Draft version July 10, 2023
Typeset using LaTeX **twocolumn** style in AASTeX62

DEx1: The First LSST-MPC Data Exchange Challenge Report

Mario Juric,[1,2] Matthew J. Holman,[3,4] Siegfried Eggl,[1,2] Michael Lackner,[4] Joachim Moeyens,[1,2] Margaret Pan,[3,4] Matthew Payne,[3,4] and Federica Spoto[3,4]

[1] *DIRAC Institute and the Department of Astronomy, University of Washington, Seattle, USA*
[2] *Vera C. Rubin Observatory Construction Project*
[3] *Harvard-Smithsonian Center for Astrophysics, 60 Garden St., MS 51, Cambridge, MA 02138, USA*
[4] *Minor Planet Center*

(Dated: July 10, 2023)

ABSTRACT

This note describes the result of the first data exchange test (DEx1) between the Minor Planet Center and the Rubin Observatory. It covered four key goals: a) assess Rubin's ability to generate valid ADES-formatted submissions, ii) assess MPC's current and expected future ability to ingest LSST-sized submissions, iii) exercise/understand the submission process and iv) establish relationships between the MPC and Rubin teams. To do so, we have simulated the first 17 nights of Rubin Solar System object discoveries, generated ADES files, and submitted them to the MPC. These were (manually) processed by the MPC to both compute orbits for new discoveries, and extend arcs for re-observations of known objects. Based on the simulations used here, we found the LSST is expected to discover approximately 0.5M new objects in the first month of operations. Designations of such objects will have unprecedentedly high cycle counts (e.g., 2023 UX_{5678}), which cannot be written in packed form following the present scheme. The packed provisional designation scheme will therefore have to be updated to accommodate (ideally) $O(1M)$ new discoveries in any half-month (or abandoned). Other than that issue, assuming necessary automation is implemented and further computational resources added, this test found no fundamental obstacles in MPC being able to process the LSST data. Future tests will focus on automation and injecting further realism in the simulated dataset.

Keywords: editorials, notices — miscellaneous — catalogs — surveys

1. INTRODUCTION AND GOALS

Over its 10-yr program, the Rubin Observatory's Legacy Survey of Space and Time (LSST) is expected to discover over 6 million asteroids, with over 500,000 individual detections and measurements taken and processed, in real time, every night. Rubin's goal is to make these observations and discoveries of Solar System objects available to the scientific community with minimum latency and in a maximally useful fashion. This will be accomplished by promptly submitting them to the MPC after successful linking and/or attribution.

To assure readiness of both Rubin Observatory and the Minor Planet Center to exchange data (see Figure 1 for a simplified diagram), we agreed to conduct a series of increasingly complex tests culminating in a full operations-like dress rehearsal sometime in LSST commissioning (not earlier than 2022). This is a brief report summarizing the outcome of the "Data Exchange Text No.1" (DEx1).

1.1. *DEx1 Goals Summary and Objectives*

The main goal of DEx1 was to simulate a dataset expected from the first 2-4 weeks of LSST operations, and exercise the process of sending those data to the MPC, and them being processed by the MPC (to fit orbits to new objects and extend arcs of previously known ones).

The specific objectives of this test were to:

1. Assess Rubin's ability to generate valid ADES-formatted submissions.
2. Assess MPC's current and expected future ability to ingest LSST-sized submissions.
3. Exercise/understand the submission process.
4. Establish relationships with the MPC team.

with one defined stretch goal:

1. Test the capability to re-fit orbits and generate a new orbit catalog

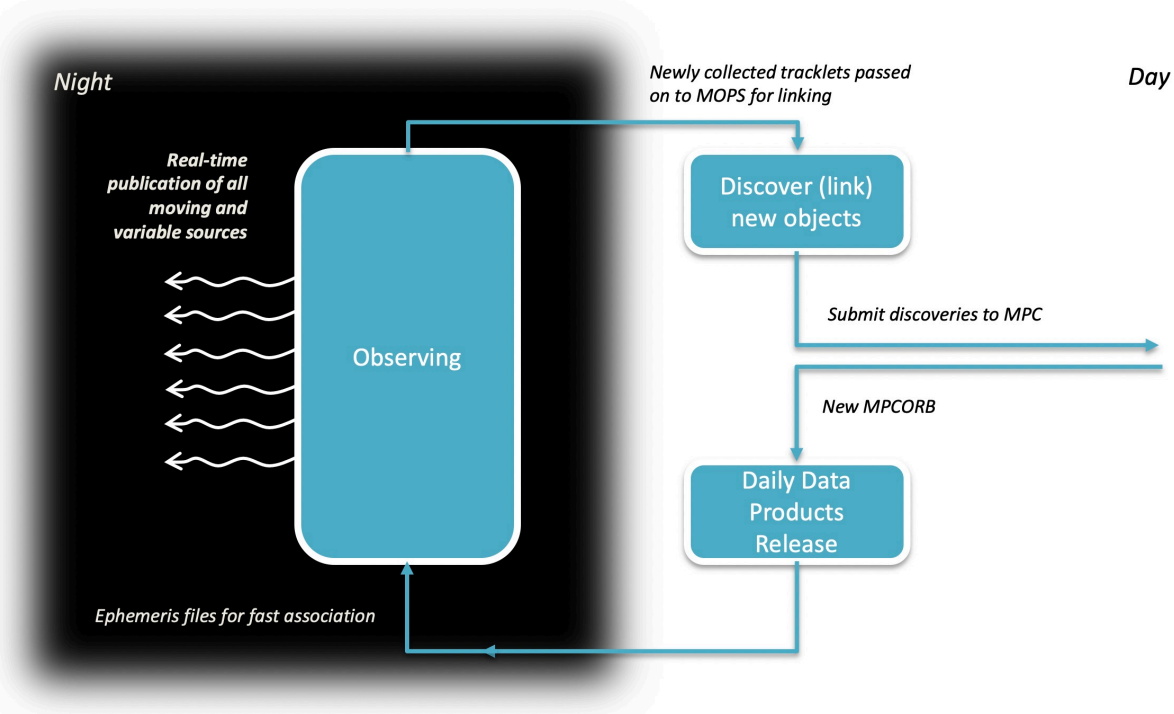

Figure 1. A diagram illustrating the LSST-MPC 24-hr processing loop. The nominal time budgeted for MPC's processing of new data is 4 hours. Consult http://ls.st/ldm-151 for a more detailed breakdown and discussion of individual components.

2. TEST INFRASTRUCTURE, SOFTWARE AND ENVIRONMENT

Datasets: We simulated the first 17 nights of a recently proposed LSST campaign, namely baseline_2snaps_v1.5_10yrs. The aforementioned OpSim database served as input to a modified version of the objectsInField survey simulator which includes the actual LSST footprint, realistic astrometric ($1\sigma = 50$mas) and photometric uncertainties as well as realistic colors for simulated Solar System objects.

The simulated Solar System catalog was comprised of known objects extracted from a current version of the MPCORB catalog (as of September 9, 2020) and supplemented with a faint population ($H > 15$mag) drawn from the Synthetic Solar System Model (S3M, Grav et al. 2011). Density plots of semimajor axis vs. eccentricity of the simulated populations are shown in Figure 4.

Of the simulated Solar System Objects, 888,474 unique objects were observed resulting in 6,931,558 observations over 17 nights (Figure 2. The majority of the observed objects were main-belt asteroids (MBAs) with 6,336,214 observations, followed by roughly 20,000 Jupiter Trojans with 52,232 observations and 2645 near-Earth objects (NEOs) with 16,193 observations.

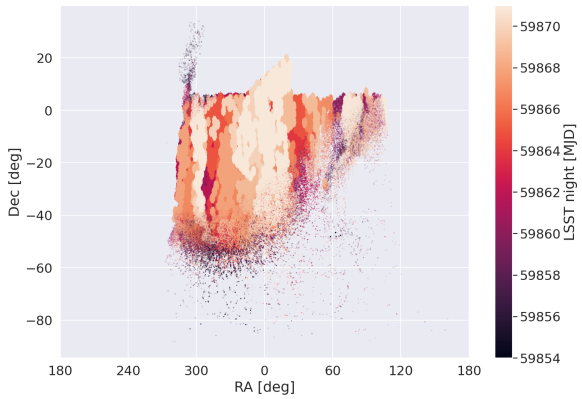

Figure 2. Simulated SSO observations during the first 17 nights of OpSim baseline_2snaps_v1.5_10yrs.

Software (LSST): We have used the development versions of LSST Solar System Processing Pipelines, part of the broader LSST Data Management suite. These were run on local machines at the University of Washington.

Software (MPC): MPC is currently using a new version of the OrbFit code[1], which is not public yet. In particular, the software used for the computation of the orbit of the new object candidates is a new code devel-

[1] http://adams.dm.unipi.it/orbfit/

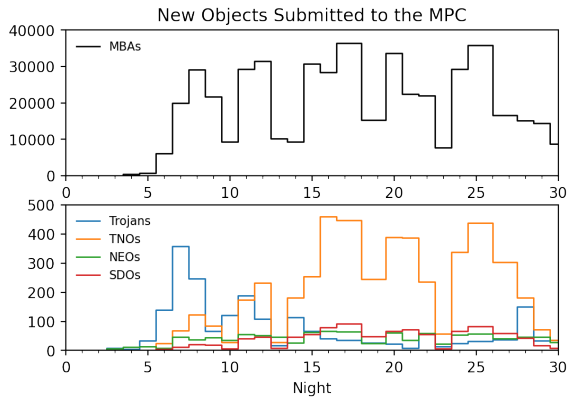

Figure 3. A figure showing the number of (mock) newly discovered objects in the first month of a simulated LSST survey. No objects are discovered in the first few nights, as a consequence of LSST's linking algorithm requiring at least three nights of data for identification of a reliable candidate. The subsequent modulation of the discovery rate is due to variations in the coverage of the ecliptic on any given night.

oped in the last months at the MPC, that will be used for the NEOCP in the near future. The code is still undergoing a testing phase, but it's performing quite well and it's very close to be the code will be permanently used for the short-arc orbit determination, e.g. new objects.

Infrastructure (MPC): Two distinct PostgreSQL version 12 database environments were used. The first was a replica of the MPC's current development database of archived submissions, 180 million observations, simulating database state at scale during import/load. The second was an Amazon Web Services `r5a.2xlarge` EC2 instance (8 processors, 64GB RAM) with the same schema but no prior data, used to simulate the MPC side of the overall process (import/load, identify known/extend new, orbit update), accessible by both teams.

Processes: Most steps in the tested workflow were executed manually. The objective of this challenge was to ensure that the core data exchange workflow steps and associated data interfaces are understood, and testing automation would add complexity and distract from that objective. Tests of automation will be a subject of future challenges.

3. TEST 1: 21-DAY DATASET

3.1. *Test description and objectives*

This test consisted of generating ADES files for the first 20 days of the simulated survey, making a submission, and having the MPC process the submission to generate new object identifications and updates to orbits of already known objects.

The test workflow is illustrated in Figure 5. The object identifications were tested for correctness, but not used beyond that (i.e., to attribute objects in the next night of observations).

3.2. *Generated dataset*

We generated a series of 21 ADES from the simulated dataset discussed in section 2. These contained mock observations of already known objects (from the MPCORB part of the simulated catalog), as well as new objects expected to be discovered by LSST's tracklet linking algorithm (Figure 3).

3.3. *Test execution*

The `curl` interface for PSV-ADES submission was tested and assured to work, but the bulk of the generated files were delivered to the Minor Planet Center via direct download (for simplicity). There, they were converted to XML-serializes ADES, and ingested into the mock MPC database, simulating the ingestion procedure which will occur in operations. The reported objects were divided into two subsets: new object candidates and observations of known objects, and processed separately. The new object candidates were divide into seven different batches. The new object candidates were verified and their orbits computed using OrbFit. The benchmarks of OrbFit runs are listed in Table 1. The table contains the total number of objects in each batch, the total number of observations in each batch, the execution time and the failed cases. Those failures correspond to cases for which the whole orbit determination process failed.

# of objects	# of observations	Execution time	# of failures
30308	455050	129m 11s	75
30308	455028	132m 13s	78
30308	401964	142m 15s	124
30308	421253	138m 23s	108
30308	434351	135m 13s	98
30308	413372	139m 10s	112
30307	368947	125m 11s	136

Table 1. Benchmarks of OrbFit runs for new object candidates. The columns contain the total number of processed objects, the total number of observations corresponding to the processed batch, the execution time and the total number of objects for which the orbit determination failed.

The observations of known objects have been used to extend their arcs and re-fit the orbits, also using orbfit. The time to execute this element was similar to the time to run orbfit on new orbits. The resulting associations and orbits were formatted in MPCORB format, and up-

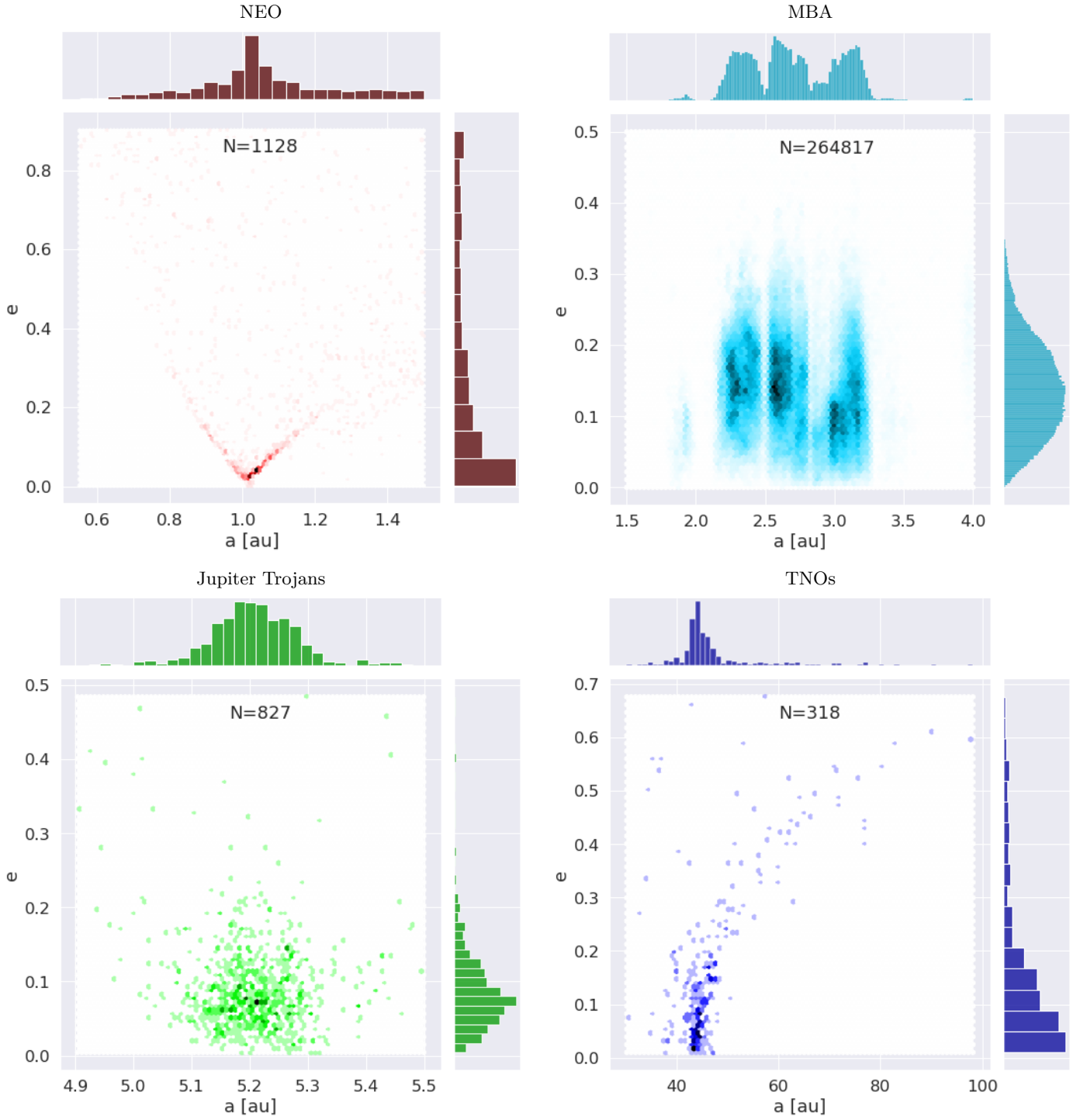

Figure 4. Semimajor axis vs orbital eccentricity for known and newly discovered Solar System objects after 17 nights of simulated LSST observations. The number N displayed in the graphs represent the number of objects falling into the plot domain. Those numbers are proxies for the actual number of discoveries.

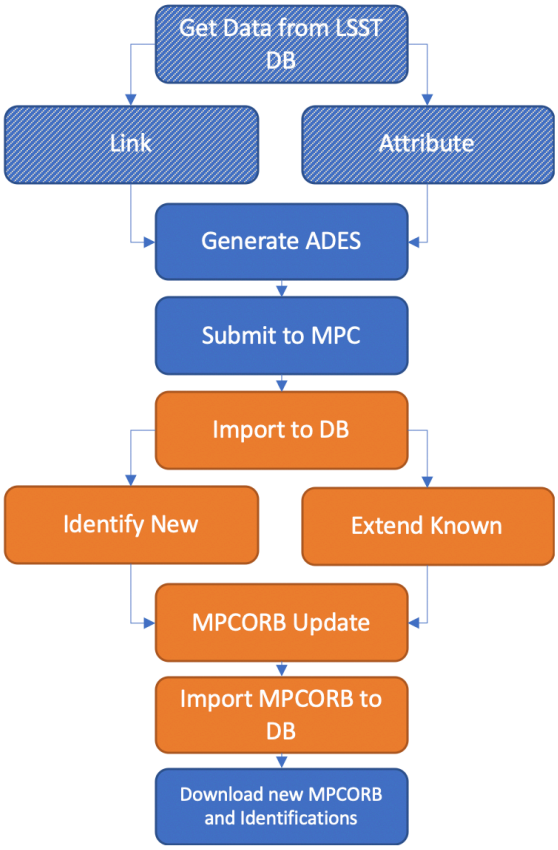

Figure 5. Workflow for Test 1. Solid-colored elements used existing prototypes of software or infrastructure, while hatched elements used simulated stand-ins.

loaded to the database. From there, they were queried by the LSST team.

3.4. *Test results*

This test was successful in full. All generated files were successfully submitted to the MPC and individually run through the workflow shown in Figure 5.

Work on attempting to reach the stretch goal and "close the loop" continues and will likely form a part of the Data Exchange Test 2.

4. DISCUSSION AND KEY CONCLUSIONS

Key findings and recommendations:

- All planned tests were attempted and – after fixing a few minor uncovered issues – ultimately successful. Stretch goals were partially executed, with full completion left for a subsequent exercise.

- The MPC is likely to be computationally capable to ingest LSST-sized datasets in the first months of LSST, assuming error-free operation. Available MPC computational capacity is already close to what's necessary to turn around the LSST submissions within the notional 4hr budget. No fundamental issues or bottlenecks were identified; assuming the necessary workflow control software is implemented, scaling to required capacity is likely to be a simple matter of adding more cores.

- The MPC should ideally plan for the ability to burst computational capacity on order of 10x relative to what it has presently available. This would enable comfortable data processing margins, as well as the ability to deal with unexpectedly large submissions (e.g., such catch-up submission after a multi-day outage). We note that this capacity is not required 24/7, but just over the few hour period when LSST data are being processed. Elastic cloud resources may be an ideal and cost-effective way to acquire it.

- Assuming it is fully operational since day one, the LSST is expected to discover approximately 0.5M new objects in the first month of operations. Designating those objects as they're discovered will result in designations with unprecedentedly high cycle counts (e.g., 2023 UX_{5678}), which cannot be written in packed form following the present scheme[2]. The packed provisional designation scheme will have to be updated to accommodate (ideally) O(1M) new discoveries in any half-month. Any changes should probably be rolled out soon, to give the community enough time to adapt existing tools that rely on the current format.

- Present ADES format specification does not allow a globally unique observation identifier – `obsId` – to be generated and submitted for unlinked detections by the observer[3]. Instead, the `obsId` field is assigned by the MPC, and is to be communicated by the observer in an unspecified manner. This creates practical difficulties for the observer in matching the MPC-assigned IDs with any IDs present in their database (especially when the submitted ADES files are automatically generated and considered ephemeral). We propose the ADES specification is updated to allow the `obsId` field to be generated by the observer (following a certain set of rules to ensure

[2] The present packing scheme allows for cycle counts up to 629, for only 16,354 discoveries in any half-month – https://www.minorplanetcenter.net/iau/info/PackedDes.html

[3] See Note 1 on Page 10 of https://minorplanetcenter.net//iau/info/IAU2015_ADES.pdf

global uniqueness), rather than the MPC. Alternatively, another observer-specified ID field – e.g., `localObsId` – could be added. This change would make the MPC database system as well as the observer's databases significantly less complex and more robust.

- The LSST should develop a realistic hybrid MPCORB+simulated populations catalog for future experiments. The simple $H > 15$ cut used to construct the merged catalog in this set of experiments inadvertantly excluded all simulated KBOs from the datasets.

- The ability to obtain data by directly querying the MPC database has been tremendously useful. While this introduces a tight coupling with the database schema (i.e., any change in the database schema may break the code that depends on it), the accelerated development enabled by it may make it a worthwhile tradeoff.

- The MPC and Rubin teams worked well together, with no collaboration issues. Having staff with small bodies expertise on the Rubin team enabled a quick "impedance match" to be established between the teams. Ability to directly message via a dedicated Slack channel has proven immensely useful. While originally planned to be done in-person, the entire data exchange challenge was conducted remotely.

Caveats and future work:

- This test did not address operations in more realistic conditions, such as having erroneous associations or other mistakes that may require manual intervention. Those will be addressed in a future challenge.

- Future tests will focus on automating most elements of the data exchange workflow, and testing of automated data exchange interfaces and processes. The verification of automation and the ability to scale is likely to be our next challenge.

- We have not tested resubmissions of large amounts of data, such as those which will be occurring after annual LSST data releases when improved astrometry and photometry become available. This is left for future tests.

This material is based upon work supported in part by the National Science Foundation through Cooperative Agreement 1258333 managed by the Association of Universities for Research in Astronomy (AURA), and the Department of Energy under Contract No. DEAC02-76SF00515 with the SLAC National Accelerator Laboratory. Additional LSST funding comes from private donations, grants to universities, and in-kind support from LSSTC Institutional Members. MJ and JM wish to acknowledge the support of the Washington Research Foundation Data Science Term Chair fund, and the University of Washington Provost's Initiative in Data-Intensive Discovery.

REFERENCES

Grav, T., Jedicke, R., Denneau, L., et al. 2011, PASP, 123, 423, doi: 10.1086/659833