The LSST DESC Data Challenge 1: Generation and Analysis of Synthetic Images for Next Generation Surveys

J. Sánchez,[1,2] C. W. Walter,[3] H. Awan,[4] J. Chiang,[5] S. F. Daniel,[6] E. Gawiser,[4] T. Glanzman,[5] D. Kirkby,[1] R. Mandelbaum,[7] A. Slosar,[8] W. M. Wood-Vasey,[9] Y. AlSayyad,[10] C. J. Burke,[11,12] S. W. Digel,[5] M. Jarvis,[13] T. Johnson,[5] H. Kelly,[5] S. Krughoff,[14] R. H. Lupton,[10] P. J. Marshall,[5] J. R. Peterson,[11] P. A. Price,[10] G. Sembroski,[11] B. Van Klaveren,[5] M. P. Wiesner,[15] and B. Xin[16]

The LSST Dark Energy Science Collaboration

[1] *Department of Physics and Astronomy, University of California, Irvine, Frederick Reines Hall, Irvine, CA, U.S.A.*
[2] *Fermi National Accelerator Laboratory, P.O. Box 500, Batavia, IL, U.S.A.*
[3] *Duke University, Department of Physics, Durham, NC, U.S.A.*
[4] *Department of Physics & Astronomy, Rutgers, The State University of New Jersey, 136 Frelinghuysen Rd, Piscataway, NJ U.S.A.*
[5] *SLAC National Accelerator Laboratory,2575 Sand Hill Rd, Menlo Park, CA, U.S.A.*
[6] *Department of Astronomy, University of Washington, Seattle, WA, U.S.A.*
[7] *McWilliams Center for Cosmology, Department of Physics, Carnegie Mellon University, Pittsburgh, PA, U.S.A.*
[8] *Brookhaven National Laboratory, Upton, NY, U.S.A.*
[9] *Pittsburgh Particle Physics, Astrophysics, and Cosmology Center (PITT PACC). Physics and Astronomy Department, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.*
[10] *Princeton University, Princeton, NJ, U.S.A.*
[11] *Department of Physics and Astronomy, Purdue University, West Lafayette, IN, U.S.A.*
[12] *Department of Astronomy, University of Illinois at Urbana-Champaign, Urbana, IL, U.S.A.*
[13] *Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA, U.S.A.*
[14] *LSST Project Office, Tucson, AZ, U.S.A.*
[15] *Benedictine University, 5700 College Road, Lisle, IL, U.S.A.*
[16] *Large Synoptic Survey Telescope, Tucson, AZ, U.S.A.*

1 July 2020

ABSTRACT

Data Challenge 1 (DC1) is the first synthetic dataset produced by the Rubin Observatory Legacy Survey of Space and Time (LSST) Dark Energy Science Collaboration (DESC). DC1 is designed to develop and validate data reduction and analysis and to study the impact of systematic effects that will affect the LSST dataset. DC1 is comprised of r-band observations of 40 deg^2 to 10-year LSST depth. We present each stage of the simulation and analysis process: a) generation, by synthesizing sources from cosmological N-body simulations in individual sensor-visit images with different observing conditions; b) reduction using a development version of the LSST Science Pipelines; and c) matching to the input cosmological catalog for validation and testing. We verify that testable LSST requirements pass within the fidelity of DC1. We establish a selection procedure that produces a sufficiently clean extragalactic sample for clustering analyses and we discuss residual sample contamination, including contributions from inefficiency in star-galaxy separation and imperfect deblending. We compute the galaxy power spectrum on the simulated field and conclude that: i) survey properties have an impact of 50% of the statistical uncertainty for the scales and models used in DC1 ii) a selection to eliminate artifacts in the catalogs is necessary to avoid biases in the measured clustering; iii) the presence of bright objects has a significant impact (2- to 6-σ) in the estimated power spectra at small scales ($\ell > 1200$), highlighting the impact of blending in studies at small angular scales in LSST;

Key words: catalogues – cosmology: dark energy – methods: observational – software: simulations

1   INTRODUCTION

The increase in statistical power of recent cosmological experiments makes the modeling and mitigation of systematic uncertainties key to extracting the maximum amount of information from these surveys. End-to-end simulations (Brun et al. 1978; Agostinelli et al. 2003; Sjöstrand et al. 2006) provide a unique framework to model systematics and streamline processing and analysis pipelines since they provide a complete understanding of the inputs and outputs. With the increasing availability of computational resources, end-to-end simulations have started to become more prevalent in imaging surveys (Bruderer et al. 2016; Zuntz et al. 2018), and similar efforts are being undertaken in spectroscopic surveys such as the Dark Energy Spectroscopic Instrument (DESI Collaboration et al. 2016).

One the science goals of the Rubin Observatory Legacy Survey of Space and Time (LSST) is the study of dark energy (Ivezić et al. 2019) using probes such as weak lensing (WL), galaxy clustering, clusters of galaxies, supernovae and strong lensing. The increased sensitivity of LSST relative to precursor Stage III dark energy experiments (Albrecht et al. 2006) necessitates more stringent control over systematic uncertainties for probes like galaxy clustering and WL. The development of end-to-end simulations enables validation and verification of the processing and analysis pipelines. For example, image simulations can be used to evaluate the performance of different shape-measurement algorithms, deblending algorithms, and other processing algorithms. In addition, the expected data volume of LSST, ~ 50 PB of raw data and ~ 40 billion objects (Ivezić et al. 2019) after 10 years, motivates the use of simulated data sets for the development of data handling and analysis pipelines.

The Dark Energy Science Collaboration (DESC[1]) has planned a series of Data Challenges (DCs) carried out over a period of years, aimed at successively more stringent and comprehensive tests of analysis pipelines, to ensure adequate control of systematic uncertainties for analysis of the LSST data. An additional goal of these DCs is development of the infrastructure for analyzing, storing, and serving substantial data volumes; even while using the outputs of the LSST Science Pipelines as inputs into analysis pipelines, the analysis pipelines will need to handle quantities of data greater than that of ongoing surveys, even after a single year of the LSST. Moreover, it is anticipated that non-negligible subsets of the data may need to be reprocessed to generate systematic error budgets (e.g., assessing sensitivity of the results to certain stages of the analysis process by changing some parameters in the analysis). These goals will be achieved in practice by a combination of reprocessing of precursor datasets, which have the advantage of being fully realistic, and completely synthetic datasets, that have both a known ground truth and the ability to enable and disable various effects and thus study them in a more controlled environment.

The goals of the DCs dictate a gradual increase in the sophistication and volume of the simulated data. In this paper, we present and analyze simulated images from DC1, the first such data challenge planned within DESC. The nominal goal is to produce synthetic data corresponding to 10 years of integration in the *r*-band over a contiguous patch of the sky covering approximately 40 deg². Only one of the six LSST filters is represented, and the area is a small fraction ($\sim 0.2\%$) of the total LSST area. We describe how the simulation was achieved and characterize the resulting products. We validate the basic photometric and astrometric calibration of these products and check the performance of the pipeline against the requirements set by LSST and DESC in their respective Science Requirements Documents (Ivezić et al. 2013; The LSST Dark Energy Science Collaboration et al. 2018). To check suitability of this dataset for galaxy clustering measurements, we perform a two-point clustering analysis in harmonic space, assess the impacts of observing conditions, foregrounds, and detector characteristics as potential sources of systematic effects and how the observing strategy can mitigate their impact. Additionally, we assess the importance of sample selection and how the presence of artifacts and the presence of bright stars can bias the clustering results. The DC1 data products encompass single-visit and coadded calibrated exposures (i.e., flattened, background subtracted, etc.) and source catalogs that, together, have a volume of ~ 225 TB.

This paper is structured as follows: In Section 2 we summarize the factors that informed the design of this data challenge, the inputs, and observing strategies used for DC1. In Section 3 we briefly summarize the tools used to generate and process synthetic images for DC1. Section 4 describes the matching procedure to relate the simulation inputs with their corresponding outputs. In Section 5, we illustrate the data products generated and perform several validation tests. In Section 6 we summarize the procedure to obtain a clean sample of galaxies suitable for clustering analyses. In Section 7, we present the clustering analyses on the simulated data products. Finally, in Section 8, we offer concluding remarks.

2   DATA CHALLENGE DESIGN

As mentioned in Section 1, the design of DC1 is driven by a combination of several needs: developing infrastructure for processing and serving data in a way that is useful to DESC and building and testing analysis pipelines including different strategies to mitigate systematic uncertainties affecting various dark energy probes. The philosophy behind the design of the DESC data challenges is to increase the complexity and level of realism of the datasets in each subsequent iteration. Thus, DC1 is limited in scope and focus, testing a subset of the systematics affecting the different probes that DESC will use.

DC1 serves as a stepping stone to eventually produce simulated surveys covering hundreds to thousands of square degrees in DC2 and beyond. The area of DC1 is sufficient to enable tests of two-point clustering statistics up to ~ 1 degree scales. To ensure the volume of simulated images is tractable, DC1 only includes images in a single band (*r*-band), but goes to full LSST 10-year depth.

The full simulation workflow is depicted in Figure 1. Briefly, we use as inputs the positions, shapes and fluxes from a galaxy mock catalog from CatSim (Connolly et al. 2010, 2014), which we describe in more detail in Section 2.1, and the observing conditions and strategies described in

---

[1] http://lsstdesc.org/.
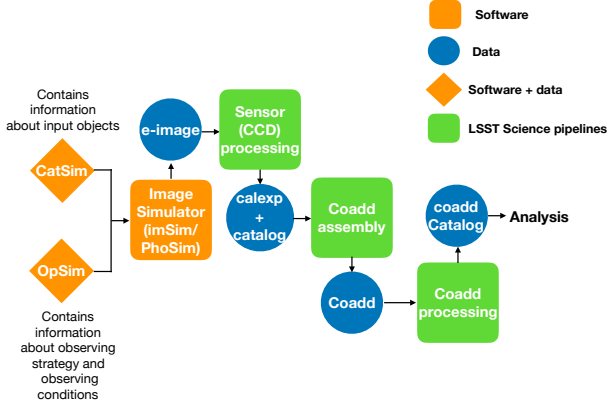
Figure 1. Workflow diagram for full data simulations. The image simulation step produces synthetic raw observations from a known truth catalog based on N-body simulations and simulated observing conditions. These data are processed by the LSST Science Pipelines to first generate calibrated single exposure images. These are calibrated both astrometrically and photometrically and are fed again into the LSST Science Pipelines that produce the image co-adds and a catalog of detected static sources.

Section 2.2 using LSST Operations Simulator (OpSim Delgado et al. 2014). These are passed to our image simulation packages described in Section 3.1 that produce raw e-images (i.e., full sensor images in electrons per pixel without any added instrumental effects such as cross-talk, bleeding, etc.). These e-images are then processed by the LSST Science Pipelines (Jurić et al. 2015; Bosch et al. 2018). The processing is described in Section 3.2. These produce the calibrated exposures, coadds and catalogs that we use for our analysis.

2.1 Image generation: input catalog

Image simulations allow us to assess the detection and deblending performance of a given image-processing pipeline. For example, if we produce images using an object catalog with random positions uniformly distributed across the sky, as well as uniformly random shapes and fluxes, we can get information about detection efficiencies as a function of flux. However, the effects of source blending would not be realistic as we would not be able to capture some correlations present in real data. On the other hand, using N-body simulations as the input to generate artificial images allows us to study all the aforementioned effects. We used the CatSim (Connolly et al. 2010, 2014) catalog as our input to include realistic correlations between galaxies and to be able to test our analysis pipelines. CatSim is a set of catalog-level simulations provided by the LSST Simulations Team representing a realistic distribution of both Milky Way and extragalactic sources. In particular, the extragalactic catalog contains galaxies spanning the redshift range $0 < z < 6$ in a 4.5 deg×4.5 deg footprint. The magnitude and redshift distributions are shown in Figure 2. The galaxies in CatSim are generated by populating the dark matter haloes from the Millennium simulation (Springel et al. 2005) using a semi-analytic baryon model described in De Lucia
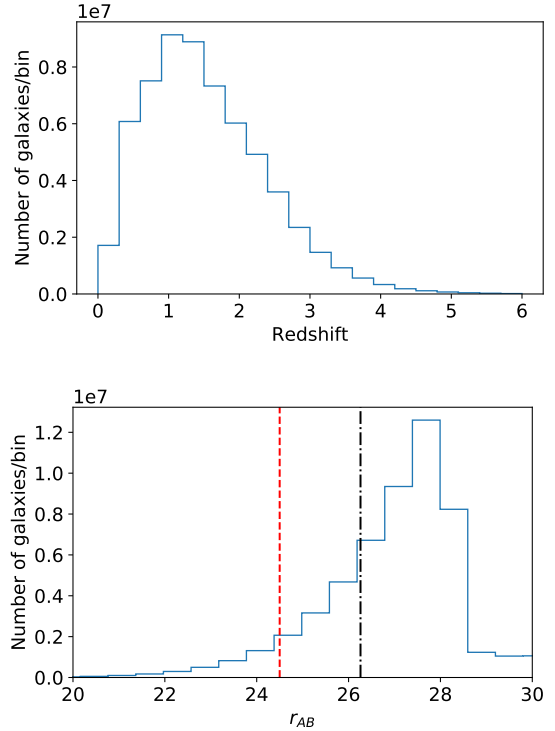
Figure 2. Redshift (top) and magnitude (bottom) distribution for the galaxies used as inputs for the Data Challenge 1 simulations. In the magnitude distribution we include, as references, the typical 5-σ point source detection depth for a single exposure (red dashed line) and the median 5-σ point source detection depth in the deepest coadded DC1 simulation (black dashed-dotted line).

et al. (2006) including magnitudes BVRIK, LSST-ugrizy, and bulge-to-disk ratios. For all sources, a spectral energy distribution (SED) is fit to the galaxy colors using Bruzual & Charlot (2003) spectral synthesis models. Fits are undertaken independently for the bulge and disk and include inclination-dependent reddening. Morphologies are modeled using two Sérsic profiles (Sérsic 1963) and a single point source (for the AGN). Half-light radii for the bulge components are derived from the absolute-magnitude vs. half-light radius relation given by Gonzalez et al. (2011). Stars are represented as point sources and are drawn from the Galfast model (Jurić et al. 2008). More information about these catalogs can be found at the LSST Simulations webpage[2]. In principle, the semi-analytic model used in CatSim can give us insight about the small-scale regime that LSST may be sensitive to. However, at fainter magnitudes ($r > 25$) the semi-analytic model underpredicted the overall number density of objects, and unclustered galaxies were added in order to obtain the correct number densities (Connolly, private comm.) which can potentially dilute the signal. We do not expect that using a different galaxy inpainting model would change the conclusions of this work since we do not attempt to make any measurement of cosmological or halo-occupation-distribution parameters.

For DC1, we chose a nominal field centered at RA $\approx 93°$

---

[2] https://www.lsst.org/scientists/simulations/catsim.

and Dec $\approx -29°$. This field has a Galactic latitude of $b \approx -20°$ and a dust extinction per magnitude of interstellar reddening $0.05 \leq E(B-V) \leq 0.35$ and thus represents a typical region in the LSST wide-fast-deep survey (Ivezić et al. 2019). The CatSim catalog was tiled to generate a 40 deg^2 footprint, covering 4 LSST full focal plane pointings. This approach introduces a periodicity that induces extra correlations in our sample; however, this is not a major issue as with the relatively small area of the DC1 simulation we are unable to measure correlations on large scales (> 1 deg) with any useful precision.

After tiling, the input catalog contains approximately 63.1 million sources, of which 97% are galaxies whose redshift and magnitude distributions are depicted in Figure 2 and the remaining objects are stars. We simulate r-band observations to the LSST full depth (10 years, 30-second exposures) within the DC1 footprint using an observing cadence generated with OpSim[3] (Delgado et al. 2014), which contains simulated pointing position, observation date and filter. It also contains information about simulated observing conditions, such as seeing, sky-brightness, and moon position. For DC1 we use 4 pointings with ≈ 184 visits per pointing over 10 years from this database.

2.2 Dither strategy

OpSim's output contains a realization of the LSST observing cadence and the survey footprint. Since OpSim divides the sky into hexagonal tiles, the nominal telescope pointings lead to overlapping regions across adjacent tiles that are observed more often than the non-overlapping part of the field of view (FOV), resulting in depth non-uniformity on the scale of ~1 degree. This non-uniformity can introduce systematic uncertainties in the two-point statistics of galaxies (Awan et al. 2016). In an effort to mitigate these effects, and following the same approach that will be taken with LSST data, we implement *dithers* – offsets in the nominal telescope pointings. Specifically, here we use *large*, i.e., as large as half the FOV, random translational and rotational dithers implemented for every visit. Note that these dithers differ from those recently discussed in Lochner et al. (2018). This translational dither strategy is chosen based on a more extensive study of the various (translational) dither strategies in Awan et al. (2016), where random dithers for every visit are found to be among the most effective.

We consider both an undithered and a dithered observing strategy. For the dithered strategy, some visits contain sensors that fall outside of the DC1 region; these sensors were not simulated in order to save computational resources. However, the sensors that partially overlapped our nominal field of view were simulated. In total, we simulate $\approx 184,600$ sensor-visits for the dithered simulation and $\approx 151,000$ for the undithered simulation.

3 IMAGE GENERATION AND PROCESSING

The artificial generation of astronomical images is a complex and computationally demanding process. In recent years, there have been major efforts in the community to create software that enables the generation of astronomical images, with various choices made in terms of level of complexity, fidelity, and computational efficiency, such as PhoSim[4] (Peterson et al. 2015), and UFIG[5] (Bruderer et al. 2016). For DC1, we model the input sources using two independent approaches.

Firstly, we use PhoSim, a fast photon Monte Carlo code that enables the generation of images with a high level of realism and can use LSST-specific information (e.g., the geometry of the CCDs and the focal plane, the system throughputs in the different bands, etc.) to generate LSST-like images. We use PhoSim version 3.6 with a custom configuration that enables some approximations (photon bundling) to be made when generating the sky background in order to reduce the overall computing time needed to produce the images.

We also employ imSim[6] (Walter et al., *in prep.*), which internally uses GalSim (Rowe et al. 2015) as a library for image rendering, and also uses LSST-specific information to generate synthetic images. For DC1 we use an early prerelease version of imSim, version v.0.1.0, which performs a series of approximations that allow us to complete the image simulations significantly faster ($\sim 60\times$) than PhoSim v3.6, at the expense of a loss in realism. These approximations include simplifications in the point-spread function (PSF) model (PhoSim performs ray-tracing through the atmosphere, while this version of imSim uses simple parametric models for the PSF) and omission of all sensor effects, which were deemed acceptable given the goals of DC1.

Due to the computational resources needed to run PhoSim we could only generate one campaign of the dithered DC1-PhoSim data, whereas we could produce the dithered and undithered campaigns for imSim. Furthermore, comparison of the results for the PhoSim and imSim images has proven less informative than intended due to some unusual features of unknown origin in the sky backgrounds in the PhoSim images[7]. For these reasons, we focus on the analysis and comparison of dithered versus undithered imSim images for the rest of this work.

3.1 DC1 imSim configuration

For DC1, we use imSim to simulate each CCD of the focal plane individually, and generate a single image with a 30-second exposure time. We omit sensor effects and variability in the optical model across the focal plane. We set the gain to be 1. Our sky brightness model is based on the Krisciunas & Schaefer (1991) model provided by OpSim, refined with

---

[3] Specifically, we use the `minion_1016` database: https://www.lsst.org/scientists/simulations/opsim/opsim-v335-benchmark-surveys.

[4] https://bitbucket.org/phosim/phosim_release/wiki/Home.
[5] https://cosmology.ethz.ch/research/software-lab/ufig.html.
[6] https://github.com/LSSTDESC/imSim.
[7] Several aspects of the implementation of the sky background rendering – including the photon-bundling approximations – are updated in subsequent versions of PhoSim, so this finding may not carry over to later PhoSim versions.

the detailed wavelength dependence of the phenomenological model from Yoachim et al. (2016). The PSF model is a Gaussian for the system (telescope, CCD and other elements that may be in the optical path other than the atmosphere) with a dependence on airmass, X, of the full-width half-maximum, $\text{FWHM}_{\text{sys}} = (0.4'')X^{0.6}$, to approximate the degradation in the image quality due to, e.g., gravity loading[8]. An airmass-dependent Kolmogorov profile[9] is used to model the atmosphere. In order to be consistent with the sky brightness model we use an airmass, X, model that depends on the angular distance to the zenith, Z, from Krisciunas & Schaefer (1991):

$$X = (1 - 0.96 \sin^2 Z)^{-0.5}. \qquad (1)$$

For DC1, imSim is used to generate three different types of objects: stars, which are modeled as PSF-like objects; galaxies, which are modeled as composite (bulge plus disk) Sérsic profiles (Sérsic 1963) using the parameters given by CatSim; and AGNs which are also modeled as point sources and, for simplicity, without any variability.[10] The brightness for these sources is computed using the magnitudes from CatSim, which are converted to counts using the latest version of the LSST throughputs[11]. In DC1, we clip the objects at magnitude 10 in order to improve the computational efficiency. Saturation is emulated by clipping the maximum number electrons per pixel in the CCDs at 100,000.

The final products of the image generation process are $4k \times 4k$ pixel images (in FITS format). We generated more than 300,000 sensor-visit images in total (including both the *dithered*, and *undithered* fields). The average time to simulate each CCD is ~ 4000 seconds and the total production time is $\sim f0,000$ CPU-hours.

3.2 Image processing

The outputs of these simulations are processed using the LSST Science Pipelines (Ivezić et al. 2019; LSST Science Collaboration 2009; LSST Dark Energy Science Collaboration 2012; Jurić et al. 2015; Bosch et al. 2018) version 13.0^{12}, which we will refer to as the Data Management (DM) stack. The DM stack is an open-source, high-performance data processing and analysis system intended for use in optical and infrared survey data[13]. A brief schematic of some of the steps in the image processing pipeline can be seen in Figure 1 as green squares. The raw, uncalibrated single exposures are used as inputs. The software performs the reduction, detection, deblending and measurement on individual visits. It then combines the single-visit images to produce the so-

called coadds[14]. This is done by computing a weighted average of resampled overlapping sensor images in a given region of the sky that we call a `patch`. An illustration of how this process works is shown in Figure 3. After assembling the coadded images, the DM stack performs measurements on them to produce a catalog. The DM stack provides calibrated images and source catalogs for the individual visits and coadds stored in FITS files. In total, we detect and measure ~ 10.6 million (9.7 million for the undithered simulation) objects with position, flux and shape information. We activated optional extensions for the pipeline to include `CMODEL` fluxes (see Bosch et al. 2018 for more details) and `HSM` shapes (Hirata & Seljak 2003; Mandelbaum et al. 2005). In Figure 3 we show an example of a single-visit processed image and a coadded image, which corresponds to 184 dithered overlapping single-visit images. We can see the stark difference in the detectable number of objects just by eye (~ 5 times more objects with $\text{SNR} > 5$), due to the increased depth (the coadd images are ~ 3 magnitudes deeper).

The reduction pipeline is essentially the same as for Hyper Suprime-Cam (HSC). This allows us to use the HSC selection criteria (Mandelbaum et al. 2018, Sec. 5.1) as the basis for our analysis, and can potentially enable direct comparisons between datasets for further validation.

The total processing time for the DC1 simulated images is $\approx 29,000$ CPU-hours.

4 MATCHING INPUTS AND OUTPUTS

Using end-to-end simulations, one can potentially trace each measured photon from its corresponding source and fully characterize the image generation and measurement processes. In practice, this is very difficult due to the large data volume and the fact that the data reduction pipeline is built around pixelated images rather than tagged photon counts.

Nevertheless, the output catalog is a noisy representation of the underlying input (truth) catalog and we need to find a way to connect the two. The simplest way to associate members of two catalogs is by using the positions of the objects in the sky. This approach has been extensively used in the literature (e.g., de Ruiter et al. 1977; Benn 1983; Wolstencroft et al. 1986) and performs reasonably well when blending is low (i.e, when there are few overlapping sources in the image). However, when the blending fraction is high, this approach might be insufficient. In this case, matching other quantities like flux, color and/or shape can become useful (Budavári & Szalay 2008; Budavári & Loredo 2015). However, when adding other quantities, the matching process can become slower and result in a lower matching completeness. For more details about challenges relating two different catalogs, we refer the reader to Budavári & Loredo (2015).

We compare two different matching strategies: positional matching, where for each detected object we find the closest object in the truth catalog, which we will refer to as *pure spatial matching* and will denote as S; and positional matching with magnitude matching, which we will refer to as

---

[8] See LSE-30 http://ls.st/lse-30 p. 80.
[9] http://galsim-developers.github.io/GalSim/_build/html/_modules/galsim/kolmogorov.html.
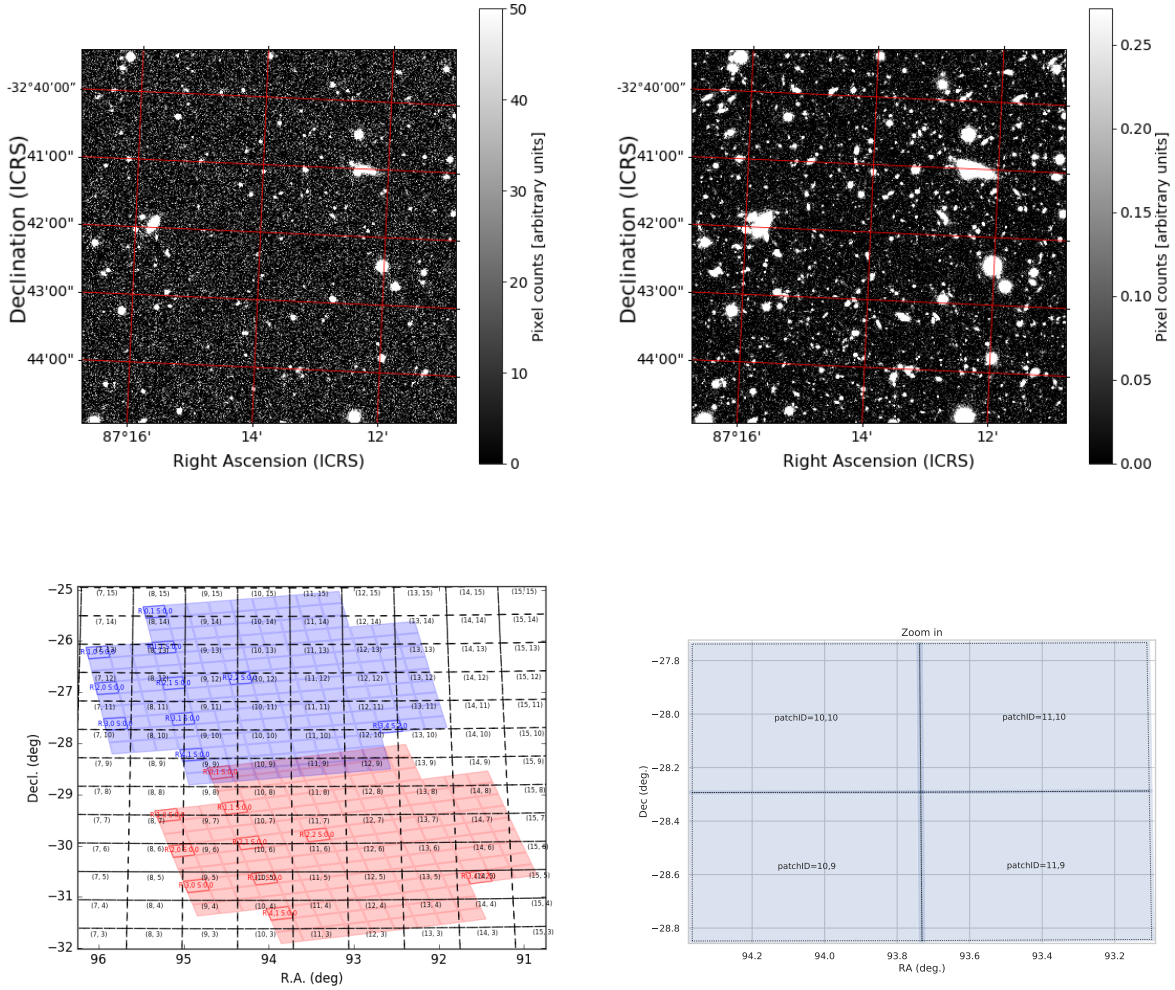[10] Newer versions of imSim have the ability to generate more complex galaxy morphologies (e.g., they can include random Gaussian knots).
[11] https://github.com/lsst/throughputs.
[12] https://pipelines.lsst.io/releases/v13_0.html.
[13] The code can be found at dm.lsst.org and pipelines.lsst.io.

[14] For more information about coaddition, see Section 3.3 in Bosch et al. (2018)

Figure 3. Example of a 1000 × 1000 pixel cutout from a calibrated exposure (top left), i.e. background subtracted, reduced single-epoch image; a full depth coadd (top right). We can see the stark difference in the number of objects that are detectable by eye. The red grid corresponds to lines of constant Right Ascension (vertical lines) and constant declination (horizontal lines). We also show a visualization of the coaddition scheme used by the Science Pipelines (bottom left and right). In the bottom left panel the different images in each sensor visit (red and blue squares) are mapped to patches (blank squares delimited by dashed lines) where the coaddition is performed independently. These patches have small 1 arcmin-wide overlaps with each other, and each patch is identified by a pair of numbers. The bottom right panel shows nine different patches as light blue squares, and we can see their overlap as the darker blue areas.

as *spatial+magnitude matching* and denote as S+M, where for each detected object we find objects from the truth catalog that lie within a three pixel radius ($0.6''$). After this, we select the object that is closest in magnitude as long as the difference in r-band magnitude, Δr, is less than a certain threshold. In our case, we conservatively choose $|\Delta r| < 1.0$. Using this approach, if none of the neighbors fulfill these conditions, the detected source is considered unmatched.

For both approaches, we build a `KDTree` (Pedregosa et al. 2011) using the positions of detected objects flagged with `detect_isPrimary=True` which ensures that the source cannot be further deblended by the pipeline and was detected in the inner part of a patch. The inner part of a patch is the part of the sky exclusive to the said patch. In order to speed up the processing and to reduce the usage of computational resources we build the `KDTree` using sources from 30 randomly selected patches ($\sim 10\%$ of the total num-

ber of patches) in the dithered simulation containing 975,605 detected sources fulfilling the aforementioned condition (we will refer to these as primary outputs or primary detected sources). Using this sample, we find that 95.2% of the sample is matched using the S+M matching. The undithered simulation yields similar results and conclusions.

One interesting metric is the fraction of primary detected sources that have been matched to the same object in the input catalog, which we will denote as f_{multi}. This is an unusual occurrence, but a situation where this can happen is when bright objects appear "shredded", i.e., one bright object is detected as multiple fainter sources. This happens because noise fluctuations of the said bright object are considered as single sources. We find that these kind of matches are ~ 100 times more likely to happen using the S matching ($f_{\text{multi}} = 3 \times 10^{-3}$) than in the S+M matching ($f_{\text{multi}} = 2 \times 10^{-5}$). This is expected because, in the
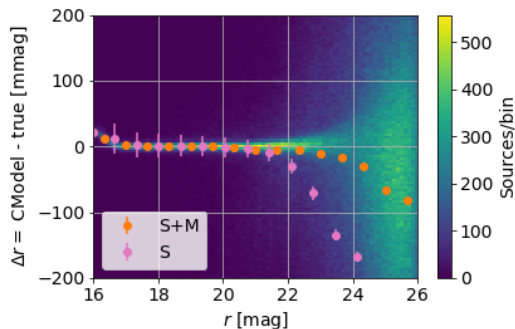
Figure 4. Median per-bin photometric residual as a function of measured magnitude for pure spatial matching (pink), and spatial+magnitude matching (orange) using 15 magnitude bins between r-band magnitude 16 and 26. We also show the 2D histogram of the residuals for the matched objects using S+M matching as a function of magnitude.

cases where the primary detected source is a random fluctuation or a shredded source, it is unlikely that the measured flux is close to the flux of the true, neighboring source thus producing an unmatched source in the case of using S+M matching. We also find that the two approaches select the same matching source in the truth catalog for only 68% of the primary detected sources for which we found a match. These differences have several potential explanations: for example, objects with a poorly determined centroid position that are close to other objects in the truth catalog (remember that the truth catalog contains objects up to $r = 28$), objects with poorly determined fluxes (low SNR), etc.

We compare the photometric residuals, $\Delta r = $ CMODEL - r_{true} (i.e. measured minus input magnitudes), using both approaches in Figure 4. The resulting median photometric residual seems strongly biased in the case of using S matching. However, in the case of S+M matching, the median residuals and their uncertainties are smaller, as expected given the limit in the magnitude difference. We can see that in this case the biases are still significant, partially because of the fact that faint sources just below the detection threshold are detected only if they have positive noise fluctuations, and we lose some faint sources above the detection threshold that have noise fluctuations that make them appear dimmer, biasing the overall residual distribution. On top of that, spurious matches also contribute to this effect. This can be mitigated using a smaller tolerance in magnitude difference (for example using 0.5 mag instead of 1 mag) but that would result in an overall reduction of the number of matched primary detected sources. Finally, observing the magnitude difference between inputs and outputs of individual matched sources using the S+M technique, we can see that this distribution is centered around zero, except for very bright sources ($r < 17$), where saturation prevents us from accurately determining the fluxes.

Different matching techniques have different potential applications and strengths (Budavári & Loredo 2015). In our case, we want to use these matching techniques to provide a clean (flux-limited) sample to perform two-point clustering analyses. Given that magnitude precision and accuracy will be important for our sample selection, the **spatial+magnitude** matching technique will be sufficient

to clean the sample from spurious and poorly measured sources. However, more complicated matching techniques may be necessary for other use cases.

5 OUTPUT CATALOGS AND VALIDATION

In order to check the level of realism and the accuracy of the processed catalogs we perform several validation tests. These tests check two different aspects: the level of realism and consistency of the simulated products with the inputs, and the performance of the simulation and processing pipelines. As a guide to check the status of our end-to-end pipeline, with a special focus on the performance of the processing pipeline, we use the LSST Science Requirements Document[15] (Ivezić et al. 2013), which we will refer to as the LSST-SRD[16]. It describes science-driven requirements for LSST data products in its Key Performance Metrics (KPMs). The results in this section are presented for the dithered simulation; however, unless stated otherwise, the procedures and results of the validation checks are similar for the undithered simulation.

After performing some basic sanity checks (e.g., of the magnitude distribution, the number density of detected objects, and the footprint coverage) we test our simulations against the different KPMs by processing the output individual-visit and coadd catalogs with the validate_drp[17] package. For quick reference, a brief description of the requirements from the LSST-SRD studied in this section can be found in Appendix A. In addition, we also validate our dataset using some PSF-related requirements in the DESC Science Requirements Document (The LSST Dark Energy Science Collaboration et al. 2018), which we will refer to as the DESC-SRD.

Note that some of the performance requirements for LSST are not met by design in our simulations, e.g., requirements involving colors or more than one filter. We will ignore such requirements in this work. On the other hand, our images automatically meet some criteria due to the design choices (e.g. the pixel size is fixed in our simulations).

5.1 Astrometric performance

Consistency with a reference frame (absolute astrometry) is required in order to compare with external catalogs. One can think about the case of cross-correlating with an external catalog, e.g. CMB maps; if the astrometric solutions are not consistent the cross-correlation signal will be biased. Much more stringent constraints for absolute astrometric accuracy come from the orbit computations for solar system objects that will be studied with LSST, for which the maximum acceptable mean deviation with respect to the reference is 100 milliarcseconds (mas). We check this requirement by measuring the difference between input and output positions, shown in Figure 5, finding a median (and mean) deviation of 38 mas. This bias is due to the uncorrected proper motion of stars used for calibration. The proper motions of

---

[15] https://ls.st/LPM-17.
[16] We use the LSST-SRD version 11.
[17] This is an open-source package that tests data release products against some KPMs: dmtn-0008.lsst.io, https://github.com/lsst/validate_drp.
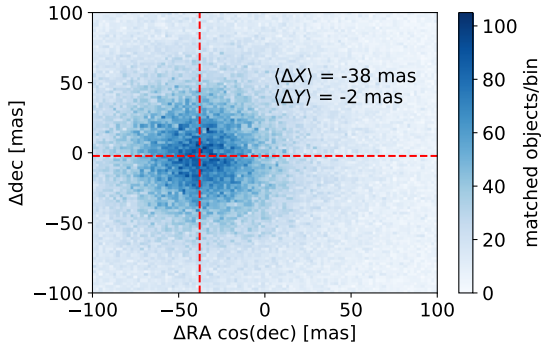
Figure 5. Astrometric residuals of detected objects (stars and galaxies) of 500 randomly selected visits in the dithered simulation. We find a bias of 38 mas in the measured right ascension due to uncorrected proper motion. We obtain similar results using the undithered simulation.

stars were included in the simulations; however, the catalog used for calibration had the location of these stars fixed at a specific epoch (J2000). This will be solved in future data challenges.

On the other hand, we require that the positions of objects in different visits be consistent (astrometric repeatability). For example, the coaddition process relies on the astrometric solutions for the single-epoch images in order to get deeper images. If the astrometric solutions between different epochs are systematically offset or their errors are too high, we can obtain galaxy profiles that are too wide, or inconsistent relative distances between objects at different epochs, which will affect our clustering and weak lensing science. The study of parallax and proper motion of stars set a more stringent requirement on the astrometric repeatability of LSST. In particular, we check pairs of bright stars ($17.0 < r < 21.5$) separated by $4. < D < 6$ arcmin, and compare the distribution of the variation in their separation distance at different epochs. We require that the root mean square (RMS) of this distribution should be below 20 mas, and that the fraction of outliers, i.e., pairs where the variation of the separation is larger than 40 mas should be below 20%. We repeat this for pairs with $19 < D < 21$ arcmin following the same criteria, and pairs with $199 < D < 201$ arcmin, where we require that the RMS should be below 30 mas, and less than 30% of the pairs vary their separation distance by more than 50 mas. We find that our dataset passes all of these astrometric requirements, guaranteeing that the positions measured in DC1 will be useful for clustering analyses.

5.2 Photometric performance

In order to have well calibrated supernova lightcurves, as well as accurate photometric redshifts, we need consistent photometric measurements across different visits, since artificial variations due to miscalibration can lead to biased cosmological estimates. In particular, LSST requires that the RMS around the mean magnitude value for bright stars should be below 8 mmag, and that no more than 20% of bright point-like sources deviate by more than 15 mmag. We check these criteria by comparing the measured magni-
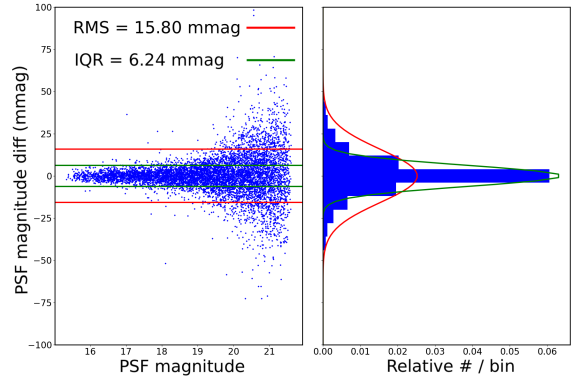
Figure 6. (left) The magnitude differences of pairs of measurements of stars across visits for stars with a typical SNR > 100 as.a function of their measured psf magnitude, i.e., the magnitude measured in a PSF-like aperture. (right) The histogram of these differences. The Gaussian root mean square (RMS) is shown in red while the interquartile range is shown in green. Note that the distribution is more peaked than a Gaussian. The interquartile range (6.2 mmag) is *smaller* than the Gaussian RMS (15.8 mmag). This means that the distribution has extended tails but most objects have very accurate magnitude measurements.

tudes of bright (SNR > 100) point-like objects across different visits as shown in Figure 6. This test validates that the pipeline is reconstructing fluxes of objects consistently across epochs, and also that different epochs are produced consistently in our image simulations. We estimate that the RMS in our case is 15.8 mmag, larger than the requirement; however, we see that the RMS is affected by the presence of outliers and decide to compare with the scaled interquartile range[18], for which we obtain an IQR = 6 mmag; We also check the fraction of photometric, bright point-like sources that deviate by more than 15 mmag, finding only 17% over this threshold, in compliance with the requirements (20%). The smaller interquartile range compared to the RMS in combination with the small fraction of outliers suggest that the majority of objects in DC1 have well-measured photometry, while a small fraction of outliers have poorly determined photometry. We do not expect this to cause any biases in clustering or weak lensing measurements. In the case of DC1, these criteria are not crucial since we do not perform any photometric redshift measurements given that we only have single-band imaging.

5.3 Zeropoint uniformity

The presence of clouds, calibration problems and noise fluctuations can lead to changes in the estimated magnitude zeropoint. Accurate estimation of photometric redshifts, distance moduli to supernova, and proper separation of stellar populations require stable zeropoints across the sky. In particular these science cases require that the uncertainty in the zeropoints, σ_{zp}, should show an RMS lower than 15 mmag

---

[18] we use a rescaled IQR$\equiv$ 75th percentile minus 25th percentile, and then divided by $2\sqrt{2}\mathrm{Erf}^{-1}(1/2)$ (the interquartile range of a Gaussian distribution with $\sigma = 1$) in order to have a robust estimation of σ.
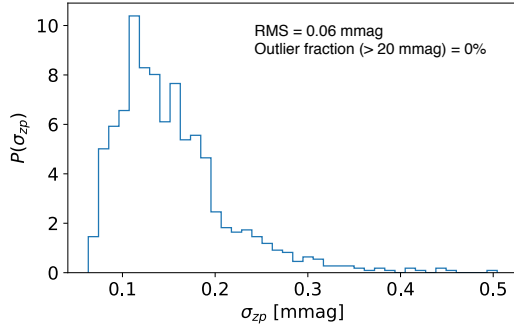
Figure 7. Normalized distribution of the error in the zeropoint value, σ_{zp} for 1000 randomly selected sensor-visits. The distribution is quite asymmetrical. The RMS is 0.06 mmag, fulfilling the requirements (15 mmag). We also find that none of the visits deviate from the median by more than 20 mmag, in compliance with the requirements.

Figure 8. Measured depth in visits with Altitude > 80 deg, $0.68'' <$ seeing $< 0.71''$ and Sky-brightness (in r-band) ≥ 21 (blue histogram) compared to the predicted depth by OpSim (solid orange histogram). The median of this distribution (dashed line) is very close to the LSST-SRD minimum depth, 24.3 (red vertical line), the 20th percentile is also shown and we can appreciate that it is larger than 24.0 as established by the LSST-SRD. All the visits fulfilling the criteria above have the same predicted depth by OpSim.

and no more than 20% of the images should have a deviation larger than 20 mmag (see the LSST-SRD for more details). We randomly select 1000 sensor-visits and check the distribution of σ_{zp}, depicted in Figure 7. We find that the RMS of the distribution is 0.06 mmag, fulfilling these requirements by a very wide margin. This is mainly a consequence of the lack of clouds and other fluctuations that may affect the zeropoint calibration (e.g., temperature/gain fluctuations in the sensors, etc.). We also find that none of the sensors have an error in the zeropoint that deviates from the median more than 20 mmag. Finally, we compare input and output magnitudes, for both stars and galaxies using `CMODEL` magnitudes (Bosch et al. 2018). We compute the median of the difference between them and obtain 17 mmag, which is smaller than the maximum allowed in the LSST-SRD (20 mmag).
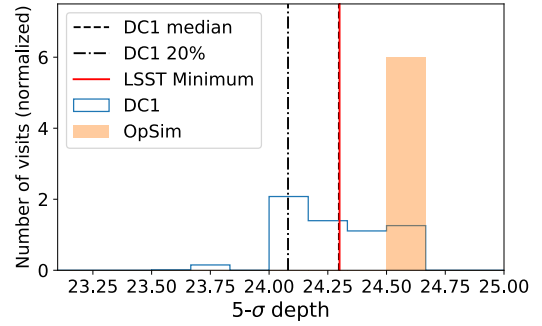
These tests show that the processing pipeline performs as needed, and that our images are generated with consistent zeropoints.

5.4 Depth requirements

The statistical power of LSST for dark energy studies relies largely on the achieved depth of the survey. A lower depth translates to an overall smaller number density of galaxies, thus increasing the statistical error budget. A careful balance between depth and area is important since a smaller survey area increases the overall statistical uncertainty. Taking this into consideration the LSST-SRD sets a minimum acceptable per-visit image depth of 24.3 mag in r-band, given a fiducial sky brightness of 21 mag/arcsec^2, exposure time of 30 s, airmass=1 and fiducial seeing (FWHM) of 0.7 arcseconds. In order to mimic this we select the visits that fulfill the following (LSST-SRD):

- Altitude > 80 degrees.
- $0.68'' <$ seeing (FWHM) $< 0.71''$.
- Sky-brightness (in r-band) ≥ 21 mag.

We obtain a total of 520 sensor-visits fulfilling these criteria. We then compute the median 5-σ depth using the magnitude errors (as described in Section 6.3) and compare with the predicted depth by OpSim. After this, we check

that the median of the depth distribution is deeper than the minimum required depth as defined in the LSST-SRD.

Clustering measurements rely on measuring changes in the number density of galaxies as a function of position to obtain cosmological information. Depth variations, usually due to changes in the sky-background levels, seeing and survey strategy (dithering), lead to variations in the observed number density and, if left unaccounted for, they can lead to biases in the inferred cosmological measurements. This is why we also check that no more than 20% of the visits have a depth lower than 24.0 by computing the lower 20th percentile in the depth distribution. The results of these checks are depicted in Figure 8. We find that the median of the depth distribution in the selected visits, 24.297 ± 0.009 mag, is compatible with the minimum of 24.3 mag. We find as well that the 20th percentile, 24.1, is larger than the minimum. We also check that in a given visit, the variation in the field of view is within the requirements. The LSST-SRD establishes that, in a representative visit no more than 20% of the field of view have a depth 0.4 mag. brighter than the nominal (24.3). We select visit 2218486 since its median depth is 24.3. We find that the 20th percentile is 24.29, fulfilling the criteria. This test demonstrates that we generate images with the required depth and in concordance with our inputs.

5.5 PSF requirements

Weak lensing measurements rely on the estimation of coherent patterns in the shapes of galaxies in order to extract cosmological information. These shapes are distorted by the PSF of the system (atmosphere+telescope+detector). A PSF with a large ellipticity module, compared to a relatively circular PSF, may have the same relative residual level, however, given the greater absolute value of the ellipticity, the absolute residual level is also higher and has a larger impact in the weak lensing signal. This is why the LSST-SRD sets criteria regarding the maximum modulus of the PSF ellipticity, $|e|$, for visits with the same criteria used for the depth requirements, mostly driven by weak lensing analyses. We
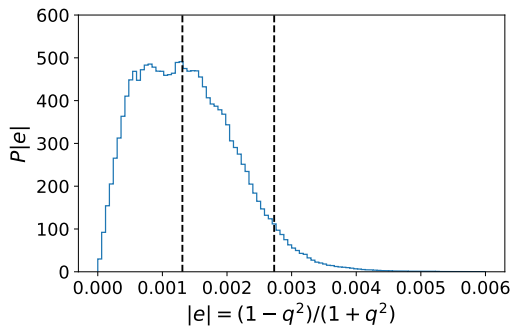
Figure 9. PSF ellipticity distribution accumulated for 520 sensor-visits measured at the positions of detected objects. The median (≈ 0.001) and 90th percentile (≈ 0.003) are shown as the dashed lines. Note that these values are an order of magnitude lower than the upper limits specified in by the LSST-SRD (0.05 and 0.1, respectively).
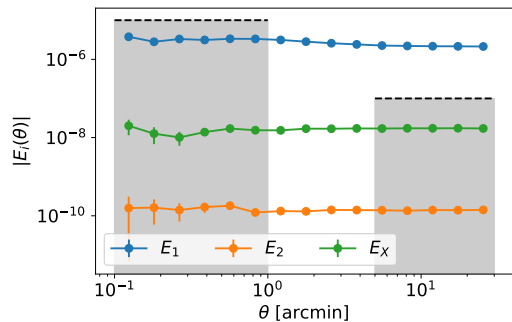

Figure 10. Auto and cross-correlation functions, E_1 (blue), E_2 (orange), E_X (green) of the PSF residuals as a function of the aperture angle θ. The median value of E_1, E_2, and E_X should be within the shaded regions in order to pass the requirements. All the correlations pass the requirements for $\theta \leq 1$ arcmin but E_1 does not meet the requirement for $\theta \geq 5$ arcmin.

use the distortion definition for $|e|$ (Miralda-Escude 1991):

$$|e| = \frac{a^2 - b^2}{a^2 + b^2}, \tag{2}$$

where a, b are the semi-major and semi-minor axes of the PSF. We test exposures with PSF-FWHM $\approx 0.69''$ and no more than 10 degrees away from zenith and check that the median ellipticity is no larger than 0.05 and that no more than 10% of the images exceed 0.1. Our analytic (and circularly-symmetric) PSF models should, by design, fulfill these criteria. However, we must test whether the reconstructed PSF also fulfills them. The PSF was reconstructed using the PSFEx (Bertin 2011) implementation in the LSST software stack. We tested this in the processed data by using the same 520 sensor-visits used to check the depth requirements described above. We checked the modulus of the PSF ellipticity at the position of detected stars, using their measured ellipticity in these visits, and accumulating them in the histogram shown in Figure 9. We obtained that the median ellipticity is ≈ 0.001 and the 90th percentile is ≈ 0.003, below the maximum values allowed by the LSST-SRD (0.05 and 0.1, respectively). The second data challenge (DC2) will use more realistic PSF models and we expect these margins to be different.

For weak lensing analyses, correct modeling of the PSF is crucial (Hirata et al. 2004) and both the LSST-SRD and the DESC-SRD specify explicit requirements about PSF residuals. In particular, the LSST-SRD states that using the full survey data the auto- and cross-correlations (E_1, E_2, E_X) of the PSF residuals over an arbitrary field of view should be below (3×10^{-5}) for $\theta \leq 1$ arcmin, and below (3×10^{-7}) for $\theta \geq 5$ arcmin.

To check these criteria we calculate E_1, E_2, E_X using

the definitions of the LSST-SRD:

$$e_1 = \frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \tag{3}$$

$$e_2 = \frac{2\sigma_{12}^2}{\sigma_1^2 + \sigma_2^2}, \tag{4}$$

$$E_1(\theta) = \langle \delta e_1^{(i)} \delta e_1^{(j)} \rangle, \tag{5}$$

$$E_2(\theta) = \langle \delta e_2^{(i)} \delta e_2^{(j)} \rangle, \tag{6}$$

$$E_X(\theta) = \langle \delta e_1^{(i)} \delta e_2^{(j)} \rangle. \tag{7}$$

The quantities σ_1^2, σ_2^2 are the second-order moments of a source along some set of perpendicular axes and σ_{12}^2 is the covariance, $\delta e_1, \delta e_2$ are the residuals, and the angle brackets indicate averaging over all pairs of stars i, j at a given angular separation θ.

In practice, we compute the PSF-corrected moments of high signal-to-noise (SNR > 100) stars across the field of view using `TreeCorr` (Jarvis et al. 2004). Our findings are shown in Figure 10. We can see that E_2 and E_X fulfill the requirements. However, E_1 does not meet the large scale requirements. This difference between E_1 and E_2 exists because E_1 is calculated in the direction of the pixels, whereas E_2 is computed in the diagonal and is subject to smaller biases. The orientation of these axes is determined somewhat arbitrarily; we could choose a different set of axes for E_1 and E_2 so both would meet the requirements. Nevertheless, PSF modeling is an area of active development within LSST Data Management and we expect improvements in future data challenges. We can compare the results in Figure 10 to the measurements in Figure 8 of Mandelbaum et al. (2018) for ρ_1 (Rowe 2010):

$$\rho_1 = \langle (\delta e_1 + i\delta e_2)(\delta e_1 - i\delta e_2) \rangle, \tag{8}$$

where i denotes the imaginary unit. So $\rho_1 \approx E_1 + E_2$. We see that both E_1 and E_2 are smaller than the measured ρ_1 for the range of scales shown in the HSC analysis by Mandelbaum et al. (2018) ($\theta \geq 3$ arcmin). However, even in the more complex case of the HSC PSF, the residual correlations for $\theta \geq 5$ arcmin are below the requirements (the published HSC work do not show the residuals for scales below 1 arcmin). Finally, the DESC-SRD requires that the
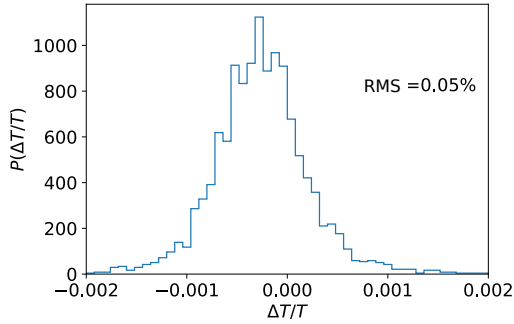
Figure 11. Normalized distribution of the relative difference in the trace of the second-order moments, $\Delta T/T$ between input and output PSF. We see that the standard deviation of the distribution is 0.05%, in compliance with the requirement (0.1%).

systematic uncertainty in the PSF model defined using the trace of the second moment matrix should not exceed 0.1% for full-depth (Y10) DESC weak lensing analysis. We randomly select 3,000 visits, obtain the input and measured PSF, and measure the trace of the second order moments, T with GalSim. We then compute the relative difference, $\Delta T/T$, obtaining the results depicted in Figure 11. We find that our dataset shows that the standard deviation of the distribution is 0.05%, lower than the requirement.

As seen in this section, our images and catalogs satisfy most of the requirements by a good margin, demonstrating our ability to both generate and process high-quality data.

6 DATA SELECTION AND MASKING

In this section, we describe how we take advantage of the fact that we have full knowledge about the simulated sources in order to get a "clean" data sample for clustering tests. We also describe the catalog mask and how we generate maps of different observational effects (seeing, sky-brightness, etc.) present in the simulation. Unless explicitly stated, the procedures and selections made in this section are performed in both the dithered and undithered simulations.

6.1 Sample selection

As we previously mentioned, the presence of sources with poorly determined fluxes, positions or shapes can affect the clustering and lensing signals. For example, if we consider the ellipticity distribution, a small fraction of sources with unrealistically large ellipticity can significantly change the inferred variance and mean of the distributions and bias any cosmological constraints. Similarly, if we consider that the positions of bright stars are uncorrelated with the positions of galaxies, the detection of spurious sources near bright stars can lead to biased clustering statistics at all scales. These are just two of the many examples that showcase the importance of having a clean, well-understood sample to perform cosmological analyses.

In this subsection we are going to use the S+M technique to identify processing flags or thresholds in variables that may allow us to get a clean sample for clustering. Given that the pipeline used for data reduction is essentially the same as the one used in Mandelbaum et al. (2018), we could

potentially perform similar cuts. However, the lack of multi-band coverage and of some catalog quantities, such as the so-called blendedness, lead us to propose our own selection cuts, although we follow the criteria in Mandelbaum et al. (2018) as guidance.

The methodology to perform the selections is simple: we check the primary detected sources that have no match using the S+M technique and we compute the fraction of objects that are flagged, $f_{u,i} = N_{\mathrm{flag_i,u}}/N_{\mathrm{total,u}}$, where the subscript u stands for unmatched, and compare it to the corresponding fraction of flagged matched primary detected sources, $f_{m,i} = N_{\mathrm{flag_i,m}}/N_{\mathrm{total,m}}$, where the subscript m stands for matched, for each of the flags, flag$_i$, in the catalog. If the ratio $f_{u,i}/f_{m,i}$ is larger than 50 for a particular flag and $f_{m,i} < 0.01$, i.e., less than 1% of the matched primary sources have that flag, it means that the presence of that flag is a good indicator of problematic sources. Thus we eliminate the sources with those flags. We also repeat the same procedure looking for the absence of a certain flag or whether a quantity is frequently measured as not-a-number, `NaN`.

We notice that some of the flags very efficiently distinguish unmatched from matched objects. For example, `base_ClassificationExtendedness_flag = True`, which means that there was a failure at the time of deciding whether a source was extended or point-like, eliminates more than 30% of the objects with no match, while barely affecting the matched objects. Three other flags would be fairly efficient at filtering out unmatched objects but, if we were to use them, we would lose $\approx$ 50% of our sample. These include `modelfit_CModel_flags_smallShape==False` which means that the initial parameter guess did not result in a negative radius (if `True` the initial guess for the radius would be negative). Intuitively, we expect this flag to be `False` for well-behaved objects and thus, we should not use this for our selection. Another case where the fraction of flagged unmatched (and matched) objects is very large is `modelfit_CModel_flags_region_usedFootprintArea==True`, which means that the pixel region for the initial fit was defined by the area of the footprint. This flag is not necessarily indicative of problems with the measured object. Finally, we see that `modelfit_CModel_flags_region_usedPsfArea==False` also affects a very large fraction of unmatched and matched objects. These objects are such that the pixel region for the initial fit was not set to a fixed factor of the PSF area, which is not indicative of any problems with the source.

As a result, we eliminate from our sample all the sources that fulfill at least one of the following conditions:

- `detect_isPrimary = False`. As discussed earlier, this means that the source has not been fully deblended or is outside of the inner region in a coadd.
- `base_NaiveCentroid_flag = True`. This means that there is a general failure during the source measurement.
- `base_SdssShape_flag_psf = True`. This means that there is a failure in measuring the PSF model shape in that position.
- `ext_shapeHSM_HsmSourceMoments_flag_not_contained = True`. This means that the center of the source is not contained in its footprint bounding box.
- `modelfit_DoubleShapeletPsfApprox_flag = True`.

This means that there is a general failure while performing the double-shapelet approximation to the PSF model at the position of this source (see Appendix 2 in Bosch et al. (2018) for more details).

• `base_PixelFlags_flag_interpolated = True`. This means that there are interpolated pixels in the source's footprint.

• `base_PixelFlags_flag_interpolatedCenter = True`. This means that the center of a source is interpolated.

• `base_PixelFlags_flag_saturatedCenter = True`. This means that the center of a source is saturated.

• `base_ClassificationExtendedness_flag = True`. This means that there is a general failure when using the `extendedness` classifier.

• `modelfit_CModel_flags_region_usedInitialEllipseMin = True`. This means that the pixel region for the final model fit is set to the minimum bound used in the initial fit.

• `base_SdssShape_x/y = NaN`. This means that the centroid position (either in the x or y axes) is measured as `NaN`.

• `base_SdssCentroid_x/yErr = NaN`. This means that the error in the centroid position (either in the x or y axis) is measured as `NaN`.

After these cuts we keep 8.25 million objects in the dithered catalog and 7.51 million objects in the undithered catalog. We will refer to this sample as the *benchmark sample*. After the cuts, the ratio of unmatched objects decreases by $\approx 60\%$ from $\sim 5\%$ (of the catalog before cuts) to $\sim 3\%$ (of the benchmark sample), while we retain 99.7% of the matched objects. In DC1 we are limited by only having r-band information. The addition of other imaging bands will provide independent information that will decrease the fraction of noise fluctuations that make it into the catalog and will allow selection of an even cleaner sample (e.g., by performing selection cuts in color-color diagrams).

We now focus on how many of these objects are matched as a function of magnitude and signal-to-noise ratio. In Figure 12, we can see that the fraction of unmatched objects grows very quickly for $r > 26$, and for SNR< 6. Therefore, $r < 26$ and/or SNR> 6 appear to be sensible selection criteria to ensure good quality data. The peak at $r \sim 15$ is likely a consequence of bright objects that have been clipped in the simulation process to emulate saturation; these objects have measured magnitudes of $r \sim 15$, however, their true magnitudes are brighter by one or more magnitudes and thus, appear as unmatched by the S+M matching strategy.

We select our *clustering sample* using the following criteria:

• Full set of cuts of the benchmark sample.
• $20 \leq$ `r_mag_CModel` ≤ 25.5.

Note that we do not use the SNR> 6 criterion, this is because the magnitude cut nearly guarantees it. These selection cuts ensure a low fraction of unmatched objects and also, as we will justify in following sections, high purity of our galaxy sample (low stellar contamination) once we select galaxies via the `base_ClassificationExtendedness_value` classifier.
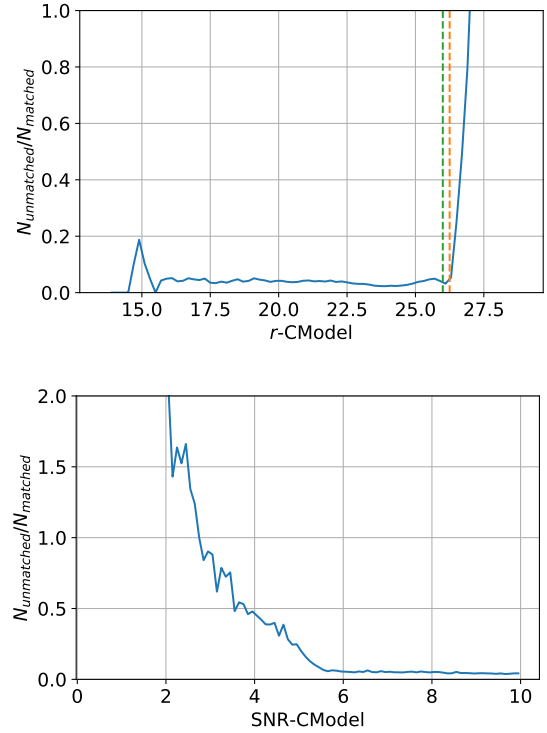
Figure 12. Top: Ratio of unmatched to matched sources in the benchmark sample as a function of magnitude. The dashed vertical lines show the median depths for the dithered (orange) and undithered (green) fields. Bottom: Ratio of unmatched to matched sources in the benchmark sample as a function of SNR.
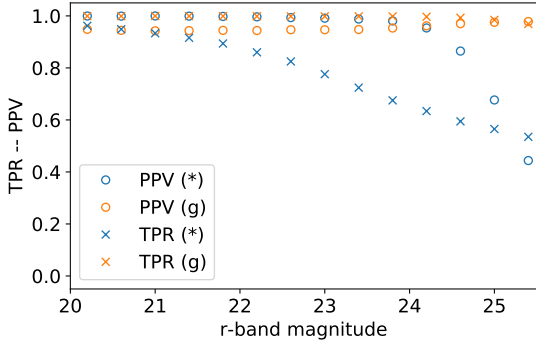
6.2 Star/galaxy classification

For weak lensing and clustering analyses, it is important to have a pure galaxy sample and good control over the fraction of stars that are classified as galaxies. Our pipeline includes the variable `base_ClassificationExtendedness_value` (see Bosch et al. (2018) for more details), which we will refer to as `extendedness`, and can be used as a proxy to separate stars from galaxies as shown in Mandelbaum et al. (2018) and Bosch et al. (2018). In this work we say that an object has been classified as a galaxy if `extendedness=1`, and that the object has been classified as a star if `extendedness=0`. To evaluate the performance of `extendedness` as star/galaxy classifier in DC1, we use the clustering sample in the dithered field, although we find similar results using the undithered field, and match it to our input catalog using the S+M method. After that, we follow Sevilla-Noarbe et al. (2018) and compute the true positive rate (TPR), usually referred to as *completeness*, and positive predictive value (PPV), usually called *purity*, defined as:

$$\text{TPR} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}, \quad (9)$$

$$\text{PPV} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}. \quad (10)$$

For galaxies, true positives are objects classified as galaxies and matched to galaxies; false negatives are objects classified as stars but matched to galaxies; and false positives are objects classified as galaxies but matched to stars. This

Figure 13. Performance of the `base_-ClassificationExtendedness_value` as star-galaxy classifier as a function of magnitude. This classifier achieves a high galaxy purity and completeness. We show the true positive rate (TPR) for stars (blue crosses) and galaxies (orange crosses) and the positive predictive value (PPV) for stars (blue open circles) and galaxies (orange open circles). Both the TPR and PPV for galaxies are high in the range of magnitudes that we are studying.

way, we know the total stellar (or galaxy) contamination as a function of measured magnitude. The results are depicted in Figure 13. We see that at the fainter end ($r \approx 25$), the PPV (purity) of the stellar sample using the `extendedness` classifier starts to decrease, getting as low as 50% for the last bin in our analysis, while the PPV for the galaxy sample remains stable across the selected range of magnitudes. We obtain a total $f_{star} = 1.4\%$. For a more restricted magnitude threshold of $r < 25$ we can get $f_{star} = 0.7\%$. Note that these fractions are larger than those presented in Bosch et al. (2018). This is primarily because of our broader PSF, making the `extendedness` classifier perform a little bit worse, and that we do not include any cuts on resolution. However, this level of stellar contamination is acceptable for the purposes of this work. Note that this selection focuses on galaxy purity and completeness and should be modified for stellar studies.

6.3 Depth maps and footprint masking

In order to estimate the depth in the coadd catalogs we generate a map of our footprint in flat-sky approximation (i.e., using the plate carrée projection), with resolution of 1.74 arcminutes, containing the sources in the *benchmark sample*. This resolution allows us to accurately estimate the power spectra up to $\ell \sim 6000$, where the power spectra will be mostly dominated by shot-noise. Then, for each cell in the map, we bin the objects in magnitude and compute the median SNR in each bin, after this we find the magnitude bin closest to SNR=5, using `CMODEL` fluxes and their quoted uncertainties.

These maps are shown in Figure 14. We can see that the dithered simulation is indeed very uniform ($> 50\%$ of its footprint lies in the same depth bin) showing the success of the dither strategy. Additionally, we can see that the undithered simulation has a smaller median depth.

We also check the depth by computing the detection
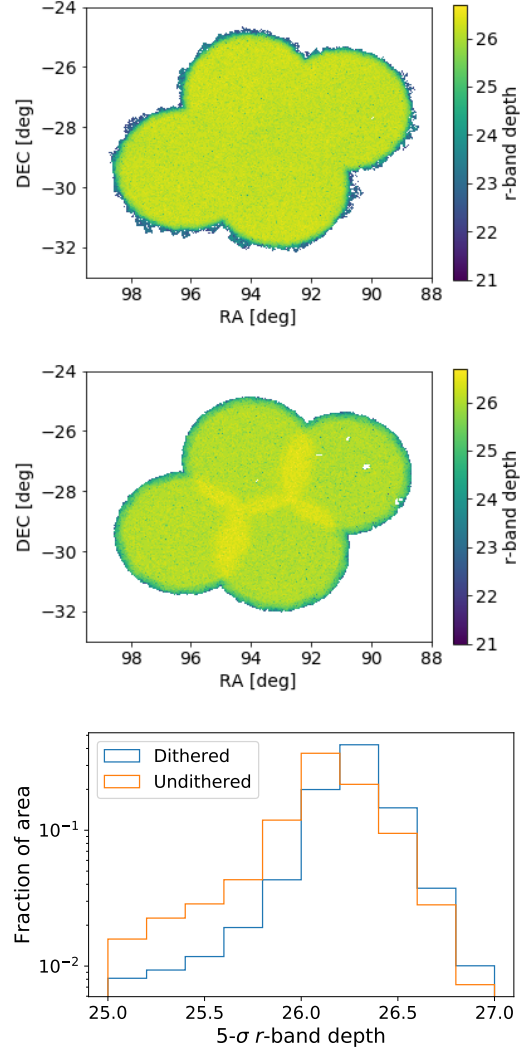
Figure 14. 5-σ depth maps for the dithered (top) and undithered (middle) fields. There is an increased depth in the overlapping parts of the LSST field of view in the undithered field but the median depth is lower. We see some holes in the undithered footprint due to missing data. The maps have a resolution of 1.74 arcmin. We also show the 1D distributions of depth (bottom) for both fields for easier comparison.

efficiency (completeness[19]) of stars and galaxies as a function of magnitude. To do so, we use the objects in the input catalog and select those that lie within the simulated footprint. After this, we compute the number of detected objects in the *clustering sample* classified as stars and classified as galaxies as a function of their `CMODEL` magnitude, and divide by the number of stars and galaxies in the input catalog as a function of the true magnitude. The results can be seen in Figure 15. We see that there is a high detection efficiency for galaxies $> 80\%$ up to $r \approx 25.5$.

Given the results in the two previous subsections and this subsection, we decide to use only the galaxies that lie

---

[19] Note that completeness is defined differently in the star/galaxy classification section 6.2.
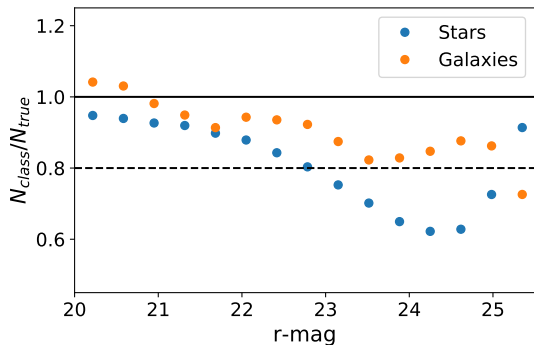
Figure 15. Ratio of number of objects in the clustering sample classified as stars and number of input stars as a function of magnitude. We also compute this same ratio for galaxies. This is basically a measurement of the detection efficiency (completeness) of stars and galaxies in the clustering sample. Note that the ratio can go above one since there will be stars classified as galaxies (and galaxies classified as stars) and artifacts that will pass our cleaning cuts.

in cells with limiting magnitude $r \geq 25.5$ and that have been visited at least 92 times, which corresponds to 50% of the nominal full-depth number of visits for the full 10-year LSST (Ivezić et al. 2019). On top of that, we select those objects with magnitudes in the range $20 \leq r \leq 25.5$. This cut ensures high detection efficiency ($> 80\%$) and it allows us to eliminate most of the spurious detections in the sample (2.8%). In addition it results in a low stellar contamination ($f_{star} \approx 1.4\%$). After these cuts and with selection of objects with `extendedness=1`, we obtain 4.5 and 4.0 million objects for the dithered and undithered fields respectively. This selection cut, however does not change the fraction of unmatched objects explored in the previous subsection.

6.4 Bright star masking

Bright objects produce significant effects in an image that affect the detection and measurement of neighboring objects. Some examples of these effects include saturation, large diffraction spikes (not included in our simulations), scattered light (also not included in our simulations), obscuration of neighboring sources. Masking regions around these sources creates a more complicated footprint but greatly simplifies the analysis of systematic effects. In order to avoid possible biases by masking bright galaxies we will only analyze the impact of bright stars on the nearby detected objects.

In order to evaluate the effect of bright star masking, we follow the procedure described in Coupon et al. (2018). Using the positions of bright objects classified as stars (`base_ClassificationExtendedness_value==0`) that lie within the considered footprint, and with input magnitudes in the range $m_1 < r < m_2$, we count all objects from the clustering sample within a given radius θ and compute the average number of neighbors, $N_{neighbors}$. We repeat this for different radii and magnitude ranges ($r < 17; 17 \leq r < 18; 18 \leq r < 20; 20 \leq r < 22$). Finally, we repeat this process for all stars in the input catalog in the footprint and compute $N_{neighbors,tot}$ and compute the ratio
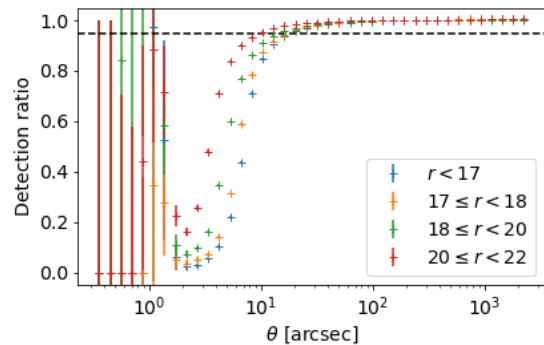
Figure 16. Ratio of the median number of primary detected objects neighboring a star in a certain magnitude range in the input catalog to the median number of objects detected near any star in the input catalog, as a function of the distance to the star θ. Different colors represent different magnitude ranges for the stars in the input catalog considered.
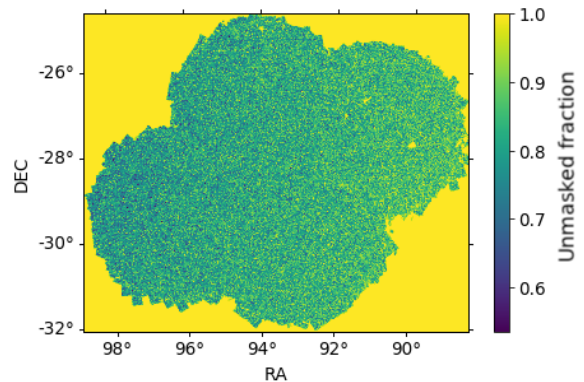
Figure 17. Map showing the unmasked fraction in each pixel. We use a high-resolution ($\approx 6.4''$) map to mask around bright stars ($r < 22$) and then we down-sample the map to a lower resolution ($\approx 1.7'$) and remove the pixels where the masked area near bright stars is higher than 75% of the pixel. We weight the galaxy counts in each pixel of the map by the inverse of this mask fraction to compensate for the area loss.

$N_{neighbors}/N_{neighbors,tot}$. This ratio is depicted in Figure 16 as a function of distance. We then generate the bright star mask as follows:

(i) In Figure 16 we identify the radius $\theta = r_{mask,fit}$, at which $N_{neighbors}/N_{neighbors,tot} = 0.95$ for each magnitude range.

(ii) We mask around bright stars using the value of $r_{mask,fit}$ corresponding to their magnitude, creating a high-resolution mask ($6.4''$-side pixels).

(iii) We downgrade the resolution of this mask to ≈ 2 arcmin and eliminate pixels that are more than 75% masked. This results in an area loss of $\approx 13\%$.

The resulting mask can be seen in Figure 17.

6.5 Blending

As previously mentioned, our output catalogs do not include any estimates of overlap between sources, or *blended-*

ness ([Bosch et al. 2018](#)). Highly-blended objects are more likely to have biased estimations of the centroid positions, shapes, and fluxes. This can lead to overall biases in the estimated photometric redshifts and cosmological parameters. Given that the mean seeing in DC1 is larger than in HSC, the impact of blended objects will be larger. [Mandelbaum et al.](#) ([2018](#)) mitigated the impact of blended objects by removing objects with more than 43% of their flux in their footprint coming from overlapping sources (blendedness $> 10^{-0.375}$), which affects only 1% of the objects; we expect this number to be larger for the DC1 simulations. Using dedicated image simulations from Sánchez et al., in prep. with a seeing similar to the seeing in DC1 ($1.04''$), in r-band, we find that if we select objects with $r < 25.5$ and SNR ≥ 1, the fraction of objects with blendedness $> 10^{-0.375}$ is $\approx 6.3\%$. If we raise the minimum SNR threshold to 6, this fraction is lowered to $\approx 2.6\%$. This means that our sample will have a fraction of these objects anywhere in the range (2.6% – 6.3%) but closer to 2.6% since the fraction of objects with SNR ≤ 6 is $\approx 0.3\%$. In any case, we do not expect that the inclusion of these objects in our two-point measurements affects the range of scales that we consider in this work. However, blending affects photometric redshifts, potentially biasing them. For DC1 photometric redshifts are not available (since we only have one band). A more rigorous study of the impact of blending on small-scale clustering measurements, and photometric redshifts is beyond the scope of the current work.

7 TWO-POINT CLUSTERING RESULTS

In previous sections we have focused on how to select a galaxy sample clean enough to perform a clustering analysis, our clustering sample. Now we want to use this sample to answer two questions: (i) is this sample actually good enough to perform a clustering measurement? (ii) given that reducing the number of objects in our catalog in order to have a cleaner sample results in an increased statistical uncertainty, we would like to understand if these cuts are actually necessary or can be relaxed. We answer these questions by measuring the angular power spectra of two different samples: (i) our so-called "clustering sample", (ii) a "polluted" sample, where we relax some of the selection criteria. Beyond these questions, these clustering measurements help us verify the simulation and the clustering analysis pipelines.

Apart from analyzing the impact of selection in our two-point clustering results, we also study how different observing conditions correlate with the observed number density in our sample, and whether the dithering strategy used in the simulation mitigates these correlations. We are going to consider maps of the following observing conditions:

• Extinction: The CatSim catalog provides the value for the magnitudes corrected for extinction using the map from [Schlegel et al.](#) ([1998](#)), which we refer to as the SFD map.

• Stellar contamination: In this case, we build a flat-sky map with all stars in the input catalog.

• Sky-background/Sky-brightness: We use the observed background level in each exposure and assign that value to the pixels in the flat-sky map that lie within that exposure. After this we calculate the mean value in each pixel to build the map with the same resolution as the mask (≈ 2 arcmin).

• Sky-noise: We use the observed noise background level in each exposure and proceed as in the previous case to build a map.

• Seeing: We proceed as before and use the observed seeing in each exposure and build a map.

• Number of visits: We count the number of exposures overlapping with each pixel of our flat-sky maps.

These maps are shown in Figures [18](#) and [19](#). We see that the spatial distributions of the different observing conditions are very different between the two simulations, even though the ranges in each of the observing conditions are very similar.

The power spectra computation is performed using `NaMaster` ([Alonso et al. 2019](#)). The systematics correction is also performed with `NaMaster` via mode deprojection ([Elsner et al. 2016](#); [Alonso et al. 2019](#)), which assumes that there is a linear dependence between the observed number density of galaxies and the contaminants. For our study, we compute the density contrast, $\delta_{g,i}$, in a given cell/pixel, i, of our maps as follows:

(i) Ignoring the area lost due to the presence of nearby bright stars, i.e., $\delta_{g,i} = \rho_i/\langle\rho\rangle - 1$.

(ii) Compensating for the area lost due to the presence of nearby bright stars, i.e., $\delta_{g,i}^* = \rho/w_i/\langle\rho/w_i\rangle - 1$.

where ρ_i is the density in a given pixel of the map, w_i is the inverse of the completeness mask, and $\langle\cdots\rangle$ represent ensemble averages. We choose $\Delta\ell = 352$ and compute the power spectra in the range $0 \leq \ell \leq 6000$. This choice for $\Delta\ell$ is not optimal for cosmological analyses, but it gives us a reasonably large number of bandpowers to visually check the estimated power spectra. The results for the power spectra are shown in Figure [20](#), where we can see that both the dithered and undithered catalogs yield similar results. The error bars shown and corresponding covariances are estimated using two complementary methods: On one hand by computing the Gaussian covariance with `NaMaster`, and on the other hand by considering 155 jackknife equal-area regions in our footprint. We find that both approaches give similar results and we choose to use the results from the jackknife computation.

In Figure [20](#) we also compare with the theoretical prediction for the power spectra computed with `CCL` ([Chisari et al. 2019](#)) as follows:

$$C_\ell^{\mathrm{TH}} = \frac{2}{2\ell+1} \int dz \left(\frac{dn(z)}{dz}\right)^2 b^2(z) H^2(z) \times \\ P\left(k = \frac{\ell+1}{r}, z\right), \tag{11}$$

where $P(k, z)$ is the theoretical power spectrum, $b(z)$ is the galaxy bias and $\frac{dn}{dz}$ is the number density as a function of redshift, and $H(z)$ is the Hubble parameter. In particular, we use the Halofit ([Takahashi et al. 2012](#)) power-spectrum with the Millenium cosmological parameters ([Springel et al. 2005](#)) ($\Omega_m = 0.25$, $\Omega_b = 0.045$, $\Omega_\Lambda = 0.75$, $n = 1$, $\sigma_8 = 0.9$, $h = 0.73$), and the dn/dz obtained using the true redshifts of galaxies matched in the input catalog. We use a bias, $b(z) = b_0/D_+(z)$, inversely proportional to the linear growth factor ([Peebles 1980](#)), $D_+(z)$, and see that, qualitatively speaking, there is a good agreement between the measurements and the prediction, as shown in Figure [20](#). We obtain a best-fit value for $b_0 = 0.43$ using the scales $\ell > 1/f_{sky}$,
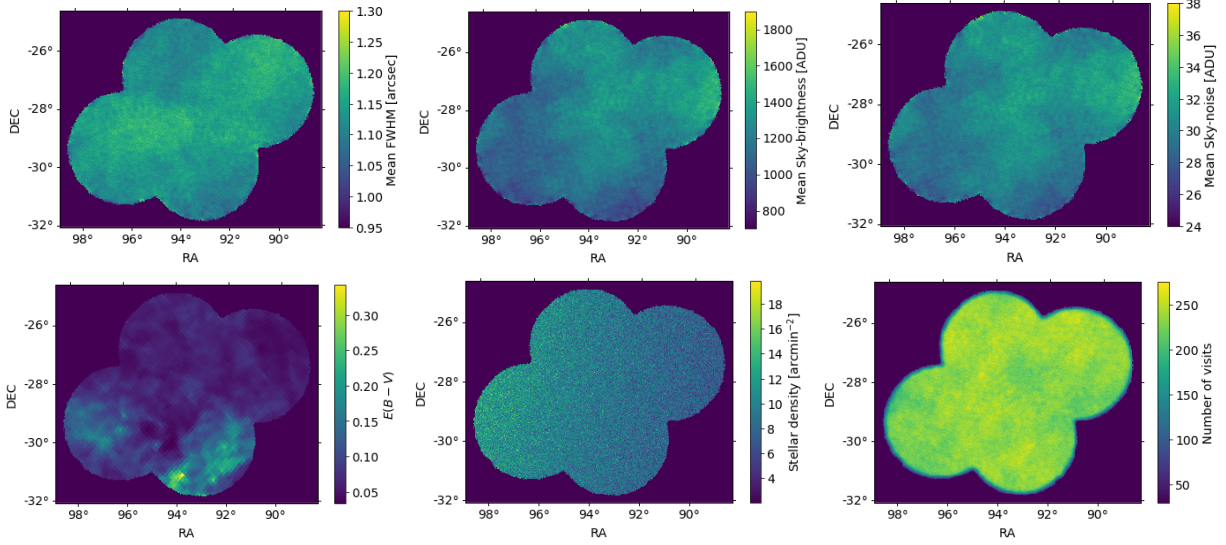
Figure 18. Maps showing the different foregrounds considered in our analysis of the dithered field. From top left to bottom right: Mean PSF FWHM, mean sky-brightness, mean sky-noise, mean extinction, stellar density and number of visits in each pixel in the flat-sky maps with the same resolution as the depth maps in Figure 14. We only show their values in the regions where the 5-σ r-band depth is larger than 25.5.
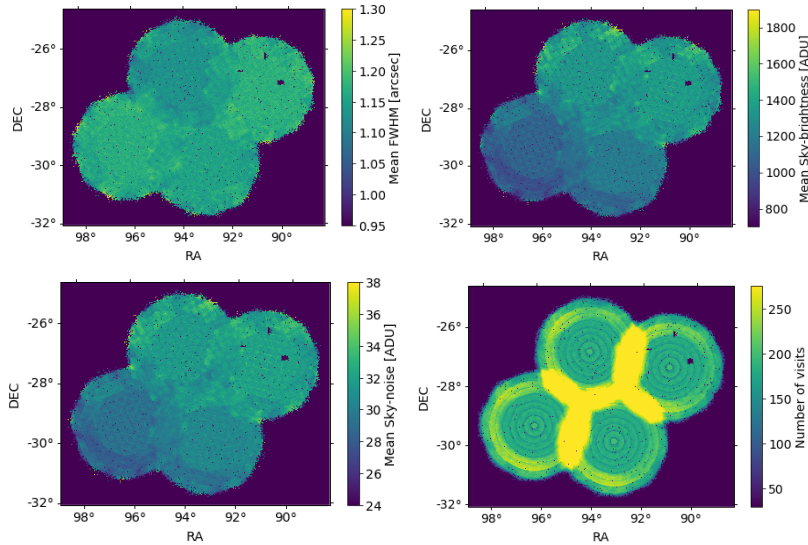

Figure 19. Same as Figure 18 for the undithered dataset. From top left to bottom right: Mean PSF FWHM, mean sky-brightness, mean sky-noise and number of visits. The maps are only shown in regions where the 5-σ r-band depth is larger than 25.5. Note that we use the same extinction and stellar density maps changing the geometry of the mask since these are not affected by dithering. The circular pattern is due to the fact that the field rotates slightly in the undithered visits.

which are well into the non-linear regime ($k \gtrsim 0.9$ Mpc^{-1} at the mean redshift of the clustering sample $\bar{z} \approx 1.51$). Additionally, we obtain the best-fit bias value, $b_0 = 0.38$, for the scales $\ell < 1/f_{sky}$, where this linear bias model is expected to be more accurate. The low value obtained for the galaxy bias (in both regimes) is a consequence of the addition of randomly positioned (i.e., non-clustered) galaxies at the fainter end ($r > 25$) in order to match the expected number density for LSST (Connolly, private comm.) that we mentioned in Section 2.1. In any case, we do not expect to be able to fully describe the measured power spectra, given the highly nonlinear nature of the scales considered in our

analysis. Unfortunately, due to the limited of area we cannot further explore the differences between the linear and non-linear regimes, but we expect to be able to study the non-linear regime in future data challenges.

In addition, we also see in Figure 20 that the overall impact of the systematics, evaluated as the difference between the estimated power spectra using mode deprojection and without mode deprojection, is smaller than the statistical uncertainty, σ_ℓ, (about 50% the size of σ_ℓ) and that they similarly affect both dithered and undithered simulations. We do not find any statistically significant difference between the correction due to systematics for the dithered
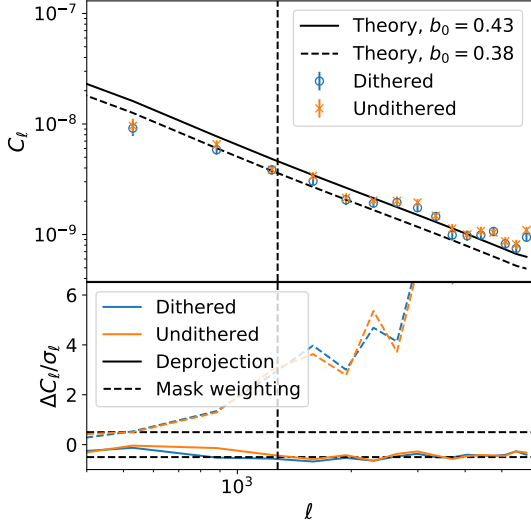
Figure 20. Top panel: Measured power spectra for undithered (orange ×) and dithered (open blue circles) datasets. These power spectra include the correction due to systematics using mode deprojection with `NaMaster`. The error bars are computed using jackknife. A theoretical prediction calculated with `CCL` is shown as the solid black line to demonstrate the overall qualitative agreement between the measurements and the data, we show the best-fit bias results using $\ell < 1/f_{sky}$ only ($b_0 = 0.38$, dashed line), and using $\ell > 1/f_{sky}$ only ($b = 0.43$, solid line). The vertical black dashed line corresponds to $\ell = 1/f_{sky} = 1192$ ($k \sim 0.9$ Mpc^{-1} at the mean redshift of the clustering sample $\bar{z} \approx 1.51$) which is the minimum ℓ that we are sensitive to given our footprint size. **Bottom panel:** Size of the correction in the power spectra, ΔC_ℓ, relative to their uncertainty, σ_ℓ, due to deprojection (solid lines) of different observing conditions, and the correction due to weighting by the bright star mask (dashed lines) for the dithered (blue) and undithered (orange) simulations. We see that the largest impact comes from the presence of bright objects, and that it is important to account for the area lost by masking via weighting. The horizontal dashed lines correspond to a correction of 50% of the statistical uncertainty, to provide visual guidance. The shadowed area in both panels corresponds to the region where $\ell < 1/f_{sky}$ which is not used for our analysis.

and undithered simulations. This is a consequence of several factors including: (i) the conservative cuts that we impose on our data to ensure well-behaved clustering statistics; (ii) that we only deproject using the mean value for the different observing conditions; and the lack of effects, such as vignetting, present in real images. For example, vignetting would affect the number of detected objects close to the edges of the focal plane in the undithered simulation, reducing the uniformity of the survey. However, this effect would be uniform across the footprint in the dithered case. In addition to this, if we decided to include results for fainter magnitudes by going deeper, the lack of uniformity of the undithered field would enhance the impact of the observing conditions over the clustering signal. We can also see that, in the ℓ range considered in our analysis, the presence of bright stars is the dominant systematic effect. In the close neighborhood to bright stars our ability to detect faint sources diminishes. These faint sources are blended in the core or the tails of the brighter objects resulting in a lower mean number

of detected sources, as shown in Figure 16. The correction by weighting by the fraction of area covered in each pixel has a considerable impact at small-scales, being larger than the statistical uncertainty in this regime. Note that even though we restrict our analysis to pixels that have a relative area loss of less than 25% the effect is very significant (more than 4 times larger than the statistical uncertainty at the scales considered). This showcases again the importance of considering the impact of blending in the small-scale regime for LSST and this issue should be carefully studied in future Data Challenges. The correction due to the presence of bright stars is comparable in both simulations. We expect that for future versions of the LSST Science Pipelines this effect will be smaller due to improvements in the deblending and measurement algorithms.

Throughout this work we have used a well-behaved clean sample (our clustering sample) to demonstrate that we can successfully perform clustering measurements. However, one may think that the requirements for our sample are too restrictive. Some of the requirements described in Section 5 are indeed beyond what clustering analyses need and, as we mentioned, are driven by other science cases that will be studied with data from LSST. If artifacts are not carefully removed from the galaxy sample, we obtain significantly different clustering results. The selection cuts performed in Section 6 were motivated by the reduction of the fraction of measured sources that could not be matched to the input catalog, on top of the usual selection of a flux limited sample to ensure good uniformity across the footprint. We analyze now the power spectrum of galaxies that fulfill only the two following conditions: `r_mag_CModel` ≤ 25.5 and `base_ClassificationExtendedness_value` $= 1$. Essentially, we are adding unmatched objects to our clustering sample. We will call this sample the polluted sample. It contains 1.5% more objects than the clustering sample. We proceed as in the previous section and compute the power spectrum deprojecting systematics and taking into account the bright object mask. We obtain the results shown in Figure 21, where we can see that, although the difference in the number of objects is very small, the difference in the measured power spectra is at the $\sim 2\sigma$ level for $\ell > 3,000$. In fact, we compute $\chi^2 = \Delta C_\ell C_{\ell\ell'}^{-1} \Delta C_{\ell'}$, with $\Delta C_\ell = C_{\ell,\text{final}} - C_{\ell,\text{raw}}$ for $\ell, \ell' > 1/f_{sky}$, and obtain 24.1 for 13 degrees of freedom, which means that the probability of both measurements being statistically compatible is $\sim 3\%$. Note that if we rescale the covariance to the full expected area of LSST useful for large-scale-structure studies, $\sim 14,300$ deg^2 (The LSST Dark Energy Science Collaboration et al. 2018); the χ^2/ndof becomes $\sim 9,800/13$. This means that these power spectra would be statistically incompatible, given the statistical power of LSST. In addition, we checked that these differences are not due to the change in the overall redshift distribution, $N(z)$, introduced by the newly added objects with `r_mag_CModel` < 20. These objects do change the overall $N(z)$ at the $\sim 1\%$ level. However, using the theoretical prediction given the input simulation parameters, we find that these changes lead to a $\Delta C_\ell/\sigma_\ell < 1\%$ which is much smaller than the effect seen here. This, together with the different behavior in the power spectrum of the "contaminants" shown in Figure 21, demonstrates that the objects that we rejected by our sample selection were, indeed, arti-
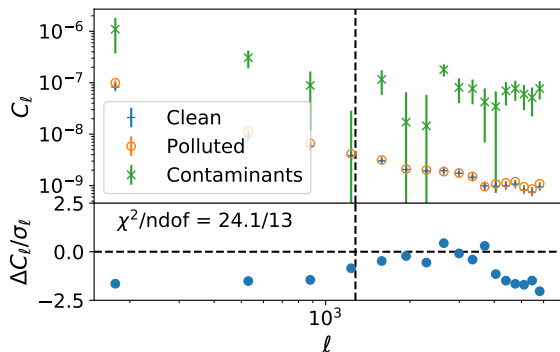
Figure 21. Top: Power spectra of the clustering sample (blue +), the polluted sample (orange open circles), and the contaminants by themselves (green ×). Error bars are computed using a jackknife procedure. Bottom: Difference between the power spectra, $\Delta C_\ell = C_{\ell,\text{final}} - C_{\ell,\text{raw}}$ relative to the statistical uncertainty in the power spectra, σ_ℓ. The vertical dashed line represents the minimum multipole used to compute the different $\ell = 1/f_{sky}$. The horizontal dashed line is drawn at zero to help visualize the difference between the power spectra. The shadowed area corresponding to $\ell < 1/f_{sky}$ is not used for our analysis.

facts that can lead to biased clustering results at small scales such as those under study in this work.

8   CONCLUSIONS

End-to-end simulations are powerful tools for testing the overall performance of current and future cosmological experiments like LSST. They allow us to validate and improve various aspects of the data processing and analysis, as well as to model and improve our control of systematic uncertainties. The access to ground truth allows us to test certain aspects of the processing and analysis pipelines that would be otherwise very challenging to test with real data (e.g., impact of undetected sources in the fluxes of detected overlapping sources).

In this paper, we present an end-to-end simulated imaging dataset that resembles single-band, full-depth (10-year) LSST data (for the wide-fast-deep survey), for the first data challenge, DC1, in the LSST DESC. This dataset was generated by synthesizing sources from cosmological N-body simulations in individual sensor-visit images with different observing conditions. Two separate runs of this dataset were generated with different dither strategies, the dithered run and the undithered run. The images from both runs were processed with the LSST Science Pipelines. We find $\sim 10\%$ more objects in the dithered simulations due to their ~ 0.1 magnitude deeper median depth. We perform several quality assurance tests on the resulting data products. Both datasets pass the tests performed.

We study different ways to relate the output catalogs to the inputs: The first method uses information about positions only, and the second involves both positions and magnitudes. For clustering analyses, adding information about magnitudes results in a lower incidence of spurious matches and is sufficient for DC1. However, neither of these methods provides a noiseless match between inputs and outputs.

The usage of matching strategies helps us define a galaxy sample suitable for clustering analyses. After cleaning the catalog, we find a small fraction ($\approx 3.6\%$) of artifacts, i.e., objects with no counterpart in the input. We anticipate that additional information from multi-band coverage and photometric redshifts, will help us to further refine the selection.

We perform a two-point clustering analysis of the simulated data. The results of this analysis indicate that the simulated foregrounds have a low impact, smaller than the statistical uncertainty, in both datasets. This is probably due to the simplicity of our foregrounds, and more complexity will be added in future Data Challenges. We do not find statistically significant differences in the impact of systematics between the dithered and undithered datasets given the area of DC1. We also see that for $\ell > 1150$ the presence of bright objects has a larger impact on the power spectra, $\approx 200 - 600\%$ of the statistical uncertainty, highlighting the impact of masking and blending with bright stars in LSST for small-scale analyses. We also demonstrate that the careful selection performed is necessary given that the presence of even a small fraction of artifacts in the sample biases the clustering results significantly.

Finally, we have been able to perform an end-to-end test of our processing and analysis pipelines. The methodology presented in this work will serve as the basis for future DESC Data Challenges, where we will aim to perform multi-band studies in a larger area, use complementary image generation strategies (PhoSim), and increase the complexity of the foregrounds included.

APPENDIX A: LSST SRD REQUIREMENTS

In this section we summarize the different requirements that we test for from the LSST Science Requirements Document version 11 that can be found at https://docushare.lsst.org/docushare/dsweb/Get/LPM-17. These requirements are driven by different science cases of LSST (for example, astrometric requirements are driven by transient science even though accurate astrometry is also important for weak lensing and galaxy clustering although not at the same level of accuracy).

A1   Astrometric requirements

• AA1: Astrometric accuracy check. Minimum absolute astrometric accuracy. We compute it as the median of the difference between input and measured centroid positions. The maximum median value is 100 mas.

• AMx: Astrometric repeatability check. Maximum RMS of the separation between pairs of stars separated by 5 arcmin (AM1: 20 mas), 20 arcmin (AM2: 20 mas), and 200 arcmin (AM3: 30 mas).

• AFx: Astrometric repeatability check. Maximum outlier fraction that deviate more than 40 mas (50 mas for AF3) for the separation between pairs of stars separated by 5 arcmin (AF1: 20 %), 20 arcmin (AF2: 20%), and 200 arcmin (AF3: 20%).

KPM/Requirement	Pass/Fail Criterion	DC1 test result	Section	Figure
AA1 (milliarcsec)	100	20	5.1	5
AM1 (milliarcsec)	20	8		
AF1 (%)	20	13		
AM2 (milliarcsec)	20	4		
AF2 (%)	20	9		
AM3 (milliarcsec)	30	7		
AF3 (%)	20	2		
PA1 (millimag)	8	6	5.2	6
PF1 (%)	20	17		6
PA3 (millimag)	15	0.06	5.3	7
PF2 (%)	20	0		7
PA6 (millimag)	20	17		7
D1 (mag)	24.3	24.3	5.4	8
Z1 (mag)	24.0	24.1		8
DB1 (mag/r-band)	24.3	24.3		8
Z2 (mag)	0.4	0.1		
SE1	0.04	0.001	5.5	9
SE2	0.1	0.003		9
SR1 (arcsec)	0.80	0.64		
SR2 (arcsec)	1.31	1.01		
SR3 (arcsec)	1.81	1.79		
TE1	3×10^{-5}	3×10^{-6}		10
TE2	3×10^{-7}	9×10^{-8}		10
WL4-Y10 (%)	0.1	0.05		11

Table A1. Summary of Key Performance Metrics (KPMs) and DESC-SRD requirements that we check to test the quality of our end-to-end simulation pipeline. Pass/Fail criterion contains the threshold pass/fail values for each KPM/Requirement and DC1 test result shows the measured value for a given KPM/requirement. We are only able to test one criterion of the DESC-SRD, which is the uncertainty in the trace of the second order moments of the PSF for 10-year depth images (WL4-Y10)

A2   Photometric requirements

• PA1: Photometric repeatability check. Maximum RMS/IQR of the magnitude distribution of objects between different visits. The maximum value allowed is 8 millimags.
• PF1: Photometric repeatability check. Maximum outlier fraction that deviate more than 15 millimag (PA2) from the mean measured magnitude. The maximum outlier fraction allowed is 20%.
• PA6: Photometric accuracy check. Minimum absolute photometric accuracy. We compute this as the median difference between the input and measured fluxes for stars. The maximum allowed is 20 millimag.

A3   Zeropoint uniformity requirements

• PA3: Zeropoint error uniformity. Maximum allowed for the RMS of the photometric zeropoint error. The maximum allowed is 15 millimags.
• PF2: Zeropoint error uniformity. Maximum outlier fraction that deviate more than 15 millimag (PA4) in the zeropoint error distribution. The maximum allowed is 10%

A4   Depth requirements

• D1: Minimum depth check. Minimum value for the median of the 5σ r-band depth for single visits with seeing 0.7 arcseconds, airmass 1.0 and 30 seconds exposure time. The minimum allowed is 24.3.

• Z1: Minimum depth check. Minimum value for the 20-th percentile (DF1) of the 5-σ r-band depth distribution for single-visits with seeing 0.7 arcseconds, airmass 1.0, sky-brightness fainter than 21 in r-band, and 30 seconds exposure time.
• DB1: Minimum depth check. DB1 is effectively the same requirement as D1, generalized to other bands (in the case of DC1 it is exactly the same as D1).
• Z2: Minimum depth check. Maximum variation within the field of view for the brightest 20-th percentile (DF2) of the depth in a representative single visit. The maximum variation allowed is 0.4.

A5   Image quality requirements

• SE1: Maximum PSF ellipticity check. Maximum median value of the PSF ellipticity modulus. The maximum allowed is 0.05.
• SE2: Maximum PSF ellipticity check. Maximum value for the 90th percentile (EF1) of the PSF ellipticity modulus. The maximum allowed is 0.1.
• SRx: Image quality test. Minimum radii to encircle at least 80% (SR1), 95% (SR2), and 99% (SR3) of the flux for a fiducial delivered seeing of 0.69 arcseconds. The values are 0.80, 1.31, and 1.81 arcseconds for SR1, SR2, and SR3 respectively.
• TE1: PSF ellipticity residual correlation check. Maximum value for the median PSF ellipticity correlations

E_1, E_2, E_3 defined in equations (8-10) for $\theta \leq 1$ arcmin. The maximum allowed is 3×10^{-5}.

- TE2: PSF ellipticity residual correlation check. Same as TE1 but for $\theta \geq 5$ arcmin. The maximum allowed is 2×10^{-7}.

Author contributions are listed below:

J. Sánchez: led study.

C. W. Walter: DC1 production planning and oversight, led imSim development and image generation.

A. Slosar: Participated in analysis and preliminary tests.

D. Kirkby: Participated and advised in analysis.

J. Chiang: One of the main developers of imSim, participated in image generation and data distribution.

T. Glanzman: Generated artificial images.

S. F. Daniel: Developed imSim and CatSim.

H. Awan: Participated in analysis and designed dithering strategy.

E. Gawiser: Participated in analysis and designed dithering strategy.

W. M. Wood-Vasey: Ran validation software and wrote part of the document.

Y. AlSayyad: LSST Science Pipelines

C. J. Burke: Contributions to PhoSim

S. W. Digel: Workflow and validation

M. Jarvis: Contributions to imSim and validation.

T. Johnson: Workflow, data generation and processing.

H. Kelly: DMstack librarian at NERSC.

S. Krughoff: Contributions to LSST Science Pipelines and DC1 workflow.

R. H. Lupton: LSST Science Pipelines, overall QA.

R. Mandelbaum: Provided feedback on DC1 design, analysis, and paper draft.

P. J. Marshall: Led the "Twinkles" DC1 pathfinder project.

J. R. Peterson: PhoSim Development for DC1.

P. A. Price: LSST Science Pipelines.

G. Sembroski: PhoSim Development for DC1.

B. Van Klaveren: DC1 workflow.

M. P. Wiesner: Completed a study on astrometry in PhoSim making it possible to improve astrometry in PhoSim for DC1.

B. Xin: Contributions to imSim and PhoSim.

Data Availability

The data underlying this article were produced by DESC at the National Energy Research Scientific Computing Center (NERSC) and will be shared on reasonable request to the corresponding author with permission of DESC. The total data volume for DC1 is 250 TB, where ~ 34 GB correspond to 10 year full-depth coadd catalogs.

REFERENCES

Agostinelli S., et al., 2003, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 506, 250

Albrecht A., et al., 2006, ArXiv Astrophysics e-prints,

Alonso D., Sanchez J., Slosar A., 2019, MNRAS, 484, 4127

Awan H., et al., 2016, ApJ, 829, 50

Benn C. R., 1983, The Observatory, 103, 150

Bertin E., 2011, in Evans I. N., Accomazzi A., Mink D. J., Rots A. H., eds, Astronomical Society of the Pacific Conference Series Vol. 442, Astronomical Data Analysis Software and Systems XX. p. 435

Bosch J., et al., 2018, Publications of the Astronomical Society of Japan, 70, S5

Bruderer C., Chang C., Refregier A., Amara A., Bergé J., Gamper L., 2016, ApJ, 817, 25

Brun R., Hagelberg R., Hansroul M., Lassalle J. C., 1978, Simulation program for particle physics experiments, GEANT: user guide and reference manual. CERN, Geneva, https://cds.cern.ch/record/118715

Bruzual G., Charlot S., 2003, MNRAS, 344, 1000

Budavári T., Szalay A. S., 2008, ApJ, 679, 301

Budavári T., Loredo T. J., 2015, Annual Review of Statistics and Its Application, 2, 113

Chisari N. E., et al., 2019, ApJS, 242, 2

Connolly A. J., et al., 2010, in Modeling, Systems Engineering, and Project Management for Astronomy IV. p. 77381O, doi:10.1117/12.857819

Connolly A. J., et al., 2014, in Modeling, Systems Engineering, and Project Management for Astronomy VI. p. 915014, doi:10.1117/12.2054953

Coupon J., Czakon N., Bosch J., Komiyama Y., Medezinski E., Miyazaki S., Oguri M., 2018, Publications of the Astronomical Society of Japan, 70, S7

DESI Collaboration et al., 2016, preprint, (arXiv:1611.00036)

De Lucia G., Springel V., White S. D. M., Croton D., Kauffmann G., 2006, MNRAS, 366, 499

Delgado F., Saha A., Chandrasekharan S., Cook K., Petry C., Ridgway S., 2014, in Modeling, Systems Engineering, and Project Management for Astronomy VI. p. 915015, doi:10.1117/12.2056898

Elsner F., Leistedt B., Peiris H. V., 2016, MNRAS, 456, 2095

Gonzalez O. A., Rejkuba M., Zoccali M., Valenti E., Minniti D., 2011, A&A, 534, A3

Hirata C., Seljak U., 2003, MNRAS, 343, 459

Hirata C. M., et al., 2004, MNRAS, 353, 529

Ivezić Ž.., et al., 2013, LSST Science Requirements Document, http://ls.st/LPM-17

Ivezić Ž.., et al., 2019, ApJ, 873, 111

Jarvis M., Bernstein G., Jain B., 2004, MNRAS, 352, 338

Jurić M., et al., 2008, ApJ, 673, 864

Jurić M., et al., 2015, preprint, (arXiv:1512.07914)

Krisciunas K., Schaefer B. E., 1991, PASP, 103, 1033

LSST Dark Energy Science Collaboration 2012, preprint, (arXiv:1211.0310)

LSST Science Collaboration 2009, preprint, (arXiv:0912.0201)

Lochner M., et al., 2018, arXiv e-prints, p. arXiv:1812.00515

Mandelbaum R., et al., 2005, MNRAS, 361, 1287

Mandelbaum R., et al., 2018, PASJ, 70, S25

Miralda-Escude J., 1991, ApJ, 380, 1

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Peebles P. J. E., 1980, The large-scale structure of the universe

Peterson J. R., et al., 2015, ApJS, 218, 14

Rowe B., 2010, MNRAS, 404, 350

Rowe B. T. P., et al., 2015, Astronomy and Computing, 10, 121

Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525

Sérsic J. L., 1963, Boletin de la Asociacion Argentina de Astronomia La Plata Argentina, 6, 41

Sevilla-Noarbe I., et al., 2018, MNRAS, 481, 5451

Sjöstrand T., Mrenna S., Skands P., 2006, Journal of High Energy Physics, 5, 026

Springel V., et al., 2005, Nature, 435, 629 EP

Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, ApJ, 761, 152

The LSST Dark Energy Science Collaboration et al., 2018, arXiv e-prints, p. arXiv:1809.01669

Wolstencroft R. D., Savage A., Clowes R. G., MacGillivray H. T., Leggett S. K., Kalafi M., 1986, MNRAS, 223, 279

Yoachim P., et al., 2016, in Observatory Operations: Strategies, Processes, and Systems VI. p. 99101A, doi:10.1117/12.2232947

Zuntz J., et al., 2018, MNRAS, 481, 1149

de Ruiter H. R., Willis A. G., Arp H. C., 1977, A&AS, 28, 211