



# Imaging the 511 keV Positron Annihilation Sky with COSI

Thomas Siegert<sup>1</sup> , Steven E. Boggs<sup>1,2</sup>, John A. Tomsick<sup>2</sup> , Andreas C. Zoglauer<sup>2,3</sup>, Carolyn A. Kierans<sup>4</sup>, Clio C. Sleator<sup>2</sup>, Jacqueline Beechert<sup>2</sup>, Theresa J. Brandt<sup>4</sup> , Pierre Jean<sup>5</sup>, Hadar Lazar<sup>2</sup>, Alex W. Lowell<sup>1</sup>, Jarred M. Roberts<sup>1</sup>, and Peter von Ballmoos<sup>5</sup>

<sup>1</sup> Center for Astrophysics and Space Sciences, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0424, USA; [tsiegert@ucsd.edu](mailto:tsiegert@ucsd.edu)

<sup>2</sup> Space Sciences Laboratory, University of California, Berkeley, 7 Gauss Way, Berkeley, CA 94720-7450, USA

<sup>3</sup> Berkeley Institute for Data Science, University of California, Berkeley, CA 94720-7450, USA

<sup>4</sup> NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA

<sup>5</sup> IRAP, 9 Av colonel Roche, BP44346, F-31028 Toulouse Cedex 4, France

Received 2020 April 12; revised 2020 May 20; accepted 2020 May 21; published 2020 July 1

## Abstract

The balloon-borne Compton Spectrometer and Imager (COSI) had a successful 46-day flight in 2016. The instrument is sensitive to photons in the energy range 0.2–5 MeV. Compton telescopes have the advantage of a unique imaging response and provide the possibility of strong background suppression. With its high-purity germanium detectors, COSI can precisely map  $\gamma$ -ray line emission. The strongest persistent and diffuse  $\gamma$ -ray line signal is the 511 keV emission line from the annihilation of electrons with positrons from the direction of the Galactic center. While many sources have been proposed to explain the amount of positrons,  $\dot{N}_{e^+} \sim 10^{50} e^+ \text{ yr}^{-1}$ , the true contributions remain unsolved. In this study, we aim at imaging the 511 keV sky with COSI and pursue a full-forward modeling approach, using a simulated and binned imaging response. For the strong instrumental background, we describe an empirical approach to take the balloon environment into account. We perform two alternative methods to describe the signal: Richardson–Lucy deconvolution, an iterative method toward the maximum likelihood solution, and model fitting with predefined emission templates. Consistently with both methods, we find a 511 keV bulge signal with a flux between  $0.9$  and  $3.1 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{ s}^{-1}$ , confirming earlier measurements, and also indications of more extended emission. The upper limit we find for the 511 keV disk,  $<4.3 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{ s}^{-1}$ , is consistent with previous detections. For large-scale emission with weak gradients, coded aperture mask instruments suffer from their inability to distinguish isotropic emission from instrumental background, while Compton telescopes provide a clear imaging response, independent of the true emission.

*Unified Astronomy Thesaurus concepts:* [Gamma-rays \(637\)](#); [High altitude balloons \(738\)](#); [Interstellar medium \(847\)](#); [Cosmic rays \(329\)](#); [Astronomy data modeling \(1859\)](#)

## 1. Introduction

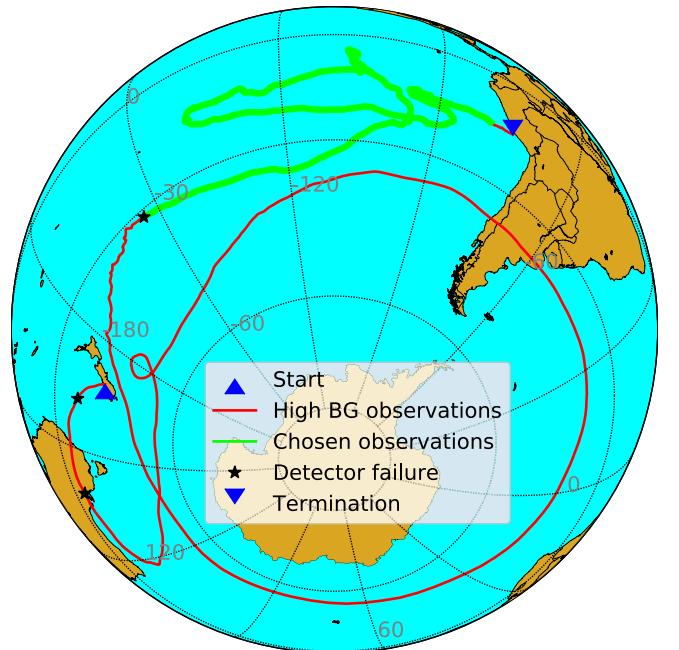
The “511 keV positron puzzle” is one of the long-standing unresolved problems in current astrophysics (see, e.g., Prantzos et al. 2011, for the latest review). In the center of the Galaxy, the strongest, persistent, diffuse  $\gamma$ -ray line signal originates from the annihilation of electrons with positrons (Johnson & Haymes 1973; Leventhal et al. 1978). The true origin of these positrons, however, is unknown and difficult to determine. While the emission itself is bright, on the order of  $10^{-3}$  photons  $\text{cm}^{-2} \text{ s}^{-1}$  (e.g., Purcell et al. 1997; Churazov et al. 2005, 2011; Knoedlseder et al. 2005; Jean et al. 2006; Weidenspointner et al. 2008; Bouchet et al. 2010; Skinner et al. 2014; Siegert et al. 2016b, 2019a), the annihilation morphology alone is believed to show only the annihilation sites and not the positron sources. The propagation of positrons away from candidate sources possibly leads to a smearing effect, which in turn might result in the diffuse 511 keV emission associated with the warm and partially ionized interstellar medium (e.g., Guessoum et al. 2006; Prantzos 2006; Higdon et al. 2009; Jean et al. 2009; Alexis et al. 2014; Panther 2018). Nevertheless, it is still reasonable to assume that not all positrons escape their production sites and annihilate in situ (e.g., Milne & Leising 1997), which could lead to a quasi-diffuse emission built from many point-like sources, such as flaring stars (Bisnovatyi-Kogan & Pozanenko 2017) or low-energy pair-plasma production in X-ray binaries (Bouchet et al. 1991; Sunyaev et al. 1992; Guessoum et al. 2006; Weidenspointner et al. 2008; Siegert et al. 2016a).

The distinction between true diffuse emission and the cumulative effect of a population of point-like sources is difficult to measure in  $\gamma$ -rays because the sensitivity of today’s instruments suffers from strong instrumental background, and the apertures can only provide a spatial resolution of the order of degrees. The pioneering instruments OSSE aboard CGRO (Johnson et al. 1993) and SPI aboard INTEGRAL (Vedrenne et al. 2003; Winkler et al. 2003) provided valuable insights into the true morphology of the positron annihilation emission. OSSE, with its four scintillation collimators (spatial resolution  $3.8^\circ \times 11.4^\circ$ , spectral resolution  $\approx 7\%$  at 511 keV), provided a first image reconstruction of the Galactic 511 keV line, showing a bright bulge and a possibly truncated disk (Purcell et al. 1993, 1997). After initial observations from balloon experiments found the Galactic emission to be apparently variable with time, results from OSSE finally resolved the signal to truly be extended and steady (Lingenfelter & Ramaty 1989; Purcell et al. 1997). The possible monopolar emission toward the north Galactic pole that was reported by OSSE, however, has not been verified by other instruments. SPI has been operating in space for 18 yr, and with its high-purity germanium (Ge) detectors, the 511 keV line and other positron annihilation emission features have been finely resolved (0.4% spectral resolution; e.g., Churazov et al. 2005, 2011; Jean et al. 2006; Weidenspointner et al. 2008; Siegert et al. 2016b, 2019a). SPI’s  $2^\circ 7$  resolution is achieved

by a coded aperture mask, and it could possibly identify individual 511 keV point sources. Such “smoking-gun” evidence is still missing. Instead, after several years of observation, SPI found the long-sought Galactic disk in positron emission (Bouchet et al. 2010; Skinner et al. 2014; Siegert et al. 2016b), which was expected from the proposed origins of positrons related to star formation. A study of possible “granularity” in the emission has been restricted to the bright bulge region (see discussion in Knoedlseder et al. 2005), but a clear characterization is still missing. Neither spiral arms nor individual positron production sites have been consistently detected. Nevertheless, different Galactic sources, such as massive stars (e.g., Oberlack et al. 1996; Diehl et al. 2006; Kretschmer et al. 2013; Pleintinger et al. 2019), core-collapse supernovae (e.g., Iyudin et al. 1997; Vink et al. 2001; Grebenev et al. 2012; Grefenstette et al. 2014, 2017; Boggs et al. 2015; Siegert et al. 2015; Tsygankov et al. 2016), and thermonuclear supernovae (e.g., Morris et al. 2006; Churazov et al. 2014, 2015; Diehl et al. 2014, 2015; Isern et al. 2016) have been shown to produce  $\beta^+$ -unstable nuclei, and microquasars have been claimed to produce pair plasma (Bouchet et al. 1991; Sunyaev et al. 1992; Siegert et al. 2016a).

A development toward a better understanding of this puzzle is provided by the usage of modern Compton telescopes in combination with high-resolution detectors. The Compton Spectrometer and Imager (COSI; Tomsick et al. 2019) is designed as a compact Compton telescope, which utilizes multiple Compton scatters in cross-strip Ge detectors to identify the direction of incoming photons. COSI mounts 12 detectors, each measuring  $8 \text{ cm} \times 8 \text{ cm} \times 1.5 \text{ cm}$ , in a  $2[x] \times 2[y] \times 3[z]$  configuration, leading to a total active volume of  $972 \text{ cm}^3$ . Five sides of the detector array are surrounded by a CsI anticoincidence shield, leading to a field of view of  $\approx\pi \text{ sr}$ . COSI is a nonpointing, i.e., free-floating, survey instrument, operating as a payload of a superpressure balloon. After shorter previous flights (see, e.g., Bandstra et al. 2011, for an overview), COSI observed the southern sky for 46 days between 2016 May and July (Kierans et al. 2016). The current COSI design leads to a spatial resolution of  $\approx 5^\circ$ , with a spectral resolution of  $\approx 0.7\%$  ( $\approx 3.5 \text{ keV FWHM}$ ) at 511 keV. With an upgraded future version in space, COSI would be a leading next-generation  $\gamma$ -ray telescope with superior background rejection and thus increased sensitivity. This is further supported by having more detectors and therefore a larger active volume, resulting in better event reconstruction (e.g., von Ballmoos et al. 1989; Boggs & Jean 2000) and better spatial resolution.<sup>6</sup>

In order to show the unique capabilities of compact Compton telescopes, in this study we perform a rigorous imaging analysis of the 511 keV positron annihilation line in the Milky Way, using the data from the 2016 balloon flight of COSI. This paper is structured as follows: In Section 2, we describe the 2016 balloon campaign, the data space intrinsic to Compton telescopes, and our specific data selection and preparation. In Section 3, we show the spatial analysis of the 511 keV line, provide our general approach for modeling the COSI data (Section 3.1), and give details about the imaging and



**Figure 1.** COSI flight path around Earth from its launch in Wanaka, New Zealand ( $45^\circ \text{ S}$ ,  $169^\circ \text{ E}$ , UTC 2016 May 16 23:35), until termination in Peru ( $16^\circ \text{ S}$ ,  $72^\circ \text{ W}$ , UTC 2016 July 2 19:54). The green line shows the chosen and analyzed data set. High-background observations (red; see also Figure 2) are excluded from the analysis. The failures of three main detectors are marked by black stars.

background response of a Compton telescope in a balloon environment (Sections 3.2 and 3.3). Imaging is performed both by an iterative deconvolution approach using a modified version of the Richardson–Lucy algorithm (Section 4.1) and in a full-forward modeling manner (Section 4.2), based on the imaging results to identify significant structures. Section 5 closes with a comparison to previous measurements and an outlook for future analyses.

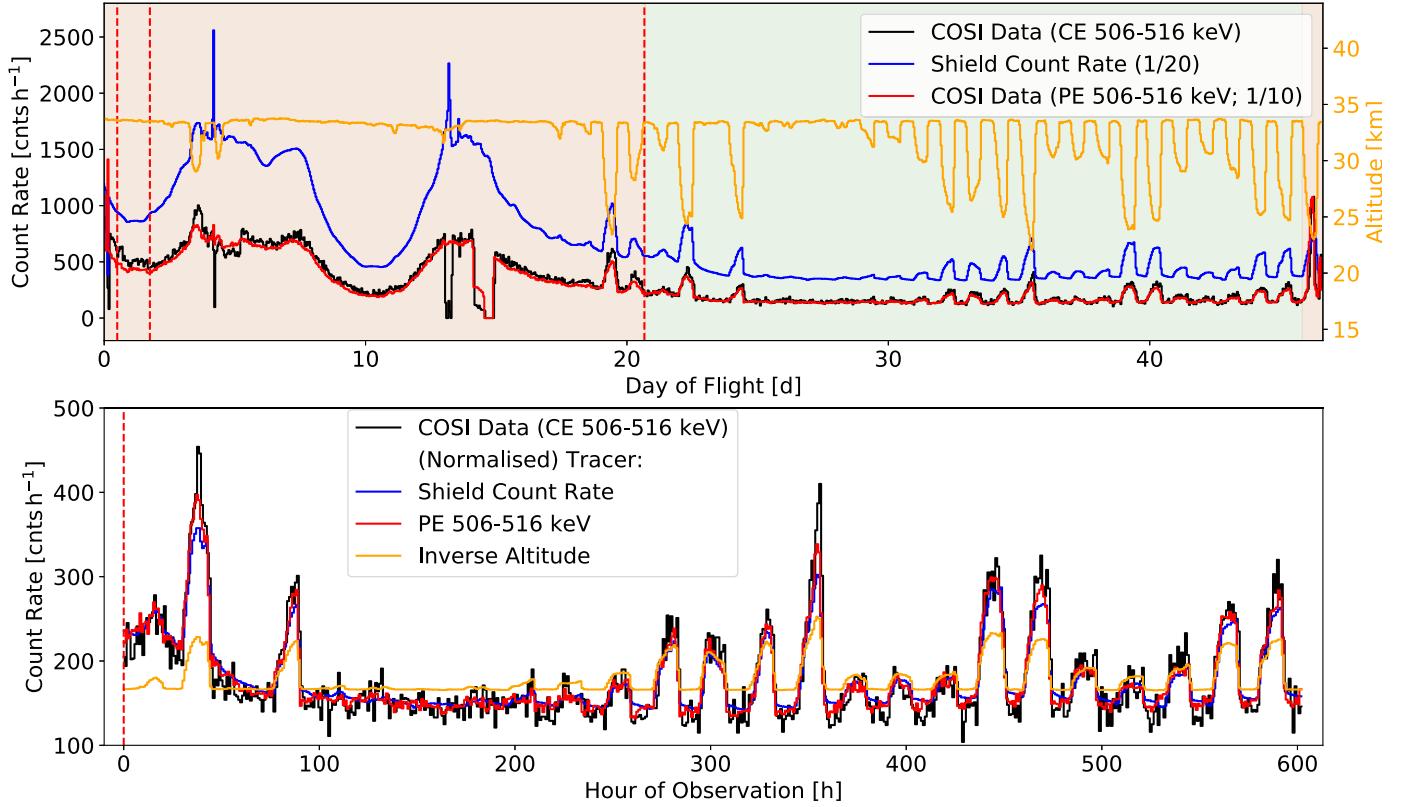
## 2. 2016 Campaign and Data Set

### 2.1. 2016 Balloon Flight

The 46-day balloon flight of COSI in 2016 started on May 17 in Wanaka, New Zealand, and was terminated 200 km northwest of Arequipa, Peru, on July 2. The nominal flight altitude was about 33 km, with anomalous altitude drops related to day and night cycles (see Section 2.2.1). During the flight, three detectors failed, reducing the sensitivity of the instrument by  $\approx 40\%$  (see Section 3.2). Because two of the malfunctions occurred in the top layer of COSI, the reduction is not proportional to the number of detectors. The flight path of the balloon is shown in Figure 1, indicating the time and position of the detector failures, as well as the selected data set for our analysis (see Section 2.2). The circumpolar winds carried the payload around Antarctica once in  $\approx 14$  days before the balloon drifted toward the equator and finally landed on the west coast of South America. Details about the 2016 balloon flight can be found in Kierans et al. (2016, 2020).

The red path indicates times/regions in which the instrumental background rates were high and that are excluded in our data set (see Section 2.2 for details). In Figure 2, we show the measured count rate of 511 keV photons (506–516 keV), detected via multiple scatters (Compton events (CE); black

<sup>6</sup> Note that the angular resolution of Compton telescopes is ultimately restricted to  $\approx 1^\circ$  owing to the intrinsic motion of electrons in the Ge lattice, leading to an inevitable Doppler broadening (Zoglauer & Kanbach 2003). Beyond this resolution, either narrow collimators or Laue lenses would be required.



**Figure 2.** Measured 511 keV Compton event count rate (CE; black) during the 2016 COSI flight (top) and for the chosen data set (bottom), as a function of time. The red and green shaded areas indicate the regions of high and low background, respectively. Red dashed lines indicate times of detector failures (see also Figure 1). In the top panel, also the shield count rate (blue, scaled by 1/20), the 511 keV photoevent rate (PE; red, scaled by 1/10), and the altitude (right axis, orange) are shown. The bottom panel compares qualitatively the count rate of the chosen data set (see Section 2.2) with potential background tracers (see Section 3.3.3), normalized to the average 511 keV count rate during the selected data set (green shaded region).

histogram), as well as complementary other rates. The green path in Figure 1 coincides with the green shaded region in Figure 2, identifying the chosen times for our data set. During days 0–21 of the flight (red shaded region), the 511 keV count rate varies between 200 and 1000 counts  $\text{hr}^{-1}$ , and no strong correlation with the flight altitude (orange, right axis) is seen. As illustrated in Figure 1, the balloon was floating at higher latitudes, which influences the geomagnetic cutoff rigidity and consequently the background rate. After day 29, frequent altitude drops lead to an increase of the CE count rate, as well as the photoevent rate (PE; red) and CsI shield (blue) count rate. These nearly one-to-one correlations will be used to empirically determine the variation of the instrumental background in Section 3.3.3, i.e., determining appropriate background tracers. The latter are shown in the bottom panel of Figure 2, normalized to the average 511 keV count rate during the selected data set (green shaded region).

## 2.2. COSI Data Space, Preparation, and Selection

As a Compton telescope, COSI records individual triggers in the position-sensitive active detector volume upon which event reconstruction is performed using the deposited energy and the kinematics of Compton scattering (e.g., von Ballmoos et al. 1989; Boggs & Jean 2000; Zoglauer et al. 2007). The stored parameters are then inherent to this measurement principle and include the total photon energy  $E$ ; the three scattering angles,  $\phi$

(Compton scattering angle;  $\in [0^\circ, 180^\circ]$ ),  $\psi$  (polar scattering angle;  $\in [0^\circ, 180^\circ]$ ), and  $\chi$  (azimuthal scattering angle;  $\in [-180^\circ, 180^\circ]$ ); and an absolute time tag. In addition, the aspect of COSI is saved independently as the pointing of the detector in  $x$  and  $z$  (optical axis) in both Galactic (longitude/latitude;  $l/b$ ) and horizon coordinate system (specifically to perform the Earth Horizon Cut; see Section 2.2.2).

### 2.2.1. Binned COSI Data

The COSI data space therefore consists of a tag for the time and energy of each event in the 3D  $\{\phi\psi\chi\}$  data space. In this work, we avoid treating each photon individually and define a binned data space in scattering angles. This is typically referred to as the “COMPTEL (or Compton) data space” (von Ballmoos et al. 1989; Diehl et al. 1992). Any narrow binning of the angles, e.g., with a bin size of  $1^\circ$  (corresponding to  $180 \times 180 \times 360 = 11,664,000$  bins), immediately results in an enormous number of data points to handle, and in fact would lead to a treatment similar to that of an unbinned analysis. As the spatial resolution of COSI is about  $5^\circ$ , this provides a natural choice for the angular binning since we expect a signal of about  $7\sigma$  (Kierans et al. 2020) to be distributed over the bulge region, thus avoiding an unmanageably large image data space. We divide the Compton scattering angle,  $\phi$ , into 36 regular  $5^\circ$  bins. The remaining ( $\psi/\chi$ )-sphere is cut into 1650 irregular 2D bins with equal solid angles (see Zoglauer et al. 2006). The resulting

$\{\phi\psi\chi\}$  data space thus contains 59,400 scattering angle bins (see below for further reduction).

The Ge detectors resolve an instrumental 511 keV line with an FWHM of about 3.5 keV. The observed astrophysical broadening of the narrow 511 keV component is about 2.0 keV (e.g., Jean et al. 2006; Churazov et al. 2011; Siebert et al. 2019a), resulting in a combined Doppler broadening of  $\approx 4$  keV. Thus, 99.7% ( $3\sigma$ ) of the expected counts of the 511 keV line are included in a band of  $\approx 10$  keV. We select only photons that fall into the energy interval [506, 516] keV for our data set (one energy bin). For a resolved COSI spectrum around the positron annihilation line, we refer to Figure 6.5 in Kierans (2018) and the spectral analysis in Kierans et al. (2020).

Since COSI is, to first order, zenith pointing, and is additionally moving around Earth, the time intervals used for the analysis should not be too long, because different exposures with and without signals will be combined together in time. They should also not be too short, as the limited number of counted photons would lead to an unnecessarily large data space. Here we adopt a time binning of 1 hr, resulting in 603 time bins of active observations, and weight different impacts on the imaging and background response accordingly within each hour (see Section 2.2.3).

We can further reduce the number of data space bins since many bins are never occupied in the (selected) data set (Section 2.2.2), the background (Section 3.3.2), or the imaging response (Section 3.2). This leads to a reduced data space,  $\{\phi\psi\chi\}_R$ , with 4243 scattering angle bins. The total number of bins in the predefined data space is thus  $4243 \times [E] \times 603 [T] = 2,558,529$ .

### 2.2.2. Event Selections

In the above-described data space, we further select events that follow more detailed quality criteria: While a lower altitude increases the background count rate, these drops (see Figure 2) happen mainly during observations of the Galactic center, i.e., when the strongest signal is expected in 511 keV. We therefore do not restrict our data set to a specific altitude interval and rather use the full response (see Section 3.2) to estimate the expected count rate (see, however, Appendix A for alternative selections as a reliability cross-check). This avoids “optimizing for the signal” (e.g., Koehler 1993; Nickerson 1998; Pohl 2004, pp. 79–96) as can typically happen in background-dominated measurements with an apparently “known” outcome.

We further use all exposed regions during the 603 selected hours, as this allows the background to be properly defined using regions that are expected to be empty, as well as to search for 511 keV disk emission. For the individual events, we only select those with a kinematic Compton reconstruction chain length (number of interactions in the detectors) of 2–7. Events with three or more scatters provide redundant information in the reconstruction, leading to higher fraction of correctly reconstructed events (Zoglauer 2006). The angular resolution of COSI at 511 keV is dominated by the position resolution owing to the strip pitch in the Ge detectors. As a consequence, events for which the first and second interactions are farther apart have better angular resolution. Using minimum distances of 0.5 cm between the first two interactions and 0.3 cm between subsequent interactions inside the detectors is found to be a good compromise between improving the angular resolution and reducing the detector efficiency. The Compton scattering

**Table 1**  
Event Selections Used for the 511 keV Imaging Analysis

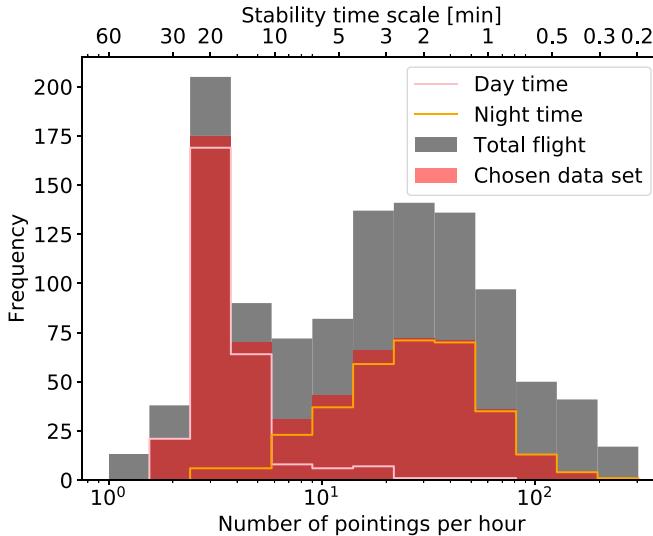
Parameter	Selection
Energy (keV)	[506, 516]
Time (MJD)	[57,545.78, 57,570.86]
Number of interactions	[2, 7]
Interaction distance (cm)	>0.5 (first 2); >0.3 (any)
Compton Scattering Angle	[0°, 60°]
Altitude (km)	[22, 35] (all; see Appendix A)
Pointing (coordinates)	full sky (full exposure)
Earth Horizon Cut	yes

angle itself provides a quality measure, as potential back-scatters ( $>90^\circ$ ) are difficult to reconstruct. We further select  $\phi$  according to the imaging response quality for larger angles (see Section 3.2), being less and less populated for angles larger than  $60^\circ$ . Since the Earth Horizon Cut (see below) removes any events above  $90^\circ$  and significantly reduces the numbers between  $60^\circ$  and  $90^\circ$ , we restrict  $\phi$  to  $\leq 60^\circ$ . The Earth Horizon Cut rejects Compton events that, projected back onto the celestial sphere, would be intersecting with the Earth horizon. This largely avoids albedo radiation, i.e., a physical background to our 511 keV measurements. The specific event selections are summarized in Table 1. The total number of photons in our data set is then  $N_{\text{ph}} = 107,880$ . Thus, only  $\approx 4.2\%$  of the data space is populated, and many bins carry zero counts. This requires a proper statistical treatment using Poisson statistics (see Section 3.1).

### 2.2.3. Balloon Stability and Pointing Definition

In each of the 603 observation hours, the balloon gondola’s absolute position (aspect) is changing. This means that either the observation direction ( $z$ -axis) or the detector plane ( $xy$ ) changes from one instance in time to another. This has to be taken into account when applying the instrument response for different times, and also within a single time bin of 1 hr. We define pointings of COSI observations, i.e., over which the imaging response is applied, by a stability criterion of the gondola: the times until the normal vectors of any instrument plane change by more than  $5^\circ$  are accumulated and saved as weighting factors for the imaging response within individual time bins. Such a treatment considers the steady slew of the instrument, as well as intrinsic rotation and tumbling of the payload.

In total, this evaluates to 35,938 pointings for the whole flight and 11,922 for the 603 1 hr time bins of our selection. Figure 3 shows the distribution of pointing lengths for the complete 46-day flight (gray) and the chosen data set (red). Clearly, the distribution is bimodal, which arises from the day and night times: during daylight, a rotator below the balloon steers the payload such that the solar panels are optimally exposed by the Sun. This provides a smooth behavior of the instrument aspect and is only slightly disturbed by altitude changes (first peak; pink histogram). The stability timescale peaks at 20 minutes (corresponding to  $\approx 5^\circ(20 \text{ minutes})^{-1} = 15^\circ \text{ hr}^{-1}$ , i.e., the rotation speed of Earth), so that only a few pointings are required to define the 1 hr time bins. At night times, the rotator is turned off and the payload more freely rotates about its zenith, which leads to a stronger influence of the environment. The



**Figure 3.** Number of stable pointings per hour of observation as given by the criterion in Section 2.2.1 with a  $5^\circ$  threshold. The gray (red) shaded histogram shows the full (selected) data set. Separating the observations in day and night time explains the bimodality of the distribution: during day times, the balloon orientation changes only every  $\sim 20$  minutes, i.e., about the rotation velocity of Earth of  $15^\circ \text{ hr}^{-1}$ . At night, rotation, tumbling, and vast altitude changes make individual pointings unstable so that the response (see Section 3.2) during 1 hr has to be reweighted more often.

stability timescale peaks around 2 minutes, so that on average  $\approx 30$  pointings have to be included in 1 hr.

The distributions of  $\phi$ ,  $\psi$ , and  $\chi$  for each observation hour have been investigated to allow for a similar reweighting of the background response as a function of time, altitude, and position on Earth. These distributions are constant with respect to all observation-specific parameters, so that we can safely assume the background response to be independent of the instrument aspect. We would expect that the background response also shows a weak dependence on the balloon altitude, but which has not been observed. Introducing such a dependence would probably be required for longer flights when these trends become important. We note that the amplitude of the background still shows the expected correlation with balloon altitude, which will be taken care of when defining the background model (see Sections 3.3 and 3.3.2 for further details).

### 3. Spatial Analysis

In this section, we will describe two approaches for inferring information about the spatial distribution of 511 keV emission in the Galaxy from COSI data. First, we will introduce the basic principle for full-forward modeling in the COSI-specific data space (Section 3.1), where we include the effect of the dynamic aspect of the balloon gondola in the imaging response (Section 3.2) and the variability of the instrumental background with altitude (Section 3.3). For a rather model-independent approach to determine the emission morphology, we use an adapted version of the Richardson–Lucy deconvolution algorithm in Section 4.1. This provides a baseline for the use of empirical functions in a full-forward fitting approach to reliably characterize the flux and extent of the Galactic 511 keV emission as seen by COSI (Section 4.2). By using these two

methods, we can cross-check different modeling assumptions and provide consistency and systematics estimates.

#### 3.1. General Approach

We model the number of counts in a data space bin,  $\{\phi\psi\chi_t\}$ , as a linear combination of sky model,  $m_{\phi\psi\chi_t}^{\text{SKY}}$ , and background model components,  $m_{\phi\psi\chi_t}^{\text{BG}}$ , such that

$$\begin{aligned} m_{\phi\psi\chi t} &= m_{\phi\psi\chi t}^{\text{SKY}} + m_{\phi\psi\chi t}^{\text{BG}} \\ &= \int_{d\Omega} \cos(b) db dl \sum_{p_t \in t} R_{\phi\psi\chi}^{\text{SKY}}(Z, A, h) \\ &\quad \cdot p_t(Z, A) \leftrightarrow (l, b) \cdot M(l, b; \theta_s) \\ &\quad + R_{\phi\psi\chi}^{\text{BG}} \cdot T_t(\theta_b). \end{aligned} \quad (1)$$

In Equation (1),  $R_{\phi\psi\chi}^{\text{SKY}}(Z, A, h)$  is the imaging response of COSI as a function of zenith ( $Z$ ), azimuth ( $A$ ), and altitude ( $h$ ), which is mapping the sky model,  $M(l, b; \theta_s)$ , to the COSI data space,  $\{\phi\psi\chi_t\}$ , by integrating over the exposed sky region,  $d\Omega$ , weighted by the pointings' time,  $p_t \in t$ , defined in each time bin,  $t$ , which also links the internal zenith/azimuth coordinate system to Galactic coordinates,  $(Z, A) \leftrightarrow (l, b)$ . The description of the spatial distribution of photons in image space is parameterized either by a differential flux value per individual pixel (Richardson–Lucy deconvolution; Section 4.1) or by a set of sky model parameters,  $\theta_s$ , which can include the shapes, extents, and flux normalizations of a multitude of morphologies, such as individual point sources or extended emission (Section 4.2). The background response,  $R_{\phi\psi\chi}^{\text{BG}}$ , describes the expected distribution of photons in the data space and is constant in time and altitude. The absolute rate of background can change with time such that its temporal variability is included by a tracer function,  $T_t(\theta_b)$ , and parameterized by a set of background parameters,  $\theta_b$ , which can include time nodes and various amplitudes (see Section 3.3).

In this way, the total model counts are predicted as parameterized through  $\theta_s$  and  $\theta_b$ , such that  $m_{\phi\psi\chi t}(\theta_s, \theta_b)$  will be unitless (number of photons). Because this describes a counting experiment, the distribution of photons in each data space bin follows the Poisson statistics, and therefore the total model is determined by maximizing the Poisson likelihood,

$$\mathcal{L}(d|m) = \prod_{\phi\psi\chi t} \frac{m^d \exp(-m)}{d!}, \quad (2)$$

with  $d$  being the measured counts in each data space bin  $\{\phi\psi\chi_t\}$ . The general description of  $M(l, b; \theta_s)$  predicts differential fluxes in units of photons  $\text{cm}^{-2} \text{s}^{-1} \text{sr}^{-1}$ . Applying the imaging response,  $R_{\phi\psi\chi}^{\text{SKY}}(Z, A, h)$  (in units of  $\text{cm}^2$ ), to a sky model for a certain pointing duration,  $p_t$  (in units of s), is computationally very expensive for a particular combination of spatial and amplitude parameters. This would be required in each step of a likelihood maximization. However, the same spatial parameters (e.g., the position or the width of a 2D Gaussian; see Section 4.2.1) predict the same relative numbers in the  $\{\phi\psi\chi_t\}$  data space. For this reason, the amplitude (i.e., flux normalization) can be separated, as this parameter only scales the expected number in each bin up and down, but will not change the expected patterns. This means that the amplitude,  $\alpha_s$ , for each

sky model  $s$  can be handled independently of the “already-convolved sky models,”

$$\begin{aligned} m_{\phi\psi\chi t}^{\text{SKY},s} &= \int d\Omega \sum_{p_t \in t} R_{\phi\psi\chi}^{\text{SKY}}(Z, A, h) \cdot p_t \cdot M_s(l, b; \theta_s) \\ &= \alpha_s \cdot \int d\Omega \sum_{p_t \in t} R_{\phi\psi\chi}^{\text{SKY}}(Z, A, h) \cdot p_t \cdot M_s(l, b; \theta_s^*) \\ &= \alpha_s \cdot m_{\phi\psi\chi t}^{\text{SKY},s,*}. \end{aligned} \quad (3)$$

In Equation (3), the set of sky model parameters is separated into a fixed set of parameters,  $\theta_s^*$ , and the amplitude:  $\theta_s = \{\theta_s^*, \alpha_s\}$ . In this way, for a specific (set of) model(s),  $m_{\phi\psi\chi t}^{\text{SKY},s,*}$  is only calculated once, and the flux is determined for (a set of) predefined, fixed, parameters;  $m_{\phi\psi\chi t}^{\text{SKY},s,*}$  is termed “convolved sky model.” This methodology will also be used when using the Richardson–Lucy deconvolution algorithm and its modification for accelerated convergence (Section 4.1.1).

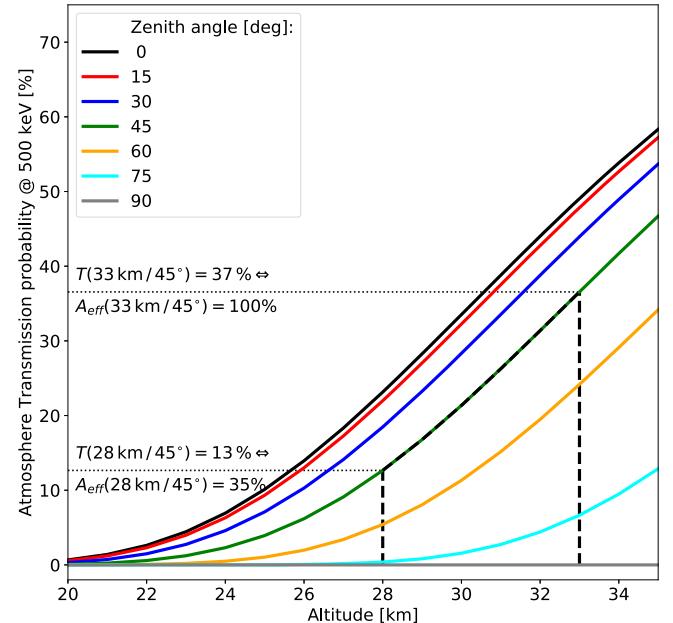
In contrast to the imaging response that is derived from simulations (see Section 3.2), the background response is determined purely empirically and will therefore be treated as being unitless. Details about how the background modeling is approached are given in Section 3.3.

### 3.2. Imaging Response

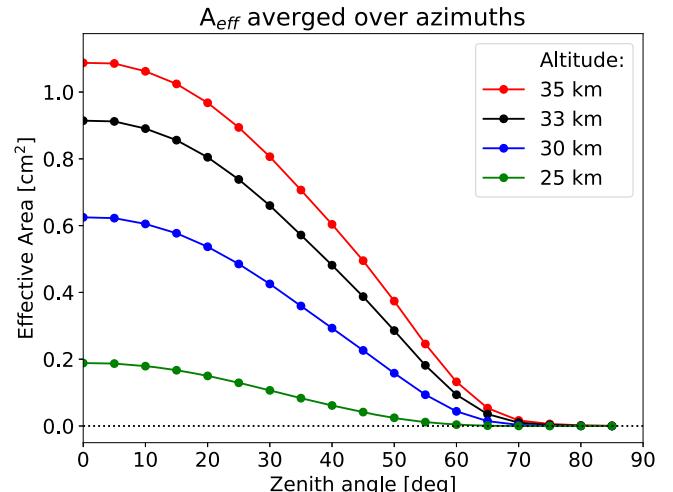
We use the Medium-Energy Gamma-ray Astronomy library (MEGAlib; Zoglauer et al. 2006) to simulate the expected number of photons at 511 keV as a function of the intrinsic zenith and azimuth coordinate system. Such a simulation requires a detailed mass model of COSI and has to take into account the three dead detectors, as well as the attenuation of the atmosphere at a specific altitude. The latter has a large impact on the resulting effective area as a function of zenith because more air mass has to be passed at the same zenith angle for lower altitudes.

The simulation setup places the mass model of COSI with nine functioning detectors in the center of an isotropically emitting sphere, at a nominal altitude of  $h = 33$  km (defining the transmission probabilities). The total number of simulated photons is  $\approx 2.65 \times 10^{12}$ , which took about 3.5 million CPU hours of computation time at the National Energy Research Scientific Computing Center’s supercomputer *Cori*. The simulated events then pass through a well-benchmarked detector effects engine (Sleator et al. 2019), making them appear like actual data (e.g., strip numbers instead of positions, AD units instead of energy). The simulated events then pass through the same calibration and analysis pipeline as the real data. After event reconstruction (Zoglauer 2006), the events are binned according to a predefined spacing in a 5D data space, defined by the zenith and azimuth angles in detector coordinates,  $(Z, A)$ , as well as the Compton data space,  $\{\phi\psi\chi\}$ . Here, on average, a  $5^\circ$  spacing is used. Finally, a 5D sky response is created:  $R_{\phi\psi\chi}^{\text{SKY}}(h = 33 \text{ km}; Z, A)$ .

Since the balloon altitude is changing between about 22 and 34 km in our selected data set, the sixth dimension of altitude has to be included as well. Instead of performing multiple simulations with ever-increasing computing time, we use the simulated response at 33 km to build a grid of relative transmissivities for zenith and azimuth as a function of altitude. In Figure 4, the altitude-dependent atmospheric transmission probability at 500 keV photon energies is shown for different zenith angles. In the indicated example, the absolute

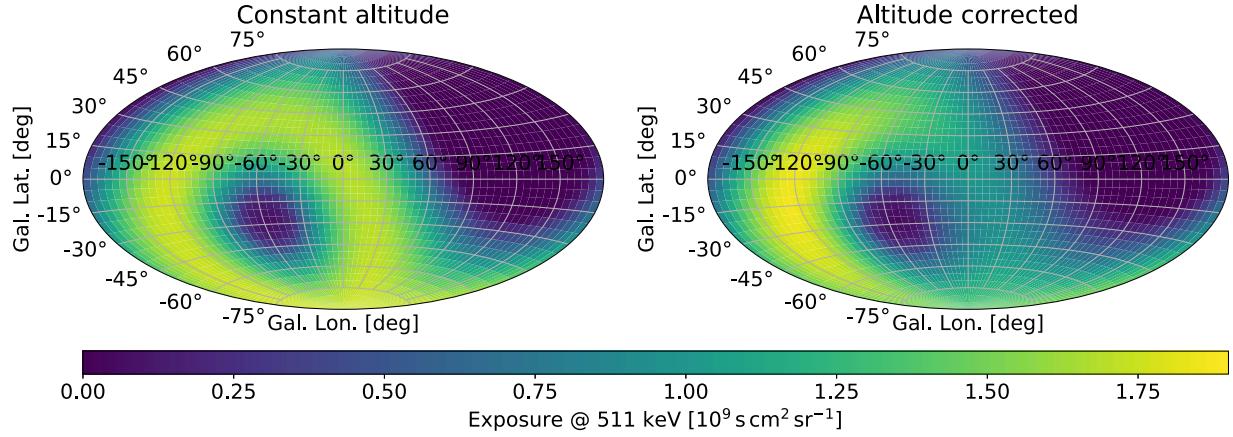


**Figure 4.** Renormalization of the aspect-dependent response at 500 keV as a function of different altitudes. The nominal response was calculated for a floating altitude of 33 km, corresponding to a transmission probability through the atmosphere for zenith angles of  $45^\circ$ , for example, of  $\approx 37\%$ . For the same aspect angle, the resulting effective area at 28 km altitude is reduced to 35%. See text for details.

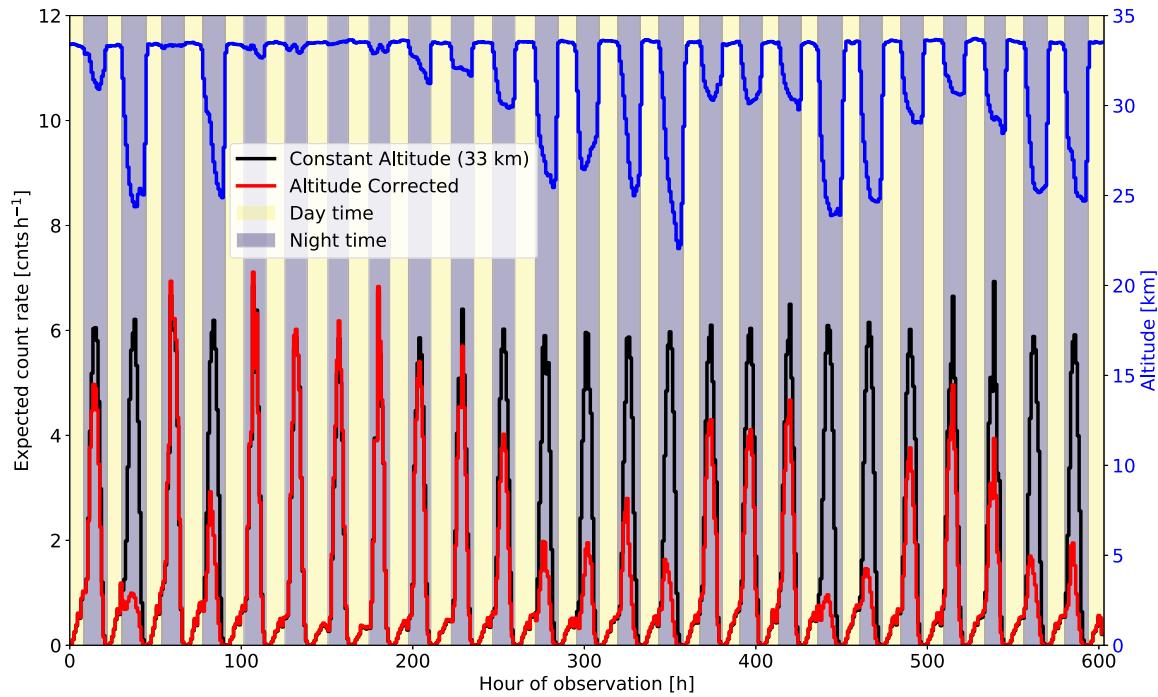


**Figure 5.** Absolute effective area at 511 keV, averaged over  $360^\circ$  of azimuths, as a function of zenith angle for different balloon altitudes. The altitude- and time-averaged effective area of the expected 511 keV signal in the Milky Way for the chosen data set is  $0.59 \text{ cm}^2$ .

transmission probability (transmissivity) at nominal altitude (33 km) for a zenith angle of  $45^\circ$  is 37%, which corresponds to a relative effective area of 100% (relative to the value at nominal altitude). At the same zenith angle, but at a considerably lower altitude, for example, 28 km, the absolute transmission probability is only 13%, for which the effective area is to be rescaled by  $13\% \times 100\% / 37\% = 35\%$ . We create a grid of altitudes from 20 to 35 km in 1 km steps and zenith angles between  $0^\circ$  and  $90^\circ$  in  $5^\circ$  steps to determine a renormalization for the absolute effective area of COSI around 500 keV photon energies. The resulting azimuth-averaged effective area is shown in Figure 5 for different altitudes.



**Figure 6.** Exposure map of the selected data set ( $T_{\text{obs}} = 603 \text{ hr} \approx 2.2 \text{ Ms}$ ) at 511 keV photon energy in units of  $10^9 \text{ s cm}^{-2} \text{ sr}^{-1}$ , assuming a constant altitude of 33 km (left) and correcting for atmospheric absorption as a function of altitude (right). The Galactic center is about 40% less exposed when taking the altitude change into account (see also Section 3.2).



**Figure 7.** Expected count rate for the 511 keV emission model of Siebert et al. (2016b) with (red) and without (black) correction for varying balloon altitudes (right axis, blue) as a function of time. Note that each 1 hr time bin consists of  $\approx 4000$  bins in the predefined (reduced) COSI  $\{\phi\psi\chi\}$  data space. Day and night times are indicated with bright and dark shading, respectively (Earth latitude/longitude dependent). The strong altitude drops mainly occur during night times, i.e., when the Galactic center is in COSI's field of view. The total loss due to low altitudes is about 38% and only affects the emission from the bulge. The expected disk emission (smaller bumps during day times) benefits from slightly higher altitudes than nominal during day times ( $\approx 5\%$ ).

It is evident from Figure 5 that the effective area is drastically changing with both zenith angle and altitude. While the effective area naturally decreases with zenith owing to the finite projected geometric area, the largest impact is still originating from the larger air mass that photons have to pass, reducing the effective area for larger zeniths even further. Also above adequate flight altitudes,  $\gtrsim 30$  km, the effective area at zenith varies by  $\approx 40\%$ . With nine functioning detectors at 33 km altitude, COSI's effective area at zenith is about  $0.91 \text{ cm}^2$ .

The altitude changes are mainly connected with day and night cycles. The strongest signal at 511 keV is expected to come from the Galactic bulge region. However, the bulge is mainly exposed at night, i.e., when the balloon's altitude drops; hence, the total exposure (in units of  $\text{cm}^2 \text{ s}$ ) is not uniform in

COSI's field of view throughout the flight. Figure 6 shows the total exposure of the chosen data set for a constant altitude (left) and corrected for the true motion (right). Especially at the Galactic center, the exposure is decreased by about 40%. But since this is the region in which most of the signal is expected, and because the more exposed regions in which no signal is expected provide a good basis for background estimates, all 603 hr of observation after the third detector failure are kept (see Section 2.2). In Figure 7, we show the expected number of photons from the empirical model for the 511 keV emission as found by Siebert et al. (2016b; see also Section 4.2.2), for a constant altitude (black) and the altitude-corrected response. Clearly, during night times (dark shaded areas), the altitude (blue) frequently drops, for which the effective area is reduced.

These times are expected to contribute most to the measured sky counts.

We want to note that assessing the quality of the response creation through simulations is difficult to benchmark under laboratory conditions for imaging diffuse emission on top of a large and varying background. Nevertheless, Sleator et al. (2019) performed a comprehensive study of detector effects that influence how individual event messages are recorded and the resulting spectral response and angular resolution of COSI. Among other effects, this included charge sharing between adjacent strips, charge loss, cross talk between electronic channels, and accurate threshold settings regarding timing and energy for veto systems. These effects are taken into account in the imaging response creation.

### 3.3. Background Modeling

The instrumental background in soft  $\gamma$ -ray telescopes is the dominant contributor to the measured count rate. A rough prediction of the expected background spectrum and its intensity can be made from expensive simulations using the full mass model of both the payload and the mount and the complete environmental conditions.

In space, a large portion of the instrumental background comes from the interaction of primary cosmic-ray and solar particles with the satellite and instrument materials (e.g., Gehrels 1985; Boggs et al. 2002; Jean et al. 2003; Cumani et al. 2019). Secondary, then lower-energy, particles ( $\sim$ MeV) lead either to nuclear excitations followed by de-excitation through the emission of  $\gamma$ -rays or other nuclear reactions, building short- and long-lived radioactive nuclei, which then also emit  $\gamma$ -rays after having decayed (activation, radioactive buildup). This constitutes a family of prompt and delayed  $\gamma$ -ray line emission. A dominant instrumental continuum background is created by  $\beta$ -particles depositing their energy inside detectors, or electromagnetic cascades induced by high-energy cosmic rays. Because the general cosmic-ray flux at Earth and consequently the instrumental background rate depend strongly on the 11 yr solar cycle, and furthermore on the unpredictable occurrence of solar flares, a physical background model, applicable for each time and position, is not feasible (Diehl et al. 2018; Siegert et al. 2019b).

The problem of instrumental background is complicated even more in a typical balloon environment: even though the atmosphere becomes more transparent at higher altitudes and for larger energies, cosmic-ray particles interact with the atmosphere and create a strong soft  $\gamma$ -ray continuum, i.e., the atmosphere is shining in hard X-rays and  $\gamma$ -rays (Earth albedo; Cumani et al. 2019). Furthermore, the bremsstrahlung from secondary electrons, for example, depends on the exact altitude of the payload (density of air, passed air mass, zenith/azimuth) and the position on Earth (geomagnetic cutoff). In both space and the atmosphere, primary emission photons, such as from the Galactic plane or extragalactic background light, will lead to downscattered  $\gamma$ -ray photons that might be seen as instrumental background.

Attempts to model the expected background behavior, especially at 511 keV photon energies, typically lead to a robust order-of-magnitude estimate. Proposing an absolute number of background counts for each observation, even given all appropriate environmental conditions and instrument-specific properties, would still require an uncertainty attached to these model predictions. These are difficult to determine.

The description of the low-energy background ( $\lesssim$ 10 MeV) at balloon altitudes, and in particular for the 511 keV line, by Ling (1975) is nevertheless useful to perform simulations and assess the adequacy of background modeling and parameter inference. This model was later also shown to provide a good description of one of the first measurements of atmospheric 511 keV  $\gamma$ -rays with a balloon-borne Ge(Li) spectrometer (Ling et al. 1977). Alternatively, calculations for the atmospheric cosmic-ray spectrum and the resulting electromagnetic emission as a function of longitude, latitude, and altitude are available from Sato (2016). These predictions are based on least-square fits to smooth analytical functions to describe the cosmic-ray spectra and might not capture the required flexibility of the actual measurement. Assessing the suitability of absolute background models in a changing balloon environment is beyond the scope of this paper.

For these reasons, we build an empirical, three-component, background model that is parameterized in variability and amplitude, in order to be fitted simultaneously with a model or along iterative image deconvolutions to describe the celestial emission. The expected number of background photons in the COSI data space  $\{\phi\psi\chi t\}$  is consequently modeled as

$$\begin{aligned} m_{\phi\psi\chi t}^{\text{BG}} &= R_{\phi\psi\chi}^{\text{BG}} \cdot T_t(\theta_b) \\ &= R_{\phi\psi\chi}^{\text{BG}} \cdot \sum_{b \leftarrow (b_i, b_f) \in \mathcal{B}} \beta_b \cdot \Theta(t - b_i) \cdot T_t \cdot \Theta(b_f - t) \\ &= R_{\phi\psi\chi}^{\text{BG}} \cdot \sum_{b \leftarrow (b_i, b_f) \in \mathcal{B}} \beta_b \cdot \mathcal{R}(t, b_i, b_f) \cdot T_t, \end{aligned} \quad (4)$$

where  $R_{\phi\psi\chi}^{\text{BG}}$  is the background response (Section 3.3.2),  $T_t$  is a tracer function (Section 3.3.3) that provides a first-order background variability estimate, and  $\mathcal{B}$  is a set of time nodes (Section 3.3.4) that subdivides the tracer function into a predefined number of subsets with amplitudes,  $\beta_b$ , for each time interval between two time nodes  $b_i$  and  $b_f$  with  $b_i < b_f$ . Those are then fitted simultaneously with the sky model amplitude  $\alpha_s$  (see Equation (3)). Cutting the tracer function into smaller portions allows for a second-order correction to the background variability, as it may take uncaptured variations into account. The set of background parameters is  $\theta_b = \{\beta_b, \mathcal{B}\}$ , and  $\Theta$  is the Heaviside function, such that  $\Theta(t - b_i) \cdot \Theta(b_f - t) = \mathcal{R}(t, b_i, b_f)$  is the rectangle (boxcar) function that returns 1 for  $b_i \leq t \leq b_f$  and 0 otherwise. This then defines a linear combination of  $|\mathcal{B}|$  background models, with a fixed relative variation between each starting and ending time node, and zero otherwise. The covariance between these individual blocks is naturally low and mainly determined by the contribution of the sky emission in each block.

#### 3.3.1. Finding a Good Background Model

We evaluate the performance of different background response, tracer, and time-node combinations by performing the above-described maximum likelihood fits for all cases. We choose several combinations among a large number of possibilities that appear most plausible as background response (Section 3.3.2), background tracer (Section 3.3.3), and background rescaling time nodes (Section 3.3.4) to explore our background model. Even though the background dominates the signal in any case, we require an optimization of the model

accounting for both background and sky components. For this, we include the best-fit 511 keV sky model by Siegert et al. (2016b) and allow the sky amplitude to change. The choice of this model compared to other models, for example, the full-sky model by Skinner et al. (2014) or a simple 2D Gaussian to only represent the bulge, has no influence on the derived background model parameters. This is reasonable since the instrumental background is anyway dominating the total signal and any first-order image proposition is rescaled to the actual number of counts in the chosen COSI data set by our fitting approach.

Since the likelihood naturally increases by introducing more parameters (“fits better”), we make use of the Akaike Information Criterion (AIC; Akaike 1974; Burnham & Anderson 2004a, 2004b), which penalizes “overfits” by taking into account the number of fitted parameters,  $n_{\text{par}}$ , such that

$$\text{AIC} = 2n_{\text{par}} - 2\mathcal{L}(\hat{\theta}_s, \hat{\theta}_b), \quad (5)$$

where  $\mathcal{L}(\hat{\theta}_s, \hat{\theta}_b)$  is the likelihood of Equation (2), evaluated at the best-fit parameters,  $\hat{\theta}_s$  and  $\hat{\theta}_b$ , for sky and background model, respectively.

In general, the lower the AIC, the “better” the model. We note that the AIC is not an absolute “goodness-of-fit” criterion but allows for a restricted set of tested models to identify the most probable (Burnham & Anderson 2004a). Since the data set is very sparsely populated, any use of an approximate  $\chi^2$  goodness-of-fit measure will be flawed. Instead, we will use posterior predictive checks (PPCs; Guttman 1967; Rubin 1981, 1984; Gelman et al. 1996) to evaluate the adequacy of our fits (see Section 4 for further details). In the following, we describe different parts of our background model setup in more detail.

### 3.3.2. Background Response

The background response,  $R_{\phi\psi\chi}^{\text{BG}}$ , is not uniquely defined. In general, it provides an expected number of counts in the  $\{\phi\psi\chi\}$  data space, which should be independent of time. This does not mean that the amplitude of the background is constant in time, but the appearance in the COSI data space is.<sup>7</sup> An exhaustive simulation using the complete mass model could potentially provide a first-order background response; however, the true environment, conditions, and circumstances will alter this expected behavior. As these parameters are constantly changing, determining an absolute background response for each instance in time through simulations is infeasible. For these reasons, we infer a background response empirically from the data.

Order-of-magnitude simulations show that the expected instrumental background compared to the 511 keV sky signal is about a factor of 100. Thus, integrating the measured count rate over long times, i.e., different aspect angles and altitudes, will smear out any contribution of the sky from which a background response can be created. Any background-dominated measurement can thus be used to define a response empirically via

$$R_{\phi\psi\chi}^{\text{BG}} = \sum_t \sum_{e \in \mathcal{E}} d_{\phi\psi\chi te}. \quad (6)$$

<sup>7</sup> Note that it will also have a dependence on energy. Since we are only taking 511 keV photons into account, we omit the dimension of energy.

**Table 2**  
Energy Bands for Background Response Creation

Energy Band (keV)	Comments
[506, 516]	Line only <sup>a</sup>
[460, 560]	Line + continuum
[460, 500]	Adjacent low-energy
[520, 560]	Adjacent high-energy
[375, 500]	Adjacent low-energy, broad
[520, 645]	Adjacent high-energy, broad
[460, 500] and [520, 560]	Adjacent continuum, no line
[375, 500] and [520, 645]	Adjacent continuum, broad, no line

**Note.**

<sup>a</sup> Best-fit background response that is used throughout this work.

In Equation (6),  $d_{\phi\psi\chi te}$  describes the data, i.e., photons with their identifiers  $\phi$ ,  $\psi$ ,  $\chi$  in the instrument-specific data space, the time (of arrival)  $t$ , as well as the photons’ reconstructed energy  $e$ . The sum over all times and a selected energy interval,  $\mathcal{E}$ , sorts each measured photon in the appropriate  $\{\phi\psi\chi\}$ -bin. We normalize any such-constructed background response to 1.0.

The energy interval  $\mathcal{E}$  has to be chosen such that (1) there is enough statistics available for the background response to predict the relative number of counts in each  $\{\phi\psi\chi\}$ -bin, (2) the correct processes in the instrument that lead to the 511 keV are presented, and (3) possible contaminations of sky emission are either completely smeared out or masked. We construct a total of eight background responses from different energy bands, listed in Table 2, to determine the best representation of our data, which always includes a possible sky contribution.

This approach is similar to the empirical background modeling by Siegert et al. (2019b) for the SPI telescope, but in the instrument-specific data space of COSI,  $\{\phi\psi\chi\}$ , instead of SPI’s 19 Ge detectors shadowed by a coded mask. We note that this approach of defining a background response can be refined even further by separating different (physical) processes inside the instrument, for example, distinguishing between the 511 keV line and its underlying continuum, or also for smaller energy bin sizes. Such an elaboration, however, requires a lot of statistics in the individually defined data space bins and might be unreliable for the current COSI data set. A running average across energies or using general linearized models might be used in future background response generations for fine spectroscopy.

While there is also a dependence on the other background parameters, such as the chosen tracer or the additional background time nodes (see next sections), using the energy band [506, 516] keV for creating the background response provides the best fits compared to all other cases (see Appendix Figure C2).

### 3.3.3. Variability Tracer

The intrinsic variability of the above-mentioned processes that lead to instrumental background radiation cannot be predicted from physically motivated models. For this reason, tracers of this variability are determined. These may be any function in time that could be related to the background-generating processes, for example, onboard or external monitors measuring the cosmic-ray flux, a voltage meter, the CsI veto-shield count rate, or the balloon altitude. As a further

step, these functions may be orthogonalized and combined with different weightings to capture additional variability (e.g., Halloin 2009). Alternatively, one “best-performing” tracer function may be cut further as depending on time or, for example, based on the intrinsic variability of the measured count rate (see Section 3.3.4).

In this study, we use three background tracer functions that are supposedly closely related to the measured CE rate at 511 keV: the CsI shield rate (SR), the (inverse of) the balloon altitude ( $h^{-1}$ ), and the photoabsorption event count rate in the analyzed band between 506 and 516 keV (PE).

The SR provides a well-sampled, i.e., high statistics, general trend of any possible process that might lead to background emission. The shield is sensitive to energies  $\gtrsim 80$  keV (Kierans 2018; Sleator 2019), but with no energy information, it also counts a large number of events that are unrelated to the specific 511 keV range. The  $\gamma$ -ray background is higher at lower altitudes, and particularly for 511 keV, lower altitudes result in more cosmic-ray particle showers that include  $\beta$ -particles and secondary decay positrons. Therefore, the inverse of the altitude may be an appropriate tracer for 511 keV. While the 511 keV PEs also include photons from the expected sky emission, the total contribution to the count rate is less than 0.1%. Thus, as the 511 keV PE rate is about 10 times larger than the 511 keV CE rate, these single-site interactions might provide a sufficient tracer of the multiple-site events.

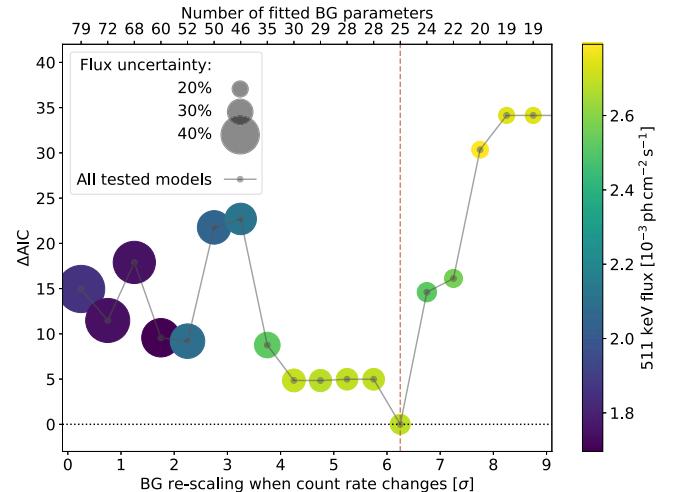
For a zero-order estimate of the predictability of any tracer, we calculate the Pearson correlation coefficient,  $\rho(\text{CE}, X)$ , between the measured 511 keV CEs per hour and any tracer ( $X$ ). The strongest correlation is found between CE and PE with  $\rho(\text{CE}, \text{PE}) = 0.958$ , followed by  $\rho(\text{CE}, \text{SR}) = 0.948$  for the CsI SR, and  $\rho(\text{CE}, h^{-1}) = 0.862$ . While all chosen tracers strongly correlate with the CE count rate, this still should be taken as only an indication for a possible tracer, because there are also photons from the sky included in the CEs (and PEs) that might also be correlated with these functions.

The number of fitted background parameters, and how they are set (time nodes) in addition to a contribution from the sky, influences the fit adequacy. Nevertheless, the PE tracer performs on average better than the other two.

### 3.3.4. Amplitude Renormalization

A function that could predict only the background variability and that is orthogonal to any celestial emission would require only one parameter  $\beta_b$  in Equation (4), i.e., one set of time nodes before the first and after the last time bin of our observations. We cannot, however, be sure that any of the tracer functions behaves in this desired way in the first place, for which reason we define three different possibilities to set varying time nodes to rescale the background tracers.

Such a rescaling is generally useful when the full sky is included in the data set and the true emission is unknown. Because COSI has a  $\sim 1\pi$  field of view and is not performing targeted observations, i.e., the exposure changes smoothly with time, any point-source location will not be visible at all times. This means that any nonperfect background model tracer that intends to describe the pointing-to-pointing variation (or here hour-to-hour variation) will overpredict the number of background counts whenever the source is not in the field of view. For this reason, it might be useful to set time nodes for the background to rescale (introduce another background parameter) whenever the source is in the field of view. However,



**Figure 8.** Performance of the background model combination: PE tracer, 506–516 keV BG response, BG amplitude renormalization using Bayesian block. The minimum AIC indicates when the optimal number of BG parameters (top axis) is reached, avoiding at the same time “bad fits” (too few parameters) and “overfitting” (too many parameters). For a threshold of  $\approx 6\sigma$  in the change of the measured 511 keV count rate (see Figure 2), the optimum is found by using 25 BG parameters (red dashed line). The fitted sky model fluxes are color-coded with their estimated uncertainties shown by the size of the symbols. See text for more details.

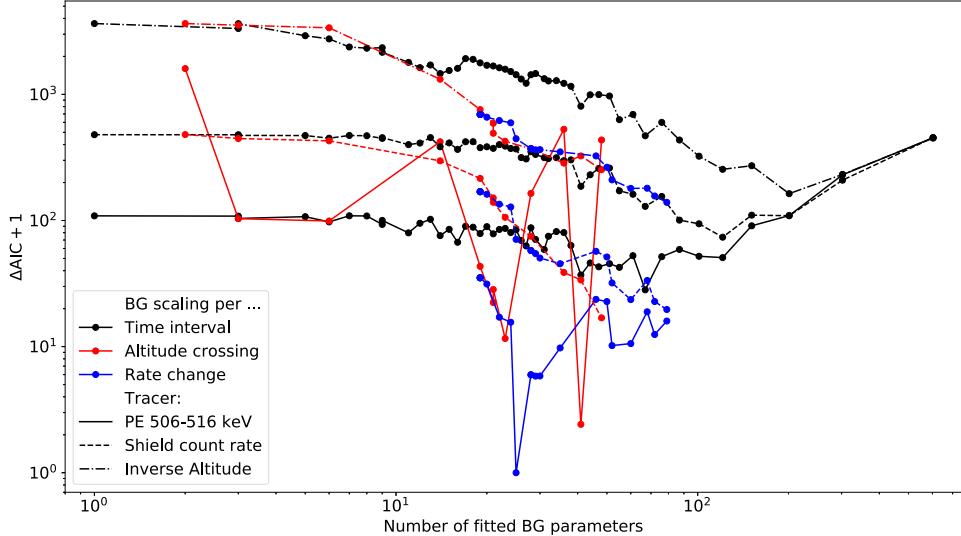
when either the position of the source is not known or the emission is of general diffuse nature with unknown extents, a more general approach to set these time nodes has to be chosen.

As it can be assumed that the background changes are not traced completely from the function  $T_t$  alone, a natural choice comes from the time dimension. We divide the 603 observation hours into equidistant time intervals, ranging between 1 time interval (i.e.,  $\theta_b = \{\beta_1, \{0, 603\}\}$ , thus 1 background parameter) and 603 time intervals (i.e.,  $\theta_b = \{\beta_1, \{0, 1\}, \beta_2, \{1, 2\}, \dots, \beta_{603}, \{602, 603\}\}$ , thus 603 background parameters). This defines 48 different cases.

The altitude can serve as a second-order predictor for when the background should be rescaled. As the altitude changes between 22 and 34 km, we define time nodes whenever the balloon crosses a certain mark, here in unit steps of 1 km. This defines 12 different cases with a number of background parameters between 2 and 48, now set at times according to the altitude crossings. This means that even though the number of background parameters in the time interval case and in the altitude case can be equal, the resulting likelihood might be different.

As a third alternative to when to set additional time nodes, we use Scargle’s Bayesian blocks (Scargle 1998; Scargle et al. 2012). This method determines “change points” of a count rate according to a false alarm probability threshold. We define 20 different thresholds,  $\tau$ , for the Bayesian block algorithm, according to a survival probability,  $S(\tau) = 1 - 2 \cdot \int_0^\tau dx \mathcal{N}_x(0, 1)$ , of the standard normal distribution,  $\mathcal{N}_x(0, 1)$ , between  $\tau = 0.25\sigma$  ( $S(0.25) \approx 80\%$ ) and  $\tau = 9.75\sigma$  ( $S(9.75) \approx 1.8 \times 10^{-22}$ ) in  $\Delta\tau = 0.5\sigma$  steps. Since different thresholds can lead to the same change points, this defines 16 unique cases.

In Figure 8, we show the performance of the best-fitting background model combination: 506–516 keV background response, PE tracer, Bayesian block rescaling time nodes. Clearly, the more background parameters (top axis, right to



**Figure 9.** Performance of all background model combinations using the 506–516 keV BG response. Clearly, choosing to rescale the BG amplitude at time nodes that correspond to strong changes in the count rate (Bayesian blocks, blue) performs best. The background is also adequately determined by using changes with altitude (red), however requiring about twice the number of parameters. Equidistant time intervals (black) show a smoother behavior but in general perform worse. The trend among different tracers is clear, with the 511 keV PE performing best (solid lines), followed by the shield rate (dashed lines) and the altitude (dashed-dotted lines). A summary for different choices of the BG response is given in Appendix C (Figure C2).

left) are included in the fit, the better the resulting likelihood (AIC). After including more than 25 background parameters ( $\tau = 6.25\sigma$ ), however, the large number of fitted parameters is penalized by the AIC. We note that for smaller thresholds (and in general for larger number of parameters), the AIC is not a smooth function, and also the resulting flux (color-coded in Figure 8) is not directly related to  $n_{\text{par}}$ . The uncertainties on the flux naturally increase with the number of fitted parameters.

A summary of  $\approx 30$  background model combinations using the 506–516 keV response is shown in Figure 9. The PE tracer (solid lines) performs best, independent of the chosen background rescaling time nodes. This is reassuring that our methodology is consistent. Depending on the time nodes set, the AIC minimum is found in the range between 25 and 64 background parameters. For other tracers, the minima move to a larger number of parameters. This evaluation has been performed with the full-sky 511 keV model from Siebert et al. (2016b). We again note that different sky models in this procedure, for example, using only a 2D Gaussian component to represent the bulge, alter the absolute likelihood values. Nevertheless, the number of background parameters that are required in this data set is consistently found to be between 25 and 64. This appears reasonable since the celestial contribution is always small and the data set is dominated by instrumental background. From using different background model combinations (see Figures 9 and C2), then with also more parameters, we estimate a systematic uncertainty in our derived flux values of 30%. We note that this is not, and can never be, a full exploration of all<sup>8</sup> possibilities to set time nodes and to choose among tracers. We instead chose among a plausible set of combinations and investigated which produces the most probable outcome, always including a first-order sky model.

<sup>8</sup> For the 603 time bins in our data set, the number of all possible, independent, background time nodes is  $\approx 10^{181.22}$ . The number of possible background tracers is infinite and hence describes an open set, for which no “absolute best fit” can be found.

## 4. Imaging

In Sections 3.1 and 3.2, we introduced the imaging response in general and as applied to our specific data set. For a fixed set of observations, as used here for the 603 hr of 511 keV measurements, the sum  $\sum_{p_t \in R_{\phi\psi\chi}^{\text{SKY}}(Z, A, h)} p_t$  from Equation (3) can be isolated and work as a data set-specific response,  $R_{\phi\psi\chi t}^{\text{SKY}}(Z, A, h)$ . This response then carries entries for the 4243 nonzero bins in the COSI data space  $\{\phi\psi\chi\}$ , times 603 entries in the time domain, times the chosen zenith/azimuth-binning, here  $36 \times 72 = 2592$ . The response thus requires at least an allocation of 53 GB memory alone.

We use this response to (1) perform an image reconstruction using a modified version of the Richardson–Lucy algorithm (Section 4.1) and (2) calculate a set of empirical functions to describe the 511 keV sky as seen with COSI (Section 4.2).

### 4.1. Richardson–Lucy Deconvolution

In order to investigate the celestial contributions of the current data set without relying on a priori assumptions, we perform an iterative image reconstruction using the Richardson–Lucy deconvolution technique (Richardson 1972; Lucy 1974). This algorithm has been successfully used in MeV γ-ray astrophysics (Knoedlseder et al. 1996, 1999, 2005) and can provide a less biased picture of the underlying morphology. It might further reveal structures, shapes, and regions that might not be tested by a pure empirical model fitting approach. We note, however, that this method cannot replace physical modeling of the 511 keV line, and individual features should not be overinterpreted. In particular, we expect on the order of  $10^3$  celestial 511 keV photons (see Kierans et al. 2020), which would be distributed over the number of pixels (here: 2592  $5^\circ \times 5^\circ$  pixels, i.e., degrees of freedom). With an expected significance of about  $7\sigma$  from COSI data (Kierans et al. 2020), only about 16 (sic!) significant ( $3\sigma$ ) pixels would be present.

The general algorithm has been proven to converge to the maximum likelihood solution of the problem (Shepp & Vardi 1982), which, however, tends to find noise peaks in the background-dominated data of MeV instruments (see Knoedlseder et al. 1999). The basic version of the Richardson–Lucy algorithm is described by the iterative update of an initial image, typically set to an isotropic low-flux map, by forward and backward application of the response, such that

$$M_j^{k+1} = M_j^k + \delta M_j^k = M_j^k + M_j^k \left( \frac{\sum_i \left( \frac{d_i}{\epsilon_i^k} - 1 \right) R_{ij}}{\sum_i R_{ij}} \right). \quad (7)$$

In Equation (7),  $M_j^k$  is the  $k$ th image (“map,” with image space indexed by  $j$ ) proposal and iteratively updated by  $\delta M_j^k$ , in which the observation-specific response,  $R_{ij}$  (with data space indexed by  $i = \{\phi\psi\chi\}$ ), is applied to an expectation,  $\epsilon_i^k = \sum_j R_{ij} M_j^k + \epsilon_i^{\text{BG}}$ , given the data set  $d_i$ . The expected number of background counts is  $\epsilon_i^{\text{BG}}$ . The application of the imaging response from image space  $j$  into the data space  $i$  would be forward folding (how does the instrument see an image), whereas the application from data space into image space would be equivalent to a backward projection of (all) data space counts onto the sphere of the sky. The latter also includes the background photons (whose absolute portions are fixed in the standard algorithm, Equation (7)), so that a single back-projection would merely show the instrument itself. For this reason, the total expectation in the data space has to be updated in several iterations. We note that a back-projection of residual counts, for example, from model fitting (Section 3.1), might identify hot spots in the image dimension that are not captured by the used sky models.

Since the standard Richardson–Lucy algorithm, Equation (7), typically uses a fixed background model, and because the delta image is typically updating only in marginal steps, which makes the algorithm very slow (i.e., low flux differences in specific regions; see Kaufman 1987; Lucy 1992), we modify the standard algorithm to also take into account the uncertain background.

#### 4.1.1. Description of Algorithm Used

It was shown in the case of MeV  $\gamma$ -ray imaging (Knoedlseder et al. 1999) that the standard Richardson–Lucy algorithm can be accelerated by applying a multiplicative factor,  $\lambda^k$ , to the delta image in each iteration, which will be determined by a maximum likelihood fit (see Section 3.1). It must be guaranteed that each image pixel  $j$  of the  $(k+1)$ th iteration is still positive, for which reason  $\lambda^k$  is constrained to  $\lambda^k > -M_j^k / \delta M_j^k$ . Note that the delta image can and must contain negative pixels.

Weakly exposed regions in the data sets of Poisson-count-limited experiments, such as COSI, are prone to artifacts, as individual fluctuations in the very sparsely populated data space can lead to unnaturally high expectations  $\epsilon_i$ . For this reason, we apply the noise damping approach from Knoedlseder et al. (2005) and introduce a factor  $w_j = \sqrt{\sum_i R_{ij}}$  for weighting the delta image, and we apply a 2°5 Gaussian filter to reduce the effective number of degrees of freedom in the image reconstruction.

A third modification to the Richardson–Lucy algorithm is provided in this work by allowing the background to vary between iterations: A fixed background model expectation,  $\epsilon_i^{\text{BG}}$ , for example, from an acceptable maximum likelihood fit using a first-order sky model, will result in a reconstruction that strongly depends on this first image and the resulting total number of background photons in each data space bin. Consequently, the reconstruction will be a distortion of the best-fit maximum likelihood solution image and introduces some granularity, but which may just be “filling the residuals” with sky emission. Such an approach is naturally flawed because only specific data space bins may be repopulated owing to the forward application of the response, as the background is fixed. This is equivalent to subtracting a background model and neglecting to consider that this model also carries its own uncertainties. In our modified algorithm, we redetermine the 25 background rescaling parameters,  $\beta_b^k$ , in each iteration, together with the acceleration parameter  $\lambda^k$ , so that the updated image is built from how much background is required to explain the data—and not assuming it in the first place.

Finally, our full modified Richardson–Lucy deconvolution version is written as

$$M_j^{k+1} = M_j^k + \lambda^k \left[ w_j M_j^k \left( \frac{\sum_i \left( \frac{d_i}{\epsilon_i^k} - 1 \right) R_{ij}}{\sum_i R_{ij}} \right) \right]_{25} \quad (8)$$

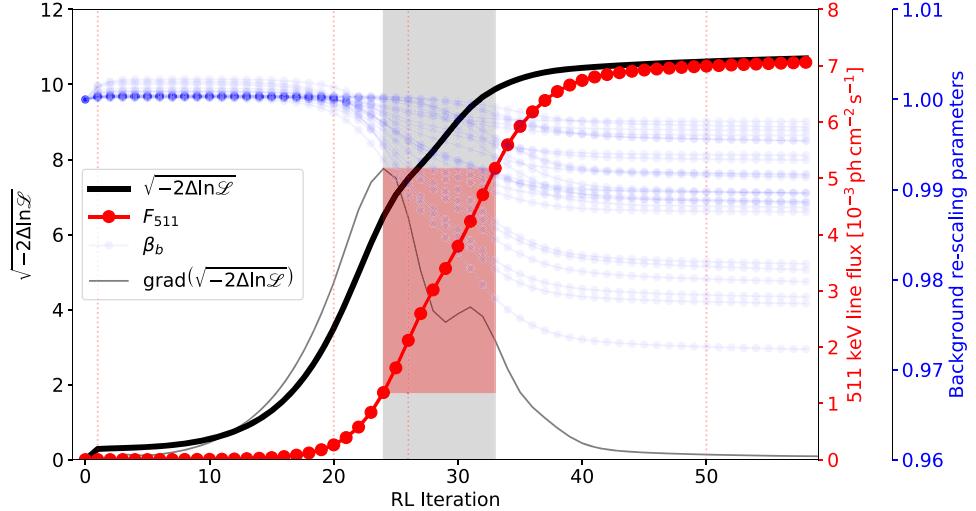
with

$$\epsilon_i^k = \sum_j R_{ij} M_j^k + \sum_{b \in \mathcal{B}} \beta_b^k \hat{R}_i^{\text{BG}}, \quad (9)$$

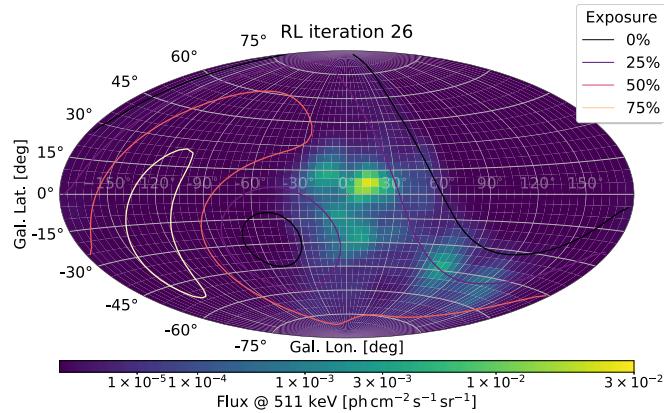
where  $\hat{R}_i^{\text{BG}}$  is the best-fit background model response from Section 3.3.2, together with set  $\mathcal{B}$ , containing the 25 required time intervals to guarantee an adequate fit.

#### 4.1.2. Images

The general problem with any such iterative procedure is to find when to stop the algorithm or determine which image to pick as best representing the data. In fact, there are no definite answers to these questions, as also each solution is in itself uncertain and just represents one realization of the set of parameters. We use the gradient of the shape of the test statistics,  $\sqrt{-2\Delta \ln \mathcal{L}}$ , between the current image proposal and a background-only description (iteration 0) to identify plausible iterations that describe the COSI 511 keV data adequately (see Figure 10). In the case of priors that set the correlation lengths of the pixels, for example, to regularize the frequency of noise in the Poisson-count-dominated data, Allain & Roques (2006) used a trade-off between the likelihood and the prior to extract an adequate solution (“*L-curve*”). Our regularization is approximately given by the Gaussian smoothing kernel and thus constant. This means that the inflection points of the likelihood function alone provide a first-order criterion. We find that iteration 24 is the first inflection point, followed by iteration 26 showing the largest positive curvature. Another inflection point is found at iteration 28, and the last largest positive curvature until convergence to the maximum likelihood (noise-dominated) solution at iteration 33 (see Figure 10).



**Figure 10.** Properties of the modified Richardson–Lucy algorithm for COSI 511 keV data as varying with iteration. The black curve shows the test statistics of the current image proposal vs. a background-only description of the data. The first sharp step is due to the large acceleration parameter found for the first iteration. Without the acceleration parameter, the first sharp step would typically take several tens to hundreds of iterations. The total map-integrated 511 keV flux is shown in red (first right axis), and the background parameters for each step in blue (second right axis). The gradient of the used test statistics (gray, arbitrary units) can be used to define a region of iterations that adequately describe the 511 keV data, defined by the first inflection point toward positive curvature and largest positive curvature before converging to the noise-dominated maximum likelihood solution. Iteration 26 is the one that represents the first maximum positive curvature. See text for details.



**Figure 11.** Iteration 26 of our modified version of the Richardson–Lucy deconvolution algorithm, Equation (9), together with the exposure regions, including 0% (black contours), 25% (purple), 50% (red), and more than 75% (white) with respect to the maximum exposure. Iteration 26 represents the case at which the likelihood ratio function (with respect to a background-only fit) shows the largest positive curvature (see Figure 10), typically chosen as the best trade-off between granularity of the map and likelihood, with a total integrated 511 keV flux of  $2.1 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$ . See text for further discussion.

Thus, iteration 26 provides a map with a compromise between noise and granularity. We show iteration 26 of the modified Richardson–Lucy algorithm in Figure 11. Clearly, there is emission around the center of the Galaxy that is also found to be uncorrelated with the exposure map (contours). This is reassuring that the algorithm works as expected. We note that beyond iteration 33, the low-frequency noise takes over and can enhance individual emission features, especially in regions with  $\lesssim 25\%$  of exposure. Between iterations 24 and 33, the total 511 keV flux varies between  $1.1$  and  $5.1 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$ , with iteration 26 showing  $2.1 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$ . The integrated flux in the central region of the reconstructed image (angular radius  $\leq 40^\circ$ ) is

$1.9 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$ . All these values are consistent with previous measurements, considering the full sky.

We want to remind that individual emission features should not be overinterpreted, in particular because the significance of the full-sky emission in this data set is  $\approx 7\sigma$ – $10\sigma$  (see Figure 10), distributed over hundreds<sup>9</sup> of pixels. For example, the apparently bright spots around  $b \approx -45^\circ, 60^\circ \lesssim l \lesssim 120^\circ$ , are very close to the completely unexposed regions of the sky (inside black contours), and therefore these might only be image artifacts. This may then also be related to the reliability of the imaging response for larger zenith angles ( $\gtrsim 45^\circ$ ) and the statistics in the response-generating simulation. Additionally, a stronger-than-expected dependence on the altitude may lead to skewed correction factors, again especially at large zenith angles.

The high-latitude features only appear in later iterations, whereas the central component is immediately present after only a few iterations. Likewise, the three distinct emission features around the Galactic center are probably due to the reconstruction method itself, favoring distinct emission spots rather than correlated pixels, and it should be considered only as describing the general extent of the emission. Nonetheless, it appears here that the emission is more extended than what was found in earlier measurements. We investigate the emission extent in more detail in Section 4.2 using empirical emission templates to obtain uncertainties on the parameters that describe the morphology. Finally, we note that any such reconstruction always depends on the choice of the background model.

From the image deconvolution, we find no evidence of a 511 keV disk. Such a feature would only be visible for negative longitudes, as COSI’s exposure is restricted to  $l \lesssim 60^\circ$ . We further discuss reliability of the image reconstruction in Appendix A and provide examples of simulated data sets including different flux levels in Appendix B.

<sup>9</sup> By smoothing the delta images, the effective number of degrees of freedom (data points) will be smaller than the number of used pixels.

#### 4.2. Model Fitting

As described above, the Richardson–Lucy algorithm is prone to produce noise peaks in low exposure regions, especially for later iterations. These could be alleviated by the usage of more elaborate image reconstruction methods. For example, the Multiresolution Regularized Expectation Maximization (MREM) method, which is based on the Richardson–Lucy algorithm, tries to damp the low-frequency noise in the delta images through wavelet thresholding (see, e.g., Knoedlseder et al. 1999, for an application to COMPTEL  $^{26}\text{Al}$  data). Alternatively, the Maximum Entropy method applies a prior in image space, measuring entropy of an image proposition by its distance to a default image, and thus counteracting the likelihood solution (e.g., Knoedlseder et al. 1996). Because the signal strength in our current data set is in general very low, we want to quantify the current findings by a more restrictive approach. Using predefined templates that are parameterized by only a few parameters, we can provide a robust estimate of the emission parameters and furthermore compare to previous findings.

Consequently, the final equation to describe the COSI 511 keV data is assuming already-convolved sky models, Equation (3), marked by an asterisk, as well as the best-fitting background response from Section 3.3, marked by a hat (still with free amplitude parameters). Thus,

$$\begin{aligned} m_{\phi\psi\chi t} = & \sum_{s \in \mathcal{M}} \alpha_s \cdot m_{\phi\psi\chi t}^{\text{SKY}, s, *} \\ & + \sum_{b \leftarrow (b_i, b_f) \in \hat{\mathcal{B}}} \beta_b \cdot \hat{R}_{\phi\psi\chi}^{\text{BG}} \cdot \hat{\mathcal{R}}(t, b_i, b_f) \cdot \text{PE}_t \end{aligned} \quad (10)$$

includes the scaling parameters  $\alpha_s$  (solid-angle integrated sky flux) for each map in the set of chosen sky models to be fitted,  $\mathcal{M}$ , and the 25 background model rescaling parameters,  $\beta_b$ , whose associated time nodes,  $(b_i, b_f) \in \hat{\mathcal{B}}$ , had been calculated by the Bayesian block algorithm with a change point threshold of  $\hat{\tau} = 6.25\sigma$ . For a single sky map, the total number of fitted parameters is thus 26.

For an intuitive check of the absolute values of resulting background parameters, we normalize each background block (time span) to the number of measured counts inside this block. This leads to a background rescaling parameter of  $\beta_b = 1.0$  if the contribution of celestial emission is zero, and should be  $<1.0$  if the sky response suggests a contribution different from zero. Thus, if an expected signal is visible throughout all exposures, the background parameters should all deviate from 1.0 (=background-only) in the fit.

We include priors for the sky and background scaling parameters, based on our image reconstruction (Section 4.1.2), previous measurements with SPI and other instruments, and the expected contribution of background counts to the total signal, such that

$$\pi(\theta | d_{\phi\psi\chi t}) \propto \mathcal{L}(d_{\phi\psi\chi t} | \theta) \pi(\theta) \quad (11)$$

is the joint posterior distribution of all parameters. We sample the posterior by using the No U-Turn Sampler (NUTS; Hoffman & Gelman 2011, 2014) built in Stan (Carpenter et al. 2017). In Equation (11),  $\mathcal{L}(d_{\phi\psi\chi t} | \theta)$  is the likelihood given in Equation (2), and  $\pi(\theta)$  are the prior distributions.

In previous studies, the 511 keV line flux in the bulge region of the Galaxy was consistently found to be of the order of

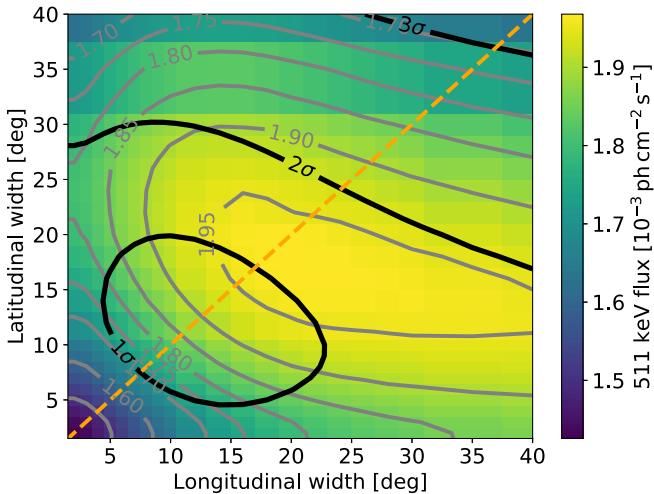
$\sim 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$ . The full-sky emission is less well determined, and the fluxes range from 1.7 to  $3.5 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$  (e.g., Knoedlseder et al. 2005; Skinner et al. 2014; Siebert et al. 2016b), with a tendency for higher fluxes with increasing exposure and INTEGRAL/SPI observations of the Milky Way disk and higher latitudes. We note that the absolute flux values from OSSE can be considerably smaller (Purcell et al. 1997). We choose a prior on the 511 keV line flux that is normalized to the flux of the convolved sky map in each case with  $F_{511}^{\text{bulge}} = 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$  for bulge-only maps (Section 4.2.1), and varying for the full-sky maps ( $F_{511}^{\text{full sky}}$ , Section 4.2.2). In this way, we can set a truncated normal prior for the sky amplitude  $\alpha \sim \mathcal{N}_{\alpha>0}(\mu = 1, \sigma = 2/3)$ . The choice of the large prior width originates from the unknown systematics in the response creation of COSI, for which the absolute efficiency at 511 keV can easily be off by several tens of percent. In general in this study, for the sky model flux, the prior functions as a scale to the problem and forces positivity to the signals. Kierans et al. (2020) extracted a positron annihilation spectrum with COSI from the central  $16^\circ$  around the Galactic center and found a 511 keV flux of  $(3.9 \pm 0.4) \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$ . Note that while the absolute flux of the 511 keV bulge emission has been measured by several balloon-borne and satellite-based instruments (e.g., Purcell et al. 1997; Prantzos et al. 2011, and references therein), the different systematics inherent to each measurement can lead to several tens of percent of additional margin. Based on the Richardson–Lucy image deconvolution, we find a 511 keV flux of  $\approx 2 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$  (Section 4.1.2). Consequently, within  $3\sigma$ , the prior width provides a factor of 2 uncertainty in the absolute measurement.

For the background, the choice of the priors is set to a truncated normal distribution,  $\beta_b \sim \mathcal{N}_{\beta>0}(\mu = 1, \sigma = 0.1)$ , for each block  $b$ , as large variations between different blocks would be unexpected (as already normalized to the count rate) and still provide enough leverage for the sky contribution to exceed previous expectations. We note that the truncation for positive fluxes (or background contributions) does not prevent the fit from reaching zero sky flux and furthermore restricts the signal to be positive, as naturally expected from an emission process.

To address the adequacy of our fits, we use PPCs. The posterior predictive distribution is given by

$$\pi(\tilde{d}_{\phi\psi\chi t} | d_{\phi\psi\chi t}) = \int d\hat{\theta} \mathcal{L}(\tilde{d}_{\phi\psi\chi t} | \hat{\theta}, d_{\phi\psi\chi t}) \pi(\hat{\theta} | d_{\phi\psi\chi t}), \quad (12)$$

where  $\pi(\hat{\theta} | d_{\phi\psi\chi t}) \propto \mathcal{L}(d_{\phi\psi\chi t} | \hat{\theta}) \pi(\hat{\theta})$  is the joint posterior distribution of all fitted parameters  $\hat{\theta}$ , and  $\mathcal{L}(\tilde{d}_{\phi\psi\chi t} | \hat{\theta}, d_{\phi\psi\chi t})$  would be the predictive distribution of “replicated” (simulated) data,  $\tilde{d}_{\phi\psi\chi t}$ , from the inferred parameters, given the original data set (Guttman 1967; Rubin 1981, 1984; Gelman et al. 1996). In this way, the data generating process is used to predict future data in the same data space, which can then be compared to the current data set and possibly uncover systematic deviations in the assumed model. While the PPC provides a probability distribution for each data point in the complete  $\{\phi\psi\chi t\}$  data space, a comparison in summary statistics is found sufficient (Gabry et al. 2019), as the behavior should change, if at all, smoothly in either dimension. In this study, we use the partial sums over either dimension of the data space to compare with the PPC. For example,  $\tilde{d}_\phi = \sum_{\psi\chi t} \tilde{d}_{\phi\psi\chi t}$



**Figure 12.** Likelihood profile (black contours) of the grid search in Section 4.2.1 as a function of 2D Gaussian widths in longitude ( $\sigma_l$ ) and latitude ( $\sigma_b$ ), centered on  $(l_0/b_0) = (0/0)$ . The likelihood is maximized for  $\sigma_l \approx \sigma_b = 12^{+8}_{-5}$  deg, resulting in a best-fit flux (color-coding, gray contours) of  $(1.90 \pm 0.45) \times 10^{-3}$  photons cm $^{-2}$  s $^{-1}$ .

describes the time-averaged distribution of Compton scattering angles for all polar and azimuth scattering angles.

#### 4.2.1. Empirical Description

Similar to previous studies, we intend to describe the diffuse 511 keV line emission and flux empirically by a number of 2D Gaussian functions. Here, we use a grid of asymmetric 2D Gaussians,  $G(l, b; \sigma_l, \sigma_b)$  with longitudinal width  $\sigma_l$ , latitudinal width  $\sigma_b$ , normalized to a full-sky-integrated line flux  $F_{511}$ , and centered on the Galactic center  $(l_0/b_0) = (0/0)$ :

$$G(l, b; \sigma_l, \sigma_b) = \frac{F_{511}}{2\pi\sigma_l\sigma_b} \exp\left(-\frac{1}{2}\left[\frac{l^2}{\sigma_l^2} + \frac{b^2}{\sigma_b^2}\right]\right). \quad (13)$$

Our chosen grid is equally spaced in  $\sigma_l$  and  $\sigma_b$ , respectively, from 1° to 10° in 1° steps, to 30° in 2° steps, and to 40° in 5° steps. This defines a grid of  $22 \times 22 = 484$  individually tested maps. In each fit, the sky amplitude  $\alpha$  and the background parameters  $\beta_b$  are redetermined simultaneously. This results in a likelihood profile as shown in Figure 12.

The 511 keV emission in the Galactic bulge region is found to be diffuse, as the width parameters favor values larger<sup>10</sup> than COSI's spatial resolution of  $\approx 5^\circ$ . There is no strong asymmetry found in the shape of the 2D profiles, so that we reduce the asymmetric Gaussian function to a symmetric one,  $\sigma_l = \sigma_b \equiv \sigma_{\text{sym}}$  (dashed orange line in Figure 12), and find a best-fit extent of  $\hat{\sigma}_{\text{sym}} = 12^{+8}_{-5}$  deg. This value is in agreement with the extension  $\approx 14^\circ$  estimated in Kierans et al. (2020). We refer to this best-fit model as  $G_{12}$  throughout the next sections. At this point in the grid, the fitted 511 keV flux is determined to be  $(1.90 \pm 0.45) \times 10^{-3}$  photons cm $^{-2}$  s $^{-1}$ . Note that there is a dependence of the extent on the flux uncertainties. This can be estimated from the tangents of flux contours in Figure 12

<sup>10</sup> We note that diffuse emission on smaller scales than the angular resolution can be fitted, as the imaging response broadens any emission profile by its 5° resolution. Only a point source would appear as a 5° emission feature; a 2D Gaussian with an FWHM of 2°, for example, would be seen as a  $\approx 5^\circ$  emission feature, approximated by Gaussian quadrature. With high enough statistics, such a deviation can be identified.

with the  $\Delta\mathcal{L} = 1\sigma$  contours (see, e.g., Appendix A.1 in Siebert et al. 2016b) and results in an uncertainty of  $(^{+0.07}_{-0.20}) \times 10^{-3}$  photons cm $^{-2}$  s $^{-1}$ . These values are also consistent with our findings from the image reconstruction, Section 4.1.2, in both flux and extent, again reassuring that our formalism is robust. We use simulations to assess the reliability of this likelihood profile and its accuracy. The results of these simulations are shown in Appendix B for different 511 keV fluxes. In addition, we can now quantify the emission uncertainties and find that emission features beyond  $\approx 40^\circ$  are not required to explain the COSI 511 keV data. We discuss possible differences with literature values and implications in Section 5.

#### 4.2.2. Full-sky Emission

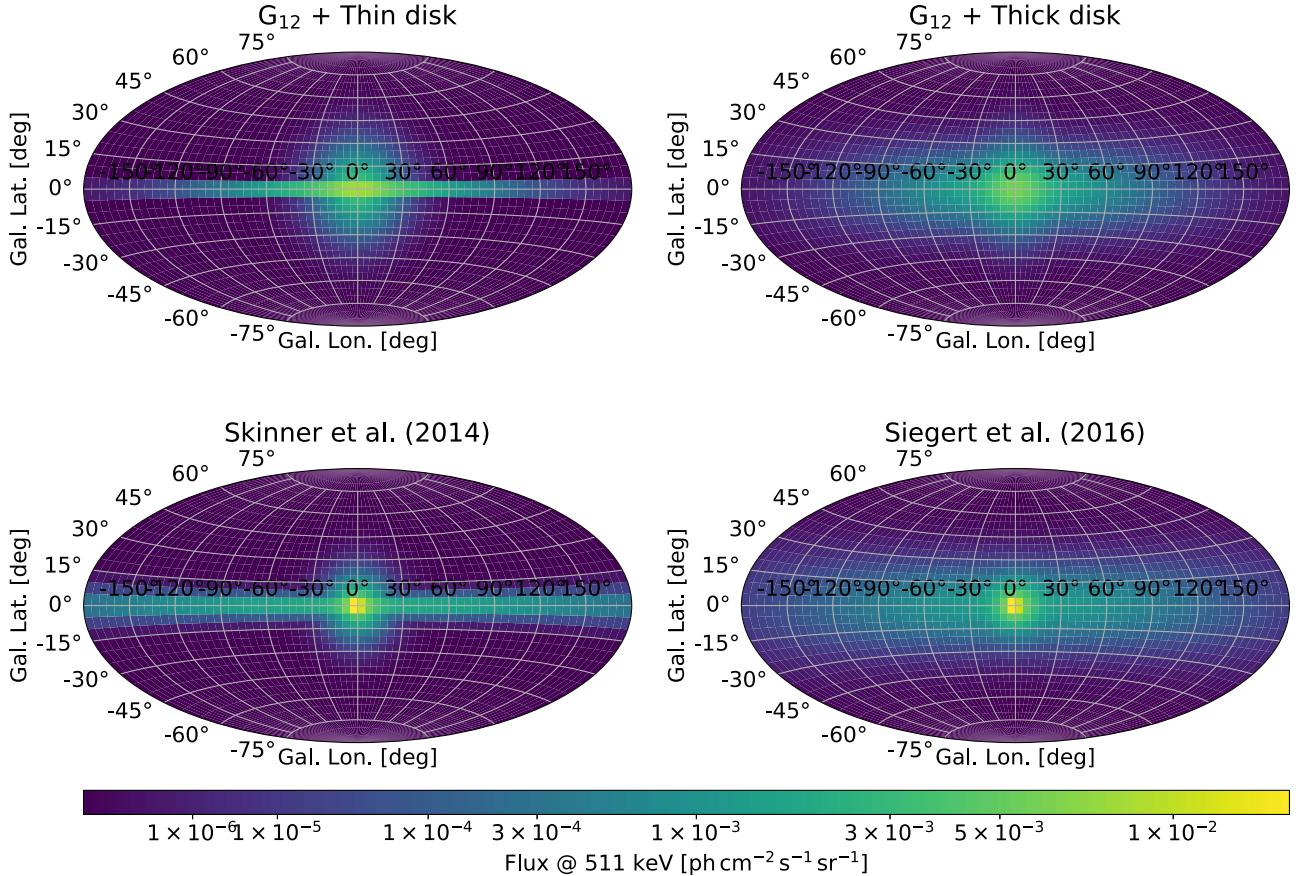
Previous studies (e.g., Purcell et al. 1997; Knoedlseder et al. 2005; Weidenspointner et al. 2008; Bouchet et al. 2010) suggested that there is an additional disk-like structure in the Galactic-wide 511 keV emission beyond what would typically be attributed to the Galactic bulge or bar. Observations with more exposure confirmed the longitudinally extended morphology (Skinner et al. 2014; Siebert et al. 2016b) and opened a discussion of whether this component is really the Galactic thin or thick disk, or if the component actually points toward a more halo-like structure.

With the ultimately more definite imaging response of Compton telescopes, this question can be investigated further. We use different model components, either in addition to the best-fit bulge from Section 4.2.1 or complete full-sky descriptions from the recent literature, to investigate whether there is a disk-like component present in COSI 511 keV data.

We use the model  $G_{12}$  and add a second disk-like component to the fit, modeled as additional 2D Gaussian with either a small ( $\sigma_b = 2^\circ$ ; see Skinner et al. 2014) or a large ( $\sigma_b = 10^\circ$ ; see Siebert et al. 2016b) latitudinal extent, and a longitudinal extent of  $\sigma_l = 40^\circ$  in both cases. The quoted disk fluxes in the literature range between 0.0 and  $2.9 \times 10^{-3}$  photons cm $^{-2}$  s $^{-1}$ , depending on the instrument and the total exposure (e.g., Purcell et al. 1997; Prantzos et al. 2011; Siebert et al. 2019a). Later observations with more than 10 yr of INTEGRAL/SPI exposure consistently found a disk-like component, with flux values in the range of  $(1.0\text{--}2.0) \times 10^{-3}$  photons cm $^{-2}$  s $^{-1}$ , for which reason we set the normalization of any such second component to  $1.5 \times 10^{-3}$  photons cm $^{-2}$  s $^{-1}$  and use again a truncated normal prior of  $\alpha_{\text{DISK}} \sim \mathcal{N}_{\alpha>0}(\mu = 1, \sigma = 2/3)$ .

We find no significant excess in either combination and provide a  $3\sigma$  upper limit on the 511 keV flux of  $< 3.1$  and  $< 4.3 \times 10^{-3}$  photons cm $^{-2}$  s $^{-1}$  for the thin and thick disk, respectively. Note that the 99.85% percentile (one-sided  $3\sigma$  bound) of the chosen prior includes  $4.5 \times 10^{-3}$  photons cm $^{-2}$  s $^{-1}$ , showing that the upper limits are dominated by the contributions from the likelihood. We find that including a second component reduces the flux of the central bulge component by  $\approx 25\%$  in each case, which points to a nonzero contribution of a disk-like component. This is reassuring, as the bulge flux now appears closer to literature values. We can, however, not claim a detection of an additional component beyond the Galactic bulge, as described by the  $G_{12}$  model.

For additional full-sky model tests, we use the four-component models by Skinner et al. (2014) and Siebert et al. (2016b) with fixed relative amplitudes to have complete descriptions across the entire sky. The number of sky model



**Figure 13.** Tested full-sky emission morphologies. The top panel shows the best-fit 2D Gaussian ( $G_{12}$ ) from Section 4.2.1, Figure 12 with symmetric width of  $12^\circ$ , plus thin disk (left, vertical extent  $2^\circ$ ) and a thick disk (right, vertical extent  $10^\circ$ ). The bottom panels show the multicomponent models as found by Skinner et al. (2014, left) and Siegert et al. (2016b, right).

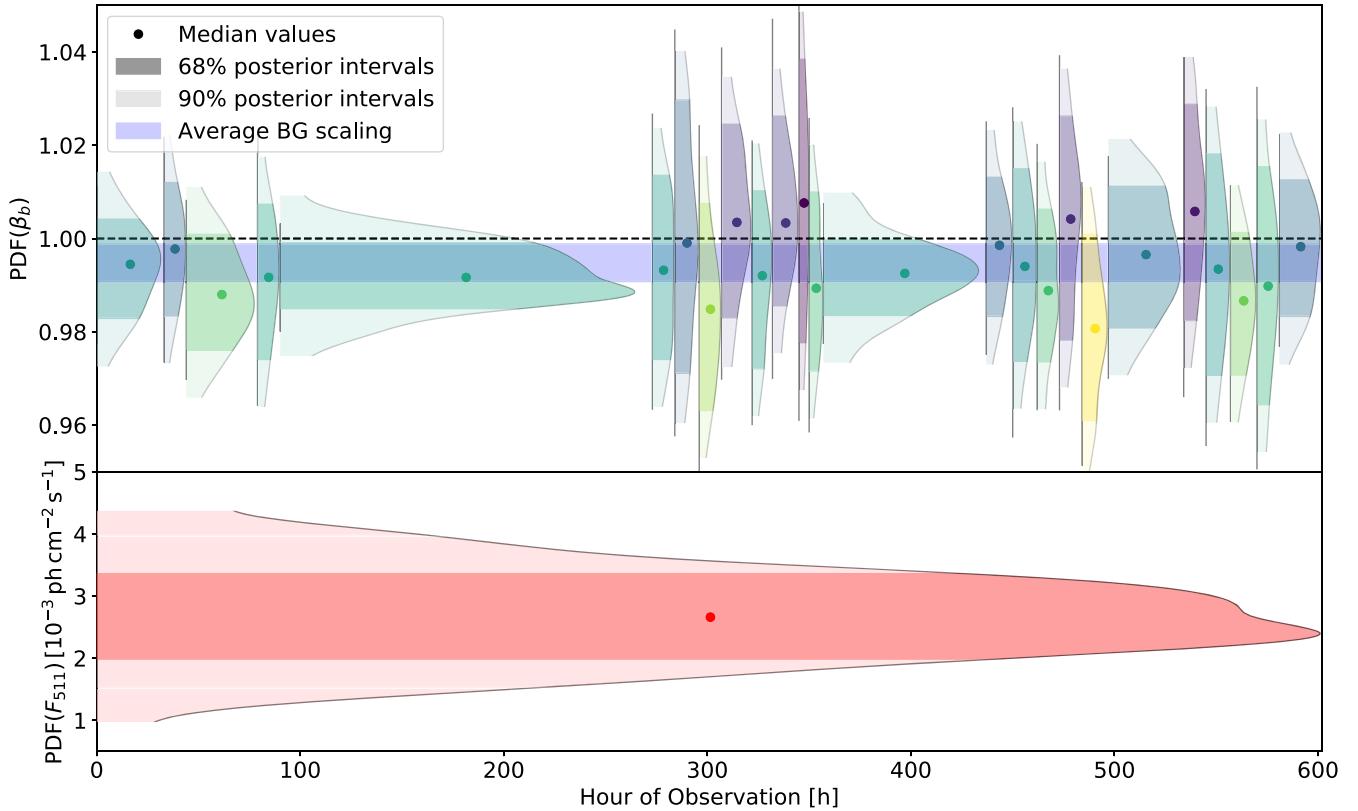
parameters for these tests is thus reduced to 1 again. This provides an estimate of the total Galactic 511 keV flux as seen by COSI during its 2016 flight. In Figure 13, a summary of the used full-sky models is shown.

For the Siegert et al. (2016b) model, we show exemplarily the posterior distributions of the fitted sky model scaling parameter, as well as the 25 background scaling parameters as they vary with time in Figure 14. The full-sky 511 keV flux is found to be  $(2.7 \pm 0.7) \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$  (bottom panel), consistent with previously found estimates. The posterior distributions for the background rescaling parameters are shown in the top panel, each of them shown according to the time intervals between which the Bayesian block method set the time nodes (see Section 3.3.4 and Figure 15). If there was no sky contribution present, i.e.,  $F_{511} = 0$ , the  $\beta_b$ -values should consistently scatter around 1.0. While each fitted background parameter is individually consistent with 1.0, there is a clear trend for a reduced background level during the 603 observation hours, as indicated by the blue shaded band.

In Figure 15, we show the PPC of this model in the time domain for a fit quality check (additional PPCs in the remaining COSI data space, i.e.,  $\{\phi\psi\chi\}$ , are shown in Appendix C). The top panel shows the COSI 511 keV data as a black histogram, together with the model posterior of sky (blue) and background (red), and the PPC as summarized into these times bins as green shading. Naturally, the total count rate is dominated by the background, and consequently so is the PPC. In all panels, the best-fit Bayesian block time nodes,  $\hat{\mathcal{B}}$ ,

are indicated by dashed orange lines. In the middle panel, the absolute residuals in count space ( $d_t - \tilde{d}_t$ ) are shown, together with the PPC as scattering around 0, however with changing variance according to Poisson statistics. In order to normalize the absolute residuals and to provide a common frame of comparison, we calculate the  $z$ -score for each time bin  $((d_t - \tilde{d}_t)/\sqrt{\text{var}(\tilde{d}_t)})$ . Clearly, the data points scatter around 0, with seldom outliers beyond the 99th percentile of the PPC. In this fit, only 12 out of 603 values are found outside this range, which is consistent with expectations. In Appendix C, we show the PPC in other COSI data space dimensions, also finding that our model describes the data adequately. We note that in various scattering angle bins, the number statistics is very small, which naturally leads to asymmetric residuals owing to the nature of the Poisson statistics.

The results for the Skinner et al. (2014) model, i.e., a thin disk, are of similar nature, providing a total 511 keV line flux of  $(2.3 \pm 0.9) \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$ . There is no significant likelihood difference found between the two models. This is expected since the exposure for any disk-like morphology is restricted to longitudes  $-150^\circ \lesssim l \lesssim -90^\circ$  and  $30^\circ \lesssim l \lesssim 60^\circ$ , i.e., small subregions for which the disk luminosity is expected to drop significantly compared to the bulge (see Figure 6). We summarize our fitted models, corresponding flux values, and likelihoods in Table 3. In addition, the question of how extended the 511 keV disk actually is is still under debate, as well as whether a more structured morphology, e.g., including spiral arms, is present. See Section 5 for further discussion.



**Figure 14.** Posterior probability distributions of the background rescaling parameters,  $\beta_b$  (top), and the resulting 511 keV flux,  $F_{511}$  (bottom), assuming the Siegert et al. (2016b) model. The horizontal width of each posterior resembles the time ( $x$ -axis) for which a specific parameter is active. The vertical width includes the 68th (dark shaded,  $\approx 1\sigma$ ) and 90th (light shaded,  $\approx 1.7\sigma$ ) percentiles of the sampled distributions. The colors indicate how far the median of the  $\beta_b$ -values deviates from a background-only time interval (i.e., 1.0). See text for further details.

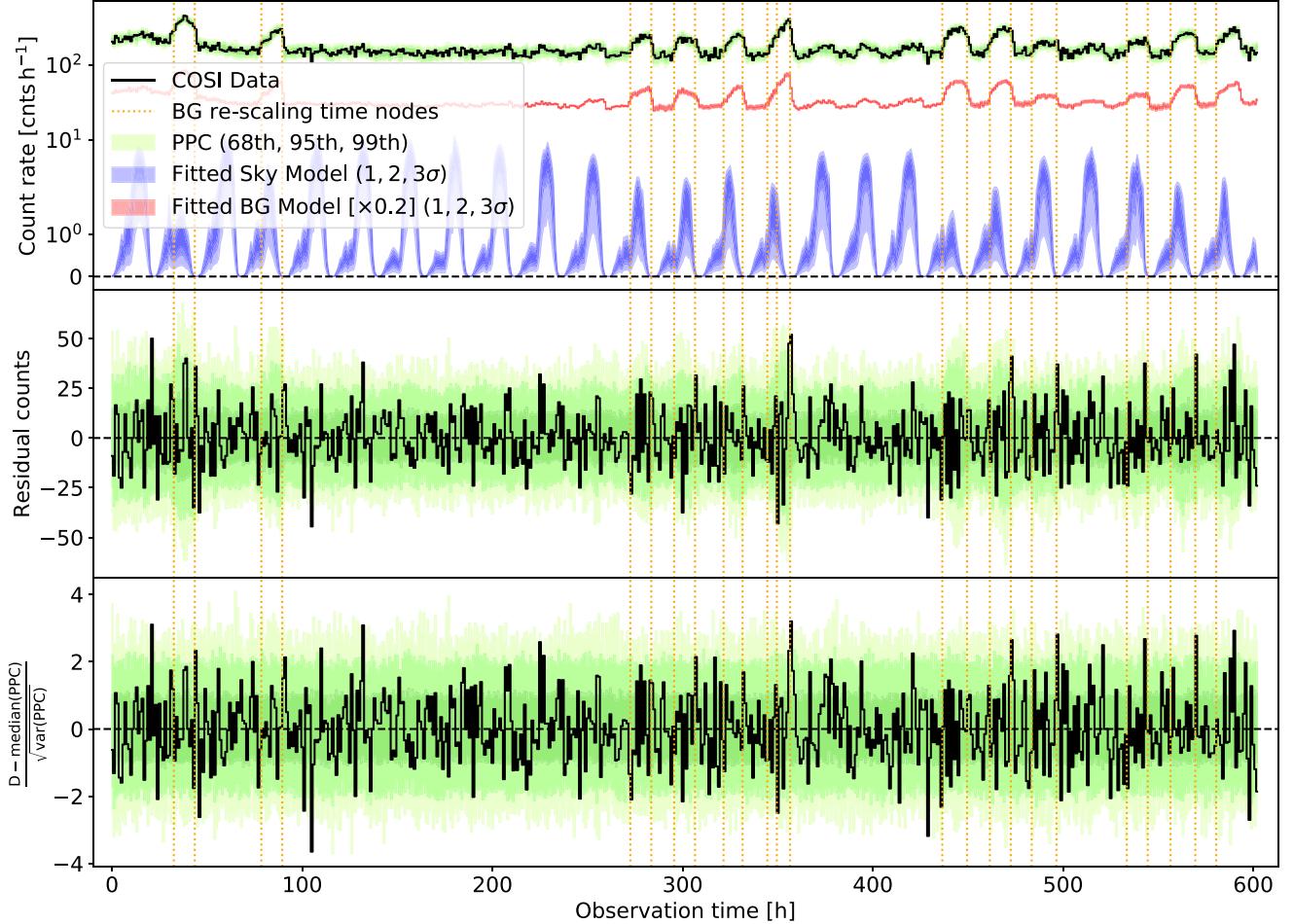
## 5. Discussion

### 5.1. Positron Annihilation Puzzle

The measured extent of the central 511 keV emission ( $\text{FWHM} \approx 28^{+19}_{-12}$  deg) is found to be at least 2–3 times larger compared to previous measurements by INTEGRAL/SPI (e.g., 8° FWHM by Knoedlseder et al. 2005), however consistent with WIND/TGRS measurements by Harris et al. (1998), finding  $24^{+8}_{-9}$  deg. Our fitted extent is in agreement with COSI data analysis from Kierans et al. (2020) ( $\approx 33^\circ$ ), who focused on spectral analyses. We note, however, that such large-scale emission regions naturally (due to the nature of the fit) capture more flux than smaller regions. In addition, this large 2D Gaussian might capture not only the “narrow bulge” emission component but rather a superposition of the true emission morphology, including the disk and/or a halo, which might not have been seen in other instruments. For example, Skinner et al. (2012), on the one hand, noted that a halo component would be favored by instruments like OSSE, TGRS, or SMM with large field of views, compared to a SPI-only description of the data. On the other hand, as already shown by Alberne et al. (1981), Leventhal et al. (1986), Lingenfelter & Ramaty (1989), or Purcell et al. (1997), this field-of-view issue (larger field of views lead to larger fluxes) has been addressed correctly in both analyses. This means that a correct forward implementation of the effective area as a function of zenith and azimuth for each instrument should yield the same results, because the field of view is already included and cannot come into play a second time. Rather, the analysis methods

themselves have to be carefully investigated, as it is typically assumed, for example, for SPI, that everything “outside” the Galactic bulge and disk (e.g., at high latitudes; Bouchet et al. 2015) will provide no contribution to the expected counts. For this reason, halo components would not be visible. In addition, halo-like emission, or any emission with a shallow gradient or isotropy, is almost impossible for a coded-mask instrument to observe, because the coding would result in an equal response for all times. If not accounted for, this ultimately would be disregarded as being due to background. Similar statements apply for collimators as well. A possible step toward observing large-scale diffuse emission with collimating or coded-mask telescopes would be occultation observations, for example, being shadowed by Earth. While the sensitivity of current  $\gamma$ -ray telescopes is probably too low to detect halo-like or isotropic emission at 511 keV in general, Compton telescopes, such as COSI, provide a direct response to single photons so that low-gradient emission could be identified. This would provide a major step in estimating the true extents of soft  $\gamma$ -ray emission in the MeV regime in general.

The 511 keV flux measurement of  $(1.90 \pm 0.45) \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$  for the best-fit 2D Gaussian component to describe the bulge is about twice as large as in previous measurements, but consistent within uncertainties. Adding a disk-like component reduces this value to about  $1.5 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$  (see Table 3), which is more consistent with earlier measurements. We find no evidence, however, for an additional disk component, probably due to its low surface



**Figure 15.** Model fit quality (summary statistics) in the time domain: shown are the COSI data (black histogram), together with the 68th, 95th, and 99th percentile of the PPC (green bands, see Equation (12)), summed over the COSI data space  $\{\phi\psi\chi\}$ , for each time bin, together with the fitted background model (red; for illustration purposes scaled by 0.2) and sky model (blue). From top to bottom, the data space, the absolute residuals, and the  $z$ -scores are shown. The chosen background time nodes are indicated by vertical lines. Clearly, the fit appears adequate, as only 12 out of 603 time bins fall outside the 99th percentile of the PPC. See Figure C3 for additional summary statistics in other data space dimensions.

**Table 3**  
Summary of Model Fitting Results

Model	Bulge	Disk	Total	$\Delta \ln \mathcal{L}$
G <sub>12</sub>	$1.9 \pm 0.4$	...	$1.9 \pm 0.4$	-954.9
G <sub>12</sub> + thin disk	$1.5 \pm 0.5$	<3.1	$2.6 \pm 0.7$	-953.3
G <sub>12</sub> + thick disk	$1.4 \pm 0.6$	<4.3	$2.9 \pm 0.8$	-953.6
Sieger et al. (2016b)	...	...	$2.7 \pm 0.7$	-958.3
Skinner et al. (2014)	...	...	$2.3 \pm 0.9$	-960.0

**Note.** Fluxes for bulge, disk, and total components are given in units of  $10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$ , with  $1\sigma$  uncertainties, including the uncertainties in the extension of the spatial distribution. Upper limits are given as the 99.85th percentile of the posterior ( $3\sigma$ ). The log-likelihood is shown as a relative value with respect to a constant. While each map performs about equally well, it should be noted that the smaller bulge components by Skinner et al. (2014) and Sieger et al. (2016b) are slightly disfavored by the COSI data.

brightness nature, and provide upper limits that are only about twice the measured values from recent SPI studies.

Our Richardson–Lucy deconvolution algorithm finds a 511 keV flux in the central regions of the Galaxy of about  $2 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$ , consistent with our model fits, and affirming a robust analysis. We apply additional filters to our data set (Appendix A), from which we find that observations

that include only night times or when the balloon altitude was above 30 km result in the most noise-free maps. In addition, we find that the second third of the 603 observing hours provides the cleanest image, in which also the emission peak appears closer to the Galactic center.

We find a general consistency between COSI measurements and earlier studies regarding the absolute 511 keV flux estimates, but with tendencies toward higher fluxes and more extended emission. This could be due to systematic mismatches between simulations for the imaging response and true effective area, unaccounted-for systematics in the background modeling procedure, or, alternatively, because of the better imaging capabilities of Compton telescope apertures in general, being able to capture also flux from emission regions with shallow gradients.

## 5.2. The Future of Compton Telescopes

A reliable, robust, and versatile background modeling for soft  $\gamma$ -ray telescopes is in general difficult to achieve. Similar to earlier CGRO/COMPTEL procedures (e.g., Knoedlseder et al. 1996; van Dijk 1996; Bloemen et al. 1999), and together with the experiences from the INTEGRAL/SPI spectrometer (Diehl et al. 2018; Sieger et al. 2019b), we developed a method to infer a flexible background model for COSI, inferred

from the measurements themselves. As opposed to a rather well-defined space environment with fixed observation patterns, a free-floating balloon-borne telescope poses additional difficulties in estimating the time-dependent background contributions. We showed that it is possible to infer imaging information in a full-forward modeling manner, by allowing the strongly variable background to be determined simultaneously with the celestial 511 keV  $\gamma$ -ray signal.

We found that, naturally, the largest impact on the background count rate at 511 keV is the balloon altitude, as more air mass increases the interaction rate of cosmic rays with the atmosphere, leading to more secondary particles and  $\gamma$ -rays. Additionally, southern latitudes lead to an increased 511 keV count rate owing to the strong latitudinal dependence of the geomagnetic cutoff rigidity (see, e.g., Ling 1975; Ling et al. 1977; Kierans 2018). The short-term variability can be predicted from independent count rates of COSI's CsI veto-shield, for example, or by the photoevents (single-site events), which provide an exceptionally good predictor for 511 keV Compton event photons. Scargle's Bayesian blocks algorithm (Scargle 1998; Scargle et al. 2012) provides a useful tool to identify additional changes of the background rate that are missed by any variability tracer. Finally, our complete background model describes a semiempirical and modular approach to tackle the unknown MeV background, being based on the instrument-specific data space, as well as expertise from balloon enterprises.

In this work, we thus confirmed earlier studies regarding the Galactic 511 keV emission and provided a scheme to approach the individual difficulties that such a complex instrument in a complex environment inherits. The ultimate measurement of the 511 keV emission would nevertheless be best from space. This can be accomplished with the COSI-SMEX space mission, which is currently in a Phase A study.<sup>11</sup> In nearly equatorial low-Earth orbit, a COSI satellite would have significant advantages compared to COSI on its balloon platform: COSI-SMEX's reduced strip pitch will lead to better angular resolution and better background identification due to fewer incorrectly reconstructed events. In addition, more events with close-by interactions can be used in the analysis, and, together with more detectors (16 instead of 9 in this work), less stringent event cuts due to better shielding will lead to a significantly larger effective area. The lower and more stable background conditions in space, along with no atmospheric absorption and a larger field of view, will ultimately result in a better sensitivity than any previous MeV  $\gamma$ -ray telescope.

Considering also the longer mission duration (2 yr, plus possible extensions), COSI-SMEX would readily be able to answer the still unsolved questions about the true 511 keV morphology and allow the study of individual regions in 511 keV and the connections to its sources. Due to its high-purity Ge detectors, it would still resolve  $\gamma$ -ray lines for high-resolution spectroscopy (Tomsick et al. 2019).

Compton binned-mode response developments were sponsored under NASA APRA grant NNX17AC84G. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under contract No. DE-AC02-05CH11231, for COSI response simulations. The COSI instrument developments and balloon flights are

supported through NASA APRA grants NNX14AC81G and 80NSSC19K1389. This work is also supported in part by CNES. C.A.K. is supported by a NASA Postdoctoral Program Fellowship. T.S. is supported by the German Research Society (DFG-Forschungsstipendium SI 2502/1-1).

*Facility:* COSI.

*Software:* MEGAlib (Zoglauer et al. 2006), numpy (Oliphant 2006), matplotlib (Hunter 2007), astropy (Astropy Collaboration et al. 2013), scipy (Virtanen et al. 2020), stan/pystan (Carpenter et al. 2017), arviz (Kumar et al. 2019).

## Appendix A Reliability of the Image Reconstruction

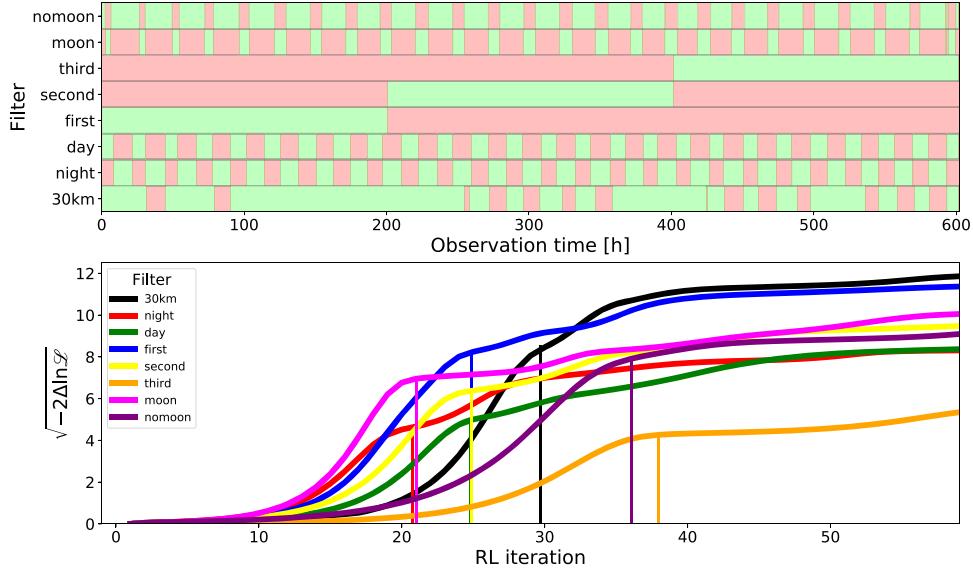
In order to investigate the reliability and stability of our modified Richardson–Lucy deconvolution algorithm, we apply different filters to the full data set to estimate systematics and check for consistency.

From our 603 hr long data set, we select eight subsets to investigate the robustness of the deconvolution algorithm and how this is related to the data quality and environmental conditions: because lower altitudes increase the background count rate, we select times when the balloon was above 30 km. We study the influence and possible contributions of the Sun by splitting the data set into either only night or only day times, defined by the sunset at each Earth latitude and time. For a temporal distinction, we use the first, second, and third 201 hr of observation as individual data sets. Finally, to study whether the Moon albedo is strong enough to show an imprint in the current COSI measurements, we select times when the Moon was in (moon) and outside (nomoon) the field of view.

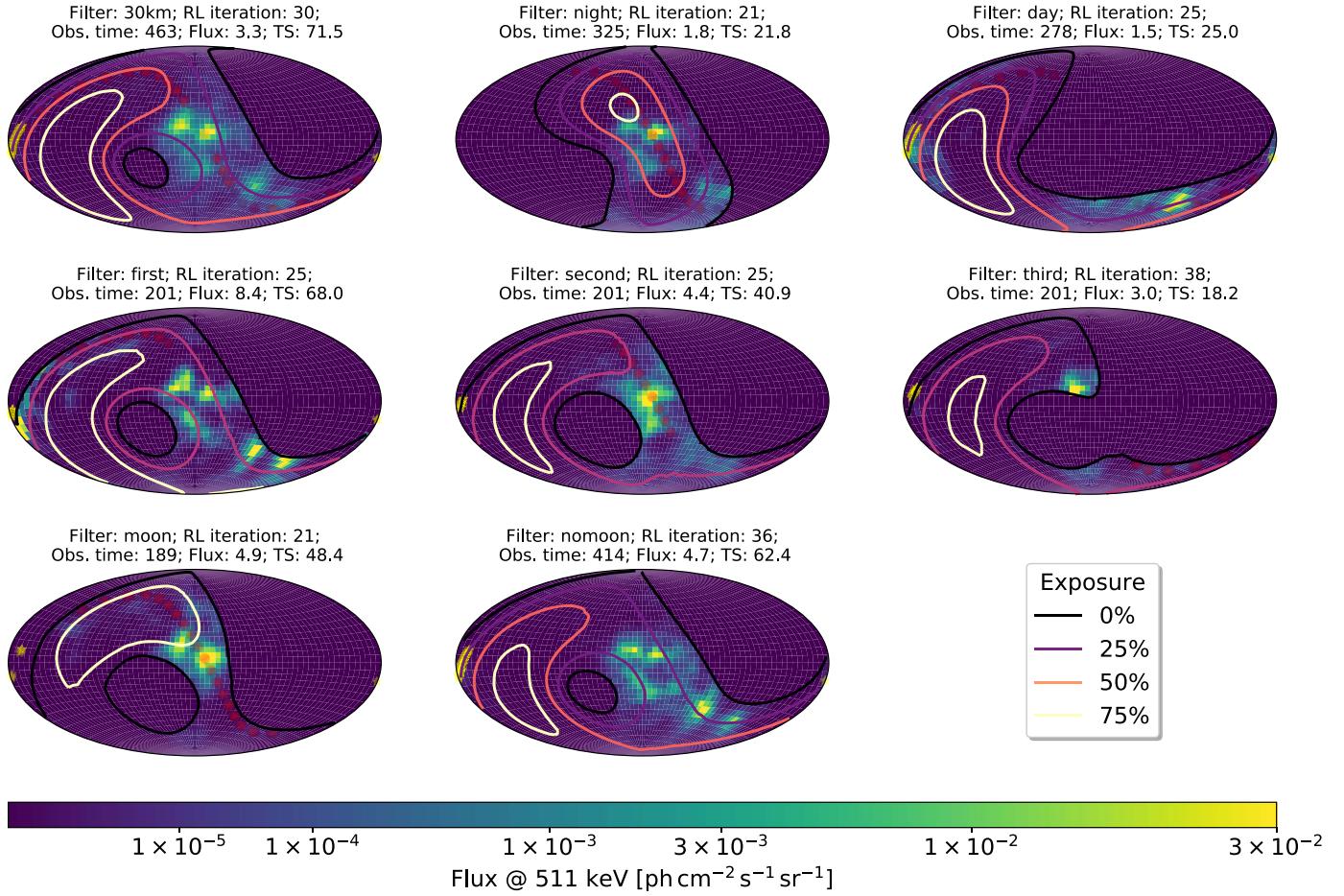
An overview of the selected times is given in the top panel of Figure A1. For each data subset, we use the same modified Richardson–Lucy algorithm as presented in Section 4.1.1 and apply the same choice for which iteration to select as representative. In the bottom panel of Figure A1, the test statistics as a function of iteration is shown, together with the chosen iteration. The range of acceptable deconvolutions spans iterations  $\approx 20$  to  $\approx 40$ , similar to the full data set. The significance of the chosen maps versus a background-only fit ranges between  $4\sigma$  and  $9\sigma$ .

In Figure A2, we show the resulting maps for each filter, together with the respective exposure maps. Clearly, when the bulge region is not masked out owing to specific selections, the bulge always appears bright in 511 keV emission. This is reassuring that the deconvolution algorithm consistently finds the signal and does not pick individual times to assign counts in different sky regions. We note, however, that also the noise peaks at high negative latitudes appear for some selections. The significance of any individual feature is between  $1\sigma$  and  $3\sigma$ , as tested by masking out the feature and then performing a maximum likelihood test to determine the significance of the additional component. The selections second and moon provide the cleanest images, with significances between  $6.4\sigma$  and  $7.0\sigma$ , and 511 keV fluxes between  $4.4$  and  $4.9 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$ . The fluxes are considerably higher than for the total data set; however, they naturally come with a larger uncertainty since the exposure time is one-third or less. Night time (night) selections on the altitude (30 km) and the first 201 observation hours result in images very similar to the full data set (see Figure 11).

<sup>11</sup> <https://www.nasa.gov/press-release/nasa-selects-proposals-to-study-volatile-stars-galaxies-cosmic-collisions>



**Figure A1.** Overview of additionally filtered data sets with green (accepted) and red (rejected) times (top), and significance of the current Richardson–Lucy iteration vs. a background-only fit, for all eight filters with chosen iteration indicated as a vertical line (bottom).



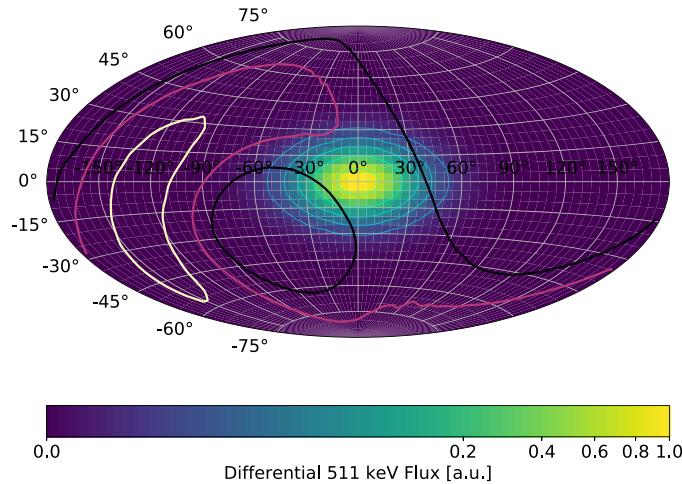
**Figure A2.** Chosen Richardson–Lucy iterations for the eight additionally filtered data sets. Color scheme and exposure map contours are similar to those of Figure 11. In each panel, the filter, the chosen iteration, the observation time in hours, the integrated 511 keV flux in units of  $10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$ , and the test statistics,  $\text{TS} = -2\Delta \ln \mathcal{L}$ , of the map vs. a background-only fit are provided. The positions of the Moon and Sun during the respective data set are indicated in red and yellow.

As a result, we consistently recover emission from the Galactic bulge region and determine the significance of individual emission hot spots outside this region to be between  $1\sigma$  and  $3\sigma$  (see exposure maps in each panel of Figure A2). We therefore cannot claim any additional detection beyond the central region of the Milky Way with COSI measurements from the balloon campaign in 2016.

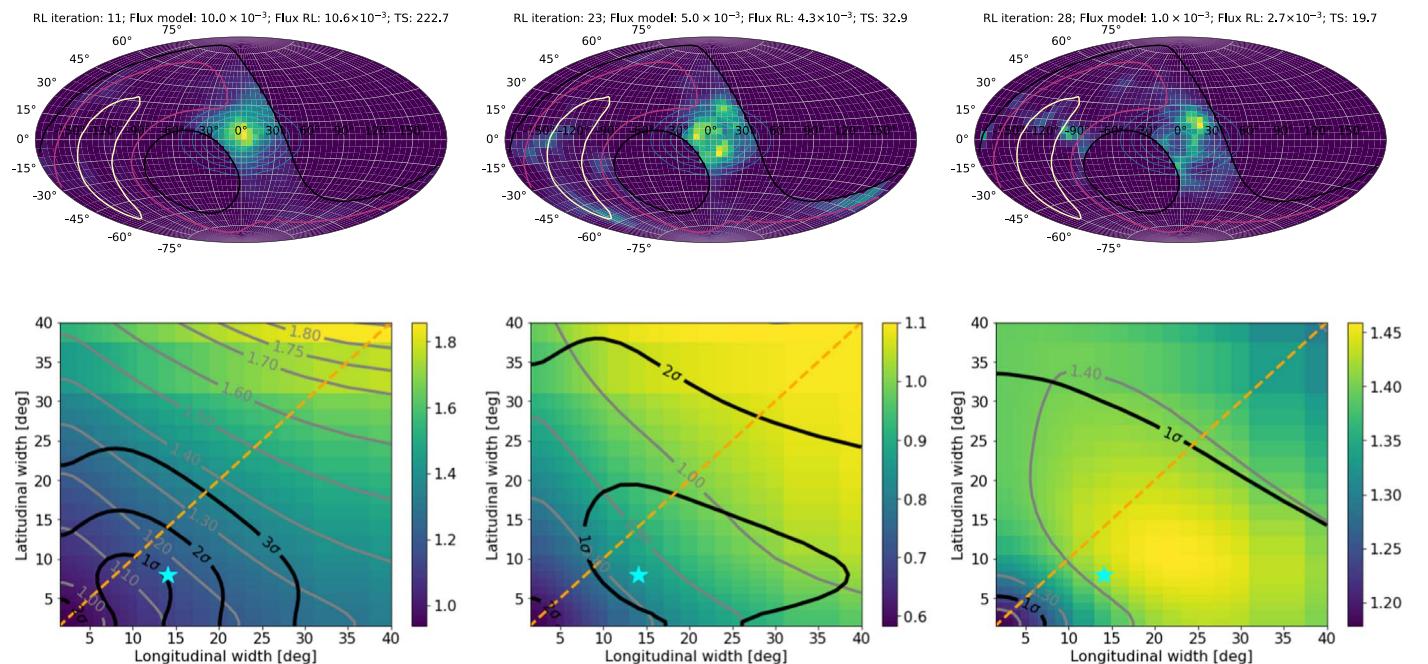
## Appendix B Simulating Data Sets

For statistical and visual comparisons, we simulate a known celestial signal on top of a known background model. We use a

subset of 201 observation hours (see second subset from Appendix A) including the typical characteristics of the 2016 COSI balloon flight to better assess the general quality of image reconstructions with our modified Richardson–Lucy algorithm and maximum likelihood fits. The background is modeled using the defined background response from Section 3.3.2 and using a median filter with a width of 5 hr of the total count rate to define an absolute number. The sky is modeled according to a 2D Gaussian function, located at the Galactic center with longitudinal and latitudinal widths of  $14^\circ$  and  $8^\circ$  ( $1\sigma$  values), respectively. We simulated three different map-integrated fluxes of  $10$ ,  $5$ , and  $1 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$  to obtain characteristic reconstructions and likelihood profiles for varying significances. As the sky model emission extends beyond



**Figure B1.** Simulated sky model (left) and count rates for background and different total fluxes (right). The exposure map of the simulated data set is indicated (similar to Figure A2), showing that emission would be expected outside the exposed regions (black).



**Figure B2.** Reconstruction (top) and maximum likelihood (bottom) results for three different simulated flux levels. From left to right, the model fluxes are  $10$ ,  $5$ , and  $1 \times 10^{-3}$  photons  $\text{cm}^{-2} \text{s}^{-1}$ , respectively. The true model parameters are indicated in cyan in all panels. The flux normalizations (colored) are shown in units of the total flux. The optimal fit is therefore  $1.0$ .

unexposed regions, this approach describes both (i) how structured the resulting map in the image reconstruction algorithm will be and (ii) how accurately a template fitting approach can determine the emission extents. The true sky model, together with the exposure map and the expected count rates for background and the three different fluxes, is shown in Figure B1.

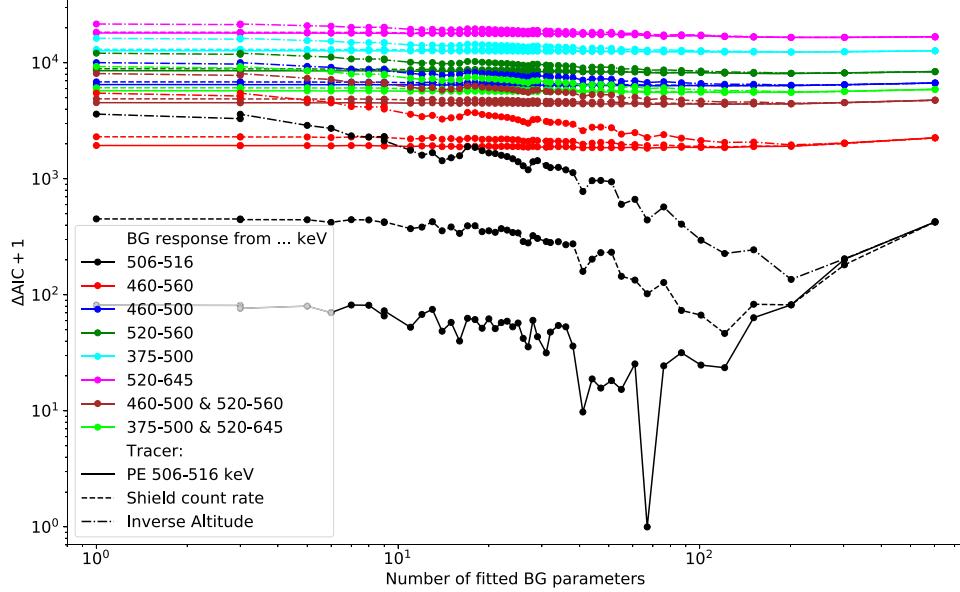
In particular, we draw Poisson samples of the combined models, background plus convolved sky, and run the same image reconstruction as described in Section 4.1.1, including a background fit in each iteration. For determining the image extent, we perform the same profile likelihood as in Section 4.2.1. We note that using exactly one-third of the full data set does not result in a factor of 9 less sensitive measurements, as the frequent altitude drops further decrease the instrument's sensitivity. A summary of reconstructed images and likelihood profiles is shown in Figure B2.

Clearly, the strongest case (left panels in Figure B2) is reliably recovered using our methods, and nearly no image artifacts emerge. The resulting image appears more concentrated when reconstructed with the Richardson–Lucy

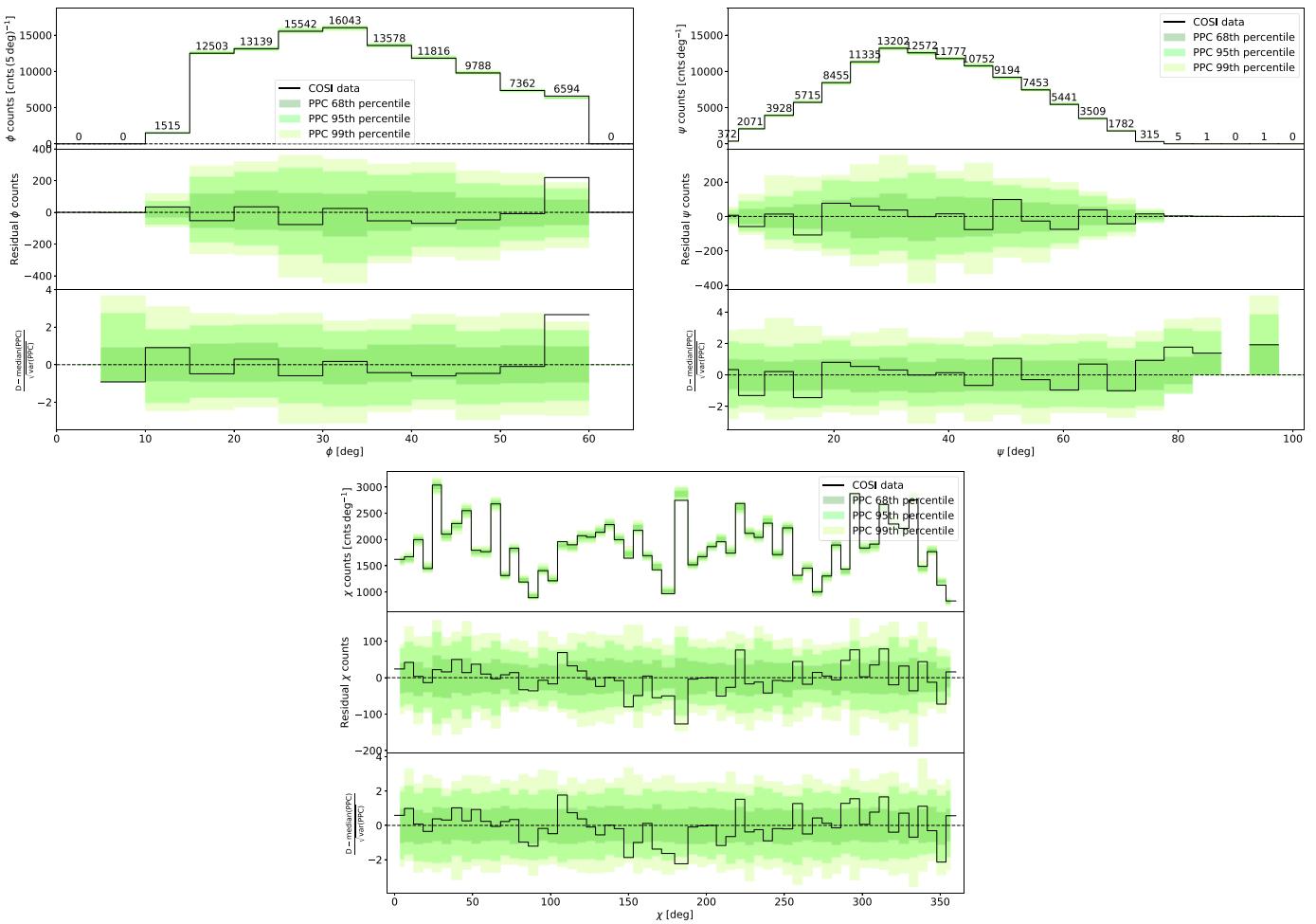
algorithm. Nevertheless, the correct emission extents are recovered within  $1\sigma$  as expected. The middle panels are similar to the real data case: here, the emission appears more structured and stronger artifacts can appear. The uncertainties in emission extent flux are increased, mainly because the information from the underexposed regions is not enough and the flux is too low. The last case (right panels) represents a marginal detection of the signal. Still, the emission is found in the regions close to the Galactic center; however, more artifacts emerge, which results in a skewed flux distribution, as well as an overestimate of the total flux. This is mainly driven by the dominance of the instrumental background over the sky signal.

## Appendix C Additional Figures

Here we show additional information on how different background response-tracer-combinations perform (Figure C1), as well as the fit adequacy in the remaining COSI data space dimensions through PPCs (Figure C2).



**Figure C1.** Performance of all background model combinations.



**Figure C2.** Same as Figure 15, but for the Compton scattering angle  $\phi$  (top left), polar scattering angle  $\psi$  (top right), and azimuthal scattering angle  $\chi$  (bottom). The number above the summed data bins indicates the photons summed over time and the remaining COSI data space angles. Note the asymmetric residuals between small  $\phi$  and large  $\psi$  due to the Poisson character of the counting experiment, leading to a heavily skewed distribution for low numbers.

ORCID iDs

Thomas Siegert  <https://orcid.org/0000-0002-0552-3535>  
John A. Tomsick  <https://orcid.org/0000-0001-5506-9855>  
Theresa J. Brandt  <https://orcid.org/0000-0003-4087-1786>

## References

- Akaike, H. 1974, *ITAC*, **19**, 716

Albernhe, F., le Borgne, J. F., Vedrenne, G., et al. 1981, *A&A*, **94**, 214

Alexis, A., Jean, P., Martin, P., & Ferrière, K. 2014, *A&A*, **564**, A108

Allain, M., & Roques, J. P. 2006, *A&A*, **447**, 1175

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, *A&A*, **558**, A33

Bandstra, M. S., Bellm, E. C., Boggs, S. E., et al. 2011, *ApJ*, **738**, 8

Bisnovatyi-Kogan, G. S., & Pozanenko, A. S. 2017, *Ap*, **60**, 223

Bloemen, H., Morris, D., Knoedlseder, J., et al. 1999, *ApJL*, **521**, L137

Boggs, S. E., Harrison, F. A., Miyasaka, H., et al. 2015, *Sci*, **348**, 670

Boggs, S. E., & Jean, P. 2000, *A&AS*, **145**, 311

Boggs, S. E., Jean, P., Slassi-Sennou, S., et al. 2002, *NIMPA*, **491**, 390

Bouchet, L., Jourdain, E., & Roques, J. P. 2015, *ApJ*, **801**, 142

Bouchet, L., Mandrou, P., Roques, J. P., et al. 1991, *ApJL*, **383**, L45

Bouchet, L., Roques, J. P., & Jourdain, E. 2010, *ApJ*, **720**, 1772

Burnham, K. P., & Anderson, D. R. (ed.) 2004a, *Model Selection and Multimodel Inference* (New York: Springer)

Burnham, K. P., & Anderson, D. R. 2004b, *Sociol. Methods Res.*, **33**, 261

Carpenter, B., Gelman, A., Hoffman, M. D., et al. 2017, *J. Stat. Software*, **76**, 1

Churazov, E., Sazonov, S., Tsygankov, S., Sunyaev, R., & Varshalovich, D. 2011, *MNRAS*, **411**, 1727

- Churazov, E., Sunyaev, R., Isern, J., et al. 2014, *Natur*, **512**, 406  
 Churazov, E., Sunyaev, R., Isern, J., et al. 2015, *ApJ*, **812**, 62  
 Churazov, E., Sunyaev, R., Sazonov, S., Revnivtsev, M., & Varshalovich, D. 2005, *MNRAS*, **357**, 1377  
 Cumani, P., Hernanz, M., Kiener, J., Tatischeff, V., & Zoglauer, A. 2019, *ExA*, **47**, 273  
 Diehl, R., Bennett, K., Collmar, W., et al. 1992, NASCP, **3137**, 95  
 Diehl, R., Halloin, H., Kretschmer, K., et al. 2006, *Natur*, **439**, 45  
 Diehl, R., Siegert, T., Greiner, J., et al. 2018, *A&A*, **611**, A12  
 Diehl, R., Siegert, T., Hillebrandt, W., et al. 2014, *Sci*, **345**, 1162  
 Diehl, R., Siegert, T., Hillebrandt, W., et al. 2015, *A&A*, **574**, A72  
 Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. 2019, *J. R. Stat. Soc. Ser. A*, **182**, 389  
 Gehrels, N. 1985, *NIMPA*, **239**, 324  
 Gelman, A., Meng, X.-L., & Stern, H. 1996, *Stat. Sin.*, **6**, 733  
 Grebenev, S. A., Lutovinov, A. A., Tsygankov, S. S., & Winkler, C. 2012, *Natur*, **490**, 373  
 Grefenstette, B. W., Fryer, C. L., Harrison, F. A., et al. 2017, *ApJ*, **834**, 19  
 Grefenstette, B. W., Harrison, F. A., Boggs, S. E., et al. 2014, *Natur*, **506**, 339  
 Guessoum, N., Jean, P., & Prantzos, N. 2006, *A&A*, **457**, 753  
 Guttman, I. 1967, *J. R. Stat. Soc. Ser. B*, **29**, 83  
 Halloin, H. 2009, piorthomodel, Explanatory Guide and Users Manual, Version 2.0, Vol. 1 (Garching: Max Planck Institut für extraterrestrische Physik), 85748  
 Harris, M. J., Teegarden, B. J., Cline, T. L., et al. 1998, *ApJL*, **501**, L55  
 Higdon, J. C., Lingenfelter, R. E., & Rothschild, R. E. 2009, *ApJ*, **698**, 350  
 Hoffman, M. D., & Gelman, A. 2011, arXiv:1111.4246v1  
 Hoffman, M. D., & Gelman, A. 2014, *J. Mach. Learn. Res.*, **15**, 1593  
 Hunter, J. D. 2007, *CSE*, **9**, 90  
 Isern, J., Jean, P., Bravo, E., et al. 2016, *A&A*, **588**, A67

- Iyudin, A. F., Diehl, R., Lichten, G. G., et al. 1997, in ESA SP-382, The Transparent Universe, ed. C. Winkler, T. J. L. Courvoisier, & P. Durouchoux (Noordwijk: ESA), 37
- Jean, P., Gillard, W., Marcowith, A., & Ferrière, K. 2009, *A&A*, 508, 1099
- Jean, P., Knödlseder, J., Gillard, W., et al. 2006, *A&A*, 445, 579
- Jean, P., Vedrenne, G., Roques, J. P., et al. 2003, *A&A*, 411, L107
- Johnson, W. N., Kinzer, R. L., Kurfess, J. D., et al. 1993, *ApJS*, 86, 693
- Johnson, W. N. I., & Haymes, R. C. 1973, *ApJ*, 184, 103
- Kaufman, L. 1987, *IEEE Trans. Med. Imaging*, 6, 37
- Kierans, C. 2018, PhD thesis, Univ. California, Berkeley, <https://escholarship.org/uc/item/1244t3h7>
- Kierans, C., Boggs, S., Chiu, J. L., et al. 2016, in Proc. 11th INTEGRAL Conf. Gamma-Ray Astrophysics in Multi-Wavelength Perspective (Trieste: SISSA), 75
- Kierans, C. A., Boggs, S. E., Zoglauer, A., et al. 2020, *ApJ*, 895, 44
- Knoedlseder, J., Dixon, D., Bennett, K., et al. 1999, *A&A*, 345, 813
- Knoedlseder, J., Jean, P., Lonjou, V., et al. 2005, *A&A*, 441, 513
- Knoedlseder, J., von Ballmoos, P., Diehl, R., et al. 1996, in SPIE's 1996 Int. Symp. on Optical Science, Engineering, and Instrumentation, ed. B. D. Ramsey & T. A. Parnell (Bellingham, WA: SPIE), 386
- Koehler, J. J. 1993, *Organ. Behav. Hum. Decis. Processes*, 56, 28
- Kretschmer, K., Diehl, R., Krause, M., et al. 2013, *A&A*, 559, A99
- Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. 2019, *JOSS*, 4, 1143
- Leventhal, M., MacCallum, C. J., Huters, A. F., & Stang, P. D. 1986, *ApJ*, 302, 459
- Leventhal, M., MacCallum, C. J., & Stang, P. D. 1978, *ApJL*, 225, L11
- Ling, J. C. 1975, *JGR*, 80, 3241
- Ling, J. C., Mahoney, W. A., Willett, J. B., & Jacobson, A. S. 1977, *JGR*, 82, 1463
- Lingenfelter, R. E., & Ramaty, R. 1989, *ApJ*, 343, 686
- Lucy, L. B. 1974, *AJ*, 79, 745
- Lucy, L. B. 1992, *AJ*, 104, 1260
- Milne, P. A., & Leising, M. D. 1997, in AIP Conf. Proc. 410, The Fourth Compton Symp., ed. C. D. Dermer, M. S. Strickman, & J. D. Kurfess (Melville, NY: AIP), 1017
- Morris, D. J., Bennett, K., Bloemen, H., et al. 2006, *NYASA*, 759, 397
- Nickerson, R. S. 1998, *Rev. Gen. Psychol.*, 2, 175
- Oberlack, U., Bennett, K., Bloemen, H., et al. 1996, *A&AS*, 120, 311
- Oliphant, T. E. 2006, A Guide to NumPy, Vol. 1 (USA: Trelgol Publishing)
- Panther, F. 2018, *Galax*, 6, 39
- Pleintinger, M. M. M., Siegert, T., Diehl, R., et al. 2019, *A&A*, 632, A73
- Pohl, R. 2004, Cognitive Illusions, A Handbook on Fallacies and Biases in Thinking, Judgement and Memory (Hove: Psychology Press)
- Prantzos, N. 2006, *A&A*, 449, 869
- Prantzos, N., Boehm, C., Bykov, A. M., et al. 2011, *RvMP*, 83, 1001
- Purcell, W. R., Cheng, L. X., Dixon, D. D., et al. 1997, *ApJ*, 491, 725
- Purcell, W. R., Grabelsky, D. A., Ulmer, M. P., et al. 1993, *ApJL*, 413, L85
- Richardson, W. H. 1972, *JOSA*, 62, 55
- Rubin, D. B. 1981, *J. Educ. Stat.*, 6, 377
- Rubin, D. B. 1984, *AnSta*, 12, 1151
- Sato, T. 2016, *PLoS ONE*, 11, e0160390
- Scargle, J. D. 1998, *ApJ*, 504, 405
- Scargle, J. D., Norris, J. P., Jackson, B., & Chiang, J. 2012, Bayesian Blocks: Detecting and characterizing local variability in time series, *Astrophysics Source Code Library*, ascl:1209.001
- Shepp, L. A., & Vardi, Y. 1982, *IEEE Trans. Med. Imaging*, 1, 113
- Siegert, T., Crocker, R. M., Diehl, R., et al. 2019a, *A&A*, 627, A126
- Siegert, T., Diehl, R., Greiner, J., et al. 2016a, *Natur*, 531, 341
- Siegert, T., Diehl, R., Khachatryan, G., et al. 2016b, *A&A*, 586, A84
- Siegert, T., Diehl, R., Krause, M. G. H., & Greiner, J. 2015, *A&A*, 579, A124
- Siegert, T., Diehl, R., Weinberger, C., et al. 2019b, *A&A*, 626, A73
- Skinner, G., Diehl, R., Zhang, X., Bouchet, L., & Jean, P. 2014, in Proc. 10th INTEGRAL Workshop: "A Synergistic View of the High-Energy Sky" (INTEGRAL 2014) (Trieste: SISSA), 054, <http://pos.sissa.it/cgi-bin/reader/conf.cgi?confid=228>
- Skinner, G., Jean, P., Knoedlseder, J., et al. 2012, in Proc. "An INTEGRAL View of the High-energy Sky (the first 10 years)"—9th INTEGRAL Workshop and Celebration of the 10th Anniversary of the Launch (INTEGRAL 2012) (Trieste: SISSA), 112
- Sleator, C. 2019, PhD thesis, Univ. California, Berkeley, <https://escholarship.org/uc/item/0zn566rj>
- Sleator, C. C., Zoglauer, A., Lowell, A. W., et al. 2019, *NIMPA*, 946, 162643
- Sunyaev, R., Churazov, E., Gilfanov, M., et al. 1992, *ApJL*, 389, L75
- Tomsick, J. A., Zoglauer, A., Sleator, C., et al. 2019, arXiv:1908.04334
- Tsygankov, S. S., Krivonos, R. A., Lutovinov, A. A., et al. 2016, *MNRAS*, 458, 3411
- van Dijk, R. 1996, PhD thesis, Univ. Amsterdam
- Vedrenne, G., Roques, J. P., Schönfelder, V., et al. 2003, *A&A*, 411, L63
- Vink, J., Laming, J. M., Kaastra, J. S., et al. 2001, *ApJL*, 560, L79
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *NatMe*, 17, 261
- von Ballmoos, P., Diehl, R., & Schoenfelder, V. 1989, *A&A*, 221, 396
- Weidenspointner, G., Skinner, G. K., Jean, P., et al. 2008, *Natur*, 451, 159
- Winkler, C., Courvoisier, T. J. L., di Cocco, G., et al. 2003, *A&A*, 411, L1
- Zoglauer, A., Andritschke, R., & Schopper, F. 2006, *NewAR*, 50, 629
- Zoglauer, A., Boggs, S. E., Andritschke, R., & Kanbach, G. 2007, *Proc. SPIE*, 6700, 670001
- Zoglauer, A., & Kanbach, G. 2003, *Proc. SPIE*, 4851, 1302
- Zoglauer, A. C. 2006, PhD thesis, Technical Univ. Munich