

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

J. Sánchez^{*}, H. Awan, E. Gawiser, A. Slosar, D. Kirkby, J. Chiang, C. W. Walter, the DE

8 January 2018

ABSTRACT

The success of the Large Synoptic Survey Telescope (LSST) as a dark energy experiment will depend on controlling systematic effects for the various cosmological probes. Simulations are critical for developing the methodology to estimate and mitigate these systematics. In the first Data Challenge from the LSST Dark Energy Science Collaboration, we evaluate the potential systematic effects that will affect observables, with an emphasis on galaxy clustering. We utilized two approaches to simulate LSST images, one of which involves fully realistic ray-tracing. Simulated images were then processed, combined, and analyzed using the current version of the LSST Data Management pipeline. Here we characterize the resulting systematics and implement corrections. Our results demonstrate that we can generate realistic LSST-like simulated images and control the systematic effects, after processing these images, at a sufficient level to enable major advances in our knowledge of dark energy and cosmology. The methodology presented here can be easily translated to current and future imaging surveys. **Rewrite?**

Key words: LSS , Data challenge, Systematics

1 INTRODUCTION

One of the most critical aspects in Stage IV experiments is the characterization of their instrumentation and systematic effects (Albrecht et al. 2006). In the case of LSST (Ivezic et al. 2008; LSST Science Collaboration 2009; LSST Dark Energy Science Collaboration 2012) this becomes especially difficult given its wide variety of cosmological probes. In this paper we present a methodology to characterize the LSST large-scale structure (LSS) transfer function and an analysis of potential systematic effects present in LSS analyses. **rewrite**

2 DC1 DATA

In this paper we are going to analyze data from the LSST-DESC Data Challenge 1 (DC1) **Add reference**. In particular, we are interested in the images from the DC1 dataset generated with a Fourier optics/analytic modeling using the ImSim package¹, which internally calls the GalSim package (Rowe et al. 2015) to generate the images for the different objects. The input catalog for these simulations is the

CatSim² catalog. This catalog is based on the dark matter haloes from the Millenium simulation and a semi-analytic baryon model described in (De Lucia et al. 2006). The galaxies are described as bulge+disk with no cosmic shear. This catalog covers a redshift range from $0 < z < 6$. The redshift and magnitude distribution can be seen at Figure 1. It has a higher number density of galaxies than the expected for LSST, since it contains sources fainter than its detectable limits. This simulation also contains stars represented as point sources and drawn from the Galfast model (Jurić et al. 2008) with peculiar motion. CatSim also contains associates each source with an SED.

In total the input catalog contains approximately 63.1 million sources of which, 61.3 million are galaxies and 1.8 million are stars. Using this catalog we generate images covering an area of approximately 40 square degrees (4 pointings) in r-band to LSST full depth (10 years). The final footprint can be seen in Figure 2

Two sets of images were generated using imSim: One with no dithering, which we will refer to *undithered*; and the other one which includes the dithering strategy in the `minion.1016.new.dithers` OpSim³ database presented in ref. (Awan et al. 2016). We will refer to the latter dataset as

^{*} e-mail:francs1@uci.edu

¹ <https://github.com/LSSTDESC/imSim>

² <https://www.lsst.org/scientists/simulations/catsim>

³ <https://www.lsst.org/scientists/simulations/opsim>

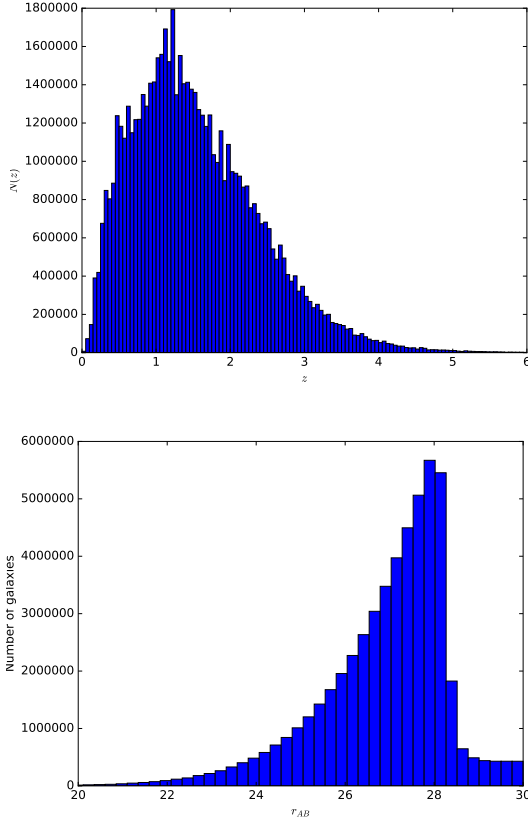


Figure 1. Redshift (top) and magnitude (bottom) distribution for the galaxies used as inputs for DC1.

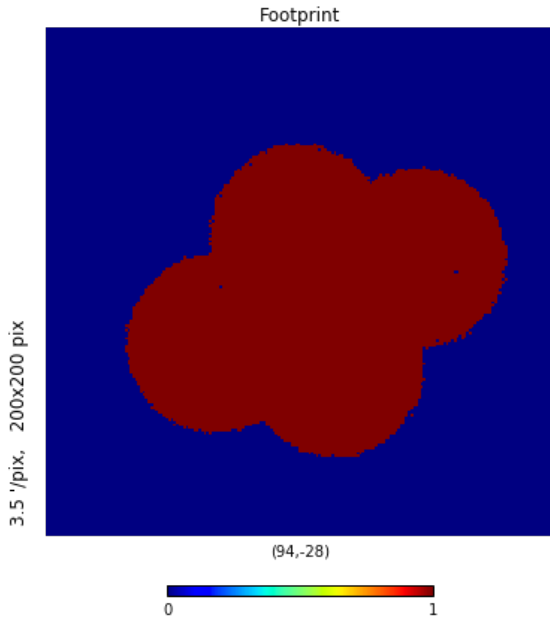


Figure 2. Footprint of the DC1 dataset. We simulate 4 LSST full focal plane pointings which roughly corresponds to 40 square degrees.

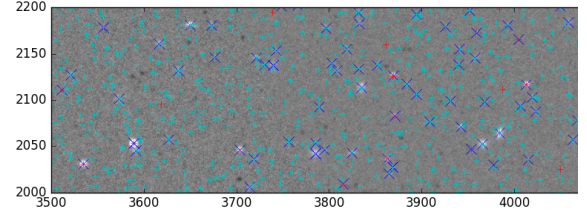


Figure 3. Example of a 550×200 pixel simulated field. The axes show the pixel number in the CCD. The input galaxies are marked with cyan +, the input stars are marked with red +, and the detected sources are marked with blue x. We can appreciate the vast difference in density between the simulation and the output given that CatSim contains objects fainter than what LSST can detect.

the *dithered* run. We do not simulate images out of the footprint in Figure 2. Using these two datasets allows us to make a fair comparison between them, and see the advantages and disadvantages of each approach.

The outputs of these simulations are then processed using the LSST data management (DM) stack (Ivezic et al. 2008; LSST Science Collaboration 2009; LSST Dark Energy Science Collaboration 2012; Bosch et al. 2017) **Is there any other reference for the stack?** The DM stack is intended to be the software used to process the data produced by LSST. Although it is still under development, most of the pieces are already meeting the science requirements or very close to do that. Running the DM stack on this dataset also constitutes a necessary test for this invaluable piece of LSST. The stack can be divided into three main steps: the coadding step where the images are put together to form a deeper image than each individual exposure; the detection step, where the coadded image is analyzed searching for peaks on the counts; and the measurement step, where these peaks are characterized. The DM stack performs these three steps automatically and provides calibrated images and a source catalog with ~ 10.6 million objects with position, flux and shape information.

An example of a coadd image can be seen at Figure 3.

3 GENERATING DEPTH MAPS

In order to estimate the depth in the coadd catalogs we select the stars by using `base_ClassificationExtendedness_value==0`. This ensures selecting PSF-like objects. After this, we have two different approaches:

(i) In the first approach we generate a HEALPix (Górski et al. 2005) map containing the PSF-like objects detected with a signal-to-noise ratio (SNR) higher or equal than 5 (`base_PsfFlux_flux/base_PsfFlux_fluxSigma>=5`), and we assign to each pixel the value of the dimmest object contained in it.

(ii) The second procedure also generates a HEALPix (Górski et al. 2005) map containing the

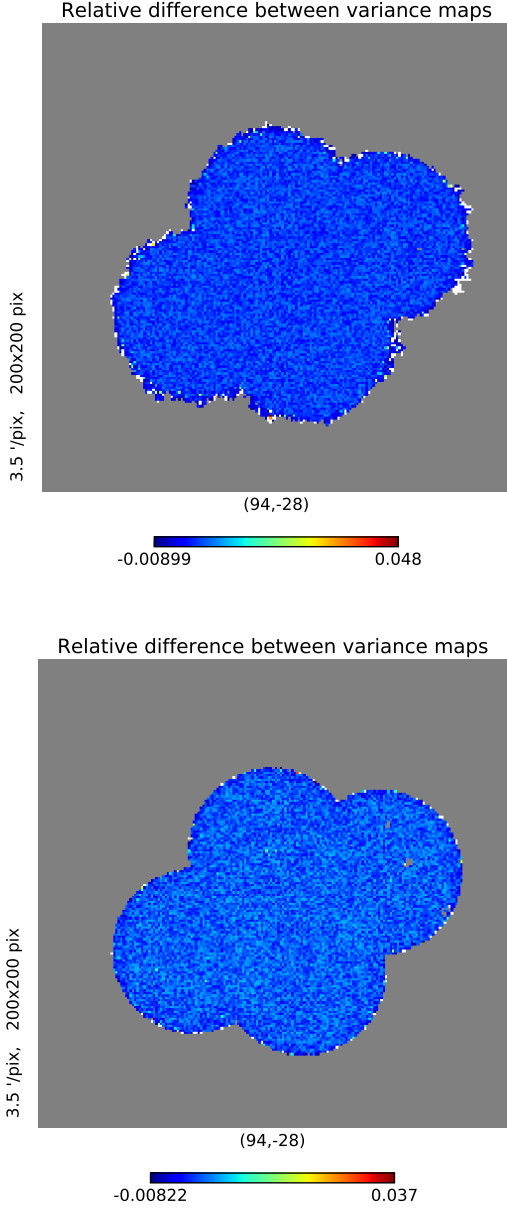


Figure 4. Relative difference between the depth calculated using the two methods presented in the text for the dithered (top) and undithered (bottom) fields.

PSF-like objects. Then in each pixel, we compute the median SNR as a function of the magnitude and get the magnitude at which SNR is the closest to 5.

These two procedures yield very similar results (within $\sim 4\%$) as it can be seen in Figure 4. We checked that the maps built selecting galaxies instead of stars are compatible as well. We will select our footprint according to the depth map using the second methodology which can be seen in Figure 5.

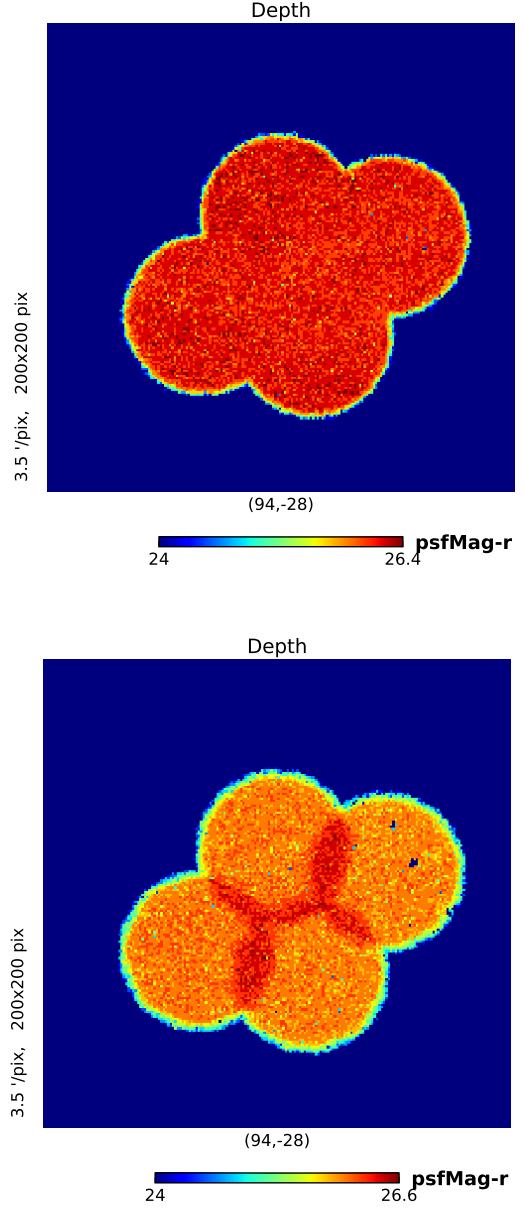


Figure 5. $5\text{-}\sigma$ depth for the dithered (top) and undithered (bottom) fields.

3.1 Bright objects masking and data selection

Bright objects produce significant effects in the image that affects the detection and measurement of neighboring objects. Some examples of the effect of these objects are: saturation of the CCD, large diffraction spikes, obscuration of neighboring sources, etc. Thus, masking a region around these sources makes a more complicated footprint but largely simplifies the analysis of systematic effects. In order to evaluate the effect of this sources, we use the stars from the input catalog, we divide them in different magnitude bins, and we count the detected objects in a given radius. However, the number of objects in a given radius is dependent of the number of stars in each bin and subject to statistical fluctuations. Therefore, we use instead the aver-

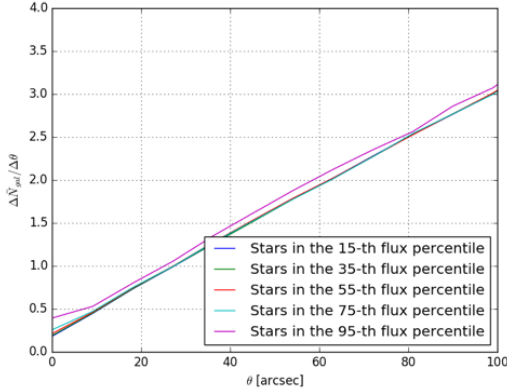


Figure 6. Mean increment in the number of detected objects around stars in the 95th flux percentile (magenta), the 75th percentile (cyan), the 55th percentile (red), the 35th percentile (green) and the 15th percentile (blue). We see that for the brightest percentiles there is no obscuration present but an overdensity of targets. These are mainly noise peaks identified as point sources.

age on this number for the different magnitude bins. This analysis is largely simplified if we use the numerical derivative of this quantity. In Figure 6 we can see that there is an excess of sources at distances lower than 10 arcseconds for the brightest stars. This is mainly due to the existence of fake sources around these bright objects. The brighter sources also have a larger noise, the DM deblender models the brightest source and subtracts this model looking for fainter sources that can be blended together with it. However, sometimes there are large noise peaks that have not been subtracted. These large noise peaks are detected in later stages as individual (point-like) sources giving as a result these fake sources. However, we do not see any obscuration present at the scales that we are interested (1 arcmin and higher) so it seems that no masking is required but we need to select the objects for our analysis carefully.

We checked the detection efficiency of galaxies in our sample. In order to do so we selected galaxies in the catalog using the variable `base_ClassificationExtendedness_value` as a star/galaxy separator as in ref. (Bosch et al. 2017). Objects where this variable is 1 are more likely to be galaxies, whereas the objects where it is 0 are more likely to be stars. We also made some quality cuts by selecting the objects with `detect_isPrimary==True`. This ensures that the object has been fully deblended and that the detection was not close to the edge of a coadded image. The results can be seen in Figure 7 where we computed the ratio of detected objects classified as galaxies and the input number of galaxies, ϵ as a function of magnitude. We checked this ratio for both the dithered and undithered catalogs and in pixels where the depths were higher than 25.7 and 26.25. We used `CMODEL_MAG` (see ref. (Bosch et al. 2017) for more details) as the reference magnitude for the detected objects and the true magnitude for the input galaxies.

Looking at Figure 7 it seems that selecting galaxies with `base_ClassificationExtendedness_value==1`, `detect_isPrimary==True` and `CMODEL_MAG > 21` is enough to get rid of most potential problems with fake detections.

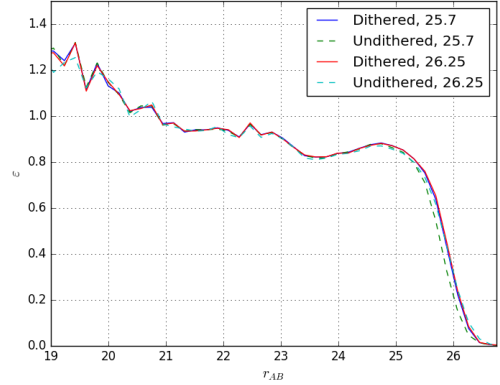


Figure 7. Number of detected objects classified as galaxies divided by the number input galaxies as a function of magnitude. We see that the detection efficiency is larger than 80% for the most part and a considerable fraction of objects with magnitude lower than 20 are being misclassified.

On top of that, it seems that selecting `CMODEL_MAG < 25.3` ensures good level of completeness ($\sim 80\%$). Uniformity is ensured by selecting the galaxies that lie within HEALPixels where the depth is greater or equal than 25.3. After these selection cuts we end up with 4.4 million galaxies for the dithered field and 4.3 for the undithered field.

4 CLUSTERING RESULTS

Add something as intro for this section. In this section we analyze the two point clustering statistics for both the dithered and undithered catalogs in real and harmonic space and check the consistency between the input and measured observables.

4.1 2-point correlation function

Using the samples presented in previous sections, we measure the 2-point correlation function using the package `TreeCorr` (Jarvis et al. 2004) with the Landy & Szalay estimator (Landy & Szalay 1993).

$$w(\theta) = \frac{DD - 2DR + RR}{RR} \quad (1)$$

where DD , DR , and RR are the number of pairs of objects taking from the data D or the random catalog R that covers the footprint. We use 15 angular log-spaced bins between $\theta = 0.05^\circ$ and $\theta = 10^\circ$. The number of bins has been chosen so that the resulting covariance matrix is nearly diagonal. The covariance matrices are calculated using the delete-one jackknife technique (Shao 1986; Norberg et al. 2009). We divide the footprint in $N_{JK} = 100$ regions. These regions are defined using the K-means algorithm from the package `kmeans_radecc`⁴. The covariance matrix is then computed as

⁴ https://github.com/esheldon/kmeans_radecc

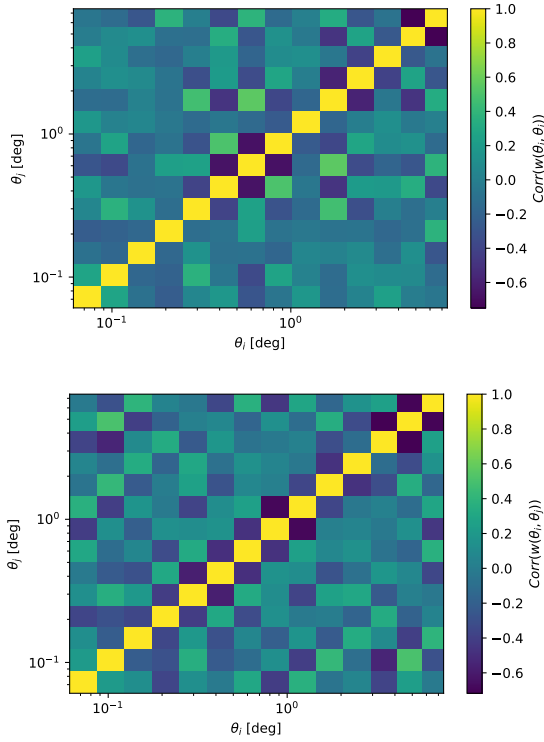


Figure 8. Correlation matrices for the dithered (top) and undithered catalogs (bottom).

$$\text{Cov}_{JK}(\theta_i, \theta_j) = \frac{N_{JK} - 1}{N_{JK}} \sum_{k=1}^{N_{JK}} \Delta w_k(\theta_i) \Delta w_k(\theta_j) \quad (2)$$

$$\Delta w_k(\theta_i) = w_k(\theta_i) - \bar{w}(\theta_i) \quad (3)$$

Where $w_k(\theta_i)$ is the value of the correlation function when deleting the k -th region at the scale θ_i , and $\bar{w}(\theta_i)$ is the average correlation function at that same scale. We compute the correlation function on the dithered and undithered catalogs using their respective footprints, and we compare with the truth catalog in the dithered footprint using the same magnitude cuts. The results can be seen at Figure 11.

We also show the correlation matrix,

$$\text{Corr}_{ij} = \frac{\text{Cov}_{ij}}{\sqrt{\text{Cov}_{ii} \text{Cov}_{jj}}} \quad (4)$$

in Figure 8 where we can see that in both cases the matrices are almost diagonal.

4.2 Angular power spectrum

We measure the angular power spectrum using the **NaMaster**⁵ package. This package computes the cross-power-spectra of masked fields with an arbitrary number of contaminants using a pseudo- C_ℓ approach (Hivon et al. 2002; Elsner et al. 2017). As we do for real space, we select the number of ℓ bins so that the covariance matrix is almost

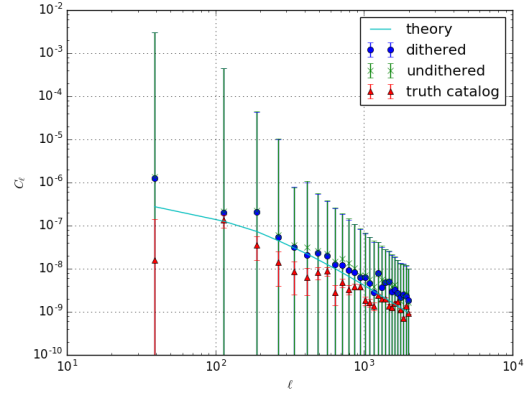


Figure 9. Measured power spectra for the input (red), undithered (green) and dithered (blue) datasets with **NaMaster**. The error bars are computed using the Gaussian approximation. All datasets seem to be compatible with the theoretical prediction (cyan) within $1\text{-}\sigma$.

diagonal. In this case we calculate the power spectrum in the range $0 < \ell < 2000$ and $\Delta\ell = 75$. We calculate the covariance matrices with two different approaches: One consists in the same delete-one jackknife technique performed in real space; in the other approach we compute the Gaussian covariances (Dodelson 2003; Cabré et al. 2007)

$$\text{Cov}_{\ell\ell'} = \frac{2}{f_{sky}} \left(\frac{C_\ell^2}{2\ell + 1} + \frac{1}{\bar{n}^2} \right) \delta_{\ell\ell'} \quad (5)$$

where \bar{n} is the number density (objects per steradian). We compute the theoretical prediction for the power-spectra with **CCL**⁶ and use them to calculate the covariance matrices given that the predictions are noise-free.

$$C_\ell^{TH} = \frac{2}{\pi} \int dz \left(\frac{dn(z)}{dz} \right)^2 b^2(z) \int dk k^2 P(k, z) j_\ell^2(kr(z)) \quad (6)$$

Where $P(k, z)$ is the power spectrum, $b(z)$ is the bias and $\frac{dn}{dz}$ is the number density as a function of redshift. We use the Millenium cosmological parameters (Springel et al. 2005) ($\Omega_m = 0.25, \Omega_b = 0.045, \Omega_\Lambda = 0.75, n = 1, \sigma_8 = 0.9, h = 0.73$), and the $\frac{dn}{dz}$ built with the input catalog using the same magnitude cuts. The results can be seen in Figure 9. It seems that there is a small bias between the measured power-spectra and the input data power-spectrum (and between the dithered and undithered data), however, the error bars in the data are large enough to make this bias negligible.

4.3 Systematic effects

In this section we analyze the different systematic effects affecting the DC1 data. We will consider the following observational quantities as sources for systematic uncertainties:

- **Extinction:** The CatSim catalog provides the value for the magnitudes already corrected for extinction using the SFD map (Schlegel et al. 1998). We use this map to cross-correlate with our galaxy catalogs.

⁵ <https://github.com/damonge/NaMaster>

⁶ <https://github.com/LSSTDESC/CCL>

- **Stellar contamination:** In this case we just build a HEALPix map using the input CatSim stellar catalog to cross-correlate with our galaxy catalogs.

- **Sky-background:** We use the observed background level in each exposure and assign that value to the HEALPix with $N_{side} = 2048$ that corresponds to the pointing position. After this we calculate the median value in each HEALPix and use them to compute the cross-correlations with the galaxy catalogs. The caveat of this approach is that we are not propagating the geometry of the focal plane.

- **Sky-noise:** We use the observed noise background level in each exposure and proceed as in the previous case to build a HEALPix map with which we will cross-correlate.

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

These maps are shown in Figure A1 and Figure A2.

In the case of the real space measurements, we proceed the same way as ref. (Crocce et al. 2016) to compute the impact of the different potential sources for systematic uncertainty. We compute the auto and cross-correlations of these maps and our data samples to obtain the “true” correlation function and since the cross-correlations between different systematic maps are negligible (*i.e.* they are independent) we can write,

$$w(\theta)_{true} = w(\theta)_{obs} - \sum_{sys} \frac{w_{sys,cross}^2}{w_{sys,auto}} \quad (7)$$

where we sum over all relevant systematics, $w_{sys,cross}$ is the cross-correlation between the considered map and the observed data, and $w_{sys,auto}$ is the auto-correlation of the considered maps. For the stellar contamination we also follow the procedure presented in ref. (Crocce et al. 2016). Given a stellar fraction f_{star} the “true” galaxy correlation function is given by

$$w_{gal} = (1 + f_{star})^2 \left(w_{obs} - f_{star}^2 w_{sg} - \frac{f_{star}^4}{(1 + f_{star})^2} \right) \quad (8)$$

where w_{sg} is the cross-correlation between our stellar map and the observed galaxies.

Comparing with the input catalog we select those objects classified as galaxies whose centroids lie within 2 pixels of a star from the input catalog. From those, we select the objects that have a magnitude difference smaller than 30 mmags. Doing this we estimate a stellar contamination of $f_{star} = 7.2\%$ in our sample.

We compare the correction term for the different systematics to the measured signal obtaining the results in Figure 10, where we can appreciate that the dithering strategy is working reducing the impact of the systematics, keeping them under control and at percent level for this “small” area, improving the signal-to-noise in addition to other benefits explored in ref. (Awan et al. 2016). This is also seen Figure 11 where we clearly see that the uncorrected and corrected correlation function for the dithered field are essentially the same.

4.4 Pushing the dataset further

5 CONCLUSIONS

In this paper we presented the methodology to characterize the transfer function of LSST. This methodology can be

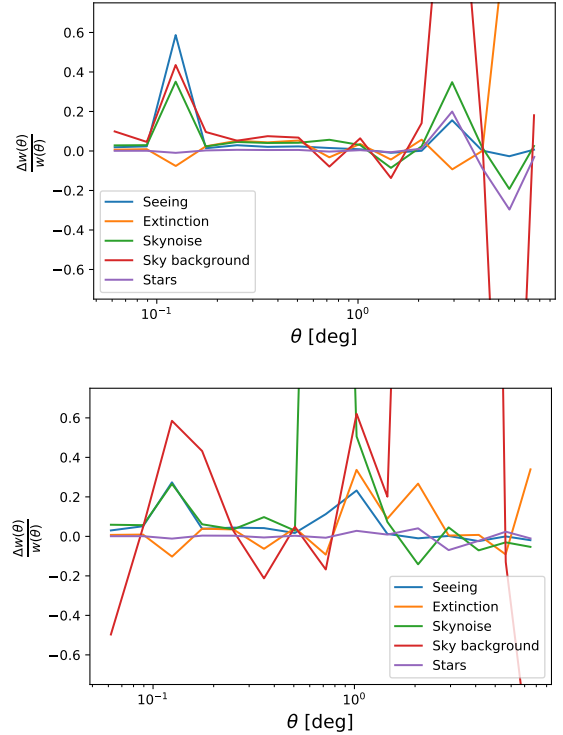


Figure 10. Correction due to the different potential sources of systematic uncertainty relative to the value of the measured correlation function of the dithered (top) and undithered (bottom) datasets.

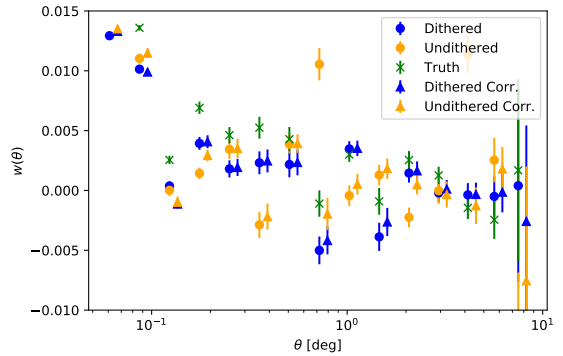


Figure 11. Results for the two-point correlation function in the input (green), dithered (blue) and undithered (orange) datasets. We can see that after the correction (triangles) for the systematic effects the agreement is better than without the correction (circles) between the input and output data especially for the undithered case (for the dithered case the corrections are negligible).

easily exported to different photometric surveys. First, we performed a quality assessment of the data produced by the pipeline. Then, we generated the mask to perform our clustering analyses. Finally we compared the input and output two point clustering statistics to characterize the transfer function. **Add more things when we have more results.**

Acknowledgments

Here is where you should add your specific acknowledgments, remembering that some standard thanks will be added via the `acknowledgments.tex` file.

This is the text imported from `acknowledgments.tex`, and will be replaced by some standard LSST DESC boilerplate at some point.

REFERENCES

- Albrecht, A., Bernstein, G., Cahn, R., et al. 2006, ArXiv Astrophysics e-prints, astro-ph/0609591
- Awan, H., Gawiser, E., Kurczynski, P., et al. 2016, ApJ, 829, 50
- Bosch, J., Armstrong, R., Bickerton, S., et al. 2017, ArXiv e-prints, arXiv:1705.06766
- Cabr , A., Fosalba, P., Gazta aga, E., & Manera, M. 2007, MNRAS, 381, 1347
- Crocce, M., Carretero, J., Bauer, A. H., et al. 2016, MNRAS, 455, 4301
- De Lucia, G., Springel, V., White, S. D. M., Croton, D., & Kauffmann, G. 2006, MNRAS, 366, 499
- Dodelson, S. 2003, Modern cosmology (San Diego, CA: Academic Press)
- Elsner, F., Leistedt, B., & Peiris, H. V. 2017, MNRAS, 465, 1847
- G rski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
- Hivon, E., G rski, K. M., Netterfield, C. B., et al. 2002, ApJ, 567, 2
- Ivezic, Z., Tyson, J. A., et al. 2008, ArXiv e-prints, arXiv:0805.2366
- Jarvis, M., Bernstein, G., & Jain, B. 2004, MNRAS, 352, 338
- Juri , M., Ivezi ,  ., Brooks, A., et al. 2008, ApJ, 673, 864
- Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64
- LSST Dark Energy Science Collaboration. 2012, ArXiv e-prints, arXiv:1211.0310
- LSST Science Collaboration. 2009, ArXiv e-prints, arXiv:0912.0201
- Norberg, P., Baugh, C. M., Gazta aga, E., & Croton, D. J. 2009, MNRAS, 396, 19
- Rowe, B. T. P., Jarvis, M., Mandelbaum, R., et al. 2015, Astronomy and Computing, 10, 121
- Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
- Shao, J. 1986, Annals of Statistics, 14, 1322
- Springel, V., White, S. D. M., Jenkins, A., et al. 2005, Nature, 435, 629 EP

APPENDIX A: MAPPING OBSERVATIONAL EFFECTS

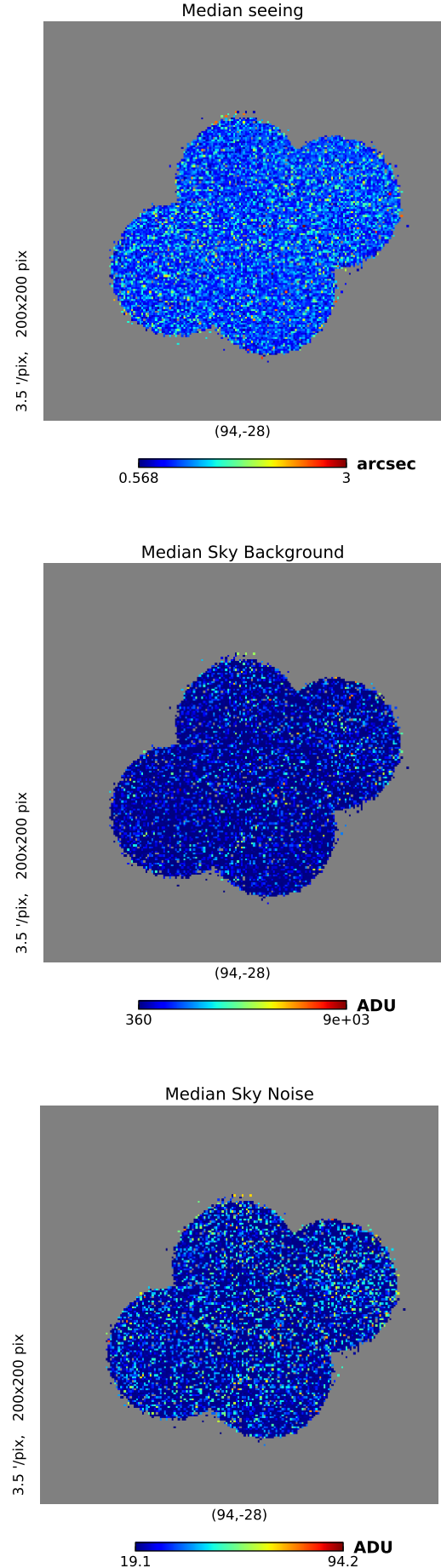


Figure A1. HEALPix maps showing the different observational effects that might be potential cause of systematic uncertainties.

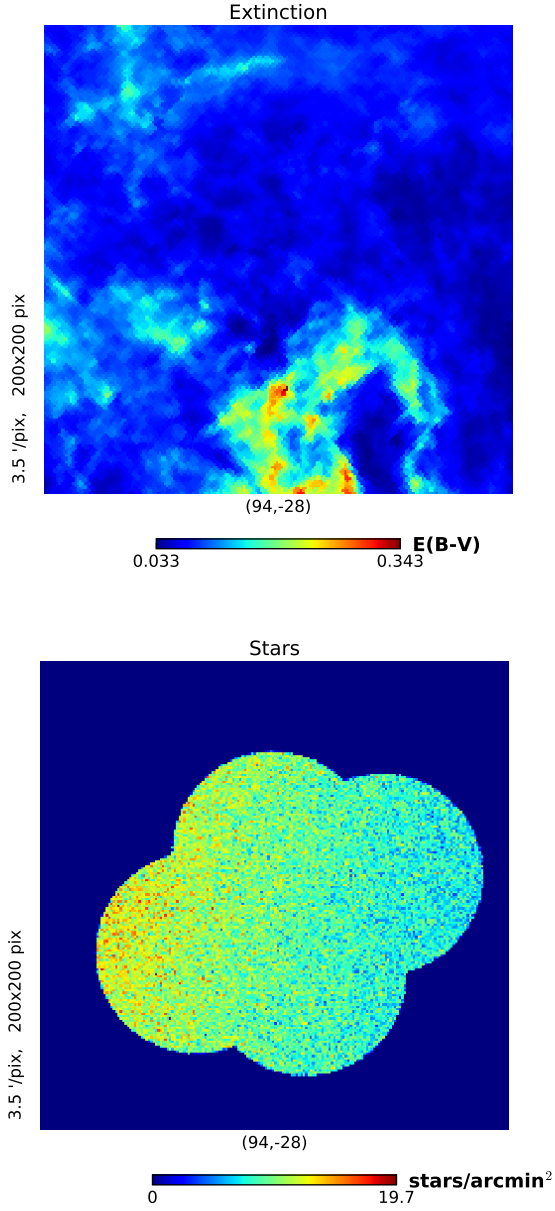


Figure A2. HEALPix maps showing the different observational effects that might be potential cause of systematic uncertainties.