Supplemental material

EVOLUTION AND STABILITY OF THE DATA

The added data in this analysis comprises a total of 314.6 hours of tracking and 6227 hours of background data. The data taking periods under axion sensitive conditions for each of the three datasets are illustrated in Fig. 1 as blue regions. The periods without solar tracking (white) were used to perform any necessary interventions and also to increase the background statistics. The time evolution of some relevant parameters corresponding to each of the periods is depicted in Fig.2. During the two argon datasets (datasets 1 and 2), the gain was relatively unchanged with time and, most importantly, the software efficiency was very stable, meaning that despite the gain variations we were able to efficiently identify X-ray-like events in every run. The first few runs in dataset 1 have a slightly lower efficiency as the parameters of the electronics were still being optimized. It is worth noting that these efficiencies are calculated with respect to the number of events with a single track. However, 1-track events are 95% of the total events in the argon case, and 97% in the xenon case at 5.9 keV. For the xenon dataset there were strong gain variations. Our hypothesis is that these were due to the gas recirculation system affecting the gas quality in different ways:

- The moisture and oxygen filters get saturated with time, and have to be changed periodically.
- The moisture filter was found to be a radon emanator. This was introducing alpha particles in the gas that produced trips, so that the mesh voltage had to be modified accordingly and this directly effected the gain. No clear effect of the alphas in the low energy background was found.
- The total volume of the recirculation system was $\sim$ 7 liters. The mylar window permeation is different for each of the components of the gas (Xe, Ne and isobutane), leading to a small but steady change in the gas composition, which increases the relative amount of isobutane with time.

Each time the gain was low, some intervention was done: either changing filters, injecting fresh gas or modifying the voltages. Despite these gain variations, the efficiency of the X-ray cuts was close to 90% during most of the runs, which means we were still able to identify X-ray-like events properly. However, during December 2020, the gain dropped to nearly 20% of the initial gain. As shown in Fig. 3, during this period the energy resolution at 5.9 keV also increased to up to 50%, and the energy threshold steadily went up to 2 keV. Out of the 109 solar tracking runs taken with xenon mixtures, 6 have an energy threshold > 1 keV.

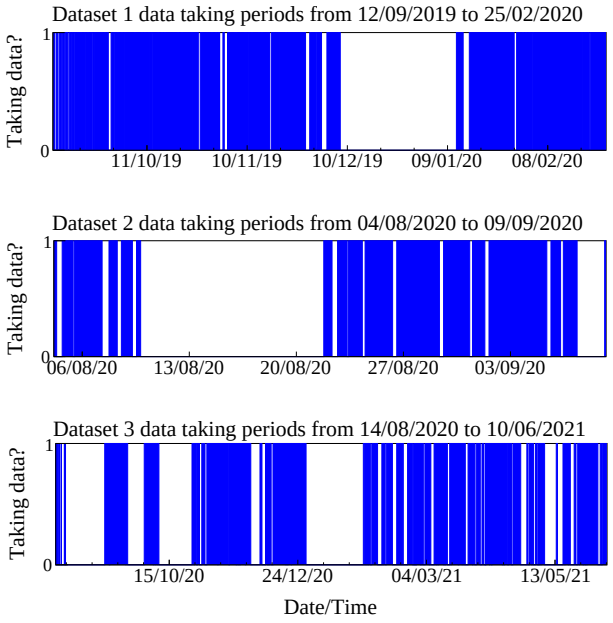

FIG. 1. Data taking periods. Each subplot shows the full time extent of each dataset. The blue shaded regions are periods in which both tracking and background data were taken. During some of the white regions, background data was also taken.

DATA ANALYSIS AND EVENT DISCRIMINATION

The data analysis has been performed with the REST-for-physics (Rare Event Searches Toolkit for Physics) framework [1], and it can be fully replicated using the official version 2.4.0. The raw data in the form of voltage pulses are read from the detector and turned into a set of digitized waveforms, which are dubbed *signal events* in REST. The shape of the raw signals depends on parameters such as the shaping time, trigger delay or time duration of the acquisition window. Enough bins without signal are left at the beginning of the window, to be able to compute a baseline. After some processing and taking into account the detector and readout descriptions, these events turn into *hits events* with a given physical position and energy (and relative time). These hits can be grouped into *track events* if they fulfill a set of conditions that classify them as belonging to the same physical event, such as the maximum 3-dimensional distance (XYZ, or equivalently XY-time). This analysis chain is depicted in Fig. 4.

The use of a stripped readout allows for topological analysis of the *track events*, which enables the definition

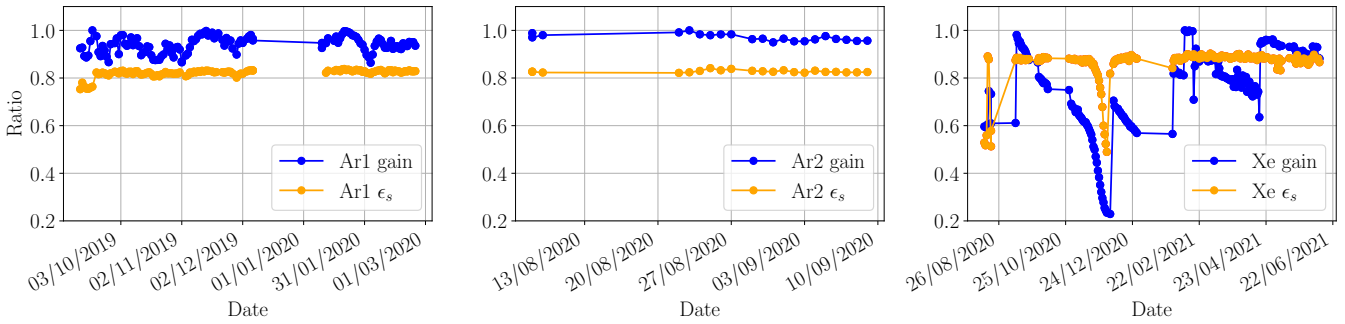

FIG. 2. Time evolution of the gain based on the position of the ^{55}Fe peak of the daily calibrations, expressed in relative terms, and of the software efficiency ϵ_s. Left for dataset 1, centre for dataset 2, and right for dataset 3.

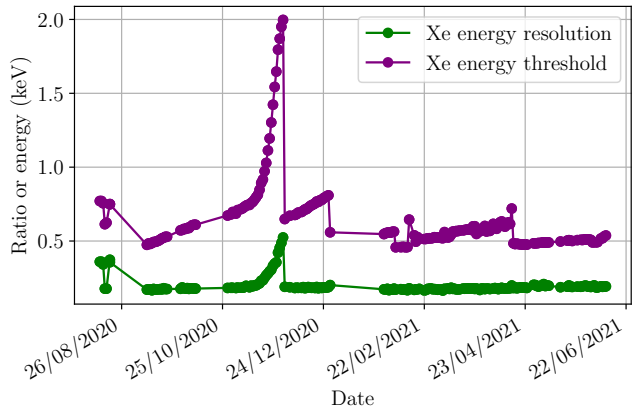

FIG. 3. Time evolution of the energy resolution and energy threshold in dataset 3. Each point corresponds to a calibration run, but not all of them have a tracking run associated. For example, of all the points with energy threshold above 1 keV, only 6 do have data taken during tracking.

of selection algorithms based on the topological shape of the events. As seen in Fig. 4, X-ray like events are small, symmetric, point like and consist of a single track, whereas background events have multiple tracks and are highly asymmetric (e.g. muon or alpha particle tracks). The process of Micromegas data cut optimization used in this work is explained in [2]. A further cut based on muons is applied. If an event in the Micromegas happens just after a muon signal is detected, the event is removed. Random coincidences are expected to be less than 0.5%, giving this cut a very high efficiency while simultaneously reducing the background level by a factor ~ 1.7 .

EFFICIENCIES AND DETECTOR RESPONSE

The efficiencies, illustrated in Fig. 5, are energy dependent and can be classified as follows:

- Detector efficiency takes into account how many X-rays of each energy go through the mylar window and are absorbed in the gas. This efficiency is later corrected by the detector response as explained below.

- Software efficiency is defined for each energy range as the rate of calibration events that remain after applying the Micromegas and veto cuts. The energy threshold is included in this efficiency by setting it to 0 for energies below the energy threshold.

- Telescope efficiency is the efficiency of the optic measured at the MPE PANTER X-ray test facility in Munich in July 2016, and it is optimized for axion searches, maximizing the telescope throughput at low energies.

- Total efficiency is the product of all the efficiencies.

Monte Carlo simulations of the detector response have been performed with REST-for-physics [1] and Geant4 [3]. A flat X-ray flux between 0 and 12 keV from a point source placed in the center of the copper pipe, 100 mm away from the the detector window, was simulated.

The output spectrum from the simulation is a result of the photons that were transmitted through the 4 microns aluminized Mylar window and interacted with the gas in the chamber, thus depositing their energy fully or partially. If one only considers events that deposit all their energy to count towards the efficiency, there is an efficiency loss because if the energy of the incoming photon is high enough (e.g. $\gtrsim 3$ keV in Ar), it can produce a fluorescence peak and deposit less energy than the photon originally carried.

To consider also these events towards computing the efficiency, the output of the simulations is used to build a detector response matrix $\mathcal{M}_{DR}$ with 2-dimensional bins of size 0.1×0.1 keV^2 shown in Fig. 6, where the energy of primary events is mapped to the deposited energy. It also encodes the Mylar window transmission, the gas absorption and the already mentioned higher order energy shifting effects. One can see that most of the events lie on the

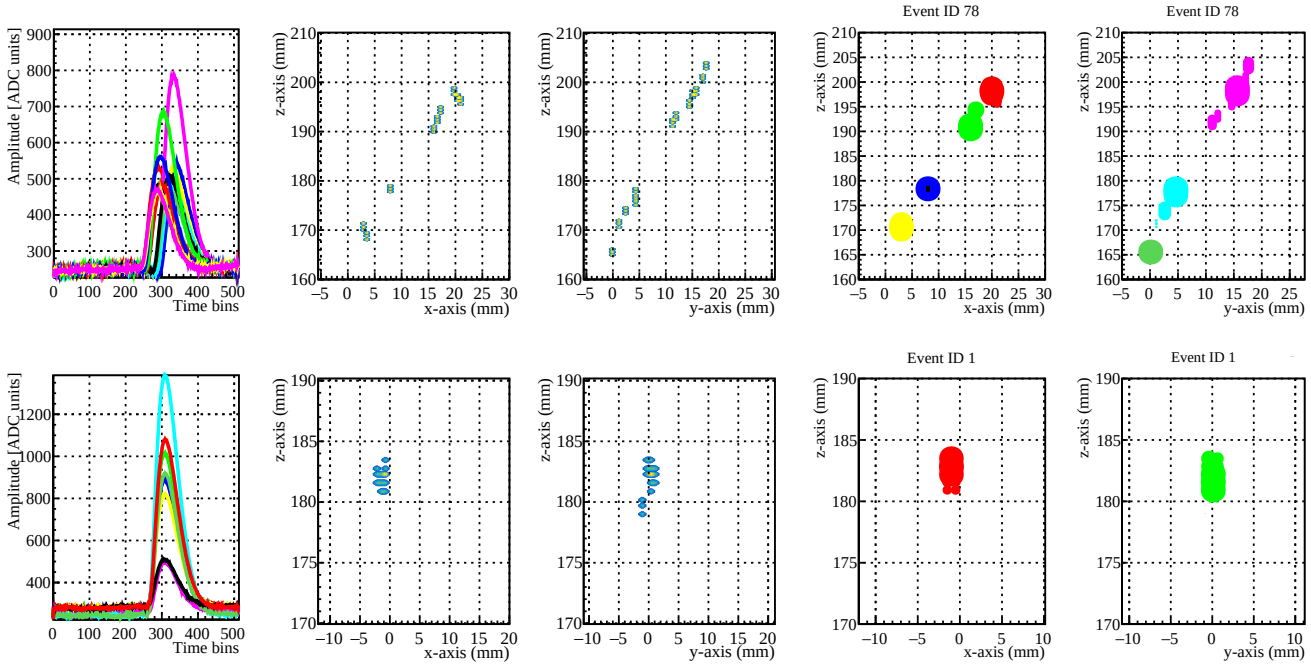

FIG. 4. Data analysis chain with the REST-for-physics framework, the top line for a background event and the bottom for a calibration event. From left to right, the *signal event*, the *hits event* in the XZ plane and in the YZ plane, and the *track event* again in the XZ and YZ planes. Each colour in a *track event* represents a different cluster.

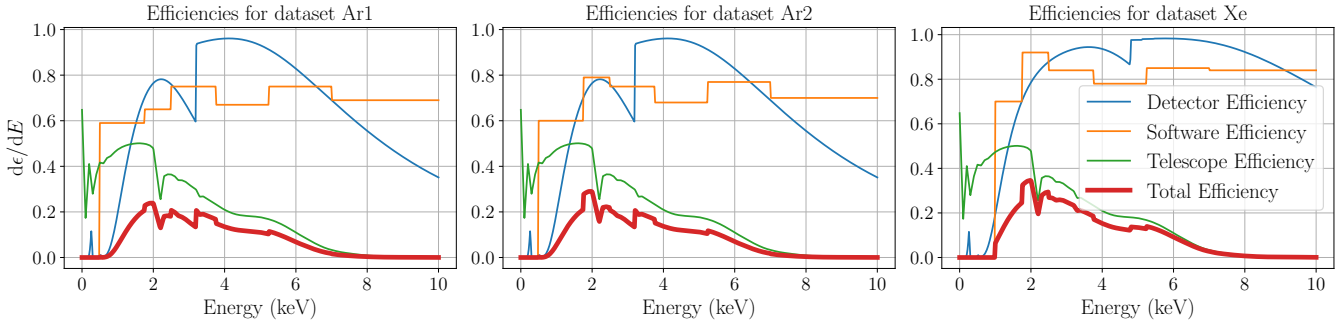

FIG. 5. Depiction of the total efficiency for each dataset and the separate components: detector (meaning gas absorption and window transmission) in blue, software in orange and telescope in green. The energy threshold is taken into account as part of the software efficiency, and is represented as a sharp drop at 0.5 keV for Ar and 1 keV for Xe. The combined total efficiency is drawn in red. A horizontal line at 1 would represent 100% efficiency.

main diagonal, but there are also clear fluorescence lines from copper Kα and Kβ at 8 and 8.9 keV that appear as horizontal lines, and an accumulation of events at any point 3.2 keV below the main diagonal, being 3.2 keV the binding energy of the innermost Ar electron. Once this matrix is built, it can be convoluted with the incoming flux, in our case given by the Primakoff spectrum folded with the axion-photon conversion probability $P_{a\to\gamma}$ and the optics efficiency ϵ_o. This flux is thus a vector of the number of X-rays at energy E_i reaching the detector. A linear combination of this vector and $\mathcal{M}_{DR}$ gives the detected number of events with their corresponding energies in the gas. Finally one needs to incorporate the software efficiency. Based on measurements taken at 6 different energies in the CAST X-ray lab, we were able to keep this efficiency between 80% and 90% in the energy range of interest, and between 60% and 70% in the lowest energy range (Fig. 5).

Regarding the detector efficiency, the use of Xe-based gas mixtures, which have an inherently higher efficiency than Ar, as well as the removal of a differential cold window which was not required for vacuum operation, have also helped to increase the overall setup efficiency. All these improvements have affected the final result in a

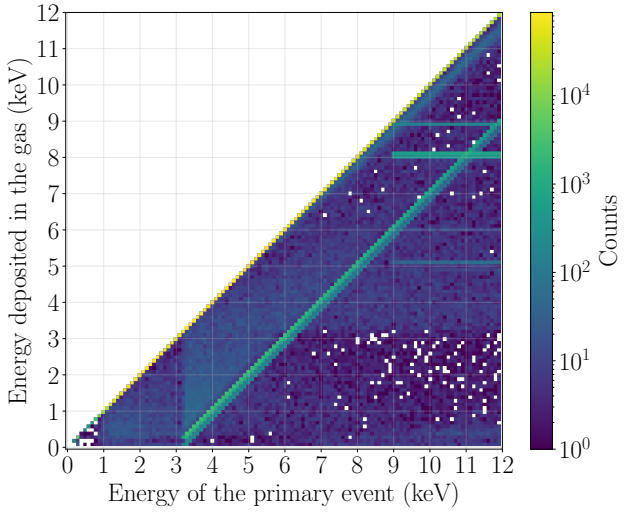

FIG. 6. Visualization of the detector response matrix with $0.1 \times 0.1\,\mathrm{keV}^2$ bins.

positive way, pushing the upper limit on $g_{a\gamma}$ lower than it would have been possible with past strategies.

The final efficiency is about twice as high as the efficiency obtained in the last results [4], and as seen in Fig.7, the improvement is especially noticeable at low energies where the expected solar axion flux is much higher. This has a significant impact in the final result.

SYSTEMATIC EFFECTS

We estimate the systematic effects for the most relevant sources of uncertainty which we describe here. The results are listed in Table I.

Magnetic field B and magnet length L – The CAST magnet was run at a very stable current of 13 kA with negligible deviations from that value. The corresponding magnetic field is obtained by fitting data relating current and magnetic field (see e.g. [5]). A linear fit is performed and the error in the fit parameters is used to compute the uncertainty by error propagation taking correlation into account. The estimated uncertainty is $B = 8.805 \pm 0.034\,\mathrm{T}$. The nominal value for the magnet length provided by the magnet group is $L = (9.26 \pm 0.05)\,\mathrm{m}$. A 0.93346 m long haloscope cavity was installed inside the bore, reducing the effective magnetic field length.

Background level – The background level used for the calculations corresponds to the central readout area circle with $r = 10\,\mathrm{mm}$, with a bin width of 1 keV and errors $\sqrt{N}$, where N is the number of background counts. The background spectra that were included in the limit calculation are shown in Fig. 8. The systematic uncertainty is computed by considering the lowest/highest possible background, which is estimated by redefining the bin

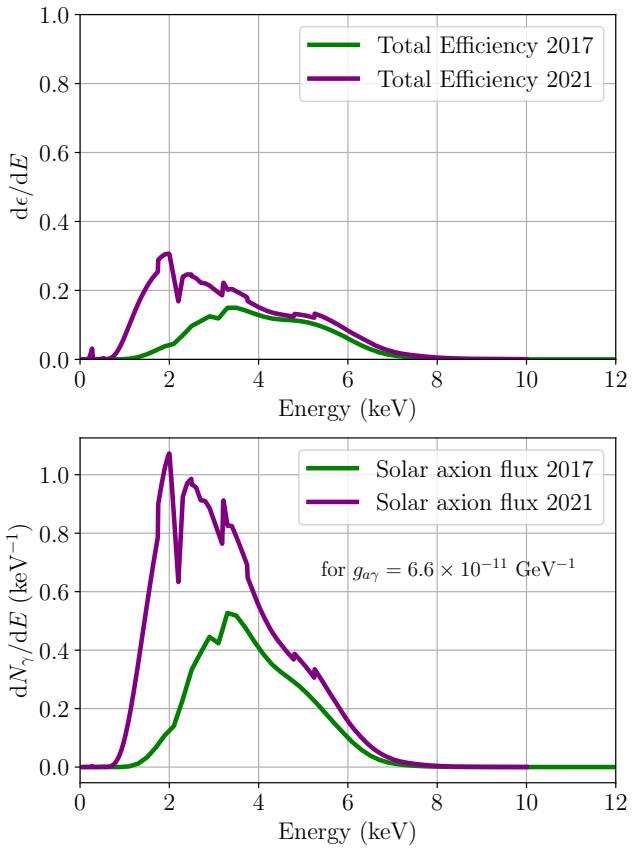

FIG. 7. (top) Plot of the average total efficiency of the three datasets, where a horizontal line at $\mathrm{d}\epsilon/\mathrm{d}E = 1$ represents 100% efficiency (similar to Fig. 5). This is compared to the efficiency obtained in the 2017 data [4], highlighting our improved efficiency, especially at low energies where the solar axion flux is expected to peak. (bottom) Spectral distribution of the expected solar axion signal counts N_γ, corrected for efficiencies and integrated over time and area of the bore. The expected signal is twice as high due to the increased efficiency shown in the top panel.

TABLE I. Systematic errors for different sources of uncertainty, quantified as the relative shift (upwards and downwards) of the upper limit on $g_{a\gamma}$ due to each particular source.

Source of uncertainty	Systematic effect in $g_{a\gamma}$	
Magnetic field strength B	−0.16%	+0.22%
Magnetic field length L	−0.27%	+0.27%
Background level	−0.22%	+0.22%
Background area	−1.56%	+0.22%
Software efficiency	−1.11%	+1.17%
Spot position	−0.22%	+0.12%
Pointing accuracy	−0.00%	+0.27%
Theoretical axion flux	−0.39%	+0.33%
Solar model type	−0.00%	+1.29%
Total systematic uncertainty	−2.00%	+1.85%

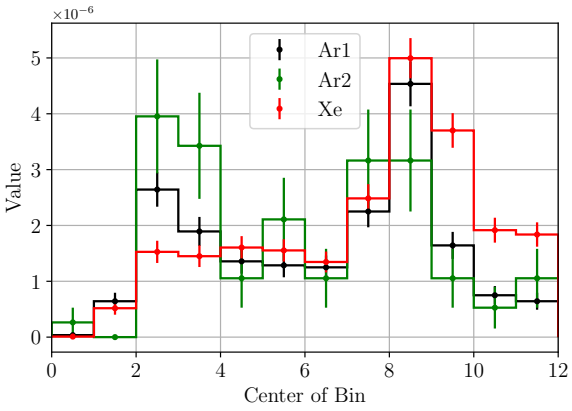

FIG. 8. Background level for the central region with $r = 10\,\text{mm}$ in $\text{keV}^{-1}\,\text{cm}^{-2}\,\text{s}^{-1}$ for Ar dataset 1 in black, Ar dataset 2 in green and and Xe in red. The error bars were used to compute the possible deviations.

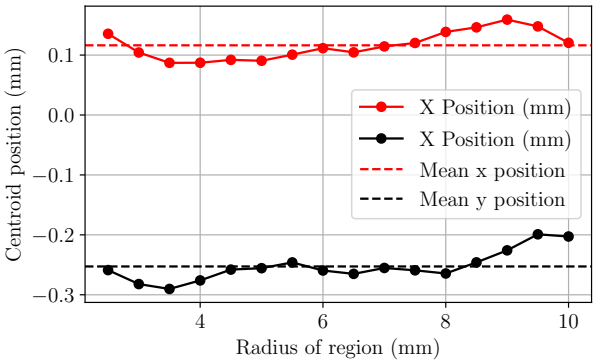

FIG. 9. Position of the centroid for different considered areas. The mean value was considered nominal, and the deviations were used to compute the uncertainty.

heights to match the lower/upper error bar endpoints. We assume that the background is spatially uniform, but if we consider only a small area this assumption might not hold due to the low counts. Therefore, the background definition will be affected by the considered area. We thus compute the uncertainty in the result caused by using the background level of the inner readout area circle with $r = 4\,\text{mm}$, and applying it to that same area and to the nominal $r = 10\,\text{mm}$ area.

Software efficiency – An algorithm has been designed to define the X-ray cuts for different energy ranges, which also affects the software efficiency in a positive way. This method yields a higher efficiency than other approaches used in the past and it maximizes the figure of merit. The results have been tested using the 3 keV Ar escape peak and were found to be consistent, so a conservative 5% uncertainty in the software efficiency is taken into account.

Alignment – The center of the spot was defined by computing the centroid of the data taken with the X-ray calibration source at CAST. The variation of the position of the centroid for circular readout areas with radii in the range $r = (2.5, 10)\,\text{mm}$ is shown in Fig. 9. The standard deviation of the centroid is 0.025 mm, which produces a negligible change in the limit. Therefore, a more conservative value of 0.1 mm, based on the range of the centroid values, was considered.

Pointing accuracy– The pointing accuracy of the CAST experiment is 0.01°. The CAST bore has an angular size of 0.5° while the solar axion signal comes from the inner 20% of the Sun, resulting in 0.1°. Thus, the solar core is always contained within the bore aperture. According to measurements taken in PANTER on- and off-axis, this results in a negligible displacement of the centroid on the readout plane. However, displacement can lead to a loss of up to 1% efficiency.

Solar axion flux– Uncertainties in the theoretical solar axion flux and their effects in solar axion searches are studied in [6]. The flux can be expected to have a statistical fluctuation of $\sim 1.5\%$, and it is also model dependent. Helioseismological models, based on the internal structure and dynamics of the Sun using information from internal sound waves, yield Primakoff fluxes that are consistently $\sim 5\%$ higher than those predicted by photospheric models. As axions are expected to be produced in the inner 20% of the Sun, the former type of model is more appropriate, and thus we compute the model uncertainty by considering a flux 5% lower. The particular expression used in this work is derived in [7].

Total uncertainty– As the uncertainty sources are independent, the total uncertainty is calculated simply as the root sum of squares $\sqrt{\sum_{i=1}^{n} u_i^2}$ of the uncertainties u_i of the i sources, and it is defined by the asymmetric range -2.00% and +1.85%.

The background definition, the solar axion flux and the software efficiency are the most relevant uncertainties, and the rest are mostly negligible. Still, all these contributions are minor compared to the statistical error inherent to rare-event experiments.

EXPECTED SENSITIVITY

To compute the expected sensitivity of the experiment one needs to simulate sets of candidates (i.e., X-ray like events during tracking) that follow the background distribution in a given readout area, for a given exposure time and energy interval. In a binned case, a representative dataset such as the Asimov dataset [8] provides an efficient solution to compute an expected limit. However, in an unbinned case a set of simulations has been performed instead, which also allows to study the possible variation due to statistical fluctuations. The expected sensitivity

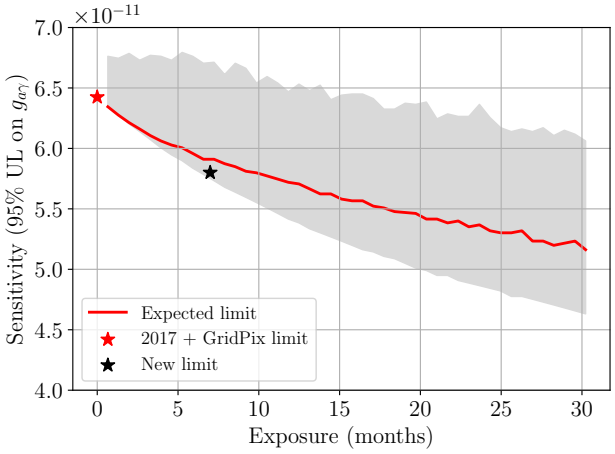

FIG. 10. Expected sensitivity of the experiment using Monte Carlo simulations given the current setup, background level, efficiencies, etc., and starting from the limit already set by the 2013-2018 data taking campaigns. The red line represents the expected limit computed as the median of the 1000 limits simulated in steps of 10 days of data taking. The time represented in the x-axis is the data acquisition time at CAST and the sensitivity is computed assuming 1.5 h of axion sensitive conditions per day. The grey shaded area is the band of possible limits for each exposure time, where the upper bound is the 99$^{\text{th}}$ percentile.

of the experiment can be simulated as follows:

- For each exposure time step, draw a set of candidates that follow a Poisson distribution based on the background rate. This step is done 1000 times in time steps of 10 days, assuming 1.5 h of data taking under axion sensitive conditions per day.
- Compute the limit that each of these set of candidates would produce.
- For each exposure time step, compute the median of the simulated limits as the expected limit.

The results of the simulations are shown in Fig. 10, where the red line is the expected limit for a given exposure time and the grey band represents the range of possible values. In the current case, the expected sensitivity for 7 months of data taking is $g_{a\gamma} < 0.59 \times 10^{-10}\,\text{GeV}^{-1}$, which is in agreement with our result.

[1] K. Altenmüller *et al.*, REST-for-Physics, a ROOT-based framework for event oriented data analysis and combined Monte Carlo response, Comput. Phys. Commun. **273**, 108281 (2022).

[2] K. Altenmüller *et al.*, Background discrimination with a Micromegas detector prototype and veto system for BabyIAXO, Front. Phys. **12**, 1384415 (2024).

[3] S. Agostinelli *et al.*, Geant4—a simulation toolkit, Nucl. Instrum. Meth. A **506**, 250 (2003).

[4] V. Anastassopoulos *et al.* [CAST Collaboration], New CAST limit on the axion-photon interaction, Nature Phys. **13**, 584 (2017).

[5] J. K. Vogel, Searching for solar axions in the eV-mass region with the CCD detector at CAST, Ph.D. thesis, Albert-Ludwigs-Universität Freiburg, Freiburg, April 2009.

[6] S. Hoof, J. Jaeckel and L. J. Thormaehlen, Quantifying uncertainties in the solar axion flux and their impact on determining axion model parameters, JCAP **2021**, no. 9, 006 (2021).

[7] S. Andriamonje *et al.* [CAST Collaboration], An improved limit on the axion-photon coupling from the CAST experiment, JCAP **0704**, 010 (2007).

[8] G. Cowan, K. Cranmer, E. Gross and O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, Eur. Phys. J. C **71**, 1554 (2011).