**Abstract**

We describe gPhoton, which enables more flexible analysis of GALEX ultraviolet data, including high time resolution photometric analyses, by sky projecting and hosting approximately 1.1 trillion individual photon events and providing a tool set for generating calibrated light curves and images with a large large number of user-definable parameters. The project includes a standalone, pure-Python calibration pipeline that reproduces core functionality of the original mission pipeline, an approximately 130 Tb publicly accessible database of photon level data hosted by the Mikulski Archive at Space Telescope, and command line tools to enable reuction and use of the photon data for common classes of analyses, including the creation of calibrated light curves of individual sources or calibrated images and movies of sourcs and fields. The gPhoton software and source code are publicly available under a permissive license and will continue to be developed and maintained in the future. We describe the motivation, design, and implementation of the calibration pipeline, database, and tools, with emphasize on divergence from the mission pipeline and challenges of working with large volumes of photon-level data. We also highlight, with examples, new types of scientific investigation made possible by this project.

**Introduction**

The Galaxy Evolution Explorer (GALEX) was a NASA Small Explorer (SMEX) telescope that surveyed the sky in the ultraviolet over approximately ten years, between launch on 28 April 2003 and spacecraft termination on 28 June 2013. The spacecraft, instruments, data and calibration are well described in previous publications [CITE] and the mission’s technical documentation. [CITE] We will restrict discussion to topics which have not appeared elsewhere in the literature, are of particular importance to the gPhoton project, or are otherwise necessary for completeness.

GALEX carried two microchannel plate detectors (MCP), simultaneously exposed via dichroic, with 1.25 degree field of view (FOV) in ultraviolet (UV) bands centered around 1528 A (Far Ultraviolet or “FUV”) and 2271 A (Near Ultraviolet or “NUV”). The detectors could observe in either direct imaging or slitless spectroscopic (grism) modes. Observations were made over 1500-1800 second periods on the night side of each orbit (“eclipses”). To avoid detector burn-in or local gain sag effects caused by depletion of electrons in the multiplier plate, and to smooth over local irregularities in detector response, the telescope did not stare fixedly at locations on the sky but moved the boresight continuously relative to the target position. Several boresight patterns or “modes” were used over the course of the mission with potential consequenes for the nature of the resulting data. In the most basic “dither” mode, the spacecraft boresight would trace out a tight spiral pattern with a radius of approximately 1 arcminute; this mode was used most often for Deep or Medium Imaging Surveys (DIS, MIS) in which each full eclipse was spent observing a single region of the sky. In the All-sky Imaging Survey (AIS) mode, the spacecraft boresight would jump between 12 positions (“legs”) on the sky for short integrations of ~100 seconds each, with the detector set to a non-observing, low voltange state during the transition between each position, resulting in one independent observation (“visit”) per leg. The “petal pattern” mode, used to spread the flux from brighter targets more widely across the detector, was similar to the AIS mode except that the 8 legs were tightly clustered within the approximate area of a single FOV, and the detector continued to observe in the transition periods.

MCPs are non-integrating photon detectors that can record position and time information individually for every detected event, and the GALEX detectors were capable of recording data with a time resolution of five microseconds, though the vast majority of observations were made in a “compressed” mode at five millisecond resolution. Due primarily to computer storage and processing constraints, the GALEX data was only released and archived as per-observation or multi-observation (coadd) integrated image maps with exposure depths of hundreds to thousands of seconds. The high time resolution data was not released to the community except in a handful of cases by special request [CITE], and very little analysis was performed by the mission team to understand the detector performance or calibration for integrations shorter than about 100 seconds.

Advances in storage and processing capabilities since the beginning of the mission over fifteen years ago now make storage, distribution, and analysis of the photon level data technologically feasible. But by the end of the mission, the GALEX calibration pipeline software (“mission pipeline”) had grown to sufficient complexity and dependence on its operating environment that attempts by multiple parties to get the software to run outside of the aging GALEX network of computers upon which it had been developed were unsuccessful. We have undertaken the gPhoton project, described herein, to migrate key functionality of the mission pipeline into a “standalone” pipeline to process raw spacecraft detector telemetry into calibrated lists of time-tagged photon events with five millisecond resolution, archive ~1.1 trillion such events recorded by GALEX, and create a suite of software tools to create calibrated light curves and images at user-specified spatial and temporal scales in service to the exploration of the data set for short time domain variability. While gPhoton *does* reproduce much of the core functionaity of the mission pipeline in an open soure and simple to use package, it is not intended as either a full migration or faithful port of the mission pipeline. As will be described, archived output files from the mission pipeline have been utilized where deemed expedient, and some deviations from the original calibration and reduction methodology have been made in service to both computational efficiency and the unique properties of photon level data.

**“****Course Sun Point”**

The FUV detector failed in May of 2009. On 4 May 2010, an event referred to by the mission team as “the Course Sun Point anomaly” (or CSP, referring to the safe mode entered by the spacecraft at that time) resulted in image degredation of the NUV detector, manifest as severe streaking in the detector Y-direction, most probably due to a failed capacitor. Although the effect was largely corrected through subsequent calibration and onboard adjustments [CITE], observations taken between 4 May and 23 June 2010 have substantially worse point spread functions.

**CAUSE Phase**

NASA support for the mission ended in February of 2011. At that time, ownership of the spacecraft was transferred to the California Institute of Technology for a mission phase called the “Complete the All-sky UV Survey Extension” (CAUSE), during which operating costs were solicited from individuals or institutions and detector field and source brightness were relaxed, enabling substantial additional observations of bright regions of the sky not observable during the primary mission, and spacecraft slew rate limits were also relaxed, permitting a new high coverage “scan mode” of observation that swept across several degrees of sky in a single integration. Ownership of the CAUSE-phase data resides with the indivudual observation sponsers, and only a small fraction of it has been made available to MAST [citation]. Though the new capabilities of the gPhoton project and its attendant tools may be of particular value in using and interpreting CAUSE phase data — in particular scan mode observations or observations of very bright or dense fields — the current paper describes work only with data up through the end of the NASA supported mission, corresponding to General Release (GR) 7 of the GALEX data to the MAST archive, and only observations collected in the direct imaging mode. Future work may address both scan mode and spectroscopic data.

Data were downlinked from the spacecraft and assembled into monolithic telemetry files (*-tlm*). The *ingest* stage of the pipeline split these into various types of encoded raw detector event and spacecraft state (*-scst*) data which inluded coarse aspect solutions from the onboard star tracker at 1 second resolution as well as spacecraft houskeeping records. The most important class of encoded raw detector data for our purposes, containing nominal scientific observations, were the *-raw6* files. The -raw6 were decoded with a sequence of bitwise manipulationsinto lists of raw detector positions (x and y) with timestamps for all detector events. These raw positions were futher adjusted with “static” (detector-space) calibrations for wiggle, walk, nonlinearity and distortion (described more completely in *Morrissey*). For post-CSP data (after eclipse number 37460), the calibration was slightly modified both because several onboard detector constants changed or were modified and because an additional processing step was added for the “YA” value of the raw position data to correct for detector streaking caused by this anomaly. [Describe the post-CSP corrections better.]

**The gPhoton Toolkit**

The gPhoton project comprises four major software tools and a publicly accessible database at MAST, described in detail below.

**gPhotonPipe**

The gPhotonPipe standalone calibration pipeline implements a subset of features from the full mission pipeline to perform lower (detector) level calibration and aspect correction of detector events. This command line function accepts the raw scientific data file (*-raw6*), the spacecraft state file (*-scst*), and one or more refined aspect solution files (*-asprta*). It returns a “photon list file” in comma seperated value (CSV) format in which each row corresponds to a detector event and records information on raw and calibrated detector positions, projected (de-dithered) sky positions, a number of intermediate values, and a flag column that encodes, among other things, the quality of the aspect solution used in the projection (as communicated from the aspect file) and whether the event falls in a known detector hotspot region. The photon list files produced by gPhotonPipe are analogous but not identical to the extended photon list (*-x*) files occasionally produced by the mission (but not archived).

Note that all of the inputs to the standalone calibration pipeline — the raw6, scst, and asprta files — are products of the mission pipeline that are archived at and available through MAST; by using these archived mission products directly, we avoid the need to recreate either the *ingest* or *aspect correction* stages of the mission pipeline. If an aspect file is *not* supplied, then gPhotonPipe will attempt to query the MAST database of asprta data; this method is probably preferrable for most users, but manually specifying the aspect file will be useful for researchers who wish to further refine, modify, or fix the mission-provided aspect solutions.

In cases where events fall outside of the main detector FOV — as erroroneous data or stim events — or are not aspect-correctable because they fall in a time period that is in a gap between refined aspect solutions — in between eclipses or visits, or simply due to a failure of aspect refinement — the events are not aspect-corrected. The right ascension and declination values for such events are assigned a value of *NULL* in the photon list file. The distinction between this “NULL data” and nominal or “non-NULL data” is important, as it is treated differently in both the database and high level calibration, as will be described.

**gFind**

*gFind* is a search tool that provides a mechanism for the user to query the available in-database coverage depth (if any) of a particular target. Given a target position, gFind returns the estimated (raw) exposure depth of available data over the whole mission, broken down into time ranges corresponding roughly to discrete observations of the target. Rather than using the visit-based bookkeeping of the mission — which distinguished between observation modes and survey type — gFind attempts to use the data itself. A given position on the sky is considered to likely be “observed” if valid data exists in a time range where the position falls within one FOV (or user-defined effective FOV) radius of the spacecraft boresight as given by the mission-provided refined aspect solution. Distinct time ranges are identified based on a user-adjustable parameter that defines the maximum allowable gap between events for data to be considered contiguous (i.e part of the same observation); this parameter defaults to 1 second, the time resolution of the aspect solutions. A separate user-adjustable parameter sets the minimum exposure depth required for a time range considered to be a “valid” exposure at all (below which threshhold it is not reported).

**gAperture**

gAperture extracts and calibrates event level data from the database to produce photometry and light curves given a host of user-specified parameters including target position, extraction aperture, background annulus, integration depth, and time ranges. Rather than performing photometric measurements on pixelized and integrated images, as the mission pipeline did, gAperture conducts cone searches on the sky positions of individual photon events for the purposes of aperture photometry and background subtraction, allowing analysis at the native spatial resolution of the data. Each photon event so selected is then weighted by the detector flat value at the spot on the detector on which it occurred, using the mission-produced flats. All events within a given observation (or “time bin”) are then weighted by the effective exposure time.

**Effective Exposure Time**

The effective exposure time is computed as the raw exposure time (the difference between the observation start and end time) less any amount of time considered “shuttered” and then scaled by the global dead time. Shuttered periods are those periods of significant length — for our purposes, considered to be 0.05 seconds or more — during which no valid observational data is available. These might be periods during which the spacecraft was not actually observing the requested region of sky, but can also include data dropouts or periods during which a valid aspect solution is not available for any number of reasons which might include being near the beginning or end of an observation or a failure of the aspect refinement stage of the mission pipeline.

For aperture photometry, the effective exposure is computed at the requested sky position, and then applied uniformly across all events in both the aperture and background annulus. This approximation is more efficient than calculating the exposure across the whole region and fails only when the annulus or background contains a masked part of the detector (e.g. hotspots as described below) or crosses the edge of the effective FOV.

**Exposure Dead Time**

Micro channel plates are subject to a global exposure “dead time” effect caused by the inability of the detector to process more than one event at a time. The effect therefore scales as a function of total global detector count rates: as count rates go up, the fraction of exposure lost to dead time likewise increases. The effect is linear as a function of global count rate up to approximately [rates in NUV and FUV]. Other MCP instruments have used this global count rate relationship as a means to estimate the dead time correction factor. The GALEX detectors, however, were equipped with four built-in electrical pulsers (“stims”) located off the main detection windown and producing a known rate of events (between all four) of 79 counts per second. The mission pipeline extracted these stim events as a direct proxy for the exposure dead time effect. That is, the ratio of the measured stim count rate to the canonical stim count rate is equivalent to the ratio of the effective exposure time to the commanded exposure time. However, one must also consider the contribution of counting statistics in this estimate. The shortest standard GALEX integrations on single targets were the ~100 second observations in the AIS survey. Assuming *zero* dead time, the 1 sigma error in the stim rate due simply to Gaussian counting statistics would be around 0.9 cps or ~1%. This level of error is negligible as compared to other sources, and, indeed, was not even propagated by the mission pipeline. For shorter integrations, however, like those intended to be produced by gPhoton, the errors in the exposure time produced by estimating dead time from stims become quite meaningful. For a ten second and one second observations, respectively with *zero* dead time, the 1 sigma errors are +/- 2.81 cps (~2.5%) and +/- 8.9 cps (~10.1%). The problem compounds for real observations in which the fractional dead time can range from 10% to 30%, producing 1 sigma errors over a ten second integration of ~3.8% to ~4.3%, respectively.

The solution used by gPhoton is to — as other missions have done — use the linear relationship between global count rate and dead time (for reasonably low global count rates) to produce an empirical exposure time correction as a function of global count rate. GALEX has typical global count rates of 10,000 cps or more, making the 1 sigma error (due to counting statistics) truly negligible even for short integrations. The mission team did produce such an empirical dead time formula early on, but it suffered from a few problems that requried that we revisit the problem and produce our own analysis. For one, while the result was recorded, the actual methodology was not, making it impossible to reproduce. Also, because it was early in the mission, extremely high count rate fields (in the non-linear response regime) of the detector had not been observed and were therefore not included in the analysis. Finally, the original analysis produced only a single model that — because their behavior was deemed so similar — simultaneously covered both detectors.

We plotted global detector count rates against stim count rates. In the denominator, we used exposure times that had been corrected by shutter but not dead time. For analysis, we assumed the foreground “signal” to follow a linear model with two parameters (slope and intercept) and Gaussian counting errors in both x and y. For a background, we also used a linear model with unknown parameters. The mixture model was sampled by MCMC against the data for ~1,000 observations in each band to produce maximum likely model parameters for stim count rate as a function of global count rate in both bands, which could then be converted to a fractional dead time by comparing the stim count rate against the reference rate. Rather than simply trusting that the quoted stim reference rate of 79 cps was true for both bands, we used the maximum likely y-intercept (at global count rate of zero) as the reference rate. This produced different models that differened in each band by about 3%, which is significant for GALEX.

Prior GALEX documentation has suggested that users of the photon-level data assume that the observation-level exposure time is distributed evenly across the whole observaiton [CITE]. This assumption, though, is not true, and sometimes quite significantly so. Not only can shutter be distributed non-linearly over an observation, but the global count rate can change quite substantially even over an observation of a single target because the detector is constantly moving on the sky (causing small changes in the global count rate due to the field brightness) at the same time as the spacecraft is travelling through the shadow of the Earth, creating global brightness shifts due to airglow. If one plots global count rates as a function of time for a typical ~1600 MIS observation, there will be a characteristic “smile” shape; this shape is inverted (a “frown”) in a plot of effective exposure as a function of time (because higher global count rates crate more detector deadtime, resulting in less effective exposure time). Researchers interested in the variability of targets within single observations should recompute the exposure time for every sub-integration (“bin”) in order to account for this bias, and gAperture does so.

Within the misison pipeline, the detector flat was projected into the sky according to the spacecraft aspect solution and then integrated and interpolated into a high resolution relative response (rrhr) map which could then be applied to the integrated count maps to correct for detector response and effective exposure (as a function of sky position), as well as any static or dynamic masking (of hotspots, ghosts, or edge reflections). The gPhoton pipeline, instead, applies the detector flat at the detector level by *weighting* each individual photon event by the value of the flat at the detector location upon which it fell.

(1) The first method is a simple annulus estimation method where the total surface flux within an annulus surrounding the extraction aperture as scaled to the area of the aperture and subtracted. This method can be bias by the presence of relatively bright sources either within or near the annulus, although the bias can sometimes be mitigated by careful selection of annulus parameters (and assuming that the field is not crowded). The direct annulus method has the benefit of correcting for changing local background over an observation due to zodiacal light or airglow. It produces good agreement with MCAT values *on average* for integrations of AIS depth or longer, but at larger dispersion (particularly in the direction of higher magnitudes where stars in or near the annulus have caused an erroneously high background estimate).

(2) The second method is estimate the local sky background based upon backgrounds of nearby sources as reported in the official GALEX source catalog (MCAT). In this method, the *coadd* MCAT is searched for sources within a small region of the target position. The arithmetical mean of reported flux of all returned MCAT entries is then used as the estimated background at the target location. This method results in gAperture fluxes that are in good agreement with the MCAT on average and with good dispersion, but does not correct for either inter- or intra-visit variations in the sky background.

We suggest that the first method be used for studies of intravisit variability. The second method may be useful for researchers wishing to compare gPhoton results to previous work that used data from the GALEX catalogs. Note that the first method could also produce false variability if the source(s) in the aperture are not changing, but one or more sources in or near the annulus *are*; when this is a possibility, we suggest that researchers check the MAST catalogs as well as a gMap produced image of the targeted region for nearby sources, and run an independent photometric analysis for those.

We also tested two variations of the annulus method that might mitigate the presence of stars in or near the background annulus, both of which were abandoned for not producing poor agreement with catalog fluxes as well as being highly sensitive to somewhat arbitrary input parameters. The first method — which we called “swiss cheese” — mimicked that applied by Morissey in the analysis of the GALEX standard star LDS749B [CITE]: nearby bright stars (as reported in the MCAT) were masked (out to several sigma, assuming Gaussian profiles), with those data and sky regions excluded from subsequent calculation. The second method was a direct port of the “sigma clipping” algorithm used by the mission pipeline [and described in Ted’s white paper [CITE]], which iteratively excluded and interpolated over relatively bright regions, using full Poissonian statistics to account for low diffus UV background.

**gMap**

*gMap* creates integrated images or movies for targeted regions of sky and specific time ranges up to and including full depth coadds. Users can request either “count” images — which have not been corrected for exposure time or response, and can be used for astrometry or diagnostics — or “intensity” images — which are fully calibrated and suitable for photometric analysis. The images produced by gMap are analogous to the imaging data products of the same name that were produced by the mission pipeline and archived at MAST, but with additional flexibility in image properties via user-tweakable parameters (e.g. dimensions, depth, edge trimming). If given a sequence of time ranges or an integration depth (or “bin size”), gMap will also produce either count or intensity *movies*, which were never available as full resolution output of the mission pipeline (although sequences of fractional depth images were sometimes created by special request). All image data are written in the Flexible Image Transport System (FITS) format with headers populated using the World Coordinate System (WCS) standard.

As with relative response correction in gAperture, rather than generating a relative response *map*, the individual events are simply weighted by the flat value assigned to the detector regions on which they fell. [Exposure time correction is not implemented yet!]

**Photon Database**

To maximize flexibility of the high time resolution data, photon list files produced by running gPhotonPipe on the full corpus of GR6/7 were ingested into databases at MAST, used by gFind, gAperture, and gMap. For performance optimization purposes, the event level data is spread over multiple databases. The non-NULL (aspect corrected) data are spread across ten different databases based on declination, with bounding ranges selected to approximately balance the number of rows per database. Each database is then further divided into “zones” or strips of 30” declination. We make use of the fast zone matching algorithm described in [Gray 2006] for loading and querying the database. The distribution of the ten databases on the sky is shown in [Figure] along with the specific declination ranges of each. The assignment to a database or partition is strict, leaving open the possibility that data from a single observation might be spread across two tables. The majority of normal queries access only a single database, but queries spanning database are handled on the server side, transparent to both the software and end users. NULL data which were not aspect corrected reside in a single database.

[What are the indexes of the databases?]

Sky positions are always reported as a two element vector where the first value reports right ascension and the second value reports declination, both in (J2000?) decimal degrees. Time ranges (or “bins”) are defined as two element vectors where the first element is the start time and the second element is the end time. The gPhoton project uses the concept of “GALEX time” throughout, defined in seconds as UNIX Time - 315964800. Search ranges (in both space and time) are generally taken to be inclusive of the lower value and exclusive of the higher value in order to eliminate duplicate counting of data at boundaries.

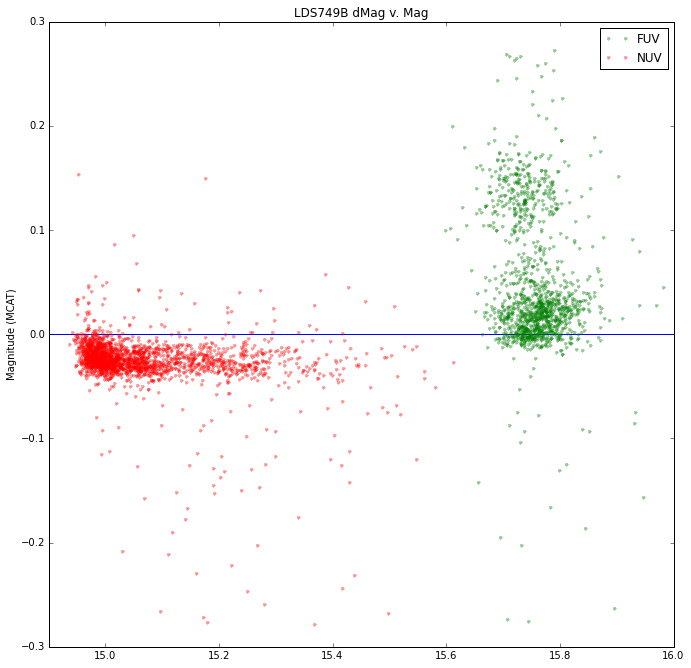
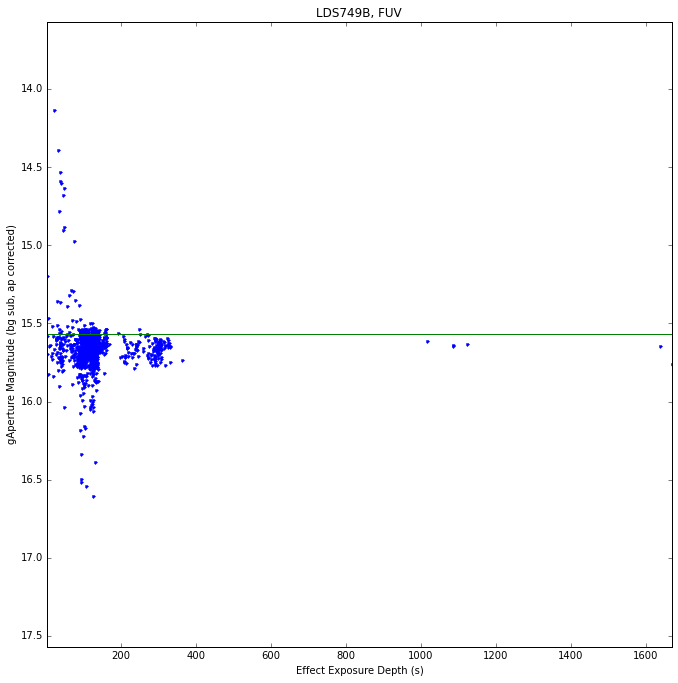
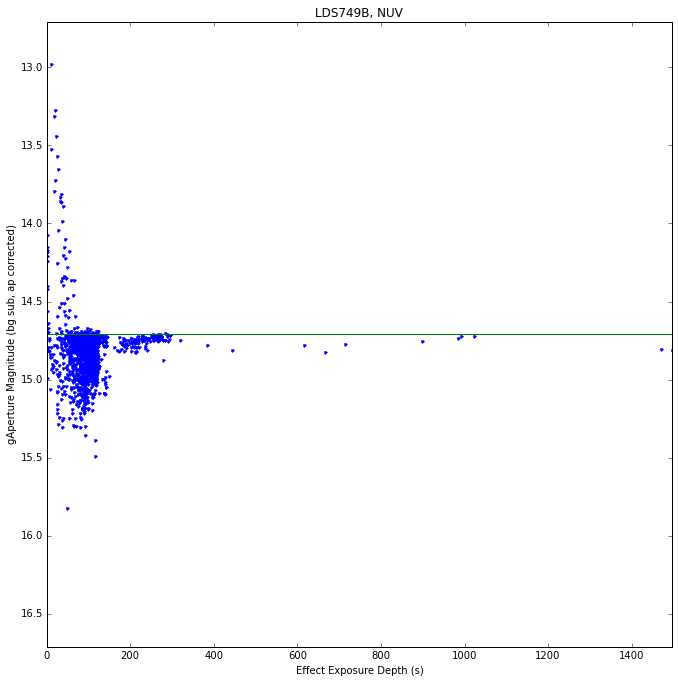
By default, the database tools (gFind, gAperture, and gMap) define an *effective* detector FOV of 1.1 degrees across. This is done in order to conservatively trim data observed on the edges of the GALEX MCPs, which suffer from poorly behaved and understood issues of sensitivity and spatial distortion. Some such data may be useful to cautious and knowledgable investigators, though, so the effective detector size is adjustable from the command line. This conservative trimming does *not* eliminate problems caused by the targeted region of interest (including the photometric aperture, background annulus, or image footprint) extending past the edge of the effective detector; photometry should only be trusted for sources that are clear of both the *physical* and *effective* FOV boundaries.

As explained above, the gPhoton calibration and data reduction is not identical to that of the mission pipeline in all respects. As such, we should not expect perfect agreement.

**Absolute Photometric Precision**

**LDS749B**

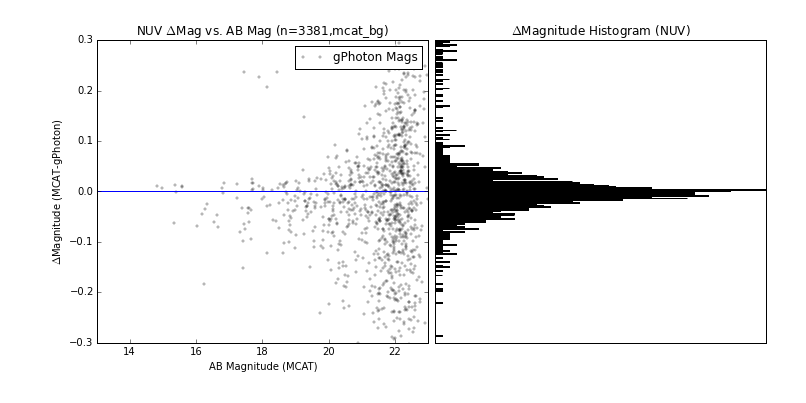
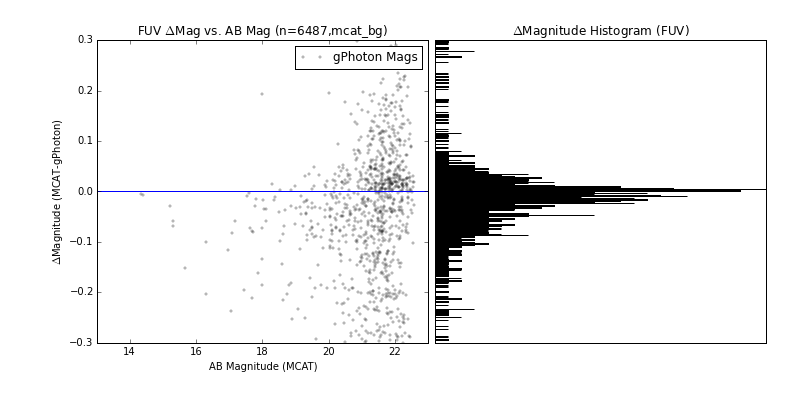
As described in [Morrissey], the calibration standard star of choice for GALEX was the white dwarf LDS749B. A total of [amount] of exposure depth in observations were made of this source. We use gAperture to compute magnitudes for all of these observations in both bands and compare them against the standard magnitudes used by GALEX. We have good agreement in general. And then compare again with only long exposures near the center of the detector.



**Other White Dwarfs**

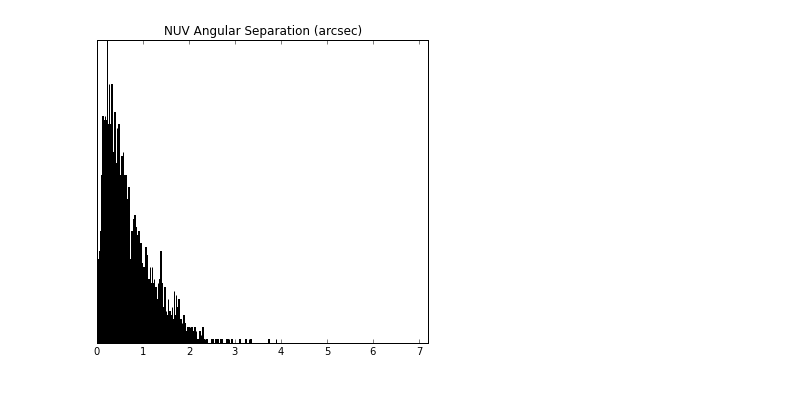
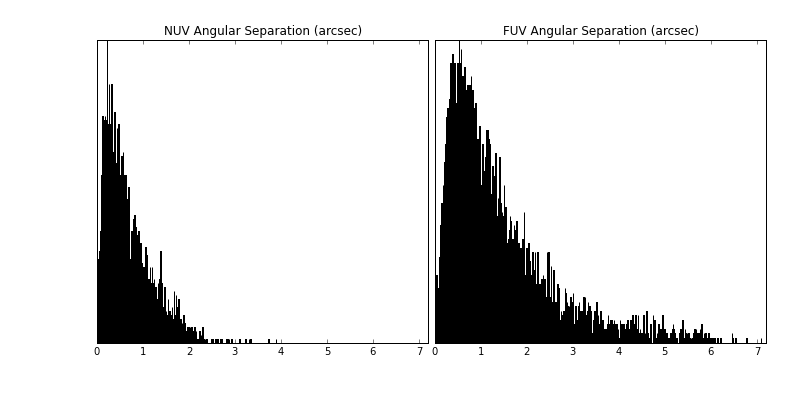
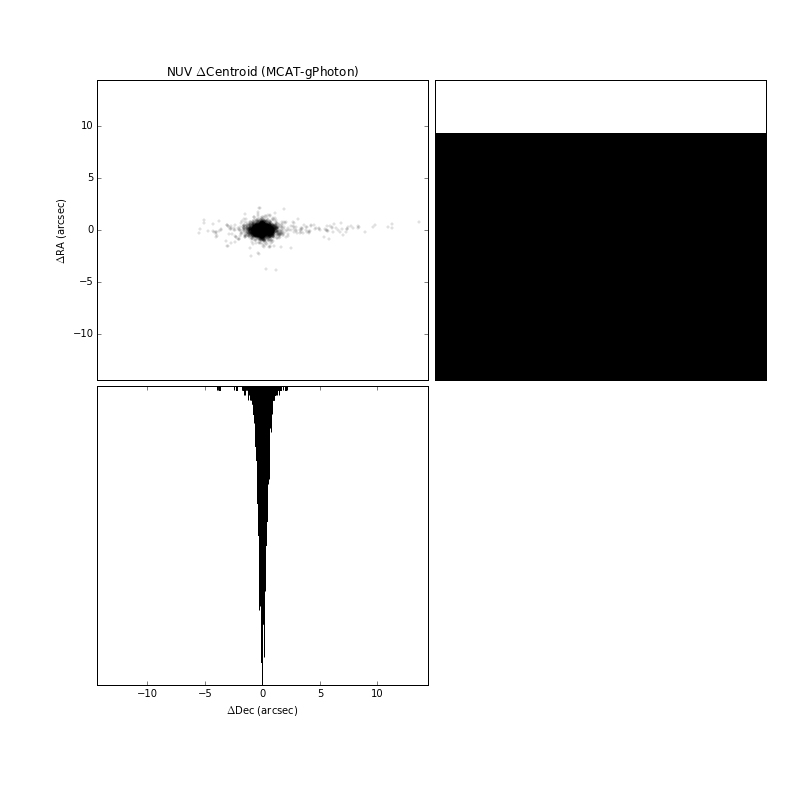
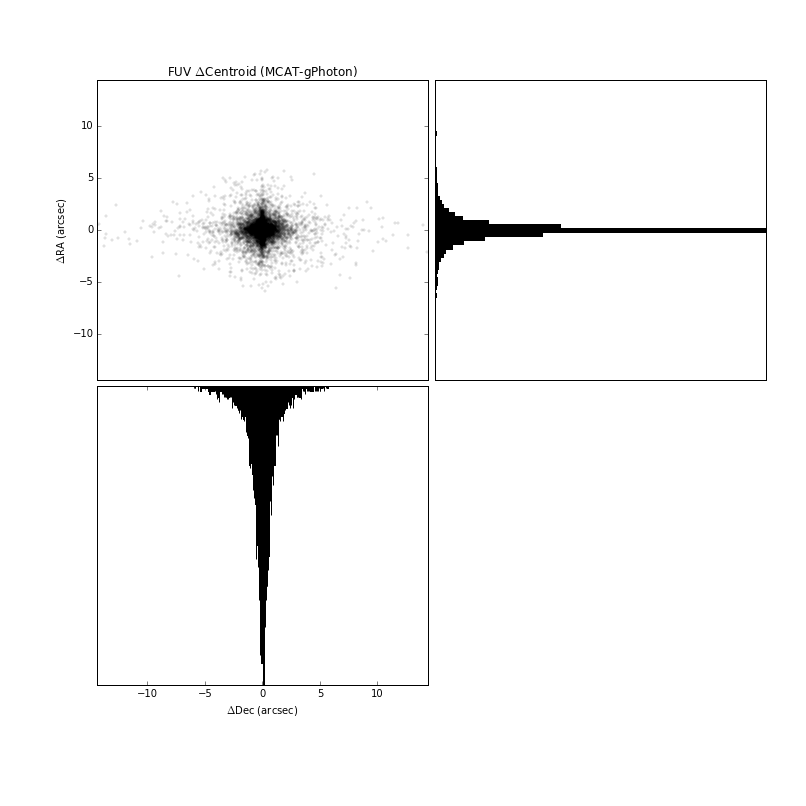
**Relative Photometric Precision**

We plot the difference in gAperture magnitude and MCAT magnitude against the MCAT magnitude for [some number of sources]. All gAperture measurements were done at source positions and over time ranges as reported in the MCAT, with an aperture radius of 0.1 arcminutes (equivalent to APER4 in the MCAT). To ensure sufficient signal to noise, only observations with depths of greater than 100 seconds were used. To prevent any particularly deep fields or exposures from dominating the result, only fields with a total (estimated raw) exposure time of less than 5000 seconds were used.



**Astrometric Accuracy**

We compare the mean photon position within the aperture to the MCAT source position. One source of uncertainty here is that the CoB was calculated by the mission pipeline on a pixel-based map, whereas gPhoton used the photons directly. Another possible source of error is that you should really fit a 2D Gaussian bounded by an off-center circle to find the CoB (and maybe the analysis should be redone like this).



**M Dwarf Flares**

**Active Galactic Nuclei**

**Cepheid Variables**

**Eclipsing Binaries**

**Conclusion**

The calibration and reduction of data imprints, *a priori*, the assumptions, limitations, and biases of the data preparers onto the final (higher level) products. Even when the preparation of the higher level data is well documented and understood by future researchers, the priorities or interests of those end users may not be the same as the data creators or archivists. In such cases, of course, their recourse would be to go back to the raw or minimally reduced data and create a modified or wholly new procedure for reducing the data. This can be onerous, time consuming, or impossible depending on type of data, quality of the documentation, and availability of the original instrument team to answer inevitable questions. The burgeoning practice of “reproducible research” seeks to address this to some extent by making the *actual methods* (as opposed to mere descriptions of the methods) available to researchers.

The gPhoton project is a trial into a new paradigm for data archiving where the correlation between raw and minimally reduced and calibrated or high level data is explicitly laid bare, with the machinery put in place *and archived* for future researchers to make modifications to the reduction methodology more-or-less on the fly and with minimal overhead or barriers to entry.