DES-2020-0538
FERMILAB-PUB-20-669-E-SCD

Draft version December 24, 2020
Typeset using LaTeX **twocolumn** style in AASTeX63

Dark Energy Survey Year 3 Results: Measuring the Survey Transfer Function with Balrog

S. Everett,[1] B. Yanny,[2] N. Kuropatkin,[2] E. M. Huff,[3] Y. Zhang,[2] J. Myles,[4,5,6] A. Masegian,[7] J. Elvin-Poole,[8,9] S. Allam,[2] G. M. Bernstein,[10] I. Sevilla-Noarbe,[11] M. Splettstoesser,[12] E. Sheldon,[13] M. Jarvis,[10] A. Amon,[5] I. Harrison,[14,15] A. Choi,[8] W. G. Hartley,[16] A. Alarcon,[17] C. Sánchez,[10] D. Gruen,[4,5,6] K. Eckert,[10] J. Prat,[18] M. Tabbutt,[19] V. Busti,[20,21] M. R. Becker,[17] N. MacCrann,[22] H. T. Diehl,[2] D. L. Tucker,[2] E. Bertin,[23,24] T. Jeltema,[1] A. Drlica-Wagner,[18,2,7] R. A. Gruendl,[25,26] K. Bechtol,[19] A. Carnero Rosell,[27,28] T. M. C. Abbott,[29] M. Aguena,[20,21] J. Annis,[2] D. Bacon,[30] S. Bhargava,[12] D. Brooks,[31] D. L. Burke,[5,6] M. Carrasco Kind,[25,26] J. Carretero,[32] F. J. Castander,[33,34] C. Conselice,[15,35] M. Costanzi,[36,37] L. N. da Costa,[21,38] M. E. S. Pereira,[39] J. De Vicente,[11] J. DeRose,[40,1] S. Desai,[41] T. F. Eifler,[42,3] A. E. Evrard,[43,39] I. Ferrero,[44] P. Fosalba,[33,34] J. Frieman,[2,7] J. García-Bellido,[45] E. Gaztanaga,[33,34] D. W. Gerdes,[43,39] G. Gutierrez,[2] S. R. Hinton,[46] D. L. Hollowood,[1] K. Honscheid,[8,9] D. Huterer,[39] D. J. James,[47] S. Kent,[2,7] E. Krause,[42] K. Kuehn,[48,49] O. Lahav,[31] M. Lima,[20,21] H. Lin,[2] M. A. G. Maia,[21,38] J. L. Marshall,[50] P. Melchior,[51] F. Menanteau,[25,26] R. Miquel,[52,32] J. J. Mohr,[53,54] R. Morgan,[19] J. Muir,[5] R. L. C. Ogando,[21,38] A. Palmese,[2,7] F. Paz-Chinchón,[55,26] A. A. Plazas,[51] M. Rodriguez-Monroy,[11] A. K. Romer,[12] A. Roodman,[5,6] E. Sanchez,[11] V. Scarpine,[2] S. Serrano,[33,34] M. Smith,[56] M. Soares-Santos,[39] E. Suchyta,[57] M. E. C. Swanson,[26] G. Tarle,[39] C. To,[4,5,6] M. A. Troxel,[58] T. N. Varga,[54,59] J. Weller,[54,59] R.D. Wilkinson,[12]

(DES Collaboration)

[1] *Santa Cruz Institute for Particle Physics, Santa Cruz, CA 95064, USA*
[2] *Fermi National Accelerator Laboratory, P. O. Box 500, Batavia, IL 60510, USA*
[3] *Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA*
[4] *Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA*
[5] *Kavli Institute for Particle Astrophysics & Cosmology, P. O. Box 2450, Stanford University, Stanford, CA 94305, USA*
[6] *SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA*
[7] *Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA*
[8] *Center for Cosmology and Astro-Particle Physics, The Ohio State University, Columbus, OH 43210, USA*
[9] *Department of Physics, The Ohio State University, Columbus, OH 43210, USA*
[10] *Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA*
[11] *Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain*
[12] *Department of Physics and Astronomy, Pevensey Building, University of Sussex, Brighton, BN1 9QH, UK*
[13] *Brookhaven National Laboratory, Bldg 510, Upton, NY 11973, USA*
[14] *Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK*
[15] *Jodrell Bank Center for Astrophysics, School of Physics and Astronomy, University of Manchester, Oxford Road, Manchester, M13 9PL, UK*
[16] *Département de Physique Théorique and Center for Astroparticle Physics, Université de Genève, 24 quai Ernest Ansermet, CH-1211 Geneva, Switzerland*
[17] *Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL 60439, USA*
[18] *Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, USA*
[19] *Physics Department, 2320 Chamberlin Hall, University of Wisconsin-Madison, 1150 University Avenue Madison, WI 53706-1390*
[20] *Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, CP 66318, São Paulo, SP, 05314-970, Brazil*
[21] *Laboratório Interinstitucional de e-Astronomia - LIneA, Rua Gal. José Cristino 77, Rio de Janeiro, RJ - 20921-400, Brazil*
[22] *Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK*
[23] *CNRS, UMR 7095, Institut d'Astrophysique de Paris, F-75014, Paris, France*
[24] *Sorbonne Universités, UPMC Univ Paris 06, UMR 7095, Institut d'Astrophysique de Paris, F-75014, Paris, France*
[25] *Department of Astronomy, University of Illinois at Urbana-Champaign, 1002 W. Green Street, Urbana, IL 61801, USA*
[26] *National Center for Supercomputing Applications, 1205 West Clark St., Urbana, IL 61801, USA*
[27] *Instituto de Astrofisica de Canarias, E-38205 La Laguna, Tenerife, Spain*
[28] *Universidad de La Laguna, Dpto. Astrofísica, E-38206 La Laguna, Tenerife, Spain*
[29] *Cerro Tololo Inter-American Observatory, NSF's NOIRLab, Casilla 603, La Serena, Chile*

Corresponding author: Spencer Everett
sweverett@ucsc.edu

This document was prepared by DES Collaboration using the resources of the Fermi National Accelerator Laboratory (Fermilab), a U.S. Department of Energy, Office of Science, HEP User Facility. Fermilab is managed by Fermi Research Alliance, LLC (FRA), acting under Contract No. DE-AC02-07CH11359.

[30] Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth, PO1 3FX, UK
[31] Department of Physics & Astronomy, University College London, Gower Street, London, WC1E 6BT, UK
[32] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra, Spain
[33] Institut d'Estudis Espacials de Catalunya (IEEC), 08034 Barcelona, Spain
[34] Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, 08193 Barcelona, Spain
[35] University of Nottingham, School of Physics and Astronomy, Nottingham NG7 2RD, UK
[36] INAF-Osservatorio Astronomico di Trieste, via G. B. Tiepolo 11, I-34143 Trieste, Italy
[37] Institute for Fundamental Physics of the Universe, Via Beirut 2, 34014 Trieste, Italy
[38] Observatório Nacional, Rua Gal. José Cristino 77, Rio de Janeiro, RJ - 20921-400, Brazil
[39] Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA
[40] Department of Astronomy, University of California, Berkeley, 501 Campbell Hall, Berkeley, CA 94720, USA
[41] Department of Physics, IIT Hyderabad, Kandi, Telangana 502285, India
[42] Department of Astronomy/Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721-0065, USA
[43] Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA
[44] Institute of Theoretical Astrophysics, University of Oslo. P.O. Box 1029 Blindern, NO-0315 Oslo, Norway
[45] Instituto de Fisica Teorica UAM/CSIC, Universidad Autonoma de Madrid, 28049 Madrid, Spain
[46] School of Mathematics and Physics, University of Queensland, Brisbane, QLD 4072, Australia
[47] Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
[48] Australian Astronomical Optics, Macquarie University, North Ryde, NSW 2113, Australia
[49] Lowell Observatory, 1400 Mars Hill Rd, Flagstaff, AZ 86001, USA
[50] George P. and Cynthia Woods Mitchell Institute for Fundamental Physics and Astronomy, and Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843, USA
[51] Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton, NJ 08544, USA
[52] Institució Catalana de Recerca i Estudis Avançats, E-08010 Barcelona, Spain
[53] Faculty of Physics, Ludwig-Maximilians-Universität, Scheinerstr. 1, 81679 Munich, Germany
[54] Max Planck Institute for Extraterrestrial Physics, Giessenbachstrasse, 85748 Garching, Germany
[55] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[56] School of Physics and Astronomy, University of Southampton, Southampton, SO17 1BJ, UK
[57] Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831
[58] Department of Physics, Duke University Durham, NC 27708, USA
[59] Universitäts-Sternwarte, Fakultät für Physik, Ludwig-Maximilians Universität München, Scheinerstr. 1, 81679 München, Germany

(Received XXX, 2020; Revised XXX, 2020; Accepted XXX, 2020)

Submitted to ApJS

ABSTRACT

We describe an updated calibration and diagnostic framework, `Balrog`, used to directly sample the selection and photometric biases of Dark Energy Survey's (DES) Year 3 (Y3) dataset. We systematically inject onto the single-epoch images of a random 20% subset of the DES footprint an ensemble of nearly 30 million realistic galaxy models derived from DES Deep Field observations. These augmented images are analyzed in parallel with the original data to automatically inherit measurement systematics that are often too difficult to capture with traditional generative models. The resulting object catalog is a Monte Carlo sampling of the DES transfer function and is used as a powerful diagnostic and calibration tool for a variety of DES Y3 science, particularly for the calibration of the photometric redshifts of distant "source" galaxies and magnification biases of nearer "lens" galaxies. The recovered `Balrog` injections are shown to closely match the photometric property distributions of the Y3 GOLD catalog, particularly in color, and capture the number density fluctuations from observing conditions of the real data within 1% for a typical galaxy sample. We find that Y3 colors are extremely well calibrated, typically within ∼1-8 millimagnitudes, but for a small subset of objects we detect significant magnitude biases correlated with large overestimates of the injected object size due to proximity effects and blending. We discuss approaches to extend the current methodology to capture more aspects of the transfer function and reach full coverage of the survey footprint for future analyses.

Keywords: sky surveys — cosmology — dark energy — astronomical simulations

1. INTRODUCTION

Wide-field imaging surveys have revolutionized modern astronomy. Some of the primary science goals of these projects are to extract precise constraints on cosmological models and galaxy evolution using measurements made from hundreds of millions of galaxies for ongoing surveys such as the Dark Energy Survey[1] (DES; Abbott et al. 2016), the Kilo Degre Survey[2] (KiDS; de Jong et al. 2013), and the Hyper Suprime-Cam Survey[3] (HSC; Aihara et al. 2018), and even billions of sources for upcoming Stage IV experiments such as Euclid (Amiaux et al. 2012) and the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezić et al. 2019). For the largest surveys, the resulting constraints have become so precise that percent-level spatial variations in the survey's depth can cause biases that dominate the statistical errors (see for instance Huterer et al. 2006; Blake et al. 2010; Ross et al. 2012; Leistedt et al. 2016; Weaverdyck & Huterer 2020). Small biases – as small as one part in 10^4 in some cases – in the measurements of sizes, shapes, and fluxes of sources can have similarly important impacts on the science results (Massey et al. 2013).

The cumulative effect of the many selection effects and measurement biases of an astronomical survey is captured by its *transfer function*. This function maps how the photometric properties of astronomical sources are distorted by real physical processes such as interstellar extinction or by our imperfect measurements at every step from detector calibration to object catalog creation. As most cosmological measurements from survey data are based on the same processed images and source catalogs, this mapping is crucial for accurately estimating the true cosmic signals imprinted on the sky such as the spatial clustering of galaxies (see Blumenthal et al. 1984; Tegmark et al. 2006; Elvin-Poole et al. 2018 for a few examples) and weak lensing of galaxy light profiles by the intervening matter field (similarly, see Brainerd et al. 1996; Mandelbaum 2018; Troxel et al. 2018).

Unfortunately, many of these effects are in practice difficult to characterize or even identify. For example, the object catalogs derived from survey images are produced by a complex process: Calibration, detection, measurement, and validation involve a number of non-linear transformations, thresholds applied to noisy quantities, and post-facto cuts made on the basis of human judgment. Despite significant efforts to characterize some of these effects in the past (see Connolly et al. 2010 and Chang et al. 2015 for the LSST and DES pipelines respectively), this complexity makes each contribution to the transfer function extremely difficult to model – and even small errors in the estimated survey completeness can substantially bias measurements such as the amplitude of galaxy clustering or important calibration efforts like the photometric redshift inference of weak lensing samples (Aihara et al. 2011; Massey et al. 2013; Hildebrandt 2016; Fenech Conti et al. 2017).

Simulating the survey data from scratch can accurately capture some, but not all, of this complexity. Spatial variations in the effective survey completeness depend not just on the observing conditions but also on the ensemble properties of the stars and galaxies being studied. Systematic errors in the sky background estimation and biases in the measurements of galaxy and stellar properties can couple to fluctuations in the galaxy density field, leading to a completeness that depends on the signal being measured. Finally, there are a wide variety of non-astrophysical features that can affect the measurement quality and completeness such as artificial satellite trails, pixel saturation, or the diffraction spikes of bright stars. Not only are these effects difficult to model or simulate at high fidelity, but attempts to do so can introduce model-misspecification bias which can underestimate the true uncertainty (Lv & Liu 2010; Pujol et al. 2020).

Injection simulations can accurately capture many of these effects. Synthetic objects added to the real data automatically inherit the background and noise in the images as well as the biases arising from measurement in proximity to their real counterparts. Injecting realistic star and galaxy populations, convolving their light profiles with an accurate model for the point-spread function (PSF), and applying accurate models for effects not directly probed (such as Galactic reddening and variable atmospheric transparency) results in a population of simulated sources that inherits the same completeness variations and measurement biases as the real data. Mock catalogs made in this way can be used to discover, diagnose, and derive corrections for systematic errors and selection biases at high precision.

Generating full-scale mocks via injection is computationally demanding for a modern wide-field survey. The injection simulations described in Suchyta et al. (2016) for the early releases of DES data did not attempt to

[1] https://www.darkenergysurvey.org/
[2] http://kids.strw.leidenuniv.nl/
[3] https://www.naoj.org/Projects/HSC/

pass the injected galaxies through every part of the measurement process, opting to inject only onto the coadd images. The `Obiwan` tool, currently developed to model completeness variations for the Dark Energy Spectroscopic Instrument (DESI: Martini et al. 2018), focuses only on the emission-line galaxies that are the primary DESI targets (Kong et al. 2020). The SynPipe package (Huang et al. 2018) has been used to characterize measurement biases for the HSC pipeline and includes single-epoch processing, but only on a small fraction of the survey's available imaging. Despite injection pipelines having shown great promise, the difficulty in distinguishing intrinsic uncertainties in their sampling of the transfer function from actual measurement biases (as well as high computational cost) have until now kept them from being used to directly calibrate cosmological measurements.

This paper describes the generation of the `Balrog`[4] injection simulations for the first three years of DES data (referred to as Y3), covering a randomly selected 20% of the total Y3 footprint. Sources drawn from DECam (Flaugher et al. 2015) measurements of the DES Deep Fields (DF) (Hartley, Choi et al. 2020) are self-consistently added to the single-epoch DES images which are then coadded and processed through the full detection and measurement pipeline. This extensive simulation and reduction effort allows us to characterize, in detail, the selection and measurement biases of DES photometric and morphological measurements as well as the variation of those functions across the survey footprint. In addition, using an input catalog with measurements from the same filters as the data resolves many of the issues in capturing the same photometric distributions as real DES objects seen in Suchyta et al. (2016) – particularly for color. The resulting catalogs generally follow completeness and measurement bias variations in DES catalogs to high accuracy, with mean color biases of a few millimagnitudes and number density fluctuations varying with survey properties within 1% for a typical cosmology sample.

As the measurement pipelines for the DF and DES wide-field (WF) are complex and quite technical, so too are parts of this paper. However, we also motivate interesting science cases for the presented response catalogs for both calibration and direct measurement purposes including the photometric calibration of weak lensing samples, magnification effects on lens samples, and the impact of undetected sources on image noise. For readers more interested in using `Balrog` for potential science applications or as a general diagnostic tool, this is discussed in detail in Sections 4 and 5.

This paper is organized as follows: In Section 2 we introduce the significantly updated `Balrog` pipeline which now emulates more of the DES measurement stack and uses a completely new injection framework for source embedding into single-epoch images. Section 3 describes the injection samples and methodological choices for the Y3 `Balrog` simulations including a new scheme for handling ambiguous matches. In Section 4 we compare the recovered `Balrog` samples to the fiducial Y3 object catalog (Y3 GOLD; see Sevilla-Noarbe et al. (2020) for details), as well as present the photometric response of the main star and galaxy samples. We also examine the performance of a typical Y3 GOLD star-galaxy separation estimator and investigate a set of catastrophic photometric modeling failures that enter science samples with dramatically overestimated fluxes (sometimes by multiple orders of magnitude). We then discuss novel applications of an injection catalog in cosmological analyses including the photometric redshift calibration of Y3 "source" galaxies and the effect of magnification on "lens" galaxy samples – in addition to a few unexpected discoveries such as noise from undetected sources and issues with background subtraction. Finally, we close in Section 6 with a discussion of our results, methodological limitations, and future directions before concluding remarks in Section 7.

2. THE `BALROG` PIPELINE

`Balrog` was introduced in Suchyta et al. (2016) as a software package[5] that injects synthetic astronomical source profiles into existing DES coadd images to capture realistic selection effects and measurement biases for the Science Verification (SV) and Year 1 (Y1) analyses. However, as the precision of the subsequent DES cosmological analyses has increased, so too has the need for even more robust systematics control and more precise characterization of the survey transfer function. The main limitations of the original methodology were that (1) injections into the coadd rather than single-epoch images skip many important aspects of the measurement pipeline whose effects we want to capture, and (2) the injected objects were drawn from fitted templates to sources in the space-based Cosmological Evolution Survey (COSMOS: Scoville et al. 2007) rather than measurements consistent with DECam filters – introducing discrepancies in the recovered colors. While the latter is

[4] `Balrog` is *not* an acronym. The software was born out of the original authors delving "too greedily and too deep" (Tolkien 1954) into their data, hence the name.

[5] https://github.com/emhuff/Balrog

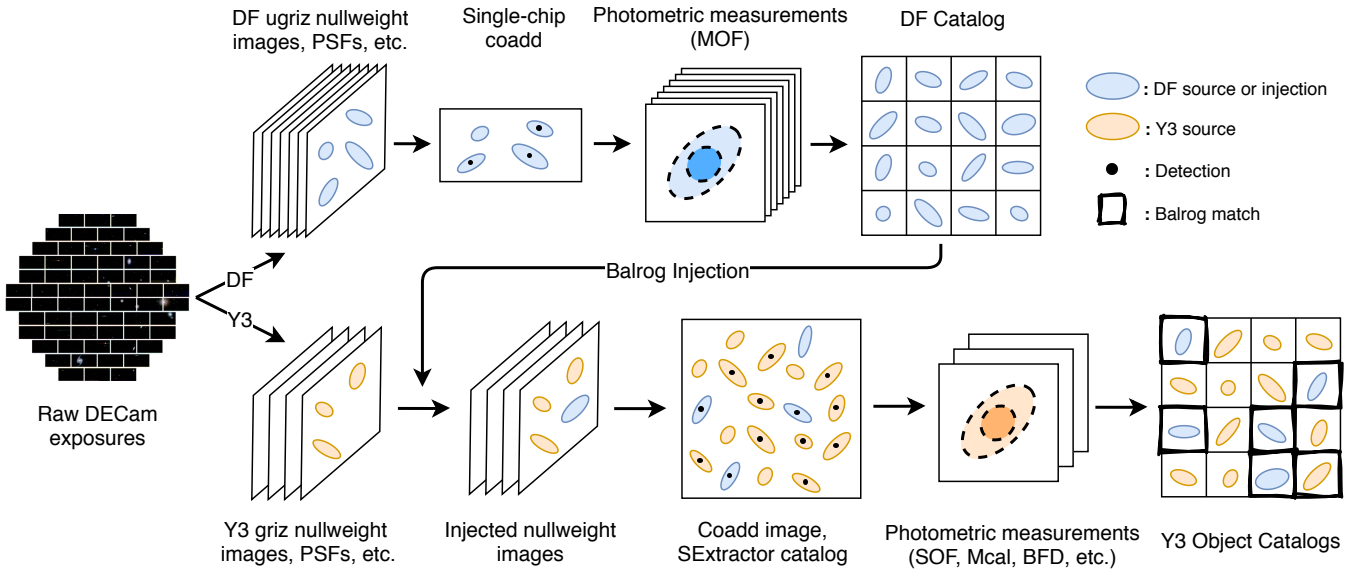


Figure 1. A high-level overview of how the Deep Fields (DF) and Y3 image processing pipelines interact to create the `Balrog` catalogs. The raw DECam exposures are used as the basis for both tracts, with the much deeper DF data being represented by the larger image stacks. The null-weight images, weight maps, PSF models, and zero point solutions are computed from the raw exposures after calibrations are applied and are the starting point of the sampled transfer function. The DF exposures are not dithered and thus single-CCD coadds are created in place of the much larger Y3 coadds. The fiducial DF catalog is created by fitting CModel profiles to detections with Multi-Object Fitting (MOF), which simultaneously models the light profiles of detected neighbors. These fitted model profiles (after a few limited cuts discussed in Section 3.4) are used as the `Balrog` injection catalog which are added to the Y3 null-weight images directly. Afterwards, the injected null-weight images are processed in a nearly identical way to the real images including coaddition, detection, and photometric measurements. Finally, we match the output object catalog to truth tables containing the injected positions. As all sources are remeasured, there is some ambiguity in the matching; this is discussed further in Section 3.5. See Hartley, Choi et al. (2020) and Morganson et al. (2018) for further DF and Y3 pipeline details respectively.

solved by using the new Y3 DF catalog (Hartley, Choi et al. 2020), the former required significant additional complexity in the simulation framework to consistently inject objects across all exposures and bands.

To address this, we have developed a completely new software framework that is described and validated in the remainder of this section. An overview of the Y3 `Balrog` process is shown in Figure 1, with simplified summaries of the DF and Y3+`Balrog` measurement pipelines. Briefly, we use the significantly deeper DECam measurements of sources in the DES DF as a realistic ensemble of low-noise objects to inject into the Y3 calibrated single-epoch images. We then rerun the DES measurement pipeline on the injected images to produce new object catalogs that contain the `Balrog` injections. Finally, we match the resulting catalogs to truth tables containing the injection positions to provide a mapping of DF truth to WF measured properties.

All astronomical image injection pipelines such as `Balrog` have two distinct elements: emulation of a survey's measurement pipeline and source injection into the processed images. As our methodology for the former is intrinsically specific to DES while the latter is a fairly generic problem, development on the new Y3 `Balrog` was split into the two corresponding pieces discussed in detail in Sections 2.1 and 2.2 below.

2.1. *DESDM Pipeline Emulation*

The DES survey data are processed through a set of pipelines by the DES Data Management team (DESDM) which perform basic astronomical image processing as well as applying state-of-the-art galaxy fitting, PSF estimation, and shear measurement codes. The standard processing steps applied to the DES Y3 data are described in detail in Morganson et al. (2018). Ideally, to ensure that identical codes and versions were used at each stage of processing, one would implement `Balrog` as part of the standard data reduction. However, this was not an option for DES Y3 as the updated `Balrog` methodology did not exist until after the Y3 data were completely processed (this is now true for a future Year 6 (Y6) `Balrog` analysis as well). Therefore it was necessary to replicate the DESDM processing pipeline stack as closely as possible. While this usually amounted to calling the relevant codes and scripts with identical configurations and software stack components, sometimes

minor changes were required due to differences in computing environments or practical considerations such as processing time. These differences will be noted whenever relevant.

A modular design for the measurement pipeline[6] was chosen both for ease of testing and for the ability to do non-standard production runs (see sections 5.2 and 5.3 for examples). The individual `Balrog` processing stages for a single DES coadd tile ($44' \times 44'$) are as follows:

1. **Database query & null-weighting** – Find all single-epoch immasked (the DES designation for flattened, sky subtracted, and masked) images in the *griz* bands that overlap the given DES Y3 tile. Download all exposures, PSFs, photometric and astrometric solutions from the DESDM Y3 processing archive. A masking process called "null-weighting" is applied to these immasked images which sets weights of pixels with certain flagged features (e.g. cosmic rays) to 0. These null-weight images are the starting point of the later injection step.

2. **Base coaddition & detection** – Remake the tile coadds from the single-epoch exposures with no objects injected using `SWarp` (Bertin et al. 2002) and the detection catalogs with `SExtractor` (Bertin & Arnouts 1996). Construct Multi Epoch Data Structure (MEDS; Jarvis et al. 2016) files with cutouts of the coadd and single epoch images used for additional photometric measurement codes. This allows us to cross-check our measured catalogs with Y3 GOLD to ensure that we recover the same detections and base photometry, as well as easily investigate proximity effects on the injections. Can be skipped to save processing time if desired.

3. **Injection** – Consistently add input objects in all relevant exposures and bands using the local PSF model in each exposure with corrections to the flux from the image zeropoints and local extinction – along with any other desired modifications such as an applied shear or magnification. This is discussed in detail in Section 2.2.

4. **Coaddition & detection** – Same as 2 but with the injected null-weight images. The resulting photometric catalogs contain existing real objects, injections, new spurious detections, and blends between the two.

5. **Single-Object Fitting (SOF)** – Fit a composite bulge + disk model that is the sum of an exponential and a de Vaucouleurs profile (CModel) to every source, while masking nearby sources.

6. **Multi-Object Fitting (MOF)** – Fit sources with CModel, but group nearby detections into friends-of-friends (FOF) groups that have all of their properties fit iteratively to account for proximity effects. Only available for some `Balrog` runs due to its computational expense.

7. **Metacalibration** – Fit a simple Gaussian profile to detections and then remeasure after applying four artificial shears (Sheldon & Huff 2017). This is useful for the creation of weak lening samples where correcting for shear-dependent systematics is more important than absolute flux calibration (Huff & Mandelbaum 2017).

8. **Gaussian APerture (GAp) fluxes** – Fit a robust, scale-length-independent alternative to model fitted photometry. Object flux is calculated within a Gaussian-weighted aperture with full-width at half-maximum (FWHM) of $4''$. Described further in Section 3.5.

9. **Bayesian Fourier Domain (BFD)** – Estimate the shear of sources without explicitly fitting a shape using the methodology described in Bernstein & Armstrong (2014). Available only for a few specialized runs.

10. **Match and compute GOLD value-adds** – Match input injections to output detections while accounting for ambiguous matches (see Section 3.5). Merge truth and measured table quantities. Compute Y3 GOLD value-added quantities including flags, object classifiers, masks, and magnitude corrections (though only the dereddening component is used for `Balrog` magnitude corrections; see §2.1.1).

The resulting photometric catalogs of measured `Balrog` sources can then be used to measure the DES wide-field response of various input quantities or used directly as randoms with realistic selection effects (see Suchyta et al. 2016 and Kong et al. 2020 for examples). In addition, an "injection catalog" is created which contains information for all injected sources, detected or not, for investigations into detection and completeness properties. The emulation steps 3 through 10 can be repeated for multiple injection realizations of a given tile to obtain sufficient sampling for the needed science case.

[6] https://github.com/kuropat/DES_Balrog_pipeline

However, as discussed in Section 3, for Y3 analyses we opted for a single realization with relatively high injection density due to the large computational cost of each realization.

2.1.1. *Differences from the DESDM Pipeline*

While `Balrog` strives to emulate the DESDM pipeline from null-weight images to science catalogs at high fidelity, there are some discrepancies due to practical limitations. The most significant are:

- **Reuse of existing single-epoch images, PSF models, photometric zeropoints, and WCS (astrometric solution):** Our input catalogs are assumed to be the true "top of the galaxy" measurements. Due to this we do not recalculate the photometric and astrometric zeropoints for any exposures which have additional objects added to them; the Y3 DESDM solution is carried forward unchanged. This means that we cannot probe the individual systematic error contributions of steps in the DESDM pipeline before this stage, such as the PSF modeling or image detrending.

- **Incomplete `SExtractor` parameter list:** We chose to measure only a subset of the Y3 `SExtractor` parameters that were anticipated to be important for downstream analyses in order to save processing time. In particular, we did not compute any model fitted magnitudes including `MAG_PSF` which is needed for the WAVG quantities described in Morganson et al. (2018). Ultimately, the overall time saved was small and we plan to save all `SExtractor` quantities for future runs.

- **MOF is skipped for the cosmology sample:** While MOF photometry is available for the Y3 GOLD catalog, most Y3 cosmological analyses use the variant SOF which skips the multi-object deblending step in favor of masking neighbors. This approach is significantly faster, fails less often, and has negligible impact in photometric performance (E. Sheldon, private communication). As MOF is not needed for Y3 cosmology calibration and contributed roughly a quarter of all `Balrog` runtime (see Table 2), we elected to skip this step for the main samples.

- **Zeropoint and chromatic corrections are not applied:** The Y3 photometric calibration introduces new chromatic corrections that achieve subpercent uniformity in magnitude by accounting for differences in response arising from varying observing conditions and differences in object SEDs (see Sevilla-Noarbe et al. 2020). However, the mean Y3 GOLD chromatic corrections are significantly below 1 millimagnitude (mmag) for all but g-band (0.45 mmag). As this is a subdominant effect that requires significant computation to correct in each injection realization, we do not account for these corrections before injecting into images. In addition, the SED-independent "gray" corrections that account for variations in sky transparency and instrumentation issues like shutter timing errors were not accounted for in the injection zeropoints. This was not intentional and will be included in all future `Balrog` runs. However, these corrections are also quite small, with the mean absolute Y3 GOLD gray zeropoint correction below 1 mmag for all bands except for z-band (1.2 mmag). As we do not modulate the truth fluxes with these corrections during injection, it is not necessary to apply these corrections *after* measurement either.

- **Partial GOLD Catalog Creation**: Due to the staged approach in the creation of Y3 GOLD with value-added products being incorporated as they were being developed, the exact same procedure for compiling the `Balrog` catalog could not be followed strictly as it would have produced an unnecessary and severe overhead in the production time. Scripts that approximately replicate this process were provided by DESDM, though they only reproduce the columns that were deemed to be most relevant to Y3 key science goals. Slight modifications had to be made to quantities such as `FLAGS_GOLD` and the object classifier `EXTENDED_CLASS_SOF` where the required MOF columns were not available; these differences are mentioned when relevant throughout the paper.

While not technically a difference in the *pipeline* emulation itself, we note here that PSF models used for injections (PSFEx; Bertin 2011) were found to be slightly too large in Zuntz et al. (2018) for bright stars in Y1 due to the brighter-fatter effect (see Antilogus et al. 2014). However, we still used PSFEx for our injection PSFs as the new Y3 PIFF PSF model described in Jarvis et al. (2020) was not yet implemented into the `GalSim` configuration structure that was required for our injection design, which is discussed below.

2.2. *Injection Framework*

As mentioned in the beginning of this section, incorporating single-epoch injection into `Balrog` required a new software design to handle the significant increase in simulation complexity beyond what was done in Suchyta

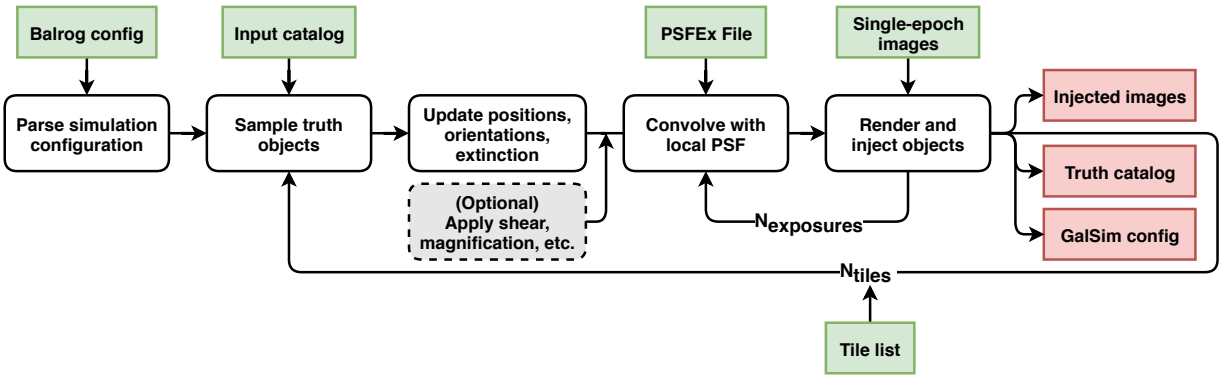


Figure 2. High-level overview of the injection processing for a single realization. Green boxes are inputs to the injection framework while red boxes are outputs. The length of each loop is determined by the number exposures and tiles considered in the full simulation. While the main runs used for Y3 cosmology calibration modify only the position, orientation, and flux normalization of the truth inputs, there are many optional transformations that can be applied such as a constant shear or magnification. The main output of our injection package is a multi-document configuration file with detailed injection specifications that is then executed by `GalSim`, with each step being executed in the physically correct order. Additional realizations replicate all steps, other than the initial configuration parsing, and produce unique outputs.

et al. (2016) for the SV and Y1 analyses. Development on the injection framework was partitioned into its own software package[7] as the injection step is fairly generic and of potential interest to other analyses outside of DES Y3 projects – as well as upcoming Stage IV dark energy experiments such as LSST. Briefly, our injection framework maps high-level simulation choices into individual object and image-level details consistent between all single-epoch images for the simulation toolkit `GalSim` (Rowe et al. 2015) to process. With this design, `Balrog` automatically inherits much of the modularity, diverse run options, and extensive validation of `GalSim`. A schematic overview of the injection process is shown in Figure 2. The remainder of this section will describe the implementation details of each step, along with some of the various user options for this new software package.

2.2.1. *Injection Configuration*

The `Balrog` configuration serves as the foundation for the final, much larger `GalSim` configuration file produced for each tile by the injection pipeline which follows the `GalSim` configuration conventions that are extensively documented. Global simulation parameters that apply to all injections are defined here such as the input object type(s) (see §2.2.2), position sampling method, injection density, and number of injection realizations. During injection processing, the requisite simulation details needed to inject the sampled input objects consistently across the relevant survey images are appended to this file to create a multi-document `GalSim` configuration file with each document corresponding to a single CCD exposure.

Configuration settings specific to a typical `Balrog` run have been wrapped into custom `GalSim` `image` and `stamp` types, both called `Balrog`:

- `image: Balrog` - This image type is required for a full `Balrog` run. It parses all novel configuration entries and defines how to add `GalSim` objects to an existing image with consistent noise properties. It also allows the `Balrog` framework to be run on blank images for testing.
- `stamp: Balrog` - An optional stamp type that allows `GalSim` to skip objects whose fast Fourier transform (FFT) grid sizes are extremely large and can occasionally cause memory errors when using photometric model fits to DES DF objects.

We also provide a much simpler `image` class called `AddOn` which adds any simulated images onto an initial image without the full `Balrog` machinery. Appendix A gives an example configuration that was used for the two main cosmology runs. Some configuration details can also be set on the command-line call to `balrog_injection.py` for ease of use as long as they do not conflict with any settings in the configuration file; see the code repository for more details on running the simulations.

2.2.2. *Input Sample and Object Sampling*

In principle any native `GalSim` input and object type can be used for injection. However, the object sampling, truth property updating, and truth catalog generation steps require knowledge about underlying struc-

[7] https://github.com/sweverett/Balrog-GalSim

ture of the input data (e.g. parametric models vs. image cutouts). We handle this ambiguity through the use of `BalInput` and `BalObject` parent classes that define the necessary implementation details to connect `GalSim` to `Balrog`. These classes can be used to register any needed injection types to `Balrog` including custom `GalSim` classes. Subclasses provided for injection types used in DES Y3 runs are described below:

- `ngmixGalaxy`: Most of the photometric measurements in Y3 DES science catalogs are based on Gaussian mixture model fits by `ngmix`[8] described in Sheldon (2014). Each parameterization is converted to a sum of `GalSim Gaussian` objects that represent the Gaussians components used in the original fit. `Balrog` can currently inject the following `ngmix` model types: a single Gaussian (`gauss`), a CModel combined bulge + disk (`cm`) that is the sum of an exponential and de Vaucouleurs profile, and a slightly simpler CModel with fixed size ratio between the two components (`bdf`, for Bulge-Disk with Fixed scale ratio). As `ngmix` allows for objects with negative size before convolution with a PSF, these negative values are clipped to a small non-zero value (`T`=10^{-6}, corresponding to a size scale of $\sim 10^{-3}$ arcsec) to avoid rendering failures.

- `DESStar`: A synthetic star sample with realistic density and property distributions across the DES footprint was created to a depth of 27 magnitude in g. These objects are treated as delta functions convolved with the local PSF. These magnitudes are referenced as δ-mag in later figures. Further details about this star catalog are described in Section 3.2.

- `MEDSGalaxy`: Single-epoch image cutouts of detected DES objects are stored in MEDS files for each band. These image cutouts can be used directly for injection after deconvolving with the original PSF solution and re-convolving with the local injection PSF. While used in testing, this injection type was not used in the main `Balrog` runs for science calibration due to issues arising from injecting stamps with larger associated PSFs than the injection image. In addition, there was not time to complete the requisite validation of stamp and mask fidelity for all injections before the runs had to start.

[8] https://github.com/esheldon/ngmix

`Balrog` can inject multiple object types in the same run by setting the `gal` field in the configuration as a `List` type; this is identical to `GalSim` configuration behaviour. The relative fraction of each injection type is then set in the `pos_sampling` field described below.

2.2.3. *Updating Truth Properties, Setting PSF, and Optional Transformations*

Most `Balrog` runs sample objects from an existing catalog. Some of the object properties are modified to fit the needs of the simulation such as the positions, orientations, and fluxes. Updates to positions and orientations are automatically applied to the output truth catalogs while flux corrections due to local extinction and zeropoint offsets are not, though we save the applied extinction factor. Different behaviour for these quantities as well as any additional changes can be defined when creating the relevant `BalObject` subclass.

Position sampling is determined by the configuration parameter `pos_sampling` and can be set to `Uniform` for spherical random sampling or one of the following grid choices that are regularly-spaced in image space: `RectGrid` for a rectangular lattice, `HexGrid` for a hexagonal lattice, and `MixedGrid` for one of the previous grid choices that mixes multiple injection object types on the same grid with a set relative abundance `inj_frac`. The user has control over the grid spacing as well as whether to apply random translations and/or rotations of the grid for each tile in addition to random rotations of the object profiles themselves with `rotate_objs`.

Object fluxes are scaled to match the photometric zeropoint of each single-epoch injection image. An additional extinction factor can be applied with the configuration option `extinct_objs`. If set, extinction factors in $griz$ for each tile are loaded and applied to object fluxes. Incorporating more sophisticated per-object extinction implementations is planned for a future code release but was found to be unnecessary for Y3 analyses. Any of the native `GalSim` noise models can be added to the injection stamps with the Poisson component ignoring the existing image pixel values as long as the `Balrog` (or `AddOn`) image type is used. Finally, the PSF used for each object is determined by the single-epoch PSFEx solution at the injection position. Simpler PSF models are also allowed for testing purposes but not recommended for science runs.

Additional transformations such as a constant shear or magnification factor can be applied depending on the desired science case (see 5.2 for an example using magnification in Y3). Transformations that are uniform per-tile can be added in the injection configuration with the same syntax as a typical `GalSim` configuration, while

per-object effects need to be implemented into the relevant `BalObject` subclass.

2.2.4. *Object Rendering and Injection*

All of the previous simulation choices are ultimately encoded in a very detailed configuration file that is structured to be read by `GalSim`. This design was chosen over explicit use of the software's Python API as the configs facilitate easily reproducible simulations and allow for runs that are identical except for minor modifications such as an added constant magnification factor. Each transformation from truth property to pixel value is automatically handled by `GalSim` processing in the physically correct order. After an object stamp is rendered, its pixels are summed with the initial image while ignoring any part of the profile that may go off image. Rarely a profile will require an extremely large FFT grid during PSF convolution and exceed available memory. To avoid this, we set a maximum grid length of 16,384 pix^{-1} (or $\sim$63,000 arcsec^{-1} for DES) per side and skip objects that exceed this limit. While the injection framework was designed with flexibility in mind for uses outside of the Y3 cosmology science goals (and even DES itself), there are currently some assumptions made about the structure of the input data to emulate DES Y3 that we plan on generalizing in upcoming releases.

2.3. *Pipeline Validation*

As `Balrog` is a non-generative, or discriminative, model of the transfer function, it is difficult to disentangle any intrinsic errors in the input sample or survey pipeline emulation from actual systematic effects we are trying to characterize – particularly since `Balrog` was run independently of DESDM processing for Y3. Therefore a series of increasingly complex test runs were completed in order to validate both the injection and emulation steps and characterize the pipeline fidelity at a detailed level. We initially ran `Balrog` with the injection step turned off to confirm that we recovered identical detection and photometry catalogs as Y3 GOLD when carefully accounting for the same random seeds in the fitters that were used in nominal Y3 processing. Once this was achieved, we verified that the injected profiles of objects drawn onto blank images matched single-object renderings made independently of the pipeline.

We then ran a series of tests where we ignored the existing survey image data during injection except for the estimated residual local sky background that is automatically subtracted from the exposures later in the pipeline. Objects were placed on a sparse grid to limit proximity effects from other injections with two types of noise depending on the run – either only Poisson noise for the injections or Poisson in addition to low levels of zero-mean Gaussian background sky noise. These blank image runs became progressively more complex as we added the features used in the main science runs described in Section 3 and acted as a form of regression testing. These tests are performed by setting the field `inj_objs_only` to `True` in the configuration file along with the `noise` field set to either `BKG` or `BKG+SKY`, though this mode of testing is only available for the provided `Balrog` image class, not `AddOn`.

These tests are relevant for more than pipeline validation; effects from methodological choices can also be identified and quantified while working in a simplified environment. As an example, the runs with only Poisson noise indicated that there were two subgroups of objects with statistically significant differences in magnitude response – one was well calibrated, the other with a mean offset of $\sim$7.5 mmag too faint in each of *griz*. This was ultimately discovered to be a result of different priors used for the parameter that measures the relative flux ratio between the de Vaucouleurs and exponential component, `fracdev`, for the `ngmix` profile type used to fit DF objects (`bdf`) and the one used to fit wide-field measurements (`cm`). A series of plots that show the difference in input vs. measured `fracdev` and examples of its downstream effect on the recovered magnitude and color responses for this test is shown in Figure 3.

The impact of the different `fracdev` fits on the magnitude response can be seen clearly in Figure 3d, where the difference in measured vs. true i-band magnitude as a function of injected magnitude is colored by the response in `fracdev` for a single tile. As the difference in profile definition between `cm` and `bdf` is largely due to fitting stability and has little to do with the true distribution of galaxy properties, this effectively puts a lower bound on the accuracy of the mean magnitude response that we are able to measure with `Balrog` when using the DF sample as inputs at around 3 mmag. Importantly, however, the effect is nearly identical in each of the *griz* bands and has negligible impact in the recovery of colors, as seen in Figure 3c. This example highlights some of the difficulties in choosing a "truth" definition for injections based on model fits and the importance of carefully testing the impacts of model assumptions.

The final version of the blank image test was performed with identical input and configuration to that used to produce the fiducial Y3 catalogs across 200 tiles which contain over 2.3 million injections and 1.6 million detections. Zero-mean Gaussian background noise was applied to the blank images with variance set to the CCD's `SKYVAR` value. The resulting object responses allow us to characterize the baseline performance of the photometric pipeline in ideal (though overly simplistic)

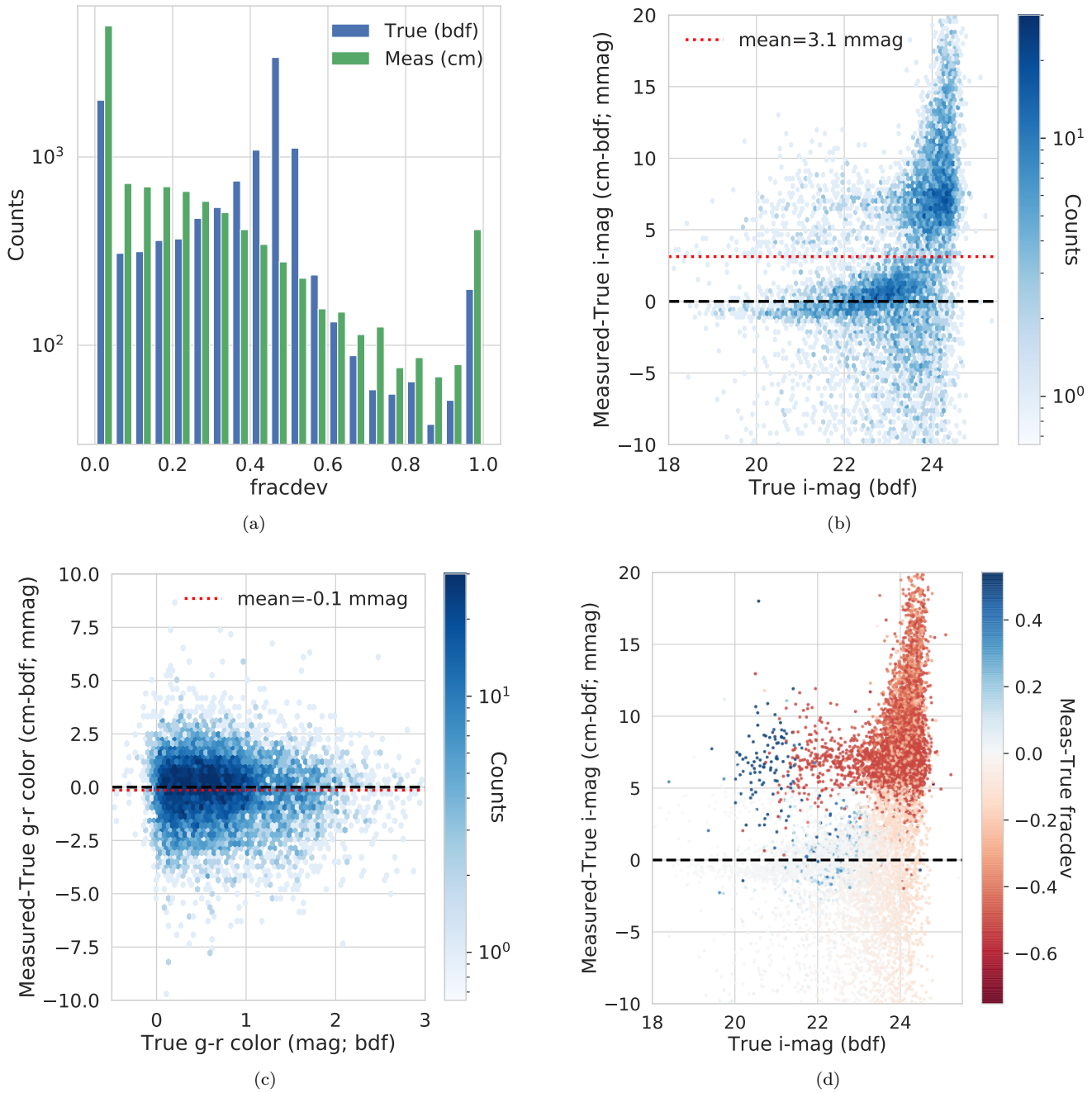


Figure 3. A series of plots highlighting aspects of the noiseless blank image test described in Section 2.3. (a) The first panel shows the difference in input `bdf_fracdev` vs measured `cm_fracdev` for detected objects. The additional peak at 0.5 for `bdf_fracdev` is a result of the slightly different model definition; for `bdf`, the relative size ratio between the bulge and disk components is forced to be 1. This constraint does not exist for `cm` and thus it has a different prior on the parameter. (b) This panel shows the i-band magnitude response of these objects, where there are clearly two different populations. The first is well-calibrated with the majority of detections well within +/-2.5 mmag of truth. The second population is biased towards fainter measurements by ∼7.5 mmag on average. (c) The $g-r$ color response for these objects. The bias in recovered magnitude is nearly identical in $griz$ and so does not translate to the recovered colors. The mean color response for $g-r$, $r-i$, and $i-z$ is 0.1, 0.3, and 0.2 mmag respectively. (d) The final panel shows that the biased magnitude population is a result of injections with input `bdf_fracdev`∼0.5 scattering to 0 or 1 to match the expected `cm_fracdev` prior. As we do not believe this differential response to be of physical origin, it contributes to a lower bound on the precision in which `Balrog` can calibrate Δmag – though importantly this does not contribute a bias to recovered colors.

conditions which in turn may provide lower limits on the intrinsic uncertainty in our sampling of the DES transfer function. The mean and median difference in recovered versus injected magnitude for *griz* is plotted in Figure 4. The vertical bars correspond to the mean of the standard deviations of *griz* magnitude responses in each truth magnitude bin, centered at the mean magnitude response.

The medians are extremely well calibrated, with only $g < 18.5$ and $22.5 < z < 23$ off by more than 5 mmag, or 0.45%, through 23rd magnitude where selection effects near the detection threshold become significant. The mean responses are consistently biased towards larger recovered flux on the bright end by ∼15 mmag due to the asymmetric tendency of SOF to measure the sizes of bright, extended objects to be too large in the presence of neighbors; this is a real effect seen in the main data runs and is discussed in greater detail in §4.3.1. Such biases are not seen in isolated SOF measurements of similar objects (E. Sheldon, private communication) and appear in this test as it was inefficient to use a grid size large enough to keep all other grid injections outside the MEDS stamps of the largest injections. This effect also keeps the magnitude error from decreasing as the intrinsic brightness increases as one would naively expect. While the magnitude bias induced by the difference in the `cm` vs. `bdf` profile definition is present in this measurement, it is negligible compared to proximity biases for extended sources and selection effects present in the noisier images.

Importantly, there is no significant band-dependence in the median magnitude responses where the recovered sample is complete, with a typical spread in median *griz* biases of ∼3 mmag for truth magnitudes ranging from 18.5 to 22 with no characteristic shape or distribution systematics. While there is a detectable band-dependence in the mean magnitude responses, it is nearly eliminated when binned in signal-to-noise (S/N) instead of magnitude to account for differences in sky noise.

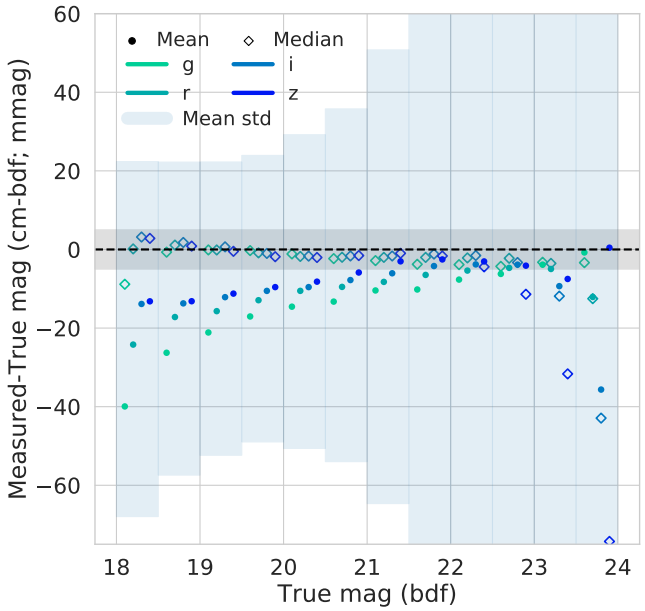


Figure 4. The mean (solid circle) and median (hollow diamond) difference in measured vs. injected magnitude (<Δmag>) as a function of input magnitude for the final blank image runs with zero-mean Gaussian background noise. The vertical bars correspond to the mean of the standard deviations of *griz* magnitude responses in each truth magnitude bin, centered at the mean magnitude response. The vertical bars represent the average of the standard deviations of *griz* magnitude responses in each bin of size 0.5 magnitudes, centered at the mean magnitude response. The overall calibration is excellent, with the median response less than 5 mmag in all bins except for $g < 18.5$ and $22.5 < z < 23$. We expect significant biases past magnitude 23 due to selection effects near the detection threshold. The mean responses show some bias however – particularly on the bright end. As discussed in the text, this is due to an asymmetric tendency for SOF to measure the fluxes of bright, extended galaxies to be too large when neighbors are contained in the object's MEDS stamps. The errors in <Δmag> do not substantially decrease past input magnitudes of 20 for the same reason. This is discussed more in §4.3.1.

3. BALROG IN DES YEAR 3

We describe here the injection samples, pipeline settings, and matching choices used to create the Y3 `Balrog` data products for the photometric performance characterization described in Section 4 and downstream science calibrations described in Section 5. For Y3, we ran `Balrog` several times with different configurations for various validation and science cases. These runs are tabulated in Table 1 which lists the following quantities: the run name, the number of simulated tiles, the total number of injected objects, the fraction of detected objects, the mean number of times a given object is injected across all tiles, the spacing between injections, and the magnitude limit used for sampling. As detection in DES is based on a composite *riz* detection coadd, we emulate the detection magnitude by averaging the dereddened *riz* fluxes of the injections.

The main runs used for cosmological analyses are called `Run2`[9] and `Run2a`. The former samples the transfer function across 1,544 randomly chosen tiles (of the

[9] The designation `Run1` was used for an earlier set of simulations that used an inferior DF catalog.

Run Name	Tiles	N Det	Det-Frac	<Inj's>	Mag Lim	Spacing	Notes
`grid-test`	200	1.6 M	0.702	3	24.5	20″	Inject into blank images with noise
`noiseless-grid-test`	196	2.6 M	0.997	4	24.5	20″	Same as above but without noise
`Run2`	1544	7.4 M	0.369	16	25.4	20″	Subset of `y3-merged` & `y3-stars`
`Run2a`	497	3.9 M	0.600	9	24.5	20″	Subset of `y3-merged` & `y3-stars`
`Run2-mag`	155	0.8 M	0.463	2	25.4	20″	2% magnification on `Run2` objects
`Run2a-mag`	497	3.9 M	0.607	9	24.5	20″	2% magnification on `Run2a` objects
`clusters`	901	39.9 M	0.930	163	23.0	10″	Tiles containing rich galaxy clusters
`blank-sky`	88	–	—	—	—	20″	Injected zero-flux objects

Table 1. Table of Y3 Balrog runs and associated parameters: the number of tiles sampled, the number of total detections (N Det), the detection fraction (Det-Frac), the mean number of injections per unique DF object (<Inj's>), the composite *riz* detection magnitude limit, and injection lattice spacing.

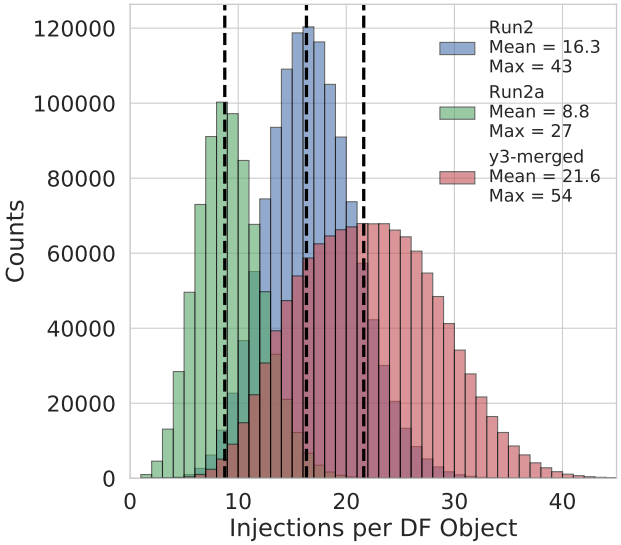


Figure 5. The number of injections per unique DF object for `Run2` in blue, `Run2a` in green, and their combination `y3-merged` in red. The mean number of injections per run is shown with dashed vertical lines and is stated along with the maximum number of injection realizations. `Run2` is composed of 1,544 tiles vs. only 497 for `Run2a`, but has a larger input catalog to sample due to the more conservative composite *riz* detection magnitude of 25.4 vs. 24.5 for `Run2a`. The resulting combination is no longer a Poisson distribution but this can be accounted for in downstream analyses by weighting by the column `injection_counts`. The typical `Balrog` object in `y3-merged` has just over 20 unique injection realizations across the sampled footprint.

10,338 Y3 tiles) to a detection magnitude limit of 25.4. This limit was chosen to capture DF objects that had at least a 1% chance of being detected as measured from a 200 tile test run. `Run2a` was a supplemental run at a shallower limiting magnitude of 24.5 across 497 tiles to increase the fraction of recovered injections for analyses that needed a larger total sample. These runs are combined for the fiducial `Balrog` catalogs `y3-merged` and `y3-stars` which are described in upcoming sections. The distributions of the number of injection realizations per input object for these runs are shown in Figure 5, and the spatial distribution of these tiles are shown compared to the full DES footprint in Figure 6. `Run2-mag` and `Run2a-mag` are identical to the above runs except for a constant added magnification of $\delta\kappa = 0.02$; these are described in more detail in Section 5.2. The `grid-test` and `noiseless-grid-test` runs were used for the validation tests shown in 2.3. The `blank-sky` and `clusters` runs were conducted separately from the main cosmology runs in order to facilitate two of the science cases discussed in Sections 5.3 and 5.5 respectively.

The processing was done on a dedicated compute cluster at Fermilab, "DEgrid", consisting of 3000 cores with 6-8GB RAM per core available. The typical core and memory provisioning along with wallclock running times for each stage of the pipeline is given in Table 2. MOF is not used for the fiducial Y3 cosmology analyses and so is excluded for `Run2` and `Run2a` – along with their corresponding magnification runs. We include the estimated computational cost to show the difficulty in scaling this methodology to full footprint coverage and WF density; we discuss this more in Section 6. All output measurement catalogs were archived including the MEDS cutout images of detected objects; the injected single-epoch images and resulting coadds were only saved for validation runs.

A few additional post-processing steps were required to match changes made to the Y3 object catalogs after the fiducial GOLD catalog creation. These consisted of a correction to the metacalibration signal-to-noise (S/N) column, redefining the `size_ratio` quantity from `mcal_T_r / psfrec_T` to `mcal_T_r / mcal_Tpsf`, and adding a shear weight to each of the metacalibration measurements described in 5.1.

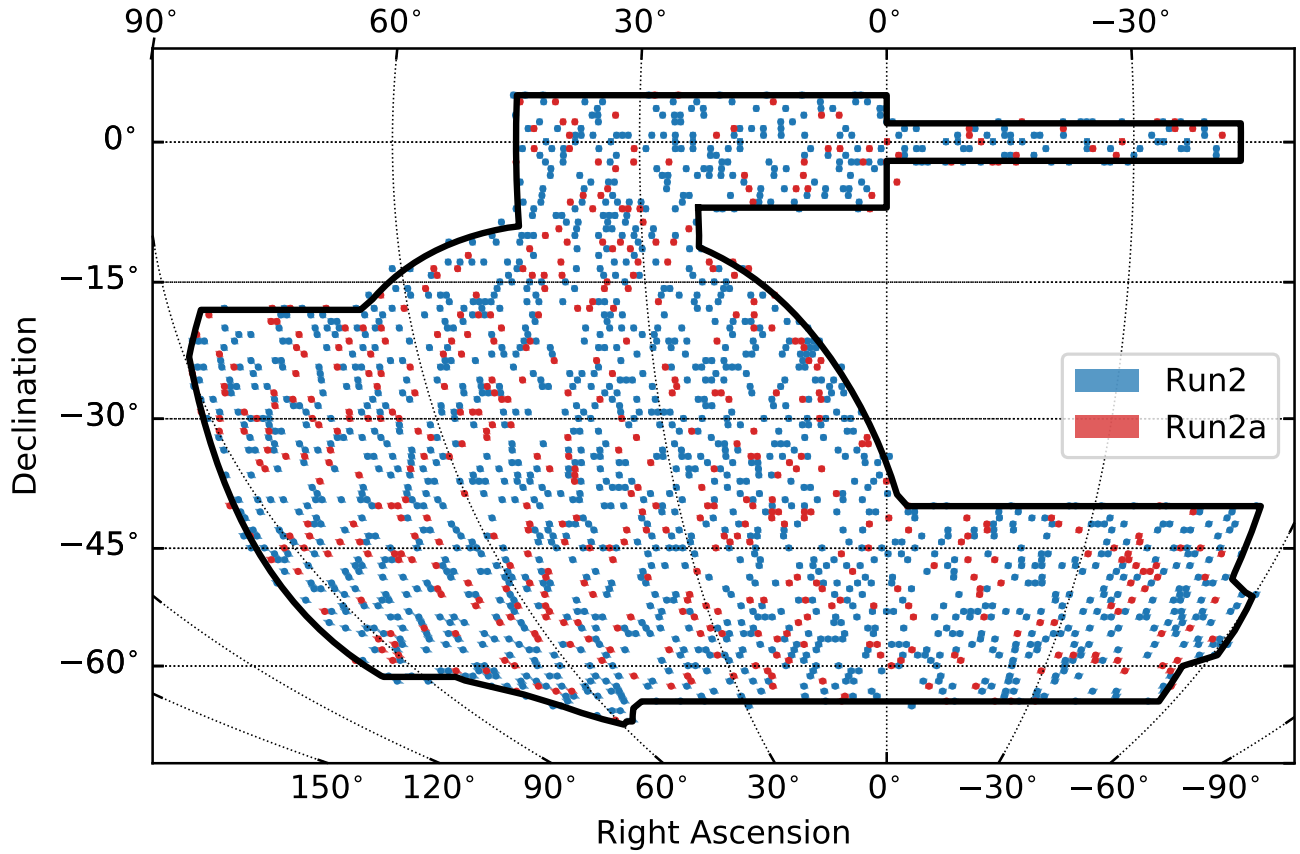


Figure 6. The spatial distribution of randomly sampled DES tiles used for `Balrog` injections. 1,544 `Run2` and 497 `Run2a` tiles are shown in blue and red respectively. The outline of the DES Y3 footprint is shown in black. Some tiles are slightly outside of the official footprint due to partial image coverage from DECam observations on the footprint edge.

Stage	Cores	RAM	Clocktime
Database Query	1	64 GB	2.0 hr
Base Coaddition/Detection/MEDS	4	64 GB	3.0 hr
Injection	16	64 GB	3.0 hr
Coaddition/Detection/MEDS	4	64 GB	5.0 hr
MOF*	32	256 GB	6.5 hr
SOF	16	64 GB	1.5 hr
Metacalibration	8	320 GB	2.5 hr
Match/Merge/Flag	2	512 GB	1 hr
Total/tile	16-32	64-512 GB	$18 - 24.5$ hr/tile

Table 2. Approximate `Balrog` stage run times and memory allocations per tile. *As MOF is not used in the fiducial Y3 cosmology analysis, this step was only run for `Run1` due to the long clocktime. The two total reported clocktimes are with MOF excluded or included in the pipeline emulation respectively.

3.1. *Input Deep Field Catalog for y3-merged*

The majority of Y3 `Balrog` analyses use injections drawn from DECam measurements of objects in the DF described in Hartley, Choi et al. (2020). In brief, this catalog of nearly 3 million sources is assembled from hundreds of repeated exposures of three DES supernovae (SN) fields and the COSMOS field. The corresponding deep single-CCD coadds have S/N of $\sim\sqrt{10}$ times their WF counterparts and thus provide a good sample of low-noise sources to draw from for explorations of systematics in the WF measurements. There are multiple versions of the DF catalogs that provide trade-offs in the average seeing quality vs. the maximum depth. In Y3 `Balrog`, we use `COADD_OBJECT_TRUTH` as it strikes a balance between using observations with 10 times the mean WF exposure time while ensuring that the composite DF FWHM be no worse than the median single epoch FWHM in the WF for each of the injection bands.

We emphasize that we are not injecting the actual *images* of DF galaxies but instead take the MOF `ngmix` parameterized model fit to each detection and generate an idealized galaxy profile based on those model parameters (with added Poissonian noise). The injection framework described in Section 2 is capable of injecting the MEDS stamps directly which in principle would account for additional diversity in galaxy morphologies and eliminate any model bias compared to the true distribution of galaxy properties. However, this requires extensive validation of the DF stamps before injection and introduces additional complications due to image masks and added noise for injections into CCDs with better seeing than the DF composite image. We plan to revisit these issues for `Balrog` in the Y6 methodology.

The DF catalog is comprised of model fits that are very similar to the WF CModel with two major differences: the two components (bulge + disk) are fit simultaneously rather than separately, and the ratio of the size of each component, `TdByTe`, is fixed to be 1. While this was chosen for increased fitting stability for the fainter DF sources, fixing the relative bulge-disk size ratio reduces the total number of free parameters in the model by one and significantly changes the distribution in the relative flux fraction `fracdev` (recall Section 2.3 for how this impacts the corresponding recovered CModel photometry in idealized conditions). Ultimately, any photometry can be used for the injection truth as long as it is an unbiased estimate of the real distribution of object properties. The `bdf` profile will be used for all Y6 DES source fitting and for Y6 `Balrog` – avoiding the small systematic difference in magnitudes between `cm` and `bdf`.

3.1.1. *DF Object Extinction*

The DF catalog has detailed photometric corrections to the fluxes including for extinction as described in Hartley, Choi et al. (2020). However, these corrections were not yet ready when `Balrog` began the cosmology runs. Thus in order to accurately account for variations in DF extinction, as well as extinction variations among tiles in the Y3 survey footprint, we enacted the following procedure to deredden the DF input objects and then re-extinct them by an appropriate amount in the injection WF tile: For the DF objects, we sample the Schlegel et al. (1998) extinction maps at five points (center and corners) in each input DF CCD (of size $9'\times18'$) and record the average of the 5 E(B-V) values. We also record the five-point average of E(B-V) for the larger (size $44'\times44'$) WF tiles. During injection, we deredden each object by the DF recorded value for its CCD of origin and apply the mean extinction value for the WF injection tile. This chip and tile-level correction is simple to implement and distorts the overall magnitude and color distribution of the DF galaxy sample from the cosmic average only slightly. However, we plan on implementing per-object extinction corrections in the Y6 methodology. The used dereddening and extinction values are preserved in the injection truth tables for later flux and magnitude corrections to enable consistent comparisons between true and measured quantities.

3.2. *Input Star Sample for y3-stars*

While the majority ($\sim$90%)[10] of the injections are sources (both stars and galaxies) from the DES DF, $\sim$10% of injections are simulated stars. The morphologies are modeled as pure delta (δ) functions convolved with the local PSFEx solution used during injection. The magnitude and color distributions are based on the local stellar population in each of the 10,338 tiles in the Y3 footprint. For example, areas of the survey with higher stellar density near the galactic plane received more bright stars than areas toward the south galactic pole in the center of the footprint. To represent color distributions fainter than the WF limit of $i\sim24$, the color distribution near $i\sim24$ was extended by two magnitudes to $i\sim26$ using models of the Galactic disk and halo (Bienaymé et al. 2018). The simulated star catalog has already been corrected for extinction, so no other preprocessing is required. The measurement pipeline has no knowledge of the difference in input star/galaxy classification and returns the same CModel fits as `y3-merged`.

[10] While most tiles were run with a 9-1 ratio between input catalogs, the first 152 tiles of `Run2` were run with an 8-2 ratio.

Besides characterizing the photometric response of stars in DES with nearly no galaxy contamination (see Section 4.2), the `y3-stars` sample is useful for quantifying the baseline performance of the DESDM pipeline for the simplest morphologies. This allows us to isolate the more complex model fitting issues for the heterogeneous `y3-merged` sample.

3.3. *Object Classification and Differences in Measurement Likelihood*

While we expect `y3-merged` and `y3-stars` will be used for calibration of DES galaxy and stellar systematics respectively, there are additional star injections in `y3-merged` as it draws from all sources in the DF that pass quality cuts. Sources in the DF catalog have been classified with a k-nearest neighbor algorithm[11] trained on a subset of objects that have near-infrared (NIR) data from the UltraVISTA survey (Hartley, Choi et al. 2020; McCracken et al. 2012). The classifier's stellar sample is not perfectly complete from magnitudes $18 < i < 24$ (an average of 93%), but it's mean weighted purity is greater than 98% over the same range. The requirement of successful detection and measured photometry for all *ugrizJHK* bands reduces the total number of objects with classification by 44.5%. The cut `NearestNeighbor_class=2` selects this star sample while `NearestNeighbor_class=1` will select the classified galaxies. The DF stars are not used in the analysis of the Y3 stellar photometric performance in this paper but are available if a larger sample is required for a given science case. However, we do use these classifications when estimating the galaxy contamination in Y3 stellar samples in Section 4.4.

We note that there is a subtle difference in the measurement likelihoods corresponding to each sample. The likelihood of the δ-sample, $\mathcal{L}^{\delta}_{\rm star}$, assumes perfect classification knowledge and is given by

$$\begin{aligned}\mathcal{L}^{\delta}_{\rm star} &= p(\boldsymbol{\theta}_{\rm meas},\, c_{\rm meas}|\boldsymbol{\theta}_{\rm true},\, c_{\rm true}{=}{\rm star}) \\ &= p(\boldsymbol{\theta}_{\rm meas},\, c_{\rm meas}|\boldsymbol{\theta}_{\rm true}),\end{aligned} \tag{1}$$

where $\boldsymbol{\theta}_{\rm meas}$ and $\boldsymbol{\theta}_{\rm true}$ are the measured and true objects' photometric parameters and $c_{\rm meas}$ and $c_{\rm true}$ are the corresponding object classifications. Alternatively, the likelihood of the DF star sample, $\mathcal{L}^{\rm DF}_{\rm star}$, accounts for the uncertainty in the truth classification:

$$\mathcal{L}^{\rm DF}_{\rm star} = p(\boldsymbol{\theta}_{\rm meas},\, c_{\rm meas}|\boldsymbol{\theta}_{\rm true},\, c_{\rm true}). \tag{2}$$

[11] This classifier was added after the `Balrog` runs completed, and so is not included as one of the truth columns. It has to be matched to the relevant Y3 DF catalogs.

This becomes particularly relevant if one wants to combine results from Sections 4.2 and 4.3 for modeling errors of the composite sample. The needed conditional probabilities that capture the stellar efficiency and galaxy contamination of `y3-merged` can be derived from the results in Section 4.4.

3.4. *Sample Selection & Injection Strategy*

While in principle we would randomly sample from all sources in the DF, there are some methodological and practical considerations that led to the following conservative cuts:

```
      flags = 0
AND   mask_flags = 0
AND   in_VHS_footprint
AND   bdf_T < 100
AND   bdf_flux / bdf_flux_err > -3
AND   bdf_det_mag < {25.4, 24.5}
```

First, we eliminate any objects flagged with model fitting errors or in manually masked regions. We also require injections be from regions with external observations in the near-infrared (IR) as these IR bands are critical for the photometric redshift calibration (5.1). We restrict the characteristic size of the injections (`bdf_T`) to be less than 100 arcsec^2 (corresponding to ∼10 arcsec) to reduce the rate of `Balrog-Balrog` blends and proximity effects on the injection grid – though this selection may result in slightly over-sampling large, highly-elliptical galaxies. In addition, this choice may be in conflict with other potential science cases such as measuring the detection efficiency and photometric response of low-surface-brightness (LSB) galaxies (Tanoglidis et al. 2020). Next, we remove objects with flux to error ratios of less than -3 in any band; this cut was needed after inspection of the DF catalog showed that there was an excess of objects with extremely negative flux values compared to WF measurements (though `ngmix` fluxes are clipped below 10^{-3} when computing magnitudes).

Finally, we apply a detection magnitude limit of 25.4 to limit the time spent on injections that have almost no chance of being detected while still using a source catalog that is ∼2 magnitudes deeper than WF. As described in the beginning of Section 3, this limit was derived from the mean dereddened *riz* `bdf_flux` of injections that had at least a 1% chance of being detected during a 200 tile test of `Run2`. We do not consider the flux in *g* in this calculation as it is not used in the detection image in DESDM processing. The `Run2a` limit of 24.5 was chosen based based on requirements for the lens magnification measurement detailed in Elvin-Poole et al. (2020) (and described further in Section 5.2). After

making this selection, the DF injection catalogs used in `Run2` and `Run2a` have just over 1.23 million and 746,000 objects respectively.

The star catalog was sampled to its full depth of 27th magnitude in g at a fraction of 10% of the total objects injected into `Run2a` and (most) `Run2` tiles. No additional cuts were made. Since the relative contribution of Galactic stars to the total object count peaks at about 21st magnitude in a standard Y3 tile, these injections do not dominate the faint end of the distribution.

Choosing the injection density per realization is a trade-off between increasing the statistical power of the catalogs, reducing the rate of `Balrog-Balrog` blends, and reaching the desired footprint coverage given available computational resources. Ideally, we would measure the response of a single source added to DES images for a high number of realizations. As this is unfeasible we instead add objects on a hexagonal lattice with $20''$ spacing using `MixedGrid` (see §2.2.3) for a single realization, corresponding to a density of ∼7.8 objects per arcmin^2 (or about 40% of the total Y3 density). We can achieve a much higher injection density than that used in Suchyta et al. (2016) as we do not randomly sample the positions which greatly reduces the self blending rate of injections. This is crucial as running a single `Balrog` tile realization in Y3 takes ∼40 times longer than in SV and Y1 due to the increased complexity of the injection framework and additional photometric measurements.

However, this relatively high density could have significant implications for a non-local deblender like the one used in MOF. In early testing, we found that this level of injection density can sometimes lead to nearly all objects in a tile becoming a single MOF FOF group. Such non-local effects are less relevant for SOF except in cases where blends of other nearby injections with large, real sources may change how the masking of the blend is handled (or for extremely large injections that would be captured in the MEDS cutout of other injections, which is why we cut on the injection size). Dealing with non-local contributions to the measurement likelihood may be an important consideration for Y6 as the object detection threshold is lower and proximity effects are more of a concern.

3.5. *Blending and Ambiguous Matches*

An important caveat in using an object injection pipeline like `Balrog` is that there is often inherent ambiguity in the matching of the new object catalogs to the injections. Remeasurement on the injection images changes the number of detections and catalog ID assignments in unpredictable ways, and light profiles that were previously considered distinct detections can be blended together into single objects. While we will show that the fraction of ambiguous cases is relatively small at our injection density in DES images (<1.5%) and can in principle be removed for our photometric tests, this ignores the increased shear noise and root mean square (RMS) of the measured ellipticity distribution for these objects which may be a dominant systematic for weak lensing measurements in deeper surveys like LSST (Dawson et al. 2015). In addition, highly non-linear detection and photometry algorithms can often respond in unexpected ways to perturbations (particularly deblenders that are intrinsically non-local) which can lead to additional spurious detections and splitting of objects. As a rule: *Any matched catalog from an injection pipeline has made assumptions about ambiguous matches and blending!* For these reasons, we save the full remeasured photometry catalogs so that different matching procedures can be applied depending on the desired science case. This is distinct from the approach in Suchyta et al. (2016) which ran remeasurement in `SExtractor`'s association mode near injection positions.

However, it is useful to have a standard catalog sample with consistent matching for downstream cosmological analyses. Unless otherwise specified, Y3 analyses using `Balrog` catalogs use a catalog which applied the following matching prescription: We define the *antecedent* of any blend as the "brightest" of the individual objects that contributes to it by some metric. Each blend thus comprises a noisy version of the antecedent as well as the non-detection of all other contributors to the blend. This approach gives a consistent and complete assignment of detection, non-detection, and antecedent to all objects of interest in the remeasured images and strikes the desired balance of including photometric scatter by blend contributors while excluding extreme outliers due to faint injections near existing bright objects. In addition, in the absence of measurement noise this scheme sets a maximum for the possible flux error of the antecedent in a two-object blend to be $|\Delta\text{mag}|\sim 0.75$; a factor of 2. An overview of how this scheme applies to the most common case of a two-object blend is shown in Figure 7.

The above prescription requires a brightness metric to determine the antecedent. We use the average of the dereddened Gaussian-weighted aperture (GAp) fluxes in each of the DES detection bands (riz). GAp fluxes are conceptually similar to GaaP fluxes described in Kuijken, K. (2008) but instead measure the aperture flux for source profiles *before* convolution with the PSF. These fluxes are computed analytically from the MOF `bdf` fits to the DF injections and the SOF CModel fits to Y3 GOLD objects using a Gaussian weight function with

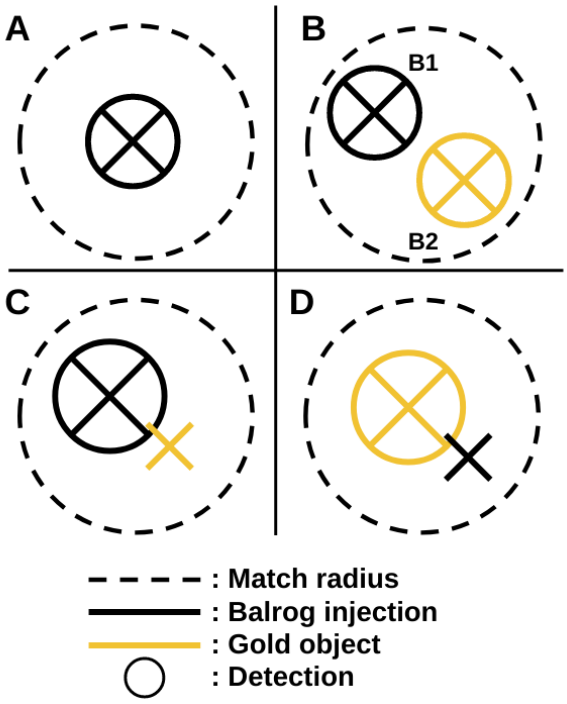


Figure 7. An overview of how ambiguous matches can arise in the case of a two-object blend. A black cross mark denotes the position of a `Balrog` injection while a gold cross mark denotes the position of a Y3 GOLD detection. A circled cross mark indicates a detection in the `Balrog` catalog while the dashed circle indicates the region inside of the search radius r_2. Case (A) is by far the most common and is unambiguously a `Balrog` injection. Case (B) has both the injection and the GOLD object detected within r_2 but is *extremely* rare; in this case we select the closer detection. Cases (C) and (D) are true blends where there is ambiguity in whether to classify it as a `Balrog` object with properties blended by the GOLD source or as a GOLD object that was blended by an injection. In this case we assign the object with the larger average *riz* GAp flux as the antecedent. Only Case (D) is removed from the `Balrog` catalogs when applying a `match_flag` cut.

FWHM of $4''$. This allows us to use an estimate derived from our best guess of the flux of the PSF-deconvolved profile near the relevant object centroids while discounting variations in measured flux due to morphological differences – particularly those arising from significant flux contributions from the wings of extended profiles. We use the average of the detection band δ fluxes for `y3-stars` as an equivalent GAp flux is not well defined. This difference only becomes relevant for the brightest star injections, though in these cases they are very likely to be the antecedent.

The matching procedure is implemented in two separate steps. First, the injection positions are matched to the closest object in the remeasured photometry catalogs within a search radius of $r_1 = 0.5''$. All objects that have a match are saved in the output `Balrog` catalogs and undergo the aforementioned post-processing steps. Afterwards, the output catalogs are matched against the Y3 GOLD catalog to compare the relative brightness of any existing detections within a second match radius r_2 for a series of radii from $0.5''$ to $2.0''$ in increments of $0.25''$. Over 96% of candidate objects have no GOLD sources within the search aperture are unambiguously a `Balrog` injection[12]. Candidates that have an existing GOLD object within r_2 with mean *riz* GAp flux below their own are considered the antecedent and given a `match_flag_{r2}_asec=1` to indicate the presence of a nearby real source. Candidates that have a match within r_2 but have a smaller mean GAp flux than the existing object are assigned `match_flag_{r2}_asec=2` and are recommended to be cut from science analyses. We encode this information as a flag instead of cuts to the fiducial catalog to allow `Balrog` users more flexibility in choosing how to handle blending and ambiguous cases as needed. In this paper, we cut on `match_flag_1.5_asec < 2` as we found that only 0.1% and 0.5% of Y3 GOLD objects were separated at distances less than $1.5''$ at i magnitudes of 21 and 22.5 respectively (or about 1.3-1.8 times the median PSF size depending on the band).

We show in Figure 8 the the difference between the recovered and injected GAp magnitude, $\Delta\mathrm{mag}_{\mathrm{gap}}$, for all recovered `Run2` objects for three choices of ambiguous matching cuts. In the left panel where no cut on ambiguous matches has been made, there is a long, asymmetric tail for negative $\Delta\mathrm{mag}_{\mathrm{gap}}$ where the recovered GAp flux is up to 10 magnitudes brighter than the input. While there can be extremely large magnitude responses to model fitted photometry in crowded fields or extreme imaging conditions (see §4.3.3), we expect GAp magnitudes to be less sensitive to these failure modes and most large discrepancies to be due to ambiguous matches. This is indeed the case: In the following panels where a match flag with r_2 of $0.5''$ and $1.5''$ are used to create the sample, the worst GAp response outliers have been removed and the fraction of detections where $|\Delta\mathrm{mag}_{\mathrm{gap}}| > 1$ falls by 41% and 65% respectively. Some remaining scatter beyond $|\Delta\mathrm{mag}_{\mathrm{gap}}| = 0.75$ is expected even for an optimal r_2 due to ambient light in dense fields, blends with extended sources, and image artifacts, though the number of objects below $\Delta\mathrm{mag}_{\mathrm{gap}} = -1$ for the $1.5''$ cut falls by over an order of magnitude for each bin of unit size.

[12] In principle there can be rare exceptions to this such as new spurious detections very close to injection positions, but we do not consider that here.

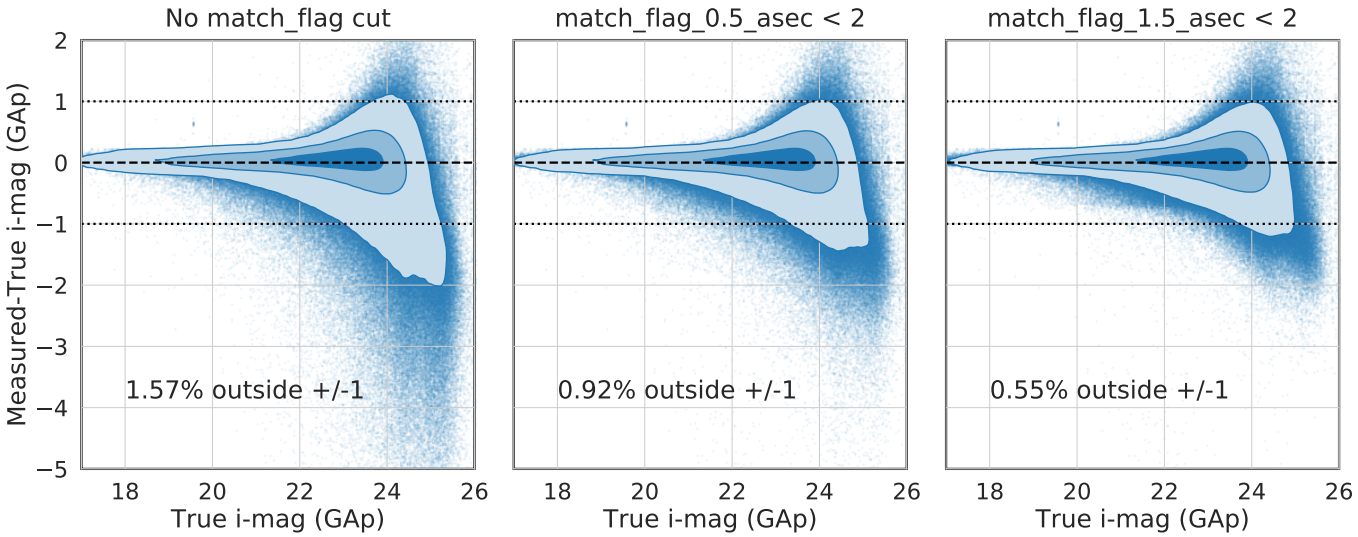


Figure 8. The effectiveness of our ambiguous matching scheme, illustrated by the difference in measured vs true i-band GAp magnitude ($\Delta\text{mag}_{\text{gap}}$) as a function of input GAp magnitude for three ambiguous matching choices. The overplotted contours contain 39.3%, 86.5%, and 98.9% of the data volume, corresponding to the volume contained by the first three σ's of a 2D Gaussian distribution respectively. The percentage of detections outside of the dashed region denoting $|\Delta\text{mag}_{\text{gap}}| < 1$ for each choice is labeled in the bottom left of each panel. The left panel shows the $\Delta\text{mag}_{\text{gap}}$ response for `y3-merged` when no cut is made to handle ambiguous matches. There is an extremely long outlier tail of injections measured to be significantly brighter than the injected flux both from real effects (See §4.3.3, though GAp fluxes are much less sensitive to these failures) and ambiguous blends. The following two panels show the same distribution after cutting on the match flag using a r_2 of $0.5''$ and $1.5''$ respectively. The outlier tail significantly decreases in size as more ambiguous blends are accounted for, with nearly three times less objects outside of $|\Delta\text{mag}_{\text{gap}}| < 1$ when using the fiducial value of $r_2 = 1.5''$.

4. DES Y3 PHOTOMETRIC PERFORMANCE

Here we present the photometric performance of the Y3 `Balrog` DF sample `y3-merged` along with the synthetic star sample `y3-stars`. While there are many photometric catalogs and science samples of interest for Y3, here we largely focus on the SOF CModel photometry of a basic Y3 GOLD sample (Sevilla-Noarbe et al. 2020) used as a starting point for more restrictive samples. Unless otherwise specified, the cuts for this sample are given by

```
    FLAGS_FOREGROUND = 0
AND FLAGS_BADREGIONS < 2
AND FLAGS_FOOTPRINT = 1
AND FLAGS_GOLD_SOF_ONLY < 2
AND EXTENDED_CLASS_SOF >= 0
AND MATCH_FLAG_1.5_ASEC < 2,
```

along with any appropriate object classification cut which will be mentioned when relevant. Note that `FLAGS_GOLD_SOF_ONLY` is used in place of the typical `FLAGS_GOLD` as we are unable to compute the first bit flag without `y3-merged` MOF runs. While ∼3.5% of Y3 GOLD objects have `FLAGS_GOLD=1`, no Y3 cosmology analyses currently use this flag bit due to the use of SOF or Metacalibration photometry in favor of MOF. Additional samples for a few interesting `Balrog` applications are discussed in more detail in Section 5.

We begin by examining how representative the `Balrog` catalog properties are compared to Y3 GOLD in Section 4.1, including a detailed look at how the number density fluctuations of both samples vary with respect to survey property maps. We then show the magnitude and color responses of `y3-stars` and `y3-merged` along with a discussion of interesting photometric failure modes in Sections 4.2 and 4.3 respectively. We then end by characterizing the performance of the `EXTENDED_CLASS_SOF` star-galaxy separator, using the extremely pure `y3-stars` sample whenever possible. As it is not practical to plot the photometric responses of all quantities of interest, one-dimensional Gaussian summary statistics for many relevant parameters are provided in Appendix C.

4.1. *Consistency with DES Data*

Even without perfect emulation fidelity, we expect the measured `Balrog` property distributions to closely resemble DES catalogs if we are indeed sampling an adequately representative transfer function and input sample. We will broadly check this agreement at various steps along the measurement path: object detection, photometric properties, and correlations with survey systematics – along with how these differences im-

pact a typical clustering signal measurement. As we are primarily interested in the consistency in the transfer function of galaxies for cosmology, we use the `y3-merged` sample throughout and mention any classification cuts when relevant.

4.1.1. *Completeness*

We begin with object detection. Of the nearly 26.5 million galaxies injected in `y3-merged`, just over 41.9% were detected during re-measurement after accounting for ambiguous matches. However, as this catalog is the merger of two runs with different magnitude limits, it is more accurate to say that 36.3% and 59.4% of objects were recovered for `Run2` and `Run2a` respectively. The fraction of injections contained in the fiducial sample drops to 14.4% and 44.2% after considering the basic flag and mask cuts described above. To simplify the comparison on the faint end we use only `Run2` for the following comparison as it is about a magnitude deeper.

The detection completeness of sources in *griz* for `Run2` (points) compared to Y3 GOLD objects in the X3 supernovae field (lines) is shown in Figure 9. The completeness is plotted as a function of reference magnitude; the injection magnitudes for `Balrog` and the DF measurements of objects in the X3 field for Y3 GOLD. As we are comparing the mean completeness of the `Balrog` sample across all `Run2` tiles to only a small region for Y3 GOLD, to make a fair comparison we estimate the uncertainty in the difference with 50 jacknife samples of the `Run2` footprint. Note that the inferred completeness is only robust until the forced magnitude limit cutoff of 25.4 indicated by the dashed vertical line; beyond this point, the sampled injection objects have inherited a selection bias that forces at least one of the other detection bands to be significantly brighter than the magnitude limit and thus is more likely to be detected.

Overall the completeness measurements are quite similar, with the only statistically significant discrepancies occurring for the brightest g-band magnitudes and the faintest i and z bin. The `Balrog` g-band completeness dips on the bright end despite the very high S/N as g is not included in the composite detection magnitude image limit, and thus objects bright in g-band but not in other bands are sometimes not detected. This is not seen as significantly in the Y3 GOLD sample which suggests that the input DF sample over represents these kinds of objects. It is more difficult to determine possible discrepancies past the detection threshold in each band without careful examination of both measurements, though their residuals are only marginally beyond 1-σ and could simply be statistical fluctuations. While it is encouraging to see similar detection properties between `Balrog` and the data, that alone is not enough to ensure sufficient similarity for science calibrations.

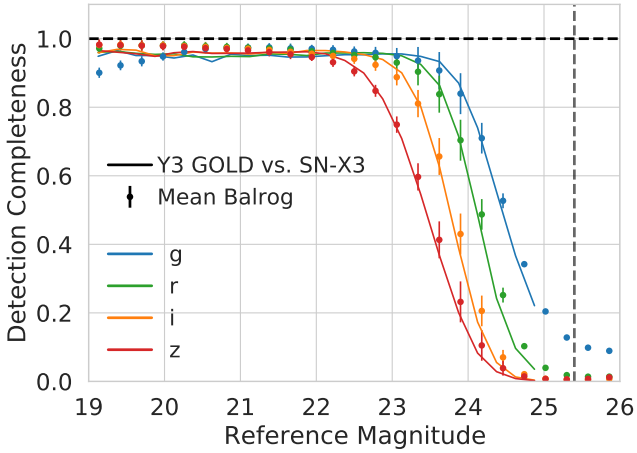


Figure 9. The fraction of objects recovered by band and input injection magnitude. Solid lines show completeness measurements comparing the wide and deep samples on the SN-X3 field as described in Section 5.2 of Sevilla-Noarbe et al. (2020). Points with error bars are the Balrog mean completeness measurements for the full sampled `Run2` footprint. Errors are the standard deviation of 50 jacknife samples of the sampled footprint, rescaled as appropriate for the area of the SN-X3 field. The dashed vertical line indicates the injection effective magnitude limit of 25.4.

4.1.2. *SOF Photometry*

We can make similar comparisons of the measured photometry. Figure 10 compares the recovered `Balrog` SOF *griz* magnitudes, $g-r$ and $r-i$ colors, and a few morphological parameters to Y3 GOLD after both samples have applied basic cuts. The comparison is in absolute counts with `Balrog` in blue and the mean of 100 GOLD bootstrap subsamples of identical size to the `y3-merged` sample in black. The standard deviation of the subsample counts in each bin are used to estimate the uncertainty and the percent errors of the binned residuals are plotted below each distribution.

Qualitatively, the distributions are extremely similar in the most dense regions of parameter space for most quantities, with the most obvious discrepancies occurring in the low-density tails of the distributions. This is particularly noticeable for the magnitudes and colors. The relative residuals confirm this: While nearly all `Balrog` magnitude bins have fractional distribution differences below 5% of the mean Y3 GOLD sample from 18 to 24, the region of interest for most Y3 cosmological analyses, `Balrog` counts in magnitudes below 18 underestimate GOLD by 10 to 50% by magnitude 16. The colors are similar, with the only discrepancy above 5%

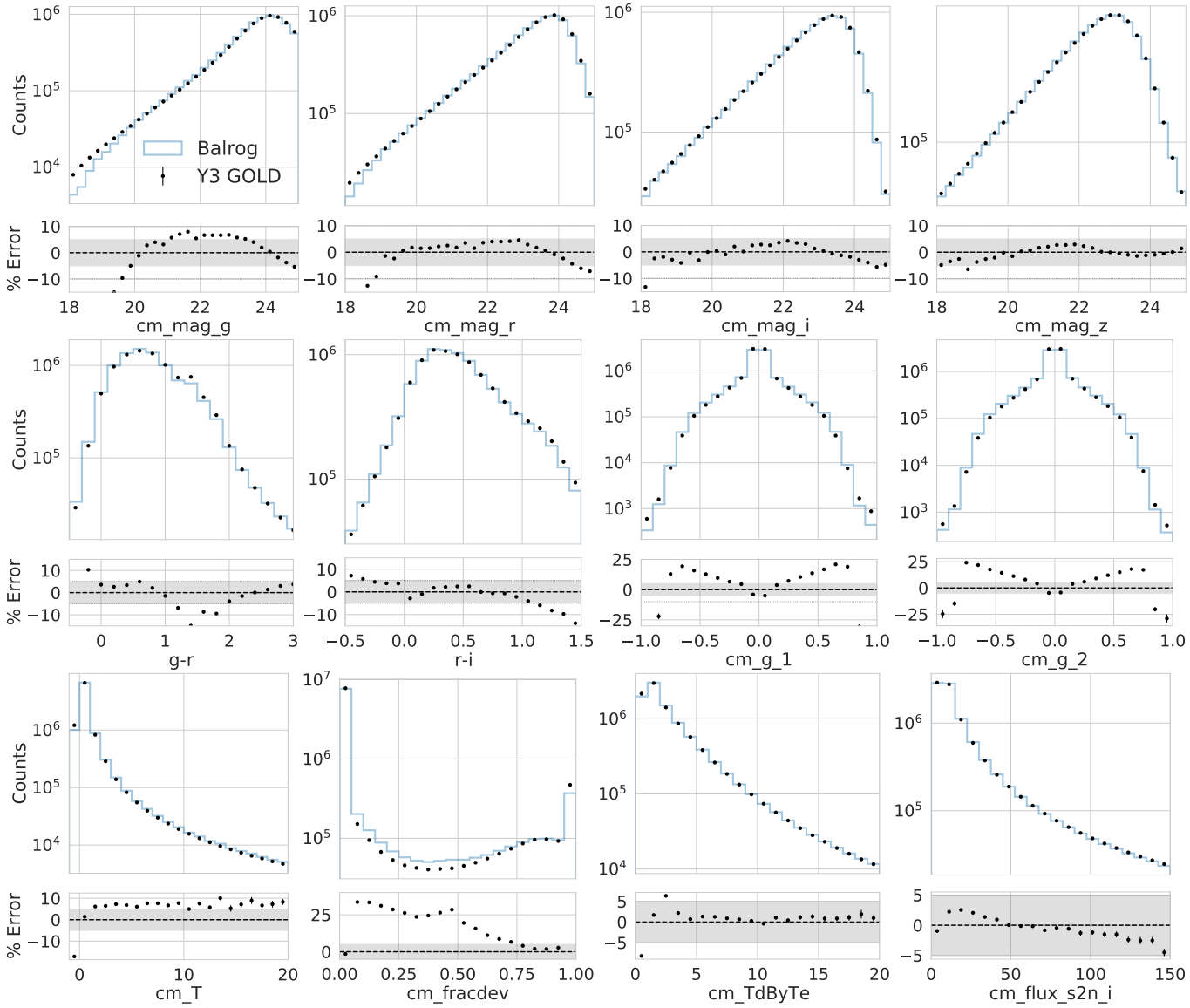


Figure 10. Comparison of the `y3-merged` sample (in blue) vs. Y3 GOLD (in black) for measured *griz* magnitudes, $g-r$ and $r-i$ color, and a variety of morphological parameters. Both samples have had the basic cuts applied as described in Section 4. To compare the distributions, we resample Y3 GOLD with replacement to match the size of the `y3-merged` catalog 100 times and plot the mean and std of these bootstrap samples in black. The percent error of the binned residuals are shown below each distribution, which have been zoomed in to show the results of the most relevant regions. The region corresponding to +/-5% has been shaded in gray. When quantities do not have hard boundaries, we include at least the 2nd-97th percentiles of the values. The residuals are very sensitive to selection cuts. For example, the discrepancies at `cm_T`< 0 and |`cm_g_{1/2}`|~ 1 are significantly smaller after cutting out suspected stars from the sample.

in the densest regions occurring at $1.3 < g-r < 1.5$; values typical of M-dwarf stars (Smolčić et al. 2004). A few other notable discrepancies are that `Balrog` appears to underestimate the number of objects with ellipticities `cm_g_{1/2}`~ 0 and negative size parameter `cm_T` relative to the Y3 GOLD sample - both of which are again values typical of stars.

We stress that these binned residuals are still a largely qualitative check on the agreement between property distributions as they are very sensitive to sample selection. For example, the relative error in `cm_T`, `cm_g_1`, and `cm_g_2` near zero are all significantly smaller after applying the stellar cut `EXTENDED_CLASS_SOF > 1` which indicates that the y3-merged sample does not capture the transfer properties of stars as well as galaxies. Yet the shape of these residuals often indicate important real differences. The change in residual sign near the detection threshold in each band indicates potential small

differences in the effective depth of the samples, and the overabundance of Balrog objects with cm_fracdev near 0.5 reflects the effect of parameter priors not matching the true underlying distribution as discussed in section 2.3.

In addition, residuals consistent with zero even under the assumption of perfect emulation fidelity requires a completely representative input sample. There are many known reasons for why our input sample fails this requirement, a few of which we discuss here: (i) The DF sample underestimates cosmic variance as it only uses objects from a tiny fraction of the sky, which is particularly a problem for the stellar population as its distribution varies across the sky much more strongly than galaxies. (ii) The photometric pipeline used to make measurements of DF objects is not identical to the one used in the WF in order to deal with non-dithered observations, an increased blending rate, the large number of exposures per detection, and instabilities in the detection of very faint sources in the presence of diffuse emission (see Hartley, Choi et al. 2020). (iii) The morphological model fits to the DF objects are subtly different (bdf vs cm) which we have shown can introduce small biases in other parameters such as the magnitude. (iv) CModel is not an appropriate photometric model for all objects in the sky. There are simple practical limitations that contribute to these discrepancies as well, such as limiting the size and magnitude distribution of objects to reduce Balrog-Balrog blends and the computational time spent on injecting near certain non-detections. We discuss these issues more in Section 6.

4.1.3. *Spatial Variation and Property Maps*

While the overall similarities in the photometries are encouraging, what is most critical is how well Balrog reproduces the measurable signals used in cosmological analyses as well as correlations with spatially varying image conditions and survey properties. These systematic trends are particularly important when measuring the galaxy clustering signal where local observing conditions can imprint fluctuations in number density that are not cosmological in origin such as variations in seeing, depth, and sky brightness (Rodríguez-Monroy et al. 2020). We now investigate the similarity of these systematic trends in Balrog and Y3 GOLD for a highly incomplete sample where the variation is more apparent, before looking at their contribution to the clustering signal itself for a cosmology-like sample in §4.1.4.

Figure 11 compares the number density of all y3-merged and Y3 GOLD galaxies with basic cuts as a function of survey property in overlapping HEALPix (Górski et al. 2005) pixels of NSIDE=2048, corresponding to an area of 2.95 arcmin^2. The survey properties are assigned from the Y3 HEALPix maps in Sevilla-Noarbe et al. (2020) (based off the methodology in Drlica-Wagner et al. 2018) that have been rescaled[13] from a $N_{\rm side}$ of 4096 to 2048 to smooth out irregularities in the pixel occupation distribution due to the regular structure and lower density of Balrog sources. The uncertainty in number density was estimated by resampling the pixels used in each sample of equal size with replacement for 100 bootstrap samples. The distribution of the rescaled survey properties for the Y3 GOLD sample are plotted in the background in green to highlight typical property values.

With a few notable exceptions, the number density of the two samples match closely in both amplitude and shape. It is especially encouraging to see Balrog capturing the high frequency structure in the dependence of a few of the more complex trends such as the local sky brightness (skybrite) and airmass. The largest differences in recovered number density occur for extremely rare values of a few properties such as the quadrature sum of zeropoint uncertainties (sig_zp) and exposure time (exp_time) and are not particularly concerning. However, there are still some more serious unresolved discrepancies in amplitude – particularly in r-band seeing and airmass. The same potential issues in input sample representativeness and photometric assumptions discussed previously apply to these measurements, but it is not immediately clear why these issues would manifest in a band-dependent fashion in seeing or why the largest discrepancies occur for an indirect parameter of the images like airmass. These differences may be indicative of features in the transfer function not currently captured by Balrog such as PSF modeling errors with unexpected chromatic effects or the unapplied injection zeropoint corrections. Such differences warrant further investigations in preparation for an improved Y6 Balrog methodology but do not themselves indicate insufficient consistency for a clustering measurement. We explore this further below.

4.1.4. *Galaxy Clustering Systematics*

Many of the core science cases of interest to cosmology involve measurements of galaxy clustering. To be useful in calibrations for this purpose, it is not enough that the number counts of Balrog and Y3 GOLD galaxies follow the same trends with image properties like those shown in Figure 11. Where the systematic error is independent of the signal (as, for example, variations in the airmass and the true galaxy density on the sky

[13] The map rescaling is done by averaging all non-empty pixels.

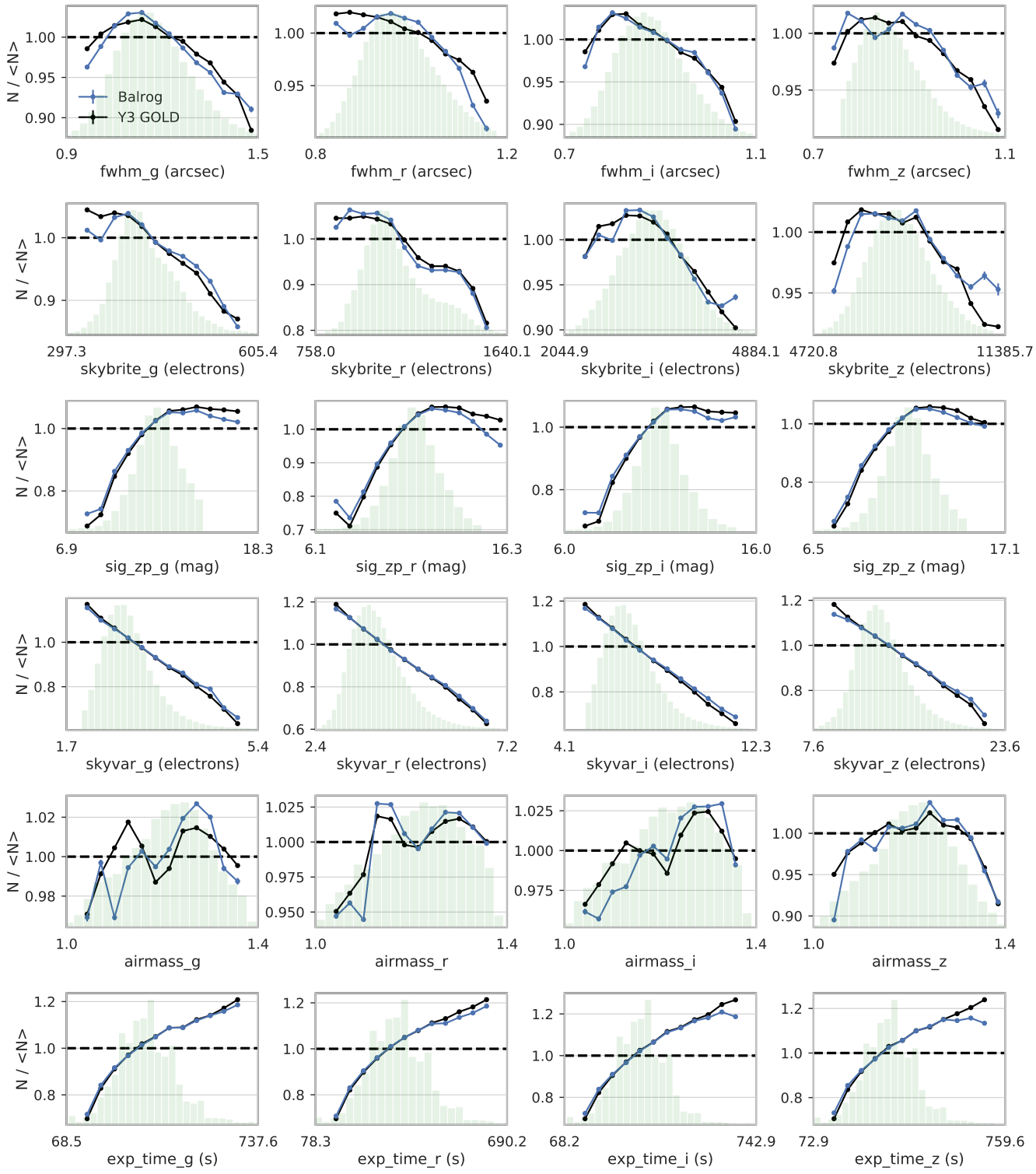


Figure 11. The trend in number density fluctuations $N/<N>$ as a function of various survey observing properties for the full (and highly incomplete) `Balrog`, in blue, and Y3 GOLD, in black, samples after basic cuts for overlapping HEALPix pixels of `NSIDE`=2048. The maps have been rescaled from `NSIDE`=4096 as described in the text to better handle the regular structure of `Balrog` injections. The distribution of survey condition values for the rescaled Y3 GOLD map is displayed in the background in green to highlight typical values. The errors have been estimated by resampling the pixels used in each sample with replacement for 100 realizations. The property maps are described in Table E.1 in Sevilla-Noarbe et al. (2020), but we briefly defined them here in order from the top: the mean PSF size, the local sky brightness, the quadrature sum of the zeropoint uncertainties, the variance of the sky brightness, the airmass, and the exposure time. `Balrog` captures many of the nonlinear features in the trend lines, though there are some unexplained band-dependent discrepancies in some property maps.

are statistically independent of one another), the resulting variations in survey depth enter, to leading order, as additive systematic errors in the two-point statistics used for cosmology. Correcting for these observational systematics is critical for unbiased cosmological inference from clustering, and the ability to use `Balrog` as object randoms with realistic measurement biases, if it sufficiently captures the clustering fluctuations of the data, offers an ideal calibration method without using the data vector directly which avoids possible overfitting (see Suchyta et al. 2016; Garcia-Fernandez et al. 2018). In addition, direct calibration with `Balrog` would eliminate the need to identify all sufficiently important survey property contributions at a desired precision (and avoid biases from any unidentified systematics) while potentially allowing for measurements on larger scales where the true signal is very small and the corrections have to be *extremely* accurate.

Here we estimate the approximate impact on the clustering signal due to systematic differences between `Balrog` and Y3 GOLD for a sample broadly similar to the MAGLIM science sample described in Porredon et al. (2020), where we cut both the Y3 GOLD and `Balrog` samples to $17.5 < i < 21.5$ in addition to the previous cuts. We make density maps based on each property map across the full Y3 GOLD footprint by interpolating the trends in `Balrog` and GOLD to fill in cells where we do not have injection samples. These maps are estimates of the MAGLIM galaxy number density fluctuations in Y3 if they could be completely described by the survey property in question[14]. We then estimate the angular power spectra of both interpolated maps for each survey property using the pseudo-C_ℓ estimation code PyMaster[15] (Alonso et al. 2019). These are then compared to the power spectra of the survey property maps themselves along with a typical nonlinear galaxy power spectrum at $z = 0.7$ computed with the `CAMB` (Lewis et al. 2000) implementation of the nonlinear power spectrum described in Mead et al. (2015). Finally, we compute the differences in power from the interpolated `Balrog` and Y3 GOLD density maps as a fraction of the galaxy power spectrum at each ℓ-scale.

Results for the best (g-band PSF FWHM) and worst (i-band `sig_zp`) performing map are shown in Figure 12. Angular clustering systematics for the remaining survey properties, generated in the same way, are shown in Appendix B. For scales comparable to or smaller than the DECam focal plane (approximately $\ell > 200$), the difference between Y3 GOLD and `Balrog` is in all cases less than 1% of the typical amplitude of the angular clustering of galaxies (plotted in black). For some quantities, such as the g-band PSF (shown in the top panel in Figure 12), the differences are several orders of magnitude smaller.

While the differences are small in absolute terms, or as compared to a realistic cosmological signal, the relative deviation between the simulated and real catalogs is in some cases quite large. It is difficult to disentangle the relative contribution to these differences from insufficient sampling across survey property values, issues in the input sample, or missing features in the sampled transfer function (such as the zeropoint corrections discussed in §2.1.1). We discuss these issues further in Section 6. However, that the absolute additive contributions are well below 1% at most relevant scales for even a single realization of a 20% sampling of the footprint gives us confidence that injection simulations like `Balrog` will be crucial for systematics calibration of clustering measurements in Y6 and the next generation of galaxy surveys with even more ambitious precision goals.

Whether `Balrog` is sufficiently similar to Y3 data ultimately depends on the science case and desired measurement precision. In addition, the magnitude of discrepancies can depend strongly on the choice of sample cuts - particularly for those effects related to star-galaxy separation and magnitude limits. However, we find that `Balrog` captures a significant amount of the variation in number density as a function of observing conditions even for *extremely* incomplete samples, and systematics control of well under 1% for the clustering measurement of a typical cosmology sample. For an additional example of how to estimate the contribution of the intrinsic uncertainty in the `Balrog` methodology to the Y3 photometric redshift calibration error budget, see Myles, Alarcon et al. (2020).

4.2. *Photometric performance of y3-stars*

As discussed in 3.2, the injections in `y3-stars` consist of pure delta functions convolved with the local PSFEx solution. The extremely high purity of this star sample with realistic transfer properties is unique to injection pipelines such as `Balrog` where we have truth information about the underlying object classification in addition to its photometry - which is not always the case for galaxy samples (discussed further in Section 4.3). This eliminates the need for a traditional star-galaxy separation metric like `EXTENDED_CLASS_SOF` and (nearly) removes any bias resulting from misclassified objects, though we still cut on `EXTENDED_CLASS_SOF` <= 1 to match what is done to create stellar samples in Y3

[14] Where only regions with `Balrog` samples are used for the estimate.
[15] https://pypi.org/project/pymaster/

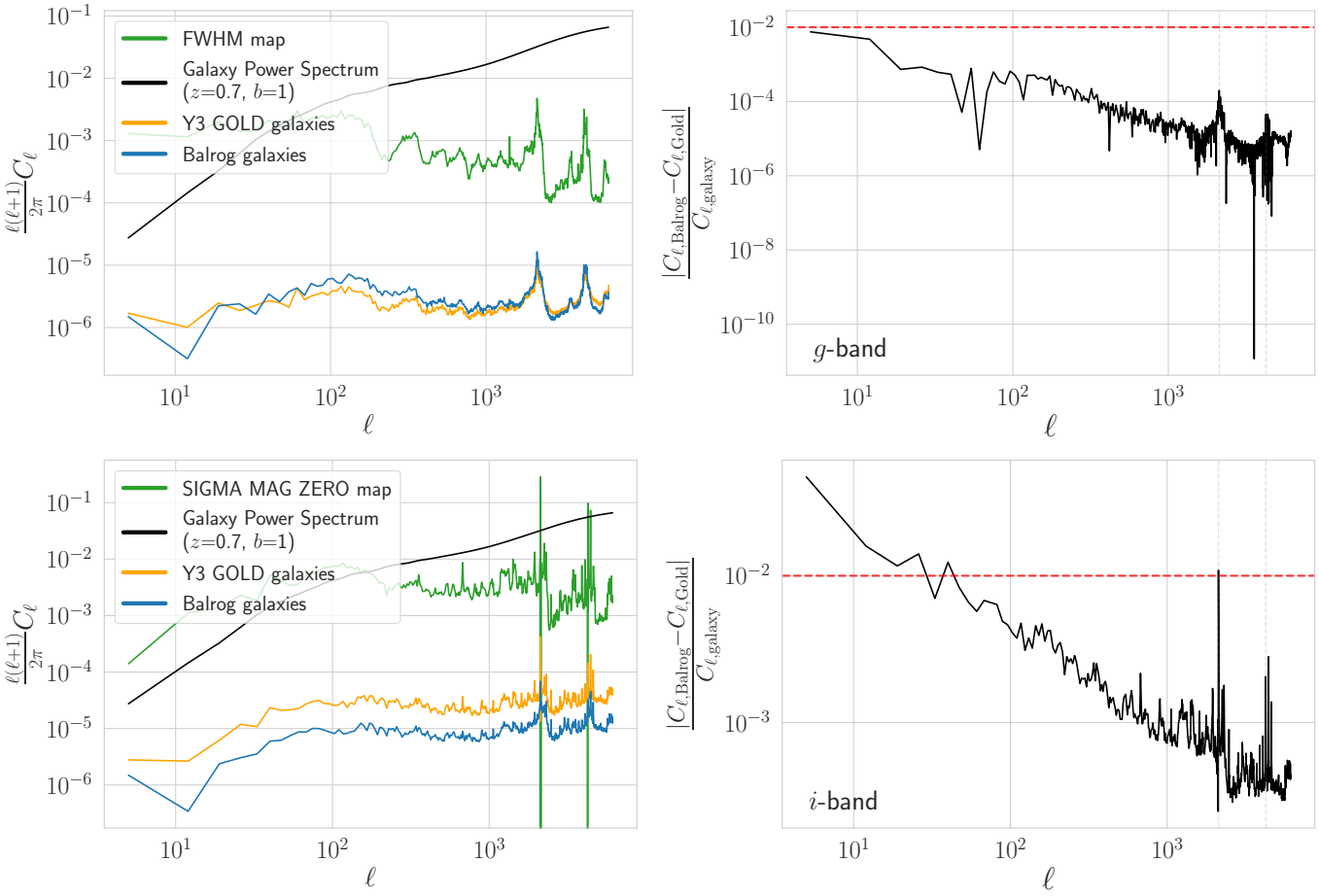


Figure 12. Examples of the survey property maps with the smallest (top row) and largest (bottom row) estimated additive systematic impact on the clustering signal from differences in number density between `Balrog` and Y3 GOLD. The left panels show the angular power spectrum of the noted survey property (in green) and the corresponding power spectra of the number densities of the `Balrog` (in blue) and Y3 GOLD (in gold) MAGLIM-like galaxies across the Y3 footprint using the interpolated trends described in §4.1.3 and §4.1.4. The reference galaxy power spectrum in black is `CAMB`'s implementation of the nonlinear matter power spectrum described in Mead et al. (2015), meant to represent a typical cosmological signal at $z = 0.7$ with linear galaxy bias parameter of 1. The right panels show the difference in power between Y3 GOLD and `Balrog` as a fraction of the fiducial cosmological power spectrum shown on the left. We draw a red dashed line indicating the 1% systematic error threshold as reference. Even in the worst case, we find that `Balrog` is able to capture the clustering amplitude due to variations in survey properties to better than 1% for $\ell > 50$ (corresponding to $\theta > \sim 3.5$) deg. Equivalent plots for many other survey property maps in all *griz*-bands are shown in Appendix B.

GOLD. The only contaminants in the main star sample come from ambiguous matches which is why we still cut on `match_flag_1.5_asec`< 2. This eliminated 1.9% of detections for this sample. Here we focus on the photometric performance and leave the discussion on stellar completeness and galaxy contamination in Section 4.4. We remind the reader that this sample probes a subtlety different measurement likelihood than that of `y3-merged` as we have knowledge of the underlying object classification, as described in 3.3.

While the underlying morphology of stellar profiles is not well described by a Sérsic model, we still use the SOF CModel fits for the stellar sample as there was a systematic calibration offset in the PSF model photometry used in Y3 measurements on the data. This has been corrected for Y6 processing but leaves us without a reliable PSF photometry for our response measurements. However, ultimately this has only a small impact on the recovered photometry for sources smaller than the PSF as these objects are fit with a `cm_T` size near 0 – effectively eliminating the Sérsic components.

4.2.1. *SOF CModel Magnitudes*

The difference in recovered CModel magnitude compared to input magnitude $\Delta\mathrm{mag}_\delta$ as a function of input magnitude for *griz* is shown in Figure 13. Density contours are plotted on top of the scatter with percentiles equivalent to the first three sigmas of a 2D Gaus-

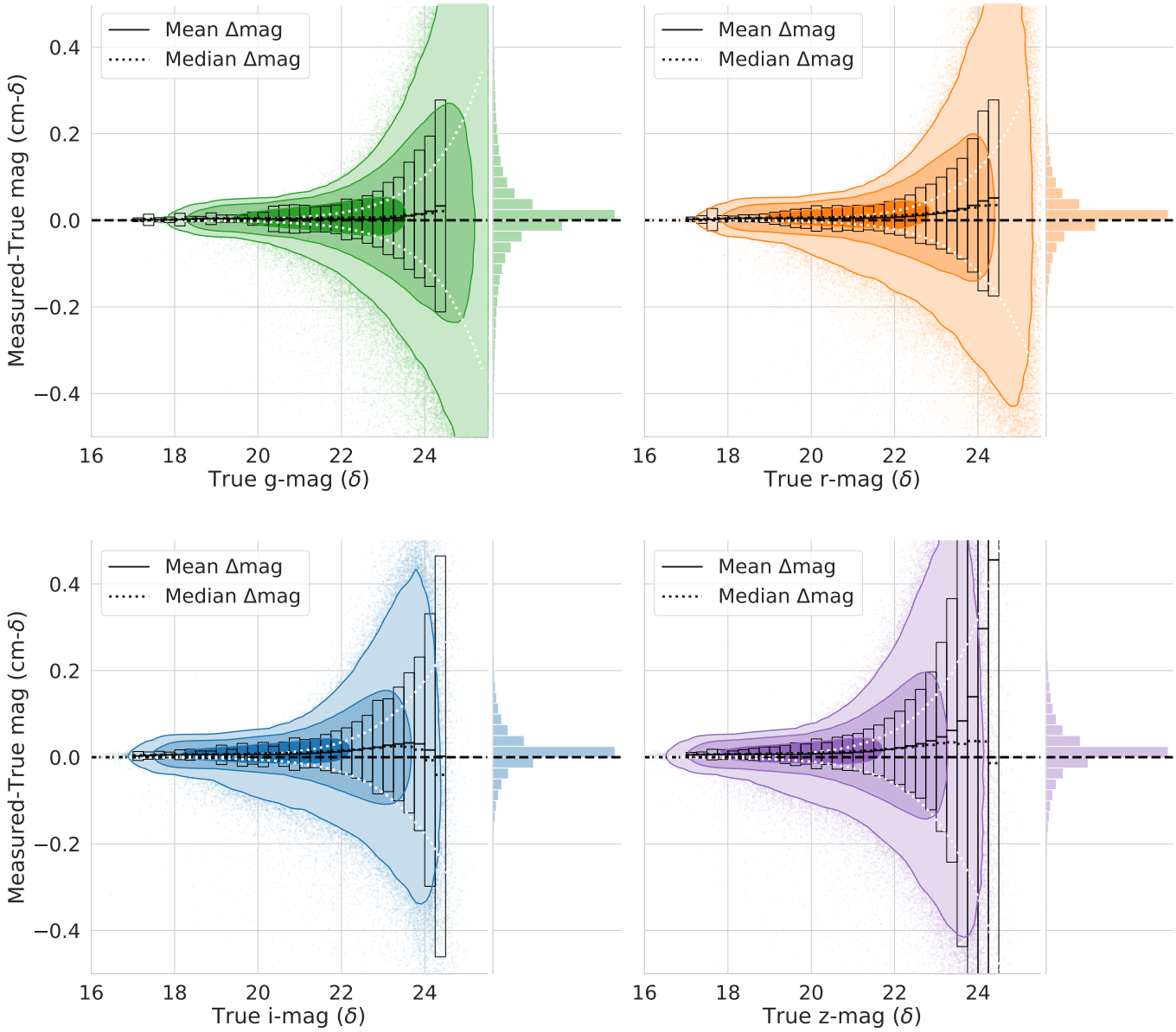


Figure 13. The distribution of differences in recovered *griz* SOF CModel magnitude vs the injected δ-magnitude (Δmag_δ) as a function of input magnitude for the `y3-stars` sample. The density is overplotted where the contour lines correspond to the percentiles of the first three sigmas of a 2D Gaussian, containing 39.2%, 86.5%, and 98.9% of the data volume respectively. The mean (solid), median (dotted), and standard deviation of the magnitude responses in bins of size 0.25 magnitude are shown in the overlaid black bars. These are compared to the reported SOF CModel errors by the dashed white lines which do not attempt to account for systematic effects. The marginal distributions of Δmag_δ are included to highlight the small relative volume of the outlier tails.

sian distribution, corresponding to 39.3%, 86.5%, and 98.9% of the total data volume. The mean response bias $<\Delta\text{mag}_\delta>$, median response $\widetilde{\Delta\text{mag}}_\delta$, and scatter $\sigma_{\text{mag}_\delta}$ in truth magnitude bins of size 0.25 magnitudes are over-potted in black bars. These summary statistics provide estimates for the statistical precision and accuracy of the SOF magnitudes, though we stress that the underlying distributions are not Gaussian. These are compared to the mean reported SOF error in the bin indicated by the dashed white curve which do not attempt to account for systematic effects.

The overall calibration of CModel for the stellar sample is quite good, with $<\Delta\text{mag}_\delta>$ and $\widetilde{\Delta\text{mag}}_\delta$ ranging from 1-10 mmag (or 0.1-0.9%) across all bands up to an input magnitude of 20 and between 2-15 mmag (0.2-1.4%) for 20 $<\Delta\text{mag}_\delta<$ 22 except for the final two z-band bins. $<\Delta\text{mag}_\delta>$ stays under 1.5% for each band in all bins where the number of objects are increasing

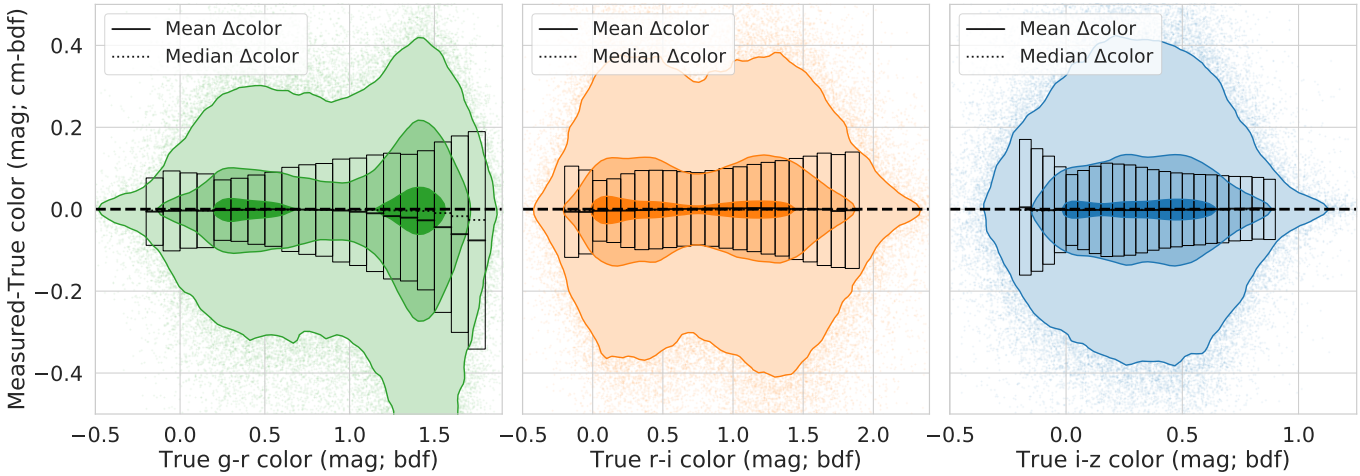


Figure 14. The distribution of differences in measured SOF CModel $g-r$, $r-i$, and $i-z$ color vs. the injected δ-color (Δc_δ) as a function of input color for the `y3-stars` sample. The density is overplotted where the contour lines correspond to the percentiles of the first three sigmas of a 2D Gaussian, containing 39.2%, 86.5%, and 98.9% of the data volume respectively. The mean (solid), median (dotted), and standard deviation of the magnitude responses in bins of size 100 mmag magnitude for $g-r$ and $r-i$ and 50 mmag for $i-z$ are shown in the overlaid black bars.

(input magnitudes of 23.5, 22.5, 22, and 22 respectively) except for the final z-band bin which is $\sim$1.7%. The responses are a bit higher than the quoted 3 mmag uniformity of Y3 GOLD stars when compared to the Gaia star catalog (Sevilla-Noarbe et al. 2020, Gaia Collaboration 2018), though the Y3 GOLD uniformity was measured only with respect to Gaia's G-band which we find to have the best photometric performance (differences of 0.5-6 mmag) over the quoted magnitude range. The Y3 GOLD measurement used a restricted $0.5 < g-i < 1.5$ color range as well which eliminates the worst outliers that we still consider here. In addition, the larger discrepancies found here could be the result of the CModel model misspecification bias discussed previously.

The response bias and scatter increase significantly after these points due to competing systematic effects as the sample becomes progressively more incomplete, with the mean responses rising to $\sim 1.5-3\%$ as they approach the detection threshold in each band. Small sample sizes and strong selection effects lead to $<\Delta\mathrm{mag}_\delta>$ and $\widetilde{\Delta\mathrm{mag}}_\delta$ biases of $\sim$4% for g and r by 24th magnitude, while the biases of the much shallower i and z rise significantly to over 10%. At the median coadd magnitude limits quoted in Table 2 of Sevilla-Noarbe et al. (2020) of 24.3, 23.0, 22.6, and 22.2 (corresponding to a S/N of 10), the mean $griz$ biases are measured to be 3.0%, 4.1%, 2.5%, and 2.2% respectively. The complete set of values for all binned summary statics are included in Table C.1. While the underlying measurement likelihood of these objects is non-Gaussian, the morphological simplicity of stars results in these summary statistics qualitatively capturing the response features well when complete. We will return to this point in 4.3 where the situation is significantly more complicated.

There is evidence of a small band dependence in both the accuracy and precision of the magnitude response. This is most evident when comparing g-band, where $\Delta\mathrm{mag}_\delta$ is never above 5 mmag (0.5%) too faint below an input magnitude of 23.25, to the z-band $\Delta\mathrm{mag}_\delta$ which is exclusively above 5 mmag too faint over the same interval. Unlike the blank image tests in Section 2.3, the $\widetilde{\Delta\mathrm{mag}}_\delta$ values for each band in a bin have a distinct, monotonically increasing shape with the spread between the bands consistently 5-10 mmag brighter than injection magnitudes of 21. However, this effect is much less pronounced when binned by the measured S/N in each band where the detection significance and local sky background is taken into account. Binned in this way, $\widetilde{\Delta\mathrm{mag}}_\delta$ is nearly identical for i and z bands for S/N greater than 20 while g and r are consistently offset by at least 5 and 2 mmag respectively. As this band-dependent response in $\widetilde{\Delta\mathrm{mag}}_\delta$ was not present in the blank image tests, it may suggest issues in the real image calibration such as the estimation of sky background which we discuss more in Sections 4.3 and 5.3.

4.2.2. *SOF CModel Colors*

Of primary interest is the accuracy of the recovered colors due to their importance for photometric calibration, star-galaxy separation, photometric redshift estimation, and the study of Milky Way structure. We plot the difference in measured SOF CModel $g-r$, $r-i$, and $i-z$ color vs. input δ-color with respect to the input color in Figure 14. The contours and summary statis-

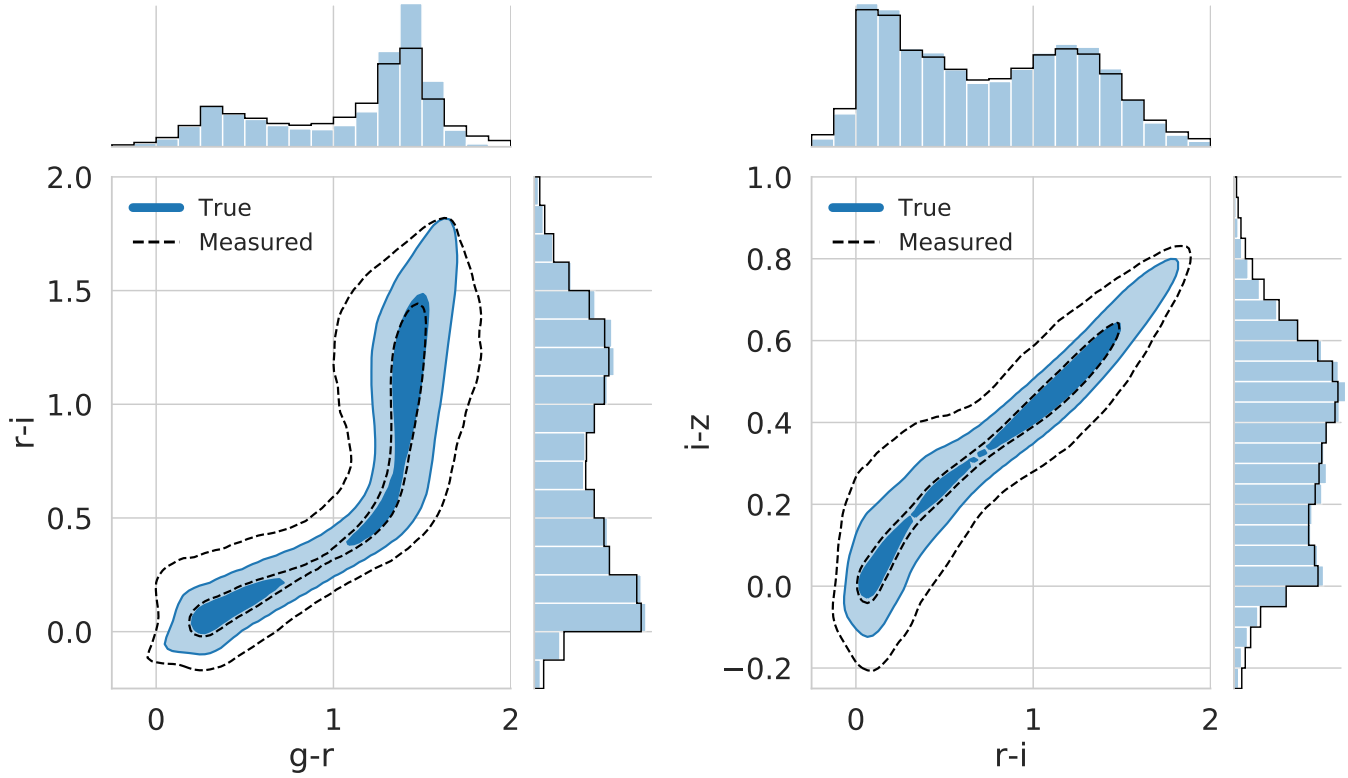


Figure 15. The $g-r$ vs. $r-i$ and $r-i$ vs. $i-z$ color-color distributions for the input colors in blue and measured colors in black. The density contour lines correspond to the percentiles of the first two sigmas of a 2D Gaussian, containing 39.2% and 86.5% of the total data volume respectively. The marginal distributions are included for comparison.

tics are computed in the same way as the magnitudes, though with a bin size of 100 mmag for $g-r$ and $r-i$ and 50 mmag for $i-z$. The color calibration for this sample is excellent. For the three colors examined here, the median color difference $\widetilde{\Delta c}_\delta$ is never greater than 5 mmag (0.5%) from injected color of -0.25 to 1.25 and is most commonly less than 3 mmag (0.3%). Beyond 1.25, $\widetilde{\Delta c}_\delta$ grows to a maximum of 25 mmag (2.3%) too blue for $g-r$ while for $r-i$ it never exceeds an absolute difference of over 3 mmag. The mean responses vary significantly due to extremely long scatter tails in both directions from the magnitude difference and are less reliable estimators of the overall performance in this case. However, they tend to be within a factor of two of the medians except for $g-r$ which increases in absolute size dramatically after 0.75 due to the long tail as can be seen in the figure. The full set of summary statistics are shown in Table C.1. Notably we do not find evidence of a systematic chromatic response in CModel color.

Next we compare the color-color diagrams for $g-r$ vs $r-i$ and $r-i$ vs $i-z$ for the input and recovered samples in Figure 15. As expected, the recovered injected colors have broader distributions due to the inherited WF noise as well as moderately large magnitude scatter near the detection threshold. However, the broadening is concentrated outside of the 1-σ contours where the agreement is extremely similar.

4.3. *Photometric Performance of y3-merged*

Unlike the synthetic δ star sample, `y3-merged` objects are sampled from fits to real sources contained in the DES DF. Thus not only are the properties of these injections far more diverse, but we do not have perfect knowledge of their true classification. However, we anticipate that most uses of this `Balrog` sample will be to calibrate galaxy samples used in cosmology analyses. In these cases, we do not care about the true classification as we want to capture the same contamination fraction as the data. For this reason we apply the cut `EXTENDED_CLASS_SOF > 1` and leave questions of star contamination to Section 4.4. Removing ambiguous matches with the cut `match_flag_1.5_asec < 2` decreased the sample by just under 1.5%.

There are numerous photometries and parameters whose response can be explored with this sample. We restrict ourselves largely to SOF CModel colors, magnitudes, and sizes here for brevity but find similar results for Metacalibration. As these quantities are important for the photometric redshift calibration modeling dis-

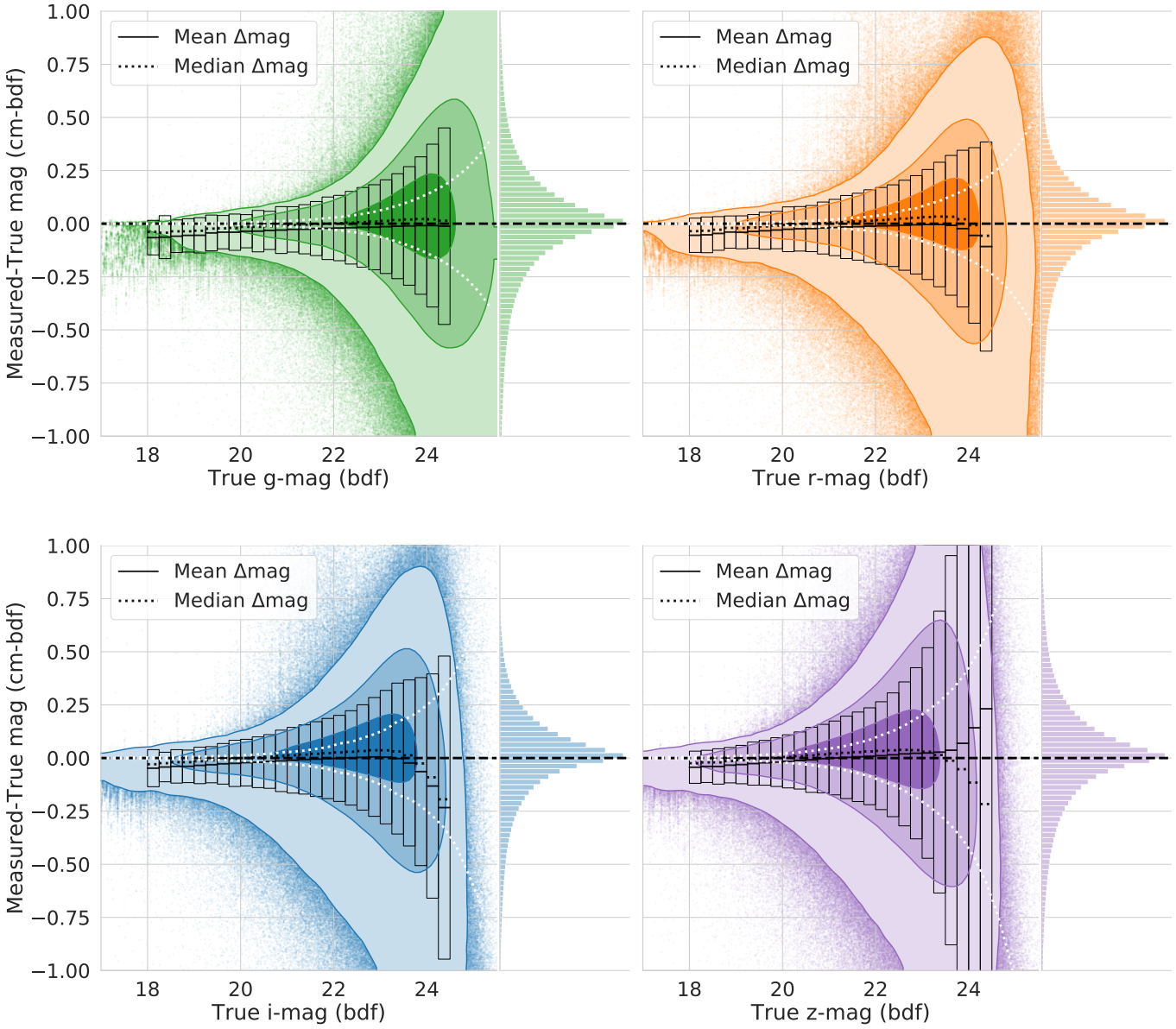


Figure 16. The distribution of differences in recovered *griz* SOF CModel magnitude vs the injected DF magnitude ($\Delta\text{mag}_{\text{DF}}$) as a function of input magnitude for the `y3-merged` sample. The density is overplotted where the contour lines correspond to the percentiles of the first three sigmas of a 2D Gaussian, containing 39.2%, 86.5%, and 98.9% of the data volume respectively. The mean (solid), median (dotted), and standard deviation of the magnitude responses in bins of size 0.25 magnitude are shown in the overlaid black bars. These are compared to the reported SOF CModel errors by the dashed white lines which do not attempt to account for systematic effects. The marginal distributions of $\Delta\text{mag}_{\delta}$ are included to highlight the small relative volume of the outlier tails.

cussed in Section 5.1, particularly the colors, we include summary statistics of the tabular results along with the SOF values in Appendix C.

4.3.1. *SOF CModel Magnitudes*

We compare the difference in recovered SOF CModel magnitude vs. true DF magnitude $\Delta\text{mag}_{\text{DF}}$ as a function of input magnitude for *griz* bands in Figure 16. As with `y3-stars`, we characterize the photometric performance of `y3-merged` measured galaxies with the summary statistics $<\Delta\text{mag}_{\text{DF}}>$, $\widetilde{\Delta\text{mag}}_{\text{DF}}$, and $\sigma_{\text{mag}_{\text{DF}}}$ in bins of truth magnitude overplotted in black bars. Unsurprisingly, the overall scatter in magnitude response for this sample is significantly larger than for the pure stellar injections due to the rich variety of injected morphologies and issues with blending of extended sources. The measured $\sigma_{\text{mag}_{\text{DF}}}$'s reflect this by being an average of over 4 times larger than the corresponding $\sigma_{\text{mag}_{\delta}}$ distribution over the same magnitude range, with the

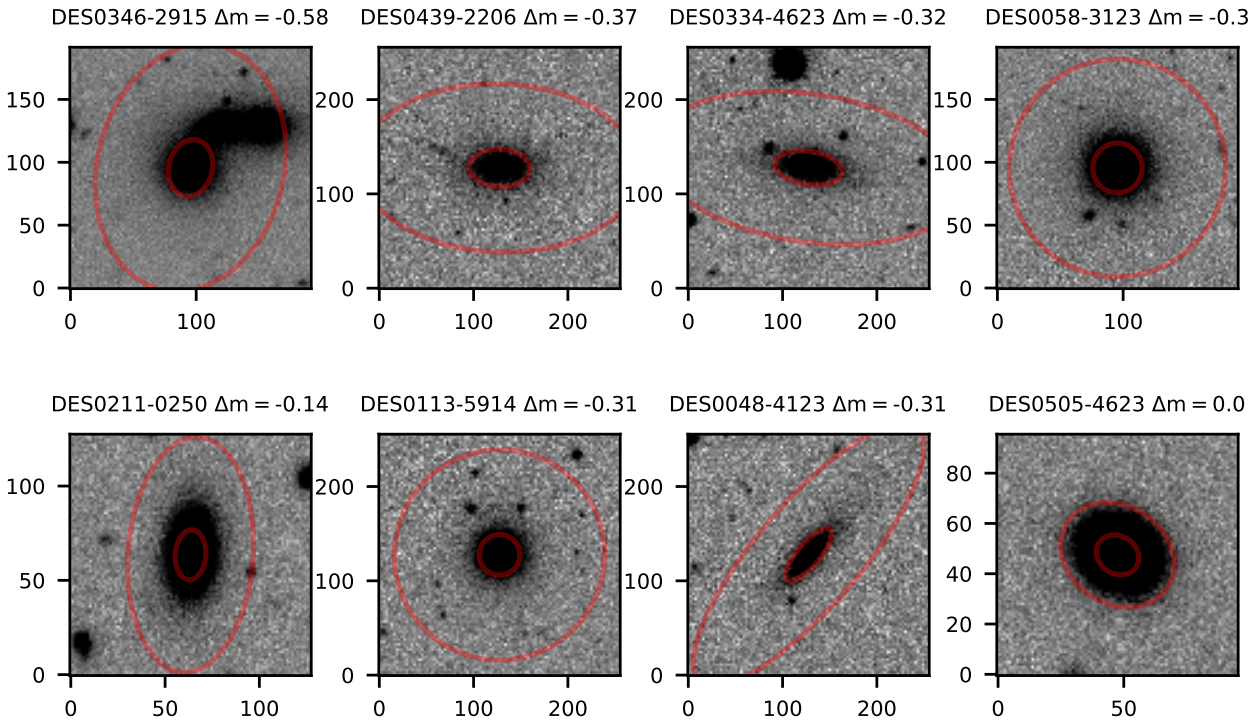


Figure 17. A few examples of injections that contribute to the long scatter tail in magnitude response of bright `y3-merged` objects due to blending of extended DF injections discussed in §4.3.1. Each injection had a true g-band magnitude between 17 and 19, and we include the tilename and magnitude response Δm at the top of each panel. The red lines correspond to the 50th and 95th percentile flux contours of the *measured* profile. The extended profiles of these injections cause the MEDS image cutout size (based on the fitted `SExtractor FLUX_RADIUS` value) to be relatively large which increases the probability of including real neighbors in the MEDS stamp. This in turn can cause SOF to significantly overestimate the `cm_T` size which leads to a much larger Δm than one would naively expect for objects with these bright magnitudes. This is discussed further in §4.3.3. The final panel shows a typical bright but compact object that is very well calibrated for comparison. Note the presence of a nearby source in the bottom that could have potentially caused the same failure mode if the box size had been slightly larger. The stretch in each panel runs from $-3\sigma_{\rm sky}$to$+10\sigma_{\rm sky}$.

ratio reaching as high as 9 for very bright objects. We then expect the mean response bias $<\Delta{\rm mag}_{\rm DF}>$ to be larger as well, but their behaviour is more interesting than the stellar sample. On the bright end below 19th magnitude, the 50th-99th percentile of objects are detected within 30 mmag (or 2.7%) of truth but there is a clear asymmetric preference for the recovered flux to be too large for the remaining objects. This result is driven by a sizeable fraction of bright, extended injections that are commonly blended with existing Y3 GOLD galaxies and are subsequently measured to have far too large of a size. The measured fluxes of these objects vary significantly depending on local conditions and create visible vertical lines in the response scatter due to their many injection realizations and relatively small population of objects with true magnitude less than 19. Image cutouts for a set of these objects along with the 50th and 95th percentiles of their measured CModel flux profiles are shown in Figure 17 – along with a more compact, typical injection at the same input magnitude that does not suffer from proximity effects or blending. These examples of large magnitude responses correlated with measured size errors are the first hint of a systematic issue with SOF fits in crowded fields that we investigate in more detail in §4.3.3.

As in the `y3-stars` sample, we detect a relatively small but clear band dependence in the mean and median responses. For all input magnitude bins brighter than 23 where the sample is nearly complete, there is a monotonic increase in the mean and median response in $griz$ with absolute spread of ∼16 mmag, or about 1.4% difference between g and z. This effect was hinted at in the response of the pure stellar sample but is far more evident here. This chromatic response is diluted but not eliminated when binning in measured S/N rather than input magnitude, with $\widetilde{\Delta{\rm mag}}_{\rm DF}$ no longer strictly monotonic and with a typical spread of 4-5 mmag for

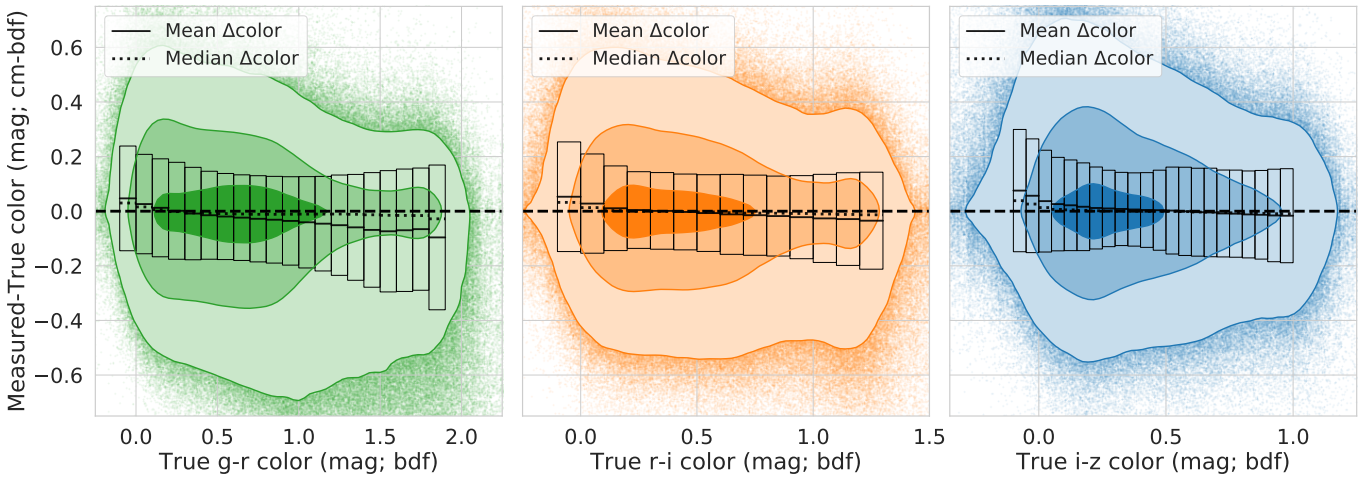


Figure 18. The distribution of differences in measured SOF CModel $g-r$, $r-i$, and $i-z$ color vs. the injected DF color (Δc_{DF}) as a function of input color for the `y3-merged` sample. The density is overplotted where the contour lines correspond to the percentiles of the first three sigmas of a 2D Gaussian, containing 39.2%, 86.5%, and 98.9% of the data volume respectively. The mean (solid), median (dotted), and standard deviation of the magnitude responses in bins of size 100 mmag magnitude for $g-r$ and $r-i$ and 50 mmag for $i-z$ are shown in the overlaid black bars.

riz-bands but 10-20 mmag when including *g*-band for S/N greater than 20.

We believe this chromatic effect is due to a systematic overestimation of the true sky background level in DES (and thus `Balrog`-injected) images. The `SExtractor` sky mode estimator is somewhat susceptible to the presence of neighboring objects in its sky annulus, especially in moderately to highly crowded fields. A mode estimate for the background appropriately allows for the fact that there will be background sources, detections, and undetected sources which is particularly important in the presence of many sources (Stetson 1987). As a precise mode estimation was once computationally impractical, traditional codes such as `SExtractor` have in practice used a Pearson-style mode estimator $\text{Mode}_{\text{est}} = 2.5 \cdot \text{Median} - 1.5 \cdot \text{Mean}$ for background estimation. This can result in a slight bias in overestimating the background which becomes larger as the field becomes more crowded and in the neighborhood of bright stars with extended wings (E. Bertin, private communication). This sky overestimation results in too faint a measurement of a galaxy's true magnitude and the effect is stronger when there is more sky noise per object signal. The fact that the sky is more crowded as one moves from bluer (g,r) to redder (i,z) bands could lead to the chromatic effect described above. That the scale of this effect is lessened by binning objects of similar S/N across bands together supports this conclusion. Note that these offsets are computed with dereddened magnitudes, which has the effect of enhancing the chromatic offset in *g*-band compared to the redder bands. Additionally, Eckert et al. (2020) analyzed the noise properties of DES images and found that there was a slight positive bias induced in the sky noise level due to faint unresolved sources in the field of essentially all images (see Section 5.3 for more details). The sign of this effect, while smaller, has the same trend and was found to only be significant for *riz* bands. We plan to investigate this further for the Y6 `Balrog` analysis and potentially propose additional magnitude corrections to account for this effect.

4.3.2. *SOF CModel Colors*

Next we investigate the color response of `y3-merged` objects in Figure 18, where we plot the difference in measured SOF CModel $g-r$, $r-i$, and $i-z$ colors vs the injected DF colors Δc_{DF} against the input colors. The density contours and overplotted summary statistics are defined in the same way as the previous plots. While the color response scatter is significantly larger than in `y3-stars`, the overall calibration is still excellent and with less extreme outlier tails than in the individual magnitude responses. The behaviour of the summary statistics is slightly more complex but we find that the median color response $\widetilde{\Delta c_{DF}}$ is typically ∼3 mmag (0.3%) too faint from -0.25 to 0 and ∼1-11 mmag too bright between 0 and 1.0 for all three colors. The responses are much noisier outside of these regions due to much smaller sample sizes. $\widetilde{\Delta c_{DF}}$ tends to be ∼15-25 mmag (1.4-2.2%) too faint below 0.25 and 15-25 mmag too bright beyond 1.0 for all colors (though a bit worse for $r-i$, reaching 12% too bright near 1.5) while $<\Delta c_{DF}>$ differences are about three times as large as $\widetilde{\Delta c_{DF}}$ in the same direction depending on the color and bin. As with the stellar injections, individual $<\Delta c_{DF}>$

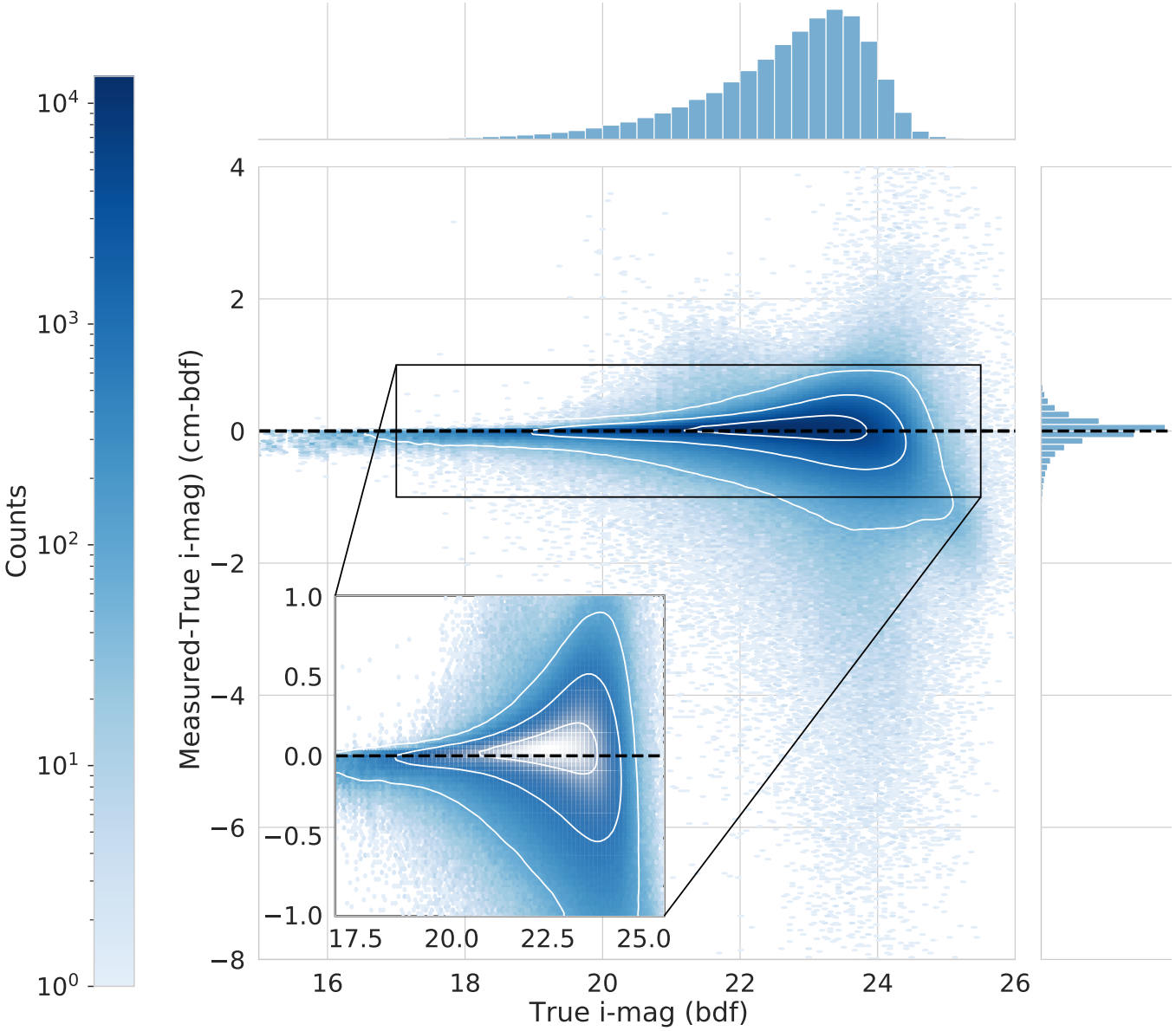


Figure 19. The distribution of differences in recovered i-band SOF CModel magnitude vs the injected DF magnitude ($\Delta\text{mag}_{\text{DF}}$) as a function of input magnitude. The inset corresponds to the i-band panel in Figure 16 where the density contours still contain 39.2%, 86.5%, and 98.9% of the data volume respectively. While most of the density is captured in the inset, it misses many of the rich features of the full magnitude response – particularly the long outlier tail of injections measured to have magnitudes up to 10 greater than truth. We explore some of the causes of this in §4.3.3.

and $\widetilde{\Delta c}_{\text{DF}}$ bin values can vary significantly due to long scatter tails and we find no evidence of a systematic chromatic response in CModel color. The full color response is summarized in Table C.4.

4.3.3. *Catastrophic Model Fitting*

While Figure 16 shows that the vast majority of magnitude responses are well calibrated and are typically much less than $\Delta\text{mag}_{\text{DF}}$ of 0.5, it ignores the very long tail of up-scattered outliers that are far larger than the measured photometric errors would predict. The responses of these outliers from blends and catastrophic photometry failures can be over an order of magnitude larger than those previously discussed as shown for i-band in Figure 19 where the contours from Figure 16 are overlaid in white.

Here the true complexity of even a small slice of the transfer function is revealed: The many competing effects are often in opposition, with biases in the opposite direction of long, asymmetric tails that vary as a function of truth magnitude in a complex way. Simple Gaussian summary statistics like $<\Delta\text{mag}_{\text{DF}}>$ and

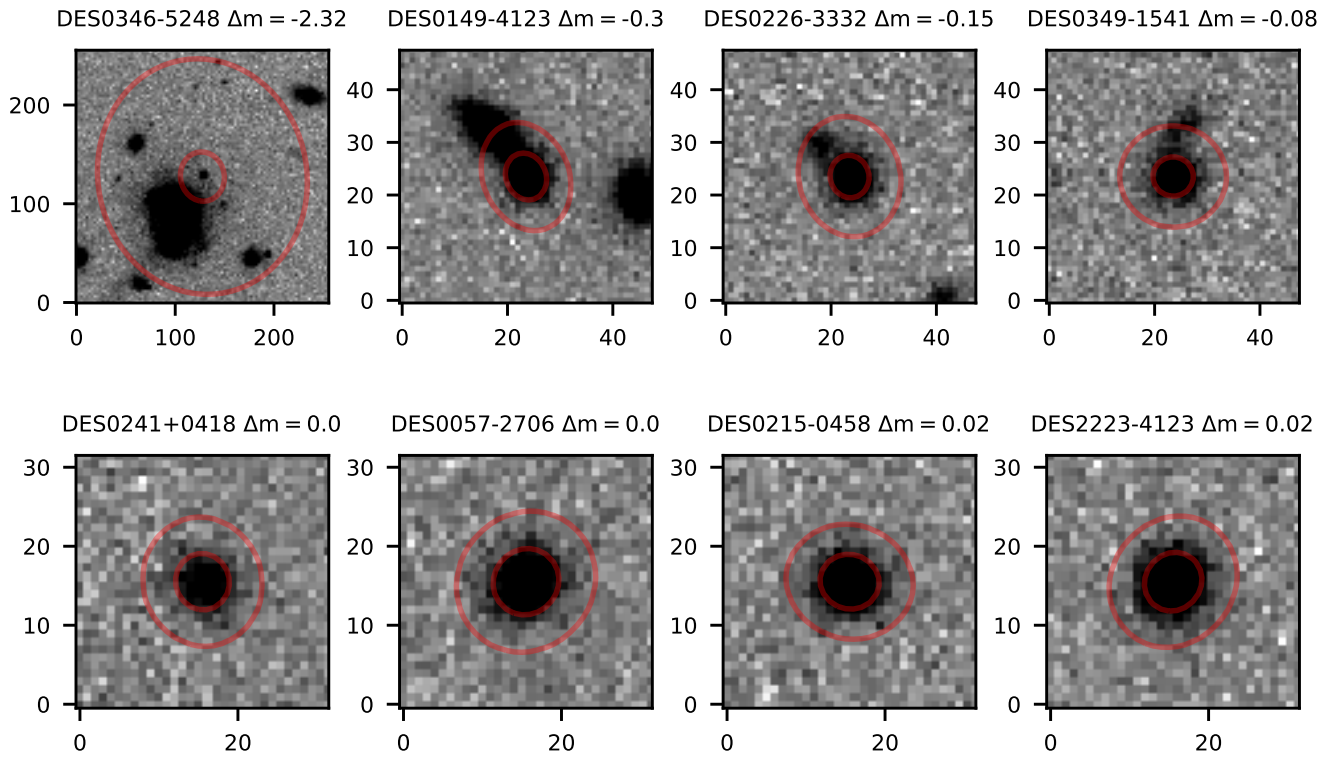


Figure 20. The MEDS image cutouts for a few injection realizations of the same DF object with true r-magnitude of 21.42 in eight distinct WF tiles (`bal_id` of 10034605248852. The red contours give the 50% and 95% enclosed light apertures for the injected object as modeled in each tile. The difference between the measured and injected magnitude Δm is listed next to each tile name, with the cutouts ordered by the magnitude response. The box sizes are in 0.263″ pixels. Not all cutouts are the same size, as the box size expands based on the initial `SExtractor FLUX_RADIUS` measurement. The true scale length of the object (after PSF deconvolution) is 0.77″. The fitted profile for the object on tile DES0149-4123 is 1.0″ and while that on tile DES0346-5248 is an unrealistic 17″, leading to an overestimate of the object flux corresponding to an error of 2.32 magnitudes). The stretch in each panel runs from $-3\sigma_{\rm sky}$to $+\,10\sigma_{\rm sky}$

$\sigma_{\rm mag_{DF}}$ are not able to appropriately capture the magnitude of these features and we argue that the `Balrog` samples themselves (or at least higher fidelity forms of data compression) should be used for most cosmological analyses that need accurate photometric error modeling. Examples of how the full richness of the transfer function can be used in photometric redshift calibration and the magnification of lens samples are given in Sections 5.1 and 5.2 respectively.

However, it is reasonable to be skeptical of magnitude responses of $\Delta{\rm mag_{DF}}\sim$2-8 (a factor of 6-1,600 in flux!) by supposedly well-calibrated photometry pipelines. To demonstrate what is causing these extremely large differences in recovered flux, we show in Fig. 20 a set of injections of the same DF object with r-band magnitude of 21.42 into eight different WF tiles where the red lines correspond to the 50th and 95th percentile flux contours. In most cases the true magnitude is recovered within the reported errors of a few percent. However, in four instances there is at least one nearby object contained in the MEDS cutout image that interferes with SOF's ability to provide a reliable fit due to either an excess of masked pixels in the cutout or residual light unassociated with the injection. The result is a fitted characteristic size `cm_T` which is much greater than its actual size. For this particular injection, the true size of the object (after deconvolution with the PSF) corresponds to a scale length of 0.77″. Yet in the four cases with nearby sources the fitted size of the object is at least 1″, resulting in a flux measurement which is significantly greater than that of the input true flux. In the worst case for tile DES0346-5248, the target object is by chance injected near a very bright pair of merging galaxies and is fitted with a scale length of over 17″ resulting flux 2.32 magnitudes brighter than the input DF value.

These photometric measurement failures correlated with errors in measured `cm_T` can be even more dramatic.

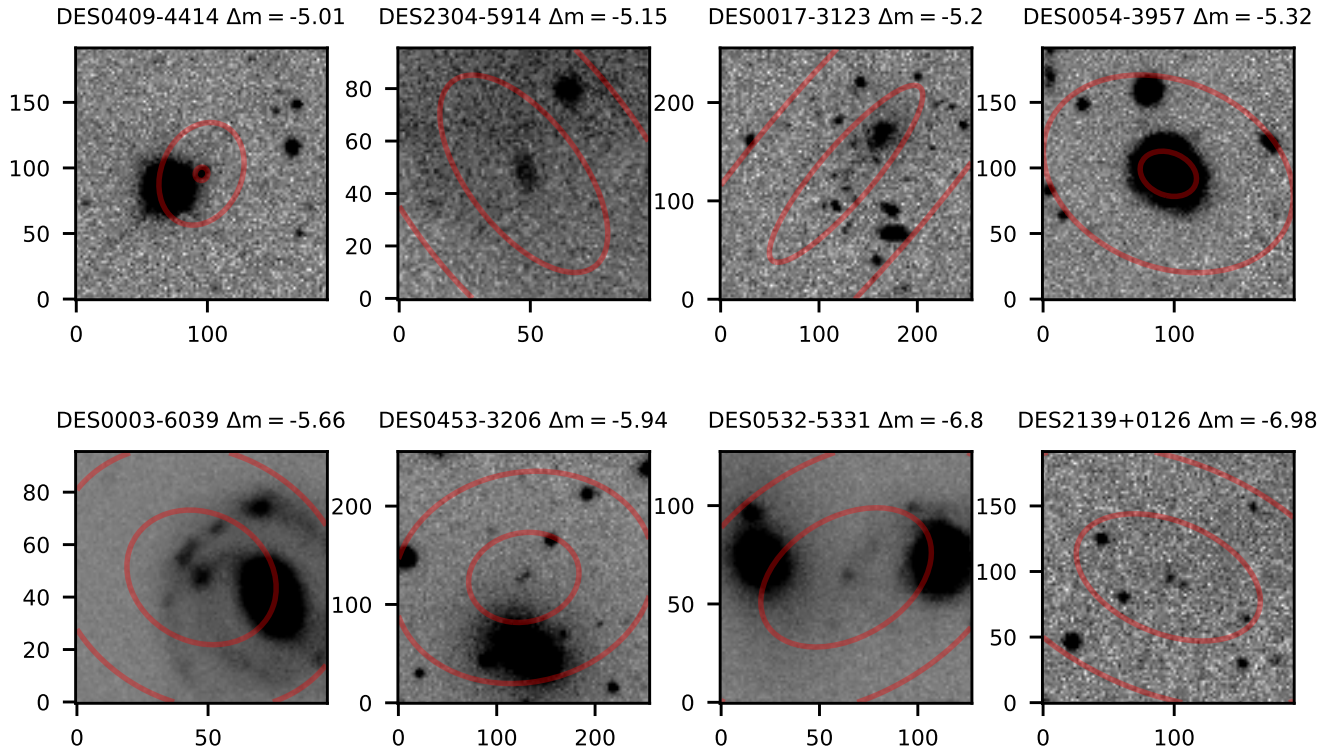


Figure 21. The MEDS image cutouts for eight `Balrog` objects with extremely large differences between the measured and injected magnitude Δm. The red lines correspond to the 50th and 95th percentile flux contours of the measured profile. These injections happened to be placed in regions of rapidly varying sky brightness, in the spiral arm of a large spiral galaxy, in a rich cluster, near a stellar diffraction spike, in between two extended galaxies, or simply in crowded fields. In all cases the fitted size is far too large for the source, which in turn leads to an overestimate of the object's flux. This processed is discussed in detail in §4.3.3. The stretch in each panel runs from $-3\sigma_{\rm sky}$to $+10\sigma_{\rm sky}$.

In Figure 21 we show eight examples of catastrophic fitting failures due to crowded fields, nearby bright stars, and unflagged image artifacts. These rare but real environments lead to `Balrog` magnitude responses from 5 to even 7 magnitudes brighter than the injected truth. We emphasize that all of these objects pass the basic Y3 GOLD science catalog quality cuts described in the beginning of Section 4.

While the exact causal relationship between complex local environments and extreme magnitude errors requires further analysis, preliminary investigations suggest the following: In crowded fields or areas with unusual image features or artifacts, the `SExtractor` `FLUX_RADIUS` (which defines a circle that contains half of the total corresponding `FLUX_AUTO` value) can get artificially inflated in size as compared to what it would return for an object in an isolated environment. As a source's MEDS cutout image size is rounded up to the next integer multiple of 16, this leads to a MEDS stamp that is significantly larger than what is needed to fit the relevant flux profile in question. This leaves large areas of the stamp with masked pixels when fit with SOF as the algorithm masks rather than models the light of other detected sources within the cutout. The resulting CModel fits then preferentially overestimate `cm_T` for this subpopulation which can greatly increase the inferred flux for a given surface brightness measurement - though we defer investigations into the exact details of the scale and frequency of this effect for a future analysis.

Even without a complete understanding of the underlying cause, the correlation between $\Delta{\rm mag}_{\rm DF}$ and Δ`T` is evident as can be seen in Figure 22. Here we have plotted the full i-band magnitude response of `y3-merged` but colored individual responses by the absolute difference in measured `cm_T` vs. input `bdf_T`. The vast majority of injections with truth i-magnitude below 23 with very small $\Delta{\rm mag}_{\rm DF}$ responses have `T` differences much less than 1 which are colored blue. Bright objects with responses substantially below the zero line have moderately large errors in recovered `T` as we discussed in §4.3.1, while fainter injections with enormous magnitude errors have correspondingly large errors in `T` – reaching as high as the parameter prior limit of 10^6 arcsec^2 (or

scale length of $\sim 10^3$ arcsec). The situation is more complicated near and past the detection threshold, about 23rd magnitude in i-band, where additional systematic effects become important.

Model fitting photometry codes are complex, nonlinear, and sometimes non-local algorithms that can have unexpected consequences – particularly for low S/N measurements, crowded fields, or when image artifacts are not appropriately weighted or masked. The journey from pixels to catalogs can at times be chaotic, and our modeling of photometric uncertainties should reflect this.

4.3.4. *Scatter from Ambiguous Matches*

Despite the efforts described in Section 3.5 there will always be some ambiguity in the matching to injected sources that can introduce large, non-physical scatter. To check this, we visually inspected hundreds of the MEDS stamps of `Balrog` objects whose absolute magnitude response was greater than 2 – and in particular the set of objects with large $\Delta \mathrm{mag}_{\mathrm{DF}}$ whose size errors were small. There were a few isolated instances of ambiguous matches where a faint injection landed in the very center of an extremely bright Y3 star whose GAp flux measurement failed. These can easily be accounted for by adapting our ambiguous matching algorithm to reject `Balrog` injections near objects with flagged GAp fluxes but this was not discovered in time to update the catalogs used in downstream measurements. However, this issue has negligible impact as we estimate only a few hundred instances in the total `y3-merged` sample.

4.4. *Star-Galaxy Separation*

We use the δ injections of `y3-stars` to estimate the stellar efficiency (or true positive rate) in blue and the classified DF sources in `y3-merged` for the contamination rate (or false discovery rate) in red for the `Balrog` star sample as a function of injection magnitude in Figure 23a. The solid, dashed, and dotted lines represent the fraction of objects classified as less than or equal to an `EXTENDED_CLASS_SOF` value of 0, 1, or 2 respectively. While `y3-merged` is required to estimate the contamination rate in order to have a realistic relative ratio between star and galaxy counts, we use the `y3-stars` sample to compute the efficiency as its truth classifications are nearly noiseless and the measurement does not need any external information about galaxy contaminants. We find that the stars are correctly classified (`EXTENDED_CLASS_SOF` <= 1) over 95% of the time below an i-band magnitude of 21.75 and 80% of the time below magnitude 22.75 before dipping to 70% efficiency near the detection threshold at $i \sim 23$. The stellar efficiency quickly drops to below 50% beyond

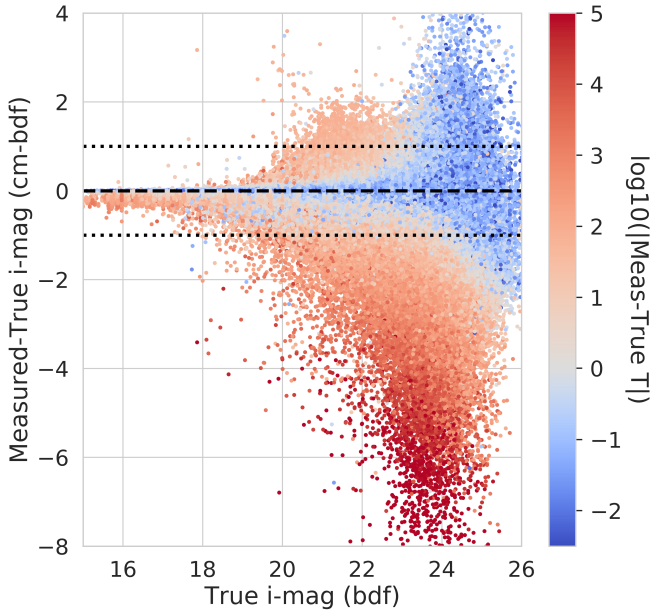
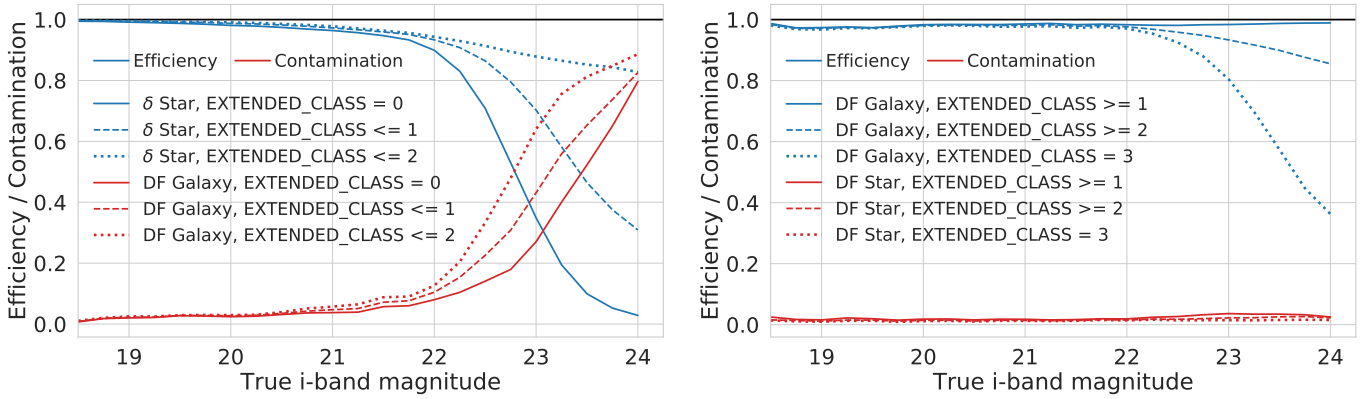


Figure 22. The full i-band magnitude response $\Delta \mathrm{mag}_{\mathrm{DF}}$ for `y3-merged` shown in Figure 19 but now colored by the logarithmic absolute error in recovered size parameter `cm_T` vs input size `bdf_T`. The response scatter is largely correlated by error in recovered size; injections with small $\Delta \mathrm{mag}_{\mathrm{DF}}$ values typically have small errors in recovered `T` as well (in blue), while nearly all of the extreme magnitude outliers have correspondingly large size errors. The correlation is less strong past the detection threshold at $i \sim 23$ where other systematic effects increase in importance.

23rd magnitude. The efficiency of high confidence stars (`EXTENDED_CLASS_SOF` == 0) follows a similar trend but reaches the previously quoted values about 0.5 magnitudes earlier. Alternatively, the rate of DF galaxies misclassified as stars stays below 10% until 22nd magnitude where there is a sharp increase until the detection limit where at low S/N it is extremely difficult to differentiate between classifications. However, we again note that the stellar efficiency measurement is less noisy due to the higher degree of confidence in accurate classification compared to the DF sample.

We make equivalent measurements for the galaxy efficiency and contamination in Figure 23b where the solid, dashed, and dotted lines now correspond to the fraction of objects classified as greater than or equal to `EXTENDED_CLASS_SOF` values of 1, 2, and 3. Here we must use sources in `y3-merged` exclusively as the ratio between stars in the δ sample and galaxies in the DF sample is not realistic as required by a contamination estimate. The efficiency is slightly lower than the stars on the bright end due to impurities in the DF knn classifier but is quite close to 100% below 22nd magnitude. The efficiency of high-confidence



(a) Star Selection

(b) Galaxy Selection

Figure 23. The efficiency (in blue) and contamination (in red) of the `Balrog` stellar sample (a) and galaxy sample (b). We use the δ injections of `y3-stars` as our population of true stars for (a) as it is a nearly pure sample, with only ambiguous matches as potential contaminates. We use the DF injections classified as galaxies from the DF k-nearest neighbor (knn) classifier described in Section 3.3 as our true galaxy sample which has intrinsic uncertainty as detailed in Hartley, Choi et al. (2020). For (b), we cannot use the δ injections as the contamination measurement requires a realistic ratio of galaxy and stars sources in the sample so we instead use the classified DF stars. Each line corresponds to the fraction of objects above or below the noted `EXTENDED_CLASS_SOF` threshold value. We do not expect the galaxy efficiency to be 100% even at magnitudes where complete due to small impurities DF knn classifier.

galaxies (`EXTENDED_CLASS_SOF` == 3) decreases sharply near the detection limit, but over 85% of DF galaxies with assigned classifications are correctly identified (`EXTENDED_CLASS_SOF` >= 2) down to 24th magnitude in i-band. The contamination rate of stars into the galaxy sample is consistently ∼2% until 22nd magnitude where it rises slightly to 4% at a magnitude of 23. This low level of contamination is largely due to the relatively small number of stars compared to galaxies at these magnitudes and is consistent with the findings quoted in Sevilla-Noarbe et al. (2020). A table of the `Balrog` classification (or "confusion") matrix as a function of input magnitude is provided in Table C.5.

5. APPLICATIONS TO DES Y3 PROJECTS

Below we present some of the most important applications of the Y3 `Balrog` catalogs, particularly those that are relevant for the DES Y3 cosmology analysis. To our knowledge, this is the first time an object injection pipeline has been used for any of the following measurements or played such a critical role in the calibration of a galaxy survey's cosmological constraints.

5.1. *Photometric Redshift Calibration*

Chief among the applications of our results is facilitating a novel inference method for the photometric redshift calibration of weak lensing samples. As shown in Buchs, Davis et al. (2019), we can extract information from the DES Y3 DF to break degeneracies in the riz[16] color-redshift relation if we have accurate estimates of the corresponding WF properties of the DF sources. In this inference method, `Balrog` plays the essential role of determining the likelihood of a given deep, many-band color to be observed at a given region of noisier color-magnitude space in DES measurements at Y3 depth. This allows us to rigorously separate the contributions from measurement noise to the true color-redshift relation when estimating the ensemble photometric redshift distribution of the lensing source sample. In practice, this inference method is facilitated by the use of two Self-Organizing Maps (SOM) which classify the galaxies in the deep and wide samples into discrete classes, called *cells*, of color and color-magnitude space. The redshift distribution of a given Y3 source is then given by

$$p(z,\hat{c},\hat{s},\theta) = \sum_c p(z|c)p_{\texttt{Balrog}}(c|\hat{c},\hat{s},\theta)p(\hat{c}|\hat{s},\theta) \qquad (3)$$

where z is redshift, c is deep SOM cell, $\hat{c}$ is wide SOM cell, $\hat{s}$ is the sample selection function, and θ is any additional conditions such as position on the sky. The middle factor $p_{\texttt{Balrog}}(c|\hat{c},\hat{s},\theta)$, a narrow slice of the full `Balrog` transfer function, expresses the likelihood of a deep color to be observed at a certain region of wide color-magnitude space. This transfer function serves to

[16] Only the riz Metacalibration fluxes are used when defining the tomographic bins.

correctly weight the well-constrained redshift distribution $p(z|c)$ of each deep SOM cell according to the probability of detecting those galaxies. As the SOM cells $\hat{c}$ are determined by Metacalibration magnitude and color, `Balrog` plays the key role of determining a distribution of observed Metacalibration magnitudes for each injected DF galaxy.

In addition to breaking degeneracies in the color-redshift relation, `Balrog`, by virtue of enabling this scheme, facilitates avoiding otherwise prohibitive selection biases resulting from the use of spectroscopic redshifts for weak lensing redshift calibrations (see, e.g. Gruen & Brimioulle 2017) because it uses spectroscopic redshifts only of galaxies for which 8 bands of DES DF photometry provide relatively well-constrained $p(z)$.

In the first application of this inference scheme to data, Myles, Alarcon et al. (2020) found that the intrinsic uncertainty in `Balrog`'s estimation of the transfer function is a negligible contributor to the overall error budget with an uncertainty on the mean redshift in each tomographic bin of $\sigma_{\bar{z}} < 10^{-3}$. This is a significant accomplishment as `Balrog` was able to decrease the systematic bias in the photometric redshift estimates without contributing a novel source of intrinsic systematic uncertainty in its sampling of the transfer function, which was not obviously the case a priori. The use of `Balrog` in photometric calibration can be further leveraged in future analyses by incorporating positional-dependent selection effects θ in the used measurement likelihood $p_{\texttt{Balrog}}(c|\hat{c}, \hat{s}, \theta)$. For further details on this method, we refer the reader to Myles, Alarcon et al. (2020).

5.2. *Magnification Bias on Clustering Samples*

Lens magnification is correlated with large-scale structure and should be taken into account in the modeling of galaxy clustering and galaxy-galaxy lensing correlation functions (Unruh et al. 2020). Magnification modifies the observed galaxy over-density $\delta_g^{\rm obs}$ through two competing effects; a geometric suppression factor $C_{\rm area}$ and a boost in detection efficiency of faint sources which increases the local number density $C_{\rm sample}$:

$$\delta_g^{\rm obs} \approx \delta_g^{\rm int} + [C_{\rm sample} + C_{\rm area}] \cdot \delta\kappa \tag{4}$$

where $\delta_g^{\rm int}$ is the intrinsic galaxy over-density before magnification by δ_κ is considered.

While a simple argument in Elvin-Poole et al. (2020) shows $C_{\rm area}$ to be -2, the contribution by $C_{\rm sample}$ for even a simple linear response to δ_κ depends on the ratio of intrinsic number density n_{int} with and without magnification as a function of measured object fluxes F,

$$C_{\rm sample}\delta_\kappa = \frac{n_{\rm int}(F, \kappa + \delta\kappa)}{n_{\rm int}(F)} \tag{5}$$

which is difficult to model explicitly as they implicitly depend on detection and measurement systematics.

To aid in this effort, supplemental runs to `Run2` and `Run2a` (designated as `Run2-mag` and `Run2a-mag` respectively) were created where the same input objects were injected with identical simulation configuration except for an additional `GalSim magnify` call that was applied to all objects uniformly. Each object was given a lensing magnification δ_κ of 2%, effectively increasing the flux and area of objects by this amount. A given galaxy sample selection can be applied to both the magnified and unmagnified runs and Equation 5 can be used to estimate the magnification bias $C_{\rm sample}$. This estimate will include not only the impact of magnification on galaxy fluxes but any selection bias (e.g. on size) introduced by the photometry or imaging systematics.

Figure 24 shows the $C_{\rm sample}$ estimates from `Balrog` for samples with a constant i-band flux limit and a simple galaxy section criteria of

```
        EXTENDED_CLASS_SOF = 3
AND     FLAGS_GOLD_SOF_ONLY & 126.
```

The same process is also applied to the real data where magnification is applied only to the galaxy fluxes. For this very simple selection the `Balrog` estimates are consistent with the data flux-only estimates, indicating any contribution from size selection or other systematics is small.

In Elvin-Poole et al. (2020) this `Balrog` methodology is applied to the lens samples used in the DES Y3 analysis including more complex color cuts and tomographic redshift binning. In this analysis, the MAGLIM lens sample (Porredon et al. 2020), which has a redshift dependent magnitude limit and tomographic binning, is found to have a $C_{\rm sample}$ from approximately 2 to 5 from low to high redshift. The redMaGiC lens sample (Rozo et al. 2016), which is a Luminous Red Galaxy (LRG) selection, has $C_{\rm sample}$ from values consistent with 0 to approximately 4 at high redshift. The `Balrog` estimates of $C_{\rm sample}$ are systematically lower than the flux-only estimates due to the additional selection effects captured by the full `Balrog` transfer function. See Elvin-Poole et al. (2020) for additional details.

5.3. *Noise from Undetected Sources*

It is important to accurately characterize image noise to get unbiased estimates of an object's photometric properties and image moments. While Poisson noise is dominant for calibrated images, there are other less-studied contributions to the image noise including undetected sources (US). Using the Bayesian Fourier Domain (BFD) method described in Bernstein & Arm-

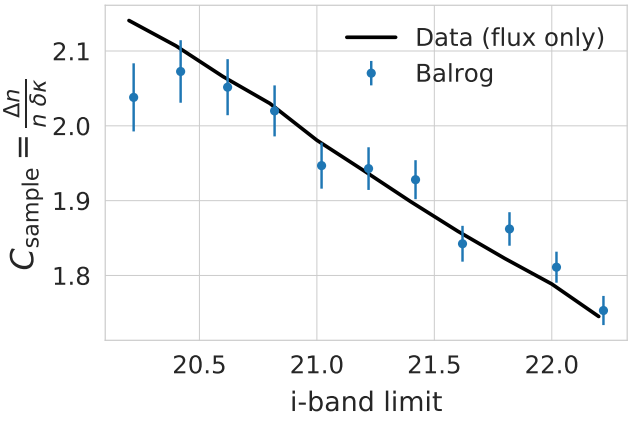


Figure 24. The magnification bias from boosted detection efficiency $C_{\rm sample}$ estimated on samples with a uniform i-band flux magnitude limit. The `Balrog` estimates in red use the magnified `Balrog` runs with a 2% magnification applied to every input object. The data estimate applies the same artificial magnification to the galaxy magnitudes in the data and reapplies the selection. The `Balrog` estimates of $C_{\rm sample}$ are consistent with the flux-only estimates for this very simple selection, indicating that the contributions from additional selection effects are small. However, the `Balrog` $C_{\rm sample}$ estimates are systematically lower than the simple data estimate for the real Y3 samples used in Elvin-Poole et al. (2020) where the selections are significantly more complicated.

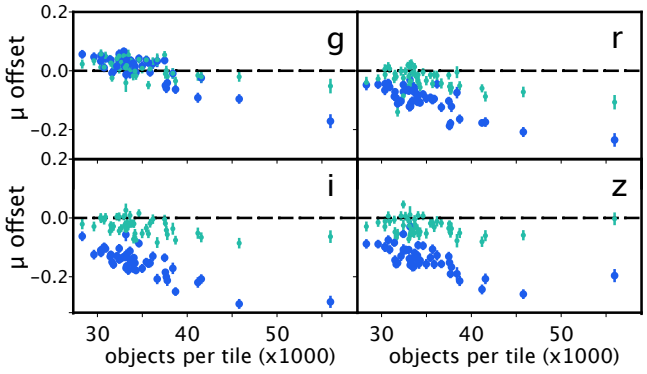


Figure 25. Reproduced from Eckert et al. (2020), Figure 3. The Gaussian mean offset μ in the BDF flux moment pull distribution as a function of object density for the 48 used `Balrog` tiles in blue. The green points show the mean offset for the tiles after a local sky subtraction which mitigates the flux bias. While g-band is relatively unaffected, the redder riz-bands show statistically significant sky oversubtraction that is correlated with object density.

strong (2014) on `Balrog` detections across 48 tiles, the variance of measured galaxy moments was found to be up to 30% in excess of Poisson predictions in Eckert et al. (2020). Furthermore, an over-subtraction of the background was detected in the riz-bands leading to a bias in the zeroth moment flux estimator as shown in Figure 25. The blue points show the mean μ of the Gaussian fit to the pull distribution of BDF flux moments for each tile as a function of object density where a clear correlation can be seen, particularly for the redder bands. The green points are the same measurements after making a local estimate of the background in each postage stamp.

In order to determine if the excess noise was due to US, a slight variant on the `Balrog` injection procedure was followed in which we injected zero-flux objects into 39 tiles at random positions and then made cutout postage stamps of these random patches of sky. The cross-power spectra of distinct exposures of the "dark" injections in $griz$ was then computed, which would yield zero signal if the noise is Poisson or read noise. A clear detection of US noise is made in each band. This empirical approach allows computed BFD moments to calibrate the moment covariance matrix on the survey images rather than relying on simulations of unknown fidelity, and naturally includes the contribution by US as a source of noise within the Bayesian calculation. See Eckert et al. (2020) for further details.

5.4. *Accurate Joint Redshift - Stellar Mass Probability Distributions with Random Forests*

In Mucesh et al. (2020), `Balrog` is used together with the Random Forest machine learning algorithm to produce well-calibrated joint redshift-stellar mass probability distributions at a fraction of the speed of traditional template-fitting methods. This was made possible because `Balrog` produces an ideal training sample: it captures both the realistic noise properties of DES WF measurements as well as the redshift and mass information of the DF injections from the the COSMOS2015 catalog (Laigle et al. 2016).

5.5. *Photometric Response Near Galaxy Clusters*

Clusters of galaxies – especially rich, crowded clusters – are known to present extra obstacles in the accurate detection and characterization of cluster members. These member galaxies often have higher detection incompleteness and significant photometric biases because of the increased rate of proximity effects. Detected sources in or near galaxy clusters in the sky can be further biased because of blending with member galaxies or contamination from intra-cluster light (Zhang et al. 2019). To aid in studies of these difficult measurement biases and selection effects, a high-density `Balrog` run was performed targeting areas near rich galaxy clusters.

A sample of 900 tiles, each containing a galaxy cluster with optical richness $\lambda > 35$ (see Rykoff et al. 2016 for a description of richness and the DES cluster catalog),

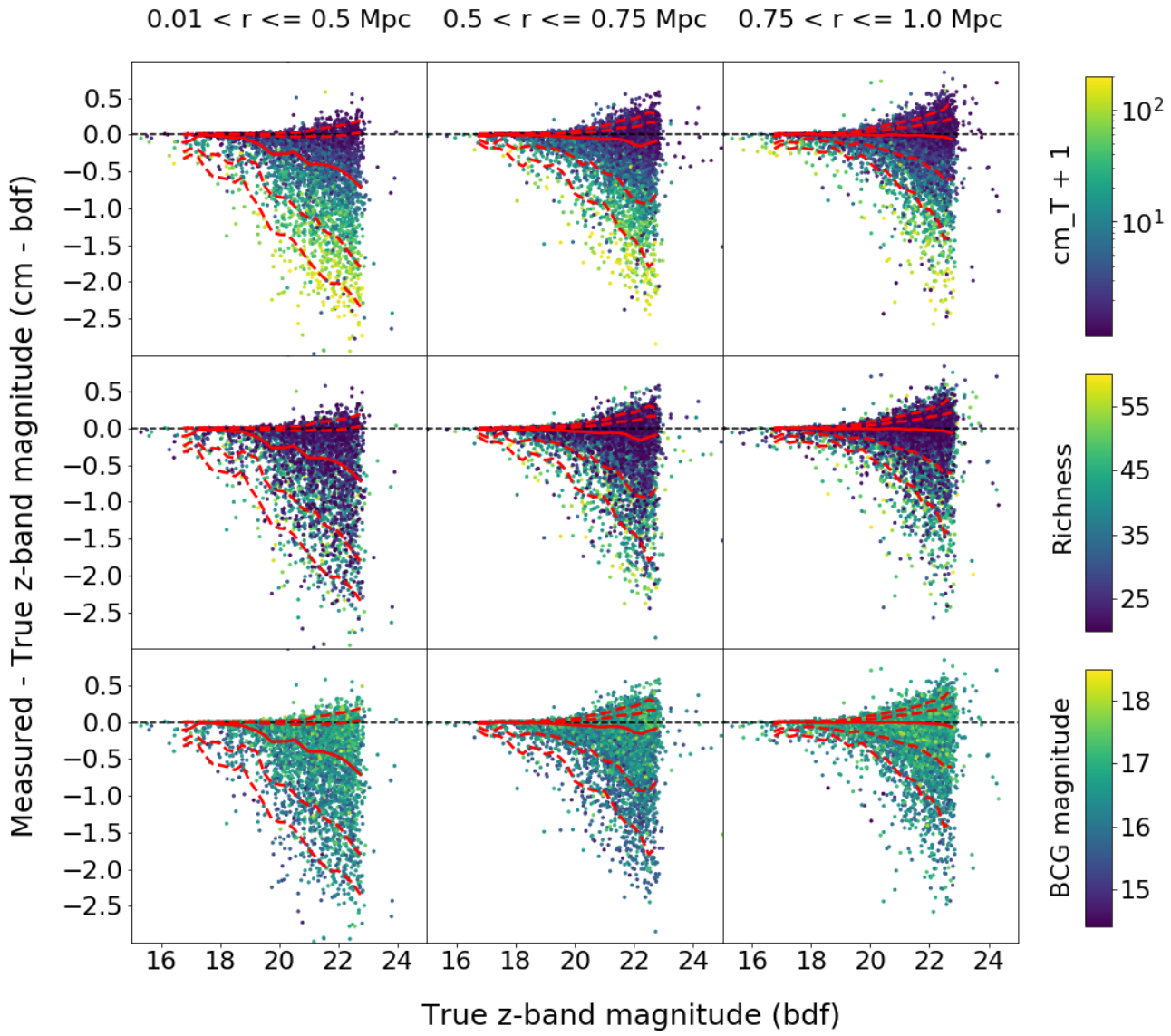


Figure 26. The difference in measured CModel z-band magnitude vs. the injected DF magnitude $\Delta\text{mag}_{\text{DF}}$ as a function of input magnitude for the high-density `clusters` run. The three columns present the $\Delta\text{mag}_{\text{DF}}$ responses binned by their radial distances to nearby cluster centers as specified at the top of the columns. The median response biases across the range of the injected magnitude are displayed as solid red lines, with the first and second $\sigma_{\text{mag}_{\text{DF}}}$ contours indicated by the dashed lines above and below. As the injections approach the center of a cluster, the median bias becomes increasingly negative indicating that the objects are measured to be progressively brighter than injected truth the closer they are to the bright central galaxy (BCG). In the three rows we color the magnitude responses as a function of (a) measured object size `cm_T`, (b) cluster richness, λ, and (c) the magnitude of the cluster's BCG. The measured object size appears to have the strongest influence over magnitude bias among the three quantities, though richer clusters also show larger $\Delta\text{mag}_{\text{DF}}$ responses. We use `cm_T`+1 for the color scale as the `ngmix T` sizes are allowed to be slightly negative. This preliminary analysis will be followed up in more detail in Masegian & Zhang et al. (in preparation).

were injected with a similar DF galaxy sample as used in `y3-merged` at a lattice separation of 10″ resulting in four times the injection density of the main cosmology runs. This higher injection density was needed to properly sample the effects of clusters on the transfer function as a function of radius from a cluster center given the number of tiles used[17].

Additionally, we used a more restrictive *riz* detection magnitude of 23 to increase the fraction of detected objects for this analysis. The magnitude responses of the injected galaxies were measured as well as their distances to the center of the nearby clusters. The sample was further subdivided by the host cluster's richness, the measured object size `cm_T`, and the magnitude of the cluster's brightest cluster galaxy (BCG) to see how these parameters affect the magnitude bias of the inserted objects. Preliminary results of this analysis are shown in Figure 26 and will be followed up in Masegian & Zhang et al. (in preparation) including a careful study of the detection efficiency as a function of various parameters including radial distance. Unsurprisingly, the magnitude responses become more negatively biased closer to cluster centers where the complex environments make accurate photometric measurements difficult and faint sources are up-scattered by the abundant residual light.

We find a similar correlation between an object's measured size and magnitude response as seen in §4.3.3. The proximity effects that cause asymmetric overestimates of `cm_T` are amplified in the very crowded cluster environments, a trend that grows even stronger closer to the cluster centers. Correlation between magnitude bias and the other examined parameters, cluster richness and the BCG magnitude, is weaker but still present – particularly for richness. All correlations appear to bias the recovered magnitudes in the same direction. The scale of these effects increases as the injections approach cluster centers. Taken together, the proximity to cluster centers, cluster richness, and BCG brightness artificially increases the number of observed objects near clusters above a fixed brightness threshold which, in turn, can collectively bias cluster measurements from a corresponding increase in cluster member galaxies. We plan on accounting for these correlations in future DES cluster analyses.

6. CURRENT METHODOLOGICAL LIMITATIONS AND FUTURE DIRECTIONS

While this latest iteration of `Balrog` has made great advances in its ability to precisely quantify difficult measurement systematics, there remain many challenges to overcome if we are to reach the level of precision required by upcoming Stage IV surveys like LSST where the increased depth, pipeline complexity, and blending rate will otherwise limit the constraining power on cosmological parameters. Some of these challenges, such as properly accounting for per-object chromatic corrections at injection time or pushing the injection step further upstream in the measurement pipeline to account for more systematic effects in the image calibration, are largely technical barriers that can be addressed with more development time. Our ambiguous matching scheme can be improved by incorporating pixel-level information on the overlap between injected and real sources similar to the blending parameter introduced in Huang et al. (2018). In addition, many of the complexities and additional development time needed for careful emulation of a survey's measurement pipeline can be nearly eliminated by having injection pipelines placed directly in the software stack of the fiducial data processing runs. While this was not possible in DES, this approach is now taken in HSC with `SynPipe` and planned for LSST. However, there are more fundamental barriers in leveraging injection pipelines to their full potential.

A primary challenge is increasing the representativeness of the input catalog. Using the DECam observations of sources in the DES DF as the basis for the input object photometry rectified many of the input sample issues described in Suchyta et al. (2016) – particularly the discrepancy in recovered `Balrog` colors as compared to Y1 GOLD that arose from interpolating the spectral energy distribution (SED) of COSMOS galaxies to match DECam filters. However, Figure 10 shows that we have further work to do. While it is difficult to disentangle intrinsic errors in the emulation of the DESDM pipeline from the input sample representativeness, there are some clear avenues for improvement. The conceptually simplest is to sample a wider population of deep objects across more deep patches of sky in order to incorporate greater cosmic variance in the injection sample. However, these deep observations are very expensive which limits the practicality of this approach. It may be possible to combine with external deep datasets, though this comes at the expense of a return to SED interpolations to match DECam filters. In addition, more detailed studies of the difficult PSF modeling in the DF may yield a stellar population more similar to the WF measurements and resolve some of the largest discrepancies between `Balrog` and Y3 GOLD for bright, PSF-like objects.

[17] While this increases the probability of unwanted proximity effects from other `Balrog` injections, we estimate that the chances of two neighboring injections with `bdf_T` > 10 (or $\sim 3.3''$) in this run to be less than 0.25%.

Another possibility is that the discrepancies between the recovered WF sample and Y3 GOLD are driven, at least in part, by the inability of CModel profiles to accurately capture the full diversity of galaxy morphologies. True galaxy profiles have many complex features such as spiral arms, star knots, and long asymmetric disruptions from mergers that we are not currently capturing with our DF injections. The most direct solution to this problem is to inject the MEDS image cutouts of the DF sources. We have already built the basic infrastructure to do so with `Balrog`, as described in §2.2.2, but there are new issues to consider. The image cutouts can include artifacts, excessive masking, truncated profiles of nearby objects, or even be blended with other sources. This may be rectified in the future by using machine learning methods such as non-negative matrix factorization or generative adversarial networks to handle the required pixel-level deblending of sources in the stamps (see Melchior et al. (2018) and Reiman & Göhre (2019) for examples respectively).

However, using the image cutouts directly would introduce undesired noise when injecting into single-epoch exposures that had better seeing conditions than the composite PSF of the single-chip DF coadd and remains an unresolved issue. In addition, precisely defining the "truth" properties of the stamps is less straightforward than for model fit injections. This will likely by handled by making accurate measurements of each relevant WF photometry type on the stamps which would eliminate inheriting non-physical parameter biases from small profile definition differences such as the resulting magnitude bias from differences in `fracdev` prior shown in Figure 3.

Perhaps the most difficult challenge to overcome is the computational cost of injection pipelines. The new single-epoch processing and additional photometric measurements in Y3 `Balrog` has increased the total mean CPU time per recovered injection to ∼80 seconds; about 12 times greater than in Suchyta et al. (2016). This large increase in runtime is only at Y3 depth corresponding to ∼4-6 epochs per injection. The situation will become significantly worse for much deeper surveys like LSST where we can expect hundreds of individual exposures for each object. Additionally, the high cost of running Y3 `Balrog` led to only a single injection realization at ∼40% Y3 density across just 20% of the total footprint. While this sampling was sufficient to capture systematics variations in the clustering amplitude to better than 1% for a MAGLIM-like sample, reaching this threshold (or beyond) for highly incomplete samples or for accurate calibrations of large-scale variations may require orders of magnitudes more injections.

Despite an expected significant increase in the total tiles sampled for Y6, achieving the many realizations of full footprint coverage required for the most ambitious `Balrog` measurements, such as providing realistic randoms for clustering and shear two-point measurements, remains impractical. One solution that we plan to explore is the use of the `Balrog` samples as a training set for an emulator that predicts additional realizations conditional on the survey property maps. A somewhat similar approach is taken in Johnston et al. (2020) where they mitigate galaxy clustering systematics by producing "organized" random catalogs with fluctuations in number density imprinted from a SOM approach that trained on maps of the variations in KiDS observing properties. Using an injection catalog like `Balrog` directly as a training sample for this approach would leverage our very high fidelity measurements of the survey transfer function to include unknown systematics not fully captured by the identified survey properties. While still more computationally expensive than a machine learning-only approach, this will allow us to build an efficient way of creating accurate random samples tuned for the desired measurement without increasing the total survey pipeline computational cost by more than a factor of two. We plan to use the presented `Balrog` catalogs to gauge the accuracy and feasibility of this approach in an upcoming analysis.

7. SUMMARY AND CONCLUSION

We have presented here the suite of DES Y3 `Balrog` simulations and resulting object catalogs used in downstream Y3 analyses. Like its Y1 predecessor, this current iteration of `Balrog` directly samples the DES transfer function by injecting an ensemble of realistic sources into real survey images to make precise measurements of the inherited systematic biases in the photometric response. However, the updated methodology (and entirely new coding framework) for Y3 `Balrog` makes significant strides beyond Suchyta et al. (2016) in replicating many of the more complex features of the DESDM pipeline including the coaddition of single-epoch images and multi-epoch photometric measurements from SOF and Metacalibration in order to probe more aspects of the true measurement likelihood. In addition, we used a more realistic input sample based on the DES DF source catalog with observations using DECam filters which eliminated the need for template fitting to COSMOS galaxies and incorporated more cosmic variance in object properties. We also implemented a novel ambiguous matching scheme to capture many of the impacts of source blending while largely eliminating the contribu-

tions from undesired dropouts that happened to land on top of existing bright sources.

This effort culminated in tens of millions of Monte Carlo samples of the DES transfer function at high fidelity across 20% of the full DES footprint to Y3 depth, capturing systematic biases from more variations in observing conditions than any previous `Balrog` analysis. The improved methodology resulted in the injected objects matching Y3 GOLD photometric properties and capturing clustering systematics correlated with survey property maps to better than 1% accuracy for a typical cosmology sample on relevant scales. Additionally, we find that `Balrog` captures the clustering amplitudes of these systematics within a few percent for even *highly* incomplete samples – an encouraging first step for future analyses that wish to leverage more of our hard-earned (and often expensive) photons.

We quantified the photometric responses of Balrog injections through the Y3 DESDM measurement pipeline, particularly for magnitudes, colors, and morphology. We find that the magnitudes of most injections are well calibrated until selection effects near the detection threshold become significant, although we have found a clear asymmetric bias for objects in crowded fields or near image artifacts to have moderately to severely overestimated sizes which correlate with large negative magnitude biases. These biases are fairly common for bright, extended objects and can become extremely large (up to $\Delta\mathrm{mag}_{\mathrm{DF}}{\sim}8$) at fainter magnitudes; though they constitute a much larger relative fraction of objects on the bright end. We demonstrated that these catastrophic photometry failures are real effects and often pass science cuts. We plan on exploring the causal relationship of this photometric failure mode further in a future analysis. While these magnitude response biases can cause significant discrepancies from more naive error estimates, fortunately their effect appears to have little impact on the recovered colors where we find typical median response biases of ∼1-3 mmag for stars and ∼5-10 mmag for galaxies in the densest regions of parameter space – an effective median color calibration offset of less than 1%.

Finally, we discussed a few of the most important applications of the presented `Balrog` catalogs to the Y3 cosmology analysis and other DES science measurements. In particular, we provided a realistic measurement likelihood in the calibration of photometric redshifts to reduce systematic biases in one of the highest sources of uncertainty in the cosmological measurement without contributing any additional uncertainty to the overall error budget. Unexpected findings such as the noise contributions from undetected sources in DES images and sky over-subtraction in the *riz*-bands described in Eckert et al. (2020), in addition to the moderate band-dependence in magnitude response and discovery of a new class of catastrophic photometry failures correlated with measured size, are indicative of the diagnostic power of object injection pipelines like `Balrog` in modern galaxy surveys.

We believe that this paper only scratches the surface in cosmological calibration potential and the identification of new systematics with injection pipelines such as `Balrog`. In particular, the combination of direct Monte Carlo sampling of the transfer function with an emulator to boost the total statistical power has the potential to facilitate many of the most difficult measurements in modern galaxy surveys. It is clear that we have yet to dig too deep.

ACKNOWLEDGEMENTS

SE and TJ acknowledge support from the U.S. Department of Energy, Office of Science, Office of High Energy Physics, under Award Numbers DESC0010107 and A00-1465-001.

Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência, Tecnologia e Inovação, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey.

The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenössische Technische Hochschule (ETH) Zürich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciències de l'Espai (IEEC/CSIC), the Institut de Física d'Altes Energies, Lawrence Berkeley Na-

tional Laboratory, the Ludwig-Maximilians Universität München and the associated Excellence Cluster Universe, the University of Michigan, NFS's NOIRLab, the University of Nottingham, The Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, Texas A&M University, and the OzDES Membership Consortium.

Based in part on observations at Cerro Tololo Inter-American Observatory at NSF's NOIRLab (NOIRLab Prop. ID 2012B-0001; PI: J. Frieman), which is managed by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

The DES data management system is supported by the National Science Foundation under Grant Numbers AST-1138766 and AST-1536171. The DES participants from Spanish institutions are partially supported by MICINN under grants ESP2017-89838, PGC2018-094773, PGC2018-102021, SEV-2016-0588, SEV-2016-0597, and MDM-2015-0509, some of which include ERDF funds from the European Union. IFAE is partially funded by the CERCA program of the Generalitat de Catalunya. Research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) including ERC grant agreements 240672, 291329, and 306478. We acknowledge support from the Brazilian Instituto Nacional de Ciência e Tecnologia (INCT) do e-Universo (CNPq grant 465376/2014-2).

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

DATA AVAILABILITY

The DES Y3 data products used in this work will be made publicly available following publication at the URL: https://des.ncsa.illinois.edu/releases.

REFERENCES

Abbott, T., Abdalla, F. B., Aleksić, J., et al. 2016, MNRAS, 460, 1270

Aihara, H., Allende Prieto, C., An, D., et al. 2011, ApJS, 193, 29

Aihara, H., Arimoto, N., Armstrong, R., et al. 2018, PASJ, 70, S4

Alonso, D., Sanchez, J., Slosar, A., & LSST Dark Energy Science Collaboration. 2019, MNRAS, 484, 4127

Amiaux, J., Scaramella, R., Mellier, Y., et al. 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8442, Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave, ed. M. C. Clampin, G. G. Fazio, H. A. MacEwen, & J. Oschmann, Jacobus M., 84420Z

Antilogus, P., Astier, P., Doherty, P., Guyonnet, A., & Regnault, N. 2014, Journal of Instrumentation, 9, C03048

Bernstein, G. M. & Armstrong, R. 2014, MNRAS, 438, 1880

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

Bertin, E. & Arnouts, S. 1996, A&AS, 117, 393

Bertin, E., Mellier, Y., Radovich, M., et al. 2002, in Astronomical Society of the Pacific Conference Series, Vol. 281, Astronomical Data Analysis Software and Systems XI, ed. D. A. Bohlender, D. Durand, & T. H. Handley, 228

Bienaymé, O., Leca, J., & Robin, A. C. 2018, A&A, 620, A103

Blake, C., Brough, S., Colless, M., et al. 2010, MNRAS, 406, 803

Blumenthal, G. R., Faber, S. M., Primack, J. R., & Rees, M. J. 1984, Nature, 311, 517

Brainerd, T. G., Blandford, R. D., & Smail, I. 1996, ApJ, 466, 623

Buchs, R., Davis, C., Gruen, D., et al. 2019, MNRAS, 489, 820

Chang, C., Busha, M. T., Wechsler, R. H., et al. 2015, ApJ, 801, 73

Connolly, A. J., Peterson, J., Jernigan, J. G., et al. 2010, in Modeling, Systems Engineering, and Project Management for Astronomy IV, ed. G. Z. Angeli & P. Dierickx, Vol. 7738, International Society for Optics and Photonics (SPIE), 612

Dawson, W. A., Schneider, M. D., Tyson, J. A., & Jee, M. J. 2015, The Astrophysical Journal, 816, 11

de Jong, J. T. A., Verdoes Kleijn, G. A., Kuijken, K. H., & Valentijn, E. A. 2013, Experimental Astronomy, 35, 25

Drlica-Wagner, A., Sevilla-Noarbe, I., Rykoff, E. S., et al. 2018, ApJS, 235, 33

Eckert, K., Bernstein, G., & et al. 2020, ApJ

Elvin-Poole, J., Crocce, M., Ross, A. J., et al. 2018, PhRvD, 98, 042006

Elvin-Poole, J. et al. 2020, To be submitted to MNRAS

Fenech Conti, I., Herbonnet, R., Hoekstra, H., et al. 2017, MNRAS, 467, 1627
Flaugher, B., Diehl, H. T., Honscheid, K., et al. 2015, AJ, 150, 150
Gaia Collaboration. 2018, A&A, 616, A1
Garcia-Fernandez, M., Sanchez, E., Sevilla-Noarbe, I., et al. 2018, MNRAS, 476, 1071
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Gruen, D. & Brimioulle, F. 2017, MNRAS, 468, 769
Hartley, W. G., Choi, A., et al. 2020, To be submitted to MNRAS
Hildebrandt, H. 2016, MNRAS, 455, 3943
Huang, S., Leauthaud, A., Murata, R., et al. 2018, PASJ, 70, S6
Huff, E. & Mandelbaum, R. 2017, arXiv e-prints, arXiv:1702.02600
Huterer, D., Takada, M., Bernstein, G., & Jain, B. 2006, MNRAS, 366, 101
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jarvis, M., Sheldon, E., Zuntz, J., et al. 2016, MNRAS, 460, 2245
Jarvis, M., Bernstein, G. M., Amon, A., et al. 2020, arXiv e-prints, arXiv:2011.03409
Johnston, H., Wright, A. H., Joachimi, B., et al. 2020, arXiv e-prints, arXiv:2012.08467
Kong, H., Burleigh, K. J., Ross, A., et al. 2020, MNRAS, 499, 3943
Kuijken, K. 2008, A&A, 482, 1053
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
Leistedt, B., Peiris, H. V., Elsner, F., et al. 2016, ApJS, 226, 24
Lewis, A., Challinor, A., & Lasenby, A. 2000, ApJ, 538, 473
Lv, J. & Liu, J. S. 2010, arXiv e-prints, arXiv:1005.5483
Mandelbaum, R. 2018, ARA&A, 56, 393
Martini, P., Bailey, S., Besuner, R. W., et al. 2018, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 10702, Ground-based and Airborne Instrumentation for Astronomy VII, ed. C. J. Evans, L. Simard, & H. Takami, 107021F
Massey, R., Hoekstra, H., Kitching, T., et al. 2013, MNRAS, 429, 661
McCracken, H. J., Milvang-Jensen, B., Dunlop, J., et al. 2012, A&A, 544, A156
Mead, A. J., Peacock, J. A., Heymans, C., Joudaki, S., & Heavens, A. F. 2015, MNRAS, 454, 1958
Melchior, P., Moolekamp, F., Jerdee, M., et al. 2018, Astronomy and Computing, 24, 129
Morganson, E., Gruendl, R. A., Menanteau, F., et al. 2018, PASP, 130, 074501
Mucesh, S., Hartley, W. G., Palmese, A., et al. 2020, arXiv e-prints, arXiv:2012.05928
Myles, J., Alarcon, A., Amon, A., et al. 2020, arXiv e-prints, arXiv:2012.08566
Porredon, A., Crocce, M., Fosalba, P., et al. 2020, arXiv e-prints, arXiv:2011.03411
Pujol, A., Sureau, F., Bobin, J., et al. 2020, A&A, 641, A164
Reiman, D. M. & Göhre, B. E. 2019, MNRAS, 485, 2617
Rodríguez-Monroy, M. et al. 2020, To be submitted to MNRAS
Ross, A. J., Percival, W. J., Sánchez, A. G., et al. 2012, MNRAS, 424, 564
Rowe, B. T. P., Jarvis, M., Mandelbaum, R., et al. 2015, Astronomy and Computing, 10, 121
Rozo, E., Rykoff, E. S., Abate, A., et al. 2016, MNRAS, 461, 1431
Rykoff, E. S., Rozo, E., Hollowood, D., et al. 2016, ApJS, 224, 1
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, The Astrophysical Journal, 500, 525
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Sevilla-Noarbe, I., Bechtol, K., Carrasco Kind, M., et al. 2020, arXiv e-prints, arXiv:2011.03407
Sheldon, E. S. 2014, MNRAS, 444, L25
Sheldon, E. S. & Huff, E. M. 2017, ApJ, 841, 24
Smolčić, V., Ivezić, Ž., Knapp, G. R., et al. 2004, ApJL, 615, L141
Stetson, P. B. 1987, PASP, 99, 191
Suchyta, E., Huff, E. M., Aleksić, J., et al. 2016, MNRAS, 457, 786
Tanoglidis, D., Drlica-Wagner, A., Wei, K., et al. 2020, arXiv e-prints, arXiv:2006.04294
Tegmark, M., Eisenstein, D. J., Strauss, M. A., et al. 2006, PhRvD, 74, 123507
Tolkien, J. 1954, The Fellowship of the Ring (Allen & Unwin)
Troxel, M. A., MacCrann, N., Zuntz, J., et al. 2018, PhRvD, 98, 043528
Unruh, S., Schneider, P., Hilbert, S., et al. 2020, A&A, 638, A96
Weaverdyck, N. & Huterer, D. 2020, arXiv e-prints, arXiv:2007.14499
Zhang, Y., Yanny, B., Palmese, A., et al. 2019, ApJ, 874, 165
Zuntz, J., Sheldon, E., Samuroff, S., et al. 2018, MNRAS, 481, 1149

APPENDIX

A. BALROG CONFIGURATION

Here we show the high-level configuration settings used for `Balrog` Run2 and Run2a, where capitalized quantities in {} refer to local file paths:

```
modules:
  - galsim.des,
  - injector,
  - ngmix_catalog,
  - des_star_catalog

input:
  des_star_catalog:
    base_dir: {INPUT_DIR}
    data_version: y3v02
    model_type: Model_16.5-26.5

  ngmix_catalog:
    catalog_type: bdf
    de_redden: True
    dir: {INPUT_DIR}
    file_name: {INPUT_FILENAME}
    t_max: 100

gal:
  type: List
  items:
  - # y3-merged DF injection
    type: ngmixGalaxy
  - # y3-stars delta-injection
    type: desStar

psf:
  type: DES_PSFEx

stamp:
  draw_method: no_pixel
  gsparams:
    maximum_fft_size: 16384
  type: Balrog

image:
  bands: griz
  extinct_objs: True
  rotate_objs: True
  n_realizations: 1
  noise: {} # Turn on Poisson noise
  nproc: 16
  pos_sampling:
    des_star_catalog:
        type: MixedGrid
        inj_frac: 0.1
    ngmix_catalog
        type: MixedGrid
        grid_spacing: 20
        grid_type: HexGrid
        inj_frac: 0.9
        offset: Random
        rotate: Random
  random_seed: {SEED}
  run_name: {Run2/Run2a}
  type: Balrog
  version: y3v02
  wcs:
    type: Fits
  xsize: 2048
  ysize: 4096
```

B. ANGULAR CLUSTERING SYSTEMATICS

Section 4.1.4 introduced a method for translating the differences between the `Balrog` and Y3 GOLD catalogs into a predicted systematic error in the angular clustering of galaxies. We first choose a sample selection which is applied to both catalogs. We then measure the dependence of galaxy counts fluctuations in both selected `Balrog` and Y3 GOLD samples on several measured image quality indicators, as in Figure 11. Finally, for each data quality indicator, we interpolate the density fluctuation trends to the full survey area and estimate the angular clustering that these trends imply. As small systematic variations in the survey depth enter, to leading order, as additive power in the measured clustering signal, a comparison of the power we measure in these interpolated maps offers a direct estimate of the importance of any deviation between our injection catalogs and the real data.

Here we show the same maps as Figure 12 for six measured survey properties in all bands, for the $17.5 < i < 12.5$ sample selection meant to emulate the Y3 MAGLIM sample. With the exception of a negligible spike in power in a few of the `SIGMA_MAG_ZERO` maps, the measured systematic errors are less than 1% of the fiducial galaxy clustering signal (calculated as described in Figure 11) on scales below approximately 1° ($\ell > 180$).

C. TABULAR RESULTS

Here we present the tabular results of many of the plots shown in Section 4. The mean ($<\Delta>$), median ($\widetilde{\Delta}$), and standard deviation (σ) of the `Balrog` *griz* magnitude responses binned in injection magnitude for the `y3-stars` and `y3-merged` samples are shown in Tables C.1 and C.2 respectively. The equivalent quantities for the color responses are shown in Tables C.3 and C.4. Measurements of the `Balrog` classification, or "confusion", matrix described in §4.4 are shown in Table C.5.

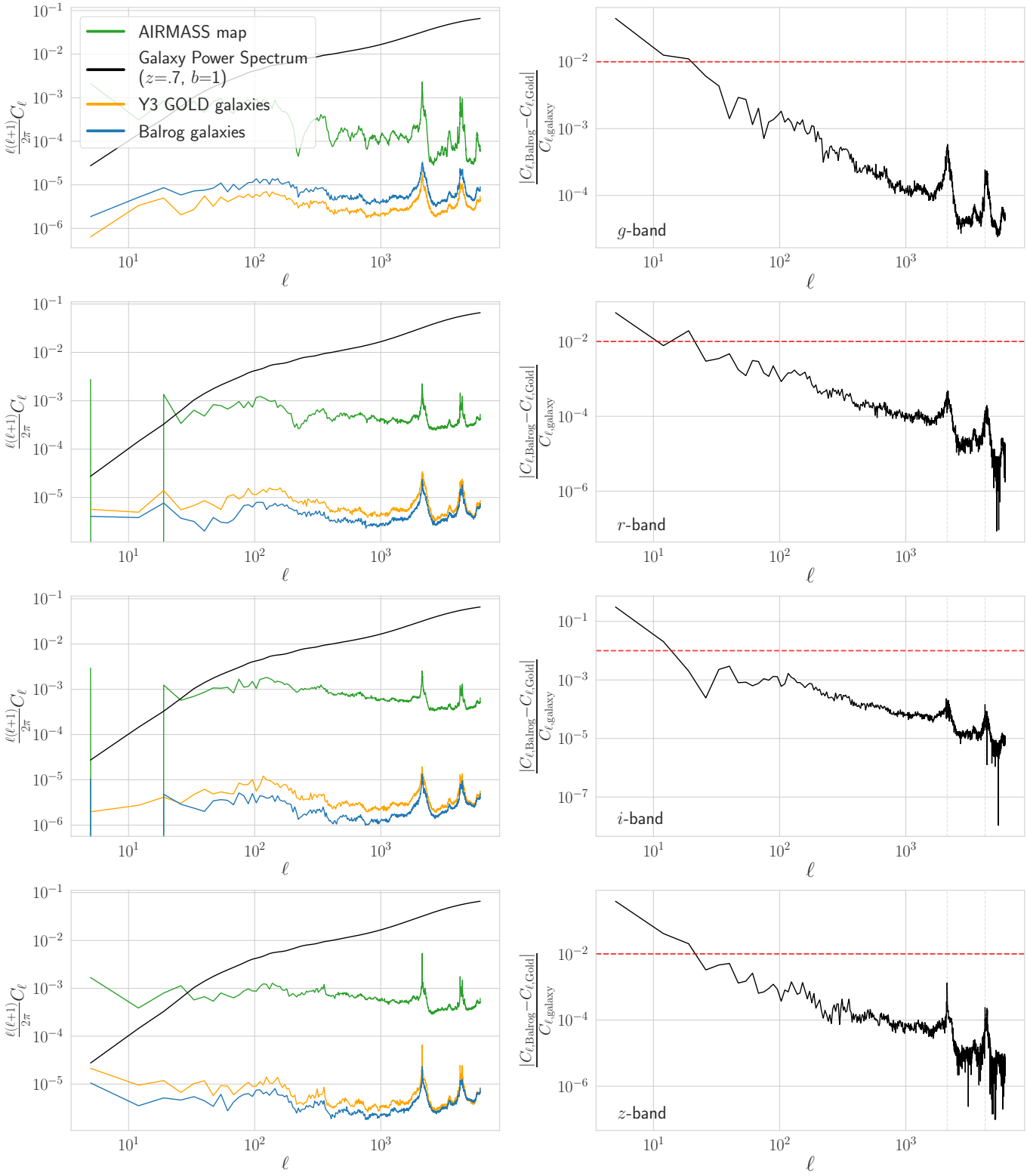


Figure B.1. Power spectra of the mean airmass, and associated interpolated `Balrog` and Y3 GOLD galaxy count variations, as in Figure 12. The left panels show the angular power spectrum of the noted survey property (in green) and the corresponding power spectra of the number densities of the `Balrog` (in blue) and Y3 GOLD (in gold) MAGLIM-like galaxies across the Y3 footprint using the interpolated trends described in §4.1.3 and §4.1.4. The reference galaxy power spectrum in black is `CAMB`'s implementation of the nonlinear matter power spectrum described in Mead et al. (2015), meant to represent a typical cosmological signal at $z = 0.7$ with linear galaxy bias parameter of 1. The right panels show the difference in power between Y3 GOLD and `Balrog` as a fraction of the fiducial cosmological power spectrum shown on the left.

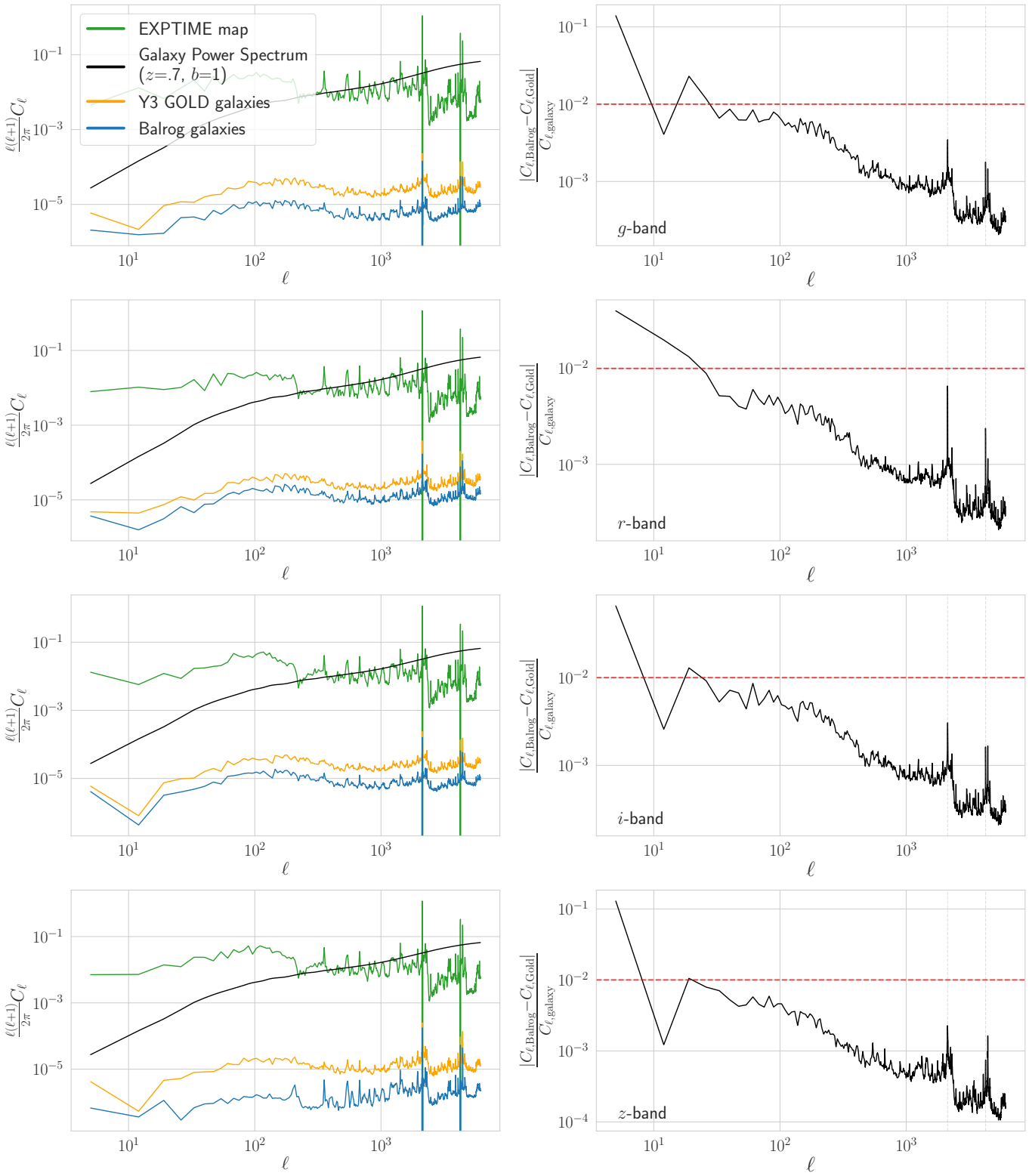


Figure B.2. Power spectra of the mean exposure time, and associated interpolated `Balrog` and Y3 GOLD galaxy count variations, as in Figure 12. The left panels show the angular power spectrum of the noted survey property (in green) and the corresponding power spectra of the number densities of the `Balrog` (in blue) and Y3 GOLD (in gold) MAGLIM-like galaxies across the Y3 footprint using the interpolated trends described in §4.1.3 and §4.1.4. The reference galaxy power spectrum in black is `CAMB`'s implementation of the nonlinear matter power spectrum described in Mead et al. (2015), meant to represent a typical cosmological signal at $z = 0.7$ with linear galaxy bias parameter of 1. The right panels show the difference in power between Y3 GOLD and `Balrog` as a fraction of the fiducial cosmological power spectrum shown on the left.

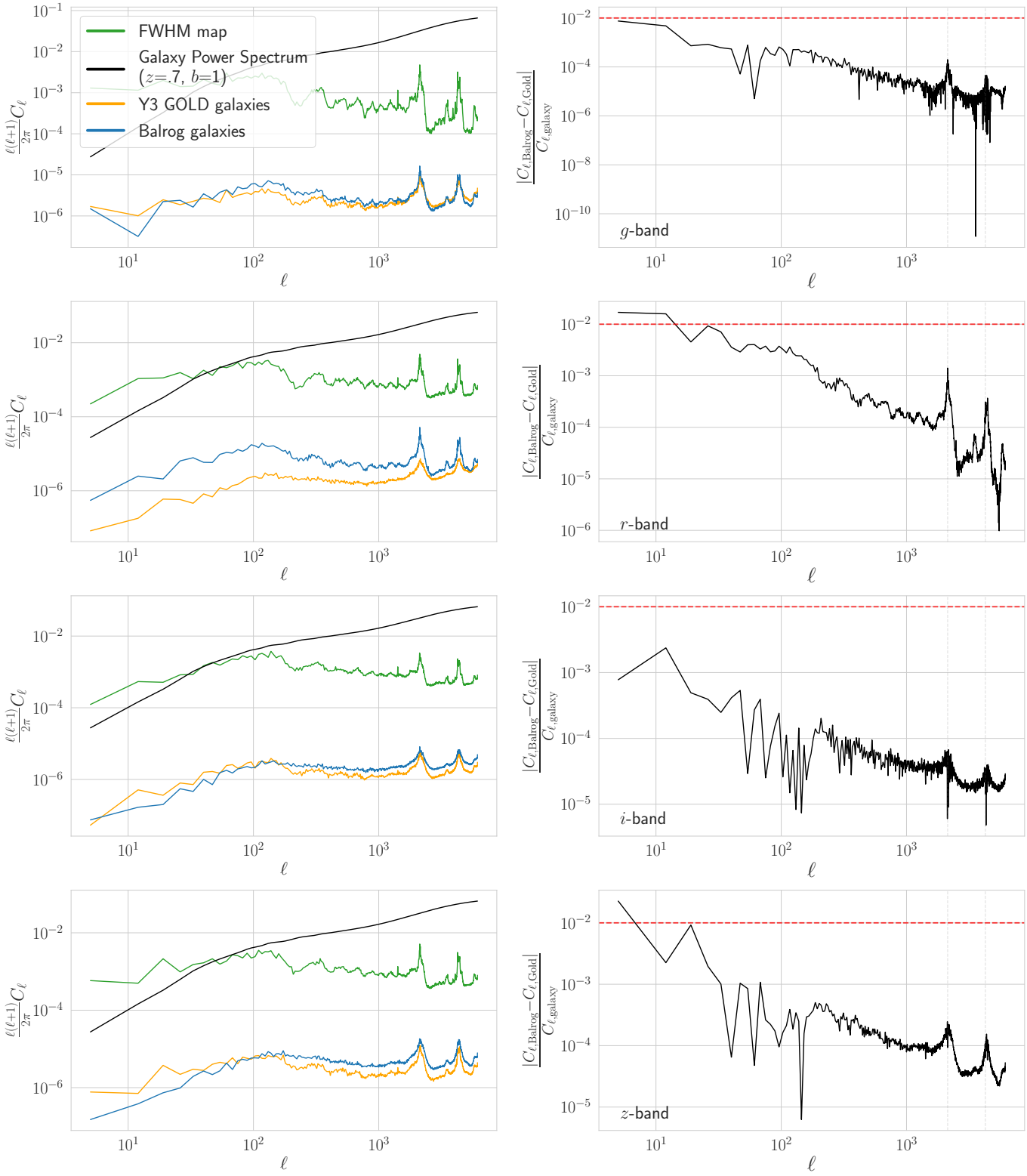


Figure B.3. Power spectra of the mean PSF FWHM, and associated interpolated `Balrog` and Y3 GOLD galaxy count variations, as in Figure 12. The left panels show the angular power spectrum of the noted survey property (in green) and the corresponding power spectra of the number densities of the `Balrog` (in blue) and Y3 GOLD (in gold) MAGLIM-like galaxies across the Y3 footprint using the interpolated trends described in §4.1.3 and §4.1.4. The reference galaxy power spectrum in black is `CAMB`'s implementation of the nonlinear matter power spectrum described in Mead et al. (2015), meant to represent a typical cosmological signal at $z = 0.7$ with linear galaxy bias parameter of 1. The right panels show the difference in power between Y3 GOLD and `Balrog` as a fraction of the fiducial cosmological power spectrum shown on the left.

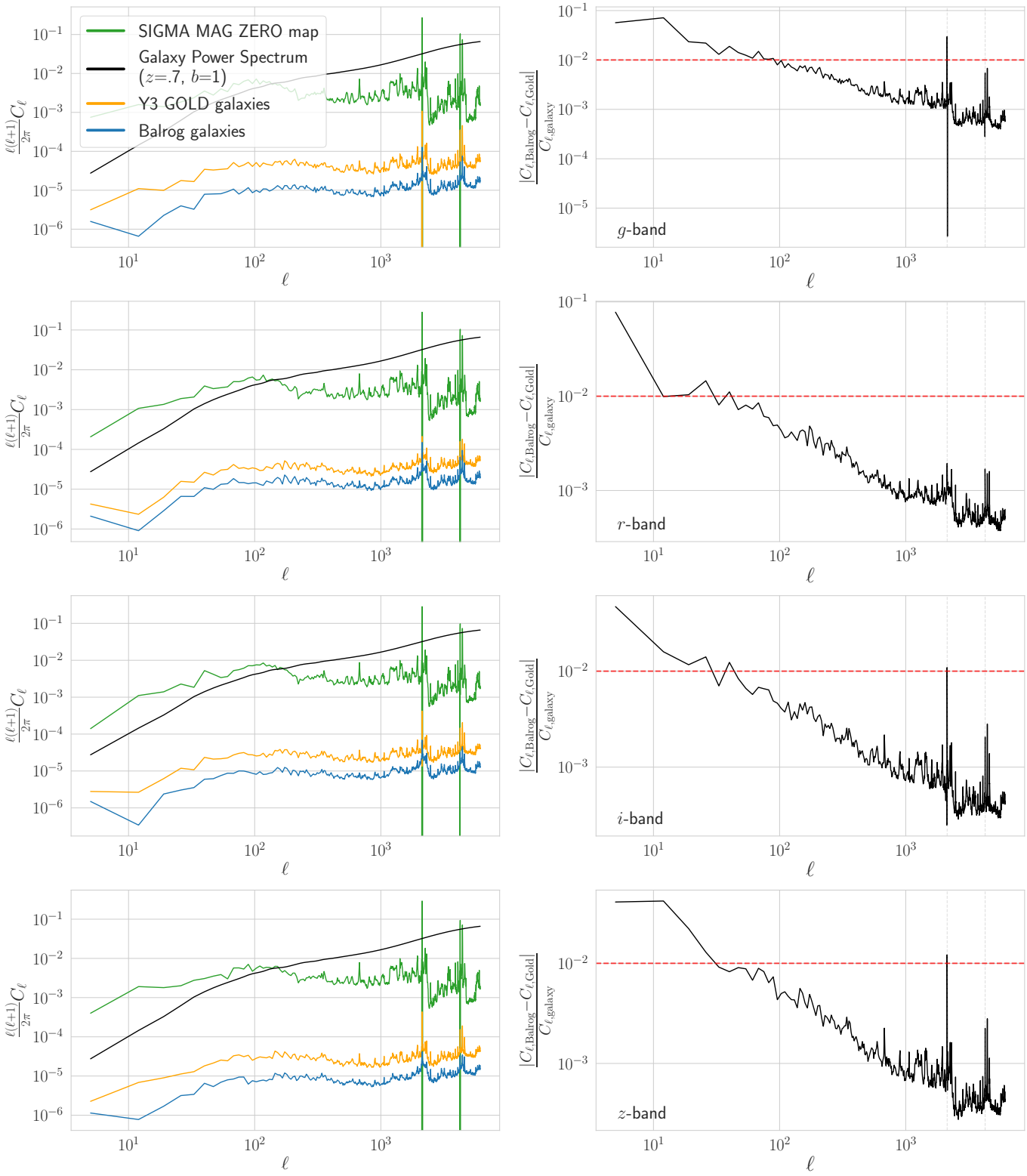


Figure B.4. Power spectra of the mean error on the grey zeropoint calibration (resulting primarily from transparency corrections) and associated interpolated `Balrog` and Y3 GOLD galaxy count variations, as in Figure 12. The left panels show the angular power spectrum of the noted survey property (in green) and the corresponding power spectra of the number densities of the `Balrog` (in blue) and Y3 GOLD (in gold) MAGLIM-like galaxies across the Y3 footprint using the interpolated trends described in §4.1.3 and §4.1.4. The reference galaxy power spectrum in black is `CAMB`'s implementation of the nonlinear matter power spectrum described in Mead et al. (2015), meant to represent a typical cosmological signal at $z = 0.7$ with linear galaxy bias parameter of 1. The right panels show the difference in power between Y3 GOLD and `Balrog` as a fraction of the fiducial cosmological power spectrum shown on the left.

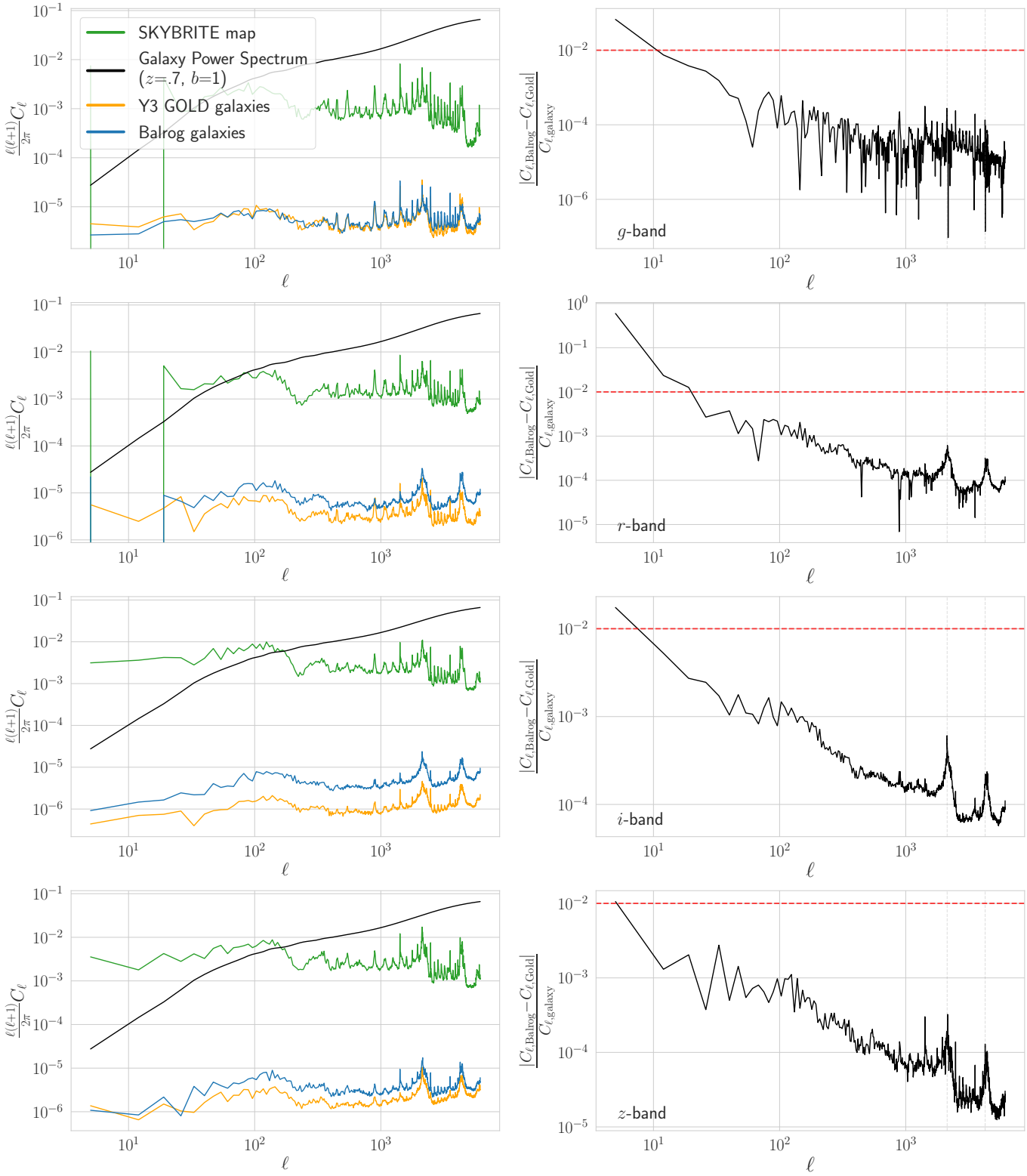


Figure B.5. Power spectra of the mean sky brightness, and associated interpolated `Balrog` and Y3 GOLD galaxy count variations, as in Figure 12. The left panels show the angular power spectrum of the noted survey property (in green) and the corresponding power spectra of the number densities of the `Balrog` (in blue) and Y3 GOLD (in gold) MAGLIM-like galaxies across the Y3 footprint using the interpolated trends described in §4.1.3 and §4.1.4. The reference galaxy power spectrum in black is `CAMB`'s implementation of the nonlinear matter power spectrum described in Mead et al. (2015), meant to represent a typical cosmological signal at $z = 0.7$ with linear galaxy bias parameter of 1. The right panels show the difference in power between Y3 GOLD and `Balrog` as a fraction of the fiducial cosmological power spectrum shown on the left.

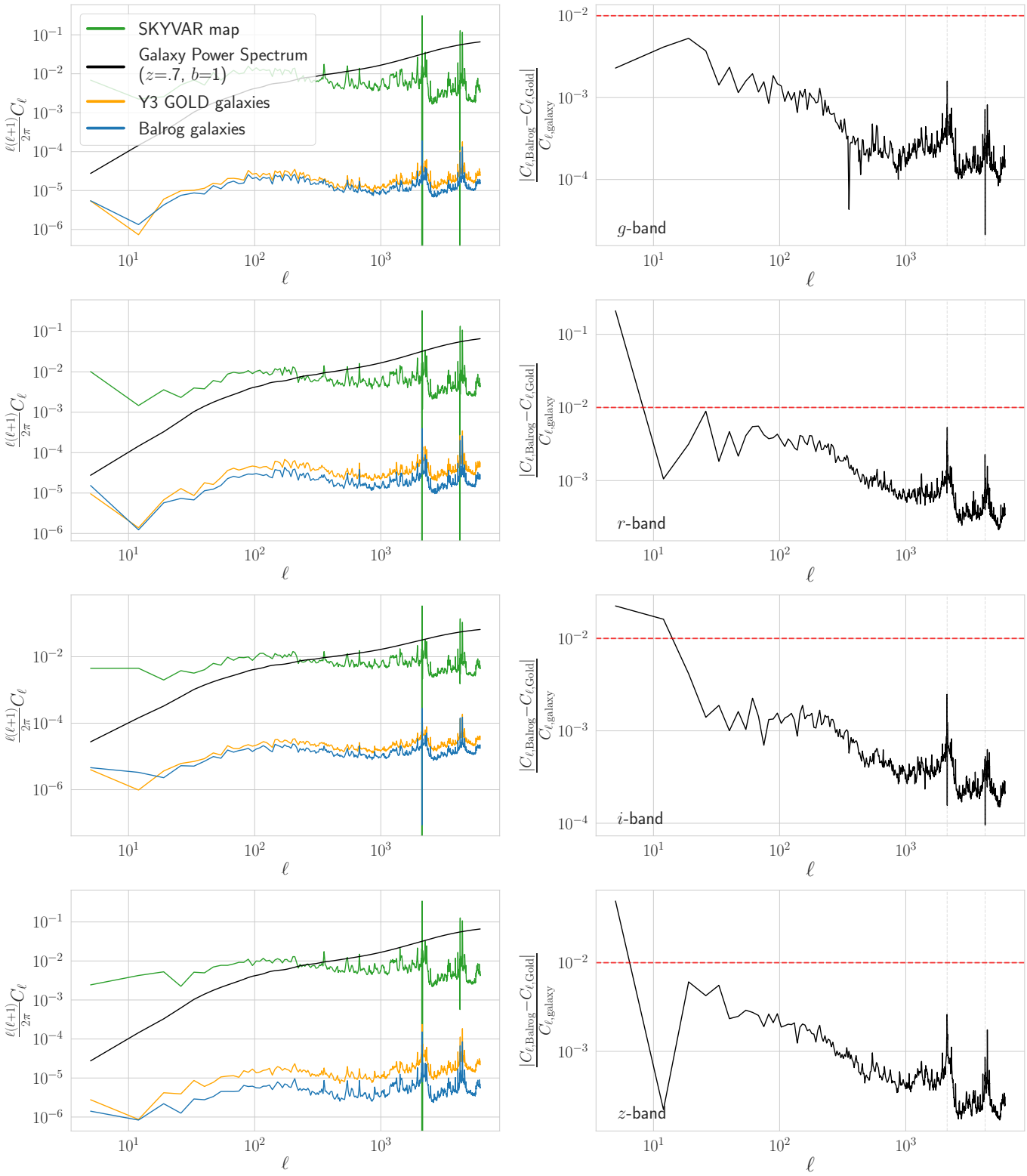


Figure B.6. Power spectra of the variance from sky background, and associated interpolated balrog and Y3 GOLD galaxy count variations, as in Figure 12. The left panels show the angular power spectrum of the noted survey property (in green) and the corresponding power spectra of the number densities of the `Balrog` (in blue) and Y3 GOLD (in gold) MAGLIM-like galaxies across the Y3 footprint using the interpolated trends described in §4.1.3 and §4.1.4. The reference galaxy power spectrum in black is `CAMB`'s implementation of the nonlinear matter power spectrum described in Mead et al. (2015), meant to represent a typical cosmological signal at $z = 0.7$ with linear galaxy bias parameter of 1. The right panels show the difference in power between Y3 GOLD and `Balrog` as a fraction of the fiducial cosmological power spectrum shown on the left.

True Mag	$<\Delta g>$ (mag)	$\widetilde{\Delta g}$ (mag)	σ_g (mag)	$<\Delta r>$ (mag)	$\widetilde{\Delta r}$ (mag)	σ_r (mag)	$<\Delta i>$ (mag)	$\widetilde{\Delta i}$ (mag)	σ_i (mag)	$<\Delta z>$ (mag)	$\widetilde{\Delta z}$ (mag)	σ_z (mag)
17.00	0.001	0.000	0.004	0.002	0.003	0.005	0.003	0.004	0.010	0.005	0.006	0.006
17.25	0.001	0.003	0.013	0.002	0.003	0.009	0.004	0.005	0.008	0.006	0.006	0.006
17.50	0.001	0.002	0.005	0.002	0.004	0.026	0.005	0.006	0.009	0.006	0.007	0.012
17.75	0.001	0.002	0.006	0.003	0.004	0.007	0.005	0.006	0.006	0.006	0.007	0.011
18.00	0.002	0.003	0.015	0.004	0.005	0.006	0.006	0.006	0.011	0.007	0.007	0.008
18.25	0.003	0.003	0.006	0.005	0.005	0.008	0.006	0.007	0.013	0.006	0.007	0.011
18.50	0.004	0.004	0.006	0.005	0.006	0.008	0.006	0.007	0.009	0.007	0.008	0.014
18.75	0.004	0.004	0.014	0.005	0.006	0.010	0.006	0.007	0.008	0.007	0.008	0.017
19.00	0.004	0.004	0.008	0.005	0.006	0.014	0.005	0.007	0.022	0.007	0.008	0.017
19.25	0.004	0.005	0.008	0.004	0.006	0.017	0.006	0.007	0.013	0.007	0.009	0.021
19.50	0.004	0.005	0.008	0.004	0.006	0.021	0.005	0.007	0.026	0.007	0.009	0.026
19.75	0.004	0.005	0.015	0.004	0.006	0.023	0.006	0.008	0.022	0.007	0.010	0.023
20.00	0.003	0.005	0.015	0.004	0.006	0.029	0.006	0.008	0.028	0.008	0.010	0.034
20.25	0.003	0.005	0.029	0.004	0.007	0.024	0.006	0.009	0.019	0.009	0.011	0.025
20.50	0.003	0.005	0.031	0.004	0.007	0.030	0.007	0.010	0.028	0.009	0.012	0.037
20.75	0.003	0.005	0.033	0.005	0.008	0.028	0.007	0.010	0.038	0.010	0.013	0.041
21.00	0.003	0.005	0.032	0.005	0.008	0.030	0.008	0.011	0.033	0.012	0.014	0.043
21.25	0.003	0.006	0.029	0.005	0.009	0.027	0.009	0.012	0.033	0.013	0.015	0.052
21.50	0.003	0.006	0.030	0.006	0.010	0.031	0.011	0.014	0.042	0.016	0.017	0.059
21.75	0.002	0.006	0.033	0.006	0.010	0.037	0.012	0.015	0.048	0.018	0.019	0.072
22.00	0.002	0.006	0.047	0.007	0.011	0.042	0.014	0.016	0.053	0.022	0.021	0.085
22.25	0.002	0.006	0.044	0.009	0.013	0.049	0.017	0.018	0.065	0.026	0.024	0.107
22.50	0.002	0.006	0.055	0.011	0.014	0.064	0.020	0.020	0.076	0.031	0.026	0.126
22.75	0.003	0.006	0.065	0.014	0.016	0.070	0.023	0.022	0.108	0.038	0.028	0.159
23.00	0.004	0.008	0.083	0.017	0.019	0.083	0.028	0.025	0.107	0.047	0.033	0.218
23.25	0.006	0.009	0.093	0.022	0.022	0.098	0.031	0.025	0.131	0.062	0.035	0.304
23.50	0.010	0.012	0.124	0.027	0.024	0.116	0.033	0.024	0.162	0.084	0.031	0.521
23.75	0.015	0.014	0.147	0.034	0.029	0.154	0.031	0.018	0.200	0.140	0.037	0.876
24.00	0.021	0.016	0.174	0.045	0.034	0.208	0.017	-0.007	0.315	0.297	0.036	1.571
24.25	0.033	0.021	0.245	0.052	0.035	0.226	0.002	-0.041	0.463	0.456	-0.014	2.170

Table C.1. The mean ($<\Delta>$), median ($\widetilde{\Delta}$), and standard deviation (σ) of the `Balrog` *griz* magnitude responses binned in injection magnitude for the `y3-stars` sample. The quoted magnitudes correspond to the left bin edge. Simple Gaussian statistics do not fully capture the complexity of the responses – see Figure 13.

True Mag	$<\Delta g>$ (mag)	$\widetilde{\Delta g}$ (mag)	σ_g (mag)	$<\Delta r>$ (mag)	$\widetilde{\Delta r}$ (mag)	σ_r (mag)	$<\Delta i>$ (mag)	$\widetilde{\Delta i}$ (mag)	σ_i (mag)	$<\Delta z>$ (mag)	$\widetilde{\Delta z}$ (mag)	σ_z (mag)
18.00	-0.066	-0.039	0.081	-0.055	-0.035	0.081	-0.048	-0.029	0.087	-0.043	-0.024	0.076
18.25	-0.063	-0.042	0.101	-0.052	-0.033	0.084	-0.042	-0.024	0.069	-0.039	-0.020	0.076
18.50	-0.059	-0.036	0.077	-0.046	-0.028	0.079	-0.039	-0.019	0.079	-0.040	-0.020	0.083
18.75	-0.055	-0.036	0.078	-0.039	-0.020	0.076	-0.039	-0.020	0.077	-0.034	-0.014	0.083
19.00	-0.055	-0.033	0.083	-0.041	-0.021	0.077	-0.035	-0.015	0.086	-0.031	-0.010	0.090
19.25	-0.044	-0.023	0.084	-0.036	-0.018	0.079	-0.031	-0.011	0.085	-0.026	-0.006	0.101
19.50	-0.040	-0.022	0.078	-0.033	-0.013	0.087	-0.027	-0.006	0.096	-0.022	-0.002	0.105
19.75	-0.040	-0.020	0.085	-0.030	-0.009	0.088	-0.025	-0.003	0.109	-0.019	0.002	0.115
20.00	-0.035	-0.015	0.078	-0.026	-0.006	0.105	-0.022	0.000	0.110	-0.016	0.005	0.125
20.25	-0.035	-0.015	0.098	-0.024	-0.003	0.105	-0.020	0.003	0.119	-0.012	0.009	0.134
20.50	-0.032	-0.012	0.090	-0.023	0.000	0.109	-0.016	0.006	0.126	-0.008	0.013	0.153
20.75	-0.030	-0.009	0.110	-0.020	0.002	0.122	-0.013	0.009	0.145	-0.003	0.017	0.161
21.00	-0.027	-0.006	0.107	-0.018	0.005	0.133	-0.010	0.013	0.155	0.001	0.021	0.174
21.25	-0.026	-0.005	0.116	-0.016	0.008	0.148	-0.007	0.017	0.163	0.003	0.025	0.194
21.50	-0.023	-0.002	0.127	-0.014	0.010	0.157	-0.005	0.020	0.176	0.006	0.028	0.211
21.75	-0.022	0.000	0.147	-0.012	0.014	0.171	-0.002	0.023	0.189	0.008	0.031	0.228
22.00	-0.020	0.002	0.154	-0.010	0.017	0.181	-0.001	0.026	0.203	0.011	0.034	0.254
22.25	-0.019	0.005	0.171	-0.009	0.020	0.192	0.001	0.030	0.222	0.015	0.036	0.291
22.50	-0.017	0.007	0.187	-0.007	0.024	0.212	0.003	0.033	0.248	0.020	0.039	0.339
22.75	-0.017	0.010	0.200	-0.005	0.028	0.231	0.005	0.036	0.279	0.022	0.037	0.403
23.00	-0.014	0.013	0.220	-0.004	0.031	0.259	0.004	0.036	0.314	0.024	0.030	0.496
23.25	-0.012	0.017	0.247	-0.004	0.034	0.293	-0.002	0.031	0.355	0.028	0.014	0.663
23.50	-0.011	0.020	0.279	-0.008	0.033	0.329	-0.023	0.013	0.391	0.037	-0.013	0.916
23.75	-0.009	0.022	0.323	-0.023	0.021	0.369	-0.064	-0.026	0.442	0.069	-0.053	1.312
24.00	-0.009	0.020	0.383	-0.055	-0.007	0.413	-0.132	-0.091	0.528	0.142	-0.115	1.874
24.25	-0.012	0.014	0.463	-0.108	-0.057	0.492	-0.233	-0.194	0.713	0.232	-0.217	2.463

Table C.2. The mean ($<\Delta>$), median ($\widetilde{\Delta}$), and standard deviation (σ) of the `Balrog` *griz* magnitude responses binned in injection magnitude for the `y3-merged` sample. The quoted magnitudes correspond to the left bin edge. Simple Gaussian statistics do not fully capture the complexity of the responses – see Figure 16.

True Color	$<g-r>$ (mag)	$\widetilde{g-r}$ (mag)	σ_{g-r} (mag)	$<r-i>$ (mag)	$\widetilde{r-i}$ (mag)	σ_{r-i} (mag)	$<i-z>$ (mag)	$\widetilde{i-z}$ (mag)	σ_{i-z} (mag)
-0.2	-0.006	-0.003	0.082	-0.006	-0.003	0.111	0.000	-0.003	0.156
-0.1	-0.004	-0.002	0.098	-0.007	-0.003	0.102	-0.002	-0.002	0.114
0.0	-0.003	-0.002	0.092	-0.004	-0.002	0.074	-0.002	-0.001	0.091
0.1	-0.004	-0.003	0.09	-0.004	-0.002	0.078	-0.002	-0.001	0.11
0.2	-0.002	-0.002	0.074	-0.003	-0.002	0.09	-0.002	-0.001	0.111
0.3	-0.001	-0.002	0.077	-0.002	-0.002	0.097	-0.002	-0.001	0.101
0.4	-0.001	-0.001	0.085	-0.001	-0.002	0.096	-0.002	-0.001	0.092
0.5	0.000	-0.001	0.09	0.000	-0.001	0.094	-0.001	-0.001	0.087
0.6	0.000	-0.001	0.103	0.001	-0.001	0.091	0.000	-0.001	0.083
0.7	-0.001	-0.001	0.109	0.001	-0.001	0.088	0.001	-0.001	0.078
0.8	-0.002	-0.001	0.113	0.002	-0.001	0.092	0.001	0.000	0.075
0.9	-0.003	-0.001	0.126	0.002	-0.001	0.097	0.001	0.000	0.081
1.0	-0.006	-0.001	0.131	0.002	-0.001	0.101	0.004	0.001	0.084
1.1	-0.010	-0.002	0.142	0.003	-0.001	0.106	0.003	0.001	0.078
1.2	-0.017	-0.003	0.154	0.002	-0.001	0.112	0.020	0.001	0.073
1.3	-0.021	-0.003	0.155	0.002	0.000	0.116	-0.024	0.000	0.177
1.4	-0.027	-0.004	0.17	0.000	0.001	0.123	0.006	-0.003	0.119
1.5	-0.044	-0.01	0.208	0.000	0.000	0.129	-0.008	-0.008	0.007
1.6	-0.061	-0.017	0.24	0.000	0.000	0.137	–	–	–
1.7	-0.076	-0.026	0.265	-0.004	-0.001	0.138	–	–	–

Table C.3. The mean ($<\Delta>$), median ($\widetilde{\Delta}$), and standard deviation (σ) of the `Balrog` $g-r$, $r-i$, and $i-z$ color responses binned in injection color for the `y3-stars` sample. The quoted colors correspond to the left bin edge. Simple Gaussian statistics do not fully capture the complexity of the responses – see Figure 14.

True Color	$<g-r>$ (mag)	$\widetilde{g-r}$ (mag)	σ_{g-r} (mag)	$<r-i>$ (mag)	$\widetilde{r-i}$ (mag)	σ_{r-i} (mag)	$<i-z>$ (mag)	$\widetilde{i-z}$ (mag)	σ_{i-z} (mag)
-0.2	0.081	0.053	0.211	0.079	0.043	0.216	0.092	0.047	0.239
-0.1	0.047	0.030	0.192	0.053	0.032	0.201	0.062	0.030	0.213
0.0	0.026	0.016	0.182	0.028	0.013	0.182	0.030	0.009	0.177
0.1	0.012	0.006	0.179	0.011	0.002	0.155	0.019	0.004	0.163
0.2	0.002	-0.002	0.178	0.004	0.000	0.140	0.011	0.001	0.145
0.3	-0.009	-0.006	0.169	0.001	-0.001	0.140	0.007	0.000	0.134
0.4	-0.015	-0.008	0.161	-0.003	-0.001	0.139	0.004	0.000	0.141
0.5	-0.019	-0.009	0.158	-0.007	-0.003	0.140	0.001	-0.001	0.160
0.6	-0.024	-0.010	0.157	-0.012	-0.005	0.146	-0.004	-0.003	0.161
0.7	-0.028	-0.011	0.158	-0.015	-0.007	0.147	-0.009	-0.005	0.159
0.8	-0.031	-0.011	0.159	-0.018	-0.007	0.146	-0.012	-0.007	0.161
0.9	-0.036	-0.011	0.162	-0.022	-0.008	0.152	-0.016	-0.009	0.171
1.0	-0.041	-0.011	0.167	-0.026	-0.010	0.161	-0.019	-0.011	0.176
1.1	-0.046	-0.011	0.173	-0.029	-0.012	0.170	-0.031	-0.016	0.193
1.2	-0.051	-0.010	0.184	-0.035	-0.013	0.178	-0.053	-0.024	0.210
1.3	-0.059	-0.011	0.194	-0.071	-0.030	0.221	-0.049	-0.024	0.215
1.4	-0.069	-0.013	0.210	-0.149	-0.091	0.276	-0.054	-0.018	0.223
1.5	-0.074	-0.015	0.222	-0.171	-0.105	0.288	-0.076	-0.028	0.236
1.6	-0.070	-0.016	0.224	-0.183	-0.112	0.300	-0.075	-0.015	0.220
1.7	-0.066	-0.016	0.224	-0.206	-0.126	0.314	-0.050	-0.007	0.240
1.8	-0.096	-0.028	0.265	-0.206	-0.127	0.334	-0.063	-0.017	0.255
1.9	-0.193	-0.092	0.358	-0.221	-0.112	0.363	-0.061	-0.003	0.220

Table C.4. The mean ($<\Delta>$), median ($\widetilde{\Delta}$), and standard deviation (σ) of the `Balrog` $g-r$, $r-i$, and $i-z$ color responses binned in injection color for the `y3-merged` sample. The quoted colors correspond to the left bin edge. Simple Gaussian statistics do not fully capture the complexity of the responses – see Figure 18.

True Mag	Star->Star (TP; %)	Gal->Star (FP; %)	Star->Gal (FN; %)	Gal->Gal (TN; %)
18.50	99.6	1.6	0.4	98.4
18.75	99.6	2.9	0.4	97.1
19.00	99.4	2.9	0.6	97.1
19.25	99.3	2.6	0.7	97.4
19.50	99.2	2.8	0.8	97.2
19.75	99.1	2.3	0.9	97.7
20.00	98.7	1.9	1.3	98.1
20.25	98.6	1.8	1.4	98.2
20.50	98.2	1.8	1.8	98.2
20.75	97.8	1.9	2.2	98.1
21.00	97.3	1.8	2.7	98.2
21.25	96.7	1.7	3.3	98.3
21.50	95.9	2.2	4.1	97.8
21.75	95.1	2.0	4.9	98.0
22.00	93.4	2.3	6.6	97.7
22.25	90.8	3.2	9.2	96.8
22.50	86.4	4.1	13.6	95.9
22.75	79.5	5.2	20.5	94.8
23.00	70.3	6.7	29.7	93.3
23.25	58.2	8.3	41.8	91.7
23.50	46.4	10.1	53.6	89.9
23.75	37.5	12.4	62.5	87.6
24.00	30.9	14.5	69.1	85.5
24.25	25.9	15.0	74.1	85.0

Table C.5. Elements of the classification (or confusion) matrix for `Balrog` sources binned by injection magnitude when normalized by percent, where the measured classification is determined by `EXTENDED_CLASS_SOF`<= 1 for stars and `EXTENDED_CLASS_SOF`> 1 for galaxies. The second through fifth columns correspond to the true positive (TP), false positive (FP), false negative (FN), and true negative (TN) rates of `Balrog` stars respectively. The very pure `y3-stars` sample is used to compute the TP and FN rates, while the nosier classifications of the DF `y3-merged` injections are used for the rest. The quoted magnitudes correspond to the left bin edge. See Figure 23.