https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

Configuration of climatological limits for surface radiation measurement quality control: A global assessment using a novel radiation climate classification

Zhiwen Wang[1], Yun Chen[2,3], Dazhi Yang[1,3], Hongrong Shi[4], Yanbo Shen[2,3], and Xiang'ao Xia[4,5]

[1]School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin, Heilongjiang, China
[2]Public Meteorological Service Centre, China Meteorological Administration, Beijing, China
[3]Key Laboratory of Energy Meteorology, China Meteorological Administration, Beijing, China
[4]Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China
[5]University of Chinese Academy of Sciences, Beijing, China

Correspondence: Dazhi Yang (yangdazhi.nus@gmail.com) and Hongrong Shi (shihrong@mail.iap.ac.cn)

Abstract. Quality control (QC) of ground-based solar radiation measurements is fundamental to ensuring the integrity of surface energy balance and climatological studies. The extremely rare limit (ERL) test, a widely implemented QC standard, is frequently noted for being overly conservative, often failing to isolate subtle instrumental or environmental anomalies. To improve QC tightness and sensitivity, this study presents a data-driven framework for configuring regime-specific climatological limits. Diverging from traditional climate classifications that do not directly account for radiative variability, we define seven distinct radiation regimes through unsupervised learning, utilizing principal component analysis and hierarchical clustering. For each identified regime, optimal test coefficients are established via a machine-learning-based optimization strategy. Specifically, we maximize the F_1 score by benchmarking the climatological limit test against an isolation forest outlier detection model. Validation using global measurements from the Baseline Surface Radiation Network demonstrates that the proposed regional limits provide a significantly tighter fit to observed data distributions compared to the original global ERL thresholds. This methodology offers a scalable and automated approach to regionalizing QC procedures, substantially enhancing the precision of global radiation monitoring networks.

1 Introduction

Surface solar radiation is an important variable in both atmospheric science and solar energy engineering (Yang and Kleissl, 2024). Radiation information is commonly obtained from three sources: ground-based measurements, satellite remote sensing, and numerical modeling (Yang et al., 2022). Among these options, ground-based measurements provide the highest accuracy and are therefore widely used to evaluate gridded products (Wandji Nyamsi et al., 2023; Elias et al., 2024) and to provide training targets for radiation models (Wiltink et al., 2025; Song et al., 2025). However, because of instrument maintenance, malfunction, degradation, and other operational issues, ground-based measurements cannot be assumed to be error-free and must be quality-controlled. Consequently, substantial effort has been devoted to developing quality control (QC) methods and

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

routines for surface radiation measurements. In this context, the core task is outlier detection, in which samples are classified as inliers or outliers, often referred to as "good" and "bad" samples.

Outlier detection has deep roots in statistics. One of the most basic methods is the boxplot introduced by Tukey (1975), in which points located more than 1.5 interquartile ranges beyond the box are flagged as anomalies. Many boxplot variants have since been proposed for more complex settings, including letter-value plots for large datasets (Hofmann et al., 2017) and bagplots for bivariate datasets (Rousseeuw et al., 1999). Beyond boxplot-type approaches, statistical outlier detection methods are commonly categorized as density-, distance-, or cluster-based; the reader is referred to Smiti (2020) for a review. Density-based methods assume that anomalies occur in low-probability regions. Distance-based methods assume that anomalies lie far from neighboring observations. Cluster-based methods partition data into groups and identify anomalies as samples that do not fit any group. Regardless of method, effectiveness is maximized when domain knowledge is incorporated.

For surface radiation measurements, domain knowledge can be grouped into two categories: information related to the radiation quantities themselves and information related to instrumentation. For example, the widely used extremely rare limit (ERL) test proposed by Long and Dutton (2002) incorporates both categories by defining upper limits for shortwave radiation quantities as functions of the cosine of the solar zenith angle, with an additional constant term that accounts for measurement uncertainty. The ERL test is expressed as

$$-2\ \mathrm{W\,m^{-2}} \le G_h \le 1.2 E_{0n} \cos^{1.2} Z + 50\ \mathrm{W\,m^{-2}}, \tag{1}$$

$$-2\ \mathrm{W\,m^{-2}} \le B_n \le 0.95 E_{0n} \cos^{0.2} Z + 10\ \mathrm{W\,m^{-2}}, \tag{2}$$

$$-2\ \mathrm{W\,m^{-2}} \le D_h \le 0.75 E_{0n} \cos^{1.2} Z + 30\ \mathrm{W\,m^{-2}}, \tag{3}$$

where G_h, B_n, and D_h denote the global horizontal irradiance (GHI), beam normal irradiance (BNI), and diffuse horizontal irradiance (DHI), respectively; Z is the solar zenith angle; and E_{0n} is the extraterrestrial irradiance, computed using the solar constant ($E_{\mathrm{sc}} = 1361.1\ \mathrm{W\,m^{-2}}$), the average sun–earth distance over the course of one revolution (R_{avg}), and the current sun–earth distance (R):

$$E_{0n} = E_{\mathrm{sc}} \times \left(\frac{R_{\mathrm{avg}}}{R} \right)^2 . \tag{4}$$

The ERL test is recommended by the Baseline Surface Radiation Network (BSRN) and is widely used in both climate and solar-energy applications (Driemel et al., 2018).

However, one key limitation of the ERL test in Eqs. (1)–(3) is that the limits are often too loose. Figure 1 shows one year (2024) of 1-min measurements from the BSRN Qiqihar (QIQ) station (Chen et al., 2025), together with the corresponding ERL values. The mismatch is clear: the ERL envelopes are overly conservative and therefore provide limited QC sensitivity. Consistent with Nollas et al. (2023), the ERL test typically rejects far fewer records than other QC tests, such as the three-component closure test. Therefore, the objective of this study is to derive improved ERL coefficients. This direction was already suggested by Long and Shi (2008), who recommended configurable climatological upper limits. Denoting the cosine of the

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

solar zenith angle by μ_0, the configurable limits are written as

$$G_h \leq a_G \times E_{0n} \times \mu_0^{b_G} + c_G, \tag{5}$$

$$B_n \leq a_B \times E_{0n} \times \mu_0^{b_B} + c_B, \tag{6}$$

$$D_h \leq a_D \times E_{0n} \times \mu_0^{b_D} + c_D, \tag{7}$$

where the nine coefficients $a_G, b_G, \ldots, c_D$ are to be determined using several years of climate-, region- or site-specific data.

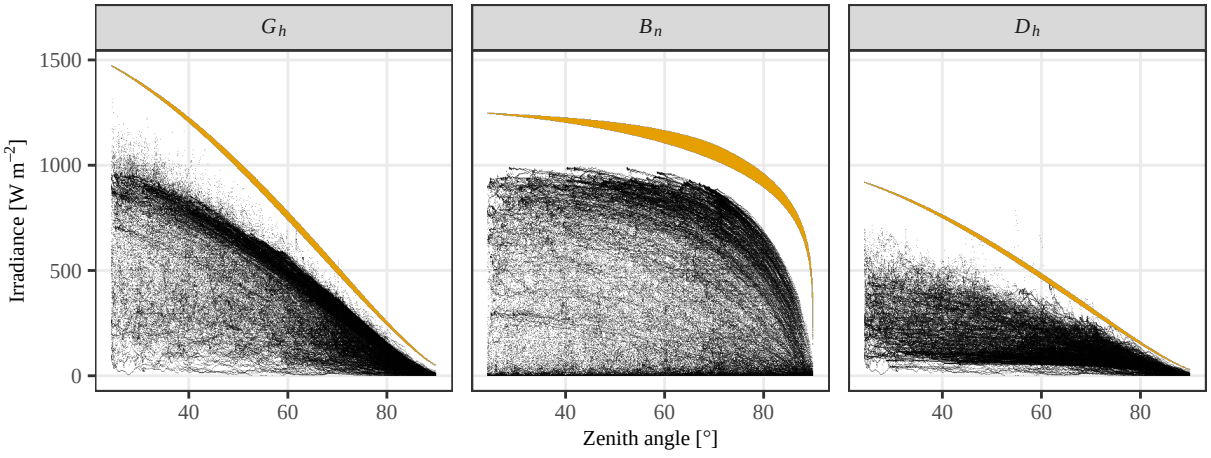

Figure 1. One year (2024) of 1-min irradiance measurements from the QIQ station (black) with their corresponding ERL values (orange).

At this stage, two questions arise: (1) how many coefficient sets should be configured for global application, and (2) how should those coefficients be determined? The first question concerns clustering. The objective is to define radiation climatic regimes from global observations and then derive one coefficient set for each regime. A possible choice is the Köppen–Geiger climate classification (Beck et al., 2018); however, that framework is based primarily on temperature and precipitation and is therefore not specifically tailored to solar-radiation variability. Accordingly, this study adopts a data-driven approach that groups BSRN radiation measurements into seven radiation climatic regimes. Clustering is performed by combining principal component analysis (PCA) and hierarchical clustering. To avoid the computational burden associated with high-dimensional time-series inputs (where each timestamp can be treated as a feature), clustering is conducted in a carefully designed feature space.

The second question concerns optimization. The objective is to determine coefficients using an explicit and reproducible criterion. The main challenge is that true anomalies are unknown during QC. In Long and Shi (2008), no specific method was provided for determining configurable coefficients. Yang et al. (2018) proposed gradually reducing coefficients until a sudden increase in rejection rate occurs; however, this strategy still involves subjective judgment. In binary classification, performance is commonly evaluated using the F_1 score, which is computed from a confusion matrix containing true positive (TP), true negative (TN), false positive (FP), and false negative (FN) counts. In this study, the confusion matrix is constructed from two

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

inlier–outlier classifications: one from the climatological-limit test (Eqs. (5)–(7)) and one from the isolation forest (iForest) method. Maximizing the F_1 score yields the highest agreement between the two QC classifications. This optimization is solved using the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with bounds (L-BFGS-B).

2 Data and method

2.1 BSRN data

The data used in this study are from the Baseline Surface Radiation Network (BSRN), which is widely regarded as a reference-quality archive of ground-based radiation measurements (Driemel et al., 2018; Ohmura et al., 1998). The observations are available from the official BSRN FTP server and from PANGAEA, which is an open-access data publisher for earth and environmental science. This study uses shortwave radiation quantities (G_h, B_n, and D_h) from 75 stations, and the corresponding station metadata are listed in Table 1. These stations span all five continents and islands across four oceans, providing broad coverage of global radiation regimes. To balance representativeness and data volume, the most recent continuous 4-year period is selected for each station. If a station does not meet this criterion, all available records are used.

Because solar radiation quantities exhibit both annual and diurnal cycles, it is customary to use normalized indices (or k indices) in data analysis. There are multiple normalization options, and the simplest is to convert irradiances to transmittances. For G_h, B_n, and D_h, the corresponding transmittances (k_t, k_b, and k_d) are computed as follows:

$$k_t = \frac{G_h}{E_0}, \tag{8}$$

$$k_b = \frac{B_n}{E_{0n}}, \tag{9}$$

$$k_d = \frac{D_h}{E_0}, \tag{10}$$

where $E_0 = E_{0n} \cos Z$ is the extraterrestrial GHI. In many references, k_t is referred to as the clearness index. Similar to the three radiation quantities, the transmittances also satisfy the closure relationship,

$$k_t = k_b + k_d. \tag{11}$$

In the clustering analysis below, the k_t, k_b, and k_d time series are used. Furthermore, all nighttime data points (defined here as $Z > 85°$) are removed. In addition, a filter requiring k_t, k_b, and k_d to be greater than zero is applied to exclude low-sun observations, which often exhibit high noise because of radiometer design limitations (Liu et al., 2025).

2.2 Radiation climatological feature selection

In this study, radiation climatological features are selected to group sites with similar radiation regimes, so that QC coefficients can be developed at the group level rather than for individual sites. Defining radiation regimes has been attempted in numerous studies and is referred to as radiation zoning (Li et al., 2013) or radiation climate classification (Dash et al., 2017). In most

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

Table 1. Metadata of the 75 BSRN stations used in this study. "Start" and "End" are reported in yymm format (year and month).

Code	Lat.	Lon.	Elev. (m)	No. mon.	Start	End	Code	Lat.	Lon.	Elev. (m)	No. mon.	Start	End
ABS	44.02	144.28	38.00	41	2103	2407	IZA	28.31	−16.50	2372.90	48	2101	2412
ALE	82.49	−62.42	127.00	48	1001	1312	KWA	8.72	167.73	10.00	48	1301	1612
ASP	−23.80	133.89	547.00	48	1501	1812	LAU	−45.05	169.69	350.00	48	1509	2006
BAR	71.32	−156.61	8.00	48	1901	2212	LER	60.14	−1.18	80.00	48	1301	1612
BER	32.30	−64.77	8.00	48	901	1212	LIN	52.21	14.12	125.00	48	1901	2212
BIL	36.60	−97.52	317.00	48	1501	1812	LMP	35.52	12.63	50.00	19	2312	2506
BON	40.07	−88.37	213.00	48	1901	2212	LRC	37.10	−76.39	3.00	48	2101	2412
BOS	40.12	−105.24	1689.00	48	1601	1912	LYU	22.04	121.56	324.00	48	1901	2212
BOU	40.05	−105.01	1577.00	48	1201	1512	MAN	−2.06	147.43	6.00	48	901	1212
BRB	−15.60	−47.71	1023.00	48	1201	1512	MNM	24.29	153.98	7.10	48	2101	2412
BUD	47.43	19.18	139.10	48	2001	2312	NAU	−0.52	166.92	7.00	48	901	1212
CAB	51.97	4.93	0.00	48	2001	2312	NEW	−32.88	151.73	18.50	30	1709	2002
CAM	50.22	−5.32	88.00	48	1301	1612	NYA	78.92	11.93	11.00	48	2101	2412
CAP	79.27	101.75	20.00	12	1601	1612	OHY	−12.05	−75.32	3314.00	41	1708	2101
CAR	44.08	5.06	100.00	48	1501	1812	PAL	48.71	2.21	156.00	48	1901	2212
CLH	36.91	−75.71	37.00	48	1201	1512	PAR	5.81	−55.21	4.00	48	1912	2311
CNR	42.82	−1.60	471.00	48	2001	2312	PAY	46.81	6.94	491.00	48	2001	2312
COC	−12.19	96.83	6.00	48	1501	1812	PSU	40.72	−77.93	376.00	48	1601	1912
DAA	−30.67	23.99	1287.00	48	1601	1912	PTR	−9.07	−40.32	387.00	48	1408	1807
DAR	−12.43	130.89	30.00	48	1101	1412	QIQ	47.80	124.49	170.00	15	2311	2501
DOM	−75.10	123.38	3233.00	48	1801	2112	REG	50.20	−104.71	578.00	48	801	1112
DRA	36.63	−116.02	1007.00	48	1901	2212	RUN	−20.90	55.48	116.00	48	2101	2412
DWN	−12.42	130.89	32.00	48	1501	1812	SAP	43.06	141.33	17.20	48	1601	1912
E13	36.60	−97.48	318.00	12	113	1213	SBO	30.86	34.78	500.00	48	802	1212
ENA	39.09	−28.03	15.20	24	1501	1612	SEL	15.78	−91.99	602.00	48	2006	2506
EUR	79.99	−85.94	85.00	48	801	1112	SMS	−29.44	−53.82	489.00	48	1301	1612
FLO	−27.60	−48.52	11.00	47	2001	2403	SON	47.05	12.96	3108.90	48	2001	2312
FPE	48.32	−105.10	634.00	48	1901	2212	SOV	24.91	46.41	650.00	48	9901	212
FUA	33.58	130.38	3.00	48	2001	2312	SPO	−89.98	−24.80	2800.00	48	1301	1612
GAN	23.11	72.63	65.00	21	1406	1901	SXF	43.73	−96.62	473.00	48	1501	1812
GCR	34.25	−89.87	98.00	48	1601	1912	SYO	−69.01	39.58	18.00	48	2001	2312
GIM	46.72	−87.41	208.00	44	2001	2312	TAT	36.06	140.13	25.00	48	2001	2312
GOB	−23.56	15.04	407.00	48	2101	2412	TIK	71.59	128.92	48.00	48	1401	1803
GUR	28.42	77.16	259.00	21	1407	1901	TIR	13.09	79.97	36.00	41	1408	1901
GVN	−70.65	−8.25	42.00	48	1801	2112	TOR	58.26	26.46	70.00	48	1701	2012
HOW	22.55	88.31	51.00	32	1410	1901	XIA	39.75	116.96	32.00	48	1001	1412
INO	44.34	26.01	110.00	26	2105	2403	YUS	23.49	120.96	3858.00	48	1901	2212
ISH	24.34	124.16	5.70	48	2101	2412							

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

studies, radiation regimes are defined by clustering monthly or daily mean radiation variables or the clearness index (e.g., de las Heras et al., 2026; Jiang et al., 2021; Anas et al., 2021). However, as data resolution increases, the clustering dimensionality—where each measurement can be treated as one dimension—quickly becomes computationally infeasible, even with dimension-reduction techniques. Therefore, this study uses *radiation climatological features* to represent high-resolution radiation data.

Three feature classes are considered: (1) zenith-angle range, (2) harmonic-analysis features, and (3) probability-density-function (PDF) features. The zenith-angle range is included because it is directly involved in QC (Eqs. (5)–(7)), where climatological limits are configured as a function of Z. Intuitively, using the same QC envelope (cf. Fig. 1) for one site with a minimum zenith angle of 50° and another with a minimum zenith angle of 0° is inadequate. Harmonic analysis describes seasonal patterns (not limited to annual and diurnal cycles) in radiation time series, which are related to the range and shape of the scatter points in Fig. 1. Harmonic analysis has also long been used as a stand-alone tool for radiation climate classification (Terjung, 1970; Horn and Bryson, 1960). In addition, the PDFs of solar-radiation transmittances (k_t, k_b, and k_d) are known *a priori* to be tied to prevailing local sky conditions. For example, arid sites are expected to show higher probability concentrations in the high-k_t range. Thus, careful PDF analysis can extract statistical descriptors of radiation regimes. Because the zenith-angle range is straightforward to obtain, the extraction of the other two feature classes is described below.

2.2.1 Harmonic-based features

For a given time series $\{y_t\}$, with $t = 1, \ldots, T$, harmonic analysis assumes that the series can be decomposed into an infinite series of harmonic components (Jakubauskas et al., 2001). Mathematically, this is expressed as

$$y_t = c_0 + \sum_{n=1}^{\infty} a_n \cos\frac{2\pi n t}{T} + \sum_{n=1}^{\infty} b_n \sin\frac{2\pi n t}{T}. \tag{12}$$

By defining the j^{th} harmonic as the j^{th} term in the Fourier series, it follows that

$$a_j \cos\frac{2\pi j t}{T} + b_j \sin\frac{2\pi j t}{T} = c_j \cos\left(\frac{2\pi j t}{T} - \phi_j\right), \tag{13}$$

where $c_j = \sqrt{a_j^2 + b_j^2}$ and $\phi_j = \arctan(b_j/a_j)$ are known as the amplitude and phase angle of the j^{th} harmonic. In practice, it is common to use a reduced form of Eq. (12) by neglecting the higher harmonics and including an error term:

$$\begin{aligned} y_t &= c_0 + \sum_{j=1}^{m} a_j \cos\frac{2\pi j t}{T} + \sum_{j=1}^{m} b_j \sin\frac{2\pi j t}{T} + \varepsilon_t \\ &= c_0 + \sum_{j=1}^{m} c_j \cos\left(\frac{2\pi j t}{T} - \phi_j\right) + \varepsilon_t, \end{aligned} \tag{14}$$

Equation (14) can be solved efficiently by least squares, yielding $(c_0, a_1, \ldots, a_m, b_1, \ldots, b_m)$, or equivalently $(c_0, c_1, \ldots, c_m, \phi_1, \ldots, \phi_m)$.

In this study, harmonic analysis is applied to daily G_h, B_n, and D_h time series. The number of harmonics (m) is empirically set to 25. Figure 2 shows results for the Boulder station (BOS), United States, and demonstrates that seasonal variability in irradiance components is captured well. Two features are extracted for each of G_h, B_n, and D_h: (1) c_0, representing the irradiance level, and (2) the range (maximum minus minimum) of the harmonic regression fit, representing seasonal irradiance contrast. Thus, harmonic analysis provides six features in total.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

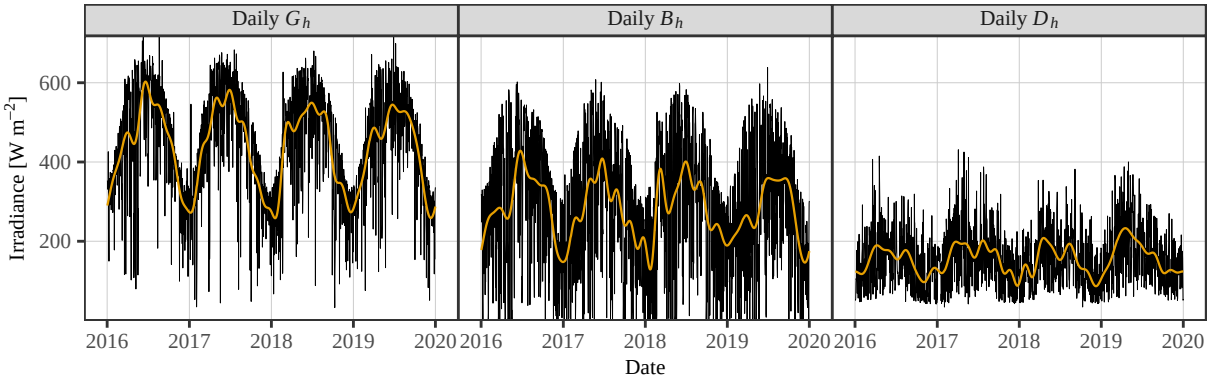

Figure 2. Harmonic analysis of daily G_h, B_n, and D_h time series at the BOS station.

2.2.2 Density-based features

The density-based features are extracted from k_t, k_b, and k_d, rather than from irradiance components themselves, because the latter random variables are "contaminated" by yearly and diurnal cycles and are thus unable to fully reflect sky conditions. There are numerous studies in the literature dealing with the statistical distribution of k_t, but far fewer on k_b and k_d. For instance, Jurado et al. (1995) proposed using a two-component normal mixture model for 5-min k_t, whereas Hollands and Suehrcke (2013) suggested a three-component normal mixture model for 1-min k_t. Regardless, a consensus in the literature is that transmittance distributions are rarely unimodal, because at least three groups of sky conditions generally exist: clear, cloudy, and overcast (Yagli et al., 2019; Lou et al., 2019). On this point, recent work by Zhang et al. (2024) revealed that component distributions are often asymmetric, which justifies the use of skew-normal distributions during modeling.

A mixture model is simply the convex combination of several component distributions, that is,

$$g(x;\boldsymbol{\Theta}) = \sum_{j=1}^{m} p_j\, f_j(x;\boldsymbol{\theta}_j), \quad \text{s.t.} \sum_{j=1}^{m} p_j = 1 \tag{15}$$

where $g(\cdot)$ is the PDF of a multi-modal random variable, $f_j(\cdot)$ is the PDF of j^{th} component, $\boldsymbol{\Theta}$ and $\boldsymbol{\theta}_j$ are parameter vectors of the respective PDFs, where $\boldsymbol{\Theta} = \left(p_1, \ldots, p_m, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_m^\top\right)^\top$. For instance, the PDF of a skew-normal distribution is

$$f(x;\boldsymbol{\theta}) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\frac{(x-\xi)}{\omega}\right), \tag{16}$$

where $\boldsymbol{\theta} = (\xi, \omega, \alpha)^\top$ is the parameter vector holding the location, scale, and shape parameters; $\phi(\cdot)$ is the PDF of a normal distribution; and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. For a three-component skew-normal distribution,

$$\boldsymbol{\Theta} = (p_1, p_2, p_3, \xi_1, \omega_1, \alpha_1, \xi_2, \omega_2, \alpha_2, \xi_3, \omega_3, \alpha_3)^\top, \tag{17}$$

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

which can be estimated using a variety of methods, such as the method of moments (MOM) and maximum likelihood estimation (MLE). It is highlighted that fitting mixture distribution requires an initial condition. In this work, MOM is used for that purpose, and MLE follows.

With the above preliminaries, Fig. 3(a) depicts the estimated densities of three different mixture models, using k_t data from the BOS station. It can be seen that the three-component mixtures have a clear advantage over the two-component mixture. In terms of feature selection, the area under the left-most component PDF and the area under the right-most component PDF are considered, for they correspond to the probability of overcast and clear-sky conditions, respectively. It should be highlighted that the component PDFs are not necessarily ordered, and the left- and right-most PDFs need to be identified through computing the mean values. For a skew-normal distribution, its mean value is given by

$$\xi + \omega\delta\sqrt{\frac{2}{\pi}}, \text{ where } \delta = \frac{\alpha}{\sqrt{1+\alpha^2}}. \tag{18}$$

In other words, we first compute the mean values of f_1, f_2, and f_3, and then sort them from smallest to largest. After that, the p_j's that correspond to the smallest and largest mean values are selected as features.

As for k_d, its analysis is analogous to that of k_t. Specifically, a three-component skew-normal mixture model is compared against two- and three-component normal mixture models in Fig. 3(c), showing evident superiority in explaining the distribution of k_d. Feature selection is also analogous, i.e., the areas under the left- and right-most component PDFs are retrieved. That said, the situation with k_b differs from that of the other two transmittances. More specifically, the value of B_n and thus k_b drops to near zero in the presence of clouds—this results in a sharp rise in probability at $k_b = 0$, which cannot be modeled using normal distributions. Considering that the probability near $k_b = 0$ gradually declines as k_b increases, there are many options for component distributions, including exponential, Weibull, gamma, and beta distributions, among others. Among the different options, the power distribution bounded on $[0, 1]$ is particularly useful for its simplicity and flexibility, and is therefore chosen. The PDF of a standard power distribution is:

$$f(x) = ax^{(a-1)}, \tag{19}$$

where a is the only parameter to be estimated. Using a power distribution as a basis, several mixture models are compared in Fig. 3(b), and the option with one power distribution and two skew-normal distributions shows the best fit. Similar to the two former cases, the areas under the power distribution and the right-sided skew-normal distribution are selected as features.

2.3 Dimension reduction and clustering

With the extracted radiation climatological features, the next step involves dimensionality reduction and clustering. This process ensures that stations from locations worldwide are allocated into sensible groups, allowing for the configuration of group-specific QC limits. In this work, we employ PCA followed by Ward's hierarchical clustering. This combination is chosen for its ability to handle feature redundancy while providing a structured, multi-level grouping of stations based on their variance profiles.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

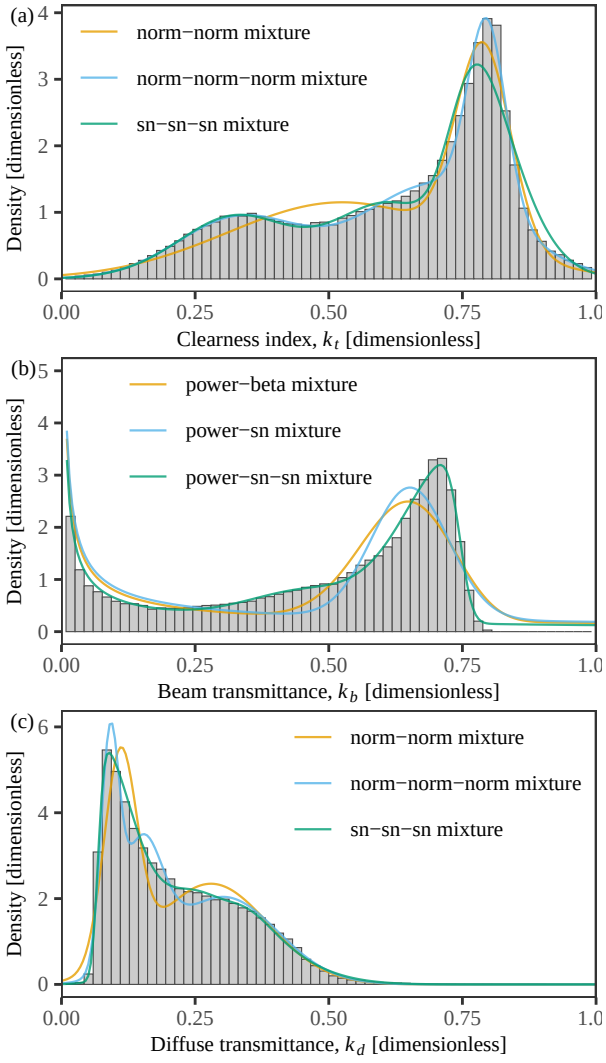

Figure 3. Histograms of k_t, k_b, and k_d at the BOS station, alongside density functions computed using selected mixture models. The short form "norm" and "sn" in the legend represent "normal" and "skew normal," respectively.

PCA transforms correlated variables into a new set of orthogonal principal components. These components are linear combinations of the original features, ordered by explained variance. To avoid errors in distance calculations during clustering, rows containing missing values are removed before analysis. Following standard practice, a 90% variance-retention threshold is applied. In this implementation, the first seven principal components are retained because they jointly satisfy this threshold.

These seven principal components are then used as inputs to hierarchical clustering with Ward's minimum-variance criterion. Unlike k-means, which requires the number of clusters to be predefined, hierarchical clustering builds a nested tree structure (dendrogram) from Euclidean distances. To determine the optimal number of clusters (k), we conduct silhouette analysis for

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

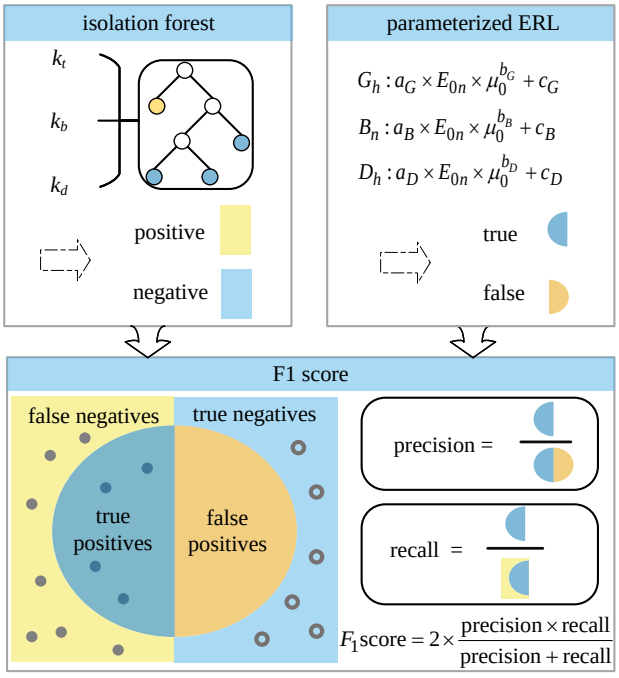

Figure 4. Workflow for configuring climatological QC limits. The procedure iteratively adjusts the parameters of two independent outlier detection algorithms to maximize the F_1 score derived from their agreement.

$k = 4$ to 10. The final value of k is selected by maximizing the mean silhouette width, which balances high intra-cluster cohesion and strong inter-cluster separation. This procedure yields a cluster configuration that effectively categorizes stations by their geographic and climatological characteristics.

2.4 QC using climatological limits and iForest

To define radiation-regime-specific climatological limits for QC, two complementary outlier-detection methods are used: the ERL test and the iForest algorithm. Both methods produce binary outlier–inlier classifications, and their agreement is quantified using the F_1 score. Because the ERL parameters directly determine the QC outcome, they are configured by maximizing F_1. This maximization problem is solved with the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with bounds (L-BFGS-B). In other words, the ERL parameter set that yields the highest F_1 score is adopted as the optimal climatological limit for each climate category. The detailed implementation is presented in the following sections, and the complete QC-configuration workflow is shown in Fig. 4.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

2.4.1 Climatological limits for true and false value determination

In the BSRN QC procedure, outliers are detected according to Eqs. (5)–(7). To establish stricter criteria for regime-specific QC evaluation, the test parameters are adjusted. Data points outside the QC limits are labeled false (F), whereas points within the acceptable range are labeled true (T). In other words, samples within the ERL limits are treated as likely valid observations, while samples exceeding the limits are treated as likely problematic observations.

2.4.2 Isolation forest for positive and negative value determination

The iForest algorithm is an unsupervised anomaly-detection method based on decision-tree principles. Its core concept is the isolation property of anomalies: anomalous points tend to lie far from the main point cloud or in regions with distinct density. Compared with conventional clustering methods, such as kernel density estimation (KDE) and Gaussian mixture models (GMM), iForest offers several practical advantages. Conventional clustering methods often rely on features such as solar zenith angle and irradiance; because these variables have different units, distance- or density-based approaches can perform unsatisfactorily. Similarly, KDE relies on grouping through latent class variables and does not fully account for observation-level characteristics, especially temporal continuity, which can limit detection performance. In contrast, iForest detects outliers using ensembles of random trees, providing a fast and efficient tree-based approach.

Let the dataset be denoted by $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, where each $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})^\top \in \mathbb{R}^d$. The iForest procedure can be summarized as follows: (1) To build one isolation tree, a random subsample $\mathcal{D}^*$ is drawn from $\mathcal{D}$. At each node, a feature index $q \in \{1, \ldots, d\}$ is selected at random, and a split value p is drawn from $[\min x_{i,q}, \max x_{i,q}]$, with $i = 1, \ldots, n$. (2) The current sample set is split into $\{\boldsymbol{x}_i : x_{i,q} \le p\}$ and $\{\boldsymbol{x}_i : x_{i,q} > p\}$, and recursion continues until a stopping condition is reached (e.g., a single sample remains or the maximum tree depth is reached). (3) This produces a binary tree in which each sample corresponds to a terminal node. The path length of sample $\boldsymbol{x}$ is defined as tree depth, denoted $h(\boldsymbol{x})$. (4) Repeating this process yields an ensemble of independent random trees, which forms the basis of iForest anomaly detection.

For a subtree with sample size n, the theoretical average path length $c(n)$ can be approximated as follows:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}, \tag{20}$$

where $2H(n-1)$ is twice the $(n-1)$th harmonic number, which can be approximated by $H(n-1) = \ln(n-1) + \gamma$, where γ is the Euler–Mascheroni constant and is approximately 0.5772. The anomaly score, based on the average path length, is defined as

$$s(\boldsymbol{x}, n) = 2^{\mathbb{E}[h(\boldsymbol{x})]/c(n)}, \tag{21}$$

$$\mathbb{E}[h(\boldsymbol{x})] = \frac{1}{T}\sum_{k=1}^{T} h_k(\boldsymbol{x}), \tag{22}$$

where T is the number of trees in the forest, n is the number of samples in a subtree, and $\mathbb{E}[h(\boldsymbol{x})]$ denotes the expected path length of sample $\boldsymbol{x}$. When $s(\boldsymbol{x}, n)$ approaches 0, the sample is more likely to be normal, corresponding to a longer path length.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

To define an anomaly threshold, kernel density estimation is applied to the anomaly score. Mathematically, the scaled kernel density of the random variable S (anomaly score) is

$$\hat{f}(s) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{s - s_i}{h}\right), \tag{23}$$

where $\hat{f}(s)$ is the density estimate, h is the bandwidth controlling smoothness, $K(\cdot)$ is the kernel function, and s_i is the ith anomaly score. The Gaussian kernel,

$$K(s) = \frac{1}{\sqrt{2\pi}} e^{-s^2/2}, \tag{24}$$

is adopted. Because the estimated density $\hat{f}(s)$ is nonnegative, the outlier threshold is defined by taking the logarithm of the density, i.e.,

$$\delta = \underset{s}{\operatorname{argmax}}\{s : \log \hat{f}(s) = 0\}. \tag{25}$$

Samples with anomaly scores exceeding the threshold ($s > \delta$) are labeled negative (N), whereas the remaining samples are labeled positive (P).

2.4.3 Quality control assessment metrics

The agreement between the ERL test and iForest classifications is evaluated using the F_1 score, which balances the competing metrics of precision and recall. Precision quantifies the correctness of positive predictions, penalizing false positives (FP). It is defined as the ratio of true positives (TP)—samples correctly identified as outliers—to all samples predicted as positive:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{26}$$

Recall, also known as sensitivity, measures the completeness of detection, penalizing false negatives (FN). It is defined as the ratio of true positives to all actual positives:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{27}$$

The F_1 score is the harmonic mean of precision and recall, providing a single metric that favors models that achieve both high correctness and high completeness in outlier identification:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{28}$$

A score of 1 represents perfect agreement, whereas 0 indicates poor performance. Maximizing the F_1 score during parameter optimization therefore enforces a balanced trade-off between minimizing false alarms and maximizing the detection of true climatological outliers.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

2.5 Optimization objective and solution method

The limited-memory BFGS (L-BFGS) algorithm is a quasi-Newton optimization method designed for large-scale problems. It approximates the inverse Hessian to determine the search direction, while using a compact history of gradients and parameter updates rather than storing a dense Hessian approximation. This limited-memory strategy requires storing only an $m \times n$ matrix, where n is the number of variables and m (typically $m < 10$) is a small number of historical updates. As a result, memory usage is substantially lower than in standard BFGS. L-BFGS is therefore well suited to high-dimensional parameter-estimation tasks. The L-BFGS-B algorithm extends L-BFGS to include simple box constraints. At each iteration, it identifies fixed and free variables (from gradient information), applies L-BFGS updates only to the free variables, and iterates until convergence (Zhu et al., 1997; Byrd et al., 1995).

3 Results and discussion

This section presents the main outcomes of the proposed framework in three parts. First, we identify and physically interpret the major global radiation climatic regimes derived from unsupervised learning. Second, we analyze the optimized regime-specific climatological-limit coefficients and examine how they reflect regional atmospheric conditions. Third, we compare QC performance between the original ERL formulation and the proposed regime-specific limits, with emphasis on boundary tightness and detection behavior.

3.1 World's major radiation climatic regimes

Using the unsupervised learning framework described in Sections 2.2 and 2.3, the 75 BSRN stations are grouped into seven distinct regimes. Figure 5 shows a heatmap of pairwise Euclidean distances (denoted as Ward distance, "WD") among the PCA-transformed station features, using principal components that explain $\geq$90% of the total variance. The color gradient represents distance magnitude, with darker shades indicating higher multivariate similarity. Notably, the diagonal blocks are much darker than the off-diagonal regions, indicating high intra-cluster cohesion and low inter-cluster similarity and thus confirming the effectiveness of the hierarchical clustering. Dendrograms along the top and left margins illustrate the nested grouping structure produced by Ward's minimum-variance method. Adjacent discrete color blocks mark the final regime assignments, partitioned according to the optimal cluster count identified by silhouette analysis. This robust regionalization provides the basis for configuring regime-specific ERL tests.

To visualize the spatial distribution of the identified radiation regimes, Fig. 6 maps the BSRN stations and colors each site by its final cluster assignment. The resulting pattern shows a physically meaningful regionalization. This coherence is especially evident for the eight polar stations, which form a distinct, unified cluster due to their shared extreme annual irradiance cycles (i.e., polar day and night). In addition, geographically distant stations exposed to similar atmospheric conditions are consistently assigned to the same regime, as illustrated by the arid stations at Desert Rock (DRA) in the United States and Alice Springs (ASP) in Australia. This agreement further supports the physical relevance of the extracted distributional features. Considering

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

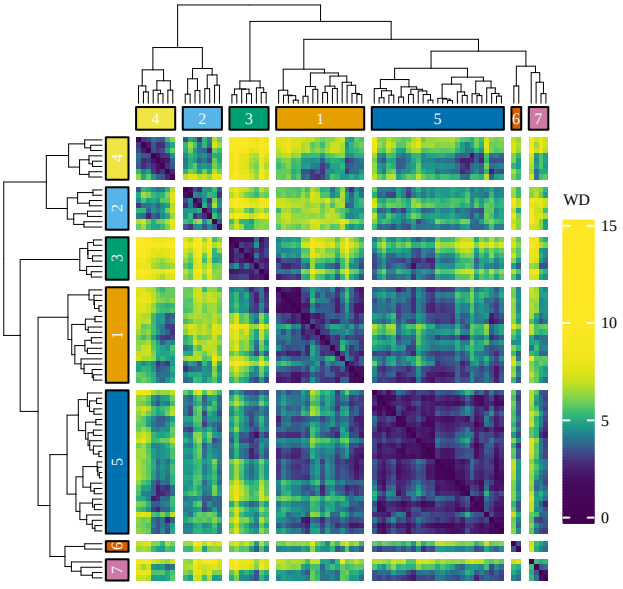

Figure 5. Heatmap and hierarchical clustering dendrogram of BSRN stations based on their radiation climatological feature space.

both the spatial organization and the intrinsic properties of these features, the seven regimes can be interpreted as distinct global climatological zones. Table 2 summarizes the regimes, including their abbreviations, representative atmosphere/surface conditions, and defining radiation traits. A detailed discussion of each regime's physical drivers and climatological context is provided below.

Regime 1 (tropical and subtropical savannas, TSS) comprises 18 stations with a mean absolute latitude of 20.8° and primarily represents global savanna and monsoon belts located between equatorial rainforests and subtropical deserts. The defining atmospheric feature of this regime is a pronounced wet–dry seasonal bifurcation driven by migration of the intertropical convergence zone and associated monsoon dynamics. Because these stations are at low latitudes, they maintain high solar elevation angles throughout the year, yielding exceptionally high theoretical maxima for G_h. However, the surface radiation budget shows strong seasonal variability. During the dry season, stable subsiding air linked to subtropical high-pressure systems allows extended periods of intense, largely unattenuated B_n. By contrast, wet-season onset brings deep convection and heavy precipitation, shifting the radiative balance so that D_h becomes dominant because of enhanced cloud scattering. Visual evidence and additional discussion are provided in Fig. A1 in Appendix A.

Regime 2 (polar and sub-polar, PSP) includes eight stations in extreme high-latitude environments, with a mean absolute latitude of 78.2°. This cluster is rooted in polar and sub-polar regions characterized by cold, dry air masses and persistent snow/ice cover. The key physical control is the extreme annual irradiance cycle—polar day and polar night—which keeps solar elevation angles low even during summer (Fig. A2 in Appendix A). As a result, the surface radiation budget is strongly modulated by the high albedo of frozen surfaces. Although B_n is strongly suppressed by long atmospheric optical paths and

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

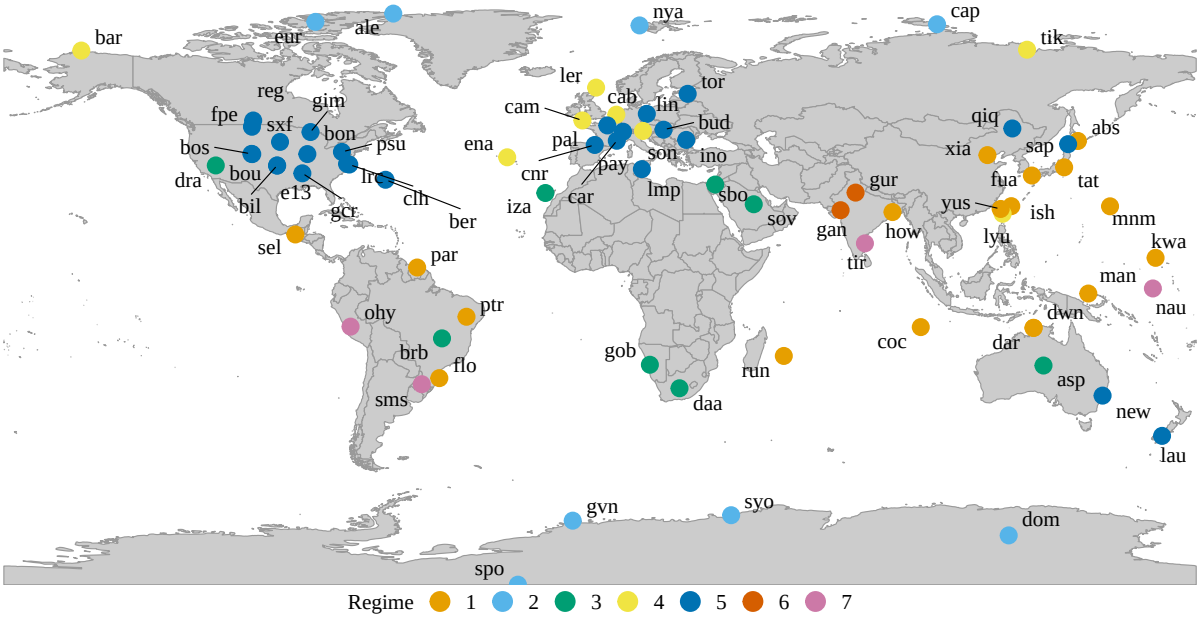

Figure 6. Geographical distribution of the BSRN stations utilized in this study, alongside their identified radiation climatic regimes.

low sun angles, D_h often dominates the radiation profile. This dominance is further reinforced by multiple scattering between the reflective surface and frequent boundary-layer ice crystals or polar stratiform clouds.

Regime 3 (arid and semi-arid deserts, ASD) comprises eight stations located mainly in subtropical high-pressure zones, with a mean absolute latitude of 26.8°. This group includes geographically distant but climatologically similar arid and semi-arid environments, such as DRA and ASP. Atmospheric conditions are controlled by persistent subsidence from the descending branches of the Hadley circulation. This large-scale subsidence maintains stable, dry air masses, minimal precipitation, and predominantly clear skies throughout the year (Fig. A3 in Appendix A). As a result, the radiation profile shows weak cloud attenuation and low atmospheric moisture. These factors produce high atmospheric transmittance, yielding consistently high B_n and G_h and very low diffuse fractions. The statistical prevalence of clear-sky days makes ASD one of the most stable and radiatively intense regimes in the surface solar radiation budget.

Regime 4 (maritime cloudy mid-latitudes, MCM) consists of eight stations across mid-to-high latitudes, with a mean absolute latitude of 51.7°. These maritime and coastal environments are strongly influenced by oceanic moisture transport and frequent subpolar low-pressure passages. Consequently, atmospheric conditions are dominated by persistent marine stratocumulus and extensive frontal cloud cover. This near-continuous cloudiness efficiently scatters incoming solar radiation, producing a substantially higher baseline diffuse fraction than in continental regions at similar latitudes (Fig. A4 in Appendix A). Although seasonal cycles in daily irradiance remain evident, persistent attenuation lowers overall G_h and strongly suppresses B_n, making D_h a defining and consistently important component of the local surface radiation budget.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

Table 2. Seven radiation climatic regimes and their dominant characteristics.

	Abbrv.	Full name	Atmosphere/surface regime	Radiation traits
1	TSS	Tropical and subtropical savannas	Seasonal bifurcation (wet/dry seasons) influenced by monsoon dynamics	High maximum G_h due to high solar zenith angles, but heavily dominant D_h during wet seasons
2	PSP	Polar and subpolar	Polar snow and ice albedo	Low solar elevation angles; extreme polar day/night cycles; high surface albedo; D_h dominant
3	ASD	Arid and semi-arid deserts	Clear-sky dominant with negligible cloud fraction	Exceptionally high B_n and G_h, with very low diffuse fractions (k_d)
4	MCM	Maritime cloudy mid-latitudes	Persistent stratocumulus cloud decks and frequent oceanic frontal clouds	Lower overall G_h compared to TML, with a much higher baseline diffuse fraction (k_d) throughout the year
5	TML	Temperate mid-latitudes	Mixed and transient cloud regimes driven by synoptic weather systems	Distinct seasonal cycles in daily irradiance; highly variable balance between B_n and D_h
6	HAA	Heavy aerosol attenuation	Intense pre-monsoon dust/pollution loading, followed by heavy South Asian monsoonal cloud cover	High clear-sky diffuse irradiance (D_h) due to extreme aerosol scattering; severe overall attenuation during the monsoon
7	LLA	Low-latitude anomalous	Highly localized tropical/subtropical environments, including equatorial marine, high-altitude tropical, and coastal zones	Highly variable irradiance profiles driven by strong local geographical forcings (e.g., oceanic convection, orographic lift) rather than broad latitudinal trends

Regime 5 (temperate mid-latitudes, TML) is the largest cluster in the BSRN dataset, containing 27 stations with a mean absolute latitude of 42.7°. Its broad distribution places these sites in the temperate mid-latitudes, where the Westerlies and frequent synoptic weather systems dominate atmospheric variability. Cloud and surface conditions are therefore highly transient, alternating between clear-sky anticyclonic periods and overcast conditions linked to passing cold and warm fronts. The TML radiation profile is characterized by a clear seasonal cycle in daily irradiance superimposed on strong day-to-day variability. Although summer solar elevations permit substantial G_h, the relative contributions of B_n and D_h shift markedly with synoptic state, making TML one of the most dynamically variable regimes in terms of short-term surface radiation behavior. Additional details are shown in Fig. A5 in Appendix A.

Regime 6 (heavy aerosol attenuation, HAA) is a highly specialized cluster containing only two stations, Gandhinagar (GAN) and Gurgaon (GUR), both in the Indian subcontinent. Unlike regimes shaped mainly by broad latitudinal gradients or synoptic variability, this local environment is defined by exceptionally high aerosol optical depth (AOD) from severe anthropogenic pollution and pre-monsoon dust loading. As a result, even under nominally clear skies, strong atmospheric scattering produces an unusually high baseline D_h and pronounced attenuation of B_n (Fig. A6 in Appendix A). During summer, this aerosol-driven attenuation is further intensified by arrival of the South Asian monsoon, which brings persistent cloud cover and deep convection that strongly suppress G_h. The algorithm's separation of these two stations into an independent regime—using ra-

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

diation observations alone and without explicit aerosol inputs—demonstrates that the extracted climatological features robustly capture extreme scattering and attenuation processes.

Regime 7 (low-latitude anomalous, LLA) consists of a small, diverse cluster of four stations—Nauru Island (NAU), Observatory of Huancayo (OHY), São Martinho da Serra (SMS), and Tiruvallur (TIR)—representing highly localized tropical and subtropical microclimates. Unlike broader zonally organized regimes, this group is controlled by strong site-specific geographic forcings rather than latitude alone (see Fig. A7 in Appendix A). For example, grouping an equatorial marine site near sea level (NAU, 7 m) with a high-altitude Andean station (OHY, 3314 m) highlights the algorithm's sensitivity to anomalous radiation signatures. Atmospheric conditions at these sites are shaped by localized processes, such as persistent deep oceanic convection or strong orographic lifting, which produce highly variable insolation profiles. Their surface radiation budgets therefore deviate from typical low-latitude behavior, with large and irregular shifts between clear-sky B_n and strongly attenuated D_h. Identification of this anomalous regime further supports the need for a data-driven regionalization framework.

3.2 Configuration result analysis

Optimization of regime-specific climatological limits is performed using a multi-stage data-driven framework designed to separate physically plausible upper boundaries from meteorological noise. To handle the large volume of 1-min observations while targeting theoretical upper limits, we adopt a stratified sampling strategy. The dataset is first partitioned into 1° zenith-angle bins. Within each bin, the top 5% of irradiance values for G_h, B_n, and D_h are extracted to represent the upper envelope. To preserve typical observational states and the broader distributional structure, this envelope-focused sample is supplemented with a 10% random sample from each bin. An iForest—configured with 100 trees in a three-dimensional feature space comprising k_t, k_d, and k_b—is deployed to assign binary classifications to each data point using a kernel-density-estimated threshold. Structural limit curves, defined by $I = a \cdot E_{0n} \cdot (\cos Z)^b + c$, are fitted to these boundaries. The L-BFGS-B optimizer is used to estimate coefficients (a, b, c) by maximizing the F_1 score relative to the iForest labels.

Table 3 reports the optimized coefficients and corresponding F_1 scores for all seven radiation regimes. The consistent F_1 range of 0.98–0.99 across all three components $(F_{1,G}, F_{1,B}, F_{1,D})$ provides strong quantitative evidence that optimization successfully identifies the physical boundaries isolated by the unsupervised framework. Coefficient patterns also reflect distinct regime physics. For G_h, the amplitude factor a_G remains relatively constrained (0.85–1.10), and the zenith-response exponent b_G is stable across clusters (1.01–1.14). A key exception is Regime 2, which shows a pronounced baseline offset ($c_G = 11.77$ W m^{-2}). This shift is consistent with strong multiple-scattering effects and high surface albedo over polar snow and ice under persistently low solar elevation conditions.

The B_n coefficients further reflect atmospheric attenuation effects. Most pristine and temperate regimes have a_B values in the range 0.89–0.99, whereas Regime 6 drops markedly to $a_B = 0.72$. This quantitatively captures the large beam-radiation suppression associated with high aerosol loading over the Indo-Gangetic Plain and demonstrates the method's ability to regionalize anthropogenic attenuation. In addition, b_B is systematically lower than b_G (0.11–0.35), consistent with the stronger zenith-angle sensitivity and faster decay of direct-beam irradiance at high solar zenith angles.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

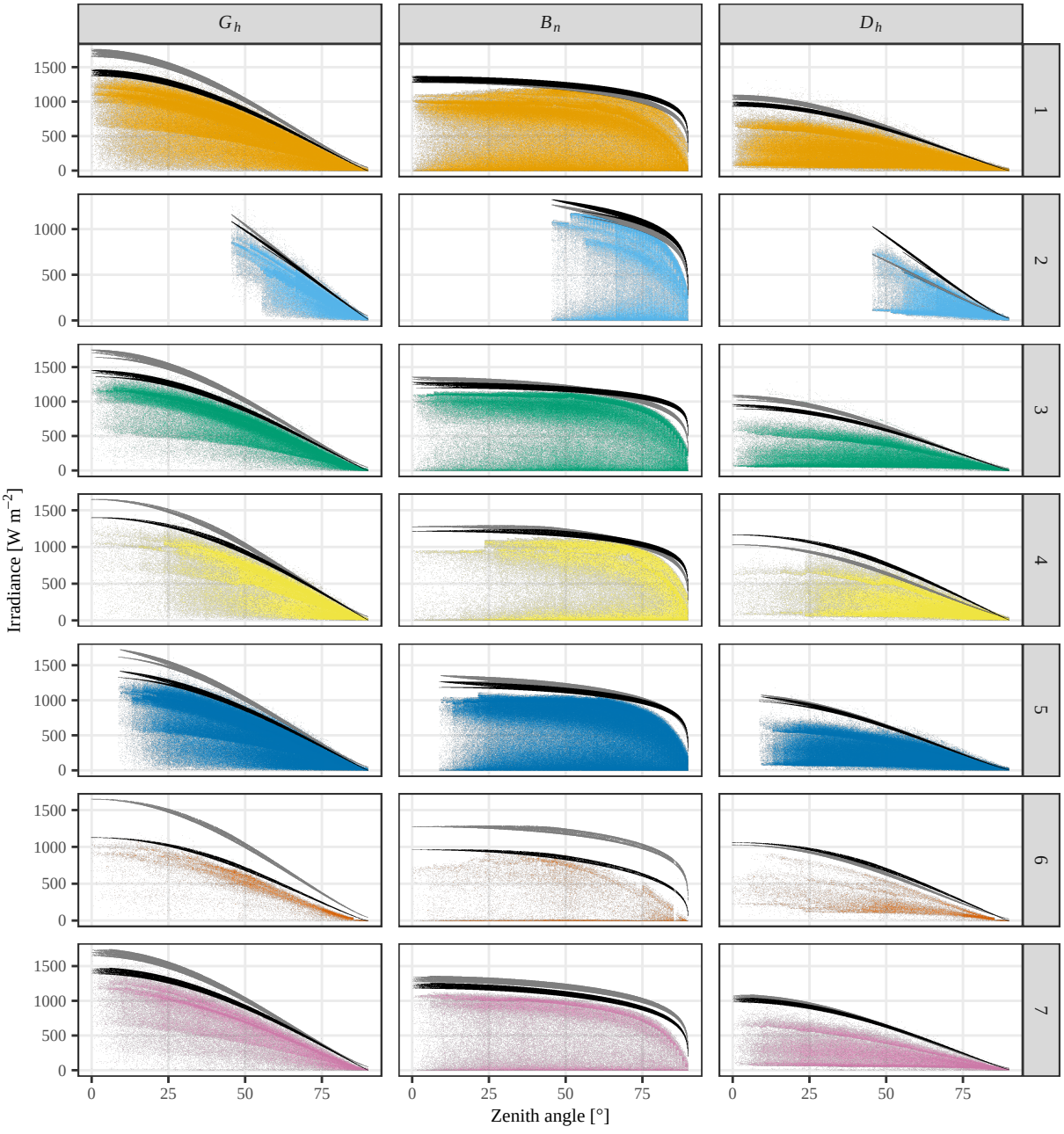

Figure 7. ERL tests for G_h, B_n, and D_h across seven radiation regimes. The configured and standard upper bounds correspond to the black and gray "eyebrows" in the figure, respectively.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

Table 3. Optimized parameters for the proposed QC limits across seven radiation climatic regimes.

	Regime	a_G	b_G	c_G	$F_{1,G}$	a_B	b_B	c_B	$F_{1,B}$	a_D	b_D	c_D	$F_{1,D}$
1	TSS	1.03	1.06	0.00	0.98	0.95	0.15	10.00	0.98	0.70	1.07	3.43	0.98
2	PSP	1.10	1.05	11.77	0.98	0.99	0.18	10.00	0.98	1.10	1.23	21.36	0.98
3	ASD	1.03	1.05	0.00	0.99	0.90	0.11	10.00	0.98	0.68	1.03	0.00	0.98
4	MCM	1.05	1.01	0.00	0.99	0.90	0.14	10.00	0.98	0.87	1.02	0.00	0.99
5	TML	1.01	1.04	0.00	0.99	0.89	0.17	10.00	0.98	0.75	1.14	7.93	0.98
6	HAA	0.85	1.14	0.00	0.99	0.72	0.35	10.00	0.99	0.80	1.04	0.00	0.99
7	LLA	1.04	1.02	0.00	0.99	0.89	0.23	0.00	0.98	0.73	1.06	7.05	0.98

Finally, D_h limits show substantial regional variability, especially in the amplitude factor a_D, which controls the theoretical upper envelope for diffuse scattering. The optimized results indicate that the upper boundary of diffuse irradiance is strongly regime dependent. For example, the stable clear-sky conditions of Regime 3 produce a comparatively low envelope ($a_D = 0.68$). In contrast, strong-scattering environments require much higher limits to avoid falsely flagging valid data; Regime 2 is a clear example, with $a_D = 1.10$ and a steeper zenith-response exponent $b_D = 1.23$. These values quantitatively represent diffuse dominance under high-albedo, multiple-scattering conditions over polar snow and ice. Overall, the coefficient contrasts confirm that a single global QC limit cannot represent the diverse scattering physics across global radiation climates.

3.3 Quality control result comparison

To evaluate the performance of the proposed climatological limits relative to the baseline BSRN ERL limits, QC outcomes are compared across all identified radiation regimes. Figure 7 presents a 7×3 matrix of scatter plots showing measured irradiance components versus solar zenith angle for all seven regimes. The superimposed "eyebrows" indicate the boundary differences between the static ERL and the configured climatological limits. For G_h, ERL thresholds are consistently broader than the proposed limits in all regimes, confirming the known tendency of ERL to be overly permissive. For B_n, the two limit sets are generally similar, except in Regime 6, where the configured limits are markedly more restrictive and better adapted to high-aerosol, low-beam conditions. For D_h, the comparison is regime dependent: ERL limits can be higher, comparable, or lower than the configured limits, which motivates further quantitative analysis.

Table 4 translates these visual differences into operational QC impacts by comparing rejection rates under the baseline ERL and the proposed climatological limits. Under standard ERL settings, QC is generally too permissive across most regions, with near-zero rejection rates in many regimes. By reducing the excess buffer above the observed physical scatter, the proposed limits tighten constraints and raise rejection rates to more informative ranges. The value of regionalization is especially clear in two regimes: polar/sub-polar (PSP) and heavy aerosol attenuation (HAA). In PSP, ERL is overly strict for the diffuse component, flagging 1.40% (10,380 points) of D_h samples. The customized limit relaxes this boundary and reduces the D_h rejection rate to 0.06%. In contrast, in the polluted HAA regime, global ERL fails to provide meaningful upper-bound control, rejecting none

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

of the G_h and D_h samples because aerosol-suppressed observations remain below the ERL ceiling. The clustered limits correct this blind spot, restoring sensitivity and increasing the G_h and D_h rejection rates to 0.47% and 0.22%, respectively.

Table 4. Component-wise rejection counts and rates. The absolute numbers are on top, and percentages bottom.

	Regime	# sample	Global (ERL)	Global (New)	Beam (ERL)	Beam (New)	Diffuse (ERL)	Diffuse (New)
1	TSS	1,748,629	133	10,391	0	0	783	6,785
			(0.01%)	(0.59%)	(0.00%)	(0.00%)	(0.04%)	(0.39%)
2	PSP	738,868	1,433	3,220	20	8	10,380	444
			(0.19%)	(0.44%)	(0.00%)	(0.00%)	(1.40%)	(0.06%)
3	ASD	804,029	82	4,879	47	38	371	5,064
			(0.01%)	(0.61%)	(0.01%)	(0.00%)	(0.05%)	(0.63%)
4	MCM	774,275	637	5,901	5	0	3,543	5,927
			(0.08%)	(0.76%)	(0.00%)	(0.00%)	(0.46%)	(0.77%)
5	TML	2,554,231	396	18,210	3	3	1,343	4,877
			(0.02%)	(0.71%)	(0.00%)	(0.00%)	(0.05%)	(0.19%)
6	HAA	83,468	0	395	1	15	0	181
			(0.00%)	(0.47%)	(0.00%)	(0.02%)	(0.00%)	(0.22%)
7	LLA	361,978	10	1,839	0	0	138	272
			(0.00%)	(0.51%)	(0.00%)	(0.00%)	(0.04%)	(0.08%)

To further explain why ERL frequently discards data in Regime 2, we conduct a false-positive analysis for the Alert (ALE) station in the Lincoln Sea (Fig. 8). An automated scan identifies a continuous 24-h period (2010-05-30) with an exceptionally high ERL rejection rate but excellent radiometric closure ($G_h \approx D_h + B_n \cos Z$). On that day, the static ERL test flags 1147 of 1440 points as erroneous. However, physical validation via the closure equation shows a mean closure error of only 1.10%, with a maximum error of 3.78% over the full diurnal cycle. This near-perfect agreement indicates that the sensors were operating correctly and that the elevated diffuse scattering was a real atmospheric signal rather than an instrumental artifact. Thus, while standard ERL fails to accommodate intense polar scattering and causes substantial data loss, the proposed climatological limits correctly envelop these valid enhancements, reducing false positives while retaining physically justified upper bounds.

4 Conclusion

Quality control (QC) of ground-based solar radiation measurements is fundamental for maintaining the integrity of surface energy balance and climatological studies. However, the widely used extremely rare limit (ERL) test is often overly conservative and can fail to isolate subtle instrumental or environmental anomalies. To improve QC tightness and sensitivity, this study presents a data-driven framework for configuring regime-specific climatological limits. Using a multidimensional unsupervised learning approach—combining principal component analysis and hierarchical clustering—BSRN stations were grouped into seven distinct radiation climatic regimes based on radiation climatological features.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

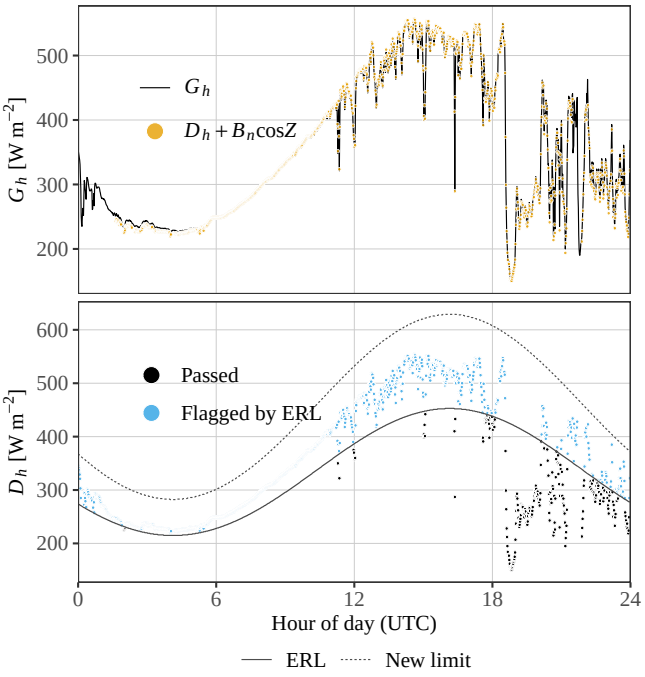

Figure 8. False-positive analysis of the ERL test during a 24-h period in Regime 2 (Polar/Sub-Polar). (Top) Radiometric closure (1.10% mean error) confirms the physical validity of the measurements. (Bottom) The ERL test incorrectly flags 1147 of 1440 valid D_h observations, whereas the proposed clustered limits successfully retain them.

For each regime, optimal coefficients for structural limit curves ($I = a \cdot E_{0n} \cdot (\cos Z)^b + c$) were determined through machine-learning-based optimization. The empirical parameters (a, b, c) were optimized within bounded constraints using the L-BFGS-B algorithm. The objective function explicitly maximized the F_1 score by benchmarking climatological-limit classifications against reference inlier–outlier labels generated by an iForest model.

Validation shows that the resulting parameter sets are more robust than the traditional ERL test, providing tighter envelopes around observations and improving QC sensitivity. The regime-specific limits adapt to local atmospheric conditions: they relax overly strict diffuse constraints in the polar and sub-polar (PSP) regime to reduce false-positive rejections linked to polar snow-albedo effects, while restoring upper-bound sensitivity in the heavily polluted HAA regime. The spatial variability of fitted parameters reflects regional differences in solar elevation, cloud regimes, and atmospheric transmissivity, supporting the physical consistency of the proposed radiation climate classification.

Overall, the proposed optimization framework mitigates the conservativeness and regional inconsistency of conventional empirical thresholds and provides a scalable, automated pathway for regionalized QC. It establishes a methodological basis for developing a unified global radiation data quality-control system. Future work will incorporate long-term meteorological reanalyses and radiative-transfer modeling to assess temporal variability and seasonal adaptability of the configured climatological limits.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

Code and data availability. The ground-based radiometric observations analyzed in this study were retrieved from the Baseline Surface Radiation Network (BSRN) repository (https://bsrn.awi.de/data/data-retrieval-via-ftp/; last access: 14 December 2025). The corresponding source code for the quality control and classification algorithms is publicly available on Gitee at https://gitee.com/dazhiyang/rad-clim-class-qc.

Appendix A: Physical characterization of the derived radiation regimes

Figure A1 supports the characterization of Regime 1 (TSS) by illustrating its seasonal bifurcation using daily maximum GHI data from the Darwin (DAR) station in 2014. The region's geographical setting yields a high and relatively stable theoretical irradiance ceiling, with the McClear clear-sky model (orange line) remaining near or above 1000 W m^{-2}. In contrast, measured daily maximum G_h values (blue points defining the gray envelope) show strong seasonal variability. During the dry season (approximately May–September), observed G_h closely follows the clear-sky limit, consistent with largely cloud-free and stable atmospheric conditions. During the monsoon season (approximately December–March), driven by intertropical convergence zone dynamics, measured maxima diverge sharply from the McClear envelope because of strong and highly variable cloud attenuation. This contrast—a smooth theoretical maximum repeatedly interrupted by deep wet-season attenuation—illustrates the regime-specific physical behavior that must be accounted for when configuring QC limits.

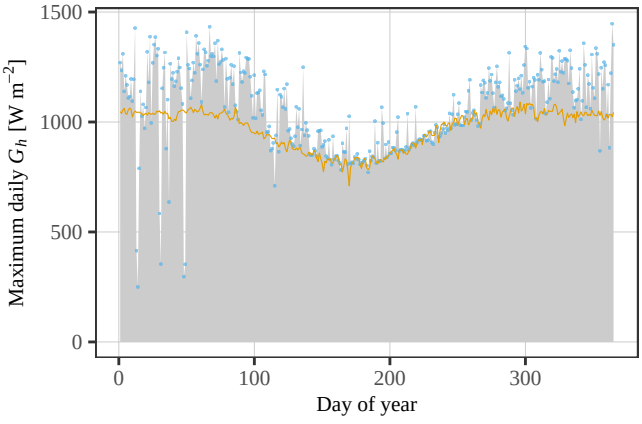

Figure A1. Visual evidence supporting the discussion of Regime 1 (TSS) in Section 3.1.

Figure A2 supports the characterization of Regime 2 (PSP) by mapping its extreme diurnal constraints using theoretical clear-sky GHI simulations for the Ny-Ålesund (NYA) station in 2021. The heatmap shows irradiance intensity by day of year (x-axis) and hour of day (y-axis), with white dashed ($\alpha = 0°$) and solid ($\alpha > 0°$ at 10° intervals) contours indicating solar elevation angle (α). Rather than stochastic daily variability, the figure displays a pronounced "hourglass" geometry governed by high-latitude solar geometry. Dark regions at the beginning and end of the annual cycle represent polar night, when irradiance remains 0 W m^{-2} throughout the day because $\alpha \leq 0°$. In contrast, the central summer period (approximately May–August) forms a continuous vertical irradiance band, indicating polar day. However, even during continuous daylight, the maximum solar elevation remains below about 35°, as shown by the contour structure. Accordingly, peak theoretical irradiance remains

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

substantially lower than at lower latitudes. This behavior highlights why Regime 2 requires dedicated QC boundaries that differ fundamentally from mid-latitude frameworks.

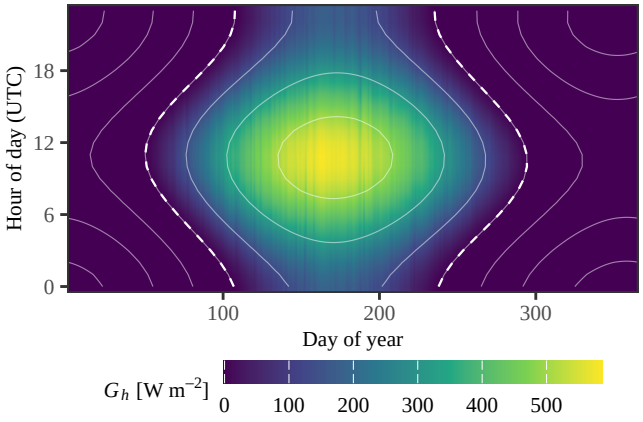

Figure A2. Visual evidence supporting the discussion of Regime 2 (PSP) in Section 3.1.

Figure A3 corroborates the environmental constraints of Regime 3 (ASD) by showing monthly probability-density distributions of the 1-min clear-sky index (κ) at the Desert Rock (DRA) station in 2021. Consistent with persistent subtropical high-pressure systems, these distributions translate the region's dry and cloud-sparse conditions into a stable statistical signature. Violin plots show that, across all months, most instantaneous κ values cluster near the dashed reference line at $\kappa = 1.0$, corresponding to near-ideal clear-sky conditions with minimal cloud attenuation. Unlike temperate or monsoon climates, which often show bimodal κ distributions from frequent cloud transitions, Regime 3 exhibits limited cloud-driven variability. The narrow spread across both winter and summer confirms that radiative transmission in this regime is largely unimpeded by transient clouds and moisture variations.

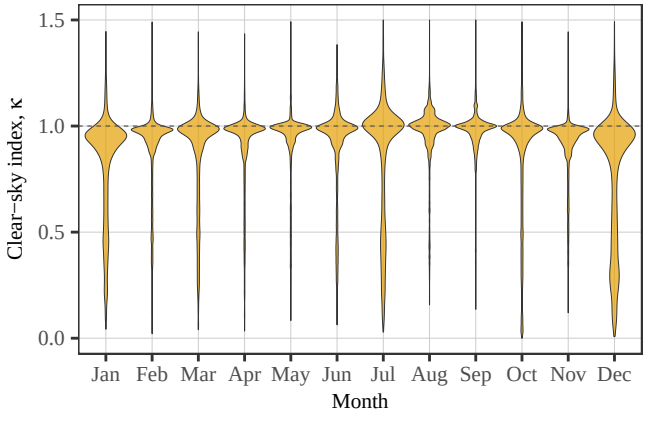

Figure A3. Visual evidence supporting the discussion of Regime 3 (ASD) in Section 3.1.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

Figure A4 quantifies the persistent radiative attenuation in Regime 4 (MCM) using kernel-density distributions of diffuse fraction ($k = D_h/G_h$). The 1-min k distribution from Lerwick, UK (Regime 4), is compared with Fort Peck, USA (Regime 5), with both sites near 55°N. As discussed in the main text, Regime 4 is strongly influenced by persistent maritime stratocumulus and frontal cloud systems. This is reflected in a clear contrast between the two distributions. The continental site shows a pronounced clear-sky peak near $k \approx 0.15$, indicating frequent direct-beam-dominant conditions. In contrast, the marine distribution lacks a comparable clear-sky peak and instead shifts strongly toward high k values approaching 1.0. This right-skewed pattern indicates sustained atmospheric scattering and confirms the diffuse-dominant nature of Regime 4.

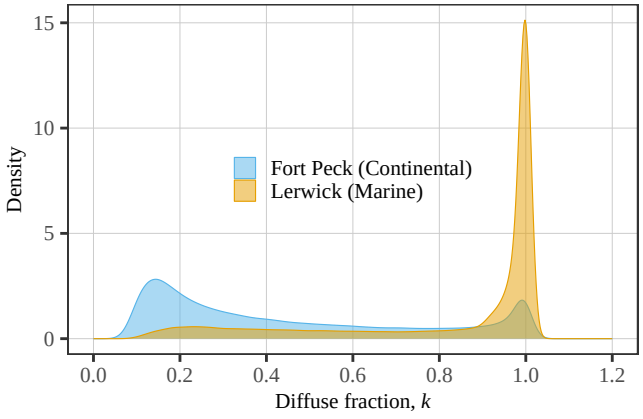

Figure A4. Visual evidence supporting the discussion of Regime 4 (MCM) in Section 3.1.

Figure A5 characterizes the radiative variability of Regime 5 (TML) using daily mean B_n and D_h time series from Payerne, Switzerland (PAY), in 2021. Under strong influence from the Westerlies, this regime is shaped by frequent synoptic weather transitions. The figure illustrates the interaction between seasonal solar forcing and short-term meteorological variability. The upper envelope of B_n follows a clear seasonal cycle controlled by mid-latitude solar elevation, while day-to-day variability remains strong. Anticyclonic periods produce clusters of high B_n values near the seasonal upper boundary, but these are repeatedly interrupted by passing fronts that sharply reduce B_n. During these frontal episodes, D_h increases rapidly, indicating fast transitions from beam-dominant to diffuse-dominant conditions. This high-frequency alternation confirms Regime 5 as one of the most dynamically variable radiation climates.

Figure A6 highlights the persistent attenuation in Regime 6 (HAA) by comparing observed B_n versus solar zenith angle for polluted Indian stations (GAN and GUR) and latitudinally comparable high-altitude clean reference sites (YUS and IZA). This pairing keeps astronomical forcing broadly comparable and isolates atmospheric-composition effects. In the figure, the standard ERL upper bound for B_n is shown as the orange dotted "eyebrow." As expected in cleaner atmospheres, the reference stations (Yushan and Izaña) produce B_n values that approach this ERL boundary. By contrast, Regime 6 stations in the Indo-Gangetic Plain show strong beam suppression across nearly all zenith angles. Even under nominally clear conditions, observations remain well below the ERL ceiling. This persistent "ceiling gap" is consistent with high aerosol optical depth (AOD) from anthropogenic pollution and dust loading, acting as a continuous atmospheric filter on surface beam irradiance.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

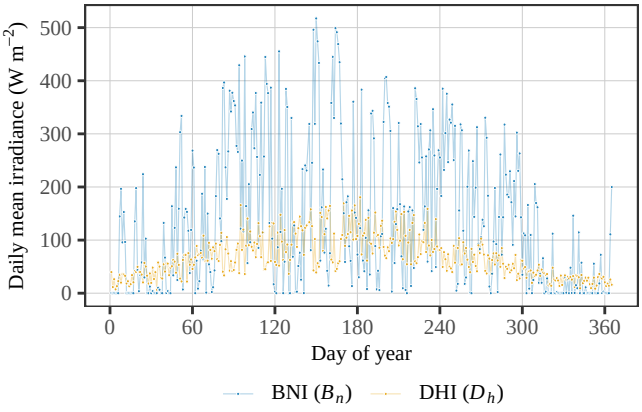

Figure A5. Visual evidence supporting the discussion of Regime 5 (TML) in Section 3.1.

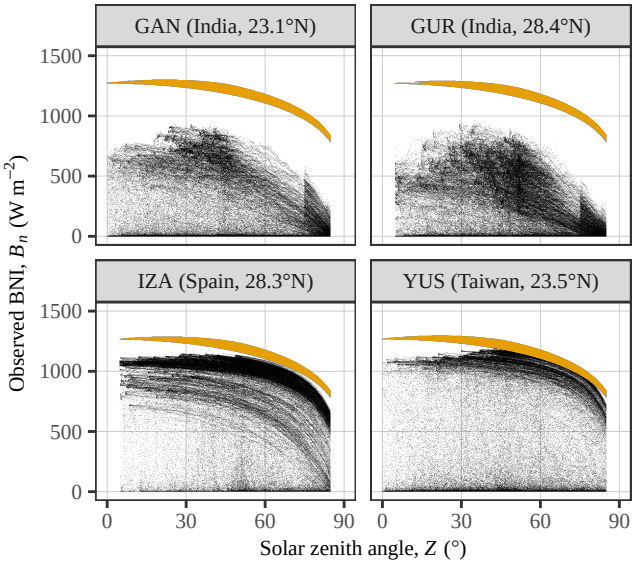

Figure A6. Visual evidence supporting the discussion of Regime 6 (HAA) in Section 3.1.

Figure A7 provides empirical support for the clustering algorithm's ability to move beyond simple geographic or latitudinal grouping by showing overlaid probability-density distributions of clearness index (k_t), diffuse transmittance (k_d), and beam transmittance (k_b) for the four Regime 7 (LLA) stations. Although these stations are geographically diverse, their distributions are statistically similar, with comparable modes and spread. This shared radiative signature persists despite large differences in altitude and local climatology, indicating that the algorithm captures structural similarities in surface radiation behavior rather than location alone. Grouping these anomalous microclimates into one cohesive regime supports the need for data-driven

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

regionalization and identifies a class that is physically distinct from typical low-latitude regimes while statistically unified in radiative response.

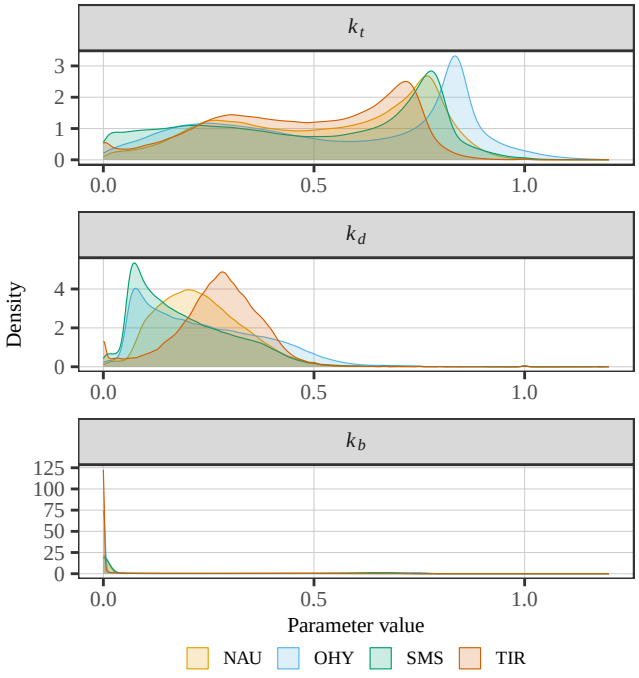

Figure A7. Visual evidence supporting the discussion of Regime 7 (LLA) in Section 3.1.

Author contributions. Zhiwen Wang: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft. Yun Chen: Methodology, Formal analysis, Writing – review & editing. Dazhi Yang: Conceptualization, Methodology, Software, Resources, Writing – original draft, Visualization, Supervision, Project administration, Funding acquisition. Hongrong Shi: Data curation, Formal analysis, Writing – review & editing, Supervision. Yanbo Shen: Validation, Writing – review & editing. Xiang'ao Xia: Validation, Writing – review & editing.

Competing interests. The contact author declares that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral concerning jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (project no. 42375192).

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

References

Anas, H., Mghouchi Youness, E., Halima, Y., Nawal, A., and Mohamed, C.: Novel climate classification based on the information of solar radiation intensity: An application to the climatic zoning of Morocco, Energy Conversion and Management, 247, 114 770, https://doi.org/10.1016/j.enconman.2021.114770, 2021.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., and Wood, E. F.: Present and future Köppen–Geiger climate classification maps at 1-km resolution, Scientific Data, 5, 180 214, https://doi.org/10.1038/sdata.2018.214, 2018.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C.: A limited memory algorithm for bound constrained optimization, SIAM Journal on scientific computing, 16, 1190–1208, 1995.

Chen, Y., Yang, D., Huang, C., Shi, H., Jensen, A. R., Xia, X., Saint-Drenan, Y.-M., Gueymard, C. A., Mayer, M. J., and Shen, Y.: Validating physical and semi-empirical satellite-based irradiance retrievals using high- and low-accuracy radiometric observations in a monsoon-influenced continental climate, Atmospheric Measurement Techniques, 18, 7315–7336, https://doi.org/10.5194/amt-18-7315-2025, 2025.

Dash, P. K., Gupta, N. C., Rawat, R., and Pant, P. C.: A novel climate classification criterion based on the performance of solar photovoltaic technologies, Solar Energy, 144, 392–398, https://doi.org/10.1016/j.solener.2017.01.046, 2017.

de las Heras, F. J. T., Isabella, O., and Vogt, M. R.: A machine learning approach to PV-climate classification, Renewable Energy, 256, 123 685, https://doi.org/10.1016/j.renene.2025.123685, 2026.

Driemel, A., Augustine, J., Behrens, K., Colle, S., Cox, C., Cuevas-Agulló, E., Denn, F. M., Duprat, T., Fukuda, M., Grobe, H., Haeffelin, M., Hodges, G., Hyett, N., Ijima, O., Kallis, A., Knap, W., Kustov, V., Long, C. N., Longenecker, D., Lupi, A., Maturilli, M., Mimouni, M., Ntsangwane, L., Ogihara, H., Olano, X., Olefs, M., Omori, M., Passamani, L., Pereira, E. B., Schmithüsen, H., Schumacher, S., Sieger, R., Tamlyn, J., Vogt, R., Vuilleumier, L., Xia, X., Ohmura, A., and König-Langlo, G.: Baseline Surface Radiation Network (BSRN): structure and data description (1992–2017), Earth System Science Data, 10, 1491–1501, https://doi.org/10.5194/essd-10-1491-2018, 2018.

Elias, T., Ferlay, N., Chesnoiu, G., Chiapello, I., and Moulana, M.: Regional validation of the solar irradiance tool SolaRes in clear-sky conditions, with a focus on the aerosol module, Atmospheric Measurement Techniques, 17, 4041–4063, https://doi.org/10.5194/amt-17-4041-2024, 2024.

Hofmann, H., Wickham, H., and Kafadar, K.: Letter-value plots: Boxplots for large data, Journal of Computational and Graphical Statistics, 26, 469–477, https://doi.org/10.1080/10618600.2017.1305277, 2017.

Hollands, K. G. T. and Suehrcke, H.: A three-state model for the probability distribution of instantaneous solar radiation, with applications, Solar Energy, 96, 103–112, https://doi.org/10.1016/j.solener.2013.07.007, 2013.

Horn, L. H. and Bryson, R. A.: Harmonic analysis of the annual march of precipitation over the United States, Annals of the Association of American Geographers, 50, 157–171, https://doi.org/10.1111/j.1467-8306.1960.tb00342.x, 1960.

Jakubauskas, M., Legates, D., and Kastens, J.: Harmonic analysis of time-series AVHRR NDVI data, Photogrammetric Engineering and Remote Sensing, 67, 461–470, 2001.

Jiang, H., Lu, N., Qin, J., and Yao, L.: Hierarchical identification of solar radiation zones in China, Renewable and Sustainable Energy Reviews, 145, 111 105, https://doi.org/10.1016/j.rser.2021.111105, 2021.

Jurado, M., Caridad, J. M., and Ruiz, V.: Statistical distribution of the clearness index with radiation data integrated over five minute intervals, Solar Energy, 55, 469–473, https://doi.org/10.1016/0038-092X(95)00067-2, 1995.

Li, M.-F., Tang, X.-P., Wu, W., and Liu, H.-B.: General models for estimating daily global solar radiation for different solar radiation zones in mainland China, Energy Conversion and Management, 70, 139–148, https://doi.org/10.1016/j.enconman.2013.03.004, 2013.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

Liu, B., Yang, D., Wang, Z., Xia, X., Qiu, H., and Shen, Y.: On the closure relationship among shortwave radiometric measurements under a cold climate during winter, Solar Energy, 285, 113 119, 2025.

Long, C. N. and Dutton, E. G.: BSRN Global Network recommended QC tests, V2, Tech. Rep. 10013/epic.38770, PANGAEA, 2002.

Long, C. N. and Shi, Y.: An automated quality assessment and control algorithm for surface radiation measurements, The Open Atmospheric Science Journal, 2, https://doi.org/10.2174/1874282300802010023, 2008.

Lou, S., Li, D., and Chen, W.: Identifying overcast, partly cloudy and clear skies by illuminance fluctuations, Renewable Energy, 138, 198–211, https://doi.org/10.1016/j.renene.2019.01.080, 2019.

Nollas, F. M., Salazar, G. A., and Gueymard, C. A.: Quality control procedure for 1-minute pyranometric measurements of global and shadowband-based diffuse solar irradiance, Renewable Energy, 202, 40–55, https://doi.org/10.1016/j.renene.2022.11.056, 2023.

Ohmura, A., Dutton, E. G., Forgan, B., Fröhlich, C., Gilgen, H., Hegner, H., Heimo, A., König-Langlo, G., McArthur, B., Müller, G., Philipona, R., Pinker, R., Whitlock, C. H., Dehne, K., and Wild, M.: Baseline Surface Radiation Network (BSRN/WCRP): New Precision Radiometry for Climate Research, Bulletin of the American Meteorological Society, 79, 2115–2136, https://doi.org/10.1175/1520-0477(1998)079<2115:BSRNBW>2.0.CO;2, 1998.

Rousseeuw, P. J., Ruts, I., and Tukey, J. W.: The bagplot: A bivariate boxplot, The American Statistician, 53, 382–387, https://doi.org/10.1080/00031305.1999.10474494, 1999.

Smiti, A.: A critical overview of outlier detection methods, Computer Science Review, 38, 100 306, https://doi.org/10.1016/j.cosrev.2020.100306, 2020.

Song, M., Yang, D., Shi, H., Chen, Y., Liu, B., Shen, Y., Ding, Z., and Xia, X.: STARNet: A deep-learning algorithm for surface shortwave radiation retrieval from Fengyun-4A, Geophysical Research Letters, 52, e2025GL116 237, https://doi.org/10.1029/2025GL116237, 2025.

Terjung, W. H.: A global classification of solar radiation, Solar Energy, 13, 67–81, https://doi.org/10.1016/0038-092X(70)90008-3, 1970.

Tukey, J. W.: Mathematics and the picturing of data, in: Proceedings of the international congress of mathematicians, vol. 2, pp. 523–531, Vancouver, 1975.

Wandji Nyamsi, W., Saint-Drenan, Y.-M., Arola, A., and Wald, L.: Further validation of the estimates of the downwelling solar radiation at ground level in cloud-free conditions provided by the McClear service: the case of Sub-Saharan Africa and the Maldives Archipelago, Atmospheric Measurement Techniques, 16, 2001–2036, https://doi.org/10.5194/amt-16-2001-2023, 2023.

Wiltink, J. I., Deneke, H., van Heerwaarden, C. C., and Meirink, J. F.: Evaluating parallax and shadow correction methods for global horizontal irradiance retrievals from Meteosat SEVIRI, Atmospheric Measurement Techniques, 18, 3917–3936, https://doi.org/10.5194/amt-18-3917-2025, 2025.

Yagli, G. M., Yang, D., and Srinivasan, D.: Automatic hourly solar forecasting using machine learning models, Renewable and Sustainable Energy Reviews, 105, 487–498, https://doi.org/10.1016/j.rser.2019.02.006, 2019.

Yang, D. and Kleissl, J.: Solar Irradiance and Photovoltaic Power Forecasting, CRC Press, https://doi.org/10.1201/9781003203971, 2024.

Yang, D., Yagli, G. M., and Quan, H.: Quality control for solar irradiance data, in: 2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), pp. 208–213, https://doi.org/10.1109/ISGT-Asia.2018.8467892, 2018.

Yang, D., Wang, W., and Xia, X.: A concise overview on solar resource assessment and forecasting, Advances in Atmospheric Sciences, 39, 1239–1251, https://doi.org/10.1007/s00376-021-1372-8, 2022.

Zhang, X., Yang, D., Zhang, H., Liu, B., Li, M., Chu, Y., Wang, J., and Xia, X.: Spatial solar forecast verification with the neighborhood method and automatic threshold segmentation, Renewable and Sustainable Energy Reviews, 202, 114 655, https://doi.org/10.1016/j.rser.2024.114655, 2024.

https://doi.org/10.5194/egusphere-2026-1256
Preprint. Discussion started: 10 June 2026

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J.: Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization,
570 ACM Transactions on mathematical software (TOMS), 23, 550–560, 1997.