CALIBRATION OF SOLAR IRRADIANCE DATA USING OPTIMAL TRANSPORT

Lucio Carlos Pimentel Paiva[1], Luiz Filipe Santana Oliveira Reis[2], Joel Kalil da Silva Pontes[1], Luis Guilherme Bastos de Castro[1], Rodrigo Raimundo Freitas Santos[1] and Leticia Bezerra de Vasconcelos Vieira[1]

[1] CDV Desenvolvimento, Fortaleza (Brazil)

[2] CDV Desenvolvimento, Petrolina (Brazil)

Abstract

This study investigates the use of optimal transport (OT) for the site adaptation of satellite-derived solar irradiance data. Using one year of ground measurements from three distinct locations for calibration and a subsequent year for validation, we evaluate the effectiveness of OT for correcting the joint distribution of the clear sky index (k_t) and diffuse fraction (k_d). We compare the performance of OT against two standard techniques: quantile mapping (QM) applied to k_t and k_d independently, and linear regression (LR) applied to global horizontal irradiance (GHI) and diffuse horizontal irradiance (DHI). DHI estimates derived from decomposition models applied to LR-corrected GHI are also assessed. The results demonstrate that OT provides a more robust calibration, better preserving the physical relationship between k_t and k_d compared to the other methods. Additionally, our findings show that both the original satellite data and the tested decomposition models systematically overestimate DHI at the studied locations.

Keywords: optimal transport, clear sky index, diffuse fraction, ground measurements, site adaptation

1. Introduction

Calibration of satellite-derived solar irradiance data against ground measurements is essential for reducing uncertainty in resource assessment and system modeling. The current study evaluates a calibration approach using optimal transport (OT). Data from three ground stations equipped with secondary standard pyranometers were used as reference values against a commercial satellite-derived irradiance dataset - Solargis. For each station, a one-year period was used to train calibration models, including OT, quantile mapping (QM), and linear regression (LR). These models were then applied to the satellite data for the subsequent year and validated against the corresponding ground measurements. Additionally, the DIRINDEX (Perez et al., 2002) and Erbs-Driesse (Driesse et al., 2024) decomposition models were applied to the GHI predictions from the LR model for comparison.

2. Methodology

2.1. Data description and pre-processing

Each of the meteorological stations is equipped with two secondary standard pyranometers for GHI and a SPN1 Delta T sunshine pyranometer for DHI measurements. Measurements were acquired every second and logged every minute for stations 1 and 2, and every ten minutes for station 3. Basic information on ground stations, such as location, pyranometer models and measurement campaign duration can be found on Tab. 1. For quality control, indicators from Tab. 2 were used prior to manual data inspection. On Tab. 2, ETN is the extraterrestrial irradiance at normal incidence, obtained from pvlib (Anderson et al, 2023), SZA is the solar zenith angle, GHI is the global horizontal irradiance at ground level, DHI is the diffuse horizontal irradiance

10.18086/swc.2025.04.09 Available at http://proceedings.ises.org

at ground level, DNI is the (calculated) direct normal irradiance, ALT is the station altitude above sea level and the normalized quantities k_t, k_n (clear sky indices) and k_d (diffuse fraction) are defined in equations 1-3. Diffuse fraction values were calculated from global and diffuse irradiance channels of the SPN1, after calibration against the main pyranometer. The solar zenith angle SZA used in this work is given by the complementary value of the solar elevation angle SEA (eq. 4), which in turn is a modification of the solar elevation given by the SPA algorithm (Reda and Andreas, 2003) as implemented on pvlib, noted SEA' on eq. 5. It is worth noticing that these modified values provide better numerical stability on DNI calculations and have a good agreement with PVsyst solar elevation angles for the sites considered. Quality control conditions from group 1 are a proxy for possible k_t consistency issues. Group 2 indicates extremely rare irradiance values. Group 3 filter data points that are likely to be above physical limits. Group 4 refers to persistence criteria, i.e the expected relationship between averages and standard deviations. Groups 1, 2 and 3 are based on Forstinger et al (2023), while group 4 is based on Journée and Bertrand (2011). Ultimately, automatic data flags issued from Tab. 2 conditions and station maintenance reports were used by the data analyst to remove invalid data. Other than these calculated conditions, data points from redundant units of the same instrument (e.g main pyranometers) for which redundant measurements diverged beyond the expected instrument uncertainty were considered outliers and thus removed.

$$k_n = DNI/ETN \quad \text{(eq. 1)}$$

$$k_t = GHI/(ETN \cdot cos(SZA)) \quad \text{(eq. 2)}$$

$$k_d = DHI/GHI \quad \text{(eq. 3)}$$

$$SZA = 90° - SEA \quad \text{(eq. 4)}$$

$$SEA = 0,\ if\ SEA' < -7°$$

$$SEA = SEA'/2 + 3.5°,\ if\ -7° \le SEA' < 7° \quad \text{(eq. 5)}$$

$$SEA = SEA',\ if\ SEA' > 7°$$

Tab. 1: Ground stations information - location, measurement campaign period, instruments and train/test periods

ID	Meas. Start	Meas. End	Pyranometers	Lat	Lon	Train period	Test period
1	Aug/21	Active	CMP10, SPN1	-21.5º	-53.8º	2022	2023
2	May/21	Jul/24	SR20-D2, SPN1	-3.8º	-41.1º	2022	2023
3	Jun/22	Active	CMP10, SPN1	-5.7º	-36.1º	Jun/22 - May/23	Jun/23 - May/24

Table 2: Quality Control Indicators

Group	Condition	Domain
1	$k_n < k_t$	$GHI > 50\,W/m^2,\ k_n > 0,\ k_t > 0$
1	$k_n < (1100 + 0.03 \cdot ALT)/ETN$	$GHI > 50\,W/m^2,\ k_n > 0$
1	$k_t < 1.35$	$GHI > 50\,W/m^2,\ k_t > 0$
1	$k_d < 1.05$	$SZA < 75°,\ GHI > 50\,W/m^2,\ k_d > 0$
1	$k_d < 1.1$	$SZA \ge 75°,\ GHI > 50\,W/m^2,\ k_d > 0$
1	$k_d < 0.96$	$SZA < 85°,\ GHI > 150\,W/m^2,\ k_d > 0,\ k_t > 0.6$
2	$-2 \le GHI \le 1.5 \cdot ETN \cdot cos(SZA)^{1.2} + 50$	All data
2	$-2 \le DHI \le 0.75 \cdot ETN \cdot cos(SZA)^{1.2} + 30$	All data
2	$-2 \le DNI \le 0.95 \cdot ETN \cdot cos(SZA)^{0.2} + 10$	All data
3	$-4 \le GHI \le 1.5 \cdot ETN \cdot cos(SZA)^{1.2} + 100$	All data

3	$-4 \leq DHI \leq 0.75 \cdot ETN \cdot cos(SZA)^{1.2} + 50$	All data
3	$-4 \leq DNI \leq ETN$	All data
4	$\frac{1}{8}\mu(\frac{GHI}{ETN}) \leq \sigma(\frac{GHI}{ETN}) \leq 0.35$	Sunrise to sunset (calculated by pvlib)
4	$\frac{1}{6}\mu(\frac{DHI}{ETN}) \leq \sigma(\frac{DHI}{ETN}) \leq 0.2$	Sunrise to sunset (calculated by pvlib)

GHI values from SPN1 were obtained by a linear regression against the main pyranometers GHI readings (eq. 6), after discarding outliers (differences above 50 W/m2) and low zenith (SZA > 80°) values. The same GHI regression coefficients were applied to the SPN1 DHI channel, as recommended by the SPN1 manufacturer. After this calibration, DHI values given by the SPN1 were clipped by its GHI values, and all GHI values for zenith angles exceeding 89.5° were overridden by the main pyranometer. Points with GHI < 50 W/m2 were assumed to have $k_d = 1$ regardless of SPN1 measurements. All k_t and k_d values were constrained to the unit interval and all negative irradiance readings were clipped to zero. After this, the series was resampled to hourly resolution.

$$\hat{GHI}_{main} = \alpha_{SPN1} \cdot GHI_{SPN1} + \beta_{SPN1} \quad \text{(eq. 6)}$$

Remaining gaps of GHI and DHI were filled by medians of grouped non-null values per hour and month. The resulting data was split into a train dataset, in order to fit the calibration techniques, and a test dataset, in order to assess its generalization capabilities. The train period corresponds to the year of 2022 for stations 1 and 2, and Jun/22 through May/23 for station 3. The test dataset corresponds to the year of 2023 for stations 1 and 2, and Jun/23 through May/24 for station 3. For the corresponding Solargis series, the pre-processing was minimal, including just the solar position values and k_t, k_d calculations.

2.2. Site calibration assessment

Three calibration approaches were compared: linear regression (LR), quantile mapping (QM) and optimal transport (OT). The baseline for predictions was the direct comparison between satellite and ground data.

The linear regression (LR) model, in its simplest form, can be described by eq. 7, where $\hat{Y}$ is the estimated value of the variable of interest Y, α and β are constants.

$$\hat{Y} = \alpha \cdot X + \beta \quad \text{(eq. 7)}$$

Quantile mapping (QM) consists of mapping the quantiles of one distribution into another, resulting in a non-linear transformation. If the CDFs of the source and target distributions are, respectively, F_1 and F_2, then the image of a point belonging to the source dataset is given by eq. 8 (Gudmundsson et al, 2012). The inverse of the CDF is called the quantile function. Using an approach similar to that of Piani et al (2010), the points from the source distribution are matched to their images to construct a non-linear mapping; these pairs are used to construct a piecewise linear function that can be applied to new data points. Source and target distributions are determined empirically from the train data.

$$\hat{Y} = F_2^{-1}(F_1(X)) \quad \text{(eq. 8)}$$

Optimal transport (OT) was first conceived as a problem of minimization of mass transport by Gaspard Monge; however, most authors refer to the Kantorovich formulation because it is more tractable (Peyré and Cuturi, 2019) from analytical and computational standpoints. In particular, eq. 8 provides the optimal transport map between two one-dimensional distributions if the source distribution does not give mass to discrete points (Thorpe, 2018) for the most basic version of the Kantorovich problem. This work uses the problem formulation given by Perrot et al (2016), as implemented by the function “ot.da.MappingTransport” of the python POT package (Flamary et al, 2021),described by eq. 9. This formulation seeks an optimal mapping L by minimizing a cost function with three terms. The first one can be interpreted as a barycentric cost, the second is the standard optimal transport loss and the third is a regularization term. The dimension or number of features is noted as d (d=2 in this work); n_s (resp. n_t) represents the number of source (target) points, X_s (X_t) is a n_s x d (n_t x d) matrix of source (target) training points, α is the n_s x d the learned coefficient matrix, C is the transport cost matrix, μ is the linear OT cost weight, η is the regularization weight, K is the n_s x n_s gaussian kernel, σ is the kernel bandwidth and T is called the transport plan matrix, although the image of source points is not directly given by T - it is given by $L(X_s)$. The image of new data points requires the re-calculation of K, but the α matrix is kept. The subscript F stands for Frobenius norm in the first term, inner product in the second term. The subscript H stands for Reproducing Kernel Hilbert Space. More information on the mathematical formulation is available in Perrot et al (2016).

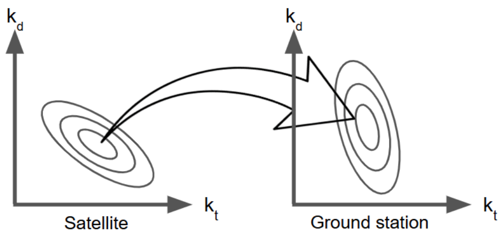

Figure 1: Optimal transport of (k_t, k_d) pairs.

$$min_{T, L \in H} || L(X_s) - n_s T X_t ||^2_F + \mu \langle T, C \rangle_F + \eta ||L||^2_H$$

$$L(X_s) = K\alpha,\ \alpha = (K + \eta I)^{-1}(n_s T X_t),\ K_{ij} = exp(-\frac{||X_{si} - X_{sj}||}{2\sigma^2}) \quad \text{(eq. 9)}$$

The site adaptation techniques are listed on Table 3. Whenever the regression target is different from the output variable, it means that the output is calculated by post-processing, e.g the estimated GHI is calculated from k_t. Furthermore, DHI is limited to GHI and (k_t, k_d) pairs are constrained to the unit square after being predicted.

Table 3: Site adjustment techniques summary.

#	Description	Features	Target	Output
0	Solargis, as-is	Satellite GHI, DHI	GHI, DHI	GHI, DHI
1	Linear regression	Satellite GHI	GHI	GHI
2	Quantile mapping	Satellite k_t	k_t	GHI
3	Quantile mapping	Satellite k_d	k_d	DHI
4	Optimal transport	Satellite k_t, k_d	k_t, k_d	GHI, DHI
5	Dirindex	Linear regression GHI output	DHI	DHI
6	Erbs-Driesse	Linear regression GHI output	DHI	DHI

3. Results

Performance was evaluated using the yearly bias, alongside R^2 and RMSE calculated from monthly totals (tables 4 and 5). RMSE values are given in kWh/m^2. While QM achieved a closer fit to the marginal distributions of k_t and k_d during training (Fig. 2), OT demonstrated a good agreement of GHI and DHI predictions and a superior ability to capture the k_d-k_t relationship when compared to QM and LR, suggesting better generalization. In particular, a visual inspection of Fig. 2 suggests that the overall shape and sample distribution in the (k_t, k_d) space is better reproduced. As a final remark, both Solargis and the tested decomposition models tend to overestimate DHI compared to ground measurements across the sites. Notice on Fig. 2 the existence of two remarkable point groups, one with $k_d = 0$ and other with $k_d = 1$. This is an artifact of linear regressions combined with post-processing clipping and reveals the limitations of linear regressions with respect to keeping the physically expected distribution of points in the (k_t, k_d) space.

Tab. 4: GHI site adaptation results

Station	Site adjustment	Train dataset			Test dataset		
		Yearly bias	R2	RMSE	Yearly bias	R2	RMSE
1	Solargis, as-is	0.33%	0.995	2.39	-0.20%	0.991	2.57
1	Linear reg.	0.10%	0.996	2.18	-0.40%	0.990	2.61
1	Quantile map.	-0.14%	0.994	2.61	-0.91%	0.981	3.65
1	OT	0.19%	0.996	2.09	-0.30%	0.980	3.80
2	Solargis, as-is	-0.36%	0.963	3.74	0.93%	0.954	4.51
2	Linear reg.	-0.07%	0.954	4.18	1.37%	0.936	5.34
2	Quantile map.	-0.81%	0.904	6.06	0.64%	0.923	5.86
2	OT	-0.15%	0.912	5.79	1.40%	0.922	5.89
3	Solargis, as-is	-2.54%	0.874	5.75	-3.39%	0.838	7.19
3	Linear reg.	1.23%	0.931	4.26	-0.59%	0.941	4.34
3	Quantile map.	0.16%	0.952	3.55	-0.78%	0.926	4.86
3	OT	0.77%	0.931	4.26	-0.51%	0.892	5.86

Tab. 5: DHI site adaptation results

Station	Site adjustment	Train dataset			Test dataset		
		Yearly bias	R2	RMSE	Yearly bias	R2	RMSE
1	Solargis, as-is	13.67%	0.665	7.97	13.68%	0.776	7.78
1	Linear reg.	0.01%	0.908	4.18	1.17%	0.959	3.32
1	Quantile map.	-2.32%	0.940	3.38	-2.48%	0.951	3.66
1	OT	-0.20%	0.893	4.51	-0.44%	0.940	4.03
1	Dirindex	7.29%	0.724	7.24	8.82%	0.855	6.27
1	Erbs-Driesse	19.96%	0.364	10.99	18.09%	0.626	10.06
2	Solargis, as-is	13.77%	0.438	10.13	15.59%	0.489	9.76
2	Linear reg.	-0.46%	0.842	5.38	0.15%	0.924	3.76
2	Quantile map.	-3.37%	0.872	4.84	-3.87%	0.926	3.71
2	OT	-0.34%	0.889	4.50	-0.94%	0.955	2.89
2	Dirindex	12.91%	0.597	8.58	10.92%	0.745	6.90
2	Erbs-Driesse	14.25%	0.396	10.50	17.07%	0.350	11.01
3	Solargis, as-is	11.70%	-0.022	9.63	9.86%	-0.778	7.97
3	Linear reg.	3.15%	0.645	5.68	-1.28%	0.479	4.31
3	Quantile map.	-2.86%	0.713	5.10	-4.37%	0.240	5.21
3	OT	-0.40%	0.660	5.56	-2.80%	0.168	5.45
3	Dirindex	8.35%	0.591	6.10	7.63%	0.014	5.93
3	Erbs-Driesse	11.53%	0.209	8.48	8.98%	-0.269	6.73

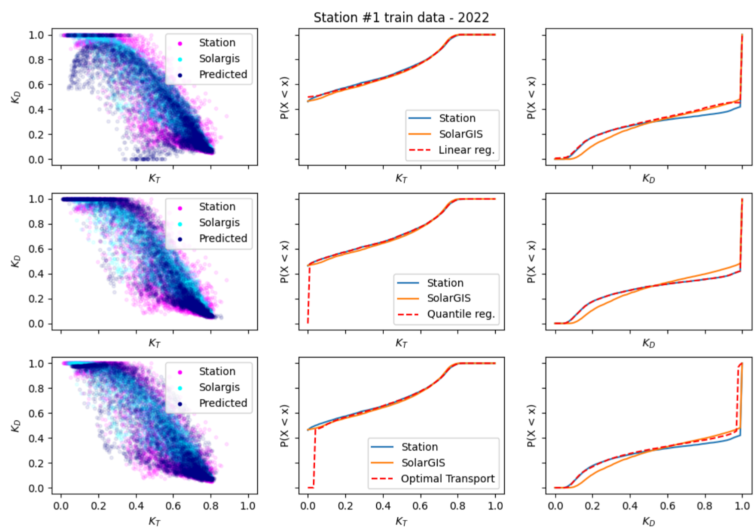

Fig. 2: Joint k_d, k_t dispersion and marginal CDFs for station 1 train data. Similar trends were observed for stations 2 and 3.

4. **References**

Anderson, K., Hansen, C., Holmgren, W., Jensen, A., Mikofski, M., and Driesse, A. “pvlib python: 2023 project update.” Journal of Open Source Software, 8(92), 5994, (2023). DOI: 10.21105/joss.05994

Driesse, A., Jensen, A., Perez, R., 2024. A Continuous form of the Perez diffuse sky model for forward and reverse transposition. Solar Energy vol. 267. DOI: 10.1016/j.solener.2023.112093

Flamary, R.; Courty, N.; Gramfort, A.; Alaya, M. Z.; Boisbunon, A.; Chambon, S.; Chapel, L.; Corenflos, A.; Fatras, K.; Fournier, N.; Gautheron, L.; Gayraud, N. T. H.; Janati, H.; Rakotomamonjy, A.; Redko, I.; Rolet, A.; Schutz, A.; Seguy, V.; Sutherland, D. J.; Tavenard, R.; Tong, A.; Vayer, T. (2021). POT Python Optimal Transport library. Journal of Machine Learning Research, 22(78), 1–8.

Forstinger, A., et al., 2023. Worldwide Benchmark of Modelled Solar Irradiance Data. IEA PVPS Task 16, Report IEA-PVPS T16-05:2023.

Gudmundsson, L., Bremnes, J.B., Haugen, J.E., and Skaugen, T.E., 2012. Technical note: Downscaling RCM precipitation to the station scale using quantile mapping–a comparison of methods. Hydrology and Earth System Sciences, 16(9), 3383-3390. doi:10.5194/hess-16-3383-2012.

Journée, M., and Bertrand, C., 2011. Quality control of solar radiation data within the RMIB solar measurements network. *Solar Energy*, 85, 72-86. doi:10.1016/j.solener.2010.10.021.

Perez, R., Ineichen, P., Moore, K., Kmiecik, M., Chain, C., George, R., & Vignola, F. (2002). A new operational model for satellite-derived irradiances: description and validation. Solar Energy, 73(5), 307-317.

Perrot, M., Flamary, R., Courty, N., and Habrard, A., 2016. Mapping Estimation for Discrete Optimal Transport. In: *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain.

Peyré, G., and Cuturi, M., 2019. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6), 355-607. doi:10.1561/2200000073.

Piani, C., Weedon, G.P., Best, M., Gomes, S.M., Viterbo, P., Hagemann, S., and Haerter, J.O., 2010. Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models. Journal of Hydrology, 395, 199-215. doi:10.1016/j.jhydrol.2010.10.024.

Reda, I.; Andreas, A. (2003). Solar Position Algorithm for Solar Radiation Applications. 55 pp.; NREL Report No. TP-560-34302, Revised January 2008.

Thorpe, M., 2018. Introduction to Optimal Transport. Lecture Notes, University of Cambridge. Available at: https://www.damtp.cam.ac.uk/research/cia/files/teaching/Optimal_Transport_Notes.pdf