CAMS Service Evolution

D4.4 Localised reliability model for radiation

Due date of deliverable	30/06/2025
Submission date	26/06/2025
File Name	CAMEO-D4-4-V1.0
Work Package /Task	W4, T4.3.2 and T4.3.3
Organisation Responsible of Deliverable	DLR
Author name(s)	Jorge Lezaca, Yves-Marie Saint-Drenan, Marion Schroedter-Homscheidt
Revision number	1
Status	Issued
Dissemination Level	PUBLIC

Funded by the European Union

The CAMEO project (grant agreement No 101082125) is funded by the European Union.
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. Neither the European Union nor the granting authority can be held responsible for them.

1 Executive Summary

CAMEO objectives are to improve the knowledge on the error structures of the CAMS Radiation Service (CRS) and to provide individual time series accuracy information as new and additional information to the CRS users.

In this study, a detailed database of ground stations with the three irradiance component observations (global, beam and diffuse) was created and quality controlled. It is the basis of a systematic error assessment and the generation of 2 uncertainty error models, one based on a parametric binning approach and the other one on a machine learning approach. The error predictor space of both approaches consists of the clear and all-sky radiation parameters, atmospheric composition parameters, cloud properties, albedo model parameters and solar geometry. Following the results of the SHAP (Shapley Additive exPlanations) study, as reported previously in D4.3, the selection of initial predictors used in the training of the error models follows the priority listing as provided by the SHAP analysis.

Preliminary tests were conducted to evaluate the two approaches to model the uncertainty of CRS. The first results are very encouraging. Both probabilistic models seem to be well calibrated and capture very well the bias of CRS in the different tests considered.

On the parametric binning-based model, the stations known to be difficult to model for the CRS showed the worst calibration and sharpness, which is a reasonable and expected result. The error model was tested to infer the uncertainty distributions on time series outputs on many days/stations which included all types of sky conditions (clear, overcasted, cloudy). For all cases tested, the width of the confidence intervals correlated well with the local variability situation, i.e., narrow intervals on clear and overcasted situations and wider intervals variable situations.

The deep learning-based model was found perfectly calibrated when tested on the same stations that were used for the training, but a significant decrease in accuracy is observed when the model is applied to stations that were not used for training. A conditional evaluation indicates that in the latter case, the model is over-dispersive at low values of the clearsky index and under-dispersive at high values of the clearsky index. This lack of spatial generalisation can be attributed to an overtraining issue. Further experiments will be conducted with a smaller network and more data to address this issue.

A significant achievement of our work is the development (for the first-time) of a detailed localised error model of the CAMS radiation service. It allows the integration of new quality information for users on each individual datapoint as well as the detailed monitoring of any service evolution activities.

List of abbreviations

Abbreviation	Definition
AOD	Aerosol Optical Depth
BHI	Beam Horizontal Irradiance
BNI	Beam Normal Irradiance
CAMEO	CAMs EvOlution
CAMS	Copernicus Atmosphere Monitoring Service
CFD	Cumulated Distribution Function
CRPS	Continuous Ranked Probability Score
CRS	CAMS Radiation Service
DHI	Diffuse Horizontal Irradiance
ECE	Expected Calibration Error
ECMWF	European Centre for Medium-range Weather Forecasts
FOV	Field Of View
GHI	Global Horizontal Irradiance
LUT	Look-Up-Table
MAE	Mean Absolute Error
MBE	Mean Bias Error
MPSD	Mean Predictive Standard Deviation
MSG	Meteosat Second Generation
PICP	Prediction Interval Coverage Probability
PINAW	Prediction Interval Normalized Averaged Width
RMSE	Root Mean Square Error
SAA	Solar Azimuth Angle
SHAP	SHapley Additive exPlanations)
STDE	STandard Deviation Error
SZA	Solar Zenith Angle
tc03	Total Column of O3
tcwv	Total Column of Water Vapor

Table of Contents

2 Introduction

2.1 Background

Monitoring the composition of the atmosphere is a key objective of the European Union's flagship Space programme Copernicus, with the Copernicus Atmosphere Monitoring Service (CAMS) providing free and continuous data and information on atmospheric composition.

The CAMS Service Evolution (CAMEO) project aims at enhancing the quality and efficiency of the CAMS service and help CAMS to better respond to policy needs such as air pollution and greenhouse gases monitoring, the fulfilment of sustainable development goals, and sustainable and clean energy.

CAMEO develops methods to provide uncertainty information about CAMS products, in particular for emissions, policy, solar radiation and deposition products in response to prominent requests from current CAMS users. CAMEO contributes to the medium- to long-term evolution of the CAMS production systems and products.

The transfer of developments from CAMEO into subsequent improvements of CAMS operational service elements is a main driver for the project and is the main pathway to impact for CAMEO.

The CAMEO consortium, led by ECMWF, the entity entrusted to operate CAMS, includes several CAMS partners thus allowing CAMEO developments to be carried out directly within the CAMS production systems and facilitating the transition of CAMEO results to future upgrades of the CAMS service.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverables

In this study, we work on a localised error model, which shall provide an error estimate for an individual point in time and space. As the CAMS Radiation Service (CRS) is a time series service for a user-defined location, it uses inputs on clouds, aerosols, water vapour, ozone, and surface albedo in various spatial and temporal resolutions to irradiance estimates at the location of interest interpolated in time and space.

Validation of the CRS was so far done only as regular quality control against ground-based observations. It provides standard mean metrics as mean bias error (MBE), mean absolute error (MAE), standard deviation of the error (STDE) and root mean square error (RMSE) against ground-based observations at as many locations as possible. Due to the general lack of high-quality ground observations, it is not possible so far to assess individual error information on every time series element, considering viewing and solar geometry conditions as well as cloudy/non-cloudy status or aerosol loaded/aerosol-free.

In this context, CAMEO followed two approaches separately: on the one hand we tried to extend the database of ground-based observations towards the spatially very dense SYNOP network to detect and quantify the importance of spatial features, On the other hand we investigated if the input data space variables are suitable as predictors for an individual data point accuracy estimate.

The extension to the SYNOP network required an in-depth quality control of single-parameter stations which provide only global horizontal irradiation (GHI) observations. Calibration and data quality problems of such stations as well as validation results with a dense spatial coverage were documented in the previous deliverable D4.1. Mainly, results showed dependencies on surface elevation and irradiance and therefore pointed towards improvements in the physical retrieval method.

In this study, we focus on stations with all three radiation components (global, direct, diffuse) measured independently and therefore, with higher accuracy. Data-driven Look-up-table and machine-learning based error models are derived and evaluated.

These methods shall provide the basis for an error estimate of individual time series data points of the CRS at the user-defined spatial location. Furthermore, the error model will be the basis for future model improvements in the CRS evolution, define priorities for method developments and serve as a monitoring tool to quantify future improvements.

2.2.2 Work performed in this deliverable

In this deliverable the work as planned in the Description of Action (DoA, WP4 T4.3.2 and T4.3.3) was performed.

2.2.3 Deviations and counter measures

Task 4.3.1 was closed without investigating the online bias correction mode based on spatially high resolved SYNOP data. The Meteo-France SYNOP data usage was found to be very time consuming and still affected by biases especially in cloud free conditions, which cannot be easily solved. Reasons are likely due to reduced maintenance of the stations compared to three-component observing high-quality stations. Any bias correction scheme would therefore likely introduce biases from the observation system instead of correcting the CAMS Radiation Service.

Furthermore, most of remaining spatial patterns can be traced back to surface elevation and mean solar radiation which are independent of their spatial structure. These effects were investigated separately, and solutions with improved parameterizations of e.g. the surface elevation treatment will be implemented in the next update of the CAMS Radiation Service, more specifically in its clearsky model McClear.

The original goal of task 4.3.1, namely to prepare and test a later operational implementation of an online bias correction, is therefore not meaningful anymore. Spatially dense data does not help us for a better product – even if trained continuously or regionally/spatially dependent. Instead, an offline bias correction will be included in the quantification of the error distribution: the complete distribution of the error of CAMS Radiation service will be predicted as a new product extension. This information will include implicitly the bias.

Nevertheless, we have achieved a methodology to assess & quantify spatially (as documented in D4.1) which is very helpful for the quality monitoring inside CAMS. We may re-run the spatial assessment once we have a new major revision of CAMS Radiation Service algorithms.

2.2.4 CAMEO Project Partners:

ECMWF	EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS
Met Norway	METEOROLOGISK INSTITUTT
BSC	BARCELONA SUPERCOMPUTING CENTER-CENTRO NACIONAL DE SUPERCOMPUTACION
KNMI	KONINKLIJK NEDERLANDS METEOROLOGISCH INSTITUUT-KNMi
SMHI	SVERIGES METEOROLOGISKA OCH HYDROLOGISKA INSTITUT

BIRA-IASB	INSTITUT ROYAL D'AERONOMIE SPATIALEDE BELGIQUE
HYGEOS	HYGEOS SARL
FMI	ILMATIETEEN LAITOS
DLR	DEUTSCHES ZENTRUM FUR LUFT - UND RAUMFAHRT EV
ARMINES	ASSOCIATION POUR LA RECHERCHE ET LE DEVELOPPEMENT DES METHODES ET PROCESSUS INDUSTRIELS
CNRS	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS
GRASP-SAS	GENERALIZED RETRIEVAL OF ATMOSPHERE AND SURFACE PROPERTIES EN ABREGE GRASP
CU	UNIVERZITA KARLOVA
CEA	COMMISSARIAT A L ENERGIE ATOMIQUE ET AUX ENERGIES ALTERNATIVES
MF	METEO-FRANCE
TNO	NEDERLANDSE ORGANISATIE VOOR TOEGEPAST NATUURWETENSCHAPPELIJK ONDERZOEK TNO
INERIS	INSTITUT NATIONAL DE L ENVIRONNEMENT INDUSTRIEL ET DES RISQUES - INERIS
IOS-PIB	INSTYTUT OCHRONY SRODOWISKA - PANSTWOWY INSTYTUT BADAWCZY
FZJ	FORSCHUNGSZENTRUM JULICH GMBH
AU	AARHUS UNIVERSITET
ENEA	AGENZIA NAZIONALE PER LE NUOVE TECNOLOGIE, L'ENERGIA E LO SVILUPPO ECONOMICO SOSTENIBILE

3 Database for the CAMEO localised error model

A well depurated ground observations database is key for an in-depth analysis and correct understanding of the deviations that the operational CRS irradiation estimates have. This database should cover a very large spatio-temporal domain to allow a reliable estimation of the contribution that different components of the CRS model produce to the total irradiation deviation.

3.1 Ground observations Catalogue

The starting point for the CAMEO reference database is the stations found on the ARMINES THREDSS server catalogue (ARMINES, 2025), which is considered as the operational ground observations database of the CRS. A considerable effort is continuously spent by the ARMINES team to collect on this catalogue ground observations from high quality irradiance measurement network all around the world. An interactive online viewer/downloader of the data is also available under http://viewer.webservice-energy.org/in-situ/. Figure 1 shows a snapshot of this web viewer. It shows the location of all available stations in the catalogue (including the stations with open and non-open data policy). Stations can be filtered by network (color of the location pointer). In the case of open data, the non-registered user can download all the available observations from the station in “csv” or “netcdf” files, and view the quality control visual dashboard of the station data. The most important metadata of the stations is also displayed and the available parameters are listed. The CAMEO development team has access to all the stations in the catalogue.

Figure 1. Web viewer for the ground observations in the THREDDS server. Left: geolocations of the stations. Right: Station information including data in csv/netcdf format, visual quality check dashboard and available parameters)

To date, the catalogue contains **295** ground observation stations from the following measurement networks:

- **BOM**: Australia Region (http://www.bom.gov.au/climate/data-services/about-data-observations.shtml#tabs=Networks-and-data)
- **BSRN**: Worldwide (https://bsrn.awi.de/)
- **enerMENA**: MENA Region, see (Schüler et al., 2016)
- **ESMAP**: Wolrwide (https://globalsolaratlas.info/solar-measurement)
- **IEA-PVPS:** Worldwide (https://iea-pvps.org/research-tasks/solar-resource-for-high-penetration-and-large-scale-applications/)

- **ISE-PVlive:** Southwest Germany (https://zenodo.org/records/5196408)
- **METEO-FRANCE** : France (https://www.aeris-data.fr/en/projects/observation-data-from-the-meteo-france-ground-based-observation-network/)
- **NREL-MIDC**: USA Region (https://midcdmz.nrel.gov/)
- **OZFLUX**: Australia and New Zeland (https://www.ozflux.org.au/monitoringsites/index.html)
- **SAURAN**: South Africa Region, see (Brooks et al., 2015)
- **SKYNET** : Worldwide (https://www.skynet-isdc.org/obs_sites.php)
- **SOLRAD**: USA Region https://gml.noaa.gov/grad/solrad/)
- **SURFRAD**: USA Region (https://gml.noaa.gov/grad/surfrad/)

All ground observation stations in this catalogue have data with a temporal resolution of 1 minute. This is the temporal resolution that we have also chosen for the reference database.

3.2 Expert quality control procedure on the ground observations

It is a fact that the ground observations found on the THREDDS catalogue come from the most renowned irradiance measurement networks available. Nonetheless, the quality of the data needed for a reliable assessment of the error of the CRS irradiation estimates cannot be automatically ensured. Indeed, as part of the CAMEO project, we took the task of developing a highly selective method to filter out data of doubtful quality from these ground observations. The main objective of these pre-selection of observations is to maximize the probability that the deviations found on our posterior analysis come from the CRS model itself and not from errors/inconsistencies on the ground observations.

The steps of the selection procedure developed as part of the CAMEO project are as follow:

1. **Field of view filtering:** as the CRS irradiation estimates are only available for the field of view (FOV) of the *Meteosat Second Generation (MSG)* and *Himawari* geostationary satellites, all ground observations outside these satellite range are discarded.
2. **3-component existence:** Allow only instances with valid measurements for the 3 main irradiance components, i.e., Global Horizonal Irradiance (GHI), Diffuse Horizontal irradiance (DHI) and Beam (or Direct) normal irradiance (BNI). In this step all stations with GHI-only observations are discarded (e.g., all stations in the Meteo-France and PVLive dense pyranometric networks)
3. **Visual quality inspection**: An expert quality assessment is performed to each of the remaining ground observations. This assessment is based on the visual inspection of the quality dashboard shown in Figure 2. This dashboard includes the visualization of the data in a time series plot (in group A), in a carpet plot or day vs hour of day (in group B), calibration inspection plots (in group C), standard BSRN 1,2, and 3 component tests (Long, 2002) plots (in group D), distribution of flagged data with the BSRN component tests (in group E) and shadow detection plots (in group F). Using this visual aid, the following steps are performed:

    a. **Discarding of shifted data**: data showing any time-shift (e.g, with respect to the clearsky daily pattern) is directly discarded
    b. **Discarding of bad calibrated data**: Data that present patterns (deviate more than 0.05 from the 1 line) on the group C plots is discarded.
    c. **Discarding of station due to flagging:** when the patterns of the BSRN tests plots in group D do not correspond to the expected patterns (as those shown in Figure 2), the stations is discarded
4. **Filtering on a 6-month basis**: remaining data is then inspected in a 6 months range base. Only 6-month ranges with complete and plausible data are retained, all other 6-

months ranges are discarded. This is done to avoid intra-seasonal bias on the reference data.

5. **Data cleaning**: From the retained observations, all time instances that remained flagged are discarded from the database (turned to NaN values).

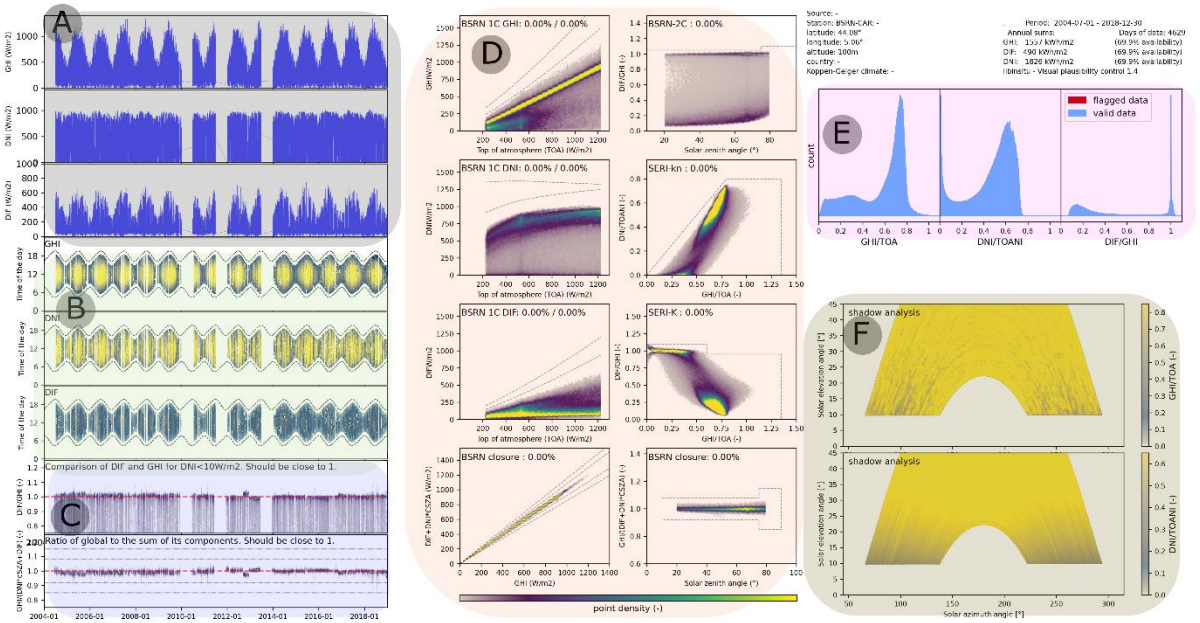

Figure 2. Visual quality dashboard for the station BSRN-CAR. It includes: time series plots (A), carpet plots (B), calibration inspection plots (C), BSRN components tests plots (D), distribution of flagged data plots (E) and shadow detection plots (F).

Figure 2 shows the visual quality check dashboard for the station BSRN-CAR because this station shows a very good quality in all the indicators presented in the dashboard. For the sake of comparison, the visual quality check dashboard for the station BSRN-KWA is shown in Figure 3. In the dashboard of Figure 3 we see clearly missing data in plot-group B. We can also confirm a bad calibration/levelling of the instruments with the plot-group C. We detect a non-expected pattern on BSRN component tests plot-group D (as seen on the BSRN closure test on the bottom-right) which is well reflected on the flag distribution on group E (red zone). Finally, we detect on the group F a shadow of about 5° in solar elevation on sunrise in the winter time (found on values of Solar Zenith Angle (SZA) between 70° and 80°).

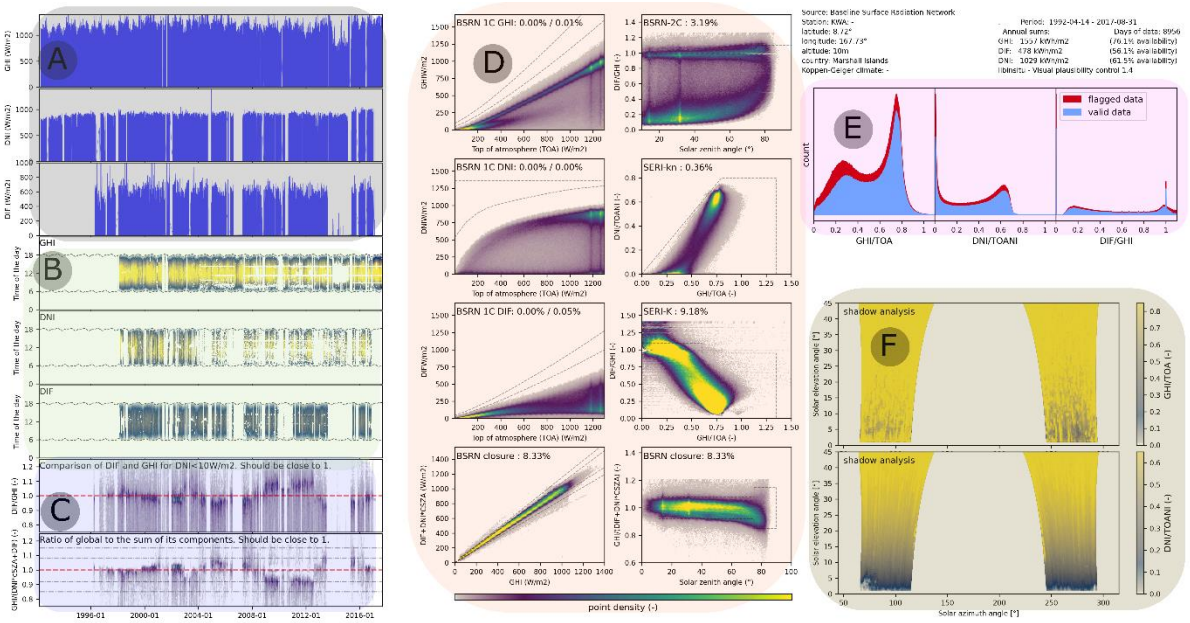

Figure 3. Visual quality dashboard for the station BSRN-KWA.
Same description as for Figure 2.

After the selection procedure described above is applied, the number of stations is reduced from **295** to **66**. The list of these retained station is shown in Table 1.

Table 1. List of stations retained for the CAMEO reference database (66 locations)

STATION NAME	Latitude [°]	Longitude [°]	Elevation [m]	Climate
ABOM-ADE	-34.952	138.521	7	Csb
ABOM-ALI	-23.798	133.888	547	BWh
ABOM-BRO	-17.949	122.234	9	BSh
ABOM-CAP	-40.672	144.688	93	None
ABOM-COC	-12.193	96.835	6	None
ABOM-DAR	-12.424	130.893	32	Aw
ABOM-GER	-28.805	114.699	30	Csa
ABOM-KAL	-30.791	121.461	368	BSh
ABOM-LEA	-22.242	114.096	6	BWh
ABOM-ROC	-23.377	150.477	12	Cfa
ABOM-TOW	-19.250	146.770	4	Aw
ABOM-TUL	-37.667	144.830	132	Cfb
ABOM-WAG	-35.160	147.456	213	Cfa
BSRN-ABS	44.018	144.280	38	None
BSRN-ASP	-23.798	133.888	547	BWh
BSRN-BUD	47.429	19.182	139	Dfb
BSRN-CAB	51.971	4.927	0	Cfb
BSRN-CAM	50.217	-5.317	88	Cfb
BSRN-CAR	44.083	5.059	100	Csb
BSRN-CNR	42.816	-1.601	471	Cfb
BSRN-DAA	-30.667	23.993	1287	BSk
BSRN-DWN	-12.424	130.893	32	Aw

BSRN-ENA	39.091	-28.029	15	N/A
BSRN-FLO	-27.605	-48.523	11	N/A
BSRN-FUA	33.582	130.376	3	N/A
BSRN-GOB	-23.561	15.042	407	BWh
BSRN-HOW	22.553	88.306	51	Aw
BSRN-INO	44.344	26.012	110	Dfb
BSRN-ISH	24.337	124.164	6	Af
BSRN-IZA	28.309	-16.499	2373	BWk
BSRN-LAU	-45.045	169.689	350	Cfb
BSRN-LER	60.139	-1.185	80	None
BSRN-LIN	52.210	14.122	125	Dfb
BSRN-LYU	22.037	121.558	324	Af
BSRN-MNM	24.288	153.983	7	None
BSRN-PAL	48.713	2.208	156	Cfb
BSRN-PAR	5.806	-55.215	4	Af
BSRN-PAY	46.815	6.944	491	Dfb
BSRN-RUN	-20.901	55.484	116	Af
BSRN-SAP	43.060	141.329	17	Dfa
BSRN-SBO	30.860	34.779	500	BWh
BSRN-TAT	36.058	140.126	25	Cfa
enerMENA-ADR	27.880	-0.274	262	BWh
enerMENA-CAI	30.036	31.009	104	BWh
enerMENA-ERF	31.491	-4.218	859	BWh
enerMENA-OUJ	34.650	-1.900	617	BSk
enerMENA-PSA	37.091	-2.358	500	Bsk
enerMENA-TAN	28.498	-11.322	75	N/A
enerMENA-TN	32.974	10.485	210	BWh
enerMENA-ZAG	30.272	-5.852	783	BWh
ESMAP-CHI	-15.548	28.248	1224	Cwa
ESMAP-CHO	-16.838	27.070	1282	Cwa
ESMAP-DAR	-6.781	39.204	190	Aw
ESMAP-KAO	-14.840	24.932	1167	Cwa
ESMAP-LUS	-15.395	28.337	1262	Cwa
ESMAP-MUT	-12.424	26.215	1317	Cwa
SAURAN-CSIR	-25.747	28.279	1400	Cwa
SAURAN-FRH	-32.785	26.845	540	Cfb
SAURAN-GRT	-32.485	24.586	660	BSh
SAURAN-NUST	-22.565	17.075	1683	BSh
SAURAN-RVD	-28.561	16.761	141	BWk
SAURAN-UBG	-24.661	25.934	1014	BSh
SAURAN-UNV	-23.131	30.424	628	Cwa
SAURAN-UNZ	-28.853	31.852	90	Cfa
SAURAN-UPR	-25.753	28.229	1410	Cwa
SAURAN-VAN	-31.617	18.738	130	BWk

Figure 4 shows the spatial distribution of the stations. We see that the spatial coverage is very good for Europe, the north/south of Africa and Australia. The coverage is much sparser in the east of south America (near the edge of the field of view of the MSG satellite) and in Japan. We have a lack of high-quality ground observations on the Saharan/sub-Sharan Africa and continental Asia.

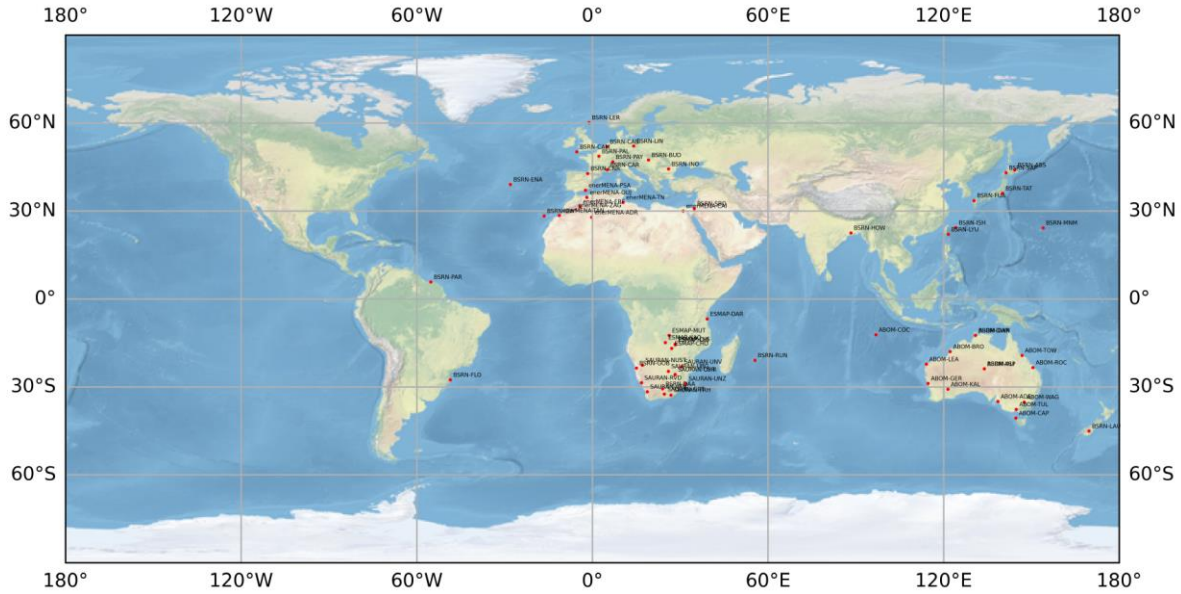

Figure 4. Ground observation stations retained for the reference database (66 locations).

The temporal availability of the reference database is shown sorted by ascending availability on Figure 5. In this figure the labels shown in the Y-axis correspond to: **number of available days - satellite - station name.** In one hand, we see that some stations retained contain more than 15 years of observations (e.g., station at the bottom: BSRN-CAB, BSRN-LIN, enerMENA-PSA). In the other hand, we see that due to the restrictive quality assessment, some stations have only 6 months of data (e.g., station at the top: BSRN-PAR, enerMENA-TAN and BSRN-HOW). In the same way, Figure 6 shows the temporal availability of the reference database sorted by satellite. From this summary, we see that the spatial and temporal availability of ground observations is greater for the MSG satellite than for the Himawari satellite.

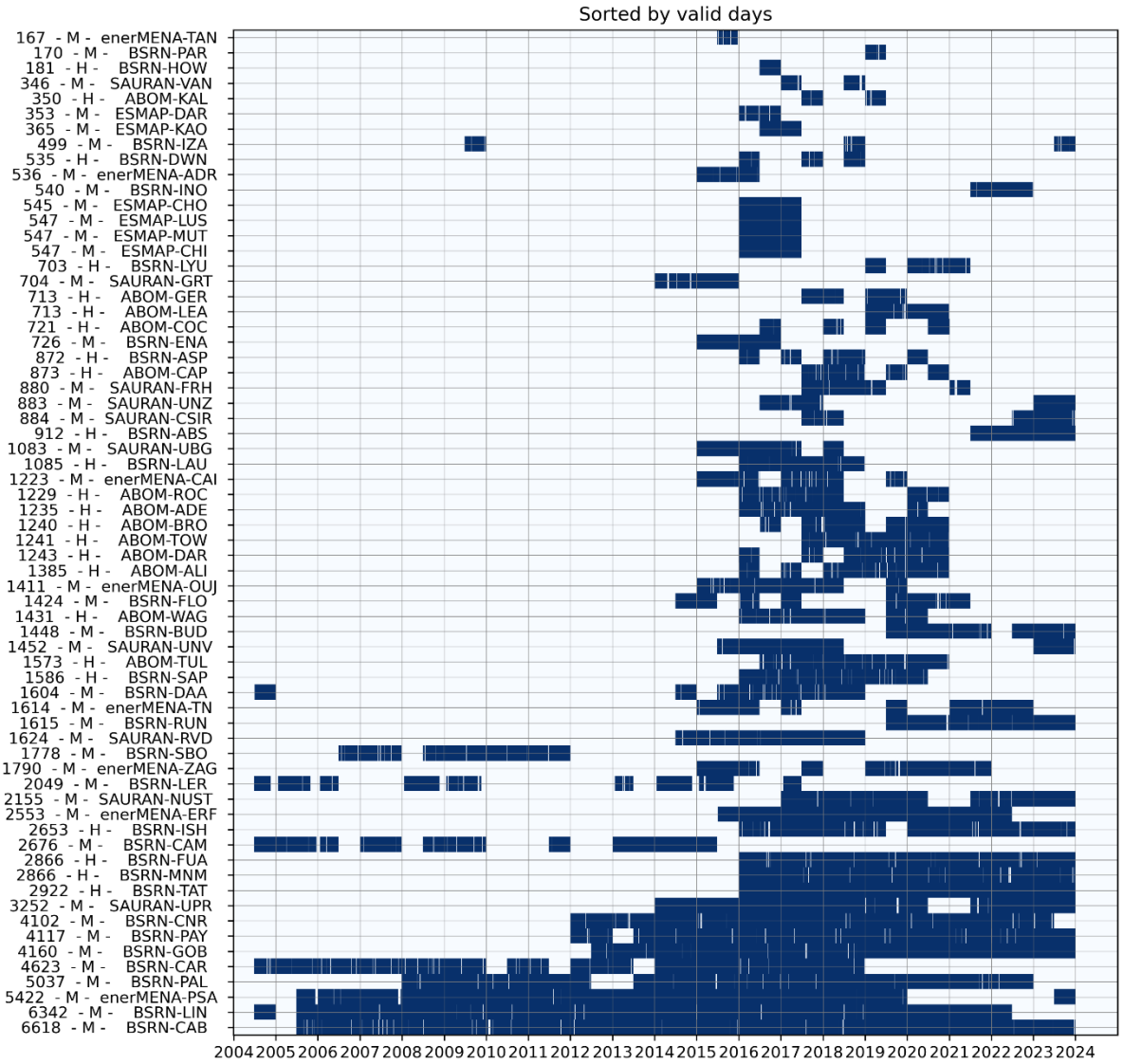

Figure 5. Temporal availability in days of the retained ground observations sorted by ascending availability. The labels on the Y axis correspond to: number of available days – satellite (M for MSG and H for Himawari) - station name.

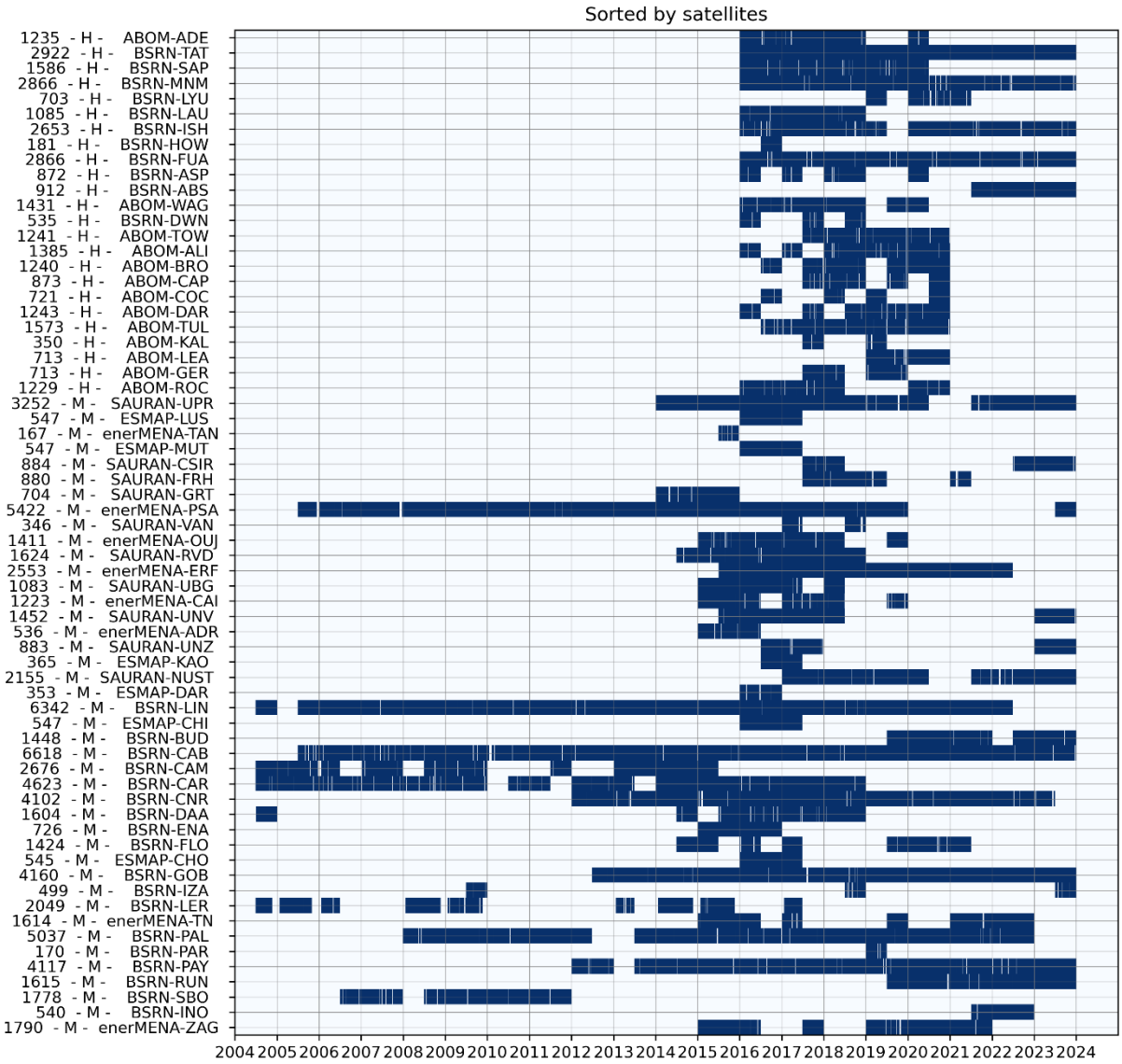

Figure 6. Temporal availability in days of the retained ground observations sorted by satellite. The labels on the Y axis correspond to: number of available days – satellite (M for MSG and H for Himawari) - station name

3.3 Creation of the database

3.3.1 CRS output data

In order to finalize the reference database, the operational expert mode output of the available timestamps for each one of the retained stations is obtained through the Vaisala CRS API (Vaisala, 2025).

The values obtained from the CRS expert mode output are:

- **Irradiation**: clear sky GHI, clear sky BHI (Beam Horizontal Irradiation), clear sky DHI, clear sky, BNI, GHI, BHI, DHI
- **Atmospheric composition**: tco3, tcwv, AOD BC, AOD DU, AOD SS, AOD OR, AOD SU, AOD NI, AOD AM, AOD SO
- **Cloud properties**: cloud optical depth, cloud coverage (probability), cloud type
- **Albedo**: fiso, fvol, fgeo, albedo

- **Other**: reliability, SZA, summer/winter split, alpha, snow probability

3.3.2 Reference dataset files creation

Once this CRS output was obtained, the dataset was created by co-aligning in time the irradiance of the ground observations retained with the output parameters of the CRS expert mode described in the previous section. This data alignment was done independently for each station. As a result, for each station a 2D numerical array is obtained, in which the first dimension is time and the second dimension is the different parameters. This array is then stored in a hdf5 file named after the station (e,g. **BSRN-CAB.h5**).

The general structure of these hdf5 files is:

```
FILE_CONTENTS {

 group      /

     dataset    /time
         attribute  /time/description
         attribute  /time/dimensions
         attribute  /time/units

     dataset    /parameters
         attribute  /parameters/description
         attribute  /parameters/dimensions
         attribute  /parameters/units

     dataset    /data
         attribute  /data/description
         attribute  /data/dimensions

}
```

Here, the ***time*** dataset is a 1D array which contains the Unix timestamps (epoch 1970) of length *N_time*, the ***parameters*** dataset is a 1D array with the names of the different parameters with length *N_params* and the ***data*** dataset is a 2D array containing the actual parameter values with a shape of (*N_time, N_params*). The grouping of the 66 hdf5 files (one per stations retained) is what we define as **the reference CAMEO dataset**.

This expert controlled dataset has been shared with all the task members and constitutes the common reference for the development of the different versions of the localised error model.

3.3.3 Reference data check

After the creation of the dataset, we proceeded to an inspection of the time series obtained. This was done to check that no error was introduced on the co-alignment of the data

3.3.3.1 Irradiance values inspection

As a first data check, we have compared the “carpet plots” (day vs time of day) of the modelled CRS irradiance estimates and the ground observations for all timestamps. Here we wanted to check plausibility of the irradiance patterns for all stations. This is shown in Figure 7 for the station BSRN-BUD.

As a result of this analysis, we have found that 3 of the locations in the HIMAWARI FOV show an un-explicable irradiance pattern during some months in the year on the diffuse and beam components. These patterns are shown in Figure 8. In this figure we see a periodical decrease of the diffuse, beam and global irradiance during the summer months. The duration of the sudden irradiance changes increases with the subsequent days until a maximum value is achieved and then decreases again until it disappears. This gives the appearance of an oval-like irradiance dip on the carpet plot on the affected station. This information and the list of

stations affected has been given to the CAMS CRS development team for further analysis. For what concerns the error model, we did not exclude the data of these stations from the reference dataset as these erroneous irradiance values make part of the operational CRS estimates (CRS v4.6) and should be considered by our localised error model.

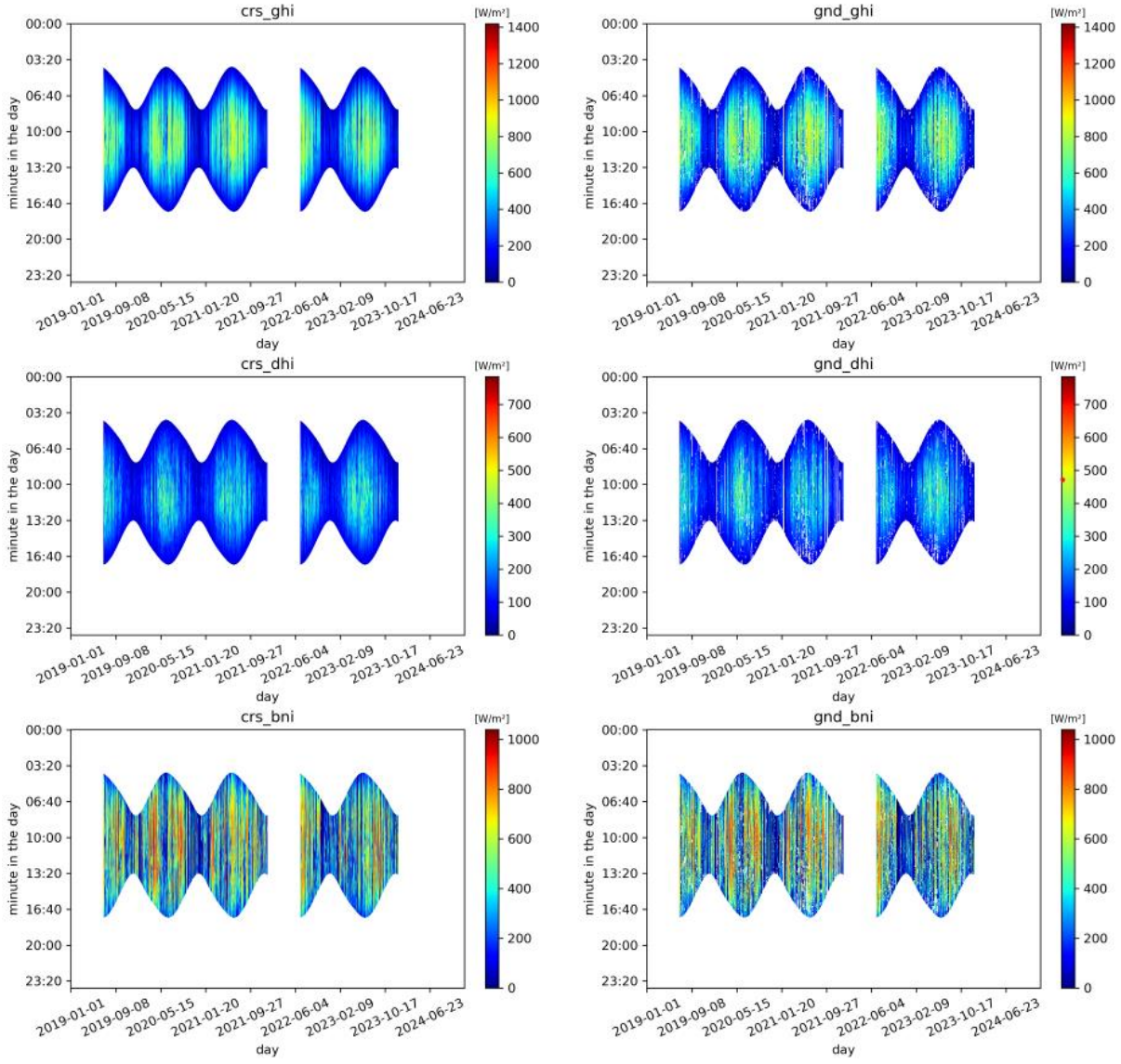

Figure 7. Reference data check: Irradiance carpet plots for the station BSRN-BUD. GHI on first row, DHI on second row and BNI on third row. CRS estimates on left column, ground observations on right column.

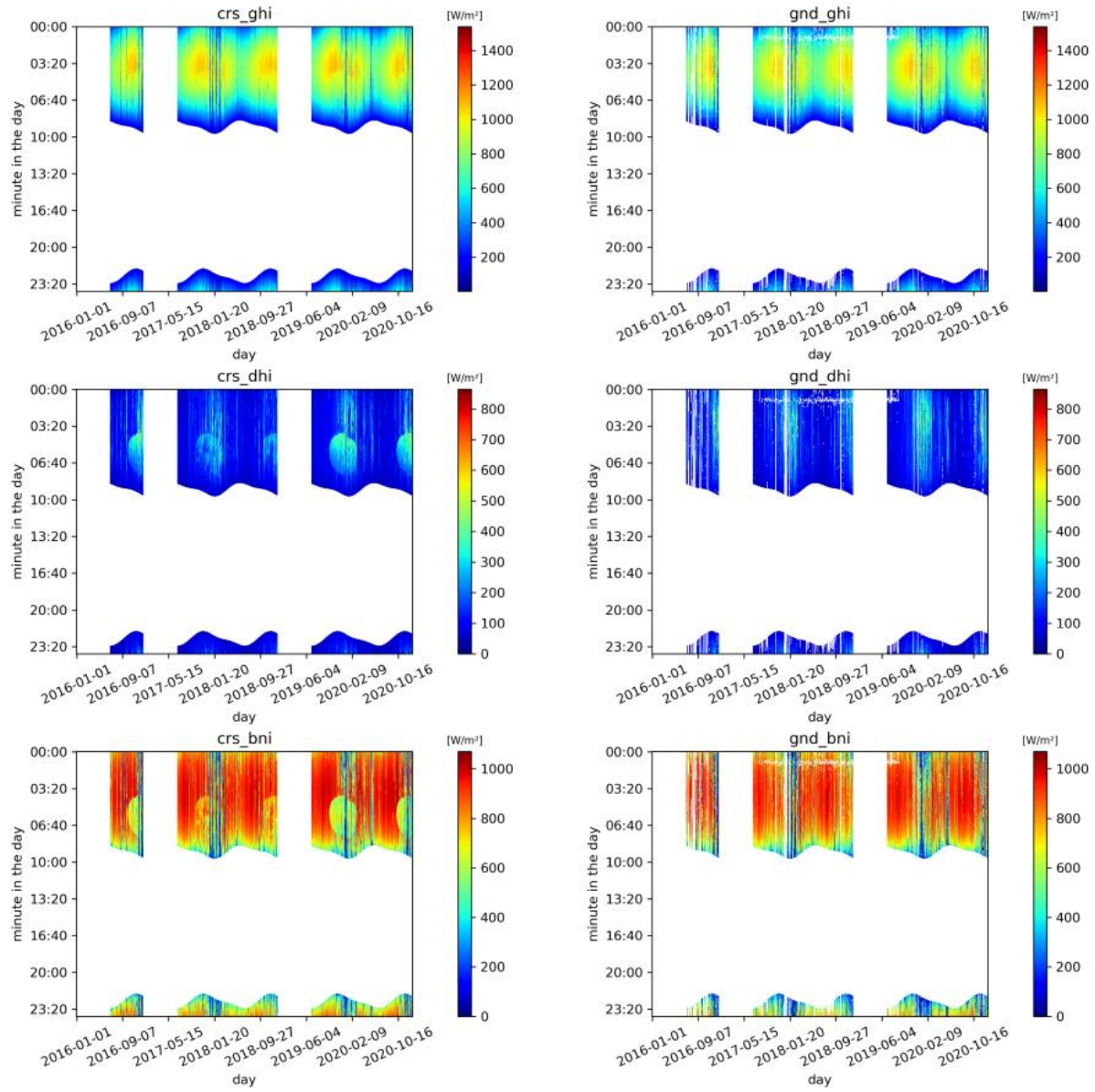

Figure 8. Reference data check: Irradiance carpet plots for the station ABOM-BRO (as in Figure 7)

No other irregularity was found on all the other stations inspected.

3.3.3.2 Deviations distributions inspection

In the same way, we have analysed the deviations found between the CRS irradiance estimates and the ground observations on all the stations in the dataset. Figure 9 shows the deviations carpet plot and the deviations distribution for the station BSRN-BUD. The distribution plots include the standard error metrics. In general, the deviations value ranges and the distributions found for the stations are found as expected. A perfectly symmetrical normal distribution of the deviations is not expected for the CRS estimates as the values do present fixed lower (0 W/m²) and upper (~1100 W/m²) irradiance limits. This can be seen specially for the diffuse and beam components.

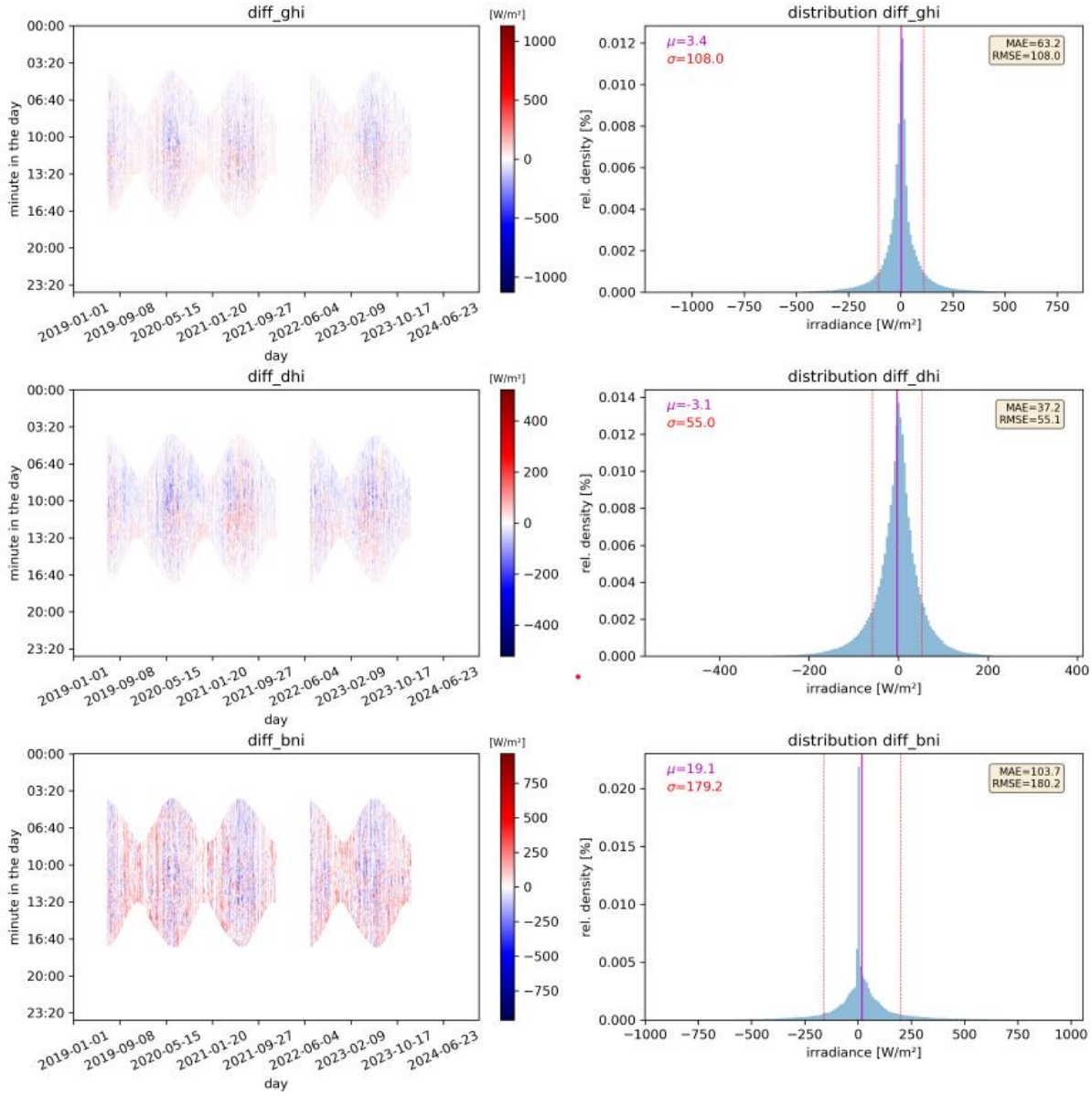

Figure 9. Reference data check: Irradiance deviations for the station BSRN-BUD. GHI on first row, DHI on second row and BNI on third row. Deviations carpet plot on the left and deviations distributions on the right. The latter plot includes error metrics MBE (μ), STDE (α), MAE and RMSE in absolute values [W/m²].

We found that for some stations the beam normal irradiance component presents a bimodal distribution, with a second peak on the negative values (overestimation). This is shown for the stations BSRN-COC and BSRN-CAP in Figure 10. The error model developed on this task will be used to assess the causes of this distribution.

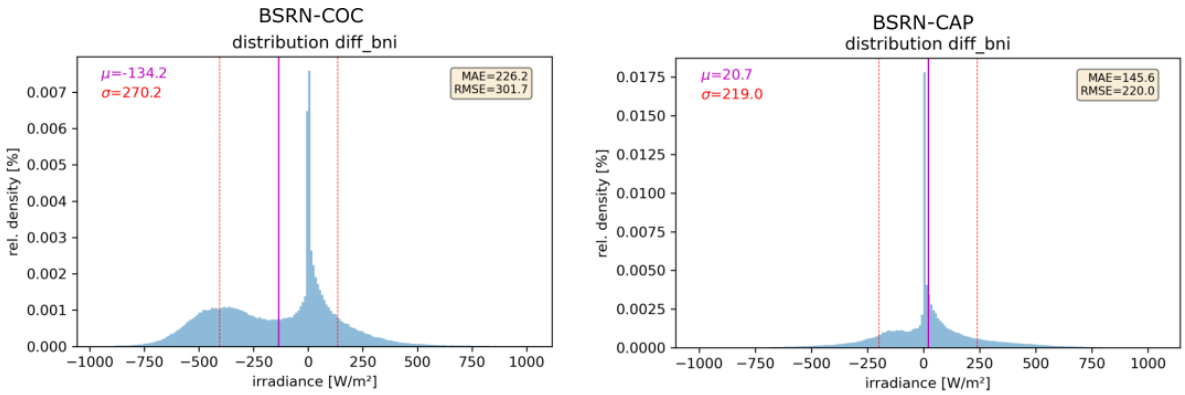

Figure 10. Reference data check: deviation distributions of the beam component for the BSRN-COC station (left) and the BSRN-CAP station (right). The error metrics MBE (μ), STDE (α), MAE and RMSE in absolute values [W/m²] are shown for each station.

3.3.3.3 Database statistical error metrics

Finally, we have checked the standard statistical error metrics for every station in the database. The MBE and STDE for each of the stations are shown in Figure 11. As expected, the metrics for BNI have higher values than for ones for GHI and the absolute value ranges are in the accordance to those found on the quarterly validation EQC reports (CAMS Radiation Serivce, 2025).

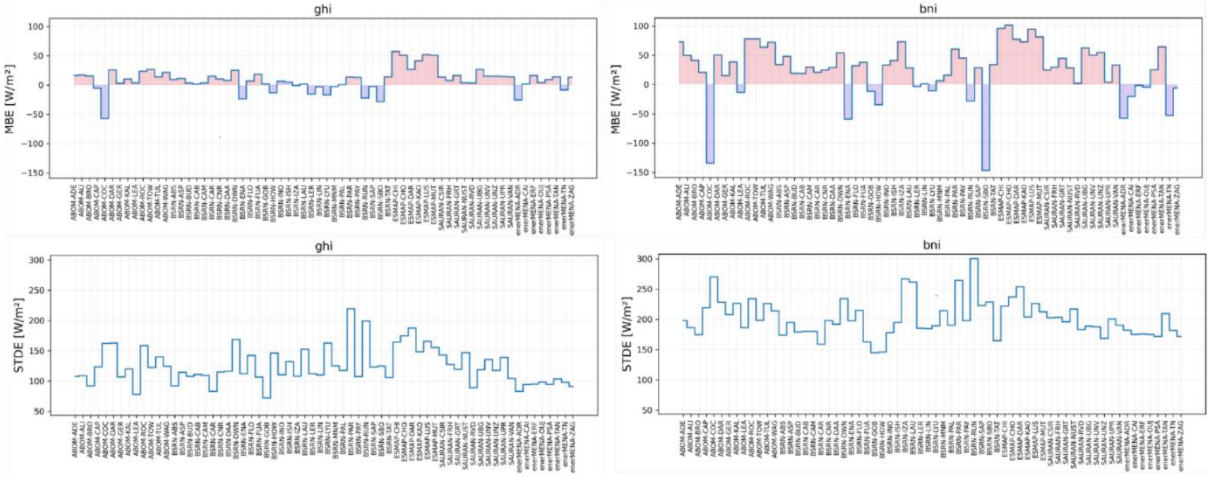

Figure 11. Dataset check: MBE and STDE for each station in the database. Top: MBE (red zone represents overestimations and blue zone represents underestimations), bottom: standard deviation, left: GHI and right: BNI.

This finalizes the sanity check of the reference CAMEO dataset. The dataset as processed until this point is the base for the development of the localised error models of the next section.

4 Localised error model 1: uncertainty inference based on parametric binning

One of the motivations of the CAMEO project is to develop a model that enables the CRS team to calculate a pixel-wise uncertainty estimate of the CRS irradiance products. Until now, the CRS irradiance products have been validated using standard aggregated error metrics at ground observation stations and aggregated over 3 months or a single year. Such results are regularly provided in the CRS quarterly validation reports (CAMS Radiation Serivce, 2025). The error metrics used in the validation reports are sufficient to get an idea of the overall performance of the CRS products but are not able to provide any error information on the individual location and time of interest. In order to breach such a limitation, as part of this project, we developed an error model optimized on the estimation of the errors for the individual timesteps at any location of the CRS domain. To develop such a model, we need to exploit a very large number of ground observations which allows us to describe the inherent errors added by the different hypothesis made on CRS input data on clouds, aerosols, water vapour, albedo as well as radiative transfer modelling algorithms. The reference dataset that we use for this development has been already processed and validated as described in section 3.

4.1 Spatio-temporal data separation

Before the reference dataset is exploited for the development of the error models, we need first to ensure that the data used for training the models represent as much as possible all the cases for which the CRS user exploits the data. The most general case is that the CRS user needs irradiance estimations on locations where no ground observations are available (resource assessment studies, PV yield calculations, etc.). This requires an estimate of the uncertainty on locations for which there is no reference as there is no ground observations available. To take this in to account, the data used for the development of our methods is separated for training and validation purposes. This is done by using N stations for the training procedure and 66 – N station for the validation/test procedures. In this way the inference on the uncertainty is done on a location where no data has been used on the training procedure. In the same way, to ensure that a data point is not use twice in the process, a time separation is also ensured. In this way, X% of the data is used for training purposes, Y% is used for validation and an unseen 100 - X - Y % is just used in a final test phase. Figure 12 shows the schematic for the spatio-temporal separation of the reference dataset.

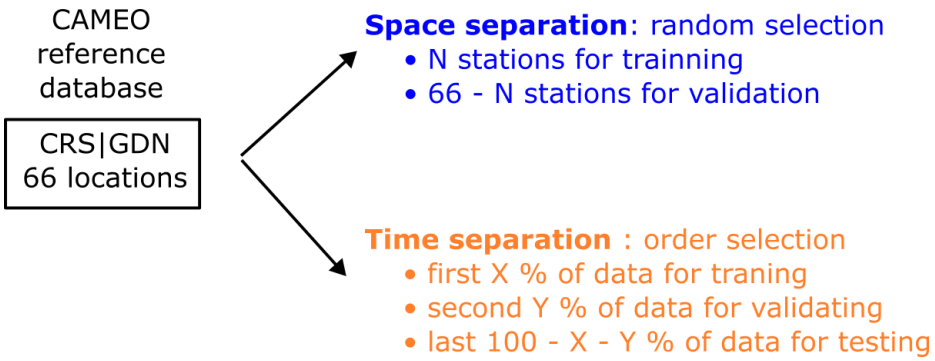

Figure 12. Spatio-temporal separation of the reference dataset.

4.2 Methodology

In order to understand the inherent characteristics of the deviations provided by the operational CRS irradiation estimates, we decided to use as a baseline an error model that characterizes the deviation distributions for every situation directly from the CRS input parameters. This baseline error model is based on the characterization of the uncertainty

distribution by means of binning the CRS model input parameters. This methodology is schematized on Figure 13. Using this method, we cluster together CRS deviations values which are calculated from the same input binning ranges (e.g, blue cube-like bin in Figure 13). The uncertainty distribution for each individual bin in the domain is estimated. This enables the attribution of a different error distribution to each bin. A look up table (LUT) is then built from these distributions. As a result, in this LUT the features are the different input parameters from the CRS model and its values correspond to the different moments that characterize the deviation distribution.

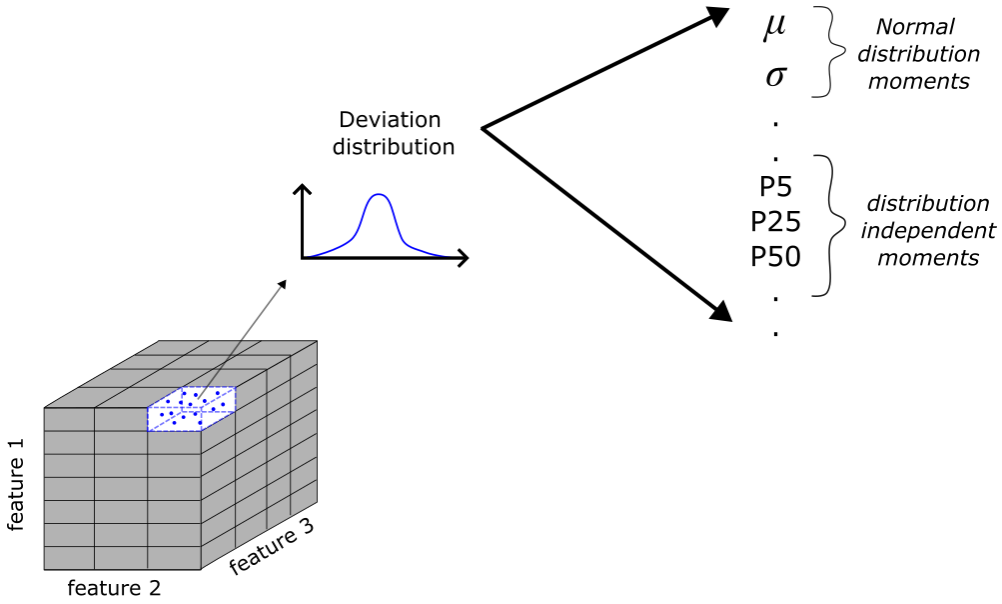

Figure 13. Baseline error model based on parametric binning. Here the features correspond to the chosen CRS model input parameters. The values inferred from the LUT correspond to the different distribution moments of the individual bins.

As a result, the input parameters of the CRS model are used directly to estimate of the uncertainty distribution on any individual location/instance.

An overarching question on this approach is which distribution function to use to describe the deviations found on the different bins (e.g., Normal distribution, Laplace distribution, etc.). This is not an easy task, as we already confirmed in the section 3.3.3.2 that the distribution of the CRS deviation can have different forms. In order to answer this question, we first use the LUT table as an inspection/monitoring tool of the CRS deviations.

4.3 Inspection run and monitoring tool

In order to visualize the base error patterns found in the CRS deviations, we create a simple LUT using the parameters of solar zenith angle (SZA), related to the position of the sun in the sky, and the clear sky index (Kc), which can be used as a simple proxy for sky conditions.

We have used for this inspection run the bin ranges for the parameters

```
SZA = {0, 10, 20, 30, 40, 50, 60, 70, 80, 90}
Kc  = {0, 0.2, 0.4, 0.6, 0.8, 1, 1.3, inf}
```

with the data of 2 stations BSRN-ABS and BSRN-BUD. These stations are known to be well modelled by the CRS.

The LUT has been then created with the standard error metrics of the CRS irradiance deviations: MBE, MAE, STDE and RMSE. Figure 14 shows the 2D visualization of such error metrics for the obtained LUT.

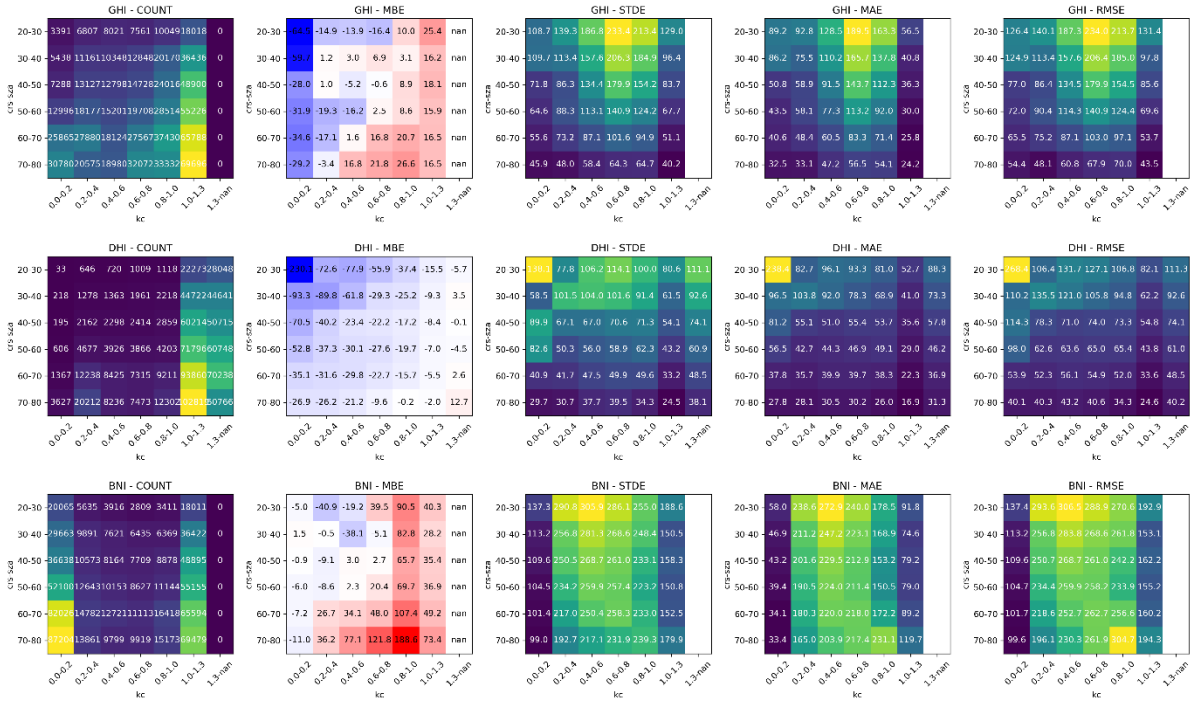

Figure 14. Inspection run of LUT on SZA and Kc for the standard error metrics of the CRS irradiance estimates. Top: GHI, middle: DHI, bottom: BNI. The columns are from left to right are count inside bin, MBE, STDE, MAE, and RMSE.

We see with this simple inspection run that we have very well populated bins (> 1000 data points) for all cases except low Kc for the DHI component. This is expected as Kc for DHI will tend to grow from a minimum value in clear sky situations to a maximum value on overcasted situation. Therefore, the probability to encounter values of Kc for DHI < 0.2 is quite low. In fact, GHI and DHI will never have a Kc = 0 because of the minimum value imposed by the diffuse radiation. As a consequence, low values of Kc for GHI and DHI must be inspected with care.

Figure 14 shows that the STDE, MAE and RMSE present very similar patterns for the same irradiance component. GHI metrics have a tendency for high errors at mid values of Kc and low values of SZA. For DHI the higher errors seem to be on lower SZA values independent of the Kc. For BNI the error patterns are still more pronounced, with increasing metrics with decreasing SZA and achieving maximal value in mid Kc values.

This simple LUT run is an effective CRS monitoring tool as it will allows to quantify the changes of errors patterns that any modification/updates on the base CRS models or its inputs bring on the CRS output estimates. We will be able to easily monitor quality improvement or deterioration of the irradiance estimates and detect directly which instances are the most affect by the deviations.

4.3.1 LUT bins distribution inspection

In the previous section we found the error metrics for each bin of the LUT but we have no information yet about the form of the error distribution for each bin. To get this information, we created a histogram of the irradiance errors for each bin in the trained LUT. These distributions are shown for the GHI component in Figure 15. This figure presents the error distributions for each LUT bin, based on 20 irradiance value intervals, represented by a continuous green line.

The theoretical normal distribution obtained with the MBE (μ) and STDE (α) of the respective bin is also shown on the dashed blue line. The values of μ and α used for the normal distribution on each bin are the ones calculated in the previously and shown in Figure 14. Both distributions are normalized for ease of comparison.

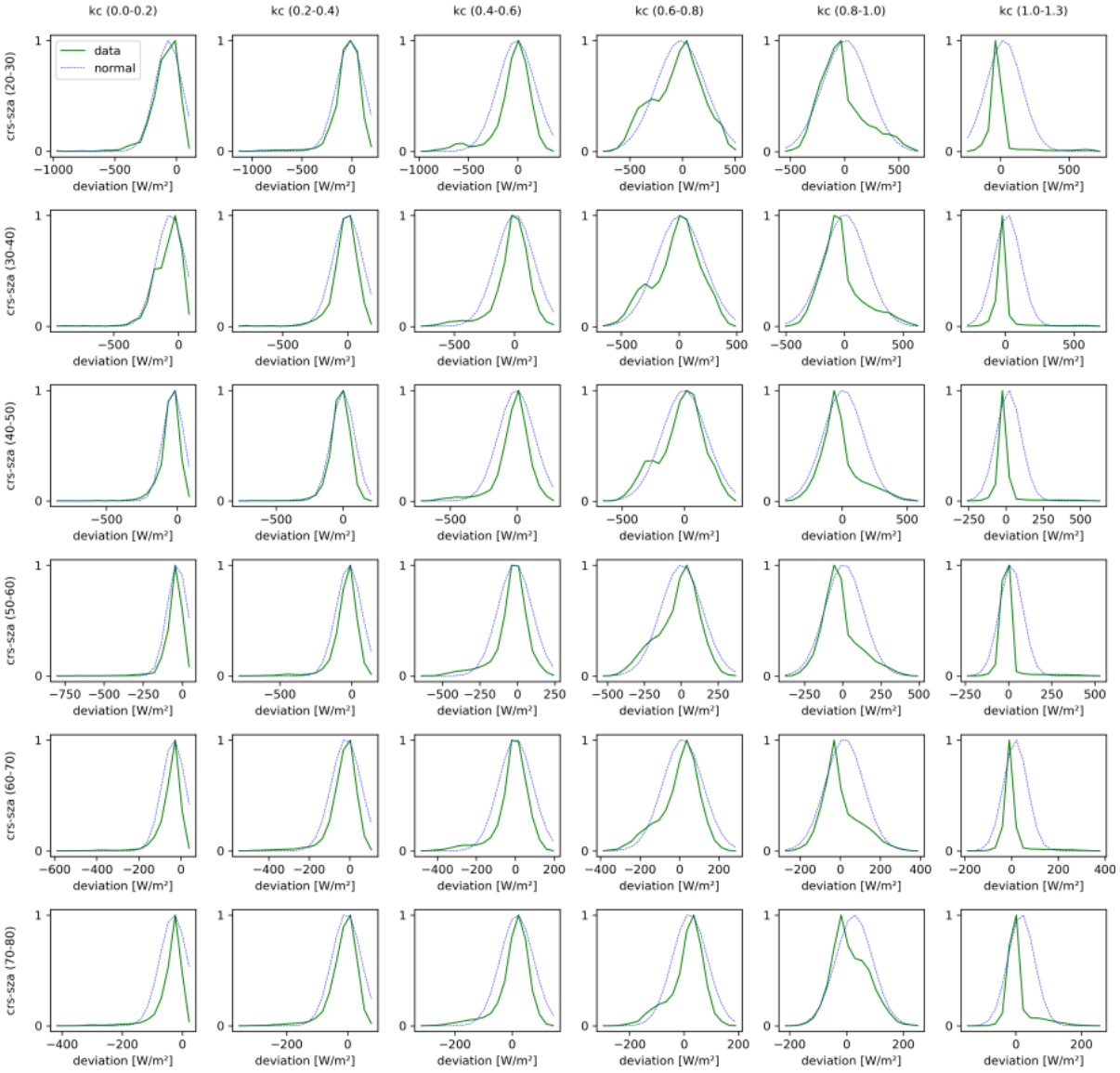

Figure 15. Normalized GHI error distribution per bin of the LUT calculated on the inspection run. For each bin the green continuous line shows the distributions found with 20 ranges on irradiance values and the dashed blue line shows the theoretical normal distribution obtained with the MBE (μ) and STDE (α) of the respective bin (shown in Figure 14). Columns: Kc ranges; Rows: SZA ranges.

Figure 15 shows that the GHI error distributions follow loosely the form of its corresponding normal distribution. First of all, the form of the distribution seems to be independent of SZA. In most of the cases the error distribution is contained inside the normal distribution. This means that if used, the normal distribution will in general model a greater variability of the errors than the one that is actually expected. Moreover, for some Kc ranges, the actual distribution does not follow a normal distribution (e.g, 0.8 to 1 and 1 to 1.3). This is due to the fact that the GHI variable is physically bounded by an upper limit (clear sky irradiance + cloud enhancements) which results by definition on a non-symmetrical distribution.

The behaviour of the error distributions for the DHI component are quite similar to those found on the GHI component. However, the distributions on the BNI component show quite a different behaviour. The distributions found to the BNI component are shown in Figure 16.

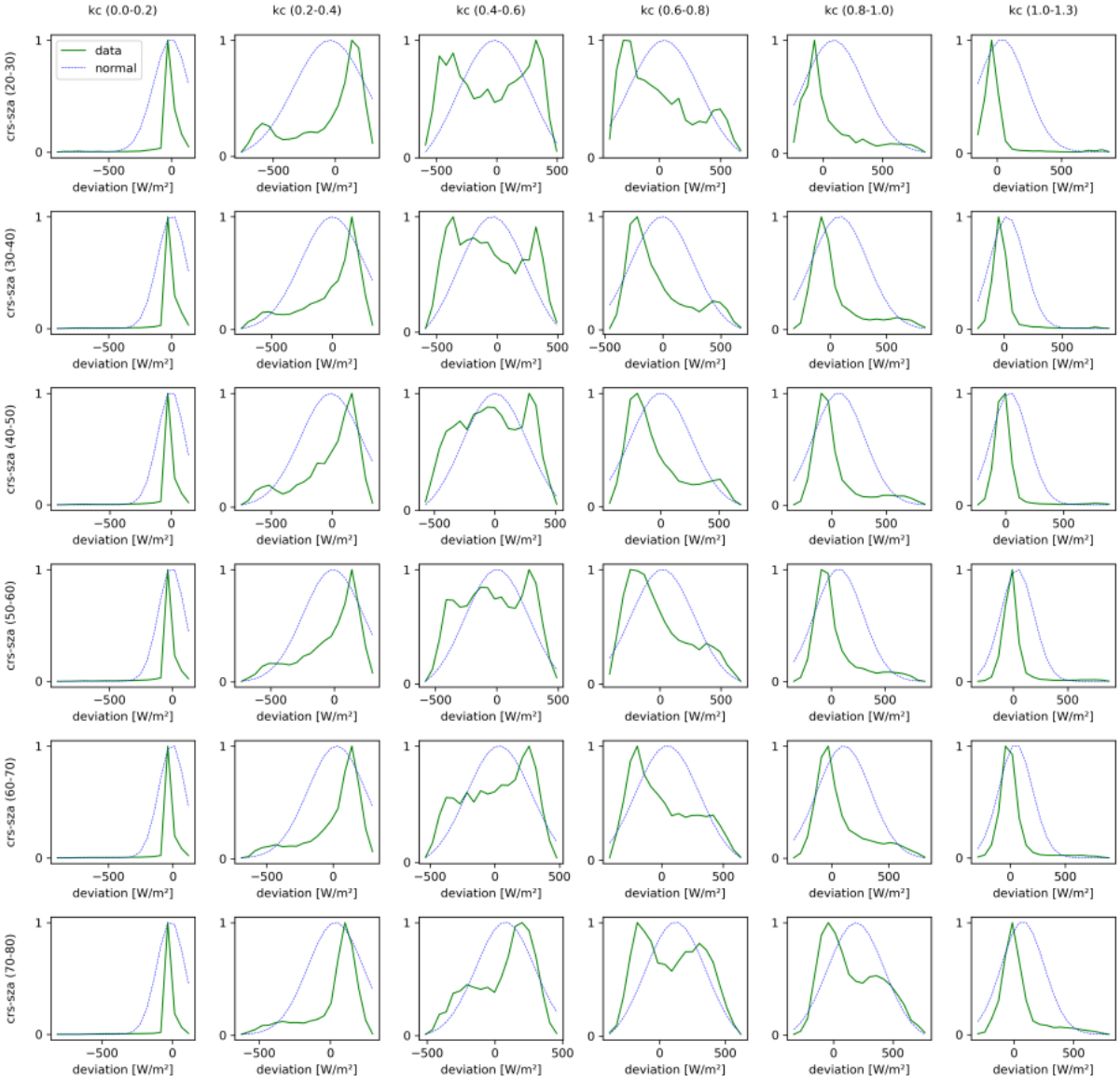

Figure 16. Normalized BNI error distribution per bin of the LUT calculated on the inspection run (same representation as the one shown in Figure 15)

For the BNI component none of the bins seems to follow a normal distribution. The asymmetry found on the lower and higher Kc ranges are also explained by the physical lower and upper limits of the BNI irradiance values respectively (limit of 0 for the lower range and clear sky irradiance for the higher range). Furthermore, the detection of thin clouds with COD below 5 (as relevant for cases with BNI above zero) as well as of small-scale clouds in sub-pixel spatial extension is restricted. It is known, that many cases are typically found in scatterplots on one of the two axes. Therefore, also in all other ranges, the expected deviation distributions are highly skewed or bi-modal. On the bi-modal distributions none of the modes appears at a deviation of 0 W/m². Therefore, the use of metrics as bias and RMSE inside each bin in the error model is restricted in such a simple LUT.

4.3.2 Location based distribution inspection

The error distributions shown until now are for the inspection run LUT that was trained using only the data from BSRN-BUD and BSRN-ABS stations, which are known to be stations that are well modelled by the CRS estimates. In a next step we want to understand the effect that the physiography of the location has on the distribution, since CRS has shown difficulties to model locations over different terrains, i.e. small island, high altitude, high latitude, etc. To do so, we first choose 2 groups of ground observation locations to analyse:

- **Group 1**: **BSRN-CAB, BSRN-PAY and BSRN-PAL.** These locations are known to be very well represented by the CRS and have not been used before.
- **Group 2: BSRN-RUN, BSRN-IZA, BSRN-ENA.** These are locations known to be difficult to model by the CRS: The **BSRN-RUN** station is located in the Reunion Island very near the coast (at around 2 km from the ocean). The **BSRN-IZA** is located on a mountain top at an altitude of 2372 m. The **BSRN-ENA** is located in a very small island in the middle of the Pacific Ocean (at around 600 m from the ocean).

To be able to compare the effect of the location, we will need train one LUT per station studied. This will allow us to obtain the error distributions of the bins per station. We proceed then by training an independent LUT for each of the stations in group 1 and group 2 (6 new LUT obtained). The parameters used on the training of the LUT are the same ones used on the inspection run described on section 4.3. The distributions found for the GHI component of the 3 stations of group 1 are shown in Figure 17.

This figure shows that the distributions for the 3 well modelled stations are very similar on all of the bins in the domain. Even the bimodal distributions found on the (0.6 - 0.8) Kc range describe the same peak positions for the 3 stations. From these results it is clear that the CRS irradiance estimates reproduce the same type of modelling errors for these 3 sites, independent from the sky situation or the sun position used.

On the other hand, the reproducibility of the same type of the modelling errors cannot be found on the distributions for the GHI component of the stations of group 2, as shown in Figure 18. These distributions present high discrepancies with respect to the distributions found in group 1, which is here represented by the distribution of the station BSRN-CAB. The discrepancies are evident on the (0.4 – 0.6) and (0.6 – 0.8) Kc ranges where the some of the group 2 stations present a bi-modal distribution while the group 1 stations do not. The same tendencies are found for the differences between group 1 and 2 in the DHI and BNI components (not shown here).

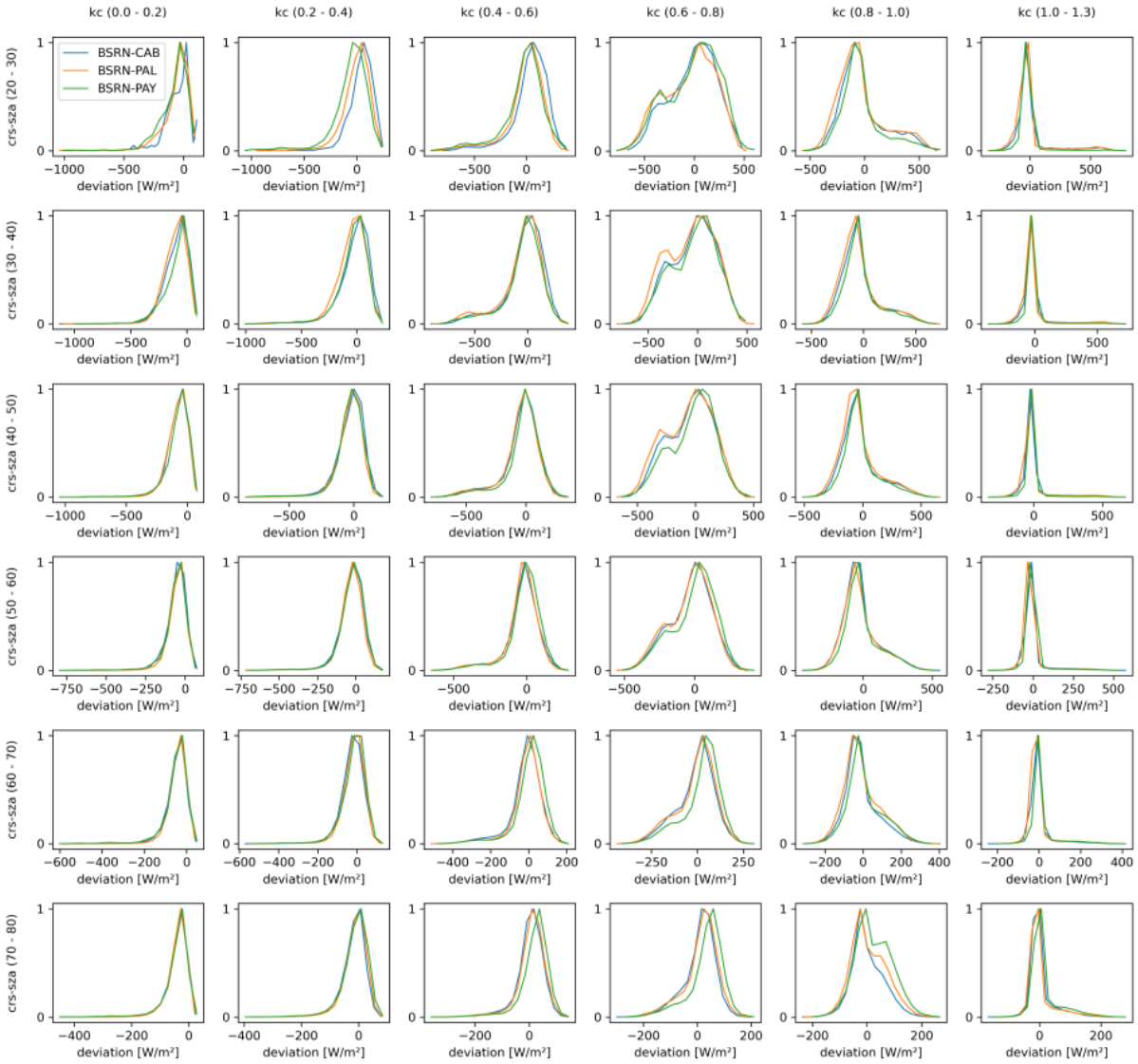

Figure 17. Normalized GHI error distributions for each of the stations in group 1 (BSRN-CAB, BSRN-PAY and BSRN-PAL). Columns: Kc ranges; Rows: SZA ranges.

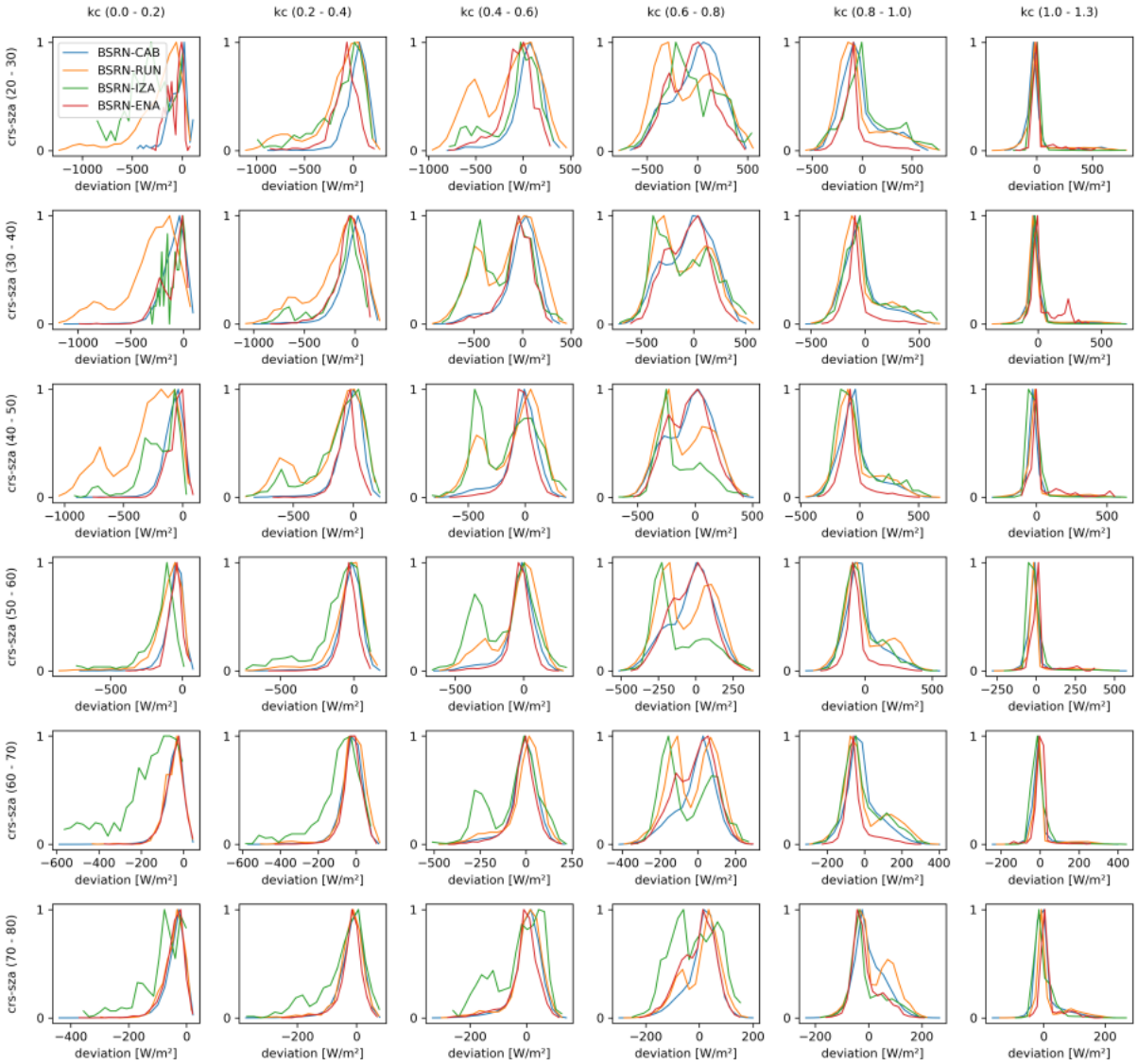

Figure 18. Normalized GHI deviation distributions for each of the stations in group 2 (BSRN-RUN, BSRN-IZA and BSRN-ENA) and one station from group 1 (BSRN-CAB). Columns: Kc ranges; Rows: SZA ranges.

We confirm from the previous results that:

- there seems to be no unique distribution that satisfies all the deviation patterns found on the CRS irradiance estimates
- the location (e.g, local terrain, orography) of the CRS estimate has a big impact on the irradiance estimate uncertainty

Following the reasoning of these findings, we have decided not to use a typical normal distribution function to model the deviations found on the operational CRS irradiance estimates. Instead, we add distribution independent parameters, i.e., percentiles, in order to achieve additionally a probabilistic-based inference of the CRS errors.

4.4 Baseline run

4.4.1 Model training

In order to select the parameters to train the baseline error model, we use the results already obtained from the SHAP (SHapley Additive exPlanations) analysis of the CRS deviations

which was reported on D4.3. From this analysis, a ranking of the CRS input parameters with respect to their contribution to the overall irradiance error (SHAP values) was found. This ranking is shown on the SHAP Beeswarm diagram on Figure 19.

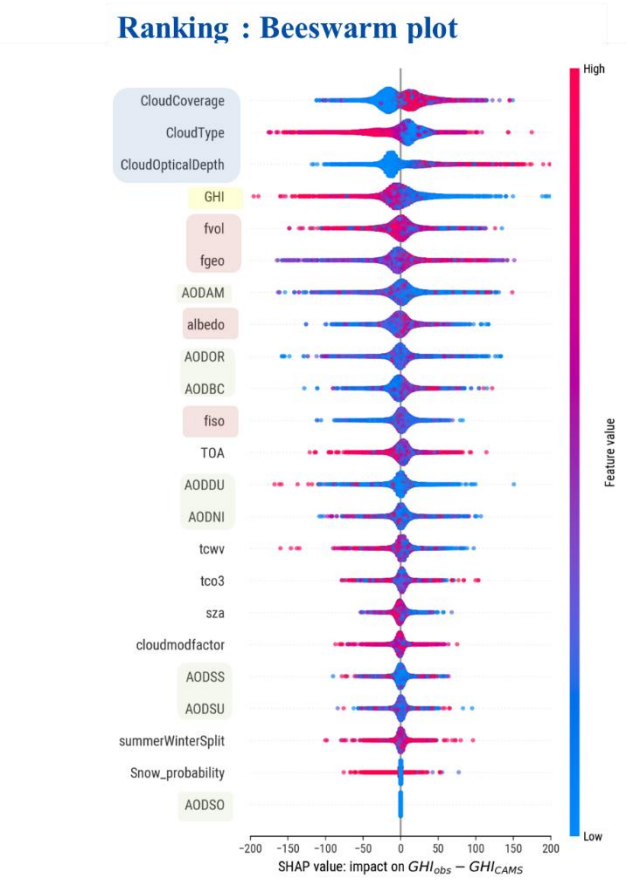

Figure 19. Beeswarm diagram for the GHI deviation obtained from the SHAP analysis on the deliverable 4.3. Cloud related parameters on blue, albedo related parameters on red, aerosol related parameters on beige and irradiance level on yellow.

This diagram provides a comprehensive view of the SHAP values found for each CRS input parameter, ranking them from high to low contribution to the overall deviation. The parameters found to have the highest contribution on biases are the cloud related parameters, followed by the irradiance magnitude and the albedo related parameters.

To avoid the complication of interpretability the use of correlated inputs induces and to avoid the scaling problem due to high number datapoints/dimensions used in the training phase, we have decided to choose firstly the 3 most relevant parameters for an initial run of the training of the error model.

It was previously found that the input parameter that shows the highest contribution to the deviations is the cloud coverage. This parameter corresponds to the cloud probability (value between 0 and 100 %) calculated for the respective pixel on the satellite image using the APOLLO_NG algorithm. Due to the high correlation between the 3 cloud parameters, the cloud type and cloud optical depth are not used on this initial implementation of the error model. The second most ranked group in the SHAP analysis is the irradiance magnitude itself. Because of this, the second parameter chosen for the training of the error model is the Kc, which corresponds to the ratio between the irradiance estimate modeled by the CRS and the irradiance expected at the surface on the cloudless situation (clear sky). Finally, we decided to use as the third parameter the SZA as it accounts for the sun position on our error model.

In future model trainings this list can be extended, but for the first assessment, we keep a small parameter space.

The bins chosen for the initial run of the error model are the ones obtained with the following spacing:

```
cloud_coverage = {from 0% to 100% every 5%} -> (20 ranges)
kc = {from 0 to 1.4 every 0.1} -> (14 ranges)
SZA = {from 0° to 80° every 2°} -> (40 ranges)
```

As it was suggested in section 4.3.2 to use distribution-independent parameters to model the CRS deviation, the error model LUT is now also populated with the 100 percentiles (P0 to P100 every 1 percentile) of the irradiance estimate errors. For comparison purposes, the error model LUT will be also populated with the normal distribution moments MBE, MAE, SDTE and RMSE.

The initial training of the error model follows then the methodology explained in section 4.1. The training is performed using

```
N = 40 (number of stations used on the training)
X = 60% (percentage of time span used for training)
Y = 30% (percentage of time span used for validation)
```

These 40 stations were chosen randomly during the training phase. A random seed is selected and saved for reproducibility of the results. The training of the model results on a 4-dimensional LUT of shape (20, 14, 40, 105) in which the first 3 dimensions correspond to ranges of the 3 input parameters described above (i.e., cloud_coverage, kc and SZA respectively) and the 4th dimension corresponds to the different metrics/moments calculated, which in our case are:

```
metrics = {from P0 to P100 every 1 percentile} + {MBE, MAE, STDE,
          RMSE} -> (105 metrics)
```

The datapoint-wise CRS irradiance uncertainty is derived for the other 26 stations in the database and the other 30% of the time instances in order to respect the spatio-temporal separation. For each CRS instance in the validation dataset an uncertainty estimation is inferred using the CRS model inputs (i.e., cloud_coverage, kc and SZA) as inputs of the trained error model. Any uncertainty interval level from the 100 percentiles can be chosen as an uncertainty estimation of the respective instance.

4.4.2 Assessment of the quality of the uncertainty estimations in different sky situations

To begin the assessment of the error estimates obtained with our model, we plot these estimates as time series per station. Figure 20 shows the GHI uncertainty inference on one example day for the station BSRN-GOB, which is a station that is known to be well represented by the CRS estimates. On the upper plot the ground observations are represented as the solid black line, the CRS estimates as a solid blue line, the 50% confidence intervals by the width between the green areas (P75 – P25) and the 90% confidence intervals by the width between the blue areas (P95 – P5). On the lower plot the normalized CRS parameters used as input of the error model are shown as dashed lines. As expected, when the station is under clear sky conditions (around 3/4^{ths} of the selected day) we find that the 50% uncertainty interval is very narrow and that it contains almost all of the ground observations. This is due to the fact that under clear sky conditions there is a very low irradiance variability, therefore very low uncertainty on the irradiance estimate and the clear-sky model is known to perform very well even at a station in an aerosol-dominated regime as this one. Under cloudy conditions (between 08:00 and 09:00) the width of the 50% uncertainty interval increases considerably.

We also see, as expected, that for these variable conditions most of the ground observations stay inside the 90% confidence interval.

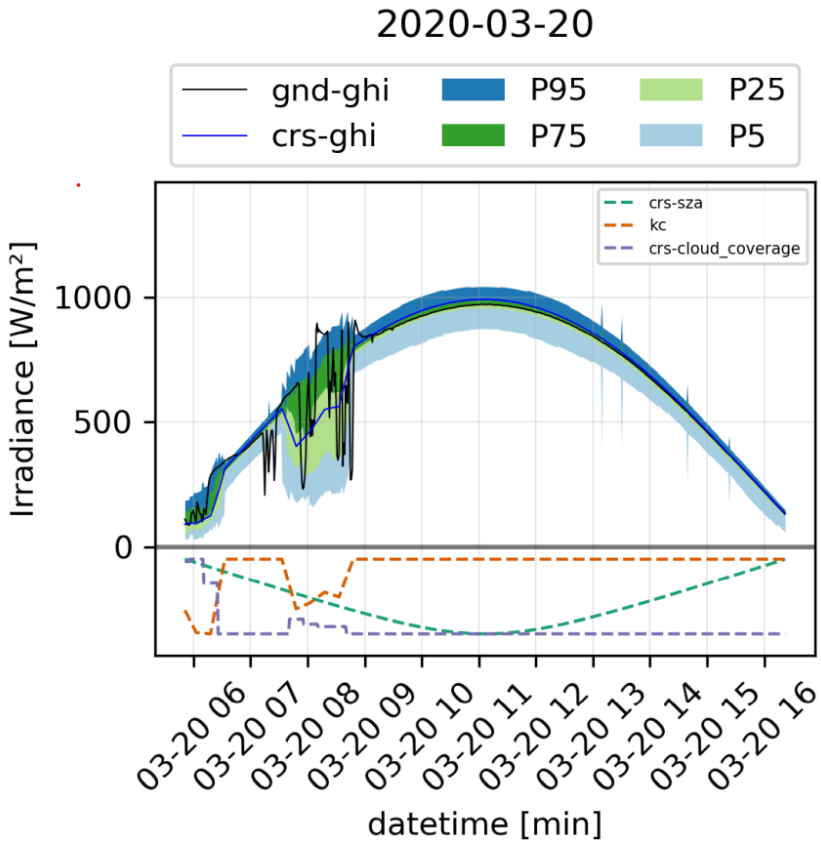

Figure 20. GHI local uncertainty derived for the station BSRN-GOB on the day 2020-03-20. Upper of the graph: ground observation in solid black line, CRS estimate in solid blue line, 50% uncertainty interval covered with the limits between the green areas (P75 – P25), 90% uncertainty interval covered by the limits between the blue areas (P95 – P5). Lower part of the graph: normalized CRS inputs to the LUT, i.e., SZA, KC and cloud coverage

In the same way, Figure 21 shows the GHI uncertainty inference for an example day at the station BSRN-RUN which is known to be a difficult location to model with the CRS estimates. We find for this station generally a very similar behaviour to the one described above for the station BSRN-GOB. But here, we find that the confidence levels are wider that those found for BSRN-GOB. This is due to the fact that the probability to have a higher error for BRSN-RUN location is higher (more difficult location for the CRS model). Here also we see that most of the ground observation values are almost all well contained inside the 90% interval.

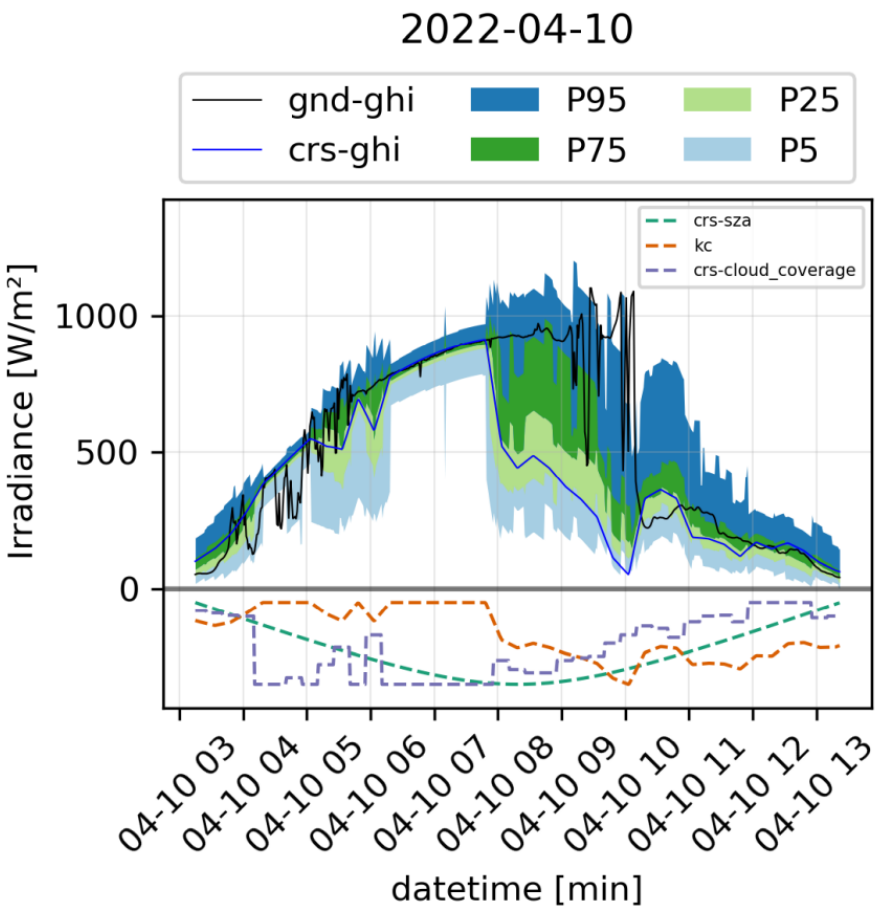

Figure 21. GHI uncertainty inference from the baseline error model for the stations BSRN-RUN on the day 2022-04-10. Same description as for Figure 20.

Finally, Figure 22 shows the GHI uncertainty inference for the station SAURAN-RUN on a very cloudy day. For this difficult day, we see that the 90% confidence interval still follows the variability of the irradiance at all times and that it contains most the observed irradiance estimates.

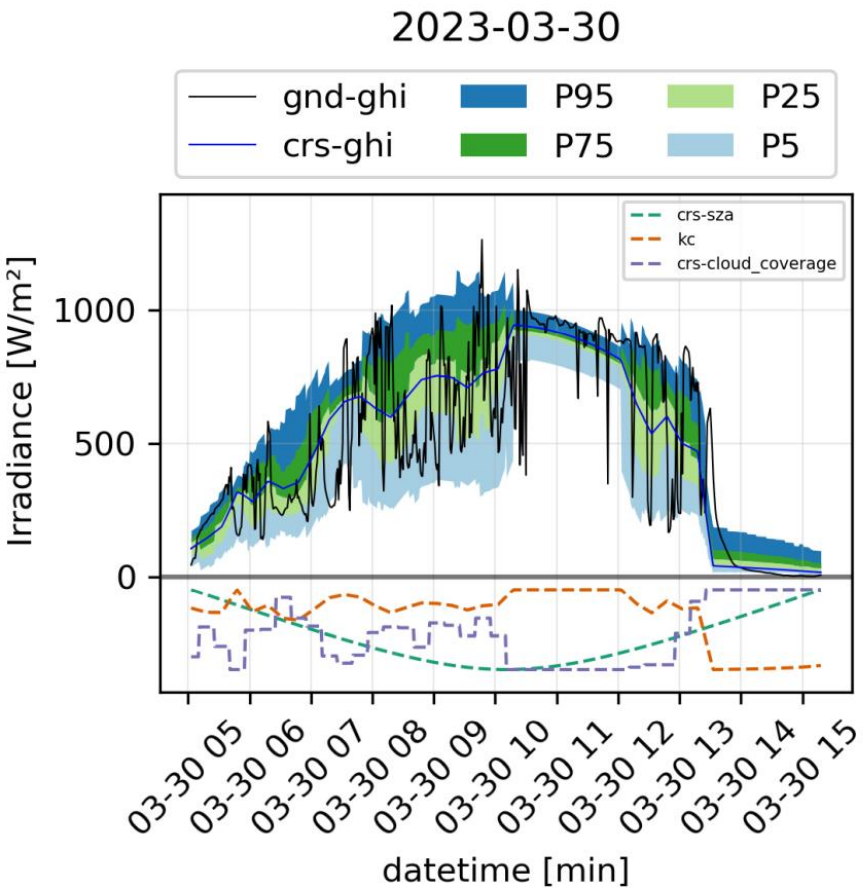

Figure 22. GHI uncertainty inference from the baseline error model for the stations SAURAN-CSIR on the day 2023-03-30. Same description as for Figure 20.

4.4.3 Generalized assessment of the Quality of the uncertainty estimation of the localised error model

In order to generalize the assessment of the quality of the localised error model, we move from the visual inspection of time series to an aggregated quantitative approach.

The evaluation of a deterministic model is typically straightforward: predictions are point estimates, and standard performance metrics such as mean squared error (MSE) or mean absolute error (MAE) are used to assess how close the predictions are to the observed values. Probabilistic models introduce an additional layer of complexity. Rather than outputting deterministic predictions, they provide a probability distribution of possible outcomes which complicates the evaluation process: the model cannot be judged solely by how close its mean prediction is to the target value, but the quality of the entire predictive distribution must be assessed.

Two fundamental aspects characterize the evaluation of probabilistic predictions:

- Reliability (or calibration): A well-calibrated model outputs predictive distributions whose uncertainty matches the observed frequencies. For instance, a 90% prediction interval should contain the true value approximately 90% of the time. Poor calibration indicates that the model is either underconfident or overconfident.

- Sharpness: Sharpness refers to the concentration or narrowness of the predictive distributions. Among different calibrated models, sharper model is preferred, as it reflects more confident predictions.

First, we look into the probabilistic calibration of the developed model using a reliability diagram. This diagram assesses the proportion of CRS irradiance errors that respect the probabilistic hypothesis (also called probabilistic contract) of our model. This diagram shows the relation between the nominal percentile level (p) and the empirical percentile level (ε^p). For a given model irradiance inference i of nominal percentile level p at a time t, i_t^p, there exists a corresponding ground observation reference data point at time t, y_t. We define a variable β_t^p as the unitary function that satisfies

$$\beta_t^p = 1\{y_t < i_t^p\} = \begin{matrix} 1 & \text{if } y_t < i_t^p \\ 0 & \text{otherwise} \end{matrix} \quad (4.1)$$

The empirical level of percentile p, ε^p, is then define as the average of β_t^p for all times

$$\varepsilon^p = \frac{1}{T}\sum_{t=1}^{T} \beta_t^p \quad (4.2)$$

Figure 23 shows the reliability diagram calculated for our error model. On the X-axis we show the nominal percentile levels p for which the model has calculated estimates using the training data and the Y-axis shows the percentage of values that respected the estimated percentile level on the validation data (empirical level ε^p). Consequently, the perfectly calibrated model will have a reliability diagram that lies on the diagonal, i.e., the probability hypothesis calculated on the training data was found in the same exact way on the validation data.

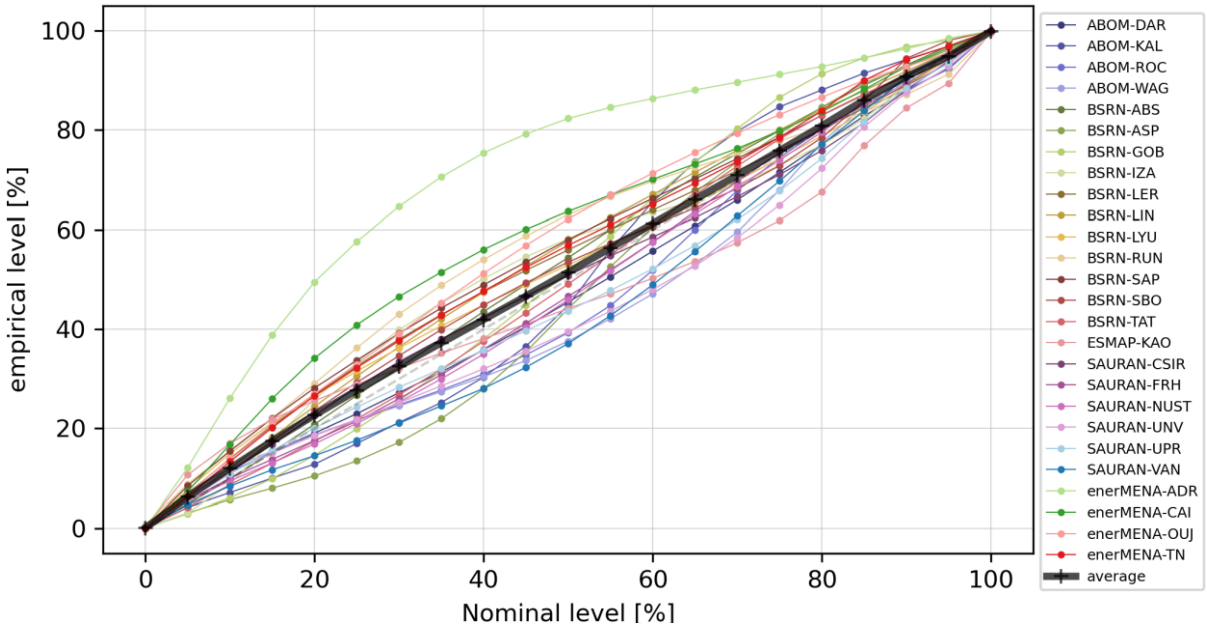

Figure 23. Reliability diagram for the GHI component of the error model. Nominal percentile level vs empirical percentile level (calculated) from the error model. Validation stations are shown with the different colored dotted lines. The average reliability of the model is shown with the bold solid black line.

Figure 23 shows the reliability curve for each validation station with individually colored dotted lines. This figure also shows the overall stations aggregated reliability curve in a bold black line, which is calculated as the mean value of the reliability curves from all locations. We see that in general the model is well calibrated. The worst calibrated station is BSRN-IZA, which is known to be a difficult station to model due to its high altitude (2372 m). All other stations seem to oscillate near the diagonal. This means that the uncertainty levels estimated by the localised error model represent quite well the uncertainty levels found on the CRS irradiance

estimates on the validation data. It is worth to note that the validation dataset used for this assessment is separated in space and time from the training data, which means that the assessment is done on locations and time spans that the model has never seen.

In order to assess the sharpness of the error model, we first estimate the sharpness on every individual validation station. As a measurement of sharpness, we calculate the mean width of the different confidential intervals for all inferences of our error model. Given the width ω of confidence interval c at a time t, ω_t^c defined as

$$\omega_t^c = i_t^{\bar{p}} - i_t^{\bar{\bar{p}}}, \tag{4.3}$$

where the confidence interval c is calculated from the upper percentile level $\bar{p}$ and lower percentile level $\bar{\bar{p}}$ as

$$c = 90 \quad \text{then } \bar{p} = 95\,,\ \bar{\bar{p}} = 5$$
$$c = 80 \quad \text{then } \bar{p} = 90\,,\ \bar{\bar{p}} = 10$$
$$\dots$$

then the width ζ for the confidence interval c , ζ^c, is defined as the mean width of ω_t^c on all t

$$\zeta^c = \frac{1}{T}\sum_{t=1}^{T} \omega_t^c \tag{4.4}$$

Figure 24 shows the sharpness of the error model by means of the average widths of the interval confidence levels c for the validation stations.

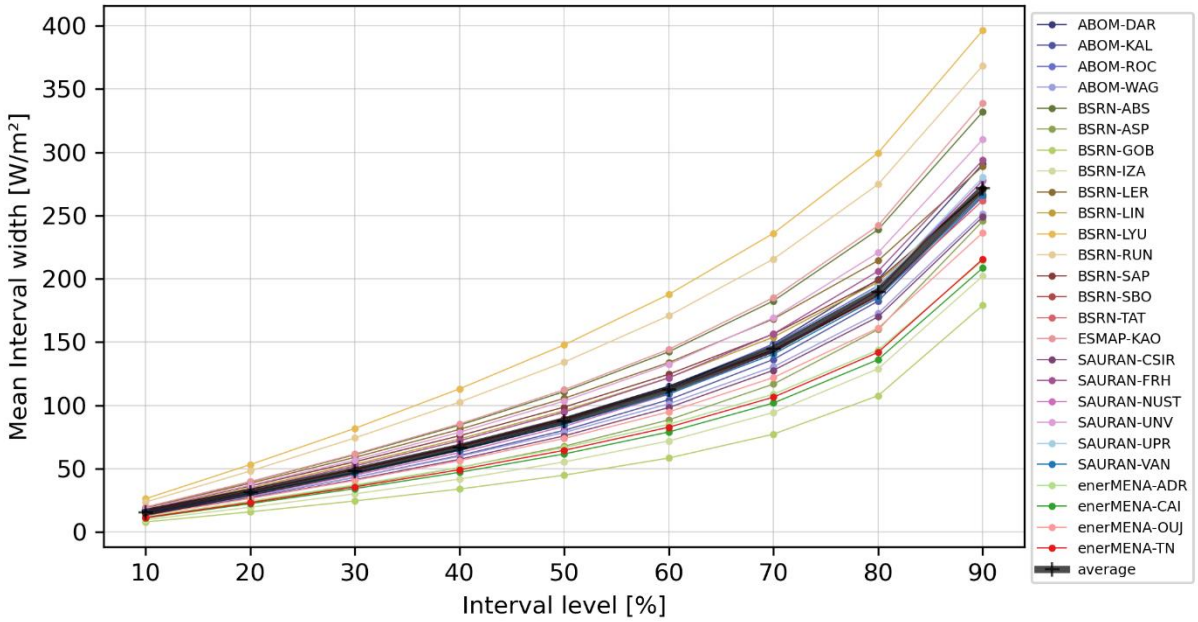

Figure 24. Sharpness diagram for the GHI component of the error model per validation station. Each colored dotted lines represents one station. The average sharpness of the model is shown with the bold solid black line

The widths calculated for each station are shown on colored doted lines. As expected we see that the interval width increases with increasing confidence level. We see a general faster increase of the widths from c =60 to c =90 which must be due to the increase in irradiance estimate errors related to partially cloudy situations (highly variable irradiance changes). It is worth noticing that there is more than a factor 2 between the sharpest station (BSRN-GOB which is a well modelled station) and the less sharp station (BSRN-LYU which is a costal station less than 1km to the ocean), which is a reasonable result. This is a clear indicator that the irradiance uncertainty is location dependent but more important still, that the error model developed here is able to correctly quantify the uncertainty levels depending on the location

characteristic. The overall error model sharpness, shown in the bold solid black line is a characteristic of the model itself an is useful for comparing with other error models or monitor the evolutions on the CRS model itself.

4.5 Inference of the localised error model for deterministic estimate corrections.

After having assessed the reliability of the probabilistic error estimate, the potential of the localised error model to deterministically correct irradiance estimate is assessed. Metrics analyzed are MBE, MAE, and STDE. The study is motivated by the general discussion to either provide a probabilistic error estimate or to use the error model to do any post-processing correction of the CRS output itself.

This correction is tested with 2 methods to assess the need for a probabilistic error description versus a gaussian distribution error description of the error model:

- **P50 correction**: correction performed by subtracting the median (P50) of the bin distribution to the CRS estimate
- **μ correction**: correction performed by subtracting the MBE (μ) of the bin distribution to the CRS estimate

The biases obtained on the 26 validation sites for the 2 correction methods are shown in Figure 25. From the geographical location of the validation sites (top diagram of Figure 25), the highest biases seem to come from the desertic sites (**BSRN-SBO** and **enerMENA-ADR)** or elevated sites (**ESMAP-KAO)**. Small distance to the coast does not seem to be correlated with high bias values on the uncorrected or corrected CRS estimates. Looking at the overall performance of the biases for the selected 26 stations (bottom diagram of Figure 25), the μ-correction tends to increases the general underestimation of the CRS estimates (to higher negative values). In the other hand, the P50-corrections tend to decrease the general bias of the CRS estimates almost to a 0 value on average. Nevertheless, both methods deteriorate metrics of many stations, and both methods have a significant number of stations where they provide decreased accuracy metrics. Only on average the effects level out.

The MAE obtained on the 26 validation sites for the 2 correction methods are shown in Figure 26. The tendencies for MAE are quite different to those obtained for the MBE. From the geographical point of view (top diagram of Figure 26) we see that the distance to the coast seems to be highly correlated with the MAE of the site, i.e., sites near the coast present higher deviations. From bottom diagram of Figure 26, we see that both corrections tend to improve the MAE metrics. The uncorrected estimates show a MAE value of 73 W/m² while the μ-correction a value of 71.4 W/m² and p50-correction a value of 69 W/m².

Overall, both corrections improve the MAE at most stations. The absolute MAE improvements are low (1.6 W/m² for the μ-correction and 4W/m² for P50 correction). Nevertheless, improvement achieved by the p50-correction doubles the improvement obtained by the μ-correction

Finally, the STDE found on the 26 validation sites are shown in Figure 27. The STDE metric presents the same geographical tendencies as the one described for the MAE. However, the STDE improvements achieved by the μ-correction almost systematically outperforms the improvements achieved by the p50-correction.

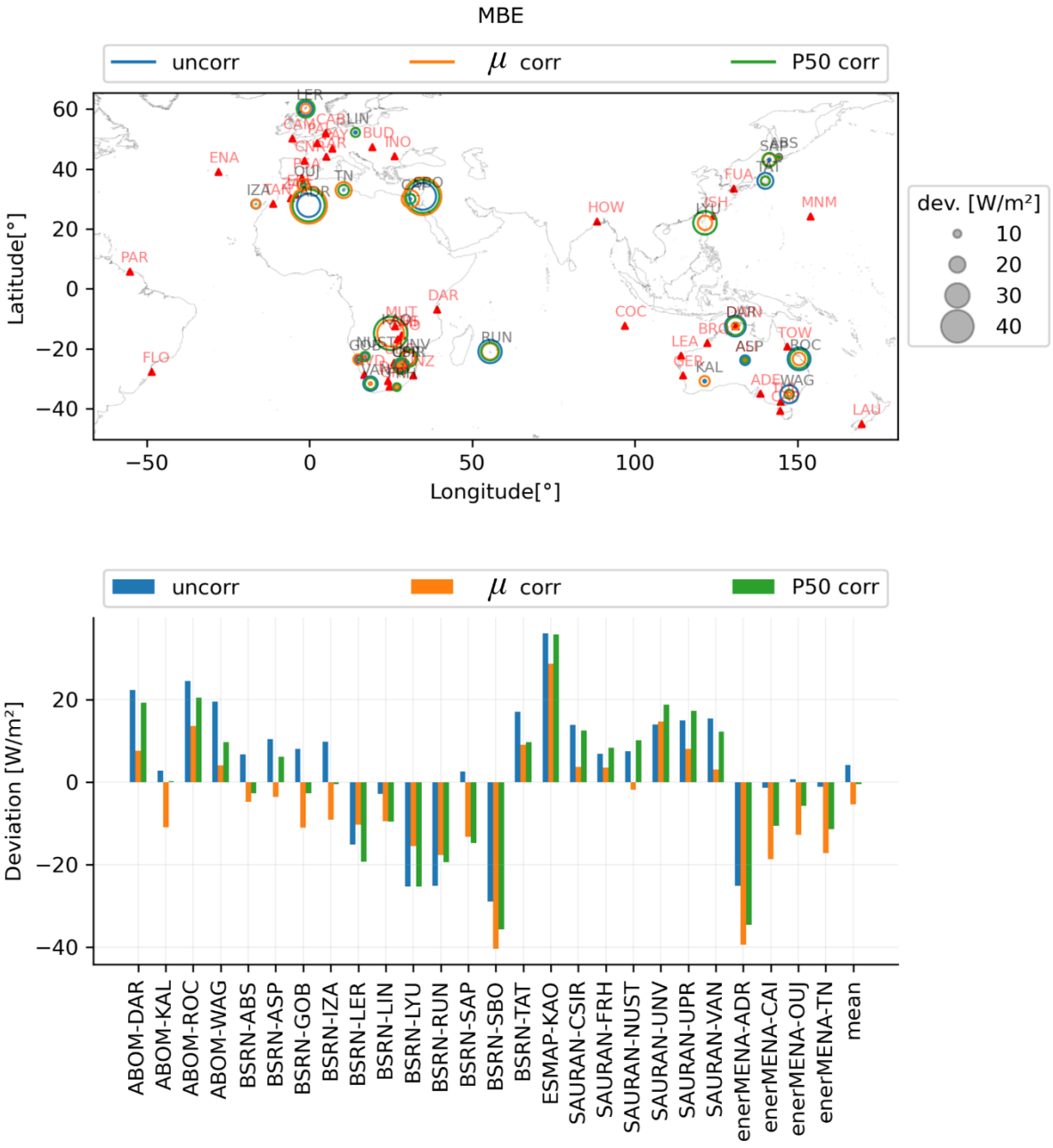

Figure 25. Bias of the deterministic GHI corrections for the 26 validation sites. Blue color represents the uncorrected GHI estimate, the orange color represent the μ-corrected GHI estimated (using the MBE of the bin) and the green color represents the P50-corrected GHI estimate (using the median of the bin). Top figure: red triangles represent the 40 sites used for training and the colored circles represents the 26 sites used for the GHI estimate inference, where the size of the circle is proportional to the deviation found for each correction. Bottom figure: values of the uncorrected, μ-corrected and P50-corrected irradiance estimates for the 26 validation sites. The last bar group ("mean") correspond to the average value for the 26 sites.

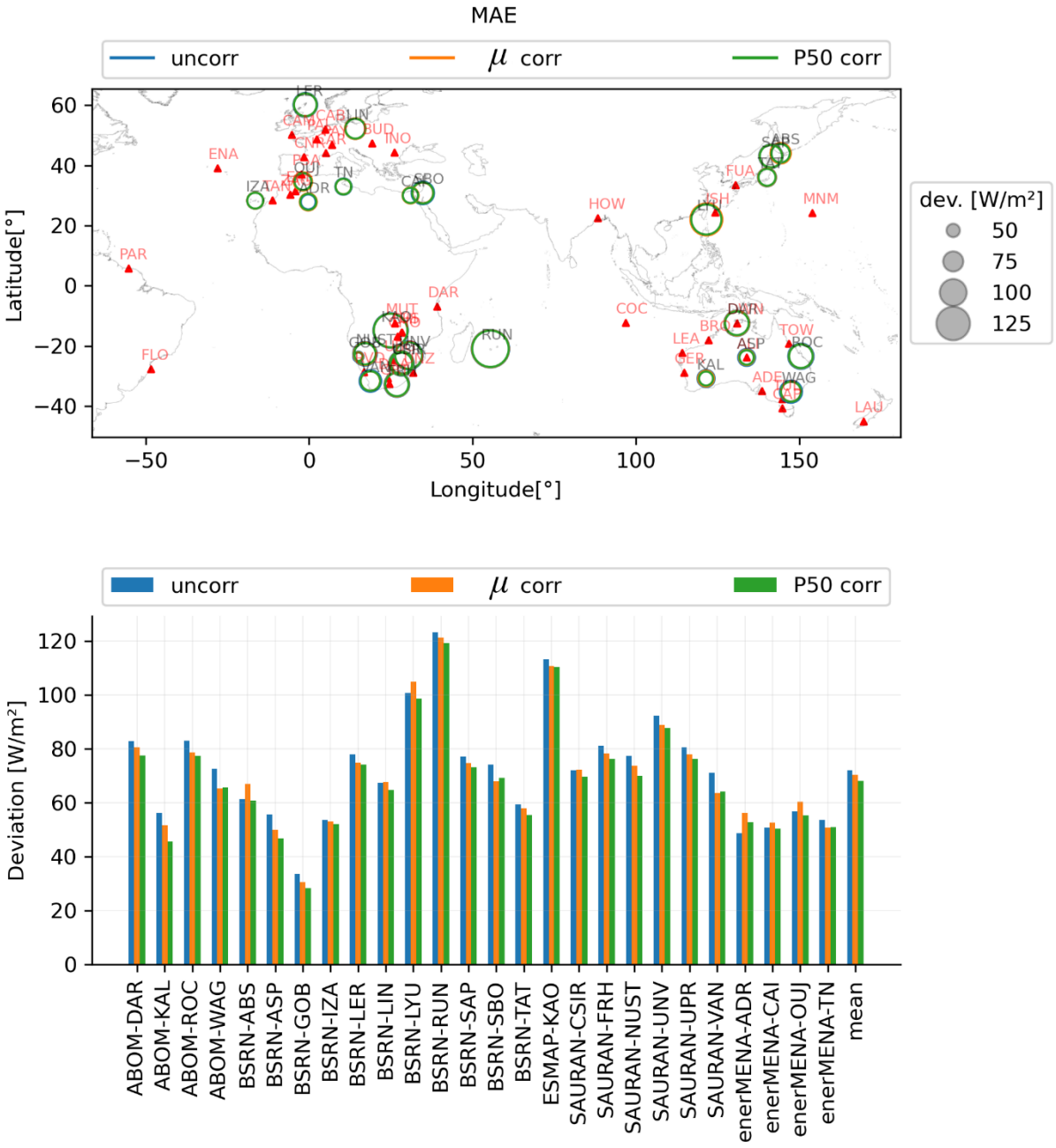

Figure 26. MAE of the deterministic GHI corrections for the 26 validation sites (same description as in Figure 25).

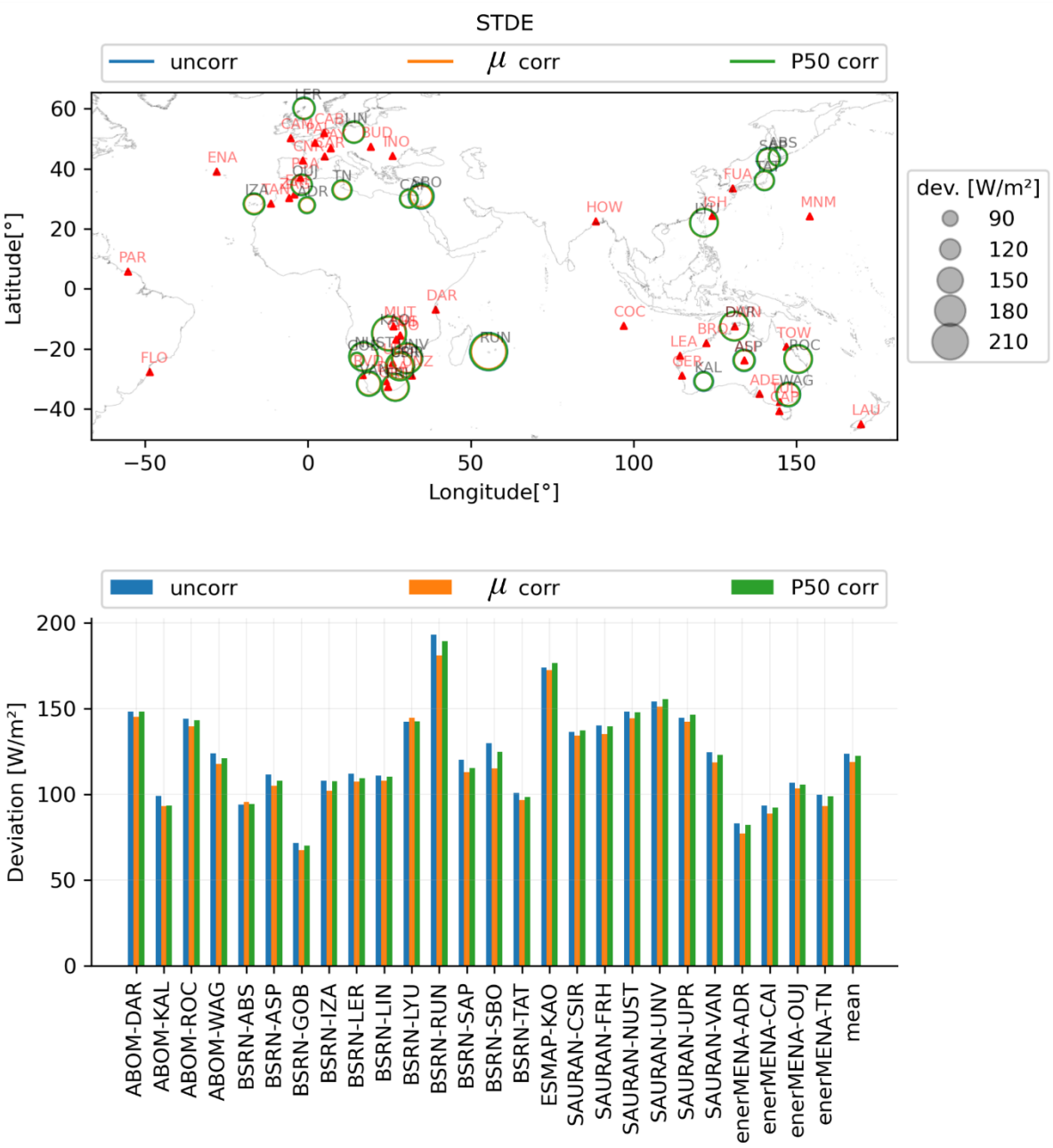

Figure 27. STDE of the deterministic GHI corrections for the 26 validation sites (same description as in Figure 25).

The results from Figure 25, Figure 26 and Figure 27 show that there is are improvements on the statistical error metrics when applying corrections based on the localised error model developed in this study. In particular, the best improvement was obtained on the MBE when using the localised P50-corrections. This method should be analysed further by the CAMS CRS development team to assess the potential of an online bias correction. Nevertheless, the overall impact is small and the value of the probabilistic distribution of the possible irradiation values may be rated as more relevant to users.

5 Localised error model 2: uncertainty inference based on Deep Learning

As an alternative to the parametric binning approach shown in section 4, a deep learning approach has also been studied to model the uncertainty distribution of the CRS. The objective of this second approach is to statistically learn the error distribution of the CRS from a reference data set using a supervised statistical model based on Machine Learning. The inherent advantage of such a model is that it is computationally efficient and can be very accurate. Its disadvantages are that the relationship with the physical modelling is not explicit, and it requires a large amount of reliable reference data. The fact that his method can be computational efficient make it a good candidate for a future implementation in the operational CRS service.

5.1 Methodology

Various methodologies can be used to estimate the uncertainty of the results of a model as a function of different related variables. Probabilistic programming models, such as Bayesian neural networks or variational inference, are based on a Bayesian formulation of the problem that allows a-priori uncertainty to be considered. They also allow to assess the aleatoric and epistemic uncertainties. However, it is difficult to implement such a model in an operational context, as sampling is required to assess the a-posteriori distribution.

Another approach is to predict the different quantiles of uncertainty using quantile regression. A neural network can be used where the pinball loss is used as cost function. Quantile regression does not require any assumptions about the shape of the modelled distribution. The quantiles of interest must be defined a priori. We decided to choose a more generic approach in which the parameters of a parametric distribution are predicted as a function of a set of related variables using a neural network. In this work, we selected the Johnson SU distribution over more commonly used alternatives (e.g., Gaussian, log-normal, or beta distributions) due to its ability to independently control the first four moments—mean, variance, skewness, and kurtosis.

The Johnson SU distribution is a flexible four-parameter probability distributions that can model a wide range of shapes, including skewed and heavy-tailed behaviours. It is parameterized by two shape parameters, along with location and scale, enabling it to approximate many real-world distributions with varying asymmetry and tail behaviour. Its probability density function is given by:

$$(x|\eta,\gamma,\lambda,\varepsilon) = \frac{\delta}{\lambda\sqrt{2\pi}} \frac{\eta}{\sqrt{1+\left(\frac{x-\varepsilon}{\lambda}\right)^2}} \, exp\left(-\frac{1}{2}\left(\gamma + \delta \, asinh\left(\frac{x-\varepsilon}{\lambda}\right)\right)^2\right) \quad (5.1)$$

Where $\gamma \in \mathbb{R}$ and $\delta > 0$ are shape parameters controlling skewness and kurtosis, while $\varepsilon \in \mathbb{R}$ and λ>0 are location and scale parameters, respectively. It can be noted that the transformation $z = \gamma + \delta \, asinh\left(\frac{x-\varepsilon}{\lambda}\right)$ maps the variable x to a standard normal variable $z \sim \mathcal{N}(0,1)$.

An illustration of the flexibility of the Johnson SU distribution is given in Figure 28, where different parameters show the ability of this family of distribution to reproduce different skewness and kurtosis.

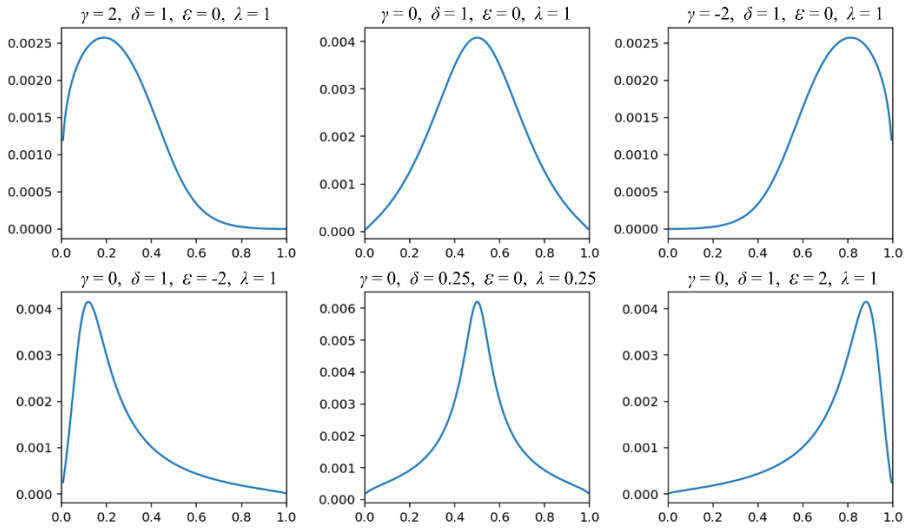

Figure 28. Form of the Johnson SU distribution for different set of parameters illustrating the ability of this distribution to reproduce different skewness and kurtosis.

The four parameters of the distribution are predicted for each time step using a neural network as a function of 6 key variables for the calculation of the solar radiation in CRS. The selected predictors are:

- The clear sky index calculated with CRS estimates (Kc)
- The cloud type
- The cloud cover
- The cloud optical depth
- SZA and Solar Azimuth Angle (SAA)

As illustrated in Figure 29, a fully connected neural network has been used with two hidden layers containing each 64 neurons. The neural network has been coupled with the Johnson SU distribution using the TensorFlow Probability (TFP) library.

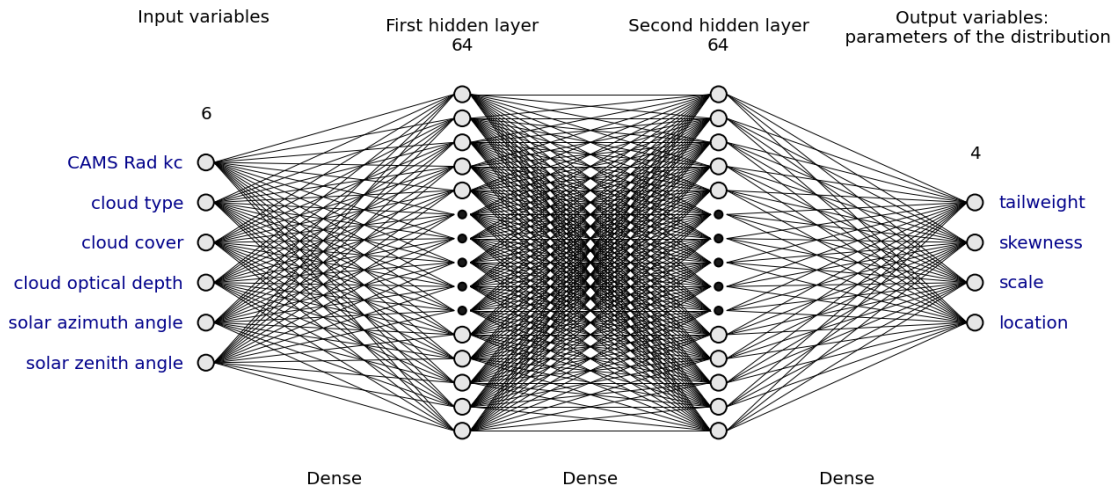

Figure 29. Schematic representation of the neural network used to predict the different parameters of the Johnson SU distribution as a function of the six selected parameters.

The neural network coupled to the Johnson SU distribution was trained using the negative log likelihood as the cost function. The maximum number of epochs was set at 200 with early

stopping based on monitoring the evolution of the Negative Log Likelihood (NLL) on a validation dataset. To avoid the effect of random initialization of the weights of the neural network on the result, the training was repeated 100 times and the run giving the lowest negative log probability on the validation dataset was selected.

The approach described was first tested on a limited number of stations, as the present work is a first proof of concept of the method. The list of stations used are given in Table 2.

Table 2 : Metadata for the BSRN ground stations used for the training and evaluation of the probabilistic model

Station name	Latitude (°)	Longitude (°)	Elevation (m)	Period used
BSRN-CAB	51.9680	4.9280	0	01/2010 – 12/2018
BSRN-CAR	44.0830	5.0590	100	01/2010 - 12/2018
BSRN-CAM	50.2167	-5.3167	88	08/2010 - 07/2017
BSRN-PAL	48.7130	2.2080	156	01/2016 - 12/2018

An important aspect of implementing a supervised model is the separation between the training/validation data in the reference dataset. As shown in Table 3, we used measurements from three stations over the period 2010-2015 for training and validation. For each station, 80% and 20% of the data were used for training and validation respectively. We considered two subsets of test data. The first subset includes the same stations used for training and validation, but separated in time (2016-2018). The second subset includes measurements from locations and time periods different from those used for training and validation. These two subsets of test data are referred to as test-T (separated in time) and test-ST (separated in time and space) respectively. They will be used to assess the model's ability to generalise over time and space. This separation is summarized in Table 3.

Table 3 : Spatio-temporal separation for the preparation of the training, validation and test phases.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Temporal split</td></tr>
<tr><td>2010-01-01 – 2015-12-31</td><td>2016-01-01 – 2018-12-31</td></tr>
<tr><td rowspan="3">Spatial split</td><td>CAB
CAR
CAM</td><td>Training dataset (80%)

Validation dataset (20%)</td><td rowspan="2">Temporal (T) test dataset</td></tr>
<tr><td>PAL</td><td>Discarded</td><td>Spatio-temporal (ST) test dataset</td></tr>
</table>

5.2 Evaluation of the deep learning-based error model

5.2.1 Quality indicators used for validation

We have previously seen the reliability and sharpness indicators are well suited for quantifying the performance of a probabilistic model. An extremely narrow distribution may rarely contain the true value and would thus be poorly calibrated. Conversely, a model may be well calibrated, but if its sharpness is too low, it will be less informative and of limited relevance. It is therefore important to consider these two aspects together, as calibration measures the statistical consistency between predictions and observed results, while sharpness measures the informativeness of predictions. Optimally, a probabilistic model should strike a balance between these two aspects: being as accurate as possible while remaining well calibrated.

The probabilistic indicators described below are used in this section to quantify this dual objective.

The prediction interval $Q_\alpha(X_i)$ at confidence α, calculated for input X_i, is defined as:

$$Q_\alpha(X_i) = [L_\alpha(X_i); U_\alpha(X_i)] \quad (5.2)$$

Where $L_\alpha(X_i)$ and $U_\alpha(X_i)$ are respectively the quantiles at levels $\alpha/2$ and $1-\alpha/2$ predicted by the probabilistic model for the input X_i.

The interval $Q_\alpha(X_i)$ represents the prediction interval within which the actual value of y_i is expected to lie with probability α, assuming the model is well calibrated.

The Prediction Interval Coverage Probability (PICP) evaluates the model reliability or calibration. It can be expressed as:

$$PICP_\alpha = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y_i \in Q_\alpha(X_i)\} \quad (5.3)$$

Where the indicator function $\mathbf{1}\{y_i \in Q_\alpha(X_i)\}$ equals 1 if y_i is included in the interval $Q_\alpha(X_i)$, and 0 otherwise.

If the model is perfectly calibrated, $PICP_\alpha$ should be equal to α.

The Expected Calibration Error (ECE) is a more general metric assessing the discrepancy between predicted and observed coverage of prediction intervals across multiple confidence levels. It evaluates how well the model's predicted quantiles correspond to empirical frequencies, as such it is a perfect mean to assess the model calibration or reliability. For a set of confidence levels $\{\alpha_b\}_{b=1}^{B} \in [0; 1]$, lower quantiles $L_{\alpha_b}(X_i)$ can be predicted by the model such that:

$$Pr\left(y_i \leq L_{\alpha_b}(X_i)\right) = \alpha_b \quad (5.4)$$

Empirically, for each level α_b, the coverage can be assessed as follows:

$$Coverage(\alpha_b) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y_i \leq L_{\alpha_b}(X_i)\} \quad (5.5)$$

Where the indicator function $\mathbf{1}\{y_i \leq L_{\alpha_b}(X_i)\}$ equals 1 if y_i is less than or equal to $L_{\alpha_b}(X_i)$, and 0 otherwise.

The ECE is then defined as the average absolute difference between the empirical coverage $Coverage_{\alpha_b}$ and the confidence levels:

$$ECE = \frac{1}{B}\sum_{b=1}^{B} |Coverage(\alpha_b) - \alpha_b| \quad (5.6)$$

This metric evaluates how well the predicted cumulated distribution function (CDF) aligns with the observed data, and a lower ECE indicates better calibration of the model's predictive distribution. This quantity can be visually interpreted using a calibration plot, which displays the empirical coverage as a function of the nominal confidence levels. In this plot, a perfectly calibrated model corresponds to the identity line (i.e., coverage equals confidence level). ECE then quantifies the average absolute vertical distance between this identity line and the model's coverage curve, summarizing the overall calibration performance.

The Prediction Interval Normalized Averaged Width (PINAW) measures the sharpness of a probabilistic model by quantifying the average normalized width of the confidence intervals $Q_\alpha(X_i)$:

$$PINAW(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\frac{U_\alpha(X_i) - L_\alpha(X_i)}{R} \tag{5.7}$$

Where $R = y_{max} - y_{min}$ is the range between the maximum and minimum values y_i. This normalization ensures that the prediction interval widths are expressed relative to the scale of the target variable, making the PINAW dimensionless and comparable across different datasets or targets.

While PINAW is a widely used measure to assess sharpness, it is inherently dependent on a chosen confidence level. This dependency limits its utility to assess the global sharpness of the model. An alternative approach is to quantify sharpness using the mean predictive standard deviation (MPSD) across forecasted distributions which is simply the average value of the distribution P_i predicted at each time step i:

$$MPSD = \frac{1}{n}\sum_{i=1}^{n}\sigma(P_i) \tag{5.8}$$

This metric captures the average spread of the predictive distributions over the dataset. A model with lower average standard deviation produces more concentrated (i.e., sharper) predictions. This measure does not rely on any specific prediction interval. Although the mean predictive variance would be a more mathematically rigorous choice, we adopt the standard deviation here for interpretability: the standard deviation has the same units as the predicted quantity, making the results more accessible and easier to interpret.

Finally, the Continuous Ranked Probability Score (CRPS) is a proper scoring rule that jointly assesses calibration and sharpness in a single metric. It quantifies the difference between a forecasted cumulative distribution function (CDF) and the observed outcome, thus providing a measure of forecast performance that accounts for both calibration and sharpness. For a predictive cumulative distribution function F and an observed value y, the CRPS is defined as:

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{+\infty}\left(F(x) - \mathbf{1}_{\{\boldsymbol{x}\geq\boldsymbol{y}\}}\right)^2 dx \tag{5.9}$$

where $\mathbf{1}_{\{\boldsymbol{x}\geq\boldsymbol{y}\}}$ is the indicator function equal to 1 if $\boldsymbol{x} \geq \boldsymbol{y}$ and 0 otherwise.

The CRPS measures the squared difference between the forecasted probability and the actual outcome across all possible threshold values. A lower CRPS indicates a better probabilistic forecast, as it corresponds to a CDF closer to the step function representing the observed value.

While global metrics described above offer high-level insights, the evaluation of a probabilistic model requires conditional analyses. These allow to detect overconfidence, underconfidence, and to understand regions where the model work well or requires improvement. A conditional evaluation as a function of the clearsky index provided by the CRS will be conducted.

5.2.2 Validation of the inference of the Machine learning based error model

Global metrics of the probabilistic model are given in Table 4 for the training, validation and the two test datasets. We can observe that the overall performances are similar for the training, validation and T-test dataset. Lower performances are observed for the ST-test dataset. These

global metrics show that the model generalizes well temporally but less spatially. These results will be further analysed in complementary analyses presented below.

Table 4 : Global metrics of the model for the training, validation and two test datasets

	Number of values	Average kc	NLL	CRPS	Coverage: ECE	Sharpness: $E(\sigma_{pred}(t))$
Training dataset	213 858	0.68	-0.7276	0.0423	0.0224	0.1601
Validation dataset	56 983	0.67	-0.7326	0.0422	0.0050	0.16424
T test dataset	120 343	0.68	-0.7405	0.0425	0.0045	0.1596
ST test dataset	47 450	0.66	-0.5392	0.0469	0.0602	0.1653

Some examples of the model output are given in Figure 30 and Figure 31. We can see in Figure 30 that the spread of the probabilistic model is lower in the absence of cloud than in broken clouds and overcast situations. In all cases, the spread seems to reflect the uncertainty of CRS, especially in broken cloud situation. We can also note that the median of predicted values by the probabilistic model is closer to the measurements than the estimates from CRS in cloud-free and overcast situation.

The cases illustrated in Figure 31 are also cloud-free, broken cloud and overcast situations, but where the probabilistic model shows some limitations. For these cases in the cloud-free and overcast situations the median predicted by the probabilistic model is less accurate than CRS estimates. In the variable conditions, some observed irradiance spikes are not included in the range of values predicted by the model, indicating that the model can be underdispersive in some situations.

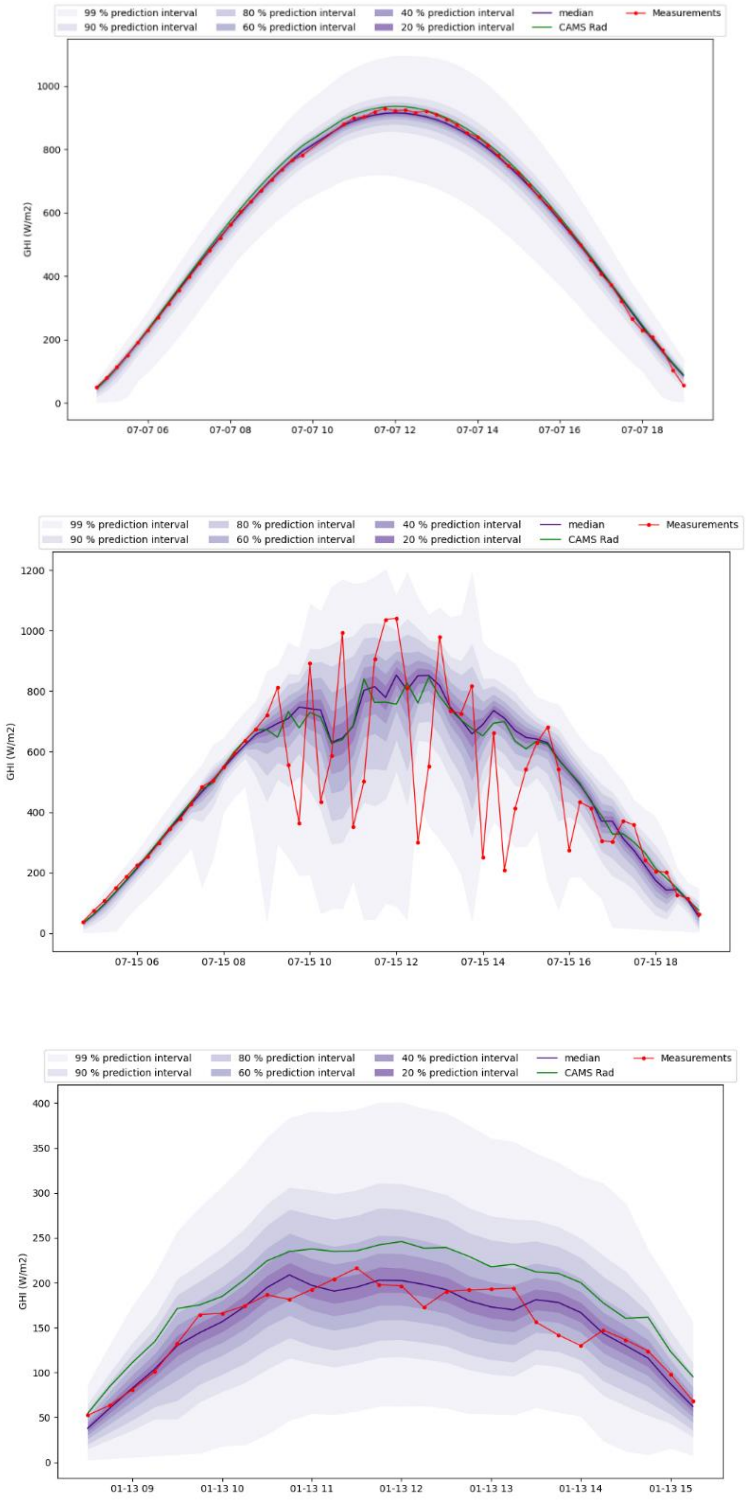

Figure 30 : Comparison of the output of the probabilistic model with CAMS Radiation service (green) and measurements (red dotted lines). The coloured area represents the confidence interval of the probabilistic model and the black line the median of the predicted distribution. The upper, middle and lower graph corresponds to a cloud-free, broken sky and overcast situation respectively.

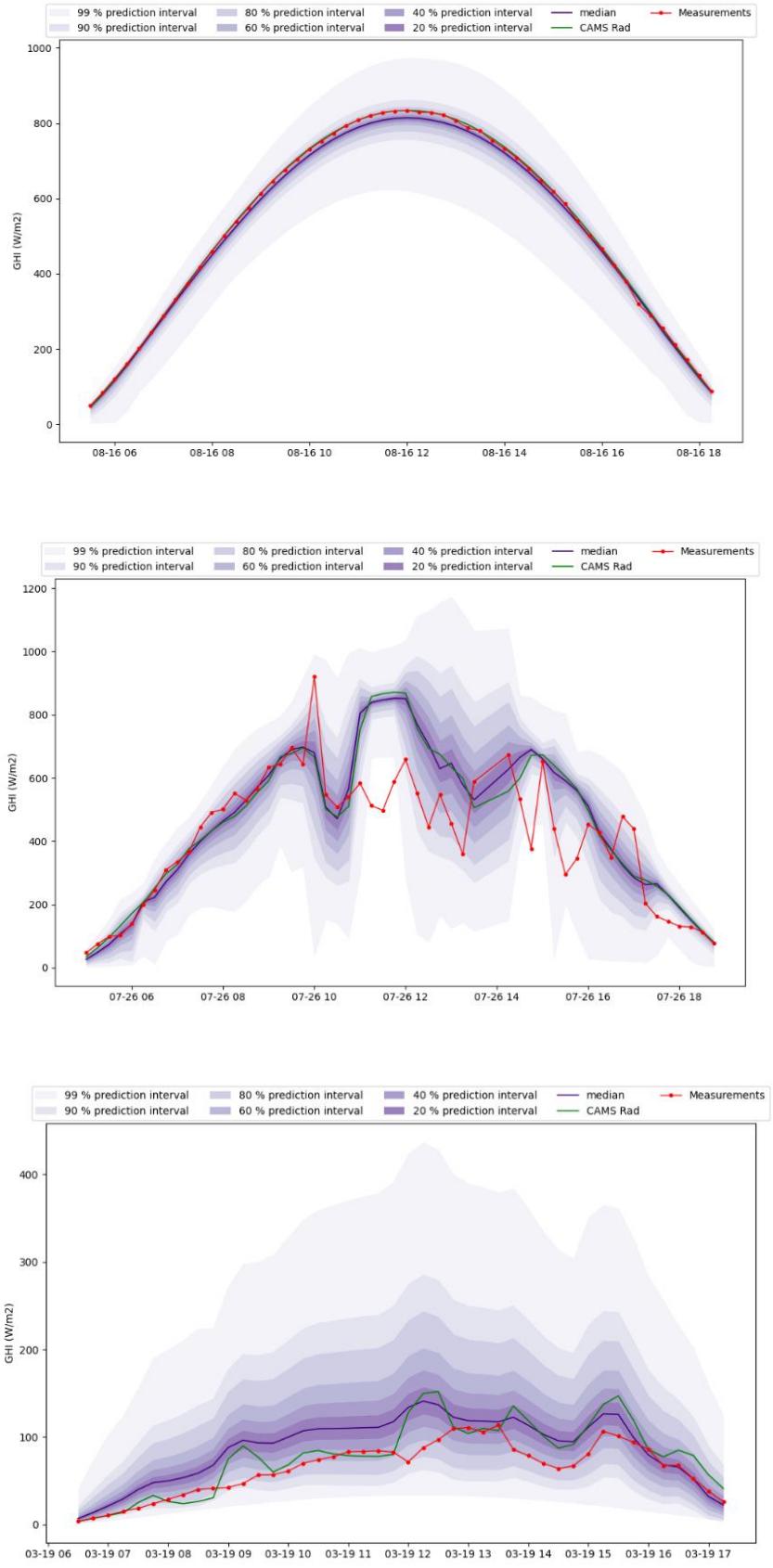

Figure 31 : Same as Figure 30 with issues observed in the probabilistic model.

To verify whether the probabilistic model captures the bias of the CRS estimates, the mean of the probabilistic model, CRS output and measurements were calculated for different classes of clearsky index calculated with CRS. The results are shown in Figure 32 for the training, validation and the two test datasets.

An underestimation of the true average kc is observed at low kc values (black curve above blue line), while an overestimation is observed at high kc values (black curve below blue line). This is true for the four different datasets. The mean predicted by the model is represented by an orange line. These values are very close to the average of measurements for all four datasets. This confirms that in, addition to predicting the distribution of the error the probabilistic model, the model captures the bias of CRS perfectly.

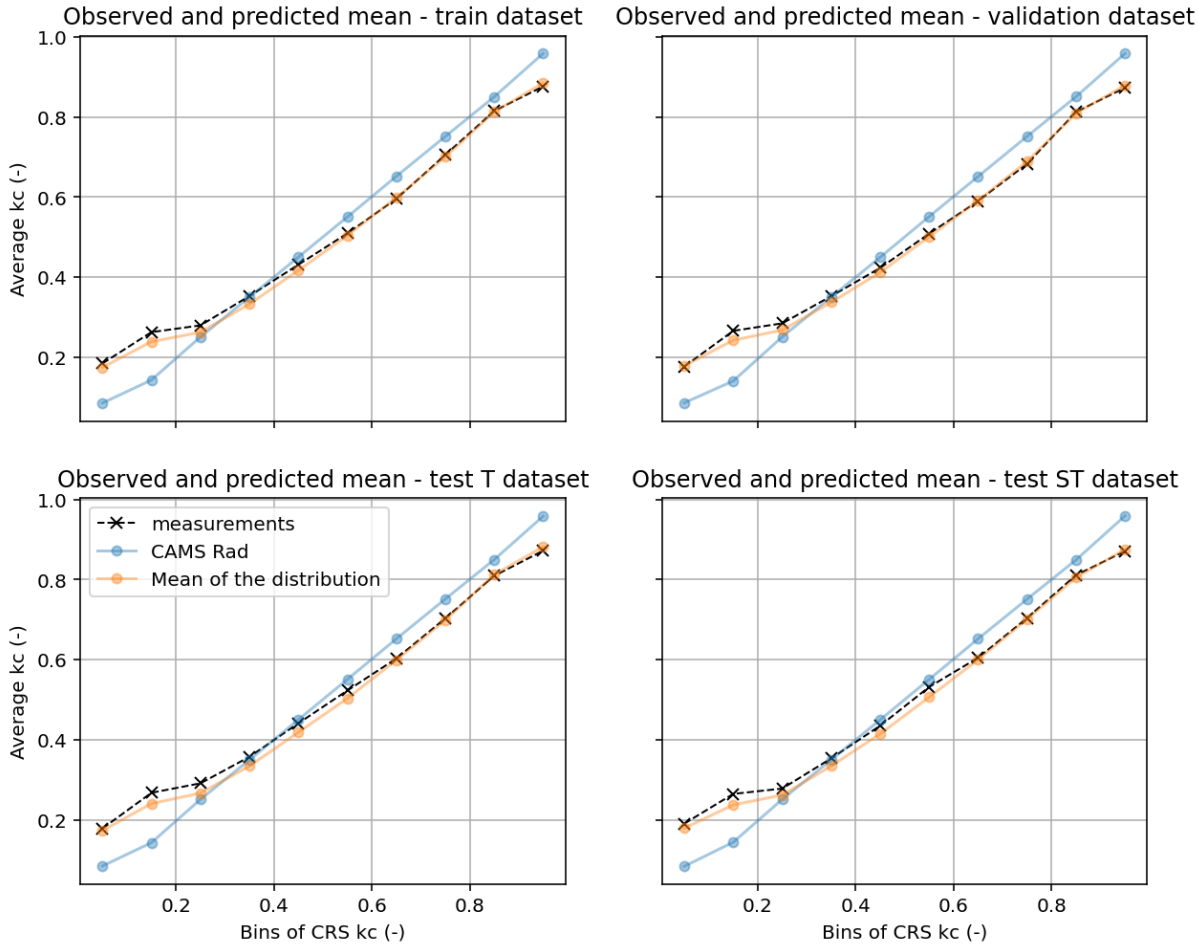

Figure 32 : Observed and predicted mean of the clearsky index for different classes of CRS clearsky index. The different plots represent results obtained on the training, validation, as well as on the two test datasets.

In the same way, the predicted standard deviation is compared to the empirical standard deviation of the CRS errors in Figure 33. For this analysis, the data has been binned according to the predicted standard deviation. Then, for each of these bins, the standard deviation of the error has been calculated. In an ideal probabilistic forecast, the predicted mean (blue line) should be equal to the empirical standard deviation of the error (orange line).

Figure 33 shows that the empirical standard deviation of the error is greater than the predicted standard deviation for the training, validation and the two test datasets, confirming the previous observations that the model is underdispersive. It can also be observed that the predicted and empirical standard deviations agree at low values of the predicted standard deviation for the training, validation and test T datasets. Conversely, the model is significantly more

underdispersive at low values of the predicted standard deviation for the ST dataset than for the three other datasets.

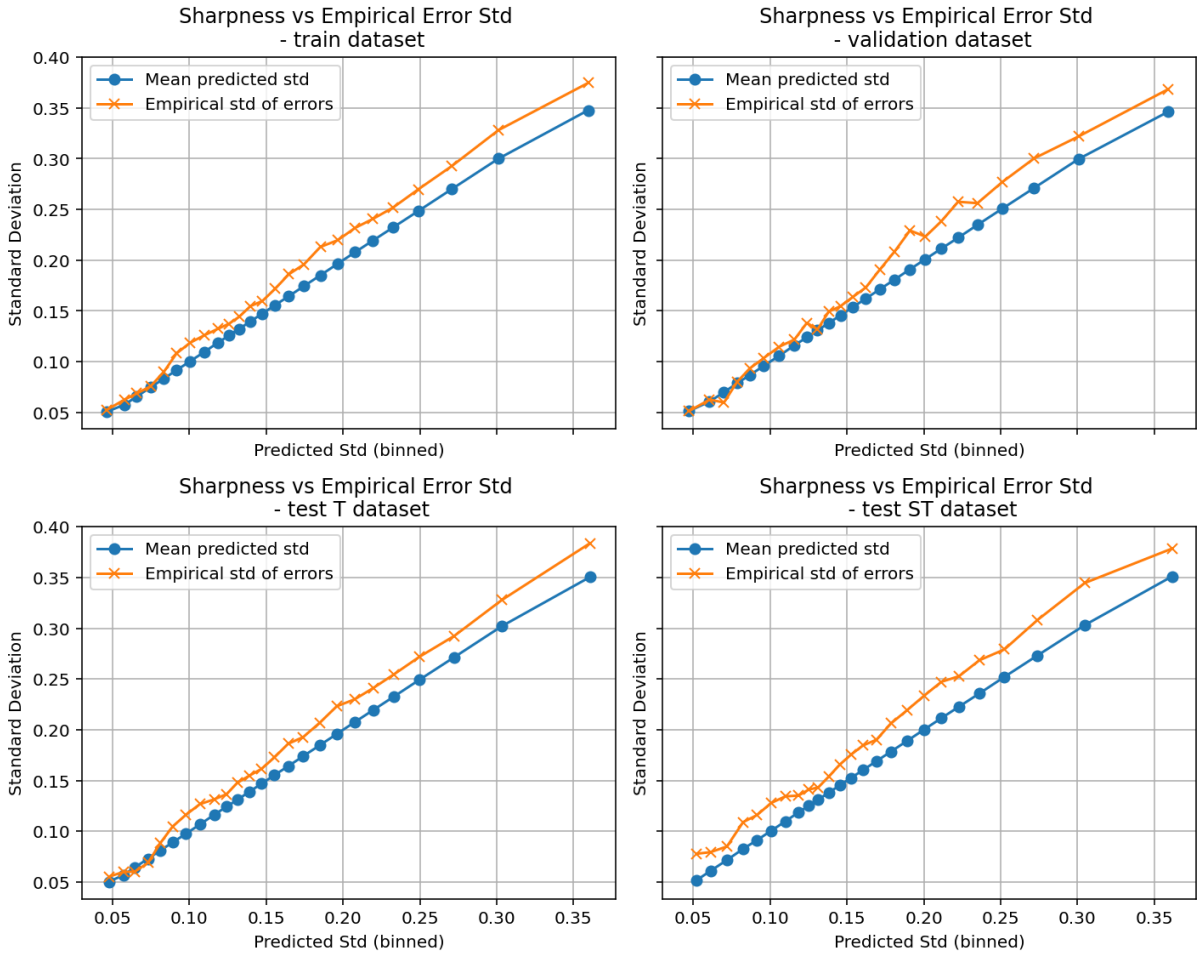

Figure 33 : Comparison of the predicted standard deviation by the model with the empirical mean of the standard deviation of the error of CRS for training, validation and the two test datasets.

To verify the calibration of the model, reliability plots for the training, validation, test-T and test-ST datasets are given in Figure 34.

Reliability plots are diagnostic tools used to assess the coverage of a probabilistic model. They evaluate whether predicted probabilities correspond to observed frequencies, i.e. the calibration of the model. A well-calibrated model produces probabilities that match the empirical frequencies of the measurements. To construct a calibration plot, predicted probabilities are grouped into bins, and for each bin, the average predicted probability is compared to the empirical frequency of the observations. Ideally, the points lie on the identity line, indicating perfect calibration. Deviations from this line reveal systematic biases: forecasts that are overconfident or underconfident.

Figure 34 figure shows that the model is well calibrated for the validation and test T datasets. There is a slight overconfidence for the training dataset and an important overconfidence for the test ST dataset. This difference in calibration for the test ST dataset is consistent with the underestimation of the predicted standard deviation observed for low values of the standard deviation of kc in Figure 33.

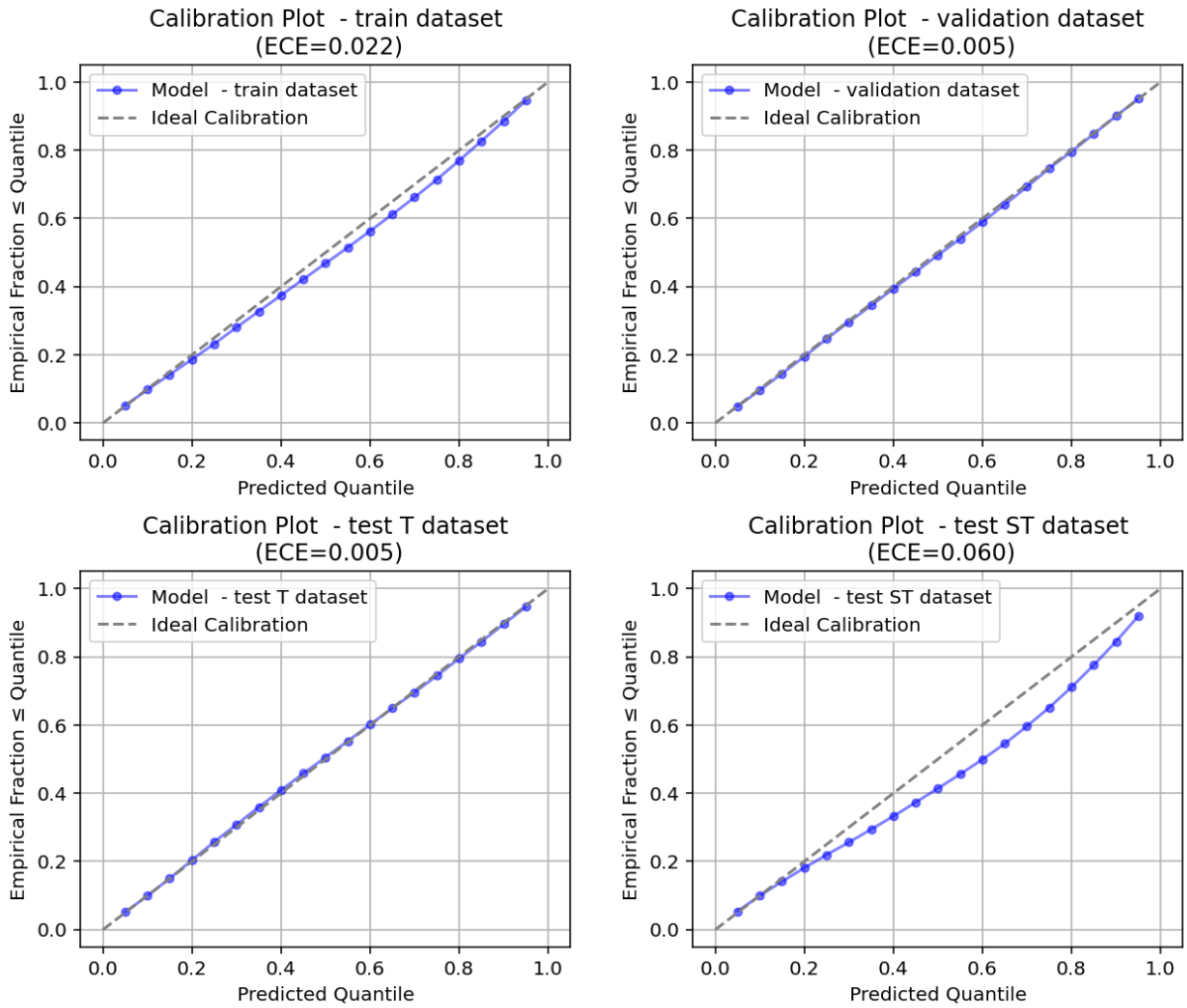

Figure 34 : Reliability plot comparing predicted probabilities to observed frequencies. The diagonal line represents ideal calibration, where predicted probabilities match the true empirical frequencies. The model data are represented by blue lines.

To analyze the calibration and sharpness of the model in more detail, PINAW and PICP were calculated for 10 bins of the clearsky index calculated with CRS and confidence levels of 95, 90, 80 and 50 %. The results are shown in Figure 35, Figure 36, Figure 37 and Figure 38 for the training, validation, test-T and test-ST datasets respectively.

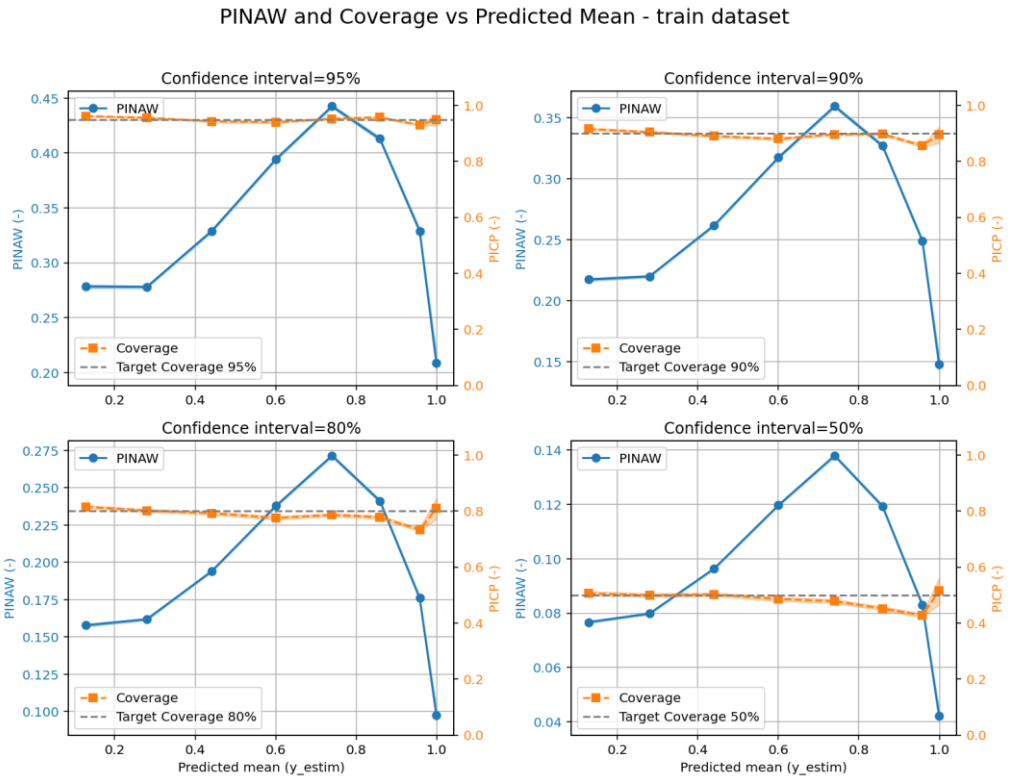

Figure 35 : PINAW (blue curve) and PICP (orange curve) as a function of the clearsky index provided by CRS (x-axis) for confidence intervals of 95, 90, 80 and 50% (upper left, upper right, lower left and lower right plots, respectively) assessed for the training dataset.

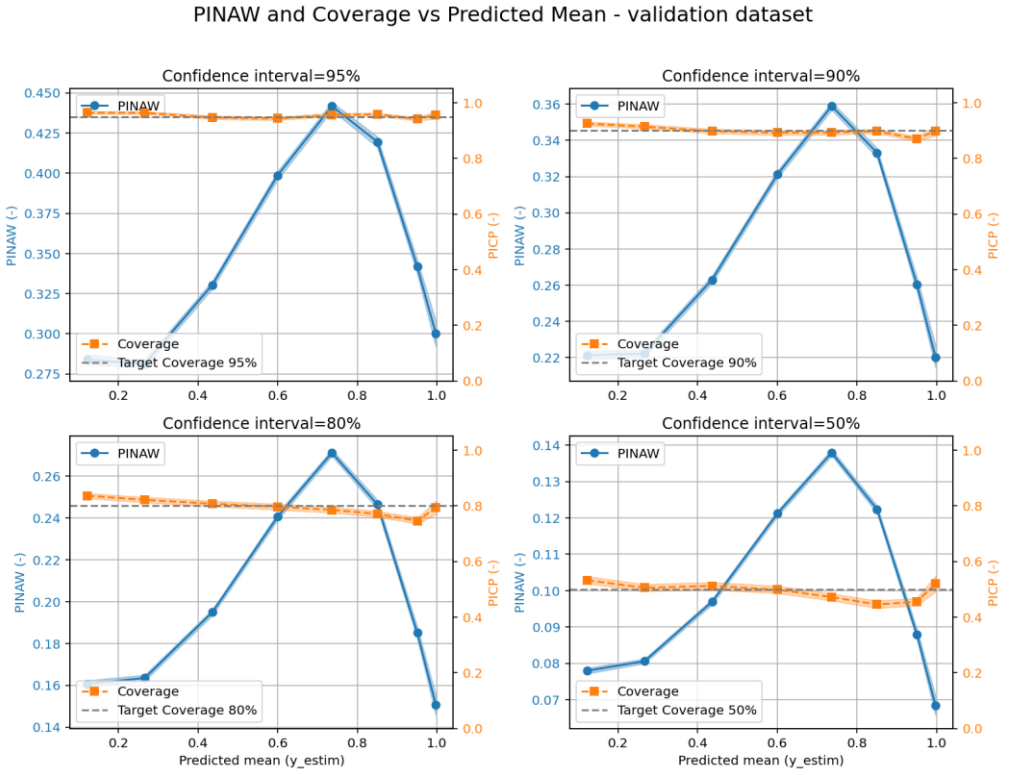

Figure 36 : same as Figure 35 for the validation dataset.

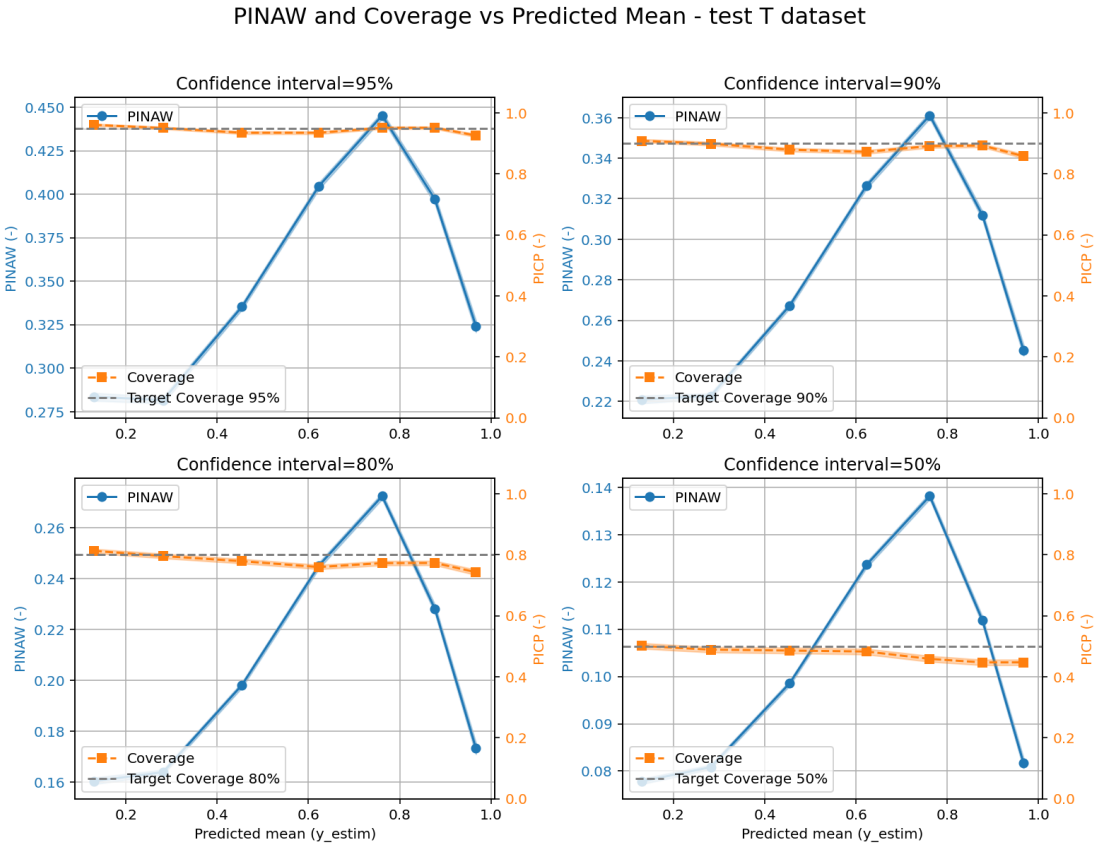

Figure 37 : same as Figure 35 for the temporal split test dataset.

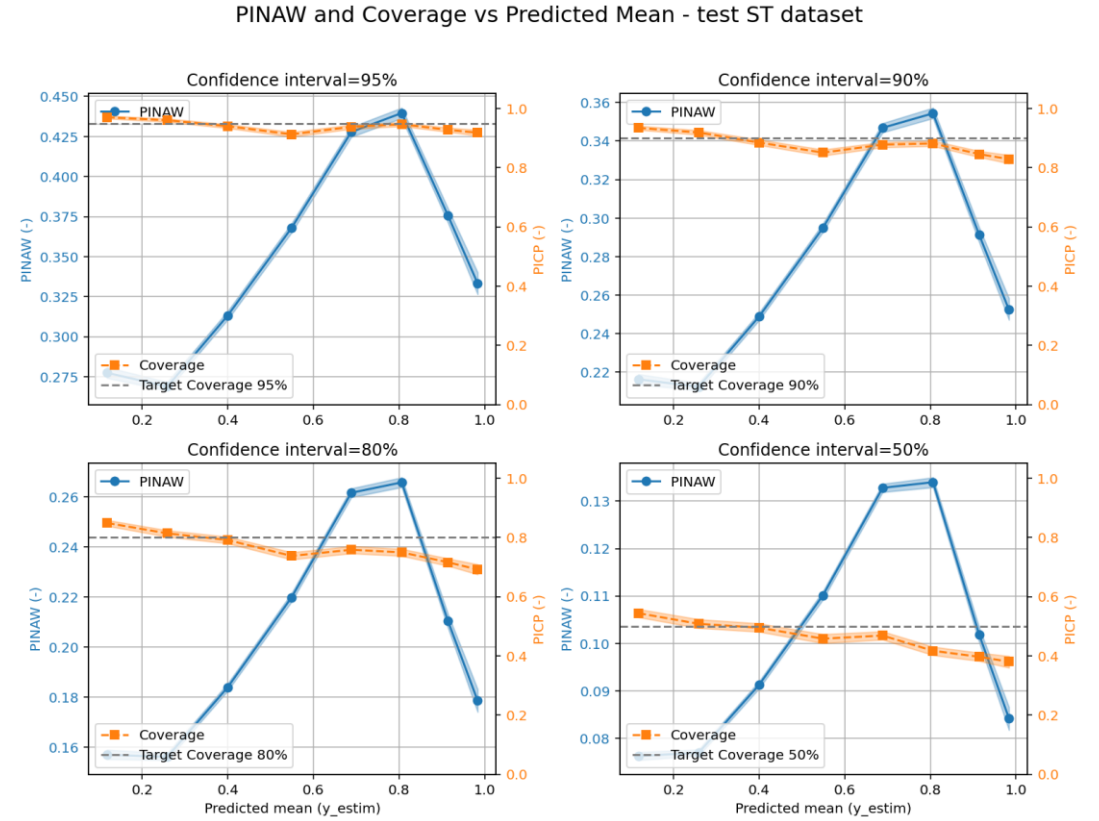

Figure 38 : same as Figure 35 for the spatio-temporal split test dataset.

We can see that – as expected - the model is well calibrated for the training and validation dataset: the PICP (orange line) is close to the different values of the target coverage (grey dashed lines). This is also true for the test-T dataset, further confirming earlier observations of good temporal generalization of the model. For the Test-ST, the coverage is less optimal: we can observe that the model is overdispersive at low kc values and underdispersive at high kc values. This is consistent with the results shown in Figure 34.

The dependence of the PINAW with the kc is similar for all datasets: the spread of the model is small for low and high values of kc and higher at intermediate values. This is consistent with the expectation that the uncertainty of CRS is the highest in broken cloud conditions that corresponds to intermediate values of kc.

We can see that PINAW values in the training dataset are significantly lower than in the validation and test datasets. This is particularly pronounced for the bin corresponding to kc=1 (cloud free situations). The fact that the PINAW is significantly lower in the training than in the test dataset, while the model is underdispersive in the test dataset, could be due to overtraining. This could be remedied by reducing the size of the network and by increasing the number of stations used for training.

6 Conclusions

A well depurated CAMEO reference database was created to be used for the development of this work and CRS developments in general. The base for the database is the collection of ground observations in the ARMINES THREDDS server. First an extended and very strict quality check procedure has been applied to all ground observations available in order to allow only the most reliable data points. The retained ground observations (66 stations) were co-aligned with the CRS operational expert mode output. The procedure followed to create this database maximizes the probability that the deviations found on any posterior analysis from the data come from the CRS model itself and not from errors/inconsistencies on the ground observations.

In this work two different approaches were developed to model the localised uncertainty distribution of the CRS radiation estimates. These approaches investigated the suitability of the input data space at the point of interest (clouds properties, aerosols, water vapour, ozone, surface albedo, solar geometry, etc.) as predictors of the local irradiance error uncertainty. The first approach is based on the parametric binning of the data input space in order to characterize the individual uncertainty distribution per bin. The distributions found are then stored in a LUT and use to infer a local uncertainty distribution of the data point of interest. In the second approach a model of localised irradiance uncertainty was develop using a neural network to predict the parameters of a flexible parametric distribution as a function of data input space. Once trained, the neural network is used to infer the parameters that describe the uncertainty distribution of the data point of interest. In this approach a Johnson SU distribution was chosen over more commonly used alternatives (e.g., Gaussian, log-normal, or beta distributions) due to its ability to independently control the first four moments—mean, variance, skewness, and kurtosis.

Preliminary tests were conducted to evaluate the two approaches to model the uncertainty of CRS. The first results are very encouraging. Both probabilistic models capture very well the bias of CRS in the different tests considered (generalisation in time and generalisation in space and time). It was also shown that both methods are able to capture a reasonable uncertainty value of the individual spatio-temporal CRS estimates.

The parametric based model evaluation used all the available data in the CAMEO refence dataset. In order to replicate the typical use case of the CRS, the data was separated in space and time for the training and inference of the uncertainty distributions. The error model seems to be very well calibrated. The stations known to be difficult to model for the CRS showed the worst calibration and sharpness, which is a reasonable and expected result. The error model was tested to infer the uncertainty distributions on time series outputs on many days/stations which include all types of sky conditions (clear, overcasted, cloudy). For all cases tested, the width of the confidence intervals correlated well with the local variability situation, i.e., narrow intervals in clear and overcasted situations and wider intervals in variable situations.

The deep learning-based model was found perfectly calibrated when tested on the same stations that were used for the training, but a significant decrease in accuracy is observed when the model is applied to stations that were not used for training. A conditional evaluation indicates that in the latter case, the model is over-dispersive at low values of the clearsky index and under-dispersive at high values of the clearsky index. This lack of spatial generalisation can be attributed to an overtraining issue. Further experiments will be conducted with a smaller network and more data to address this issue. The use of a deep-learning model requires a dedicated infrastructure (GPU) making a direct implementation of the approach in the operation CAMS radiation service difficult. However, a simple approach to use the proposed methodology in an operational context could be to pre-calculate the four parameters of the distribution for all possible combinations of the input parameters. The results could be used in a lookup table to emulate the neural network in a simple operational code very similar to McClear and McCloud.

7 Outlook

Both error models developed in this work show sensible and encouraging results and are potential candidates for an implementation on the operational CRS service. Even so, both models should continue to be tested on a larger number or different combinations of predictors from the input data space. The improvement and deterioration of the quality indicators should be evaluated with the new predictor combinations and the computational efficiency should be continuously assessed in view of an optimal operational implementation in the CRS.

The uncertainty distribution of a local irradiance estimate can be inferred from the error models developed in this work. This amount of uncertainty information could be overwhelming for the typical CRS user. Discussion have already started with the CRS development team and the directly with the users to define simple and useful uncertainty indicators to be implemented as an operational CRS product. The quest here is to find the indicators that will help the typical user to better understand the data and ultimately take better decisions. This is not an easy task, as there is always an equilibrium to be found between the amount of information given and its interpretability. This effort has already stared on the different conferences were the CRS development teams was present (e.g.,: ICEM2025, EGU2025) and will continue as a CRS team task.

8 References

ARMINES. 2025. Webservice-energy. *Webservice-energy THREDDS Catalog.* [Online] 2025. https://tds.webservice-energy.org/thredds/catalog/in-situ.html.

CAMS Radiation Serivce, CAMS. 2025. *D1.3.1 Regular Validation Report.* 2025.

Long, C.N., Dutton,. 2002. *BSRN Global Network recommended QC tests, V2.0.f.* [Available online at http://epic.awi.de/30083/1/BSRN_recommended_QC_tests_V2.pd] 2002. (Last accessed 10/06/2025).

Vaisala. 2025. CAMS Radiation Service API access. [Online] 2025. [Cited: 10 06 2025.] https://www.soda-pro.com/help/cams-services/cams-radiation-service/automatic-access#wget.

Document History

Version	Author(s)	Date	Changes
0.1	Jorge Lezaca	13/06/2025	Initial version
1.0	Jorge Lezaca	June 2025	Updated after internal review

Internal Review History

Internal Reviewers	Date	Comments
Ana Carvalho	21/06/2025	Content and formatting issues raised.
Janot Tokaya	24/06/2025	Content and formatting issues raised.