ZTF Science Data System: Progress & Plans

Frank Masci & the IPAC/Caltech ZTF Team

September 3, 2019

HEISING-SIMONS
FOUNDATION

Outline

- Updates
- Reference image reanalysis
- Second Public Data Release
- Photometric corrections (in context of *lightcurve matchfiles*)
- Miscellaneous / planned tasks
- ZTF Phase II

Some updates

- First public data release occurred on May 8, 2019. Much feedback received from community & partnership.
- Continued refinements to forced photometry service from feedback received.
- Transitioned from PostgreSQL to SQLite database for querying PS1 sources and matching to alerts.
- Set up data system to handle new northern equatorial polar fields (to reduce that infamous 20° hole).
- Improved archive “retry logic” to mitigate archive failures in realtime (network related).
- Computation of new *Deep Real-Bogus* (*drb*) metric for point source transients and inclusion in alert packets.
- Routine (monthly) generation of lightcurve tarballs from TESS sector observations.
- Reformatting of (newly appended) *matchfile* contents into ancillary text files to support ingestion into kowalski.
- Characterizing systematics in epochal PSF-fit photometry with respect to PS1 (spearheaded by Andrew Drake).
  - Developed infrastructure to apply photometric corrections to lightcurves (*matchfile* pipeline).
- Reanalysis of overall reference image quality with impact study.

Reference Image Analysis

- There was a push early in the survey to generate reference images for as much of the sky as possible in order to commence alert generation.
- Reference image usage:
  - Image subtraction → alert production, asteroid detection.
  - Accompanying source catalogs provide "seed positions" for generating lightcurves.

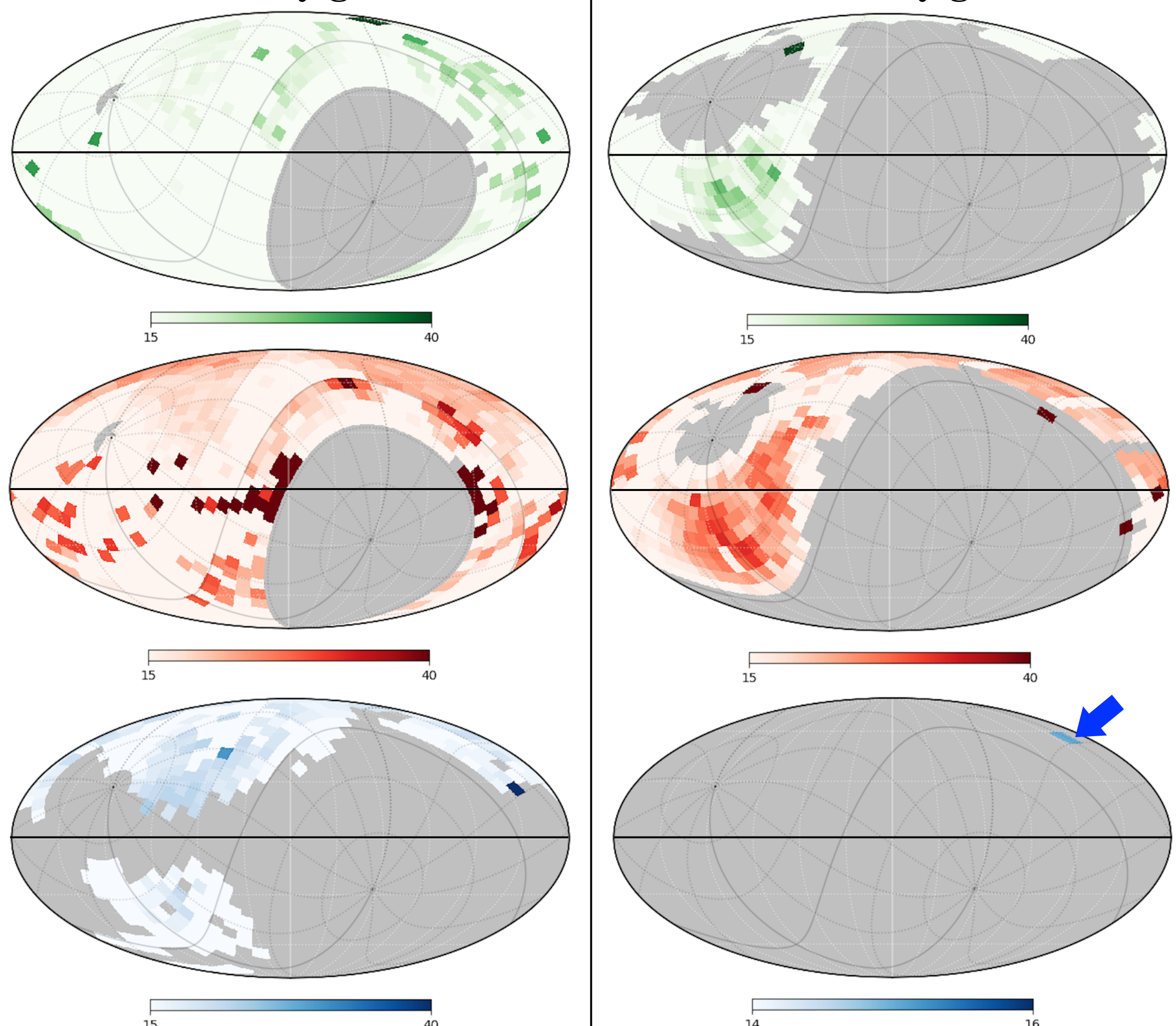
- Procedure to this day:
  - Execute reference image "checker" pipeline every morning.
  - Checks which fields / CCD-quadrants / filters are missing references.
  - As soon as $N \geq 15$ science images satisfy quality criteria, generate a reference image.
  - Archive and lock-down the reference image; never revisit. Max cutoff is 40 input images.

- Now that we have a lot more epochal data, it's worth revisiting whether we can improve reference image quality by being more restrictive on the input image selection criteria.

- Reference image quality impacts all science programs.

Reference Image Coverage: Aug 23, 2019
galactic projection (l, b = 0, 0 centered)

	Primary grid	Secondary grid



***g*-filter**

Total = 50,932 (both grids)
% of sky with ≥ 1 visit: 98.6%
% of sky with ≥ 15 visits: 99.3%

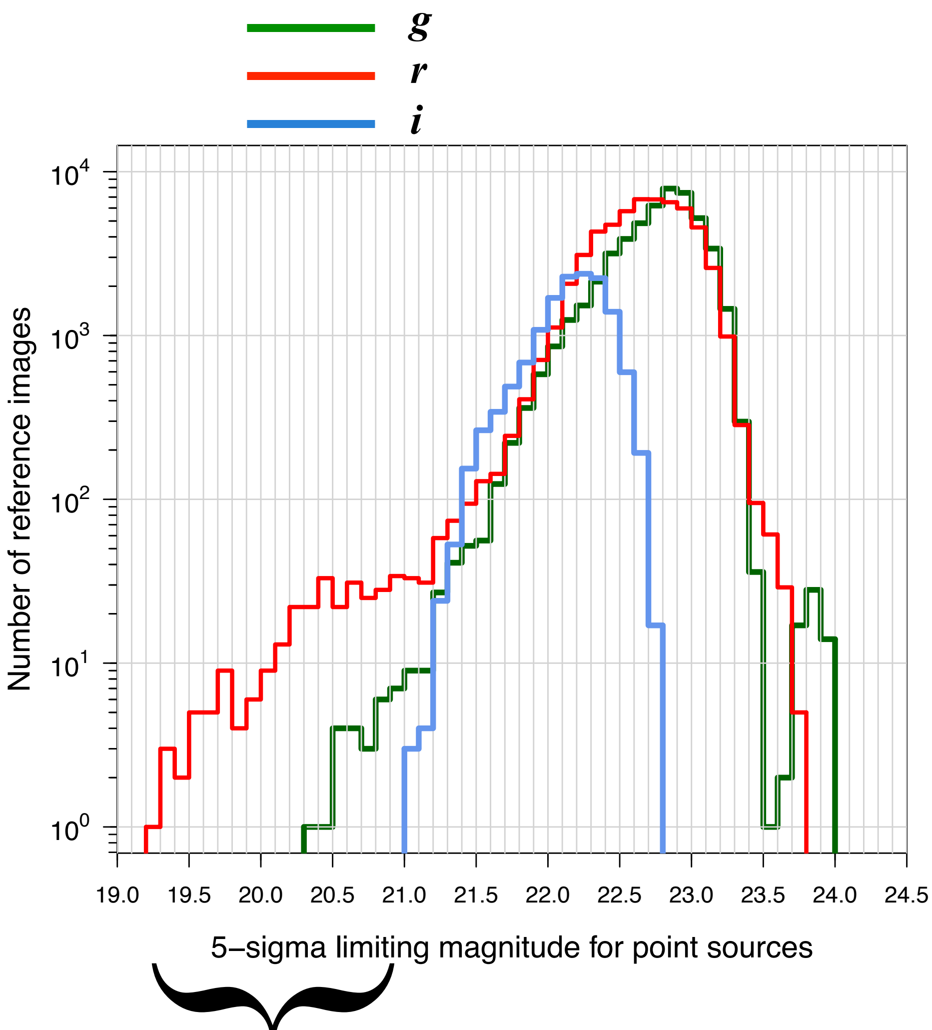
***r*-filter**

Total = 57,837 (both grids)
% of sky with ≥ 1 visit: 98.8%
% of sky with ≥ 15 visits: 99.4%

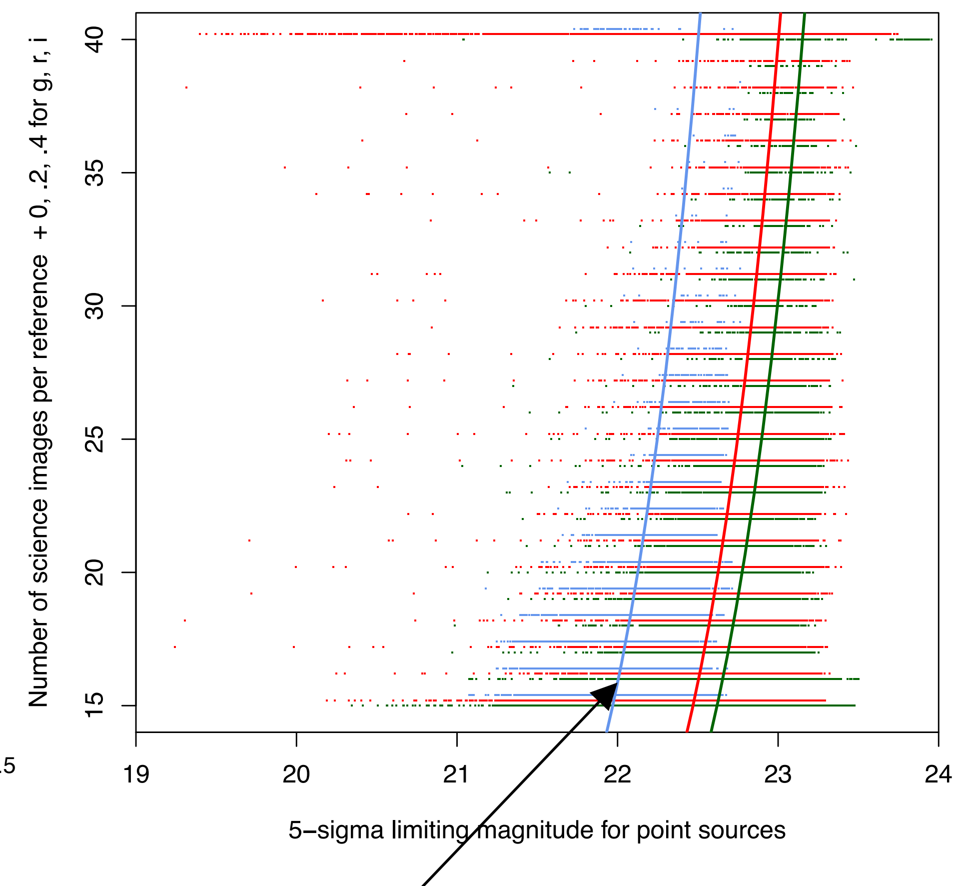
***i*-filter**

Total = 13,631 (both grids)
% of sky with ≥ 1 visit: 48.0%
% of sky with ≥ 15 visits: 87.7%

Current reference image depths



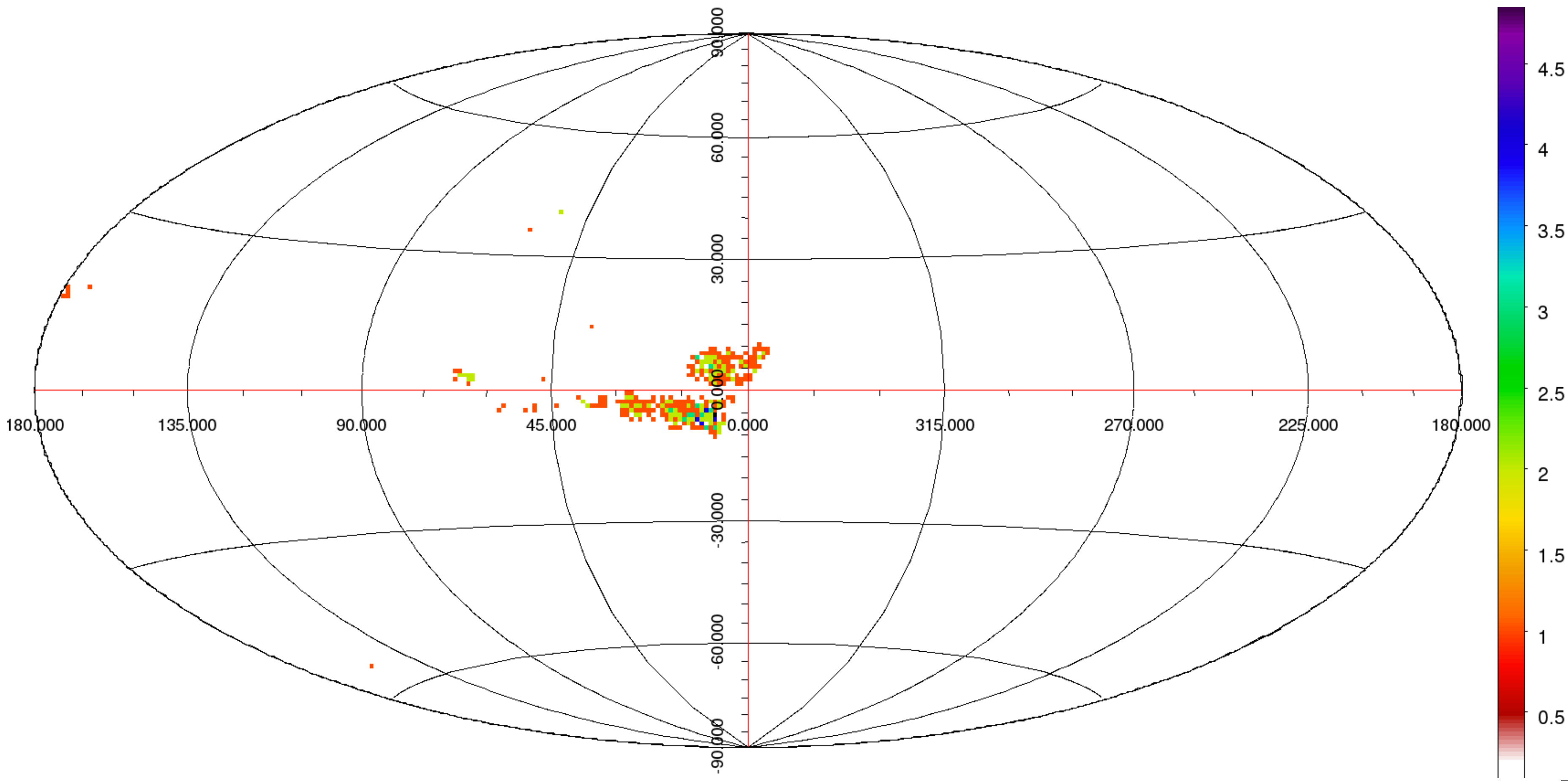
Fields with ultra-high source confusion (PTO): noise & mag-limit estimators break down



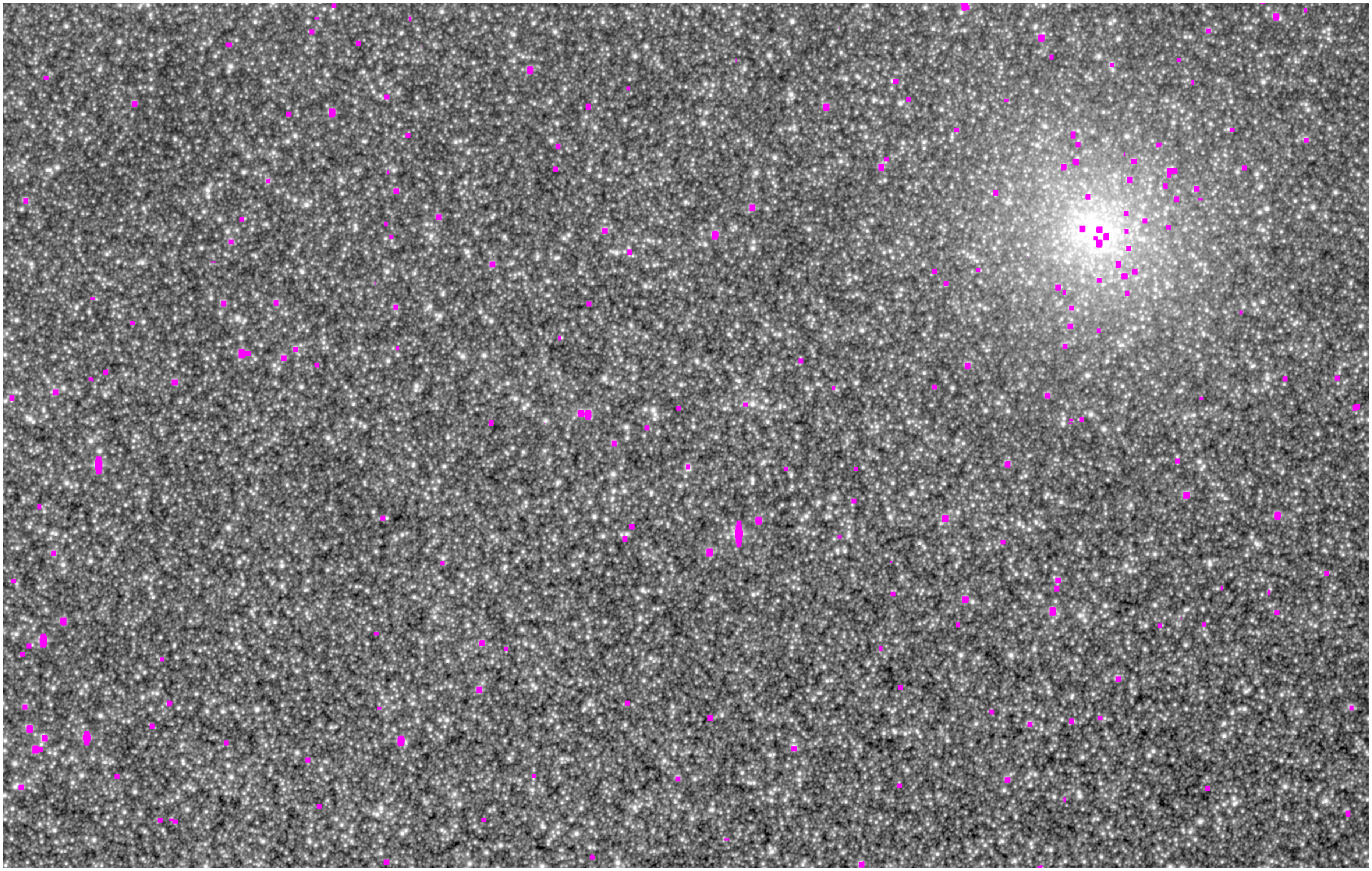
Prediction: $m_{limref} \sim m_{limsci} + 1.25 \log_{10} N$

Where are those "low-depth" reference images located?

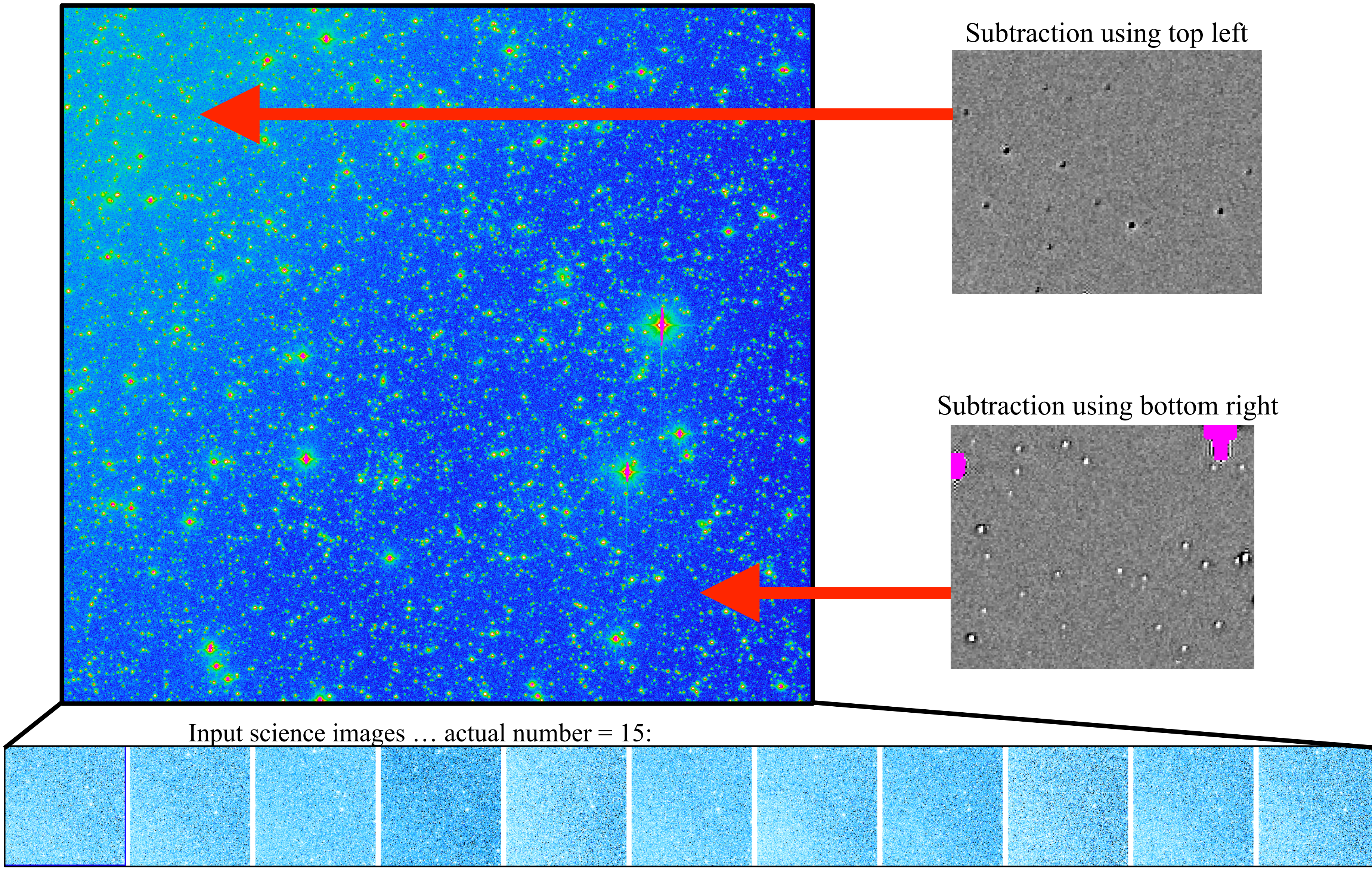
- Shown are CCD-quadrant "footprints" mapped into galactic coordinate system.
- Only those with limiting magnitudes < 21.2 mag are shown (388 references).
- Colors refer to overlaps which include effects from resampling onto a coarser grid.



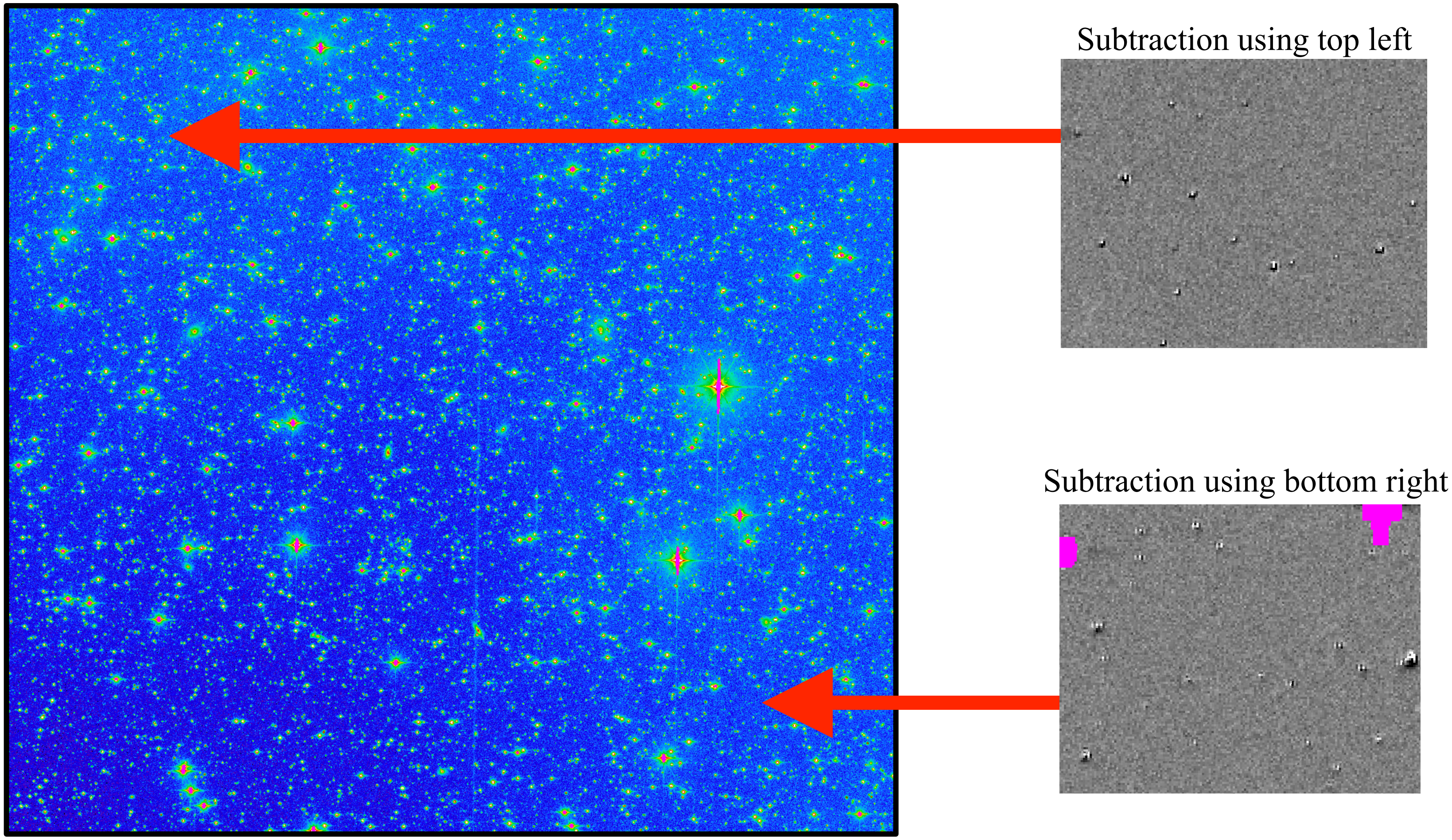
Example of a reference image near galactic center (this is a zoom!)



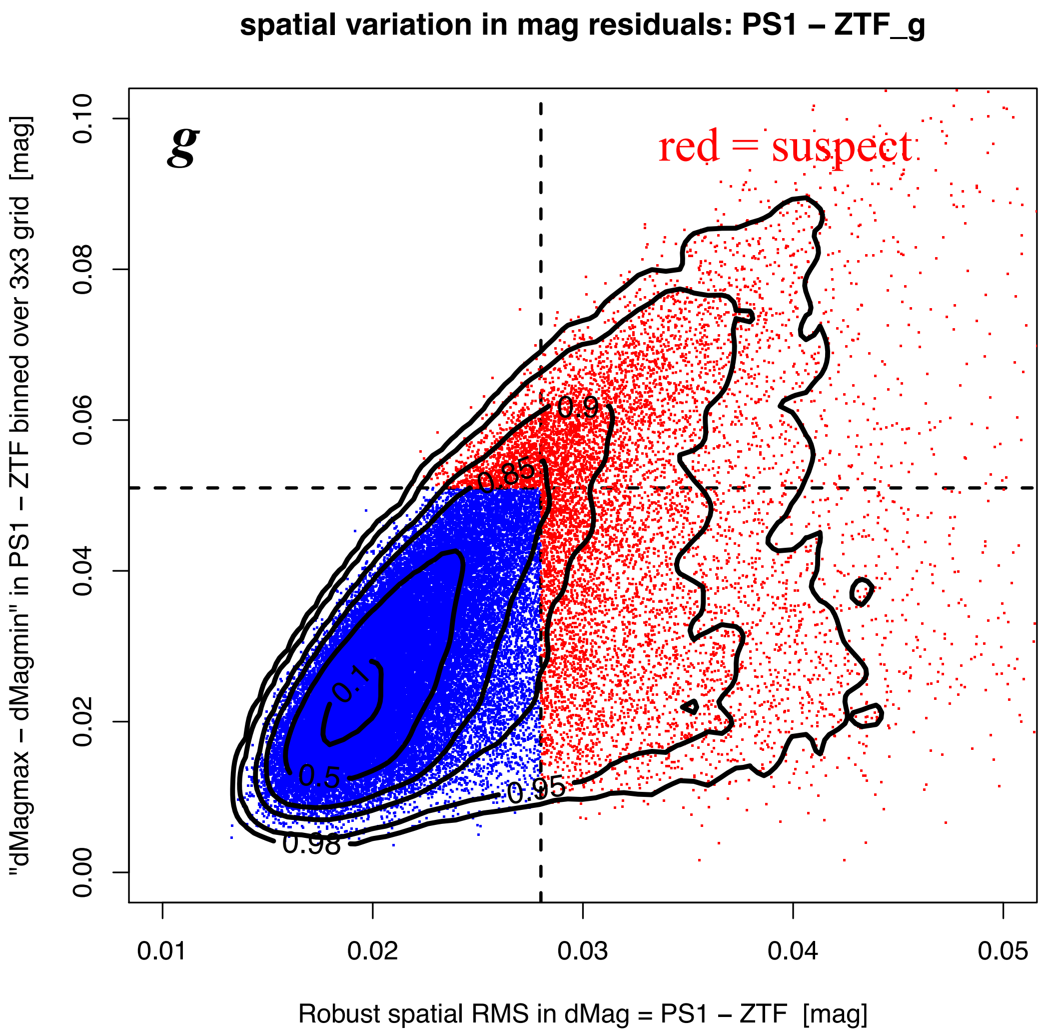
Example of a bad quality reference and its impact



Same reference image using cleaner (*flatter*) input science images



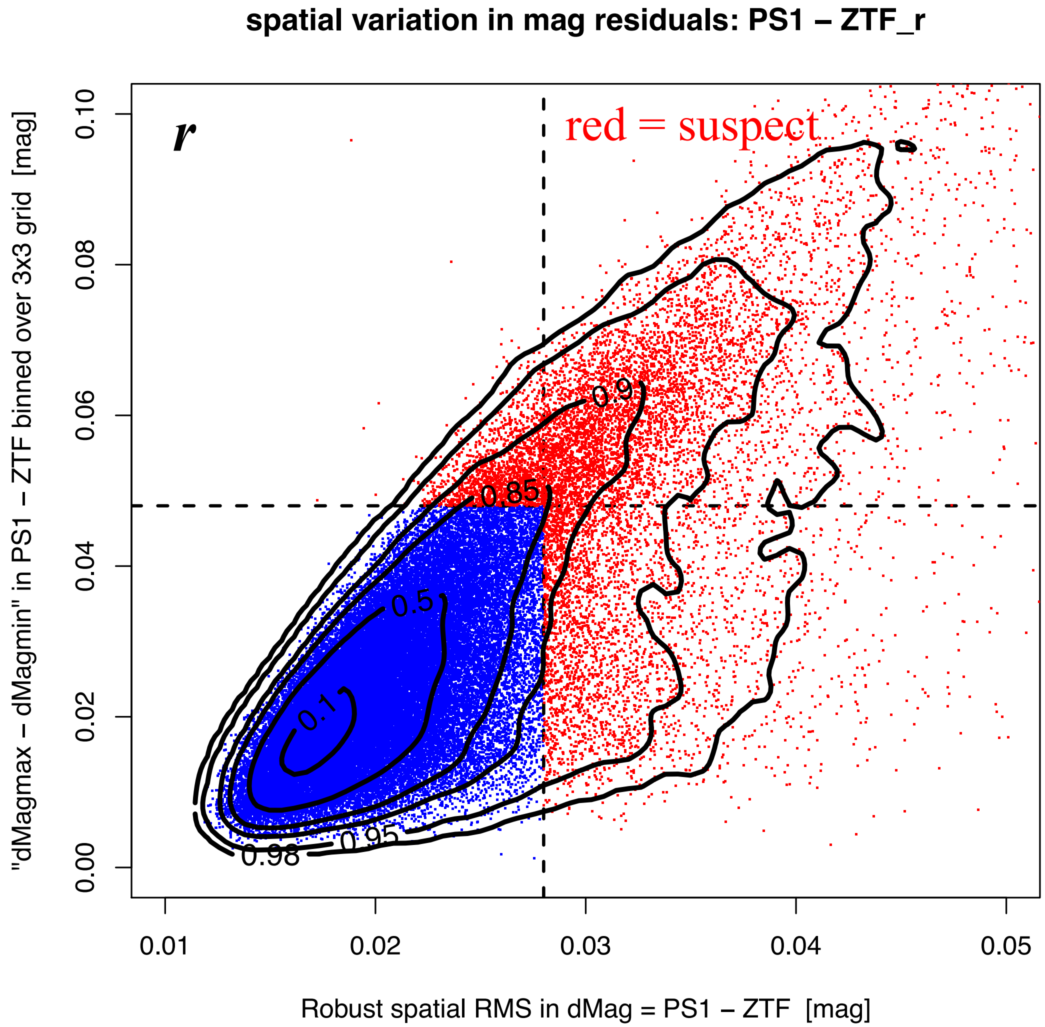
Suspect reference images in *g*



For each ref image, computed:
$\Delta = max\{\text{dMag}\} - min\{\text{dMag}\}$ where $\{\text{dMag}\} = median(\text{PS1} - \text{ZTF}_{mag})$ in 3×3 spatial bins over each image
and
Robust global *RMS* in dMag for all sources with $13.5 \leq mag \leq 18.5$.

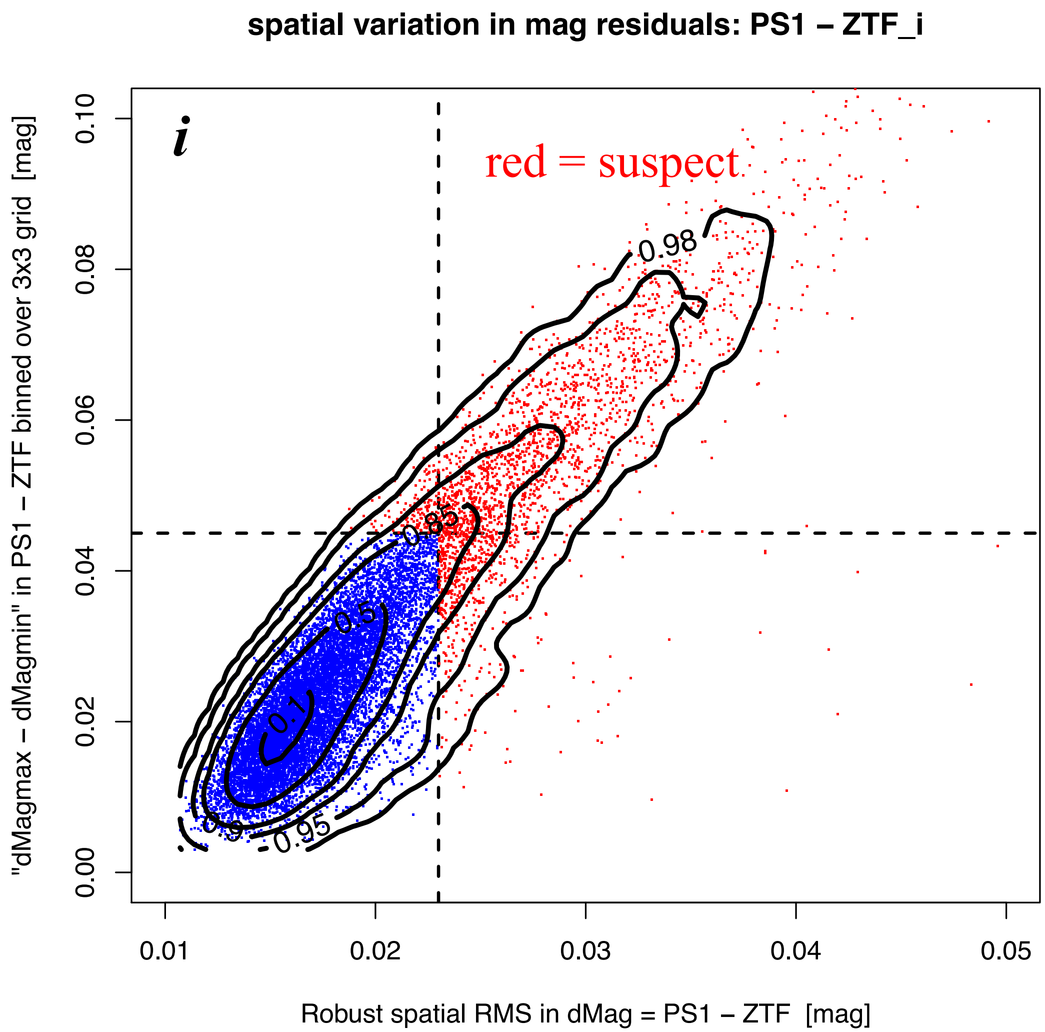
- Total number in *g* : 50,932
- Number suspect : 6,574
- Percentage suspect : ~ 12.9%
- Could be lower since metrics are dependent on confusion level and effective mag range used.

Suspect reference images in *r*



- Total number in *r* : 57,837
- Number suspect : 7,621
- Percentage suspect : ~ 13.2%
- Could be lower since metrics are dependent on confusion level and effective mag range used.

Suspect reference images in *i*



- Total number in *i* : 13,631
- Number suspect : 1,734
- Percentage suspect : ~ 12.7%
- Could be lower since metrics are dependent on confusion level and effective mag range used.

Plan: extend selection criteria for science image inputs

Current criteria, from: **https://zwicky.tf/ykv** (Section 6.7):

(i) Image quality falling in range $1.7 \leq$ FWHM ≤ 5.0 arcsec for the *g* and *R* filters, and $1.7 \leq$ FWHM ≤ 4.5 arcsec for *i* filter.
(ii) Overall quality *status* = 1 where the criteria used to set *status* = 1 (or equivalently, none of the bad `INFOBITS`) are defined in Section 10.4.
(iii) $25.3 \leq$ MAGZP(*g*) ≤ 26.5 or $25.3 \leq$ MAGZP(*R*) ≤ 26.5 or $25.25 \leq$ MAGZP(*i*) ≤ 25.85 for filters *g, R, i* respectively.
(iv) $-0.20 \leq$ CLRCOEFF(*g*) ≤ 0.15 or $-0.05 \leq$ CLRCOEFF(*R*) ≤ 0.22 or $0.05 \leq$ CLRCOEFF(*i*) ≤ 0.30 for filters *g, R, i* respectively.
(v) MAGLIM(*g*) ≥ 19.0 or MAGLIM(*R*) ≥ 19.0 or MAGLIM(*i*) ≥ 18.0 for filters *g, R, i* respectively.
(vi) Global pixel median: *gmedian*(*g*) ≤ 1900 DN or *gmedian*(*R*) ≤ 1600 DN or *gmedian*(*i*) ≤ 1200 DN for filters *g, R, i* respectively.
(vii) Global robust pixel RMS: *gpctdif*(*g*) ≤ 100 DN or *gpctdif*(*R*) ≤ 100 DN or *gpctdif*(*i*) ≤ 80 DN for filters *g, R, i* respectively.
(viii) All science exposures acquired on or after UT night-date February 5, 2018. This is when the camera was reinstalled on the telescope.
(ix) A minimum of 15 overlapping science images satisfying (i) to (viii).
(x) Following criteria **(i)** to **(ix)**, the resulting science image list is sorted in order of *increasing* FWHM after which the first N_{max} images are retained. N_{max} therefore defines the desired depth. Currently, $N_{max} = 40$.

+

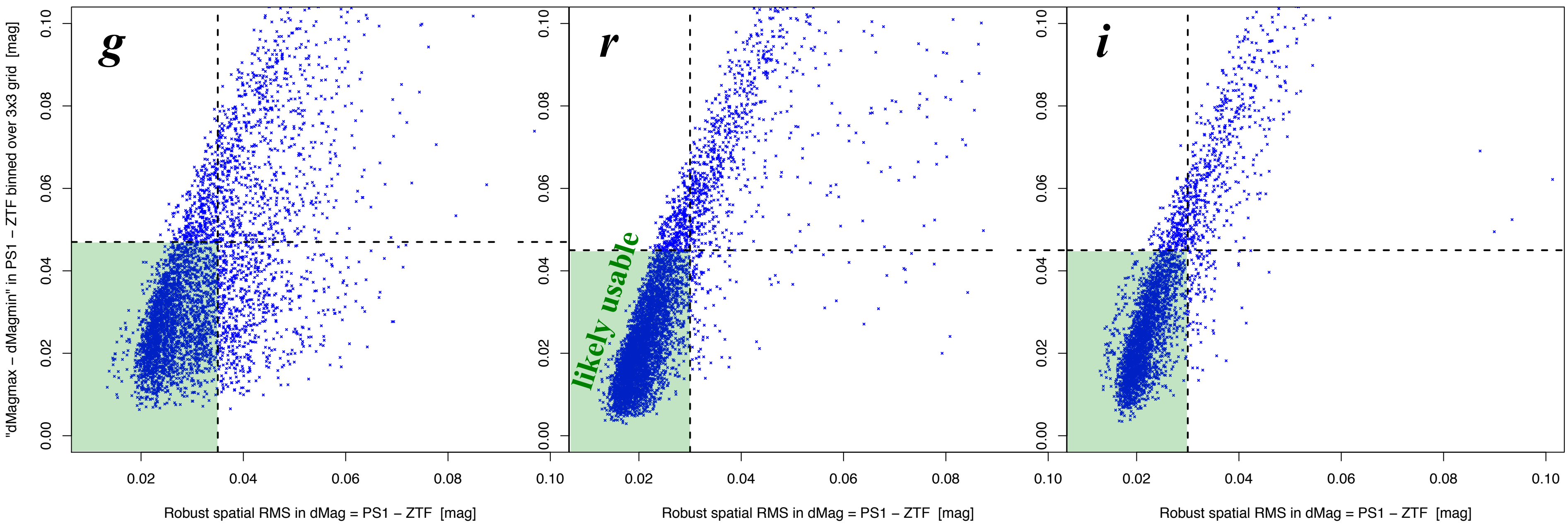
Flatness Criterion

Spatial distribution in photometric throughput in a science image using residuals w.r.t. PS1 catalog over a grid is < some threshold.

This will also filter images with significant spatial variations from varying atmospheric transparency.

Setting "throughput-flatness" thresholds
for science images

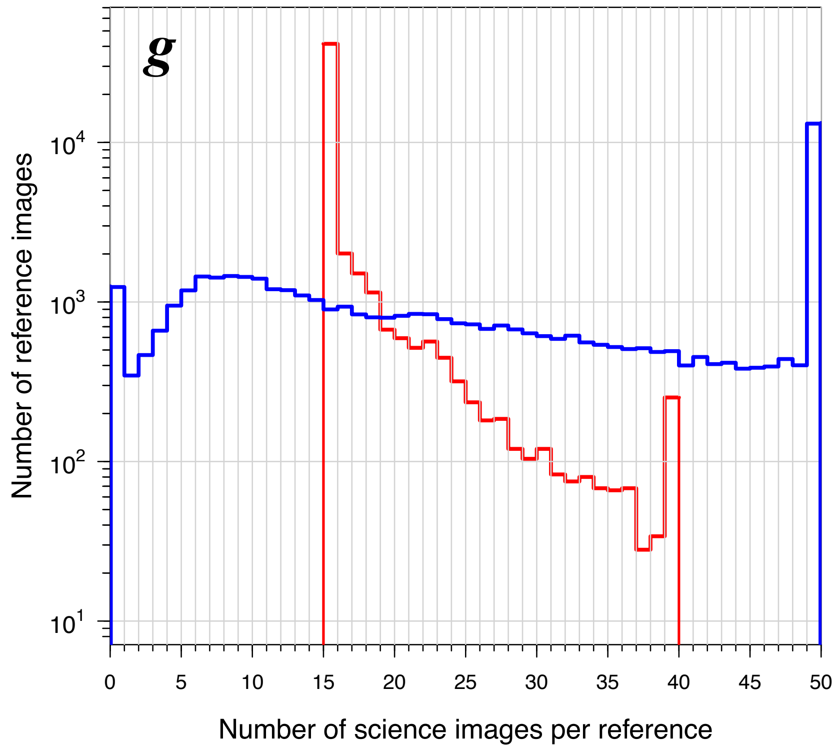
- Used same metrics as before (slide 11: Δ versus *rms*) but this time compute for a random sample of science images.
- **Goal:** explore impact on reference image statistics if impose a flatness criterion when selecting input images.



- Provisional (experimental) thresholds to select usable science images for reference image generation:
  - ***g*** : $rms \leq 0.035$; $\Delta_{minmax} \leq 0.047$
  - ***r*** : $rms \leq 0.030$; $\Delta_{minmax} \leq 0.045$
  - ***i*** : $rms \leq 0.030$; $\Delta_{minmax} \leq 0.045$
- In reality, I expect these thresholds to be field dependent (e.g., high source confusion will impact metrics).

Reference image statistics in *g* using new criterion

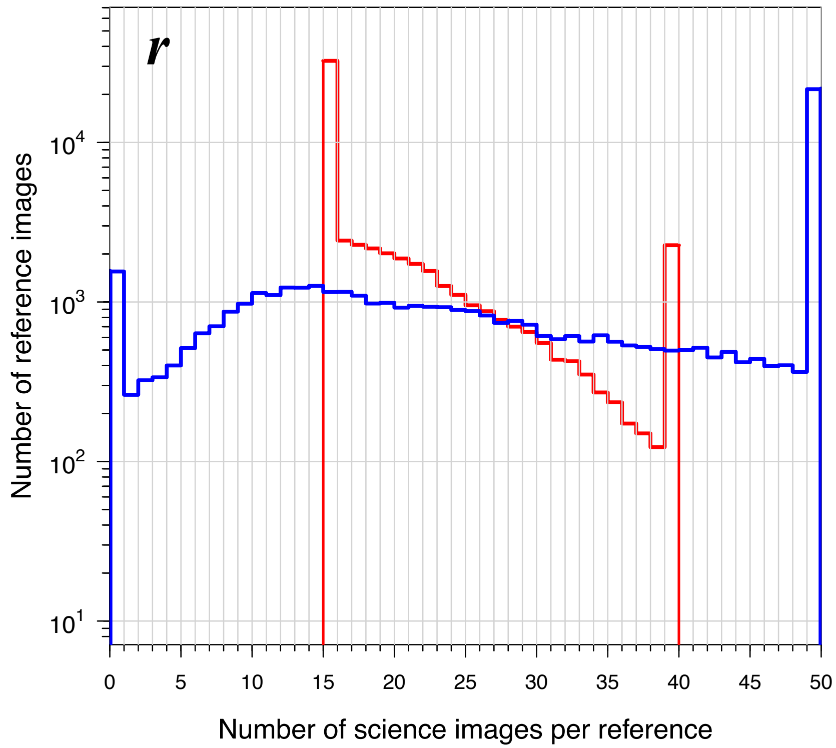
- **Red histogram:** what we have now in archive
- **Blue histogram:** what we'll get if all references were to be regenerated with *flatness* criterion included



***g*-filter:**
Nrefsnow (Nmin=15; Nmax=40) : 50,932
Nrefsnew (Nmin=15; Nmax=50) : 35,297; %lost ~ 30.7%
Nrefsnew (Nmin=10; Nmax=50) : 41,680; %lost ~ 18.2%

Reference image statistics in *r* using new criterion

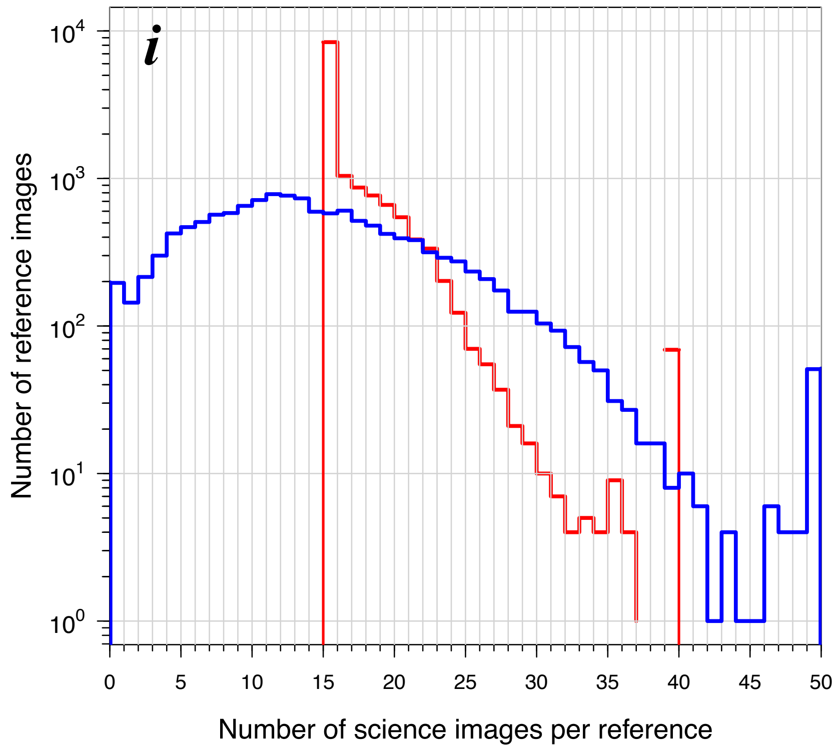
- **Red histogram:** what we have now in archive
- **Blue histogram:** what we'll get if all references were to be regenerated with *flatness* criterion included



***r*-filter:**
Nrefsnow (Nmin=15; Nmax=40) : 57,837
Nrefsnew (Nmin=15; Nmax=50) : 46,516; %lost ~ 19.6%
Nrefsnew (Nmin=10; Nmax=50) : 52,217; %lost ~ 9.7%

Reference image statistics in *i* using new criterion

- **Red histogram:** what we have now in archive
- **Blue histogram:** what we'll get if all references were to be regenerated with *flatness* criterion included



***i*-filter:**
Nrefsnow (Nmin=15; Nmax=40) : 13,631
Nrefsnew (Nmin=15; Nmax=50) : 6,420; %lost ~ 52.9%
Nrefsnew (Nmin=10; Nmax=50) : 10,149; %lost ~ 25.5%

Consequences of regenerating references

- There will be losses in reference-image sky coverage if *flatness* criterion is included.
  - can retune/relax other input filters to minimize losses.

- Lightcurves derived from differential photometry in alert packets will change depending on input timespans and level of contamination from inadvertent inclusion of real transient signal in ref image.
  - lost/irrecoverable alerts, particularly near thresholds.
  - changes in the positions of already published alerts, not only photometry.

- Source positions in reference image catalogs will change – used to seed source-matching across epochs for generating lightcurves (*source matchfile* products, **not** subtraction image photometry).
  - breaks the "appending model" when updating lightcurves at Cahill. Need to re-match (do once) and re-assign new objectIDs in databases.
  - lost sources in reference image by virtue of *"transient behavior"* over time (not reoccurring variables) => lost lightcurves.

- Changes in reference image quality => retraining of machine-learned classifiers for point-source transients and streaks (asteroids) detected in subtraction images. Difficult to quantify.

Regenerating references: moving forward

Goal: maximize reference image quality but also minimize loss in sky coverage.

Possible direction:

1. Retune input science image selection criteria (explore field dependencies / source confusion).
2. Identify & regenerate suspect references (<~ 13% per filter) with *Nmin* = 15, *Nmax* = 50 images deep
   - ➢ if have $N < 15$ images, flag existing reference in archive as "potentially updatable" – check these daily as survey proceeds and regenerate as soon as $N \geq 15$.
3. Regenerate non-suspect references only if new selection criteria yield deeper references.

Special case:

- *i*-filter – makes sense to deploy fringe corrector; reprocess all science images and re-archive; then regenerate all reference images using new criteria.
  - ➢ we can indeed support reprocessing of all *i*-filter image data at this time.

- "Re-baselining" the survey to a new reference image library makes sense in the long term due to intermittent updates to the observing system and calibrations:
  - ➢ camera/cryostat cleansing; new CCD waveforms; new electronic gains/linearity curves; focal-plane leveling; DIQ refinements from flexure correction model updates, …

Second Public Data Release

- DR2 is scheduled for December 11, 2019.

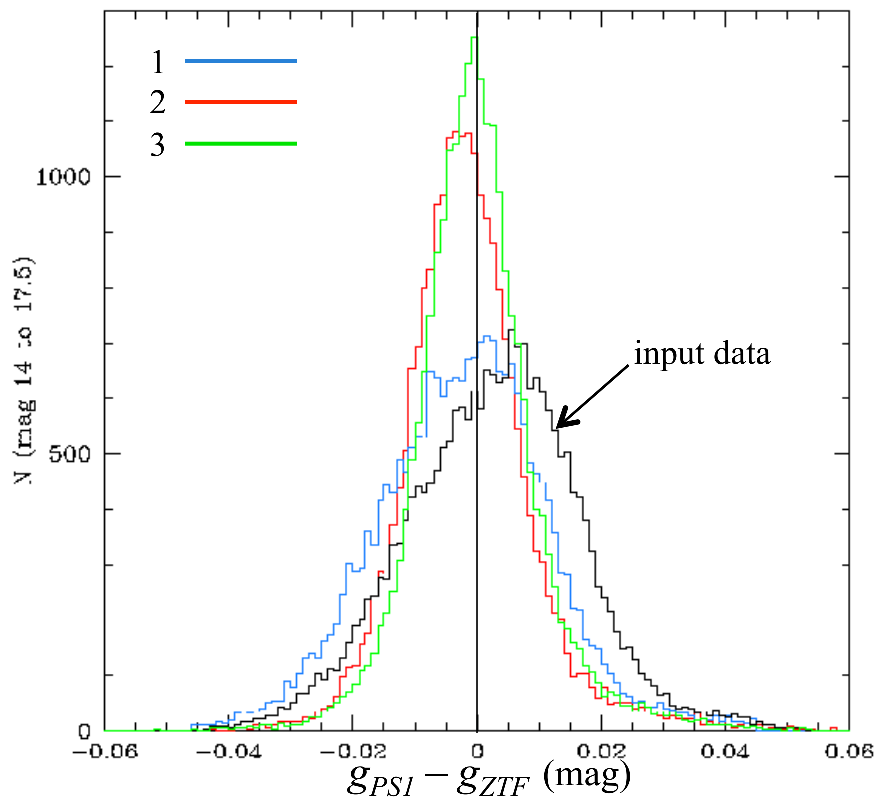
- **Content (adds onto DR1):**
  - program ID 1 (MSIP) epochs: Jan 1 – Jun 30, 2019 (*g* and *r*)
  - program IDs 2 & 3 (partnership and Caltech-time) epochs: Mar 17 – Jun 30, 2018 (*g*, *r*, and *i*)

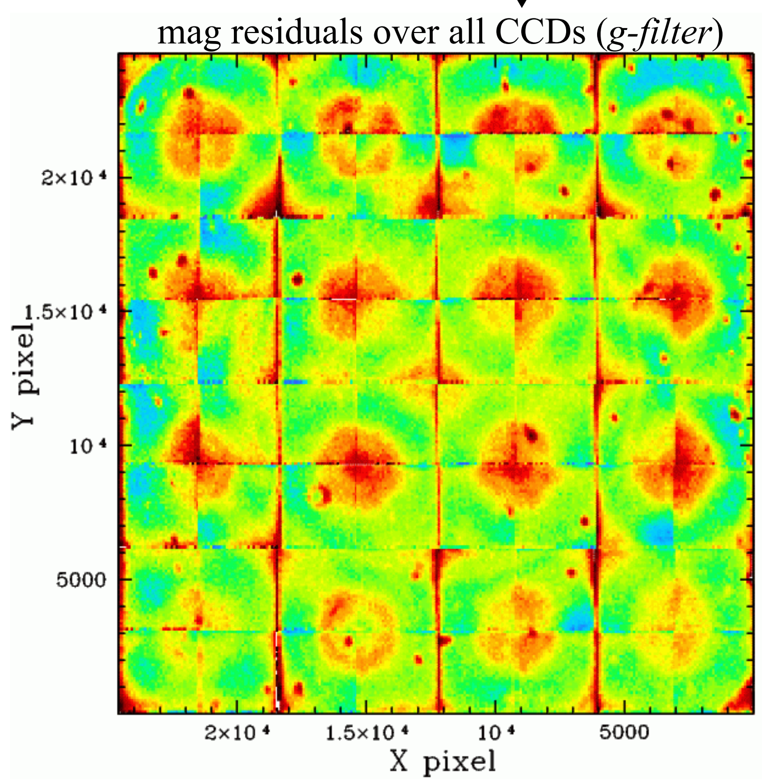
- **Release products (same as DR1):**
  - raw CCD image files & calibration image files.
  - epochal instrumentally calibrated science images, file-based catalogs, and ancillary products
  - reference images, accompanying file-based catalogs, and ancillary products
  - object source database (reference-image drawn) to facilitate lightcurve queries
  - lightcurves derived from matched epochal PSF-fit photometry – including tarballs for bulk download.

- **Improvements:**
  - new columns for lightcurve DB: *nobsrel*, *ngoodobsrel* (#epochs covering DR1 + DR2 only; not everything).
  - refinements to *catflags* (quality flags) column (inclusion of masked-pixel information).
  - corrections to (*matchfile*) lightcurve photometry (see next slide).

Corrections to epochal PSF-fit photometry

- Systematics characterized by Andrew Drake (*http://nesssi.cacr.caltech.edu/ZTF/Web/Calib.html*):
  1. magnitude-dependent biases relative to PS1: < 1.3% for *g*, *r* < 17.5.
  2. biases from position-dependent responsivity (flatfield) residuals: ± 3% (max at edges)
  3. global field-dependent biases (sky-location dependent): <~ 0.5%
- **Net result:** RMS in residuals relative to PS1 now typically <~ 7 millimag.





Miscellaneous: in progress & TBD

- Preparations for Second Public Data Release.
- **TBD:** continued reference image quality analysis with possible regeneration of a subset.
- **TBD:** deployment of *i*-filter fringe correction and reprocessing of all *i*-filter data, including references.
- New "Kafka topic" to identify ProgramID=3 alerts in TESS footprints for distribution to UW & public brokers.
- Moving-object (SSO) pipeline updates to support twilight survey and more efficient scanning of tracklets.
- Update naming of SSOs from pipeline to conform to MPC's new extended packed-naming format.
- Continue to support generation of lightcurve tarballs for TESS sector observations (every month).
- Support ingestion of lightcurve matchfile contents into kowalski – use an "appending model" approach.
- **TBD:** "backfilling" of subtraction images in archive due to late creation of references (currently done ad-hoc).
- **TBD:** bulk reprocessing in general using improved calibrations and methodologies.
- Brainstorming for ZTF Phase II.

ZTF Phase II thoughts

- Extend / improve data-retrieval and analysis tools for community (in context of MSIP deliverables).
- Inevitably, will also be of value to the partnership.

- Store contents of *avro* packets into a database to facilitate easier access / faster querying (not a broker!)
  - include info. from other databases (e.g., NED – NASA/IPAC Extragalactic Database).
  - extend photometric histories.
- Adopt a more efficient and adaptable datastore for distributing lightcurve data instead of matchfiles.
  - can be used directly with parallel processing frameworks that also leverage GPUs (large queries).
  - also optimized for cloud computing.
- Integration of archive with science-user workspaces and analysis platforms (LSST-like model).
  - custom co-addition & mosaicking, integrated with forced photometry.
- More frequent public data releases.
- Your thoughts?

Back up slides

Reminder on documentation

- **ZSDS Explanatory Supplement** (linked from ZTF public website under):
  https://www.ztf.caltech.edu/page/technical#science-data-system

- **Science Data System paper**:
  https://iopscience.iop.org/article/10.1088/1538-3873/aae8ac

- **Archive access and services:**
  https://irsa.ipac.caltech.edu/Missions/ztf.html

- **Public alert archive and usage:**
  https://ztf.uw.edu/alerts/public/

- **First Public Data release:**
  https://www.ztf.caltech.edu/page/dr1

Baseline deliverables / data access portals

1. **Instrumentally calibrated, epochal image products, bit-masks, source catalogs, PSFs, and difference images**
   Archive (IRSA)

2. **Raw image data and image calibration products used in pipelines**
   Archive (IRSA)

3. **Reference images (co-adds) from combining (1):** coverage maps, uncertainty maps, and source catalogs
   Archive (IRSA)

4. **Alert (point-source event) stream** from real-time image-differencing pipeline: packetized with metadata
   Marshal(s); Public Brokers; Archived in IRSA

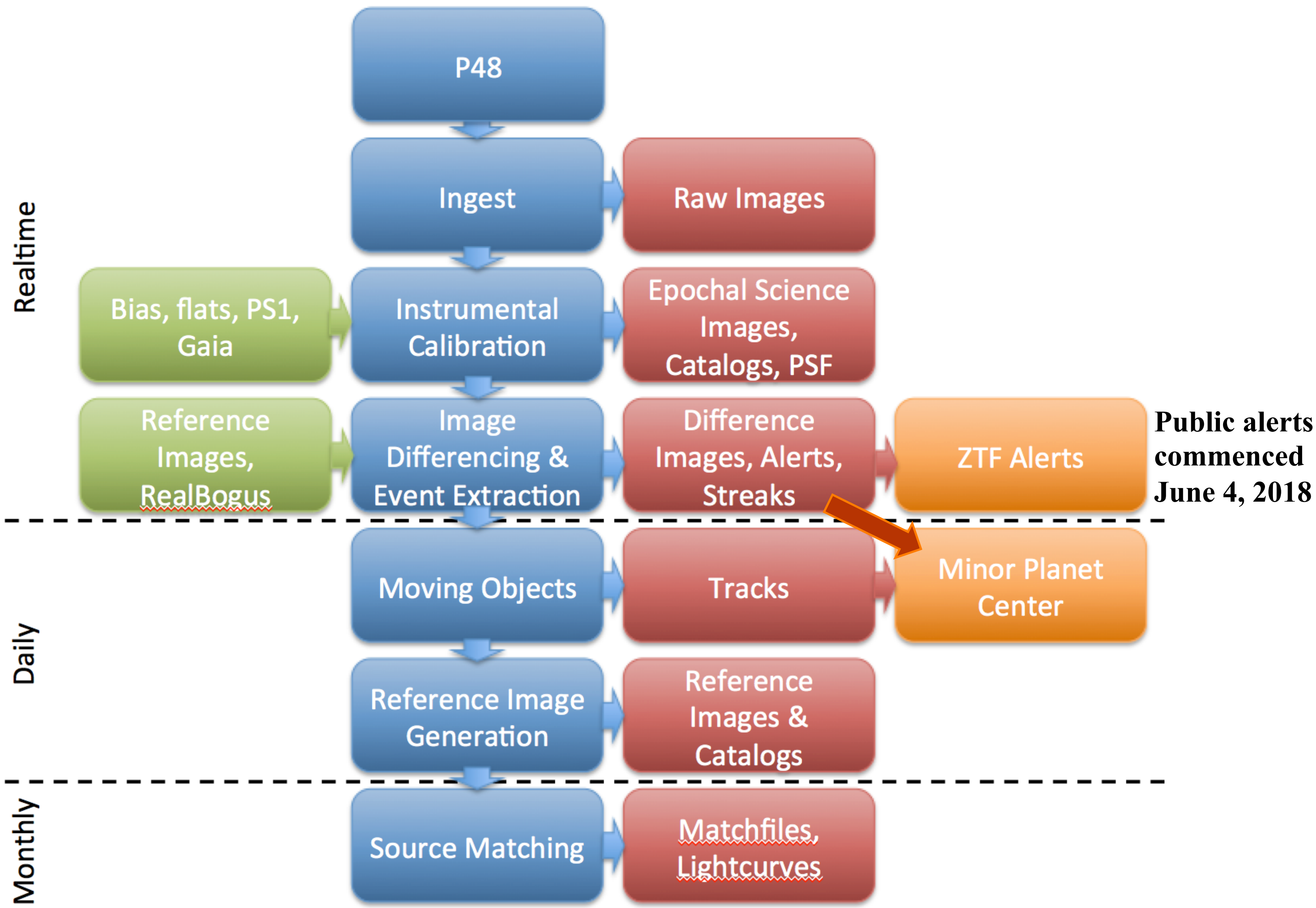
5. **Products to support real-time Solar System / NEO discovery and characterization:** both streaks and tracks
   ZTF-Depot (internal) and IAU-Minor Planet Center

6. **Lightcurves & metrics from matching sources across individual epochs using (1) to beginning of survey**
   Archive (IRSA); ZTF-Depot (raw matchfiles)

7. **Quality assurance metrics, summary statistics, and survey coverage maps:** for performance monitoring
   ZTF-Depot (internal)

8. **Documentation:** cautionary notes, recipes, and tutorials on data-retrieval and analysis
   Explanatory Supplement on ZTF Public Website; PASP paper published in Dec 2018

Pipeline summary: timeline view



Data volumes & Statistics
Mar 17, 2018 (survey start) – Mar 6, 2019

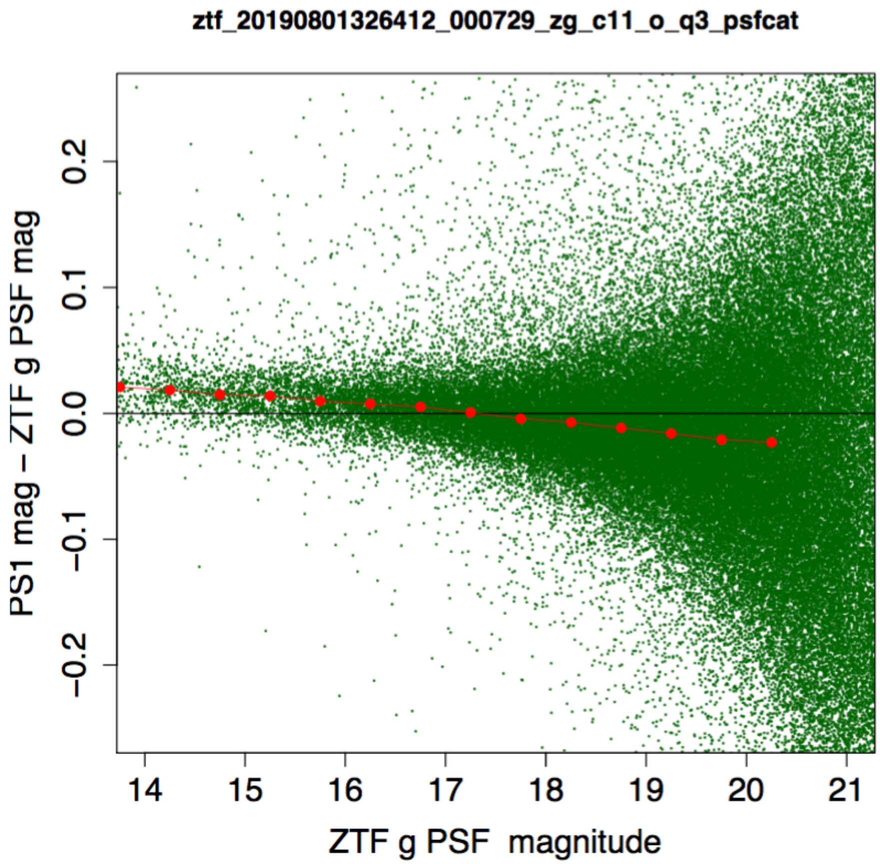
Exposure/Image Metric	*g*	*r*	*i*
Raw on-sky exposures	67,781	103,366	5,510
Survey-ready quadrant-based reference images (#quadrants N≥15 visits)	45,087 (47,934)	53,392 (56,259)	11,127 (13,193)
Lightcurve matchfiles (last made Dec. 15, 2018)	40,822	51,076	11,018
Epochal science image products archived (all CCD quadrants)	~ 10.6 Million (788.6 TB)		

Source Extraction Metric	Number
Epochal science image PSF-fit extractions	183 B
Epochal science image aperture-based extractions	113 B
Reference image PSF-fit extractions ("seeds" for lightcurves)	4.5 B
Reference image aperture-based extractions	1.5 B

Event Extraction Metric	Number
Raw candidate events from all difference images (+ and – diffs)	+ 274 M – 136 M
Alert packets generated from all difference images (+ and – diffs)	+ 58 M – 29 M
Alert packets associated with known solar system objects (≤ 3 arcsec)	2.6 M
Streaked detections from new SSOs	30
Streaked detections from known SSOs	> 12 K
Moving object tracklets not associated with known SSOs & delivered to MPC	~ 5 K
Moving object tracklets associated with known SSOs & delivered to the MPC	> 850 K

Corrections to epochal PSF-fit photometry

BEFORE



AFTER