Data Analysis Pipeline for Comparing CLARA and CERES Outgoing Longwave Radiation

Max Reiter[1]

Supervisors: Margit Haberreiter[2], Wolfgang Finsterle[2]

[1] *ETH Zurich, Department of Physics, Zurich, Switzerland*
[2] *Physikalisch Meteorologisches Observatorium Davos and World radiation Center (PMOD/WRC), Davos, Switzerland*

(Dated: April 27, 2025)

Abstract

The Earth's Radiation Budget (ERB) is essential for climate studies, and precise measurements of the OLR, coupled with measurements of incoming solar radiation and the reflected outgoing shortwave radiation (OSR), are of key importance to understand the evolution of climate change. The Compact Lightweight Absolute Radiometer (CLARA) is an electrical substitution radiometer (ESR) that uses a novel three-cavity design to improve redundancy and monitor instrument degradation. This report presents a data processing pipeline that enables a systematic comparison between CLARA and CERES datasets. The pipeline is designed to align measurements from both instruments in space and time, apply relevant filtering criteria, and provide a basis for validating CLARA against the established CERES data. This report also serves as a guide to understanding the logic and structure of the code, facilitating future improvements and applications of the framework. The report concludes with a discussion of key challenges, underlying assumptions, and areas for potential improvement. With targeted improvements, it has the potential to evolve into a comprehensive platform for CLARA data comparison, validation, and interpretation.

1 Introduction

Understanding the Earth's Radiation Budget (ERB) is essential for monitoring and modeling climate change. The ERB quantifies the balance between incoming solar radiation and outgoing terrestrial radiation at the top of the atmosphere. A disturbance in this balance—known as the Earth Energy Imbalance (EEI)—drives changes in the climate system, such as global warming. Increased greenhouse gases reduce the amount of outgoing radiation, leading to a positive EEI, which then raises surface temperatures until a new radiative equilibrium is achieved. Because EEI is both a driver and indicator of climate change, its continuous observation is critical (Haberreiter & Finsterle, 2022).

Satellites have played a key role in measuring the ERB since the early days of space-based Earth observation. Radiometers measuring shortwave and longwave radiation have evolved over decades, becoming increasingly accurate and sophisticated. This history, going back to missions in the 1950s and 60s, shows that ERB instruments were among the first to be launched for climate-related observations (House et al., 1986; Loeb et al., 2018; Wielicki et al., 1996).

The Clouds and the Earth's Radiant Energy System (CERES) project represents one of the most significant steps in this development. Initiated by NASA as part of the Earth Observing System (EOS), CERES builds on earlier ERB missions like ERBE and aims to provide highly accurate, long-term data to understand the role of clouds and surface properties in the Earth's climate system (Wielicki et al., 1996). CERES not only measures top-of-atmosphere (TOA) radiative fluxes but also, in combination with other EOS instruments like MODIS, allows for flux estimation at the surface and within the atmosphere (Wielicki et al., 1996). The CERES instruments are scanning radiometers with three spec-

tral channels that operate in various scan modes to optimize spatial and angular coverage. Their design includes coaligned telescopes and precision bolometers capable of high temporal and spatial resolution, and the data processing pipeline combines multiple input sources and algorithms to generate climate-quality data products (Loeb et al., 2018).

In contrast to CERES, the Compact Lightweight Absolute Radiometer (CLARA) was developed by PMOD/WRC to achieve high absolute accuracy in space-based radiometry. CLARA is an electrical substitution radiometer (ESR) that uses a novel three-cavity design to improve redundancy and monitor instrument degradation (Walter et al., 2017, 2018). It is compact and lightweight, making it ideal for micro-satellite missions like NorSat-1, on which it was launched. One of CLARA's core innovations is its ability to perform measurements at 30-second cadences while minimizing instrumental uncertainties through careful design and end-to-end calibration against SI-traceable standards (Walter et al., 2017). Importantly, CLARA is the first SI-traceable radiometer capable of measuring Earth's outgoing longwave radiation (OLR) from space (Haberreiter & Finsterle, 2022).

The goal of this report is to present a data processing pipeline that enables a systematic comparison between CLARA and CERES datasets. This pipeline has been developed to align measurements from both instruments in space and time, apply relevant filtering criteria, and provide a basis for validating CLARA against the established CERES dataset. In doing so, this report also serves as a guide to understanding the logic and structure of the code, facilitating future improvements and applications of the framework.

This report begins by outlining the general approach taken in the implementation of the data processing pipeline. It then moves into a detailed explanation of the code, providing insight into the logic and functionality behind each step. To demonstrate the pipeline in practice, an initial comparison between CLARA and CERES data is presented, highlighting how the framework can be used for validation. The report concludes with a discussion of key challenges, underlying assumptions, and areas for potential improvement, rounding off with a summary of the most important findings.

2 Implementation

The data analysis pipeline is designed as a modular system that processes satellite data to compare outgoing longwave radiation (OLR) measurements from two independent sources: CLARA and CERES. It consists of two modules, each responsible for a distinct part of the processing. The first module determines relevant geographical coordinates for CLARA observations whereas the second module integrates CLARA and CERES data by spatially and temporally matching the datasets and merging relevant parameters.

Each module takes specific input data, processes it, and generates output files where the output of the first module is used as input for the second. Module 1 (`create_coordinates_file()`) has position and orientation data from NorSat as input and outputs coordinates for CLARA observations (`CLARA_position`). This is in turn, together with CLARA OLR and CERES data, the input for module 2 (`combine_CLARA_CERES()`) that outputs combined OLR values along with additional metadata required for further analysis.

The pipeline is implemented in Python, with functions for different processing steps organized into separate `.py` files. These function files are dynamically called during execution, ensuring modularity and reusabliity. For an overview of all functions and their interdependency we refer to Table A.I in the Appendix.

To allow flexibility, the pipeline uses a configuration file (`config.yaml`) where users can specify processing parameters, file directories, and filtering options. This setup makes it easy to adapt the pipeline for different research objectives or minor changes to the input data files without modifying the core codebase.

2.1 Data

The pipeline relies on three key datasets: satellite position and orientation data from NorSat, CLARA OLR data, and CERES data. Each dataset serves a specific role in the processing workflow.

NorSat is a Norwegian microsatellite that houses the compact lightweight absolute radiometer (CLARA). The pipeline uses its position and orientation data to compute where CLARA is pointed at and how observations align with

Earth's surface. This operation data is provided in `.csv` format, suitable for direct processing.
The CLARA OLR data is level 2 data of the radiometer measurements and is provided as IDL save (`.save`) files. It contains the irradiance (W/m^2) data as well as the timestamp.
The CERES data used for the comparison is single scanner footprint (SSF) level 2 data from the NOAA-20 satellite. It can be accessed through the CERES data products. The pipeline requires selected CERES parameters to match them with CLARA observations. These are the Time/Location/Angles, CERES Observed TOA Radiances, and the Surface Parameters.

2.2 Configuration File

One of the primary functions of the configuration file is to enable module selection. Users can specify whether to run the full pipeline or execute only one of the two modules. This modular approach allows for efficient debugging, testing, and variation of parameters. Additionally, the configuration file controls the filtering applied to the final results, ensuring that only relevant observations are included in the output. The filters are maximum values for the time difference between the CERES/CLARA OLR timestamps and the corresponding CLARA position timestamp. It also defines whether intermediate results should be saved during processing, which can be useful for verification and further analysis.
The configuration file manages directory paths for input, output, and intermediate files. It specifies where raw data should be loaded from, where processed files should be stored, and how filenames should be structured. Naming conventions such as prefixes, suffixes, and date formats are defined here, ensuring consistency in file management.
Several key parameters governing the data processing steps are also stored in the configuration file. These include the limiting view zenith angle, which restricts observations to a certain range based on the satellite's orientation; the field of view (FOV) angle, which determines the area captured by the sensor; the number of vertices for the footprint polygon; and the time bin size, which controls how CERES data points are grouped in temporal matching.
The structure of input files is also defined in the configuration file, ensuring compatibility with the pipeline's processing logic. Column names for satellite position, orientation, and radiation flux values are specified, preventing potential mismatches between data sources.

3 Module 1 - Get Coordinates

This module determines the geographic coordinates of CLARA measurements by computing the intersection between the satellite's line of sight and the Earth's surface. it is based on a method originally described in Haberreiter and Finsterle, 2022, which served as a foundation for the implementation used here. Figure 1 illustrates the structure of this module and is included to provide a visual reference for the following explanations.
The module begins by loading the configuration file and verifying the presence of necessary input data in the given directory (1). The NorSat position and orientation data, stored in `.csv` files for single days, is then grouped by month to facilitate processing (2). For each month, the module iterates through the available daily files, extracting relevant information and computing the required coordinates (3). In a last step, for each month, the local time and local solar zenith angle is calculated before saving the monthly file (4).
The key task in this module is computing the coordinates of CLARA's observation points. Given the satellite's position and orientation, loaded in from the NorSat file (3a), the line of sight is determined and intersected with the Earth's surface (3b-1,2). The intersection point can be determined by combining the vector equations for the sphere and the line

$$\|\mathbf{x}\|^2 = r^2, \quad \text{and} \quad \mathbf{x} = \mathbf{o} + d\mathbf{u} \tag{1}$$

where $\mathbf{x}$ are the points on the sphere and the line respectively, r is the radius of the Earth, $\mathbf{o}$ is the position of the satellite, $\mathbf{u}$ is the unit direction the satellite is pointing at and d is the distance from the origin of the line and we need to find values of d s.t. the point is both on the sphere and on the line. The combination

$$d^2\|\mathbf{u}\|^2 + 2d[\mathbf{u}\cdot\mathbf{o}] + \|\mathbf{o}\|^2 - r^2 = 0 \tag{2}$$

has the form of a quadratic formula and we can solve for d

$$d = -[\mathbf{u}\cdot\mathbf{o}] \pm \sqrt{\nabla} \tag{3}$$

with

$$\nabla = [\mathbf{u}\cdot\mathbf{o}]^2 - (\|\mathbf{o}\|^2 - r^2) \tag{4}$$

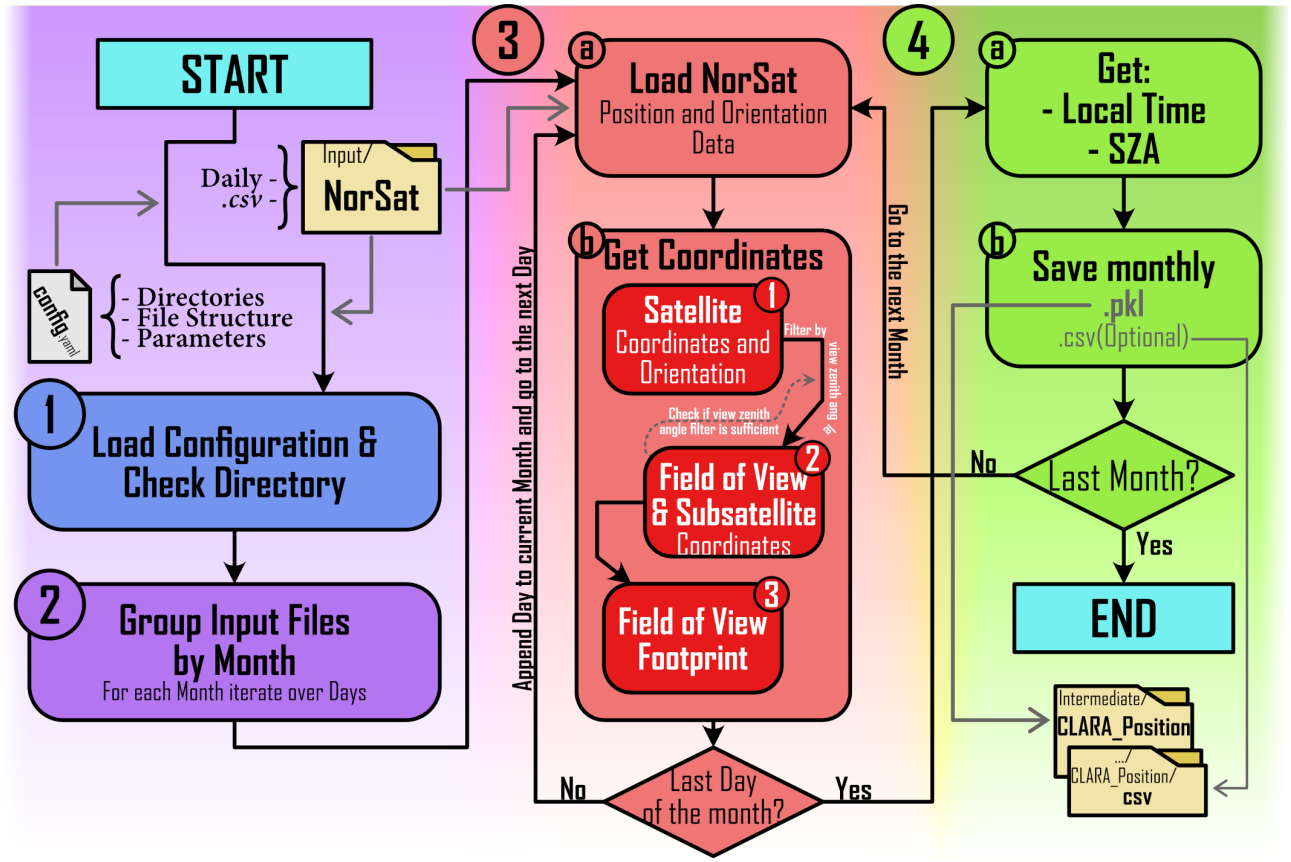

Figure 1: Flowchart of the function `create_coordinates_file()`.

where only solutions exist if $\nabla \geq 0$.

The satellite's position coordinates and the intersection coordinates are provided in the Earth-Centered Inertial (ECI) system, which needs to be converted into geodetic coordinates. This transformation is handled using the `pymap3d.ecef` package, ensuring that the calculated points align correctly with geographic locations on Earth's surface.

A crucial aspect of this process is enforcing a view zenith angle limit, which ensures $\nabla \geq 0$ and therefore only observations where Earth is within the satellite's field of view are considered. The default threshold is set to 65°, but this value is configurable through the settings file. A built-in safety check is included in the code to verify that the Earth remains within the field of view based on the defined limit, preventing cases where the satellite might be pointed too far towards space.

Another important step is determining the satellite's footprint—the region on Earth's surface that contributes to the measurement (3b-3). Conceptually, the footprint is the intersection of the satellite's *view cone* with the Earth's surface. The opening half-angle of this cone is defined in the configuration file and has a default value of 4°, which corresponds to the 50% response threshold of the instrument, as illustrated by the response curve (Figure 2). To compute the footprint, the module first determines the line of sight vector for the CLARA instrument. This vector is then pivoted outward by the specified opening half-angle, and a series of additional vectors are generated by precessing around the line of sight in small angular steps. The number of steps, which determines the number of vertices used to construct the footprint polygon, is also configurable (default is 10). Each of these vectors undergoes a line-sphere intersection calculation to determine its corresponding ground location.

One challenge in handling the footprint arises when the calculated polygon crosses the Prime Meridian, the International Date line, or is in regions near the poles. Since the footprint is stored as a polygon on a flat surface rather than on a sphere, these crossings can cause issues with continuity. To address this, the implementation automatically detects problematic cases and splits

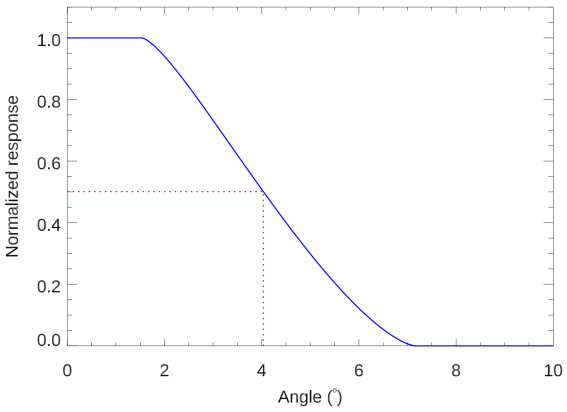

Figure 2: CLARA Response Curve. Courtesy of Margit Haberreiter

the footprint polygon into a `MultiPolygon` structure, ensuring accurate representation of the observed area.

Beyond spatial considerations, the module also calculates local time (stored as hours after midnight) and the solar zenith angle (SZA) at the observed point for each observation (4a). This information can be used for filtering out data points that are not fully on the night side of the Earth if such filtering is needed in later analyses.

After processing, the computed coordinates and associated metadata are saved as monthly `.pkl` files, which serve as input for the next module. if needed, the results can also be exported as `.csv` files, making them easier to inspect manually or analyze in external tools that do not support Python's pickle format.

4 Module 2 - Combine CLARA & CERES

This module is responsible for integrating CLARA and CERES datasets by matching them spatially and temporally and then combining relevant parameters into a single dataset. The process follows a structured workflow, beginning with loading the configuration and verifying data availability, followed by processing CLARA data, loading and matching CERES data, and finally saving the results.

To ensure that the necessary input files exist and are correctly aligned, the first step is to determine the date range covered by the available CLARA and CERES data (1). A function (`extract_file_dates()`) is used to extract the time span of each dataset which then can be used to check for overlap. This step is essential because CERES data, due to its size, is often provided in segments that do not necessarily cover an entire month as is the case for the CLARA data. Since loading excessively large files could lead to performance issues, limiting the data to the relevant time period optimizes processing.

Once the relevant CERES and CLARA OLR files have been identified (2b) they can be loaded. In tandem to loading the data, the CLAR OLR values require correction based on the instrument's field of view to convert from irradiance (W/m^2) to radiance ($W/m^2/sr$) and is included in `get_clara_olr()` (2c).

The next step is loading CERES data (3a), which is spread across multiple files, each containing different groups of variables. Since these groups are stored separately within the files, they must be loaded individually and converted into Pandas DataFrames before being combined into a single DataFrame for each group. To ensure consistency in time handling, the time values of the netCDF file are used as the primary index for all groups. This is even the case for the `Time_and_Position` group as the time values there are stored in a less intuitive format.

Once the CLARA and CERES datasets are loaded, they are matched spatially (3b). The matching process involves identifying all CERES data points that fall within the footprint of a given CLARA observation. To efficiently handle this, the CLARA footprints, which are stored as `MultiPolygons`, are first exploded into individual `Polygons` to simplify the spatial join operation. The DataFrames, both CLARA and CERES, are then converted to `GeoDataFrames` using the `Polygons`/`Points` as geometry, allowing for the spatial join operation in the first place (3b-1). A bounding box filter is then applied to reduce the number of comparisons required, improving computational efficiency. To further optimize the process, the data is converted into `Dask GeoDataFrames`, allowing for parallelized spatial operations (3b-2). A spatial join is performed to link CERES observations to CLARA footprints, with an option to save the intermediate spatial match as a `.parquet` file, as the dataset can be too large to save in a different format.

After spatial matching, a temporal match is performed by calculating the time difference Δt

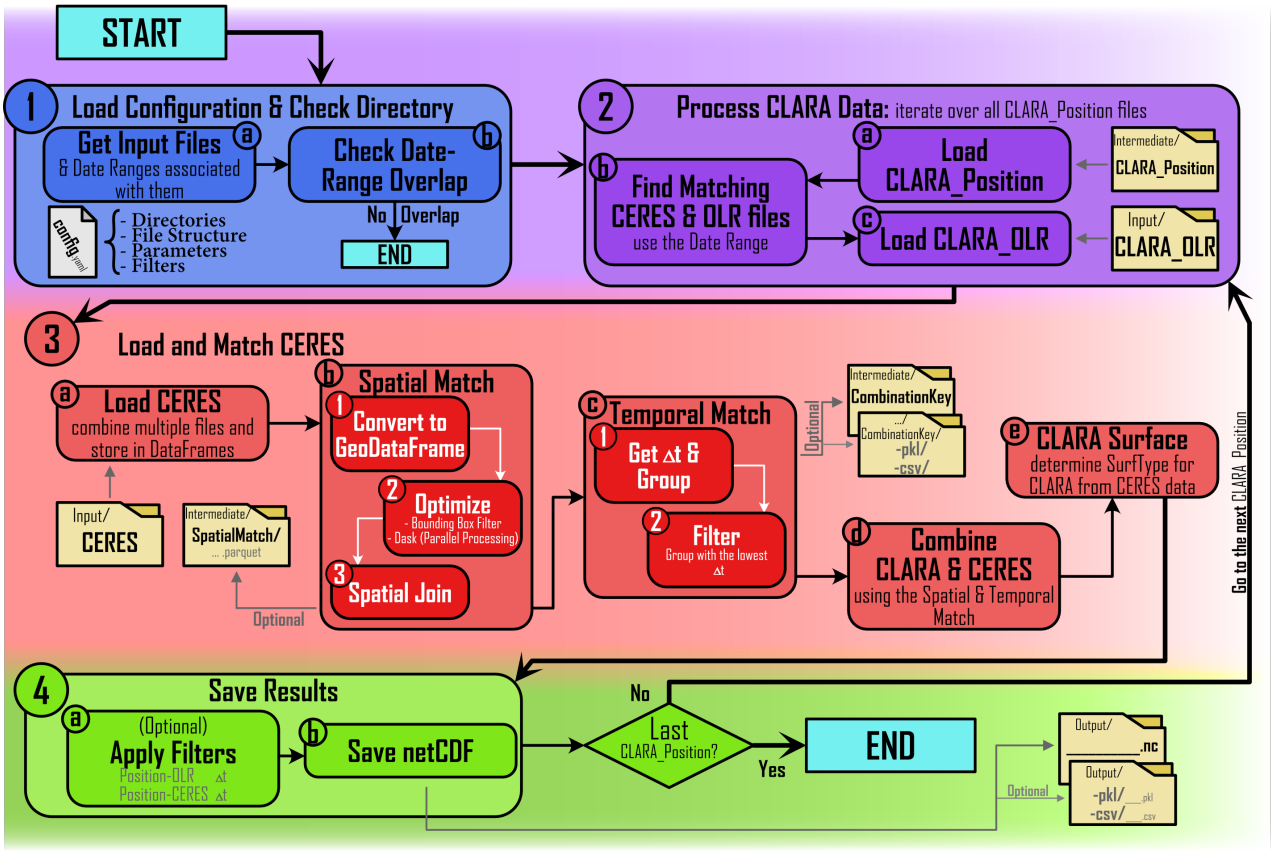

Figure 3: Flowchart of the function `combine_CLARA_CERES()`.

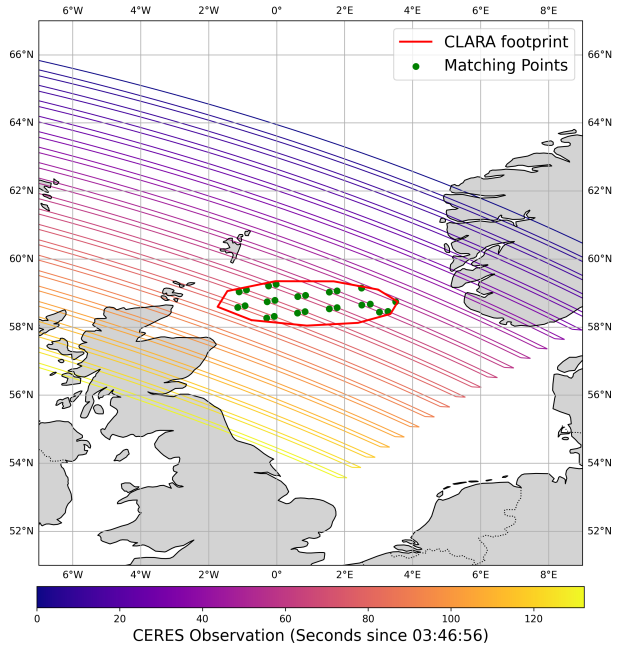

Figure 4: Illustration of the matching process. The time binned CERES Observations (gradient line) are overlapped with the CLARA footprint (red), resulting in the matching points (green).

between the CLARA observation and each associated CERES observation (3c). The data is grouped based on this time difference, using a binning approach to ensure consistent observations. Each observation is defined as a set of data points with the same rounded Δt_{CERES} value, where rounding is done according to the `time_bin_size` specified in the configuration file (5 min). To prevent observations from being split due to the rounding, an additional check is performed: if the two lowest Δt_{CERES} values differ exactly by the `time_bin_size`, both are included in the observation group. If needed, the matching process can be saved as `combination_key`, which records the links between CLARA and CERES observations for further verification or analysis. Figure 4 illustrates the matching process.

Once the matches are established, CLARA and CERES data are combined (3d). Each CLARA observation is linked to the corresponding CERES measurements using the `combination_key`. Since multiple CERES observations are associated with a single CLARA measurement, the final CERES value is computed

as the mean of the matched data points. The CLARA OLR value is selected as the one with the closets timestamp to the CLARA position data, and the time difference between them is recorded for reference ($\Delta t_{\texttt{OLR}}$).
In addition to matching radiation measurements, the module also determines the surface type associated with each CLARA footprint (3e). Using the `combination_key`, all CERES surface data points linked to a CLARA observation are extracted and combined to estimate the surface type within the footprint of the CLARA measurement.
Before saving, optional filters can be applied based on $\Delta t_{\texttt{CERES}}$ and $\Delta t_{\texttt{OLR}}$ values to remove observations with excessive time mismatches (4a). The primary output format is a `netCDF` file, which includes not only the combined data but also metadata containing all relevant configuration parameters. For an overview of all variables and metadata stored in the final file we refer to Table A.II in the appendix. If needed, the combined datasest (excluding CLARA surface information) can also be saved in `.pkl` or `.csv` formats to facilitate external analysis in environments that do not support netCDF.

5 First Comparison

This section presents an initial comparison of the processed data, demonstrating how the resulting dataset can be used for validation and parameter adjustment. By analyzing the relationship between CLARA and CERES observations, it is possible to assess the quality of the matching process and identify systematic differences that may inform refinements in the data processing pipeline.
Before any detailed analysis, an initial filtering step is applied to remove unrealistic or unreliable data points, The first criterion is based on time differences between key observation timestamps. The time difference between CLARA position data and CLARA OLR measurements ($\Delta t_{\texttt{OLR}}$) is limited so a maximum of 60 seconds, ensuring that the radiance measurement corresponds to the satellite's location within the expected exposure time. Similarly, the time difference between CLARA position data and CERES OLR measurements ($\Delta t_{\texttt{CERES}}$) is constrained to a maximum of 60 minutes, preventing the inclusion

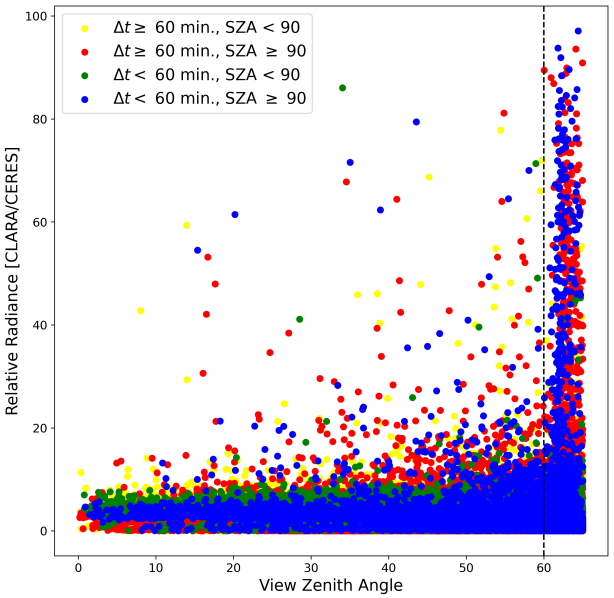

Figure 5: Plotting the Relative Radiance with respect to the view zenith angle to determine the view zenith angle filter. The different colors indicate further filtering options.

of observations where atmospheric or surface conditions may have changed significantly between measurements. These filtering steps align with the built-in options available in the pipeline and correspond to step (4a) in the Figure 3. Additionally, any physically implausible radiance values, such as negative values or those exceeding 5000 $\mathrm{W/m^2/sr}$, are removed from the dataset.
To explore possible sources of variation in the data, the relative radiance (CLARA OLR divided by CERES OLR) is plotted against different observational parameters. The most significant correlation is observed with the view zenith angle, which shows a strong increase beyond 60° (see Figure 5). As a result, data points with a view zenith angle greater than 60° are excluded to minimize errors related to extreme observation geometries. The solar zenith angle (SZA) is also considered, as it provides insight into whether the observed location is on the night side of the Earth. by filtering for SZA values greater than 90°, the analysis focuses on observations that occur in nighttime conditions, ensuring consistency in radiance measurements.
Applying these filters significantly reduces the number of available data points. For a detailed breakdown of this reduction, including the number of observations removed at each step, we refer to Table A.III in the appendix.

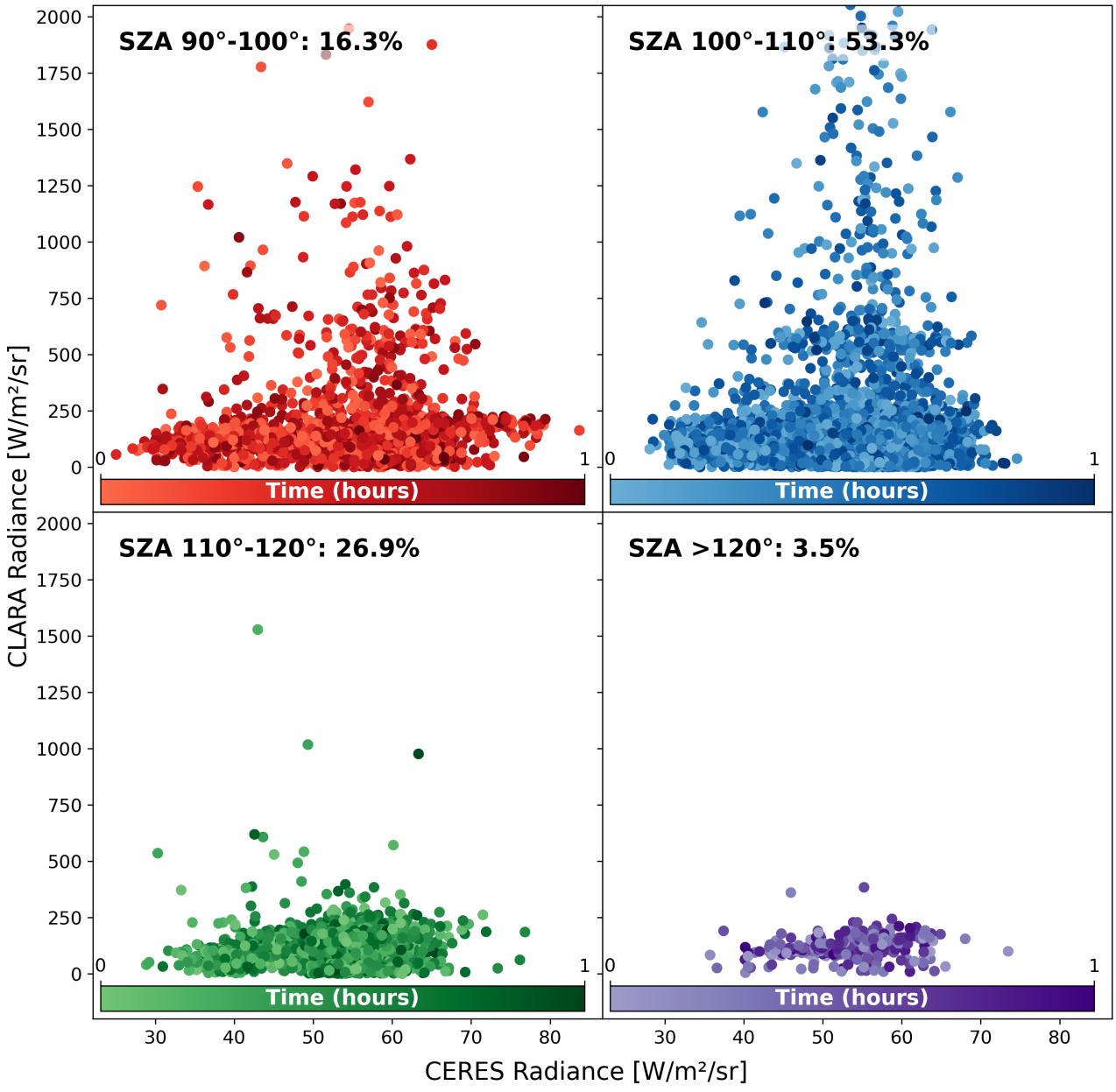

Figure 6: Scatter plot of the filtered CLARA and CERES radiance data. It is further split into solar zenith angle bins with the value and size of the bin given in the top left corner of each subplot.

After all filters are applied, a scatter plot of CLARA OLR versus CERES OLR is used to asses the correlation between the two datasets (Figure 6). In general, CLARA value tend to be higher than CERES values by a factor between 3 and 5, indicating a systematic difference that may be attributed to the calibration. Further investigation reveals a trend related to the solar zenith angle. For SZA values between 90° and 110°, there appears to be a gap in the data around CLARA OLR values of 500 $\mathrm{W/m^2/sr}$, while for SZA values greater than 110°, the data points are lower and more tightly grouped. This suggests that the increased CLARA OLR values are due to remainders of the day side of Earth. Lastly, considering $\Delta t_{\texttt{CERES}}$, the data does not show any detectable influence on the observed patterns, suggesting that the temporal matching process is robust within the defined constraints.

6 Discussion

Comparing satellite-based measurements is inherently challenging, particularly when attempting to evaluate the agreement between two different instruments with distinct observational characteristics. Ideally, such comparisons would be based on measurements taken at exactly the same time and of the same location, but in practice, achieving this level of synchrony is not straight forward. Differences in observation time can lead to changes in measured radiance, especially in dynamic atmospheric conditions where cloud cover, among other things, can shift rapidly. Spatial alignment poses a similar issue. While it is possible to identify the coordinates in the satellite's field of view at a given moment, defining what counts as the *same position* is more complex. The filed of view is not sharply bounded; instead, it is shaped by the instrument's response curve. In this case, the footprint is based on an opening angle corresponding to the 50 % response level and the differing response level within the area is not further considered. Additionally, the observed areas of CLARA and CERES may only partially overlap, making a true one-to-one comparison difficult.

Differences in view geometry add another layer of complexity. Even when observations are aligned in time and location, discrepancies in the view zenith angel between CLARA and CERES could affect the measurements.

To mitigate these issues, a series of filters were applied in section 5 to limit the dataset to observations that are more likely to be comparable. Time-based filters were used to control for the differences between CLARA position data and its associated OLR measurement ($\Delta t_{\texttt{OLR}}$), as well as the time offset between CLARA and CERES observations ($\Delta t_{\texttt{CLARA}}$). the 60-second limit for $\Delta t_{\texttt{OLR}}$ is informed by the instrument's exposure time but is still just an approximation of the position corresponding to the measurement. The broader 60-minute threshold for $\Delta t_{\texttt{CLARA}}$ is more lenient, a necessary compromise to retain enough data points for meaningful analysis, though it introduces uncertainty given how much can change in an hour.

The view zenith angle was anther key factor considered (see Figure 5). Filtering out data with high view zenith angles serves to eliminate extreme geometries where the path through the atmosphere is longer and the observed radiance is more susceptible to scattering and absorption effects in addition to the chance to measure parts of Earth still illuminated by the Sun. This filter was chosen by visually inspecting the data and identifying a threshold beyond which the values became unstable, but future considerations could benefit from a more formalized selection method. The local solar zenith angle was also used to ensure that only nighttime observations were included in the comparison. By selecting for SZA values greater than 90°, 100°, 110°, 120°, the dataset is restricted to points where the observed region is not directly or indirectly illuminated by the Sun. Although this filter significantly reduces the number of usable data points, it increases the consistancy of the observations and minimizes the influence of direct solar radiation on the measured OLR.

Throughout the entire processing and comparison workflow, several simplifying assumptions were made. The determination of ground coordinates using a line-sphere intersection presumes that Earth is a perfect sphere, which is not the case. Local topography, such as mountains or valleys, is not considered, which may introduce small spatial mismatches on the footprint estimation. Similarly, while the instrument's field of view is defined using a fixed angle corresponding to a 50 % response threshold, the actual sensitiv-

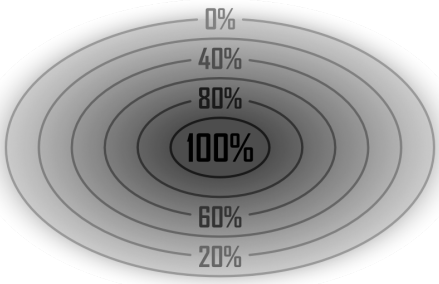

Figure 7: Qualitative illustration of the *fuzzy edge* of the FOV footprint with response values indicated.

ity of the sensor tapers gradually. This means the edges of the footprint are not as well-defined as the geometric model assumes. Finally, in the spatial matching step, the CERES data points that are linked to the CLARA footprint offer only partial coverage. this introduces additional uncertainty, as the CERES measurements may not fully represent the are observed by CLARA.
These limitation highlight the inherent complexity of satellite data integration and underscore the need for careful methodological design and transparency in assumptions. While the current implementation already provides a useful framework for comparison, acknowledging these uncertainties is essential for interpreting the results and guiding future improvements to the pipeline.

7 Conclusion

The developed pipeline serves as a valuable tool for data validation, presenting a systematic approach to examine and verify the match between CLARA and CERES observations. By processing extensive satellite datasets and executing spatial and temporal concordance, it provides an early examination of the quality and behavior of the measurements. This initial comparison aids in guiding further data analysis, identifying anomalies or patterns that may demand detailed investigation or adjustments to processing parameters. Aside from its data validation role, the codebase offers a flexible platform for future data evaluation initiatives. Its modular design facilitates the adaptation or extension of specific components, whether it's for integrating new datasets, updating matching algorithms, or refining filtering criteria. This flexibility allows for iterative enhancement: the outcome of each validation step can be reintegrated into the pipeline, guiding adjustments that improve the accuracy and reliability of future analyses. As such, it is not merely a one-time tool but rather a foundation for more extensive and specialized data comparison workflows. Despite these features, there are apparent avenues for enhancement, particularly concerning some simplifying assumptions in the current version. One improvement could involve moving past the binary footprint definition to incorporate the actual instrument response curve, accounting for the gradual sensitivity decrease towards the edge of the field of view. This approach would allow more realistic weighting of observations within the footprint area and could address uncertainties from partial overlaps. The temporal matching could also see refinement by making the maximum permitted Δt_{CERES} dynamic, adjusting it according to context variables like cloud cover. This would allow for tighter constraints under rapidly changing conditions and more leeway when atmospheric dynamics are more stable and less critical to the measurement's quality. Similarly, implementing view zenith angle corrections to simulate nadir observations might improve comparability, though it's debatable if this is essential. Previous work by Haberreiter and Finsterle, 2022 suggests that, for a Lambertian emitter, the measured radiance is independent of the viewing angle, possibly obviating the need for such corrections under these assumptions. In terms of determining surface type for CLARA observations using CERES surface data, the temporal precision of the match is arguably less significant. Given that surface properties change more gradually than atmospheric ones, a wider temporal matching window might be employed. This would increase the number of CERES points available within each CLARA footprint, enhancing spatial coverage and aiding more accurate surface type assignments. Such changes could bolster the robustness of any subsequent analysis relying on surface classification. Overall, the pipeline already provides valuable insights into the validity of CLARA data, and its modular architecture sets the stage for ongoing development. Through targeted enhancements, it holds the potential to transform into a comprehensive platform for CLARA data comparison, validation, and interpretation.

References

Haberreiter, M., & Finsterle, W. (2022). CIOMP PMOD Collaborative Research Project (tech. rep.). PMOD/WRC, Davos, Switzerland.

House, F. B., Gruber, A., Hunt, G. E., & Mecherikunnel, A. T. (1986). History of Satellite Missions and Measurements of the Earth Radiation Budget (1957-1984) (Paper 5R0785). Reviews of Geophysics, 24, 357. https://doi.org/10.1029/RG024i002p00357

Loeb, N., Su, W., Doelling, D., Wong, T., Minnis, P., Thomas, S., & Miller, W. (2018). 5.03 - earth's top-of-atmosphere radiation budget. In S. Liang (Ed.), Comprehensive remote sensing (pp. 67–84). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-409548-9.10367-7

Walter, B., Andersen, B., Beattie, A., Finsterle, W., Kopp, G., Pfiffner, D., & Schmutz, W. (2018). First tsi results and status report of the clara/norsat-1 solar absolute radiometer. Proceedings of the International Astronomical Union, 14(A30), 358–360. https://doi.org/10.1017/S1743921319004617

Walter, B., Levesque, P.-L., Kopp, G., Andersen, B., Beck, I., Finsterle, W., Gyo, M., Heuerman, K., Koller, S., Mingard, N., Oliva, A. R., Pfiffner, D., Soder, R., Spescha, M., Suter, M., & Schmutz, W. (2017). The CLARA/NORSAT-1 solar absolute radiometer: instrument design, characterization and calibration. Metrologia, 54(5), 674. https://doi.org/10.1088/1681-7575/aa7a63

Wielicki, B. A., Barkstrom, B. R., Harrison, E. F., Lee, R. B., Smith, G. L., & Cooper, J. E. (1996). Clouds and the earth's radiant energy system (ceres): An earth observing system experiment. Bulletin of the American Meteorological Society, 77(5), 853–868. https://doi.org/10.1175/1520-0477(1996)077<0853:CATERE>2.0.CO;2

A Functions

Function		Dependencies	
File: `math_and_conversion.py`			
MC.1	`safe_float_conversion()`	-	FH.5
MC.2	`latitude_to_colatitude()`	-	CO.3,4
MC.3	`colatitude_to_latitude()`	-	-
MC.4	`longitude_to_positive()`	-	CO.3,4
MC.5	`positive_longitude_to_standard()`	-	-
MC.6	`clara_radiance_conversion()`	-	FH.5
MC.7	`create_quaternion()`	-	CO.4
MC.8	`get_satellite_eci_coordinates()`	-	CO.4
MC.9	`get_satellite_pointing_vector()`	-	CO.4
MC.10	`compute_view_zenith_angle()`	-	CO.4
MC.11	`extract_polygon_coords()`	-	OP.1
MC.12	`extract_polygon_from_multipolygon()`	-	CO.3
File: `file_handling.py`			
FH.1	`extract_file_dates()`	-	M.II
FH.2	`filter_files_by_date_range()`	-	M.II
FH.3	`create_filename()`	-	M.I,II
FH.4	`load_config()`	-	M.I,II
FH.5	`read_norsat_position()`	MC.1	M.I
FH.6	`get_clara_olr()`	MC.6	M.II
FH.7	`load_and_merge_ceres()`	-	M.II
File: `coordinates.py`			
CO.1	`get_clara_fov_lonlat()`	-	CO.3,4
CO.2	`generate_fov_cone_vectors()`	-	CO.3
CO.3	`get_fov_fooprint()`	MC.2,4,12; CO.1,2	CO.4
CO.4	`get_coordinates()`	MC.2,4,7,8,9,10; CO.1,3	M.I
CO.5	`calculate_local_time_after_midnight()`	-	M.I
File: `matching_and_merging.py`			
MM.1	`match_ceres_clara_positions()`	-	M.II
MM.2	`get_nearest_time_deltas()`	-	MM.3
MM.3	`filter_ceres_clara_by_time()`	MM.2	M.II
MM.4	`add_ceres_data_to_clara()`	-	M.II
MM.5	`get_clara_surface()`	-	M.II
File: `output.py`			
OP.1	`save_netCDF()`	MC.11	M.II
File: `modules.py`			
M.I	`create_coordinates_files()`	FH.3,4,5; CO.4,5	-
M.II	`combine_CLARA_CERES()`	FH.1,2,3,4,6,7; MM.1,3,4,5; OP.1	-

Table A.I: This table provides a comprehensive list of all functions used within the pipeline, organized by the Python files in which they are contained. Each function is labeled with a unique identifier (XX.number), and the table also shows the functions they call and the functions that call them.

B NetCDF File

Variable	Units	Dimensions
time	days since 1970-01-01	(time,)
Time of observation (UTC)		
SurfType	-	(SurfType,)
Index of the most prevalent surface types		
Header/		
number_of_measurements	-	-
The total count of CLARA measurements included in this netCDF file.		
clara_max_view_zenith_angle	degrees	-
The maximum view zenith angle, in degrees, that is allowed for CLARA observations in this dataset.		
num_footprint_vertices	-	-
The number of vertices used to define the CLARA footprint area.		
clara_fov_angle	degrees	-
The half-angle, in degrees, of the CLARA field of view (FOV) cone. This defines the angular extent of the satellite's observation region.		
ceres_matching_time_bin_size	minutes	-
The time threshold, in minutes, used to group CERES matches into bins.		
filter_pos_olr_time_diff	seconds	-
The maximum allowed time difference, in seconds, between the CLARA position time and the corresponding OLR measurement time.		
filter_clara_ceres_time_diff	minutes	-
The maximum allowed mean time difference, in minutes, between the CLARA position time and the corresponding CERES observation time.		
Time_and_Position/		
instrument_fov_longitude	degrees_east	(time,)
Geodetic longitude of the instrument's surface field of view, expressed in degrees east from 0 to 360. This represents the surface intercept location observed at each time step.		
instrument_fov_latitude	degrees_north	(time,)
Geodetic latitude of the instrument's surface field of view, expressed in degrees north from -90 to 90. This location corresponds to the surface intercept point observed at each time step.		
instrument_fov_colatitude	degrees	(time,)
Geodetic colatitude of the instrument's surface field of view, expressed in degrees from 0° at the North Pole to 180° at the South Pole. This is complementary to latitude and may be used in spherical coordinate systems.		
instrument_fov_footprint	degrees	(time, footprint_vertices, lon_lat)
Geodetic coordinates (longitude, latitude) of the surface footprint vertices defining the field of view at each time step.		
UTC_time	days since 1970-01-01	(time,)
Time of each observation expressed in UTC, as days since 1970-01-01 00:00:00. This variable follows CF conventions and can be interpreted using standard calendar handling.		
local_time	hours	(time,),
Local time at the surface observation point, expressed in decimal hours after local midnight (0–24). This value is derived from UTC time and the observation longitude.		
solar_zenith_angle	degrees	(time,)
Solar zenith angle at the geodetic location and time of observation, expressed in degrees.		
Radiance/		

⋮

Continued on next page

Variable	Units	Dimensions
`longwave_radiance_clara`	$W/m^2/sr$	(time,)
Top-of-atmosphere outgoing longwave radiance measured by the CLARA instrument, expressed in watts per square meter per steradian ($W/m^2/sr$). This represents the upwelling thermal emission in the longwave spectral range at the time and location of observation.		
`longwave_radiance_ceres`	$W/m^2/sr$	(time,)
Mean top-of-atmosphere outgoing longwave radiance from CERES observations, spatially and temporally matched to the CLARA observation. Values are expressed in watts per square meter per steradian ($W/m^2/sr$).		
`Surface_Map/`		
`surface_igbp_type`	-	(time, SurfType)
International Geosphere-Biosphere Programme (IGBP) land cover type indices present within the instrument's field of view at each observation time. Each index corresponds to a specific land cover class (e.g., evergreen needleleaf forest, grassland, urban).		
`surface_igbp_type_coverage`	percent	(time, SurfType)
Percentage of each International Geosphere-Biosphere Programme (IGBP) land cover type present within the instrument's field of view at each observation time. Coverage values are expressed as percentages (0 to 100 %) of the observed area corresponding to each surface type index.		
`CLARA_Satellite/`		
`subsatellite_longitude`	degrees_east	(time,)
Longitude of the surface point directly beneath the CLARA satellite, expressed in degrees east (0–360). This value represents the location on the Earth's surface where the satellite is directly overhead at each time step.		
`subsatellite_colatitude`	degrees	(time,)
Colatitude of the surface point directly beneath the CLARA satellite, expressed in degrees (0–180), where 0° corresponds to the North Pole and 180° corresponds to the South Pole. This value represents the location on the Earth's surface where the satellite is directly overhead at each time step.		
`satellite_position_eci`	km	(time, VectorXYZ)
Position of the satellite in ECI (Earth-Centered Inertial) coordinates at each time step. The position is provided as a 3-component vector, with values for the X, Y, and Z coordinates, expressed in kilometers (km). These components define the satellite's location relative to the Earth's center.		
`satellite_pointing_vector_eci`	-	(time, VectorXYZ)
Direction in which the satellite is pointing, represented as a unit vector in ECI coordinates at each time step. The vector defines the orientation of the satellite relative to the Earth's center, with components in the X, Y, and Z directions. The vector is normalized to unit length.		
`view_zenith_angle`	degrees	(time,)
The angle between the satellite's viewing vector (or pointing vector) and the nadir vector, expressed in degrees. This represents the zenith angle of the instrument's view relative to the Earth's surface, where 0° corresponds to nadir (directly below the satellite), and larger values indicate a more oblique view.		
`CERES_Satellite/`		
`mean_solar_zenith_angle`	degrees	(time,)
The average solar zenith angle at the surface during matched CERES observations, expressed in degrees.		
`mean_relative_azimuth_angle`	degrees	(time,)
The average relative azimuth angle at the surface during matched CERES observations, expressed in degrees. The relative azimuth angle represents the angle between the Sun-satellite plane and the satellite-observation direction.		
`mean_view_zenith_angle`	degrees	(time,)
The average view zenith angle at the surface during matched CERES observations, expressed in degrees.		
`Matching_Information/`		

⋮ *Continued on next page*

Variable	Units	Dimensions
`ceres_match_times`	days since 1970-01-01	(time, match)
Timestamps of the CERES instrument observations that correspond to each CLARA observation, expressed as days since 1970-01-01 00:00:00 (UNIX epoch). These times are aligned to the matched CLARA observations.		
`num_ceres_matches`	-	(time,)
The total number of CERES observations that correspond to each CLARA observation. This value indicates how many CERES data points are spatially and temporally matched to the CLARA observation.		
`mean_ceres_match_time_diff`	seconds	(time,)
The average time difference, in seconds, between the CERES observations and the corresponding CLARA observation.		
`olr_measurement_time`	days since 1970-01-01	(time,)
The time, in days since 1970-01-01 00:00:00, of the CLARA outgoing longwave radiation (OLR) measurement that corresponds to the matched CLARA position time.		
`olr_position_time_diff`	seconds	(time,)
The time difference, in seconds, between the CLARA outgoing longwave radiation (OLR) measurement time and the corresponding CLARA position time.		

Table A.II: This table provides an overview of the variables stored in the final netCDF file, grouped by the file's internal groups. It includes the variable name, units, and dimensions, along with a brief description of each variable.

C Data reduction

Filter	No. of Datapoints	Percentage
No Filter	439,445	100.00 %
$0 \leq \mathrm{OLR}_{\mathtt{CLARA}} < 5000\,\mathrm{W/m^2/sr}$	387,663	88.22 %
$\Delta t_{\mathtt{OLR}} < 60\,\mathrm{s}$	233,944	53.24 %
$\Delta t_{\mathtt{CERES}} < 60\,\mathrm{min}$	21,927	4.99 %
SZA$\geq 90°$	9809	2.23 %
View Zenith Angle$<60°$	5802	1.32 %

Table A.III: This table illustrates the effect of each filter applied in the first comparison section 5, showing the number of data points remaining and the corresponding percentage after each filter is applied.