






# Beyond the 3rd moment: a practical study of using lensing convergence CDFs for cosmology with DES Y3

D. Anbajagane (சுயாா) <sup>1,2</sup>★ C. Chang <sup>1,2</sup> A. Banerjee <sup>3</sup> T. Abel,<sup>4,5,6</sup> M. Gatti,<sup>7</sup> V. Ajani <sup>8</sup>,  
A. Alarcon <sup>9</sup>, A. Amon,<sup>10,11</sup> E. J. Baxter,<sup>12</sup> K. Bechtol,<sup>13</sup> M. R. Becker,<sup>9</sup> G. M. Bernstein,<sup>7</sup> A. Campos,<sup>14</sup>  
A. Carnero Rosell,<sup>15,16,17</sup> M. Carrasco Kind,<sup>18,19</sup> R. Chen,<sup>20</sup> A. Choi,<sup>21</sup> C. Davis,<sup>5</sup> J. DeRose,<sup>22</sup>  
H. T. Diehl,<sup>23</sup> S. Dodelson,<sup>14,24</sup> C. Doux,<sup>7,25</sup> A. Drlica-Wagner,<sup>1,2,23</sup> K. Eckert,<sup>7</sup> J. Elvin-Poole,<sup>26</sup>  
S. Everett,<sup>27</sup> A. Ferté,<sup>6</sup> D. Gruen,<sup>28</sup> R. A. Gruendl,<sup>18,19</sup> I. Harrison,<sup>29</sup> W. G. Hartley,<sup>30</sup> E. M. Huff,<sup>27</sup>  
B. Jain,<sup>7</sup> M. Jarvis,<sup>7</sup> N. Jeffrey,<sup>31</sup> T. Kacprzak,<sup>8</sup> N. Kokron,<sup>4,5</sup> N. Kuropatkin,<sup>23</sup> P.-F. Leget,<sup>5</sup>  
N. MacCrann,<sup>32</sup> J. McCullough,<sup>5</sup> J. Myles,<sup>4,5,6</sup> A. Navarro-Alsina,<sup>33</sup> S. Pandey,<sup>7</sup> J. Prat,<sup>1,2</sup> M. Raveri,<sup>34</sup>  
R. P. Rollins,<sup>35</sup> A. Roodman,<sup>5,6</sup> E. S. Rykoff,<sup>5,6</sup> C. Sánchez,<sup>7</sup> L. F. Secco,<sup>2</sup> I. Sevilla-Noarbe,<sup>36</sup>  
E. Sheldon,<sup>37</sup> T. Shin,<sup>38</sup> M. A. Troxel,<sup>20</sup> I. Tutusaus,<sup>39,40,41</sup> L. Whiteway,<sup>31</sup> B. Yanny,<sup>23</sup> B. Yin,<sup>14</sup>  
Y. Zhang,<sup>42,43</sup> T. M. C. Abbott,<sup>42</sup> S. Allam,<sup>23</sup> M. Aguena,<sup>16</sup> O. Alves,<sup>44</sup> F. Andrade-Oliveira,<sup>44</sup> J. Annis,<sup>23</sup>  
D. Bacon,<sup>45</sup> J. Blazek,<sup>46</sup> D. Brooks,<sup>31</sup> R. Cawthon,<sup>47</sup> L. N. da Costa,<sup>16</sup> M. E. S. Pereira,<sup>48</sup> T. M. Davis,<sup>49</sup>  
S. Desai,<sup>50</sup> P. Doel,<sup>31</sup> I. Ferrero,<sup>51</sup> J. Frieman,<sup>2,23</sup> G. Giannini,<sup>52</sup> G. Gutierrez,<sup>23</sup> S. R. Hinton,<sup>49</sup>  
D. L. Hollowood,<sup>53</sup> K. Honscheid,<sup>54,55</sup> D. J. James,<sup>56</sup> K. Kuehn,<sup>57,58</sup> O. Lahav,<sup>31</sup> J. L. Marshall,<sup>59</sup>  
J. Mena-Fernández,<sup>36</sup> F. Menanteau,<sup>18,19</sup> R. Miquel,<sup>52,60</sup> A. Palmese,<sup>14</sup> A. Pieres,<sup>16,61</sup>  
A. A. Plazas Malagón,<sup>5,6</sup> K. Reil,<sup>6</sup> E. Sanchez,<sup>36</sup> M. Smith,<sup>62</sup> M. E. C. Swanson,<sup>63</sup> G. Tarle,<sup>44</sup>  
and P. Wiseman<sup>62</sup> (DES Collaboration)

*Affiliations are listed at the end of the paper*

Accepted 2023 October 9. Received 2023 October 7; in original form 2023 August 10

## ABSTRACT

Widefield surveys probe clustered scalar fields – such as galaxy counts, lensing potential, etc. – which are sensitive to different cosmological and astrophysical processes. Constraining such processes depends on the statistics that summarize the field. We explore the cumulative distribution function (CDF) as a summary of the galaxy lensing convergence field. Using a suite of  $N$ -body light-cone simulations, we show the CDFs’ constraining power is modestly better than the second and third moments, as CDFs approximately capture information from all moments. We study the practical aspects of applying CDFs to data, using the Dark Energy Survey (DES Y3) data as an example, and compute the impact of different systematics on the CDFs. The contributions from the point spread function and reduced shear approximation are  $\lesssim 1$  per cent of the total signal. Source clustering effects and baryon imprints contribute 1–10 per cent. Enforcing scale cuts to limit systematics-driven biases in parameter constraints degrade these constraints a noticeable amount, and this degradation is similar for the CDFs and the moments. We detect correlations between the observed convergence field and the shape noise field at  $13\sigma$ . The non-Gaussian correlations in the noise field must be modelled accurately to use the CDFs, or other statistics sensitive to all moments, as a rigorous cosmology tool.

**Key words:** large-scale structure of Universe – cosmology: observations.

## 1 INTRODUCTION

The structure in the Universe – namely the distribution of matter – contains significant information on all kinds of physical processes; from the largest cosmological scales that probe the initial conditions of the Universe, to the galaxy and halo scales that probe both nonlinear gravitational evolution and baryonic imprints due to astrophysical

processes, to the intragalaxy scales where the gas and stellar phase space exhibit distinct structures from the rich physics of magneto-hydrodynamics. It is clear that the observed fields are abundant with information on both cosmology and astrophysics. It is then pertinent to question how best to extract the information from these fields, i.e. how best to maximize the constraints we can place on physical phenomena through measurements of these fields.

In the scenario where the field is a mean-zero Gaussian random field that is isotropic and homogeneous, the only degree of freedom

\* E-mail: [dhayaa@uchicago.edu](mailto:dhayaa@uchicago.edu)

for the field is the covariance between the pixels/voxels in real space (or alternatively, the power spectra in Fourier space). In such a scenario, it is clear that the maximal constraining power is obtained by measuring the power spectra, i.e. the only degree of freedom. For cosmological fields, the initial conditions seeding structure formation are Gaussian to a very good approximation, as has been verified by the cosmic microwave background (CMB) observations (Planck Collaboration 2016b, 2020), and a large part of the cosmological information in the resulting late time density field is still Gaussian, i.e. encoded in the variance of the field. Thus, the power spectra are a good way to extract information from the late-time fields as well.

However, there still remains significant, additional information beyond the power spectra. Even in the fiducial  $\Lambda$ CDM case – where  $\Lambda$ CDM is the cosmological model with cold dark matter (CDM) and the cosmological constant  $\Lambda$  – and the initial conditions contain no primordial non-Gaussianities, the presence of nonlinear, gravitational evolution generates signatures beyond the power spectra. This is commonly called ‘higher-order information’<sup>1</sup> and represents information in the field that is not captured by the power spectra. Such information still encodes signatures from cosmological and astrophysical processes, and is often highly complementary to the 2-point constraints; as a result, the combination of power spectra with higher-order information leads to constraints that are better than the trivial sum of the individual parts (e.g. Fluri et al. 2018, 2019, 2022; Gatti et al. 2020; Zürcher et al. 2021; Gatti et al. 2022; Lanzieri et al. 2023).

There exists a rich body of literature on different, complementary ways to extract this non-Gaussian information from continuous scalar fields like the density field or the weak lensing convergence field. The  $N$ -point correlation functions (or their Fourier equivalents, the poly-spectra) are the most well known and widely used statistic, and measure the correlation of  $N$  points in space, where the points are separated by some distances. For  $N = 3$ , these statistics are computationally expensive to compute, and for  $N = 4$  they are mostly prohibitive unless measured in specific limiting cases. Given this, many alternative methods have been explored to capture some/all of this information in a computationally inexpensive way. Some of the most commonly known/used methods include moments (Petri et al. 2015; Chang et al. 2018; Peel et al. 2018; Gatti et al. 2020, 2022), Minkowski Functionals (Mecke, Buchert & Wagner 1994; Blake, James & Poole 2014; Petri et al. 2015; Parroni et al. 2020), density-split statistics (Friedrich et al. 2018; Gruen et al. 2018) and more. Similar statistics exist for the discrete fields, such as counts-in-cells (Baugh, Gaztanaga & Efstathiou 1995; Adelberger et al. 1998) and the  $k$ -nearest neighbour (kNN) distributions (Banerjee & Abel 2021a, b). For the weak lensing field, the 3-point information has been pursued either through the direct measurement (Fu et al. 2014; Secco et al. 2022b) or approximate summaries like the density-split statistics (Friedrich et al. 2018; Gruen et al. 2018), mass aperture moments (Secco et al. 2022b), field moments (Petri et al. 2015; Gatti et al. 2020, 2022), and integrated shear functions (Halder et al. 2021). Weak-lensing peaks (Kratovich, Haiman & May 2010; Martinet et al. 2018; Shan et al. 2018; Zürcher et al. 2022) probe a specific, fixed combination of  $N$ -point functions, as is the case with other statistics like cosmic void distribution functions (Davies et al. 2021) and persistent homology (Heydenreich, Brück & Harnois-Déraps

2021; Heydenreich et al. 2022). Field-level inference tools are also employed (Fluri et al. 2018, 2019, 2022; Jeffrey et al. 2020), while others explore machine learning-informed, but still interpretable, statistics such as scattering transforms (Cheng & Ménard 2021) and wavelet phase harmonics (Allys et al. 2020).

An outstanding question is identifying the ‘maximally’ informative statistic for summarizing, and extracting constraints from, the fully nonlinear late-time density/convergence field. This is an unsolved problem given we do not a priori know the exact cosmological information contained in the different non-Gaussian signatures (including those beyond the 3-point function) across both linear and nonlinear scales. Thus, to ensure we use all the available cosmological information in the field, it is desirable to consider statistics that capture all orders of statistical information (rather than just one order, or a specific combination of orders). The kNN distributions have been formally shown to be such a statistic for discrete tracers (Banerjee & Abel 2021a) as they capture volume integrals of all  $N$ -point auto/cross-correlation functions of the field. While these kNN distributions are constructed for discrete tracer fields, Banerjee & Abel (2023) demonstrated that the analogous statistic for continuous fields are the CDFs of the field smoothed on different length-scales.

The CDFs – or the probability distribution functions (PDFs), which are interchangeable ideas given they are connected by a linear integral transform – are the main statistic of focus in this work and have been theoretically known as a good non-Gaussian statistic for lensing fields since more than two decades ago (Jain, Seljak & White 1998; Kruse & Schneider 2000). The CDF is also an intuitive, visually informative statistic for non-Gaussian features and is often used to check and validate reconstructed lensing fields (White & Hu 2000; Chang et al. 2018; Jeffrey et al. 2021). Previous works have also shown that the lensing PDF significantly improves constraints in  $w$ CDM compared to the standard 2-point functions (Giblin, Cai & Harnois-Déraps 2023), while more works have shown the utility of the 3D matter density PDF in probing both  $w$ CDM and other extended cosmologies (Friedrich et al. 2020; Uhlemann et al. 2020; Boyle et al. 2021; Cataneo et al. 2022; Gough & Uhlemann 2022).

While the benefits of using the CDF – namely the level of cosmological non-Gaussianity it can capture – have been explored in the past, this has mostly been in the more idealistic regime where some key observational factors were not included in the analysis. Thus, while we have had a prior understanding of the benefits of using PDFs/CDFs of the lensing field, we currently have an incomplete picture of the practical challenges in using this statistic to infer cosmological constraints.

In this work, we measure the CDFs of the lensing field from the first three years (Y3) of the Dark Energy Survey (DES) data and validate that the common lensing systematics – such as point spread function (PSF) contributions, reduced shear approximation, source clustering, and baryon imprints – have an impact on this statistic that is either negligible or can be adequately mitigated. Many of these tests have been extensively performed for 2-point statistics (Gatti et al. 2021) and have also been done for some 3-point statistics (Gatti et al. 2022; Secco et al. 2022b). The CDFs are sensitive to information at all orders, and validating the impact of these observational/modelling systematics on the CDFs also provides validation for higher-order information beyond the 3-point.

This work is organized as follows: first, we introduce the formalism for the CDFs in Section 2. In Section 3, we describe the data sets and simulations used in this work, as well as the procedures used to forward-model the simulations to match the DES Y3 data. In Section 4, we define the data vector used for the rest of this work,

<sup>1</sup>Power spectra are referred to as ‘2-point statistics’ and they capture up to second-order information as they are fundamentally a variance measure and contain two orders of the field. ‘Higher-order’ here refers to higher than second-order information, which needs to be captured by beyond 2-point statistics, or sometimes referred to as ‘higher-order statistics’.

and also demonstrate the Fisher constraining power of the CDFs for DES Y3-like data. In Section 5, we measure the CDFs on the DES Y3 weak lensing maps, and quantify the signal-to-noise of the measurements. We then validate the impact of different effects – PSF contributions, source clustering, reduced shear approximation, and baryonic imprints – on this statistic and discuss any scale cuts required to mitigate these effects. Finally, we conclude in Section 6.

## 2 CDF FORMALISM

We begin in Section 2.1 by describing the formalism of the CDF statistics used in this work, including the exact measurement procedure. In Section 2.2, we briefly review the kNN distributions, which are a recently introduced statistic for discrete tracers that summarize all higher-order information, and we discuss how the analogous, continuous-field statistic is the CDF. Finally, in Section 2.3, we validate the CDFs using Gaussian fields. Note that the CDFs are closely related to other statistics in the literature and we will describe these later on in Section 6.

### 2.1 Cumulative distribution functions

The CDFs<sup>2</sup> used in this work are defined as follows. Given a set of uniform/random points in a field, with spheres of radius  $r$  around each point, the CDFs summarize the fraction of spheres that have an enclosed density – i.e. the mean density within radius  $r$  – that exceeds a chosen threshold. In 2D, the density becomes a surface density,  $\Sigma$ , and the radius is a projected aperture,  $\theta$ . The calculation of the fraction of points whose enclosed surface densities exceed a threshold can be formally written down using the following expression,

$$\text{CDF}(\theta, k) = P(\kappa_\theta > k), \quad (1)$$

where  $\kappa_\theta \equiv \kappa(< \theta)$  is the average surface overdensity within an aperture  $\theta$ . This measurement can also be trivially modified to use the surface density, rather than overdensity, just switching  $\kappa \rightarrow \bar{\Sigma}(1 + \kappa)$ , where  $\bar{\Sigma}$  is the mean surface density field. It can also be done with the surface mass, by simply multiplying the surface density with the aperture area associated with scale  $\theta$ .

For a given map, the CDF measurement is performed as follows:

First, we fill the map with a grid of points. Without loss of generality, we take these points to be located at the centre of the HEALPix pixels (with  $\text{NSIDE} = 1024$ ), as this greatly simplifies the calculations. Increasing the number of points in the grid (i.e. the number of pixels) will improve the precision of the measurement, as is the case with the traditional 2-point correlations.

Second, we pick a certain aperture scale,  $\theta$ , and for each point we compute  $\kappa_\theta$ , the convergence smoothed on scale  $\theta$ . The smoothing is done in harmonic space using a harmonic tophat filter

$$B(\ell) = 2 \frac{J_1(\ell\theta)}{\ell\theta}, \quad (2)$$

where  $J_1(x)$  is the Bessel function of the first order. The choice of tophat over a Gaussian filter is because the former allows for an easy interpretation of an enclosed quantity within a given physical scale. Our computing procedure is the same for any other choice of filter as well.

Third, we measure what fraction of the grid points satisfy the inequality in equation (1), which is the probability,  $P(\kappa_\theta > k)$ . The

<sup>2</sup>The entire formalism could also be done using PDFs instead of CDFs. The latter is simply a more natural/convenient choice when connecting to the kNN formalism, as we describe in Section 2.2.

choice of thresholds is a degree of freedom in the measurement, and we describe our choices in Section 4.1.

Fourth and finally, steps 2 and 3 are repeated for a range of scales and thresholds to extract the distribution,  $P(\kappa_\theta > k)$ , for different choices of  $\theta$ . The exact choice of scales and thresholds used in this work is described in Section 4.1.

Fig. 1 illustrates how the CDFs are constructed in a given field, and highlights some generic features of the CDFs. In the limit where the variance  $\sigma^2 \rightarrow \infty$ , we expect  $P(\kappa_\theta > k) \rightarrow 0.5$ , and where  $\sigma^2 \rightarrow 0$ , then we expect  $P(\kappa_\theta > k) \rightarrow 0$  if  $k > 0$ , and  $P(\kappa_\theta > k) \rightarrow 1$  if  $k < 0$ . In Fig. 1 we see that all curves are closer to  $P = 0.5$  on small scales where the field’s variance is high compared to the threshold values, and move towards  $P = 0$  or  $P = 1$  on large scales where the large smoothing scale suppresses the field’s variance to values lower than the thresholds. Additionally, we see  $P(\kappa_\theta > 0) \approx 0.4$  at small scales, where the distribution is log-normal (see top panels of Fig. 1) and so the median of the distribution is not the same as the mean,  $\langle \kappa \rangle = 0$ . At large scales, we find  $P(\kappa_\theta > 0) \approx 0.5$  as the distribution becomes more Gaussian.

Thinking in 3D space, the CDFs extract  $P(> \rho | R)$ , the conditional distribution of the enclosed mean density given radius, as well as  $P(R | > \rho)$ , the conditional distribution of radii or volumes given a density threshold. These two distributions can be related using Bayes’ theorem,

$$P(> \rho | R) = P(R | > \rho) \frac{P(> \rho)}{P(R)}. \quad (3)$$

Note that given the enclosed density  $\rho$  and spherical radius  $R$ , we can easily obtain a mass  $M \equiv \frac{4}{3}\pi R^3 \rho$ . So the above can be rewritten as

$$P(> M | R) = P(R | > M) \frac{P(> M)}{P(R)}. \quad (4)$$

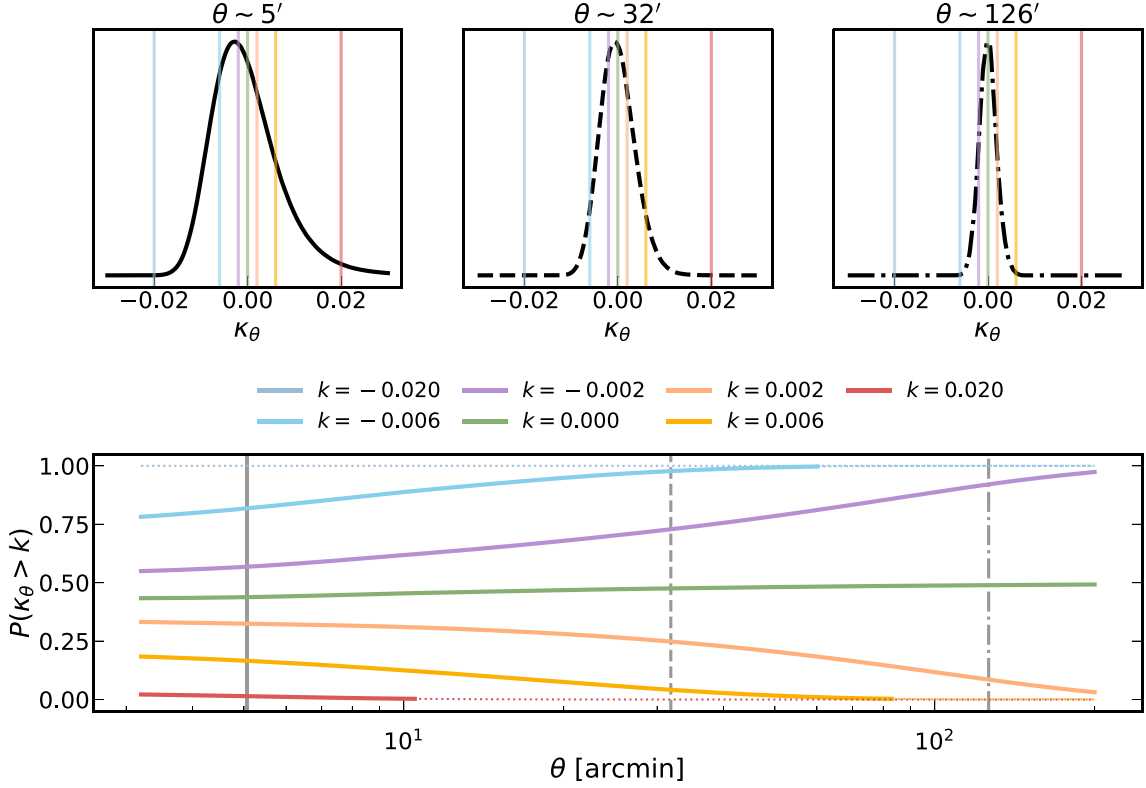
equation (3) better elucidates the connection between the CDFs and the ideas from halo collapse. The quantity  $P(> 200\rho_c | R)$  is simply the fraction of volumes that contain a halo, where the haloes are identified/defined as overdensities of at least  $\rho > 200\rho_c$ , with  $\rho_c$  being the critical density of the Universe.

We can also generalize the CDF formalism to multifield probes by computing the joint CDFs of multiple fields; this is simply,

$$P(\kappa_{\theta,1} > k_1, \kappa_{\theta,2} > k_2 | \theta), \quad (5)$$

where  $\kappa_{\theta,1}$  and  $\kappa_{\theta,2}$  are two different fields (e.g. different tomographic bins of a single type of field, or different types of fields). While we are allowed to choose different values for the thresholds  $k_1$  and  $k_2$ , we will enforce  $k = k_1 = k_2$  henceforth for simplicity in the data vector. In this work, we will consider the cross-correlation between tomographic bins as part of our measurement. Note that the 2-field version of the CDFs formally contains all the 1-field information as well. This connection is identical to how 2D PDFs contain the marginal 1D distributions within them.<sup>3</sup> We will use both 1-field and 2-field CDFs as part of our main data vector. The 3-field and 4-field CDFs will formally have additional information beyond the 1-field and 2-field CDFs, though our tests have shown there is only marginal improvement in cosmological constraints for the analysis choices described here (e.g. tomographic bin, angular scales, and thresholds).

<sup>3</sup>A simple example is the 2D CDF,  $P(\kappa_{\theta,1} > k_1, \kappa_{\theta,2} > k_2 | \theta)$  taken in the limit  $k_2 = -\infty$ . In this case,  $\kappa_{\theta,2}$  is always above the threshold  $k_2$  and so the 2D CDF reduces to a 1D CDF,  $P(\kappa_{\theta,1} > k_1, \kappa_{\theta,2} > k_2 | \theta) \rightarrow P(\kappa_{\theta,1} > k_1 | \theta)$ .



**Figure 1.** Bottom: The probability that  $\kappa_\theta$ , the average convergence within circles of apertures  $\theta$ , exceeds a chosen threshold  $k$ . We use seven thresholds and show measurements for a noiseless convergence field corresponding to the fourth tomographic redshift bin in DES Y3. The solid lines are converted to dotted ones when the CDFs fall into the 99.7 percent ( $3\sigma$ ) tail. The grey–blue line is always in the tail for this particular measurement. Top: The PDFs of  $\kappa_\theta$  for different choices of aperture,  $\theta$ . The three aperture scales that we show PDFs for are indicated by the vertical grey lines in the bottom panel. The PDFs are estimated from noiseless convergence fields and are smoothed with a Gaussian for visualization purposes. The vertical lines in these top three panels are the thresholds we use. The probability to exceed is the integral from each threshold up to  $P(\kappa = \infty)$ . For high thresholds, we have a lower probability to exceed and vice versa for low thresholds.

For some tests, we will also post-process the 2-field CDFs to isolate just the cross-covariance/correlation. This is done by performing the redefinitions described in Banerjee & Abel (2021b),

$$\psi_{1,2}(k) = \text{CDF}_{1,2}(k) - \text{CDF}_1(k)\text{CDF}_2(k), \quad (6)$$

which takes the joint probability to exceed in two different fields and removes the product of the individual probability to exceed for each field. The quantity  $\psi_{1,2}(k)$  is 0 if the fields are completely uncorrelated, and non-zero otherwise. The sign of  $\psi_{1,2}(k)$ , for any threshold  $k$ , indicates the sign of the correlation between the two fields at that threshold.

We can also extend this formalism to more than 2 fields (e.g. a triplet  $ABC$ , where each letter is a field index). While we do not consider such measurements in our analysis here, we note their potential utility both for cosmological information, but also as further compressions of the data vector. Note that there is no benefit to repeating a field twice (e.g. the triplet  $AAB$ , where  $A$  is repeated twice) if we also fix the threshold  $k$  for all the fields. The joint probability  $P(\kappa_1 > k, \kappa_2 > k, \kappa_3 > k)$  is exactly similar to  $P(\kappa_1 > k, \kappa_2 > k)$ .

While we have discussed the CDFs in terms of lensing convergence, it is not necessary to be limited to this quantity. For example, one could consider the kinetic or thermal Sunyaev–Zeldovich fields (Sunyaev & Zeldovich 1972; Carlstrom, Holder & Reese 2002), which are generated by baryons in haloes and thus inherit the non-Gaussian features of the structure traced by these haloes.

## 2.2 Connection to kNN distributions for discrete fields

The kNN distributions (Banerjee & Abel 2021a, b) are a novel way to summarize the clustering in a field of discrete tracers, such as galaxies or haloes. They have been formally shown to capture volume integrals of all  $N$ -point functions of the tracer field, but can be computed in  $\mathcal{O}(N \log N)$  time, where  $N$  is the number of tracers. Thus, they have the same computational efficiency as a 2-point correlation function, but capture integrals of all the information held in the  $N$ -point functions (2-point, 3-point, 4-point, etc.). This statistic has already been measured in observational data, particularly to quantify the signal-to-noise of all correlations (both Gaussian and non-Gaussian) in a clustered field (Wang, Banerjee & Abel 2022).

The kNNs are computed by taking a field of tracers with a known number density  $n_{\text{tr}}$ , and then generating a large set of random points in this field as one would for computing an  $N$ -point clustering function (although a set of uniform points would be a sufficient choice as well). For each point, one computes the distance to the nearest tracer neighbour. The distribution of distances to the  $k$ th nearest neighbour forms a kNN distribution. This statistic is probing the distribution  $P(V | > k_{\text{tr}})$ , i.e. the distribution of volumes that contain at least  $k_{\text{tr}}$  tracers, where  $k_{\text{tr}}$  takes integer values. Assuming spherical volumes, this can be reformulated as the distribution  $P(R | > k_{\text{tr}})$ . Given kNNs depend on the counts of tracers enclosed within a volume, it is sensitive to volume integrals of all the correlation functions. However, the fact that the sensitivity is to a volume integral of the functions means signals

from specific configurations of the  $N$ -point functions will be mixed together.<sup>4</sup>

In the limit of  $n_{\text{tr}} \rightarrow \infty$ , the number counts threshold  $>k_{\text{tr}}$  becomes a density threshold  $>\rho$ , and the conditional distribution becomes  $P(R | >\rho)$  which can be related, using Bayes' theorem, to the distribution probed by the CDFs,  $P(>\rho | R)$ . A detailed discussion on this connection between kNNs and CDFs can be found in Banerjee & Abel (2023, see their section 2.1). The analytic connection between the two statistics directly confirms that the CDFs can be *formally* shown to contain all volume integrals of higher-order functions, and this makes them better suited for summarizing a field, where we do not a priori know the exact cosmological information contained in all the non-Gaussian signatures of the field. In addition, this connection means the CDFs are the natural statistic to cross-correlate discrete and continuous fields while using the kNN formalism for the former (Banerjee & Abel 2023).

### 2.3 Consistency relations for Gaussian fields

In the Gaussian limit of  $P(\kappa_\theta) = \mathcal{N}(\kappa_\theta; \mu, \sigma)$  – where  $\mathcal{N}$  is a normal distribution with mean  $\mu$  and variance  $\sigma^2$  – there are three degrees of freedom for the CDF: the mean and variance of the map at each aperture scale, and the threshold  $k$ . The threshold is an input parameter, and the mean of the map is taken to be  $\mu = 0$  given  $\kappa$  is derived from the overdensity field and so is defined as a perturbation field with the mean background subtracted. Thus, the variance is the only unconstrained parameter, and this variance can also be measured directly on the map. Formally, a Gaussian CDF is parametrized as,

$$\text{CDF}(k) = 1 - \int_k^\infty \mathcal{N}(x - \mu, \sigma) dx = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{k - \mu}{\sigma\sqrt{2}} \right) \right]. \quad (7)$$

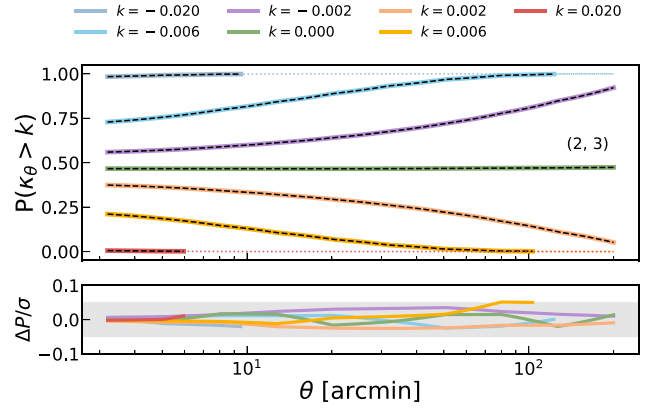
We can thus use the variance measured from the map smoothed on a given scale,  $\theta$ , to predict the CDFs at that scale. For a purely Gaussian field, the measurements and predictions must agree. The same exercise is trivially extended for the 2-field CDFs. In the Gaussian limit, the joint PDF of any set of fields is given by a multivariate normal distribution,

$$\text{PDF} = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp \left[ -\frac{1}{2} (\vec{\kappa} - \mu)^T \Sigma^{-1} (\vec{\kappa} - \mu) \right], \quad (8)$$

where the column vector  $\vec{\kappa} = \{\kappa_1, \kappa_2, \dots, \kappa_n\}$  are the kappa value in each field, and denote the point in multidimensional space where we evaluate the probability. The PDF in equation (8) can be integrated, assuming some set of thresholds for each field, to obtain the CDF. Recall that in this work we set all thresholds to the same value  $k$ . We also use  $\mu = 0$ . The unknown degrees of freedom for the distribution are then entirely in the covariance matrix. Thus if we know this covariance matrix, we can always predict the CDFs exactly.

We verify this in Fig. 2 for our analysis setup. The top panel shows the 2-field CDF measured on noiseless, simulated maps whose signal mimics the DES Y3 data used in this work (see Section 3.2 for more details). In particular, the convergence map has the same redshift distribution as the third and fourth tomographic bins. These are all Gaussian maps made by post-processing  $N$ -body products, as detailed below in Section 3.1.4. The dashed lines (prediction) are

<sup>4</sup>For the 2-point function, there is no configuration information as the correlations depend on just distance,  $r$ . For  $N$ -point correlations of  $N > 3$ , the geometry connecting the  $N$  points will contain additional information, though the exact information contained in this geometry remains an open question.



**Figure 2.** Top: The 2-field CDFs averaged over 1000 *noiseless, full-sky, Gaussian* convergence maps. The  $n(z)$  for the two fields corresponds to the third and fourth DES Y3 redshift bins. The solid lines switch to dotted when the CDF is outside the range  $[0.003, 0.997]$  (approximately corresponding to the  $3\sigma$  bounds). The black dashed lines show the predictions for the CDFs given the covariance of the two fields at a given smoothing scale,  $\theta$ ; under the assumption the fields are Gaussian, the predictions must match the measurement. Bottom: The difference between the CDF measurement and Gaussian-field predictions,  $\Delta P = P_{\text{meas}} - P_{\text{theory}}$  normalized by the uncertainty in the CDFs – where the uncertainty is cosmic variance and is the observational limit for measurement uncertainty – estimated from the 1000 realizations. The grey band shows  $\Delta/\sigma < 0.1$ . In all cases, the difference,  $\Delta P$ , is within this region and is completely negligible.

consistent with the solid ones (measurement). The bottom panel shows the Gaussian model predictions are within  $0.05\sigma$  of the measurements, where the  $\sigma$  of the data vector is just cosmic variance and thus represents the observational limit in precision.

## 3 DATA

We first describe in Section 3.1 the different simulations used in our analysis. We then detail the DES Y3 data in Section 3.2 and in Section 3.3 we describe how the simulated maps are forward modelled to imitate the DES Y3 data.

All maps used in this work are made with the HEALPIX convention of  $\text{NSIDE} = 1024$ . This corresponds to a pixel scale of 3.2 arcmin. The one exception are the products used from the COSMOGRID suite, described in Section 3.1.3, which are  $\text{NSIDE} = 512$ .

### 3.1 Simulations

While the CDF is a statistic that can be used to summarize any scalar field, in this work we are specifically interested in the lensing convergence,  $\kappa$ , which is a line-of-sight integral of the density field,

$$\kappa(\hat{\mathbf{n}}, z_s) = \frac{3}{2} \frac{H_0^2 \Omega_m}{c^2} \int_0^{z_s} \delta(\hat{\mathbf{n}}, z_j) \frac{\chi_j(\chi_s - \chi_j)}{a(z_j)\chi_s} dz_j \frac{d\chi}{dz} \Big|_{z_j}, \quad (9)$$

where  $z_s$  is the redshift of the ‘source’ plane/galaxies being lensed,  $\hat{\mathbf{n}}$  is the pointing direction on the sky,  $\delta$  is the overdensity field,  $\chi$  is the comoving distance from an observer to a given redshift,  $a$  is the scale factor,  $H_0$  is the Hubble constant,  $\Omega_m$  is the matter energy density fraction at  $z = 0$ , and  $c$  is the speed of light. We use the shorthand  $\chi(z_s) \equiv \chi_s$  and  $\chi(z_j) \equiv \chi_j$ .

We model this convergence using full-sky density maps from different  $N$ -body simulations, with each simulation serving a different purpose in this work. We detail these different simulations below.

Such simulations are uniquely suited for modelling these fields in the nonlinear regime. For quasi-linear and linear regimes, analytic models can also be utilized (e.g. Barthelemy et al. 2023).

### 3.1.1 Anbajagane23 simulations (A23)

In this work, we use a suite of  $N$ -body simulations run with the PKDGRAV3 solver (Potter, Stadel & Teyssier 2017), where the suite has been specialized for performing Fisher forecasts for widefield surveys. This simulation suite, formally denoted the ULAGAM suite but referred to in this work by the abbreviation ‘A23’ for simplicity, is described in Anbajagane et al. (2023b). We describe here just the essential features of the runs and the relevant data products used in this work. The A23 simulations are run in  $1h^{-1}$  Gpc boxes, starting at  $z = 127$ , with  $N = 512^3$  dark matter particles. The initial conditions for all simulations are obtained from the QUIJOTE suite (Villaescusa-Navarro et al. 2020), and so the simulations are essentially light-cone runs of the QUIJOTE simulations specialized for widefield survey analyses. The original QUIJOTE suite was designed for studying the Fisher information of the nonlinear structure, as well as building emulators sampling different cosmological parameters, but the data products are inadequate for producing mock light-cones of the lensing/density field. These products include snapshots and halo catalogues at only five redshifts, which is too coarse a redshift resolution for building light-cones. Hence we have rerun a subset of these simulations to create accurate full-sky lensing and density maps.

The suite contains simulations for computing the derivatives of the lensing/density field with respect to multiple cosmological parameters, of which three are of interest to us –  $\Omega_m$ ,  $\sigma_8$ , and  $w$ . For each parameter, the suite contains 100 full-sky simulations where the parameter takes values slightly higher than the fiducial, and another 100 full-sky simulations where the value is lower than the fiducial. These two sets are used to compute the derivatives of a summary statistic with respect to  $\Omega_m$ ,  $\sigma_8$ , and  $w$ . The fiducial cosmology is from Planck Collaboration (2016a), and the derivatives are computed over differences of  $\Delta\Omega_m = 0.02$ ,  $\Delta\sigma_8 = 0.03$  and  $\Delta w = 0.05$ , which are all the same settings as the QUIJOTE suite. The suite also has 2000 simulations at the fiducial cosmology which are used to compute the covariance matrix for our data vector. Since each all-sky map can have 4 completely independent DES footprints, we have a total of 8000 estimates of each summary statistic to use for the covariance, and 400 independent estimates of the derivative of the summary statistic with respect to each parameter.

While the original QUIJOTE suite was run using GADGET3 (last described in Springel 2005), we use PKDGRAV3 which has already been employed extensively to perform both theoretical studies of the lensing field as well as simulation-based analyses of data from different weak lensing surveys (Fluri et al. 2019; Gatti et al. 2022; Zürcher et al. 2022). The PKDGRAV3 solver automatically builds light-cones as it solves the gravitational dynamics of the system forward in time, and so our final outputs are the light-cone shells – i.e. HEALPIX maps – of the density field at different redshifts. The simulation box is tiled/repeated as needed to construct large enough volumes to then build full-sky light-cones to a given redshift. This repetition will bias any large-scale correlations in the light-cone, but in this work we only consider scales much smaller than the box size.

The simulations have a total of 100 time-steps/shells, with 95 shells between  $0 < z < 10$ . This gives us a high redshift resolution of between  $\Delta z \approx 0.01 - 0.05$  in that redshift range, with the exact value depending on the shell. The time-steps in this redshift range

are spaced uniformly in proper time,  $t$ , and this corresponds to different  $z$  and comoving distances depending on the cosmology. These density shells are then post-processed via equation (9), with the integral over  $z_j$  replaced by a simple discrete sum, to create a lensing convergence field at each source plane redshift,  $z_s$ . This technique uses the Born approximation, which computes the total effective deflection due to lensing but along an undeflected ray path. A more precise calculation uses full ray-tracing, which calculates these deflections while constantly deflecting/Updating the ray path. Petri, Haiman & May (2017) found the Born approximation leads to differences of  $\lesssim 5$  per cent for the third moments statistic we will use in Section 4.2, but this is subdominant to the current uncertainties of  $\approx 15$  per cent.

Note we have not performed any resolution-convergence tests. The numerical requirements for this work are less stringent as we do not use the simulations for cosmological inference, but rather for (i) performing a Fisher analysis (Section 4.2), where the relevant quantities are relative and not absolute differences in the simulations as we vary cosmological parameters, and for (ii) computing covariance matrices for our systematic checks (Section 5).

### 3.1.2 Takahashi17 simulations (T17)

The Takahashi17 simulations (Takahashi et al. 2017) are a suite of  $N$ -body simulations run at a WMAP9 cosmology (Hinshaw et al. 2013), and have a higher particle resolution than the A23 suite, with  $2048^3$  particles. They, however, have lower redshift resolution than the A23 suite with 38 shells between  $0 < z < 5$ . The shells are spaced equally in comoving distance, with widths of  $150 \text{ Mpc } h^{-1}$ , and this leads to redshift spacing of roughly  $\delta z \sim 0.05 - 0.2$ . The T17 simulations have been used to model/test higher-order statistics in many works (Gatti et al. 2020; Secco et al. 2022b; Gong et al. 2023; Heydenreich et al. 2023; Munshi et al. 2023) for modelling, validation etc. and so we measure our statistics on these simulations for completeness. There are 108 independent full-sky maps, and that gives us a total of 432 DES Y3 cutouts.

### 3.1.3 Cosmogrid

COSMOGRID is a large suite of simulations that span the  $w$ CDM parameter space, including the sum of the neutrino masses, and are designed for simulation-based modelling of widefield survey data (Kacprzak et al. 2023). They were run using PKDGRAV3, similar to the A23 simulations, and have a  $900 \text{ Mpc}/h$  box size with  $832^3$  particles. The simulations are run at 2500 points spanning the parameter space, with 7 realizations at each point. They have 140 time-steps, with 70 equally spaced steps in proper time between  $4 < z < 99$ , and another 70 equally spaced steps in proper time between  $0 < z < 4$ . The spacing is different in each of the two regimes.

In this work, we use COSMOGRID to test the impact of baryons on the lensing CDF statistic. For this purpose, we use the fiducial runs which are 200 simulations run at fixed cosmology (Kacprzak et al. 2023, see their table 2). We use both the default  $N$ -body run as well as the run post-processed using the method of Schneider et al. (2019) so the density field looks like that of a hydrodynamic simulation with baryons. We discuss this more in Section 5.6. While the raw maps are available at  $\text{NSIDE} = 2048$ , the maps post-processed to look like those of hydrodynamic simulations are provided only at  $\text{NSIDE} = 512$  – which is lower than the fiducial resolution of  $\text{NSIDE} = 1024$  used in this work – and we discuss the impact of this in Section 5.6 as well.

### 3.1.4 Gaussian maps

For the purpose of validating non-Gaussian statistics, it is useful to have maps that are purely Gaussian – i.e. are represented entirely by a power spectrum – rather than ones that contain a realistic level of nonlinearity/non-Gaussianity. We use the power spectrum measured on the  $N$ -body maps, which contain the relevant nonlinearities, to then create consistent Gaussian maps. These maps will by construction have the same nonlinear power spectra as the original maps. The method employed for doing this is the same as Giannantonio et al. (2008, see their appendix A). It involves computing all auto- and cross-spectra between the relevant fields on the simulated maps, and then using these spectra with random phases to generate spherical harmonic modes  $a_{\ell m}$  that are appropriately correlated to reproduce the input auto- and cross-power spectra. The  $a_{\ell m}$  can then be transformed to obtain a real-space map. By definition, such maps will have no higher-order information and be described entirely by their power spectra.

If we have two maps  $X$  and  $Y$ , and want to generate Gaussian maps that have the same auto and cross-power spectrum as  $X$  and  $Y$ , we obtain the  $a_{\ell m}$  via

$$\begin{aligned} a_{\ell m}^X &= \eta_{\ell m}^X T^{XX} = \eta_{\ell m}^X \sqrt{C_\ell^{XX}}, \\ a_{\ell m}^Y &= \eta_{\ell m}^X T^{XY} + \eta_{\ell m}^Y T^{YY} \\ &= \eta_{\ell m}^X \frac{C_\ell^{XY}}{\sqrt{C_\ell^{XX}}} + \eta_{\ell m}^Y \sqrt{C_\ell^{YY} - \frac{(C_\ell^{XY})^2}{C_\ell^{XX}}}, \end{aligned} \quad (10)$$

where  $\eta_{\ell m}$  is a complex random normal variable with zero mean and unit variance, and  $T_{ij}$  are coefficients derived from the power spectra, with a general form given by,

$$T^{ij} = \begin{cases} \sqrt{C^{ji} - \sum_{k=1}^{j-1} (T^{ik})^2}, & \text{if } i = j; \\ \frac{1}{T^{jj}} \left( C^{ji} - \sum_{k=1}^{j-1} T^{ik} T^{jk} \right), & \text{if } i > j. \end{cases} \quad (11)$$

and equations (10) and (11) above have been reproduced from Omori (2022, see Appendix C).

For producing real maps, the  $m = 0$  coefficients must be handled separately as they should have no imaginary component (see appendix B in Sellentin et al. 2023, for an example). Thus, we explicitly remove their imaginary component, by setting  $\text{Im}(a_{\ell m=0}) = 0$ , and then rescale the coefficients as  $a_{\ell m=0} \rightarrow \sqrt{2}a_{\ell m=0}$ .<sup>5</sup> From these final  $a_{\ell m}$  values we generate the Gaussian maps using the HEALPY routine, `alm2map`.

Note that when we post-process the Gaussian maps to mimic the DES year 3 observations (see Section 3.3), only the true convergence field is Gaussian. The procedures applied to the field to post-process it – such as non-Gaussian noise, and survey masks of complicated geometries – will still induce a non-zero non-Gaussianity in the final simulated convergence field, but these non-Gaussianities will not be cosmological in origin.

<sup>5</sup>Formally, our complex variable satisfies  $\langle \eta \rangle = 0$  and  $\langle \eta \eta^* \rangle = 1$ . Thus, the real and imaginary components of  $\eta$  have variance 0.5 each. For the  $a_{\ell m=0}$  coefficients, we remove their imaginary component, and so their real component must be rescaled for the coefficients to have the intended unit variance.

### 3.2 Dark Energy Survey Year 3 (DES Y3)

The Dark Energy Survey (The Dark Energy Survey Collaboration 2005) is an optical imaging survey of 5000 deg<sup>2</sup> of the southern sky, and is currently the largest precision photometric data set for cosmology. We use the data from the Year 3 data release (Sevilla-Noarbe et al. 2021), and in particular the galaxy shape catalogues. This is the same data set used for the fiducial 2-point correlation function shear results (Amon et al. 2022; Secco et al. 2022a) and harmonic power spectrum results (Doux et al. 2022), as well as the higher-order statistics such as the moments (Gatti et al. 2022), mass aperture (Secco et al. 2022b), and peaks (Zürcher et al. 2022). In this work, the Y3 METACALIBRATION galaxy shape catalogue (Gatti et al. 2021) is used to make a map of the ellipticities, which is then converted into a convergence map via the Kaiser Squires method (Kaiser & Squires 1993). This is the same technique used in previous works on the mass map (Chang et al. 2018; Jeffrey et al. 2021). We perform all our measurements and tests on these maps.

We also use the DES Y3 PSF and reserved star shape catalogues from Jarvis et al. (2021) to estimate the impact of PSF contributions to the signal observed by our statistic. The shape catalogues are used to make a PSF ‘mass map’ the same way the galaxy ellipticities are used, and this mass map is used to test the PSF contributions (see Section 5.3 for more detail). The same star shape catalogue was used to test PSF contributions for both the shear 2-point function (Gatti et al. 2021) and the 3-point function (Gatti et al. 2022; Secco et al. 2022b).

### 3.3 Making simulated DES Y3-like mass maps

We modify the simulated convergence/mass maps described in the above sections to include all the relevant observational effects of the DES Y3 data. Note that the main purpose of the maps is both to perform realistic forecasts of the cosmological constraints (Section 4), and to validate the contribution of different systematics to the CDFs data vector (Section 5). In this work, we do not use these simulations to get cosmology constraints from the DES Y3 data vector.

To make the mock maps, we start from the true convergence field,  $\kappa$ , and use an inverse Kaiser–Squires (KS) transform (Kaiser & Squires 1993) to obtain the two shear components,  $\gamma_{1,2}$ . The shear is the true observable of a weak lensing survey given we measure galaxy shapes. The KS transform can be quickly performed in harmonic space as

$$\gamma_E^{\ell m} + i\gamma_B^{\ell m} = -\sqrt{\frac{(\ell+2)(\ell-1)}{\ell(\ell+1)}} \left( \kappa_E^{\ell m} + i\kappa_B^{\ell m} \right), \quad (12)$$

where the subscripts denote the E-mode and B-mode shear/convergence maps respectively. In the full-sky limit, where we have no survey masks, this is an exact expression. The technique has been validated for realistic data and found to have adequate accuracy (Chang et al. 2018; Jeffrey et al. 2021).

**Redshift distribution/bins:** We use four tomographic redshift bins with source galaxy  $n(z)$  distributions matching DES Y3 (Myles et al. 2021); the mean redshifts of these bins are  $z_{\text{mean}} \in \{0.336, 0.521, 0.741, 0.935\}$ . The true shear maps corresponding to each bin are obtained via a weighted sum of the shear maps in each redshift, where the weights are the  $n(z)$  distributions.

**Noise realization:** The noise is obtained using the DES Y3 METACALIBRATION shape catalogue from Gatti et al. (2021), using the same technique as Gatti et al. (2022). The galaxy shapes are randomly rotated to remove all spatial correlations of the galaxy ellipticities,

thus removing any cosmological signal. We then place galaxies in pixels of a  $N_{\text{SIDE}} = 1024$  map, and compute the weighted average of the shear components in each pixel of the map,  $\gamma_{1,2}^{\text{noise}}(\hat{\mathbf{n}})$ , using the weights provided in the catalog. We add this noise to the true shear maps,  $\gamma_{1,2}$ , separately for each tomographic redshift bin. This ensures the Y3 data and the simulated noise maps have the exact same variations in source/survey depth, and as we will show later, these variations create a strong non-Gaussian feature in the map (Section 5.2).

**Multiplicative bias:** The measured galaxy shapes have a bias of the order 1 per cent that has been calibrated using large suites of image simulations of the DES Y3 survey (MacCrann et al. 2022). We include these bias terms,  $m$ , in the maps by simply multiplying the shears as  $\gamma_{1,2} \rightarrow (1 + m)\gamma_{1,2}$ .

**Mask:** We only use map pixels that have at least one DES Y3 galaxy in each of the four redshift bins. All pixels that do not fall into this category are discarded, and this defines the survey mask which is used in all further analyses, both for the simulations and for the DES Y3 data.

**Kaiser-Squires reconstruction:** Following the steps above, we obtain a spin-2 shear field,  $\gamma_{1,2}$ , per DES Y3 tomographic redshift bin, that has noise, multiplicative bias, and a mask applied to it. We then convert this back to a convergence field using equation (12) to obtain a noisy convergence map for each redshift bin. We only use the E-mode convergence map in our analysis. This map is then used as our final DES Y3-like map. Other, more sophisticated map-making techniques have been explored in the Y3 data as a replacement to KS reconstruction. A detailed description can be found in Jeffrey et al. (2021). The KS method remains the simplest method that is also quick and accurate. The simplicity in compute time is a particularly attractive feature here as we make  $\mathcal{O}(10^4)$  mock DES Y3 maps in this work. Note that the mass maps we generate from DES Y3 data in Section 3.2 are also created by making the shear maps  $\gamma_{1,2}$  and using the KS transformation to obtain the convergence field.

In Section 5, we will add other effects to the mock maps – such as PSFs, higher-order shear effects, and so on – to test their impact on the measured signal and quantify which effects can be safely ignored and which effects may require scale cuts on the data vector. We do not address the impact of intrinsic alignments in this work, as it is often treated as a systematic that can be modelled, and thus marginalized over, in a full cosmological analysis as opposed to an effect that contaminates the data vector and requires scale cuts. For example, Zürcher et al. (2022) present a framework to do such marginalization assuming a simple intrinsic alignments model that can be forward-modelled in the simulations, while Hoffmann et al. (2022) presents a more advanced and physically motivated way to incorporate the same into high-resolution simulations.

## 4 CDF ANALYSIS SETUP AND FISHER CONSTRAINTS

We define the CDF data vector for DES Y3 in Section 4.1 and show the Fisher information in this CDFs data vector, as well as data vectors of other closely related statistics, in Section 4.2.

### 4.1 Defining CDFs data vector

In this work, we measure all possible 1-field and 2-field CDFs for the four tomographic bins of DES Y3. This results in four 1-field ‘auto’ CDFs, and six 2-field ‘cross’ CDFs. We measure the CDFs across 10 smoothing scales, spaced logarithmically between  $3.2'$  and  $200'$ . The choice of scales matches the moments-based DES Y3 analysis

of Gatti et al. (2022). For each scale, we use 7 thresholds  $k \in \{-20, -6, -2, 0, 20, 6, 20\} \times 10^{-3}$ . These were chosen by looking at the variance of the field at the smallest and largest smoothing scale, and ensuring at least two thresholds did not asymptote to 0 or 1 at each scale. Using the Fisher forecast below we have checked that these thresholds probe most of the relevant information while being practical to implement, and we do not perform a more methodic study of the optimal threshold choices. We have, however, verified that removing any one of the seven thresholds leads to a fractional change in the constraints of 5 per cent to 10 per cent. We did not test adding more thresholds as the longer data vector leads to poorer numerical convergence, which then makes it difficult to robustly identify the increase in constraining power provided by the additional thresholds. One could also include the 3-field and 4-field CDFs in the data vector. We have verified that for the cosmology parameters considered here and for the choice of thresholds listed above, including these 3-field and 4-field CDFs do not improve the constraints relative to the 1-field plus 2-field case.

For all CDF measurements, we only focus on the range of scales where  $0.05 < \text{CDF}(k, \theta) < 0.95$ , which excludes the  $\sim 2\sigma$  region of the distribution for each threshold  $k$  and smoothing scale  $\theta$ . This removes measurements of the tails of the distribution where noise can cause spurious signals, and it also helps remove regions where the CDF has asymptoted to constant values of 0 or 1. We have confirmed that using different choices, such as  $3\sigma$  or  $4\sigma$  cuts, leads to a fractional difference of  $< 5$  per cent in the Fisher constraints. While the tails of the distribution are a sensitive probe of the non-Gaussian information, they are also much noisier and so the actual constraining power from this region of the distribution is not significant. The ‘bulk’ of the distribution – for example, the  $1\sigma$  to  $2\sigma$  region – is still quite sensitive to non-Gaussian features while being less susceptible to noise (e.g. Friedrich et al. 2020; Uhlemann et al. 2020).

Our initial data vector has size  $N = 10z\text{-bins} \times 10\text{scales} \times 7\text{thresholds} = 700$  data points. The procedure above of focusing only on  $0.05 < \text{CDF} < 0.95$  removes more data points as multiple thresholds reach asymptotic behaviour of  $\text{CDF} = 0$  and  $\text{CDF} = 1$  at large smoothing scales, especially for the lower redshift bins where the variance of the convergence field is lower.<sup>6</sup> In practice, the data vector for DES Y3-like maps has  $N = 460$  points. Note that different thresholds reach these asymptotic values at different scales. Fig. 1 illustrates this behaviour.

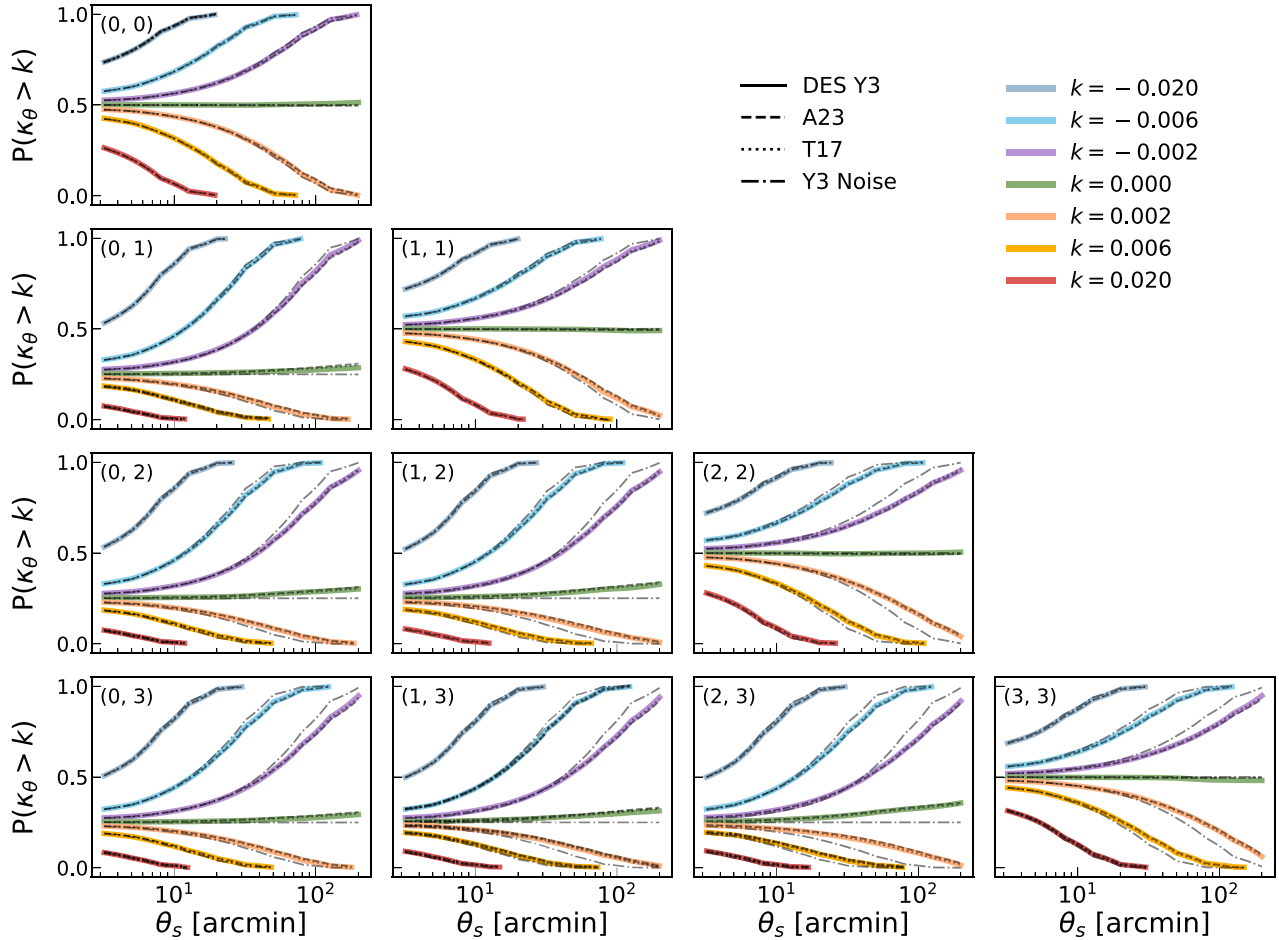
Fig. 3 presents the data vector measured on the DES Y3 data as well as different simulations described in Section 3.1. The 1-field (2-field) CDFs are shown in the diagonal (off-diagonal) panels. The coloured lines show  $P(\kappa_\theta > k)$ , the fraction of the map that exceeds a given threshold at a given smoothing scale, where each colour is a different threshold. At a fixed threshold, the probability is driven to 0 or 1 with larger  $\theta$ , and this behaviour is discussed in Section 2.1.

The threshold  $k = 0$  is special as it is the mean of the 1D marginal distributions, and so its probability for the 1-field CDFs is  $P \approx 0.5$  across all scales.<sup>7</sup> In the 2-field case the probability for  $k = 0$  is  $P(\kappa_{\theta,1} > 0, \kappa_{\theta,2} > 0) \approx 0.25$  but has scale-dependent

<sup>6</sup>The density field has a higher variance at lower redshifts, but the lensing kernel has a lower amplitude for low-redshift sources and so the variance of the convergence field increases with redshift.

<sup>7</sup>Fig. 1 shows the true convergence field is log-normal on small scale, and thus has  $P(\kappa_\theta > 0) \neq 0.5$ . However, for *noisy* convergence fields, the noise dominates the cosmological signal on small scales and this noise is a symmetric distribution (the odd moments are zero, as discussed in Section 5.2). This restores the measurements to  $P(\kappa_\theta > 0) \approx 0.5$  as mentioned.





**Figure 3.** The fiducial data vector used in this work. Coloured solid lines are measurements of the CDFs on DES Y3 mass maps, dark dashed lines are from the A23 suite, dotted lines are from T17, and the dashed-dotted lines are from just shape noise maps with no cosmological signal. All simulated maps have the same DES Y3 shape noise field, survey mask,  $n(z)$  distribution, and are put through the same convergence reconstruction method. The panels show 1-field or 2-field CDFs for different bin combinations, with the specific combination denoted in the corner of each panel. There are clear differences between the noise-only CDFs and the DES Y3 data CDFs, particularly on larger scales and in higher redshift bins, which are the expected imprints for a cosmological signal in the lensing convergence maps. The A23 and T17 simulation predictions are a decent match to the Y3 data.

deviations. This is because the correlation between the two fields alters this probability, and this correlation has a scale dependence, meaning the deviations from  $P \approx 0.25$  will also be scale-dependent as expected.

We can also see a clear visual difference between the CDFs of the shape noise field (dashed-dotted) and those of the observed convergence field. In particular, the 1-field CDFs of the (3, 3) bin show the clearest difference at larger scales. The shape noise field has a notably smaller variance than the observed convergence field, and this causes the CDFs to asymptote to 0 or 1 more quickly compared to the CDFs of the data. We also find that the T17 predictions are quite similar to those of A23, and that the simulations are generally a decent match to the data.

## 4.2 Fisher information

We use the data vectors and covariance matrices constructed from the A23 simulations to perform a Fisher forecast for three  $w$ CDM parameters that are the target of current and future lensing surveys –  $\Omega_m$ ,  $\sigma_8$ , and  $w_0$ . We measure three broad types of summary statistics for this forecast:

**Gaussian Statistics**, such as angular power spectra and the second moments of the field are well known for being sensitive to only the variance of the field, and the variance is often denoted the Gaussian part of the distribution. These statistics provide a good baseline for cosmological constraints obtained from current fiducial analyses, which primarily use such Gaussian statistics. The angular power spectra are measured in 20 bins in the range  $10 < \ell < 2048$ . The second moments are measured on the maps smoothed with a tophat across 10 scales that are logarithmically spaced in the range  $3.2' < \theta < 200'$ .

**Higher-order moments** are a natural extension to the second moments where one averages higher powers of the fields,  $\langle \kappa^N \rangle$ . The most common one is the third moment (or skewness), though the fourth moment (or kurtosis) has also been measured in lensing data before across a smaller range of angular scales  $2' < \theta < 8'$  (Van Waerbeke et al. 2013). In this work we measure the second and third moments in the range  $3.2' < \theta < 200'$ .

Finally, the **CDF** is the non-Gaussian statistic that is the focus of this work. The data vector definition is described in Section 4.1, and the measurement on DES Y3 data and some simulated mock maps is shown in Fig. 3.

Note that the data vectors of these higher-order statistics tend to be long, and this is particularly an issue when computing the covariance numerically, as the number of realizations needed for the covariance increases with the data vector size. However, the A23 simulation suite contains 8000 DES Y3-like maps, and this number is far larger than the length of any data vector computed in this work.

We can now estimate the Fisher information with the standard approach,

$$\mathbf{F}_{ij} = \sum_{m,n} \frac{d\tilde{X}_m}{d\theta_i} (\mathcal{C}^{-1})_{mn} \frac{d\tilde{X}_n}{d\theta_j}, \quad (13)$$

where  $\frac{d\tilde{X}_m}{d\theta_i}$  is the mean derivative of point  $m$  in data vector  $X$  with respect to parameter  $\theta_i$ , where the mean is computed using 400 DES Y3 realizations (see Appendix C and Fig. C1).  $\mathcal{C}^{-1}$  is the inverse of the numerically estimated covariance matrix and this is computed while accounting for the Hartlap factor (Hartlap, Simon & Schneider 2007),

$$\mathcal{C}^{-1} \rightarrow \frac{N_{\text{sims}} - N_{\text{data}} - 2}{N_{\text{sims}} - 1} \mathcal{C}^{-1}. \quad (14)$$

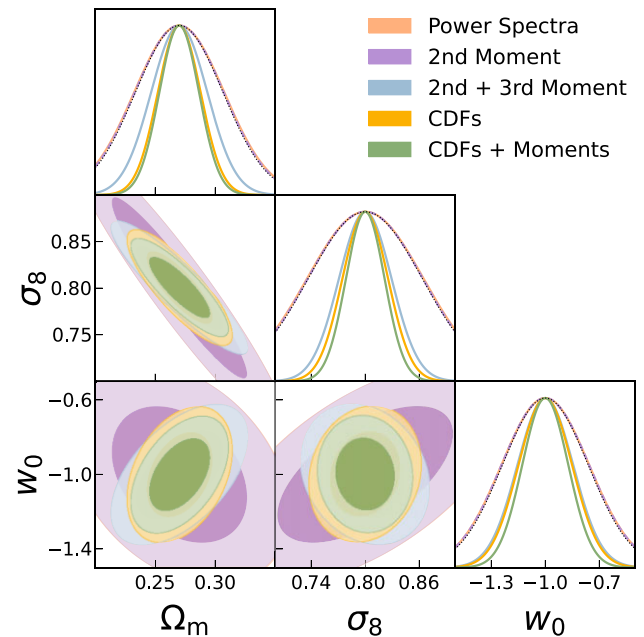
The Hartlap factor for all data vectors in this work is  $\gtrsim 0.9$ . We have verified that the Fisher information – for all the statistics we present – changes by  $< 1$  per cent even if we halve the number of realizations used to compute the covariance matrix, from  $N = 8000 \rightarrow 4000$ . Similarly, halving the number of realizations used in computing the derivatives,  $N = 400 \rightarrow 200$ , changes the Fisher information by  $< 1$  per cent for most statistics; the one exception is the CDFs, where the change in Fisher information is still at the 5–10 per cent level. However, a numerical uncertainty of this level does not change our qualitative interpretations below.

Fig. 4 shows the Fisher information of each statistic. The parameter constraints are obtained by inverting the Fisher matrix of equation (13). First, we see that the angular power spectra and the second moments have indistinguishable constraints, and this is the expected behaviour as one is simply a transformation of the other; given the  $\mathcal{C}_\ell$ , one can predict the second moments exactly via an integral, and vice versa.<sup>8</sup> We also see that the CDFs measured on a Gaussian version of the simulated Y3-like fields, shown by the grey dotted line in the diagonal panels, have constraints very consistent with those of the power spectrum and second-moment. We show in Appendix B that the statistics used in this figure all follow a Gaussian likelihood even when measured on fully nonlinear, non-Gaussian fields – which is not always the case for higher-order statistics as has been found in previous works (Park et al. 2022; Euclid Collaboration 2023).

Including the third moment alongside the second moment improves the constraints significantly for all parameters. This is primarily because of the different degeneracy directions for the different moments (Gatti et al. 2020, 2022).

The CDFs improve the FoM compared to the combination of second and third moments. This confirms that there is still usable information beyond the third moment in the convergence field, particularly in constraining  $\Omega_m$ . However, the modest improvement in going from the second + third moments to the CDFs (when compared to the increase from second moments to second + third moments) shows that there is less information from the fourth moment and beyond. We explicitly check the information content

<sup>8</sup>This assumes we measure both harmonic space and real space over a wide enough range of scales to perform the transform. The agreement between  $\mathcal{C}_\ell$  and second moments in Fig. 4 then implies we chose an appropriately wide range of scales.



**Figure 4.** The Fisher information of different statistics for  $\sigma_8$ ,  $\Omega_m$  and  $w_0$  when using DES Y3-like data. The power spectra and second moment probe only the Gaussian information and their contours overlap completely (the peach contour is hidden underneath the purple). Adding the third moment significantly improves the constraints, and the CDF, which approximately contains all moments, improves upon that a non-negligible but diminishing amount. The degeneracy direction of second + third moments and the CDFs is also visibly different, and combining them leads to a further 20–30 per cent improvement in constraints. The black dashed lines in the diagonal panels show the 1D constraints from CDFs measured on a purely Gaussian field, and these are consistent with those from the other Gaussian statistics. The constraints are tabulated in Table 1.

of the fourth and fifth moments later in Fig. 6. We have separately verified that the constraining power of the moments approach agrees better with that of the CDFs if we include the fourth and fifth moments in the former.

In general, we find that the CDFs do better than the combination of the second and third moments by around  $\approx 20$  per cent in the three parameters we focus on (Table 1). They are also more compact, meaning the data vector for the CDFs ( $N = 460$ ) is notably smaller than the data vectors for the higher-order moments – from progressively including the fourth moment ( $N = 650$ ) or fifth moment ( $N = 1210$ ) – while still providing constraints that are better than using up to the fifth moment. Combining the CDFs with the second and third moments leads to constraints that are 20–30 per cent better than using just the second and third moments. We have verified in Appendix B that the combined data vector also follows a Gaussian likelihood.

We also use the Figure of Merit (FoM), which is defined as the inverse of the area/volume of the ellipsoid formed by the parameter constraints,

$$\text{FoM}_\theta = \sqrt{\frac{1}{\det(\mathbf{F}^{-1})_\theta}}, \quad (15)$$

where  $\theta$  is the subset of parameters used to define the FoM and in our case is  $\theta \in \{\Omega_m, \sigma_8, w_0\}$ . The FoM metric provides a concise way to summarize the constraining power in a multidimensional parameter space. We list the FoM values of our data vectors in

**Table 1.** The Fisher information constraints for a joint analysis of  $\Omega_m$ ,  $\sigma_8$ , and  $w_0$ , the Figure of Merit [FoM, equation (15)], and the size of the data vectors. All FoM values are normalized by that of the Power Spectra. We show results from DES Y3 on Cosmic Shear (Amon et al. 2022; Secco et al. 2022a), second and third moments (Gatti et al. 2022), and Peaks (Zürcher et al. 2022). For KiDS 1000, we show results from cosmic shear (Asgari et al. 2021b) and a field-level analysis (Fluri et al. 2022). For HSC Y3, we show cosmic shear in real space (Li et al. 2023) and harmonic space (Dalal et al. 2023). The DES constraints from second + third moments use more conservative analysis choices (scale cuts, nuisance parameters, etc.) than the Fisher forecast here, resulting in the looser constraints.

Analysis	$\sigma(\Omega_m)$	$\sigma(\sigma_8)$	$\sigma(w_0)$	FoM	$N_{\text{dof}}$
<i>Fisher information (this work)</i>					
Power spectra	0.037	0.064	0.24	1.00	200
2nd moment	0.037	0.064	0.24	1.02	100
2nd + 3rd moments	0.023	0.029	0.15	2.95	300
CDFs	<b>0.018</b>	<b>0.025</b>	<b>0.15</b>	<b>3.47</b>	<b>460</b>
CDFs + moments	<b>0.016</b>	<b>0.021</b>	<b>0.12</b>	<b>5.01</b>	<b>760</b>
<i>DES Y3</i>					
Cosmic shear	0.051	0.083	–		
2nd + 3rd moments	0.030	0.050	–		
Peaks	0.060	0.099	–		
<i>KiDS-1000</i>					
Cosmic shear	0.050	0.080	–		
Field level	0.096	0.206	0.29		
<i>HSC Y3</i>					
Cosmic shear ( $\xi_{\pm}$ )	0.050	0.090	–		
Cosmic shear ( $C_{\ell}$ )	0.065	0.120	–		

Table 1. Including the third moments improves the FoM, relative to the second moments, by a factor of 3. Including the CDFs improves it by 15 per cent, relative to the FoM of the combination of the second and third moments. Combining the CDFs with the second and third moments improves the latter’s FoM by 65 per cent and the former’s FoM by 40 per cent.

## 5 LENSING CDFS IN DES Y3 DATA

We now discuss measurements of the CDF on the DES Y3 data in Section 5.1, including the non-Gaussian aspect of the noise field in Section 5.2, and then detail the contributions from different effects that can impact the inference process: PSFs in Section 5.3, source clustering in Section 5.4, higher-order shear effects in Section 5.5 and baryonic effects in Section 5.6. Finally, we discuss scale cuts in Section 5.7.

### 5.1 CDF measurement and signal-to-noise

In Fig. 3, we have already shown the DES Y3 measurements in solid lines, with the noise-only data vector in dotted grey lines and the A23 version of DES Y3-like map in the grey dashed lines. There is a clear cosmological signal as evidenced by the difference between the noise-only and DES Y3 measurements. Fig. 5 now shows the signal-to-noise of the cosmological component for each data point in the data vector. This is computed as the residuals normalized by the uncertainty,  $S/N = |\text{CDF}_{Y3} - \text{CDF}_N|/\sigma(\text{CDF}_{A23})$ . We then also combine the statistical significance of the individual points, accounting for the covariance between them, and find a total signal-to-noise of  $S/N = 45.3$ .

If the difference between the signal + noise and noise-only fields is a difference in only their even moments (e.g. variance and kurtosis)

then for the 1-field CDFs (the ‘autocorrelation’ part) in Fig. 5, the S/N of a positive threshold should be similar to that of a negative threshold of the same amplitude. We see some indication of this via visual inspection of the 1-field CDF of the third and fourth tomographic bin. We also see an asymmetry in the S/N, and this is a sign of an additional skewness caused by the signal field – for example, in the (0, 0) bin the amplitude of the yellow line ( $k = 0.006$ ) is higher than the light blue one ( $k = -0.006$ ). Thus, we can also visually see indications that this statistic captures non-Gaussian signatures.

Note that while we quote a signal-to-noise for the full set of residuals, we do not use it as a robust estimate of the amount of information. This is because the CDFs respond to noise and signal nonlinearly,<sup>9</sup> so a  $\chi^2$  statistic is not the ideal way to quantify deviations *if the deviations are large*, which is the case between measurements of the noise-only maps and the noisy convergence maps. The interpretation of a  $\chi^2$  in the large-deviation regime is unclear. Note that this is not a problem for our Fisher forecast as the residuals are small given the shifts in the cosmology parameters, as needed for the derivatives, are also small.

Given the results of Fig. 4, where we find the CDFs are a useful and complementary statistic for constraining cosmology, and Fig. 5, where we find the CDFs in DES Y3 have a clear cosmological signal with signs of both the Gaussian and non-Gaussian part, we would like to now test the robustness of this statistic to the relevant observational effects in the Y3 weak lensing data. We will explore exactly this in the following subsections:

(i) Naturally we would want to know how much of the cosmological information seen in Fig. 5 is non-Gaussian – this requires a more precise understanding of the non-Gaussianity in the noise field (Section 5.2).

(ii) The measured shape of galaxies will have some contributions from the PSF, which can then lead to non-cosmological spatial correlations of the galaxy ellipticities – we find this is negligible (Section 5.3).

(iii) Source galaxies, which trace the density field, will be clustered and this can impact the observed convergence field – this has a noticeable impact (Section 5.4).

(iv) The source clustering also leads to correlations between the shape noise field and the convergence field, as seen in the CDFs – we can model this correlation effectively (Section 5.4).

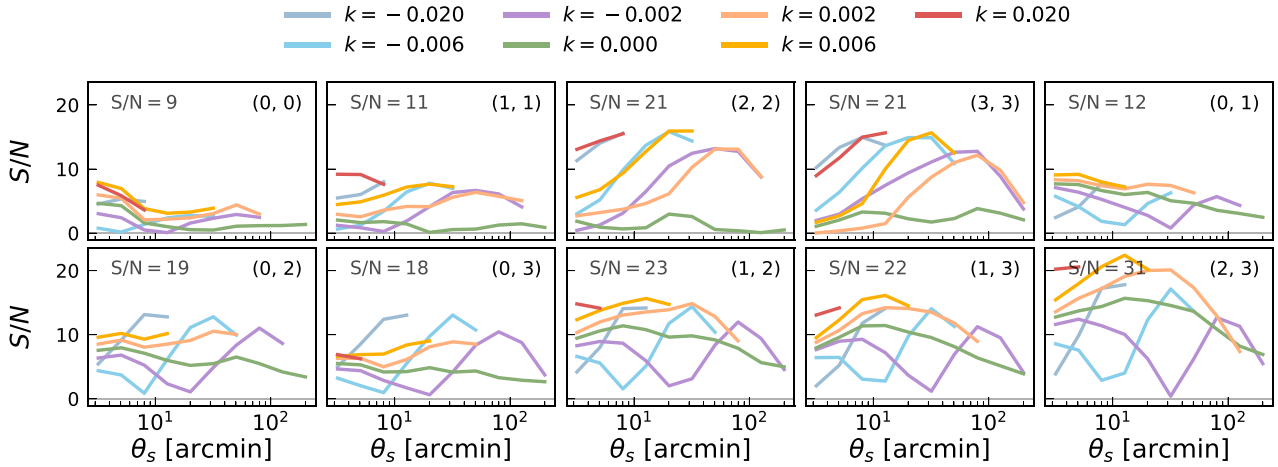
(v) The impact of ignoring higher-order shear effects when modelling the data vector – this is also negligible (Section 5.5).

(vi) The effect of baryonic physics on our statistics – as expected from previous works, this is important (Section 5.6).

(vii) Given the tests above, we detail the analysis choices one would need to make – under our current modelling ability – to robustly infer cosmology using the CDFs (Section 5.7).

The impact from other common systematic factors, such as  $n(z)$  uncertainties, multiplicative bias uncertainties, and intrinsic alignments, is not considered here. These effects can all be marginalized in the inference and modelling process when obtaining cosmological constraints via the CDFs data vector. Such marginalization has already been performed for multiple different analyses of higher-order statistics (e.g. Gatti et al. 2022; Zürcher et al. 2022).

<sup>9</sup>Even in the Gaussian case, the CDF heuristically goes as  $\int \exp[1/\sigma^2] dx$ , so changes in  $\sigma$  lead to highly nonlinear responses in the CDF.



**Figure 5.** The S/N of the DES Y3 data vector. There is a clear signal observed in the CDFs with  $S/N = 45.3$  which is slightly higher than, but generally consistent with, the S/N of the 2-point analyses in DES ( $S/N = 40.2$ , see section IV of Secco et al. 2022b). We show the S/N from individual bin combinations as text in the upper left panels. The upper right text in a panel denotes the bin combinations used in a certain CDF measurement. Note that the measurements are significantly correlated so one cannot trivially add the S/N of different bins together.

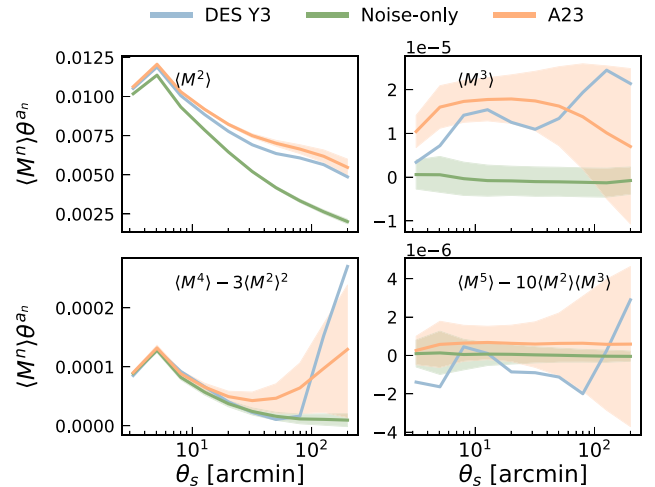
## 5.2 Non-Gaussianity of shape noise fields

To quantify the level of cosmological non-Gaussianity observed by the CDFs, one first needs to understand the non-Gaussianity in the noise field. This is particularly relevant for us as the CDFs are sensitive to all moments of the field, meaning all moments of the cosmological signal but also all moments of the noise field. For this particular investigation, we will switch to using the fields' moments to summarize the noise field and cosmological field at different orders. We do this as the moments can easily isolate the signal from different orders, which helps disentangle the information contained in the CDFs.

Fig. 6 shows the second to fifth moments of DES Y3 mass map, as well as the shape noise map, for the fourth tomographic bin. We find that there is a significant non-Gaussianity in the noise, particularly in the fourth moment and on small scales. Such a feature is naturally expected if the field of source galaxy number counts is not uniform. In the limit that the galaxy counts are uniform across the whole DES Y3 footprint, then every pixel in the map has the same number of galaxies, and thus would have the same shape noise per pixel. In reality, the number of source galaxies per pixel varies across the footprint, either from survey observing conditions or from the intrinsic clustering of sources due to structure formation (see Section 5.4 or Fig. 9). In this case, the noise variance per pixel varies across the footprint, and summing the individual Gaussian noise distributions within the pixels results in a Gaussian mixture model that is symmetric about the  $x = 0$  mean, but can have a significant non-Gaussianity in its even cumulants/moments starting from the kurtosis/fourth moment. This is also consistent with the fact that we detect no odd moments in the noise field.

We also see in Fig. 6 that for DES Y3-like data the cosmological signal exists only in the third and fourth moments. At the fifth moment, the measurement is already consistent with no signal. The noise field has a third moment that is consistent with 0 across the full range of scales. For the fourth moment, however, the noise has a larger fourth moment than the cosmological signal. We can infer this by seeing that the fourth moment of the observed field is very similar to that of the noise-only field.

The significance of the fourth moment in Fig. 6 highlights the need to accurately model the noise field, since almost all the non-



**Figure 6.** The moments of the fourth tomographic bin, as a function of smoothing scales, for the DES Y3 map, the noise-only maps, and the A23 maps. The fourth and fifth moments (bottom panels) have their disconnected components subtracted out. The bands show  $1\sigma$  uncertainties for the noise-only and A23 maps from the  $\mathcal{O}(10^4)$  realizations used in this work. The moments are re-scaled by  $\theta^{a_n}$  as a visualization choice, where  $a_n = n/2$  and  $n \in \{2, 3, 4, 5\}$  is the moment order. The second and third moments have significant information beyond the noise. The fourth moment is significant on the smallest scales, but this contribution is entirely from the noise field since the blue/orange and green lines are almost perfectly overlaid. On larger scales, there is a weak, cosmological signal. The fifth moment is fully consistent with no signal across all scales.

Gaussianity on small scales is coming from the shape noise field rather than the convergence field. Note that some previous works have also shown a strong detection of the fifth moment in the convergence field from data (Van Waerbeke et al. 2013), but they analyse the total fifth moment  $\langle \kappa^5 \rangle$ , whereas here we only consider the connected component, which is obtained as  $\langle \kappa^5 \rangle - 10\langle \kappa^2 \rangle \langle \kappa^3 \rangle$ ,

where  $\langle \kappa^2 \rangle \langle \kappa^3 \rangle$  is the disconnected component.<sup>10</sup> Accounting for this disconnected component is important when isolating the signal in the higher orders. For example, Gaussian distributions have a non-zero fourth moment that must be accounted for – by subtracting out this ‘disconnected’ piece – when measuring non-Gaussian features via the fourth moment. A similar scenario occurs for the fifth moment, where we subtract contributions from lower orders, namely the product of the second and third moments.

### 5.3 PSF contributions

So far we have assumed that spatial correlations between the measured galaxy shapes are a purely cosmological signal. However, this is not guaranteed to be the case as the ellipticities from the PSF can have spatial correlations as well. These correlations have been studied extensively for the 2-point functions (Jarvis et al. 2021), and the work from Gatti et al. (2021); Amon et al. (2022) have explicitly shown their contributions to the cosmological signal/constraints from 2-point functions are negligible. This test has also been done at the 3-point function level (Gatti et al. 2022; Secco et al. 2022b) and found the contributions continue to be negligible. We now replicate this test at the CDF level, which will test the contribution of the PSFs to *all higher-order moments*.

First, we detail the different PSF contributors to the galaxy shapes. The lensing convergence is obtained from the lensing shear maps, which in turn are obtained from individual galaxy ellipticities. The measured ellipticity of a single galaxy can be separated into distinct components,

$$\mathbf{e}^{\text{obs}} = \mathbf{e}^{\text{gal}} + \mathbf{e}^{\text{shear}} + \alpha \mathbf{e}^{\text{psf, true}} + \beta \Delta \mathbf{e}^{\text{psf, err}} + \gamma \Delta T \mathbf{e}^{\text{psf, true}}, \quad (16)$$

where  $\mathbf{e}^{\text{gal}}$  is the intrinsic ellipticity of a given galaxy,  $\mathbf{e}^{\text{shear}}$  is the ellipticity modification due to weak lensing from foreground structure,  $\mathbf{e}^{\text{psf, true}}$  is the PSF ellipticity,  $\Delta \mathbf{e}^{\text{psf, err}}$  is the PSF ellipticity error<sup>11</sup>, and  $\Delta T \mathbf{e}^{\text{psf, true}}$  is the PSF size error<sup>12</sup>. The first quantity of equation (16) is assumed to average to zero,  $\langle \mathbf{e}^{\text{gal}} \rangle = 0$ , while the PSF components can still make a non-zero mean contribution. The coefficients,  $\alpha$ ,  $\beta$ ,  $\gamma$  connect the PSF components to their effective contributions on the measured shear. The values of these coefficients can be measured directly from the data, and we use the values reported in Gatti et al. (2021, see their table 2) of  $\alpha = 0.001$ ,  $\beta = 1.09$  and  $\gamma = -0.5$ . These PSF-based ellipticities can then be used to make a ‘PSF mass map’ in the same way galaxy ellipticities are used to make the DES Y3 mass map. In practice, we make three PSF maps for each of the three PSF components in equation (16) and sum them together in the end.

We test the impact of PSFs on the CDFs by comparing measurements between two types of maps. The first type of map is the sum of the cosmological signal from the A23 simulations, the Y3-like shape noise field, and a PSF mass map for each of the three individual PSF terms of equation (16). The second type of map contains the same signal and noise fields as the first, but the PSF mass map is now created after rotating all the PSF-related

ellipticities in random directions. Thus the first map preserves any PSF-based spatial correlation signals, whereas the second map removes such correlations. Therefore, the residuals between the CDF measurements on these two maps quantify the significance of the PSF ellipticities being spatially correlated, which in turn quantifies how much this non-cosmological spatial correlation will contaminate our signal.<sup>13</sup> Note that we add the *same* PSF mass map to all tomographic bins. We make 8000 DES Y3 maps of each type, using the 8000 independent realizations in the A23 suite. All results are averages over these realizations.

We show in Fig. 7 the significance of the residuals between these two maps as measured by the CDFs, averaged over 8000 realizations. The results show that the significance of the PSF contribution is below  $0.1\sigma$  for all bins, scales, and thresholds. More importantly, we also show the cosmology signal seen by the CDFs – the same results from Fig. 5 – and find the PSF contribution is multiple orders of magnitude below the cosmological signal, which has a significance of  $3\sigma$ – $10\sigma$ . This also confirms that the PSF contributions at the DES Y3 data quality are negligible even beyond the 3-point information.

Note that there are some dipping/valley features in both the dashed and solid lines, which are locations where the residuals switched between positive and negative values.<sup>14</sup> This crossing implies there are scales where the residuals from the cosmological signal, at a given convergence threshold, are zero. This does not coincide with the scales where the same zero crossing occurs for the PSFs. So in principle, for a given threshold, there can be certain scales where the PSFs contribute more than the cosmological signal. However, this contribution would still be between 1–10 per cent of the measurement uncertainty and thus is not a concern for cosmological constraints.

### 5.4 Source clustering

Surveys observe the lensing field sampled at the location of source galaxies, and the ellipticities of these source galaxies are then used to infer the original lensing and convergence fields. The standard prediction for the convergence correlations has an additional correction because the source galaxies do not uniformly sample the lensing field and are themselves clustered given they trace the underlying, clustered density field.

This clustering of source galaxies impacts the observed convergence as follows: the  $n(z)$  of a survey details the weighting of the convergence field at different redshifts, and is computed across the whole survey footprint. However, the precise  $n(z)$  varies across the sky. For example, at redshift  $X$  in direction  $\hat{a}$ , we can have a significant overdensity of structure. This means the  $n(z)$  in the  $\hat{a}$  direction has more galaxies at redshift  $X$ , and the  $n(z)$  must be reweighted accordingly. We will refer to this effect henceforth as source clustering (SC), as was first denoted in Bernardeau (1998), though this effect has also been called source–lens clustering (Hamana et al. 2002). The effect of source clustering is not present in the fiducial post-processing technique described in Section 3.3. However, it can be

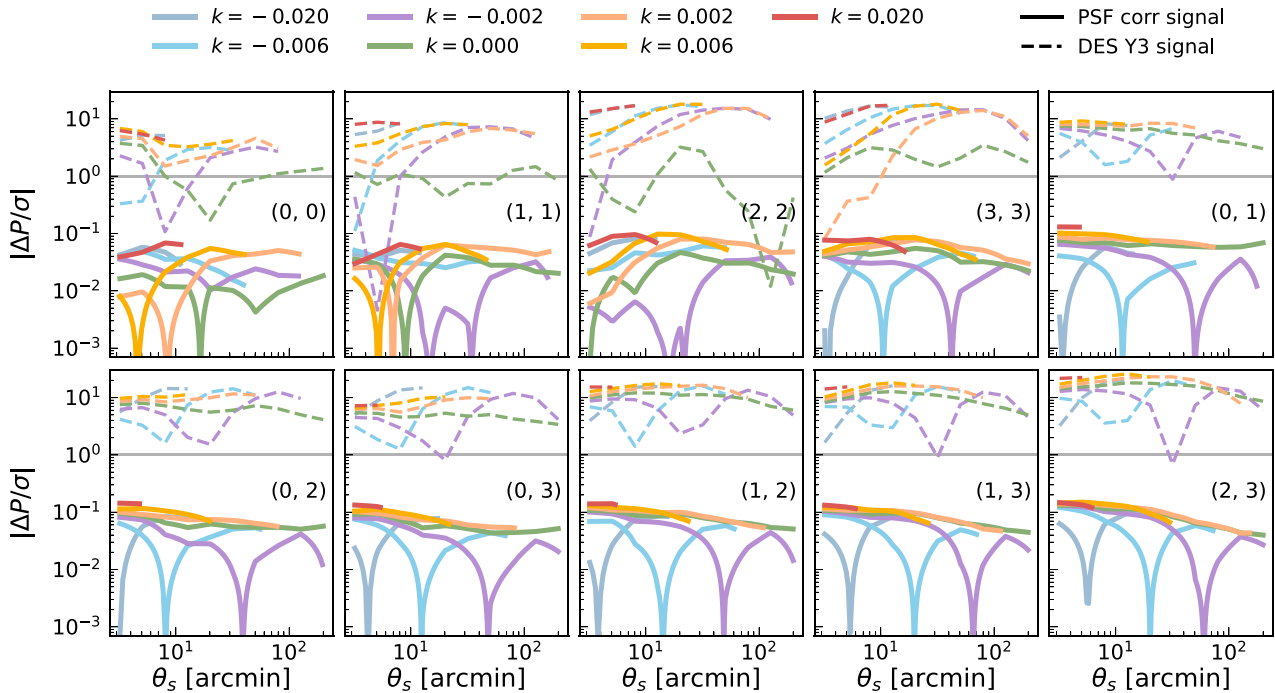
<sup>10</sup>The factor of 10 can be seen by writing all unique combinations of  $\langle \kappa_i \kappa_j \rangle \langle \kappa_k \kappa_l \kappa_m \rangle$ , which is the disconnected fifth moment, with  $i, j, k, l, m \in \{0, 1, 2, 3\}$ . There are 10 unique combinations.

<sup>11</sup>This is defined as  $\mathbf{e}^{\text{psf, true}} - \mathbf{e}^{\text{psf}}$ , which is the difference between the ellipticity of a star (the ‘true’ PSF) and that of the PSF model evaluated at the star’s position.

<sup>12</sup>This is defined as  $\Delta T = (T^{\text{psf, true}} - T^{\text{psf}})/T^{\text{psf}}$ , the fractional difference between the size of a star (the ‘true’ PSF size) and the size of the PSF model evaluated at the location of the star.

<sup>13</sup>One could also compare maps with and without the PSF mass map. However, this would simply show that the PSF shapes are elliptical, which is already a well-established fact (Jarvis et al. 2021).

<sup>14</sup>Such a feature is expected if the noise-only measurement has a certain shape to it. Other higher-order statistics, such as weak lensing peaks, also find nodes in their data vector where  $\text{signal} - \text{noise} = 0$  (Zürcher et al. 2022, see their fig. 5). This does not imply a lack of any cosmological signal, and is simply a consequence of the different shapes of the observed data vector and noise-only data vector.



**Figure 7.** The difference in CDFs measured on two DES Y3-like simulated maps. One contains the Y3 PSF mass map, and the other contains a PSF mass map obtained after rotating all the PSF-based ellipticities. The contribution of any correlations from the PSF (solid lines) is below  $<0.1\sigma$  and is statistically negligible for all thresholds (different colours). It is also 2–3 orders of magnitude below the cosmological signal in DES Y3 (dotted lines). The total signal-to-noise of PSF-induced correlations is  $0.3\sigma$ .

included through the prescription detailed in Gatti et al. (2023, see their equation 5) and previously used in Gatti et al. (2020),

$$\gamma_{\text{SC}}(\hat{\mathbf{n}}) = \frac{\int n(z)(1 + b_g \delta(\hat{\mathbf{n}}, z)) \gamma(\hat{\mathbf{n}}, z) dz}{\int n(z)(1 + b_g \delta(\hat{\mathbf{n}}, z)) dz}, \quad (17)$$

where  $n(z)$  is the source redshift distribution of the tomographic bin, averaged across the survey footprint,  $\delta(\hat{\mathbf{n}}, z)$  and  $\gamma(\hat{\mathbf{n}}, z)$  are the density and true shear maps at a given direction/pixel and redshift, and  $b_g$  is the source galaxy bias. In simple terms, equation (17) modulates the  $n(z)$  across the survey footprint by reweighting it in a direction-dependent way using the density fields. Note that Gatti et al. (2023) take  $b_g = 1$ , which we follow in this work as well, and this is a fair approximation for source galaxies which tend to be mostly blue galaxies. We make 8000 DES Y3 maps with source clustering, using the 8000 independent realizations in the A23 suite. All results are averages over these realizations.

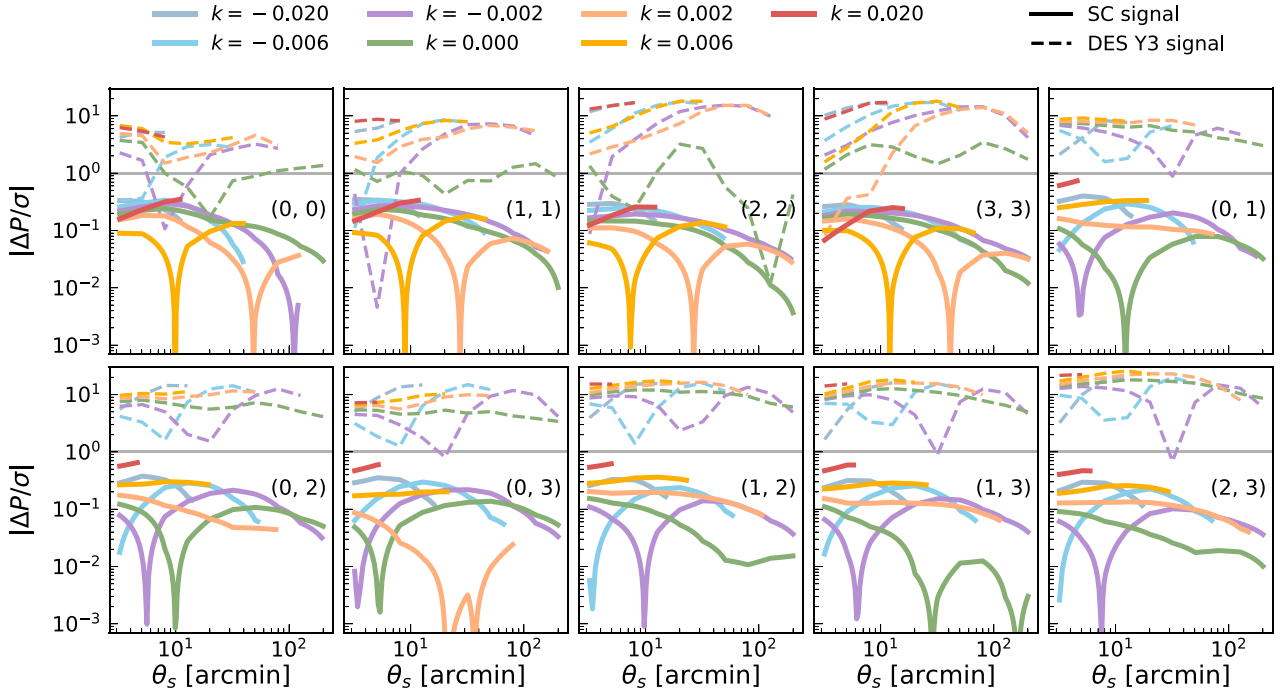
In Fig. 8, we show the difference in the CDF data vector measured on a convergence field with/without source clustering. Both sets of simulations have the same noise field, which is described in Section 3.3. Thus, Fig. 8 presents the impact of source clustering on the cosmological signal. We find here that the impact on the CDFs is at most  $0.1 - 0.5\sigma$ , and it is generally 1–10 per cent that of the cosmological signal. Gatti et al. (2020, 2023) show the impact of source clustering on the second and third moments is at the 1–10 per cent level as well. Krause et al. (2021) show that the source clustering effect on cosmic shear 2-point functions leads to negligible bias ( $<0.15\sigma$ ) in cosmological parameter constraints, but this result is obtained after performing fiducial scale cuts which remove scales where the impact of source clustering is most prominent. Thus these findings are still consistent with our statement above that source clustering is a  $0.5\sigma$  effect on small scales.

We have thus far checked the impact of source clustering on the convergence field. However, source clustering will also induce a correlation between the true convergence field and the shape noise field. Both the convergence field and the source galaxy number density field depend on the density field, and are thus correlated with one another. Given the noise depends inversely on the source galaxy number density as  $\sigma_\kappa \propto 1/\sqrt{n_{\text{gal}}}$ , the convergence field is anticorrelated with the noise field. For example, consider two redshift bins A and B, with  $z_A > z_B$ . If there is an overdensity in bin B, it would simultaneously induce a large convergence in bin A and a suppressed noise in bin B, causing an anticorrelation between the convergence field of bin A and the noise field of bin B.

Gatti et al. (2023) describe a simple modification of the noise field that models this correlation,

$$\gamma_{\text{SC, noise}}(\hat{\mathbf{n}}) = F(\hat{\mathbf{n}}) \left( \frac{\int n(z) dz}{\int n(z)(1 + b_g \delta(\hat{\mathbf{n}}, z)) dz} \right)^{1/2} \gamma_{\text{noise}}(\hat{\mathbf{n}}), \quad (18)$$

where the definitions are the same as equation (17), with  $\gamma_{\text{noise}}(\hat{\mathbf{n}})$  as the shape noise field, which is obtained as described in Section 3.3; by using the DES Y3 galaxy shape catalog, and randomly rotating the galaxy orientations. The density factor in equation (18) varies the number counts of source galaxies across the sky according to the underlying density field. This is the same source clustering effect discussed above but we now consider its effect on the shear noise field,  $\gamma_{\text{noise}}$ , rather than the true shear field,  $\gamma$ . As a consequence of the density-based reweighting, the even moments (variance, kurtosis etc.) of the modified noise field,  $\gamma_{\text{SC, noise}}(\hat{\mathbf{n}})$ , are slightly inconsistent with those of the original noise field  $\gamma_{\text{noise}}(\hat{\mathbf{n}})$ . The factor  $F(\hat{\mathbf{n}})$  is implemented as a correction for this inconsistency [see section 3 of Gatti et al. (2023) for a more detailed discussion], and is



**Figure 8.** The difference in CDFs measured on two DES Y3-like simulated maps, where one map contains source clustering and the other does not. The signal from source clustering (solid lines) is at  $0.1\sigma$ – $0.5\sigma$  and generally contributes  $\approx 5$ – $10$  per cent to the total signal. The total signal-to-noise of source clustering-induced residuals is  $1.3\sigma$ .

modelled as

$$F(\hat{\mathbf{n}}) = A\sqrt{1 - B\sigma^2(\hat{\mathbf{n}})}, \quad (19)$$

where  $\sigma^2(\hat{\mathbf{n}}) = \gamma_{\text{noise},1}^2(\hat{\mathbf{n}}) + \gamma_{\text{noise},2}^2(\hat{\mathbf{n}})$  is the shear variance, summed over both components, in a given direction/pixel and for a given noise realization. The coefficients  $A$  and  $B$  are calibrated in Gatti et al. (2023) for the four DES Y3 bins using the COSMOGRID simulations, with values  $A \in \{0.97, 0.985, 0.990, 0.995\}$  and  $B \in \{0.1, 0.05, 0.035, 0.035\}$ . We have verified that the results of Fig. 9 below are insensitive to the inclusion/exclusion of  $F(\hat{\mathbf{n}})$  in equation (18), which is expected as they focus on the *correlations* between fields, rather than the *covariance* between them.

The correction to the noise field in equation (18) is known to improve the modelling of the third moments, which are sensitive to such convergence–shape noise correlations (Gatti et al. 2023). We post-process our simulations using equations (17) and (18) to obtain convergence maps with such correlations. We then quantify the statistical significance of these correlations, as determined by the CDFs measured on these maps. The CDFs are a useful tool here as they inherit the properties of the kNN distributions, which are the discrete-field version of the CDFs and are a higher signal-to-noise estimator than the 2-point function for determining whether two fields are correlated (Banerjee & Abel 2021b, see their fig. 5).

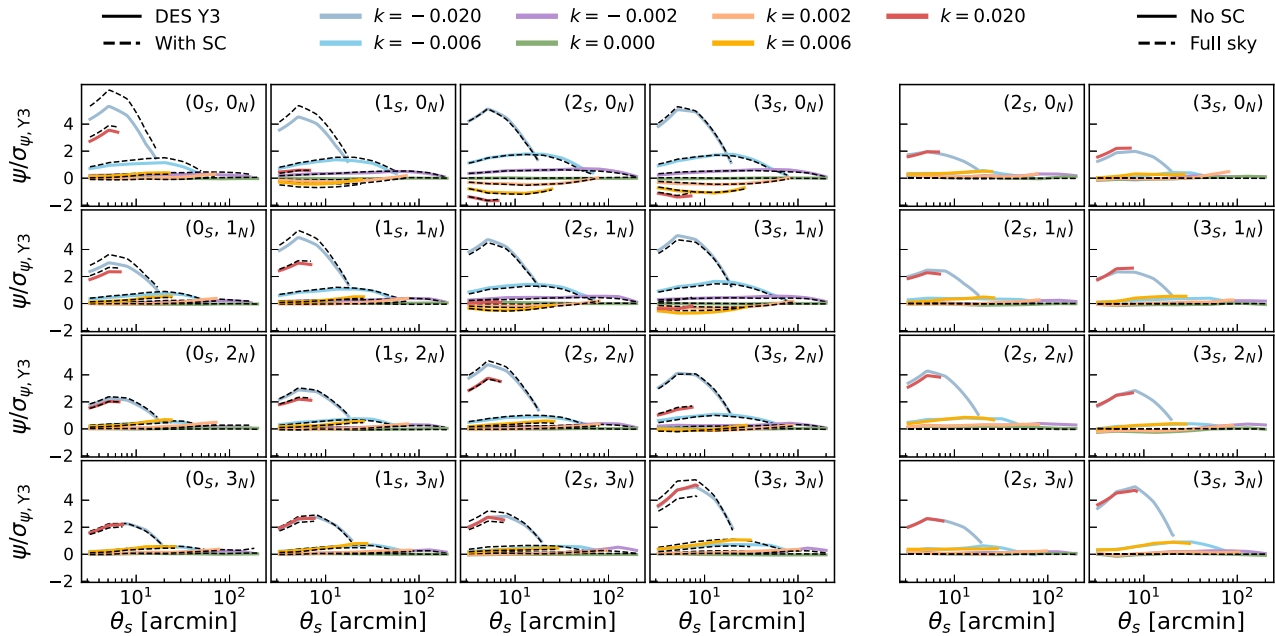
Fig. 9 shows the convergence–shape noise correlation as seen in the CDFs. Instead of the 2-field CDFs, we show the cross-component defined in equation (6) and normalize it by the uncertainty in these correlations, estimated across 1000 DES Y3 realizations. Thus, the presented quantity can be interpreted as a significance of correlation. In the left panels are the results from DES Y3 and from the A23 simulations with source clustering. The DES Y3 result is the mean data vector from correlating the same DES Y3 mass map with 1000 different noise maps. The right panels show A23 simulations

without source clustering, and finally the A23 simulations with purely Gaussian noise and no survey mask.

The exclusion of source clustering leads to a simulated model that is clearly different from what is observed in the data, and including source clustering brings the model and data into good agreement. The right panels of Fig. 9 show that even if we do not include source clustering, there are correlations between the simulated mock maps. Such correlations are expected due to the survey observing properties. The first such cause is survey depth variations, which modulate the source galaxy number density across the sky in the same way for all noise realizations and tomographic bins. The second is the presence of a common survey mask when we perform the KS reconstruction, which induces features in the map that are correlated across independent noise realizations given they all share the same mask. The black dashed lines in the right panels of Fig. 9 confirm that a full-sky analysis with Gaussian shape noise and no survey mask – which by construction has removed the survey property-based effects discussed above – has no convergence–shape noise correlations.

Focusing on the top row of the left panels, we see correlations measured by positive thresholds flip signs depending on the tomographic bin of the convergence field (indexed as  $\mathcal{A}_S$ ). In the absence of source clustering, the KS inversion artefacts cause a positive correlation between the noise and signal field. As we consider convergence fields of higher redshift bins (leaving the noise field fixed at a particular redshift bin), source clustering effects grow in amplitude and result in a 3-point anticorrelation between the noise and convergence field (Gatti et al. 2022, see their Fig. 14). This causes measurements from positive (negative) thresholds to take negative (positive)  $\psi$  values. The threshold-dependent differences in the sign of  $\psi$  highlight the non-Gaussian nature of the induced correlations.

Fig. 9 also shows that convergence–shape noise correlations are statistically significant in the data vector and so are a necessary component in forward-modelling the CDFs. This is also true of



**Figure 9.** The correlation between two fields, which are the observed convergence field – either from DES Y3 data or forward modelled from simulations – and the simulated Y3-like shape noise fields. We find a significant detection of correlation. The panels show the index of the tomographic bin for the observed field (S) and the shape noise field (N). The left panels show the DES Y3 data and the A23 simulations with source clustering. The right panels show a subset of correlations for two other types of simulations – one with no source clustering, and one with Gaussian noise and no survey mask. The simulations with no source clustering show a clear difference from those with it included. However, even without source clustering, the observed field is correlated with the noise field, and this is due to performing KS reconstruction with a survey mask. We also measure the CDFs on full sky maps that use Gaussian noise and no survey mask. In this regime, the signal and noise fields are completely uncorrelated as expected. The total signal-to-noise of the convergence–shape noise correlation, computed as the difference between the ‘With SC’ and ‘No SC’ models, is  $13\sigma$ . The ‘With SC’ model is within  $3.5\sigma$  of the Y3 measurements.

other higher-order statistics. The analysis of Gatti et al. (2022) found correlations between the signal and noise field but was able to denoise the measurements to remove this effect. This was possible as they used the third moments of the field as their statistic,  $\langle \kappa_{\text{obs}}^3 \rangle = \langle (\kappa_{\text{signal}} + \kappa_{\text{noise}})^3 \rangle$ , and so the noise-dependent terms – such as  $\langle \kappa_{\text{signal}} \kappa_{\text{noise}}^2 \rangle$  – that contributed to the measured moments,  $\langle \kappa_{\text{obs}}^3 \rangle$ , could be subtracted exactly. This can be done for moments of any order. For statistics like the CDFs, however, the data vectors depend on the noise in a nonlinear way, and a simple subtraction will not remove all convergence–shape noise correlations. In this case, we are reliant on an accurate forward model of the shape noise field.<sup>15</sup>

### 5.5 Higher-order shear effects

In equation (16), the contribution to the measured ellipticity from the cosmological component is written as  $\mathbf{e}^{\text{shear}}$ . This is then connected to the shear field,  $\gamma$ , as  $\mathbf{e}^{\text{shear}} = \gamma/(1 - \kappa)$ . In the limit of  $\kappa \ll 1$ , this is approximated to leading order as  $\gamma/(1 - \kappa) \approx \gamma$ . Thus, the measured ellipticities are assumed to directly trace the shear  $\gamma$ , and we ignore higher-order terms, the first of which is  $\gamma\kappa$ .<sup>16</sup> The effect of this approximation is generally known to be subdominant to the cosmological signal (Krause & Hirata 2010). The specific impact on the second and third moments measured in DES Y3 is also known to be negligible, especially when compared to the uncertainties in the

Y3 measurements and to other effects such as baryon imprints (Gatti et al. 2020, see their fig. 4).

In Fig. 10, we show the residuals between CDF measurements made on a mass map where the input true shear field is just  $\gamma$  and a map where the input field is actually  $\gamma/(1 - \kappa)$ . Note that by using  $\gamma/(1 - \kappa)$  rather than the approximation  $\gamma(1 + \kappa + \dots)$  we test the impact of ignoring all higher-order terms in the reduced shear approximation, rather than just the leading order correction,  $\gamma\kappa$ . We then perform the full post-processing pipeline with both map versions. We make 8000 DES Y3 maps for both versions, and our results are averages over all realizations. The differences at the data vector level are within  $<0.1\sigma$  and are subdominant to the signal by multiple orders of magnitude. The impact of this approximation increases with redshift, which is expected as the variance of the  $\kappa$  field increases for source galaxies at higher redshift, and so ignoring the  $1/(1 - \kappa)$  factor has a larger significance.

This result also provides a validation for magnification effects, which at leading order in  $\kappa$  modify the shear as  $\gamma \rightarrow \gamma(1 + q\kappa)$ , where  $q$  is some  $\mathcal{O}(1)$  constant. As was the case with the PSF contributions, these effects have been quantified up to the 3-point function for DES Y3 (Gatti et al. 2020), and we have now implicitly extended it to include higher-order moments through our focus on the CDFs.

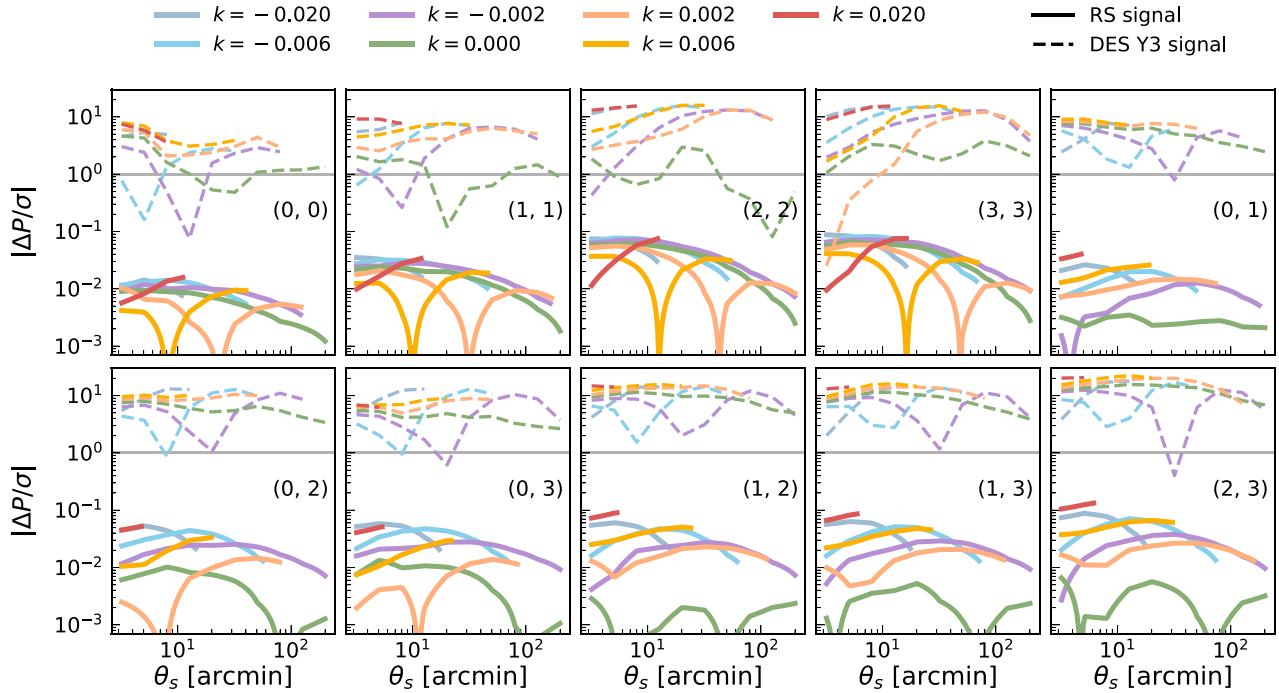
### 5.6 Baryon imprints

Finally, we check the impact of baryon modelling on this statistic. Over the past decades, it has been well-established that galaxy formation processes like gas cooling and AGN (Active Galactic Nuclei) feedback can alter the distribution of total matter within and around haloes (Blumenthal et al. 1986; Gnedin et al. 2004;

<sup>15</sup>It may still be possible to approximately denoise the CDFs, but we have not explored this possibility in this work.

<sup>16</sup>This can be seen by expanding the reduced shear expression as a Taylor series around  $\kappa = 0$ , which gives  $\gamma/(1 - \kappa) \sim \gamma(1 + \kappa + \kappa^2/2 + \dots)$ .





**Figure 10.** The difference in CDFs depending on whether or not we account for reduced shear effects,  $\Delta P = P^{\text{RS}} - P^{\text{fid}}$ . The high-redshift bins, especially when looking at the 2-field CDFs, see the largest impact given source planes at high redshift have larger values of  $\kappa$  and thus the  $1/(1 - \kappa)$  term for the reduced shear is larger. The deviations are still within  $\leq 0.1\sigma$  in all cases and are 2–3 orders of magnitude below the cosmological signal. The total signal-to-noise of reduced shear-induced residuals is  $0.3\sigma$ .

Duffy et al. 2010), which consequently will impact the weak lensing signal (Chisari et al. 2018). These baryonic imprints have a strong mass/redshift dependence (Lovell et al. 2018; Beltz-Mohrman & Berlind 2021; Anbajagane, Evrard & Farahi 2022a) and this mass/redshift-dependent impact on the halo potential can vary across simulation prescriptions (e.g. Shao, Anbajagane & Chang 2022; Anbajagane et al. 2022b).

Recently, Schneider et al. (2019) implemented a halo-based model that can alter  $N$ -body simulations – which are cheaper to run than full hydrodynamic simulations with galaxy formation – to then model the baryon imprints on the density/convergence field. This technique provides a higher-level, approximate galaxy formation model that depends only on ‘macro’ properties like the halo baryon fraction, the baryon density profiles, dark matter density profile etc. and the flexibility manifesting from the method’s approximate nature is particularly useful both for matching the range of halo property scaling relations found in the latest hydro simulations (e.g. Anbajagane et al. 2020, 2022b; Lim et al. 2021; Cui et al. 2022; Lee et al. 2022; Stiskalek et al. 2022; Anbajagane, Evrard & Farahi 2022a) and for handling differences between the evolution of gas in observations and simulations as found in different analyses (e.g. Hill et al. 2018; Amodeo et al. 2021; Pandey et al. 2022; Anbajagane et al. 2022c, 2023a).

In this section, we once again compute residuals between CDFs measured on maps from  $N$ -body simulations and maps that have been ‘baryonified’. Both sets of maps used in this section come from the COSMOGRID suite, and the baryonification was performed with the same model as Schneider et al. (2019). The parameters of the baryonification model were all given their default values, except for some of the gas model parameters which we given values of  $M_c = 13.82$  and  $\nu = 0$ . These parameters are part of a reparametrization

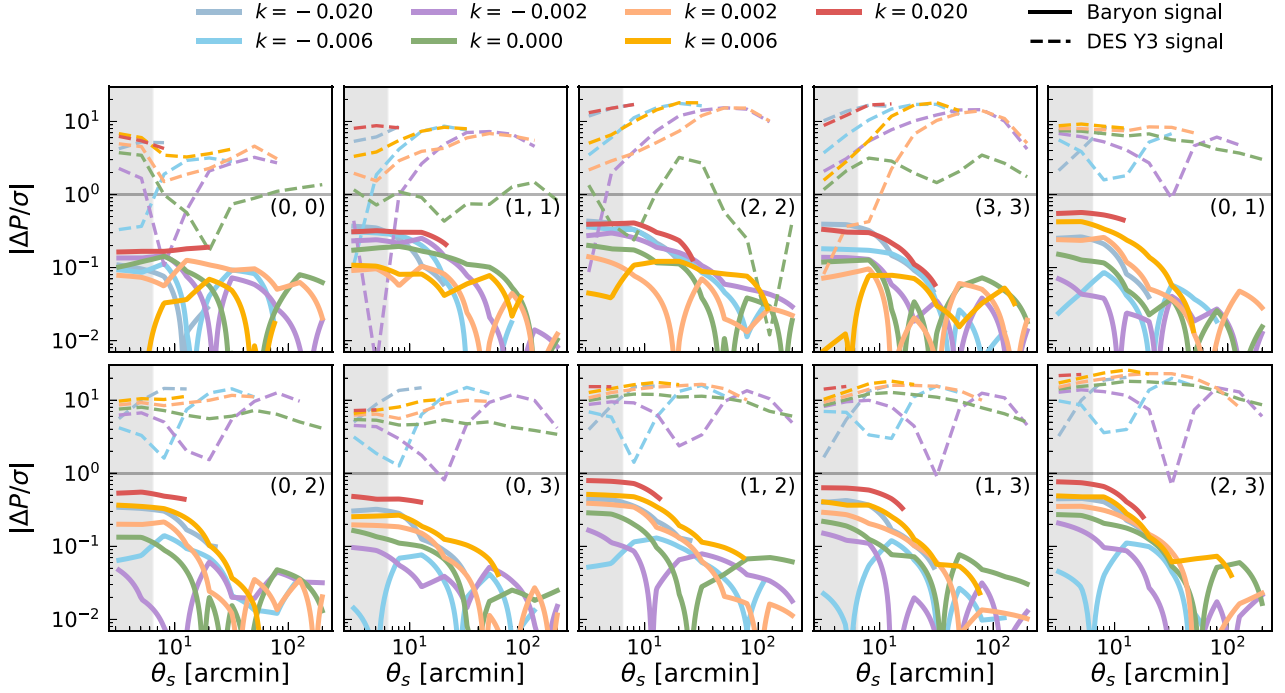
done in Fluri et al. (2022) and control the gas density profiles’ slopes. We take the true convergence fields from COSMOGRID and post-process them using the same pipeline described in Section 3.3. We make 800 DES Y3 cutouts from each set of maps. All results are averages over these realizations.

Fig. 11 shows the residuals due to baryonic imprints on DES Y3-like mock maps. In all cases, the baryon impacts are below  $1\sigma$ . However, note that the maps from COSMOGRID have a resolution of  $\text{NSIDE} = 512$ , and thus the pixel resolution is  $6.4'$  arcmin, instead of the  $3.2'$  arcmin minimum scale used in this work. Since the baryons’ dominant contribution is on smaller scales, it is likely that the *true* residuals at  $3' < \theta < 6.4'$  are actually larger than what is presented in Fig. 11 but are currently suppressed due to the pixel resolution of the COSMOGRID maps. Nevertheless, we can state that the baryon imprints for  $\theta > 10'$  have a significance that is approximately 1–2 orders of magnitude below the cosmological signal.

The impact is also highest for the extreme thresholds in the CDF – the  $k = -0.006$  and  $k = -0.020$  thresholds – and this has been seen in previous, theoretical works. Osato, Liu & Haiman (2021) compared hydrodynamic simulations with a dark matter-only counterpart and showed the lensing PDF can be impacted by more than 10 per cent at the tails of the distribution (see their fig. 5). Sunseri, Li & Liu (2023) used the same set of simulations to show that the impact of baryons on haloes, filaments, and voids affects different parts of the matter PDF.

### 5.7 Scale cuts

In the above sections, we have determined the impact of different systematics and modelling approximations on the CDF data vector. Some systematics are negligible for the whole data vector, such



**Figure 11.** The difference in CDFs measured on dark matter-only (DMO) simulations and ‘baryonified’ DMO simulations. As expected, baryon imprints are a significant effect on the data vector. The grey band shows the scales below  $\theta < 6.4'$ , which is the pixel resolution of the COSMOGRID DES Y3 maps, and is a factor of 2 larger than the other maps we consider in this work. Thus, the baryon effects we estimate below that scale are an underestimate of the true effect given the pixel resolution will suppress these effects. The total signal-to-noise of baryon imprints is  $3.5\sigma$ , though this is a lower bound given the suppression due to map resolution.

as the PSFs (Section 5.3) and the reduced shear approximation (Section 5.5), while others are prominent at a subset of scales, such as baryon imprints (Section 5.6). Thus, using the CDFs to robustly infer cosmological constraints will require us to discard some parts of the fiducial data vector – namely the parts where the amplitude of the systematics is high – and obtain constraints using the remaining fraction of the data vector.

Amongst all the systematic effects considered in this work, the most significant are the baryon imprints (Fig. 11) and the source clustering effect (Fig. 8). These will determine how the data vector is truncated. Our scale cuts are determined by requiring that the parameter bias due to unmodelled systematic effects is below a certain threshold. We compute this bias using the extended Fisher formalism of Amara & Réfrégier (2008) and Asgari et al. (2021a),

$$\Delta_p^{\text{bias}} = \sum_q (F^{-1})_{pq} \frac{d\tilde{X}_{\text{fid}}}{dp} C^{-1} (\tilde{X}_{\text{biased}} - \tilde{X}_{\text{fid}}), \quad (20)$$

where both  $p$  and  $q$  are indexes over the cosmological parameters of interest. The average bias in the data vector,  $\tilde{X}_{\text{biased}} - \tilde{X}_{\text{fid}}$ , is a quantity we have already computed and presented in the above subsections. We then summarize this bias-per-parameter,  $\Delta_p^{\text{bias}}$ , into a bias for the full N-D posterior as

$$\delta = \sqrt{\sum_{p,q} \Delta_p^{\text{bias}} (C^{-1})_{pq} \Delta_q^{\text{bias}}}, \quad (21)$$

where  $C$  is the covariance of the parameters, and so  $C^{-1}$  is just the Fisher matrix,  $F$ . Our procedure for scale cuts is simply removing data points until  $\delta < X$ , where  $X$  is some chosen threshold. We will use  $X \in \{0.3, 0.2, 0.1\}$ . The choice  $X = 0.3$  matches the tests done in the main methodology pipeline for DES Y3 (e.g. Krause et al.

2021; Amon et al. 2022; Secco et al. 2022a) while the other values are chosen to explore more stringent cuts that could be reflective of Stage IV surveys. Note that this threshold,  $X$ , is somewhat arbitrary, but that is not a concern as our goal is to see how the scale-cuts for the CDFs compare to those for the moments; as long as the same choices are applied across both statistics, the arbitrariness of the choices is not relevant.

The other component we must decide is how to determine and discard data points to achieve the condition  $\delta < X$ , as there is significant freedom in doing so. We could throw away all data points for every bin/threshold corresponding to aperture scales below a certain chosen value. However, the choice of a fixed scale threshold is suboptimal as the impact of systematics at a chosen scale varies across bins and thresholds (as seen in any of the Figures above). Thus, our choice here is a scale cut done bin-by-bin (and threshold-by-threshold, in the case of CDFs) and follows the approach of Amon et al. (2022); Secco et al. (2022a). We compute the chi-squared of a given effect in a specific tomographic bin combination (and also specific threshold, in the case of CDFs), and remove the data points corresponding to the smallest scales until we satisfy the relation,

$$(\tilde{X}_{\text{sub,based}} - \tilde{X}_{\text{sub,fid}}) C_{\text{sub}}^{-1} (\tilde{X}_{\text{sub,based}} - \tilde{X}_{\text{sub,fid}})^T < \Delta\chi_{\text{thresh}}^2, \quad (22)$$

where  $\tilde{X}_{\text{sub,based}}$  and  $\tilde{X}_{\text{sub,fid}}$  are subsets of the data vectors used in equation (20), where the subsets correspond to specific tomographic bin combination (and threshold, when using CDFs),  $C_{\text{sub}}$  is the covariance matrix of the subset, and  $\Delta\chi_{\text{thresh}}^2$  is the maximum change in  $\chi^2$  we allow for the full data vector. In practice, we vary  $\Delta\chi_{\text{thresh}}^2$  until the parameter bias goes below our required threshold. The data points that have been removed to achieve this condition define the scale cuts.

**Table 2.** The Fisher information constraints presented in this work for CDFs measured on simulations and for a joint analysis of  $\Omega_m$ ,  $\sigma_8$ , and  $w_0$ , but after implementing various types of scale cuts. From top to bottom, we do (i) simple, fixed angular scale cuts, and then cuts based on (ii) baryonic imprints and (iii) source clustering. The cuts are made by removing data points until  $\delta < X$ , where  $\delta$  – defined in equation (21) – is the total parameter bias in a full N-D parameter space. We show the size of the modified data vector in the rightmost column. The FoM is quoted relative to the FoM of the CDFs constraints with no scale cuts.

Scale Cut	$\sigma(\Omega_m)$	$\sigma(\sigma_8)$	$\sigma(w_0)$	FoM	$N_{\text{dof}}$
<i>Fixed angular cuts</i>					
$\theta > 3'$	<b>0.018</b>	<b>0.025</b>	<b>0.15</b>	<b>1.0</b>	<b>460</b>
$\theta > 10'$	0.022	0.032	0.18	0.85	410
$\theta > 20'$	0.025	0.035	0.21	0.59	270
<i>Baryonic imprints cuts</i>					
$\delta < 0.3$	0.033	0.053	0.33	0.14	129
$\delta < 0.2$	0.035	0.061	0.39	0.10	100
$\delta < 0.1$	0.036	0.062	0.42	0.08	92
<i>Source clustering cuts</i>					
$\delta < 0.3$	0.038	0.063	0.44	0.07	84
$\delta < 0.2$	0.038	0.078	0.57	0.05	71
$\delta < 0.1$	0.042	0.109	0.82	0.03	34

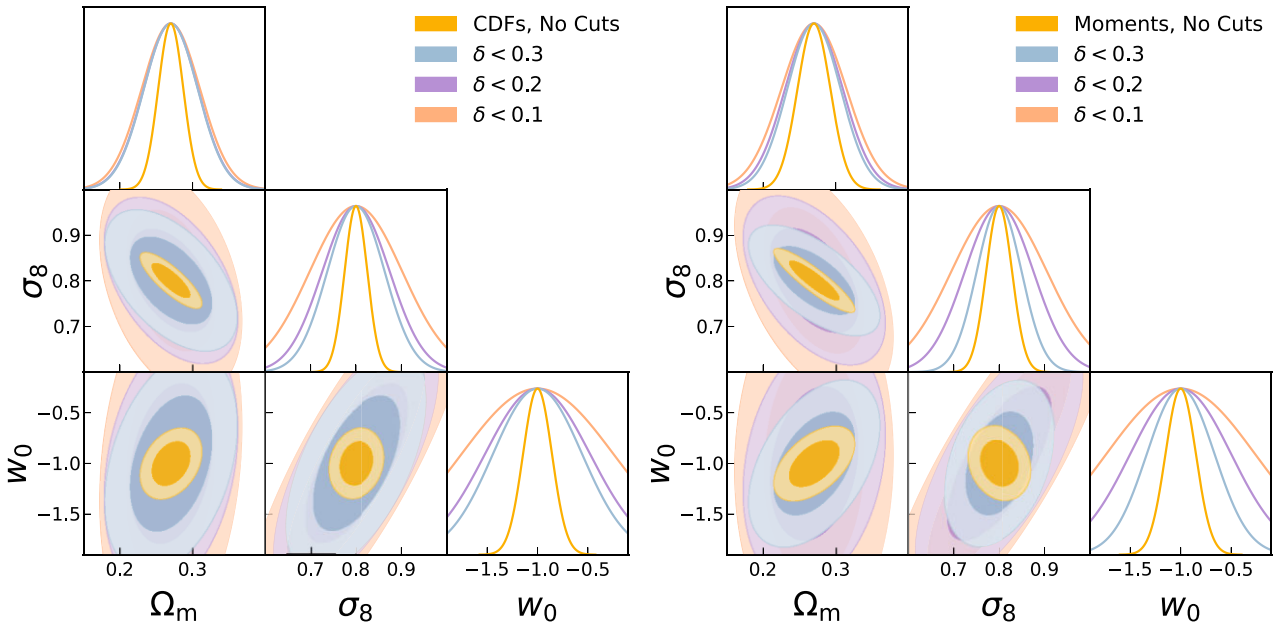
Once the scale-cuts have been defined, we recompute the Fisher constraints using the truncated data vector; the results for the CDFs are tabulated in Table 2. The table also shows constraints from generic scale cuts, where we set a fixed minimum angular scale for all tomographic bins and all thresholds. For the fixed angular scale cuts of  $3.2'$ ,  $10'$ ,  $20'$ , baryonic effects cause a parameter bias of  $\delta = 1.2$ ,  $0.6$ , and  $0.3$ , respectively. Cutting all scales below  $20'$  causes a fractional change of  $\approx 30$  per cent in the constraints. At these scale cuts, the CDFs are comparable to combining second and third Moments, and we have verified that combining the CDFs with the

moments still leads to a 30 per cent improvement in the constraints. The PSFs and reduced shear effect have no impact on scale cuts so we do not show them here. Note that, as was discussed in Section 5.6, the impact of baryonic effects is an underestimate given the baryonified COSMOGRID maps used to estimate the effect have a  $6.4'$  minimum resolution scale. Baryon effects are more impactful at smaller scales and will be more than 10 per cent of the signal if the resolution limit is corrected. However, for our goal of consistently comparing the impacts on CDFs and Moments, this suppression is not a limiting factor.

Table 2 shows that baryon imprints and source clustering both cause notable differences in the parameter constraints, especially in  $\sigma_8$  and  $w_0$ . The FoM in the 3D parameter space drops by a factor of nearly 10 after implementing these scale cuts, which highlights the growing need to improve modelling of these effects instead of robustly trimming the data vector to be insensitive to the effects. Note that while the impact of source clustering on determining the scale cuts is larger than that of the baryonic imprints – which is counter to the standard expectation – this is once again because of the suppression of baryon effects on the small scales due to the resolution scale of the COSMOGRID data products.

Fig. 12 and Table 3 also show the results from defining scale cuts using both baryon imprints and source clustering, and doing so for CDFs and for the second and third moments. This provides a self-consistent reference to compare the two data vectors. The combination of scale cuts is done by looking at both baryonic effects and source clustering, and at each data point we pick the amplitude of the effect that is highest, i.e.  $E = \max |\text{Baryons, SC}|$  for each data point. We find that the moments' constraints are comparable to the CDFs' after these scale cuts. Once we remove  $w_0$  from the analysis the scale cuts cause only a factor of 3 degradation of the FoM as opposed to the factor of 10 if we include  $w_0$ .

Generally, one may expect the CDFs to be less sensitive to these effects than the moments; reduced shear, source clustering, and baryon imprints are all effects that grow with the amplitude of the



**Figure 12.** The Fisher constraints from CDFs (left) and second + third moments (right) measured on simulations. We present four cases, where we either have no scale cuts or cut the data vector so the parameter bias in the  $\Omega_m - \sigma_8 - w_0$  contour is below a certain value; see equation (21). The CDFs and the moments have comparable constraints, which are denoted in Table 3.

**Table 3.** The Fisher information constraints presented in this work for a joint analysis of either  $\Omega_m$ ,  $\sigma_8$ , and  $w_0$  (top two) or just  $\Omega_m$  and  $\sigma_8$  (bottom two), but after implementing scale cuts to reduce the parameter bias. We show the constraints, after scale cuts, for both the CDFs and for the combination of second and third moments. We show the size of the modified data vector in the rightmost column. The FoM is quoted relative to the FoM of the CDFs constraints with no scale cuts.

Scale Cut	$\sigma(\Omega_m)$	$\sigma(\sigma_8)$	$\sigma(w_0)$	FOM	$N_{\text{dof}}$
<i>CDFs, All cuts (<math>\Omega_m, \sigma_8, w_0</math>)</i>					
$\delta\text{CDF} < 0.3$	<b>0.037</b>	<b>0.063</b>	<b>0.44</b>	0.07	84
$\delta\text{CDF} < 0.2$	0.037	0.074	0.52	0.05	75
$\delta\text{CDF} < 0.1$	0.040	0.102	0.75	0.03	49
<i>2nd and 3rd moments, all cuts (<math>\Omega_m, \sigma_8, w_0</math>)</i>					
$\delta\text{Moments} < 0.3$	0.037	0.050	0.34	0.13	109
$\delta\text{Moments} < 0.2$	0.040	0.076	0.50	0.06	86
$\delta\text{Moments} < 0.1$	0.045	0.105	0.74	0.03	58
<i>CDFs, all cuts (<math>\Omega_m, \sigma_8</math>)</i>					
$\delta\text{CDF} < 0.3$	0.033	0.050	–	0.27	99
$\delta\text{CDF} < 0.2$	0.035	0.053	–	0.20	79
$\delta\text{CDF} < 0.1$	0.038	0.057	–	0.16	52
<i>2nd and 3rd moments, all cuts (<math>\Omega_m, \sigma_8</math>)</i>					
$\delta\text{Moments} < 0.3$	0.031	0.044	–	0.34	118
$\delta\text{Moments} < 0.2$	0.036	0.051	–	0.22	91
$\delta\text{Moments} < 0.1$	0.044	0.058	–	0.12	4

density field and/or the convergence field. This means they impact the tails of the density/lensing distribution the most and leave the ‘bulk’ of the PDF – roughly the 68 per cent or the 95 per cent region centred around the median – relatively unaffected. The moments are defined as an integral over the whole distribution and so cannot isolate just parts of it. The CDFs on the other hand *can* perform such an isolation. They fundamentally only probe whether or not a pixel’s convergence is above a given threshold; thus, if the convergence is well above/below the threshold, the measurement of the CDFs is unaffected by that pixel value shifting around due to various effects. For example, the negative thresholds  $k < 0$  will be unaffected by the baryon imprints in massive haloes, as massive haloes exist in  $\kappa > 0$  regions and baryon imprints reduce the  $\kappa$  value but always keep it positive, and so the convergence around haloes will always be above the  $k < 0$ . Of course, if the  $\kappa$  values of interest are near a threshold, then any shifts will have a stronger impact on the CDF measurements at that threshold. This argument also suggests there are a particular choice of thresholds that balance constraining power while alleviating such systematics. We have not explored such an optimal selection. In Table 3, we also redo the scale cuts but now leave out  $w_0$  when computing the total parameter bias, as this is a closer match to the procedures used in Stage III surveys (e.g. Krause et al. 2021). Our qualitative findings remain the same even in this case.

## 6 CONCLUSIONS

In this work, we have explored the use of the Cumulative Distribution Functions (CDFs) of the convergence field as a summary statistic for extracting cosmological information, drawing on the development of the kNN distributions for the discrete fields. The CDFs are a convenient, succinct summary of the field that approximately capture all higher moments of the field in a significantly shorter data vector that is also quicker to compute. We explore the theoretical advantages of using these CDFs and check their sensitivity to the relevant practical challenges in extracting robust cosmology constraints from Y3-like data. The conclusions of this work are as follows:

(i) For scales of  $3' < \theta < 200'$  and tomographic bins of DES Y3, the CDFs have better constraints on  $\Omega_m$ ,  $\sigma_8$  and  $w$  when compared to those from the combination of both second and third Moments (Fig. 4). This improvement is modest, but the CDFs still have a slightly different degeneracy direction to the moments, and combining the CDFs and moments leads to the constraints improving by 20–30 per cent.

(ii) The CDFs measured on a Gaussian field provide Fisher constraints that are completely consistent with the angular power spectra and second Moments computed on the fully nonlinear, non-Gaussian field (Fig. 4). The CDFs and moments all have Gaussian likelihoods as well (Fig. B1).

(iii) The DES Y3 noise field is highly non-Gaussian, with a very significant fourth moment (Fig. 6). There is some cosmological signal at large scales in the fourth moment, but none in the fifth moment.

(iv) We create a PSF ‘mass map’ for testing PSF contributions at the map level, and show the signal from PSF shapes is 2–3 orders of magnitude below the cosmological signal (Fig. 7). This validates not only the CDFs, but also indirectly validates the minimal impact of the PSFs on information beyond the third moment (existing works have already validated them at the second and third moment level).

(v) The presence or lack of spatial correlations in the source galaxy number counts, i.e. ‘source clustering’, impacts the convergence field model at the 1–10 per cent level (Fig. 8).

(vi) The CDFs are sensitive to correlations between the convergence field and the shape noise field, induced by source clustering. We detect these correlations at  $13\sigma$ , and can adequately model them in the simulated maps (Fig. 9).

(vii) The reduced shear approximation changes the cosmological signal at the 1–5 per cent level (Fig. 10), while baryon imprints are 1–10 per cent of the cosmological signal (Fig. 11).

(viii) We perform scale cuts that limit the parameter bias due to systematic effects under a certain level. The cut CDF data vector has comparable constraining power to the cut data vector of the second and third Moments (Table 2 and Fig. 12).

Optimizing the summary of fields is a rich area of study, with a variety of approaches and outcomes. The CDFs, through their sensitivity to all moments of the field, probe both the cosmological signal at all these orders as well as any potential modelling challenges that surface at these orders (e.g. the high kurtosis of the noise field that does not impact 2-point and 3-point functions). This sensitivity to all orders becomes a more relevant trait as we extend our analyses to smaller scales, which are more nonlinear and thus more non-Gaussian. It may also become relevant in constraining – and/or marginalizing over – the impact of baryons on the density field; these effects happen pre-dominantly within haloes, and so are localized around the most nonlinear regions of the density field and thus will have non-Gaussian signatures. The CDFs might also be one of the few ways to probe the highest orders of information in the field. They are more robust given they can isolate specific parts of the distribution, and this is in contrast to the higher order moments which will be increasingly sensitive to noise/outliers in the tails of the distribution. Thus, if there is significant, usable higher-order information in the cosmological field (for example, in future surveys with different noise levels and sensitivities), the CDF may be one of the only ways to robustly access it.

While efforts have already been made to obtain cosmology from up to the third moment, we show there remains some information beyond the third moment that can likely be accessed in a robust manner, i.e. without worrying about systematics. Effects like reduced shear, source clustering, and baryons have some impact that is at

the 0.1 per cent–10 per cent level depending on the effect and the angular scale. After enacting scale cuts to reduce the bias on cosmological constraints to be within  $0.3\sigma$ , the CDF data vector still provides constraints better than those of the second and third moment data vector. We have identified that accurate modelling of the noise field at higher orders is the current limiting factor in robustly inferring cosmology from statistics like the CDFs. Alternatively, an accurate way of denoising the CDFs – which effectively bypasses requirements in modelling the noise field by removing its contribution from the data vector – would enable robust cosmology constraints with the CDFs.

Finally, we note that even though this work has specifically focused on validating the CDF as a summary statistic, the validation results have significant implications for the broader range of lensing convergence statistics discussed in the literature. The key underlying information is the distribution of convergence as a function of scale,  $P(\kappa_\theta)$ , and the CDFs are a convenient and compact way of summarizing this distribution/information. Other statistics summarize this distribution in different ways, such as lensing-in-cells<sup>17</sup> and Minkowski Functionals.<sup>18</sup> As has been discussed above, another closely connected statistic is the moments of the field,  $\langle \kappa_\theta^n \rangle$ , which are a further summary of the distribution,  $P(\kappa_\theta | \theta)$ , and computing moments to an arbitrarily high order is equivalent to computing the CDF to arbitrarily many thresholds.

As we move towards Stage IV surveys with wider survey areas and deeper observations – both leading to higher precision measurements – other systematics could become relevant. As a rough example, the LSST Year 10 data set will have  $\sim 3$  times the survey area as DES Y3 and  $\sim 5$  times the source galaxy number density as DES Y3 (The LSST Dark Energy Science Collaboration 2018), which leads to a factor of 4 increase in precision of the data vector and in the significance of any systematic we discuss in this work. The reduced shear effect (Fig. 10) – which can be safely ignored in Stage III surveys – will likely need to be included in the model for Stage IV, especially for LSST’s highest redshift bins as the amplitude of the effect grows with redshift. However, this component can be trivially included via simulation-based modelling using the same approach we used to include its effects in our simulations (Section 5.5). Source clustering will also be a necessary modelling ingredient for Stage IV surveys as its signal-to-noise will exceed 1 for LSST. While this modelling can also be done through simulation-based modelling, it requires some galaxy bias prescription (equation (17)) which would introduce a modelling uncertainty that has yet to be quantified. Additionally, we discussed that the Born approximation is adequate for modelling the weak lensing field under DES-like uncertainties. However, previous works have shown that for Stage IV data quality we will require ray-tracing when using higher-order statistics (Petri, Haiman & May 2017).

These effects above – reduced shear, source clustering, and Born approximation – impact all statistical summaries of the lensing field, including the standard 2pt and 3pt functions. Systematics that will uniquely impact the CDFs are then effects that generate a fourth moment and beyond. We have already found in this work that the

fourth moment of the noise field is a highly relevant modelling component for the CDFs. In DES Y3, this was primarily sourced by the survey depth fluctuations as well as the intrinsic, cosmological clustering of source galaxies. In general, however, any process that spatially modifies the shape noise per galaxy or the number of galaxies per pixel will generate the fourth moment. For Stage IV surveys, the precision will be high enough that effects such as spatially varying multiplicative bias – which impacts the measured variance of the shape distribution – could also be a required modelling component, but we must first quantify how much this bias will actually vary across the sky.

The validation steps performed in this work have implications for the statistics mentioned above – lensing-in-cells, Minkowski Functionals, field moments etc. For example, it is likely that PSF ellipticity correlations will be a few orders of magnitude below the cosmological signal for all of these statistics. A similar case can be made for the impact of source clustering and the reduced shear approximation. Of course, it is still ideal to perform a separate validation for those statistics to explicitly verify their robustness to these effects, but the results of this work indicate – given the statistics all summarize the same underlying distribution,  $P(\kappa_\theta | \theta)$  – that it is likely these other statistics will also be robust to these. By using the CDFs, which are approximately summarizing all higher order moments, we have tested these systematics at the map level and beyond the third moment. We hope the methodologies for map-level tests that we employed and/or introduced in this work enable more checks of the large library of higher-order statistics that are being developed for the convergence field, and thus enhance the trustworthiness of these newer statistics.

## ACKNOWLEDGEMENTS

DA is supported by NSF grant No. 2108168. CC is supported by the Henry Luce Foundation and DOE grant DE-SC0021949. We thank the referee for their insightful comments that helped improve this work.

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

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

<sup>17</sup>This is the lensing-focused analogue of counts-in-cells, where the latter is the distribution of tracer counts within a given volume,  $P(k_{\text{tr}} | V)$ . If we replace trace counts with lensing convergence, then we obtain lensing-in-cells.

<sup>18</sup>The CDFs are the same as the zeroth-order Minkowski functional, though in our formalism we also introduce a cross-correlation method – inspired by the formalism for kNNs in Banerjee & Abel (2021b) – which is traditionally not used/defined for the Minkowski Functionals.

gan, NSF's NOIRLab, the University of Nottingham, The Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, Texas A&M University, and the OzDES Membership Consortium.

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

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

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

## DATA AVAILABILITY

All DES Y3 data used in this work are publicly available at <https://des.ncsa.illinois.edu/releases/y3a2/>. The ULAGAM simulation suite, which was denoted as A23 throughout this work for brevity, is available at <https://ulagam-simulations.readthedocs.io>.

## REFERENCES

- Adelberger K. L., Steidel C. C., Giavalisco M., Dickinson M., Pettini M., Kellogg M., 1998, *ApJ*, 505, 18
- Allys E., Marchand T., Cardoso J. F., Villaescusa-Navarro F., Ho S., Mallat S., 2020, *Phys. Rev. D*, 102, 103506
- Amara A., Réfrégier A., 2008, *MNRAS*, 391, 228
- Amodeo S. et al., 2021, *Phys. Rev. D*, 103, 063514
- Amon A. et al., 2022, *Phys. Rev. D*, 105, 023514
- Anbajagane D., Evrard A. E., Farahi A., Barnes D. J., Dolag K., McCarthy I. G., Nelson D., Pillepich A., 2020, *MNRAS*, 495, 686
- Anbajagane D., Evrard A. E., Farahi A., 2022a, *MNRAS*, 509, 3441
- Anbajagane D. et al., 2022b, *MNRAS*, 510, 2980
- Anbajagane D. et al., 2022c, *MNRAS*, 514, 1645
- Anbajagane D. et al., 2023a, preprint (arXiv:2310.00059)
- Anbajagane D., Chang C., Lee H., Gatti M., 2023b, preprint (arXiv:2310.02349)
- Asgari M., Friswell I., Yoon M., Heymans C., Dvornik A., Joachimi B., Simon P., Zuntz J., 2021a, *MNRAS*, 501, 3003
- Asgari M. et al., 2021b, *A&A*, 645, A104
- Banerjee A., Abel T., 2021a, *MNRAS*, 500, 5479
- Banerjee A., Abel T., 2021b, *MNRAS*, 504, 2911
- Banerjee A., Abel T., 2023, *MNRAS*, 519, 4856
- Barthelemy A., Halder A., Gong Z., Uhlemann C., 2023, preprint (arXiv:2307.09468)
- Baugh C. M., Gaztanaga E., Efstathiou G., 1995, *MNRAS*, 274, 1049
- Beltz-Mohrmann G. D., Berlind A. A., 2021, preprint (arXiv:2103.05076)
- Bernardeau F., 1998, *A&A*, 338, 375
- Blake C., James J. B., Poole G. B., 2014, *MNRAS*, 437, 2488
- Blumenthal G. R., Faber S. M., Flores R., Primack J. R., 1986, *ApJ*, 301, 27
- Boyle A., Uhlemann C., Friedrich O., Barthelemy A., Codis S., Bernardeau F., Giocoli C., Baldi M., 2021, *MNRAS*, 505, 2886
- Carlstrom J. E., Holder G. P., Reese E. D., 2002, *ARA&A*, 40, 643
- Cataneo M., Uhlemann C., Arnold C., Gough A., Li B., Heymans C., 2022, *MNRAS*, 513, 1623
- Chang C. et al., 2018, *MNRAS*, 475, 3165
- Cheng S., Ménard B., 2021, *MNRAS*, 507, 1012
- Chisari N. E. et al., 2018, *MNRAS*, 480, 3962
- Cui W. et al., 2022, *MNRAS*, 514, 977
- Dalal R. et al., 2023, preprint (arXiv:2304.00701)
- Davies C. T., Cautun M., Giblin B., Li B., Harnois-Déraps J., Cai Y.-C., 2021, *MNRAS*, 507, 2267
- Doux C. et al., 2022, *MNRAS*, 515, 1942
- Duffy A. R., Schaye J., Kay S. T., Dalla Vecchia C., Battye R. A., Booth C. M., 2010, *MNRAS*, 405, 2161
- Euclid Collaboration, 2023, preprint (arXiv:2301.12890)
- Fluri J., Kacprzak T., Refregier A., Amara A., Lucchi A., Hofmann T., 2018, *Phys. Rev. D*, 98, 123518
- Fluri J., Kacprzak T., Lucchi A., Refregier A., Amara A., Hofmann T., Schneider A., 2019, *Phys. Rev. D*, 100, 063514
- Fluri J., Kacprzak T., Lucchi A., Schneider A., Refregier A., Hofmann T., 2022, *Phys. Rev. D*, 105, 083518
- Friedrich O. et al., 2018, *Phys. Rev. D*, 98, 023508
- Friedrich O., Uhlemann C., Villaescusa-Navarro F., Baldauf T., Manera M., Nishimichi T., 2020, *MNRAS*, 498, 464
- Fu L. et al., 2014, *MNRAS*, 441, 2725
- Gatti M. et al., 2020, *MNRAS*, 498, 4060
- Gatti M. et al., 2021, *MNRAS*, 504, 4312
- Gatti M. et al., 2022, *Phys. Rev. D*, 106, 083509
- Gatti M. et al., 2023, preprint (arXiv:2307.13860)
- Giannantonio T., Scranton R., Crittenden R. G., Nichol R. C., Boughn S. P., Myers A. D., Richards G. T., 2008, *Phys. Rev. D*, 77, 123520
- Giblin B., Cai Y.-C., Harnois-Déraps J., 2023, *MNRAS*, 520, 1721
- Gnedin O. Y., Kravtsov A. V., Klypin A. A., Nagai D., 2004, *ApJ*, 616, 16
- Gong Z., Halder A., Barreira A., Seitz S., Friedrich O., 2023, preprint (arXiv:2304.01187)
- Gough A., Uhlemann C., 2022, *Universe*, 8, 55
- Gruen D. et al., 2018, *Phys. Rev. D*, 98, 023507
- Halder A., Friedrich O., Seitz S., Varga T. N., 2021, *MNRAS*, 506, 2780
- Hamana T., Colombi S. T., Thion A., Devriendt J. E. G. T., Mellier Y., Bernardeau F., 2002, *MNRAS*, 330, 365
- Hartlap J., Simon P., Schneider P., 2007, *A&A*, 464, 399
- Heydenreich S., Brück B., Harnois-Déraps J., 2021, *A&A*, 648, A74
- Heydenreich S., Brück B., Burger P., Harnois-Déraps J., Unruh S., Castro T., Dolag K., Martinet N., 2022, *A&A*, 667, A125
- Heydenreich S., Linke L., Burger P., Schneider P., 2023, *A&A*, 672, A44
- Hill J. C., Baxter E. J., Lidz A., Greco J. P., Jain B., 2018, *Phys. Rev. D*, 97, 083501
- Hinshaw G. et al., 2013, *ApJS*, 208, 19
- Hoffmann K. et al., 2022, *Phys. Rev. D*, 106, 123510
- Jain B., Seljak U., White S., 1998, preprint (arXiv:astro-ph/9804238)
- Jarvis M. et al., 2021, *MNRAS*, 501, 1282
- Jeffrey N., Lanusse F., Lahav O., Starck J.-L., 2020, *MNRAS*, 492, 5023
- Jeffrey N. et al., 2021, *MNRAS*, 505, 4626
- Kacprzak T., Fluri J., Schneider A., Refregier A., Stadel J., 2023, *J. Cosmol. Astropart. Phys.*, 2023, 050
- Kaiser N., Squires G., 1993, *ApJ*, 404, 441
- Kratochvil J. M., Haiman Z., May M., 2010, *Phys. Rev. D*, 81, 043519
- Krause E., Hirata C. M., 2010, *A&A*, 523, A28
- Krause E. et al., 2021, preprint (arXiv:2105.13548)
- Kruse G., Schneider P., 2000, *MNRAS*, 318, 321
- Lanzieri D., Lanusse F., Modi C., Horowitz B., Harnois-Déraps J., Starck J.-L., The LSST Dark Energy Science Collaboration, 2023, preprint (arXiv:2305.07531)
- Lee E. et al., 2022, *MNRAS*, 517, 5303
- Li X. et al., 2023, preprint (arXiv:2304.00702)

Lim S. H., Barnes D., Vogelsberger M., Mo H. J., Nelson D., Pillepich A., Dolag K., Marinacci F., 2021, *MNRAS*, 504, 5131

Lovell M. R. et al., 2018, *MNRAS*, 481, 1950

MacCrann N. et al., 2022, *MNRAS*, 509, 3371

Martinet N. et al., 2018, *MNRAS*, 474, 712

Mecke K. R., Buchert T., Wagner H., 1994, *A&A*, 288, 697

Munshi D., Jung G., Kitching T. D., McEwen J., Liguori M., Namikawa T., Heavens A., 2023, *Phys. Rev. D*, 107, 043516

Myles J. et al., 2021, *MNRAS*, 505, 4249

Omori Y., 2022, preprint (arXiv:2212.07420)

Osato K., Liu J., Haiman Z., 2021, *MNRAS*, 502, 5593

Pandey S. et al., 2022, *Phys. Rev. D*, 105, 123526

Park C. F., Allys E., Villaescusa-Navarro F., Finkbeiner D. P., 2022, preprint (arXiv:2204.05435)

Parroni C., Cardone V. F., Maoli R., Scaramella R., 2020, *A&A*, 633, A71

Peacock J. A., 1983, *MNRAS*, 202, 615

Peel A., Pettorino V., Giocoli C., Starck J.-L., Baldi M., 2018, *A&A*, 619, A38

Petri A., Liu J., Haiman Z., May M., Hui L., Kratochvil J. M., 2015, *Phys. Rev. D*, 91, 103511

Petri A., Haiman Z., May M., 2017, *Phys. Rev. D*, 95, 123503

Planck Collaboration, 2016a, *A&A*, 594, A13

Planck Collaboration, 2016b, *A&A*, 594, A16

Planck Collaboration, 2020, *A&A*, 641, A7

Potter D., Stadel J., Teyssier R., 2017, *Comput. Astrophys. Cosmol.*, 4, 2

Schneider A., Teyssier R., Stadel J., Chisari N. E., Le Brun A. M. C., Amara A., Refregier A., 2019, *J. Cosmol. Astropart. Phys.*, 2019, 020

Secco L. F. et al., 2022a, *Phys. Rev. D*, 105, 023515

Secco L. F. et al., 2022b, *Phys. Rev. D*, 105, 103537

Sellentini E., Loureiro A., Whiteway L., Lafaurie J. S., Balan S. T., Olamaie M., Jaffe A. H., Heavens A. F., 2023, preprint (arXiv:2305.16134)

Sevilla-Noarbe I. et al., 2021, *ApJS*, 254, 24

Shan H. et al., 2018, *MNRAS*, 474, 1116

Shao M., Anbajagane D., Chang C., 2022, preprint (arXiv:2212.05964)

Springel V., 2005, *MNRAS*, 364, 1105

Stiskalek R., Bartlett D. J., Desmond H., Anbajagane D., 2022, *MNRAS*, 514, 4026

Sunseri J., Li Z., Liu J., 2023, *Phys. Rev. D*, 107, 023514

Sunyaev R. A., Zeldovich Y. B., 1972, *Comments on Astrophys. Space Phys.*, 4, 173

Takahashi R., Hamana T., Shirasaki M., Namikawa T., Nishimichi T., Osato K., Shiroshima K., 2017, *ApJ*, 850, 24

The Dark Energy Survey Collaboration, 2005, preprint (arXiv:astro-ph/0510346)

The LSST Dark Energy Science Collaboration, 2018, preprint (arXiv:1809.01669)

Uhlemann C., Friedrich O., Villaescusa-Navarro F., Banerjee A., Codis S., 2020, *MNRAS*, 495, 4006

Van Waerbeke L. et al., 2013, *MNRAS*, 433, 3373

Villaescusa-Navarro F. et al., 2020, *ApJS*, 250, 2

Wang Y., Banerjee A., Abel T., 2022, *MNRAS*, 514, 3828

White M., Hu W., 2000, *ApJ*, 537, 1

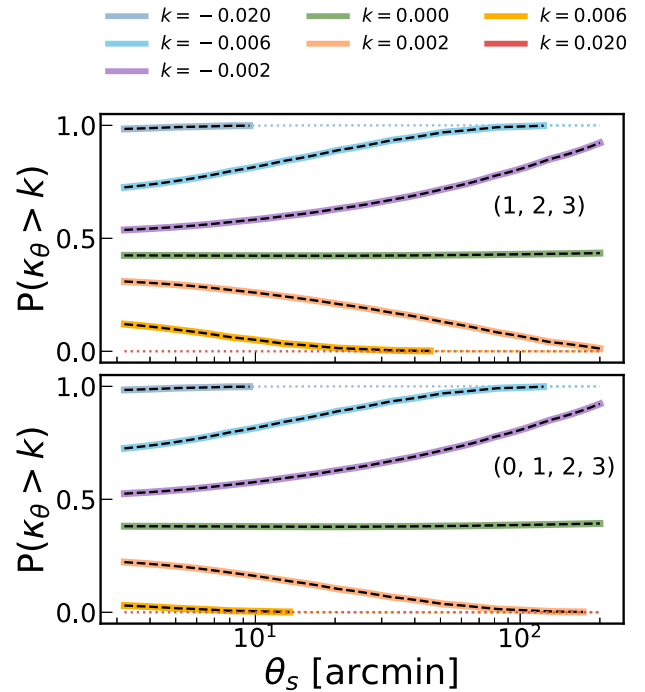
Zürcher D., Fluri J., Sgier R., Kacprzak T., Refregier A., 2021, *J. Cosmol. Astropart. Phys.*, 2021, 028

Zürcher D. et al., 2022, *MNRAS*, 511, 2075

## APPENDIX A: 3-FIELD CDFS AND BEYOND

Formally, in the Gaussian limit, the 3-field CDF contains no new information beyond those from the 2-field CDFs, since they can also be described completely by the multivariate normal in equation (8). Thus, the 3-field CDFs can be predicted exactly using the covariance of the fields as a function of smoothing scale.

We show this explicitly in Fig. A1. We make measurements of the 3-field and 4-field CDFs on Gaussian fields, and then exactly predict the measurements given the covariance matrix as a function of smoothing scale. The covariance matrix is measured directly on



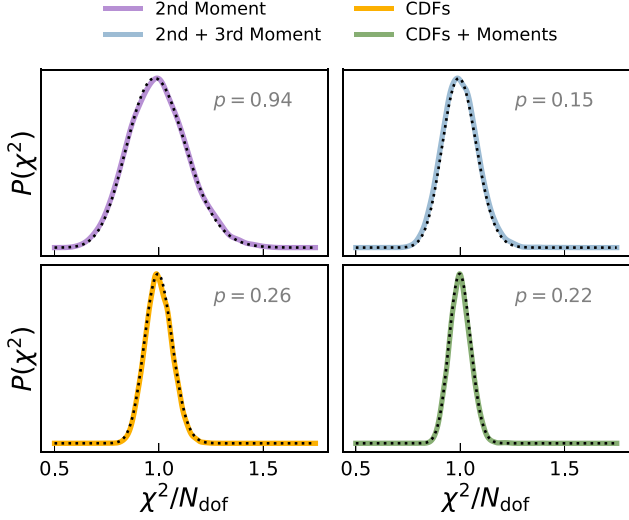
**Figure A1.** Measurements for the 3-field and 4-field CDFs on the noiseless DES Y3-like simulations (coloured lines), and a theoretical prediction in the limit of the field being Gaussian (black, dashed lines). The latter follows the same procedure of Section 2.3. The Gaussian model fits the data well, as is expected in this limit. The bin indices show the different tomographic bins used in the measurement.

the map. We have verified the residuals between the measured  $N$ -field CDFs and the prediction is within  $0.1\sigma$ , where  $\sigma$  comes solely from cosmic variance. This test is an extension of Fig. 2 for  $N$ -field CDFs of higher  $N$ .

## APPENDIX B: GAUSSIANTY OF COVARIANCE MATRIX

The process of performing a Fisher forecast, or obtaining constraints using likelihood minimization, assumes the likelihood of the data vector is Gaussian, i.e. the measurement uncertainty in the data vector is distributed as a multivariate Gaussian. We test here the validity of that assumption. We do so by first transforming every realization  $i$  of a data vector by removing its mean,  $S_i = D_i - \langle D \rangle$ , where the mean is computed over all  $i$  realizations. We then compute  $\chi^2 = S_i C^{-1} S_i$ , where  $C$  is the covariance matrix estimated using all realizations of  $D$ . In the limit that the likelihood is Gaussian, the distribution of  $\chi^2$  must follow a standard  $\chi^2$  distribution.

In Fig. B1, we show the measured and expected distributions for four different data vectors, and in all cases we find the measured distributions match the expected Gaussian-limit distributions. We also compute a Kolmogorov–Smirnov statistic to quantify the level of agreement between the measured and expected distribution (Peacock 1983). This validates that the Fisher formalism is an accurate way to estimate potential constraints from the statistics considered in this work. Some additional techniques can also be used to quantify this Gaussianity of the likelihood (Park et al. 2022; Euclid Collaboration 2023), and they are roughly similar to the approach we have taken here.



**Figure B1.** The chi-squared distributions of the data vectors (solid lines), compared with a theoretical chi-squared distribution (dotted black line) with  $N_{\text{dof}}$  given by the size of the data vector. In the Gaussian likelihood limit, the theoretical distributions will match the measured distribution. A Kolmogorov–Smirnov test shows the probability that the observed and expected distributions are similar exceeds  $p > 0.1$ . The data vectors considered in this work have a sufficiently Gaussian likelihood.

### APPENDIX C: DEPENDENCE OF DATA VECTOR ON COSMOLOGY

In Fig. C1, we show the derivative of the CDF measurement with the three cosmology parameters we have varied in Section 4.2. For brevity, we only show the derivative for the 1-field CDF of the fourth tomographic bin. At fixed threshold, the scale-dependence of the derivatives varies across the parameters, particularly at larger scales. At smaller scales, the derivatives with respect to  $\Omega_m$  and  $\sigma_8$  have larger amplitudes for the negative tail ( $k = -0.02$ ) than the positive tail ( $k = 0.02$ ). The derivative for  $k = 0$  (green line) is near-zero in the 1-field CDFs, but we have checked that it is significantly non-zero for 2-field CDFs; this difference between the 1-field and 2-field behaviour is similar to that seen in Fig. 3. Any change in the  $k = 0$  line for 1-field CDFs means the median of the distribution (and thus, the shape of the distribution) is being altered.

<sup>1</sup>Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, USA

<sup>2</sup>Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA

<sup>3</sup>Indian Institute of Science Education and Research, Pune 411008, India

<sup>4</sup>Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA

<sup>5</sup>Kavli Institute for Particle Astrophysics & Cosmology, P. O. Box 2450, Stanford University, Stanford, CA 94305, USA

<sup>6</sup>SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA

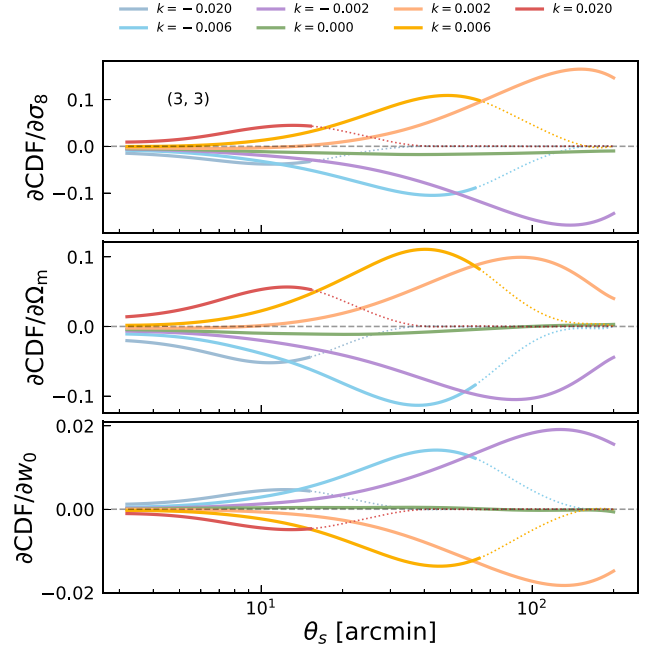
<sup>7</sup>Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA

<sup>8</sup>Department of Physics, ETH Zurich, Wolfgang-Pauli-Strasse 16, CH-8093 Zurich, Switzerland

<sup>9</sup>Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL 60439, USA

<sup>10</sup>Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK

<sup>11</sup>Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK



**Figure C1.** The derivative of the CDFs data vector with respect to  $\sigma_8$ ,  $\Omega_m$ , and  $w_0$ . For brevity, we only show results for the 1-field CDF of the highest redshift bin. The derivatives are shown as dotted lines when the CDFs fall into the 99.7 per cent tail. The dashed grey line shows the zero-derivative mark as a reference. The derivatives with different parameters have noticeable scale differences, leading to the degeneracy breaking in the parameter constraints. The negative tail ( $k = -0.02$ ) has a higher derivative amplitude for  $\Omega_m$  and  $\sigma_8$  compared to that of the positive tail ( $k = 0.02$ ).

<sup>12</sup>Institute for Astronomy, University of Hawai'i, 2680 Woodlawn Drive, Honolulu, HI 96822, USA

<sup>13</sup>Physics Department, 2320 Chamberlin Hall, University of Wisconsin-Madison, 1150 University Avenue, Madison, WI 53706-1390, USA

<sup>14</sup>Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15312, USA

<sup>15</sup>Instituto de Astrofísica de Canarias, E-38205 La Laguna, Tenerife, Spain

<sup>16</sup>Laboratório Interinstitucional de e-Astronomia – LIneA, Rua Gal. José Cristino 77, Rio de Janeiro, RJ-20921-400, Brazil

<sup>17</sup>Dpto. Astrofísica, Universidad de La Laguna, E-38206 La Laguna, Tenerife, Spain

<sup>18</sup>Center for Astrophysical Surveys, National Center for Supercomputing Applications, 1205 West Clark St., Urbana, IL 61801, USA

<sup>19</sup>Department of Astronomy, University of Illinois at Urbana-Champaign, 1002 W. Green Street, Urbana, IL 61801, USA

<sup>20</sup>Department of Physics, Duke University, Durham, NC 27708, USA

<sup>21</sup>NASA Goddard Space Flight Center, 8800 Greenbelt Rd, Greenbelt, MD 20771, USA

<sup>22</sup>Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

<sup>23</sup>Fermi National Accelerator Laboratory, P. O. Box 500, Batavia, IL 60510, USA

<sup>24</sup>NSF AI Planning Institute for Physics of the Future, Carnegie Mellon University, Pittsburgh, PA 15213, USA

<sup>25</sup>Université Grenoble Alpes, CNRS, LPSC-IN2P3, F-38000 Grenoble, France

<sup>26</sup>Department of Physics and Astronomy, University of Waterloo, 200 University Ave W, Waterloo, ON N2L 3G1, Canada

<sup>27</sup>Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr, Pasadena, CA 91109, USA

<sup>28</sup>University Observatory, Faculty of Physics, Ludwig-Maximilians-Universität, Scheinerstr. 1, 81679 Munich, Germany



- <sup>29</sup>*School of Physics and Astronomy, Cardiff University, Cardiff CF24 3AA, UK*
- <sup>30</sup>*Department of Astronomy, University of Geneva, ch. d'Écogia 16, CH-1290 Versoix, Switzerland*
- <sup>31</sup>*Department of Physics & Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
- <sup>32</sup>*Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK*
- <sup>33</sup>*Instituto de Física Gleb Wataghin, Universidade Estadual de Campinas, 13083-859 Campinas, SP, Brazil*
- <sup>34</sup>*Department of Physics, University of Genova and INFN, Via Dodecaneso 33, I-16146 Genova, Italy*
- <sup>35</sup>*Jodrell Bank Center for Astrophysics, School of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK*
- <sup>36</sup>*Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), E-28040 Madrid, Spain*
- <sup>37</sup>*Brookhaven National Laboratory, Bldg 510, Upton, NY 11973, USA*
- <sup>38</sup>*Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY 11794, USA*
- <sup>39</sup>*Département de Physique Théorique and Center for Astroparticle Physics, Université de Genève, 24 quai Ernest Ansermet, CH-1211 Geneva, Switzerland*
- <sup>40</sup>*Institut d'Estudis Espacials de Catalunya (IEEC), E-08034 Barcelona, Spain*
- <sup>41</sup>*Institut de Recherche en Astrophysique et Planétologie (IRAP), Université de Toulouse, CNRS, UPS, CNES, 14 Av. Edouard Belin, F-31400 Toulouse, France*
- <sup>42</sup>*Cerro Tololo Inter-American Observatory, NSF's National Optical-Infrared Astronomy Research Laboratory, Casilla 603, La Serena, Chile*
- <sup>43</sup>*Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA*
- <sup>44</sup>*Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA*
- <sup>45</sup>*Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 3FX, UK*
- <sup>46</sup>*Department of Physics, Northeastern University, Boston, MA 02115, USA*

- <sup>47</sup>*Physics Department, William Jewell College, Liberty, MO 64068, USA*
- <sup>48</sup>*Hamburger Sternwarte, Universität Hamburg, Gojenbergsweg 112, D-21029 Hamburg, Germany*
- <sup>49</sup>*School of Mathematics and Physics, University of Queensland, Brisbane, QLD 4072, Australia*
- <sup>50</sup>*Department of Physics, IIT Hyderabad, Kandi, Telangana 502285, India*
- <sup>51</sup>*Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029 Blindern, NO-0315 Oslo, Norway*
- <sup>52</sup>*Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, E-08193 Bellaterra (Barcelona), Spain*
- <sup>53</sup>*Santa Cruz Institute for Particle Physics, Santa Cruz, CA 95064, USA*
- <sup>54</sup>*Center for Cosmology and Astro-Particle Physics, The Ohio State University, Columbus, OH 43210, USA*
- <sup>55</sup>*Department of Physics, The Ohio State University, Columbus, OH 43210, USA*
- <sup>56</sup>*Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA*
- <sup>57</sup>*Australian Astronomical Optics, Macquarie University, North Ryde, NSW 2113, Australia*
- <sup>58</sup>*Lowell Observatory, 1400 Mars Hill Rd, Flagstaff, AZ 86001, USA*
- <sup>59</sup>*George P. and Cynthia Woods Mitchell Institute for Fundamental Physics and Astronomy, and Department of Physics and Astronomy, Texas A&M University, College Station, TX 77843, USA*
- <sup>60</sup>*Institució Catalana de Recerca i Estudis Avançats, E-08010 Barcelona, Spain*
- <sup>61</sup>*Observatório Nacional, Rua Gal. José Cristino 77, Rio de Janeiro, RJ-20921-400, Brazil*
- <sup>62</sup>*School of Physics and Astronomy, University of Southampton, Southampton SO17 1BJ, UK*
- <sup>63</sup>*National Center for Supercomputing Applications, 1205 West Clark St., Urbana, IL 61801, USA*

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.