Solar Physics (2021) 296:149
https://doi.org/10.1007/s11207-021-01893-3

The Effect of Missing Groups in the Calculation of the Solar Irradiance Deficit

Analysis of the Sunspot Areas from the SOON Network

Luis Leuzzi[1,2] · **Laura A. Balmaceda**[3,4] · **Carlos Francile**[5]

Received: 21 April 2021 / Accepted: 5 September 2021 / Published online: 18 October 2021

Abstract
Sunspot areas are one of the most important indices of solar activity. To obtain an extended time series covering multiple solar cycles, one must combine data from different observatories after a proper comparison and calibration of the individual datasets. We compare the daily and group values of sunspot areas provided by the different stations from the Solar Optical Observing Network, SOON, which are determined using similar instruments and techniques. We investigate if there are systematic differences among the stations and whether the differences in the daily values can be attributed to missing groups in the records or errors in the measurements. We find significant differences among the stations of the SOON network in terms of sizes (average daily and group values), quality of observations and coverage (considering the number of missing groups and data gaps). Our results indicate that calibration factors for daily values can be used with confidence to combine datasets from different stations. However, for some applications which require the location of the sunspot groups, the same correction factors should not be used. We estimate the irradiance deficit due to sunspot through the Photometric Sunspot Index and compare the output from similar datasets to quantify the effect of missing groups. We find differences as high as 150 ppm during the maximum of solar cycle. The effect increases for sunspot groups near the center of the disk accounting for about 80% of the observed differences.

Keywords Solar cycle · Observations · Sunspots · Statistics

✉ L. Leuzzi
leuzzi@unsj-cuim.edu.ar

L.A. Balmaceda
lbalmace@gmu.edu

C. Francile
cfrancile@unsj-cuim.edu.ar

1 Facultad de Ciencias Exactas, Físicas y Naturales (FCEFyN), Universidad Nacional de San Juan (UNSJ), San Juan, Argentina

2 CONICET, San Juan, Argentina

3 George Mason University, Fairfax, VA, USA

4 NASA Goddard Space Flight Center, Greenbelt, MD, USA

5 Observatorio Astrónomico Félix Aguilar (OAFA), UNSJ, San Juan, Argentina

1. Introduction

Studies of the long-term solar variability require the use of daily historical records of sunspot areas, along with the most recent data from several different observatories. In order to combine various datasets of sunspot areas into a homogeneous single record, a proper comparison and calibration of the data should be carried out (see, e.g., Fligge and Solanki, 1997; Balmaceda et al., 2009; Mandal et al., 2020). This is imperative because significant discrepancies may arise from the different instrumentation, as well as different observing and measuring techniques used by independent observatories.

Historically, data from the Royal Greenwich Observatory (RGO) has been used as the backbone in studies of the long-term solar activity. After the ending of the RGO records in 1976, the data from the SOON (Solar Optical Observing Network) was generally used to fill in the information in recent cycles, from Cycle 21 onward.

Past studies (Balmaceda et al., 2009; Hathaway, 2010) revealed that the measurements of the size of sunspots made by the telescopes from SOON network, present large differences (of about $\sim$40%) compared to those obtained by other observatories. In spite of the efforts, however, there is no consensus as to the magnitude of that difference (see, e.g., Foukal, 2014; Muñoz-Jaramillo et al., 2015b).

Discrepancies in sunspot areas provided by independent observatories may differ due to, among other factors, random errors introduced by the observer's bias. Moreover, the characteristics of the instruments (resolution, sensitivity, etc.) and the techniques used for both, the observation and measurements, also affect the final products (e.g., areas, location, etc.). In this context, the SOON network provides us with a great opportunity to investigate the extent of these errors by comparing data obtained with similar instrumentation and, in principle, same observing and measuring techniques. In the recent work by Giersch, Kennewell, and Lynch (2018), the authors compared daily sunspot areas from five of the SOON stations. Overall, they found that the daily measurements agree within 5% over the whole period of observations. The authors report that differences as large as 8.5% can arise from down-rounding sunspots at the limb while other sources of error are the local seeing conditions and bad quality of the drawings.

In this work, we present a detailed study of the sunspot area datasets from each of the observatories that belong to the SOON network, for the period 1982–2013. We include the comparisons of eight stations and use both daily and group measurements. The main goal of this study is to quantify the effect of missing sunspot data in the PSI calculation. We should note that the SOON network explicitly leaves out spots with areas smaller than 10 micro-hemispheres, so these areas are not considered here. Foukal (2014), however, points out that the influence of small sunspots in the PSI is of minor importance.

SOON data has been widely used in earlier studies, before more reliable datasets (such as KMAS–Kislovodsk Mountain Astronomical Station, PCSA–Pulkovo's Catalog of Solar Activity, or DPD–Debrecen Photoheliographic Data) became available to the scientific community. Moreover, in spite of the acknowledged flaws in the dataset, these sunspot areas are still used in a wealth of current applications (see, e.g., Bhowmik and Nandy, 2018, who used a correction factor of 1.4 only for areas smaller than 206 micro-hemispheres). The importance of the SOON dataset lies in that they serve as a link between the historical record of the Greenwich Royal Observatory (1874–1976) and the most recent observations (after 1976). Among the many applications of sunspot areas, the reconstruction of total solar irradiance variations is one of the most important. Certainly, uncertainties in sunspot area measurements translate into uncertainties in the irradiance estimates both in the short and

the long term (see, e.g., Balmaceda et al., 2009; Foukal, 2014). In this work, we further investigate the influence of missing sunspot groups in the available datasets when estimating the Photometric Sunspot Index (PSI).

The paper is organized as follows. In Section 2 we present a summary of the main characteristics of the SOON data: group and daily areas coverage per station, typical sizes, position of sunspot groups on the disk and quality of the observations. In Section 3 we describe the methodology used to determine the calibration factors and present the results from the comparison of group and daily areas. In Section 4 we estimate the effect of missing sunspot groups on the calculation of the Photometric Sunspot Index (PSI). The discussion and the conclusions are presented in Sections 5 and 6, respectively.

2. The SOON Network

The Solar Optical Observation Network comprises nine stations: Hollman Air Force (HOLL), Learmonth (LEAR), San Vito (SVTO), Ramey (RAMY), Mount Wilson (MWIL), Boulder (BOUL), Manila (MANI), Palehua (PALE), and Culgoora (CULG). The network is continuously monitoring the Sun over 24 hours. In a recent work, Giersch, Kennewell, and Lynch (2018) provide a detailed description of the observing procedure and the techniques used to measure sunspot areas. In this work, we use the group sunspot areas from the USAF_MWL dataset available for the period 1981–2013 at: https://www.ngdc.noaa.gov/stp/solar/sunspotregionsdata.html.

In Table 1 we summarize the main characteristics of the individual databases of each station for the period 1982–2013. The names of the stations are listed in column 1, the number of days with observations by each station is given in column 2, and the number of days with sunspot group number (S_N) different than zero is listed in column 3 for reference. The total number of sunspot groups with at least one record is indicated in column 4, while the total number of sunspot groups indicated by the NOAA identification number in the full period is specified in column 5. The start and end date of the period of observations for each station is given in columns 6 and 7, and the corresponding solar cycle numbers are shown in column 8. The three last columns 9, 10, and 11 provide the percentages of coverage of daily and group sunspot records which are discussed in detail in Section 2.1.

The stations with longest runs of observations are HOLL, LEAR, and SVTO as can be seen in column 2, covering at least three solar cycles.

2.1. Group and Daily Areas Coverage per Station

To determine the fraction of missing spots in the SOON records per station, we make the following considerations:

(a) First, in order to identify real gaps in the daily observations, we inspect the daily sunspot number S_N. Days when the daily sunspot number is 0, i.e., spotless days which are common during cycle minima, are not considered as data gaps and are therefore included in the calculations.

(b) Second, to detect individual groups missing in the records, we use the NOAA sunspot group number which is provided in the original data from SOON. This allows us to use the number of days in which a sunspot group is reported by NOAA as reference to estimate, e.g., if a sunspot group has been tracked by a SOON station for the full period in which the active region was present on the disk.

Table 1 Characteristics of the observations from the stations in the SOON network. The superscript [1] indicates the stations which continue operating.

Station (1)	Nr. days with obs. (2)	Nr. days $S_N \neq 0$ (3)	Nr. groups with obs. (4)	Nr. groups NOAA (5)	Start Date (6)	End Date (7)	Solar Cycle (8)	Coverage		
								Groups [%] (9)	Daily [%] (10)	Rep. Groups [%] (11)
LEAR	8973	11688	7282	8467	01/01/1982	12/31/2013[1]	21–24	86	77	70 ± 4.11
SVTO	6791	9953	5890	7193	10/02/1986	12/31/2013[1]	22–24	82	68	62 ± 1.08
RAMY	6048	7791	5820	6876	01/01/1982	04/30/2003	21–23	85	78	68 ± 3.35
HOLL	8632	11686	7263	8466	01/03/1982	12/31/2013[1]	21–24	86	74	68 ± 3.35
MANI	523	1499	791	1237	01/01/1982	02/06/1986	21	64	35	40 ± 7.23
BOUL	2924	4749	3066	4229	01/01/1982	04/22/1994	21–22	72	62	54 ± 1.94
PALE	3488	5577	3435	4551	01/01/1982	04/07/1997	21–23	75	62	57 ± 0.81
CULG	1509	3756	1783	2398	06/01/1986	04/12/1992	22	74	40	54 ± 1.94

We obtain the three coverage indices listed in Table 1 as follows:

- Group coverage (column 9) is determined by dividing the number of groups with at least one reported observation by the total number of sunspot groups reported by NOAA, i.e., identified with a unique number, over the period of observations. This percentage is obtained using the values in columns 4 and 5.
- Daily coverage (column 10) is calculated by dividing the number of days with reported sunspot groups (column 2) by the total number of days observed in the period with $S_N \neq 0$ (column 3).
- The coverage of reported groups (column 11) is obtained as the sum of days in which a group was actually reported by the SOON station divided by the total number of days in which the group should have been observed according to the NOAA identification number. This gives the number of sunspot groups in the SOON reports that are not tracked over the full period reported by the NOAA groups. We include the standard deviation as a measure of the dispersion with respect to this average value.

We find that many sunspot groups from the SOON records are not reported over the whole period they appear on the disk. For example, 4% of the groups are reported only one day (some of them reported by more of one station) and about 2% of the groups are found in the records only once (i.e., reported by a single station and only one day). Also, about 7% of sunspot groups are not reported by any of the stations.

The stations with the highest coverage, both for individual groups and daily areas, are HOLL, LEAR, SVTO, and RAMY as can be seen in columns 9–10, with the latter ending the activities in 2003. We can see that even for these stations, apart from the gaps in the observations (i.e., days without any data), there is also a high percentage of missing groups in the reported data (of at least 14%, as seen from column 9 in Table 1).

2.2. Sizes and Position on the Disk

In this section we present the characteristic sizes of sunspots averaged over the full period of observations for each station. Mean, minimum, maximum, and median values of individual sunspot groups are shown in Table 2. The maximum daily areas are also included, as well as the interquartile range (IQR) as a measure of the statistical dispersion of the data.

Table 2 Typical sizes per station. The values are given in ppm of the solar hemisphere.

Station	Min Area	Max Group Area	Mean Group Area	Median Group Area	IQR	Max Daily Area
LEAR	10	5400	142	60	140	8420
SVTO	10	9070	147	70	130	9800
RAMY	10	9910	151	60	140	12060
HOLL	10	9980	146	60	140	11570
MANI	10	3700	209	100	210	5380
BOUL	10	3900	159	70	150	4540
PALE	10	3600	156	70	150	5620
CULG	10	8020	155	60	160	8860

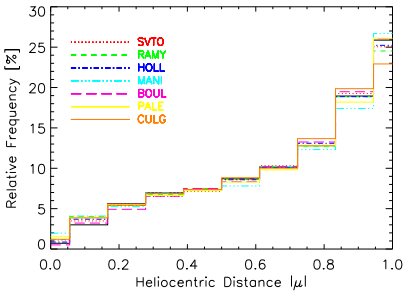

Figure 1 Fraction of sunspot groups observed on the disk per bins of 0.1μ values and per station.

The differences in the mean sizes for the different stations seem to arise from variations in the activity level both within a cycle and from one cycle to another. For example, the mean group areas for observatories with data only from Cycles 21 and 22 (such as MANI, BOUL, and CULG) are higher than those from observatories including observations in the weaker cycles, namely 23 and 24 (such as LEAR and SVTO).

MANI differs noticeably from the other datasets. While the maximum group area is almost the lowest with the exception of PALE, the mean and median of the group areas, together with the IQR value, are the highest. This point is discussed further in Section 5.

We analyze whether there is a preferential average position of the sunspots on the solar disk. For this, we estimate the percentage of sunspot groups per bins of 0.1 of the heliocentric distance μ for each station as shown in Figure 1. In this case, all individual groups reported by the stations are taken into account. The distribution is not uniform. In general, the fraction of observed sunspot groups decreases from approximately 25% at the center of the solar disk ($\sim \mu = 1$) to about 5% near the limb ($\mu < 0.3$). About 60% of the sunspot groups correspond to locations near the center $\mu > 0.7$.

2.3. Quality of Observations

The reports provided by the SOON stations include an index for the quality of the observations: *very poor* $= 1$, *poor* $= 2$, *fair* $= 3$, *good* $= 4$, and *excellent* $= 5$ (Giersch, Kennewell, and Lynch, 2018). The classification is made according to the white-light seeing conditions at the observing site and is provided in the individual reports from SOON.

The characterization of the quality of the observations is provided in Table 3. The values listed correspond to the full period of operations for each station.

Table 3 Distribution of the quality of observations per station.

Station	Quality index				
	1 [%]	2 [%]	3 [%]	4 [%]	5 [%]
LEAR	3.5	23.5	56.3	15.7	0.9
SVTO	3.0	31.1	58.3	7.1	0.5
RAMY	3.7	15.9	52.5	27.0	0.9
HOLL	2.2	20.5	54.6	22.3	0.3
MANI	0.0	27.4	72.6	0.0	0.0
BOUL	22.7	32.2	38.5	6.4	0.1
PALE	2.5	19.1	64.9	13.4	0.1
CULG	8.2	42.1	46.9	2.8	0.0

From this table, we find that RAMY is the observatory with best quality of observations, followed by HOLL, LEAR, and PALE. The poorest observations, on the other hand, are those reported by MANI, CULG, and BOUL. SVTO ranks in the middle range.

3. Comparison of Sunspot Areas

3.1. Methodology

In order to derive a calibration factor among different observatories, we follow the procedure described in Balmaceda et al. (2009). It mainly consists of estimating the slope for the regression line through the origin of two given datasets during the period of overlap using the Ordinary Least-Squares (OLS) method.

We use the bisector method (Isobe et al., 1990; Eisenhauer, 2003), which computes the regression line which bisects the two OLS(X|Y) and OLS(Y|X) curves. This method is suitable for the problem we are dealing, in which it is not clear which dataset should be treated as the independent variable and which as the dependent variable in the regression.

The slope of the regression line, or calibration factor β, can be determined as:

$$\beta = (b_1 + b_2)^{-1} \left[b_1 b_2 - 1 + \sqrt{\left(1 + b_1^2\right)\left(1 + b_2^2\right)} \right], \tag{1}$$

with b_1 and b_2 being the slopes from OLS(X|Y) and OLS(Y|X), respectively.

As suggested by Isobe et al. (1990), Akritas and Bershady (1996), the corresponding error for the slopes can be computed using *bootstrapping*. Consequently, we derive the bootstrapped 95% confidence interval (CI) for the slopes in the linear regression (i.e., calibration factor β). The estimated confidence interval is based on 1000 replications.

Following Muñoz-Jaramillo et al. (2015b), we impose a threshold limit for all datasets of one order of magnitude above the minimum size reported. In this way, we only use data values above this limit in the estimate of the slopes since small areas introduce bias.

We take LEAR, SVTO, and HOLL as the base observatories since these are the longest databases in the network. We compare the data from the rest of stations with them. Although RAMY is a station with high coverage and the best quality of observations, we do not include it as base observatory because of its shorter period of operations compared with the other three stations.

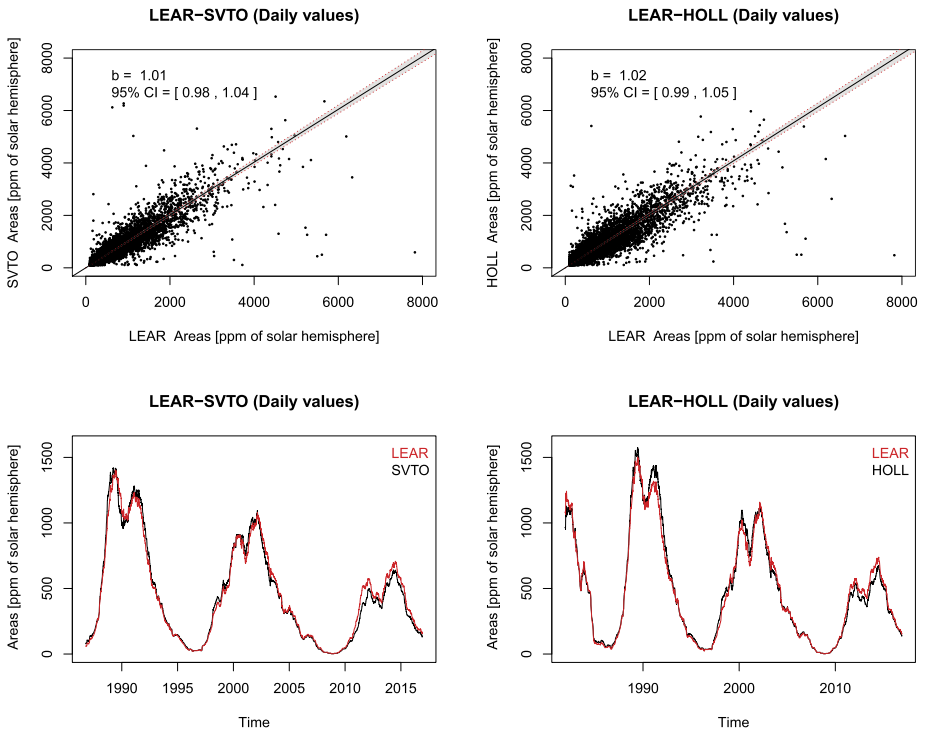

Figure 2 (Upper panels) Scatter plot for the comparison of daily sunspot areas. Solid lines represent the linear fit to the data. The gray bands show the 95% CI for the slope. (Lower panels) 12-month running means of daily sunspot areas vs. time.

Differently from Balmaceda et al. (2009), which only perform the comparison of daily areas, we apply the procedure to both daily and individual group areas.

3.2. Calibration Factors for Groups and Daily Areas

In this section, we present the results from the comparison of individual sunspot group and daily areas taking pairs of SOON stations.

In Figure 2 we show a comparison of daily sunspot areas corrected for foreshortening between pairs of different stations. Panel (a) shows SVTO vs. LEAR, and panel (b) HOLL vs. LEAR. Solid lines represent linear regressions to the data neglecting an offset (i.e., forced to pass through zero), as well as data points close to the origin. The gray bands show the 95% CI for the slope. In the lower panels of Figure 2, we show the 12-month running means of daily sunspot areas vs. time. Red curve shows the data used as basis level, and black curves are the data from the second observatory.

In Figure 3 we present a comparison of individual group areas corrected for foreshortening between pairs of different stations. Panel (a) shows SVTO vs. LEAR, and panel (b) HOLL vs. LEAR. As in the case for daily values, solid lines represent linear regressions to the data neglecting a possible offset (i.e., forced to pass through zero), as well as data points close to the origin. The gray bands show the 95% CI for the slope.

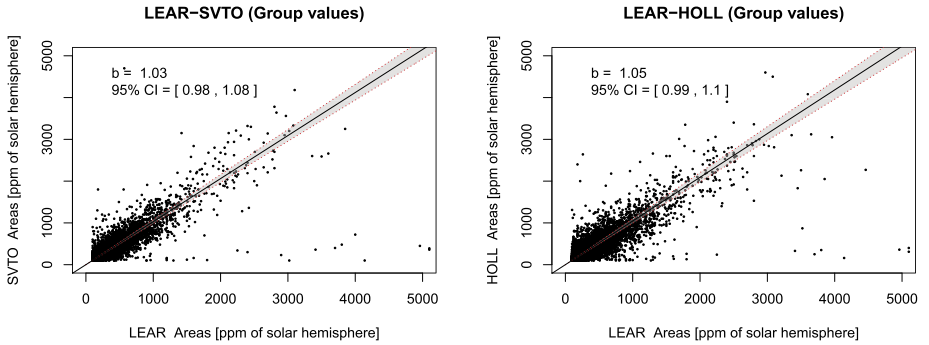

Figure 3 Scatter plot for the comparison of group sunspot areas. Solid lines represent the linear fit to the data. The gray bands show the 95% CI for the slope.

Table 4 Calibration factors β (see Equation 1) for group and daily sunspot areas.

Obs. AUX	Obs. BAS	Overlapping Period	β Group	95% CI	β Daily	95% CI
RAMY	LEAR	01/01/1982 – 04/30/2003	1.11	1.05 – 1.18	1.08	1.04 – 1.11
MANI	LEAR	01/01/1982 – 02/06/1986	1.01	0.94 – 1.09	1.10	1.04 – 1.15
BOUL	LEAR	01/01/1982 – 04/22/1994	0.89	0.86 – 0.92	0.88	0.85 – 0.90
PALE	LEAR	01/01/1982 – 04/07/1997	0.98	0.95 – 1.01	0.98	0.95 – 1.00
CULG	LEAR	05/07/1986 – 04/12/1992	0.90	0.85 – 0.94	0.87	0.83 – 0.90
RAMY	SVTO	10/02/1986 – 04/30/2003	1.06	1.00 – 1.13	1.05	1.01 – 1.10
BOUL	SVTO	10/02/1986 – 04/22/1994	0.84	0.80 – 0.87	0.85	0.81 – 0.88
PALE	SVTO	10/02/1986 – 04/07/1997	0.93	0.90 – 0.97	0.96	0.93 – 0.98
CULG	SVTO	06/01/1986 – 04/12/1992	0.85	0.82 – 0.87	0.82	0.78 – 0.85
RAMY	HOLL	01/03/1982 – 30/04/2003	1.03	0.96 – 1.10	1.03	1.00 – 1.07
MANI	HOLL	01/03/1982 – 02/06/1986	0.93	0.72 – 1.10	1.08	0.94 – 1.20
BOUL	HOLL	01/03/1982 – 04/22/1994	0.85	0.83 – 0.87	0.84	0.82 – 0.86
PALE	HOLL	01/03/1982 – 04/07/1997	0.93	0.88 – 0.97	0.93	0.91 – 0.96
CULG	HOLL	06/01/1986 – 04/12/1992	0.86	0.84 – 0.89	0.82	0.78 – 0.85
SVTO	LEAR	10/02/1986 – 12/31/2013	1.04	0.99 – 1.09	1.01	0.99 – 1.04
HOLL	LEAR	01/03/1982 – 12/31/2013	1.05	0.99 – 1.11	1.02	0.99 – 1.05
SVTO	HOLL	10/02/1986 – 12/31/2013	0.97	0.91 – 1.03	0.98	0.95 – 1.01

The calibration factors β for each pair of stations with their respective 95% confidence interval are summarized in Table 4. The values derived from both comparisons, i.e., group and daily sunspot areas, are listed.

The longest periods of overlapping observations are given by SVTO, HOLL, and LEAR as can be seen in the three last rows in Table 4. The comparisons between pairs of these stations do not yield statistically significant differences, with the correction factors lying close to 1 for both, group and daily areas.

The observatories which do present statistically significant differences with SVTO, HOLL and LEAR are BOUL, CULG, and PALE. The trends hold for both group and daily areas.

For most of the comparisons in Table 4, the calibration factors for group and daily areas differ in less than 4%. In particular, for the comparisons involving BOUL, LEAR, PALE, RAMY, SVTO, and HOLL, the difference is 1% or less. These stations are among those with the highest quality of observations and also the best coverage. The largest differences in the corrections factors are found for MANI–LEAR and MANI–HOLL, and the reason for such a discrepancy is probably the missing sunspot groups in MANI reports.

Although the comparison of group areas from LEAR and HOLL with MANI yields CI for the calibration factors including 1, the upper and lower values differ by 15% and over 30%, respectively. This might be due to the large variability in MANI data. Also, the period of overlapping is only 4 years.

4. Use of Sunspot Areas in Irradiance Variation Estimates

4.1. The Photometric Sunspot Index (PSI)

One of the most widely used applications of sunspot areas concerns the estimates of the total irradiance deficit due to passage of sunspots across the solar disk. Hudson et al. (1982) defines the P_S or PSI (*Photometric Sunspot Index*) to quantify the decrease in total solar irradiance due to the passage of sunspots onto the solar disk. The deficit in the radiative flux due to a sunspot is expressed in terms of the quiet sun irradiance with a value $S_Q = 1360.8 \pm 1.3\ \mathrm{W\,m^{-2}}$ (Kopp, Lawrence, and Rottman, 2005; Kopp and Lean, 2011) and it is defined as

$$\frac{\Delta S_S}{S_Q} = \frac{\mu A_S\,(C_S - 1)\,(3\mu + 2)}{2}, \tag{2}$$

where A_S accounts for the sunspot areas in millionths of the solar hemisphere and μ is the heliocentric distance. The factor $C_S - 1$ represents the residual intensity contrast of the sunspot with respect to the photospheric background. According to Brandt, Stix, and Weinhardt (1994), it depends on the size of the spot through the relation

$$C_S - 1 = -0.2231 - 0.0244 \log\,(A_S)\,. \tag{3}$$

The photometric index P_S is then obtained by adding up the contributions of the n sunspots present on the disk for a given day,

$$P_S = \sum_{i=1}^{n} \left(\frac{\Delta S_S}{S_Q} \right)_i . \tag{4}$$

Therefore, the P_S index depends on both the size and position of the sunspot relative to the center of the disk. When the sunspot crosses the central meridian, the deficit is maximum and produces large drops in the total irradiance.

Here we use the sunspot areas from the individual stations of the SOON network to estimate the P_S index rather than the averaged value among all the observatories as done usually (see, e.g., Balmaceda et al., 2009). We also construct a composite file combining data from two stations and discuss the variations in the estimate of the irradiance deficit caused by the missing data from individual sunspot groups in the SOON records in the next section.

Using the combined dataset, we are able to fill in the gaps corresponding to 9915 and 11809 missing groups in LEAR and HOLL datasets, respectively. Note that these numbers

correspond to the sum of all the missing records of a given group. The group coverage listed in column 11 in Table 1 increases from 70% to 92% for LEAR and from 68% to 86% for HOLL. In this way the differences in the total daily areas and therefore in the P_S index can be attributed to the missing groups in the reports by each station.

The analysis presented in the next section was conducted in `R` (R Core Team, 2017) and figures were produced using the packages: `boot` for *bootstrapping* (Canty and Ripley, 2017; Davison and Hinkley, 1997), `reshape` (Wickham, 2007), `tidyr` (Wickham and Henry, 2018), and `plyr` (Wickham, 2011).

4.2. The Effect of Missing Sunspot Groups in Irradiance Estimates

To estimate the net effect of the missing sunspot groups in the P_S calculation, we use a composite dataset. This composite time series is built by combining two of the stations into a single file. We choose LEAR and HOLL, based on the following criteria: (i) these two stations cover the whole period of observations; (ii) they both show the maximum coverage in the individual and total sunspot areas; and (iii) the comparison of their daily and individual spot areas is of the order of 5% (see Section 3.2), which we take as within the expected error. For this reason, we will not use any correction factor to multiply either of the datasets.

To build the composite file, we first identify the missing groups in both LEAR and HOLL datasets separately for each day. The sunspot group areas in a given day that are detected in one but not the other dataset are entered directly in the composite file. For the rest of the groups, i.e., those that are reported by both stations, we test different strategies and combine the observations by:

(a) taking the mean value of both measurements;
(b) taking the maximum value of both measurements;
(c) taking LEAR as the base observatory and filling in gaps with HOLL data;
(d) taking HOLL as the base observatory and filling in gaps with LEAR data.

The methodology in (a) is the generally used approach for combining group data from all the SOON stations to generate a single file (see, e.g., Balmaceda et al., 2009). In strategy (b), the combined measurement corresponds to the largest reported area on a day and it gives an upper limit in the difference between two observatories. The options (c) and (d) show the effect of missing spots in the final composite.

The differences between the daily P_S index using the composite file created using the four methodologies described above and the original HOLL and LEAR datasets averaged over 12 months using running means are displayed in Figure 4. The four different composites are shown in the respective panels (a)–(d). For comparison, the daily values from strategies (c) and (d) are shown in panels (e) and (f), respectively.

Naturally, the largest differences between the P_S composite file and LEAR and HOLL come from using the maximum areas (b), with values $|\Delta \mathrm{P_S}| < 150$ ppm. When using the mean values (a), we find $|\Delta P_S| < 50$ ppm. In both cases, the differences cannot be attributed to the missing sunspot groups only, but also to the methodology used when combining the data when available for both HOLL and LEAR. The effect of missing groups is only revealed by analyzing the plots (c) and (d).

The comparison between the P_S index estimated from the composite using HOLL as base observatory and the original HOLL dataset reveals that most missing spots belong to the maximum of Solar Cycle 22, around 1990, accounting for differences of 20 ppm in the 12-month averages. Some individual groups can account for irradiance deficits of more than 500 ppm as can be seen in the lower panels where the daily differences are plotted without

Table 5 Mean values (and standard deviation) of P_S [W m^{-2}] at different levels of solar activity.

Cycle	HOLL	LEAR	COMPOSITE
Minimum ± 1 year			
21	–	–	–
22	–0.073 (0.130)	–0.059 (0.109)	–0.077 (0.135)
23	–0.049 (0.104)	–0.035 (0.078)	–0.051 (0.105)
24	–0.011 (0.042)	–0.010 (0.037)	–0.013 (0.047)
Maximum ± 1 year			
21	–0.901 (0.654)	–0.938 (0.592)	–1.005 (0.702)
22	–0.967 (0.725)	–0.939 (0.732)	–1.101 (0.810)
23	–0.708 (0.422)	–0.721 (0.478)	–0.808 (0.495)
24	–	–	–
Full cycle			
21	–0.496 (0.649)	–0.494 (0.619)	–0.549 (0.697)
22	–0.503 (0.605)	–0.470 (0.580)	–0.560 (0.663)
23	–0.346 (0.432)	–0.340 (0.437)	–0.390 (0.482)
24	–0.170 (0.233)	–0.196 (0.292)	–0.220 (0.306)

any averaging. The differences are also noticeable around the maxima of Cycles 23 and 24, around years 2001 and 2012, respectively. For the rest of the time, the values get close to 0, but not exactly, indicating that missing groups are distributed all over the period of observations.

A similar trend is observed when comparing the P_S index from the composite using LEAR as base with the original LEAR dataset. These differences are more pronounced around 1991.

In Figure 5 we show the histograms of the sunspot sizes, position, and the estimated irradiance deficit for the missing groups in LEAR and HOLL datasets. The distributions for both stations are very similar. It is clear that most of the identified missing groups (about 60%) correspond to the smallest areas, i.e., below 50 ppm. These groups are distributed at all longitudes on the solar disk with a large fraction (at least $\sim 50\%$) being located near the solar disk ($\mu > 0.6$). The estimated median irradiance deficit due to the missing groups is ~ -8 ppm of the S_Q (with the 5th and 95th percentile values between -1 and -160 ppm of the S_Q).

At different phases of the solar cycle or even at different cycles, these values can vary considerably.

In Table 5 we quantify these differences at different solar activity levels for the extreme case (i.e., for the composite using the maximum areas). We list the mean values and standard deviation in parenthesis. The values are calculated in the period of 2 years centered at each maximum and minimum and the full solar cycle. For Solar Cycle 21, we take the first two years to characterize the maximum while the minimum is not included because it is prior to 1982.

Differences between the means of the P_S index using HOLL and the composite dataset are statistically significant during maxima, while we cannot conclude anything for the minima. The differences between the means of the P_S index using LEAR and the composite time series, on the other hand, are statistically significant for Solar Cycles 22 and 23 minima and maxima.

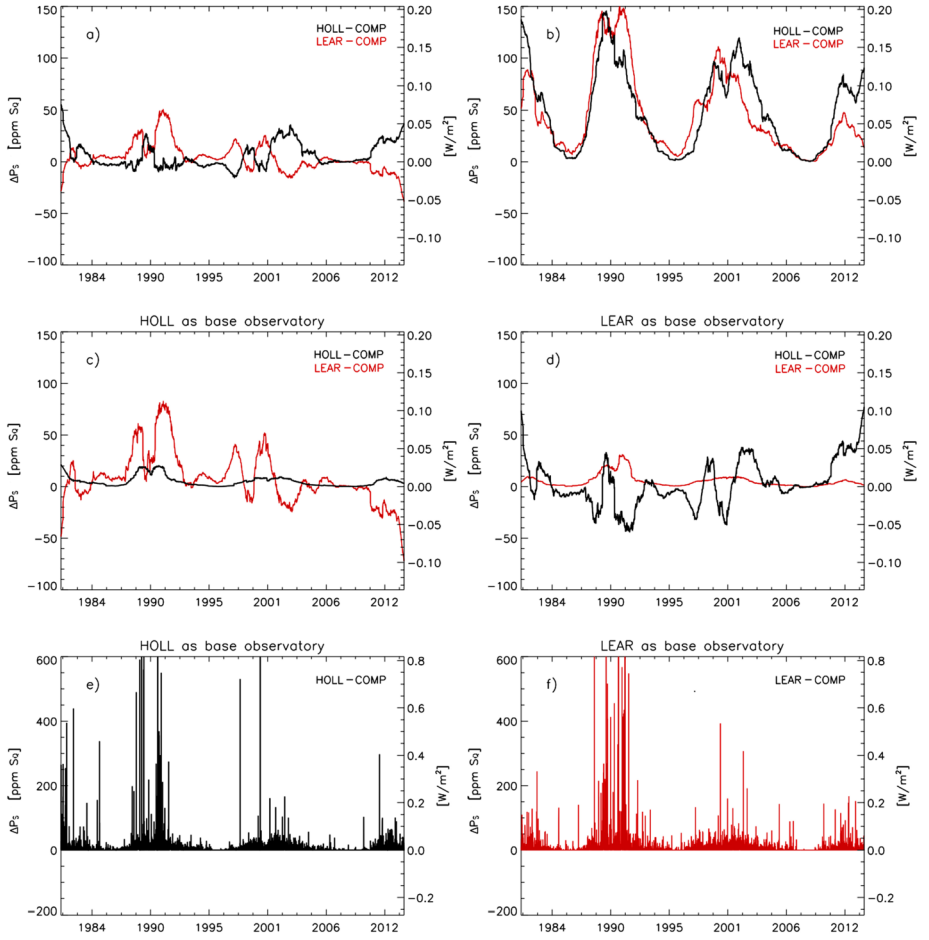

Figure 4 (**a**)–(**d**) 12-month running means of the ΔP_S using four strategies a–d (see Section 4.2), respectively, to build the composite file; (**e**)–(**f**) daily values of the difference between the PSI index using HOLL and LEAR as base observatory and the original file.

In Figure 6 we show the difference between the P_S index from the composite COMP and HOLL and LEAR datasets, respectively, per bins of the heliographic distance μ. We analyze the differences close to the limb ($\mu < 0.33$), at intermediate regions ($0.33 < \mu < 0.66$) and close to the disk center ($\mu > 0.66$). The differences increase considerably for sunspot areas near the disk center.

5. Discussion

The observatories from the SOON network were initially planned to provide real-time data. However, because of their long-term success, their products are still widely used in a large number of studies (see, e.g., Krivova, Balmaceda, and Solanki, 2007; Bhowmik and Nandy, 2018). In spite of their importance, however, the daily areas from the SOON network differ considerably from the values reported by other observatories. For example, Balmaceda et al.

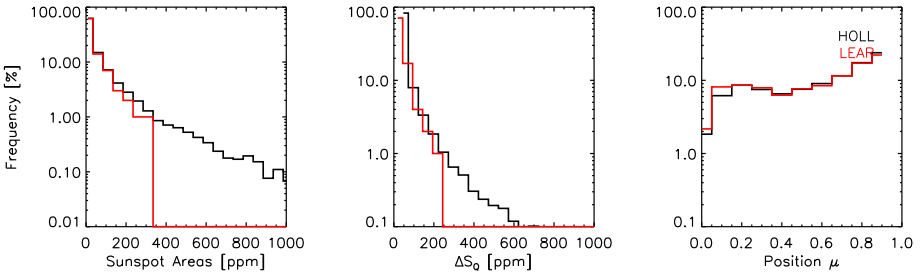

Figure 5 Histograms of the sunspot group areas (left), estimated irradiance deficit (center), and position (right) for the missing groups in HOLL (black) and LEAR (red) stations.

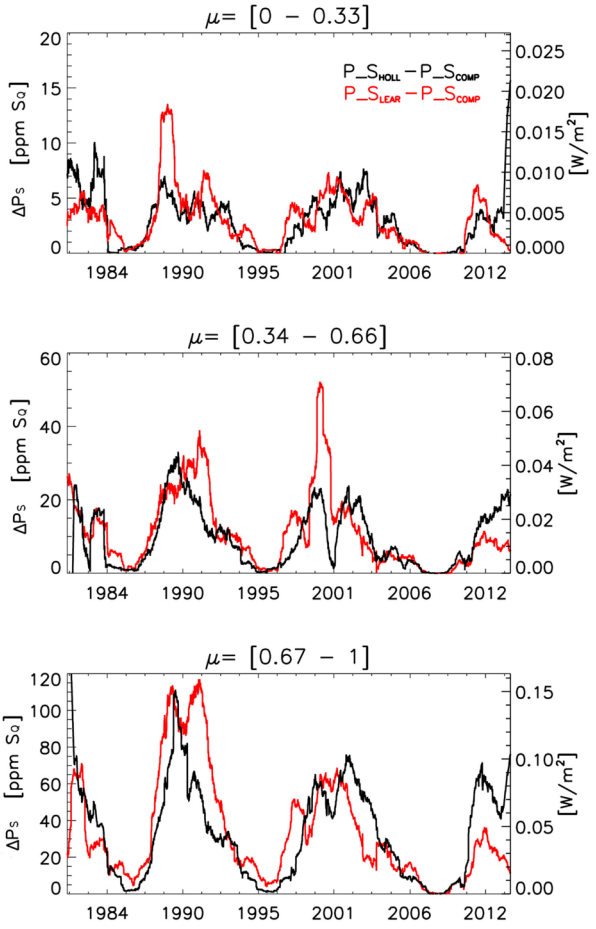

Figure 6 12-month running means of the differences between the P_S index calculated with the composite file and HOLL (black) and LEAR (red), respectively. Each panel shows the values corresponding to three ranges of positions in the disk.

(2009) and Hathaway (2012) pointed out that the daily values of sunspot areas from the SOON data should be corrected by a factor of 1.4 in order to compensate for the differences with the RGO dataset. The magnitude of this correction factor was questioned in more recent works (see, e.g., Foukal, 2014; Muñoz-Jaramillo et al., 2015a).

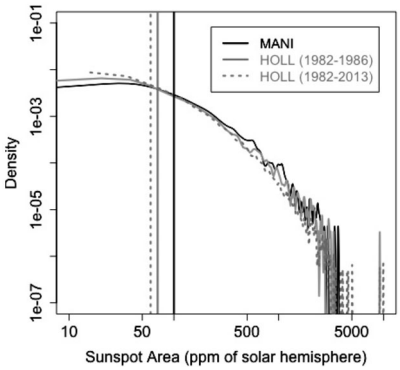

Figure 7 Comparison of sunspot area distribution functions between MANI and HOLL, for the same period (solid line) and for the entire period of HOLL (dotted line).

In this work, we investigated the differences between the data from the stations in the SOON network. We started with the hypothesis that if all the stations use similar equipment and apply the same measuring techniques, one can quantify the differences arising from the observers (bias). Therefore, we would not expect significant differences when comparing individual group areas. The comparison of daily values, on the other hand, allowed us to determine if there also exist differences due to the missing groups in the record, which can be attributed to the varying seeing conditions at the observing site.

To determine the calibration or correction factors, we used the same methodology as in Balmaceda et al. (2009) (recently also followed by Mandal et al., 2020) but applied it to both the daily values per station and the individual group values. We further improved the method by using *bootstrapping* technique to determine 95% confidence intervals for each comparison.

As mentioned above, from the comparison between individual group areas, we did not expect calibration factors significantly different from unity. However, we found that the differences between the calibration factors for group and daily areas show differences of the order of 4%. This suggests that a large fraction of the differences in the daily areas are due to the bias from the observer. In the case of SOON stations, the same correction factors can be used for both group and daily datasets since their respective correction factors agree within the 95% CI, with the exception of MANI, where the confidence intervals for the calibration factors are too wide. This might be in part due to the small size of the sample, but also to the large level of variability in the data. The areas from this dataset are not reliable and should not be included when compiling a combined dataset for the SOON network. To analyze more in detail these issues, we compare the distribution of sunspot areas from MANI (black solid line) and HOLL (gray lines) in Figure 7. For the latter, we consider two time intervals: the same interval as covered by MANI (1982–1986, gray solid line) and the full period covered by HOLL (1982–2013, gray dashed line). The vertical lines indicate the median values for each dataset. From this comparison, we can see that MANI reports less sunspot areas with sizes below 100 ppm and shows an excess of sunspots with larger areas when compared with HOLL in the same period. The median value for MANI and HOLL in this time interval are 100 and 70 ppm, respectively. The observed variability of MANI can be attributed to the following reasons:

(a) There is a large number of missing data in MANI records (36% of sunspot groups are not reported);
(b) The short period of observations of MANI corresponds to the descending phase of Cycle 21, while other observatories cover much longer periods; and

(c) The high median value in MANI sunspot areas is also representative of the stronger Cycle 21 compared to more recent cycles. This is confirmed by the median value in HOLL records that is also larger when the short period is considered (70 ppm in the period covered by MANI vs. 60 ppm for the full period of observations).

Our results are in good agreement with those by Giersch, Kennewell, and Lynch (2018). They estimated calibration factors between 0.95 and 1.05 when comparing daily values from SVTO, LEAR, HOLL, RAMY, and PALE stations. However, they did not include CULG, BOUL, and MANI in their analysis. In this work, we find that these differ significantly from the five stations in Giersch, Kennewell, and Lynch (2018). The differences can be as large as 15–16%. We suggest that only those observatories with no significant differences between them should be used when compiling a single dataset instead of averaging all available measurements as usually done (i.e., Balmaceda et al., 2009).

Finally, we used the SOON dataset to estimate the irradiance deficit due to the sunspot passage via the P_S index (Hudson et al., 1982). This index depends on the size and location of the sunspot groups present on the solar disk. In order to quantify the effect of the missing groups on the calculation of the P_S index, we compared the time series using HOLL, LEAR, and a composite dataset from these two stations. We found that the differences can be as high as 150 ppm during the maximum of solar cycle. We also found that the differences increase when considering the contribution of sunspots near the center of the disk. Missing sunspots near the center can account for about 80% of the observed differences.

6. Conclusions

In this section, we summarize the main results of our analysis.

First, we performed a simple statistical analysis to summarize the characteristics of the observations from each station. The main results are:

1. We estimated that about 7% of sunspot groups are missing in the SOON records (not reported by any of the stations); 4% of the groups are reported only one day and 2% of the groups appear only once in the full SOON dataset (i.e., reported by a single stations and only one day).
2. The stations with the best coverage of both sunspot groups and daily sums are LEAR, SVTO, RAMY, and HOLL. LEAR and HOLL also cover the whole period analyzed in this work. SVTO and RAMY cover at least three solar cycles.
3. HOLL and LEAR are also the stations with best averaged quality of observations.
4. In all the stations, a systematic deficiency of measurements is evidenced close to the limb. Near the disk center (at $\mu > 0.7$) the fraction of reported sunspot groups is $> 60\%$, significantly dropping to 5% at $\mu < 0.3$.
5. MANI can be regarded as the least reliable dataset. This station presents the poorest coverage in both group and daily observations, the shortest period of observations, and the lowest quality of observations.

The main results from the comparisons between pairs of stations are presented below:

1. We did not find significant differences in the calibration factors for a given pair of observatories when comparing group or daily sunspot areas. We would expect much lower differences between individual group areas than in the total area per day. Observatories using similar instrumental and measurement techniques, as in the case for the SOON stations, would provide similar values for the areas. For daily areas, on the other hand, the

difference should be larger due to the missing groups in the dataset. This is the case only for the comparisons with MANI.
2. The differences in the group areas vary from 1% to 16%, with the largest from the comparisons between CULG and BOUL with HOLL and SVTO. The comparison of HOLL, LEAR, and SVTO, among themselves, gives small differences in the group areas, less than 5% with a tight CI around $\beta = 1$.
3. Also for HOLL, LEAR, and SVTO, the differences in daily values give similar values and show the same trends. This suggests that the data from these three observatories can be combined into a single series without further corrections.
4. The stations with the shortest observation periods (MANI, CULG, PALE, and BOUL) provide measurements that differ largely (up to 18% in the daily values) with respect of the rest of the observatories.

A single correction factor can be used to multiply the sunspot areas from one observatory to get it to the same level of another. In this way, one can compensate for systematic errors. This kind of correction can be useful when daily values are needed, for instance, irradiance models such as SATIRE (Balmaceda, Krivova, and Solanki, 2007; Krivova, Balmaceda, and Solanki, 2007). However, for studies relying on the individual group sunspot areas, such as flux transport models (e.g., Bhowmik and Nandy, 2018), comparison of distribution functions of the size of sunspots (Muñoz-Jaramillo et al., 2015b), or estimate of irradiance deficit due to sunspots via the P_S index, this might not be correct. When estimating the P_S index, a single correction factor seems inadequate because the contribution of sunspot areas to the total irradiance variations depends on their position on the solar disk. For this particular purpose both the size and positions of sunspot groups are needed. For instance, Baranyi (2018) compared multiple datasets with Debrecen Photoheliographic Data (DPD). They found that SOON sunspot areas need a multivariate correction factor to set them on the scale of DPD. The correction factor, ranging between 1.1 and 1.9, depends on time, on the distance of the group from the disk center, and on the specific SOON station.

Acknowledgements The authors acknowledge the National Centers for Environmental Information from the National Oceanic and Atmospheric Administration which provides the data used in this work. L.L. is a post-doctoral fellow from CONICET, Argentina. The authors would also like to thank the referee for the valuable comments which helped improve the manuscript.

Declarations

Disclosure of Potential Conflicts of Interest The authors declare that they have no conflicts of interest.

References

Akritas, M.G., Bershady, M.A.: 1996, Linear regression for astronomical data with measurement errors and intrinsic scatter. *Astrophys. J.* **470**, 706. DOI. ADS.

Balmaceda, L., Krivova, N.A., Solanki, S.K.: 2007, Reconstruction of solar irradiance using the Group sunspot number. *Adv. Space Res.* **40**, 986. DOI. ADS.

Balmaceda, L.A., Solanki, S.K., Krivova, N.A., Foster, S.: 2009, A homogeneous database of sunspot areas covering more than 130 years. *J. Geophys. Res.* **114**, A07104. DOI. ADS.

Baranyi, T.: 2018, Stable sunspot area level of Debrecen photoheliographic data and multivariate correction factor of SOON data. *Solar Phys.* **293**, 142. DOI. ADS.

Bhowmik, P., Nandy, D.: 2018, Prediction of the strength and timing of sunspot cycle 25 reveal decadal-scale space environmental conditions. *Nat. Commun.* **9**, 5209. DOI. ADS.

Brandt, P.N., Stix, M., Weinhardt, H.: 1994, Modelling solar irradiance variations with an area dependent photometric sunspot index. *Solar Phys.* **152**, 119. DOI. ADS.

Canty, A., Ripley, B.D.: 2017, boot: Bootstrap r (s-plus) functions. R package version 1.3-20.
Davison, A.C., Hinkley, D.V.: 1997, *Bootstrap Methods and Their Applications*, Cambridge University Press, Cambridge ISBN 0-521-57391-2. http://statwww.epfl.ch/davison/BMA/.
Eisenhauer, J.G.: 2003, Regression through the origin. *Teach. Stat.* **25**, 76. DOI.
Fligge, M., Solanki, S.K.: 1997, Inter-cycle variations of solar irradiance: Sunspot areas as a pointer. *Solar Phys.* **173**, 427. DOI. ADS.
Foukal, P.: 2014, An explanation of the differences between the sunspot area scales of the Royal Greenwich and Mt. Wilson Observatories, and the SOON Program. *Solar Phys.* **289**, 1517. DOI. ADS.
Giersch, O., Kennewell, J., Lynch, M.: 2018, Reanalysis of solar observing optical network sunspot areas. *Solar Phys.* **293**, 138. DOI. ADS.
Hathaway, D.H.: 2010, The solar cycle. *Living Rev. Solar Phys.* **7**, 1. DOI. ADS.
Hathaway, D.H.: 2012, A statistical test of uniformity in solar cycle indices. In: *American Astronomical Society Meeting Abstracts #220*, *American Astronomical Society Meeting Abstracts* **220**, 206.01. ADS.
Hudson, H.S., Silva, S., Woodard, M., Willson, R.C.: 1982, The effects of sunspots on solar irradiance. *Solar Phys.* **76**, 211. DOI. ADS.
Isobe, T., Feigelson, E.D., Akritas, M.G., Babu, G.J.: 1990, Linear regression in astronomy. *Astrophys. J.* **364**, 104. DOI. ADS.
Kopp, G., Lawrence, G., Rottman, G.: 2005, The Total Irradiance Monitor (TIM): Science results. *Solar Phys.* **230**, 129. DOI. ADS.
Kopp, G., Lean, J.L.: 2011, A new, lower value of total solar irradiance: Evidence and climate significance. *Geophys. Res. Lett.* **38**, L01706. DOI. ADS.
Krivova, N.A., Balmaceda, L., Solanki, S.K.: 2007, Reconstruction of solar total irradiance since 1700 from the surface magnetic flux. *Astron. Astrophys.* **467**, 335. DOI. ADS.
Mandal, S., Krivova, N.A., Solanki, S.K., Sinha, N., Banerjee, D.: 2020, Sunspot area catalog revisited: Daily cross-calibrated areas since 1874. *Astron. Astrophys.* **640**, A78. DOI. ADS.
Muñoz-Jaramillo, A., Senkpeil, R.R., Longcope, D.W., Tlatov, A.G., Pevtsov, A.A., Balmaceda, L.A., DeLuca, E.E., Martens, P.C.H.: 2015a, The minimum of solar cycle 23: As deep as it could be? *Astrophys. J.* **804**, 68. DOI. ADS.
Muñoz-Jaramillo, A., Senkpeil, R.R., Windmueller, J.C., Amouzou, E.C., Longcope, D.W., Tlatov, A.G., Nagovitsyn, Y.A., Pevtsov, A.A., Chapman, G.A., Cookson, A.M., Yeates, A.R., Watson, F.T., Balmaceda, L.A., DeLuca, E.E., Martens, P.C.H.: 2015b, Small-scale and global dynamos and the area and flux distributions of active regions, sunspot groups, and sunspots: A multi-database study. *Astrophys. J.* **800**, 48. DOI. ADS.
R Core Team: 2017, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.
Wickham, H.: 2011, The split-apply-combine strategy for data analysis. *J. Stat. Softw.* **40**, 1. http://www.jstatsoft.org/v40/i01/.
Wickham, H.: 2007, Reshaping data with the reshape package. *J. Stat. Softw.* **21**, 1. http://www.jstatsoft.org/v21/i12/paper.
Wickham, H., Henry, L.: 2018, tidyr: Easily tidy data with 'spread()' and 'gather()' functions. R package version 0.8.1. https://CRAN.R-project.org/package=tidyr.