ZTF Science Data System Status

Frank Masci & the IPAC-Caltech ZTF Team

ZTF Working Group Meeting, March 2018

Outline

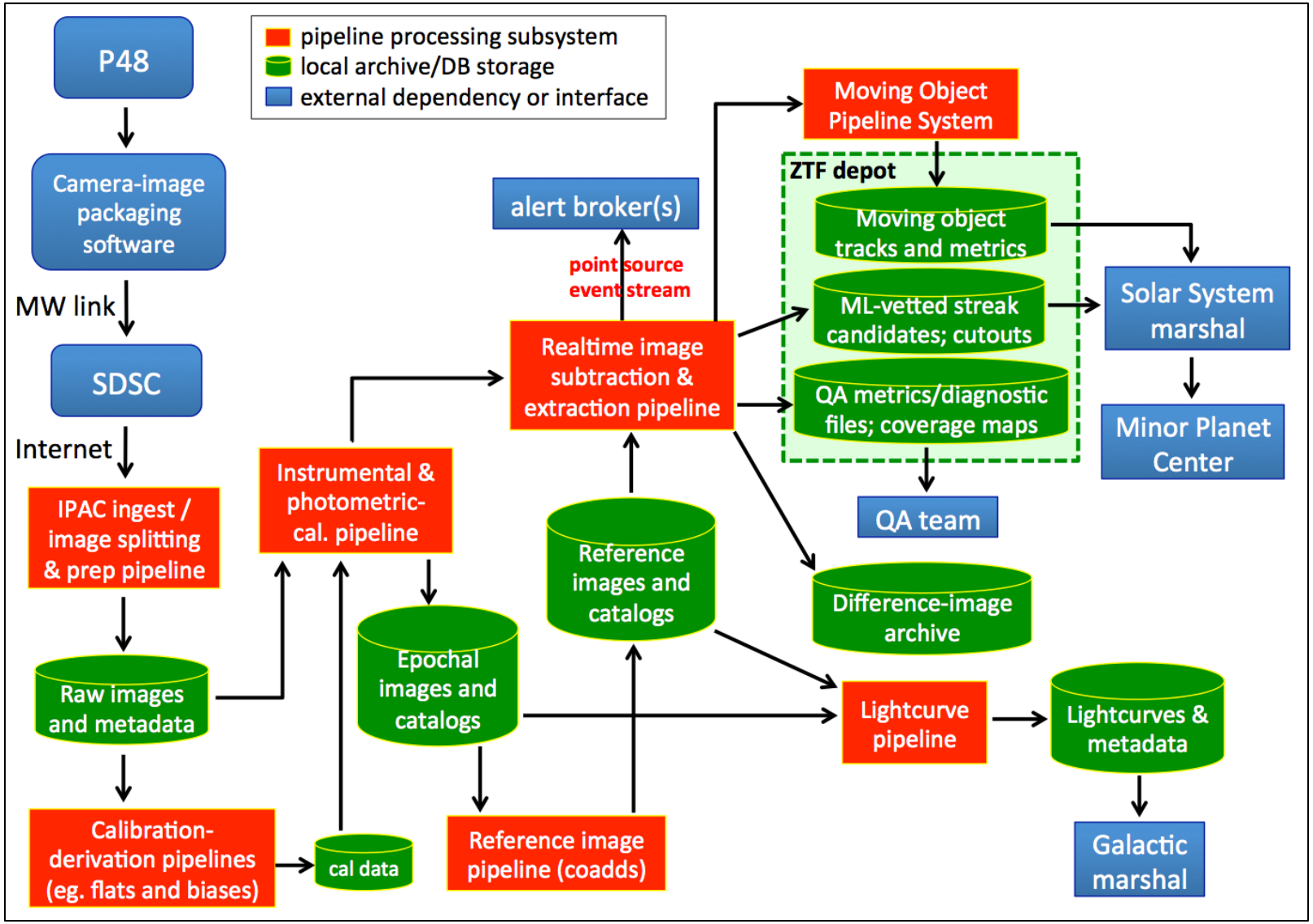
- Overall status of pipelines, services, and deliverables
- Sky coverage and summary statistics on data holdings
- Pipeline performance
- On-sky performance (astrometry and photometry)
- To-do list and wish list

Background Material

- Presentation from August 2017 Team Meeting:
  http://web.ipac.caltech.edu/staff/fmasci/home/miscscience/masci_pasadena_08.01.17.pdf

- Primary document: design, deliverables, product usage, data access, and on-sky performance:
  http://noir.caltech.edu/twiki_ptf/pub/ZTF/ZTFcommissioningaccess/ztf_pipelines_deliverables.pdf

Overall processing & data flow



Baseline deliverables & data access portals

1. **Instrumentally calibrated, <u>readout-quadrant based</u> epochal image products:**
   - images with photometric zero-points from PSF-fit photometry; bit-mask images; PSF templates
   - two source catalogs per image: PSF-fitting and aperture photometry
   - difference images, accompanying matching PSFs and QA metadata
   - ➢ archive via GUI or API at IPAC; can interface with Moving-Object Search Tool (MOST)

2. **Raw image data (CCD-based files with metadata) and image calibration products used in pipelines**
   - ➢ archive via GUI or API at IPAC

3. **Reference images (co-adds), coverage, unc maps, and two source catalogs per image**: PSF-fitting and aperture
   - ➢ archive via GUI or API at IPAC

4. **Alert (point-source event) stream** from real-time image-differencing pipeline: packetized with metadata, 30 day photometric histories, upper limits, ML-scores, provisional names, cutouts on new, reference and diff images, …
   - ➢ mirrored at UW using Kafka; access is through specific program ID channel

5. **Products to support realtime Solar System / NEO discovery and characterization:**
   - streaks (fast moving objects) from difference images: metrics, ML-scores, and cutouts
   - moving object tracks from linking point-source events; known objects are tagged
   - ➢ ZTF-depot webserver (restricted audience)
   - ➢ Moving object metadata (ephemeral astrometry and photometry) delivered to IAU Minor Planet Center

Baseline deliverables & data access portals

6. **Quality assurance metrics, summary statistics, and coverage maps for performance monitoring**
   - ZTF-depot webserver (restricted audience)

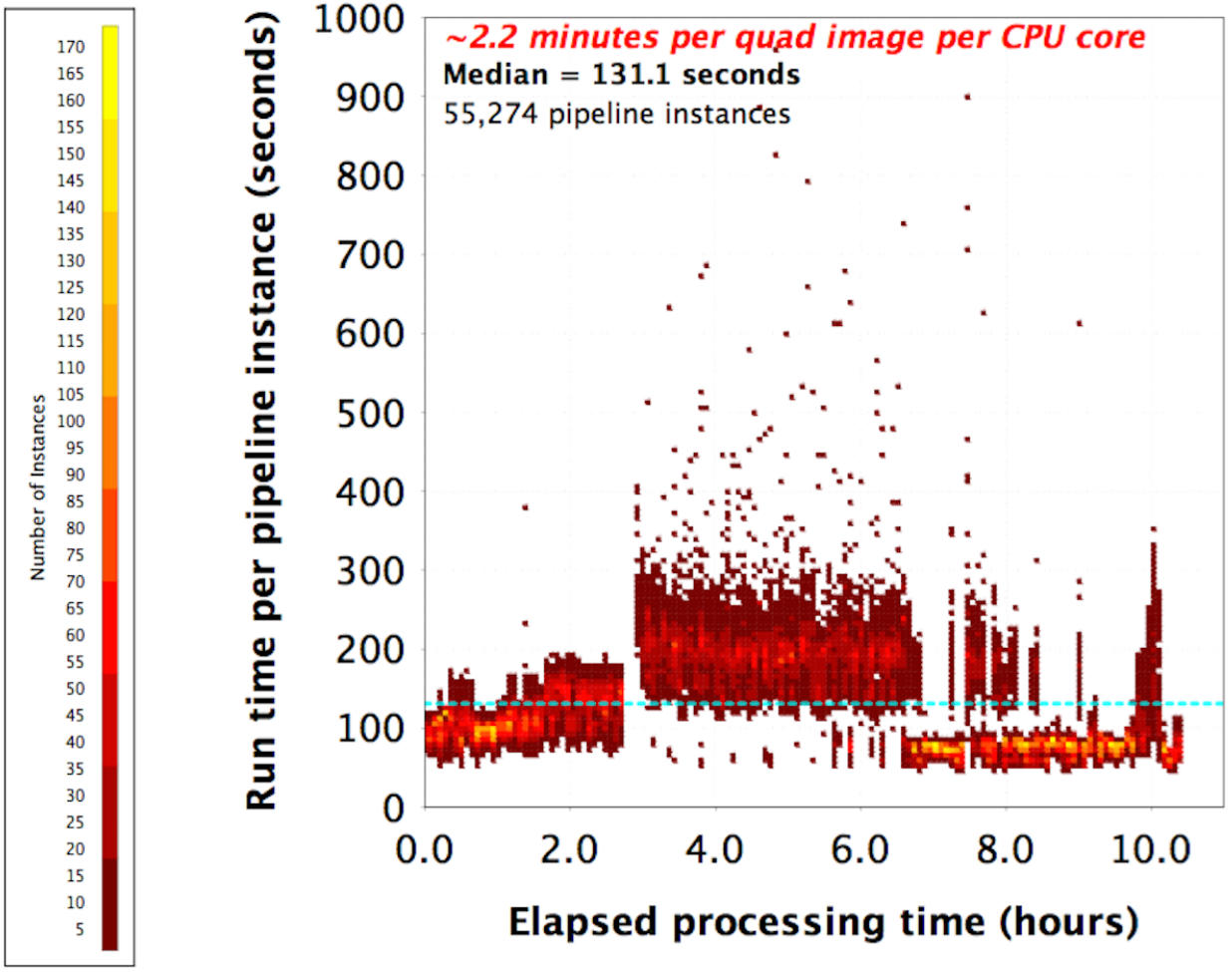
7. **Matchfiles: all lightcurves per readout-quadrant:** from source-matching of epochal PSF-fit extractions
   - restricted (galactic marshal)

8. **Service to query lightcurves & metrics using "object-based" position searches, extracted from matchfiles**
   - archive via GUI or API at IPAC; can interface with lightcurve viewer/analyzer **(in progress)**

9. **Data System Documentation:** pipeline descriptions, recipes, and tutorials on data-retrieval

10. **ZTF Public Website:**
    - *http://www.ztf.caltech.edu*
    - designed by IPAC Communications and Education (ICE) team
    - updated by project team members with privileges

ZTF real-time pipeline runtime
processing unit = one readout-quadrant image



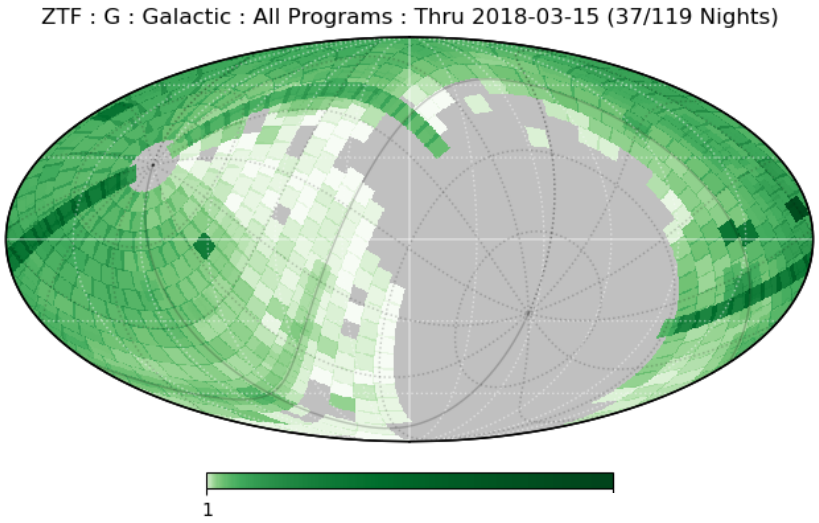
- 66 machines × 8 jobs each
- Based on fields processed on night of March 9, 2018 (UT)
- High tail: with image-diff pipeline, alert generation etc since ref images were available.
- Low tail: science image processing only (no ref images available).

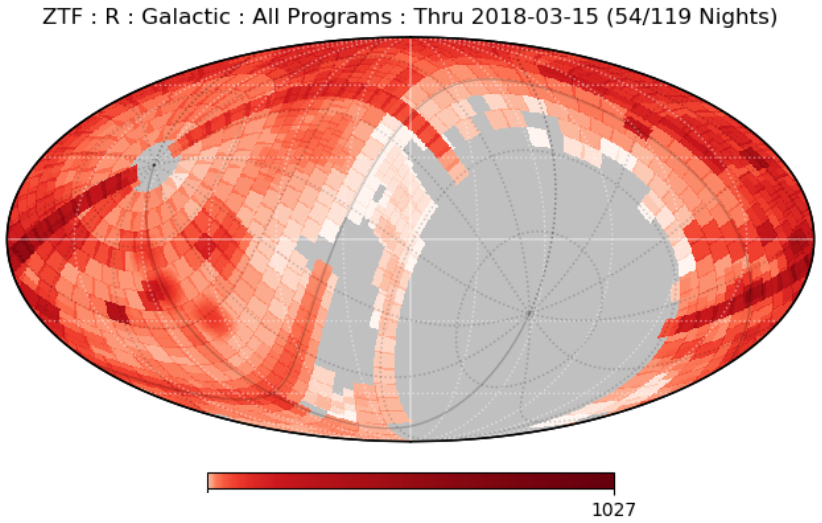
g-filter sky coverage
Oct 14, 2017 – Mar 15, 2018

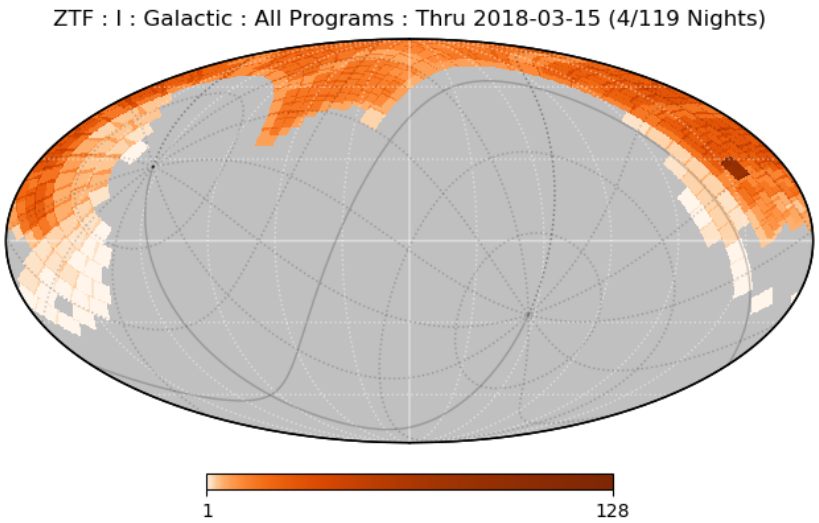
- Depth-of-coverage maps and movies are on ZTF-depot
- Split by program ID (and all combined); filter; three projections: ecliptic, equatorial, and galactic; both cumulative and incremental (per night); formats: .fits, .png, and .mov

R-filter sky coverage
Oct 14, 2017 – Mar 15, 2018

sky coverage (g, R, i)
Oct 14, 2017 – Mar 15, 2018







Accumulated Data Volumes and Statistics
Oct 14, 2017 – Mar 13, 2018

- Number of raw *on-sky* camera exposures ingested: 19,584 (*g*), 27,321 (*R*), 2,147 (*i*)
  - Volume (all raw CCD image files): ~ 23 TB
- Number of archived epochal science image products per CCD quadrant (excludes ancillary files): 2,759,765
  - Volume (with ancillary files): ~ 195 TB
- Number of "survey-ready" reference images (created from data ≥ 2018-02-05): 445 (*g*), 12,306 (*R*), 7,336 (*i*)
  - Volume (with ancillary files): ~ 2.7 TB
- Number of lightcurve matchfiles: 184 (*g*), 135 (*R*)
  - Volume: ~ 110 GB

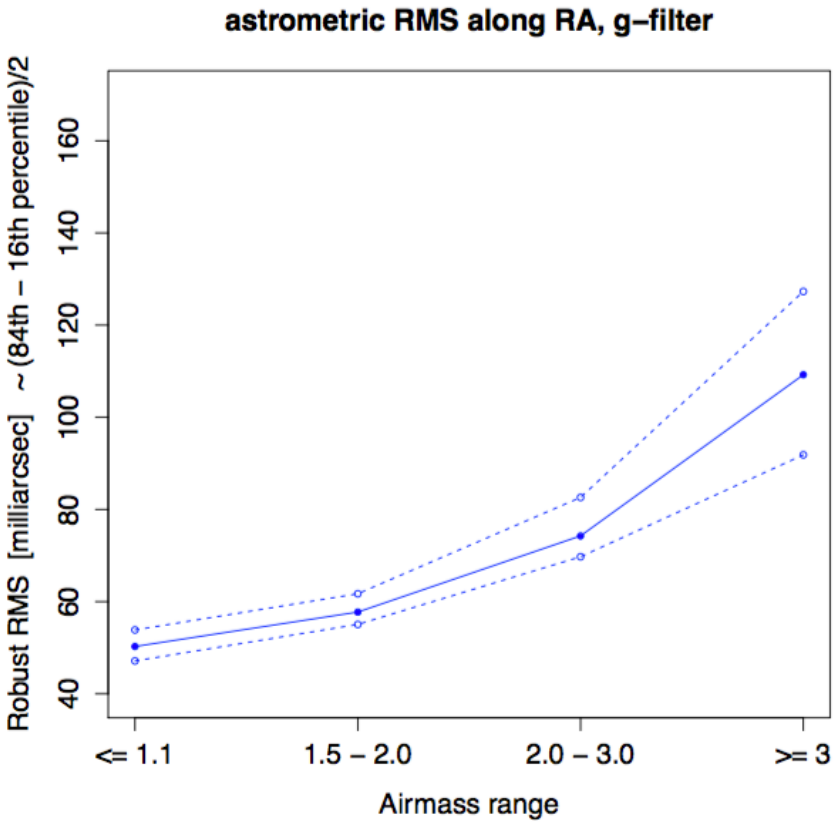
- Number of epochal science image PSF-fit extractions: ~ 14 billion
- Number of epochal science image aperture-based extractions: ~ 11 billion

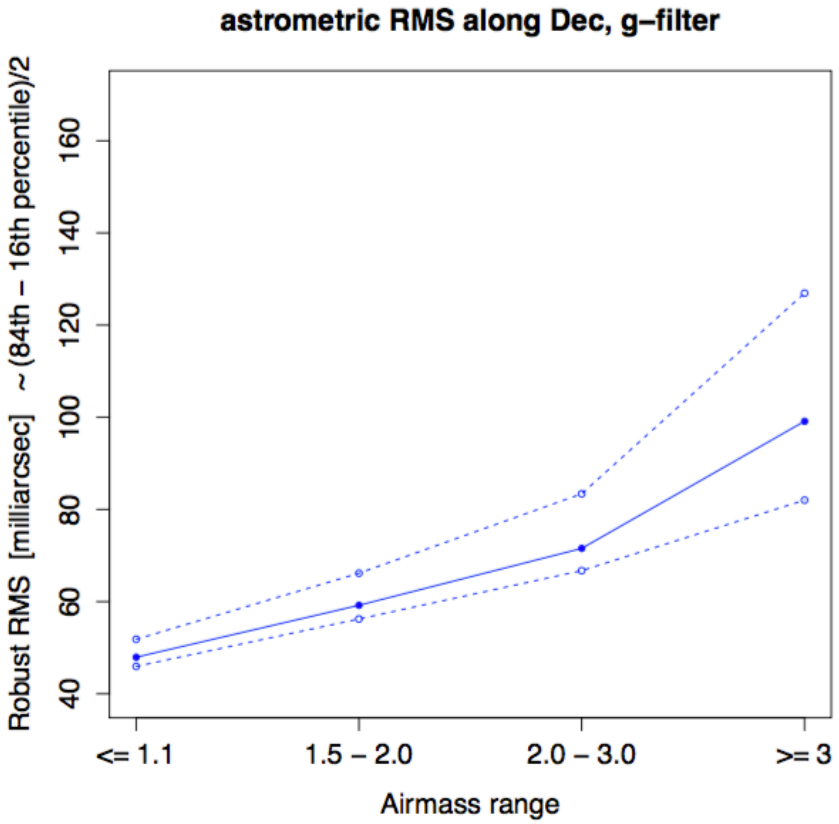
- Number of raw candidate events extracted from all difference images: ~ 10.5M *positive,* 10.1M *negative*
- Number of alert packets generated therefrom: ~ 3.2M *positive,* 3.1M *negative*
- Number of alert packets *potentially* associated with known Solar System objects (≤ 5 arcsec): ~ 945,000
- Number of alert packets with *RealBogus* score ≥ 0.9 since inception of v1 *RB* on 2018-01-12: ~ 145,000

- Number of streaking objects *not associated with known objects*: 3
- Number of streaking objects *associated with known objects*: ~ 2,938
- Number of moving object tracklets *not associated with known objects* delivered to the MPC: 1,388
- Number of moving object tracklets *associated with known objects* delivered to MPC: ~ 116,048

Astrometric Performance (*g*-filter)

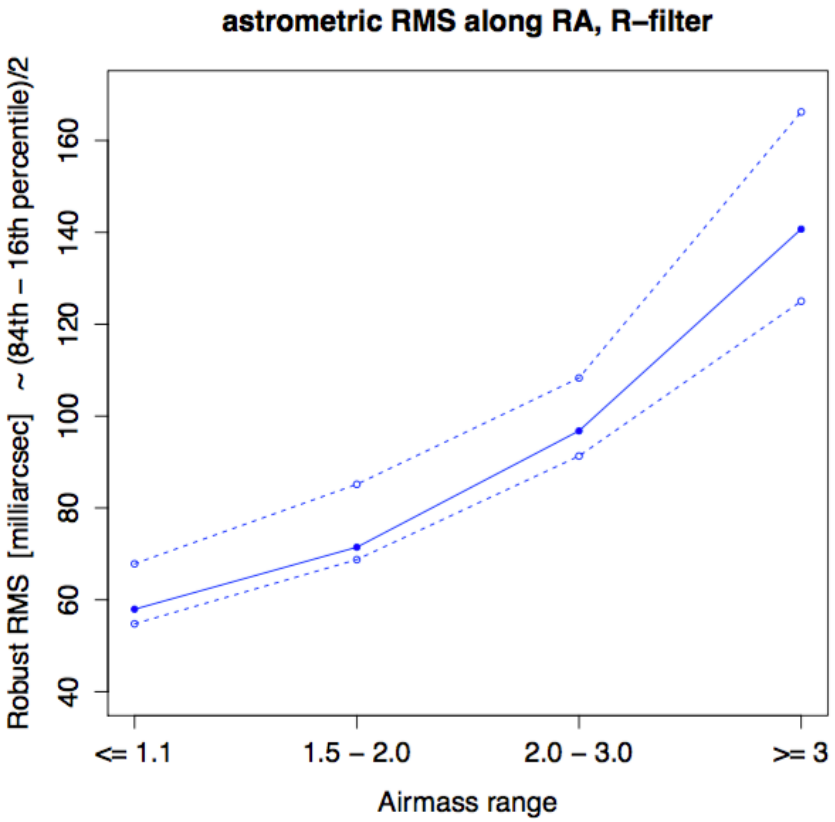
- For airmass range 1 – 2 (survey area), median RMS in *g* is ~ 45 to 65 mas per axis
- Dashed lines are minima, maxima across CCD quadrants

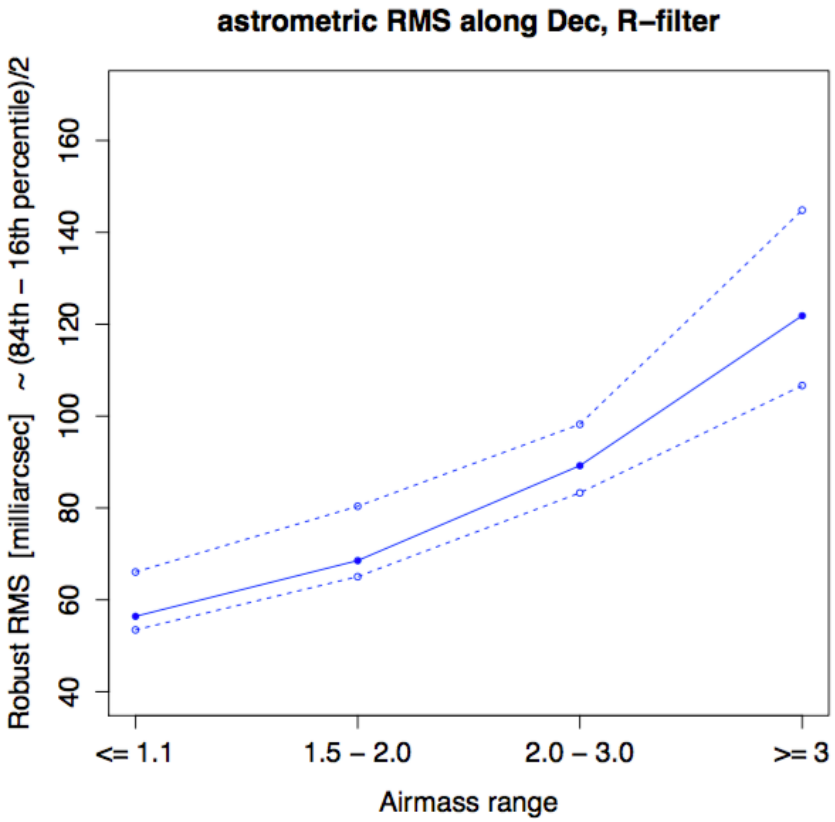




Astrometric Performance (*R*-filter)

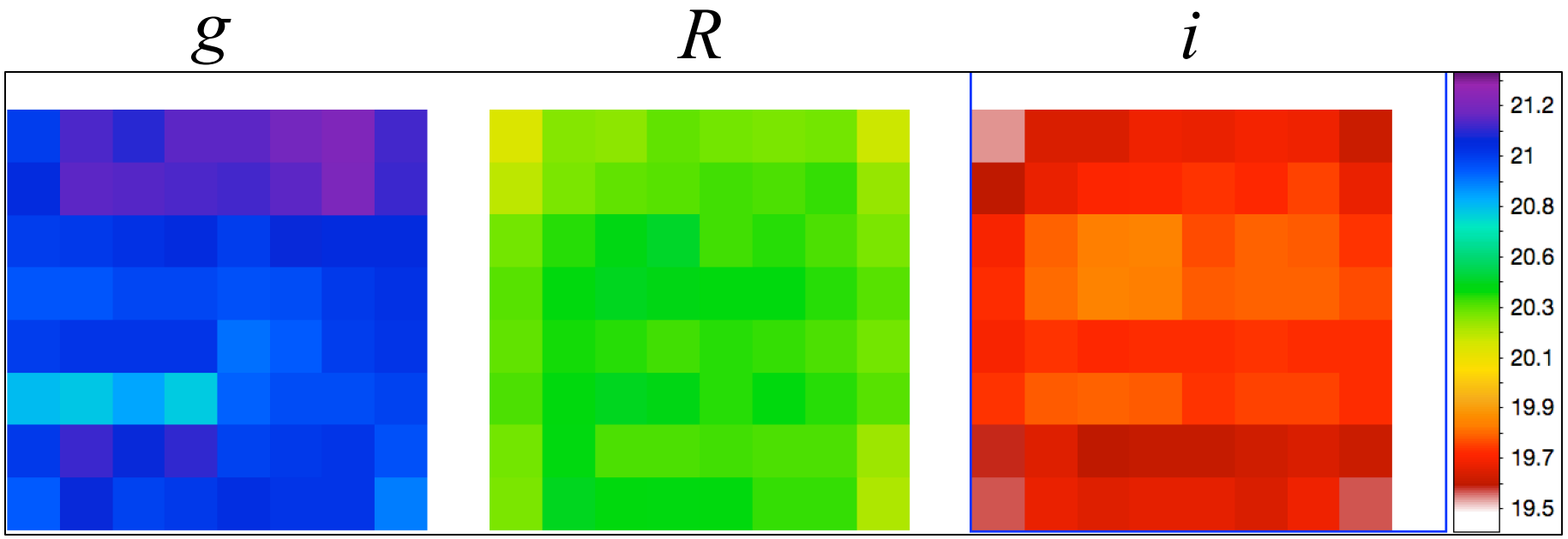
- For airmass range 1 – 2 (survey area), median RMS in *R* is ~ 55 to 80 mas per axis
- *R* has larger residuals than in *g*





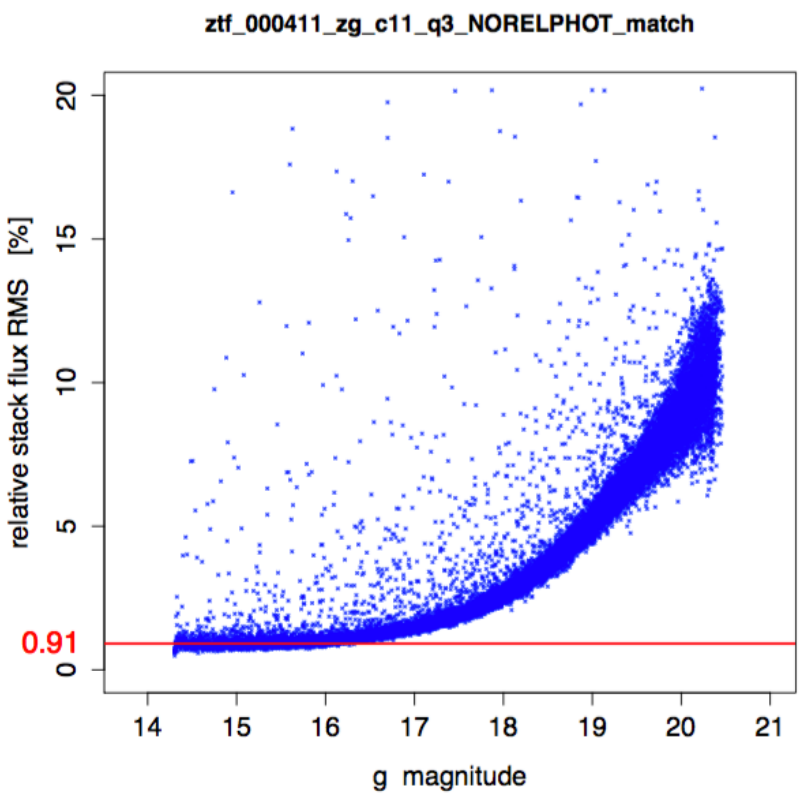
Sensitivity Limits

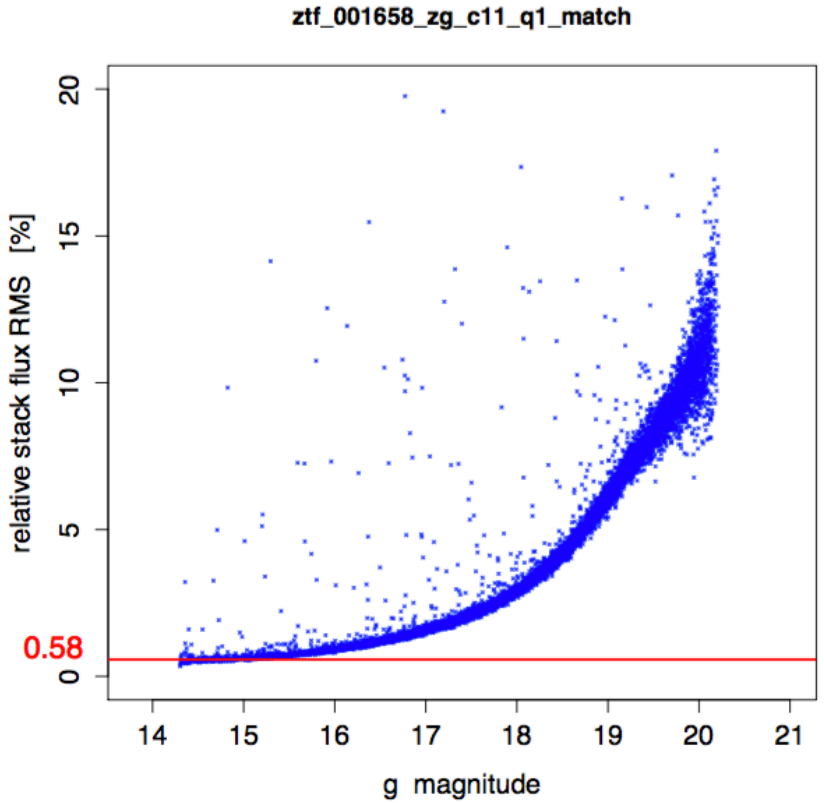
- Using PSF-fit catalog flux uncertainties, estimated 5σ depths are ~ 21.005 (*g*), 20.442 (*R*), 19.715 (*i*)
- Medians from all data acquired at airmass ≤ 1.3 and ≥ 2018-02-05
- Broken down by CCD quadrant:



Photometric precision (repeatability) from *current* matchfiles

- Photometric precision (temporal repeatability of bright sources): ~ 5.5 to 10 millimag across matchfile holdings
- Improved when mutual relative refinement of zeropoints across epochs is turned on
- However, these products are not reaching 5σ depths!
- Problem was with input PSF-fit catalogs (due to internal thresholding on quality metrics); now fixed

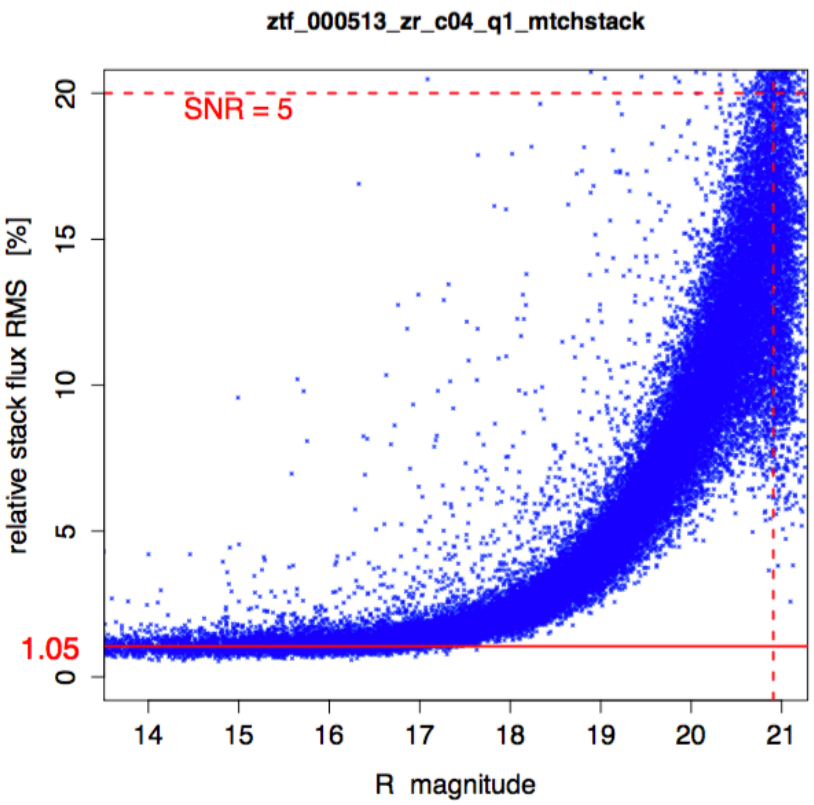




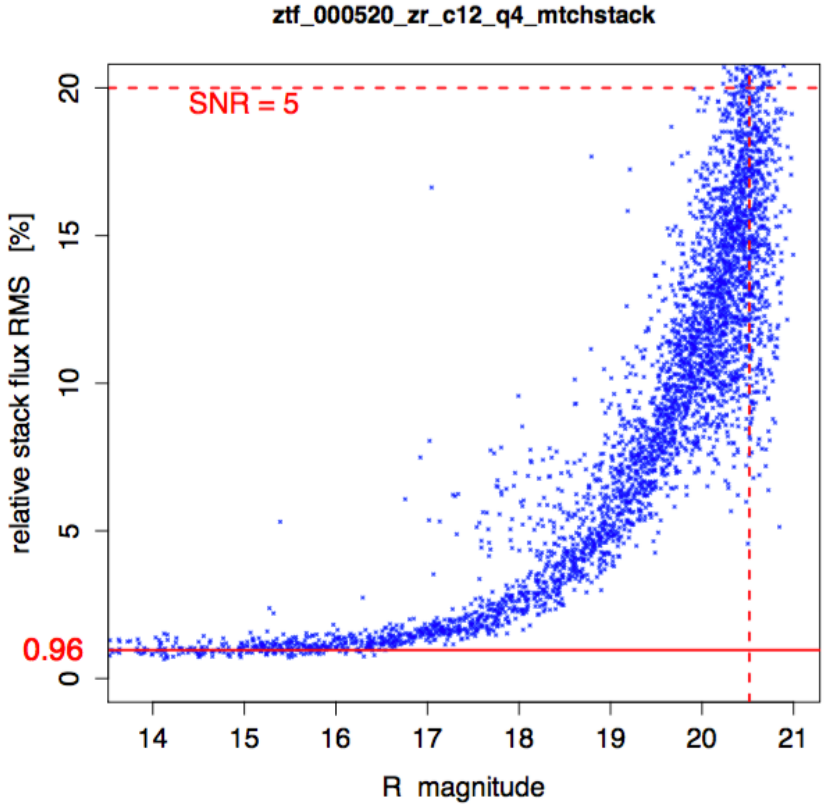
Photometric precision (repeatability) and depth following fix

- Matchfiles were not regenerated; PSF-catalogs were matched using offline software
- New results are below (no relative refinement of zeropoints across epochs)
- Depths now consistent with expectations from photometric uncertainties in PSF-fit catalogs (slide 14)

galactic plane field



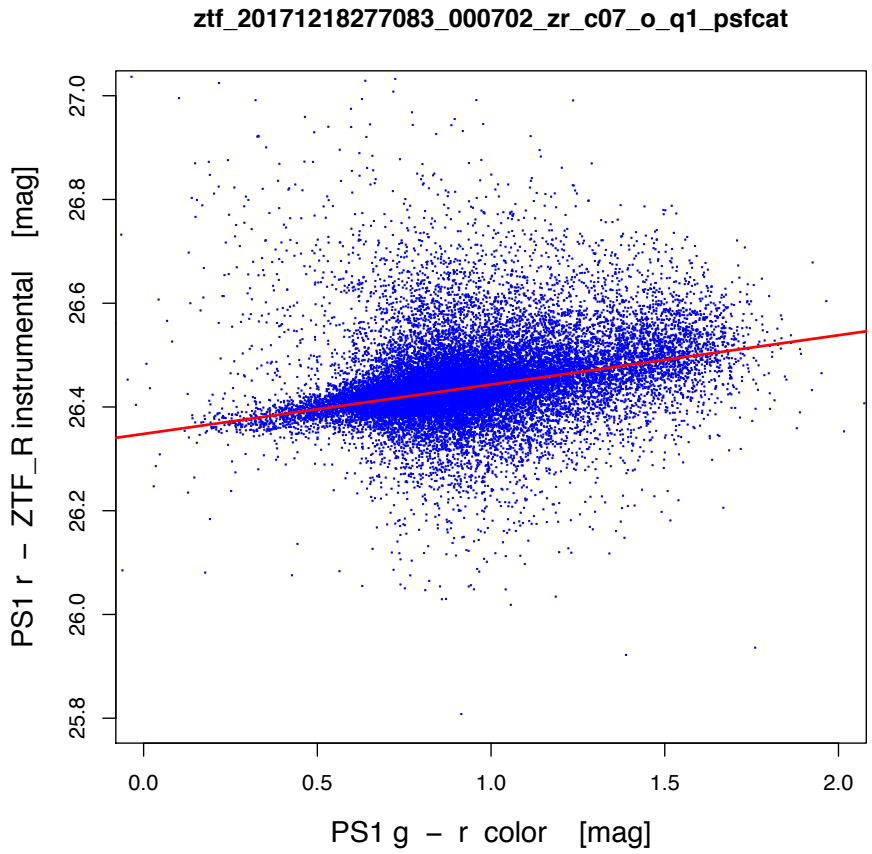
high galactic latitude field



Photometric calibration

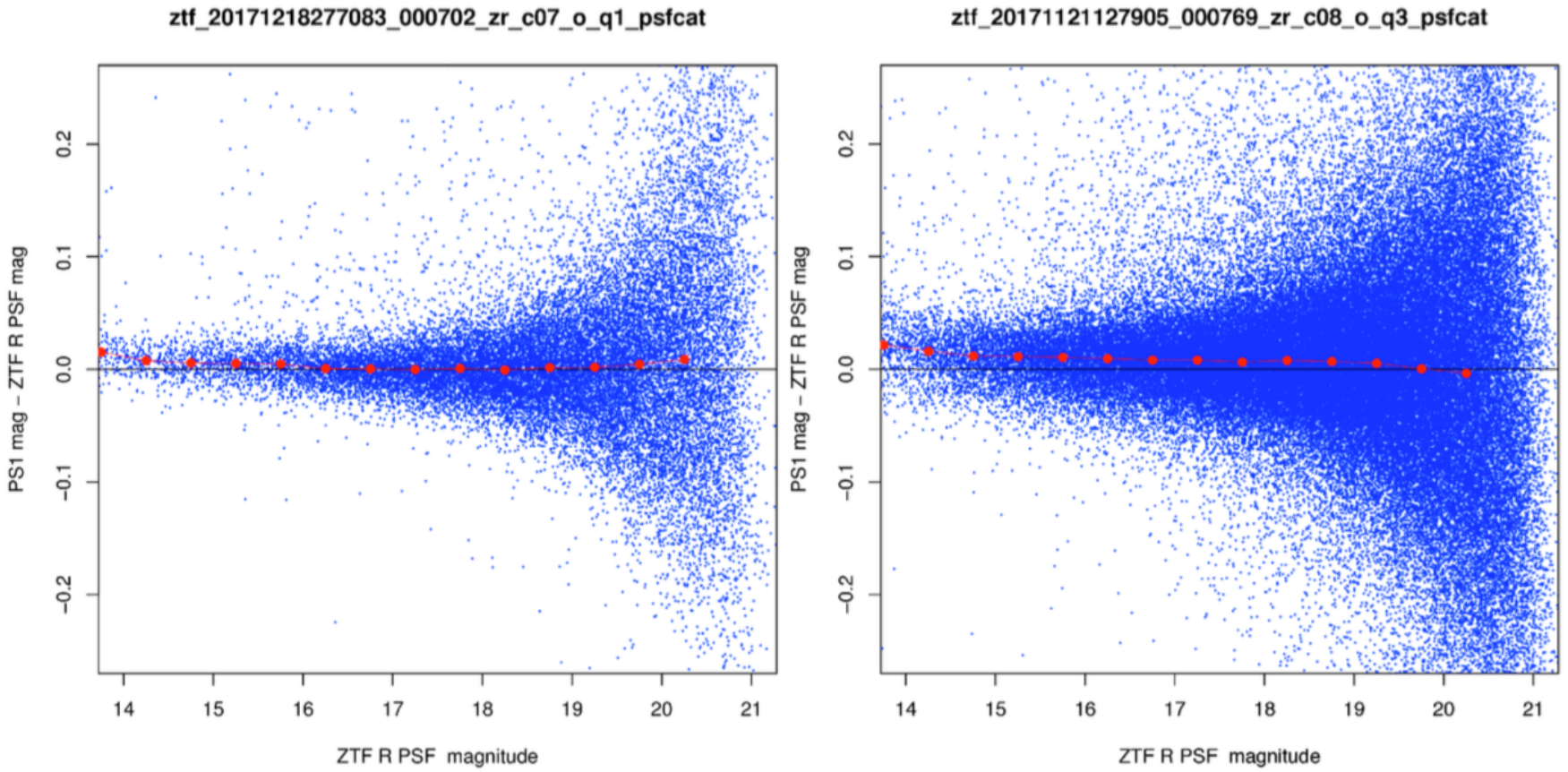
- Instrumental magnitudes are calibrated against Pan STARRS1 using a filtered, non-variable subset of calibrators
- Simple linear model is fit per CCD quadrant in each exposure to derive a *zeropoint* and *color coefficient*
- Uncertainties in fit parameters are typically < 0.1%; very stable across same-night exposures at fixed *airmass*

$$m_{ps1} - m_{instr} = ZP + c(g-r)_{ps1}$$



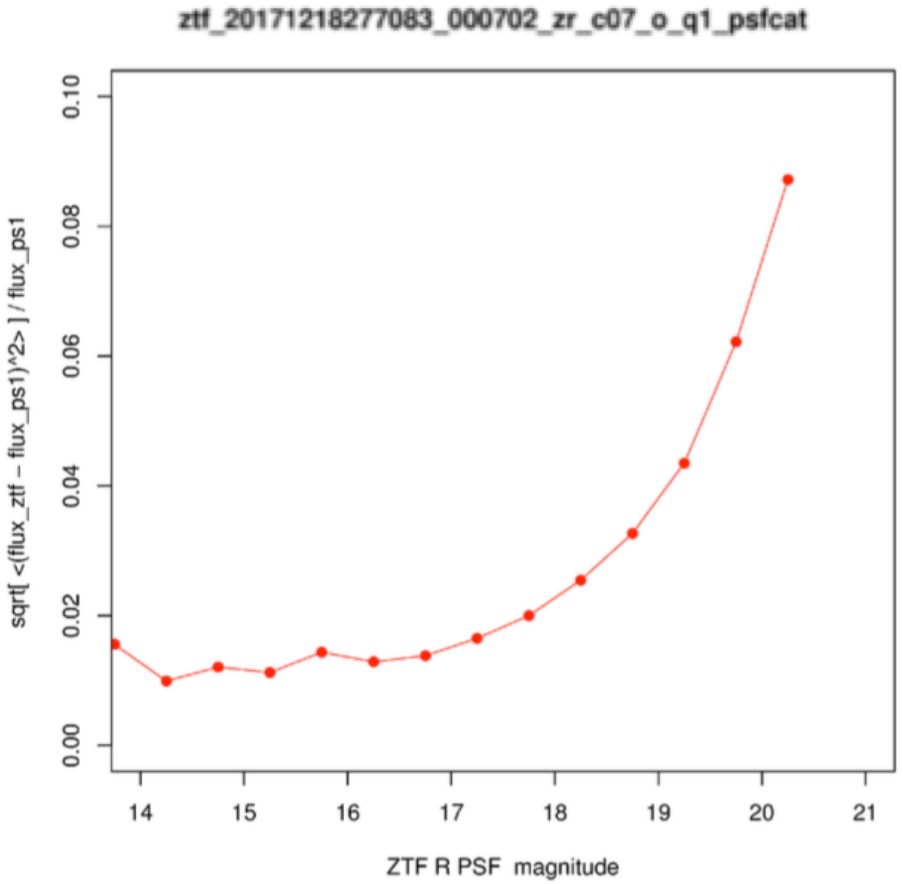
Photometric calibration check

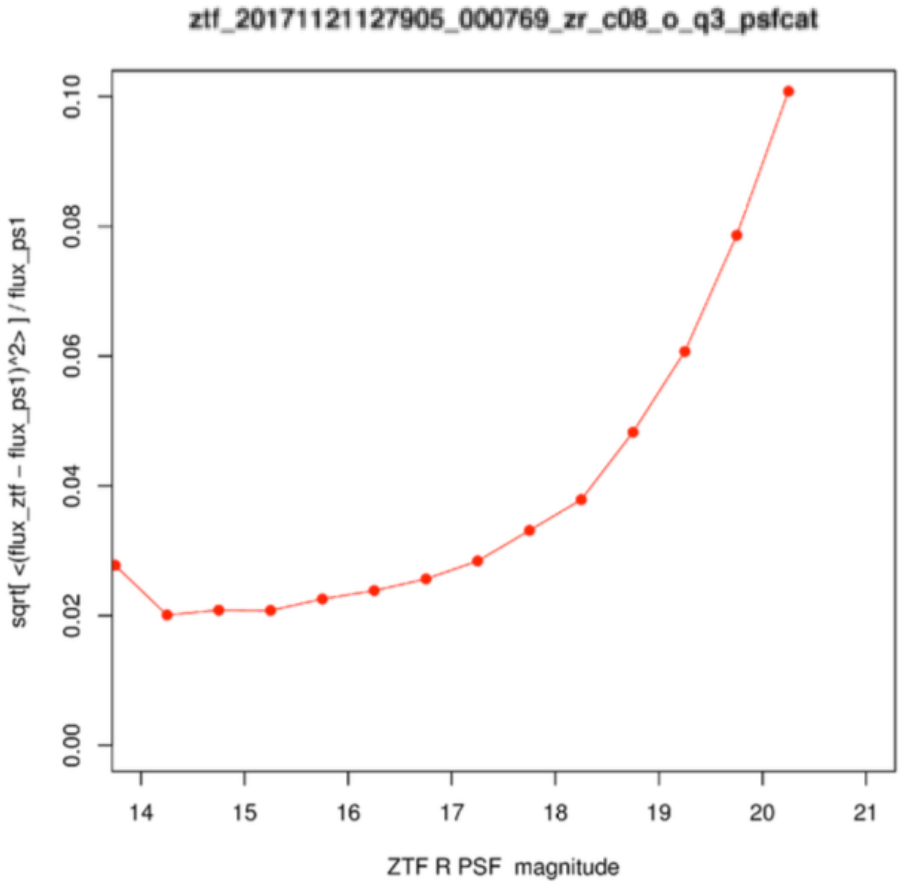
Following calibration, magnitude dependent biases of up to ~ 0.02 mag are seen in both PSF-fit and aperture-based catalogs with respect to Pan STARRS1. This is variable across fields.



Photometric calibration check

- Using previous plots, relative RMS with respect to Pan STARRS1 is ~ 1 to 2.5% at R <~ 17; variable across fields
- Contribution is mostly from scatter in instrumental photometry, not calibration error
  (compare with RMS from photometric repeatability in slide 16)





Task List (tied to baseline design)

Currently in progress:

- Ingestion of new PS1 star/galaxy scores into database to associate with alert stream in realtime (any day now)
- Refinements to matchfiles; robustness updates
- Lightcurve retrieval GUI to interface with matchfiles; includes time series viewer/analyzer
- Enable image-cutout service to also operate on archived (compressed) difference images

Scheduled for near future:

- Ghost prediction and masking (have specs in hand)
- Ability to download all ancillary products from archive using GUI (already possible via APIs)
- Long-term archiving of avro alert packets, alongside science products; remove from ZTF-depot

Dependencies (external deliverables; with ongoing integration and testing):

- Refinements to point source real-bogus classifier (crucial for alert packets)
- Streak real-bogus classifier
- Tagging *known* comets in alert stream and moving object products (working with MPC software)
- Eyes on data products!

To be discussed / wish list

To be discussed (pending analysis):

- Optimal weighting strategy for flat-field generation per CCD quad (combining different LED configurations)
- Star-flat assessment and application
- Exposure-time correction map (flat augmentation, ~ 0.2% at edges)
- Interface with a finderchart service?
- *i*-filter fringe correction

Personal wish list:

- High airmass astrometric corrections due to differential atmospheric refraction **outside survey boundary** (see plots)
- Mask CCD charge bleeds from bright/saturated sources
- All-sky depth-of-coverage maps for reference image holdings per filter
- Refined photometric (re)calibration of reference images and catalogs; pending analysis
- Global synopsis of available Gaia and PS1 calibrators in all fields and CCD quadrants in science grid; i.e., which CCD quadrants on science grid cannot be calibrated? Good to have a list for cross-checking
- Global synopsis of photometric calibration accuracy

Back up slides

High level Data System
Objectives & Requirements

From the MSIP (NSF) proposal & ZTF Management Plan (03 / 12 / 2014):

- Sustain processing & storage for three years of operations (initially: 2017 – 2019)
- Leverage the existing PTF Data System and Archive infrastructure
- Scale data processing and storage to sustain a 15 x PTF data-acquisition rate
- Generate data products similar to those as PTF and additionally:
  - Provide a lightcurve retrieval / search tool
  - Real-time transient alerts for public consumption, beginning in survey year two
- Support public release of archived products every six months, with the first occurring at the end of survey year one
- Process data and generate alerts at all galactic latitudes in the Northern visible sky

More concrete specifications on products, requirements & goals, aligned with the science proposed to NSF and by project partners (known at the time) came together in a document by E. Bellm (12 / 16 / 2015):

- Data System development was guided primarily by this document.
- Two additional requirements presented a huge challenge:
  - Include a database of sources detected by PSF-fit photometry from each epochal image.
  - Maintain a database of lightcurves generated by positionally matching all epochal image extractions.
  - Following analyses on source statistics, these are not feasible; arrived at a compromise for serving lightcurves
- Timing requirements:
  - > 95% of the images acquired at P48 need to arrive at IPAC within 10 min (goal: 5 min)
  - > 95% of the images received at IPAC must be processed with alerts published in < 10 min (goal: 5 min)

Decision made in early 2017:

- Commence public alerts earlier than planned: now in April 2018.

ZTF Pipelines and run frequency

Overall, there are 9 interdependent pipelines, grouped into four categories.

- All implemented and tested on simulated camera-image data; some pipelines also tested using real camera data.
- All baseline archival products, formats, and methods for access are finalized.

Raw data ingestion and initial processing:

1. Raw data ingest, archival of raw images and storage of metadata in database [*realtime*]
2. Raw-image decompression, splitting into readout-quadrant images, floating bias correction, QA metrics [*realtime*]

Calibration-image generation:

3. Bias-image derivation from stacking calibration images acquired in afternoon [*before/after on-sky operations*]
4. High-ν flat (pixel-to-pixel responsivity) from stacking illum. flat-screen exposures [*before/after on-sky operations*]

Real-time science-level processing:

5. Instrumental calibration of readout-quadrant images: includes astrometric and photometric calibration [*realtime*]
6. Image-subtraction with transient-event extraction (point sources & streaks), alert packets & distribution [*realtime*]

Ensemble-based (collective-image/catalog) processing:

7. Reference-image generation (co-addition of epochal images from 5) [*when sufficient good quality data available*]
8. Source-matching/lightcurves with relative photometric refinement; inputs from 5 & 7 [*every month, TBD*]
9. Moving object tracks, orbit-fitting, QA; from linking point-source events from 6 [*end of night, 3-4 day window*]

ZTF Lightcurve Pipeline

- All sources detected in epochal images are matched against the reference-image source catalog for a given field, CCD quadrant, and filter
- The "cleanest" least variable sources are used as anchors for the relative photometric calibration
- Individual image gain-correction factors are computed using a global least-squares fit across all epochs
- These gain-corrections are applied the image photometric zero-points
- The refined zero-points *are expected* to improve relative photometry to a few millimag for bright sources
- This pipeline will be triggered on timescales of typically one month (TBD), contingent on data accumulated
- All lightcurves for a single CCD quadrant and filter are stored in a "matchfile" (hdf5 pytable format)
- Accompanying each lightcurve is a set of >100 metrics: RMSs, Skews, Stetson indices …
- All lightcurves and metrics are seeded by an object ID; these objects are loaded into a database to support spatial searches; associated lightcurve is retrieved from the "matchfile" containing that object position
- Expect of order 1.3 billion objects (individual lightcurves) for ZTF
  - There will be multiple (disjoint) lightcurves per object due to the two overlapping science grids
  - There is no plan at present to splice lightcurves belonging to the same object

Data Access Policy

- Observing time during science operations will be split between three categories:
  - **Public** (NSF-funded MSIP survey: 40%)
  - **Private collaboration** (40%)
  - **Caltech TAC** (20%)
- Managed per exposure (epoch) using a *programID* propagated from scheduler to raw-image metadata

---

- Private/Caltech observers can access their data in near-realtime, soon after archive ingestion. This includes all calibration products and lightcurves from epochs tagged by their respective *programID*s queried via archive GUI.
- Public data will only be available at the public release times for general access by all.
  - raw images, processed epochal images, accompanying source-catalog files, difference images
  - reference images and catalog files
  - lightcurves constructed from public epochal data only
  - calibration data products
- Public alert packets (triggered from events detected in public exposures) will only contain public data. This includes their 30 day event histories.
- Private alert packets (triggered from events detected in private exposures) can contain unreleased public data in their 30 day event histories.
- Caltech alert packets (triggered from events detected in Caltech exposures) can contain unreleased public data in their 30 day event histories.
- No restriction on input data used to generate products for Solar System science: streaks & moving-object tracks; selected (human-vetted) products will be delivered to MPC.
- No restriction on input data used to generate reference image (co-add) products.
- No restriction on input data used to generate source match-files (lightcurve files):
  - MOU in place with the only customer of these products: Galactic Marshal
  - only privately-tagged and *already-released* public data therein to be ingested by Marshal

Expectations prior to commissioning: data volumes and source statistics

Estimates per night, based on an average on-sky duration of ~ 8hr 40min

- Number of on-sky camera exposures per night: ~ 700; calibration exposures: ~ 100
- Raw image data volume: ~ 1 TB (no compression)
- Raw incoming data rate: ~ 230 mega bits per second (no compression)
- Data product volume: ~ 3.8 TB per night (real-time products only)
- Number of **point source** transient events (flux and motion-induced): ~ 1 million (0.1x – 5x)
  - ➢ following machine-learned vetting of > 5σ events; sky-location dependent
- Number of streaks (candidates for "fast-moving" objects following ML vetting): ~ 50 to 500
- Number of single exposure extractions: ~ 1 billion (PSF-fit based); ~ 300 million (aperture-based)
  - ➢ sky-location dependent
- Number of single readout-quadrant image products (science, difference, mask, catalogs): ~ 230,000

For nominal three-year survey (number of observing nights / year: ~ 260):

- Volume of data products: ~ 3 PB
- Number of single-exposure extractions: ~ 800 billion (PSF-fit based); ~ 230 billion (aperture)
- Number of reference images (co-adds in static library for image subtraction): ~ 282,000 (~ 55 TB)