National Laboratory of the Rockies

AI-Based Rooftop Solar Access Values: Equivalence Testing Against a Drone-Acquired Reference

Dylan Hettinger

National Laboratory of the Rockies

The National Laboratory of the Rockies is a national laboratory of the U.S. Department of Energy, Office of Critical Minerals and Energy Innovation, operated under Contract No. DE-AC36-08GO28308.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

Technical Report
NLR/TP-6A20-100072
July 2026

National Laboratory of the Rockies

AI-Based Rooftop Solar Access Values: Equivalence Testing Against a Drone-Acquired Reference

Dylan Hettinger

National Laboratory of the Rockies

Suggested Citation
Hettinger, Dylan. 2026. *AI-Based Rooftop Solar Access Values: Equivalence Testing Against a Drone-Acquired Reference*. Golden, CO: National Laboratory of the Rockies. NLR/TP-6A20-100072. https://www.nlr.gov/docs/fy26osti/100072.pdf.

The National Laboratory of the Rockies is a national laboratory of the U.S. Department of Energy, Office of Critical Minerals and Energy Innovation, operated under Contract No. DE-AC36-08GO28308.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

Technical Report
NLR/TP-6A20-100072
July 2026

National Laboratory of the Rockies
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nlr.gov

NOTICE

This work was authored by the National Laboratory of the Rockies for the U.S. Department of Energy (DOE), operated under Contract No. DE-AC36-08GO28308. Funding provided by Artemis. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via www.OSTI.gov.

Cover photo by Josh Bauer, National Laboratory of the Rockies 061725.

The National Laboratory of the Rockies prints on paper that contains recycled content.

Abstract

This study tested whether AI-generated solar access values (SAVs) may be statistically equivalent to drone-acquired SAV measurements for residential rooftop design and found equivalence at the strictest practical tolerance evaluated. Solar access value quantifies the fraction of unshaded annual solar resource available at a rooftop location after accounting for shading and sky obstruction. SAV is central to residential solar and roofing design, panel layout, and proposal quality. Remote, AI-based SAV workflows can reduce field effort and accelerate quoting relative to drone-acquired surveys, but their operational use depends on whether they produce SAV estimates that are close enough to reference measurements to support equivalent design decisions. This report evaluates whether SAVs generated by an AI-workflow are practically equivalent to SAVs generated by a drone-acquired reference workflow using a matched dataset of seven buildings, 100 rooftop locations, and monthly, seasonal, and annual reporting slices across California and Texas. Because the dataset contains multiple rooftop locations per building and repeated observations through time, equivalence is assessed using a dependence-aware two one-sided tests (TOST) workflow rather than independence-based uncertainty estimates.

In the global monthly analysis, the estimated mean paired difference is 0.765 SAV for drone-acquired SAVs relative to AI-derived SAVs. The dependence-aware spatiotemporal confidence interval is [0.644, 0.904] and the bootstrap sensitivity interval is [0.612, 0.938]. Both intervals lie fully within the strictest equivalence region (-1, +1). Equivalence is uniformly supported at margins $\Delta \in \{3, 5, 10\}$ across monthly, seasonal, annual, and building-level analyses. At $\Delta = 1$, conclusions become sensitive to slice- and roof-level uncertainty. The study demonstrates credible SAV equivalence assessment requires uncertainty models that account for clustered, spatial, and spatiotemporal dependence.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

Executive Summary

Solar access value (SAV) quantifies the fraction of unshaded annual solar resource available at a candidate rooftop location after accounting for shading and sky obstruction. It is central to panel layout, production screening, and proposal quality. Site-survey drone-acquired workflows can produce high-fidelity SAV measurements but require site access and field resources. Remote, AI-based workflows can reduce that cost and accelerate quoting, but their operational use depends on whether their SAV estimates are close enough to reference measurements to support equivalent design decisions.

This report evaluates whether AI-generated SAVs, produced by a remote workflow, are practically equivalent to SAVs produced by drone-acquired reference workflow for residential rooftops. The matched dataset spans seven buildings, 100 rooftop locations, and monthly, seasonal, and annual reporting slices in California and Texas. The central question is whether any average difference is small enough to be operationally negligible for solar design and proposal workflows.

Equivalence is assessed using the two one-sided tests (TOST) framework, where we evaluate mean differences between methods for effective equivalence rather than strict equality by calculating a mean difference between pairs of observations and checking if the confidence interval is within pre-determined bounds (Schuirmann 1987; Wellek 2010; Lakens 2017). We assess four SAV equivalence bounds: $\Delta \in \{1, 3, 5, 10\}$. This range spans a very strict near-zero criterion to wider operational tolerances. Because the dataset contains multiple rooftop locations per building and repeated monthly observations at fixed locations, the analysis uses dependence-aware spatial and spatiotemporal uncertainty models rather than independence assumptions.

The primary result indicates operational equivalence. In the global monthly analysis, the estimated mean paired difference is 0.765 SAV. The dependence-aware spatiotemporal confidence interval is [0.644, 0.904] and the bootstrap sensitivity interval is [0.612, 0.938]. Both intervals are within the strictest equivalence bounds (-1, +1), indicating practical equivalence at the portfolio level even under a stringent SAV tolerance.

Analyses by month, season, annual aggregate, and building refine that conclusion. Equivalence is consistently supported at $\Delta = 3$, $\Delta = 5$, and $\Delta = 10$. At the strictest margin, $\Delta = 1$, some slices and buildings remain fully equivalent whereas others fall just outside the upper bound. These near-threshold cases are driven by uncertainty width after accounting for spatial and spatiotemporal dependence, not by large shifts in mean paired SAV differences.

Two main conclusions follow. First, the AI remote workflow shows strong practical agreement with the drone-acquired reference workflow in SAV space for the matched rooftops and reporting windows examined here. Second, credible equivalence assessment for rooftop SAV data requires uncertainty quantification that reflects clustered, spatial, and spatiotemporal dependence.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

Preface

This report documents a practical equivalence assessment of solar access values (SAVs) generated by two rooftop solar design workflows: AI-derived values and the drone-acquired workflow. The goal is to evaluate whether an AI-derived methodology provides SAV estimates that are sufficiently close to an on-site methodology to be considered operationally interchangeable for matched rooftop locations and reporting windows.

The report is intended for technical, programmatic, and applied audiences interested in solar design validation, uncertainty quantification, and method comparison under dependent data. In addition to reporting application results, the report demonstrates a dependence-aware equivalence-testing workflow appropriate for matched rooftop datasets with clustered, spatial, and spatiotemporal structure.

The analysis focuses on practical agreement rather than exact numerical agreement. Equivalence is assessed using the TOST framework across multiple SAV margins. Because the study data include repeated measurements across rooftop locations and time, the report emphasizes dependence-aware uncertainty quantification as the basis for credible equivalence conclusions.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

Acknowledgments

The authors gratefully acknowledge the reviewers at the National Laboratory of the Rockies for their technical review, feedback, and guidance, which improved the clarity and rigor of this report.

The authors also thank Exactus Energy and Artemis for providing the datasets that made this analysis possible. Their collaboration enabled the matched comparison of drone-acquired and remotely generated solar access values used throughout this study.

Generative AI was used to produce reproducible figure code from study results.

This work was authored by the National Laboratory of the Rockies for the U.S. Department of Energy (DOE), operated under Contract No. DE-AC36-08GO28308. Support for the work was also provided by Artemis under Agreement TSA-25-32760. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or Alliance. The views and opinions of authors expressed in the available or referenced documents do not necessarily state or reflect those of the United States Government or Alliance.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

List of Acronyms

AI	artificial intelligence
AR(1)	first-order autoregressive
CI	confidence interval
GIS	geographic information system
GLS	generalized least squares
IID	independent and identically distributed
ML	maximum likelihood
PV	photovoltaic
REML	restricted maximum likelihood
SAV	solar access value
TOST	two one-sided tests

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

Table of Contents

Executive Summary ..... **iv**
1 Introduction **1**
1.1 Motivation and Use of SAV Equivalence Assessments 1
1.2 Tools Compared: AI versus Drone-Acquired SAVs 2
1.3 Why Equivalence Testing Rather Than Difference Testing? 3
2 Data and Study Design **5**
2.1 Dataset Description and Matched-Pairs Construction 5
2.2 Analysis Groupings 6
2.3 Equivalence Margins and Decision Scale 6
3 Methods: Dependence-Aware TOST **7**
3.1 Diagnosing Dependence 8
3.1.1 Spatial Dependence Within Time Slices 8
3.1.2 Spatiotemporal Dependence in the Global Monthly Dataset 9
3.2 Spatial Model for Per-Time-Slice Analyses 10
3.3 Spatiotemporal Model for the Global Monthly Panel 11
3.4 Bootstrap Validation and Sensitivity Analyses 13
3.5 Decision Rule and Reporting 13
4 Results **14**
4.1 Overall Equivalence 14
4.2 Equivalence by Time Slice 14
4.3 Equivalence by Building 16
5 Discussion **19**
6 Limitations and Future Work **21**
7 Conclusions **23**
References **24**

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

List of Figures

Figure 1. Example rooftop panel layouts. Blue circles denote panel centroids, which define the spatial support for the paired-difference analysis. Approximate intercentroid distances range from roughly 1 to 12 meters (m) depending on roof geometry. 3
Figure 2. Equivalence decision for the TOST framework. Equivalence is declared only when the confidence interval lies entirely within the prespecified acceptance region. Equivalence depends on uncertainty quantification, not only on the mean difference. 4
Figure 3. Study geography for the matched SAV dataset. The case study comprise seven buildings in California and Texas and 100 rooftop locations observed over time. 5
Figure 4. Median short-range within-building correlation of paired SAV differences by reporting slice, with the shaded band showing the 10^{th}–90^{th} percentile range across buildings. Positive values indicate nearby rooftop locations tend to move together. 9
Figure 5. Implied correlation surface for the global monthly dataset as a function of within-building distance and month lag. The surface is derived from a pooled spatiotemporal semivariogram and shows dependence persists jointly across space and time. 10
Figure 6. Global mean paired SAV difference for the monthly panel. Both the primary analytic confidence interval and the bootstrap sensitivity confidence interval are fully within the acceptance region. 14
Figure 7. Building-level mean paired SAV differences with confidence intervals. Four of seven buildings have confidence intervals within the strict (-1, +1) SAV tolerance, whereas three extend modestly above the upper bound. 17

List of Tables

Table 1. Dataset inventory by building, state, and total. Monthly observations count the January–December panel only; total observations also include annual, summer, and winter aggregates. 5
Table 2. Global equivalence results from the spatiotemporal uncertainty model using the balanced monthly panel 14
Table 3. Per-time-slice equivalence results comparing AI-derived solar access value (SAV) to drone-acquired SAV estimates. All values are in SAV units (percentage points). Slices whose model-based confidence interval lies fully within (-1, +1) satisfy the strict equivalence criterion; all slices satisfy equivalence at $\Delta = 3$. 16
Table 4. Building-level equivalence results using the spatiotemporal uncertainty model and bootstrap sensitivity intervals 17

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

1 Introduction

1.1 Motivation and Use of SAV Equivalence Assessments

Non-hardware or "soft" costs account for more than half of installed residential photovoltaic (PV) system cost in the United States, with hardware costs reported as well below 50% of total rooftop system cost, and they continue to constrain broader deployment (Friedman et al. 2013; O'Shaughnessy et al. 2019; Ramasamy et al. 2023). These costs include customer acquisition, site assessment, system design, permitting, inspection, interconnection, and other process-intensive activities. They also affect project timelines: frictions in permitting, inspection, and interconnection can delay installation and increase the labor required to move a residential system from quote to operation (O'Shaughnessy et al. 2022). Soft-cost reduction is thus both a question of lower price and of more efficient workflow.

Remote and automated solar-design tools are attractive in this setting because they may reduce field effort, accelerate quoting, and standardize early-stage design inputs. These gains are especially relevant for rooftop solar access assessment, where traditional site visits require personnel time, travel, rooftop access, and in-person measurement. Solar access value (SAV) is an important quantity in that workflow: It quantifies the fraction of unshaded annual solar resource available at a candidate rooftop location after shading and sky obstruction are considered. SAV is expressed as a percentage (0%–100%), where 100% indicates a fully unshaded location and lower values reflect obstruction losses. Accurate SAV estimation is central to layout design, production screening, and proposal quality (Nangle et al. 2015). Building-scale solar resource assessment has long relied on geographic information system (GIS) models and 3-D urban geometry to characterize rooftop exposure (Hofierka and Kaňuk 2009), with LiDAR enabling detailed shading analysis (Nguyen and Pearce 2012), time-resolved radiation simulation supporting hourly production modeling (Jakubiec and Reinhart 2013), and urban design frameworks linking these inputs to emissions management scenarios (Kanters and Wall 2014).

A natural operational question follows: Can a remote AI-assisted workflow provide SAV estimates that are interchangeable with a drone-acquired field workflow? Prior validation work approached this question by comparing remote shading outputs against reference measurements at decision-relevant tolerances (Nangle et al. 2015; Simon and Dean 2015; National Renewable Energy Laboratory 2021). The important issue is not whether two workflows produce numerically identical SAV values at every location but whether any systematic difference is small enough to be negligible for screening and design decisions.

This report continues that validation line of work using a matched dataset that compares AI-produced SAVs using a remote workflow (Artemis 2026) against a drone-acquired reference workflow (Exactus Energy 2026). The statistical objective is to quantify agreement in a way that is both operationally meaningful and inferentially rigorous. The two one-sided tests (TOST) framework is well suited to this purpose: It asks whether the mean paired SAV difference is small enough to be practically negligible, not whether it differs from zero (Schuirmann 1987; Wellek 2010; Lakens 2017). In a TOST analysis of mean differences, the equivalence decision is equivalent to checking whether the relevant two-sided $(1 - 2\alpha)$ confidence interval lies fully within the equivalence region, so the validity of the conclusion depends directly on how

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

uncertainty is quantified (Schuirmann 1987; Lakens 2017). In a dataset with building-level grouping, repeated rooftop locations, and repeated monthly observations, incorrectly assuming independence can overstate effective sample size and distort the equivalence assessment (Cameron and Miller 2015; Cressie 1993; Cressie and Wikle 2011). This report retains the practical-equivalence framing of earlier SAV validation studies while using dependence-aware uncertainty models for the AI–drone comparison.

1.2 Tools Compared: AI-derived versus Drone-Acquired SAVs

The study compares two SAV-generation workflows. Exactus Energy provides drone-acquired, site-survey-based SAV measurements and serves as the field-acquired reference (Exactus Energy 2026). Artemis provides remotely generated SAV estimates from an AI-based modeling and simulation workflow (Artemis 2026). Both workflows produce SAV values that can be aligned at common rooftop locations and common reporting windows, making a matched-pairs design the natural basis for comparison.

For rooftop location *s* and reporting slice *t*, the primary analysis variable is the paired difference

$$\mathrm{d}(\mathrm{s},\mathrm{t}) = \mathrm{SAV}_{\mathrm{Drone}}(\mathrm{s},\mathrm{t}) - \mathrm{SAV}_{\mathrm{AI}}(\mathrm{s},\mathrm{t}). \tag{1}$$

The estimand of interest is the mean paired difference,

$$\mu = \mathbb{E}[\mathrm{d}(\mathrm{s},\mathrm{t})], \tag{2}$$

where $\mathbb{E}[\cdot]$ denotes expectation over the study population of sampled rooftop locations and reporting slices. Positive values indicate drone-SAVs report higher SAV than AI-SAVs on average; negative values indicate the inverse. This paired formulation removes between-location heterogeneity and focuses the analysis on agreement between the two workflows. Figure 1 shows an example of the fixed panel-location geometry underlying the matched SAV evaluations.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

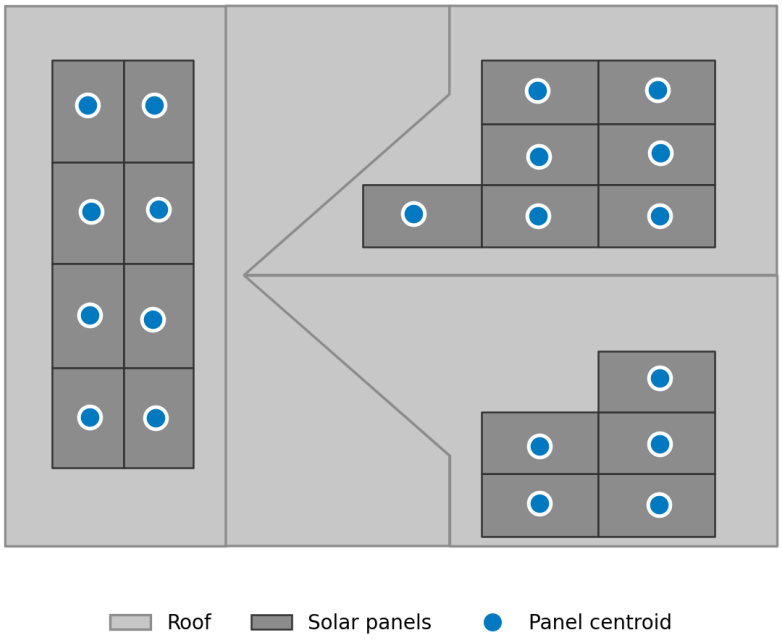

Figure 1. Example rooftop panel layouts. Blue circles denote panel centroids, which define the spatial support for the paired-difference analysis. Approximate intercentroid distances range from roughly 1 to 12 meters (m) depending on roof geometry.

The present dataset is richer than earlier SAV validation studies: It contains multiple rooftop locations per building *with precise location data*, complete monthly coverage, and seasonal and annual aggregations. Those features improve the operational relevance of the comparison, but they also mean the data are not well described by an independence assumption. Multiple points on the same roof can share obstruction geometry and local shading context, and repeated observations at the same location can share persistent seasonal or structural effects through time. The method-comparison question therefore cannot be separated from the uncertainty model used to assess equivalence.

1.3 Why Equivalence Testing Rather Than Difference Testing?

A conventional two-sided difference test asks whether the mean paired difference is distinguishable from zero. That is not the main decision problem here. Two workflows need not be numerically identical to be operationally interchangeable. What matters is whether any average difference is small enough to be negligible at tolerances relevant to design and proposal quality. Equivalence testing addresses that question directly by replacing the null hypothesis of no difference with a null hypothesis of nonequivalence (Schuirmann 1987; Wellek 2010; Lakens 2017).

Let $\Delta > 0$ denote a practical equivalence margin in SAV units. TOST is equivalent to two separate one-sided tests, each at level α, on the hypotheses $H_{0,1}: \leq -\Delta$ and $H_{0,2}: \geq -\Delta$. Together they evaluate the joint composite null

$$H_0: \mu \leq -\Delta \text{ or } \mu \geq \Delta \quad \text{versus} \quad H_1: -\Delta < \mu < \Delta. \tag{3}$$

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

Equivalence is declared only when the confidence interval for μ lies entirely within $(-\Delta, \Delta)$ (Schuirmann 1987; Lakens 2017). As Figure 2 shows, both the estimated mean difference and the width of the confidence interval affect the final decision. This framing suits SAV validation because the acceptance band is expressed directly in SAV units and can be interpreted in terms of operational tolerance, as in prior remote SAV studies (Nangle et al. 2015).

Figure 2. Equivalence decision for the TOST framework. Equivalence is declared only when the confidence interval lies entirely within the prespecified acceptance region. Equivalence depends on uncertainty quantification, not only on the mean difference.

The key methodological issue here is not only the choice of Δ but also how uncertainty is quantified for μ. If paired differences were independent, a simple analysis with independent and identically distributed (IID) sample assumptions would suffice (Casella and Berger 2002). In rooftop SAV data, nearby observations on the same building can be spatially redundant and repeated monthly measurements can exhibit temporal persistence. The effective information content can thus be much smaller than the row count suggests (Cressie 1993; Cressie and Wikle 2011). This report applies a dependence-aware TOST workflow, so equivalence is assessed using confidence intervals that reflect the clustered, spatial, and spatiotemporal structure of the matched SAV dataset.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

2 Data and Study Design

2.1 Dataset Description and Matched-Pairs Construction

This study evaluates agreement between remote AI-assisted workflow (AI-SAVs; Artemis 2026) and a site-survey reference workflow (drone-SAVs; Exactus Energy 2026) using matched SAV observations at common rooftop locations and reporting slices. Each observation corresponds to a rooftop location s on a specific building and a reporting slice t, where t denotes either an individual month or an aggregate reporting window. The study comprises 100 unique rooftop locations across seven buildings in two states. Five buildings are in California and contribute 66 rooftop locations; two buildings are in Texas and contribute 34 locations. Figure 3 shows the geographic distribution of the sampled buildings. Table 1 summarizes the dataset inventory by building, state, and study total; Figure 3 marks their locations.

Table 1. Dataset inventory by building, state, and total. Monthly observations count the January–December panel only; total observations also include annual, summer, and winter aggregates.

Subset	Buildings	Locations	Monthly Observations	Total Observations
Per Building				
McKinney, TX	1	14	168	210
Berkeley, CA	1	13	156	195
Daly City, CA	1	15	180	225
Los Angeles, CA	1	8	96	120
San Jose, CA	1	13	156	195
Cupertino, CA	1	17	204	255
Humble, TX	1	20	240	300
Per State				
California	5	66	792	990
Texas	2	34	408	510
Total				
All data	7	100	1,200	1,500

Figure 3. Study geography for the matched SAV dataset. The case study comprises seven buildings in California and Texas and 100 rooftop locations observed over time.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

Each rooftop location has complete monthly coverage for January through December, with a single representative measurement for each month. Three aggregate reporting slices are also considered: annual, summer (May–October), and winter (November–April). The full matched dataset contains 15 reporting slices per rooftop location and 1,500 paired observations in total. The balanced monthly panel yields 1,200 observations on a complete 12 × 100 location-by-month grid. These seasonal definitions follow prior SAV validation practice (Nangle et al. 2015).

The hierarchical structure of the dataset is central to the study design. Rooftop locations are nested within buildings, and the number of locations per building ranges from 8 to 20 (Table 1). The number of rows in the analysis table is therefore not the same as the number of independent units of information: Locations on the same roof can share common shading context, obstruction geometry, and surrounding built environment. The temporal structure is equally important. The same fixed rooftop locations are observed repeatedly across months, so the monthly panel contains repeated measurements on a common spatial support rather than independent temporal replicates. These features motivate the dependence-aware uncertainty models introduced in the next section.

2.2 Analysis Groupings

Results are presented at three levels of aggregation. First, the global analysis uses the balanced monthly panel pooled across all buildings and months. It targets the overall mean drone-AI difference in the full monthly dataset and provides the clearest portfolio-level assessment of equivalence when both spatial and temporal dependence are acknowledged.

Second, per-slice analyses are performed separately for each monthly, seasonal, and annual reporting slice. These assess whether agreement is stable across the calendar year and across the aggregate reporting windows used operationally. Analyzing slices separately shows whether the overall monthly conclusion is broadly representative or whether some reporting windows exhibit systematically larger offsets or wider uncertainty.

Third, by-building analyses treat each roof as its own evaluation unit. These assess whether the overall agreement pattern is consistent across buildings or whether some roofs lie closer to the operational decision boundary. The building-level results are the most informative for local robustness: A workflow can appear equivalent on average while still showing roof-specific departures that matter for conservative operational use.

Together, these three groupings provide complementary views of agreement: portfolio-level performance, temporal stability, and spatial heterogeneity across roofs.

2.3 Equivalence Margins and Decision Scale

Equivalence is assessed at $\Delta \in \{1, 3, 5, 10\}$ SAV. The strictest margin, $\Delta = 1$, is a near-zero agreement criterion corresponding to one percentage point of unshaded solar resource. The wider margins $\Delta = 3$, $\Delta = 5$, and $\Delta = 10$ provide a sensitivity analysis across progressively more permissive operational tolerances, anchored to prior SAV validation practice (Nangle et al. 2015). This multi-margin design improves interpretation. A workflow that fails at $\Delta = 1$ but passes at $\Delta = 3$ or $\Delta = 5$ is not necessarily unsuitable; it indicates practical interchangeability depends on the strictness of the tolerance.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

3 Methods: Dependence-Aware TOST

The data comprise paired rooftop locations nested within buildings and repeatedly observed through time, so the validity of the equivalence conclusion depends directly on how uncertainty is quantified for the mean paired difference. All equivalence decisions use α = 0.05, so the operational TOST criterion is containment of a two-sided 90% confidence interval within $(-\Delta, +\Delta)$; Schuirmann 1987; Lakens 2017). The analysis was implemented in Python using the pyTOST library for paired equivalence testing under dependent structures; pyTOST was developed by the lead author (Hettinger 2026).

Given

$$\mathrm{d}(\mathrm{s},\mathrm{t}) = \mathrm{SAV}_{\mathrm{Drone}}(\mathrm{s},\mathrm{t}) - \mathrm{SAV}_{\mathrm{AI}}(\mathrm{s},\mathrm{t}) \tag{4}$$

denotes the paired difference at rooftop location *s* and reporting slice *t* and

$$\mu = \mathbb{E}[\mathrm{d}(\mathrm{s},\mathrm{t})] \tag{5}$$

denotes the population mean paired difference. For equivalence margin $\Delta > 0$, TOST evaluates

$$H_0: \mu \le -\Delta \text{ or } \mu \ge \Delta \quad \text{versus} \quad \mathrm{H}_1: -\Delta < \mu < \Delta. \tag{6}$$

Equivalence is declared the two-sided $(1-2\alpha)$ confidence interval for μ lies entirely within $(-\Delta, +\Delta)$; Schuirmann 1987; Lakens 2017). The estimand and the equivalence margins are fixed throughout; what changes across analytic engines is the uncertainty model used to construct that interval.

A useful intuition is the effective-sample-size concept. If observations are positively dependent, the information content of the dataset is smaller than the row count suggests. Under a working assumption of uniform pairwise correlation $\bar{\rho} > 0$,

$$\mathrm{Var}(\bar{d}) \approx \frac{\sigma_d^2}{n}[1 + (n-1)\bar{\rho}]. \tag{7}$$

That is, variance of mean difference scales by the sample size and the correlation between samples. In effect, the data behave approximately like an independent sample of size

$$n_{\text{effective}} \approx (n)/(1 + (n-1)\bar{\rho} \tag{8}$$

As $\bar{\rho}$ increases, $n_{\text{effective}}$ can be much smaller than *n* (Kish 1965). Because TOST is equivalent to a confidence-interval containment check, underestimating $\mathrm{Var}(\bar{d})$ makes intervals too narrow and can make equivalence too easy to declare. The following methods first diagnose dependence, then propagate it into TOST confidence intervals through spatial and spatiotemporal covariance models.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

3.1 Diagnosing Dependence

The analysis begins with exploratory diagnostics for two settings: spatial dependence within individual reporting slices and joint spatiotemporal dependence in the balanced monthly panel. These diagnostics are not the final inferential models; they are descriptive tools used to assess whether nearby rooftop observations and repeated monthly observations behave like independent replicates. The logic follows standard exploratory spatial and spatiotemporal practice based on short-range correlation summaries and semivariogram concepts (Cressie 1993; Cressie and Wikle 2011; Cliff and Ord 1981).

3.1.1 Spatial Dependence Within Time Slices

For each reporting slice t, each building b is treated as an independent rooftop-scale spatial field. Within building b, all within-building point pairs (i, j) are formed and Euclidean distances

$$h_{ij} = |s_i - s_j| \tag{9}$$

are computed. Distances are grouped into within-building quantile bins so the short-distance diagnostic is defined relative to each roof's geometry.

If $[q_0, q_1)$ denotes the smallest within-building distance bin, the short-range correlation diagnostic is

$$\hat{\rho}_{b,t}^{(\text{short})} = \text{Corr}\left(d(s_i, t), d(s_j, t) \mid_{h,j} \in [q_0, q_1)\right). \tag{10}$$

This quantity shows whether nearby rooftop locations within the same reporting slice tend to move together. Under an IID model, nearby points should show no systematic tendency to have similar paired differences after conditioning on the slice. Persistent positive short-range correlation indicates spatially nearby rooftop locations are partially redundant rather than independent sources of information.

To stabilize the diagnostic, at least six points per building and at least 10 pairs in the short-range bin are required. If the first bin is too sparse, it is widened until the minimum pair threshold is met. For each slice, the resulting building-level short-range correlations are summarized by their median and a 10^{th}–90^{th} percentile band across buildings.

The results suggest short-range dependence. As shown in Figure 4, the median short-range correlation is frequently positive at the monthly level, including 0.220 in January, 0.285 in February, 0.284 in March, and 0.299 in October. The summer aggregate shows strong short-range dependence, with median 0.693 and a 10^{th}–90^{th} percentile range of [-0.006, 0.838]. Winter has median -0.079 with range [-0.236, 0.520]. Even for slices whose median is near zero, the cross-building spread remains wide, indicating some roofs still exhibit local dependence. In an IID setting, short-range correlations would fluctuate around zero without a persistent positive tendency (Cliff and Ord 1981). Their behavior here indicates nearby rooftop locations within a slice are not behaving like independent replicates, motivating covariance-aware spatial inference for the per-slice TOST analyses.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

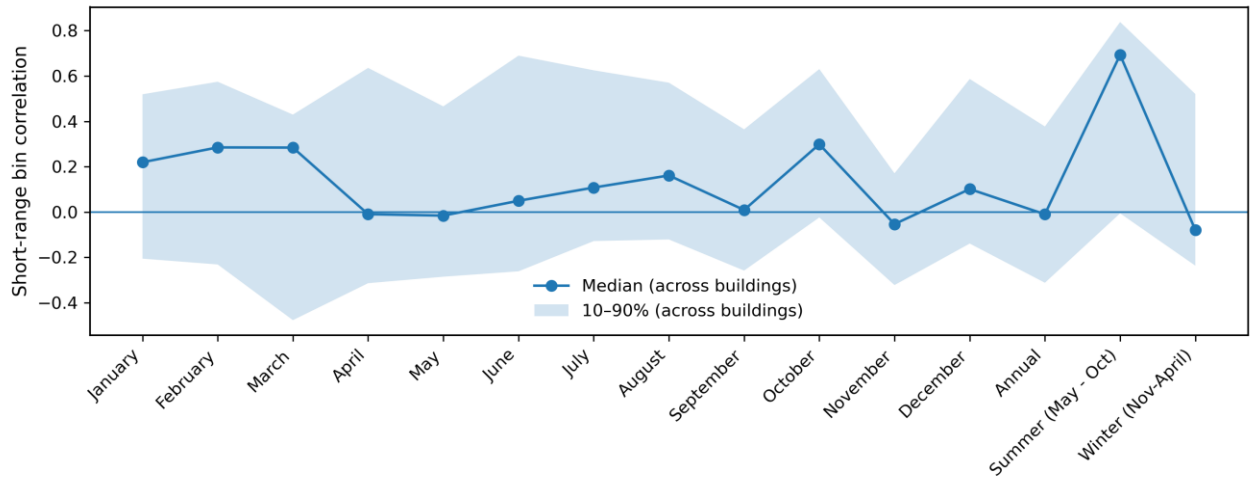

Figure 4. Median short-range within-building correlation of paired SAV differences by reporting slice, with the shaded band showing the 10th–90th percentile range across buildings. Positive values indicate nearby rooftop locations tend to move together.

3.1.2 Spatiotemporal Dependence in the Global Monthly Dataset

For the global analysis, dependence must be considered jointly across space and time because the same rooftop locations are observed repeatedly through the monthly panel. Let $d_{i,t}$ denote the paired difference at rooftop location i and month t. The pooled spatiotemporal semivariogram at spatial lag h and temporal lag u is estimated by averaging squared differences over all within-building pairs (i, j) with interlocation distance $|s_i - s_j| \approx h$ and over all month pairs (t_1, t_2) with $|t_1 - t_2| = u$:

$$\tilde{\gamma}(h,u) = \frac{1}{2N(h,u)} \sum_{\substack{(i,j,t_1,t_2):\\ ||s_i - s_j|| \approx h, |t_1 - t_2| = u}} \left(d_{i,t_1} - d_{j,t_2}\right), \tilde{\gamma}(h,u) \tag{11}$$

where $N(h,u)$ is the number of contributing pairs in the (h,u) bin (Cressie 1993; Cressie and Wikle 2011). Pooling is performed over all buildings after binning distances. $\tilde{\gamma}(h,u)$ is small when observations at nearby locations and nearby times resemble one another and large when they differ substantially.

For interpretability, this surface is converted to an implied correlation using the lag-specific sill estimate

$$\hat{\rho}(h,u) = 1 - \frac{\tilde{\gamma}(h,u)}{\widehat{\sigma_u^2}}, \quad \widehat{\sigma_u^2} = \max_h \tilde{\gamma}(h,u). \tag{12}$$

This transformation is descriptive, not model-based: It rescales the empirical semivariance so values near zero semivariance correspond to a positive association and values near the lag-specific sill correspond to near-independence. Note $\hat{\rho}(h,u)$ is confined to [0, 1] when $\tilde{\gamma}(h,u) \leq \widehat{\sigma_h^2}$ but can fall below zero at distances where the semivariance continues to rise beyond the lag-

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

specific sill. This quantity should therefore be read as an exploratory visualization of joint persistence, not as a formally estimated correlation function.

The implied-correlation surface in Figure 5 indicates dependence persists jointly across spatial separation and month lag. At lag $u = 1$, the implied correlation is 0.477 at $h = 1.38$m 38m and 0.516 at $h = 5.81$m 81m. At lag $u = 2$, the corresponding values are 0.423m and 0.492m, and at lag $u = 5$ they remain 0.346m and 0.512m. These values indicate substantial persistence over short temporal lags and short-to-moderate within-building distances, consistent with stable roof geometry and slowly varying shading conditions. The surface also shows a near-independence band around $h \approx 8.01$m 01m, where the empirical semivariance reaches the lag-specific sill and the implied correlation drops to zero by construction. Taken together, these diagnostics indicate the global monthly panel is not adequately represented by a model that is only spatial or only temporal; the dependence is genuinely joint.

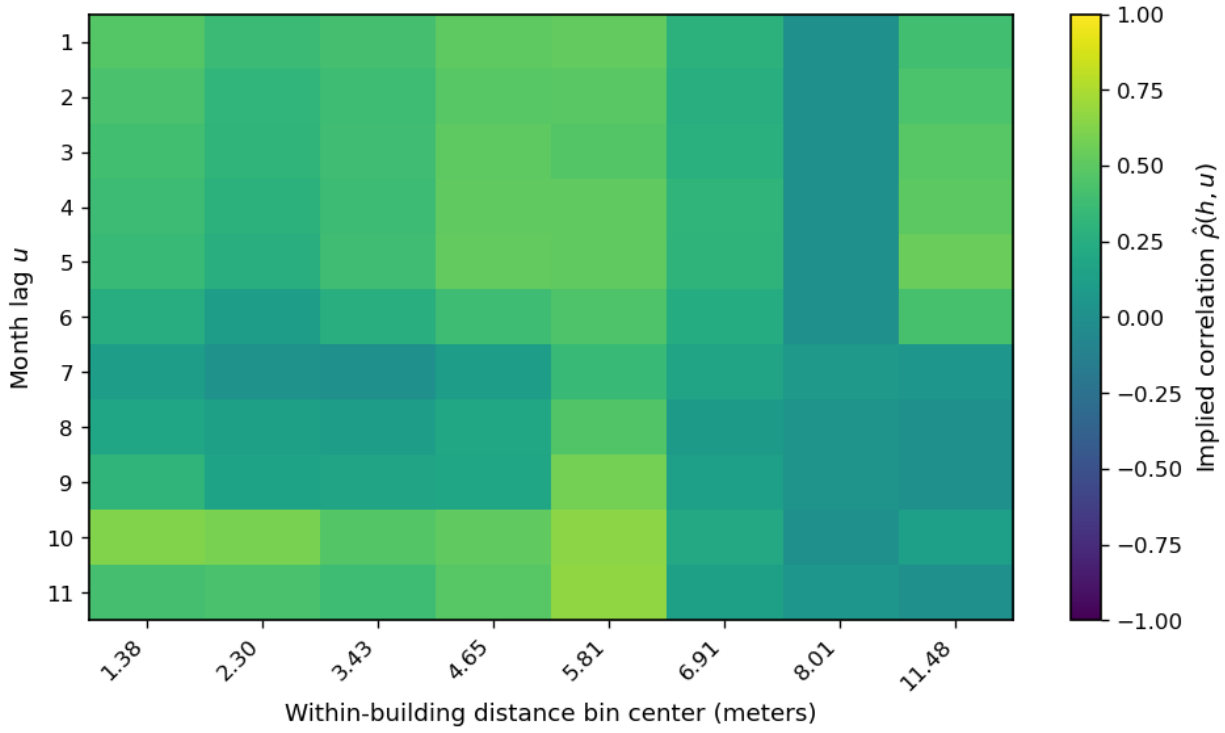

Figure 5. Implied correlation surface for the global monthly dataset as a function of within-building distance and month lag. The surface is derived from a pooled spatiotemporal semivariogram and shows dependence persists jointly across space and time.

3.2 Spatial Model for Per-Time-Slice Analyses

For a fixed reporting slice *t*, the dominant dependence pattern is within-building spatial correlation with independence across buildings. The primary per-slice analysis treats each building as an independent spatial block and models paired differences within each block using a Gaussian spatial covariance structure (Cressie 1993; Stein 1999).

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

For building b, let $\boldsymbol{d}_{b,t} \in \mathbb{R}^{n_b}$ denote the vector of paired differences in slice t, and let $\mathbf{1}_b$ denote a conformable vector of ones. The model is

$$\boldsymbol{d}_{b,t} \sim N(\mu_t \mathbf{1}_b, \Sigma_b(\theta_t)) \tag{13}$$

with covariance

$$\Sigma_b(\theta_t) = \sigma_t^2 R_{\text{Matérn}}(H_b;\ \lambda_t, \nu_t) + \tau_t^2 I_{n_b}, \tag{14}$$

where H_b is the within-building distance matrix, $R_{\text{Matérn}}(H_b;\ \lambda_t, \nu_t)$ is the Matérn correlation matrix, λ_t is the spatial range parameter, ν_t is the smoothness parameter, σ_t^2 is the partial sill, and τ^2 is the nugget variance (Stein 1999; Guttorp and Gneiting 2006). Stacking buildings yields the block-diagonal covariance matrix

$$\Sigma_t(\theta_t) = d\text{iag}\{\Sigma_1(\theta_t, \dots, \Sigma_b(\theta_t)\}. \tag{15}$$

Given $\Sigma_t(\theta_t)$, the mean paired difference for slice t is estimated by generalized least squares (GLS),

$$\hat{\mu}_t = \frac{\mathbf{1}_{n_t}^\top \Sigma_t(\theta_t)^{-1} \boldsymbol{d}_t}{\mathbf{1}_{n_t}^\top \Sigma_t(\theta_t)^{-1} \mathbf{1}_{n_t}^\top}, \tag{16}$$

Where $\mathbf{1}_{n_t}$ is the vector of ones with $\Sigma_b n_b$ equal to the total number of rooftop locations in slice t. This estimator down weights spatially redundant rooftop points (Aitken 1936). Covariance parameters are estimated by restricted maximum likelihood (REML; Patterson and Thompson 1971; Harville 1977). REML is used for the per-slice analyses because it yields unbiased variance-component estimates when the mean is estimated from the same data (Patterson and Thompson 1971; Harville 1977); the global monthly panel instead uses ML to permit likelihood-ratio-based model assessment across covariance specifications (see Section 3.3).

Inference for μ_t is based on a profile restricted log-likelihood ratio confidence interval,

$$\text{LR}_t^{\text{REML}}(\mu_t) = 2\{\ell_t^{\text{REML}}(\hat{\mu}_t, \hat{\theta}_t) - \ell_t^{\text{REML}}(\mu_t, \hat{\theta}_{t,\mu_t})\}, \tag{17}$$

with confidence set

$$C_{t\ 1-2\alpha} = \{\mu_t :\ \text{LR}_t^{\text{REML}}(\mu_t) \le \chi^2_{1,1-2\alpha}\}, \tag{18}$$

where $\chi^2_{1,1-2\alpha}$ is the $(1 - 2\ \alpha)$ quantile of the chi-squared distribution with one degree of freedom (i.e., $\chi^2_{1,0.90} \approx 2.706$ for $\alpha = 0.05$; Pawitan 2001). This spatial model is used for the annual, summer, winter, and month-specific analyses.

3.3 Spatiotemporal Model for the Global Monthly Panel

The global monthly analysis uses the full balanced January–December panel jointly. Because the same rooftop locations are observed in every month, the data form a repeated space–time panel rather than a simple cross section. A joint spatiotemporal covariance model is used so the

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

uncertainty in the mean paired difference reflects both spatial redundancy within roofs and temporal persistence across months (Cressie and Wikle 2011).

Let $\boldsymbol{d}$ denote the vector of all monthly paired differences stacked by building, location, and month. The model is

$$\boldsymbol{d} \sim N\big(\mu\mathbf{1}, K(\theta)\big). \tag{19}$$

The covariance is block diagonal across buildings,

$$K(\theta) = \mathrm{diag}\{K_1(\theta), \ldots, K_B(\theta)\}, \tag{20}$$

where within each building

$$K_b(\theta) = \sigma^2\big\{\mathrm{R}_{t(\phi)} \otimes \mathrm{R}_{\mathrm{s,b}}(\lambda, \nu)\big\} + \tau^2 I. \tag{21}$$

Here $\mathrm{R}_t(\phi) \in \mathbb{R}^{T \times T}$ (with $T = 12$ months) is a first-order autoregressive, or AR(1), correlation matrix with entries

$$[\mathrm{R}_t(\phi)]_{ij} = \phi^{|i-j|}, \quad |\phi| < 1, \tag{22}$$

$\mathrm{R}_{s,b}(\lambda, \nu) \in \mathbb{R}^{n_b \times n_b}$ is the Matérn spatial correlation matrix for the fixed rooftop support of building b, and $\otimes$ denotes the Kronecker product, so $K_b \in \mathbb{R}^{Tn_b \times Tn_b}$ (Brockwell and Davis 1991; Stein 1999; Cressie and Wikle 2011). This separable structure captures temporal persistence across repeated months at the same location and spatial similarity among nearby rooftop locations on the same roof.

The GLS estimator of the global mean paired difference (Aitken 1936) is

$$\hat{\mu}_{\mathrm{st}} = \frac{\mathbf{1}^\top K(\theta)^{-1}\boldsymbol{d}}{\mathbf{1}^\top K(\theta)^{-1}\mathbf{1}}, \tag{23}$$

with model-based variance

$$Var(\hat{\mu}_{\mathrm{st}} \mid \theta) = (\mathbf{1}^\top K(\theta)^{-1}\mathbf{1})^{-1}. \tag{24}$$

Covariance parameters are estimated by maximum likelihood (ML). ML is used rather than REML for the global panel because the balanced, large-n structure of the monthly panel (n = 1,200 observations across 100 locations and 12 months) substantially reduces the finite-sample bias in ML variance estimates relative to the small within-slice samples; more importantly, ML is required for likelihood-ratio-based comparison across candidate covariance specifications during model selection (Pinheiro and Bates 2000, §2.4). The per-slice analyses use REML, where per-slice sample sizes are smaller and the bias reduction is more material (see Section 3.2).

In the balanced-panel setting, this joint model is preferable to month-by-month estimation followed by averaging because it uses the full spatiotemporal covariance structure to quantify effective information.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

The same separable spatiotemporal specification also fit separately within each building for the roof-specific analyses. For building b, the estimand is

$$\mu_b = \mathbb{E}[d(s,t) \mid b], \tag{25}$$

and inference uses the building-restricted analogue of the GLS estimator and covariance expression above, applied to the monthly panel for that roof alone.

3.4 Bootstrap Validation and Sensitivity Analyses

TOST decisions depend directly on confidence-interval width, so sensitivity analyses assess whether the main conclusions are robust to alternative uncertainty constructions.

For per-slice analyses, sensitivity is assessed with a building-level block bootstrap. Buildings are treated as resampling units, preserving within-building spatial structure while relaxing reliance on a specific parametric covariance form. The resulting percentile interval serves as a robustness check on the analytic spatial interval and is conservative with respect to the higher-level study design because the building is the natural unit of independent replication.

For building-level analyses, sensitivity is assessed by resampling rooftop locations within the building while retaining each selected location's full monthly series. This preserves the within-location temporal structure while relaxing reliance on the fitted separable covariance model for that roof.

For the global monthly analysis, sensitivity is assessed with a parametric bootstrap under the fitted separable spatiotemporal covariance model. After estimating $\hat{\mu}_{\mathrm{st}}$ and $\widehat{K}$, bootstrap replicates are generated as

$$\boldsymbol{d}^{*(r)} \sim N\left(\hat{u}_{\mathrm{st}}\mathbf{1}, \widehat{K}\right), \quad r = 1, \ldots, R \tag{26}$$

and the GLS mean is recomputed for each replicate. The resulting empirical distribution is then used to form a percentile interval (Efron and Tibshirani 1993; Davison and Hinkley 1997). This procedure checks whether the global equivalence conclusion is stable when uncertainty is propagated through repeated simulation under the fitted joint covariance model.

3.5 Decision Rule and Reporting

For every both the main results and sensitivity analysis, the decision rule is the same. Let $[L, U]$ denote the two-sided $(1 - 2\alpha)$ confidence interval for the relevant mean paired difference. Equivalence is declared only if

$$L > -\Delta \quad \text{and} \quad U < \Delta. \tag{27}$$

The estimand and equivalence margins are held fixed throughout the report. Any change in the final pass/fail result is thus attributable to the uncertainty model. not to a change in the target of inference. The results that follow report the estimated mean paired difference, its confidence interval, and the resulting equivalence decision for the global monthly panel, the individual reporting slices, and the building-level analyses.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

4 Results

4.1 Overall Equivalence

The global analysis uses the balanced monthly panel and the spatiotemporal TOST model described in Section 3. The mean-paired difference is

$$\hat{\mu} = 0.765 \text{ SAV}, \tag{28}$$

indicating these AI-SAVs report slightly lower SAV than these drone-SAVs on average. As summarized in Table 2. and shown in Figure 6, the model-based confidence interval is [0.644, 0.904] and the bootstrap sensitivity interval is [0.612, 0.938]. Both intervals are within the strictest equivalence region (-1, +1), so the global monthly analysis supports equivalence even at $\Delta = 1$.

This result is important because it places a tight dependence-aware bound on the mean drone-AI difference. Under the fitted spatiotemporal model, the average offset is confined to a narrow positive range well below one SAV unit. The bootstrap interval is wider than the model-based interval but leads to the same decision. At wider operational tolerances, $\Delta \in \{3, 5, 10\}$, equivalence follows immediately because both intervals remain inside the corresponding acceptance regions.

Table 2. Global equivalence results from the spatiotemporal uncertainty model using the balanced monthly panel

$\hat{\mu}$	Confidence Interval Low	Confidence Interval High	Bootstrap Confidence Interval Low	Bootstrap Confidence Interval High
0.765	0.644	0.904	0.612	0.938

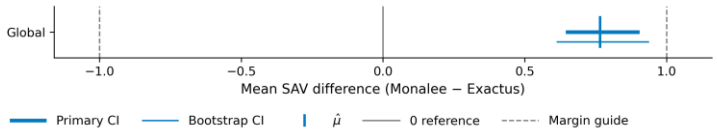

Figure 6. Global mean paired SAV difference for the monthly panel. Both the primary analytic confidence interval and the bootstrap sensitivity confidence interval are fully within the acceptance region.

4.2 Equivalence by Time Slice

Per-slice analyses provide a more detailed view of agreement across the calendar year and across the seasonal and annual reporting windows used operationally. Table 3. presents the spatial TOST results for each monthly, seasonal, and annual slice. The estimated mean paired difference is positive across all slices, ranging from 0.562 in December to 1.010 in winter, so the direction of the average AI-drone difference is stable through time. Because the comparison is relative rather than absolute, this persistent positive offset does not by itself identify which workflow is closer to ground truth and could reflect downward bias in AI-SAVs, upward bias in drone-SAVs, or a combination of smaller opposing biases in both methods.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

At operationally relevant margins, results are uniformly favorable. Every model-based interval in Table 3 lies within (-3, +3), so every month, both seasonal aggregates, and the annual aggregate satisfy equivalence at $\Delta = 3$, $\Delta = 5$, and $\Delta = 10$. The distinction among slices emerges only at the strict margin $\Delta = 1$. Under the spatial model, April, July, and August have confidence intervals fully contained within (-1, +1): [0.561, 0.933], [0.372, 0.901], and [0.392, 0.994], respectively. All other slices fail at $\Delta = 1$ because the upper confidence bound exceeds 1, though in most cases the departure is modest. January has an upper bound of 1.057, March has 1.069, September has 1.070, and December has 1.019. The largest upper bounds occur in October (1.219), summer (1.192), and winter (1.247).

These slice-specific results indicate the substantive signal is not a large practical disagreement between workflows but rather a combination of small positive mean offsets and slice-specific variation in effective information once spatial dependence is acknowledged. In every slice, the mean difference remains on the order of one SAV unit or less, and all slices satisfy equivalence under the wider margins used in prior SAV validation practice. The failures at $\Delta = 1$ are thus best interpreted as strict-threshold sensitivity to dependence-aware interval widening, not as evidence of materially different SAV estimates.

The bootstrap intervals also align closely with the model-based spatial intervals. Five slices have bootstrap intervals fully inside (-1, +1): March [0.744, 0.993], April [0.574, 0.961], May [0.501, 0.961], July [0.402, 0.829], and December [0.291, 0.853]. The remaining slices fail at $\Delta = 1$, because the upper bound lies only slightly above 1, as in January (1.028), February (1.034), and November (1.054). The largest bootstrap departures occur for summer (1.501), winter (1.248), and September (1.165). This close correspondence between model-based and bootstrap intervals supports the interpretation that the spatial covariance-aware inference is neither anomalously narrow nor implausibly conservative.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

Table 3. Per-time-slice equivalence results comparing AI-derived solar access values (SAVs) to drone-acquired SAV estimates. All values are in SAV units (percentage points). Slices whose model-based confidence interval lies fully within (-1, +1) satisfy the strict equivalence criterion; all slices satisfy equivalence at Δ = 3.

Time Slice	$\hat{\mu}$	Confidence Interval Low	Confidence Interval High	Bootstrap Confidence Interval Low	Bootstrap Confidence Interval High
January	0.970	0.880	1.057	0.935	1.028
February	0.700	0.341	1.062	0.454	1.034
March	0.881	0.698	1.069	0.744	0.993
April	0.749	0.561	0.933	0.574	0.961
May	0.742	0.454	1.043	0.501	0.961
June	0.776	0.453	1.090	0.533	1.099
July	0.625	0.372	0.901	0.402	0.829
August	0.676	0.392	0.994	0.370	1.026
September	0.868	0.486	1.070	0.645	1.165
October	0.793	0.380	1.219	0.398	1.082
November	0.974	0.873	1.074	0.894	1.054
December	0.562	0.122	1.019	0.291	0.853
Summer (May–Oct)	0.768	0.356	1.192	0.391	1.501
Winter (Nov–Apr)	1.010	0.782	1.247	0.829	1.248
Annual	0.931	0.780	1.073	0.857	1.051

4.3 Equivalence by Building

The building-level analysis assesses whether the global agreement is representative across roofs or reflects averaging over heterogeneous rooftop environments. Locations within a roof share obstruction geometry and shading regimes, whereas different roofs can vary substantially in layout, surrounding vegetation, and nearby structures. Examining equivalence separately by building determines whether the global result reflects broadly consistent agreement or averaging over localized departures. Figure 7 provides a visual summary of these roof-level intervals.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

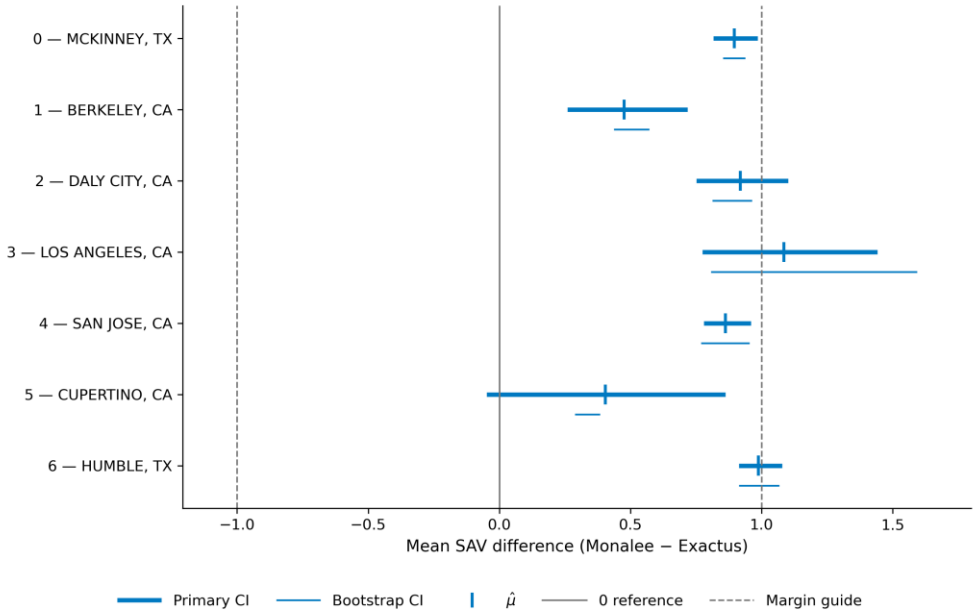

Figure 7. Building-level mean paired SAV differences with confidence intervals. Four of seven buildings have confidence intervals within the strict (-1, +1) SAV tolerance, whereas three extend modestly above the upper bound.

For each building, the mean paired difference is estimated using the building-restricted spatiotemporal framework described in Section 3. Table 4 reports the building-specific means and model-based confidence intervals from the spatiotemporal model, alongside bootstrap sensitivity intervals computed from a location-cluster resampling of the unadjusted (naive) mean difference. Because the bootstrap interval targets the naive estimator rather than the spatiotemporal fixed-effect estimate, the two intervals are not directly nested, and the model-based mean can fall outside the bootstrap interval, particularly for buildings with strong spatial dependence (e.g., Cupertino, CA). Estimated mean paired differences range from 0.404 to 1.085 SAV across buildings. All estimates are positive, indicating this drone-SAV dataset tends to produce slightly higher SAVs than this AI-SAV dataset on average, though the magnitude of the offset varies by roof.

Table 4. Building-level equivalence results using the spatiotemporal uncertainty model and bootstrap sensitivity intervals

Building	$\hat{\mu}$	Confidence Interval Low	Confidence Interval High	Bootstrap Confidence Interval Low	Bootstrap Confidence Interval High
McKinney, TX	0.895	0.816	0.985	0.853	0.937
Berkeley, CA	0.475	0.260	0.717	0.437	0.571
Daly City, CA	0.919	0.752	1.101	0.812	0.964
Los Angeles, CA	1.085	0.775	1.442	0.807	1.593

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

Building	$\hat{\mu}$	Confidence Interval Low	Confidence Interval High	Bootstrap Confidence Interval Low	Bootstrap Confidence Interval High
San Jose, CA	0.862	0.780	0.960	0.769	0.954
Cupertino, CA	0.404	-0.048	0.863	0.287	0.384
Humble, TX	0.988	0.914	1.078	0.914	1.068

At $\Delta = 1$, McKinney, Berkeley, San Jose, and Cupertino have model-based intervals within (-1, +1): [0.816, 0.985], [0.260, 0.717], [0.780, 0.960], and [-0.048, 0.863], respectively. Daly City, Los Angeles, and Humble fail at $\Delta = 1$ because the upper bounds exceed 1: [0.752, 1.101], [0.775, 1.442], and [0.914, 1.078], respectively. Los Angeles shows the largest departure from strict equivalence, with both the highest point estimate and the widest interval. Daly City and Humble are more naturally interpreted as borderline under the strictest criterion.

The bootstrap intervals provide similar results. McKinney, Berkeley, Daly City, San Jose, and Cupertino remain inside (-1, +1), with bootstrap intervals [0.853, 0.937], [0.437, 0.571], [0.812, 0.964], [0.769, 0.954], and [0.287, 0.384], respectively. Los Angeles and Humble fail under the bootstrap as well, with [0.807, 1.593] and [0.914, 1.068]. Daly City passes the bootstrap test (upper bound 0.964) but fails the model-based test (upper bound 1.101). This discordance arises because the parametric covariance model extrapolates the spatial covariance structure to configurations not well sampled on this relatively small roof, widening the model-based interval. The bootstrap, which resamples only the observed locations, avoids that extrapolation. Because the bootstrap is the pre-specified sensitivity check and does not require covariance extrapolation beyond the observed support, its interval is the more appropriate basis for the equivalence decision when the two intervals disagree for this reason; the model-based interval is reported alongside it for completeness and as the primary result for roofs where extrapolation is not a concern.

From an operational perspective, the portfolio-level monthly analysis supports equivalence even at $\Delta = 1$, but the roof-level view shows strict equivalence is not uniform across all buildings. Even so, all buildings satisfy equivalence at $\Delta = 3$, because the largest building-level upper bound (1.593) remains well below 3. The building analysis supports the same broad conclusion as the global and per-slice results: The average drone-AI difference is small in SAV units, whereas strict equivalence claims become sensitive to localized uncertainty in 3 of 7 roofs.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

5 Discussion

The results support a clear application-level conclusion. In the global monthly analysis, the dependence-aware spatiotemporal model yields $\hat{\mu} = 0.765$ SAV with confidence interval [0.644, 0.904] and bootstrap sensitivity interval [0.612, 0.938]. Both intervals lie within (-1, +1), and therefore, the global monthly panel supports equivalence even at the strictest margin considered here. The average AI – drone difference is small enough to satisfy a very strict practical tolerance after explicitly accounting for both spatial and temporal dependence.

The value of a remote SAV workflow is that it can support comparable early-stage design decisions with less field effort. Equivalence testing is the more relevant inferential framework here: It evaluates whether the mean discrepancy is small enough to be operationally negligible, not simply whether it differs from zero (Nangle et al. 2015; Schuirmann 1987; Lakens 2017). The global result indicates, at the portfolio level, this AI-derived SAV dataset satisfies that standard even under a strict ±1 SAV criterion.

The per-slice and building-level analyses provide necessary qualifications. All slices satisfy equivalence at $\Delta = 3$, $\Delta = 5$, and $\Delta = 10$. The stricter $\Delta = 1$ criterion is more selective: Only a subset of slices passes under the spatial model, whereas the rest fail because the upper confidence bound lies modestly above 1. The same pattern appears at the building level. Four of seven buildings satisfy equivalence at $\Delta = 1$; the three that do not show failures driven more by uncertainty near the decision boundary than by large shifts in the estimated mean. These AI-derived and drone-acquired datasets are consistently close in SAV space, but strict interchangeability is sensitive to the level of aggregation and to the local information content remaining after dependence is acknowledged.

That distinction matters for interpretation. A favorable global result does not imply uniform strict equivalence across every roof and reporting slice. Mean paired differences remain positive and generally small throughout the study, but confidence-interval width varies with rooftop geometry, sampling density, and dependence structure. The results support equivalence at the portfolio level and a more conditional conclusion locally: these AI-derived SAVs are operationally interchangeable with these drone-acquired SAVs on average, whereas a minority of roofs and slices lie close enough to the strict $\Delta = 1$ threshold that additional conservatism or local verification may be warranted.

The multi-margin design makes that interpretation more useful. At $\Delta = 10$, $\Delta = 5$, and $\Delta = 3$, the evidence is uniformly favorable. At $\Delta = 1$, the global monthly conclusion remains favorable, but heterogeneity becomes visible by slice and by building. The analysis thus shows the tolerance scale at which practical agreement is robust and the stricter scale at which local uncertainty begins to matter.

The comparison with earlier SAV validation work is also informative. Prior studies established the usefulness of treating remote shading validation as an equivalence problem rather than as a standard difference-detection problem (Nangle et al. 2015). This study reaches the same broad practical conclusion but in a richer matched dataset with repeated monthly observations and multiple rooftop locations per building. That richer design sharpens the methodological requirement: Because nearby rooftop points and repeated monthly observations are not

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

independent replicates, uncertainty quantification must reflect spatial and spatiotemporal redundancy rather than row counts alone (Cressie 1993; Cressie and Wikle 2011). The dependence-aware TOST workflow used here addresses that requirement directly.

This point strengthens the favorable results. Dependence-aware intervals are generally wider than intervals obtained under independence assumptions, and wider intervals make equivalence harder to declare. The global monthly panel still supports equivalence at $\Delta = 1$ after accounting for joint space–time dependence. This provides a stronger basis for operational confidence than an IID analysis would. Conversely, the near-threshold failures at $\Delta = 1$ in some slices and buildings illustrate why dependence modeling is important: When the target tolerance is strict, realistic uncertainty quantification can materially change whether equivalence is declared.

The bootstrap sensitivity analyses support the same overall interpretation. At the global level, the bootstrap interval remains inside (-1, +1). At the per-slice and building levels, the bootstrap results preserve the broad structure of the findings: All units pass comfortably at wider margins and only a subset pass at $\Delta = 1$. Differences between model-based and bootstrap conclusions are concentrated in borderline cases, which is the expected behavior when intervals lie near a decision threshold. These AI-derived SAVs tend to report slightly lower SAVs than these drone-acquired SAVs on average, and the magnitude of that offset is usually small in operational terms. This signal is stable across uncertainty constructions. Where the two interval types disagree for a borderline unit — as with Daly City — the bootstrap interval is treated as the operative result because it conditions only on observed location configurations and requires no covariance extrapolation; this interpretive hierarchy was established as part of the sensitivity-analysis design and is not a post-hoc accommodation to a particular building's result.

Overall, this study supports two conclusions. First, these AI-derived and drone-acquired SAVs show strong practical agreement across this matched rooftop dataset, especially at operationally realistic margins. Second, dependence-aware uncertainty quantification materially improves the interpretation of that agreement by distinguishing between robust global equivalence and localized near-threshold cases.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

6 Limitations and Future Work

Several limitations are important for interpreting the scope of these results.

First, the study is based on seven buildings and 100 rooftop locations in California and Texas. The repeated monthly panel provides substantial within-study information, but the number of higher-level sampling units remains modest. Building-level results should be interpreted as evidence about the roofs represented here, not as a complete characterization of all residential rooftop conditions.

Second, the drone-acquisition method is treated as the reference workflow, but it is not ground truth. The comparison addresses practical agreement between a field-survey drone-acquired workflow and an AI-assisted remote workflow, not absolute accuracy against directly observed irradiance or shading measurements. Future work could strengthen the validation chain by linking both workflows to an external observational benchmark where such data are available.

Third, the dependence-aware models used here are simplified representations of the underlying rooftop shading process. The per-slice analyses assume within-building spatial dependence and between-building independence, whereas the global monthly analysis uses a separable spatiotemporal covariance model with AR(1) temporal structure and Matérn spatial structure (Cressie 1993; Stein 1999; Cressie and Wikle 2011). These are reasonable choices for the present data, but more flexible mean models, nonseparable spatiotemporal covariance models, or roof-specific covariance parameterizations could change interval width, especially for buildings near the strictest equivalence threshold.

Fourth, the strongest conclusions depend on the equivalence margin. The results are highly stable at $\Delta = 3$, $\Delta = 5$, and $\Delta = 10$, whereas heterogeneity becomes more visible at $\Delta = 1$. Future work would benefit from linking SAV margins more explicitly to downstream design consequences such as panel placement, expected production, system sizing, and proposal quality. That would make the practical meaning of a given SAV tolerance more transparent.

Fifth, the analysis is based on one temporal resolution and one reporting structure: a complete monthly panel together with annual and seasonal aggregates. Other operational datasets may involve missing months, irregular observation windows, different seasonal definitions, or asynchronously updated rooftop models. Extending the workflow to irregular or partially observed panels would broaden its applicability.

These limitations suggest several natural directions for future work. The first is expanded validation on a larger and more diverse building sample spanning additional roof types, climates, obstruction regimes, and urban contexts. A broader sample would improve precision at the building level and clarify whether the borderline cases observed here are isolated or recurring.

A second direction is extension of the equivalence target beyond SAV itself. Future work could also examine equivalence for downstream quantities such as production estimates, panel-level shading losses, or proposal-facing performance metrics. That would help determine whether small SAV differences are amplified, attenuated, or practically canceled in later stages of the design workflow.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

A third direction is continued methodological development of dependence-aware equivalence testing for rooftop and other geospatial design datasets. Additional work could explore more flexible covariance structures, alternative bootstrap designs (Cameron, Gelbach, and Miller 2008; Lahiri 2003), and hierarchical formulations that pool information across buildings while preserving roof-specific inference (Harville 1977).

Finally, future work should consider operationalizing dependence-aware equivalence testing as part of routine quality assurance and model-governance practice. As remote-design workflows evolve, new geographies are added, or acquisition pipelines change, repeated matched validation with prespecified SAV margins and dependence-aware uncertainty quantification would provide a principled way to monitor whether practical agreement remains stable across space, time, and deployment settings.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

7 Conclusions

This report evaluated whether AI-derived SAVs are practically equivalent to drone-acquired SAVs in a matched residential rooftop dataset spanning seven buildings, 100 rooftop locations, and monthly, seasonal, and annual reporting slices. The analysis used the TOST framework with dependence-aware uncertainty quantification, so equivalence decisions reflected the clustered, spatial, and spatiotemporal structure of the data.

The primary result is favorable. In the global monthly analysis, the estimated mean paired difference was positive but small ($\hat{\mu} = 0.765$ SAV), and both the model-based spatiotemporal confidence interval [0.644, 0.904] and the bootstrap sensitivity interval [0.612, 0.938] lay fully within the strictest equivalence region (-1, +1). These AI-derived and drone-acquired SAVs are practically equivalent under dependence-aware inference even at a stringent SAV tolerance.

The more detailed analyses refine that conclusion. Equivalence was consistently supported at Δ = 3, Δ = 5, and Δ = 10 across monthly, seasonal, annual, and building-level evaluations. At the strictest margin Δ = 1, equivalence became sensitive to slice- and roof-level uncertainty, with some intervals extending modestly above the upper bound. These near-threshold cases indicate local uncertainty matters when the acceptance criterion is strict, even when overall agreement remains strong.

Two conclusions follow. First, AI-based workflows show strong practical agreement with in-field workflows in SAV-acquisition for the matched rooftops and reporting windows examined here. Second, dependence-aware uncertainty quantification materially improves the credibility and interpretability of that conclusion by distinguishing between robust global equivalence and localized borderline cases. For operational SAV validation, that is the appropriate inferential standard, not simply whether two workflows differ but whether any difference is small enough to be negligible under realistic assumptions about the information content of the data.

More broadly, this study shows equivalence testing provides a useful framework for validating remote solar-design workflows when the objective is practical interchangeability rather than exact numerical equality. The contribution is both applied and methodological: It supports remote, AI-derived SAVs as a practically comparable SAV workflow in this case study and demonstrates the value of dependence-aware TOST for matched rooftop datasets with repeated spatial and temporal structure.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

References

Aitken, A. C. 1936. "On Least Squares and Linear Combination of Observations." *Proceedings of the Royal Society of Edinburgh* 55: 42–48.

Artemis. 2026. *Artemis – Energy & Roofing Software*. Accessed April 2026. https://artemispower.com/.

Brockwell, Peter J., and Richard A. Davis. 1991. *Time Series: Theory and Methods*. 2nd ed. New York: Springer.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90(3): 414–427.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50(2): 317–372.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury.

Cliff, A. D., and J. K. Ord. 1981. *Spatial Processes: Models and Applications*. London: Pion.

Cressie, Noel A. C. 1993. *Statistics for Spatial Data*. Rev. ed. New York: John Wiley & Sons.

Cressie, Noel A. C., and Christopher K. Wikle. 2011. *Statistics for Spatio-Temporal Data*. Hoboken, NJ: John Wiley & Sons.

Davison, Anthony C., and David V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Exactus Energy. 2026. *Residential Site Surveys*. Accessed April 2026. https://www.exactusenergy.com/residential.

Friedman, Barry, Kristen Ardani, David Feldman, Ryan Citron, Robert Margolis, and Jarett Zuboy. 2013. *Benchmarking Non-Hardware Balance-of-System (Soft) Costs for U.S. Photovoltaic Systems, Using a Bottom-Up Approach and Installer Survey - Second Edition*. Golden, CO: National Renewable Energy Laboratory. NREL/TP-6A20-60412.

Guttorp, Peter, and Tilmann Gneiting. 2006. "Studies in the History of Probability and Statistics XLIX. On the Matérn Correlation Family." *Biometrika* 93(4): 989–995.

Harville, David A. 1977. "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems." *Journal of the American Statistical Association* 72(358): 320–338.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

Hettinger, Dylan. 2026. pyTOST: Dependence-Aware Two One-Sided Tests for Equivalence Under IID, Clustered, Spatial, and Spatiotemporal Uncertainty. Computer software. Zenodo. https://doi.org/10.5281/zenodo.19393895.

Hofierka, Jaroslav, and Ján Kaňuk. 2009. "Assessment of Photovoltaic Potential in Urban Areas Using Open-Source Solar Radiation Tools." *Renewable Energy* 34(10): 2206–2214.

Jakubiec, J. Alstan, and Christoph F. Reinhart. 2013. "A Method for Predicting City-Wide Electricity Gains from Photovoltaic Panels Based on LiDAR." *Solar Energy* 93: 127–143.

Kanters, Jouri, and Maria Wall. 2014. "The Impact of Urban Design Decisions on Net Zero Energy Solar Buildings in Sweden." *Urban, Planning and Transport Research* 2(1): 312–332.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley & Sons.

Lahiri, S. N. 2003. *Resampling Methods for Dependent Data*. New York: Springer.

Lakens, Daniël. 2017. "Equivalence Tests: A Practical Primer for t-Tests, Correlations, and Meta-Analyses." *Social Psychological and Personality Science* 8(4): 355–362.

National Renewable Energy Laboratory. 2021. *Comparing the OpenSolar Remote Shading Analysis Tool to On-Site Shading Measurements*. Golden, CO: National Renewable Energy Laboratory. NREL/FS-5R00-81316.

Nangle, John, Jeff Dean, and Otto Van Geet. 2015. *Analysis of Solar Census Remote Solar Access Value Calculation Methodology*. Golden, CO: National Renewable Energy Laboratory. NREL/TP-7A40-63098. https://doi.org/10.2172/1176739.

Nguyen, Ha T., and Joshua M. Pearce. 2012. "Incorporating Shading Losses in Solar Photovoltaic Potential Assessment at the Municipal Scale." *Solar Energy* 86(5): 1245–1260.

O'Shaughnessy, Eric, Gregory F. Nemet, Jacquelyn Pless, and Robert Margolis. 2019. "Addressing the Soft Cost Challenge in U.S. Small-Scale Solar PV System Pricing." *Energy Policy* 134: 110956.

O'Shaughnessy, Eric, Shiyuan Dong, Jeffrey J. Cook, Jesse Cruce, Kristen Ardani, Emily Fekete, and Robert Margolis. 2022. "Effects of Local Permitting and Interconnection Requirements on Solar PV Installation Durations." *Energy Policy* 161: 112734.

Patterson, H. D., and R. Thompson. 1971. "Recovery of Inter-Block Information When Block Sizes Are Unequal." *Biometrika* 58(3): 545–554.

Pawitan, Yudi. 2001. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.

Pinheiro, José C., and Douglas M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.

Ramasamy, Vignesh, Jarett Zuboy, Michael Woodhouse, Eric O'Shaughnessy, David Feldman, Jal Desai, Andy Walker, Robert Margolis, and Paul Basore. 2023. *U.S. Solar Photovoltaic System and Energy Storage Cost Benchmarks, With Minimum Sustainable Price Analysis: Q1 2023*. Golden, CO: National Renewable Energy Laboratory. NREL/TP-7A40-87303.

Schuirmann, Donald J. 1987. "A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability." *Journal of Pharmacokinetics and Biopharmaceutics* 15(6): 657–680.

Simon, Joseph, and Jesse Dean. 2015. *Evaluation of the Aurora Application Shade Measurement Accuracy*. Golden, CO: National Renewable Energy Laboratory. NREL/FS-7A40-65558.

Stein, Michael L. 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.

Wellek, Stefan. 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.

This report is available at no cost from the National Laboratory of the Rockies (NLR) at www.nlr.gov/publications.