Draft version August 27, 2013
Preprint typeset using LaTeX style emulateapj v. 04/17/13

THE THIRD GRAVITATIONAL LENSING ACCURACY TESTING (GREAT3) CHALLENGE HANDBOOK

Rachel Mandelbaum[1], Barnaby Rowe[2,3], James Bosch[4], Chihway Chang[5], Frederic Courbin[6], Mandeep Gill[5], Mike Jarvis[7], Arun Kannawadi[1], Tomasz Kacprzak[2], Claire Lackner[8], Alexie Leauthaud[8], Hironao Miyatake[4], Reiko Nakajima[9], Jason Rhodes[3,10], Melanie Simet[1], Joe Zuntz[11], Bob Armstrong[7], Sarah Bridle[11], Jean Coupon[12], Jörg P. Dietrich[13,14], Marc Gentile[6], Catherine Heymans[15], Alden S. Jurling[16,17], Stephen M. Kent[18], David Kirkby[19], Daniel Margala[19], Richard Massey[20], Peter Melchior[21], John Peterson[22], Aaron Roodman[5], & Tim Schrabback[9]

Draft version August 27, 2013

ABSTRACT

The GRavitational lEnsing Accuracy Testing 3 (GREAT3) challenge is the third in a series of image analysis challenges, with a goal of testing and facilitating the development of methods for analyzing astronomical images that will be used to measure weak gravitational lensing. This measurement requires extremely precise estimation of very small galaxy shape distortions, in the presence of far larger intrinsic galaxy shapes and distortions due to the blurring kernel caused by the atmosphere, telescope optics, and instrumental effects. The GREAT3 challenge is posed to the astronomy, machine learning, and statistics communities, and includes tests of three specific effects that are of immediate relevance to upcoming weak lensing surveys, two of which have never been tested in a community challenge before. These effects include realistically complex galaxy models based on high-resolution imaging from space; spatially varying blurring kernel; and combination of multiple different exposures. To facilitate entry by people new to the field, and for use as a diagnostic tool, the simulation software for the challenge is publicly available, though the exact parameters used for the challenge are blinded. Sample scripts to analyze the challenge data using existing methods will also be provided. See http://great3challenge.info and http://great3.projects.phys.ucl.ac.uk/leaderboard/ for more information.

Subject headings: gravitational lensing: weak, methods: data analysis, methods: statistical, techniques: image processing

1. INTRODUCTION

In our currently accepted cosmological model, the baryonic matter from which stars and planets are made accounts for only 4% of the energy density of the Universe. In order to explain many cosmological observations, we have been forced to posit the existence of dark matter (which we detect through its gravitational attraction) and dark energy (which causes a repulsion that is driving the accelerated expansion of the Universe, the discovery of which led to the 2011 Nobel Prize in Physics). While we infer the existence of these dark components, the question of what they actually are remains a mystery.

Gravitational lensing is the deflection of light from distant objects by all matter along its path, including dark matter (Fig. 1). Lensing measurements are thus directly sensitive to dark matter. They also permit us to infer the properties of dark energy (Hu 2002), because the accelerated expansion of the Universe that it causes directly

[1] McWilliams Center for Cosmology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
[2] Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK
[3] Jet Propulsion Laboratory, California Institute of Technology, MS 300315, 4800 Oak Grove Drive, Pasadena, CA
[4] Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA
[5] KIPAC, Stanford University, 452 Lomita Mall, Stanford, CA 94309, USA
[6] Laboratoire d'astrophysique, Ecole Polytechnique Fédérale de Lausanne (EPFL), Observatoire de Sauverny, CH-1290 Versoix, Switzerland
[7] Department of Physics and Astronomy, University of Pennsylvania, 209 South 33rd Street, Philadelphia, PA 19104, USA
[8] Kavli Institute for the Physics and Mathematics of the Universe (WPI), Todai Institutes for Advanced Study, the University of Tokyo, Kashiwa, Japan
[9] Argelander-Institut für Astronomie, Universitat Bonn, D-53121 Bonn, Germany
[10] California Institute of Technology, MC 350-17, 1200 East California Boulevard, Pasadena, CA 91125, USA
[11] Jodrell Bank Centre for Astrophysics, School of Physics and Astronomy, The University of Manchester, Manchester M13 9PL, UK
[12] Institute of Astronomy and Astrophysics, Academia Sinica, P.O. Box 23-141, Taipei 10617, Taiwan
[13] Universitäts-Sternwarte München, Scheinerstr. 1, 81679 München, Germany
[14] Excellence Cluster Universe, 85748 Garching b. München, Germany
[15] Scottish Universities Physics Alliance, Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ, UK
[16] NASA Goddard Space Flight Center, Greenbelt, MD 20771
[17] Department of Physics and Astronomy, University of Rochester, Rochester NY, 14618, USA
[18] Center for Particle Astrophysics, Fermi National Accelerator Laboratory, Batavia, IL 60510, USA; operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.
[19] Department of Physics and Astronomy, University of California, Irvine, CA 92697, USA
[20] Institute for Computational Cosmology, Durham University, South Road, Durham DH1 3LE, UK
[21] Center for Cosmology and Astro-Particle Physics & Department of Physics, The Ohio State University, Columbus, OH, 43210, USA
[22] Department of Physics, Purdue University, West Lafayette, IN 47907, USA

opposes the effects of gravity (which tends to cause matter to clump into ever larger structures) and influences light propagation through its impact on the geometry of the universe.

This measurement entails detecting small but spatially coherent distortions (known as weak shears) in the shapes of distant galaxies, which provide a statistical map of large-scale cosmological structures. Weak lensing measurements have already placed some constraint on the growth of structure, typically with 10% statistical errors, or as small as 5% for the most recent analyses (Heymans et al. 2013; Jee et al. 2013). Because of the sensitivity of weak lensing to the dark components of the Universe, the astronomical community has designed upcoming surveys to measure it very precisely, and thereby constrain cosmological parameters. In addition to several experiments beginning in 2013, there are even larger experiments that are planned to start at the end of this decade. In the Astro2010 Decadal Survey of US astronomy (Blandford et al. 2010), the most highly endorsed large experiments both from the ground (the Large Synoptic Survey Telescope, or LSST) and space (the Wide-Field Infrared Survey Telescope, or WFIRST-AFTA) are ones with a significant emphasis on weak lensing cosmology. The European Space Agency recently decided that of several possible large space-based astronomical surveys, they will proceed with the Euclid mission, which likewise has a major emphasis on lensing.

However, the increasing size of these experiments, and the decreasing statistical errors, comes with a price: to fully realize their promise, we must understand systematic errors increasingly well. The coherent lensing distortions of galaxy shapes are typically $\sim 1\%$ in size, far smaller than galaxy intrinsic ellipticities (~ 0.3) and, more problematically, smaller than the coherent distortions due to light propagation through the atmosphere and telescope optics (the point-spread function, or PSF). Removing the effects of the PSF and measuring lensing shears for galaxies that are only moderately resolved and have limited signal-to-noise is a demanding statistical problem that has not been solved adequately for upcoming surveys. Systematic errors related to shape measurement must be reduced by factors of 5-10 in the next decade. The weak lensing community has gained substantially from a practice of carrying out blind challenges in order to test shear measurement methods. By using simulated data in which the ground truth is known, but with realistically complicated galaxies and PSFs, we can estimate the systematic errors associated with current methods, and use our new knowledge of their failure modes to spur further development in the field.

1.1. *Previous challenges and the context for GREAT3*

The current and past GREAT challenges have all been supported by the PASCAL network. The GREAT08 Challenge (Bridle et al. 2009, 2010) set a highly simplified version of the problem, using known PSFs, simple galaxy models, and a constant applied gravitational shear. The GREAT10 Challenge (Kitching et al. 2011, 2012, 2013) increased the realism and complexity of its simulations over GREAT08 by using cosmologically-varying shear fields and greater variation in galaxy model parameters and telescope observing conditions. Since imperfect knowledge of the PSF can also bias shear measurements, GREAT10 tested PSF modelling in a standalone Star Challenge. GREAT08 and GREAT10 were preceded by a number of internal challenges within the astrophysics community, known as the Shear Testing Programme, or STEP (Heymans et al. 2006; Massey et al. 2007b), which demonstrated the highly non-trivial nature of the shear measurement problem. Both GREAT08 and GREAT10 generated significant (factors of 2-3) improvement in the accuracy of weak lensing shape measurement, while also providing a greater understanding of the major limitations of existing methods.

The key goals of the GREAT3 challenge are to facilitate further work in understanding existing methods of PSF correction, to suggest ways that they can be developed and improved in the future, and to spur the creation of new methods that solve the limitations of existing ones. We aim to address the challenges in this field in two ways: (1) We provide a suite of simulated galaxy images for making controlled tests of outstanding issues in lensing shear measurement, focusing on crucial issues not addressed in previous challenges and adding new levels of realism; and (2) we provide the simulation code, GalSim (Rowe et al., *in prep.*), as a fully documented, open source (licensed under the GNU General Public License, or GPL) development toolkit in a modern language framework (object-oriented Python wrapping around C++, §5). GalSim is already public[23], but the exact challenge input parameters are blinded. We anticipate that the open source status of this simulation code will facilitate more rapid improvement of existing methods and development of new ones. Using real galaxy images (from the *Hubble Space Telescope*) and detailed, physically motivated PSF models as inputs will ensure that the space of possible simulations is overwhelmingly large, as in reality. This development will be crucial for weak lensing to achieve its unique potential for understanding the nature of dark energy and matter in the Universe. In this way we accurately reflect the real problem of galaxy shape measurement, which combines well understood gravitational physics with input galaxies and observing conditions about which we have significantly less information.

As for previous lensing challenges, the simulations are statistically matched to the size of the largest upcoming weak lensing surveys such as Euclid[24] (Laureijs et al. 2011), LSST[25] (LSST Science Collaborations & LSST Project 2009), and WFIRST-AFTA[26] (Spergel et al. 2013). The goal of participants is to measure gravitational shears sufficiently precisely that systematic errors in the measurements are below the statistical errors, so that the error budget is not dominated by systematics. The challenge is split into branches that reflect different issues in the field and types of observations; participants may enter as many or as few branches as they wish.

An overview of how to use this handbook for the GREAT3 Challenge is as follows. §2 presents an overview of the physics behind gravitational lensing and astronomical imaging, to motivate the major issues in shear estimation. To estimate the gravitational shear in the galaxy

[23] https://github.com/GalSim-developers/GalSim
[24] http://sci.esa.int/euclid
[25] http://www.lsst.org/lsst/
[26] http://wfirst.gsfc.nasa.gov

image, the standard procedure is to measure each galaxy shape and infer the overall shear from these; an overview of existing approaches to shape measurement is in Appendix A. §3 contains a summary of the specific issues that the GREAT3 Challenge is designed to address. The structure of the Challenge and how it is run is detailed in §4. In §5 we describe the simulation generation and design, and relate it to the issues from §3. Finally, we summarize the simplifications of the GREAT3 challenge in §6.

2. PHYSICS BACKGROUND

Here we describe the basic physics behind gravitational lensing and astronomical imaging. The processes described in this section are shown in Fig. 1.

2.1. *Lensing shear*

Gravitational lensing distorts observed images of distant galaxies, in a way that depends on the distribution of mass around the line of sight. This distortion can be described as a general coordinate transformation, but for the overwhelming majority of distant galaxy light sources, the transformation is well approximated as being locally linear. This limit is known as weak gravitational lensing.

Weak gravitational lensing can be described as a linear transformation between unlensed coordinates (x_u, y_u; with the origin at the center of the distant light source) and the lensed coordinates in which we observe galaxies (x_l, y_l; with the origin at the center of the observed image),

$$\begin{pmatrix} x_u \\ y_u \end{pmatrix} = \begin{pmatrix} 1-\gamma_1-\kappa & -\gamma_2 \\ -\gamma_2 & 1+\gamma_1-\kappa \end{pmatrix} \begin{pmatrix} x_l \\ y_l \end{pmatrix}. \quad (1)$$

Here we have introduced the two components of the complex-valued lensing shear $\gamma = \gamma_1 + \mathrm{i}\gamma_2$, and the lensing convergence κ. The shear describes the *stretching* of galaxy images due to lensing. The convention is such that a positive (negative) γ_1 results in an image being stretched along the x (y) axis direction. Likewise a positive (negative) γ_2 results in an image being stretched along the line $y = x$ ($y = -x$). The convergence κ describes a change in apparent size for lensed objects: areas of the sky for which κ is positive have apparent changes in area (at fixed surface brightness) that make lensed images appear larger and brighter than if they were unlensed, and a modified galaxy density.

Often, as we do not know the distribution of sizes of distant galaxies well, it is common to recast the transformation (1) as

$$\begin{pmatrix} x_u \\ y_u \end{pmatrix} = (1-\kappa) \begin{pmatrix} 1-g_1 & -g_2 \\ -g_2 & 1+g_1 \end{pmatrix} \begin{pmatrix} x_l \\ y_l \end{pmatrix}, \quad (2)$$

in terms of the *reduced* shear, $g_1 = \gamma_1/(1-\kappa)$ and $g_2 = \gamma_2/(1-\kappa)$. In many applications the $(1-\kappa)$ term is not estimated from the data (although see, e.g., Casaponsa et al. 2013), and so it is the image stretching described by the reduced shear that is in fact observed in galaxies (hence the use of this notation in Bridle et al. 2009). We often encode the two components of shear and reduced shear into a single complex number, e.g. $\gamma = \gamma_1 + \mathrm{i}\gamma_2$, $g = g_1 + \mathrm{i}g_2$. In most cosmological applications $g \simeq \gamma$ is a reasonable approximation; however, the GREAT3 simulations with cosmologically varying shear fields do also contain a corresponding κ variation.

The lensing shear causes a change in estimates of the *ellipticity* of distant galaxies. If sources with intrinsically circular isophotes (contours of equal brightness) could be identified, the observed sources (post-lensing) would have elliptical isophotes that we can characterize by their minor-to-major axis ratio b/a and the orientation of the major axis ϕ. For $|g| < 1$, these directly yield a value of the reduced shear

$$|g| = \frac{1-b/a}{1+b/a} \quad (3)$$

which, combined with the orientation ϕ, gives the two orthogonal components of shear $g_1 = |g|\cos 2\phi$, $g_2 = |g|\sin 2\phi$.

In practice we cannot identify distant galaxy sources that are circular prior to lensing, nor do distant galaxies have elliptical isophotes. However, it is possible to estimate properties that transform in similar ways to the simplified case presented above, and from which we can extract statistical estimates of shear. One method is to model the light from galaxies using a profile that does have a well-defined ellipticity. We can write this ellipticity as a complex number $\varepsilon = \varepsilon_1 + \mathrm{i}\varepsilon_2$, with magnitude $|\varepsilon| = (1-b/a)/(1+b/a)$ and orientation angle determined by the direction of the major elliptical axis. Under an applied shear with $|g| \leq 1$, this definition of ellipticity transforms as

$$\varepsilon = \frac{\varepsilon^{(s)} + g}{1 + g^*\varepsilon^{(s)}} \quad (4)$$

(see Bartelmann & Schneider 2001 for the strong shear $|g| > 1$ result). Here we have labelled the ellipticity of the source prior to lensing as $\varepsilon^{(s)}$. For $g \ll 1$, eq. (4) becomes $\varepsilon \simeq \varepsilon^{(s)} + g$. For a population of source ellipticities that are randomly oriented so that $\langle \varepsilon^{(s)} \rangle = 0$, the ensemble average ellipticity after lensing gives an unbiased estimate of the shear: $\langle \varepsilon \rangle \simeq g$.

Another common choice of shape parametrization is based on second brightness moments across the galaxy image,

$$Q_{ij} = \frac{\int \mathrm{d}^2x I(\mathbf{x}) W(\mathbf{x}) x_i x_j}{\int \mathrm{d}^2x I(\mathbf{x}) W(\mathbf{x})}, \quad (5)$$

where the coordinates x_1 and x_2 correspond to the x and y directions respectively, $I(\mathbf{x})$ denotes the galaxy image light profile, $W(\mathbf{x})$ is an optional[27] weighting function (see Schneider 2006), and where the coordinate origin $\mathbf{x} = 0$ is placed at the galaxy image center (commonly called the centroid). A second definition of ellipticity, sometimes referred to as the *distortion* to distinguish it from the ellipticity that satisfies equation (4), can then be written as

$$e = e_1 + \mathrm{i}e_2 = \frac{Q_{11} - Q_{22} + 2\mathrm{i}Q_{12}}{Q_{11} + Q_{22}}. \quad (6)$$

[27] Optional for the purpose of this definition; but in practice, for images with noise, some weight function that reduces the contribution from the wings of the galaxy is necessary to avoid moments being dominated by noise.

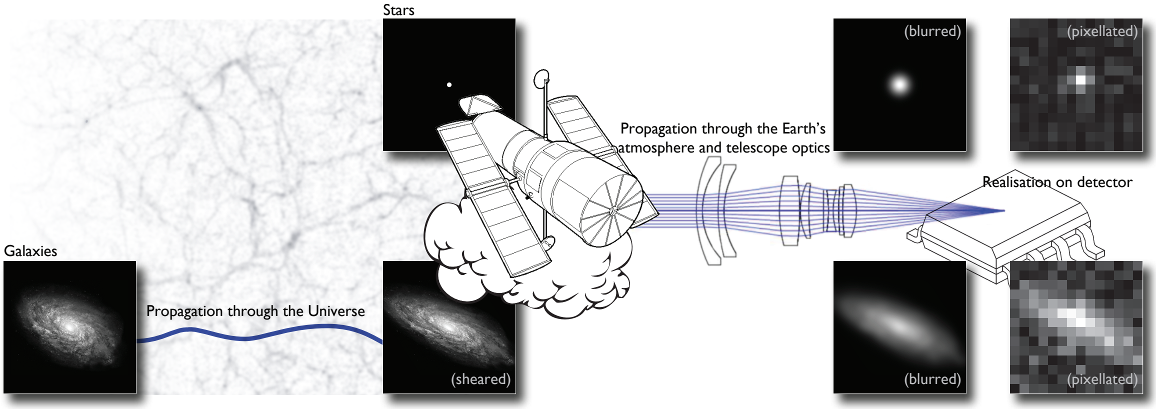
Stars
Galaxies
Propagation through the Universe
(sheared)
Propagation through the Earth's atmosphere and telescope optics
(blurred)
(pixellated)
Realisation on detector
(blurred)
(pixellated)

FIG. 1.— An illustration of the process of gravitational lensing and other effects that change the apparent shapes of galaxies in the astronomical imaging process. (Based on Figure 8 from Kitching et al. 2011).

The ellipticity ε can also be related to the moments, like the distortion, but replacing the denominator in Eq. 6 with $Q_{11} + Q_{22} + 2(Q_{11}Q_{22} - Q_{12}^2)^{1/2}$.

If the weighting function $W = 1$ (unweighted moments) or $W = W[I(\mathbf{x})]$ (a brightness-dependent weight function) then an image with elliptical isophotes of axis ratio b/a has

$$|e| = \frac{1 - b^2/a^2}{1 + b^2/a^2}. \tag{7}$$

Under a shear, e transforms from a source (pre-lensing) distortion $e^{(s)}$ as

$$e = \frac{e^{(s)} + 2g + g^2 e^{(s)*}}{1 + |g|^2 + 2\Re[g e^{(s)*}]}, \tag{8}$$

so that in the weak shear limit, $e \simeq e^{(s)} + 2[1 - (e^{(s)})^2]g$. For a population of source distortions that are randomly oriented so that $\langle e^{(s)} \rangle = 0$, the ensemble average e after lensing gives an unbiased estimate of approximately twice the shear that depends on the population root mean square (RMS) ellipticity, $\langle e \rangle \simeq 2[1 - \langle (e^{(s)})^2 \rangle]g$.

2.2. *Shear fields*

Although gravitational lensing distortions at the locations of individual galaxies can typically be approximated as linear, the shear and convergence vary with position across the sky. This variation is due to the non-uniform distribution of massive structures in the universe. Estimates of this variation, which are discrete estimates of the underlying shear field, are used in various ways to improve our models of the universe. In the following Section we will primarily focus on shear fields[28].

One well-motivated place to look for shears is around structures that can be directly observed, for example around galaxies likely to lie in dark matter-rich regions (a study known as *galaxy-galaxy lensing*), or around foreground galaxy clusters. Around a central lens object, the tangential shear induced by gravitational lensing is approximately constant at fixed radius; thus measuring constant shear is a goal of galaxy-galaxy and cluster-galaxy lensing measurements. Estimates of the shear around such objects have been compared to parametric models of the matter content to provide great insight into the way that visible matter is traced by underlying mass.

The statistics of lensing shear, and its spatial correlation as a function of angular scale on the sky, is another key prediction of many models of the universe, and this is therefore the other goal of our measurement. Comparison of these models to observations is expected to greatly improve our understanding of dark matter and dark energy, and to constrain models of modified gravity that have been posed as another way to explain the accelerated expansion of the Universe (e.g., Peacock et al. 2006; Albrecht et al. 2009). Fig. 2 shows a typical cosmological shear field in a 10×10 deg^2 region (the size of GREAT3 images).

A map of galaxy shears is a spin-2 (headless vector) field. As such, it can be decomposed into two components, called the E-mode and B-mode by analogy with electric and magnetic fields. The E-mode can be represented as the gradient of a scalar field, and has no curl; the B-mode can be represented as the curl of a vector field, and has no divergence.

Cosmological shear fields include almost exclusively E mode signals - with tangential shears induced around density peaks - and only negligible B modes[29], a fact that is often exploited in reality to test for systematic errors (observed B modes are taken as a sign of systematic error).

2.3. *Measuring shear fields*

The shear fields around galaxies and galaxy clusters are generally used to constrain either parametric models of the unseen mass distribution around these objects (e.g., Leauthaud et al. 2012; Velander et al. 2013), or a non-parametric map of the same (e.g., Dietrich et al. 2012; Van Waerbeke et al. 2013). When estimating shears around astrophysical objects for the purpose of fitting a parametric model, it is common to estimate the average shear in annuli of separation from the center of the foreground object. For more general map making, shear

[28] See e.g. Huff & Graves (2011); Schmidt et al. (2012), & Casaponsa et al. (2013) (and references therein) for recent developments in lensing magnification.

[29] Some B-mode contributions can occur due to multiple lens planes, source clustering, and other effects, all of which are quite small (e.g., Schneider et al. 2002; Hirata & Seljak 2004; Vale et al. 2004).

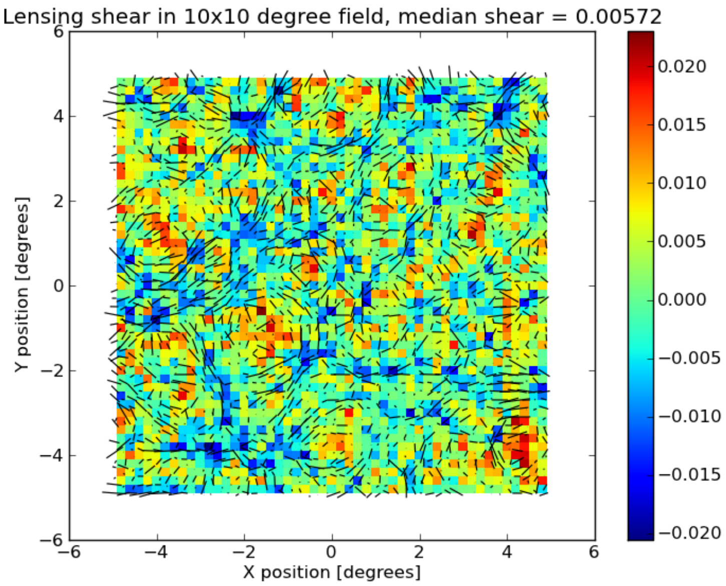

FIG. 2.— An example of a cosmological shear field in a 10 × 10 deg^2 region of sky, with the same statistical properties as realistic cosmological shear fields. At each point on the grid, the size of the arrow shows the magnitude of the shear (for reference, the plot title gives the median shear value), the orientation shows the shear direction at that location, and the color shows the convergence κ. As shown, the shear field exhibits coherent alignments over large scales, with tangential orientation around mass overdensities.

estimates from individual galaxies are typically averaged in cells across the sky, and the smoothed shears are then used to estimate the projected density.

However, another important application of weak lensing shear estimates is to probe the statistical properties of the shear field, as a function of angular separation on the sky. Different models of the universe predict differing statistical distributions of shear as a function of angular scale. Recent estimates of the spatial correlations between shears, and the evolution of these correlations over cosmic time (Jee et al. 2013; Heymans et al. 2013), used catalogs of shear estimates across a whole survey area as a probe of the growth of matter structure. In these 'cosmic shear' analyses, the shear correlation function estimated from catalogs of shears (see Appendix B) is the data vector for the estimation of cosmological parameters.

The shear power spectrum, related to the correlation function by a Fourier transform, is also a quantity of interest for describing the statistics of cosmic shear. In the GREAT10 challenge, the goal was to estimate the power spectrum directly from the shear using the discrete Fourier transform due to the galaxies being positioned at fixed grid locations. However, in practice the use of power spectra presents challenges because of the non-regular spacing of galaxies on the sky, and the presence of holes in coverage due to bright foreground objects or camera defects. In GREAT3, we adopt a correlation-function based metric for the simulations containing variable shear. As a useful side product, this also allows us to sensitively probe contamination of the shear field due to both variable and constant point-spread functions (PSFs). In the following Section we discuss the importance of the PSF in weak lensing science.

2.4. *Variable PSF*

After the light from distant galaxies is sheared, it passes through the turbulent atmosphere (when observing with a ground-based telescope), and through telescope optics and a detector. While the initial shear is the desired signal, these later effects (which can typically be modeled as convolution with a blurring kernel called the PSF) systematically modify the images. The blurring due to the atmosphere is typically larger than that due to optics, and varies relatively rapidly in time compared to typical exposure times for astronomical imaging (e.g., Heymans et al. 2012; Hamana et al. 2013). In contrast, the PSF due to the optics varies relatively slowly with time. The optical PSF is commonly described as a combination of diffraction plus aberrations (possibly up to quite high order). Both the atmospheric and optical PSF have some spatial coherence, qualitatively like lensing shear, though the scaling with separation is not identical.

The effect of the PSF on the galaxy shapes that we wish to measure is twofold: first, applying a roughly circular blurring kernel tends to dilute the galaxy shapes, making them appear rounder by an amount that depends on the ratio of galaxy and PSF sizes. Correction for this dilution can easily be a factor of 2 for typical galaxies, for which we wish to measure shears to 1%. Second, the small but coherent PSF anisotropies can leak into the galaxy shapes if not removed, mimicking a lensing signal.

Stars in the images are effectively point sources before blurring by the PSF, and hence are measures of the PSF. However, the PSF must be estimated from them and then interpolated to the positions of galaxies. For a summary of some common methods of PSF estimation and interpolation, see Kitching et al. (2013).

2.5. *Summary of effects*

Fig. 1 summarizes the main effects that go into a weak lensing observation. The galaxy image is distorted as it is deflected by mass along the line-of-sight from the galaxy to us. This is the desired signal. It is then further distorted by the atmosphere (for a ground-based telescope), telescope optics, and pixelation on the detector; these effects collectively form the PSF and are equivalent to convolution[30] with a blurring kernel. The images have noise, which can cause a bias when solving the non-linear problem of estimating the original shear, and there are also detector effects (not shown here). Given that upcoming datasets will have hundreds of million or billions of galaxies, removing these nuisance effects to sub-percent precision is a necessary but formidable challenge.

3. IMPORTANT ISSUES IN THE FIELD

The goal of this challenge is to address three major open issues in the field of weak lensing, as determined by a consensus among the community. These could conceivably be limiting systematic errors for weak lensing surveys beginning this year, but their importance has not been systematically quantified in a community challenge. In the interest of making a fair test of these issues, we exclude other issues that were deemed to be of lesser importance for now (§6). The GREAT3 challenge consists of experiments that can test each of the issues below separately, so that people who are interested in only certain issues can still participate.

3.1. *Realistic galaxy morphologies*

Multiple studies have shown that no method of shape measurement based on second moments can be completely independent of the details of the galaxy population (e.g., morphology and substructure), because the shear couples the second moments to the higher-order moments (Massey et al. 2007a; Bernstein 2010; Zhang & Komatsu 2011). This issue is particularly pressing given that several state-of-the-art shape measurement methods (see Appendix A) are based on fitting relatively simple galaxy models or are based on a decomposition into basis functions that cannot necessarily describe galaxy profiles in detail (Voigt & Bridle 2010; Melchior et al. 2010). More complex decompositions into basis functions often can describe more complex galaxies, but at the expense of introducing many tens or > 100 parameters, making them impractical for typical images with typical signal-to-noise ratios $S/N \sim 10$–20 (see §5.3 for the formal definition of this quantity). In addition, methods that measure galaxy distortions (§2.1) require an estimate of the intrinsic RMS galaxy distortion to convert to an ensemble shear, resulting in another type of dependency on the underlying nature of the galaxy population.

As an illustration of this problem, Fig. 3 shows several typical galaxies in high-resolution data from the *Hubble Space Telescope* (*HST*). Only a few tens of percent ($\sim 20\%$) of galaxies can be perfectly fit by simple galaxy models such as those commonly used by model-fitting methods today (e.g., top left); nearly half can be fit by them, but with additional substructure clearly evident (e.g., bottom left); and a few tens of percent ($\sim 30\%$) are true "irregulars" that cannot be fit by simple models at all (right panel). The GREAT08 and GREAT10 challenges used simple galaxy models, which motivates us to explore the impact of realistic galaxy morphology on shape measurement in the GREAT3 challenge, thus constraining "underfitting biases"[31]. Nearly all lensing data is lower resolution than what is shown in Fig 3; however, for this particular scientific application, we have reasons to believe that what we do *not* know (the unresolved, detailed galaxy morphology) *does* hurt us at some unknown level. One goal of the GREAT3 challenge is to quantify the extent to which that statement is true.

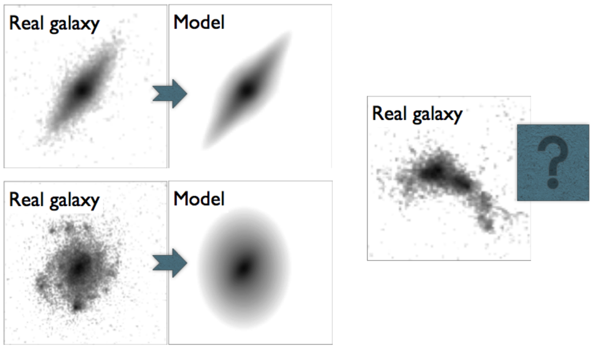

FIG. 3.— Real galaxies from the *HST* as observed by the Advanced Camera for Surveys (ACS) in the COSMOS survey (Koekemoer et al. 2007; Scoville et al. 2007b,a). The top left shows a galaxy that is well-fit by a simple parametric model from Lackner & Gunn (2012). The bottom left shows a galaxy that is reasonably well-fit but with additional substructure evident. The right side shows a true "irregular" galaxy that is not well-fit by simple parametric models with ~ 10 parameters.

The galaxies used for these simulations therefore come from *HST*. The technique for rendering the appearance of these galaxies with an added lensing shear is in Mandelbaum et al. (2012) and §5.1 of this handbook.

3.2. *Variable PSFs*

As discussed in §2.4, realistic PSFs have complex profiles and spatial variation due to the turbulent atmosphere (in ground-based measurements) and the instrument (the telescope and the camera). Different approaches have been used to study these PSF characteristics using data and simulations (e.g., Jarvis et al. 2008; Heymans et al. 2012; Chang et al. 2013; Hamana et al. 2013). We would like to test the impact of realistic PSFs on weak lensing measurement, both for the case of (a) a realistically complex PSF profile that is provided for participants, and (b) the case where the PSF has spatial variation that the participants must infer from a provided star field. The latter test is complicated by the low density of high-S/N stars that can be used to infer the PSF, making it hard to track high-frequency modes.

Case (a) can be motivated by Fig. 4, which shows a realistically complex PSF due to telescope optics and a simple model that is commonly used to represent it. As shown, the former is more complicated than the latter, and it is plausible that shape measurement methods could behave differently for the two cases. For simulated

[30] This equivalence is valid in the limit that these functions are continuous. For data that are discretely sampled, it is important to ensure that they are Nyquist sampled, otherwise the statement that pixelation can be treated as a convolution is false.

[31] These are biases in an M-parameter fit that arise when the true image has $N > M$ parameters, and some of the $N - M$ additional parameters correlate with the shear; e.g., Bernstein 2010.

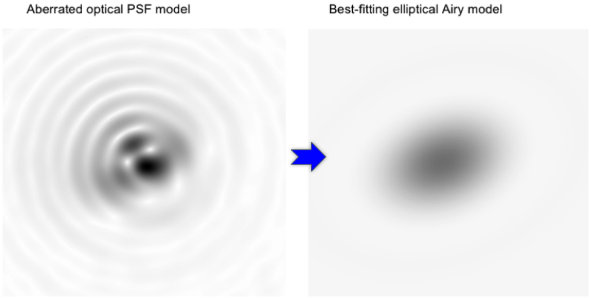
Aberrated optical PSF model
Best-fitting elliptical Airy model

FIG. 4.— Examples showing the simplicity of common approximations to optical PSFs, i.e., without an atmospheric contribution. **Left panel:** A realistic PSF, generated using lowest-order aberration theory with values that are typical for a well-aligned ground-based telescope, or for a (perhaps temporarily) misaligned space-based instrument. **Right panel:** The best-fitting (least-squares) approximation to this PSF using an elliptical Airy disk, a parametrized PSF model used in the GREAT10 galaxy and star challenges. The images are normalized to the same linear scale.

data from ground-based telescopes, there is also a convolution by the far broader atmospheric PSF[32].

Case (b), where the participants are required to infer the PSF, is similar to the GREAT10 Star Challenge. However, in that case the participants were judged on the accuracy of their PSF reconstruction. In GREAT3, the metric is the accuracy of shear field reconstruction, i.e., we test how PSF determination errors propagate into the recovered shear field. The value of this test is that different PSF reconstructions at a particular RMS accuracy could actually involve different spatial patterns in the residuals that affect shear field recovery in different ways, so ultimately we must quantify the performance of PSF estimation in terms of its impact on shear measurement.

3.3. *Combination of multiple exposures*

Most datasets used for weak lensing measurement are not single images, but rather multiple short exposures that are slightly offset from each other ("multi-epoch" data). Part of the data reduction procedure involves combining them to estimate the galaxy shapes - either via "co-addition" to form a stacked image (e.g. Lin et al. 2012; Jee et al. 2013), or by applying some simultaneous fit procedure that treats each exposure separately (e.g. Miyatake et al. 2013; Miller et al. 2013). Previous challenges have included a single deep image. In GREAT3, we include a test of how methods handle multiple images.

If the individual exposures are all Nyquist sampled and there are no major distortions or holes in the data (due to defects, cosmic rays, etc.), the combination of multi-exposure data is in principle straightforward, making this test less interesting. However, for a fraction of the data from ground-based telescopes, and all data from upcoming space missions, the data in individual exposures is *not* Nyquist sampled, which means that it is only possible to create a critically-sampled image by combining the multiple dithered (offset by sub-pixel amounts) images (e.g., Rowe et al. 2011). This is a more complicated algorithmic issue, and while our challenge does not address all aspects of it (e.g., it is even more complicated when there are holes in the data) we make a basic test of image combination.

When the PSF is very different in some exposures than others, it is possible to imagine gaining an advantage by up-weighting higher-resolution data. Hence it is possible that a method that does the most basic, fundamentally correct image combination could do worse than a method that is more clever in how the exposures are combined.

3.4. *Challenge philosophy*

The GREAT3 challenge is structured as a series of experiments to evaluate three key issues separately before combining them. Since our goal is to address how important these issues are for extant shape measurement methods (and encourage the development of new methods that might address these issues better), we deliberately omit some complications that were not chosen by the GREAT3 collaboration as top priorities. For a list of omitted issues, see §6, and note that the simulation software is capable of generating simulations that can address most of them.

One important note is the increased complexity compared to GREAT08 and GREAT10, for which the simpler questions being asked demanded simulations with (typically) δ-function distributions in galaxy and/or PSF parameters. If a GREAT3 participant needs simulations at that level of simplicity to test their code, they can either download the GREAT08 or GREAT10 simulations, or generate new (simple) simulations with public simulation software. Thus we are deliberately including more realistic distributions of galaxy parameters, but still in a format that allows for controlled experiments of the impact of realistic galaxy morphology, real PSFs and their variation, and combination of multiple exposure data. The challenge structure described below (§4.1) reflects this goal.

Like previous GREAT challenges, GREAT3 is meant to be inclusive of different data types. In that spirit, it includes both ground- and space-based data (of which participants may choose to analyze either or both); within those datasets, the images have some variations of key parameters so that they do not appear to come from the same instrument. Likewise, it has both constant- and variable-shear data, as some methods have assumptions that favor one or the other, and both are scientifically useful.

4. THE CHALLENGE

4.1. *Branch structure*

To achieve the goals outlined in §3, the GREAT3 challenge consists of five experiments:

- The control, which includes none of the three effects in §3 (but is a non-trivial test of how shear estimation methods cope with a galaxy sample that has a continuous distribution in size and S/N).
- Three experiments that each include only one of the effects of interest.
- One experiment that includes all three effects together.

For each experiment, there are branches with four data types: 2 shear types (constant and variable) and 2 obser-

[32] In the long-exposure limit, the PSF can simply be treated as the convolution of the optical and atmospheric PSF contributions.

vation types (ground and space). With four data types and five experiments, we have 20 branches (Fig. 5).

Within each branch, the physical setup is similar to that in previous challenges: there are 200 images per branch, each with a grid of 100 × 100 galaxies, and the goal of the participants is to infer some statistic of the shear field for each image. The images represent 10 × 10 degree fields. These images are not completely independent: each branch of the 14 branches with variable shear and/or PSF has 10 fields representing distinct regions of the sky, but each field contains 20 slightly offset *subfields* with different galaxies that sample the same shear field (in the case of variable shear). See Fig. 6 for an illustration of how subfields and fields are related. Thus participants must estimate the shear correlation function for each of the 20 fields for the variable shear experiment, combining all galaxies in all subfields when estimating the correlation function, which can be done using software supplied with the data (see §4.2). The subfields within a field sample the same PSF pattern (see §5.6). For the 6 branches that have constant shear and constant PSF, the branch contains 200 fields, each with a single subfield per field. Thus the constant shear branches have 200 separate shear values. For the multi-epoch simulations, each *epoch* of a given subfield has a different PSF; however, a particular epoch has the same PSF for all subfields in the field. The branches are meant to represent the same underlying galaxy population, modulo issues that arise when the PSF size varies (which means that galaxies that are smaller might be simulated in one image but not another, see §5.1).

4.2. *Overall information*

Given the challenge branch structure in Fig. 5, we estimate a total data volume of 2 TB (still to be finalized). This figure is dominated by the 8 multi-epoch branches. Participants may choose to submit results to any or all of those branches at their own discretion, and likewise can download any subset of the data that they wish. The preferred method of getting the data is via download from our server or its US mirror, however for a limited number of people for whom this is not feasible, we can supply a hard drive with the simulations.

The challenge is carried out as a competition, with a separate leader board for each branch evaluated according to metrics described in §4.4, and an overall leader board with rankings determined based on a combination of results from the individual leader boards as described in Appendix C.2. There are prizes for the first and second place winning teams (§C.1) of the overall challenge leader board.

Detailed rules for the challenge are listed in Appendix C. Here, we summarize the online resources related to the challenge:

- Webpage with leader boards, information on downloading the simulated data, basic information about shear conventions, submission format, and simulation file formats: http://great3.projects.phys.ucl.ac.uk/leaderboard/
- GREAT3 web page with basic information, announcements of meetings: http://great3challenge.info
- GREAT3 public code repository, which includes code that participants can use to automatically calculate shear correlation functions in a format needed for submission, a FAQ, and an issue page that participants can use to ask questions about the challenge: https://github.com/barnabytprowe/great3-public
Eventually this will also include example scripts that can analyze all of the challenge data with some simple, existing method.

Participants may optionally sign up for a mailing list for announcements related to the challenge data; information about this is available on the leader board website. Questions about the challenge can be sent to challenge@great3challenge.info.

4.3. *Timeline*

At the time of the submission of this handbook to arXiv, there was a beta release of several challenge branches and a working leader board. During a two month time period, people are encouraged to download, play with the data, and raise any questions and issues on the public GitHub repository described in §4.2. After that point, the entire challenge dataset will be released publicly. The challenge will run for 6 months from that time, after which there will be a final meeting.

4.4. *Evaluation of submissions*

Evaluation of submissions within each branch uses metrics described here, where the metric depends on whether the branch has constant or variable shear.

4.4.1. *Constant shear*

For simulations with constant shear, each subfield has a particular value of constant shear applied to all galaxies. Participants must submit an estimated shear for each subfield in the branch, and the metric calculation uses those estimated shears as follows:

Following a parametrization used in the STEP challenges and elsewhere (Heymans et al. 2006; Huterer et al. 2006; Massey et al. 2007b) we can relate biases in observed shears $g^{\rm obs}$ to an input true shear $g^{\rm true}$ using a linear model in each component:

$$g_i^{\rm obs} - g_i^{\rm true} = m_i g_i^{\rm true} + c_i \tag{9}$$

where $i = 1, 2$ denotes the component of shear, and m_i and c_i are referred to as the multiplicative and additive bias, respectively. From the user-submitted estimates of the mean $g_i^{\rm obs}$ for each of 200 subfields in a branch, the metric calculation begins with a linear regression to provide estimates of m_i, c_i given the known true shears. This is done in a coordinate frame rotated to be aligned with the mean PSF ellipticity in the field, since otherwise (with randomly oriented PSF ellipticities) the c values will not properly reflect contamination of galaxy shapes by the PSF anisotropy. There is a subtlety in this calculation, which is that methods that apply weights to the galaxies will not in general give the same weight to a galaxy and its 90-degree rotated pair (§5.3), resulting in imperfect shape noise cancellation. At some level, the weights will typically correlate with the PSF ellipticity,

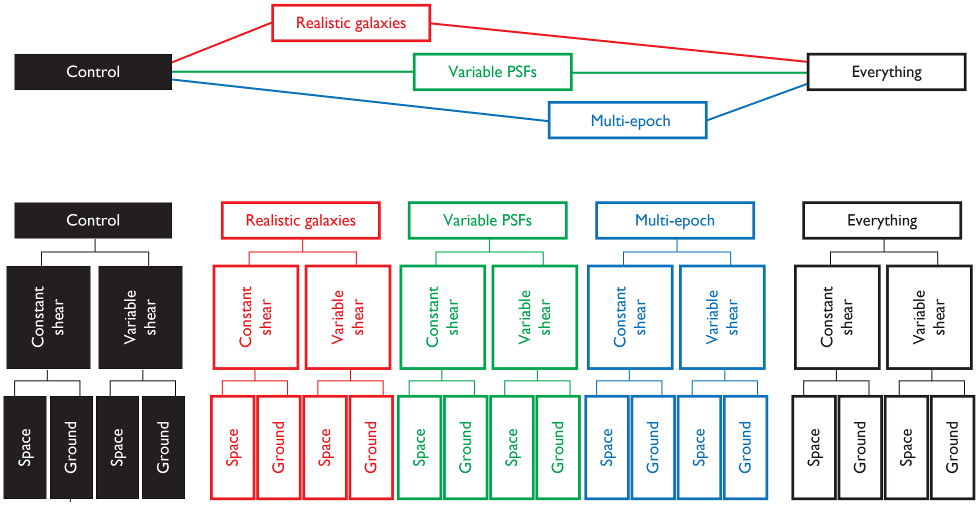

FIG. 5.— A schematic of the GREAT3 branch structure.

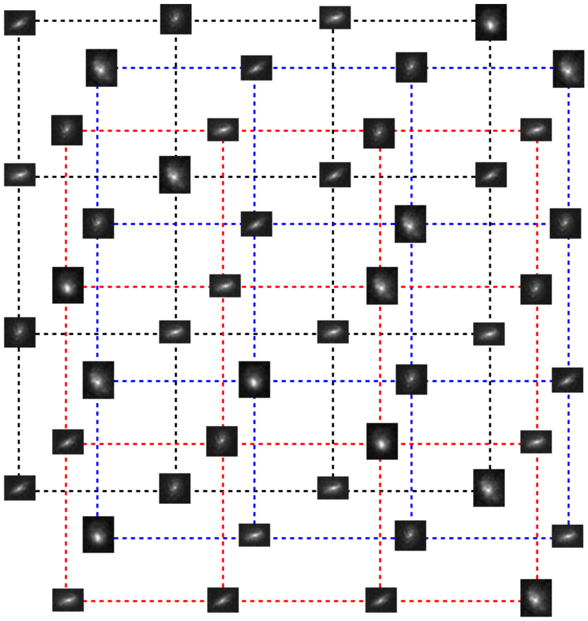

FIG. 6.— An illustration of how subfields relate to fields, for a simple case with 3 subfields per field, each containing a 4 × 4 grid of galaxies. The image shows galaxies that are all part of the same field (a region of sky containing a particular cosmological shear field and PSF). Purely for the sake of convenience, rather than due to any real difference between these sets of galaxies, we distribute the images in subfields that consist of regular grids shown as dashed lines of different colors. In our actual case of 20 subfields per field (for variable shear and/or PSF) and 100 × 100 galaxies, we have randomly chosen different sets of 20 offset subfield positions for each field.

thus giving rise to a spurious "c" value that is due to selection bias rather than due to failure to correct for the PSF anisotropy properly (e.g., Miller et al. 2013). Methods with aggressive weighting schemes may be more susceptible to this issue. However, as this issue should arise in real data as well, it seems like a true issue rather than one that occurs in simulations alone, so we do not attempt to correct for it.

Note that for variable PSF branches with constant lensing shear, we are somewhat less sensitive to additive systematics, because if the average PSF ellipticity is zero then even in the presence of huge additive systematics, there is no well-defined PSF direction for the field and the additive systematics cancel out[33].

Having estimated these bias parameters m_i, c_i, we then construct the metric for constant shear branches, which we call Q_c. This is done by comparison of m_i, c_i to 'target' values $m_{\rm target}$, $c_{\rm target}$. The values of these targets are imposed by the statistical uncertainties for upcoming weak gravitational lensing experiments: in GREAT3 we adopt $m_{\rm target} = 2 \times 10^{-3}$ and $c_{\rm target} = 2 \times 10^{-4}$, motivated by the most recent estimate of requirements for the ESA *Euclid* space mission (Massey et al. 2013). The constant shear branch metric is then defined as

$$Q_{\rm c} = \frac{2000}{\sqrt{\sum_{i=1,2} \left(\frac{m_i}{m_{\rm target}}\right)^2 + \left(\frac{c_i}{c_{\rm target}}\right)^2}} \tag{10}$$

This metric is normalized such that we expect a value $Q_{\rm c} \simeq 1000$ for methods that meet our chosen targets on m_i and c_i. The response of the metric to m and c larger than the fiducial values is shown in Table 1.

4.4.2. *Variable shear*

For simulations with variable shear, the key test is the reconstruction of the shear correlation function. This differs from GREAT10, which used a metric based on reconstruction of the power spectrum. We adopt a correlation function-based metric because the power spectrum-based metric requires the subtraction of shot noise (§5.3) that contributes at all values of k and depends on the details of the shape measurement method (Kitching et al. 2012). Subtraction of the shot noise term has associated uncertainty, and the real-space correlation function is a cleaner quantity since that shot noise only contributes at zero angular separation. Also, the correlation function-based metric is more sensitive to additive shear systematics in the case of a constant PSF. The correlation function has other complications, particularly the fact that the simplest correlation functions to calculate

[33] This is not the case for variable shear branches, given our use of a correlation function-based metric.

For fixed $c = c_{\rm target}$			For fixed $m = m_{\rm target}$		
m	$Q_{\rm c}$	$Q_{\rm v}$	c	$Q_{\rm c}$	$Q_{\rm v}$
0.002	1000	1000	0.0002	1000	1000
0.00632	470	950	0.000632	450	930
0.02	150	680	0.002	150	530
0.06325	50	250	0.006325	50	80
0.2	20	70	0.02	20	9

TABLE 1
Approximate response of our metrics in Eqs. 10 and 11 for constant and variable shear, respectively, to multiplicative (top) and additive (bottom) shear biases. These results depend on the noise in shape measurements, which includes the impact of pixel noise plus any failure in shape noise cancellation. Thus, the results in reality may differ slightly.

do not cleanly separate into E and B modes, which is necessary to separate lensing shear signals from our input B-mode shape noise (see Appendix H). However, there is a straightforward prescription for E vs. B mode separation from correlation functions that does not depend on the shape measurement method, making it a good candidate for use in a variable shear field metric, which we now describe.

Submission of results for variable shear branches begins with calculation of correlation functions (Appendix B), this being done by the participant. Software to calculate the correlation function in the proper format for submission is distributed in the GREAT3 code repository[34], though participants may use their own software if they wish. The submission consists of estimates of a quantity called the *aperture mass dispersion* (e.g., Schneider 2006; Schneider et al. 1998), which can be constructed from simple ξ_+ and ξ_- correlation function estimators, and allows a separation into contributions from E and B modes (see Appendix B for details). We label these E and B mode aperture mass dispersions M_E and M_B in the discussion that follows.

The submissions take the form of an estimate of $M_{E,j}$ for each of 10 fields labelled by index j: this estimate is therefore constructed using *all twenty* subfields in a given field. This choice is to provide a large dynamic range of spatial scales in the correlation function, which helps the metric probe a greater range of shear signals. The $M_{E,j}$ can be estimated by the provided software in $N_{\rm bins}$ logarithmically spaced annular bins of galaxy pair separation θ_k, where $k = 1, \dots, N_{\rm bins}$, from the smallest available angular scales in the field to the largest.

These $M_{E,j}(\theta_k)$ are to be submitted for each field $j = 1., \dots, N_{\rm fields}$, where $N_{\rm fields}$ is the total number of fields in the branch. The metric $Q_{\rm v}$ for the variable shear branches is then constructed by comparison to the known, true value of the aperture mass dispersion for the realization of E-mode shears in each field. These we label $M_{E,{\rm true},j}(\theta_k)$. The variable shear branch metric is then calculated as

$$Q_{\rm v} = \frac{\eta}{\frac{1}{N_{\rm fields}N_{\rm bins}} \sum_{j=1}^{N_{\rm fields}} \sum_{k=1}^{N_{\rm bins}} \left| M_{E,j}(\theta_k) - M_{E,{\rm true},j}(\theta_k) \right|} \tag{11}$$

where η is a normalization factor designed to yield $Q_{\rm v} \simeq 1000$ for a method achieving $m = m_{\rm target}$ and $c = c_{\rm target}$ in both shear components (similar to the normalization for the constant shear metric). Using simulations, we chose $\eta \simeq 8.41 \times 10^{-5}$ for this desired scaling. However, this value for η depends on an estimate for the typical statistical uncertainty on a single shear measurement due to pixel noise (i.e. that noise which remains after galaxy intrinsic shape noise is placed in the B-mode). To estimate $\eta = 8.41 \times 10^{-5}$ we used a value $\sigma_n = 0.05$ for the independent uncertainty on each component of shear; as we get better information about σ_n from submissions to the challenge, we can adjust η to help improve the correspondence between $Q_{\rm v}$ and shear measurement targets such as those of Massey et al. (2013). For this choice of η, Table 1 shows the response of $Q_{\rm v}$ to multiplicative and additive shear systematics. As shown, it is less sensitive than $Q_{\rm c}$. This is primarily due to the fact that typical shears in the variable shear branch are several times smaller than those in the constant shear branch.

5. SIMULATIONS

The simulations for this challenge were all produced using GalSim, a publicly available[35] image simulation tool that has been developed as a community project in part for GREAT3, but with additional capabilities. The software package is fast, modular, and written in C++ and Python. Since it is described in detail in the documentation on the webpage and an upcoming paper (Rowe et al., *in prep.*), here we simply present evidence that GalSim can accurately simulate galaxies with an applied shear – see Appendix D for details.

Below we describe the ingredients that go into the simulations.

5.1. *Galaxies*

The galaxy population in the GREAT3 simulations is meant to represent a realistic galaxy population in its distribution of size, magnitude, and morphology. Here we describe how we achieve this goal.

We use data from the *HST* to ensure the realism of the size and magnitude (and thus S/N) distribution, as it provides the highest resolution images available with a uniform coverage over a reasonable area. The way this is done depends on whether the branch in question tests the question of realistic galaxy morphology. If it is, then we use actual images of galaxies from the training sample, with a modification of the procedure described in Mandelbaum et al. (2012) to remove the *HST* PSF (unlike in that work, fully rather than partially) in Fourier space, apply the lensing shear and magnification, convolve with the target PSF, then return to real space and resample to the target pixel scale. This method generates an image of what the galaxy would have looked like at lower resolution, provided that the target band limit $k_{\rm lim,targ}$

[34] https://github.com/barnabytprowe/great3-public

[35] https://github.com/GalSim-developers/GalSim

relates to the original *HST* band limit $k_{\rm lim,HST}$ via

$$k_{\rm lim,targ} < \left(1 - \sqrt{\kappa^2 + \gamma^2}\right) k_{\rm lim,HST}. \qquad (12)$$

For weak shears and convergences, the above condition is easily satisfied by all upcoming lensing surveys, even those from space.

For the simulations that are *not* meant to test the effects of realistic galaxy morphologies, we still use the *HST* data to determine a distribution of galaxy properties, based on simple parametric fits (Sérsic) to the *HST* images. We then use the best-fitting models rather than the images themselves. This means that a comparison with the simulations that use real galaxy images will show the importance of underfitting bias.

To limit the volume of simulations needed to constrain biases in shear estimation to levels needed for upcoming lensing surveys, we must cancel out the dominant form of noise in lensing observations, the intrinsic shape noise (see §5.3). Cancellation of shape noise requires that each measurement use all simulated galaxies without any exclusions. Since many methods eliminate a non-negligible fraction of the galaxies with $S/N < 20$, we therefore only include a distribution of galaxies with S/N values down to 20 (which in practice results in a median $S/N \sim 32$ due to the predominance of faint galaxies). Even then, noise will result in imperfect shape noise cancellation due to chance failures to measure galaxies or differently assigned per-galaxy weights; for more details, see §5.3. This S/N distribution is for single-epoch branches; for multi-epoch branches, the S/N values are such that the *total* S/N over all epochs is ≥ 20. Likewise, we exclude galaxies that are so small as to be nearly unresolved in the simulations, since many methods will have difficulty measuring their shapes. The exclusion is done on a per-simulation basis, so the galaxies used in the space and ground simulations differ, with the latter being a subset of the former. These pragmatic choices, meant to avoid simulated datasets that are $\geq$ 10 TB in size, do mean that we are in practice avoiding a galaxy population that is of great interest, given the large number of galaxies at $S/N < 20$ and the known issues with noise biases (Hirata et al. 2004; Kacprzak et al. 2012; Kitching et al. 2012; Melchior & Viola 2012; Refregier et al. 2012). Future work should explore the interaction between modeling bias, the main subject of our tests here, and noise bias, which is already the subject of a substantial body of work.

Galaxy populations evolve with redshift, including an increasing abundance of irregular-type morphologies and decrease in the number of elliptical galaxies at high redshift, where there are more young, star-forming galaxies and recent mergers (e.g., Bundy et al. 2005). This redshift evolution of the galaxy population translates into a depth-dependence; for deeper data, there are more high-redshift galaxies and therefore more irregulars. Thus it is relevant to ask what is the effective depth of the simulated dataset. Here we are limited by the *HST* dataset that we use. If we use real images as the basis for simulations, then the noise in those images also undergoes the same steps as the galaxy (deconvolution, shearing and magnification, convolution with the target PSF, and resampling to the target pixel scale). The noise in the result can be predicted from the original one (since the aforementioned processes do well-defined things to the noise) and has a direction that correlates with both the input shear and the target PSF. Moreover there are generally non-negligible pixel-to-pixel correlations. While we can add noise that is anti-correlated and anisotropic to achieve isotropic, uncorrelated noise in the simulations (a process that we call "noise whitening", see Rowe et al., *in prep.* for more details), this also imposes a further limitation on the depth of the simulated images[36]. A simple calculation for reasonable PSFs is that the effective limiting magnitude for simulated space-based data is actually 0.6 magnitudes brighter than that in the *HST* training sample given the need to whiten the substantial correlated noise[37] that is already present, whereas for simulated ground-based data it is 0.15 magnitudes brighter than the limiting magnitude in *HST* (because the correlated noise gets more washed out in the simulation process due to the larger PSF). However, we defer the final answer to this question of the depth of the simulated galaxy population to §5.8.

Details of the *HST* training sample and the fits to parametrized models can be found in Appendix E.

5.2. *Point-Spread Functions (PSFs)*

Here we describe the PSF model used in the challenge. In all simulated images, there is an optical PSF; the simulations of ground-based data also have an atmospheric PSF. The two main aspects to consider for all simulated PSFs are (1) the profile of the PSF and (2) the spatial variation of the PSF profiles across the focal plane. Both of these factors are determined by the physical mechanisms that give rise to the PSF. In §5.2.1 and §5.2.2, we describe the physical basis for the optical and atmospheric PSFs in the "variable PSF" experiment in this challenge.

The simulated 10×10 deg^2 images are much larger than the field-of-view (FOV) of typical telescopes. Thus, we do not simulate a variable PSF model for an image that size. Instead, for the branch with variable PSFs, we divide each image into square tiles, and simulate a PSF model in each one of them; this underlying PSF model is the same for all subfields within the same field, modulo a position shift (which is provided; participants do not have to estimate it). For the experiments with a constant PSF, we use a simplified version of our variable PSF models, selecting a single PSF for each 10×10 deg^2 subfield. For the single epoch experiments, we make simulated images corresponding to a single deep exposure rather than to an idealized co-added image, because co-added PSF profiles can have complicated features that would make comparison between the experiments difficult.

5.2.1. *Optical PSFs*

All telescopes effectively convolve the images they observe with a PSF of finite size, due to diffraction and optical aberrations. We refer to this contribution as the

[36] Technically the noise whitening procedure means that we treat the noise in the original images as part of the galaxy. This motivates us to use data for which the added (simulated) noise dominates over the existing noise in the images.

[37] Correlated noise is present in these images for several reasons, including the correction for charge-transfer inefficiency and the resampling from the instrument pixel scale to a smaller pixel scale when combining multiple exposures.

"Optical PSF". In GREAT3, the contributions to this PSF can be grouped into the following categories:

- The shape of the pupil through which light diffracts: this includes obscurations such as secondary mirror or instrument at prime focus, and obscurations due to any struts or supports for these.
- Aberrations, which can be split into three terms: the "design residual" (aberrations in a perfectly realised telescope design); "figure errors" (aberrations due to warping or manufacturing imperfections in mirrors); and "alignment" (how well mirrors are positioned relative to each other).
- Pointing jitter, the variation in the telescope pointing during an observation, which gives rise to a blurring kernel of a size and ellipticity that is highly design-dependent.
- Detector response: a primary term here is the diffusion of charge within detectors before readout, but other (possibly non-linear) effects may be present, depending on the device in question. Strictly speaking, any non-linear, signal dependent, or space variant effects cannot be represented in the simple convolutional model of image formation, and must be applied on top of that.

In all simulations with variable PSFs, we have a fixed pupil shape and detector response across a given FOV for all tiles in the same image. Typically the pupil consists of a circular aperture upon which is superimposed a co-centric, circular obscuration, and additional supports that are rectangular in shape and extend radially from the central obscuration.

Aberrations vary across the simulated FOV, however, with a different prescription for the ground and space-based data. Unfortunately, we found that it is extremely difficult to make a realistically complex optical PSF model, including full spatial variation, that is not instrument-specific in some way. While our goal is to make the GREAT3 challenge as generic as possible, we nevertheless had to adopt optical PSF models that are specific to certain instruments for the "variable PSF" branches. The specific optical PSF models used for that branch are described in Appendix F, and some example PSF images are in Fig. 7.

For the simulations with constant PSF models, we adopt simple variants of the models described in Appendix F. For example, the space-based optical PSF model we use for the "constant PSF" branches is generalized compared to that for WFIRST-AFTA in several ways. In particular, the size of the basic diffraction-limited PSF is determined by the ratio of wavelength of the light to primary mirror diameter. We choose a range of allowed values for this parameter including the values for several upcoming surveys; a range of obscuration by the secondary mirror; several different sets of configurations for the struts. Some additional aberrations to represent deviations from the design residual are included. These are evenly distributed among all the aberrations we consider for the space-based model, and for the ground-based model all aberrations are represented, but defocus is most important (motivated by the realistic ground-based optical PSF model). As is commonly the case, the size of the additional aberrations is a factor of several higher for the ground-based PSF than for the space-based PSF.

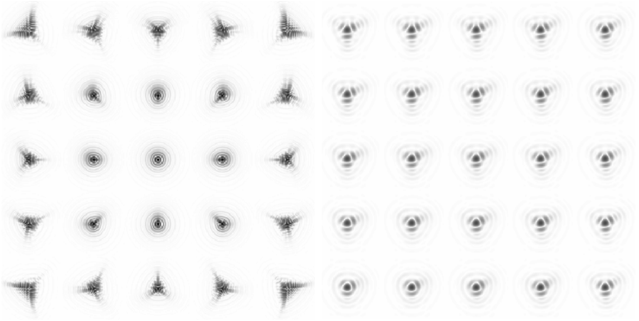

FIG. 7.— **Left:** The optical PSF (no atmospheric contribution) for the ground-based "variable PSF" branch at 5 × 5 grid positions across a simulated FOV, going all the way to the edge where aberrations are large. **Right:** Same as left, for the space-based model. Both are shown on a logarithmic scale. These include some (stochastic) added aberrations at a level used for the challenge. The space-based optical PSF model is more constant across the field than the ground-based model because of different assumed field-dependent aberrations.

5.2.2. *Atmospheric PSFs*

Atmospheric turbulence is the primary contributor to the PSF in ground-based data. Our model for the ground-based PSF is that of a large ($\geq$ 2 m) ground-based telescope taking long exposures without adaptive optics. To properly simulate the profile and the spatial variation of the atmospheric PSFs, we consider the following construction based on a combination of high-fidelity atmospheric turbulence simulations and observational data. Further technical details regarding the design of our atmospheric PSFs can be found in Appendix G.

We invoke the LSST Image Simulator[38] (PhoSim, LSST Science Collaborations & LSST Project 2009; Connolly et al. 2010; Peterson et al. in prep.), a high-fidelity photon ray-tracing image simulation tool, for this purpose. PhoSim adopts an atmospheric turbulence model similar to that used in the adaptive optics (AO) community (Roggemann & Welsh 1995; Hardy 1998), with several novel implementations to adapt to the wide-field nature of modern survey telescopes. The PhoSim atmospheric model has been shown to properly represent observational data (Peterson et al. in prep.). Since we were concerned only with studying the effects of the atmosphere, we ran PhoSim in a special mode with the LSST optics removed.

First, we consider the general profile of the atmospheric PSF. To first order, this includes the PSF radial profile, the PSF size, and any anisotropy of the PSF shape. The atmospheric PSFs generated from PhoSim with exposure times appropriate for the challenge (> 1 minute) has a radial profile that is consistent with the long-exposure limit atmospheric PSF predicted by a Kolmogorov model. The

[38] https://dev.lsstcorp.org/trac/wiki/IS_phosim

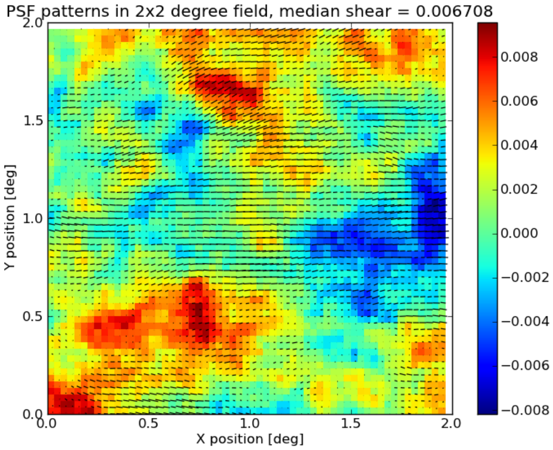

FIG. 8.— A single random realization of an atmospheric PSF anisotropy pattern in a 2×2 deg^2 field, for a 2 minute exposure at a 4-meter telescope. The plot title gives the median PSF shear. The color scale indicates the fractional change in size of the atmospheric PSF as a function of position.

PSF profile can be written as (Fried 1965):

$$\mathrm{PSF}(\vec{\theta}) = FT\left\{\exp\left[-\frac{1}{2}6.88\left(\frac{\bar{\lambda}D|\vec{f}|}{r_0}\right)^{5/3}\right]\right\} \qquad (13)$$

where "FT" represents a Fourier transform between angular position $\vec{\theta}$ and spatial frequency $\vec{f}$, $\bar{\lambda}$ is the average wavelength, D is the aperture size, and r_0 is the Fried parameter.

Given a survey design, the Kolmogorov PSF takes one parameter, r_0, which can be rephrased in terms of the commonly-used "atmospheric seeing", defined as the full-width at half maximum (FWHM) of the atmospheric PSF. We adopt a distribution of FWHM values estimated at the summit of Mauna Kea in one optical filter (R, $\langle\lambda\rangle \sim 651$ nm) at zenith[39]. The quartiles of this distribution are 0.49, 0.62, and 0.78 arcsec; the mean value is 0.66 arcsec. For a single exposure, we draw a value of atmospheric PSF FWHM from this distribution. These are not purely random; since single-epoch experiments only have 10 different PSFs in the entire branch, and we want to properly cover this distribution, we draw randomly from within percentiles, i.e., one field has a PSF drawn randomly from below the tenth percentile in the distribution, another is from the tenth to twentieth percentile, and so on. Finally, these Kolmogorov PSFs are assigned an ellipticity (Eq. 6) to represent the small anisotropy in the atmospheric PSF. The ellipticity values are based on a large number of PhoSim simulations.

Next, we consider the spatial variation of the PSF model parameters (size and ellipticity), quantified by a 2-point correlation function. We find that the spatial variation of atmospheric PSF parameters in PhoSim can be well described by a functional form with two parameters. We generate the spatially varying PSF parameters as a Gaussian random field that corresponds to this correlation function, with the two model parameters allowed to vary in a reasonable range. An example of an ellipticity field and the spatial variation of PSF size generated from this procedure (described in more detail in Appendix G) is shown in Fig. 8.

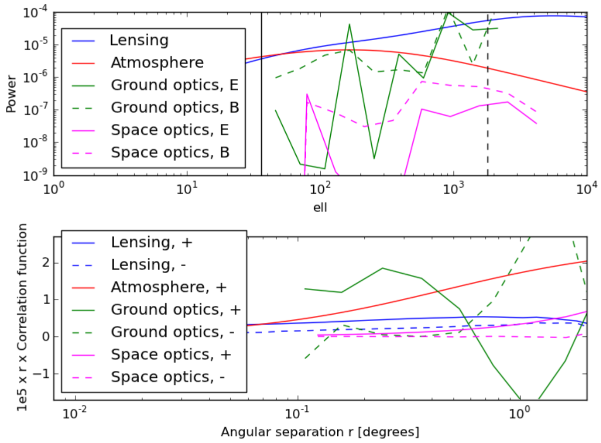

FIG. 9.— **Top:** Dimensionless power of lensing shears, atmospheric PSF shears, and optical PSF shears (where the latter is computed by tiling several adjacent pointings for our ground-based optical PSF model – without additional tilt, misalignment, or defocus – after convolving with a circular 0.7″ Kolmogorov blur for the ground-based model). For lensing, the power spectrum that is shown is E mode, and the B mode power spectrum is zero. For the atmospheric PSF, the power spectrum shown is the same for E and B. The solid black line shows the minimum accessible ℓ value given the size of our images. The dashed black line shows the maximum ℓ value given the grid spacing on our images; however, since multiple images sample the same shear field, the true maximum ℓ is actually larger. **Bottom:** Correlation functions for the three cases shown above, after multiplying by separation on the sky in degrees. Here we show ξ_+ and ξ_-, but the latter is identically zero for our atmospheric PSF model.

Our choice to use a sheared Kolmogorov profile (without any higher-order distortions) is a simplification compared to reality, but the simulations suggest that for reasonable exposure times and telescope sizes, it is quite close to correct. Hence we consider our prescription to be realistically complex enough for an interesting test.

Fig. 9 shows a comparison between the power spectrum and correlation functions of the lensing shears, the atmospheric PSF anisotropies, and the ellipticity of the optical space- and ground-based PSF model (in the latter case, after convolving with a circular, typical-sized atmospheric PSF). Here we have omitted the aberrations other than the design residual to get an idealized version of the results for the optical PSF model. This plot shows the most important scales for the various systematics compared to the weak lensing shears. For example, we see that the lensing power spectrum is below that of the atmospheric PSF anisotropies on large scales (small ℓ). However, for nearly all relevant scales on our grid, the atmospheric PSF anisotropy correlation function is above that of the lensing shear. The optical PSF anisotropy is primarily relevant on small scales (small angular separation or high ℓ). However, because of the tiling of multiple fields of view, it can be important on large scales, particularly for the space-based model.

5.2.3. *Star fields*

[39] Figure 1 from http://www2.keck.hawaii.edu/optics/ScienceCase/TechSciInstrmnts/Products_SeeingVarMaunaKea.pdf.

In the constant PSF experiments, we provide several noiseless images of the PSF for each image, at the same resolution as the galaxy images. One of those images is centered within a pixel; the others are randomly offset by some amount to be determined by participants, who may use them if they wish to recover information about the PSF on sub-pixel scales. For the space-based images with single exposures, the PSF is Nyquist sampled, and hence those offset images carry no additional information.

In the variable PSF experiment, we provide star fields that can be used for PSF estimation, one star field per subfield (however the 20 subfields in a field have the same underlying PSF). If our 20 slightly offset grids of 100×100 galaxies that belong to the same field cover the same 10×10 deg^2 area of the sky, and we want to simulate a realistic stellar density[40] of 2 arcmin^{-2} down to $S/N = 50$ for an image that goes to r magnitude of 25, that means each star field has $\sim 1.3 \times 10^4$ randomly-located stars[41]. The magnitude distribution for the star fields is motivated by the model in Jurić et al. (2008). Some methods use only very high S/N stars resulting in ~ 1 arcmin^{-2}, but those that can go to lower S/N will find a higher usable stellar density and may be able to better trace the small-scale fluctuations in the PSF. For the experiment containing all effects, each epoch will have its own star field for PSF estimation, since the PSF varies per epoch.

5.3. *Noise model*

In a weak lensing measurement, two important sources of noise are "shape noise" (the intrinsic, randomly-oriented galaxy shapes that we must average out to measure the small, coherent lensing shears), and "shape measurement error", the noise in individual galaxy shape measurements due to the noise in each pixel. For typical galaxy populations, the shape noise dominates over measurement error for all but the very lowest signal-to-noise galaxies, where the two might become comparable. Together these sources of error are often called "shot noise".

Previous challenges have incorporated schemes to cancel out the shape noise, thus substantially reducing the volume of simulations needed to test shear measurement methods very accurately from of order 10 TB to ~ 1 TB. Shape noise cancellation is imperfect due to measurement error, but it is still reasonably effective down to observations with $S/N \sim 20$ (for S/N defined in Eq. 16). For galaxies with lower S/N, the noise typically leads to a substantial measurement failure rate that renders shape noise cancellation very ineffective. Shape noise cancellation also eliminates some selection biases.

Given the limitation imposed by our desire to keep the simulation volume under control, in GREAT3 we employ shape noise cancellation, with a lower limit on the galaxy S/N of 20. In the constant-shear simulations, shape noise cancellation is carried out by having the same galaxy included twice, with orientations rotated by 90 degrees before shearing and PSF convolution (Massey et al. 2007b). Given the typical S/N and intrinsic ellipticity distribution for the galaxies in our parent sample, the shape noise cancellation scheme reduces the errors on measured shears by a factor of 3 (equivalent to 9× simulation volume). We have tested the effect of completely random galaxy omissions (e.g., due to convergence failure for some shape measurement method), and found that for simulated data with typical S/N, the errors on the measured shear increase from the optimal case (perfect shape noise cancellation) by 8%, 30% and 50% for the case of 5%, 10%, and 20% missing galaxies, respectively. This is still well below the 200% increase that corresponds to no shape noise cancellation, so even for a significant random failure rate the errors on the shear (and therefore metric) increase, but not so much that the results become useless.

In the variable-shear simulations, as for GREAT10, the lensing shear is entirely E-mode power (as in reality) and shape noise is only in the B-mode (this is completely unrealistic, but useful). This task is more complicated than for GREAT10 given our use of a real galaxy population; see Appendix H for a description of how we carry out shape noise cancellation.

It thus remains to describe our model for pixel noise. In real data, pixel noise is largely Poisson (since the CCDs are counting photons) but with a small Gaussian component from detector read noise. In many datasets, Poisson noise is dominated by the sky rather than the objects, except for very bright ones that constitute a small fraction of the objects used for shape measurement. Moreover, the sky level is often high enough that its Poisson noise is essentially Gaussian. We therefore employ a Gaussian noise model only. Our definition of galaxy S/N, which we use to decide which galaxies go into the simulations, is the same as for GREAT08 (Bridle et al. 2010). We define the signal as a weighted integral of the flux,

$$S = \frac{\sum W(\mathbf{x}) I(\mathbf{x})}{\sum W(\mathbf{x})} \tag{14}$$

and its variance is

$$\mathrm{Var}(S) = \frac{\sum W^2(\mathbf{x}) \mathrm{Var}(I(\mathbf{x}))}{(\sum W(\mathbf{x}))^2}. \tag{15}$$

In the limit that the sky background dominates, $\mathrm{Var}(I(\mathbf{x}))$ is a constant, so we can just call it $\mathrm{Var}(I(\mathbf{x})) = \sigma^2$, the pixel noise variance. We adopt a matched filter for W, i.e., $W(\mathbf{x}) = I(\mathbf{x})$. Putting those assumptions into Eqs. 14 and 15 gives

$$S/N = \frac{\sqrt{\sum I^2(\mathbf{x})}}{\sigma}. \tag{16}$$

While we do not have noise-free images for the real galaxies (for calculating the sum over squared intensities), we can use the model fits to the galaxy images as noise-free images for this purpose.

This S/N definition is also used for the stars in the star fields when defining a S/N limit. For stars, the assumption that background dominates is not very realistic.

Also, as described in §5.1, the original training data in the "realistic galaxy" branches has noise in it already, so we only add enough noise to ensure that the resulting noise correlation function is the target one, i.e., Gaussian noise with σ defined by Eq. 16, without pixel-to-pixel correlations.

[40] Considered as an average over stellar densities for different galactic latitudes at reasonable galactic altitudes.

[41] However, a small exclusion radius is placed around each one to avoid blending effects.

Many image processing steps that are carried out on real data, especially from space telescopes, can lead to correlated noise, due to stacking of multiple exposures. For simplicity we include only uncorrelated noise in GREAT3.

5.4. *Image rendering*

GalSim provides two primary methods of rendering images of a galaxy that has been sheared/magnified and convolved with a PSF: via discrete Fourier transform (DFT), and via photon-shooting. The latter method was used by the software for the GREAT08 and GREAT10 challenges, and involves representing shears, magnifications, and convolutions as offsets of photons that were originally drawn according to the light distribution of the intrinsic galaxy profiles.

However, for the GREAT3 challenge, we have adopted DFT as our method of image rendering, for the following reason: to use real galaxy images as the basis for our simulations (§5.1), we need to remove the original PSF from the *HST* images. There is no way to represent deconvolution in a photon-shooting approach, and so for consistency, all branches of the GREAT3 challenge (even those that use parametric galaxies) are generated using DFT.

However, since the two methods use different approximations, our tests of the image rendering in Appendix D include a comparison of DFT versus photon-shooting as a way to validate the results.

5.5. *Constant versus variable shear*

The challenge consists of two shear types. Half of the challenge branches contain images with a single constant value of shear for the image, and the other half contain images that have a variable shear field. The justification for this division is that some lensing measurements, like galaxy-galaxy lensing, can be carried out by averaging some roughly constant shear value within annuli around lens object(s), whereas cosmic shear measurements involve estimating the variable shear field caused by large-scale structure. Both types of measurements are scientifically useful. Additionally, some shear estimation methods may work better in one regime than the other; stacking methods (e.g., Lewis 2009; Bridle et al. 2010) are simplest to interpret in the constant shear regime, whereas methods that assume something about the statistical isotropy of the galaxy shape distribution may fail in a constant shear field.

For the constant shear branches, simulations have a single constant value of shear drawn randomly from a hidden distribution in $|g|$ with some minimum and maximum value, with purely random position angles.

In the variable shear branches, we start with a shear power spectrum with reasonable shape for a typical cosmology, and with slightly high amplitude in order to increase sensitivity of Q_v to multiplicative biases. Then, we include a nuisance function that gives scale-dependent modifications of order $\sim$ 10% on the range of scales traced by our grid of galaxies. In a single dimension, the angular grid extent of $L = 10$ degrees ($L_\mathrm{rad} = \pi/18$ radians) means that the minimum relevant ℓ value is $\ell_\mathrm{min} = 2\pi/L_\mathrm{rad} = 36$.

This shear power spectrum is given as input to GalSim, which uses it to generate a realization of a Gaussian random shear field, and also generates self-consistent convergences. The resulting values of shear γ and convergence κ are used to shear the galaxy according to the reduced shear g (§2) and to magnify the galaxy according to the magnification[42] $\mu = [(1-\kappa)^2 - \gamma^2]^{-1}$.

The GalSim "lensing engine" that carries out this process works in the flat-sky limit. It uses Fourier transforms, with a Fourier-space grid that is of equal size to the real-space grid, and hence the power is assumed to be zero for $|k| < 2\pi/L$ and $|k| > \sqrt{2}\pi/\Delta x$ (see Appendix I for details). This artificially reduces shear correlations on large scales by a significant amount compared to those in a realistic shear field (Rowe et al., *in prep.*). To address this limitation, we use an extended real-space grid for calculating shears, which lowers the minimum k represented in the power spectrum and preserves the shear correlations on scales corresponding to our box size.

Because of various effects that modify the power spectrum at levels up to a few percent (e.g., reduced shear, random chance in a single realization of the shear field, flat sky approach, and the limited Fourier space grid used to generate the shears), we do not compare submitted shear correlation functions with the *input* ideal ones, but rather with correlation functions that we estimate using the true reduced shears output from GalSim before they are actually applied to the galaxies.

5.6. *A note about physical scales*

In real images, galaxies may be quite close together (given typical number densities of ~ 20 arcmin^{-2}), yet in the case of variable shear fields we usually only estimate shear correlations for galaxies that are significantly farther apart than the average separation between galaxies. This fact has motivated GREAT10 and now GREAT3 to consider galaxy grids that are 10×10 deg^2 with 100×100 galaxies, not spending time simulating galaxies that are very close together.

However, for variable PSFs, much of the interesting PSF variation happens on smaller scales than the 0.1 deg grid spacing. This has motivated us to make each variable PSF branch contain ten fields of 20 subfields that sample the same cosmological shear field and PSF field, thus sampling the PSF field more densely than the cosmological shear field. This also aids us in the calculation of the metric, §4.4.

5.7. *Space versus ground*

Much of the difference between space-based and ground-based data comes from the different PSFs, as described in §5.2. The PSFs in space-based data are smaller and more stable over time than ground-based PSFs. However, there is an additional difference that is included in the GREAT3 challenge, related to the sampling of the images.

Data from existing optical space telescopes, as well as planned future telescopes, are typically undersampled due to the relatively large pixel scale compared to the PSF size. Sub-pixel dithers are used to recover Nyquist

[42] This procedure only includes changes in observed galaxy sizes and fluxes; it does not include the other important effect of magnification (the modification of the number density of objects due to the change in solid angles and the fact that galaxies get scattered across cuts in flux and apparent size by the magnification process).

sampled data after combining multiple exposures. However, since the combined image typically has a smaller pixel scale than the original image, the combined image has other features such as correlated noise (and depending on how the image combination is carried out, there might be some aliasing - see, e.g., Rowe et al. 2011).

In the control, realistic galaxy, and variable PSF experiments, the simulated data do not have multiple exposures. Thus, if we are simulating space-based data, we need some way to have that single exposure be Nyquist sampled. Our choices are (a) to simulate some realistic co-add over multiple single exposures, including effects like correlated noise, or (b) to simulate what would happen if our detectors had smaller pixels that allowed them to be Nyquist sampled while also having uncorrelated noise. We opt for choice (b). In the multi-epoch and full experiments, the individual exposures have pixel scales that are larger and hence not Nyquist sampled until all exposures are combined. In practice, we use pixel scales of $0.1''$ and $0.05''$ for simulated multi-epoch and single-epoch data, respectively.

In contrast, ground-based data is rarely undersampled, and we adopt a single pixel scale of $0.2''$ for the simulated ground-based data, regardless of whether it is single- or multi-epoch.

5.8. *Deeper data*

Many lensing surveys that are planned for the near- and far-future are designed with both "deep" and "wide" components. The "deep" components are typically a small subset (few percent) of the area of the "wide" component, but include enough observations to increase the S/N by a factor of several. These deep fields enable training methods to learn something about galaxy populations, which can then be used when interpreting the data in the (more cosmologically-interesting) wide survey.

To facilitate tests of such training methods, the GREAT3 challenge has additional simulations for each branch (corresponding to 2.5% of the volume of that branch, i.e., 5 images) that represent data that are one magnitude deeper (2.5 times higher S/N) than the rest of that branch, but are otherwise drawn from the same underlying galaxy population. The deeper data are not to be used to estimate shears, and results for them should not be submitted; they are purely for use as a training dataset.

In a real dataset, the deep survey would include a large fraction of galaxies that are not even detected in the wide survey. However, since we do not want most of the galaxies in the GREAT3 deep data to be useless, we only simulate the ones that would be observed in the rest of the GREAT3 challenge at $S/N > 20$. The galaxies that are simulated in the deep dataset still have resolution cuts imposed according to the PSF size in the deep dataset. The population is therefore identical, but with $S/N > 50$ in the deep dataset, which means that the effective deep data fraction is actually 5–7.5% rather than 2.5% This volume of deep data is actually relatively high compared to many planned surveys, but since the amount of deep data needed is still an open question, a test with this amount of deep data is quite useful.

Our interest in simulating a galaxy population in the challenge that goes to $I < 25$ with limiting $S/N = 20$, but to also have a subset of simulations in which the effective S/N limit *for the same population* is $S/N = 50$, poses a difficulty for our training dataset. The depth of our *HST* training dataset (§5.1 and Appendix E) is such that at $I = 25$, the images we observe have S/N below 50. We are forced to conclude that if we wish to have a limiting S/N of 50 in the deep simulations for a magnitude-limited parent sample from *HST*, we must use $I < 23.5$. This is relatively shallow compared to many extant and future lensing surveys, and hence somewhat undesirable. To ameliorate this issue, we developed a simple procedure to use the $I < 23.5$ sample to mimic the observed properties of an $I < 25$ sample by simple changes in flux and radius; a description and tests of this procedure are in Appendix E.3. This procedure does not preserve the *intrinsic* properties of the galaxies such as their redshift distribution, luminosity distribution, or intrinsic size distribution. However, it allows us to use the $I < 23.5$ sample to match the quantities that dictate the shear systematics for an $I < 25$ sample - namely the S/N, *observed* size, and observed morphology. This also helps address the concern raised in §5.1 that noise in the original *HST* images is treated as part of the galaxy; with the scheme described here, the added noise dominates over the noise that was already present for all galaxies.

6. SIMPLIFICATIONS

As described in §3.4, our goal of making simulations to test particular effects has led to some simplifications. Here we briefly mention several classes of problems that are left for future work:

- **Non-gridded galaxies:** Since our simulations include galaxies on grids at known locations, we do not test for issues due to blends (overlapping galaxy profiles).
- **Selection biases:** Imposition of selection criteria that might lead to the probability of a galaxy being selected to correlate with the shear or PSF direction are not tested in this scheme. However, if the weights used for a particular galaxy depend on the shear or PSF, then that form of selection bias will show up in the challenge results.
- **Wavelength-dependent effects:** Real PSFs are color-dependent at some low level. Since star and galaxy SEDs are not the same, this results in a different effective PSF for galaxies as for stars. Further complications arise due to color gradients within galaxies. None of these effects are included in GREAT3, but may be quite important for upcoming lensing surveys (Semboloni et al. 2013).
- **Instrument/detector specific effects:** There are a whole host of instrument and detector effects that are not included in GREAT3, for example cosmic ray hits, saturation, bad pixels or columns, scattered light, charge transfer inefficiency, and distortion. Because these effects are manifested in different instruments in different ways, it is difficult to simulate them in a generic context.
- **Star/galaxy separation:** In a realistic data analysis, it is necessary to determine from the data itself

which objects are galaxies (to be used for shear estimation) and which are stars (to be used for PSF modeling). Cross-contamination between the samples in either direction can cause biases in shear estimation. Since the GREAT3 challenge provides galaxy fields for shear estimation, and star fields for PSF estimation, star/galaxy separation is not explicitly tested by this challenge.

- **Background estimation:** The images in the GREAT3 challenge have essentially had the sky background level subtracted. In practice, the sky background level is unknown and spatially varying; incorrect removal (especially contamination by the extended light of galaxies) can lead to unsubtracted sky level gradients that mimic shear signals.
- **Low S/N:** As described in §5.1, we have strong logistical reasons to not include a substantial galaxy population below $S/N = 20$ in the GREAT3 challenge. However, it is likely that issues such as noise bias and underfitting bias are not completely decoupled, so it will be important in future to investigate their relationship to each other.
- **More complex noise model:** In real images, the noise may be more complex than the simple model adopted here (§5.3). Aside from the issue of spatially-varying sky background and Poisson noise from the galaxy flux, some steps in image processing can induce correlations in noise levels between pixels.
- **Redshift-dependent effects:** The GREAT3 challenge does not include redshift-dependent shears or allow for estimation of a redshift-dependent shear calibration.
- **Flexion:** The GREAT3 challenge assumes the shear is constant across each galaxy, so it does not include higher-order distortions such as flexion.

ACKNOWLEDGMENTS

The GREAT3 collaboration benefited greatly from discussions with the WFIRST project office, including David Content; Christopher Hirata; the Euclid Consortium; and the LSST imSim team, including En-Hsin Peng; and Peter Freeman. We thank the PASCAL-2 network for its sponsorship of the challenge. This work was supported in part by the National Science Foundation under Grant No. PHYS-1066293 and the hospitality of the Aspen Center for Physics.

This project was supported in part by NASA via the Strategic University Research Partnership (SURP) Program of the Jet Propulsion Laboratory, California Institute of Technology; and by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This article only reflects the authors' views.

RM was supported in part by program HST-AR-12857.01-A, provided by NASA through a grant from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Incorporated, under NASA contract NAS5-26555. BR and SB acknowledge support from the European Research Council in the form of a Starting Grant with number 240672. CH acknowledges support from the European Research Council under the EC FP7 grant number 240185.

REFERENCES

Abahamid, A., Jabiri, A., Vernin, J., et al. 2004, A&A, 416, 1193

Albrecht, A., Amendola, L., Bernstein, G., et al. 2009, ArXiv e-prints: 0901.0721, arXiv:0901.0721

Bartelmann, M., & Schneider, P. 2001, Phys. Rep., 340, 291

Bernstein, G. M. 2010, MNRAS, 406, 2793

Bernstein, G. M., & Armstrong, R. 2013, ArXiv e-prints: 1304.1843, arXiv:1304.1843

Bernstein, G. M., & Jarvis, M. 2002, AJ, 123, 583

Blandford, R., Haynes, M. P., Huchra, J. P., et al. 2010, New Worlds, New Horizons in Astronomy and Astrophysics (The National Academies Press)

Blanton, M. R., Eisenstein, D., Hogg, D. W., Schlegel, D. J., & Brinkmann, J. 2005, ApJ, 629, 143

Bosch, J. 2010, AJ, 140, 870

Bridle, S., Shawe-Taylor, J., Amara, A., et al. 2009, Annals of Applied Statistics, 3, 6

Bridle, S., Balan, S. T., Bethge, M., et al. 2010, MNRAS, 405, 2044

Bundy, K., Ellis, R. S., & Conselice, C. J. 2005, ApJ, 625, 621

Casaponsa, B., Heavens, A. F., Kitching, T. D., et al. 2013, MNRAS, 430, 2844

Chang, C., Kahn, S. M., Jernigan, J. G., et al. 2013, MNRAS, 428, 2695

Connolly, A. J., Peterson, J., Jernigan, J. G., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7738, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series

Content, D. A., Armani, N. V., Baker, C. L., Jackson, C. E., & M., K. D. 2013, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8860, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series

de Jong, R. S. 1996, A&AS, 118, 557

Dietrich, J. P., Werner, N., Clowe, D., et al. 2012, Nature, 487, 202

Duncan, C., Joachimi, B., Heavens, A., Heymans, C., & Hildebrandt, H. 2013, ArXiv e-prints: 1306.6870, arXiv:1306.6870

Ellerbroek, B. L., & Rigaut, F. J. 2000, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4007, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. P. L. Wizinowich, 1088–1099

Fried, D. L. 1965, Journal of the Optical Society of America (1917-1983), 55, 1427

Gruen, D., Seitz, S., Koppenhoefer, J., & Riffeser, A. 2010, ApJ, 720, 639

Hamana, T., Miyazaki, S., Okura, Y., Okamura, T., & Futamase, T. 2013, ArXiv e-prints: 1304.4992, arXiv:1304.4992

Hardy, J. W. 1998, Adaptive Optics for Astronomical Telescopes, ed. Hardy, J. W.

Heymans, C., Rowe, B., Hoekstra, H., et al. 2012, MNRAS, 421, 381

Heymans, C., Van Waerbeke, L., Bacon, D., et al. 2006, MNRAS, 368, 1323

Heymans, C., Grocutt, E., Heavens, A., et al. 2013, MNRAS, 432, 2433

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

Hirata, C. M., & Seljak, U. 2004, Phys. Rev. D, 70, 063526

Hirata, C. M., Mandelbaum, R., Seljak, U., et al. 2004, MNRAS, 353, 529

Hu, W. 2002, Phys. Rev. D, 66, 083515

Huff, E. M., & Graves, G. J. 2011, ArXiv e-prints: 1111.1070, arXiv:1111.1070

Huterer, D., Takada, M., Bernstein, G., & Jain, B. 2006, MNRAS, 366, 101

Jarvis, M., Bernstein, G., & Jain, B. 2004, MNRAS, 352, 338
Jarvis, M., Schechter, P., & Jain, B. 2008, ArXiv e-prints: 0810.0027, arXiv:0810.0027
Jee, M. J., Tyson, J. A., Schneider, M. D., et al. 2013, ApJ, 765, 74
Jurić, M., Ivezić, Ž., Brooks, A., et al. 2008, ApJ, 673, 864
Kacprzak, T., Zuntz, J., Rowe, B., et al. 2012, MNRAS, 427, 2711
Kaiser, N. 2000, ApJ, 537, 555
Kaiser, N., Squires, G., & Broadhurst, T. 1995, ApJ, 449, 460
Kilbinger, M., Schneider, P., & Eifler, T. 2006, A&A, 457, 15
Kitching, T., Balan, S., Bernstein, G., et al. 2011, Annals of Applied Statistics, 5, 2231
Kitching, T. D., Balan, S. T., Bridle, S., et al. 2012, MNRAS, 423, 3163
Kitching, T. D., Rowe, B., Gill, M., et al. 2013, ApJS, 205, 12
Koekemoer, A. M., Fruchter, A. S., Hook, R. N., & Hack, W. 2002, in The 2002 HST Calibration Workshop : Hubble after the Installation of the ACS and the NICMOS Cooling System, ed. S. Arribas, A. Koekemoer, & B. Whitmore, 337–+
Koekemoer, A. M., Aussel, H., Calzetti, D., et al. 2007, ApJS, 172, 196
Kolmogorov, A. 1992, Dokl. Akad. Nauk SSSR, 30, 301
Lackner, C. N., & Gunn, J. E. 2012, MNRAS, 421, 2277
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, ArXiv e-prints: 1110.3193, arXiv:1110.3193
Leauthaud, A., Massey, R., Kneib, J., et al. 2007, ApJS, 172, 219
Leauthaud, A., Tinker, J., Bundy, K., et al. 2012, ApJ, 744, 159
Lewis, A. 2009, MNRAS, 398, 471
Lin, H., Dodelson, S., Seo, H.-J., et al. 2012, ApJ, 761, 15
LSST Science Collaborations, & LSST Project. 2009, ArXiv e-prints: 0912.0201, http://www.lsst.org/lsst/scibook, arXiv:0912.0201
Mandelbaum, R., Hirata, C. M., Leauthaud, A., Massey, R. J., & Rhodes, J. 2012, MNRAS, 420, 1518
Mandelbaum, R., Seljak, U., Hirata, C. M., et al. 2008, MNRAS, 386, 781
Markwardt, C. B. 2009, in Astronomical Society of the Pacific Conference Series, Vol. 411, Astronomical Society of the Pacific Conference Series, ed. D. A. Bohlender, D. Durand, & P. Dowler, 251–+
Massey, R., & Refregier, A. 2005, MNRAS, 363, 197
Massey, R., Rowe, B., Refregier, A., Bacon, D. J., & Bergé, J. 2007a, MNRAS, 380, 229
Massey, R., Stoughton, C., Leauthaud, A., et al. 2010, MNRAS, 401, 371
Massey, R., Heymans, C., Bergé, J., et al. 2007b, MNRAS, 376, 13
Massey, R., Hoekstra, H., Kitching, T., et al. 2013, MNRAS, 429, 661
Melchior, P., Böhnert, A., Lombardi, M., & Bartelmann, M. 2010, A&A, 510, A75
Melchior, P., & Viola, M. 2012, MNRAS, 424, 2757
Miller, L., Heymans, C., Kitching, T. D., et al. 2013, MNRAS, 429, 2858
Miyatake, H., Nishizawa, A. J., Takada, M., et al. 2013, MNRAS, 429, 3627
Ngan, W., van Waerbeke, L., Mahdavi, A., Heymans, C., & Hoekstra, H. 2009, MNRAS, 396, 1211
Noll, R. J. 1976, Journal of the Optical Society of America (1917-1983), 66, 207
Peacock, J. A., Schneider, P., Efstathiou, G., et al. 2006, ESA-ESO Working Group on ”Fundamental Cosmology”, Tech. rep., arXiv:astro-ph/0610906
Peterson, J. R., et al. in prep.
Poyneer, L., van Dam, M., & Véran, J.-P. 2009, Journal of the Optical Society of America A, 26, 833
Refregier, A., Kacprzak, T., Amara, A., Bridle, S., & Rowe, B. 2012, MNRAS, 425, 1951
Rhodes, J., Refregier, A., & Groth, E. J. 2000, ApJ, 536, 79
Rhodes, J. D., Massey, R. J., Albert, J., et al. 2007, ApJS, 172, 203
Roggemann, M. C., & Welsh, B. 1995, Imaging through turbulence, ed. Roggemann, M. C. & Welsh, B.
Rowe, B., Hirata, C., & Rhodes, J. 2011, ApJ, 741, 46
Sasiela, R. J. 1994, Electromagnetic wave propagation in turbulence. Evaluation and application of Mellin transforms
Schmidt, F., Leauthaud, A., Massey, R., et al. 2012, ApJ, 744, L22
Schneider, P. 2006, in Saas-Fee Advanced Course 33: Gravitational Lensing: Strong, Weak and Micro, ed. G. Meylan, P. Jetzer, P. North, P. Schneider, C. S. Kochanek, & J. Wambsganss, 269–451
Schneider, P., van Waerbeke, L., Jain, B., & Kruse, G. 1998, MNRAS, 296, 873
Schneider, P., van Waerbeke, L., & Mellier, Y. 2002, A&A, 389, 729
Scoville, N., Abraham, R. G., Aussel, H., et al. 2007a, ApJS, 172, 38
Scoville, N., Aussel, H., Brusa, M., et al. 2007b, ApJS, 172, 1
Semboloni, E., Hoekstra, H., Huang, Z., et al. 2013, MNRAS, 432, 2385
Sérsic, J. L. 1968, Atlas de galaxias australes (Cordoba, Argentina: Observatorio Astronomico, 1968)
Simard, L., Mendel, J. T., Patton, D. R., Ellison, S. L., & McConnachie, A. W. 2011, ApJS, 196, 11
Sirianni, M., Clampin, M., Hartig, G. F., et al. 1998, in Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, Vol. 3355, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. S. D'Odorico, 608–612
Spergel, D., Gehrels, N., Breckinridge, J., et al. 2013, ArXiv e-prints: 1305.5422, arXiv:1305.5422
Taylor, G. I. 1938, Royal Society of London Proceedings Series A, 164, 476
Tokovinin, A., & Travouillon, T. 2006, MNRAS, 365, 1235
Vale, C., Hoekstra, H., van Waerbeke, L., & White, M. 2004, ApJ, 613, L1
van Waerbeke, L. 2010, MNRAS, 401, 2093
Van Waerbeke, L., Benjamin, J., Erben, T., et al. 2013, MNRAS, arXiv:1303.1806
Velander, M., van Uitert, E., Hoekstra, H., et al. 2013, ArXiv e-prints: 1304.4265, arXiv:1304.4265
Vernin, J., Agabi, A., Avila, R., et al. 2000, AURA report
Viola, M., Melchior, P., & Bartelmann, M. 2011, MNRAS, 410, 2156
Voigt, L. M., & Bridle, S. L. 2010, MNRAS, 404, 458
Vorontsov, A. M., Paramonov, P. V., Valley, M. T., & Vorontsov, M. A. 2008, Waves in Random and Complex Media, 18, 91
Zhang, J., & Komatsu, E. 2011, MNRAS, 414, 1047
Zuntz, J., Kacprzak, T., Voigt, L., et al. 2013, ArXiv e-prints: 1302.0183, arXiv:1302.0183

APPENDIX

EXISTING APPROACHES TO SHEAR MEASUREMENT

Initially, the field of weak lensing was dominated by methods that involved applying a correction to the weighted second moments of the galaxy image to account for smearing by the PSF. These early methods include KSB (Kaiser et al. 1995) and implicitly make unrealizable assumptions about the nature of the galaxy and PSF: that they have concentric isophotes (Massey et al. 2007b), and small intrinsic ellipticities (Viola et al. 2011). Since then, the weak lensing community has made significant progress in developing additional PSF-correction methods. Like KSB, some of those methods also start with measuring moments of the galaxy and PSF, with some prescription for correcting the former to account for the latter (e.g., Kaiser 2000; Rhodes et al. 2000; Hirata & Seljak 2003).

Other methods are based on forward modeling of the intrinsic galaxy profile, including some methods that carry out < 10 parameter fits for an astrophysically-motivated galaxy model (e.g., Zuntz et al. 2013 and Miller et al. 2013), and others that decompose the galaxy images into an orthonormal basis set (e.g., Bernstein & Jarvis 2002, Massey & Refregier 2005, and Ngan et al. 2009), requiring many more parameters but also allowing a lot more flexibility for describing complex galaxies. Additionally, several methods have gone in newer directions such as Fourier-space approaches and non-parametric methods (Bernstein 2010; Zhang & Komatsu 2011; Bernstein & Armstrong 2013). For measurements of constant shears, image stacking methods (which were highly successful in the GREAT08 Challenge, Lewis 2009; Bridle et al. 2010), have a clear potential application.

Several studies have assessed the limitations of previous methods and devised ways of compensating for them (e.g., Bernstein & Jarvis 2002, Hirata & Seljak 2003, Bosch 2010, and Viola et al. 2011). A relatively new development in the GREAT10 challenge (Kitching et al. 2012) was the introduction of several methods using techniques from machine learning and computer science, such as the use of training methods (neural network and lookup table approaches, e.g., Gruen et al. 2010).

Because of the wealth of information about these methods in the literature and in the summaries of the GREAT08 and GREAT10 challenges, we refer the interested reader to the relevant papers and references therein for more details on modern shape measurement methods.

ESTIMATING THE SHEAR CORRELATION FUNCTION

For the variable shear branches, we cannot use the average shear as a useful metric to decide whether a given set of shear measurements match the input shear field. In fact, the average input shear is zero by construction. Instead, we use the two-point correlation function of the shear field. This statistic is commonly used in weak lensing cosmic shear studies as the lowest order description of the shear field in a given patch of sky. For a Gaussian field, as is used in this challenge, it encapsulates all of the measurable information about the shear field.[43]

As the name implies, the "two-point" correlation function involves an average over all pairs of two shear measurements. The math is simplest if we treat the shears as complex numbers, $g = g_1 + \mathrm{i}g_2$. Because of the complex nature of shear, there are actually two shear correlation functions, ξ_+ and ξ_-, defined as follows:

$$\xi_+(r) = \langle g(\mathbf{x}) g^*(\mathbf{x}+\mathbf{r}) \rangle \tag{B1}$$

$$\xi_-(r) = \langle g(\mathbf{x}) g(\mathbf{x}+\mathbf{r}) e^{-4\mathrm{i}\alpha} \rangle \tag{B2}$$

where the averages are over all pairs of measured shear values, α is the polar angle of $\mathbf{r}$ and * indicates complex conjugation.

Both ξ_+ and ξ_- are complex-valued by construction, but they are both effectively real in practice. In fact, ξ_+ is identically real if the average is allowed to count each pair of galaxies twice, letting the two shear values swap places for the second counting. The expectation value of ξ_- is real for shear fields that are parity invariant. That is, if the shear field is statistically identical after being reflected along some axis, then the imaginary part of ξ_- has an expectation value of 0, and deviations from this value in a particular realization of a shear field can be discarded as meaningless.

To measure the shear correlation function, we use a public, open-source software package, called `corr2`[44]. It uses a ball-tree algorithm to avoid having to calculate the product of every pair of galaxies individually. Essentially, it calculates the shear products for groups of galaxies that have nearly the same separation vector, and thus belong in the same final bin. For more details on the algorithm, see Jarvis et al. (2004).

Another relevant property of shear fields is that they can be divided into so-called E-mode and B-mode components (see §2.2). As discussed in §5.3, in our simulated shear fields, the lensing shear is constructed to be purely E-mode, whereas the shape noise (due to galaxy intrinsic shapes) is almost purely[45] B-mode. Thus, separating the measurements into E-mode and B-mode components allows us to mostly remove the largest source of noise in the measurement, which lets us use far fewer galaxies than we would otherwise need to achieve a given statistical precision.

The method we use to perform the separation is called the "aperture mass statistic", as discussed in §4.4.2. The information in the shear field can be divided into E-mode and B-mode components via the following formulae:

$$M_E(\theta) = \frac{1}{2}\int_0^\infty r\,\mathrm{d}r \left[\xi_+(r) T_+\left(\frac{r}{\theta}\right) + \xi_-(r) T_-\left(\frac{r}{\theta}\right) \right] \tag{B3}$$

$$M_B(\theta) = \frac{1}{2}\int_0^\infty r\,\mathrm{d}r \left[\xi_+(r) T_+\left(\frac{r}{\theta}\right) - \xi_-(r) T_-\left(\frac{r}{\theta}\right) \right] \tag{B4}$$

[43] The actual cosmic shear field of the universe is not Gaussian, so higher order statistics such as three-point correlation functions and shear peak statistics, among others, are also used to characterize the non-Gaussian features in the shear field.

[44] Available at https://code.google.com/p/mjarvis/

[45] "Almost" because it turns out not to be possible to make the shape noise pure B-mode while maintaining other features that we wanted to have, such as a Gaussian input field and the galaxy shape distribution matching the real galaxy shapes.

where

$$T_+(x) = \frac{x^4 - 16x^2 + 32}{128} \exp\left(-\frac{x^2}{4}\right) \tag{B5}$$

$$T_-(x) = \frac{x^4}{128} \exp\left(-\frac{x^2}{4}\right) \tag{B6}$$

The integrals formally go from $r = 0$ to ∞; however, the weight functions T_+ and T_- go to zero quickly for large values of x. At $r = 0$, T_+ goes to a constant, so E vs. B mode separation formally requires integration over the correlation functions to zero separation[46]. For GREAT3, we get around this difficulty by knowing the correlation function of the true input shear field at scales smaller than the closest pairs of galaxies in the simulation. When we receive a submission consisting of the measured ξ_+ and ξ_-, we can use the true values for the parts of these statistics that are unmeasurable from the data. The measured correlation functions are used for the bulk of the range of integration, so the correction is small.

In practice, the measured correlation functions are measured only at specific logarithmically binned values. Thus we convert Eqs. B3 and B4 into sums over those binned values using the simplest possible approximation (constant $\xi(r)$ within each bin). While this procedure would be problematic for a cosmological analysis, leading to deviations from the true underlying aperture masses that are more than several percent, it is not a problem for the challenge because we can apply the same procedure to the true input shears before comparing with the M_E and M_B submitted by participants.

CHALLENGE RULES

Here we describe the rules related to participation in the challenge.

Teams

Participants can register on the leader board webpage using a user name and e-mail address (no full name required). The e-mail address must be a real one, as it will be used to communicate information related to the challenge (but will not be shared/used for any other purposes). This is particularly important since the simulations may be updated as needed during the challenge if problems are found, and participants will need to know about these modifications. At the close of the challenge, participants will be asked to reveal their identities and participate in the writing of a results paper. The first and second place winners must reveal their identities in order to receive the prizes, and they are strongly encouraged to present their method at the GREAT3 final meeting, for which travel support will be available. During the course of the challenge, participants are encouraged to describe their method(s) on the wiki at the public GitHub repository described in §4.2. However, before writing papers based on the GREAT3 challenge results while the challenge is still ongoing, please write to challenge@great3challenge.info to consult with the leaders of the challenge. After the challenge ends, participants are encouraged to write papers based on the results, preferably citing the official challenge results paper.

Submissions are to be made by teams, which can include any number of people on them. Likewise, people may be on any number of teams. Teams are permitted to submit results labeled as different methods. On any given branch, only the top-ranked method for any team will appear on the leader board, though the other submissions are stored for later reference and interpretation of results.

Submissions should be considered new methods when the algorithms have some new element involved. A team with several methods (in terms of algorithm or basic assumptions) may be ranked on the leader boards for different branches with different methods[47]. The points for those top rankings with different methods are considered when determining where that team is ranked on the overall leader board (§C.2). Forming a new team to submit "new" methods without any significant differences[48] in order to push other teams off the leader boards is grounds for disqualification from the challenge; we reserve the right to disqualify teams for other malicious behavior as well. However, there is an element of choice here: for a truly different method, people may decide whether they wish to submit it as the same team or as a different one. On any given branch, teams are limited to one submission per day. During the course of the challenge, participants are welcome to form new teams by opening issues on the public GitHub repository described in §4.2; for example, someone who only has shape measurement code might try to form a team with someone who has PSF estimation code in order to participate in the "variable PSF" branch.

Teams that include ≥ 1 participant on the "GREAT3 executive committee" (which gives them access to privileged information about the simulations) are flagged as such. This means that while they appear on the individual leader boards, they do not receive an official ranking (e.g., if their metric is at the top, they will appear in the top position, but the first unflagged person is the one who is ranked as first place on that board when it comes to determining points for the overall leader board, §C.2). Such teams appear with starred scores on individual branch leader boards,

[46] See Kilbinger et al. (2006) for more discussion of this difficulty for cosmological surveys.

[47] Note that this is a change from GREAT08 and GREAT10, which ranked methods rather than teams. The reason for the change is that, given the large variation in simulated data types (constant and variable shear, space and ground data), we want to allow the possibility that one team might have two or more "specialist" methods that only handle certain data types, but do so very well. In our scheme, the high rankings for those methods can be combined to allow this team to win.

[48] See the leader board website for examples of what constitutes a different method.

and do not appear on the overall leader board at all. The list of executive committee members is on the FAQ at the public GitHub site[49]; however, it is the responsibility of the committee members to identify themselves as such at the time their team is formed. If an executive committee member wishes to join a team after the time of its formation, they should e-mail the challenge e-mail address given in §4.2 to change that team's status to "flagged".

Overall leader board

To create an overall leader board, we award points to each team based on their rankings on the individual branches. Each team is awarded points based on their best-ranked 5 branches (or less than 5, if they submit to less than 5 branches); we award 1,000 points for a fifth-place finish, 2,000 points for fourth, 4,000 points for third, 8,000 points for second, and 16,000 points for first. The team with the highest total number of points is the winner. In the case of a tie, the total number of points from *all* submitted branches will be totaled for the tied teams, and the team with the most total points wins. If the teams are still tied, then as a second-level tie-breaker they will be ranked by the earliest submission time stamp among the branches that contributed points to the tiebreaker (any branch in which the team placed in the top 5), with earlier time stamps winning over later time stamps. We performed 10,000 Monte Carlo simulations of possible challenge scores, and found that our ranking method did not strongly favor specific distributions of scores, although it rewards those who are consistent across all branches slightly more than those who specialize in an experiment or observation type. Limiting to the best-ranked 5 branches helps reduce the impact of the number of submissions per team: around 60% of the simulated teams submitted to 10 or fewer branches, and they made up around 40% of winners, compared to 15% of winners when we did not limit to the best-ranked 5 branches per team. With this method, almost 60% of the first-place teams did not place first in any individual branch.

VALIDATION OF GALSIM SHEARING PRECISION

Here we provide evidence that GalSim can be used to create sheared images of parametric and realistic galaxies using the DFT method of image rendering (§5.4) for the purpose of testing weak lensing algorithms.

Our first test is for Sérsic profiles. GalSim can render Sérsic profiles in two ways: via photon-shooting or the DFT approach. For the former approach, the primary approximation is the use of a lookup table to represent the radial profile when sampling the photons. The shearing, convolution with a PSF, and binning into pixels is in principle exactly represented with photon offsets. For DFT, there are more approximations: we have to represent the Fourier-space profile as a lookup table, but shearing also assumes that we are in the regime where DFTs can substitute for continuous Fourier transforms. There are thus additional caveats for the DFT approach, and they are in principle independent of the issues that can arise from photon-shooting. Thus our first test for Sérsic profiles is whether sheared Sérsic profiles agree when generated using these two methods. For this purpose we make images in both ways, measure their shears using adaptive moments, and define a STEP-like calibration bias (see §4.4),

$$\gamma_{\rm DFT} - \gamma_{\rm phot} = m_{\rm DFT}\gamma_{\rm phot} + c_{\rm DFT}. \tag{D1}$$

Our target level of accuracy for shear testing with GREAT3 is that we would like to test for calibration biases and additive systematics at the level needed for Euclid (Massey et al. 2013), $m \sim 2\times10^{-3}$ and $c \sim 2\times10^{-4}$. Thus we would like our simulation software to produce spurious shears that are a factor of ≥ 10 below that, i.e., $m_{\rm DFT} < 2\times10^{-4}$.

We carried out a test of shearing accuracy for each of the two shear components, alongside a similar test for the correct rendering of galaxy size. Several values of Sérsic indices n were investigated, for a range of galaxy half light radii and intrinsic ellipticities $\varepsilon^{(s)}$ drawn from a random sample of 30 single component Sérsic model fits to the COSMOS training data sample. For each galaxy a circular profile was first sheared to create an object with ellipticity $\varepsilon^{(s)}$, convolved with a COSMOS-like PSF, then rendered as an image via both DFT and photon shooting. Differences between moment estimates of the resulting ellipticities are plotted in Fig. 10. As shown, the values of $m_{\rm DFT}$ demonstrate that we can consistently represent galaxy shears at the few $\times10^{-6}$ level for $n = 1.5$, with $m_{\rm DFT}$ rising as high as $\sim 1\times10^{-4}$ for the highest $n = 6.2$ (note that for GREAT3 we use $n \leq 6$). These values are safely below our target values of 2×10^{-4}. Since errors in DFT and photon-shooting are completely independent, it is highly improbable that this good agreement is due to chance, and it supports our claim that we can accurately shear galaxies rendered via DFT for the GREAT3 challenge.

As a parallel investigation, we also estimated the size of the galaxies in the images described above, using adaptive moments. By fitting a slope to the differences between DFT and photon shooting results, we can also estimate the accuracy at which weak lensing magnifications can be simulated using GalSim. We found a slope of $m_{\rm DFT} = (4.7 \pm 0.7)\times10^{-5}$ for the Sérsic $n = 1.5$ galaxy sample, $m_{\rm DFT} = (-1.4 \pm 3.6)\times10^{-5}$ for the $n = 4.5$ sample, and $m_{\rm DFT} = (-3.4 \pm 8.1)\times10^{-5}$ for the $n = 6.2$ sample. These results are safely below the 2×10^{-4} target adopted for multiplicative-style biases in the simulation of shear. Indeed, the signal-to-noise expected for cosmological magnification measurements has been estimated as $\lesssim 50\%$ relative to shear (e.g. van Waerbeke 2010; Schmidt et al. 2012; Duncan et al. 2013) motivating a corresponding relaxation of requirements by a factor $\gtrsim 2$ (although this figure is dependent both on the dataset and analysis technique used, see e.g. Huff & Graves 2011). These results suggest that the representation of galaxy sizes with GalSim therefore falls comfortably within requirements for future surveys.

Next, we show that we can accurately shear more complex, realistic galaxy images. For the GREAT3 challenge, we must remove the *HST* PSF, shear and magnify, and convolve with the target PSF. In this case there is no ground

[49] https://github.com/barnabytprowe/great3-public/wiki/Frequently-Asked-Questions

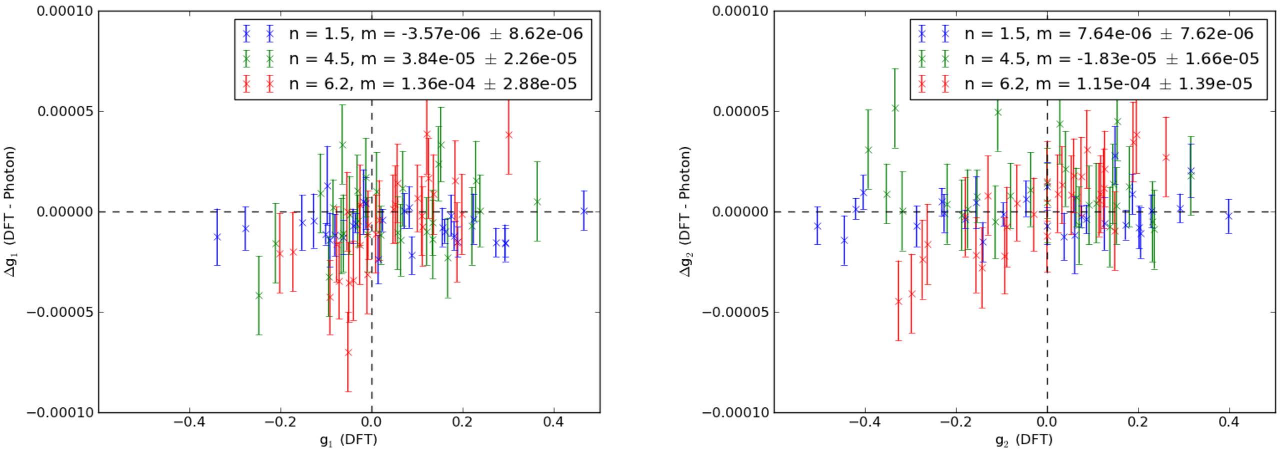

FIG. 10.— Difference between measured shears (upper panel: g_1; lower panel: g_2) for Sérsic profiles simulated using the two methods of image rendering in GalSim, photon-shooting and DFT, plotted against the shear measured from the DFT image. Results are shown for 30 galaxies with realistic size and shape distribution, and several Sérsic n values shown in the legend. The points can be fit to lines to measure $m_{\rm DFT}$ as defined in Eq. D1, and the best-fit values are shown in the legend.

truth. Instead, we begin with a simpler test for which we do have ground truth: we treat the *HST* PSF for our training galaxies as part of the galaxy itself. In that case, we can compare the shape of the original image (*HST* PSF included) with the shape when we shear it by a small, known amount. We carry out this test for simulated Sérsic profile images and for realistic *HST* galaxies, and ensure that the recovered shears are as expected despite the need to carry out interpolation to do the shearing. In this case, since the original galaxies are not round, we calculate the difference between the observed shear (after applying a shear) and the expected one given the intrinsic shear and the applied one, and define calibration bias due to interpolation $m_{\rm interp}$ as

$$\gamma_{\rm obs} - \gamma_{\rm expected} = m_{\rm interp}\gamma_{\rm applied} + c_{\rm interp}. \tag{D2}$$

We find for $m_{\rm interp}$ for the two shear components is $(-3.1 \pm 0.3) \times 10^{-4}$ and $(-1 \pm 2) \times 10^{-5}$. The former is around our threshold, and the latter is well below it. These numbers are all using the default interpolants and settings in GalSim; future work might involve refining these, but they are clearly adequate for the levels of calibration bias that can be detected in GREAT3. $c_{\rm interp}$ is of order 10^{-10}, consistent with zero within the errors.

In addition, we check for leakage between shear components (i.e., that applying one shear component does not result in an incorrect level of shear in the other component). For example, we define cross terms such as

$$\gamma_{\rm obs,1} - \gamma_{\rm expected,1} = m_{\rm interp}^{(1,2)}\gamma_{\rm applied,2} \tag{D3}$$

and likewise for leakage between magnification and shear. We find that $m_{\rm interp}^{(1,2)}$ and $m_{\rm interp}^{(2,1)}$ are of order 1×10^{-5}.

Our final test is to show that we can successfully carry out the process of "reconvolution" (Mandelbaum et al. 2012) using GalSim - that is, when we say we are representing what some galaxy looks like with an added shear $\gamma_{\rm applied}$ when viewed at lower resolution, is that statement correct? This test was carried out using simulated Sérsic profiles at high resolution, putting them through the reconvolution process and ensuring that the result looks like what we simulate by taking the original Sérsic profile and viewing it directly at low resolution. We quantify any error in the effectively applied shear due to the reconvolution process as $m_{\rm reconv}$, defined by

$$\gamma_{\rm reconv} - \gamma_{\rm direct} = (1 + m_{\rm reconv})\gamma_{\rm applied}. \tag{D4}$$

$m_{\rm reconv}$ was determined for 270 galaxies randomly selected from the training sample described in Appendix E, for a space-based and a ground-based target PSF. As for previous tests, our target value is $m_{\rm reconv} < 2\times10^{-4}$. Since galaxies with different light profiles might be more or less difficult to accurately render using reconvolution, we consider not only the mean $\langle m_{\rm reconv}\rangle$ but also its standard deviation, as an indicator of possible galaxy types for which the method fails to work sufficiently accurately even if it works for most galaxies. For the default GalSim settings used for the GREAT3 simulations, we find $\langle m_{\rm reconv}\rangle$ is completely consistent with zero, with a standard deviation of 3×10^{-5}, well below our target value of $m_{\rm reconv}$ of 2×10^{-4}. This result shows that any profile inaccuracies due to the reconvolution process do not interfere with our ability to accurately render what a galaxy looks like with a particular shear, even for different galaxy types.

The results in this section use the default set of parameters for DFT and photon-shooting accuracy in GalSim; more detailed investigations will be presented in Rowe et al., *in prep.*.

REAL GALAXY DATASET

Here we describe the dataset used to simulate a realistic galaxy population in the GREAT3 challenge.

HST *training sample*

The training sample that is compiled here comes from the COSMOS survey, using galaxy selection criteria from Mandelbaum et al. (2012), as summarized below.

The COSMOS *Hubble Space Telescope* (*HST*) Advanced Camera for Surveys (ACS) field (Koekemoer et al. 2007; Scoville et al. 2007b,a) is a contiguous 1.64 degrees^2 region centered at R.A.=10:00:28.6, Dec.=+02:12:21.0 (J2000). Between October 2003 and June 2005 (*HST* cycles 12 and 13), the region was completely tiled by 575 adjacent and slightly overlapping pointings of the ACS Wide Field Channel. Images were taken through the wide F814W filter ("Broad I"). We use the 'unrotated' images (as opposed to North up) to avoid rotating the original frame of the PSF. The raw images are corrected for charge transfer inefficiency (CTI) following Massey et al. (2010). Image registration, geometric distortion, sky subtraction, cosmic ray rejection and the final combination of the dithered images are performed by the multidrizzle algorithm (Koekemoer et al. 2002). As described in Rhodes et al. (2007), the multidrizzle parameters have been chosen for precise galaxy shape measurement in the co-added images. In particular, a finer pixel scale of $0.03''$/pix was used for the final co-added images (7000 × 7000 pixels).

The following cuts are then applied on catalogs derived from the COSMOS images; for more details on the flags, see Leauthaud et al. (2007):

- $F814W < 25.2$: This cut corresponds to a S/N limit of ~ 20. However, as discussed in §5.8, we only use those galaxies at < 23.5 for GREAT3, applying simple transformations (§E.3) to mimic a fainter sample.
- MU_CLASS = 1: This requirement uses the relationship between the object magnitude and peak surface brightness to select galaxies, and to reject other objects.
- CLEAN = 1: This cut is required to eliminate galaxies with defects due to very nearby bright stars, or other similar issues.
- GOOD_ZPHOT_SOURCE = 1: This cut requires that there be a good photometric redshift, which typically is equivalent to requiring that the galaxy not be located within the masked regions of the ground-based $BVIz$ imaging used for photometric redshifts. We impose this cut here because we wish to test the galaxy population going into our simulations to ensure that it is representative of reality, and having a photometric redshift estimate is an important part of those tests.

Following the procedure in Mandelbaum et al. (2012) for a brighter subset of the data, postage stamps were cut out around the position of each galaxy. The background level was subtracted, and additional objects besides the central one were masked with a correlated noise field with the same properties as the noise in the rest of the image. As in that work, in order to remove the effects of the COSMOS PSF, we use PSF models from a modification of version 6.3 of the Tiny Tim ray-tracing program[50]. These models represent PSFs for different primary/secondary separation, since that separation is the main determinant of the PSF ellipticity; while imperfect particularly at long wavelengths (Sirianni et al. 1998), the Tiny Tim PSFs are close enough to reality to use in our simulations. Future work will include empirically-estimated PSFs.

As described in e.g. Mandelbaum et al. (2008), the COSMOS field is small enough that, when measuring quantities as a function of redshift in small redshift bins, large-scale structure in the field induces non-negligible noise in the results. This should also be the case when using it as a training sample to estimate shear calibration as a function of redshift: the intrinsic ellipticity distribution can differ in dense and underdense environments due to their different galaxy populations, so for narrow redshift slices, the shear calibration would reflect those different populations. However, here we are using the COSMOS sample to measure the shear calibration for some redshift-averaged population, such that the large-scale structure fluctuations in narrow Δz slices effectively cancel out. As a result, we do not impose any density-dependent weighting on the sample.

Parametric fits

We fit the galaxies in the training set with parametric models. The functional form is given by a Sérsic profile (Sérsic 1968). The radial surface brightness profile is

$$I(R) = I_{1/2} \exp\left[-b_n \left((R/R_{\rm eff})^{1/n} - 1\right)\right] \quad , \tag{E1}$$

$$R = \Big[((x - x_0)\cos\phi + (y - y_0)\sin\phi)^2 \tag{E2}$$

$$+ ((y - y_0)\cos\phi - (x - x_0)\sin\phi)^2 / q^2\Big]^{1/2} \quad ,$$

where $R_{\rm eff}$ is the half-light radius, $I_{1/2}$ is the surface brightness at the half-light radius, n is the Sérsic index, and b_n is a normalization factor dependent on the Sérsic index. The radius, R, defines an ellipse, with minor-to-major axis ratio

[50] http://www.stsci.edu/software/tinytim/

$q = b/a$. We fit each galaxy twice: once with a Sérsic profile, and once with a de Vaucouleurs ($n = 4$) bulge profile plus an exponential ($n = 1$) disk profile. The fitting method is described in detail in Lackner & Gunn (2012). The Sérsic profile contains 7 free parameters: $R_{\rm eff}$, $I_{1/2}$, n, the central position (2 parameters), the axis ratio of elliptical isophotes, and the position angle. We place some constraints on the fitted parameters. The surface brightness must be positive, the Sérsic index is between 0.1 and 6 (following Blanton et al. 2005), the axis ratio $0.05 \leq q \leq 1$, and the size of the galaxy must be smaller than the the size of the postage stamp. As in Lackner & Gunn (2012), the Sérsic models cutoff smoothly at large radii. The cutoff radius varies smoothly from 4 half-light radii for $n = 1$ to 8 half-light radii for $n = 4$.

The bulge+disk models have 10 free parameters, since we fix the Sérsic indices of both components ($n = 4$ for the bulge and $n = 1$ for the disk) and require that the bulge and disk share the same centroid. In addition, we require that the bulge half-light radius is less than that of the disk. Previous studies have shown that varying the bulge Sérsic index does not yield statistically significantly better fits for the typical galaxy in this sample (Simard et al. 2011; Lackner & Gunn 2012; de Jong 1996).

The best-fit parameters are found using a 2-dimensional Levenberg-Marquardt minimization, `mpfit2dfun` in IDL (Markwardt 2009). The fitter minimizes the weighted sum of the squared differences between the galaxy image and PSF-convolved model. This χ^2-minimization method assumes the pixel values are uncorrelated, which is not true for the *HST* postage stamps. The weights are given by the inverse variance in each pixel, including sky noise and photon noise from the source. Although this is the optimal weighting scheme for a least χ^2 fit, it does introduce changes in the weighting scheme as functions of galaxy brightness. Faint galaxies are fit with constant, sky-noise-dominated weights, while bright galaxies are down-weighted in the central regions. The initial values for the minimization are obtained from an exponential profile fit to the galaxy.

We have tested the fits by creating mock single-Sérsic and bulge+disk images at various resolutions and S/N using GalSim. The fitter recovers the correct input parameters for the relevant range of resolution and S/N, although the uncertainties grow as the S/N decreases (Lackner, et al. in prep). For simulated galaxies where the ratio of bulge flux to total flux (B/T) is between 0.3 and 0.7, the error in B/T varies from 0.10 to 0.17 as the S/N decreases from 100 to 50, typical for galaxies in the HST training sample. The errors in bulge half-light radius are typically $25 - 44\%$, while the errors in the disk half-light radius are always smaller, $10 - 30\%$. These uncertainties depend most strongly on S/N, not B/T, when B/T is far from both 0 and 1. For single-component simulated galaxies, the uncertainties in galaxy sizes are smaller still, ranging from $2 - 10\%$, depending on S/N, input galaxy size, and Sérsic index. The errors in size are usually 3 times larger for galaxies with Sérsic index near 4 than for bulge-less exponential disks. In all cases, down to $S/N = 50$, the mean offsets between the measured values and the input parameters are within 1 standard deviation.

For our simulations, we use the bulge+disk model described above, except for cases where B/T is below 0.1 or above 0.9; or where the bulge radius or axis ratio runs up against the fit limits (e.g., $q_{\rm bulge} = 0.05$ precisely). In those cases, the galaxy is dominated by a single component, and it occasionally happens that the subdominant component has extremely large radius and low surface brightness (absorbing some sky gradient), or otherwise poorly constrained parameters. Thus we do not use the bulge+disk fits for these edge cases, and instead use the single Sérsic fits with free n. We also require that the median absolute deviation or MAD be lower for the 2-component fits, otherwise we just use the single component fits. After all cuts, we use two-component fits for the $\sim 1/3$ of the sample for which they seem justified, and single Sérsic fits for the rest. Note that the fits do include populations of galaxies with $0.1 < B/T < 0.2$ with bulges with unusual properties (e.g., $q \sim 0.1$); visual inspection suggests that these are not fit failures but rather the fitter attempting to represent bars or the beginnings of spiral arms using a "bulge" component. Thus we do not attempt to remove these fits.

Fig. 11 shows some properties of the sample based on these fits. Of particular note is the middle right panel, which compares the position angles for bulge and disk shapes in the two-component fits. As shown, the bulge and disk have a significant tendency to be aligned with each other, but some non-negligible offsets are allowed and thus will be represented in our simulations even for the control experiment.

Mimicking a fainter sample

Since we wish to use a sample with robust two-component fits to make the simulations, but want to simulate a galaxy sample that is typical for deeper surveys ($F814W \sim 25$), we have used a sample with $F814W < 23.5$ to mimic a deeper sample with a limit of < 25.2. In details of intrinsic properties like redshift, this would be a difficult task. However, we largely wish to reproduce the *observed* properties of the sample that determine shear calibration, including the distributions of S/N, apparent size, intrinsic ellipticity, and morphology. To test our ability to do this, we use the fewer-parameter single-Sérsic fits (and the B/T from the double Sérsic fits described in the above subsection) and demonstrate that a very simple prescription enables us to achieve our goal.

Fig. 12 shows the distributions of B/T, Sérsic n and half-light radius, and photo-z from the catalogs of Leauthaud et al. (2012). The histograms of these properties are shown as a function of magnitude in bins shown in the upper left panel, and for the sample overall in the solid black line. As shown, the histograms of B/T, Sérsic and n are largely independent of magnitude[51]. However, we see the expected trends that fainter galaxies are (a) at higher redshift and (b) smaller in size. As noted previously, (a) does not affect shear calibration per se, but rather the true shear

[51] Another important property, the ellipticity distribution, is also consistent with being independent of apparent magnitude (modulo noise, which increases scatter towards high ellipticity in a well-understood way; e.g., Leauthaud et al. 2007). We have not plotted this

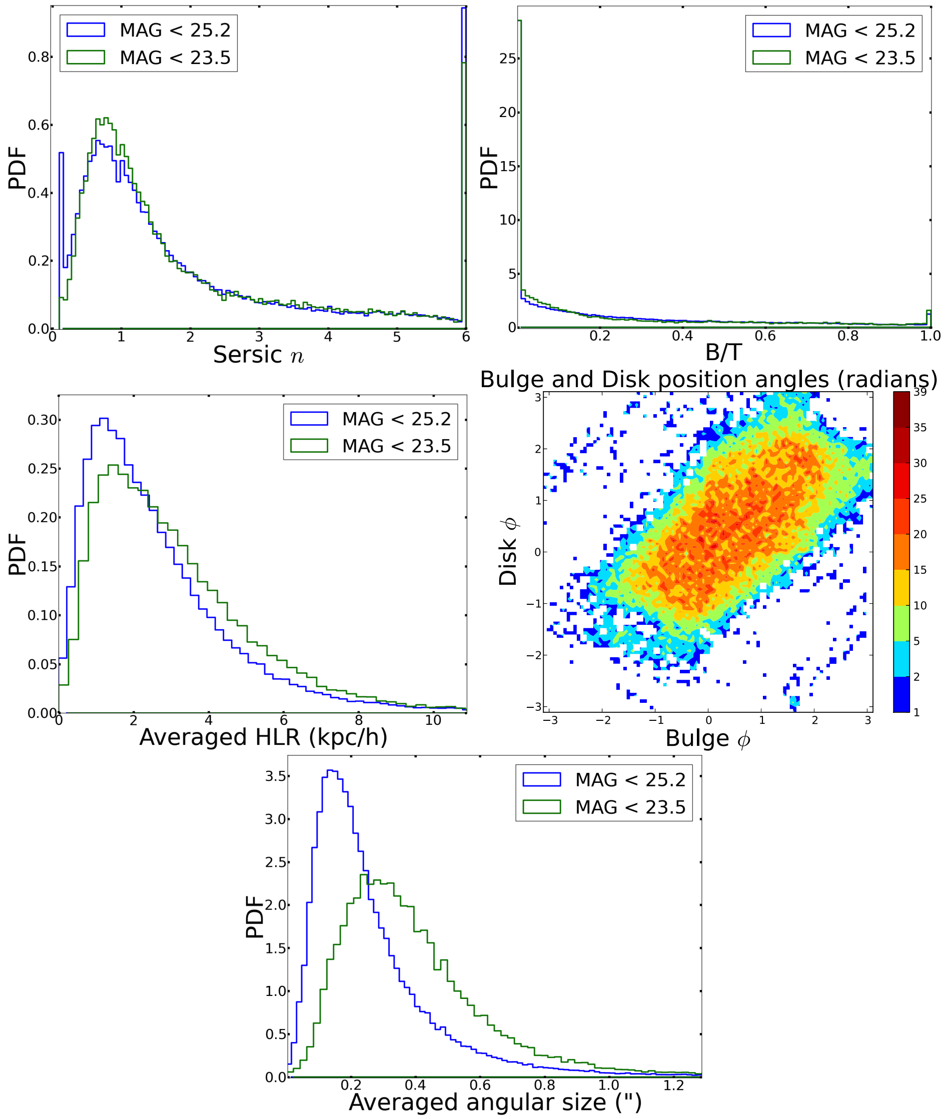

FIG. 11.— For two magnitude cuts, we show distributions of various galaxy properties from the fits described in appendix E.2. **Top left:** Distribution of Sérsic n values, with a slight tendency to pile up at the lower and upper limits for the fainter sample (due to noise). **Top right:** Distribution of bulge-to-total flux ratio B/T from the two-component fits. **Middle left:** Distribution of physical half-light radius, which tends to smaller values for the fainter sample because that one has more low luminosity objects. **Middle right:** For the sample limited at 25th magnitude, for galaxies with two significant components, this plot shows density contours for the joint distribution of bulge and disk position angles. Bottom:Distribution of angular half-light radius.

experienced by a galaxy; hence if we are trying to calibrate some average shear calibration, we do not need to reproduce distributions of photo-z. Therefore, our key challenge is to get the size and S/N distribution of the < 23.5 sample to look like that of the < 25.2 sample.

We find that a simple mapping that involves reducing the flux and decreasing sizes by a factor of 0.6 is sufficient to make the < 23.5 sample look statistically like the < 25.2 sample in terms of the distributions of apparent size, S/N, n, B/T, and ε. The 2d distributions of properties for the < 25.2 sample (Fig. 13) are almost completely reproduced by the < 23.5 sample if we make this transformation. The one exception to this statement is a slight difference in the ellipticity distribution; the "fake" sample has fewer high-ellipticity objects. However, since the existence of those high-ellipticity objects in the faint sample is consistent with being caused by noise, it may actually be a benefit that our "fake" deep sample does not contain them.

Thus, for all branches, we always apply this transformation factor of 0.6 to the observed sizes in the $I < 23.5$ sample to mimic a deeper sample.

We have not confirmed that this scheme reproduces the fraction of irregular galaxies, since we have no good way of quantifying irregularity for the faint sample. It is likely that our procedure slightly under-represents the population of irregulars, which is known to increase at higher redshift. This means that our conclusions about realistic galaxy

quantity, because the ellipticity distribution is so central to shear inference, and as for real data, challenge participants must carry this out themselves from the simulated data.

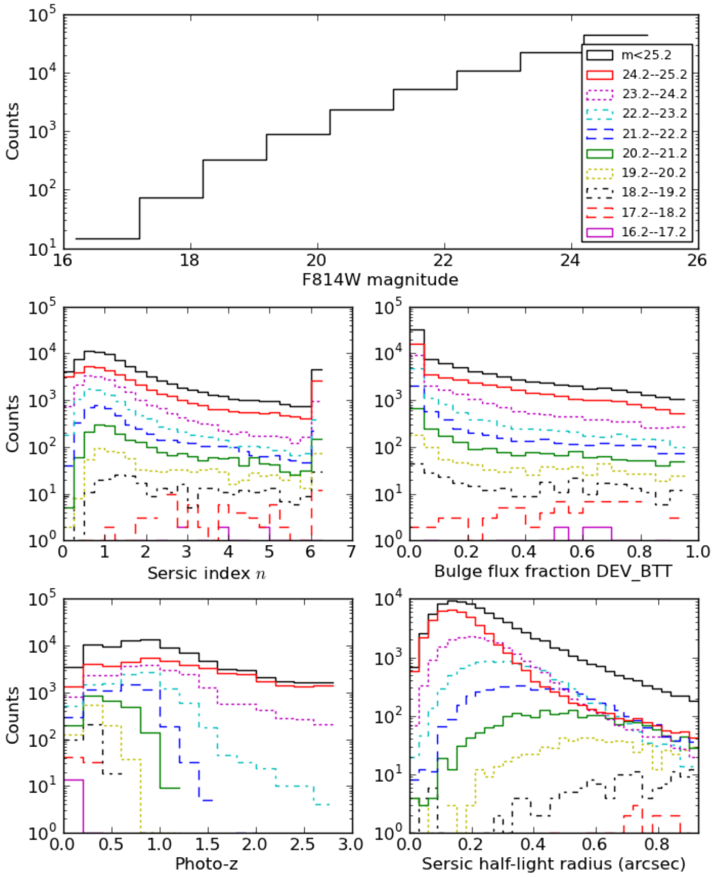

FIG. 12.— Histograms of the properties of training sample galaxies for magnitude bins defined in the uppermost panel.

morphology might slightly underestimate that in reality. However, this seems preferable to the alternative of using rather noisy galaxy images as the basis for our simulations, effectively considering the non-negligible noise as part of the galaxies.

OPTICAL PSF MODELS

Here we describe the specific optical PSF models used for the "variable PSF" experiment, for simulated ground- and space-based data.

For the simulations that mimic observations from a space telescope, we have secured an approximate description of the design residual (§5.2.1) of a prototype telescope model for the 2.4m WFIRST-AFTA mission[52] (Content et al. 2013). The model consists of a Zernike polynomial description of wavefront errors up to order $j = 11$ in the notation of Noll (1976), and therefore contains trefoil and third-order spherical aberration (but no higher order aberrations). This Zernike approximation to the design residual was provided at a series of fixed locations in the WFIRST-AFTA FOV, and we interpolate the Zernike terms between these locations to provide a fully continuous approximate model of a space telescope. Additional aberrations, to model those due to misalignment or figure errors, were included as additions to these Zernike terms (also up to order $j = 11$ only). Values of these additional Zernike aberrations were chosen such that the ensemble root mean square wavefront error added was $\lambda/13$, where λ is the wavelength of the light being observed. This is a relatively stringent operational definition of a diffraction-limited optical system, and a target for aberrations due to misalignment and figure errors for space missions such as WFIRST-AFTA.

In addition to these aberrations, the WFIRST-AFTA prototype model includes 6 non-radial struts (i.e., ones that do not go directly across the center of the aperture). GalSim is currently only able to simulate PSFs with radial struts, so that is the model we use for GREAT3. Since jitter can be directional but typically not with a preferred direction over long time-scales, we model jitter as convolution with a Gaussian with RMS of 0.005–0.015$''$ per axis, with ellipticity

[52] http://wfirst.gsfc.nasa.gov/

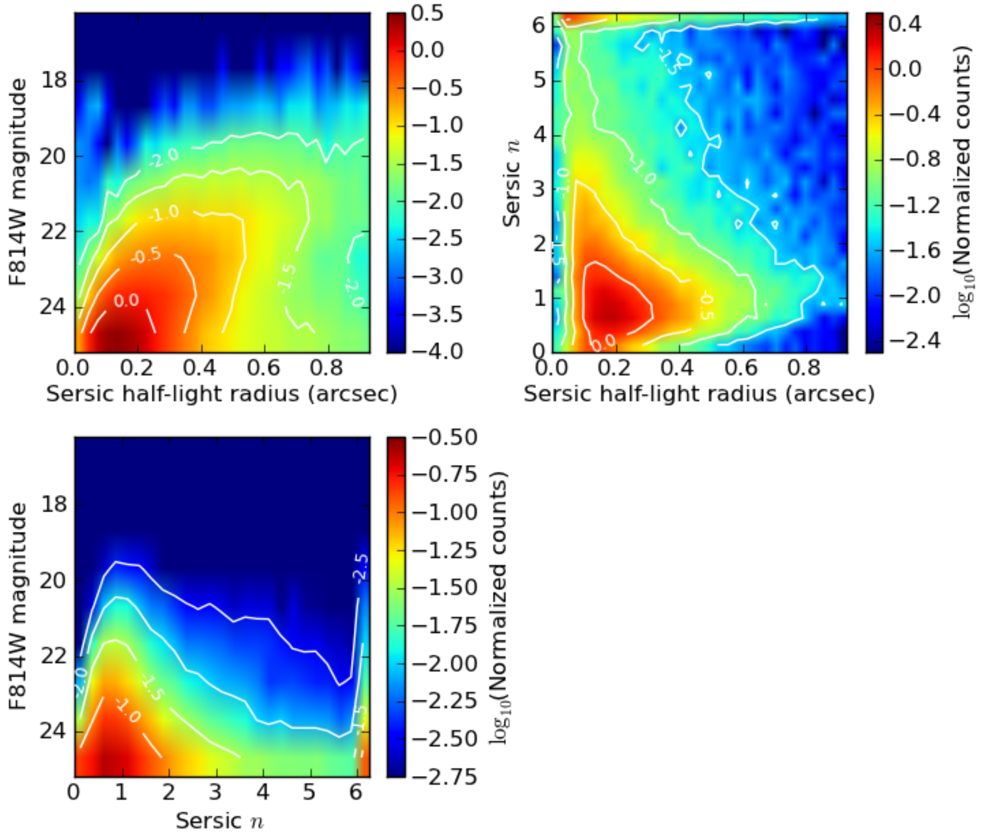

FIG. 13.— Two-dimensional contour plots showing the relationship between $F814W < 25.2$ training galaxy properties, shown on a logarithmic scale.

from 0 to 0.3 but random direction. In contrast, charge diffusion often has some preferred direction, so we model it as a Gaussian with $\sigma = 0.05$-0.2 pixels, with ellipticity from 0-0.2, always in the same direction. Like the additional aberrations, the jitter and charge diffusion parameters are chosen for each field as a whole, and for a given epoch, they are the same for all subfields within the field.

We need to determine a size for the tiles within the 10×10 deg^2 images that will represent individual fields of view for the PSF model. The WFIRST-AFTA model is defined within a 0.42×0.42 deg^2 FOV; we artificially stretch these length scales to 0.5×0.5 deg^2, which means we can tile a 10×10 deg^2 region with 400 PSF tiles in the space-based simulations.

For the simulations that mimic observations from a ground-based telescope, we use an approximate description of the design residual of an early model[53] for the Dark Energy Camera (DECam) at the Blanco Telescope in Chile. This model differs in some respects from the one that was actually used, but it is nonetheless a reasonable optical PSF model for an instrument on a 4m telescope. As for the WFIRST-AFTA PSF model, we restrict ourselves to a Zernike polynomial description of wavefront errors up to order $j = 11$ at a series of fixed locations in the FOV, between which we interpolate the Zernikes. We add additional aberrations to the ground-based PSF model to represent misalignment and tilt, based on a model for DECam determined using extra-focal imaging (code to be included in a future version of GalSim). This model is defined over a 1.56×1.56 deg^2 field of view; however, for convenience, we stretch all length scales so that it is 2×2 deg^2, which allows us to use 25 optical PSF tiles within a 10×10 deg^2 image.

DESIGN AND IMPLEMENTATION OF THE ATMOSPHERIC PSF MODEL

The PhoSim atmospheric model

Here we give more details on the atmospheric model used for PhoSim, which we use as the basis for GREAT3 as described in §5.2.2. In this model, a set of frozen Kolmogorov screens (Kolmogorov 1992) are distributed vertically above the telescope (representing the column of air above the telescope). For the work described here, the atmospheric model assumes 7 atmospheric layers at altitudes of 16km, 8km, 4km, 2km, 1km, 0.5km, and 0.02km (ground layer),

[53] S. Kent and M. Gladders, priv. comm.

each having different strengths. During the time of the exposure, the screens move according to the wind conditions at different altitudes. As photons propagate through different parts of the screen at different times, their trajectories are perturbed by an amount depending on the wavelength and the value of the screen at that location. This simulates the refraction of light as it passes though air of different densities (and thus refractive index). Atmospheric dispersion is included by scaling this perturbation according to the wavelength and zenith angle, as the screens represent a thicker layer of air when the telescope is pointed away from zenith.

The "frozen screen approximation" is justified since the time scale for the shapes of turbulent cells to change significantly is much longer than the time required for those cells to pass through the field of view, given the typical wind speeds of a few meters per second (Taylor 1938; Poyneer et al. 2009). These atmospheric screens are constructed according to a full three-dimensional van Karman power spectrum (see, e.g., Sasiela 1994) with assigned parameters including the structure function, inner scale, outer scale, wind speed and wind direction. Adopting the model of Tokovinin & Travouillon (2006), PhoSim uses 7 atmospheric layers (including ground layer), each layer has the effective physical size of $\sim 2.6 \times 2.6$ km^2 and resolution of $\sim 1 \times 1$ cm^2. Since storing all the information in these large turbulent screens while ray-tracing is practically impossible, PhoSim adopts the technique used by Vorontsov et al. (2008) and splits the full van Karman power spectrum into three, each containing a smaller range of scales. PhoSim then generates three much smaller screens with these piecewise power spectra and only registers the value of the "combined screen" on the fly as the photon hits a specific pixel on the screen.

Since the specific model parameters vary from site to site, PhoSim uses parameters based on atmospheric data taken close to the LSST site, Cerro Pachon, Chile (2738 m above sea level, $70°44'01''$W, $30°14'17''$S) in order to simulate LSST data. Since this site also hosts the 8-meter Gemini-South telescope and the 4-meter SOAR telescope, which are both equipped with adaptive optics instruments, relatively complete atmospheric data and literature can be found (Vernin et al. 2000; Abahamid et al. 2004; Ellerbroek & Rigaut 2000). Note, however, that dynamic information about the variation in timescales shorter than a day is currently lacking.

Estimation of atmospheric PSF parameters

In this section we explain how we derived the ellipticity values and the spatial variation of the ellipticity and size of the atmospheric PSF using PhoSim.

We used PhoSim to make simulated images of exposure time[54] 10s, 20s, 60s, 120s, each covering 0.5×0.5 deg^2, with stars that have $S/N \sim 1000$ on a regular grid of 0.5 arcmin grid spacing. For each exposure time we made 10 $r-$band images with different seeing/wind/structure functions based on a random seed. We then carried out several tests on the resulting images.

We calculate the star ellipticities, e (Eq. 6), and their correlation functions; examples are in Fig. 14. The zero-lag value of the correlation function is simply the variance of the shapes, i.e., $\langle e_1^2 + e_2^2 \rangle$ averaged over all simulated stars in the exposure. The salient features of this plot are

- There is substantial range in the amplitude and slope of this correlation function between individual realizations.
- The amplitude of the ellipticity variance is relatively small, typically in the range 10^{-4} to 10^{-3}.
- The shapes are coherent to quite large scales, an effect that has been seen even for simulations of a larger area than shown here. (See Fig. 9 for an example of how the shape correlation functions compare to those for lensing shear or optical PSFs.)

In practice, we adopted a functional form that can describe these atmospheric PSF correlations in a set of images covering a larger 2×2 deg^2 field, and on even larger scales in PhoSim. That functional form has two parameters: the overall PSF ellipticity variance and a length scale determining how quickly the correlations die off with scale. For a given exposure in the variable PSF experiment, two random numbers are chosen for those parameters with a flat distribution (also consistent with the simulations) to determine a PSF ellipticity correlation function. The amplitude also scales inversely with the exposure time and the telescope diameter. For the GREAT3 simulations, we choose random values of exposure times of 60–180s.

The spatial variation of the PSF size also follows a similar correlation function. Typical fractional fluctuations in size are a few tenths of a percent.

GalSim implementation of the atmospheric PSF

To simulate the atmospheric PSF for the challenge, we use the GalSim software that takes a lensing shear power spectrum and uses it to simulate galaxy shears[55]. The physics in the two cases is the same, except that for the atmosphere, there is equal power in E and B modes, whereas lensing only generates E modes. Moreover, the fluctuations in PSF size across the FOV are sourced by the same physical source of the E-mode anisotropies, so we can use the "convergences" from the GalSim outputs as fractional changes in PSF size. Using this software requires us to convert the PSF anisotropy correlation function that we use to a power spectrum, via numerical integration. We tabulate

[54] Note that real observations are typically anywhere from one to several minutes long, though PhoSim typically runs with 15s exposures to match the LSST observation plan.

[55] This is significantly less computationally intensive than making a large number of PhoSim simulations directly, and in the long-exposure limit we can make simple models that capture the relevant physics at the level needed for this challenge.

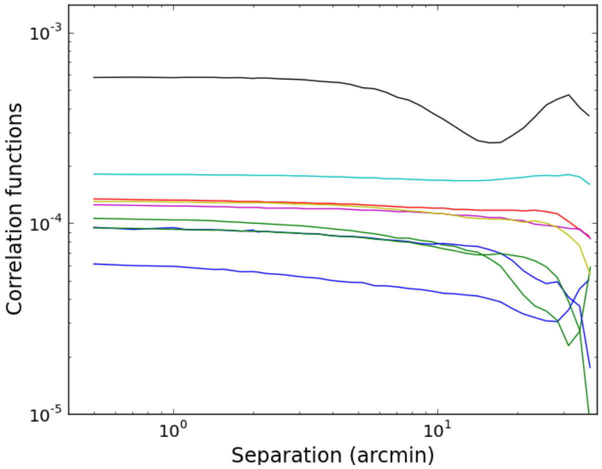

FIG. 14.— For several simulated atmospheric PSF fields from the LSST PhoSim with 60s exposure times, we show the correlation function of PSF shapes defined as ξ_+ (Eq. B1). ξ_- (Eq. B2) is consistent with zero and therefore not shown. Different line colors are different realizations with randomly-chosen atmosphere parameters.

the power spectra for logarithmically-spaced values of correlation length; for some random value of correlation length, we use the nearest one for which the power spectrum was tabulated. The GalSim lensing code can then generate a random realization of a gridded shear field with very large spatial extent (to avoid issues with cutoffs in correlations at the edges of our image, §5.5) using the chosen power spectrum divided into half E-mode and half B-mode power. We use this gridded shear field along with simple assumptions described at the end of §5.2.2 to generate the atmospheric PSF as a function of position in the field of view.

B-MODE SHAPE NOISE

As described in §5.3, in order to maintain a reasonable simulation volume for the GREAT3 challenge, we need a way to remove the intrinsic galaxy shape noise from the quantity of interest, the reconstructed shear correlation function. However, the scheme described there of using 90-degree rotated galaxy pairs does not work for spatially varying shear fields (see, e.g., Appendix A of Kitching et al. 2011). As in the GREAT10 challenge, we adopt a scheme to ensure that the intrinsic shape noise only shows up in the B mode shear correlation function, whereas the lensing shear is only E mode.

However, the situation in GREAT3 is somewhat more complex because the galaxy $p(|\varepsilon^{(s)}|)$ is determined for us by the galaxy training sample that we are reproducing. Whereas in GREAT10 it was possible to generate a Gaussian random field of pure B-mode intrinsic ellipticities with an appropriate variance, in GREAT3 we only have the freedom to choose the orientations, not the ellipticity magnitudes, of the galaxies which we simulate. Without altering the $p(|\varepsilon^{(s)}|)$ for our training sample, which we wish to avoid doing as the $p(|\varepsilon^{(s)}|)$ is an important characteristic of realistic galaxy populations for weak lensing, it is impossible to avoid some leakage of shape noise into E-modes where it increases the uncertainty on GREAT3 submission results.

The extent of this B-mode leakage can be reduced using a prescription we now describe. First, an estimate of ε is made for every galaxy in the COSMOS training sample, using the second moments of high resolution images of the model fits described in Appendix E. Taking these estimates of $\varepsilon^{(s)}$ for the training sample, we calculate the variance in each component of ellipticity, $\mathrm{Var}[\varepsilon_1^{(s)}]$ and $\mathrm{Var}[\varepsilon_2^{(s)}]$. These determine the variance σ_I of the pure B-mode, constant power spectrum, Gaussian random field to use as a 'target' for the intrinsic galaxy shapes in the simulations, $\sigma_I = \mathrm{Var}[\varepsilon_1^{(s)}] + \mathrm{Var}[\varepsilon_2^{(s)}]$.

We label as b the resultant ellipticities for a given realization of this target B-mode Gaussian field (using the ε convention for ellipticity, see §2.1). We seek to put down source galaxies from the training set with $\varepsilon^{(s)}$ as close to b as needed to ensure negligible B-mode leakage. The $p(|b|)$ is, by definition, a Rayleigh distribution with $\sigma = \sigma_I$. Comparison with histograms of the training sample $|\varepsilon^{(s)}|$ showed reasonable, but not perfect, agreement between the distributions. This provided encouragement that it might be possible to generate a field of galaxies with nearly-pure B-mode intrinsic shapes by appropriate selection from the training sample, followed by rotation (we are free to rotate our source galaxies to align their ellipticities with b). The procedure adopted was then as follows:

1. For each simulation field, a realization of pure B-mode ellipticities is generated as a Gaussian random field, yielding a target ellipticity b_j at each of $j = 1., \ldots, N$ galaxy positions in the field. We note that here the subscript j does not denote shear component.

2. A sample of N galaxy models are drawn from the full training sample, with replacement. These models have estimated ellipticities $\varepsilon_k^{(s)}$.

3. The ranked ordering of b_j by ascending $|b_j|$ is determined by sorting; the ranked ordering of $\varepsilon_k^{(s)}$ by ascending $|\varepsilon_k^{(s)}|$ is determined similarly.

4. At each galaxy position with target ellipticity b_j the source galaxy for which $|\varepsilon_k^{(s)}|$ took the same ordered rank as $|b_j|$ is selected, and assigned to this position.

5. This source galaxy is then rotated so that its ellipticity $\varepsilon_k^{(s)}$ is aligned with b_j.

This procedure yielded samples of source galaxies with intrinsic ellipticities that were acceptably close to being a pure B-mode signal, while maintaining the real $p(|\varepsilon^{(s)}|)$ from the training set. Simulations using the COSMOS training sample demonstrated a leakage into the E-mode that was a factor of 7-8 smaller in variance than the expected shot noise σ_n^2 due to noisy pixels (e.g., §4.4.2), which is a tolerable contribution to the overall uncertainty.

SIMULATING SHEAR FIELDS ON FINITE GRIDS

The Fourier space analogue of the shear correlation function is the power spectrum $P(k)$, which describes the variance of a shear field in Fourier modes as a function of the angular wavenumber $k = |\mathbf{k}|$ on the sky.

An approximate simulation of a random shear field according to a specified power spectrum is straightforward using the Discrete Fourier Transform (DFT). Inherent in the approach is that the underlying shear must be approximated as a Gaussian random field, and values of the shear are provided only at grid points of fixed spatial separation Δx, which we label γ_{ij}.

The DFT of these shears, $\tilde{\gamma}_{ij}$ for discrete wavenumbers k_{ij}, can be generated as complex random variables subject to the constraint that $\left\langle |\tilde{\gamma}_{ij}|^2 \right\rangle = (\Delta k)^2 \, P(k_{ij})$, where Δk is the grid spacing in Fourier space ($\Delta k = 2\pi/L$, where L is the spatial extent of the grid in real space, and we assume a square grid for simplicity). Drawing Gaussian random deviates so that

$$\tilde{\gamma}_{ij} = \Delta k \sqrt{\frac{P(k_{ij})}{2}} \left[N(0,1) + \mathrm{i} N(0,1) \right] \tag{I1}$$

satisfies this constraint. Applying the inverse DFT to such a realization yields γ_{ij} with a power spectrum that can be directly related to a periodic sample of the desired $P(k)$. Provided Δx and Δk are sufficiently small, this will be a good approximation to the desired shear field. For more details, see https://github.com/GalSim-developers/GalSim/blob/master/devel/modules/lensing_engine.pdf.