Progress Report

Ian Evans
On behalf of the Chandra Source Catalog Project Team

Chandra Users' Committee Meeting
October 23, 2014

Summary of Progress Since Last CUC Meeting and Future Plan

- Completed algorithm development and implementation through source detection and evaluation (through "*stacker*" pipeline)
- Established simulation pipeline and evaluated results
- Purchased and deployed new hardware (2013 CUC recommendation)
- Recently completed major pre-production test run
  - 150 observation stacks
- Assembled and trained a volunteer science team (mostly outside CSC team) to review pipeline output
  - Appreciate support of CXC director in doing this
- Propose to provide early release list of detections (based on 2013 CUC recommendation)
  - Include *at a minimum* position + errors, likelihood, aperture photometry proxy
- Plan going forward
  - Complete science test and incorporate feedback
  - Complete current "open" tasks
  - Start production — by end of 2014 (time to complete: several months)
  - Provide initial list of detections (FITS format) to community — 2nd quarter 2015
  - Run phase 2–3 pipelines — 2nd half of 2015 (time to complete; 1–2 months)
  - Perform final (human review) QA — late 2015 (time to complete: ~ 1 month)
  - Release full CSC version 2.0, all databases and data products

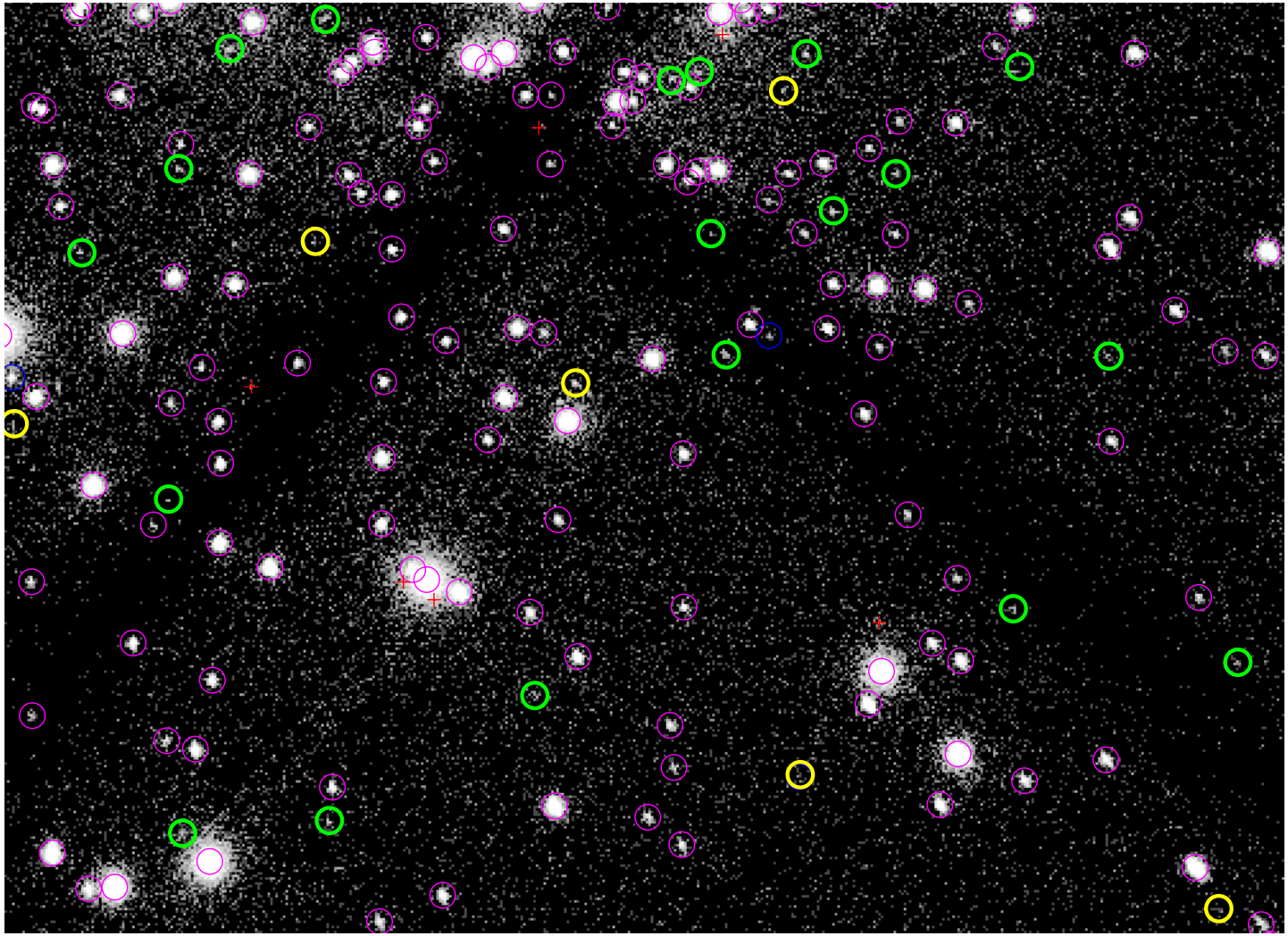
Science Highlights Since Last CUC Meeting

- Continued evaluating data from numerous pipeline test runs and simulations
  - Identified significant background-related issues/investigated possible solutions
    - Adaptively smoothing the background addressed the core problem
  - Evaluated performance of MLE point source fits (source position and amplitude)
    - Appears to meet production requirements — ready for final science tests
  - Evaluating ability to detect and classify "compact" (~1″–30″) non-point sources
    - Have possible solution; detection efficiency is low
- Updated the simulation pipeline/constructing simulations as needed
  - Injects sources with known properties into the data
  - Simulations are useful for diagnosing issues and are required to set key production pipeline parameters (e.g., likelihood thresholds)
- Defined method to derive position error estimates using MCMC algorithm
  - Still evaluating test results, but appears to produce good results for point sources
  - Working on optimizing extended source fits
  - Not a lien on production start — can be run as a "follow-on" pipeline after *stacker* pipeline *if required*
- Continuing development of specifications for post-detection phase pipelines
  - Completed aperture photometry specifications
  - Developed draft specifications for spectral fits
- Aperture photometry paper (Primini and Kashyap) accepted for publication by ApJ (Nov. 10 issue)

Software Highlights Since Last CUC Meeting

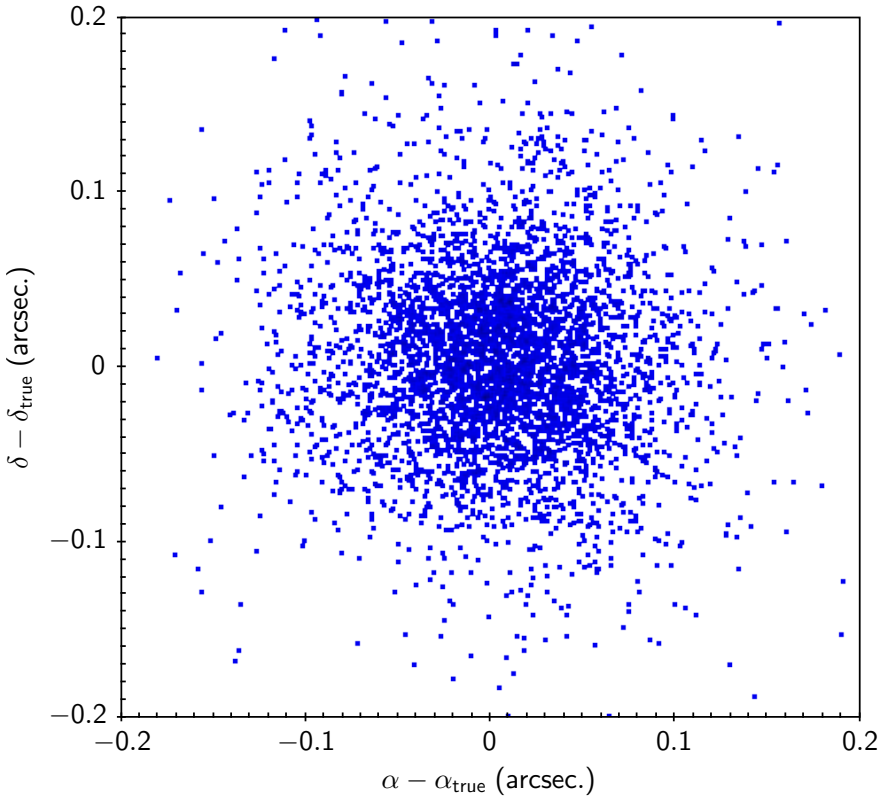
- Continued developing/refining pipelines through *stacker* pipeline
  - Corrected stacked-observation astrometry to exposure weighted aspect solution
  - Developed new adaptive smoothing algorithm for background maps
    - Ensures smoothing scale is sufficiently long on-axis (where PSF is small)
  - Determined position error ellipses from per-detection fit statistic surfaces using pyBLoCXS MCMC algorithm in *Sherpa*
  - Extensively investigated and refined algorithms based on science testing
  - Identified and resolved bugs/unexpected issues as necessary
- Infrastructure updates
  - Evaluated, received, and tested new processing cluster
    - 8 Dell R910 servers (total 320 cores, 4 TB memory, 43.2 TB local disk)
  - Established simulation thread
  - Rationalized archival file data products, contents, and file headers
  - Integrated quality assurance pipelines and manual quality assurance GUI
  - Migrated CSC archive infrastructure to 64 bit Linux
  - Added automated population of web pages to assist review of test run results
- Science testing
  - Continued to support extensive science testing through `stacker` pipeline
  - Continued managing pipeline test runs and simulations
  - Continued performing preliminary assessments of test runs prior to science review

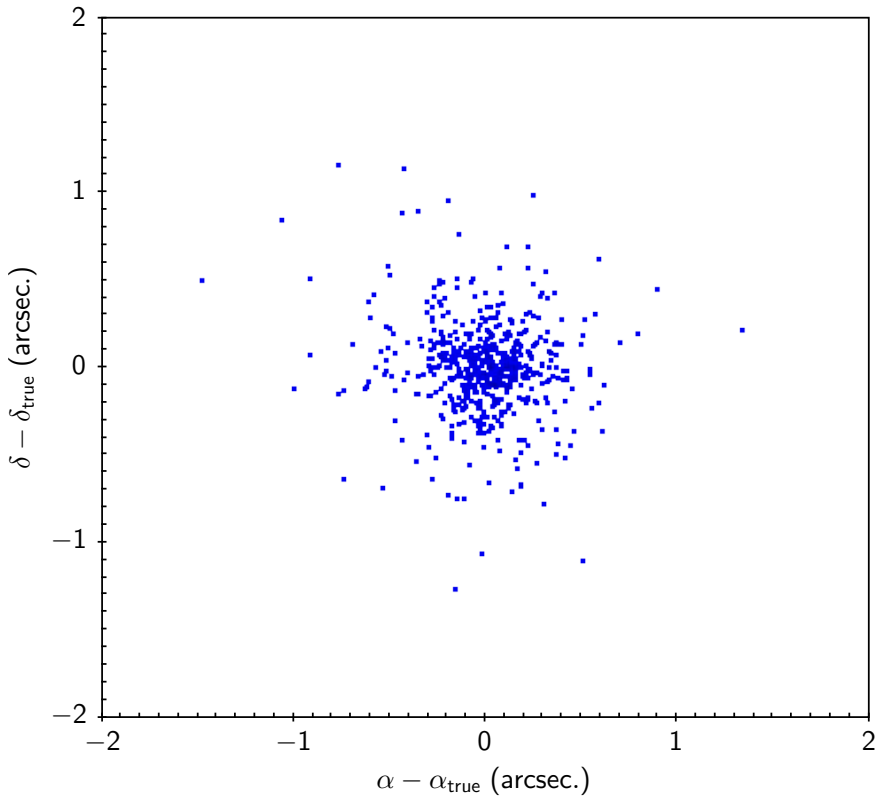
Release 2: Source Detection



The image compares source detections between rel. 1 and the expected performance for rel. 2, in a region of the Orion nebula (M42). The image is ~4′ across. Bright sources detected in both releases are indicated by magenta (rel. 2 likelihood, $\mathcal{L} \geq 15$) or blue ($9 \leq \mathcal{L} < 15$) circles. Green circles indicate rel. 2 sources with $\mathcal{L} \geq 15$ not found in rel. 1, while yellow circles indicate rel. 2 sources with $9 \leq \mathcal{L} < 15$ not found in rel. 1. Red crosses mark detections from rel. 1 that were either not detected in rel. 2 or that have $\mathcal{L} < 9$ (and are therefore flagged as false).

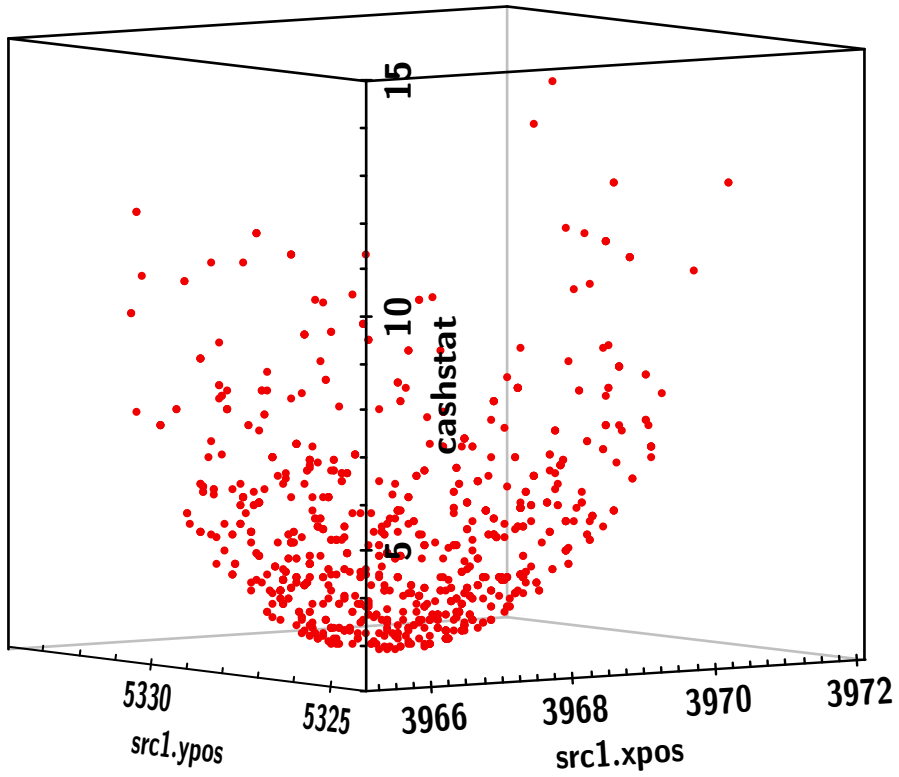
Release 2: Source Positions





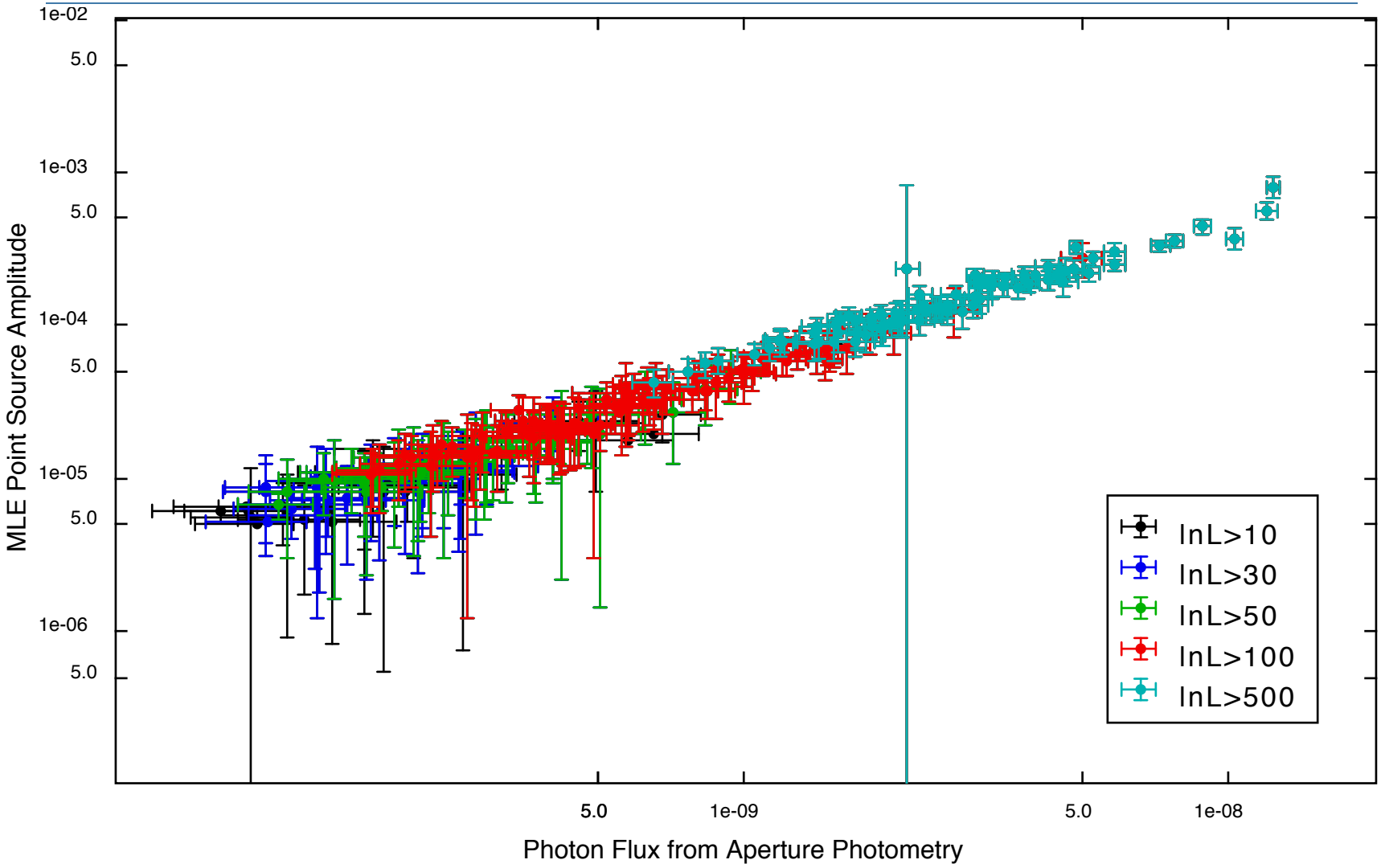
For simulated point sources with more than 50 counts and $\theta < 3'$ (left), the mean error in the fitted position is ~0.005″ in both α and δ, with an rms of ~0.05″. For 1″ extended sources (right), the mean error is comparable, but the rms is ~0.2″. *Note that the scale for the plot on the right is a factor of 10 larger than that for the plot on the left.*

Release 2: Source Positions (cont.)



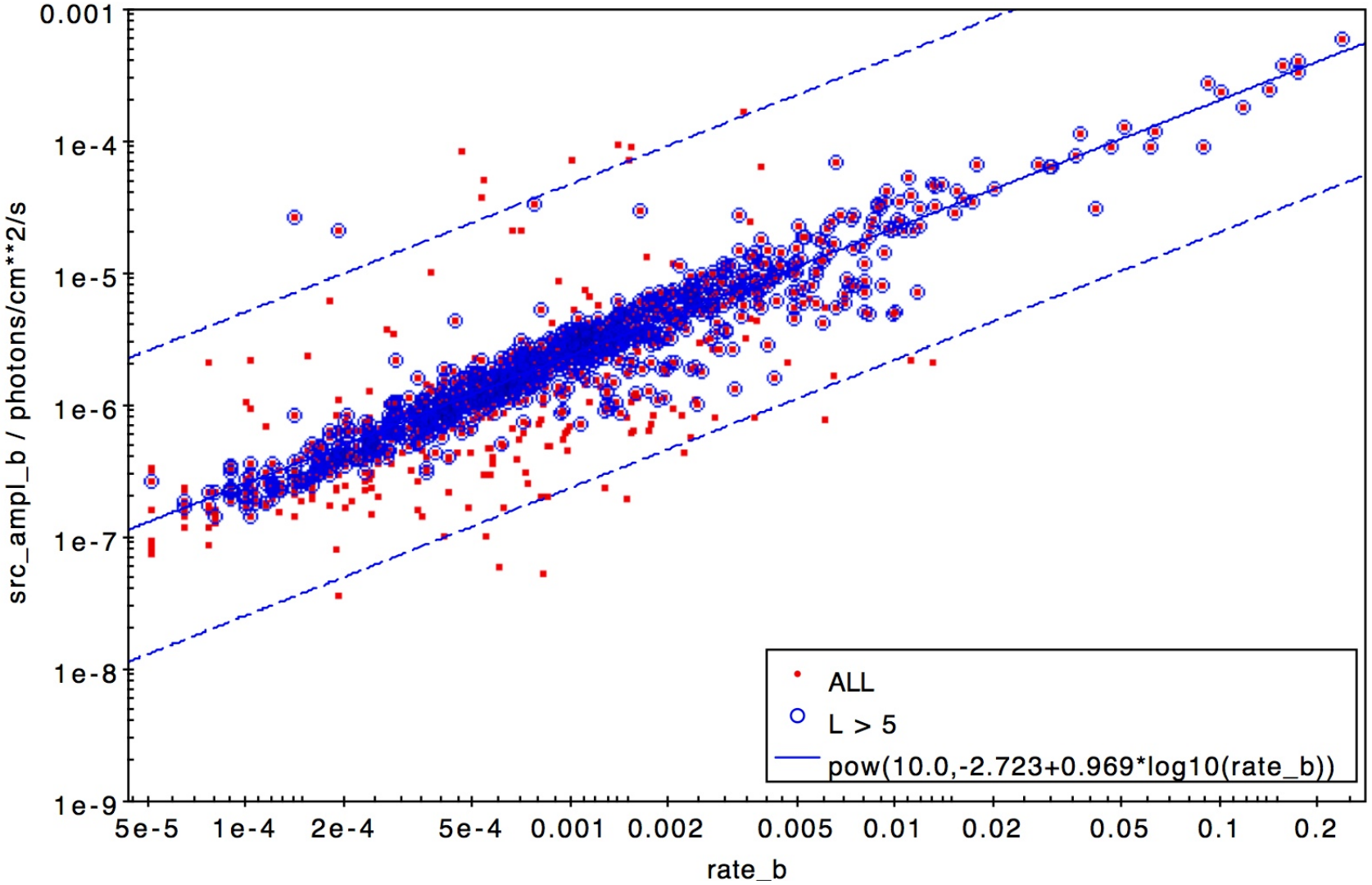
Source position errors in release 2 will be reported as 95% error ellipses computed using Sherpa's get_draws pyBLoCXS MCMC based algorithm. The actual draws for each source and energy band will be recorded in archived data products. The example above shows the accepted draws for a point source located ~10′ off-axis with ~50 net counts in the broad energy band in a single observation.

Release 2: Aperture Photometry



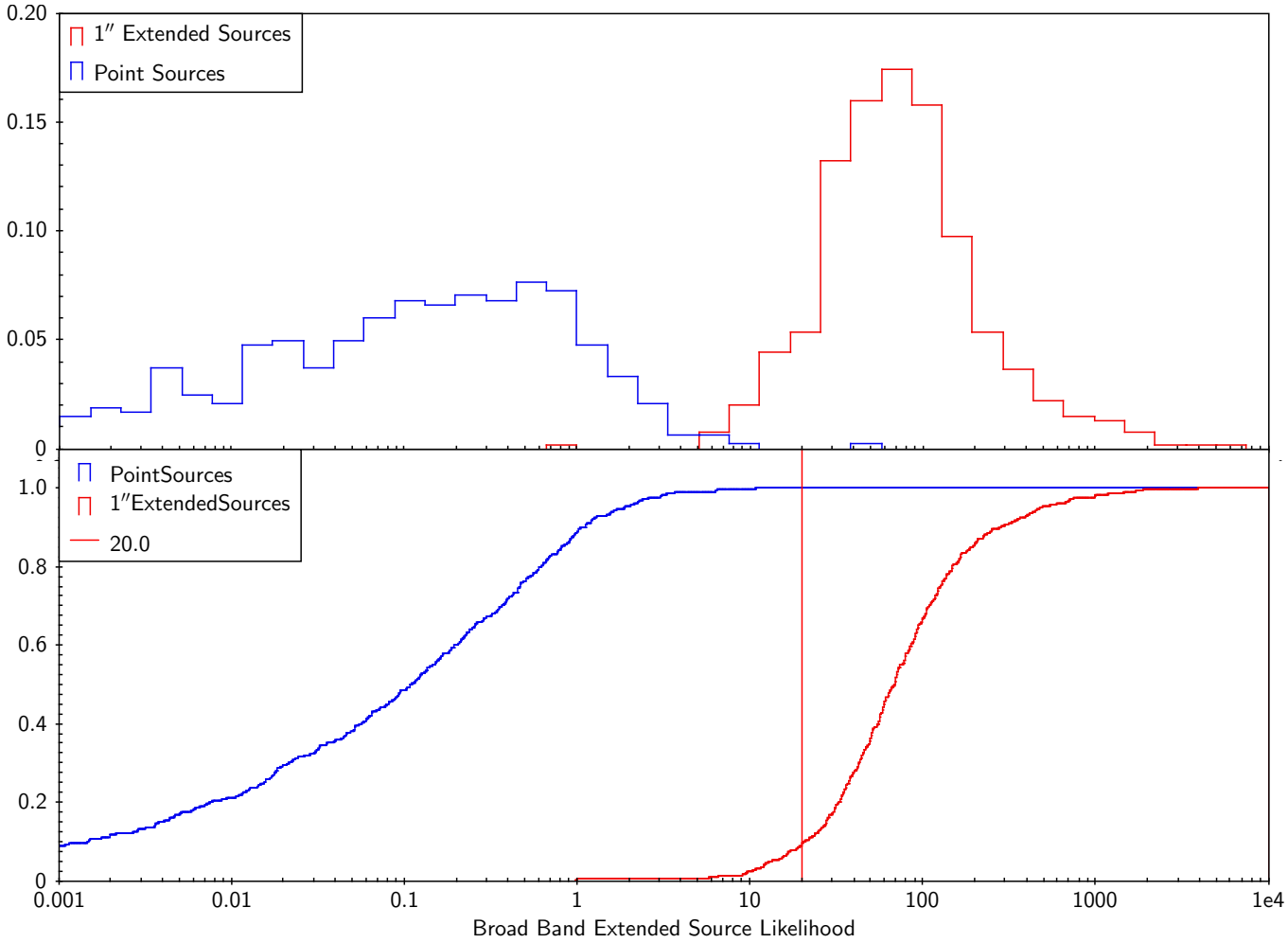
Correlation between broad-band amplitude of MLE point source fit and photon flux computed using full aperture photometry formalism for a sample of point sources. The correlation indicates that the MLE amplitude may be used as a proxy for aperture photometry of point sources in the short term.

Release 2: Aperture Photometry (cont.)



Correlation between broad-band amplitude of MLE point source fit and event rate in the source aperture for a sample of point sources extends the correlation to the faintest detectable sources.

Release 2: Extended Source Identification



For sources with more than 50 counts, with $\theta < 3'$, the distribution of extent likelihood (i.e., the likelihood that an extended source fit is preferred over a point source fit) is well-separated for actual point and 1″ extended sources (top). For this range of θ and counts, an extent likelihood threshold of 20 yields a Type 1 error (mistaking a point source for an extended source) of << 1%, with a Type 2 error (missing a true extended source) of only a few percent (bottom).

Release 2: Schedule Update

Schedule Status

- Schedule has slipped by ~6 months since the last CUC meeting
  - Resource competition continues largely unchanged or is marginally worse
  - Loss of key science/software staff has significantly impacted the schedule, particularly in areas of background determination and source fitting
  - Unexpected issues with background determination, source detection and fitting took longer than anticipated to investigate and resolve due to loss of key staff
  - Science algorithms through *stacker* pipeline now mostly in place
  - Catalog processing performance issues have been resolved with new hardware

Example Schedule Impacts

- Acquired higher performance hardware following CUC recommendation
  - Lack of obvious replacement candidate required careful evaluation of performance improvement to justify high cost
  - Testing effort, purchase, and installation had a ~3 month impact on parts of the schedule
  - Proven very worthwhile since testing and operations present a heavy load
- Unexpected network file server crashes
  - To minimize risk to CXC operational threads, discontinued using catalog test system until a software update could be developed by the vendor and installed
  - This ultimately took ~2 months

Release 2: Production Schedule

Operations plan

- ***Phase 1:*** **run through *stacker* pipeline**
  - Creates all per-stack source detections
  - Steps to phase 1 production
    - CAT 4.2.6 release
    - Run "mini-test" with good results
    - Run test of 150 stacks
    - Complete science review I and feedback
    - Set initial production parameters based on simulations
    - CAT 4.2.7 release
    - Run test of 400 stacks (performance test)
    - Complete science review II and feedback
    - Refine production parameters based on simulations and run of 400
    - CAT 4.2.8 release
    - Complete reprocessing of 2012 Sep–2014 Jun aspect data to fix ~0.4″ offset (scheduled for November)
  - Production run through *stacker*, including auto and manual QA
  - Run *a posteriori* position error determination pipeline *if required*
  - Start production ~end 2014 (time to complete: several months)
  - *Propose to make the per-stack source detections available to the user community in FITS binary table format*

Release 2: Production Schedule (cont.)

Operations plan (cont.)

- ***Phase 2:*** **run *master_match* pipeline**
  - Merges source detections from overlapping stacks and *assigns source names*
    - Pipeline to be developed
    - Source matching based on existing code from release 1
    - New functionality to compute photometric upper limits
      - » Based on existing code from the release 2 pre-*stacker*, but spec not yet available
  - Start production 2nd half 2015 (time to complete: ~ week)

Release 2: Production Schedule (cont.)

Operations plan (cont.)

- ***Phase 3a:*** **run *source* pipeline**
  - Computes source properties
    - Pipeline to be developed
    - Mostly based on release 1 code updated with enhanced functionality
      - » Aperture photometry — significant update to release 1 algorithms
        spec/tool development mostly complete
      - » Bayesian blocks observation partitioning — new functionality
        science prototype/draft spec exists
      - » Hardness ratios — revised to use PDFs from aperture photometry
        straightforward, but spec not yet available
      - » Spectral model fluxes and spectral fits — minor update to release 1
        draft spec exists (needs review)
      - » Temporal variability analysis — intra-observation minor update to release 1; inter-observation revised to use PDFs from aperture photometry
        straightforward, but spec not yet available
  - Start production 2nd half 2015 (time to complete: ~ month)

Release 2: Production Schedule (cont.)

Operations plan (cont.)

- ***Phase 3b:*** **run *master* pipeline**
  - Merges properties for each unique source on the sky
    - Pipeline to be developed
    - Similar to merge processing from release 1 *master* pipeline
    - Mostly uses components from the *source* pipeline with groups redefined, but spec not yet available
  - Start production late 2015 (time to complete: ~ week)

- ***Phase 4:*** **perform final (human review) QA**
  - All pipelines are followed by appropriate automated QA processing
    - If automated QA detects a pipeline or data anomaly that requires human review, manual QA is invoked
  - Manual QA can be forced after several critical pipelines if needed
  - Almost all manual QA takes place during the actual pipeline production runs
  - A set of "sanity check" final QA analyses are performed at the end of each production phase
  - The *phase 4* final review approves release 2.0 of the catalog
  - Start production late 2015 (time to complete: ~ month)

Initial List of Detections

- 2013 CUC recommendation
  - Consider staged releases, where certain parts of the sky might be released ahead of others
- We propose to make available an *initial list of per-stack source detections* as soon as practicable
  - As a minimum, the list will include detection position and error, likelihood, and point source fit amplitude (proxy for aperture photometry) from MLE
  - Additional categories of properties that *could* be included are
    - Fit statistics and associated fit metrics
    - Parameterized fits to "compact" (extended by a few arcsec) detections
    - *wavdetect* and *mkvtbkg* candidate point/compact detections input to MLE
    - Extended source convex-hull polygons
  - However, these properties are primarily intended for debugging and QA
    - *With the exception of source position and amplitude for isolated point sources, they may not correlate well with actual physical source properties, and many of the candidate detections will be false*
- Input on whether any of these additional properties from these categories should be included in the initial list of per-stack detections is welcome
  - We would minimally validate the additional properties prior to release

Release 2: Software Releases

CSC 2.0 Catalog Production-Related Releases

Release	Date	Content
CAT 4.1.2	04 Dec	Additions to PLs (tune algorithm params, crater detection, add ultrasoft band); 2nd tier MLE updates; data product refinements
CAT 4.1.3	31 Jan	Reviewed and updated file headers; number of tuning upgrades to PL parameters; set w band PSF scale param; increased bkg source exclude param by 10%; added source position error ellipse to MLE output
CAT 4.1.4	06 Feb	Migrate archive CAT builds to 64 bit; add new archive server capability to handle CSC Rel. 2 data; few CSCview bugfixes
CAT 4.1.5	15 Feb	Initial integration of QA; MLE bkg changes
CAT 4.1.6	04 Apr	Fixed the issue with bright positions in dim bkgds; integrated crates updates and bug fixes into PL; increased speed when writing mrgsrc3 files with many rows
CAT 4.1.7	12 Apr	Migrate CIAO 4.6 code base and OTS to CSC
CAT 4.2	06 Aug	QA thru *stacker*; reprocessing support; interleave-mode; performance tuning; upgrade compiler to gcc 4.8
CAT 4.2.1	29 Jul	CAT integ issues (QA, repro); several upgrades from science PLs

CSC 2.0 Catalog Production-Related Releases (cont.)

Release	Date	Content
CAT 4.2.2	12 Sep	New adaptive smoothing for backgrounds; fix to MLE for off center points; ability to add user exclusion regions; update to repro script to no longer create sexp3; MLE fix from regression in positions; MLE get_draws added
CAT 4.2.3	29 Sep	MLE valid flag and NULL case handling; lower risk updates/ bugfixes from PL
CAT 4.2.4	30 Sep	Upgrade CAT archive servers to linux
CAT 4.2.5	01 Oct	Science test *must-haves*; lower priorities if they fit schedule
CAT 4.2.6	17 Oct	Address SAOTrace call issue; 0 counts on CCD (corner case)
CAT 4.2.7	~15 Nov CF	Lower priority RFEs/bugfixes; feedback from science test I
CAT 4.2.8	~01 Dec CF	Remaining science priorities; feedback from science test II **Production run through *stacker* PL**
CAT 4.3	2nd half 2015	*master_match* and *source* PLs; limiting sensitivity population **Production run *master_match* and *source* PLs**
CAT 5.0	Late 2015	*master* PL; populate databases **Production run *master* PL, final QA review, and release**

Current Catalog Release: Statistics

Summary

- Catalog version: 1.1; Released: 2010 Aug 10
  - 106,586 master sources
  - 158,071 source detections
  - 5,110 observations with at least one detected source
- Subset of master source properties are available via HEASARC Browse/ Xamin, NED, and Vizier services
  - Usage statistics reported below do not include accesses via these services

Usage Statistics

Release 1.1	Current Reporting Period 2014 Apr 01 – 2014 Sep 31		Previous Reporting Period 2013 Oct 01 – 2014 Mar 31	
	Number	**% Non-CfA**	**Number**	**% Non-CfA**
CSCview catalog browser initializations	163 /month	86%	158 /month	82%
CSCview catalog browser properties searches	140 /month	87%	180 /month	82%
Command-line (CLI) searches	3455 /month*	81%	687 /month	20%
VO cone searches	2986 /month**	~100%	5209 /month	99%
CSC Sky in Google Earth	575 visits/month		520 visits/month	

* Excludes 198K searches (~ all non-CfA) from 2014 May; ** Excludes 99K searches (~ all non-CfA) from 2014 June

Current Catalog Release: Support

CSCview

- Released two updated versions of CSCview including several bug fixes and enhancements
  - Third update currently being tested

3rd Party Interfaces to CSC Release 1.1

- Master-source table "basic summary properties" (only) have been available via NED since 2011
- CDS has recently made these data available through Vizier
- HEASARC has made the master-source table available through their browse/Xamin interface
  - HEASARC does not provide access to the source-by-observation table or any catalog FITS file-based data products
  - HEASARC has changed the names of many of the columns, as well as the name and designation of the broad energy band
    - They are no longer consistent with either the full CSC release at CXC or with the published CSC paper (however the changes are documented)