Comparing and Automatically Optimizing the Performance of Systematic Error Correctors for TESS Light Curves

David Rapetti
Universities Space Research Association, NASA Ames Research Center

In collaboration with:
Jon Jenkins
NASA Ames Research Center
Joseph Twicken, Douglas Caldwell
SETI Institute, NASA Ames Research Center

INTRODUCTION AND MOTIVATION

- Accurate and precise removal of systematic errors from TESS light curves (lcs) is critical for exoplanet, stellar, and asteroseismology studies
- Various approaches exist to correct for systematics while preserving intrinsic stellar variability (Luger et al. 2016, 2018; Hedges et al. 2021; Smith et al. 2012; Stumpe et al. 2014; Aigrain et al. 2017)
- However, no comprehensive analyses have been carried out to properly compare these approaches and determine their validity on different target types in both an individual and statistical manner
- Current correctors have been usually demonstrated on their own and on limited sets of hand-picked targets
- To apply these corrections more generally, it is important to compare multiple correctors on larger samples
- We are particularly interested in the ability to remove scattered light contamination from the Earth and the Moon, which is a key systematic for TESS

TOOLKIT FOR COMPARING SYSTEMATIC ERROR CORRECTING ALGORITHMS

- We present the code SysCoCoPy: Systematic Corrector Comparisons in Python (Rapetti et al., in prep.) which:
  - is designed to optimize and compare corrector performances
  - is aimed for public release in conjunction with the publication to allow reproduction and further analyses
  - builds upon the popular Python software package Lightkurve (Lightkurve Collaboration et al. 2018)
  - allows the user to select case (target and sector) samples on which to apply the corrections
  - includes products in the form of figures and metrics
  - implements an automatic corrector parameter optimization

- Our current correctors are: RCQ (Regression Corrector with Quaternions), PLD (Pixel Level Decorrelation), and CBV (Cotrending Basis Vectors). We also compare them with PDC (Presearch Data Conditioning) lcs from SPOC

- Our first goal is to examine the ability of RCQ and PLD to effectively remove scattered light features in the presence of instrumental systematics and a range of stellar properties and variability

- An initial statistical analysis of 826 cases selected based on ranges of stellar properties yields meaningful metric distributions for each corrector

CURRENT CORRECTORS: RCQ, PLD, CBV

RCQ (Regression Corrector with Quaternions): This is an adaptation of the correction method in Hedges et al. (2021). We use the RegressionCorrector class of Lightkurve and employ:

i. A spline polynomial fit aimed at accounting for stellar variability
ii. Principal Component Analysis (PCA) eigenmodes to model the sky background light
iii. The mean, standard deviation, and skewness of the spacecraft quaternions to describe pointing jitter and mechanical momentum dumps

PLD (Pixel Level Decorrelation): This is an adoption of the PLD technique (Deming et al. 2015; Luger et al. 2016, 2018) as implemented in the PLDCorrector class of Lightkurve and utilizes:

- The same (i) and (ii) elements in RCQ (parameter values might differ) for stellar variability and background light
- PLD for pointing jitter and mechanical effects

CBV (Cotrending Basis Vector): This corrector utilizes the CBV technique of SPOC's PDC method as adapted in the CBVCorrector class of Lightkurve. (We added a spline polynomial fit such as those above, but current testing indicates degeneracies between the spline and CBV parameters advising us to not use the spline at this time.)

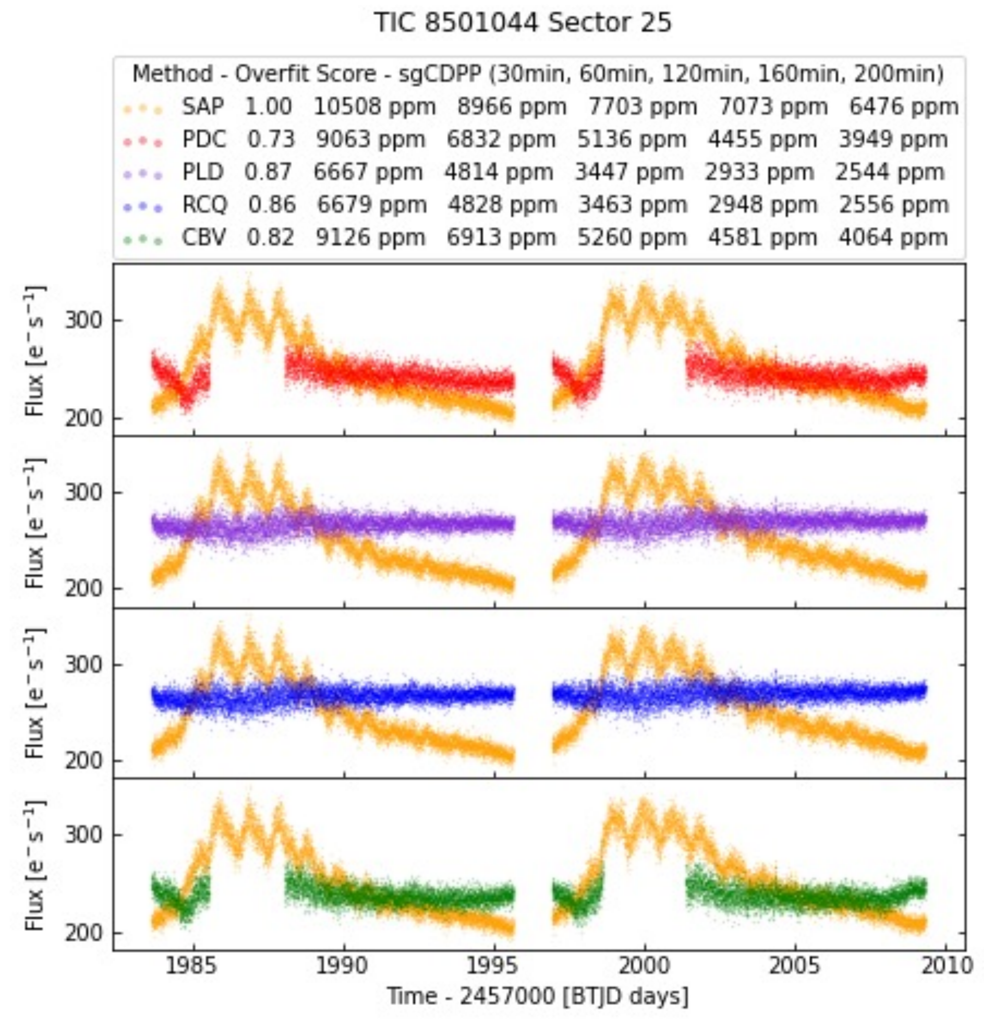
AUTOMATIC OPTIMIZATION OF CORRECTOR PARAMETERS

- For PLD, RCQ, and CBV, we optimize the parameters and their values shown in the table
- We selected them by testing a 24-cases sample chosen to have pathological scattered-light, as found by SPOC
- Besides those below, we use default Lightkurve parameter values, with PLD order of 1 for TESS, and propagate errors
- For each case and corrector, we minimize a goodness-of-fit metric to select the best set of parameter values from the table
- The current metric is the harmonic mean (HM) of the various duration sgCDPPs shown in the next figures. This HM metric was selected to balance the post-correction photometric scatter at different time scales while penalizing outliers

Parameter	PLD	RCQ	CBV
Number of spline knots	(6, 12, 24, 48)	(6, 12, 24, 48)	(0)
PLD aperture mask	(pipeline, threshold)	N/A	N/A
Number of PCA terms	(8, 16, 32)	(3, 6, 9)	N/A
Added background	True	True	False
CBV type	N/A	N/A	(MultiScale.1-2-3/Spike, SingleScale/Spike)
CBV indices	N/A	N/A	(1-8/all, 1-8/all)
CBV α regularization	N/A	N/A	(0.00001, 0.001, 1)

Rapetti, Jenkins, Twicken, Caldwell & Smith (in prep.)

PRELIMINARY RESULTS: COMPARING SYSTEMATIC ERROR CORRECTORS



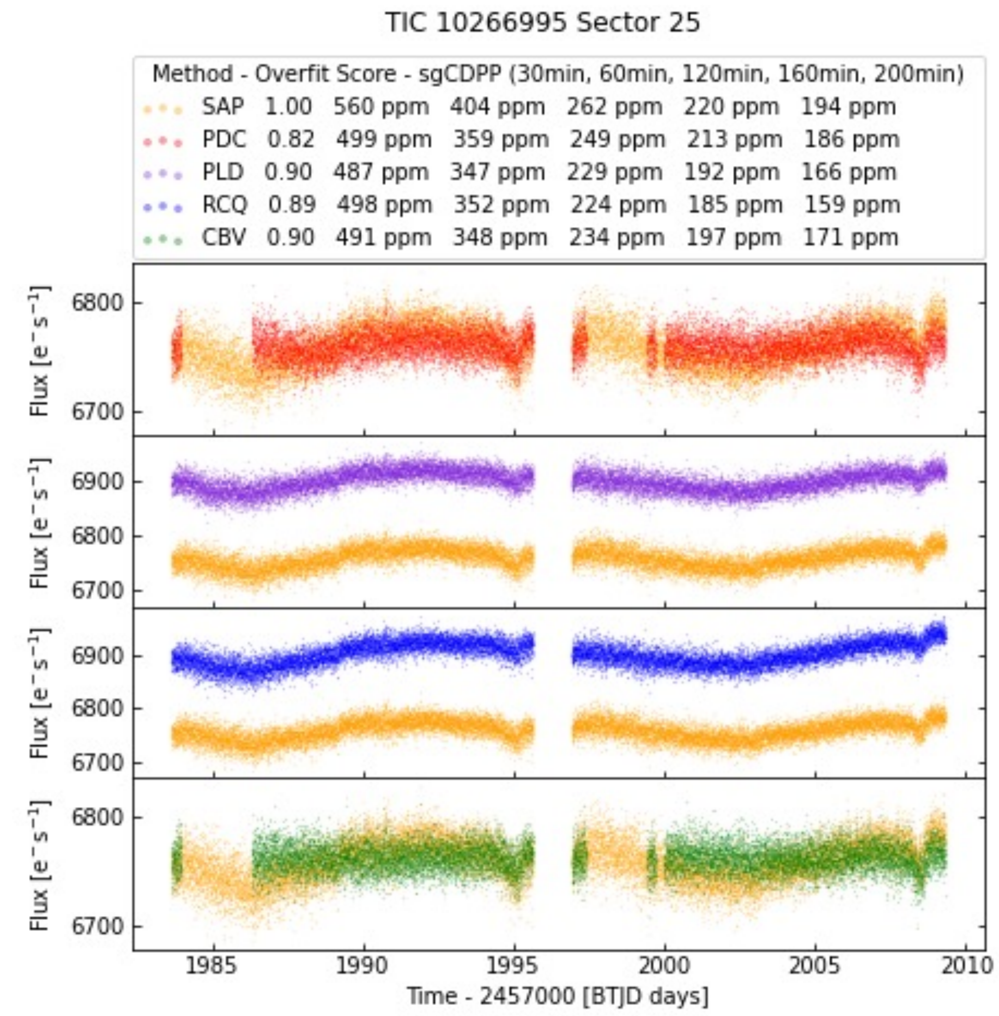
Rapetti, Jenkins, Twicken, Caldwell & Smith (in prep.)

- This a product of SysCoCoPy. Using our automatic corrector parameter optimization, this is an example of a correctors comparison between PLD, RCQ, CBV, and PDC, with the SPOC uncorrected SAP light curve optionally shown for reference

- PLD and RCQ can successfully recover significant portions of light curves excluded by PDC due to scattered light contamination, as shown here for a severe case from the pathological sample used to calibrate the optimization

- To estimate the remaining scatter after corrections, we employ a Lightkurve method that uses a a Savitsky-Golay (sg) high-pass filter as a Combined Differential Photometric Precision (CDPP) proxy, with a window length 5 times the maximum timescale, for durations: 30, 60, 120, 160, and 200 minutes

PRELIMINARY RESULTS: COMPARING SYSTEMATIC ERROR CORRECTORS



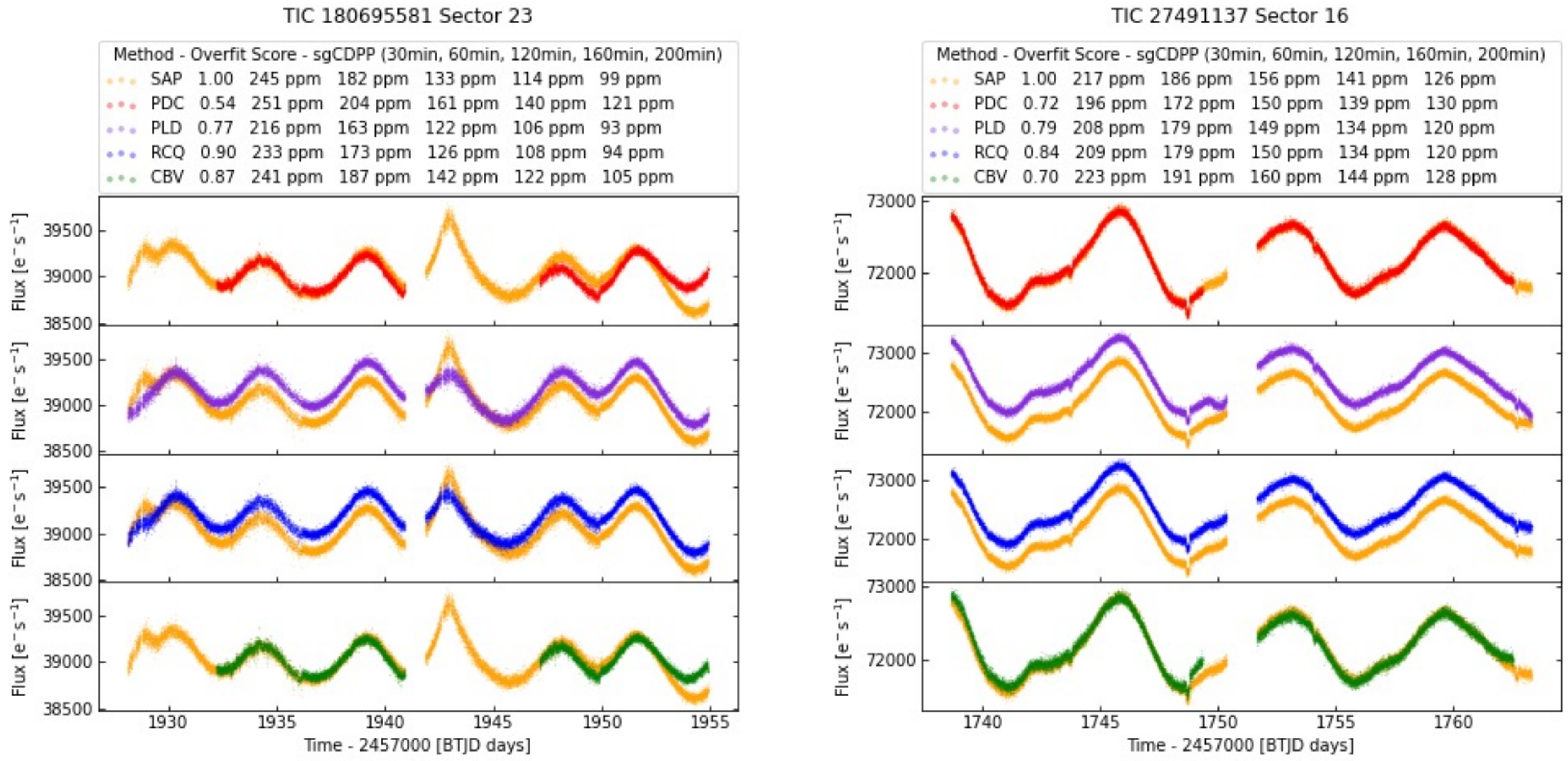
Rapetti, Jenkins, Twicken, Caldwell & Smith (in prep.)

- This a product of SysCoCoPy. Using our automatic corrector parameter optimization, this is an example of a correctors comparison between PLD, RCQ, CBV, and PDC, with the SPOC uncorrected SAP light curve optionally shown for reference

- PLD and RCQ can successfully recover significant portions of light curves excluded by PDC due to scattered light contamination, as shown here for a severe case from the pathological sample used to calibrate the optimization

- To estimate the remaining scatter after corrections, we employ a Lightkurve method that uses a a Savitsky-Golay (sg) high-pass filter as a Combined Differential Photometric Precision (CDPP) proxy, with a window length 5 times the maximum timescale, for durations: 30, 60, 120, 160, and 200 minutes

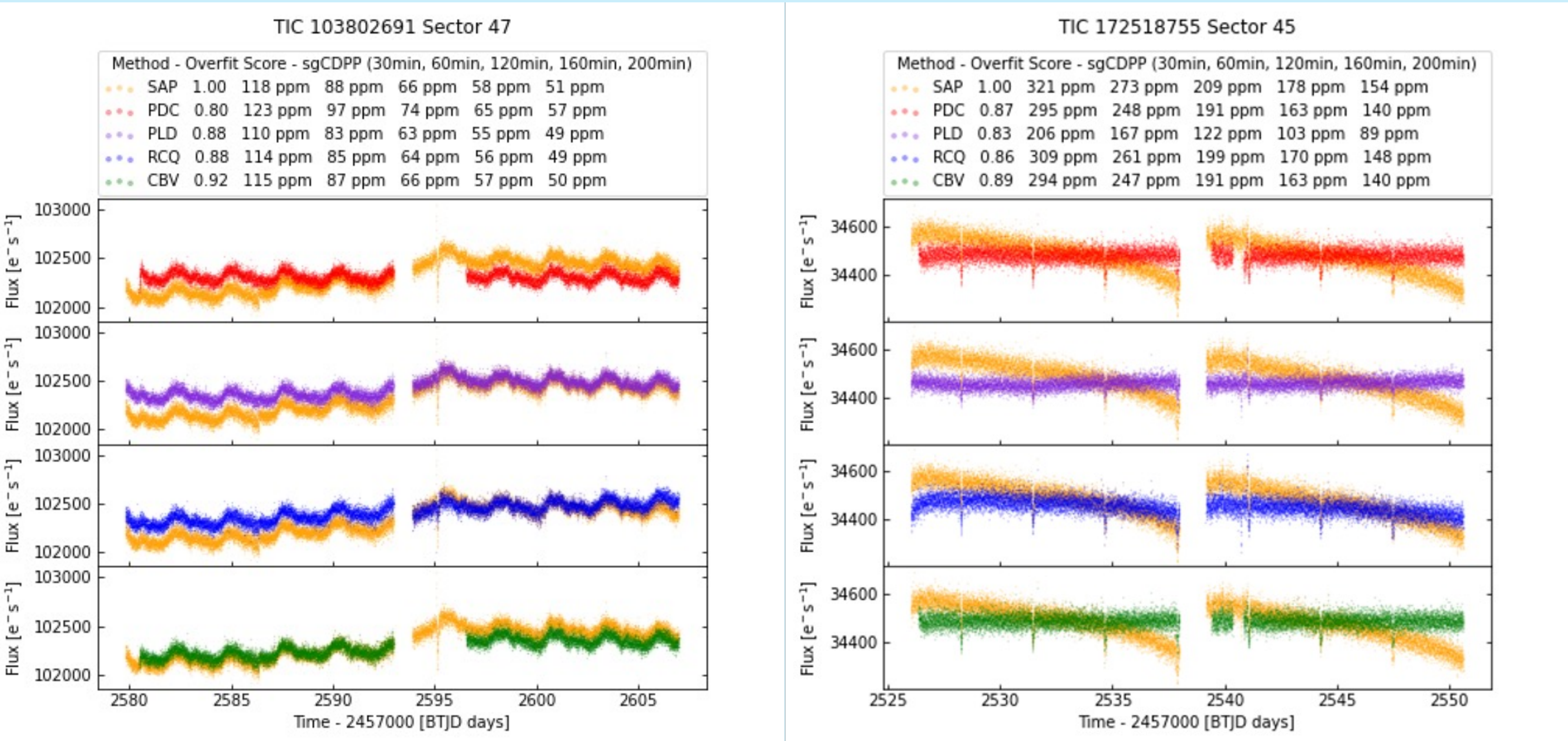
PRELIMINARY RESULTS: COMPARING SYSTEMATIC ERROR CORRECTORS



Rapetti, Jenkins, Twicken, Caldwell & Smith (in prep.)

- Two cases from another sample with 76 cases selected to include a range of stellar variabilities
- Left: TOI-1807, exhibiting a light curve with large gaps in PDC and CBV at the start of the orbits recovered by PLD and RCQ
- Right: TOI-2076, a rotationally variable star with 3 confirmed planets, showing promise for improving transit modeling

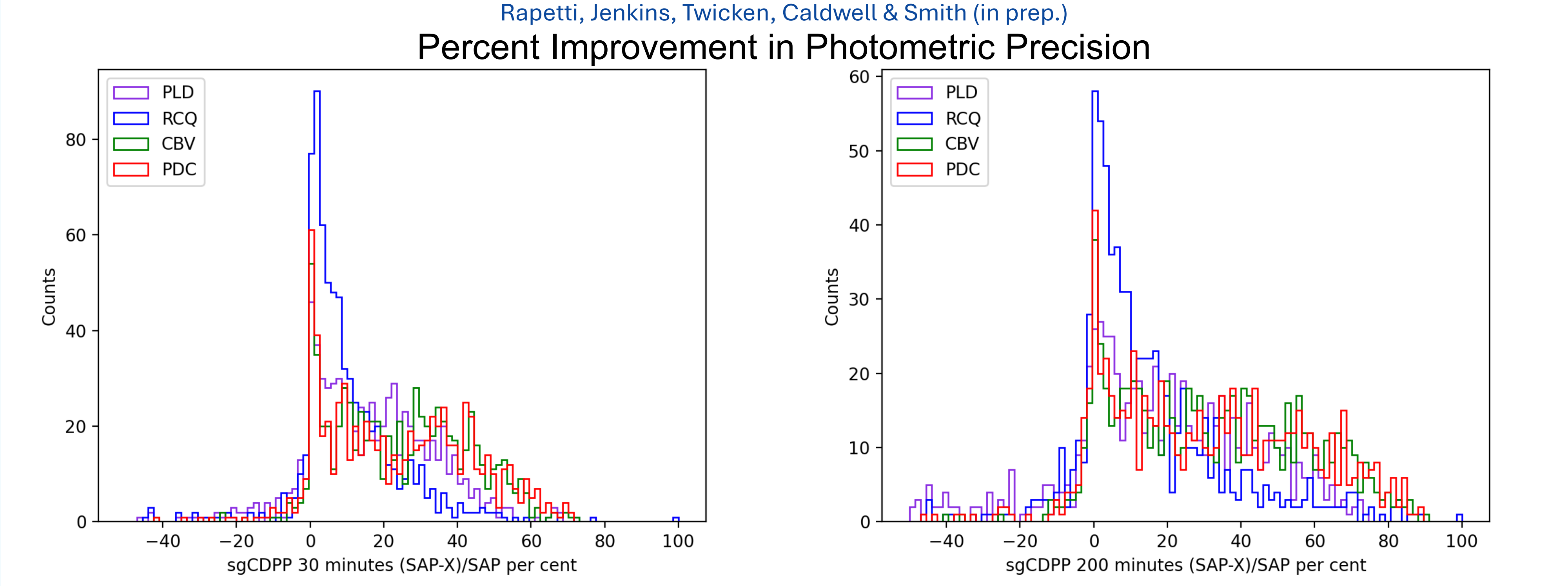
PRELIMINARY RESULTS: COMPARING SYSTEMATIC ERROR CORRECTORS



Rapetti, Jenkins, Twicken, Caldwell & Smith (in prep.)

- Two more cases from the sample with 76 cases including varied stellar variabilities
- Left: Additional example displaying potential data recovery by at least PLD in a large gap at the start of the second orbit in PDC and CBV
- Right: TOI-1789, with the transit at the start of the second orbit being better detected by PLD and RCQ

PRELIMINARY RESULTS: STATISTICAL DISTRIBUTIONS OF METRICS FOR EACH CORRECTOR

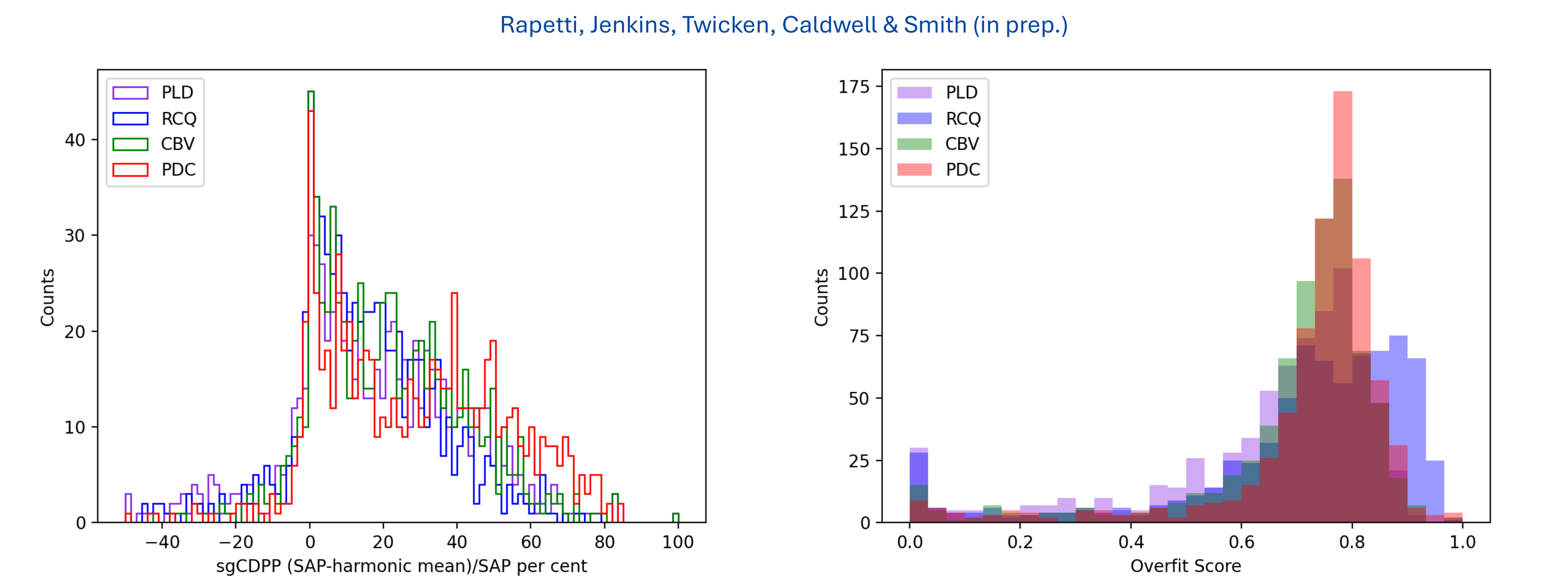


Our currently largest sample of 826 was selected to be more representative of the typical TESS target

Sample ranges:
2717 < Teff < 10000 K
4 < log g < 5
7.98 < T_{mag} < 8

- These are other products of SysCoCoPy that rather than individual show statistical comparisons
- We study the change of sgCDPPs in corrected lcs with respect to SAP lcs and find:
  - A peak around 0% with a broad positive tail and a sharp, long low-level negative tail
  - RCQ has the largest peak around 0% and the smallest positive tail, followed by PLD
  - CBV and PDC present similar relatively bi-modal shapes with a second peak around 30-40% improvement
  - The distributions widen for 200 minutes in both directions

PRELIMINARY RESULTS: STATISTICAL DISTRIBUTIONS OF METRICS FOR EACH CORRECTOR



Our currently largest sample of 826 was selected to be more representative of the typical TESS target

Sample ranges:
2717 < Teff < 10000 K
4 < log g < 5
7.98 < T_{mag} < 8

- These are other products of SysCoCoPy that rather than individual show statistical comparisons
- Left: sgCDPP change for the HM metric. Right: Over-fitting metric distributions. We find that:
  - In the left, RCQ does not present the highest peak anymore. PLD, RCQ and CBV perform more comparably, and PDC outperforms all of them, with a second mode around 40% improvement
  - The over-fitting scores appear relatively consistent with a peak ~0.8 and a longer tail downwards, except for RCQ with a bimodality centered ~0.8 and a second mode ~0.9

CONCLUSIONS AND BENEFITS OF OUR CODE

- We presented the code SysCoCoPy (Rapetti et al., in prep.) which compares corrector performances on selected samples, provides individual and statistical products, and is to be publicly released

- We implemented an initial automatic corrector parameter optimization based on the minimization of the harmonic mean of sgCDPP at different time scales

- We find that PLD and RCQ can automatically remove scattered light contamination in the presence of instrumental systematics and a range of stellar properties and variability in certain cases

- We applied our automatic PLD optimization to TOI-1835 to support the discovery of a resonant sextuplet (Luque et al., 2023). PDC-discarded cadences due to scattered light were recovered to detect additional transits where predicted by their resonance. For two, this served to confirm the latter with a second transit

- From visual inspection, the present automatic optimization does not avoid certain evident failures for PLD and RCQ that might be resolved either manually (not feasible for large samples) or by amplifying the grid of parameter values. This motivates ongoing and planned extensions

- sgCDPP and over-fitting score distributions somewhat compare between correctors (although catastrophic failures of PLD/RCQ might not be captured by the metrics), but with an overall higher performance of PDC