Making of a solar spectral irradiance dataset I: observations, uncertainties, and methods

Micha Schöll, Thierry Dudok de Wit, Matthieu Kretzschmar, Margit Haberreiter

J. Space Weather Space Clim., **6**, A14 (2016)
DOI: 10.1051/swsc/2016007

SWSC

TECHNICAL ARTICLE OPEN ACCESS

Making of a solar spectral irradiance dataset I: observations, uncertainties, and methods

Micha Schöll[1,a,*], Thierry Dudok de Wit[1,a], Matthieu Kretzschmar[1,a], and Margit Haberreiter[2]

[1] LPC2E/CNRS & Université d'Orléans, 3A Av. de la Recherche Scientifique, 45000 Orléans, France
[a] Currently at Physikalisch-Meteorologisches Observatorium and World Radiation Center.
[*]Corresponding author: mschoell@gmx.com
[2] Physikalisch-Meteorologisches Observatorium and World Radiation Center, Dorfstrasse 33, CH-7260 Davos Dorf, Switzerland

Received 16 February 2015 / Accepted 10 January 2016

ABSTRACT

Context. Changes in the spectral solar irradiance (SSI) are a key driver of the variability of the Earth's environment, strongly affecting the upper atmosphere, but also impacting climate. However, its measurements have been sparse and of different quality. The "First European Comprehensive Solar Irradiance Data Exploitation project" (SOLID) aims at merging the complete set of European irradiance data, complemented by archive data that include data from non-European missions.
Aims. As part of SOLID, we present all available space-based SSI measurements, reference spectra, and relevant proxies in a unified format with regular temporal re-gridding, interpolation, gap-filling as well as associated uncertainty estimations.
Methods. We apply a coherent methodology to all available SSI datasets. Our pipeline approach consists of the pre-processing of the data, the interpolation of missing data by utilizing the spectral coherency of SSI, the temporal re-gridding of the data, an instrumental outlier detection routine, and a proxy-based interpolation for missing and flagged values. In particular, to detect instrumental outliers, we combine an autoregressive model with proxy data. We independently estimate the precision and stability of each individual dataset and flag all changes due to processing in an accompanying quality mask.
Results. We present a unified database of solar activity records with accompanying meta-data and uncertainties.
Conclusions. This dataset can be used for further investigations of the long-term trend of solar activity and the construction of a homogeneous SSI record.

Key words. Solar spectrum – Data analysis – Statistics and probability – Time series analysis – Algorithms

1. Introduction

Changes in the spectral solar irradiance (SSI) strongly affect the upper atmosphere, but also impact climate. In particular, a consistent dataset of SSI changes serves as an input for climate models (Lean et al. 1995; Ermolli et al. 2013; Thuillier et al. 2014) as well as a central parameter for space weather predictions. These models are in need of datasets spanning several decades of continuous and radiometrically accurate measurements. However, satellites measuring SSI usually have a life span of a few years up to a decade, before instrument degradation, equipment failure, or end of mission financing stops the data flow. Not a single instrument has measured SSI for two or more solar cycles, and even for those instruments that have measured an entire solar cycle, e.g. SUSIM,[1] instrument stability is not sufficient to properly assess the existence of long-term trends.

To merge SSI observations and reconstruct missing observations, we need models. These models are often driven by solar proxies. Frequent choices include the sunspot number, which has been measured since the beginning of the 17th century, solar radio flux measurements, with a popular example being the 10.7 cm radio flux available since 1947, and spectral ratios such as the MgII core-to-wing ratio (Viereck et al. 2004). There are two kinds of models: empirical ones, where one or more solar proxies are directly fitted to data using some criteria of best-fit, and semi-empirical ones, which involve physical modeling of some solar phenomena such as surface flux transport of the solar photospheric magnetic field. In this latter case, the proxy data are coupled to some underlying physical quantity. For example, sunspot area is used to infer magnetic field strength, which in turn is used to estimate the SSI (Domingo et al. 2009; Ermolli et al. 2013).

Both empirical (e.g. Tobiska 2004) and semi-empirical (e.g. Yeo et al. 2014) designs have been used for various SSI models. Proxy-based models can be very accurate for the calibration time period, but they are usually calibrated and compared to a specific instrument. For example, the semi-empirical NRLSSI (Lean 2000) model is calibrated against UARS/SOLSTICE. Relying on a single instrument makes it difficult to distinguish between the solar signal and instrumental artifacts; it also limits the long-term temporal calibration to two solar minima.

A homogeneous uncertainty estimation is essential for the delivery of a single composite. Such a single composite is one of the explicit goals of the SOLID project. In Bayesian statistics, the influence of the data is proportional to the uncertainty of the data. As such, it is important that the same methods are used for each instrument to determine the uncertainty. However, even though data providers usually present the accompanying uncertainties, their uncertainty budget often accounts for different sources of uncertainty. These include the standard deviation over short (sub-daily) time spans, modeled instrument response functions, aspects of manufacturing, and

[1] The expansion of acronyms used in this work is given in Appendix A.

calibration constants. All these sources determine the final uncertainty budget. Ideally, proxy-based models would be simultaneously calibrated to different instruments covering several time spans. Yet, there are many obstacles that have prevented scientists from doing so, including instrumental artifacts, uncertainty of the instrument stability, and the offset to the true zero. Addressing these challenges requires a common estimate of uncertainties, and a common data format that would allow for a combination or comparison of different instrumental data would certainly be helpful.

DeLand & Cebula (2008) already provide an SSI composite of the ultraviolet (UV) based on six instruments, specifically the SBUV instrument on Nimbus-7, SBUV/2 on NOAA-9 and NOAA-11, SME on board SOHO, and the SUSIM and SOLSTICE instruments on UARS, together with synthetic proxy-based models in the case of missing instrumental data. Out of these instruments they selected one specific instrument for different temporal (November 1978 to August 2005) and spectral (120–400 nm) regions. The uncertainties were directly derived from the original data. However, the instruments chosen limited the temporal and spectral ranges, and, due to the methods of combining the data, instrumental artifacts were not accounted for in the composite.

To obtain homogeneous SSI datasets, our approach is thus as follows. We develop a coherent methodology to estimate the precision and stability for all datasets, and we evaluate the uncertainties induced by each method. We rescale all data into a common time grid with daily resolution and remove non-physical values. We take advantage of spectral coherency to estimate missing data in one wavelength band from a combination of other bands that display a similar temporal variability. We detect instrumental artifacts by taking advantage of an autoregressive model and combine the model with proxies as a qualitative indicator. Finally, we use proxy interpolation to handle data gaps.

Our final product consists of three datasets per instrument: first, the original, unified dataset converted to common units including independent precision estimates; second, a homogenized and stratified "close-to-original" dataset that does not change any data except for re-gridding in time, accompanied by independent uncertainty estimates and quality flags; and third, a "best-estimate" dataset where missing data and identified outliers have been replaced by values obtained from proxy-based models. This systematic approach to attain homogeneous SSI datasets in a unified format is part of the *First European Comprehensive SOLar Irradiance Data Exploitation project* (SOLID, http://projects.pmodwrc.ch/solid), whose objective is to attain a long-term SSI composite of daily resolution covering the satellite space-age era by utilizing all available data. This can be used by modelers of the Earth atmosphere and climate, researchers in stellar physics, planetary science or astrobiology, and as a cross-comparison for other models and instruments.

Subsequently, Section 2 provides, after a short historic overview of SSI measurements, a description of all data incorporated in SOLID. Section 3 presents our uncertainty estimations. Section 4 details our pipeline approach, consisting of the pre-treatment of the data (Sect. 4.1), the interpolation of missing data (Sect. 4.2), the re-gridding of the data (Sect. 4.3), the outlier detection (Sect. 4.4), and the proxy-based interpolation (Sect. 4.5). Section 5 specifies the content, availability, and the future of the SOLID database, while Section 6 concludes our paper. For better readability, we include most of our tables in Appendix A, followed by a formal description of the data format in Appendix B and our algorithms in pseudo-code in Appendix C.

2. Data

Scientists have long struggled with making long-term SSI observations. As Schmidtke (2014) describes in his historical overview of SSI measurements, the first attempts at obtaining SSI were not suitable for long-term recording of the variable solar spectrum. While E. Regener and V. H. Regener, in 1932, used balloons flying up to an altitude of 33 km to obtain a spectrum of the visible to the far ultraviolet (UVC, 200 nm–280 nm) region, Baum et al. (1946, cited by Schmidtke) used V2 rockets to obtain the first SSI measurements in the extreme ultraviolet (EUV) band. Though innovative, their methods, based on photographic film, were restricted by the necessity of instrumental retrieval. Therefore, they represented only a snapshot of solar activity. Del Zanna & Andretta (2011) collected an exhaustive list of rocket flights, measuring the EUV. It was the development of photoelectric diodes in the 1960s that allowed for satellite-based long-term measurements. Using these diodes, the EUVS experiment aboard the AE-E satellite showed, for the first time in history, the variability of the EUV from 20 to 185 nm over the course of 3 years from a solar minimum in 1976 to a solar maximum in 1979 (Hinteregger 1981). This was the beginning of long-term SSI measurements. The first instrument observing a full solar cycle was UARS/SUSIM (Upper atmosphere research satellite/Solar Ultraviolet Spectral Irradiance Monitor), 1991–2005 (Rottman et al. 1993).

Starting with data from the AEE/EUVS experiment in 1967, we have collected and processed data from 26 instruments over a wavelength range from 4 nm to 2.4 μm. Our space-based observations cover the time range of 1977–2015. We also provide nine reference spectra, with our first spectrum by Arvesen et al. (1969) and our latest one by Woods et al. (2009), as well as 14 proxies of solar activity. We utilize these proxies to enhance instrumental data by detection of possible outliers (Sect. 4.4) and by interpolation of missing data (Sect. 4.5). Figure 1 presents an overview of all instruments, reference spectra, and proxies available in the SOLID database. Not all existing data have been included, either due to bad quality or missing availability. Notably, the Airglow Solar Spectrometer Instrument (ASSI) data from the San Marco mission and the EUVS from the Orbiting Solar Observatory (OSO) III mission are not included. Other data, such as the broadband filter radiometer LYRA on PROBA2, will be used to verify the composite.

To establish a common data format, we follow the Net Common Data Format (NetCDF) Climate Forecast standard, version 1.6,[2] which provides standardized variable names, well-defined field names for global meta-information such as author, data source, and creation date, meta-information for each variable such as units, flag description for masks, and cross-references to other variables.

In Appendix A, we present an overview of all the data together with references, sources, and key information such as time spectral ranges below in tabular form. Table A.1 lists all instruments included in the SOLID database, Table A.3 contains all nine reference spectra, Table A.4 displays all proxies, and Table A.5 presents the three composite time series in

[2] http://cfconventions.org/

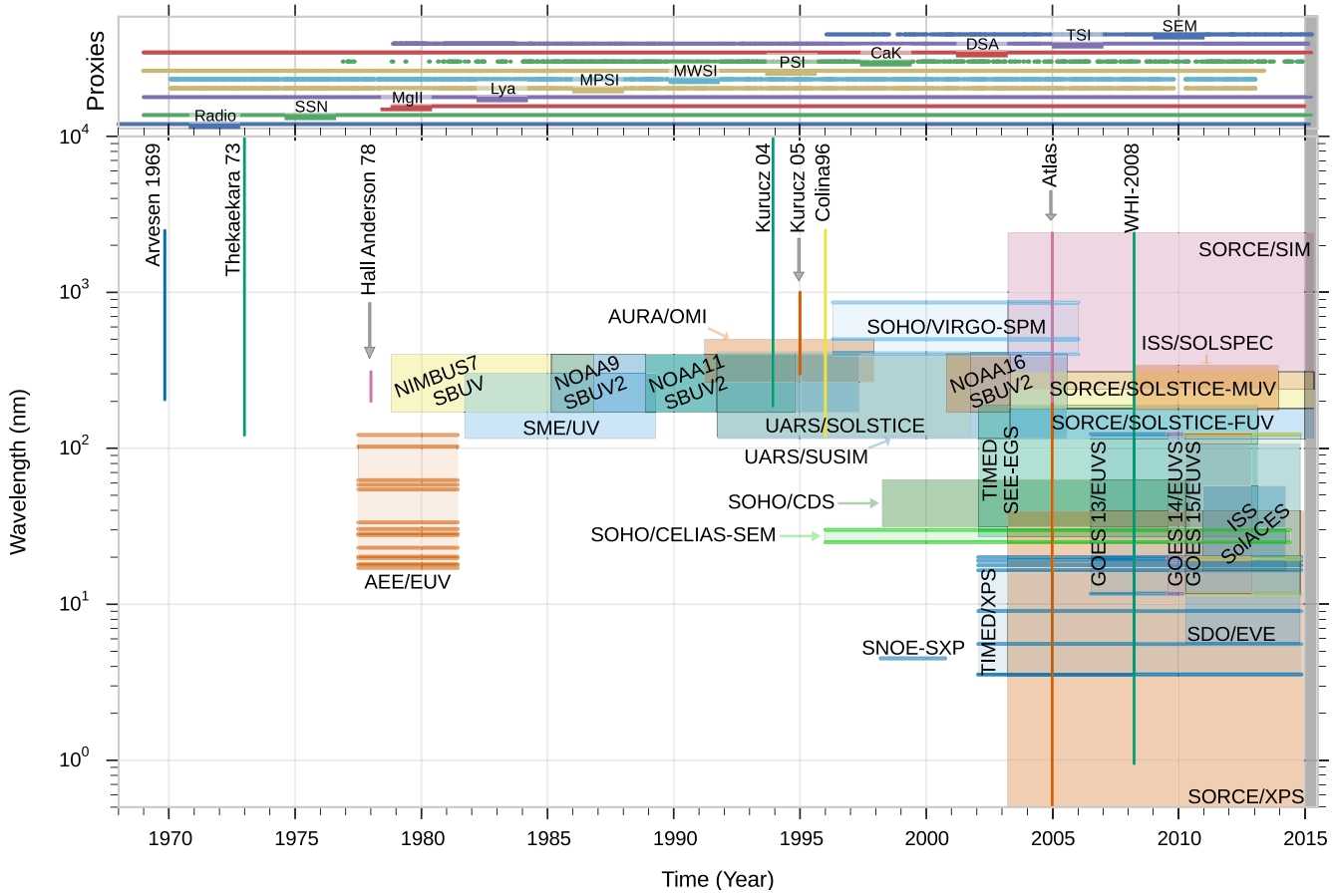

Fig. 1. A graphical summary of Table A.1 (Instruments), A.3 (References), A.4 (Proxies), and A.5 (Composite datasets) representing almost all the datasets incorporated in SOLID versus time and wavelength. The upper panel presents all available proxies, together with two spectral composites. The lower blue line, labeled Radio, represents the 3.2, 8, 10.7, 15, and 30 cm radio fluxes from the Nobeyama and Penticton radio-observatories. The time range shown corresponds to the available times in the SOLID database, which truncates all proxies but the radio fluxes at 1969. The lower panel presents the temporal and spectral ranges for all instruments in the database. Vertical lines represent the time point and spectral range of the reference spectra. Instruments with fewer than 20 channels are shown as light boxes with a horizontal line at the wavelength of the central frequency of each channel. While the 21st century is well covered both in temporal and spectral dimensions, previous data are only available for selected wavelengths. Only the UV between 170 nm and 400 nm has been covered continuously since 8 November 1978.

the database. Two of the composite time series, the Woods et al. (2000) Lyman α composite and the Fröhlich (2006) TSI composite, are also used as proxies for SSI. The third composite, the aforementioned DeLand SSI composite has been included in the database, but will not be used for the construction of our composite. Instead we will use the original instrumental data. Altogether, we currently have 53 datasets in the database. Any new dataset incorporated into SOLID will also be made available at the SOLID homepage.

3. Uncertainty estimation

As discussed above, the definition of uncertainties and the inclusion of different uncertainty sources differ for each instrument. Hence, it is not surprising that the final uncertainty estimates vary considerably between instruments. A particularly pronounced example are the measurements of total solar irradiance, as shown in Figure 2, where uncertainties vary over three orders of magnitude and the highest uncertainties are given for the first fully-calibrated instrument, TIM, due to the inclusion of accuracy in its uncertainty estimates. In conclusion, any meaningful interinstrument comparison of uncertainties must take into account their sources and definitions.

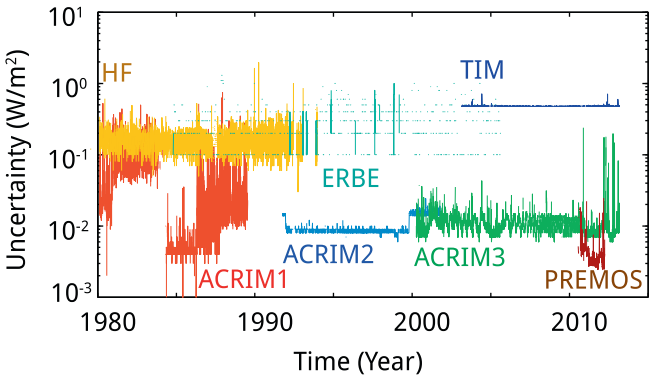

Fig. 2. Instrument uncertainties for different TSI instruments. They differ by up to three orders of magnitude with the highest uncertainties for a modern instrument, TIM. This is due to different definitions used for what an instrumental uncertainty is. For that reason, these values cannot be meaningfully compared.

A few words on terminology are necessary first. While precision corresponds to what is commonly known as the random error or noise, stability and accuracy both make up the systematic error. Stability, here, describes a time-dependent estimate

Table 1. Different sources and properties of uncertainties, together with synonyms and their inverses. The unit of the uncertainties depends on the unit of the data, which is abbreviated as u_d. Inverse is used in the sense that the larger the value of the uncertainty, the lower the value of the inverse. We also indicate whether we present our own estimation of the uncertainty and whether the uncertainty is time dependent. Time-dependent uncertainties are provided with the same temporal resolution as the data.

Name	Synonyms/*inverses*	Unit	Estimated in SOLID	Time-dependent
Precision	Repeatability/*Noise, Random Error*	u_d	Yes	Yes
Stability	Long-term Stability/*Drift, Shift*	u_d/yr	Yes	Yes
Accuracy	Exactness/*Bias, Offset, Systematic Error*	u_d	No	No

of the instrumental drift, in other words, a trend; accuracy stands for the offset from the "true zero" at the beginning of the mission, i.e. the bias of the absolute value. As the same concept often comes with different names, we provide our own naming convention in Table 1.

Since uncertainty estimates given by data providers vary enormously in definition and quality, it is necessary to provide an independent and homogeneous uncertainty estimate, which can be meaningfully compared between instruments. We provide two such estimates for precision and stability. There is no way to independently assess accuracy without resorting to external data, except as to refer to the absolute ground and on-board calibration. Therefore, only instrument teams can provide these estimates and we do not attempt to estimate them independently. If they are provided, we include them in the SOLID database.

As discussed above, uncertainty estimates provided by the PI (principal investigator) cannot be directly compared with each other. However, they do contain valuable information, such as satellite off-pointing or different integration time, and as such, they are also provided. Each PI presents these estimates differently: They can be provided as absolute or relative values for each data-point, as time- or wavelength-independent fractional values or percentages or as a global estimate. We considered two possibilities: either the uncertainty estimates are provided for each data-point of the time series, in which case we convert them (if needed) into absolute values, or the provided uncertainties are time- or wavelengths-independent, in which case they are expressed in fraction of the data. In the latter case, we did not multiply the uncertainty by each data-point (which would have provided an uncertainty estimate in the same format as in the first case) to reduce the file size. This results in the somewhat cumbersome situation that the uncertainties may differ in units and dimensions from one dataset to the other: in some cases, they have the same units as the timeseries, in other cases, they are unit-less fractions. The correct unit is always given in the corresponding field attribute.

As for precision, we provide both the original precision estimates given by data providers and our own estimate, which we describe here. Most data display either Poisson or Gaussian noise statistics or a mix of the two. Noise is generally estimated from the high-frequency component of the data, where its impact is often strongest. As a consequence, estimators usually involve some time differencing, followed by a measurement of the dispersion, for example by taking the standard deviation. Today, the wavelet methods (e.g. Mallat 2009) are among the most powerful means for estimating the noise level. The general idea, which was formalized by Donoho & Johnstone (1995), is to decompose the time series into different time scales by means of a discrete wavelet transform, then to consider wavelet coefficients w_i that correspond to the smallest time scale only, and finally to use median $(|w_i(x)|)/0.6748$ as an estimate of the noise level. The use of the median allows one to exclude the few unusually large wavelet coefficients that may be associated with sudden transients, such as sunspot darkenings, and to focus only on the bulk of the fluctuations. The 0.6748 correction factor allows this quantity to be equated with the standard deviation in the case of additive white noise. For details regarding the wavelet decomposition we refer to Donoho & Johnstone (1995). The choice of the mother wavelet, i.e. the unscaled wavelet, is not critical here, as long as its regularity is sufficient. In our case, the 4th order Daubechies wavelet gives satisfactory performance.

While this estimator works well for quantifying random errors stemming from counting processes (i.e. with a Poisson distribution) and, more generally, for white noise, it does not provide a good estimate of the colored noise level, whose power spectral density is not flat. Noise color is defined by the power β of the spectral density per unit frequency. White noise is defined by a flat spectral density ($\beta = 0$), while integrated white noise, also known as Brownian noise, corresponds to $\beta = -2$ (Pink noise corresponds to $\beta = -1$). On the other side of the spectrum we have differentiated white noise, namely violet noise, that is defined by $\beta = 2$ and describes noise that varies more strongly in the short term. Blue noise is defined as $\beta = 1$.

Using this definition of noise, we modify the Donoho noise estimator to take into account all the wavelet coefficients (named Donoho-FULL). This underestimates blue and violet noise levels. Our solution is to adapt the Donoho noise estimator by combining the Donoho and the Donoho-FULL estimator using a weighted average, yielding a good estimate of white, blue, and violet noise, as shown in Figure 3. The weights are determined empirically such that they provide a best fit over $\beta \in [0, 2]$. This does not only estimate the white noise level correctly, but also blue and violet noise levels which may be present in the case of a positive feedback loop.

However, we are not aware of a reliable method to determine Brownian noise level without resorting to external data. As such, we cannot take into account possible Brownian contribution, despite the fact that most observations are likely to have a negative β. Therefore, for consistency, and although we know that this is not optimal, we assume our data to be free from Brownian noise and use our adapted Donoho estimator as a systematic means for comparison. Our stability estimator does use external data in the form of proxies. Hence, the red noise can be estimated as part of our stability estimate.

Another important limitation of our precision estimator is that its upper limit equals the variability of the data. In other words, our signal-to-noise ratio never gets below 1, as shown in the lower panels of Figure 4 in the visible wavelength range.

Thus, our precision estimation consists of this adapted Donoho estimator, calculated with a running window of a maximum length of 100 days, and the uncertainties induced by interpolation (Sects. 4.2 and 4.3) and the outlier detection

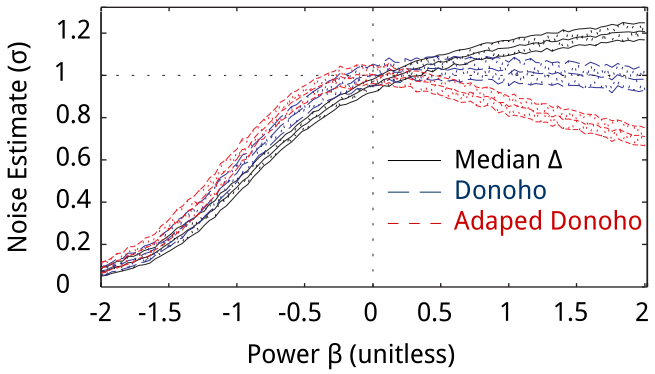

Fig. 3. Comparison of the three presented noise estimators for different colors of noise. We show the original Donoho noise estimator (solid black), a wavelet estimator that takes into account all wavelet scales (Donoho-FULL, short-dashed red), and our adapted Donoho estimator (long-dashed blue). The β value, representing the power of the spectral density, is varied from -2 (Brownian noise) to 2, where 2 is violet noise (1 would be blue noise). Each point represents the estimated noise level of a one thousand point randomly generated time series of the specified power. The spread indicates the standard deviation of the estimator from one hundred trial runs. While the classical Donoho estimator overestimates all noise for $\beta > 0$, but using all wavelet components underestimates the noise of the same data, the adapted Donoho estimator provides a better estimate.

routine (Sect. 4.4). The maximum window length of 100 days has been chosen to allow for temporal changes in the precision, while keeping the statistical properties stable. The full algorithm is given in Section C.2.4. We include estimates of these uncertainties for each data-point in the final uncertainty budget. The combined precision is presented in the NetCDF attribute `solid_precision` (Sect. B.3.6).

As for stability, that is the uncertainty on the drift, the problem is a bit different. Instruments that have on-board monitoring systems (using calibration lamps or star observations) can provide an estimate of the drift with its uncertainty. Not all datasets, however, contain such stability estimates. Furthermore, the stability estimates for some instruments rely on assumptions based on the correlation between irradiance and solar proxies (e.g. the SBUV datasets). Although we cannot provide a completely independent estimate of the stability, there is a need for a common and homogeneous stability estimate for each dataset. The estimation of the stability (i.e. long-term uncertainties) is a very challenging task and can only be formally done through instrument intercomparison. Meanwhile, we can estimate medium-term uncertainties (on a yearly time scale) and use this as a first guess for stability. We do this by fitting each time series with a combination of proxies (Magnesium II, daily sunspot area, and the radio fluxes at 3.2 cm, 10.7 cm, 15 cm, and 30 cm) and by distinguishing the short (<108 days) and long (>108 days) time scales, the 108-day limit corresponding to four solar rotations. The fitting is designed in a flexible way (e.g. by including proxies that capture various solar features) in order not to force the irradiance to follow specific solar proxies and to keep the independent nature of the measurements. The stability at each time step is defined as the difference in the yearly slope of the observed and fitted time series at each wavelength. This method and its

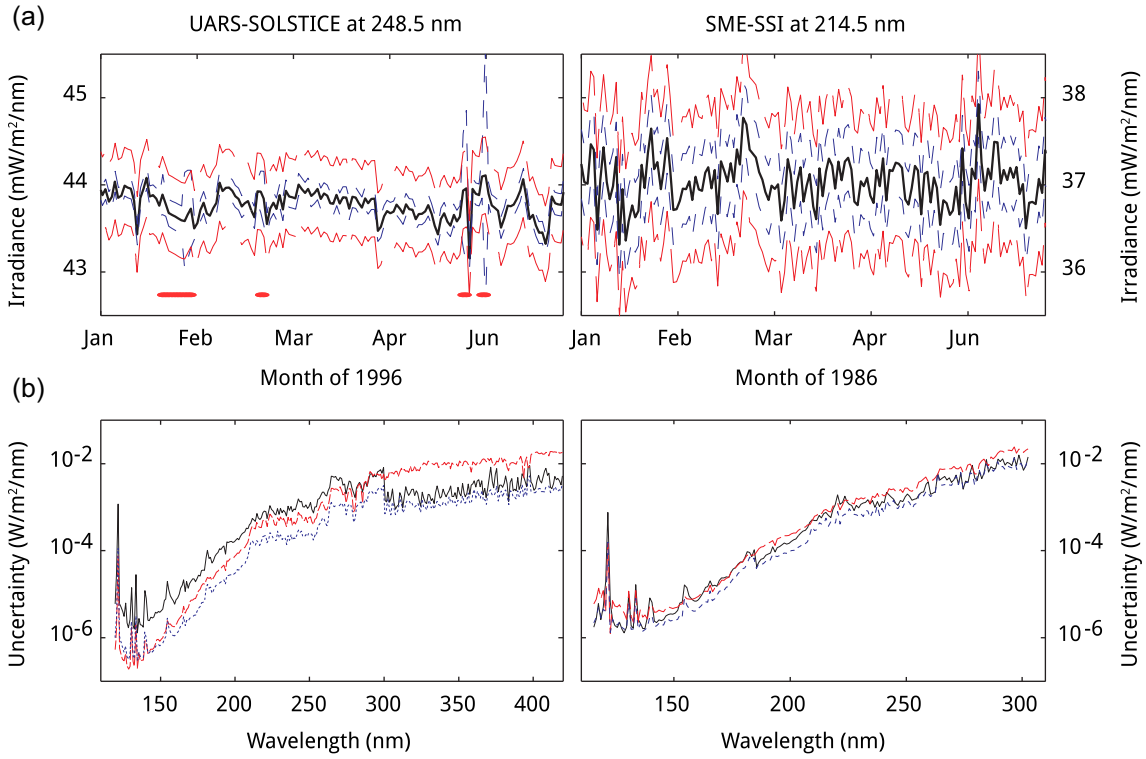

Fig. 4. (a) Measured SSI (black) of UARS SOLSTICE at 248.5 nm (left) and SME at 214.5 nm, along with the precision provided by PIs (long-dashed red) and our own estimate of precision (short-dashed blue). We also mark the times where we have detected outliers (red circles). Two strong outliers are visible in June, while two possible, but unlikely outliers are around 20 February 1996. The interpolated data from end of January have also been correctly classified as outliers. (b) The $1-\sigma$ standard deviation of the spectral variability (solid black) for each wavelength and the provided (long-dashed red) and estimated (short-dashed blue) averaged uncertainties per wavelength of the two instruments from (a), but for the full time-range (3 Oct 1991 to 29 Sept 2001 for UARS/SOLSTICE and 8 Oct 1981 to 12 Apr 1989 for SME/UV). As discussed in the text, our own estimates are, by definition, bounded by the variability of the data.

results are described in a forthcoming paper (Kretzschmar et al., in preparation).

The different components of uncertainty need to be combined into a single value. Here we present two different methods to calculate the absolute and relative uncertainty. Equation (1) determines the absolute uncertainty for a point in time, that is, taken the given value, it determines how accurate the data are in absolute terms, given accuracy, precision, and stability, while Eq. (2) calculates the uncertainties between two points in time to determine the accuracy of a trend.

$$\varepsilon_\lambda(t) = \sqrt{a_\lambda^2 + p_\lambda^2(t) + \left(\int_0^t s(t)\mathrm{d}t\right)^2}, \quad (1)$$

$$\varepsilon_\lambda(t_1, t_2) = \sqrt{p_\lambda^2(t_1) + p_\lambda^2(t_2) + \left(\int_{t_1}^{t_2} s(t)\mathrm{d}t\right)^2}, \quad (2)$$

with p representing precision, a accuracy as estimated by the PI, s stability, and t time, where $t = 0$ corresponds to the time of calibration.

As the discussion above shows, the accompanying uncertainties are an integral part of the data. To sum up, we provide the uncertainties as given by data providers, converted to standardized units, as well as our own independent estimates of precision in the form of the adapted Donoho estimator and stability of the data as estimated by the proxy-based model.

4. Methods

Here we present the pipeline for the data preparation for each individual dataset, providing methods that can be applied to all datasets and uncertainty estimations that allow different datasets to be compared in a meaningful way. Our pipeline approach entails the following steps: pre-processing, interpolation of missing data, re-gridding in time, outlier detection, and, finally, a proxy-based interpolation of erroneous or missing data. We provide uncertainty estimates for each step in the processing pipeline.

Figure 5 presents a graphical overview of all steps to construct the final composite, marking all steps discussed in this paper.

Appendix B includes a complete and formal description of the data format, while Appendix C provides the formal description of all methods in meta-code. The following sections describe the methods and the underlying rationale for their selection.

4.1. Pre-treatment

Our first step in the processing queue consists of the pre-treatment of the data: setting up a unified format, re-gridding the data on a two-dimensional time-wavelength grid, and removing the most extreme outliers. First, we convert the data format into one common format using the NetCDF Climate Forecast standard, version 1.6.

Next, we arrange the data on a two-dimensional time-wavelength grid with the axes conforming to wavelength in nanometers and time in days since 1980 without further data adjustments. As we will describe below (Sect. 4.3), we regularize the temporal grid at a later point in our pipeline.

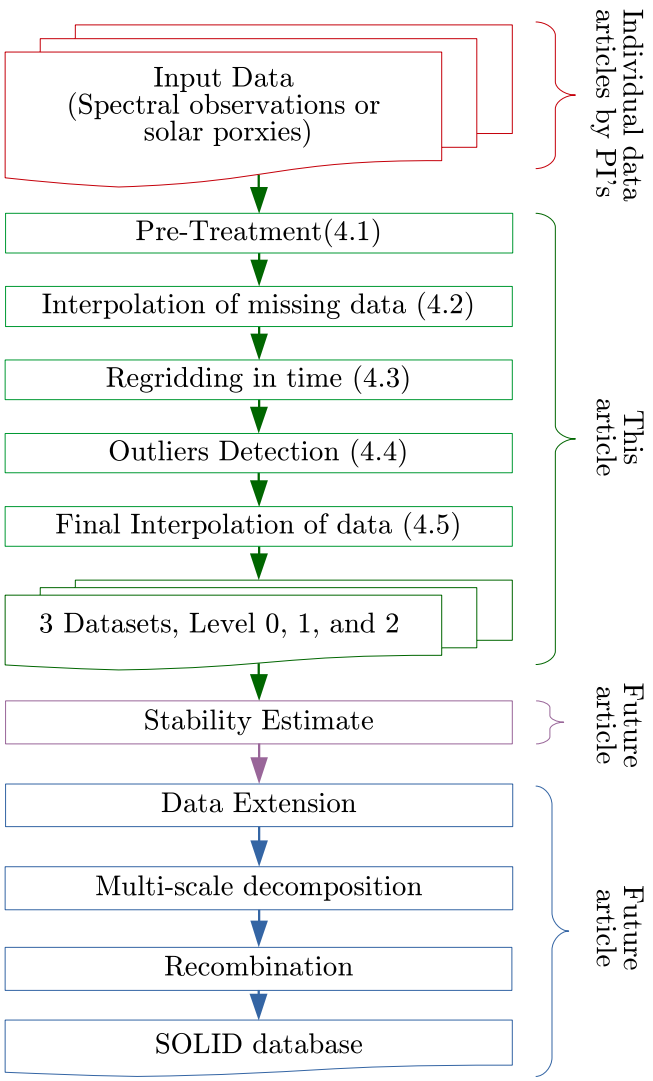

Fig. 5. A graphical overview of all steps to construct the final composite. The input data (red) are provided by instrument teams (References can be found in Appendix A). This work discusses the construction of the homogeneous individual datasets (green, numbers link to the corresponding sections), while the stability estimation (magenta) is discussed in a separate paper. The combination of all these datasets will be provided in the future (blue).

This approach is justified since most statistical methods either require or are easier to use when applied to regularly gridded data without any missing data.

Finally, it is necessary to remove obvious errors, that is values that are well outside the physical realm, as some methods, notably least-square fits, can be heavily influenced by those outliers. Figure 6 displays an example of a time series with such outliers. We flag and remove all data that are farther than 16 σ away from the mean. Even though the choice of 16 σ is somewhat arbitrary, it eliminates the most extreme non-physical outliers, while guaranteeing that we do not eliminate large amplitude transients, which are more frequent in the UV bands. The goal here is not to detect all outliers, but to eliminate the possibility of a single value dominating the statistics of a dataset.

All these changes are tracked in an accompanying quality map of the same dimension as the data field (formal description in Sect. B.3.5) with its entries representing the sum of

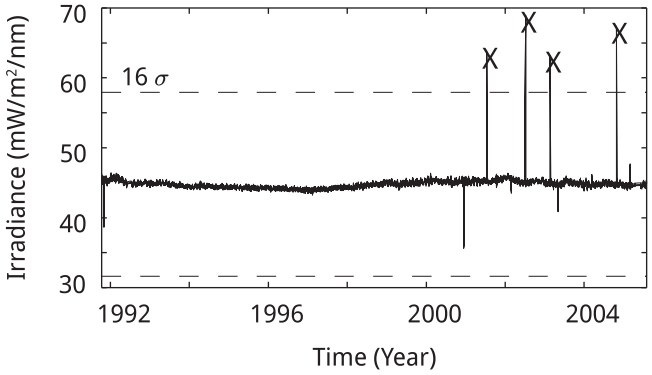

Fig. 6. UARS/SUSIM at 239.5 nm as provided by the instrument team and the 16 σ thresholds (dashed). All points outside of the thresholds are marked as outliers, replaced by linear interpolation. Further possible outliers are selected and handled in Section 4.4. We define their uncertainties as the difference from the original data to the interpolated points. In the case of SUSIM, these outliers occur directly after instrumental data gaps, giving further credence to the possibility of instrumental, non-physical, outliers.

binary flags. The numeric value of each flag is a unique power of two – 1, 2, 4, etc. – and the entry on the quality map is the sum of all the values of the present flags. For example, a value of 12 represents two flags, corresponding to the values 4 and 8. Any processing applied to the data is indicated in the quality map.

4.2. Interpolation of missing data

Most instruments incorporated in SOLID feature missing data with notable exceptions being SME/UV and SNOE/SXP. Missing data result from various problems, including faulty transmissions, eclipses, temporary device failures, and outages when the satellite is put in safe mode. Whatever its origin, we need to systematically deal with it for each instrument.

To reconstruct missing data, we rely on the high coherency of spectral solar variability. This coherency implies a strong redundancy of the observed variability, in the sense that temporal variations in a given spectral band can be adequately reconstructed by linear combination of typically 1–5 time series. This property has been exploited by Dudok de Wit (2011) to build an interpolation scheme that relies on Singular Value Decomposition (SVD, Kalman 1996). What follows is a description of this method. First, we temporarily flag and replace all missing data by a two-dimensional linear interpolation. Theoretically, any value can be used. Other possible choices are the average of the data or random sampling from the underlying distribution of the data. The improvement gained by a good estimate is a faster convergence of our method. Once we have a preliminary estimate, we recursively replace the missing data with a first rank SVD approximation until convergence is achieved. When converged, we increase the rank of the SVD by one and replace all missing data with the second rank SVD approximation until convergence. This is done repetitively up to the 10th rank. Missing data are also marked as such in the quality map, using the flag `Interpolated`. Figure 7 presents an example of the resulting interpolation of SORCE/SOLSTICE at 144 nm. While most data gaps can easily be interpolated due to neighboring channels providing data, gaps due to instrumental outage, which affect the whole instrument, are also flagged as missing and temporarily interpolated linearly. For the final data product, we use proxies to provide the temporal information. However at this step we refrain from inducing any external data into the dataset. The error induced by this method is estimated via bootstrapping.

Table A.2 presents an overview of all data adjustments, including the interpolation of data. These adjustments result in a dataset without any missing value on a two-dimensional time-wavelength grid with accompanying uncertainties.

4.3. Re-gridding in time

This section describes our regridding of the datasets onto a regular time grid. This procedure is necessary for several reasons. First, the autoregressive model utilized in our outlier detection, described in Section 4.4, requires a regular time grid. Also, a regular time grid is helpful to directly compare data with each other and to apply multiscale decomposition methods, as planned for future work.

Luckily, most datasets are already available with daily resolution, centered at noon. Only three datasets, namely AEE/EUV, ENVISAT/SCIAMACHY, and ISS/SolACES, need to be adjusted in time. Another dataset, NIMBUS7/SBUV, has multiple data values for the same time. We take the average of these and adjust the uncertainties by calculating the standard deviation. Table 2 provides the statistics of these datasets.[3]

We re-grid these datasets to a constant time grid of daily resolution centered at 12:00 UTC via a linear interpolation scheme and estimate the interpolation-induced uncertainty by a linear error regression of the variability. For this, we assume an independent error propagation, and, since the wavelet noise estimator is an accurate short-term noise estimate, we use the wavelet estimator of the original nearest neighbor as the non-weighted starting point. This estimate is weighted by the distance of the re-gridding. A proxy-based interpolation scheme is used for the final interpolation (Sect. 4.5); linear interpolation is used in intermediate steps for its cheap computational costs and simple statistical properties.

This procedure presents a classic chicken-and-egg problem. While the autoregressive model, utilized in our outlier detection, requires a regular time grid, the regularization of this time grid, based on interpolation, dilutes precisely these outliers. Though this problem may be solved through an iterative approach, we prefer to first interpolate the data including the outliers at the risk of diluting the latter and to later remove them. Our procedure yields similar results, but is more efficient than the iterative approach. Furthermore, since our final interpolation, described in Section 4.5, does not distinguish between missing data and outliers, this problem does not affect the final data values, but only the accompanying estimates of the precision.

4.4. Outlier detection

The following section describes our outlier detection routine based on an autoregressive model (AR, Chatfield 2003). Figure 8 provides a quick overview of this procedure, consisting of the aforementioned AR model together with a proxy-based quality estimator to differentiate physical outliers from instrumental artifacts.

Before addressing the procedure above in greater detail, it is necessary to discuss the nature of outliers. Having removed

[3] This time shift affects all data-points and is the only change applied to the original data not marked as such in the quality map `solid_flags`.

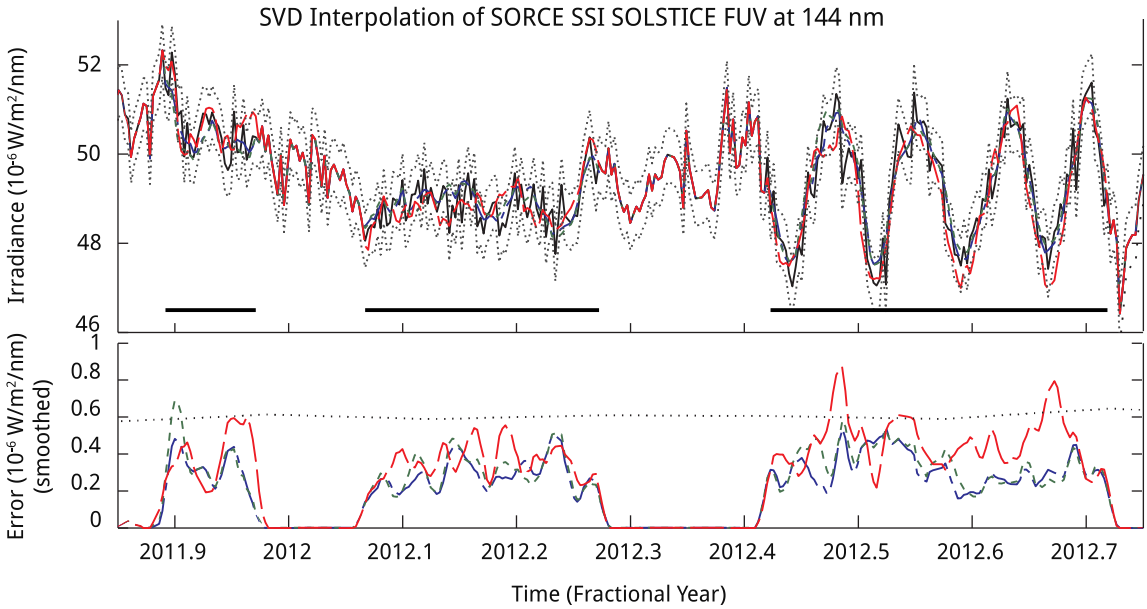

Fig. 7. Coherency interpolation at different points in time for the 144 nm channel of SORCE SOLSTICE A, which contains 65 channels in total (115 nm–179 nm). Each color represents a reconstruction of the 144 nm channel using a sequence of datasets of which some contain gaps for the days marked by a black horizontal line that were filled using the SVD interpolation. We show the original data (solid thick black), interpolated data where only this channel has been removed and re-interpolated (long-short dashed, blue), data where not only the 144 nm channel has been removed, but also its lower and upper 10 channels, i.e. everything from 134 nm to 154 nm (short-dashed green), and data where everything but the outermost channels (the two channels at 115 nm and 179 nm) have been removed for the selected days (long-dashed red). We also show our estimated precision (dotted black). The lower plot displays the error introduced by the SVD interpolation, smoothed over four days for better visibility. Due to the high spectral coherency, the error in the reconstruction is not significantly affected when removing not only the neighboring two, but 21 channels. Removing all but two channels does increase the error up to 100%, yet most data are still below our own estimation of the precision.

Table 2. An overview of all datasets that are interpolated in time. We present minimum, median, maximum, and standard deviation of the absolute time shifts. Time is given in hours (hh:mm).

Name	Min.	Med.	Max.	Sth
AEE/EUV	0:00	6:35	11:55	3:19
SCIAMACHY	0:57	5:57	9:23	0:37
SolACES	0:01	4:26	11:57	2:58

the most extreme outliers during pre-treatment (Sect. 4.1), we now aim at distinguishing instrumental artifacts from peculiar data corresponding to actual physical phenomena such as solar flares and huge sunspots. As they may have the same magnitude, they cannot be distinguished by a simple threshold. We also want to include sunspots but flag solar flares due to their short time spans, and, as such, their inclusion into each instrumental dataset strongly depends on internal characteristics like integration time and data selection, i.e. whether, for example, averaged daily data or a single daily measurement has been provided.

We resort to an autoregressive model to detect both instrumental artifacts and physical outliers. The approach is conceptually similar to that developed by Mann & Lees (1996), and it assumes that each record can be modeled by a linear time-invariant model. The discrepancy between this model and the observations is then used for detecting outliers. An AR model is a linear model of $p + 1$ parameters, which assumes that each point is linearly dependent on the p previous points, $y_t = \varepsilon_t + \sum_{i=1}^{p} b_i y_{t-i}$, with y_t corresponding to the data at time t, b the coefficients of the AR model, and ε_t the error at time t. These models are specified by their order p, and the parameter b can be estimated by a least-squares fit to y (Priestley 1981). The higher the model order, the better the fit, but also the higher its complexity. To determine the optimal order of the model, we apply the corrected Akaike information criterion (AIC_C, Burnham & Anderson 2002, see below Eq. (3) for a version applied to AR), which weights the goodness-of-fit versus the complexity of the model. The lower C in AIC_C, representing "corrected", accounts for the possibility of overfitting a data series of small size by multiplying the number of model parameters, $p + 1$ with a correction factor.

$$\widehat{\mathrm{AIC}}_C = \overbrace{n(\log\sigma_p^2 + 1)}^{\text{goodness-of-fit}} + \overbrace{2(p+1)}^{\text{complexity}} \overbrace{\frac{n}{n-p-2}}^{\text{correction factor}}, \qquad (3)$$

with n the sample size, p the order of the tested AR model, and σ_p the standard deviation of the AR model of order p.

For most models, the estimated AIC_C reaches its maximum for 4th order models. Hence, here, we construct a stationary 4th order AR model. However, it should be noted that, especially for the visible light spectrum, this parameter can increase significantly and that an adaptive parameter may improve the predictability and, thereby, the quality of the model.

Our model assumes that the Sun is a steady system perturbed by magnetic active regions that remain stable over the course of a few days. However, the Sun may also display fast transients that cannot be predicted by an autoregressive model, e.g. rapidly emerging sunspots and flares. Here, we account for the possibility of such physical outliers using an external reference in the form of solar proxies to relax the outlier detection criteria. Yet solar flares, as another type of physical outliers, are removed from our datasets, as our solar proxies do not contain

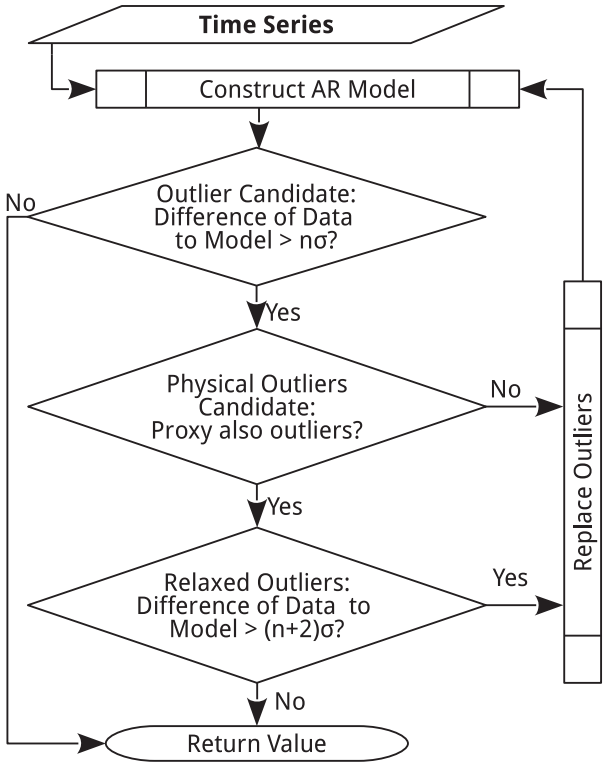

Fig. 8. Flowchart of the outlier detection with the possibility of physical outliers. First, we construct an autoregressive model of the time series. Then we flag all values whose difference to the model is greater than a given threshold n as possible outlier candidates. To account for the possibility of physical outliers, we relax this condition if proxies flag the same data as outliers.

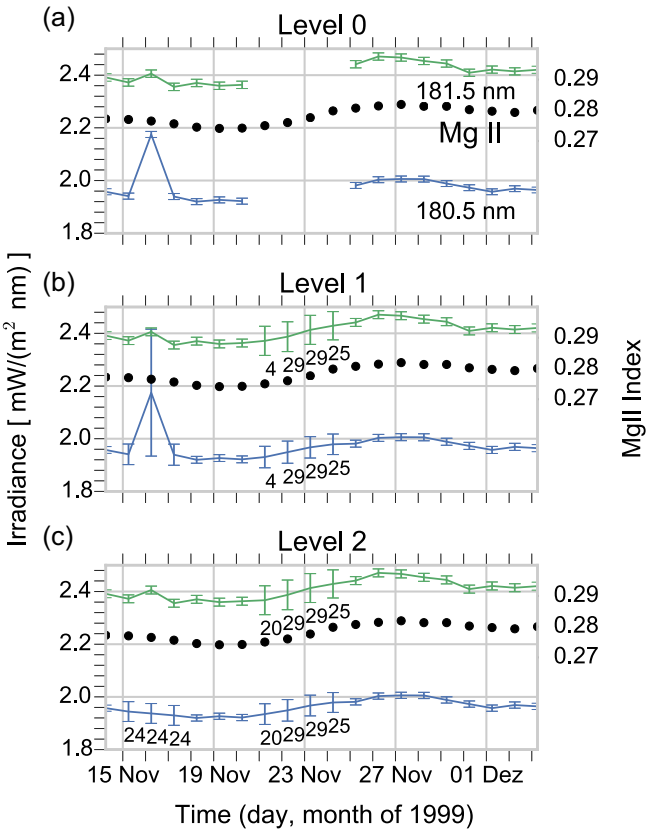

Fig. 9. A sample of the three provided levels of UARS/SUSIM, the original data, named level 0 (a), re-gridded and missing data interpolated as level 1 (b), and our best-estimate, level 2, where outliers have been removed and re-interpolated (c). One solar proxy, Mg II, is shown for comparison (black dots). The error bars represent our estimate of the precision, including the induced uncertainty of each method. Also shown are the values of all set flags. For example, a value of 24 means that bits 3 (2^3) and 4 (2^4) are set, marking this data-point as both being detected as an outlier (8) and re-interpolated in the final interpolation (16).

flare information and the flare detection strongly depends on the instrument.

We flag all instrumental outliers and increase their uncertainty in three steps, proportional to our confidence that they are actual outliers. First, we use the aforementioned 4th order AR model to flag all outlier candidates, both instrumental and physical, and calculate the standard deviation of the differences, preliminarily flagging all data where the difference of the model to the data is greater than $n\sigma$, with $n = 4$. Second, to account for the possibility of outliers being of physical origin, the parameter n is increased by 2 if at least 70% of our solar proxies are also flagged as outliers. Third, all such flagged data, as well as data that have been modified in a previous method, are temporarily replaced by linear interpolation. This procedure is repeated until convergence is reached. The linear interpolation is replaced by a more advanced scheme in a later step. Typically, this process converges within three steps, but the number of steps necessary for convergence increases when long data gaps have been interpolated in a previous step. We have empirically determined the required parameters, that is n, the increase by 2, and the percentage threshold of the proxies.

Finally, we convert the replaced data into uncertainties while keeping the original data in place, with the final step, coherency interpolation, only applied after all processing is finished. This is described below in Section 4.5.

The estimated uncertainty induced by outliers is defined as the absolute difference between the original and predicted data.

Applying the outlier detection routine as discussed above to our datasets yields results which we found to be physically consistent. We present an application of our routine to SOHO-SEM in Figure 10. As this figure shows, after the removal of the most extreme outliers, our method still detects not only several obvious instrumental artifacts, but also data-points that are below any meaningful direct threshold. In some cases, as in the case of SUSIM (Fig. 6), they appear before a data gap, giving further credence that they are due to instrumental artifacts. Our routine also correctly flags all linearly interpolated data as non-physical. Some values are possibly flagged as outliers erroneously. However, the proposed replacement barely differs from the original value, and, hence, it only slightly affects the induced uncertainty.

4.5. Final interpolation of missing data

Finally, we interpolate all data that have been flagged using the coherency interpolation scheme as described above in Section 4.2, this time including proxies as additional channels. While this interweaves instrumental data with proxy data for all data-points that are missing from the original dataset, it attains temporal variability during times of instrumental outages when no channel has data.

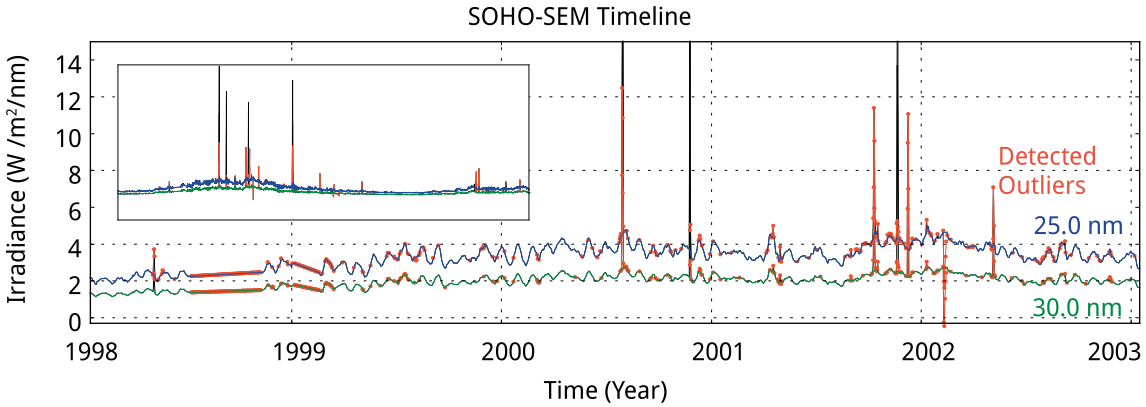

Figure 10. Time series of SOHO-SEM data after application of the AR outlier removal routine (blue [25 nm], green [30 nm]). We show the original data (black) without any processing done and the input data to the AR routine (red). Red dots mark points that are detected as outliers. This method picks obvious errors (e.g. around February 2002 there are some negative values), linearly interpolated data (mid-end of 1999, January 1999), and solar flares. However, it also detects some points that are not obviously (and necessarily) outliers. Their uncertainty is increased by the difference to the interpolated data. The inlet displays the whole time series, January 1996 to June 2014.

5. Results

Based upon these methods, we provide three kinds of datasets: first, the original data converted into the common data format; second, the original data accompanied by uncertainty estimates and interpolated data; and, third, a "best-estimate" data product with possible outliers replaced by proxy interpolation. Figure 9 displays these three datasets, together with the accompanying flags and precision estimation.

As our database now contains three different types of datasets, the question "Which one is best for a given purpose?" naturally arises. This depends largely on the chosen method, as the subsequent examples illustrate. As a simple least-square fit can be dominated by outliers, it is of vital importance that instrumental artifacts be removed. Thus, the third type of dataset should be used for least-square fitting.[4] However, if a weighted leastsquare fit or, more generally, any method that takes into account the uncertainties of data is used, the second type of dataset is more appropriate. The first type of dataset serves primarily as a reference and starting point, as it does not include independent uncertainty estimates.

All datasets are available at ftp://www.pmodwrc.ch/pub/projects/solid/database. The database will be updated regularly as new data become available, including modeled data, which will complete temporal and spectral coverage. Using standardized methods to homogenize the data makes the inclusion of new data, or new versions of already available data, easily achievable.

6. Conclusion

As part of SOLID, we provide homogeneous datasets of available space-based SSI measurements with uncertainty estimates and accompanying meta-data in a unified format, following standard conventions.

Several challenges, however, remain. Even though we have put much effort into developing an independent and homogeneous uncertainty estimate, we do recognize that we have not yet solved the problem of combining these with the PI provided uncertainty estimates. As for now, we provide both sets of estimates, and it is up to the user to combine them. Furthermore, our methods are designed to be generic and applicable to all instruments incorporated in SOLID, but consequently, they do not reflect all properties specific to each instrument. Some datasets, for instance, are accompanied by flags describing the origin, the processing method, and the quality of the data or the number of measurements per datapoint. Though provided in our database, these flags are not part of our processing.

Notwithstanding these challenges, our datasets are a starting point to merge all available data into one single homogeneous dataset. A follow-up paper will present a new approach to create such an SSI composite, which will also eventually be available at the above address. In any case, the present database will help to provide an SSI reconstruction of the satellite era, and, thereby, it will foster diverse research in solar science, space weather, and climate studies.

Acknowledgements. The authors acknowledge that the research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7 2012) under Grant Agreement No. 313188 (SOLID, http://projects.pmodwrc.ch/solid). The authors would like to thank all data providers who provided their data on-line, David Bolsée for the SOLSPEC data, Robert Schäfer for SolACES data, Gérard Thuillier for the ATLAS reference spectra, and Marty Snow for the helpful discussions during the SOLID meetings in Bremen and London. This research has made use of NASA's Astrophysics Data System. The editor thanks two anonymous referees for their assistance in evaluating this paper.

References

Anderson, G.P., and L.A. Hall. Solar irradiance between 2000 and 3100 Angstroms with spectral band pass of 1.0 Angstroms. *J. Geophys. Res.*, **94**, 6435–6441, 1989, DOI: 10.1029/JD094iD05p06435.

Arvesen, J.C., R.N. Griffin, and B.D. Pearson, Jr. Determination of extraterrestrial solar spectral irradiance from a research aircraft, *Appl. Opt.*, **8**, 2215, 1969, DOI: 10.1364/AO.8.002215.

Bailey, S.M., T.N. Woods, C.A. Barth, S.C. Solomon, L.R. Canfield, and R. Korde. Measurements of the solar soft X-ray irradiance by the Student Nitric Oxide Explorer: first analysis and underflight calibrations. *J. Geophys. Res.*, **105**, 27179–27194, 2000, DOI: 10.1029/2000JA000188.

[4] Alternatively, another fitting method such as the median fit alleviates the problem of outliers and can be used with any type of dataset.

Balmaceda, L.A., S.K. Solanki, N.A. Krivova, and S. Foster. A homogeneous database of sunspot areas covering more than 130 years. *J. Geophys. Res.*, **114**, A07104, 2009, DOI: 10.1029/2009JA014299.

Baum, W.A., F.S. Johnson, J.J. Oberly, C.C. Rockwood, C.V. Strain, and R. Tousey. Solar ultraviolet spectrum to 88 kilometers. *Phys. Rev.*, **70**, 781–782, 1946, DOI: 10.1103/PhysRev.70.781.

Burnham, K.P., and D.R. Anderson. *Model selection and multimodel inference*, 2nd edn., Springer, New York, ISBN: 0387953647, 2002.

Burrows, J.P., E. Hölzle, A.P.H. Goede, H. Visser, and W. Fricke. SCIAMACHY – scanning imaging absorption spectrometer for atmospheric cartography. *Acta Astronaut.*, **35**, 445–451, 1995, DOI: 10.1016/0094-5765(94)00278-T.

Cebula, R.P., M.T. DeLand, and E. Hilsenrath. NOAA 11 solar backscattered ultraviolet, model 2 (SBUV/2) instrument solar spectral irradiance measurements in 1989–1994. 1. Observations and long-term calibration. *J. Geophys. Res.*, **103**, 16235–16250, 1998, DOI: 10.1029/98JD01205.

Chatfield, C. *The Analysis of Time Series: An Introduction*. 6th edn., Chapman and Hall/CRC, Boca Raton, Florida, ISBN: 9781584883173, 2003.

Colina, L., R.C. Bohlin, and F. Castelli. The 0.12–2.5 micron absolute flux distribution of the sun for comparison with solar analog stars. *Astron. J.*, **112**, 307–307, 1996, DOI: 10.1086/118016.

Del Zanna, G., and V. Andretta. The EUV spectrum of the Sun: SOHO CDS NIS irradiances from 1998 until 2010. *A&A*, **528**, A139, 2011, DOI: 10.1051/0004-6361/201016106.

DeLand, M.T., and R.P. Cebula. NOAA 11 Solar Backscatter Ultraviolet, model 2 (SBUV/2) instrument solar spectral irradiance measurements in 1989–1994. 2. Results, validation, and comparisons. *J. Geophys. Res.*, **103**, 16251–16274, 1998, DOI: 10.1029/98JD01204.

DeLand, M.T., and R.P. Cebula. Spectral solar UV irradiance data for cycle 21. *J. Geophys. Res.*, **106**, 21569–21584, 2001, DOI: 10.1029/2000JA000436.

DeLand, M.T., and R.P. Cebula. Creation of a composite solar ultraviolet irradiance data set. *J. Geophys. Res.*, **113 (A12)**, A11103, 2008, DOI: 10.1029/2008JA013401.

DeLand, M.T., R.P. Cebula, and E. Hilsenrath. Observations of solar spectral irradiance change during cycle 22 from NOAA-9 Solar Backscattered Ultraviolet Model 2 (SBUV/2). *J. Geophys. Res.*, **109**, D06304, 2004, DOI: 10.1029/2003JD004074.

Domingo, V., I. Ermolli, P. Fox, C. Fröhlich, and M. Haberreiter, et al. Solar surface magnetism and irradiance on time scales from days to the 11-year cycle. *Space Sci. Rev.*, **145**, 337–380, 2009, DOI: 10.1007/s11214-009-9562-1.

Donoho, D.L., and I.M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, **90 (432)**, 1200–1224, 1995, DOI: 10.1080/01621459.1995.10476626.

Dudok de Wit, T. A method for filling gaps in solar irradiance and solar proxy data. *A&A*, **533**, A29, 2011, DOI: 10.1051/0004-6361/201117024.

Eparvier, F.G., D. Crotser, A.R. Jones, W.E. McClintock, M. Snow, and T.N. Woods. The Extreme Ultraviolet Sensor (EUVS) for GOES-R. *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 7438 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 4, 2009, DOI: 10.1117/12.826445.

Ermolli, I., K. Matthes, T. Dudokdewit, N.A. Krivova, and K. Tourpali, et al. Recent variability of the solar spectral irradiance and its impact on climate modelling. *Atmos. Chem. Phys.*, **13**, 3945–3977, 2013, DOI: 10.5194/acp-13-3945-2013.

Evans, J.S., D.J. Strickland, W.K. Woo, D.R. McMullin, S.P. Plunkett, R.A. Viereck, S.M. Hill, T.N. Woods, and F.G. Eparvier. Early Observations by the GOES-13 Solar Extreme Ultraviolet Sensor (EUVS). *Sol. Phys.*, **262**, 71–115, 2010, DOI: 10.1007/s11207-009-9491-x.

Floyd, L., G. Rottman, M. DeLand, and J. Pap. 11 years of solar UV irradiance measurements from UARS. In: A. Wilson, Editor. *Solar Variability as an Input to the Earth's Environment*, vol. 535 of ESA Special Publication, ESA Publications Division, Noordwijk, 195–203, 2003.

Fröhlich, C. Solar irradiance variability since 1978. Revision of the PMOD composite during solar cycle 21. *Space Sci. Rev.*, **125**, 53–65, 2006, DOI: 10.1007/s11214-006-9046-5.

Fröhlich, C., J. Romero, H. Roth, C. Wehrli, B.N. Andersen, et al. VIRGO: Experiment for helioseismology and solar irradiance monitoring. *Sol. Phys.*, **162**, 101–128, 1995, DOI: 10.1007/BF00733428.

Harrison, R.A., E.C. Sawyer, M.K. Carter, A.M. Cruise, R.M. Cutler, et al. The coronal diagnostic spectrometer for the solar and heliospheric observatory. *Sol. Phys.*, **162**, 233–290, 1995, DOI: 10.1007/BF00733431.

Hinteregger, H.E.. Representations of solar EUV fluxes for aeronomical applications. *Adv. Space Res.*, **1**, 39–52, 1981, DOI: 10.1016/0273-1177(81)90416-6.

Hinteregger, H.E., D.E. Bedo, and J.E. Manson. The EUV spectroheliometer on atmosphere explorer. *Radio Science*, **8**, 349–359, 1973, DOI: 10.1029/RS008i004p00349.

Kalman, D. A singularly valuable decomposition: the SVD of a matrix. *College Math J.*, **27**, 2–23, 1996, DOI: 10.1.1.113.1193.

Keil, S.L., T.W. Henry, and B. Fleck. NSO/AFRL/Sac Peak K-line Monitoring Program. In: K.S. Balasubramaniam, J. Harvey, and D. Rabin, Editors. *Synoptic Solar Physics*, vol. 140 of Astronomical Society of the Pacific Conference Series, American Scientific Publishers, Valencia, California, 301, 1998.

Kurucz, R.L. Remaining line opacity problems for the solar spectrum. *Revista Mexicana de Astronomia y Astrofisica*, **23**, 187, 1992.

Kurucz, R.L. High Resolution Irradiance Spectrum from 300 to 1000 nm, *ArXiv Astrophysics e-prints*, 2006.

Lean, J. Evolution of the Sun's spectral irradiance since the maunder minimum. *Geophys. Res. Lett.*, **27**, 2425–2428, 2000, DOI: 10.1029/2000GL000043.

Lean, J., J. Beer, and R. Bradley. Reconstruction of solar irradiance since 1610: implications for climate change. *Geophys. Res. Lett.*, **22**, 3195–3198, 1995, DOI: 10.1029/95GL03093.

Mallat S., Editor. *A wavelet tour of signal processing*, 3rd edn., Academic Press, Boston, ISBN: 978-0-12-374370-1, 2009, DOI: 10.1016/B978-0-12-374370-1.00001-X.

Mann, M.E., and J.M. Lees. Robust estimation of background noise and signal detection in climatic time series. *Clim. Change*, **33 (3)**, 409–445, 1996, DOI: 10.1007/BF00142586.

Marchenko, S.V., and M.T. DeLand. Solar spectral irradiance changes during cycle 24. *Astrophys. J.*, **789**, 117–117, 2014, DOI: 10.1088/0004-637X/789/2/117.

McClintock, W.E., G.J. Rottman, and T.N. Woods. Solar-Stellar Irradiance Comparison Experiment II (Solstice II): instrument concept and design. *Sol. Phys.*, **230**, 225–258, 2005, DOI: 10.1007/s11207-005-7432-x.

Mount, G.H., and G.J. Rottman. The solar absolute spectral irradiance 1150–3173 A – May 17, 1982. *J. Geophys. Res.*, **88**, 5403–5410, 1983, DOI: 10.1029/JC088iC09p05403.

Priestley, M.B. *Spectral analysis and time series*. Academic Press, London, 1981.

Rottman, G., J. Harder, J. Fontenla, T. Woods, O.R. White, and G.M. Lawrence. The Spectral Irradiance Monitor (SIM): early observations. *Sol. Phys.*, **230**, 205–224, 2005, DOI: 10.1007/s11207-005-1530-7.

Rottman, G.J., C.A. Barth, R.J. Thomas, G.H. Mount, G.M. Lawrence, D.W. Rusch, R.W. Sanders, G.E. Thomas, and J. London. Solar spectral irradiance, 120 to 190 nm, October 13, 1981–January 3, 1982. *Geophys. Res. Lett.*, **9**, 587–590, 1982, DOI: 10.1029/GL009i005p00587.

Rottman, G.J., T.N. Woods, and T.P. Sparn. Solar-Stellar Irradiance Comparison Experiment 1. I – Instrument design and operation. *J. Geophys. Res.*, **98**, 10–667, 1993, DOI: 10.1029/93JD00462.

Schmidtke, G. Extreme ultraviolet spectral irradiance measurements since 1946, 2014, Under review.

Schmidtke, G., B. Nikutowski, C. Jacobi, R. Brunner, C. Erhardt, S. Knecht, J. Scherle, and J. Schlagenhauf. Solar EUV Irradiance Measurements by the Auto-Calibrating EUV Spectrometers (SolACES) aboard the International Space Station (ISS). *Sol. Phys.*, **289**, 1863–1883, 2014, DOI: 10.1007/s11207-013-0430-5.

SILSO World Data Center. The International Sunspot Number. *International Sunspot Number Monthly Bulletin and online catalogue*, 1970–2015.

Snow, M., W.E. McClintock, G. Rottman, and T.N. Woods. Solar Stellar Irradiance Comparison Experiment II (Solstice II): examination of the solar stellar comparison technique. *Sol. Phys.*, **230**, 295–324, 2005, DOI: 10.1007/s11207-005-8763-3.

Tapping, K.F. The 10.7 cm solar radio flux (F10.7). *Space Weather*, **11**, 394–406, 2013, DOI: 10.1002/swe.20064.

Thekaekara, M.P. Extraterrestrial solar spectrum, 3000–6100 Å at 1-Å intervals. *Appl. Opt.*, **13**, 518–522, 1974, DOI: 10.1364/AO.13.000518.

Thuillier, G., L. Floyd, T.N. Woods, R. Cebula, E. Hilsenrath, M. Hersé, and D. Labs. Solar irradiance reference spectra. In: J.M. Pap, P. Fox, C. Frohlich, H.S. Hudson, J. Kuhn, J. McCormack, G. North, W. Sprigg, and S.T. Wu, Editors. *Solar variability and its effects on climate*, Geophysical Monograph 141, American Geophysical Union, Washington, DC, 171, 2004.

Thuillier, G., S.M.L. Melo, J. Lean, N.A. Krivova, C. Bolduc, et al. Analysis of different solar spectral irradiance reconstructions and their impact on solar heating rates. *Sol. Phys.*, **289**, 1115–1142, 2014, DOI: 10.1007/s11207-013-0381-x.

Tobiska, W.K. SOLAR2000 irradiances for climate change research, aeronomy and space system engineering. *Adv. Space Res.*, **34**, 1736–1746, 2004, DOI: 10.1016/j.asr.2003.06.032.

Viereck, R., F. Hanser, J. Wise, S. Guha, A. Jones, D. McMullin, S. Plunket, D. Strickland, and S. Evans. Solar extreme ultraviolet irradiance observations from GOES: design characteristics and initial performance. *Proc. SPIE*, **6689**, 66890K.1–66890K.10, 2007, DOI: 10.1117/12.734886.

Viereck, R.A., L.E. Floyd, P.C. Crane, T.N. Woods, B.G. Knapp, G. Rottman, M. Weber, L.C. Puga, and M.T. DeLand. A composite Mg II index spanning from 1978 to 2003. *Space Weather*, **2**, S10005, 2004, DOI: 10.1002/2004SW000084.

Wieman, S.R., L.V. Didkovsky, and D.L. Judge. Resolving differences in absolute irradiance measurements between the SOHO/CELIAS/SEM and the SDO/EVE. *Sol. Phys.*, **289**, 2907–2925, 2014, DOI: 10.1007/s11207-014-0519-5.

Wilson, R.M., and D.H. Hathaway. On the relation between sunspot area and sunspot number. *NASA STI/Recon Technical Report N*, **6**, 20186, 2006.

Woods, T.N., P.C. Chamberlin, J.W. Harder, R.A. Hock, M. Snow, F.G. Eparvier, J. Fontenla, W.E. Mcclintock, and E.C. Richard. Solar Irradiance Reference Spectra (SIRS) for the 2008 Whole Heliosphere Interval (WHI). *Geophys. Res. Lett.*, **36**, L01101, 2009, DOI: 10.1029/2008GL036373.

Woods, T.N., F.G. Eparvier, R. Hock, A.R. Jones, D. Woodraska, et al. Extreme Ultraviolet Variability Experiment (EVE) on the Solar Dynamics Observatory (SDO): overview of science objectives, instrument design, data products, and model developments. *Sol. Phys.*, **275**, 115–143, 2012, DOI: 10.1007/s11207-009-9487-6.

Woods, T.N., D.K. Prinz, G.J. Rottman, J. London, P.C. Crane, et al. Validation of the UARS solar ultraviolet irradiances: comparison with the ATLAS 1 and 2 measurements. *J. Geophys. Res.*, **101**, 9541–9570, 1996, DOI: 10.1029/96JD00225.

Woods, T.N., E.M. Rodgers, S.M. Bailey, F.G. Eparvier, and G.J. Ucker. TIMED solar EUV experiment: preflight calibration results for the XUV photometer system. In: A.M. Larar, Editor. *Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research III*, vol. 3756 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Proceedings of SPIE, Bellingham, Washington, 255–264, 1999.

Woods, T.N., and G. Rottman. XUV Photometer System (XPS): Solar Variations during the SORCE Mission. *Sol. Phys.*, **230**, 375–387, 2005, DOI: 10.1007/s11207-005-2555-7.

Woods, T.N., G.J. Rottman, R.G. Roble, O.R. White, S.C. Solomon, G.M. Lawrence, J. Lean, and W.K. Tobiska. Thermosphere-Ionosphere-Mesosphere Energetics and Dynamics (TIMED) Solar EUV Experiment. In: J. Wang, and P.B. Hays, Editors. *Optical spectroscopic techniques and instrumentation for atmospheric and space research*, vol. 2266 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Proceedings of SPIE, Bellingham, Washington, 467–478, 1994.

Woods, T.N., W.K. Tobiska, G.J. Rottman, and J.R. Worden. Improved solar Lyman α irradiance modeling from 1947 through 1999 based on UARS observations. *J. Geophys. Res.*, **105**, 27195–27216, 2000, DOI: 10.1029/2000JA000051.

Yeo, K.L., N.A. Krivova, S.K. Solanki, and K.H. Glassmeier. Reconstruction of total and spectral solar irradiance from 1974 to 2013 based on KPVT, SoHO/MDI, and SDO/HMI observations. *A&A*, **570**, A85–A85, 2014, DOI: 10.1051/0004-6361/201423628.

Cite this article as: Schöll M, Dudok de Wit T, Kretzschmar M & Haberreiter M. Making of a solar spectral irradiance dataset I: observations, uncertainties, and methods. *J. Space Weather Space Clim.*, **6**, A14, 2016, DOI: 10.1051/swsc/2016007.

Appendix A. Tables

We have moved most of our tables to this separate section for better readability purpose. Here we present all instruments (Table A.1), an overview of all flags and their percentage applied to each dataset (Table A.2), and the list of all reference spectra (Table A.3) and proxies (Table A.4).

Table A1. The list of all instruments in the SOLID database, together with the data version if provided and the wavelengths and time ranges. Below each dataset name we expand the acronym of the mission and instrument name and provide a reference publication and the source URL if available. The dates correspond to the start and end dates of the database. These can differ from the mission dates. This is always the case for ongoing missions.

Name Sources and References	Version	Wavelength (nm)	Time interval
AEE/EUV Atmospheric Explorer-E/Solar EUV monitor; Hinteregger et al. (1973)		17.1–121.6	01 Jul 1977–09 Jun 1981
AURA/OMI Aura *(Latin for breeze)/Ozone* Monitoring Instrument; Marchenko and DeLand (2014)		265.0–499.5	08 Apr 1991–07 Dec 1997
ENVISAT/SCIAMACHY Environmental Satellite/SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY; Burrows et al. (1995), www.iup.uni-bremen.de/~weber/SOLARDATA/SCIAMACHYSSI.7z		212.5–2,385.0	02 Aug 2002–07 Apr 2012
GOES 13/EUVS Geostationary Operational Environmental Satellite 13/Extreme Ultraviolet Sensor; Evans et al. (2010), ftp://satdat.ngdc.noaa.gov/sem/goes/data/new_avg/2014/new_euv_temp/		11.7–123.2	07 Jul 2006–27 Oct 2014
GOES 14/EUVS Geostationary Operational Environmental Satellite 14/Extreme Ultraviolet Sensor; Eparvier et al. (2009); Viereck et al. (2007), ftp://satdat.ngdc.noaa.gov/sem/goes/data/new_avg/2014/new_euv_temp/		11.7–123.2	24 Jul 2009–20 Nov 2012
GOES 15/EUVS Geostationary Operational Environmental Satellite 15/Extreme Ultraviolet Sensor; Eparvier et al. (2009); Viereck et al. (2007), ftp://satdat.ngdc.noaa.gov/sem/goes/data/new_avg/2014/new_euv_temp/ *GOES 15/E is scaled to SORCE/SOSLTICE*		11.7–123.2	07 Apr 2010–26 Oct 2014
ISS/SolACES International Space Station/SOLar Auto-Calibrating EUV/UV Spectrophotometers; Schmidtke et al. (2014), *private comm. (Robert Schafer)*		16.5–57.5	04 Jan 2011–24 Mar 2014
ISS/SOLSPEC International Space Station/SOLar SPECtral Irradiance Measurements; Eparvier et al. (2009), *private comm. (David Bolsee),* calibration in progress		175.2–340.6	05 Apr 2008–10 Dec 2013
NIMBUS7/SBUV Nimbus *(Latin for rain cloud)* 7/Solar Backscatter Ultraviolet; DeLand and Cebula (2001), http://sbuv2.gsfc.nasa.gov/solar/		170.0–399.0	08 Nov 1978–28 Oct 1986
NOAA9/SBUV2 National Oceanic and Atmospheric Administration 9/Solar Backscatter Ultraviolet Model 2; DeLand et al. (2004), http://sbuv2.gsfc.nasa.gov/solar/		170.0–399.0	14 Mar 1985–05 May 1997
NOAA11/SBUV2 National Oceanic and Atmospheric Administration 11/Solar Backscatter Ultraviolet Model 2; Cebula et al. (1998); DeLand & Cebula (1998), http://sbuv2.gsfc.nasa.gov/solar/		170.0–399.0	05 Dec 1988–15 Oct 1994
NOAA16/SBUV2 National Oceanic and Atmospheric Administration 16/Solar Backscatter Ultraviolet Model 2; DeLand & Cebula (2008), http://sbuv2.gsfc.nasa.gov/solar/		170.0–406.2	10 Nov 2000–30 Apr 2003
SDO/EVE Solar Dynamics Observatory/EUV Variability Experiment; Woods et al. (2012), http://lasp.colorado.edu/eve/data_access/evewebdata/products/level3/	5	5.8–106.2	29 Apr 2010–21 Oct 2014
SDO/EVE-1 nm Solar Dynamics Observatory/EUV Variability Experiment-1 nm binned (SOLID product)	5	5.5–106.5	29 Apr 2010–21 Oct 2014

(*continued on next page*)

Table 1. (*continued*)

Name Sources and References	Version	Wavelength (nm)	Time interval
SME/UV Solar Mesosphere Explorer/Ultraviolet Solar Monitor Experiment; Rottman et al. (1982); Mount & Rottman (1983), http://lasp.colorado.edu/lisird/tss/sme_ssLcsv		115.5–302.5	08 Oct 1981–12 Apr 1989
SNOE/SXP Student Nitric Oxide Explorer/SNOE Solar X-ray Photometer; Bailey et al. (2000), http://lasp.colorado.edu/home/snoe/data/		4.5–4.5	11 Mar 1998–30 Sep 2000
SOHO/CDS Solar and Heliospheric Observatory/Coronal Diagnostic Spectrometer; Harrison et al. (1995); Del Zanna and Andretta (2011), *private comm. (Guilio Del Zanna)*	3.1 (1998)	31.4–62.0	23 Apr 1998–14 Jun 2010
SOHO/CELIAS-SEM Solar and Heliospheric Observatory/Charge, Element, and Isotope Analysis System-Solar Extreme Ultraviolet Monitor; Wieman et al. (2014), http://www.usc.edu/dept/space_science/sem_data/sem_data.html	3.1 (1998)	25.0–30.0	01 Jan 1996–05 Jun 2014
SOHO/VIRGO-SPM Solar and Heliospheric Observatory/Variability of Solar Irradiance and Gravity Oscillations-Solar Photometers; Fröhlich et al. (1995), ftp://ftp.pmodwrc.ch/pub/data/irradiance/virgo/SSI/spm_level2_d_170496_010206.dat The long-term trends of the SPM are not understood well enough to assess the solar cycle variability of the spectral channels.		402.0–862.0	17 Apr 1996–11 Jan 2006
SORCE/SIM Solar Radiation and Climate Experiment/Spectral Irradiance Monitor; Rottman et al. (2005), http://lasp.colorado.edu/home/sorce/data/	21	240.0–2,412.3	14 Apr 2003–12 May 2015
SORCE/SOLSTICE-FUV Solar Radiation and Climate Experiment/SOlar Stellar Irradiance Comparison Experiment-Far UV; McClintock et al. (2005); Snow et al. (2005), http://lasp.colorado.edu/home/sorce/data/	13	115.0–179.0	14 May 2003–12 May 2015
SORCE/SOLSTICE-MUV Solar Radiation and Climate Experiment/SOlar Stellar Irradiance Comparison Experiment-Middle UV; McClintock et al. (2005); Snow et al. (2005), http://lasp.colorado.edu/home/sorce/data/	13	180.0–309.0	14 May 2003–12 May 2015
SORCE/XPS Solar Radiation and Climate Experiment/XUV Photometer System; Woods & Rottman (2005), http://lasp.colorado.edu/home/sorce/data/	10	0.5–39.5	10 Apr 2003–08 Dec 2014
TIMED/SEE-EGS Thermosphere Ionosphere Mesosphere Energetics and Dynamics/Solar EUV Experiment-EUV Grating Spectrograph; Woods et al. (1994), http://lasp.colorado.edu/home/see/data/	11	27.1–189.8	08 Feb 2002–16 Feb 2013
TIMED/SEE-XPS Thermosphere Ionosphere Mesosphere Energetics and Dynamics/Solar EUV Experiment-XUV Photometer System; Woods et al. (1999), http://lasp.colorado.edu/home/see/data/	11	1.0–9.0	22 Jan 2002–09 Nov 2014
UARS/SOLSTICE Upper Atmosphere Research Satellite/SOlar Stellar Irradiance Comparison Experiment; Woods et al. (1996), http://lasp.colorado.edu/lisird/tss/uars_solstice_ssi.csv	11	9.5–419.5	03 Oct 1991–29 Sep 2001
UARS/SUSIM Upper Atmosphere Research Satellite/Solar Ultraviolet Spectral Irradiance Monitor; Rottman et al. (1993); Floyd et al. (2003), http://wwwsolar.nrl.navy.mil/uars/v22/	22	115.5–410.5	12 Oct 1991–01 Aug 2005

Table A2. An overview of the amount of data adjusted by each method. Each flag corresponds to a specific method: interpolation of missing data (1, Sect. 4.2), averaging multiple data-points (2, Sect. 4.3), outlier detection (3, Sect. 4.4), and proxy interpolation (4, Sect. 4.5). The flags are further described in Section B.7. Generally, all interpolated data are detected as outliers and interpolated by our coherency interpolation scheme. The CDS instrument does not provide daily observations, and, as such, has most data points interpolated. The last column corresponds to the number of channels in the instrument.

Name	Percentage of flagged data-points				Channels #
	1	2	3	4	
AURA/OMI	0	0	2.54	98.33	26
GOES 13/EUVS	0	0	2.66	68.06	3
GOES 14/EUVS	0	0	1.21	57.35	3
GOES 15/EUVS	0	0	0.90	6.93	3
ISS/SolACES	0	0	0	77.15	42
ISS/SOLSPEC	0.15	0	0	73.58	486
NIMBUS7/SBUV	0.19	2.40	0.71	30.43	230
NOAA11/SBUV2	0.07	0	1.26	18.98	230
NOAA16/SBUV2	6.63	0	6.91	15.44	1616
NOAA9/SBUV2	1.75	0	3.29	14.99	230
ENVISAT/SCIAMACHY	41.98	0	42.45	44.60	826
SDO/EVE	3.83	0	6.20	10	5020
SME/UV	0	0	0.25	0.26	188
SNOE-SXP	0	0	0.64	7.59	1
SOHO/CDS	0	0	0.38	99.05	58
SOHO/CELIAS-SEM	11.45	0	15.35	20.67	2
SOHO/VIRGO-SPM	4.68	0	7.13	7.13	3
SORCE/SIM	0	0	1.62	8.76	1217
SORCE/SOLSTICE-FUV	0	0	0.34	8.03	65
SORCE/SOLSTICE-MUV	0	0	0.53	8.72	130
SORCE/XPS	0	0	4.14	20.97	40
TIMED/SEE-EGS	1.24	0	1.53	2.29	1519
TimedSEESSI	1.70	0	2.40	3.92	190
TIMED/SEE-XPS	76.07	0	79.66	83.07	8
UARS/SOLSTICE	4.79	0	6.68	6.68	301
UARS/SUSIM	1.62	0	2.34	14.05	296

Table A3. A list of all reference spectra currently available in the SOLID database.

Name	Reference	Wavelength (nm)	Time
Arvesen 1969	Arvesen et al. (1969)	205–2,495	01 Nov 1969
Thekaekara 73	Thekaekara (1974)	115–400,000	01 Jan 1973
Hall Anderson 78	Anderson & Hall (1989)	200–310	01 Jan 1978
Kurucz 04	Kurucz (1992)	200–200,000	01 Jan 1994
Kurucz 05	Kurucz (2006)	299–1,000	01 Jan 1995
Colina96	Colina et al. (1996)	119.5–2,500	01 Jan 1996
ATLAS Comp.1	Thuillier et al. (2004)	0.5–2,397	26 Dec 2004
ATLAS Comp. 3	Thuillier et al. (2004)	0.5–2,397	26 Dec 2004
WHI-2008 Reference Spectra	Woods et al. (2009)	0.9–2,399	27 Mar 2008

Table A4. Total of 14 available proxies. For the SOLID database we truncate all proxies at 1969.

Name	Full Name References and web sites	Provider	Time interval
ISN	International Sunspot Number SILSO World Data Center (1970–2015) http://sidc.oma.be/silso/ (full daily coverage since Dec. 1848)	WDC-SILSO	1818 – present
$f_{10.7}$	10.7 cm Radio Flux Tapping (2013), ftp://ftp.geolab.nrcan.gc.ca/data/solar_flux/daily_flux_values/,	Penticton	1947 – present
$f_{3.2}$	3.2 cm Radio Flux	Nobeyama	1950s – present
f_8	8 cm Radio Flux		
f_{15}	15 cm Radio Flux		
f_{30}	30 cm Radio Flux http://solar.nro.nao.ac.jp/norp/html/daily_flux.html		
MgII	Mg II Core-to-Wing Ratio Viereck et al. (2004), ftp://laspftp.colorado.edu/pub/solstice/composite_mg2.dat	LASP, Boulder	1978 – present
MPSI	Magnetic Plage Strength Index http://www.astro.ucla.edu/~obs/150_data.html	Mt. Wilson Obs.	1970–2013
MWSI	Mount Wilson Sunspot Index http://www.astro.ucla.edu/~obs/150_data.html	Mt. Wilson Obs.	1970–2013
SEM 0-order	Intensity in 26–34 nm Band Dominated by the Hell line	SOHO/SEM	1995 – present
CaK	Normalized Intensity of Ca IIK Keil et al. (1998), ftp://ftp.nso.edu/idl/cak.parameters, http://nsosp.nso.edu/cak_mon	Sacramento Peak	1976 – present
DSA	Daily Sunspot Area Wilson & Hathaway (2006), http://solarscience.msfc.nasa.gov/greenwch/	Greenwich Obs.	1874 – present
PSI	Photometric Sunspot Index Balmaceda et al. (2009)		1874–2013

Table A5. Composite time series, their wavelength ranges, and time intervals. Below each entry we provided a reference and the source URL.

Name Sources and references	Wavelength (nm)	Time interval
DeLand SSI Composite DeLand & Cebula (2008), http://lasp.colorado.edu/lisird/cssi/	120.5–399.5	08 Nov 1978–01 Aug 2005
Ly α composite Woods et al. (2000), http://lasp.colorado.edu/lisird/lya/	121.0–122.0 0	2 Feb 1947 – present
Total Solar Irradiance Fröhlich (2006), http://www.pmodwrc.ch/pmod.php?topic=tsi/composite/ SolarConstant ftp://ftp.pmodwrc.ch/pub/data/irradiance/composite	Integrated	17 Nov 1978 – present

Appendix B. Data format

All data are provided in NetCDF CF-1.6 (see Footnote 2) format. In the following we give the detailed information provided in NetCDF, including global attributes, the dimensions of the data, the notation of variables used, instrument precision, instrument stability, instrument accuracy, and the use of flags.

B.1. Global Attributes

Each NetCDF file has one set of global attributes, i.e. attributes that are not bound to a specific variable. For our file format they are as follows:

creation_date Creation date of the dataset, given in "YYYY-MM-DD HH:MM:SS UTC+00"
title Name of the dataset
institution Name of the institution where the dataset was created, currently always LPC2E/CNRS.
reference Reference to the relevant publication that describes the original dataset.
reference_ads The reference key corresponding to the astrophysics data system key.

source The source of the data, this is usually a URL from where to retrieve the data.

instrument_version The data version as provided by the instruments team.

Conventions Convention used to name the variables and their properties. Currently this is the NetCDF Climate and Forecast (CF) Metadata Conventions, version 1.6 (see Footnote 2), given as "CF-1.6".

history A history of the dataset, i.e. when was it converted to which level.

SOLID_data_type The data type can be one of Instrument, Proxy, Reference, Composite, or Model.

SOLID_creation_date Date of the file creation of the SOLID data file.

SOLID_version SOLID data version. This specifies version of the used processing program.

SOLID_system Name of the platform where the program was run, e.g. 'Matlab 8.1.0.604 (R2013a)'

SOLID_level Processing level of the data, an integer between 0 and 5, corresponding to the processing steps described in this paper.

B.2. Dimensions

All our datasets have two dimensions, time and wavelength. For ease of use and generic application, the same scheme has also been applied to proxy data. In these cases, the wavelength dimension is the number of proxies in the data file and the wavelength entries are negative indices, that is -1 for the first proxy, -2 for the second, and so on. The names of the proxies are stored in the field proxy_names.

time The number of distinct time points.

wavelength The number of distinct wavelengths in the data file.

B.3. Variables

First, a general description of a variable with the main common attributes, followed by specific variables.

B.3.1. Variable Name(dimension1, dimension2, ...)

The name of the variable containing the following attributes. Here we give a short description of the attribute names generic to all variables. In the subsequent sections, we provide the actual values of these attributes.

standard_name The standard name as mandated by CF-1.6. This is a unique and precise version of the variable name.

long_name The long name as mandated by CF-1.6. This is the human-readable format.

units The unit of the variable. This may depend on the dataset.

valid_min, valid_max Minimum and maximum values allowed for this variable.

missing_value A value that defines a missing value in the dataset.

B.3.2. time(time)

A one-dimensional array containing the time given in days since January first, 1980.

standard_name 'time'.

long_name 'Time'.

units 'days since 1980-01-01 00:00:00 UTC+00'.

calendar The calendar used, this is always 'standard'.

B.3.3. wavelength(wavelength)

This variable contains the second dimension of the data variable. For irradiance datasets this contains the wavelength in nanometers, otherwise it contains negative indices. While the variable name wavelength does not make sense in the latter case, we keep it for consistency, but use the correct values for the attributes.

standard_name 'radiation_wavelength' or 'data_index'.

long_name 'Wavelength' or 'Index'.

units Wavelength are given in nanometers, 'nm', unless the dataset does not contain irradiances. This is the case for e.g. proxy data, in which case the unit is "negative index" and the variable wavelength contains -1, -2, ..., -n.

valid_min If the variable contains irradiance data, it is '0', otherwise it is not set.

B.3.4. data(wavelength, time)

The actual data stored as a matrix of dimension time $\times$ wavelength. The values of most properties depend on the actual dataset. However, it does differ in the case of e.g. a proxy dataset.

standard_name For an irradiance dataset: 'downwelling_spectral_irradiance_in_vacuum', otherwise data dependent.
long_name 'Spectral Irradiance' for spectral irradiance, otherwise data dependent.
units For irradiance datasets we use 'W/(m2 nm)', otherwise data dependent.
valid_min For most datasets this is '0'.
ancillary_variables 'solid_flag solid_precision solid_stability instrument_precisioninstrument_stability instrument_accuracy' or a subset of these variables. This depends on the actual data.
SOLID_name The proxy tables have named indices that are given in this attribute, one for each wavelength index.

B.3.5. solid_flag(wavelength, time)

This is an ancillary variable of data, containing the quality flag mask of the dataset. Each processing step corresponds to an entry in the data mask. A description of all the values is presented in Section B.7. To determine the presence of a flag, one may either use the bitwise 'and' operator on integers and test for non-zero.

standard_name 'downwelling_spectral_irradiance_in_vacuum status_flag'.
long_name 'Spectral Irradiance Quality'.
units 'level'.
valid_range '0, 15'.
flag_masks '1, 2, 4, 8' as unsigned integers.
flag_meanings 'Interpolated Multiple_Time_Values AR_Fitted FinalInterpolation'.

B.3.6. solid_precision(wavelength, time)

This is an ancillary variable of data, containing our estimated precision of the dataset.

missing_value '-1.e+99'.
_FillValue '-1.e+99'. We use the same values for _FillValue and missing_value to indicate that the data has no default value.
standard_name 'downwelling_spectral_irradiance_in_vacuum standard_error'
long_name 'Spectral Irradiance Precision'.
units same as the units attribute of the data variable.
description 'Estimated Precision'.

B.3.7. solid_stability(wavelength, time)

This is an ancillary variable of data, containing our estimated stability of the dataset. Each entry corresponds to an entry in the data mask.

_FillValue '-1.e+99'.
missing_value '-1.e+99'.
standard_name 'downwelling_spectral_irradiance_in_vacuum standard_error'
long_name 'Spectral Irradiance Stability'.
units same as the units attribute of the data variable per year.
description 'Estimated Stability'.

B.4. instrument_precision(wavelength, time)

The precision as provided. Either time or wavelength dimensions can be missing. In these cases, the missing dimensions have to be broadcasted.

_FillValue '-1.e+99'.
missing_value '-1.e+99'.
standard_name 'downwelling_spectral_irradiance_in_vacuum standard_error'.
long_name 'Spectral Irradiance Precision'.
units Either the same as units attribute of the data variable, or, if a dimension has been collapsed, it is given without units, in which case it is relative to the data.
description 'Provided Precision'.

B.5. instrument_stability(wavelength, time)

The stability as provided. Either time or wavelength dimensions can be missing. In these cases, the missing dimensions have to be broadcasted.

_FillValue '-1.e+99'.
missing_value '-1.e+99'.
standard_name 'downwelling_spectral_irradiance_in_vacuum standard_error'.
long_name 'Spectral Irradiance Stability'.
units Either same as the units attribute of the data variable divided by year, or, if a dimension has been collapsed, it is '1/year', in which case it is relative to the data per year.
description 'Provided Stability'.

B.6. instrument_accuracy(wavelength, time)

The accuracy as provided. Either time or wavelength dimensions can be missing. In these cases, the missing dimensions have to be broadcasted.

_FillValue '-1.e+99'.
missing_value '-1.e+99'.
standard_name 'downwelling_spectral_irradiance_in_vacuum standard_error'.
long_name 'Spectral Irradiance Accuracy'.
units Either same as the units attribute of the data variable, or, if a dimension has been collapsed, it is given without units, in which case it is relative to the data.
description 'Provided Accuracy'.

B.7. Flags

A list of all flags used in the variable solid_flags described in Section B.3.5. The flags are represented as binary values and stored as sums, i.e. if the quality flag for a specific time and wavelength is 6, this data point is composed of multiple time values and has been re-fitted by the autoregression model. To obtain the flag of value f, compute $\text{flag}_{t,\lambda}(f) \leftarrow \lfloor 2^{1-f} q_{t,\lambda} \rfloor \bmod 2 = 1$

1: Missing data or obvious outlier This flag is set if the data has been interpolated due to missing data or data that is well outside the physical range.
2: Multiple time values Some data provide multiple values for the same time. If so, the average of those is used.
4: Regrid in Time This value indicates a re-gridding of the data in time
8: Outlier The autoregression fit model set this flag if this data has been flagged as an outlier and replaced by linear interpolation.
16: Proxy Interpolate This data-point has been replaced by a proxy-based model.

Appendix C. Methods

C.1. General Processing Routines

Here we define a short description of helper routines that are used for the main routines below.

$\#x$ Number of elements in array x
Shape(x) The size of the matrix in the form [n, m, ...]
Onset(x) A matrix of ones with the same size as another matrix x, similar functions are Zeros, True, Nan, False.

C.2. Basic Processing Routines

C.2.1. Noise Estimators

Here we present several estimators for the precision that can be used to give an internal noise estimate. This estimator's purpose is to estimate the noise introduced due to time shifting. In this work we used the wavelet noise estimator (Sect. C.2.4), which is a composite of the Donoho noise estimator (Sect. C.2.2) and the all component wavelet noise estimator (Sect. C.2.3). We did not use the proxy autoregression wavelet estimator (Sect. C.2.5) because despite the additional computational cost and, more importantly, its reliance on an external data source, it did not return significant different results.

C.2.2. Donoho Noise Estimator

The original wavelet noise estimator by Donoho. It calculates the median of the wavelet coefficients and normalizes it with regard to white noise. It works well for white noise, however, as discussed in Section 3 it underestimates low varying colored noise, e.g., brown and pink noise while overestimating fast varying noise, e.g., blue and violet noise.

def DonohoNoiseEstimator(d, $wn \leftarrow$ ('daubechies', 8))
Require: From WaveLab v8.05: MakeONFilter() and FWTPO()
  $q \leftarrow$ MakeONFilter(wn)
  $p_l \leftarrow \lfloor \log_2(\#d) \rfloor$
  $d^s \leftarrow d_{1..p_l,\#d..(\#d-p_l)}$
  $w \leftarrow$ FWTPO(d^s, 1, q)
  return median($|w_{p_l/2..p_l}|$)/0.6748

C.2.3. All Component Wavelet Noise Estimator

An adapted Donoho wavelet noise estimator. It calculates the median of all wavelet components coefficients and normalizes it with regard to white noise. As in the case of the Donoho wavelet noise estimator, it works well for white noise, however it underestimates all colored noises, e.g., brown, pink, blue, and violet noise.

def FullWaveletNoiseEstimator(d, $wn \leftarrow$ ('daubechies', 8))
Require: WaveLab v8.05
  $q \leftarrow$ MakeONFilter(wn)
  $p_l \leftarrow \lfloor \log_2(\#d) \rfloor$
  $d^s \leftarrow d_{1..p_l,\#d..(\#d-p_l)}$
  $w \leftarrow$ FWTPO(d^s, 1, q)
  return median($|w|$)/0.6748

C.2.4. Wavelet Noise Estimator

By combining the Donoho's wavelet noise estimator with the high-frequency noise estimator, it is possible to get a good estimate for all high-frequency components. This is the estimator used in this work.

def WaveletNoiseEstimator(d, $wn \leftarrow$ ('daubechies', 8))
Require: DonohoNoiseEstimator, FullWaveletNoiseEstimator
  $n^d \leftarrow$ DonohoNoiseEstimator(d, wn)
  $n^w \leftarrow$ FullWaveletNoiseEstimator(d, wn)
  return (n^d + 1:2n^w)/2.2

C.2.5. Proxy autoregressive Wavelet Noise Estimator

The simple wavelet noise estimator may overestimate the noise due to physical short-term variability and due to physical outliers. This can be compensated by removing physical signals using a multiple input single output autoregressor before estimating the noise via the wavelet noise estimator.

We use radio flux data as the additional multiple input component. While it would be possible to also use the neighboring wavelength as additional input, this approach has two drawbacks, first it makes the noise estimate dependent on the number of measured wavelengths by the instrument (and the spectral resolution) and it will remove noise that influences several wavelengths at once.

def WaveletautoregressiveNoiseEstimator(d, $wn \leftarrow$ ('daubechies', 8), $p \leftarrow$ radio))
Require: WaveletNoiseEstimator, Autoregressive MISO
  $p' \leftarrow p -$ mean(p)
  $d' \leftarrow d -$ mean(d)
  $d'^{\text{fit}} \leftarrow$ Autoregressive MISO(d, p', q = 3, mode = 'forward')
  return WaveletNoise($d' - d'^{\text{fit}}$)

C.3. SVD-Interpolation

def SVD-Interpolation(ts, $\epsilon \leftarrow 10^{-8}$)
  $d \leftarrow$ ts.Data
  $m \leftarrow$ IsNanOrNeighboor(d)
  $d_m \leftarrow$ Interp($d_{\neg m}$)
  for $i \leftarrow$ 1..10**:**
    converged $\leftarrow$ **False**
    while not $\leftarrow$ converged**:**
      (u, s, v) $\leftarrow$ svd$_i$(d)
      $d_m \leftarrow$ (usv)$_m$

converged $\leftarrow$ ($\|d_m - d\|_\infty < \epsilon \, \|d\|_\infty$)
return ts

def IsNanOrNeighboor(d)
$m' \leftarrow$ IsNan(d)
$m \leftarrow m'$ **bitor** $\begin{bmatrix} m'_{2..,*} \\ F\,F\cdots F \end{bmatrix}$ **bitor** $\begin{bmatrix} F\ \ F\cdots F \\ m'_{..\#_1 m'-1,*} \end{bmatrix}$
return m

C.4. Autoregression Outlier Detector

Here we describe a method to flag and remove outliers by an autoregressive model. For each wavelength we construct a 4th order AR model, calculate the error and flag, and replace all values temporarily by linear interpolation whose error lies outside a given σ. The σ can be adaptive by providing a proxy to detect real outliers, that is outliers caused by solar phenomena. The same procedure is applied to the proxy first. All outliers that are detected in 70% of the proxies increase σ by 2 for the detected outliers in the proxy data.

def AutoregressiveOutlierDetector(ts, $n \leftarrow 3$, proxy $\leftarrow$ radio proxies)
Require: $n > 0$
Require: shape(n) ϵ ([1], shape(y)]
if shape(n) = [1]:
$n \leftarrow n$ ones(y) ▷ Explode n to the *same size* as y
if proxy $\neq$ **empty** :
$p^m \leftarrow$ AutoregressiveOutlierDetector(ts $\leftarrow$ proxy, proxy $\leftarrow$ []) ▷ *First argument is the timeseries* 'proxy'
$n_{p^m} > 3 \leftarrow n_{p^m>3}+2$. ▷ *Relax condition for outliers detection*
$q^m \leftarrow$ ts.Quality = Interpolated) ▷ Interpolated = *1*
for λ **in** ts.Wavelength:
(ts.Data$_\lambda$, q_λ) $\leftarrow$ RmOutliers(ts.Data$_\lambda$, n, q^m_λ)
(ts.Quality $\leftarrow$ ts.Quality **bitor** q
return ts
def RmOutliers(y, $n \leftarrow 3$, mask $\leftarrow \emptyset$)
Require: $n > 0$
Require: shape(mask) ϵ {$\emptyset$, shape(y)}
Require: shape(n) ϵ {[1], shape(y)}
$f^* \leftarrow$ inf
$q \leftarrow$ False(y)
repeat
$f \leftarrow f^*$
(ϵ, f) $\leftarrow$ |y – arfit(y)|
$m \leftarrow (\epsilon > n\sigma_y)$ **or** mask
$y(m) \leftarrow$ interp$_{\text{linear}}(y(\neg m))$
$q(m) \leftarrow$ AR-Fitted ▷ AR-Fitted = 4
until $\left|\frac{f}{f^*} - 1\right| < 100\varepsilon_{\text{float}}$ **or** maximum iteration limit is reached
return (y, q)