

5.1.1. Accuracy and bias performance

Table 2 presents comprehensive performance metrics for all estimation methods. Physics-based approaches outperform statistical methods in mean

error and bias under optimal parameter conditions. PVLib with ground truth parameters (BDPV) achieves the lowest mean pRMSE (7.74%), closely approaching the Oracle model's theoretical minimum (7.23%). However, this advantage deteriorates with parameter uncertainty: PVLib with fixed assumptions (Constant) increases to 9.30% pRMSE, while incorrect assumptions can produce maximum errors up to 54.25%.

RTE maintains stable accuracy (7.98% mean pRMSE) with substantially lower maximum errors (16.65%), demonstrating operational robustness. The statistical significance of performance differences is established through comprehensive testing detailed in Appendix C.5.

5.1.2. Operational robustness assessment

Statistical methods exhibit reduced sensitivity to extreme system configurations and metadata inaccuracies. RTE's bias range (-417.33 W to 558.35 W) demonstrates controlled variability, while PVLib with remote sensing parameters (DPVM) shows broader bias distribution (-2075.66 W to 1494.28 W). Although DPVM achieves the lowest mean bias (41.29 W), this variability indicates substantial sensitivity to parameter accuracy.

Physics-based methods require precise system characterization (tilt, azimuth, installed capacity) to exceed statistical baselines. Statistical approaches compensate for unknown system details through population-level patterns, enhancing robustness in data-limited operational environments.

5.1.3. Performance implications for grid operations

RTE's method performs comparably to physics-based models while maintaining superior parameter robustness. Comparison of remote sensing (DPVM) versus fixed parameter (Constant) approaches reveals minimal accuracy gains: median errors of 8.25% versus 8.39%. This indicates that current automated parameter extraction provides marginal operational benefits, suggesting that enhanced remote sensing capabilities are required for physics-based methods to consistently outperform statistical approaches.

Statistical methods assume low representativeness error between reference and target systems. Analysis in Appendix C estimates this contributes 3–4 percentage points (approximately half of total observed error of 8%). Removing this representativeness penalty would yield performance around 4%, aligning with literature benchmarks of 3.8-4.2% under optimal conditions (Saint-Drenan et al., 2018). Increasing system diversity suggests this representativeness gap may widen, potentially degrading statistical method

performance over time. Notably, our implementation of the upscaling and RTE methods achieves performance metrics consistent with those reported in similar validation studies, confirming the reliability of our methodological approach.

5.2. Spatial and Temporal Error Characteristics

Comprehensive error analysis across geographic, temporal, and aggregation scales provides operational insights for TSO deployment and identifies system-level performance characteristics relevant to grid management applications.

5.2.1. Geographic error distribution

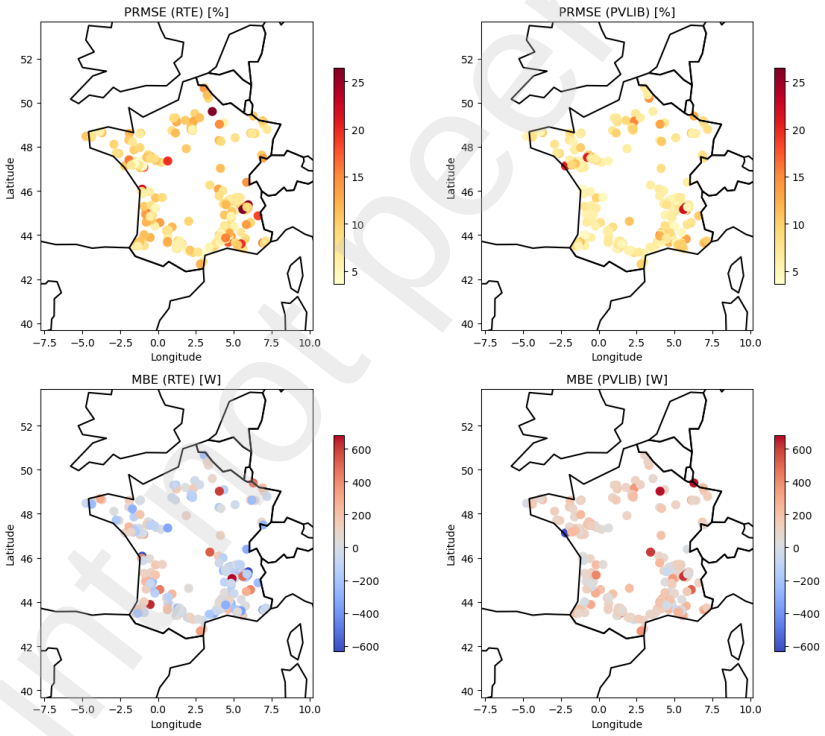


Figure 5: Spatial error distributions for RTE (left) and PVLib (right): estimation errors (top) and systematic bias (bottom). Absence of clear geographic patterns demonstrates consistent national-scale performance.

Figure 5 presents spatial distributions of estimation errors (pRMSE) and biases (MBE) across measurement locations. Neither RTE nor PVLib exhibit

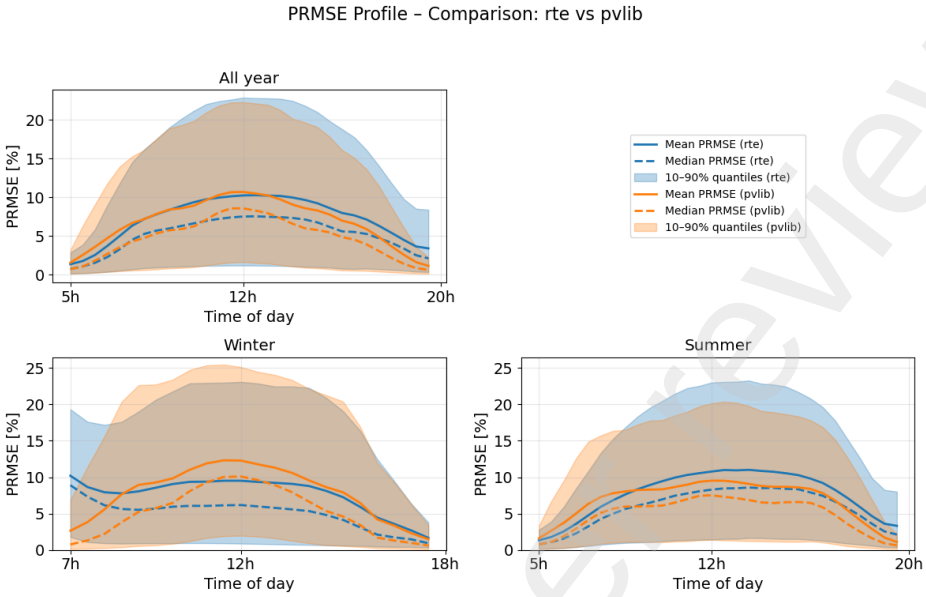
systematic geographic error patterns, demonstrating consistent performance across diverse locations and climatic conditions. This spatial uniformity indicates that large-scale deployment will not produce systematic regional estimation bias, supporting national-scale operational implementation.

5.2.2. Temporal performance analysis

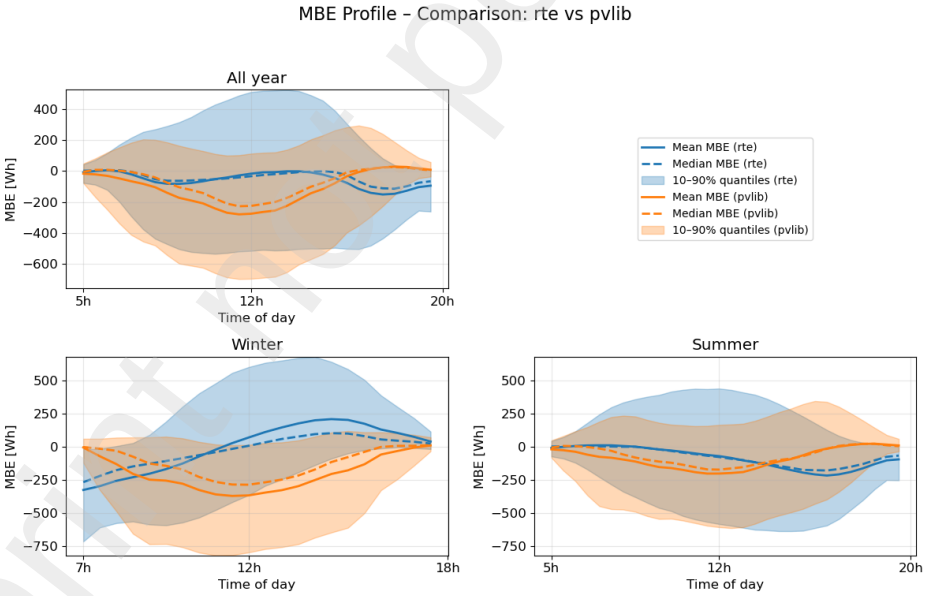
Figure 6 presents temporal decomposition of estimation errors and biases across seasonal and diurnal cycles. RTE and Upscaling exhibit minimal systematic bias due to compensating upward and downward errors, reflecting representativeness differences between reference systems (utility-scale, optimized configuration) and target systems (rooftop installations with suboptimal orientations).

PVLib demonstrates systematic negative bias, particularly during winter conditions, consistent with sensitivity to inaccurate tilt and azimuth parameters. This underestimation pattern suggests physics-based methods are sensitive to the southwest orientations typical of residential installations.

The temporal analysis reveals operational distinctions: RTE and Upscaling provide reliable daily energy yield estimates but limited intra-day production profile accuracy. PVLib offers superior temporal resolution but requires precise system metadata to avoid systematic bias. This trade-off has implications for different TSO operational requirements.



(a) Estimation error (pRMSE) temporal patterns



(b) Systematic bias (MBE) temporal patterns

Figure 6: Temporal error characteristics across seasons and daily cycles: RTE maintains consistent performance while PVLib shows systematic winter underestimation.

5.2.3. System aggregation performance

Figure 7 demonstrates error reduction with increasing system aggregation. Both methodological approaches benefit from aggregation effects, with errors decreasing and converging as system count increases. This validates prior findings for physics-based methods and confirms that aggregation mitigates individual system parameter uncertainties.

For TSO operations, this demonstrates that both approaches are suitable for regional-scale estimation, though parameter availability remains the determining factor for relative performance. The convergence suggests that method selection can prioritize operational requirements (robustness versus accuracy) rather than fundamental capability differences.

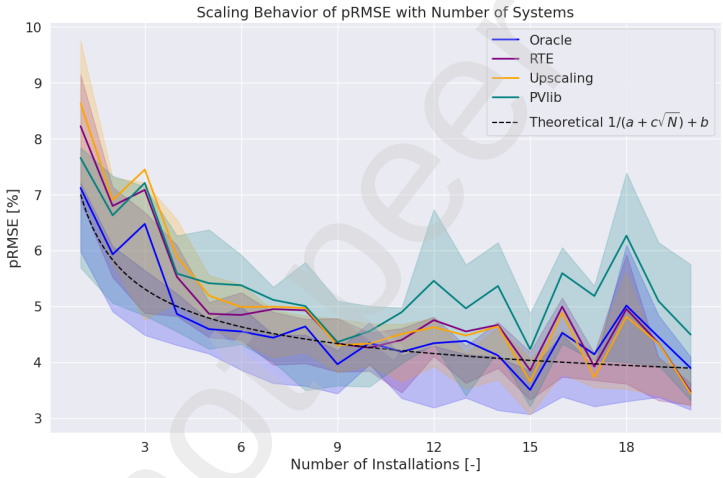


Figure 7: Error reduction through system aggregation: both RTE and PVLib demonstrate improved accuracy with increasing system count, validating regional-scale application suitability.

5.3. Sensitivity and Enhancement Effects

5.3.1. Shading impact assessment

Detailed analysis in Appendix C.4 demonstrates that incorporating local shading effects in physics-based models reduces estimation errors by up to 25%. This enhancement requires 3D environmental data but no additional system parameters, making it scalable for operational deployment. However, shading corrections increase sensitivity to system parameterization accuracy, reinforcing the robustness trade-offs identified in the primary analysis.

5.3.2. Parameter sensitivity quantification

Individual system analysis reveals physics-based method sensitivity hierarchy: installed capacity estimation (highest impact), followed by azimuth angle, then tilt angle. This sensitivity ordering differs from utility-scale analyses, reflecting the unique characteristics of small residential systems and system-level versus aggregated error analysis.

5.4. Operational Implications and Performance Summary

The evaluation quantifies French TSO operational method accuracy and establishes independent physics-based benchmarks. Key findings demonstrate that RTE's approach for reconstructing unmetered rooftop PV power profiles achieves accuracy comparable to physics-based methods, supporting prior conclusions (Huxley et al., 2022) that production estimation uncertainty represents a manageable component of overall rooftop PV operational uncertainty.

Physics-based methods can achieve marginally superior accuracy under optimal conditions but exhibit significantly larger error variability and reduced operational robustness. RTE's method demonstrates superior robustness, highlighting practical trade-offs between peak accuracy potential and minimal system characterization requirements for operational deployment.

Systematic bias analysis reveals different error characteristics: physics-based methods tend toward underestimation due to capacity parameter uncertainties, while RTE's approach produces compensating morning/afternoon errors reflecting representativeness differences between reference and target system populations. Representativeness errors contribute substantially (3–4 percentage points) to statistical method uncertainty.

Both methodological approaches demonstrate effective scaling properties: aggregated estimation accuracy exceeds individual system performance, confirming suitability for regional TSO applications. This aggregation benefit partially mitigates individual system parameter uncertainties for physics-based methods while maintaining statistical method robustness advantages.

From an implementation perspective, the two approaches offer different capabilities for addressing evolving grid challenges. The physics-based method provides a foundation for modeling complex scenarios such as self-consumption patterns, where individual system production estimates can serve as inputs to consumption behavior models. In contrast, statistical upscaling methods inherently aggregate consumption and production patterns, making it difficult to separate gross generation from net injection when

self-consumption is present. As self-consumption installations now represent 60% of new rooftop PV capacity, the physics-based approach offers greater adaptability for future grid integration challenges. The individual system-level estimates enable more sophisticated modeling frameworks that can account for storage systems, demand response, and behavioral consumption patterns—capabilities that become increasingly important as distributed energy systems evolve beyond simple grid injection models.

6. Discussion and conclusion

Conclusions and main results. This work provides the first quantitative benchmarking of the French TSO's method for estimating rooftop PV production against a physics-based approach leveraging minimal information regarding rooftop PV systems. Using a ground-truth dataset of metered systems across France, we evaluated the accuracy, robustness, and operational trade-offs of both methods.

Our central finding is that the TSO's method performs better than expected, matching the performance of the physics-based method initially introduced by Kasmi et al. (2024). Prior to this work, concerns expressed in the 2021 RTE report (RTE France and IEA, 2021) suggested that current estimation methods might be significantly inaccurate, but no quantitative assessment existed. This study provides the first empirical validation, revealing that both methods achieve operationally acceptable accuracy with aggregated estimation errors of approximately 4% and individual system-level performance characterized by pRMSE of 7.98%. Three conditions explain this performance in France: (i) the density of measurable systems is high enough to minimize interpolation errors, (ii) the installed capacity of distributed PV systems is well known (whether at individual or aggregate levels, as it represents the main source of estimation uncertainty), and (iii) self-consumption and behavioral changes in distributed PV usage remain negligible.

When condition (i) is not met, the physics-based method of Kasmi et al. (2024) presents a feasible alternative, as its performance is comparable with the TSO's approach and is not sensitive to spatial sparsity. Condition (ii) shifts the challenge from the estimation methodology to the accuracy of the underlying PV registries, underlining the critical importance of maintaining accurate and up-to-date inventories of rooftop PV installed capacity and retrieving their characteristics accurately. This finding aligns with prior work

in the UK, which similarly demonstrated that estimation uncertainties are dominated by installed capacity assessment rather than production estimation methodologies (Huxley et al., 2022). Finally, condition (iii) is already obsolete, as self-consumption systems (surplus injection and autoconsommation modes) currently account for approximately 60% of connected rooftop PV systems and now represent 100% of rooftop PV growth, creating immediate challenges for both TSO and physics-based approaches.

Limitations and outlook. While this analysis offers insights at national scale, it is subject to several limitations. First, the size of the ground-truth sample, though large, may not capture all heterogeneity in rooftop PV systems, particularly in regions with atypical deployment patterns.

Second, our evaluation assumes no self-consumption, which is increasingly unrealistic as distributed energy resources proliferate. Addressing this limitation will require future work to integrate net-load estimation at the substation level—where aggregate consumption and generation data can be combined—leveraging scenario-based assumptions about self-consumption shares. The physics-based method presented here represents a significant step towards such integration, but future studies should consider explicit modeling of self-consumption and behavioral variability.

The rapid evolution toward self-consumption installations raises questions about the temporal validity of our findings. However, at aggregate scales relevant for TSO operations (e.g., substation level), the visibility of self-consumption patterns in load curves may be limited, potentially preserving the applicability of both estimation approaches. Moreover, the physics-based method's adaptability to evolving system configurations provides greater resilience to changing deployment patterns compared to statistical approaches dependent on historical reference data. Future work should investigate the aggregate-level impacts of self-consumption on estimation accuracy and develop hybrid approaches that account for consumption behavior at the substation scale.

Broader applicability. These findings have broader relevance beyond the French context. Physics-based approaches can be particularly valuable in regions where metered systems are sparse or where the assumption of representativeness between observed and unobserved systems does not hold. While the French TSO's method benefits from dense measurement networks, its performance depends on the proximity and similarity of monitored plants—an

assumption that may weaken as utility-scale PV systems increasingly employ trackers and distributed systems diversify in design and usage patterns. In such cases, physics-based models offer a robust alternative with comparable performance independent of measurement density. Conversely, in data-rich environments, hybrid approaches combining statistical extrapolation and physics-based estimation may provide the most resilient solution for ensuring rooftop PV observability.

Regarding method selection, physics-based approaches should be preferred for estimating rooftop PV production. These methods already demonstrate satisfactory performance and can be further improved by reducing uncertainties in system parameters and installed capacity. The TSO's method, despite its simplicity and robustness, is increasingly limited by representativeness errors that may worsen as both utility-scale and distributed PV systems diversify.

Overall, this work underscores the growing importance of accurate and dynamic PV registries, alongside the need to anticipate behavioral changes in distributed PV operation. Further research should evaluate optimal aggregation levels, as individual system-level estimation may be computationally intensive relative to operational requirements.

Acknowledgements

The authors gratefully acknowledge the BDPV (Base de Données Photovoltaïque) association and its volunteer contributors for sharing high-quality production measurements from rooftop PV systems across France. This initiative was essential for enabling independent validation of operational estimation methods.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Claude Sonnet 4 (Anthropic) in order to improve writing clarity, refine paragraph structure, and assist with literature organization. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

E. Arnaudo, G. Blanco, A. Monti, G. Bianco, C. Monaco, P. Pasquali, and F. Dominici. A Comparative Evaluation of Deep Learning Techniques for Photovoltaic Panel Detection from Aerial Images. *IEEE Access*, pages 1–1, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3275435.

X. Chen and O. Ardakanian. Data efficient energy disaggregation with behind-the-meter energy resources. *Sustainable Energy, Grids and Networks*, 32:100813, Dec. 2022. ISSN 2352-4677. doi: 10.1016/j.segan.2022.100813. URL https://www.sciencedirect.com/science/article/pii/S2352467722001102.

C. M. Cheung, S. R. Kuppannagari, R. Kannan, and V. K. Prasanna. Towards Improved Real-Time Observability of Behind-Meter PhotoVoltaic Systems: A Data-Driven Approach. In *Proceedings of the Tenth ACM International Conference on Future Energy Systems*, pages 447–455, Phoenix AZ USA, June 2019. ACM. ISBN 978-1-4503-6671-7. doi: 10.1145/3307772.3331019. URL https://dl.acm.org/doi/10.1145/3307772.3331019.

J. de Hoog, M. Perera, K. Bandara, D. Senanayake, and S. Halgamuge. Solar PV Maps for Estimation and Forecasting of Distributed Solar Generation. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021.

A. Dobos. PVWatts Version 5 Manual. Technical Report NREL/TP-6A20-62641, 1158421, NREL, Sept. 2014.

N. M. Haegel, R. Margolis, T. Buonassisi, D. Feldman, A. Froitzheim, R. Garabedian, M. Green, S. Glunz, H.-M. Henning, B. Holder, and others. Terawatt-scale photovoltaics: Trajectories and challenges. *Science*, 356(6334):141–143, 2017. Publisher: American Association for the Advancement of Science.

H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer,

L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146 (730):1999–2049, 2020. ISSN 1477-870X. doi: 10.1002/qj.3803. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803.

W. Holmgren, C. Hansen, and M. Mikofski. pvlib python: a python package for modeling solar energy systems. *Journal of Open Source Software*, 3 (29):884, Sept. 2018. ISSN 2475-9066. doi: 10.21105/joss.00884.

O. Huxley, J. Taylor, A. Everard, J. Briggs, K. Tilley, J. Harwood, and A. Buckley. The uncertainties involved in measuring national solar photovoltaic electricity generation. *Renewable and Sustainable Energy Reviews*, 156:112000, Mar. 2022. ISSN 13640321. doi: 10.1016/j.rser.2021.112000.

International Renewable Energy Agency (IRENA). Renewable Capacity Statistics 2025. Technical report, International Renewable Energy Agency, Abu Dhabi, 2025.

G. Kasmi. *Enhancing the Reliability of Deep Learning Models to Improve the Observability of French Rooftop Photovoltaic Installations*. Theses, Université Paris sciences et lettres, Apr. 2024. Issue: 2024UPSLM027.

G. Kasmi, L. Dubus, Y.-M. Saint-Drenan, and P. Blanc. Towards Unsupervised Assessment with Open-Source Data of the Accuracy of Deep Learning-Based Distributed PV Mapping. In T. Corpetti, D. Ienco, R. Interdonato, M.-T. Pham, and S. Lefèvre, editors, *Proceedings of MACLEAN: MAChine Learning for EArth ObservatioN Workshop colocated with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2022), Grenoble, France, September 18-22, 2022*, volume 3343 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.

G. Kasmi, Y.-M. Saint-Drenan, D. Trebosc, R. Jolivet, J. Leloux, B. Sarr, and L. Dubus. A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata. *Scientific Data*, 10(1): 59, Jan. 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-01951-4.

G. Kasmi, A. Touron, P. Blanc, Y.-M. Saint-Drenan, M. Fortin, and L. Dubus. Remote-Sensing-Based Estimation of Rooftop Photovoltaic Power Production Using Physical Conversion Models and Weather Data. *Energies*, 17(17):4353, Aug. 2024. ISSN 1996-1073. doi: 10.3390/en17174353.

J. Lin, J. Ma, and J. Zhu. A Privacy-Preserving Federated Learning Method for Probabilistic Community-Level Behind-the-Meter Solar Generation Disaggregation. *IEEE Transactions on Smart Grid*, 13(1):268–279, Jan. 2022. ISSN 1949-3061. doi: 10.1109/TSG.2021.3115904. URL `https://ieeexplore.ieee.org/document/9548947`.

E. Lorenz, T. Scheidsteger, J. Hurka, D. Heinemann, and C. Kurz. Regional PV power prediction for improved grid integration. *Progress in Photovoltaics: Research and Applications*, 19(7):757–771, Nov. 2011. ISSN 1062-7995, 1099-159X. doi: 10.1002/pip.1033.

N. Martin 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, Dec. 2001. ISSN 0927-0248. doi: 10.1016/S0927-0248(00)00408-6.

N. Martín and J. M. Ruiz. A new model for PV modules angular losses under field conditions. *International Journal of Solar Energy*, 22(1):19–31, Jan. 2002. ISSN 0142-5919. doi: 10.1080/01425910212852. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01425910212852.

N. Martín and J. M. Ruiz. Annual angular reflection losses in PV modules. *Progress in Photovoltaics: Research and Applications*, 13 (1):75–84, 2005. ISSN 1099-159X. doi: 10.1002/pip.585. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pip.585.

K. Mayer, B. Rausch, M.-L. Arlt, G. Gust, Z. Wang, D. Neumann, and R. Rajagopal. 3D-PV-Locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3D. *Applied Energy*, 310:118469, Mar. 2022. ISSN 03062619. doi: 10.1016/j.apenergy.2021.118469.

M. Milligan, E. Ela, B.-M. Hodge, B. Kirby, D. Lew, C. Clark, J. DeCesaro, and K. Lynn. Are Integration Costs and Tariffs Based on Cost-Causation?

Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2011.

K. Pan, Z. Chen, C. S. Lai, C. Xie, D. Wang, X. Li, Z. Zhao, N. Tong, and L. L. Lai. An unsupervised data-driven approach for behind-the-meter photovoltaic power generation disaggregation. *Applied Energy*, 309: 118450, Mar. 2022a. ISSN 0306-2619. doi: 10.1016/j.apenergy.2021.118450. URL https://www.sciencedirect.com/science/article/pii/S0306261921016755.

K. Pan, Z. Chen, C. S. Lai, C. Xie, D. Wang, Z. Zhao, X. Wu, N. Tong, L. Lei Lai, and N. D. Hatziargyriou. A Novel Data-Driven Method for Behind-the-Meter Solar Generation Disaggregation With Cross-Iteration Refinement. *IEEE Transactions on Smart Grid*, 13(5):3823–3835, Sept. 2022b. ISSN 1949-3053, 1949-3061. doi: 10.1109/TSG.2022.3171656. URL https://ieeexplore.ieee.org/document/9766031/.

Z. Qu, A. Oumbe, P. Blanc, B. Espinar, G. Gesell, B. Gschwind, L. Klüser, M. Lefèvre, L. Saboret, M. Schroedter-Homscheidt, and L. Wald. Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat-4 method. *Meteorologische Zeitschrift*, 26(1):33–57, Feb. 2017. ISSN 0941-2948. doi: 10.1127/metz/2016/0781.

R. Ross. Design Techniques for Flat-Plate Photovoltaic Arrays. In *15th Photovoltaic Specialists Conference*, pages 1126–1132, Orlando, FL, 1981.

RTE France. Bilan Électrique 2024. Technical report, RTE France, Paris, France, 2025.

RTE France and IEA. Conditions and Requirements for the Technical Feasibility of a Power System with a High Share of Renewables in France Towards 2050. Technical report, Paris, 2021.

Y.-M. Saint-Drenan. *A probabilistic approach to the estimation of regional photovoltaic power generation using meteorological data*. PhD thesis, Sept. 2016.

Y.-M. Saint-Drenan, S. Bofinger, R. Fritz, S. Vogt, G. Good, and J. Dobschinski. An empirical approach to parameterizing photovoltaic plants for power forecasting and simulation. *Solar Energy*, 120:479–493, Oct. 2015. ISSN 0038092X. doi: 10.1016/j.solener.2015.07.024.

Y.-M. Saint-Drenan, G. Good, M. Braun, and T. Freisinger. Analysis of the uncertainty in the estimates of regional PV power generation evaluated with the upscaling method. *Solar Energy*, 135:536–550, Oct. 2016. ISSN 0038092X. doi: 10.1016/j.solener.2016.05.052.

Y.-M. Saint-Drenan, L. Wald, T. Ranchin, L. Dubus, and A. Troccoli. An approach for the estimation of the aggregated photovoltaic power generated in several European countries from meteorological data. In *Advances in Science and Research*, volume 15, pages 51–62. Copernicus GmbH, May 2018. doi: 10.5194/asr-15-51-2018. URL `https://asr.copernicus.org/articles/15/51/2018/`. ISSN: 1992-0628.

S. Schierenbeck, D. Graeber, A. Semmig, and A. Weber. Ein distanzbasiertes Hochrechnungsverfahren für die Einspeisung aus Photovoltaik. 2010.

H. Shaker, H. Zareipour, and D. Wood. A data-driven approach for estimating the power generation of invisible solar sites. *IEEE Transactions on Smart Grid*, 7(5):2466–2476, 2015. Publisher: IEEE.

Y. Trémenbert, G. Kasmi, L. Dubus, Y.-M. Saint-Drenan, and P. Blanc. PyPVRoof: a Python package for extracting the characteristics of rooftop PV installations using remote sensing data, Sept. 2023. arXiv:2309.07143 [eess].

A. Walch, M. Rüdisüli, R. Castello, and J.-L. Scartezzini. Quantification of existing rooftop PV hourly generation capacity and validation against measurement data. *Journal of Physics: Conference Series*, 2042(1):012011, Nov. 2021. ISSN 1742-6588, 1742-6596. doi: 10.1088/1742-6596/2042/1/012011.

F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 00994987. URL `http://www.jstor.org/stable/3001968`. Publisher: [International Biometric Society, Wiley].

Y.-K. Wu, Y.-H. Lai, C.-L. Huang, N. T. B. Phuong, and W.-S. Tan. Artificial Intelligence Applications in Estimating Invisible Solar Power Generation. *Energies*, 15(4):1312, Jan. 2022. ISSN 1996-1073. doi: 10.3390/en15041312. URL `https://www.mdpi.com/1996-1073/15/4/1312`. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

J. Yu, Z. Wang, A. Majumdar, and R. Rajagopal. DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States. *Joule*, 2(12):2605–2617, Dec. 2018. ISSN 2542-4351. doi: 10.1016/j.joule.2018.11.021.

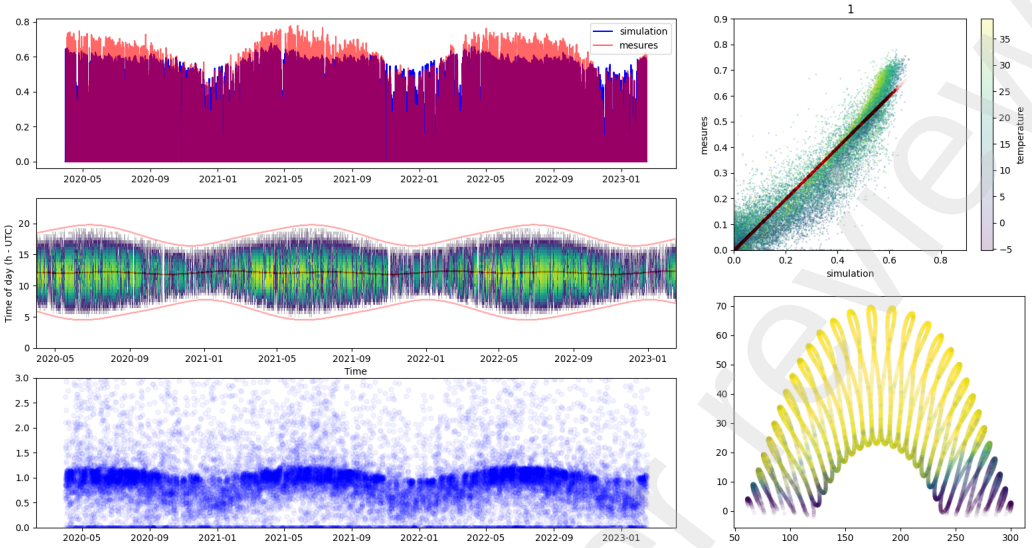
Appendix A. Implementation Details

Appendix A.1. Ground Truth Data Filtering Process

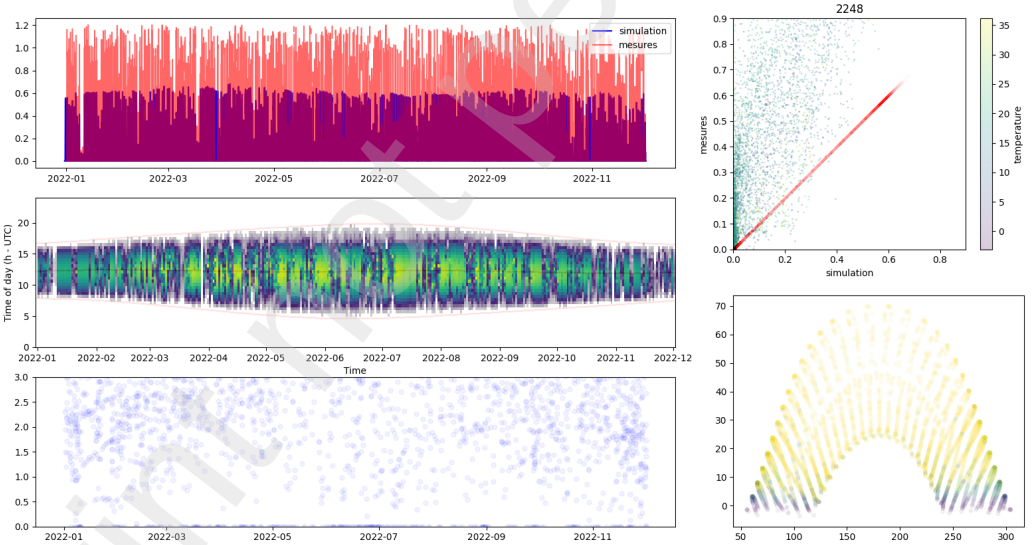
Note: The following description of the data curation process is adapted from Kasmi et al. (2024).

In addition to the quantitative quality checks detailed in Section **??**, we performed qualitative inspection of individual system performance using diagnostic reports, as illustrated in Figures A.8a and A.8b. This procedure consists of fitting a detailed simulation model to each installation and comparing its expected production with recorded measurements. Unlike the simplified simulation model employed elsewhere in this study, this diagnostic tool accounts for all known parameters of the PV system to evaluate consistency at the system level.

Figure A.8a shows an example of an installation successfully passing the quality check, with close alignment between simulated and measured production over time. Conversely, Figure A.8b highlights a system (ID #2248) flagged for exclusion due to significant discrepancies. In this case, the mismatch likely stems from either inaccurate geolocation (leading to erroneous weather inputs) or system malfunction resulting in consistently lower-than-expected production. Such diagnostic plots proved critical for identifying and filtering out installations with unreliable data, ensuring robustness of the benchmarking dataset.



(a) Example of a rooftop PV system passing the quality control (QC) process. Simulated production (blue) aligns closely with recorded measurements (orange), supporting data reliability.



(b) Example of a rooftop PV system failing the quality control (QC) process. Large discrepancies between simulated and recorded production suggest metadata inaccuracies or system malfunction.

Figure A.8: Quality control analysis of two rooftop PV systems. The first system (top) passes the QC checks, while the second system (bottom) fails due to poor agreement between simulation and measurements.

Appendix A.2. Upscaling Approach Implementation Details

To mitigate bias in the upscaling estimation process, we implement a filtering step that excludes systems with zero or missing output from the interpolation. This ensures that the aggregated estimate remains representative of active systems and prevents artificial deflation of estimated power. This filtering procedure is applied at each time step independently.

Formally, let m denote the total number of available systems in the dataset, and let n denote the number of nearest neighbors used for interpolation (with $n \leq m$). For each time step, let $\mathcal{I}$ denote the subset of systems with positive output:

$$\mathcal{I} = \{i \in \{1, \ldots, m\} \,|\, y_t(x_i) > 0\},$$

where $y_t(x_i)$ is the power output at time t and location x_i. From this subset, we select the n nearest neighbors to the target location x_0:

$$\tilde{\mathcal{I}} = \text{indices of the } n \text{ closest points within } \mathcal{I} \text{ to } x_0.$$

If insufficient neighbors with positive output exist ($|\tilde{\mathcal{I}}| = 0$), the estimate is set to zero:

$$\hat{y}_t(x_0) = 0.$$

Otherwise, the estimate is computed as a weighted average of the neighbors' outputs, where the weights decay with distance according to a power parameter p:

$$w(x_0, x_i) = \frac{d(x_0, x_i)^{-p}}{\sum_{j \in \tilde{\mathcal{I}}} d(x_0, x_j)^{-p}} \quad \text{for } i \in \tilde{\mathcal{I}},$$
$$\hat{y}_t(x_0) = \sum_{i \in \tilde{\mathcal{I}}} w(x_0, x_i) \cdot y_t(x_i),$$

where $d(x_0, x_i)$ denotes the distance between points x_0 and x_i.

Figure A.9 illustrates the impact of applying this filtering step. Before filtering, a significant number of estimated outputs are exactly zero, which artificially biases the error metrics and underestimates the true performance of the method. After filtering, these zero-value estimates are removed, resulting in error estimates that more accurately reflect the model's capabilities.

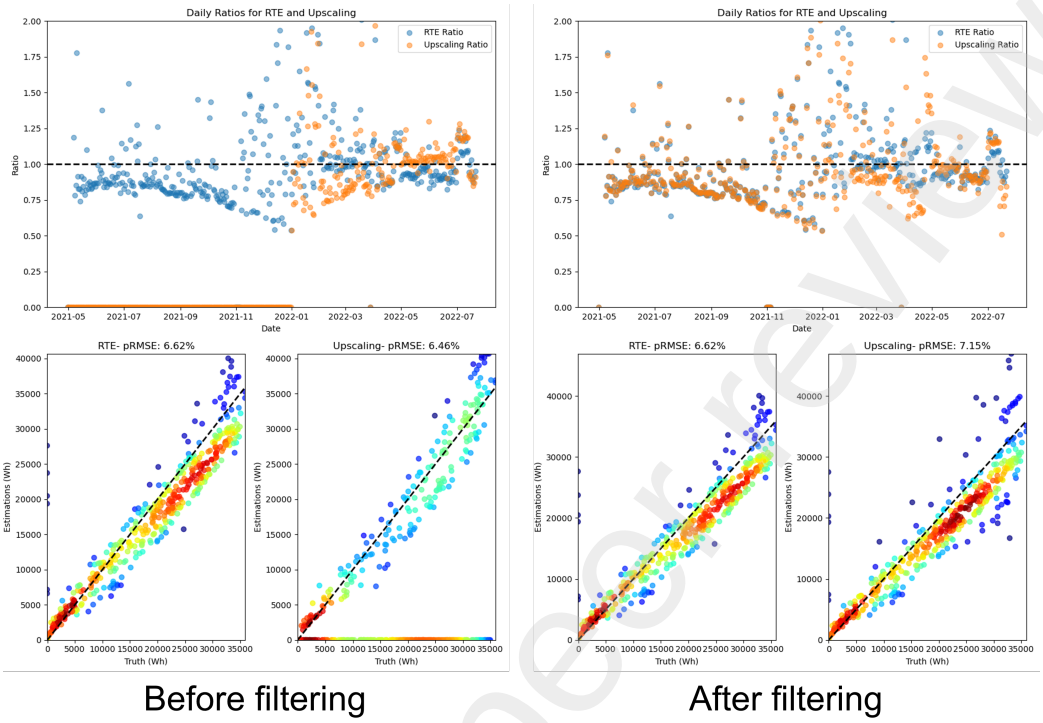


Figure A.9: Effect of the filtering step on the distribution of upscaled power estimates. The filtering removes zero-valued estimates, leading to more realistic error evaluation.

Appendix A.3. TSO Method Implementation Details

Installed capacities versus monthly energies. The TSO relies on monthly energies retrieved from distribution system operators as the basis for fitting power estimation curves, rather than the installed capacity of systems (despite this information being available for each system). For easier comparison with the vanilla upscaling approach and physics-based methods, which all take as input the installed capacity of the target system, we implemented the TSO's method using installed capacity rather than monthly energies.

As shown in Table A.3, switching between monthly energies and installed capacity does not substantially change the results. The biases are smaller when using monthly energies, but the error is less robust than when installed capacity is used. As further discussed in Section Appendix B.2, our emulation of the TSO's method is comparable to actual data. We emulated the method because we cannot precisely identify the ground truth systems in the TSO's data, with the exception of anecdotal cases. Moreover, there is

no certain way of ensuring the match between our telemetered systems and those registered in RTE's system, as system identifiers do not match and other identifying information may be inconsistent or unavailable.

Table A.3: Comparison of RTE's method performance when adjusting power curves using monthly installed capacity (as done in production) compared to adjustment by installed capacity.

Metric	Method	Mean	Min	Max	Median
Error (pRMSE) [%]	Monthly energy	9.17	3.77	273.51	7.70
	Installed capacity	9.54	5.88	18.97	9.20
Bias (MBE) [kW]	Monthly energy	-6.74	-128.37	101.34	-5.44
	Installed capacity	52.24	-420.53	583.93	50.20

Appendix A.4. Physics-Based Method Implementation

Note: The following description of the physics-based method derivation is adapted from Kasmi et al. (2024).

Calculation of Plane-of-Array (POA) Irradiance. The plane-of-array (POA) irradiance corresponds to the solar radiation incident on a surface oriented according to the tilt and azimuth angles of the photovoltaic array. It represents the amount of solar energy effectively received by the module surface based on its alignment with the sun. This irradiance can be decomposed into three principal components:

- The direct (or beam) component: solar radiation traveling directly from the sun to the surface without scattering;
- The diffuse component: radiation scattered by atmospheric particles and molecules, arriving from all directions except the direct solar beam;
- The reflected component: sunlight reflected from nearby surfaces such as the ground or surrounding structures onto the PV module.

The sum of these components yields the total POA irradiance. Computation is performed using the `pvlib` Python library (Holmgren et al., 2018), which requires as inputs solar zenith and azimuth angles, top-of-atmosphere sun position, global horizontal irradiance (GHI), diffuse horizontal irradiance (DHI), direct normal irradiance (DNI), and the module's tilt and azimuth angles.

Calculation of Module Temperature. Since PV module efficiency decreases as temperature rises, the module temperature is estimated following the approach of Ross (1981), as given by Equation (A.1):

$$T_{\text{module},t} = T_{2m,t} + \frac{k_{\text{therm}} G_{\text{POA},t}}{G_{\text{stc}}} \tag{A.1}$$

where $T_{2m,t}$ denotes the ambient temperature at 2 meters height, $G_{\text{POA},t}$ is the plane-of-array irradiance at time t, $G_{\text{stc}} = 1000\ \text{W/m}^2$ is the irradiance under standard test conditions, and k_{therm} is a thermal coefficient representing the linear increase of module temperature with irradiance (Kasmi et al., 2024).

Calculation of Effective POA Irradiance. To account for optical losses caused by reflection at the module surface, we apply the Incident Angle Modifier (IAM) model developed by Martin and Ruiz Martin and Ruiz (2001); Martín and Ruiz (2002, 2005). This model adjusts the POA irradiance based on the solar incidence angle θ_{AOI}, acknowledging that reflectance losses depend on this angle. For monocrystalline modules, the angular losses $AL(\theta_{AOI})$ are expressed as:

$$AL\left(\theta_{AOI}\right) = 1 - \frac{\bar{T}(\theta_{AOI})}{\bar{T}(0)} = 1 - \frac{1 - \exp\left(-\cos(\theta_{AOI})/a_r\right)}{-1/\exp(1 - a_r)} \approx 1 - \frac{1 - \bar{R}(\theta_{AOI})}{1 - \bar{R}(0)} \tag{A.2}$$

where $\bar{T}(x)$ and $\bar{R}(x)$ denote the weighted transmittance and reflectance at angle x, respectively, and a_r is an empirical angular losses coefficient Kasmi et al. (2024).

The IAM factor f_{I_α} is then defined as the ratio of the module's short circuit current at incidence angle θ_{AOI} to that at normal incidence:

$$f_{I_\alpha} = \frac{I_{sc}(\theta_{AOI})}{I_{sc}(0)\cos(\theta_{AOI})} \approx \frac{1 - \bar{R}(\theta_{AOI})}{1 - \bar{R}(0)} \tag{A.3}$$

Applying these modifiers to each POA component, the effective POA irradiance is calculated as:

$$POA_{\text{eff}} = f_{\text{beam}} \times G_{\text{POA}} + f_{\text{diff,sky}} \times D_{\text{POA}} + f_{\text{diff,ground}} \times R_{\text{POA}} \tag{A.4}$$

where each $f_\bullet$ is the corresponding IAM for the direct, diffuse sky, and reflected components. Figure A.10 provides a schematic overview of these solar radiation components.

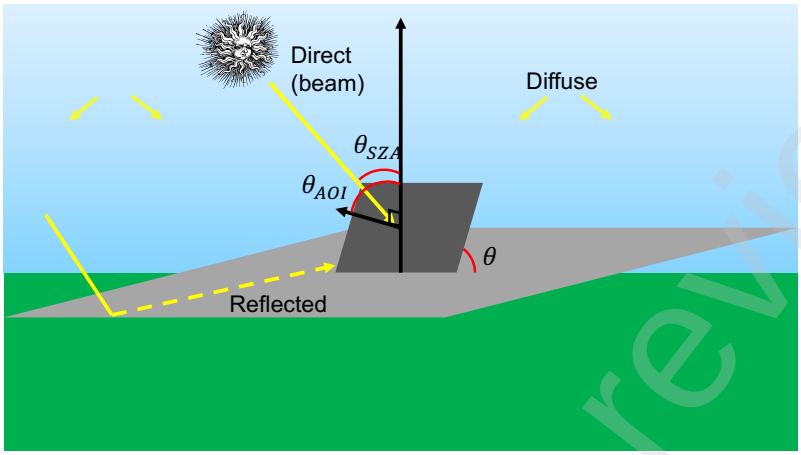


Figure A.10: Schematic representation of the modeled POA irradiance components. Angles indicated include the angle of incidence (AOI) and solar zenith angle (SZA). The flat surface (light gray) contrasts with the tilted surface (dark gray) with tilt angle θ. Source: Kasmi et al. (2024)

Appendix B. Robustness Checks

Appendix B.1. Time Lag Correction Between Multiple Time Series Using Cross-Correlation

Proposed Method: Cross-Correlation for Time Alignment. Let $x(t)$ and $y(t)$ be two time series sampled at regular intervals, where $y(t)$ may be a delayed version of $x(t)$. The objective is to estimate the lag τ that best aligns $y(t+\tau)$ with $x(t)$ by maximizing their similarity.

The cross-correlation function between $x(t)$ and $y(t)$ is computed as:

$$R_{xy}(\tau) = \sum_t x(t) \cdot y(t+\tau).$$

In practice, we use the normalized cross-correlation defined by:

$$\rho_{xy}(\tau) = \frac{\sum_t (x(t) - \bar{x})(y(t+\tau) - \bar{y})}{\sqrt{\sum_t (x(t) - \bar{x})^2} \cdot \sqrt{\sum_t (y(t+\tau) - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ denote the mean values of x and y, respectively.

The estimated lag $\hat{\tau}$ is the value of τ within the search window $[-\tau_{\max}, \tau_{\max}]$ that maximizes $\rho_{xy}(\tau)$:

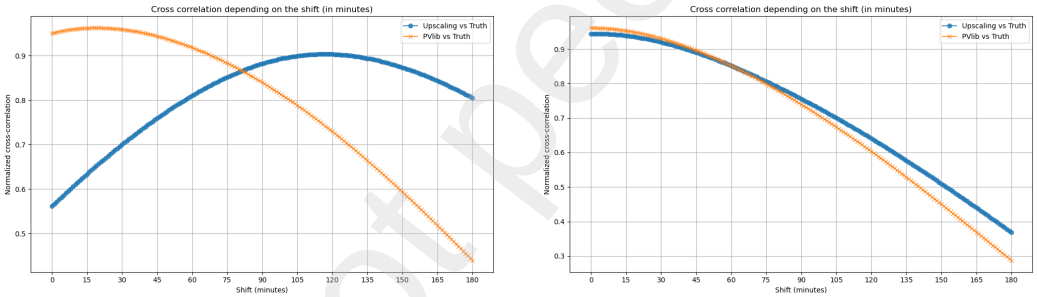
$$\hat{\tau} = \arg \max_{\tau \in [-\tau_{\max}, \tau_{\max}]} \rho_{xy}(\tau).$$

Applying this lag yields the aligned time series:

$$y_{\text{aligned}}(t) = y(t + \hat{\tau}).$$

Practical Considerations.

- **Windowing**: For long or non-stationary series, compute cross-correlation over sliding windows to capture local delays.
- **Preprocessing**: Normalize or detrend data to reduce spurious correlations.
- **Boundary Handling**: Use zero-padding or truncate overlapping regions to handle signal shifts.
- **Filtering**: Exclude zero or inactive periods (e.g., only consider $x(t) > 0$) to focus on physically relevant intervals.



(a) Cross-correlation before time alignment

(b) Cross-correlation after time alignment

Figure B.11: Normalized cross-correlation functions between ground truth and estimation methods before and after correcting for the detected time lag. The peak shift in (a) reveals a 120-minute lag that is corrected in (b), aligning the time series for subsequent analysis.

Application to Our Case. We apply this approach to align three time series: the ground measurements, the PVLib estimations, and the RTE ground truth. Figure B.11 displays the normalized cross-correlation between the ground truth and PVLib, and between the ground truth and RTE (or equivalently the upscaling estimation from the same data source) before lag correction.

In the illustrated example, a lag of 120 minutes is evident, corresponding to a timezone offset between UTC and UTC+2. This lag was corrected optimally for each system prior to further analysis.

Appendix B.2. Alignment between our emulation of RTE's method and the actual results

We were able to manually identify a few individual systems in RTE's data. For these systems, we compared our emulation of the estimation method with the actual value. The aim is to check that our emulation produces sensible results and is faithful to the original method, which is currently deployed in production.

Figure B.12 plots an example comparison. We can see that for this system, the true method and its emulation produce comparable results. On this system in particular, it appears that we are overestimating the performance of RTE's method slightly (the error and bias are lower with the emulation than with the true method), but we have too few samples to draw robust conclusions. Qualitatively, it appears that the emulation is underestimating power production during summer.

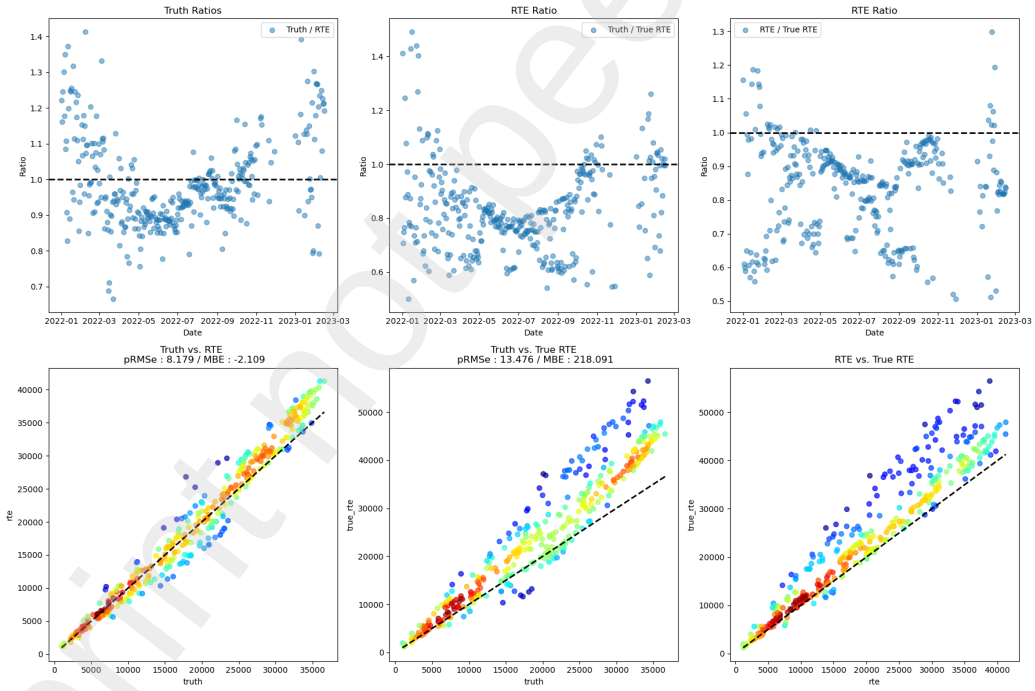


Figure B.12: Comparison of our emulation of RTE's method with the actual method on a specific sample.

Appendix B.3. Effect of source fleet size and distance on statistical methods

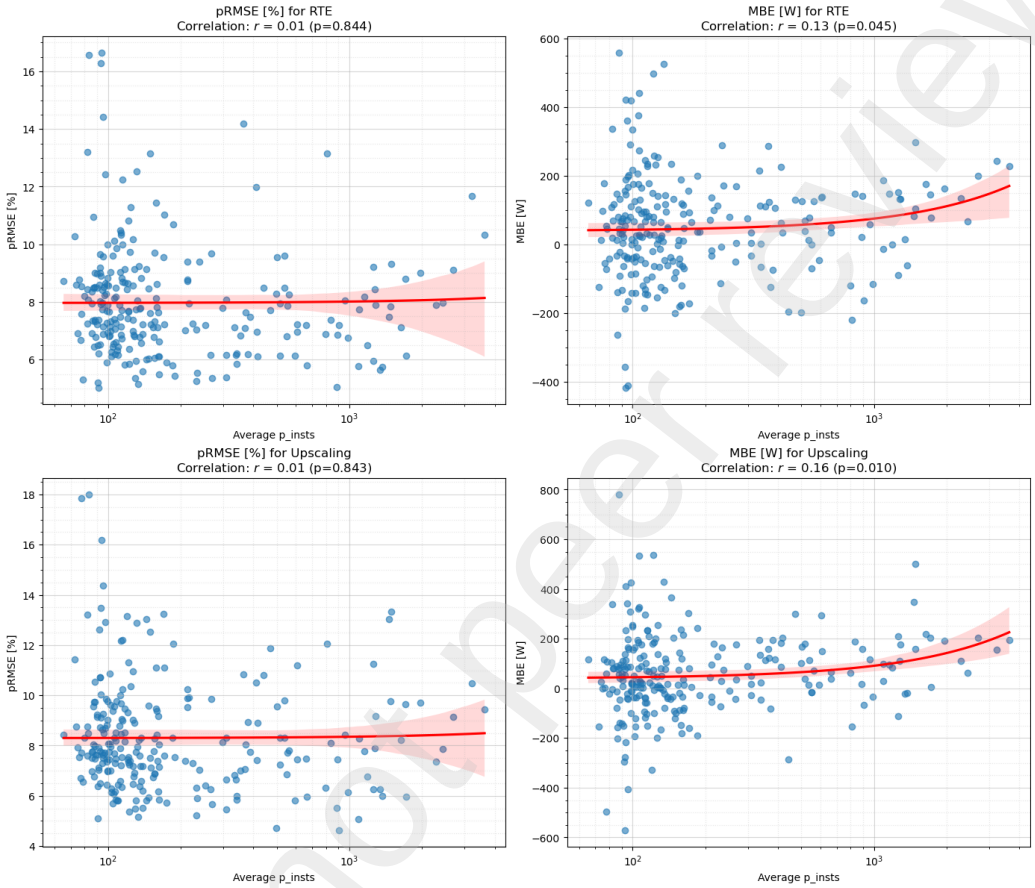


Figure B.13: Error (left column) and bias (right column) of RTE (first row) and Upscaling (second row) as a function of the average individual system size in the source systems. The installed capacities are plotted on log scale.

Effect of average source fleet size. Figure B.13 plots the error and bias of the Upscaling method and RTE's median filtering as a function of the average size of the source plants. We can see that in our sample, most plants have an average (individual) installed capacity lower than 500 kW_p, indicating small power plants or even some rooftop plants. We can see that the error is not affected by the average fleet size. The bias, on the other hand, exhibits a slight upward tendency, which could be explained by the fact that larger systems tend to have a higher average load factor (due to system optimization), thus

leading to overestimation of power production. However, in our sample, this correlation is only weakly significant (the p-value of the regression coefficient is equal to 0.05 for RTE and 0.01 for Upscaling).

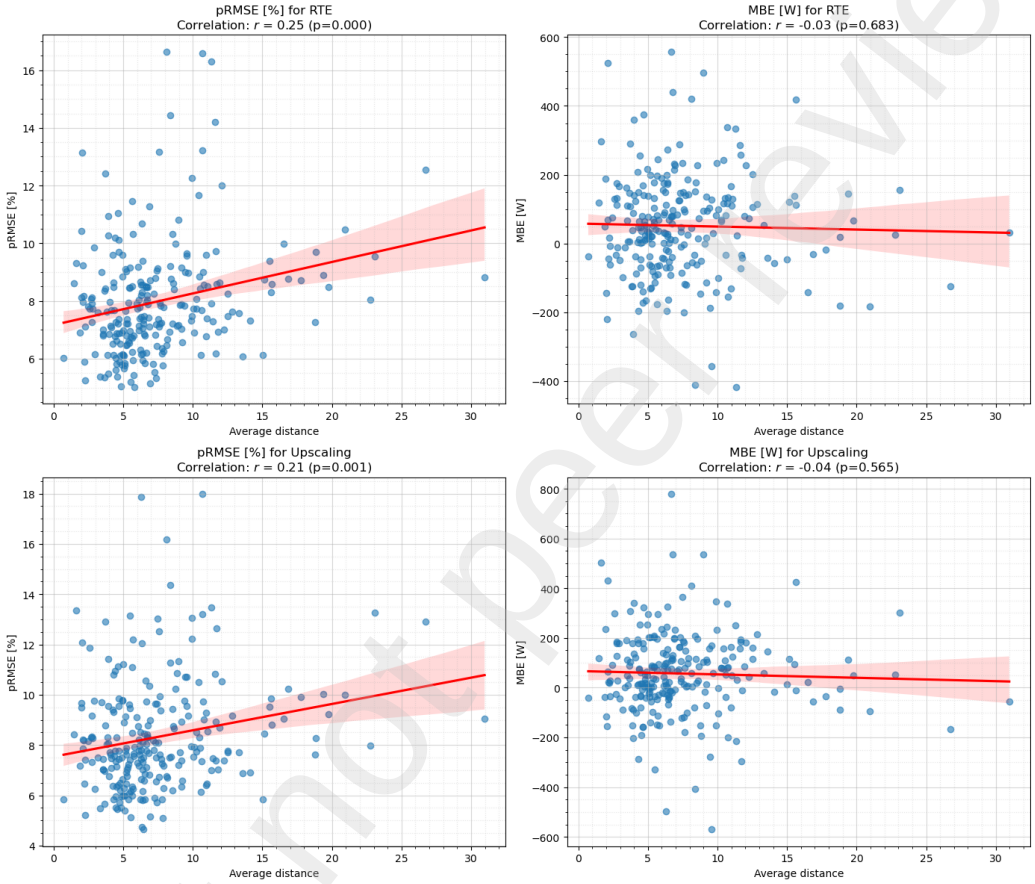


Figure B.14: Error (left column) and bias (right column) of RTE (first row) and Upscaling (second row) as a function of the average distance between source systems and the target system.

Effect of geographical distance. Regarding distance, our results unequivocally document a negative effect of distance on error. Despite the average distance in our sample being relatively modest (between 5 and 10 km), we can see a strict increase in error when distance increases. On the other hand, the bias, which has *a priori* no reason to be affected by distance, is as expected not influenced by distance.

Appendix C. Additional Results

Appendix C.1. Individual system visualizations

The main section of the paper focused on analysis of the methods at the aggregated scale. In this section, we provide complementary analysis at the individual system scale. Figure C.15 illustrates the results for a system randomly chosen from our sample.

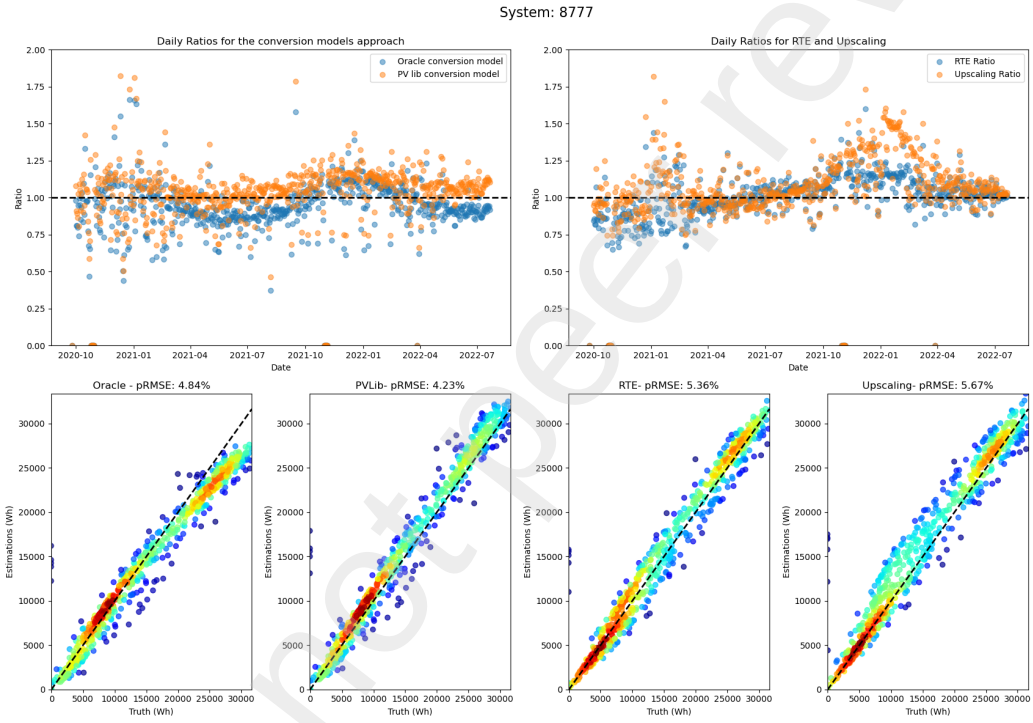


Figure C.15: Analysis report of a system from our sample.

In the upper panels, we plot the daily ratios between estimated and actual aggregated power production over time. A ratio of one indicates perfect match between estimation and ground truth. Ratios below one reflect underestimation, whereas ratios above one correspond to overestimation of daily production.

We observe that for the physics-based methods, ratios tend to be higher in winter. This behavior arises from two main factors: first, daily production levels are lower in winter, which makes ratios more sensitive to small abso-

lute errors; second, these physics-based methods do not account for shading effects, which are more significant during winter due to lower solar elevation.

The lower panels show scatter plots of estimated versus actual daily production. The density of points is color-coded, with warmer colors indicating higher densities. In these plots, the dashed $x = y$ line represents perfect agreement. For Oracle and PVLib, the estimations generally align well with ground truth, though Oracle shows slight negative bias (underestimation), especially at higher production levels. PVLib displays little to no bias and exhibits strong concentration of points along the $x = y$ line for high production values. For RTE and Upscaling, the spread is more pronounced at medium-range values, and while no strong systematic bias is evident, the variability suggests less consistent performance across the entire production range.

In contrast, Figure C.16 depicts a malfunctioning system. Here, all methods consistently underestimate production, indicating that the estimation of installed capacity is likely incorrect for this system. In the upper panels, ratios remain significantly below one throughout the year, confirming systematic underestimation.

In the lower panels, the scatter plots deviate markedly from the $x = y$ line. For both PVLib and Oracle, the relationship flattens beyond a certain production threshold, and the negative bias becomes increasingly pronounced at higher values. This pattern suggests that the models fail to scale properly for high production days, possibly due to underestimation of the system's nominal capacity.

Appendix C.2. Analysis of Error Patterns

Error distributions. Figure C.17 presents the distribution of mean bias error (MBE) across the sample of rooftop PV systems for each estimation method. The statistical approaches (RTE and Upscaling) exhibit limited overall bias but higher variance in error distribution. In contrast, the physics-based approach (PVLib) tends to show systematic upward bias in power production estimates. Standard errors remain broadly comparable across methods, though a few extreme values emerge in specific configurations, as explored in the geographical analysis.

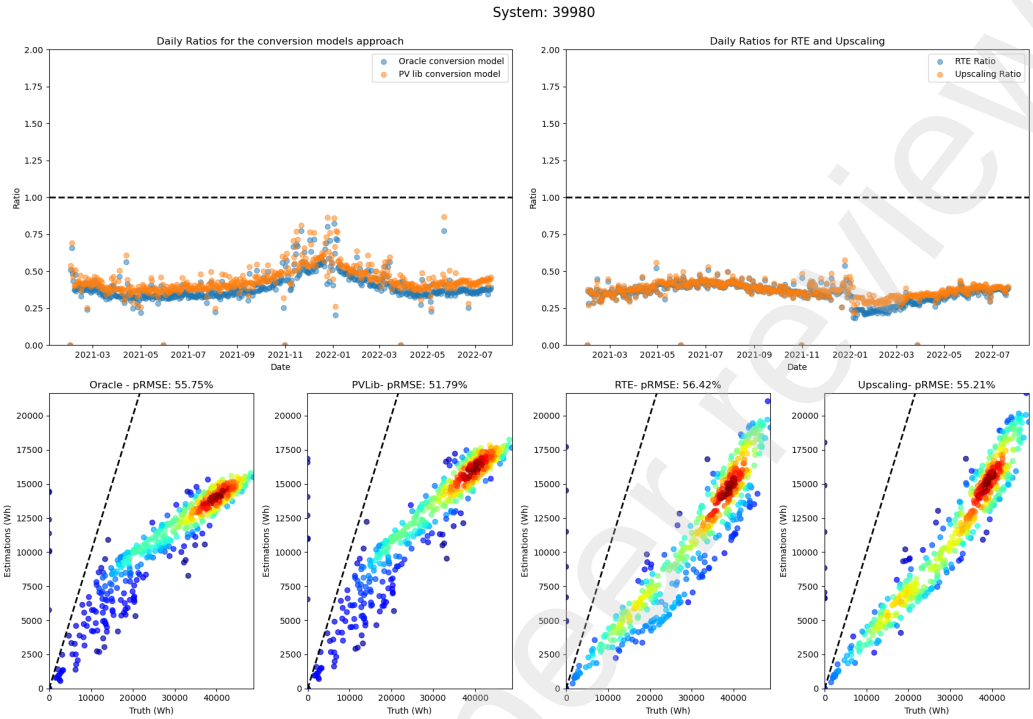
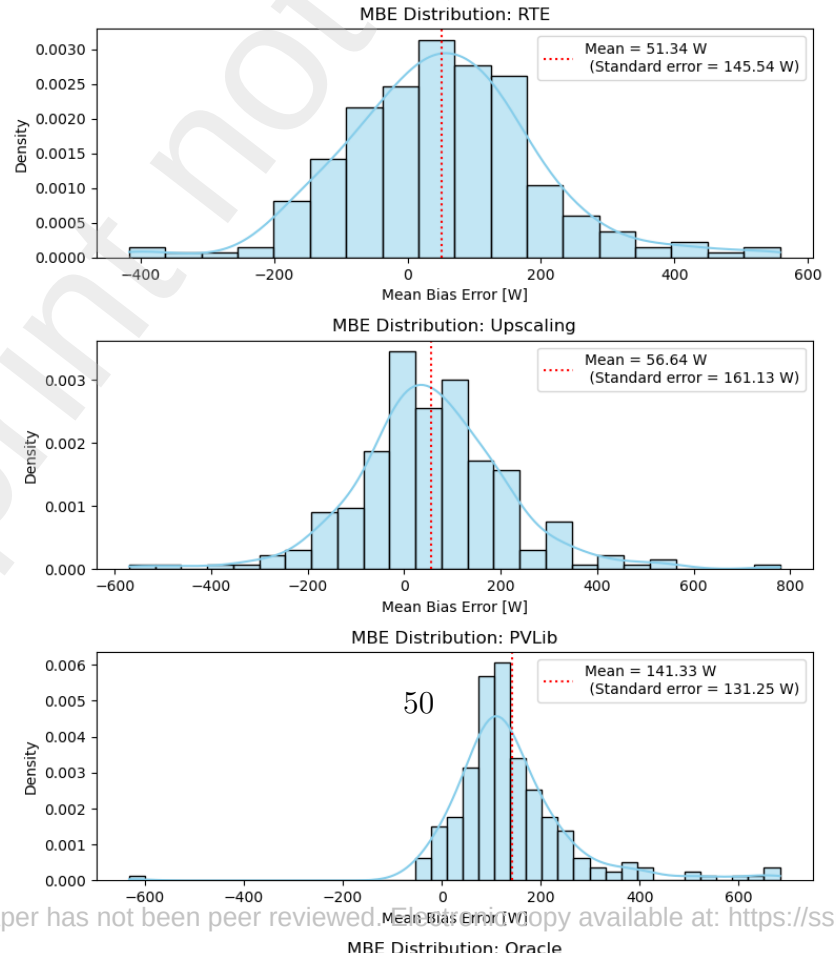


Figure C.16: Example of a failing system.



Error correlations between methods. Figure C.18 shows the cross-correlations of bias and error metrics between methods. Two distinct clusters emerge: statistical methods (RTE and Upscaling) share similar error patterns, while physics-based variants (PVLib) form a separate group. This distinction reflects the underlying methodological differences and suggests that the two classes of approaches capture complementary aspects of system behavior. Notably, there are no cases where all methods fail simultaneously, which supports the absence of systematic biases in the ground truth data.

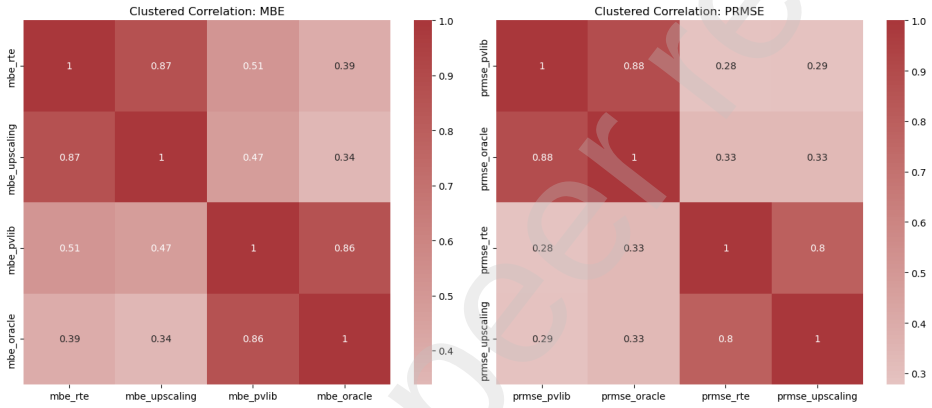


Figure C.18: Cross-correlations of mean bias error (MBE) and normalized root mean square error (pRMSE) between estimation methods. Statistical methods cluster together, while physics-based methods exhibit distinct error patterns, highlighting methodological complementarity.

Appendix C.3. Additional analysis of the Upscaling and RTE methods

Estimation of representativeness error. The representativeness error, defined as the discrepancy between the characteristics of source and target plants (Saint-Drenan et al., 2016), plays a critical role in the performance of power estimation methods. In this study, we approximate small rooftop PV systems as essentially small power plants for which parameters can be optimized.

To evaluate the irreducible error of the methods, we run both the RTE and upscaling approaches by taking metered plants as targets. These targets are otherwise used as source plants in the main analysis. This check allows us to assess a baseline level of error that cannot be eliminated due to inherent differences between systems.

The results, summarized in Table C.4, show that errors tend to be lower when source and target plants share more similar characteristics. The observed irreducible error floor is approximately 4 percentage points in terms of percentage root mean square error (pRMSE). This highlights a fundamental limitation in these estimation approaches stemming from representativeness constraints.

Metric	Method	Mean	Min	Max	Median
Error (pRMSE) [%]	RTE	3.26	0.98	19.21	2.57
	Upscaling	3.75	0.25	28.08	2.93
Bias (MBE) [kW]	RTE	0.15	-24.10	42.71	0.18
	Upscaling	0.42	-26.11	47.16	0.20

Table C.4: Performance metrics of RTE and Upscaling methods when target plants are similar to source plants. The error is expressed as percentage RMSE (pRMSE), and bias as mean bias error (MBE) in kW.

A significant factor contributing to this irreducible error is the discrepancy in azimuth distributions between metered plants and the broader population of rooftop installations. Figure C.19 illustrates this difference, showing that metered plants predominantly face southward, whereas the azimuth distribution of residential rooftops is more varied.

This azimuth mismatch may further increase in the future as adoption of tracking systems grows among large plants, improving their optimization and yields, while the increasing diversity of rooftop installations, driven by expanding solar adoption in varied urban contexts, may reduce the average optimization level for small systems. This divergence underscores the challenges in scaling estimation methods trained on metered data to the broader, heterogeneous rooftop PV population.

Appendix C.4. Effect of shading on physical model estimation accuracy

In addition to the main results presented in Section 5.1, we conducted evaluation on a set of 11 photovoltaic systems to assess the impact of incorporating shading effects using a physical shading model. These systems were manually selected based on their exposure to shading, which was characterized by annual patterns rather than daily variations. Such shading sources include, for example, buildings or trees that intermittently obstruct sunlight at certain times of the year, particularly during winter.

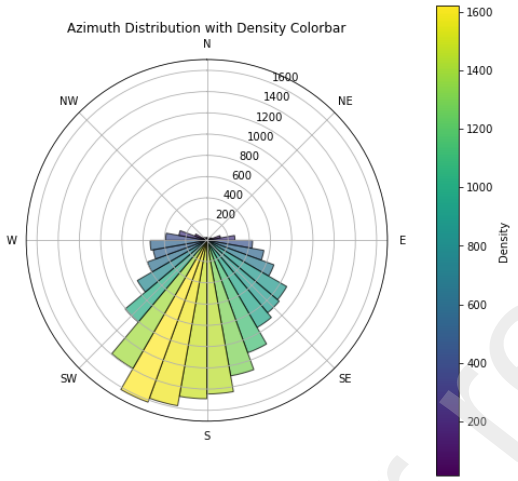


Figure C.19: Distribution of azimuth angles for metered PV plants compared to residential rooftops, highlighting the predominance of south-facing installations in the metered dataset.

Specifically, the set comprises 4 systems without shading, 3 with moderate shading, and 3 with strong shading. The shading severity was quantified by analyzing daily deviation between power estimates provided by an Oracle model and corresponding ground truth measurements.

For fair comparison, we limited the analysis to the subset of systems for which the RTE method was also simulated. The results are summarized in Table C.5. On average, incorporating shading via the physical model led to a reduction of approximately 25% in the error metric (here measured with pRMSE). However, the improvements were not uniform: in the subset of systems with moderate shading, accounting for shading effects actually resulted in slight increase in error.

A key advantage of the shading-aware approach is that it does not require detailed information about the PV system itself (such as tilt, azimuth, or installed capacity). Instead, it relies on modeling the surroundings to compute tilted irradiance that accounts for shading.

These results suggest that accounting for local shading can enhance prediction accuracy. Nevertheless, such improvements should be balanced against the potential increased sensitivity of the method to parameter uncertainties, which may reduce robustness overall.

Shading intensity	**None**	**Moderate**	**Strong**	
				Average error
RTE method	7.59	16.33	16.99	11.84
PVLib (shading not accounted for)	4.12	2.94	5.36	3.76
PVLib (accounting for shading)	2.78	3.23	4.07	3.00

Table C.5: Comparison of error (pRMSE) across shading conditions for different methods. Each column reports the average error computed on the 4 (None) or 3 (Moderate, Strong shading) systems.

Appendix C.5. Statistical significance of the error

To evaluate whether differences observed between predictive methods are statistically significant, we employ the Wilcoxon signed-rank test (Wilcoxon, 1945). This non-parametric test compares paired samples—in this case, the performance metrics (e.g., errors or biases) of two prediction methods across the same set of installations—to assess whether their distributions differ in a systematic way. Unlike parametric tests such as the paired Student's t-test, the Wilcoxon test does not assume normality of the differences and is therefore more robust to skewed or non-Gaussian data, which is common in error distributions of predictive models. It tests the null hypothesis that the median difference between paired observations is zero, providing a reliable measure of whether one method significantly outperforms another across the dataset. We chose this test because of its suitability for the paired and often non-normally distributed nature of the evaluation metrics.

Table C.6 presents the results. Bold values indicate that the reported statistic estimates a significant difference between the two error distributions. We can see that even if the overall performance is of the same order of magnitude across all methods, differences between them are significant in almost all cases, indicating that the differences in errors and biases correspond to actual differences and not statistical variations.

While this result is valid at the individual scale, it should be noted that a few percentage points difference in terms of accuracy can result in sizeable differences once scaled to the regional or national level.

Appendix C.6. Sensitivity analysis of PVLib method to system characterization

This work includes the study of different parameterizations of a conversion model. Table C.6 highlighted that the performance differences are significant

Table C.6: Wilcoxon statistic matrix reporting paired test statistics for both pRMSE (error) and MBE (bias). The matrix is symmetric, with empty diagonal entries. Bold values indicate significant differences (p ¡ 0.05). Upper and lower triangles display the same values for easier comparison. The upper part before the midrule corresponds to pRMSE, the lower part after midrule corresponds to MBE.

	Method	Truth	DPVM	Oracle	Constant	Naive	RTE	Upscaling
Error (pRMSE)	Truth		**6240.0**	**9942.0**	**7213.0**	**3328.0**	**12174.0**	**10507.0**
	DPVM	**6240.0**		**5147.0**	**11145.0**	**8203.0**	**8820.0**	**9569.0**
	Oracle	**9942.0**	**5147.0**		**4207.0**	**2281.0**	**7994.0**	**6735.0**
	Constant	**7213.0**	**11145.0**	**4207.0**		**7230.0**	**8192.0**	**10046.0**
	Naive	**3328.0**	**8203.0**	**2281.0**	**7230.0**		**4957.0**	**6103.0**
	RTE	**12174.0**	**8820.0**	**7994.0**	**8192.0**	**4957.0**		**11239.0**
	Upscaling	**10507.0**	**9569.0**	**6735.0**	**10046.0**	**6103.0**	**11239.0**	
Bias (MBE)	Truth		**7811.0**	**0.0**	**2199.0**	12489.0	**4536.0**	**5417.0**
	DPVM	**7811.0**		**11110.0**	**781.0**	**750.0**	12626.0	13001.0
	Oracle	**0.0**	**11110.0**		**5873.0**	**6243.0**	13543.0	**12769.0**
	Constant	**2199.0**	**781.0**	**5873.0**		**1.0**	**6136.0**	**6157.0**
	Naive	12489.0	**750.0**	**6243.0**	**1.0**		**8567.0**	**9136.0**
	RTE	**4536.0**	12626.0	13543.0	**6136.0**	**8567.0**		13725.0
	Upscaling	**5417.0**	13001.0	**12769.0**	**6157.0**	**9136.0**	13725.0	

from one method to another. We now study to which system parameters (tilt, azimuth, or installed capacity) the conversion model is most sensitive. To this end, we rely on sensitivity analysis.

Sensitivity metrics. We use two different metrics for our sensitivity analysis. The first one is One-At-a-Time (OAT) sensitivity analysis, which quantifies the effect of varying a single input parameter x_i on the model output $Y = f(\mathbf{x})$, while holding other parameters $\mathbf{x}_{-i}$ fixed at nominal values. The OAT sensitivity measure μ_i^* is computed as:

$$\mu_i^* = \frac{\Delta Y}{\Delta x_i} = \frac{f(x_1, \ldots, x_i + \Delta x_i, \ldots, x_k) - f(\mathbf{x})}{\Delta x_i}.$$

This approach provides a local and partial view of sensitivity but neglects parameter interactions. In practice, the step size Δx_i is taken to be a variation corresponding to $\pm 10\%$ of the parameter's nominal value (e.g., if a system has an installed capacity of 3.5 kW_p, then the step will be ± 0.35 kW_p). In our case, the nominal values are the true values of the system (i.e., corresponding to the standard conditions). We use this approach to be able to

compare our results with the results reported by Saint-Drenan (2016) using the same methodological basis.

In addition to OAT, we leverage Global Sensitivity Analysis (GSA). GSA evaluates the contribution of each input parameter to the output variance across the full input space. A common GSA metric is the Sobol' first-order index S_i, defined as

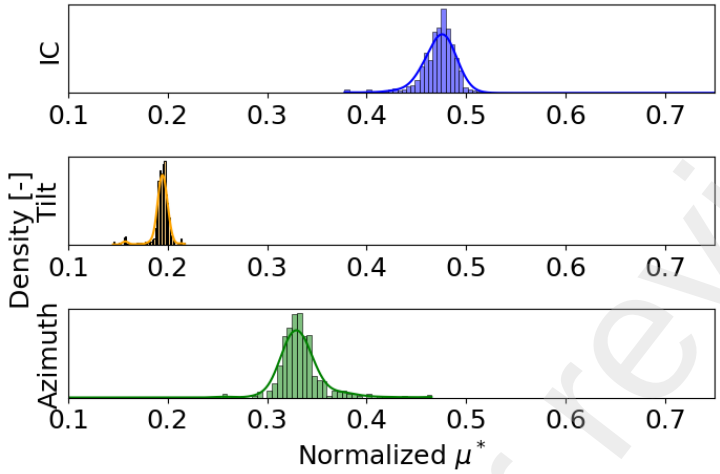
$$S_i = \frac{\mathrm{Var}_{x_i}\big(\mathbb{E}_{\mathbf{x}_{-i}}[Y \mid x_i]\big)}{\mathrm{Var}(Y)},$$

which represents the fraction of output variance explained solely by x_i. Total-effect indices S_{T_i} further capture interactions with other inputs:

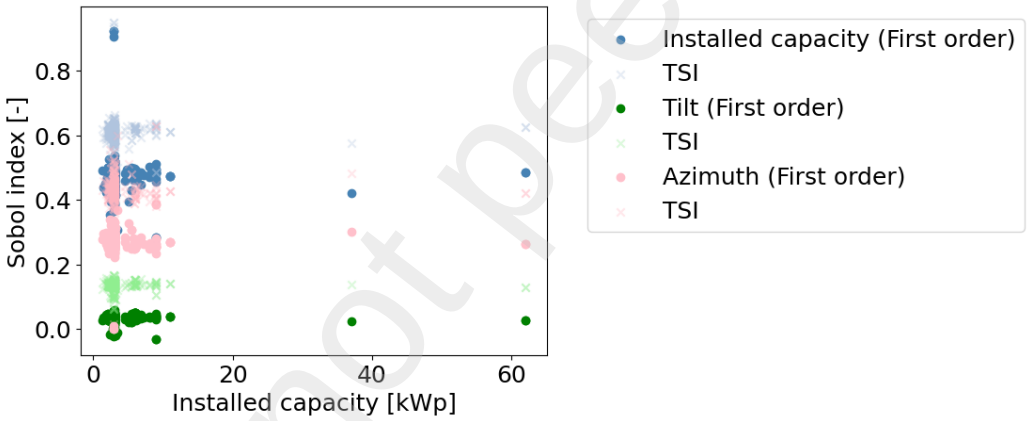
$$S_{T_i} = \frac{\mathbb{E}_{\mathbf{x}_{-i}}\big[\mathrm{Var}_{x_i}(Y \mid \mathbf{x}_{-i})\big]}{\mathrm{Var}(Y)}.$$

Results. Figure C.20 presents the results of our sensitivity analysis, with both the Global Sensitivity Analysis (GSA) and One-at-a-Time (OAT) approaches yielding consistent findings. At the individual system level, installed capacity emerges as the primary source of uncertainty, which is consistent with the findings of Huxley et al. (2022). However, when accounting for parameter interactions through GSA, the relative importance of installed capacity becomes less prominent compared to other factors, particularly the tilt angle.

Both methods indicate a reversed ranking between tilt and azimuth sensitivities compared to those reported by Saint-Drenan (2016). This is consistent with the smaller system size considered here, where factors like increased shading or installation constraints may amplify the influence of azimuth relative to tilt.



(a) OAT sensitivity analysis



(b) GSA sensitivity analysis

Figure C.20: Results for the OAT and GSA sensitivity analyses. Each point corresponds to a system.