Dark energy survey internal consistency tests of the joint cosmological probes analysis with posterior predictive distributions

C. Doux, E. Baxter, P. Lemos, C. Chang, A. Alarcon, A. Amon, A. Campos, A. Choi, M. Gatti, D. Gruen, et al.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **503,** 2688–2705 (2021) doi:10.1093/mnras/stab526
Advance Access publication 2021 February 24

Dark energy survey internal consistency tests of the joint cosmological probes analysis with posterior predictive distributions

C. Doux [1★] E. Baxter,[2] P. Lemos [3] C. Chang [4,5] A. Alarcon,[6] A. Amon,[7] A. Campos,[8] A. Choi,[9] M. Gatti,[10] D. Gruen,[7,11,12] M. Jarvis,[1] N. MacCrann,[13] Y. Park,[14] J. Prat,[4] M. M. Rau,[8] M. Raveri,[5] S. Samuroff,[8] J. DeRose,[15,16] W. G. Hartley,[17] B. Hoyle,[18,19,20] M. A. Troxel,[21] J. Zuntz,[22] T. M. C. Abbott,[23] M. Aguena,[24,25] S. Allam,[26] J. Annis,[26] S. Avila,[27] D. Bacon,[28] E. Bertin,[29,30] S. Bhargava,[31] D. Brooks,[3] D. L. Burke,[7,12] M. Carrasco Kind,[32,33] J. Carretero,[10] R. Cawthon,[34] M. Costanzi,[35,36] L. N. da Costa,[25,37] M. E. S. Pereira,[38] S. Desai,[39] H. T. Diehl,[26] J. P. Dietrich,[18] P. Doel,[3] S. Everett,[16] I. Ferrero,[40] P. Fosalba,[41,42] J. Frieman,[6,25] J. García-Bellido,[27] D. W. Gerdes,[38,43] T. Giannantonio,[13,44] R. A. Gruendl,[32,33] J. Gschwend,[25,37] G. Gutierrez,[26] S. R. Hinton,[45] D. L. Hollowood,[16] K. Honscheid,[9,46] E. M. Huff,[47] D. Huterer,[38] B. Jain,[1] D. J. James,[48] E. Krause,[49] K. Kuehn,[50,51] N. Kuropatkin,[26] O. Lahav,[3] C. Lidman,[52,53] M. Lima,[24,25] M. A. G. Maia,[25,37] F. Menanteau,[32,33] R. Miquel,[10,54] R. Morgan,[34] J. Muir,[7] R. L. C. Ogando,[25,37] A. Palmese,[5,26] F. Paz-Chinchón,[13,33] A. A. Plazas,[55] E. Sanchez,[56] V. Scarpine,[26] M. Schubnell,[38] S. Serrano,[41,42] I. Sevilla-Noarbe,[56] M. Smith,[57] E. Suchyta,[58] M. E. C. Swanson,[33] G. Tarle,[38] C. To,[7,11,12] D. L. Tucker,[26] T. N. Varga,[19,20] J. Weller,[19,20] and R. D. Wilkinson[31] (DES Collaboration)

Affiliations are listed at the end of the paper

Accepted 2021 February 18. Received 2021 February 15; in original form 2020 October 23

ABSTRACT

Beyond ΛCDM, physics or systematic errors may cause subsets of a cosmological data set to appear inconsistent when analysed assuming ΛCDM. We present an application of internal consistency tests to measurements from the Dark Energy Survey Year 1 (DES Y1) joint probes analysis. Our analysis relies on computing the posterior predictive distribution (PPD) for these data under the assumption of ΛCDM. We find that the DES Y1 data have an acceptable goodness of fit to ΛCDM, with a probability of finding a worse fit by random chance of $p = 0.046$. Using numerical PPD tests, supplemented by graphical checks, we show that most of the data vector appears completely consistent with expectations, although we observe a small tension between large- and small-scale measurements. A small part (roughly 1.5 per cent) of the data vector shows an unusually large departure from expectations; excluding this part of the data has negligible impact on cosmological constraints, but does significantly improve the p-value to 0.10. The methodology developed here will be applied to test the consistency of DES Year 3 joint probes data sets.

Key words: gravitational lensing: weak – methods: statistical – dark energy – large-scale structure of Universe.

1 INTRODUCTION

Several recent cosmological measurements appear to be in mild to severe tension in the context of the standard cosmological constant and cold dark matter (ΛCDM) model. For instance, the value of H_0 inferred from the cosmic microwave background (CMB; Planck Collaboration VI 2020) and from the cosmic distance ladder (Riess et al. 2019) are discrepant at roughly the 5σ level (e.g. Bernal, Verde & Riess 2016; Feeney, Mortlock & Dalmasso 2018; Aylor et al. 2019). Similarly, the value of σ_8 inferred from the CMB and from large-scale structure, in particular weak lensing measurements, are discrepant at roughly the 3σ level (e.g Battye, Charnock & Moss 2015; MacCrann et al. 2015; Raveri 2016; Hildebrandt et al. 2017; DES Collaboration 2018; Raveri & Hu 2019; Asgari et al. 2020; Joudaki et al. 2020; Park & Rozo 2020; Heymans et al. 2021). These tensions could be indicative either of a breakdown in the standard cosmological model, or of systematics impacting various analyses. Given these possibilities, identifying and quantifying cosmological tensions is of prime importance. We make a somewhat artificial distinction between *external* tensions – those between different experiments – and *internal* tensions – those between the measurements of a single experiment. In practice, correlations between the data are common in the case of internal tensions, but are rarer for external tensions.

In this work, we explore internal tensions in the Dark Energy Survey (DES; The Dark Energy Survey Collaboration 2005) Year 1 (Y1) measurements of two-point functions of large-scale structure

★ E-mail: cdoux@sas.upenn.edu

(DES Collaboration 2018). The DES is a 6-yr optical imaging survey of 5000 square degrees of the southern sky. The analysis of DES Collaboration (2018) derived measurements of galaxy positions and galaxy shear from first year observations of DES and used these to measure three two-point functions: the autocorrelations of galaxy shear and of galaxy positions, and the cross-correlation of galaxy position with galaxy shear. Cosmological constraints were then obtained by fitting this so-called 3 x 2 pt combination of correlation functions.

There are a few reasons why tests of internal consistency of the DES data are important and timely. First, as mentioned previously, measurements of σ_8 from surveys of large-scale structure tend to be lower than the value inferred from primary anisotropies in the CMB. If this tension is due to a true breakdown in ΛCDM – such as a departure from the expected growth of structure – then DES data alone might be expected to be internally inconsistent assuming ΛCDM. Secondly, all weak lensing surveys necessarily suffer from systematic errors (see e.g. Chang et al. 2019), thus introducing uncertainties which must be accounted for. Systematic errors in the data, such as unaccounted photometric redshift biases, are likely to manifest as internal inconsistency. Finally, one of the aims of this analysis is to develop the methodology and specific data tests that will be applied to the forthcoming analysis of Year 3 (Y3) data from DES.

We address the question of whether the DES Y1 3 x 2 pt measurements are self-consistent using posterior predictive methods. The posterior predictive distribution (PPD; see e.g. Gelman et al. 2004 for a review) is the distribution of possible new data, conditioned on observed data, given an underlying model. By comparing the PPD to the observed data, we can assess the degree to which the observed data are internally consistent in the context of ΛCDM. Several recent works have adopted the PPD as a means to examine consistency of cosmological data sets (e.g. DES Collaboration 2019b; Feeney et al. 2019). Köhlinger et al. (2019) proposed a test of internal consistency based on three different tests, that occur at various levels of the analysis: a global test, based on ratios of Bayesian evidences (Marshall, Rajguru & Slosar 2006), a parameter difference test that occurs in parameter space, and a PPD test on data space. Later, Handley & Lemos (2019) showed that the test based on the evidence ratio is proportional to the prior volume, and substituted it with a test based on the *suspiciousness* statistic (extended to correlated data sets in Lemos et al. 2020). Here, we focus on identifying potential subsets of the DES Y1 data in tension with each other, for which the PPD tests are particularly well-suited since they operate entirely in data space.

Our approach is to split the DES Y1 3 x 2 pt data into subsets motivated by considerations of possible systematics, as well as by considerations of possible extensions to ΛCDM. We first evaluate the goodness of fit of these subsets of data to ΛCDMusing the standard PPD formalism. Next, we use the PPD to perform *consistency* tests where we evaluate the goodness of fit of some subset of the data *conditioned* on the observed data from another, disjoint subset. For instance, we consider the likelihood of the measured two-point functions at large scales, conditioned on their observed values at small scales. This test in effect determines whether the large and small-scale measurements are consistent.

The paper is organized as follows. In Section 2, we describe the DES Y1 3 x 2 pt measurements; in Section 3, we give an overview of the PPD framework and application to DES Y1 data; in Section 4, we present the results of the application of this framework to the DES Y1 measurements; in Section 5, we lay out our plan for the upcoming DES Y3 analysis; we conclude in Section 6.

2 THE DES Y1 3 X 2 PT MEASUREMENTS

In this section, we briefly describe the 3 x 2 pt data vector from the DES Y1 analysis; more details can be found in DES Collaboration (2018). From the DES imaging data, galaxy positions and shears are measured. The galaxy samples are divided into two samples: 'lenses' and 'sources.' The lens galaxies are selected using the `redMaGiC` algorithm (Rozo et al. 2016), and have tightly constrained redshifts. The source galaxy sample, which extends to higher redshift than the lenses, have shapes measured using `METACALIBRATION` (Sheldon & Huff 2017) and `Im3SHAPE`, as described in Zuntz et al. (2018). Using the galaxy position measurements and the galaxy shear measurements (for the sources only), three two-point correlations functions were computed: shear–shear, position–shear, and position–position. Each two-point correlation was computed as a function of angular separation, θ. Since *cosmic shear* is a spin-2 field, the shear–shear correlation was divided into two components, $\xi_+(\theta)$, and $\xi_-(\theta)$. For notational convenience, we refer to the position–position correlation, or *clustering*, as $w(\theta)$, and to the galaxy–shear correlation, also referred to as *galaxy–galaxy lensing*, as $\gamma_t(\theta)$. The lens and source samples were divided into five and four tomographic redshift bins, respectively, and the auto and cross-correlations between the bins were measured. While all correlation functions were computed at 20 fixed (logarithmic) angular bins between 2.5 and 250 arcmin, measurements at small scales were not included in the analysis due to the presence of effects that could make the DES Y1 model an inaccurate description of the data in the small-scale regime (such as non-linear galaxy bias, baryonic effects on the matter power spectrum, etc). Details of the measurements of these correlation functions can be found in Elvin-Poole et al. (2018), Troxel et al. (2018), and Prat et al. (2018). The full 3 x 2 pt data vector includes all two-point measurements, in all redshift bin combinations, across a range of angular scales.

3 POSTERIOR PREDICTIVE DISTRIBUTION

In this section, we present an overview of the PPD formalism (Section 3.1), discuss the choice of test statistic (Section 3.2), its application to DES Y1 data for *goodness-of-fit* and *consistency* tests (Section 3.3) and our sampling strategy (Section 3.4). In particular, we identify a potential caveat associated with the standard choice of χ^2 statistic when testing the consistency of two experiments whose posteriors have little overlap. We illustrate this problem with a toy model in Section 3.2 and present a solution applicable to DES Y1 data in Section 3.3.

3.1 Overview

The posterior predictive distribution (PPD) is the distribution of possible (unobserved) data realizations from an experiment, given the posterior on parameters, Θ, of the model, M, obtained from the observed data. For observed data, $d_{\rm obs}$, the model posterior is $P(\Theta|d_{\rm obs}, I)$, where I represents all prior information, such as the form of the likelihood and priors. The PPD function for unobserved data $d_{\rm rep}$ is then[1]

$$P(d_{\rm rep}|d_{\rm obs}, I) = \int {\rm d}\Theta P(d_{\rm rep}|d_{\rm obs}, \Theta, I) P(\Theta|d_{\rm obs}, I). \quad (1)$$

[1]We follow the notation of Gelman et al. (2004) in referring to the *replicated*, unobserved PPD repetitions of the data as $d_{\rm rep}$.

In the case that $d_{\rm rep}$ and $d_{\rm obs}$ are conditionally independent given Θ, then $P(d_{\rm rep}|d_{\rm obs}, \Theta, I) = P(d_{\rm rep}|\Theta, I)$, i.e. the data likelihood. We will consider both this case and the case where $d_{\rm rep}$ and $d_{\rm obs}$ are not conditionally independent below. Hereafter, we drop I for conciseness.

The analytic computation of the the integral in equation (1) is cumbersome. Furthermore, the quantity on the left is a probability density, and therefore it requires to be normalized to provide a meaningful statement about the statistical significance of the internal tensions in a data set. One straightforward possibility is to take the ratio $R = P(d_{\rm rep}|d_{\rm obs})/P(d_{\rm rep})$, where we can identify the denominator as the Bayesian evidence for $d_{\rm rep}$. However, as pointed out by Lemos et al. (2020), this is equivalent to the Bayes' ratio (Marshall et al. 2006), which is not suitable for the case of wide and uninformative priors. In practice, rather than compute the integral in equation (1), one typically draws realizations of $d_{\rm rep}$ from $P(d_{\rm rep}|d_{\rm obs}, \Theta)$ at each point Θ in a Markov chain sampling the posterior $P(\Theta|d_{\rm obs}, I)$. These realizations of $d_{\rm rep}$, i.e. samples from the PPD, can then be compared directly to the observed data $d_{\rm obs}$ to look for signs of discrepancy.

Below, we will explore graphical and numerical approaches to comparing $d_{\rm rep}$ and $d_{\rm obs}$. Indeed, a powerful application of the PPD framework is to simply plot the observed PPD realization on top of the actual data, as done in Fig. 2 for instance. If the true data look significantly different from the PPD realizations, that could be a sign that the data are inconsistent with the assumed model. However, this approach does not allow one to quantify consistency and is mainly used here to provide insight into useful splits of the full data vector. A common way to quantify the level of tension between $d_{\rm rep}$ and $d_{\rm obs}$ is to use a test statistic, $T(d, \Theta)$, that can be computed for both $d_{\rm rep}$ and $d_{\rm obs}$, and which may be a function of the parameters Θ. A p-value can then be associated with the comparison between $d_{\rm rep}$ and $d_{\rm obs}$ via

$$p = P(T(d_{\rm rep}, \Theta) > T(d_{\rm obs}, \Theta)|d_{\rm obs}). \quad (2)$$

In other words, p is the probability of getting a higher test statistic for PPD realizations than $T(d_{\rm obs}, \Theta)$ by random chance. A very low p-value (say less that 0.01) would then be indicative that $d_{\rm obs}$ was unlikely, while a high p-value (say greater than 0.99) could be indicative of a problem in the model, such as an overestimate of the noise covariance. Following DES Collaboration (2018), we will adopt a p-value threshold of $p = 0.01$: if $p > 0.01$, we will consider the data to be in reasonable consistency. In practice, the p-value can be computed easily from a Markov chain sampling the posterior $P(\Theta|d_{\rm obs})$. At each set of parameters Θ_i in the chain, one draws a realization $d_{{\rm rep},\,i}$ of $d_{\rm rep}$ using the (possibly conditional) likelihood $P(d_{\rm rep}|d_{\rm obs}, \Theta_i)$. Values of $T(d_{\rm obs}, \Theta_i)$ and $T(d_{{\rm rep},\,i}, \Theta_i)$ are then computed. The p-value is simply the fraction of samples Θ_i for which $T(d_{{\rm rep},\,i}, \Theta_i) > T(d_{\rm obs}, \Theta_i)$. One of the challenges of a PPD analysis is selecting an appropriate test statistic, $T(d, \Theta)$, as seen in the next section.

Compared to other tension metrics, the PPD has several advantages. First, unlike evidence ratio-based tests (e.g. Marshall et al. 2006), the PPD does not require specifying an alternative model. This is often desirable, since the choice of alternate model is not always clear, especially in the case of cosmological analyses. Secondly, the comparison between data and PPD realizations occurs entirely in data space rather than in parameter space. This means that the PPD is particularly well suited to identifying particular parts of the data that may be unusually discrepant with the model. Finally, if the posterior is likelihood-dominated (as opposed to prior-dominated), the PPD realizations will not depend on the prior volume. This is not the case for e.g. evidence ratios, for which the choice of prior outside the likelihood-dominated region is important.

3.2 Choice of test statistic for computing p-value

As discussed above, assigning a p-value to the results of a PPD test requires a choice of test statistic. For high-dimensional data – in particular Gaussian data – a common choice is χ^2, i.e.

$$T(d, \Theta) = (d - \mu(\Theta))^{\intercal}\mathbf{C}^{-1}(d - \mu(\Theta)), \quad (3)$$

where $\mu(\Theta)$ is the model evaluated at parameter values Θ and $\mathbf{C}$ is the covariance matrix. The use of χ^2 as the test statistic, however, can bias the p-value low when testing the consistency of two experiments – which can be disjoint subsets of a data vector – that constrain very different volumes of the parameter space. In this case, the replicas of one experiment conditioned on the other can naturally yield p-values that are not uniformly distributed in [0,1] over (consistent) data realizations, but are skewed towards lower values. Thus, these should be interpreted with caution when using them for consistency tests, as we illustrate with a toy model in the following subsection (see also Gelman 2013, for a discussion about non-uniform posterior predictive p-values). In Section 3.3, we show how PPD tests can be repeated on simulated DES Y1 data to calibrate p-values and distinguish this effect from true tensions. We will therefore rely on calibrated p-values, denoted $\tilde{p}$, to test internal consistency.

3.2.1 A toy model example

For purposes of illustration, we consider two independent experiments, A and B, that both make two-dimensional measurements, respectively $d_A = (d_1^A, d_2^A)$ and $d_B = (d_1^B, d_2^B)$. We now suppose both experiments to be normally distributed with means μ_A and μ_B, unit variance per component and covariances ρ_A and ρ_B between components, such that the covariance of d_A is

$$\mathbf{C}_A = \begin{pmatrix} 1 & \rho_A \\ \rho_A & 1 \end{pmatrix}, \quad (4)$$

and similarly for d_B. Problems with using χ^2 as a test statistic emerge when the parameter space is at least two-dimensional. We therefore consider a two-parameter model for the data, that is specified by parameters $\Theta = (\theta_1, \theta_2)$, with true values $\Theta_0 = (0, 0)$, and such that $\mu_A = \mu_B = \Theta$. In other words, the expectation values of the data points are the parameter values. Assuming flat priors over parameters, the likelihoods and posteriors are normal distributions we can easily sample from. Finally, we imagine having measurements for each experiment that coincide with their fiducial values, $d_A^{\rm obs} = d_B^{\rm obs} = (0, 0)$, such that they are perfectly consistent.

(i) *Goodness-of-fit test.* In the left-hand panels of Fig. 1, we perform a *goodness-of-fit* test for experiment A alone using the PPD $P(d_A^{\rm rep}|d_A^{\rm obs})$, which is the distribution of possible future *replicated* measurements of d_A, given $d_A^{\rm obs}$. We generate a sample of parameters Θ drawn from the posterior $P(\Theta|d_A^{\rm obs})$ (blue points in the upper left-hand panel) and, at each point Θ_i, we draw a sample from the PPD (green points) by sampling the likelihood $P(d_A^{\rm rep}|\Theta_i)$. Here, we assumed $\rho_A = 0.5$. Note that the PPD realizations include the uncertainty on the posterior as well as the uncertainty from the data likelihood; this is why the distribution of green points is broader than that of the blue points. In the lower left-hand panel, we show the histogram of the difference of test statistic values from this mock analysis, $T(d_A^{\rm obs}, \Theta_i) - T(d_A^{\rm rep}, \Theta_i)$, using the test statistic from equation (3) with covariance $\mathbf{C}_A$. Comparing the test statistics from

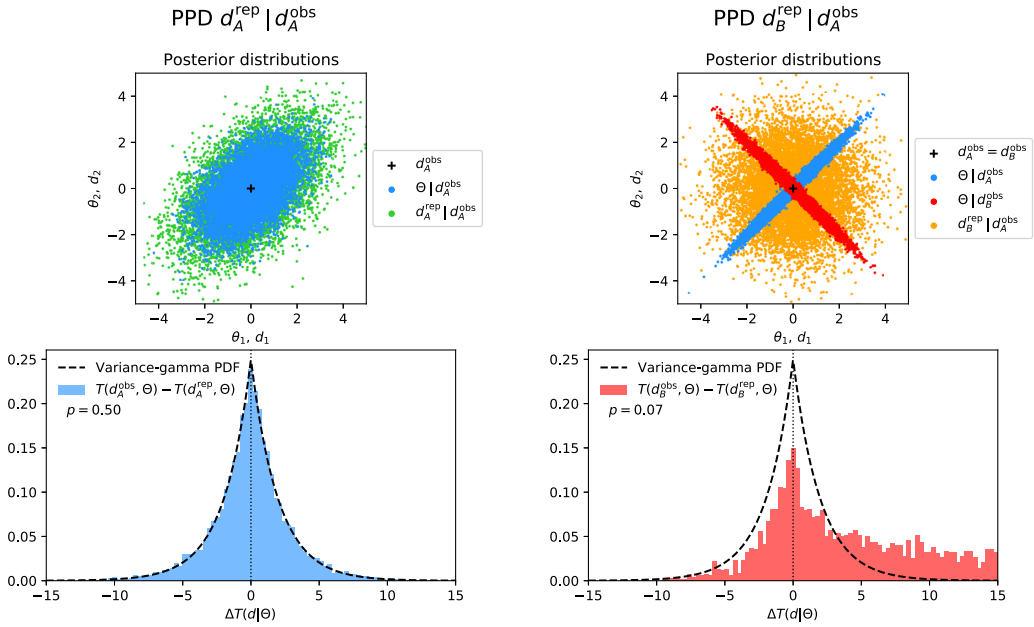

Figure 1. Illustration of the challenges of using χ^2 as a test statistic for calculating a p-value from the PPD. Top panels show draws from the parameter posteriors, while bottom panels show histograms of the difference of computed test statistics (with the probability density function of the difference of independent χ^2 statistics, given by the variance-gamma distribution developed by Madan & Seneta 1990). In the panels at left-hand, we consider PPD realizations $d_A^{\rm rep}$ from the same experiment used to generate the observed data, i.e. $d_A^{\rm obs}$. In the panels at right-hand, we consider PPD realizations of a different experiment, $d_B^{\rm rep}$, than the original data, $d_A^{\rm obs}$. In the latter case, we can make the p-value arbitrarily small by reducing the fractional overlap in the posteriors from A and B.

$d_A^{\rm obs}$ and from $d_A^{\rm rep}$ yields p-value of $p = 0.5$, which is reasonable, given that we know the data are consistent, and the PPD is working as expected.

(ii) *Consistency test.* In the right-hand panel of Fig. 1, we perform a *consistency* test between experiments A and B using the PPD $P(d_B^{\rm rep}|d_A^{\rm obs})$. Since experiments A and B are independent, the likelihood is $P(d_B|d_A, \Theta) = P(d_B|\Theta)$, i.e. the normal distribution with mean $\mu_B = \Theta$ and covariance $\mathbf{C}_B$. The problem we highlight here occurs when the overlap of the posteriors from experiments A and B is small. Therefore, we set $\rho_A = 0.99$ and $\rho_B = -0.99$, such that the posteriors of experiments A and B, shown respectively in blue and red in the upper right-hand panel, have little overlap around the true parameters. At each point in the posterior sample for $d_A^{\rm obs}$, we draw realizations from the data likelihood for experiment B in order to generate $d_B^{\rm rep}$ (orange points). This sample can be thought of as combining the posterior uncertainty from experiment A (i.e. the spread of the blue points), with the likelihood uncertainty from experiment B (i.e. the spread of the red points). We see now that, on average, the orange points are far from the red points (as measured with χ^2 using $\mathbf{C}_B$), since the parameter values allowed by the posterior of experiment A are typically far from the posterior of experiment B. The histogram of the difference of test statistics, $T(d_B^{\rm obs}, \Theta_i) - T(d_B^{\rm rep}, \Theta_i)$, is asymmetric with a tail at higher values, as shown in the lower right-hand panel. As a consequence, the p-value for this comparison will be low; we find $p \approx 0.07$. We can make the p-value even lower by increasing ρ_A and decreasing ρ_B.

3.2.2 *Conclusions from toy model*

We have therefore shown how two experiments that we know are consistent (they are generated from the same true parameter values) can be made to have an arbitrarily low p-value if their posteriors do not overlap significantly in parameter space. We emphasize that this difficulty emerges because of the choice of test statistic, not because of some fundamental shortcoming of the PPD. The PPD realizations in orange in the right-hand panel of Fig. 1 are true possible realizations of future measurements of experiment B, given our knowledge from experiment A. The reason why p-values appear to be biased low in the second kind of test is that $T(d_B^{\rm rep}, \Theta_i)$ is likely to be smaller than $T(d_B^{\rm obs}, \Theta_i)$, since $d_B^{\rm rep}$ *was generated assuming the true parameter values are* $\Theta_{\rm i}$. In principle, it may be possible to design some test statistic that does not suffer from this complication; we leave this to future work.

In practice, however, the fact that *consistency* tests between two experiments A and B ($d_B^{\rm rep}|d_A$) – as opposed to the *goodness-of-fit* tests of a single experiment A ($d_A^{\rm rep}|d_A$) – tend to bias the p-values low – i.e. skew the distribution of p-values over data realizations towards low values – means that the PPD tests measure degrees of consistency conservatively. This can be seen as an advantage since we want to be careful about claiming internal consistency. Moreover, most tests we will perform in the following sections do not suffer from this problem as much as our toy model because posteriors are generally not as discrepant. We will, however, consider a method to calibrate p-values in the next section in order to eliminate confusion between this effect and real tensions, and rely on calibrated $\tilde{p}$-values for all tests, in order to facilitate the interpretation of our results.

3.3 Considerations for application to DES Y1 data

3.3.1 *DES Y1 likelihood and PPD tests*

The DES 3 x 2 pt analysis, described in Krause et al. (2017), adopts a Gaussian likelihood for the 3 x 2 pt data vector (see e.g. Sellentin &

Heavens 2018; Louca & Sellentin 2020, for a discussion of this approximation). The Gaussian likelihood, $\mathcal{L}$, is determined by the expectation value of the data vector at parameters Θ, $\mu(\Theta)$, and by a covariance matrix $\mathbf{C}$

$$\mathcal{L}(d_{\rm obs}|\Theta) = \mathcal{N}(\mu(\Theta), \mathbf{C}) \tag{5}$$

$$\propto \exp\left[-\frac{1}{2}(d_{\rm obs} - \mu(\Theta))^{\intercal}\mathbf{C}^{-1}(d_{\rm obs} - \mu(\Theta))\right], \tag{6}$$

where $\mathcal{N}(\mu, \mathbf{C})$ the probability distribution function of a multivariate Gaussian variable. We now distinguish two types of tests, depending on whether $d_{\rm rep}$ and $d_{\rm obs}$ refer to the same subset of the data vector (goodness-of-fit test) or different, disjoint subsets (consistency test).

(i) *Goodness-of-fit test.* In this case, $d_{\rm rep}$ is considered to be future, independent realizations of the same observable as $d_{\rm obs}$, in which case, $d_{\rm rep}$ does not depend on $d_{\rm obs}$. Therefore, $P(d_{\rm rep}|d_{\rm obs}, \Theta) = \mathcal{L}(d_{\rm rep}|\Theta)$ and the PPD p-value in equation (2) can be thought of as a Bayesian goodness-of-fit test, analogous to the classical χ^2 goodness-of-fit test, but including uncertainty over model parameters.

(ii) *Consistency test.* In this case, $d_{\rm rep}$ and $d_{\rm obs}$ correspond to disjoint, but correlated subsets of the full 3 x 2 pt data vector. For instance, $d_{\rm obs}$ can consist of cosmic shear and clustering measurements while $d_{\rm rep}$ can refer to galaxy–galaxy lensing observations, i.e. $d_{\rm obs} = \xi_{\pm}(\theta)$, $w(\theta)$, and $d_{\rm rep} = \gamma_{\rm t}(\theta)$. In this example, the conditional likelihood $P(d_{\rm rep}|d_{\rm obs}, \Theta)$ is the distribution of possible realizations of $\gamma_{\rm t}(\theta)$ data, given that the measurements of $\xi_{\pm}(\theta)$ and $w(\theta)$ are known and that the model parameters are known to be Θ. We will, however, consider other ways of splitting the data vector. Given our assumption of a multivariate Gaussian likelihood for the data, the distribution $P(d_{\rm rep}|\Theta, d_{\rm obs})$ is also a multivariate Gaussian, the mean, and covariance of which can be computed as follows. For a partition of the full data vector into disjoint subsets d_1 and d_2, we have

$$\begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}\right), \tag{7}$$

where μ_i is the mean of d_i, and $\mathbf{C}_{ij}$ represents the covariance matrix between d_i and d_j. The distribution of d_2 conditioned on d_1 is then also a multivariate Gaussian given by

$$P(d_2|d_1) = \mathcal{N}\big(\mu_2 + \mathbf{C}_{21}^{-1}\mathbf{C}_{11}^{-1}(d_1 - \mu_1),\ \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\big). \tag{8}$$

We will use this expression when computing the PPD for disjoint subsets of the full 3 x 2 pt data vector.

3.3.2 Calibration of PPD tests applied to DES Y1

Given the potential for p-values to be biased low for consistency tests, where $d_{\rm obs}$ and $d_{\rm rep}$ represent different observables (or other splits of the full data vector), we wish to be able to tell whether an observed p-value is biased low due to the effect described in Section 3.2 or if it is rather indicative of an actual tension in the data. To do so, we will attempt to calibrate p-values by sampling the distribution of such p-values for simulated data vectors which we know to be generated consistently. Given a fiducial cosmology (taken here as the best-fit of the full 3 x 2 pt analysis), we generate noisy data vectors by sampling the Gaussian likelihood and measure the p-values for the same PPD tests as those applied to real data. The comparison of the distribution of p-values for simulated data vectors to the p-value for observed data will provide a *calibrated* $\tilde{p}$-value, which is the fraction of simulated data vectors yielding a lower p-value than the observed data.

It would be prohibitively expensive in computing time to run a Markov chain for each simulated (noisy) data vector. Therefore, for each PPD test, we instead run a single chain on the fiducial (noiseless) data vector and use importance sampling to reweight samples in the chain for each simulated data vector, before recomputing PPD test statistics. The successive steps are the following:

(i) We run a standard Markov chain for the fiducial data vector, $d_{\rm fid}$, to generate a sample of parameters $\Theta_i \sim P(\Theta|d_{\rm fid})$ with weights w_i (see Section 3.4).

(ii) For each simulated data vector $d_{{\rm sim},\,j}$, we repeat the following steps[2]:

(a) We compute the importance weights, given by the ratios of the posteriors for simulated and fiducial data vectors, $\alpha_{ij} = P(d_{{\rm sim},\,j}|\Theta_i)/P(d_{\rm fid}|\Theta_i)$, and multiply them by the original weights w_i to get updated weights $w_{ij} = w_i\alpha_{ij}$;

(b) We compute the PPD test statistics $T(d_{{\rm sim},\,j}, \Theta_i)$;

(c) We draw samples from the PPD by generating a realization $d_{{\rm rep},\,i}$ at each parameter sample Θ_i (conditioned on $d_{{\rm sim},\,j}$ if calibrating a consistency test), and compute test statistics $T(d_{{\rm rep},\,i}|\Theta_i)$;

(d) We compute the p-value, p_j, given statistics $T(d_{{\rm sim},\,j}, \Theta_i)$ and $T(d_{{\rm rep},\,i}|\Theta_i)$ with weights w_{ij} using equation (2).

(iii) We use the distribution of p_j obtained from $N_{\rm sim} = 10^5$ independent simulated data vectors to calculate the calibrated $\tilde{p}$-value for a given test with p-value p, such that

$$\tilde{p} = \frac{1}{N_{\rm sim}}\sum_j \mathbb{1}(p - p_j), \tag{9}$$

where $\mathbb{1}$ is the Heaviside function.

In practice, we find that, given our sampling strategy (see Section 3.4), the importance sampling procedure results in relatively low errors on the estimated (raw) p-values, based on effective number of samples (of order few hundreds), thus validating the method. We will therefore report calibrated $\tilde{p}$-values for each test and replace our consistency criterion by $\tilde{p} > 0.01$.

Finally, we mention that we will consider many different tests of the data below. By performing multiple tests, we are necessarily more likely to obtain evidence for tension by random chance. One method to correct for this effect is the Bonferroni correction (Dunn 1959), which scales down the p-value threshold by a factor equal to the number of hypothesis tests. However, this correction can be overly severe, particularly when the data are correlated (as is the case here). Since the number of goodness-of-fit tests that we apply is fairly small (essentially only four), we will generally ignore corrections to our p-value threshold for multiple hypothesis tests. In some sense, this is a conservative approach since it makes us more likely to be worried about possible tensions. While we report many p-values for individual redshift bins (and more) below, we will not consider the data to be in internal tension if a few of these p-values fall below our p-value threshold. Instead, we will take the approach of considering subsets of the data with low p-values to warrant more exploration. In other words, we are not worried about the possibility of type-I errors, where a true null hypothesis is rejected, since we mainly use these p-values as a means to investigate agreement between the model and data.

[2]Note that, since the likelihood is Gaussian, these steps amount to recomputing various differences of log-likelihoods, which can be easily parallelized and performed within a few minutes for a sample of 10^5 simulated data vectors.

3.4 Sampling methodology

As described above, we generate PPD realizations from a posterior by drawing from the (possibly conditional) likelihood at each point in a posterior chain. The posterior chains are generated using `PolyChord` (Handley, Hobson & Lasenby 2015a,b). `PolyChord` uses Nested Sampling to calculate both the Bayesian evidence and the posterior distribution. It overcomes some of the issues of ellipsoidal nested sampling by using slice sampling (Neal 2000; Aitken 2013). This makes it the optimal sampler for the DES likelihood, as discussed in DES Collaboration (in preparation).

As described in DES Collaboration (2018), we sample over the standard ΛCDM parameter space, combined with the parameters describing systematic errors in the DES data. The DES-specific parameters describe multiplicative shear biases in the shear measurements, redshift biases in the assumed source and lens galaxy redshift distributions, and linear galaxy bias parameters. We additionally allow the neutrino mass to vary. Details of the assumed priors can be found in DES Collaboration (2018).

Once the parameter chain for a given posterior is generated, we re-process the chain using `CosmoSIS` (Zuntz et al. 2015) to generate the realizations from the likelihood at each step in the chain. Since `PolyChord` generates weighted samples of the posterior, these weights are also applied to all PPD realizations (and multiplied by importance weights for calibration purposes).

4 APPLICATION OF POSTERIOR PREDICTIVE CHECKS TO DES Y1 DATA

We now apply the PPD formalism developed above to various splits of the DES Y1 measurements. We consider several splits, motivated by considerations of potential systematic errors and possible beyond ΛCDM physics.

For visual clarity, when plotting the data we subtract the best-fitting 3 x 2 pt model from both the data and the PPD realizations, and normalize relative to the diagonals of the covariance matrix. In other words, we plot

$$\delta X_i = \frac{d_i - \mu_i^{\rm MAP}}{\sqrt{\mathbf{C}_{ii}}}, \tag{10}$$

where d_i is either the true data or the PPD realizations of the data, $\mu_i^{\rm MAP}$ is the best fit to the *full* 3 x 2 pt data vector (with fiducial scale cuts), and $\mathbf{C}$ is the covariance matrix of the 3 x 2 pt data vector. The choice to plot δX_i rather than d_i has no impact on the comparison between data and realizations and makes the plots easier to visualize.

Table 1 summarizes $\tilde{p}$-values for the full data vector as well as each individual probe for each test considered in the following Sections 4.1–4.6.

4.1 Goodness of fit of the full 3 x 2 pt data vector

We first consider the PPD for the full 3 x 2 pt data vector, shown in Fig. 2. This first test is useful to determine whether the model favoured by the data is actually a good fit to the data. However, unlike the classical χ^2 test which only uses the best-fitting model, the PPD goodness-of-fit test marginalizes over model parameter uncertainties.

The different insets in Fig. 2 split the datavector into the different observables ($\xi_\pm$, $\gamma_{\rm t}$, and $w(\theta)$), while the different panels split the correlation functions by redshift bin combination. The distribution of PPD realizations is shown as the blue bands, while the actual data is shown as the red points. The faded out points represent measurements that were not included because of angular scale cuts.

As discussed in Section 3, the computation of a p-value using the test statistic of equation (3) is motivated for goodness-of-fit tests like that considered in Fig. 2. We also perform the calibration test, repeating the same PPD goodness-of-fit test for 10^5 simulated, noisy data vectors sampled at the 3 x 2 pt best-fitting cosmology, and using importance sampling to rapidly compute corresponding p-values. We show their distribution in Fig. C1. We then compute a calibrated $\tilde{p}$-value, given by the ratio of simulated p-values below the one measured from data. We find similar values for this specific test, which indicates that, as expected, our choice of statistics has little impact on goodness-of-fit tests.

We compute a calibrated $\tilde{p}$-value of 0.065 ($p = 0.046$ uncalibrated) for the full 3 x 2 pt data vector, indicating an acceptable fit given our criterion of $\tilde{p} > 0.01$. We report calibrated $\tilde{p}$-values for individual probes in Table 1. Additionally, we find calibrated $\tilde{p}$-values of 0.373 when considering both components of cosmic shear, and 0.111 when considering $\gamma_{\rm t}$ and $w(\theta)$. Each panel of Fig. 2 also indicates the calibrated $\tilde{p}$-value computed for that particular subset of the data (still using the PPD realization from the full 3 x 2 pt data vector). Given the potential pitfalls of multiple hypothesis testing, we use these individual $\tilde{p}$-values mostly to rank the different bin combinations by largest discrepancy with the model. The lowest $\tilde{p}$-value is obtained for the (2,4) redshift bin combination of ξ_-. We discuss this particular subset of the data more in Section 4.7.

4.2 Goodness of fit of individual two-point functions

We now consider goodness-of-fit tests of the different two-point function components of the full 3 x 2 pt data vector separately, i.e. we test for the goodness-of-fit of cosmic shear $\xi_\pm$, galaxy–galaxy lensing $\gamma_{\rm t}$ and clustering w, one at a time. This differs from the test considered in Section 4.1 in that we condition on the posteriors for each subset of the 3 x 2 pt data vector, rather than on the posterior from the analysis of the full 3 x 2 pt vector. This allows us, first, to test each probe separately, and then to exclude the case where the 3 x 2 pt test $\tilde{p}$-value is dominated by one or two probes, which could mask a poor fit to the third. Given the similarities between these tests and that of Section 4.1, we relegate the associated plots to Appendix A.

Fig. A1 shows the graphical PPD test for cosmic shear. There are no obvious discrepancies between the PPD realizations of cosmic shear and the actual data. We compute a calibrated $\tilde{p}$-value for cosmic shear of $\tilde{p} = 0.386$, indicating no evidence for tension between the measurements and PPD realizations. As for the 3 x 2 pt goodness-of-fit test, though, the (2,4) bin combination of ξ_- shows a low p-value with $\tilde{p} = 0.013$. Fig. A2 shows the PPD realizations for galaxy–galaxy lensing. The PPD realizations in this case look generally consistent with the data. Considered as a whole, though, the $\gamma_{\rm t}$ measurements exhibit a good fit, with $\tilde{p} = 0.262$. Fig. A3 shows the PPD realizations for clustering. Again, the realizations show good agreement with the data. The p-value for $w(\theta)$ is $\tilde{p} = 0.057$.

4.3 Testing for internal consistency between the two-point functions

We next consider PPD realizations of the form $P(d_1^{\rm rep} \mid d_2^{\rm obs})$, where d_1 represents one of $\xi_\pm$, $\gamma_{\rm t}$, w, and d_2 represents the remaining elements of the 3 x 2 pt data vector from the two other probes. Such tests are interesting for several reasons. First, tests of this form can be used to split parts of the 3 x 2 pt data vector that depend on different known systematics, such as shear calibration bias. For

Table 1. Summary of calibrated $\tilde{p}$-values obtained for all tests. The 'Test' column specifies the test. The second and third columns show the observables considered for sampling ($d_{\rm rep}$) and conditioning ($d_{\rm obs}$). The fourth column shows the $\tilde{p}$-value for all observables considered in the test with the uncalibrated p-value in parenthesis, while the rest of the columns show the $\tilde{p}$-value when considering individual observables.

			PPD test calibrated $\tilde{p}$-values				
Test	$d_{\rm rep}$	$d_{\rm obs}$	$d_{\rm rep}\|d_{\rm obs}$	$\xi_+\|d_{\rm obs}$	$\xi_-\|d_{\rm obs}$	$\gamma_{\rm t}\|d_{\rm obs}$	$w\|d_{\rm obs}$
Goodness-of-fit tests							
Full 3 x 2 pt	$\xi_+, \xi_-, \gamma_{\rm t}, w$	$\xi_+, \xi_-, \gamma_{\rm t}, w$	0.065 (0.046)	0.537	0.182	0.238	0.071
Cosmic shear	ξ_+, ξ_-	ξ_+, ξ_-	0.386 (0.396)	0.533	0.192	–	–
Galaxy–galaxy lensing	$\gamma_{\rm t}$	$\gamma_{\rm t}$	0.262 (0.245)	–	–	0.262	–
Clustering	w	w	0.057 (0.050)	–	–	–	0.057
Consistency tests: data type splits							
Cosmic shear *versus* galaxy–galaxy lensing and clustering	ξ_+, ξ_-	$\gamma_{\rm t}, w$	0.396 (0.299)	0.600	0.162	–	–
Galaxy–galaxy lensing *versus* cosmic shear and clustering	$\gamma_{\rm t}$	ξ_+, ξ_-, w	0.336 (0.142)	–	–	0.336	–
Clustering *versus*cosmic shear and galaxy–galaxy lensing	w	$\xi_+, \xi_-, \gamma_{\rm t}$	0.050 (3.6×10^{-5})	–	–	–	0.050
Consistency tests: other							
Cosmic shear: bin 1 *versus* no bin 1	ξ_+, ξ_-	ξ_+, ξ_-	0.639 (0.532)	0.648	0.543	–	–
Cosmic shear: bin 2 *versus* no bin 2	ξ_+, ξ_-	ξ_+, ξ_-	0.392 (0.344)	0.379	0.366	–	–
Cosmic shear: bin 3 *versus* no bin 3	ξ_+, ξ_-	ξ_+, ξ_-	0.547 (0.372)	0.771	0.287	–	–
Cosmic shear: bin 4 *versus* no bin 4	ξ_+, ξ_-	ξ_+, ξ_-	0.376 (0.293)	0.593	0.095	–	–
3 x 2 pt: large *versus* small scales	$\xi_+, \xi_-, \gamma_{\rm t}, w$	$\xi_+, \xi_-, \gamma_{\rm t}, w$	0.034 (0.016)	0.091	0.741	0.167	0.030
Cosmic shear: ξ_- *versus* ξ_+	ξ_-	ξ_+	0.186 (0.151)	–	0.186	–	–

example, additive shear systematics can impact measurement of $\xi_\pm$, but are expected to not impact $\gamma_{\rm t}$ significantly, and have no impact on clustering. Therefore, if we set $d_1 = w$ and $d_2 = \xi_\pm, \gamma_{\rm t}$, then d_2 is impacted by potential shear biases while d_1 is not. Similarly, if $d_1 = \xi_\pm$ while $d_2 = \gamma_{\rm t}, w$, then d_2 will be impacted by potential biases in the lens galaxy redshifts, while d_1 will not. Secondly, departures from ΛCDM might be expected to appear in some observables, but not in others. For instance, a split for which d_2 depends on lensing while d_1 does not would show tension in models where lensing is impacted by beyond-ΛCDM physics while clustering is not. For example, this is the case for the (Σ, μ) parametrizations of modified gravity that perturb the Poisson equation differently for light and matter, as explored with DES Y1 data (DES Collaboration 2019a) and other galaxy surveys and CMB data (Simpson et al. 2013; Mueller et al. 2018; Ferté et al. 2019; Planck Collaboration VI 2020).

We note that the consistency tests considered in this section are similar to the test of consistency between $\xi_\pm$ and $\{\gamma_{\rm t}, w\}$ considered in DES Collaboration (2018), where these two subsets of the 3 x 2 pt data vector were found to be consistent. However, the test presented in DES Collaboration (2018) used an evidence ratio to identify consistency. Furthermore, the evidence ratio test naturally lives in model space rather than data space. Consequently, it is difficult to use the evidence ratio test to evaluate particularly discrepant elements of the data vector. The PPD test on the other hand, does this naturally.

Fig. 3 shows the PPD realizations for clustering, conditioned on the observed cosmic shear and galaxy–galaxy lensing measurements. Observational systematics, such as galactic dust, are expected to impact clustering more than cosmic shear and galaxy–galaxy lensing. Galactic dust would tend to obscure galaxies, modulating the number density (and thus $w(\theta)$), but likely not having a large impact on inferred shears. Furthermore, as pointed out by Leauthaud et al. (2017), gravitational lensing measurements around BOSS-selected galaxies show a lower amplitude signal than expected based on the clustering properties of the galaxies. If a similar discrepancy was born out in DES data, this test would be expected to reveal it.

The results in Fig. 3 show that $w(\theta)$ appears generally within the bounds of the PPD realizations. Given the large covariance between the $w(\theta)$ points, this appearance can be deceiving: we find that the uncalibrated p-values for several of the $w(\theta)$ redshift bins are quite low, in the range of 0.01–0.02. However, we report a calibrated $\tilde{p}$-value of 0.050, revealing agreement between clustering measurements and expectations from shear and galaxy–galaxy lensing measurements. We therefore note that this test is an example of the case presented in Section 3.2, where the part of the data vector that is being tested has a different parameter dependence than the part that is used for conditioning. Namely, $w(\theta)$ is more sensitive to linear galaxy bias than $\{\gamma_{\rm t}, \xi_\pm\}$ and exhibits different parameter degeneracies. The theory data vectors and consequently PPD realizations for $w(\theta)$ conditioned on $\{\gamma_{\rm t}, \xi_\pm\}$ therefore have amplitudes spread over a wide range, which is visible in Fig. 3. Consequently, the uncalibrated p-value in this case will be driven low as discussed in Section 3.2 and shown in Fig. C1.

We relegate the two other PPD comparisons of this type to Appendix B. Fig. B1 shows the PPD realizations for galaxy–galaxy lensing, conditioned on the observed clustering and cosmic shear measurements. In this case, we find that all of the data points appear to be quite consistent with the PPD realizations, and the $\tilde{p}$-values all appear to be reasonable, with an overall $\tilde{p}$-value of 0.336. Fig. B2 shows the PPD realizations for cosmic shear, conditioned on the observed clustering and galaxy–galaxy lensing measurements. Nevertheless, we note that there appears to be a weak trend for the ξ_+ data points at large scales to be low relative to the PPD realizations. Furthermore, as seen previously, the (2,4) bin of ξ_- yields a low p-value. We find in this case that most of the data appear reasonable given the PPD realizations, with an overall $\tilde{p}$-value of 0.396.

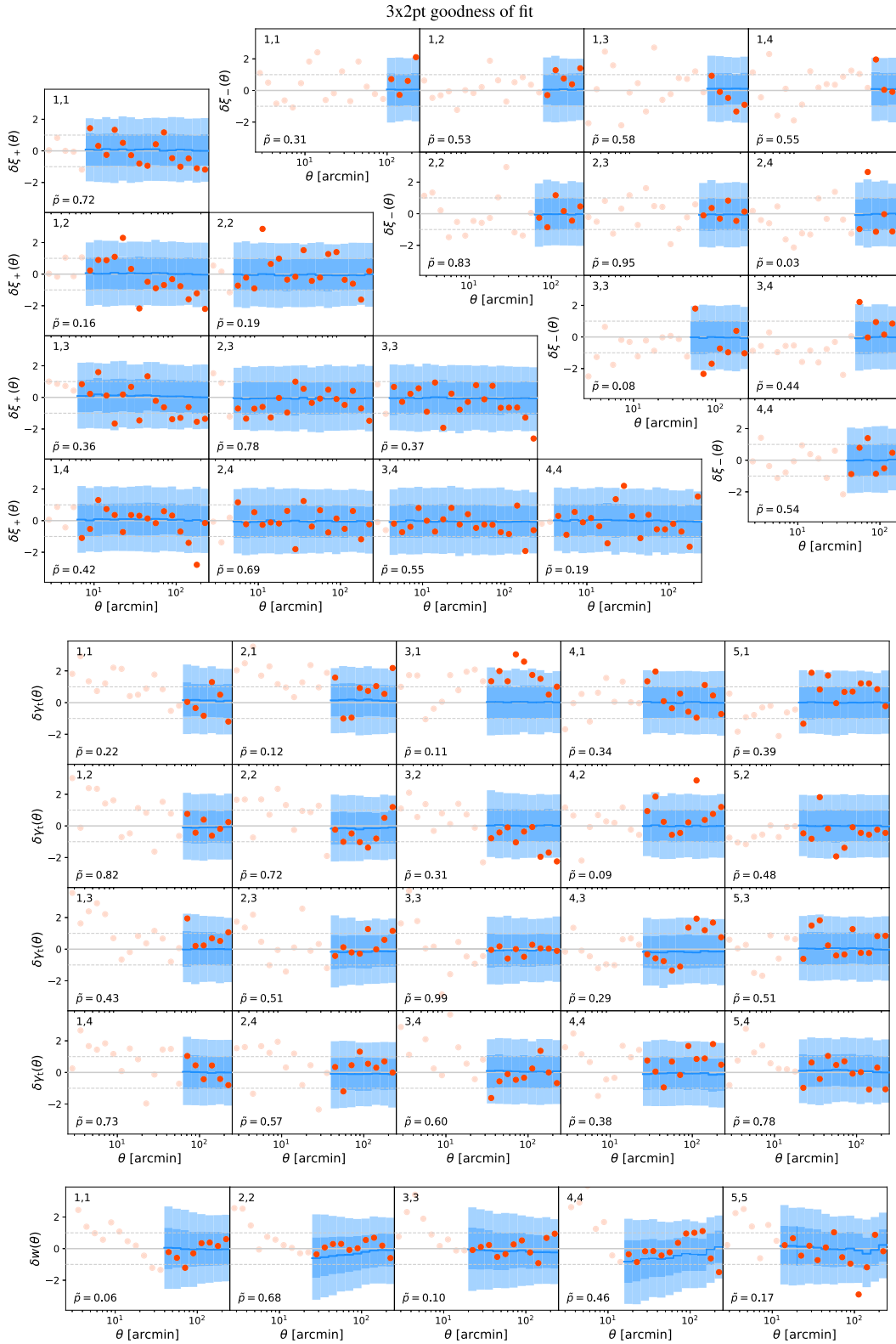

Figure 2. PPD goodness-of-fit test for the 3 x 2 pt data vector. The 68 and 95 per cent confidence bands on the PPD realizations of 3 x 2 pt, conditioned on the posterior from the analysis of the 3 x 2 pt data is shown as the blue bands. Red points represent the actual data. The bands and the data points are plotted relative to the best-fitting 3 x 2 pt theory curve, and are normalized by the diagonal of the 3 x 2 pt covariance, such that data error bars have unit size. The different insets split the datavector into the different observables ($\xi_\pm$, γ_t, and $w(\theta)$), while the different panels split the correlation functions by redshift bin combination. The calibrated $\tilde{p}$-value for each redshift bin combination of each observable is indicated in the bottom left-hand corner of the corresponding panel and calibrated $\tilde{p}$-values per observable and for the entire set are reported in Table 1.

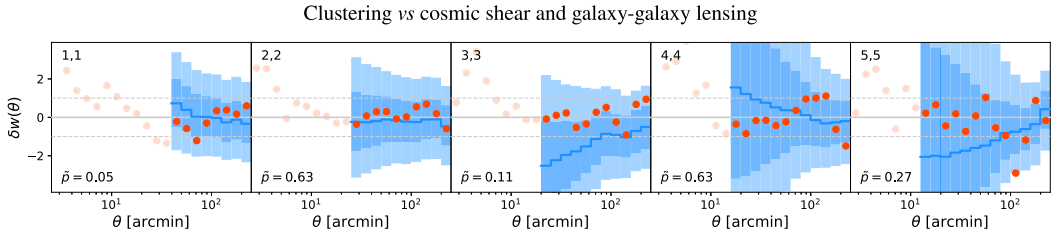

Figure 3. PPD for clustering, conditioned on the posterior from cosmic shear and galaxy–galaxy lensing. See Fig. 2 for explanation of bands.

4.4 Testing for cosmic shear redshift-dependent inconsistency

Splits of the data by redshift bin are motivated both as a probe of departures from ΛCDM and by concerns of systematic errors. Models with an evolving dark energy equation of state, for instance, would predict redshift-dependent departures from ΛCDM that could be revealed by such tests. Similarly, shear measurements at high redshift could be more impacted by issues such as source blending and PSF modelling uncertainty, while low-redshift measurements are more impacted by modelling errors of the non-linear matter power spectrum and intrinsic alignments, which might then lead to tension between low- and high-redshift measurements.

One complication of splitting the data on redshift bin is that certain model parameters – such as galaxy bias or the multiplicative shear bias parameters – impact only one of the redshift bins. These parameters will then be unconstrained when conditioning on the other redshift bins. For such parameters, the PPD realizations will then involve drawing from the parameter priors. This is not problematic for parameters like multiplicative shear bias and photometric redshift bias, which are prior dominated anyways, but is problematic for linear galaxy bias. If a bias parameter is unconstrained, the PPD realizations will necessarily span a much broader range than the data, making the graphical PPD tests difficult. We avoid this issue by focusing on redshift splits of the cosmic shear data vector, which is unaffected by galaxy bias.

Fig. 4 shows the PPD realizations for single bins of cosmic shear, conditioned on the realizations from the other redshift bins. For instance, the upper plot shows the PPD consistency test for bins (1,1), (1,2), (1,3), and (1,4), conditioned on measurements of all other redshift bin combinations. The PPD realizations and $\tilde{p}$-values generally appear reasonable, although again the (2,4) bins of ξ_- consistently exhibits a $\tilde{p}$-value close to 0.01. However, all overall calibrated $\tilde{p}$-values are well above 0.01, indicating no sign of tensions between redshift bins in DES Y1 cosmic shear data.

4.5 Testing for large and small-scale systematics

Like splits of the data on redshift bin, splits of the data on angular scale are motivated by both considerations of physics and systematic errors. Departures of galaxy clustering from the assumed linear bias model, for instance, could lead to tension between the small and large-scale measurements. The measurements of clustering at large scales are expected to be particularly susceptible to data systematics, such as dust extinction or varying observing conditions (fluctuations in depth, airmass, exposure time, width of the point spread function). We therefore consider PPD realizations of the large-scale components of the 3 x 2 pt data vector, conditioned on the small-scale measurements. We choose $\theta = 100$ arcmin as the splitting point, leaving four data points for each two-point function in the large-scale parts.

Fig. 5 shows the PPD realizations for the large-scale 3 x 2 pt, conditioned on the small-scale measurements. We obtained an overall $\tilde{p}$-value of 0.034, relatively low compared to the full 3 x 2 pt goodness of fit. Although some individual two-point functions show low p-values, only the (3,4) bin of ξ_+ shows a p-value below 0.01, and we do not observe any noticeable trend related to redshift based on p-values alone. When considering clustering alone (still conditioned on small-scale measurements of all observables), we obtain a $\tilde{p}$-value of 0.030. Individual bins of the clustering two-point functions show lower p-values at higher redshifts, which could point to systematic effects affecting large-scale clustering measurements, such as survey observing conditions. We note this is another example of consistency tests where the uncalibrated p-value is indeed biased low, as shown in Fig. C1.

4.6 Testing for cosmic shear systematics

We finally consider tension within the cosmic shear measurements by splitting the two components ξ_+ and ξ_-. This test is motivated by potential systematic effects as well as modelling considerations. On the systematics side, PSF leakage and shear-dependent selection biases can generate a B-mode pattern that affects each component differently (ξ_+ is related to the sum of E- and B-mode power spectra, while ξ_- is related to their difference). On the modelling side, we note that ξ_+ and ξ_- receive contributions from different physical modes at given angular separation θ. In particular, ξ_- receives contributions from smaller scales impacted by non-linear evolution and baryonic physics, which justify stricter scale cuts for ξ_- than ξ_+. Since ξ_+ has a higher signal-to-noise ratio than ξ_- for the scales we consider, we will apply the test to ξ_- conditioned on the ξ_+ posterior.

Fig. B3 shows the PPD test in this case. We measure a $\tilde{p}$-value of 0.186, indicating good agreement between both components of cosmic shear measurements.

4.7 The impact of the (2,4) bin of ξ_-

We consistently find in several tests described above that the (2,4) redshift bin combination for ξ_- yields a low $\tilde{p}$-value, even after calibration. These tests are correlated (since they share parts of the same input data), so the fact that multiple tests show similarly low $\tilde{p}$-values is not surprising. Moreover, as pointed out in Section 3.2, we have looked at many $\tilde{p}$-values for different redshift bin combinations, so the fact that one of them shows a low $\tilde{p}$-value is not unexpected. Given that ξ_- is sensitive to smaller scales and more susceptible to systematics than ξ_+, the fact that this bin shows the largest discrepancy between the observed data and that predicted by the PPD motivates us to verify its impact on the DES Y1 3 x 2 pt constraints. When measuring the goodness of fit of the full 3 x

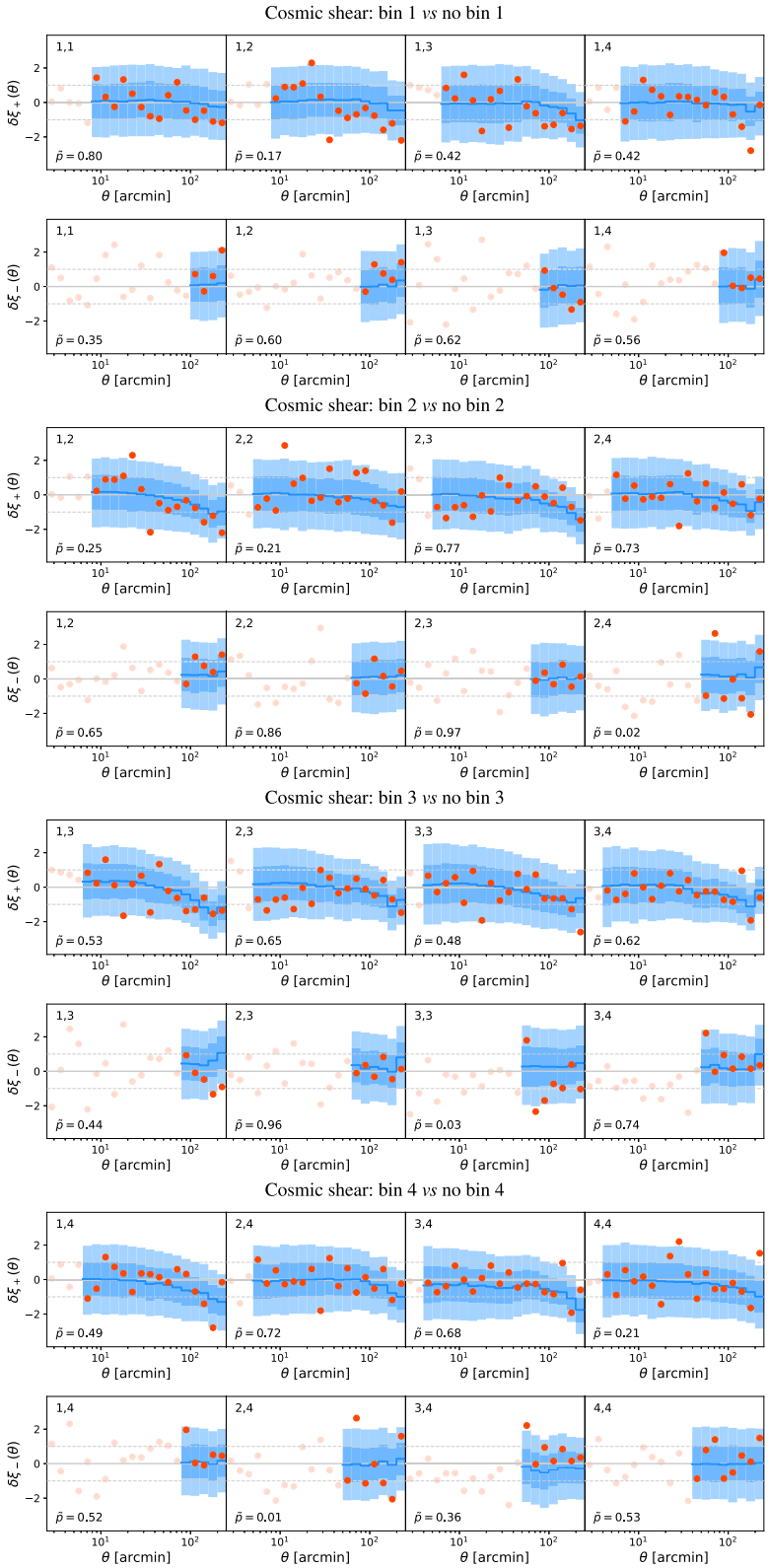

Figure 4. PPD for cosmic shear where one redshift bin is tested at a time, conditional to the posterior obtained from cosmic shear by removing auto and cross-correlations with that bin. See Fig. 2 for explanation of bands.

2 pt vector, removing this bin increases the agreement between the model and the data, yielding a $\tilde{p}$-value of 0.13 ($p = 0.10$ uncalibrated), compared to 0.065 for the full data vector. Fig. 6 shows the impact on the DES Y1 3 x 2 pt cosmological constraints of removing the (2,4) redshift bin combination of ξ_-. We find that the impact on the cosmological constraints is negligible, essentially within the uncertainty of the sampling algorithm. We conclude that, while it may be the case that the (2,4) bin combination of ξ_-

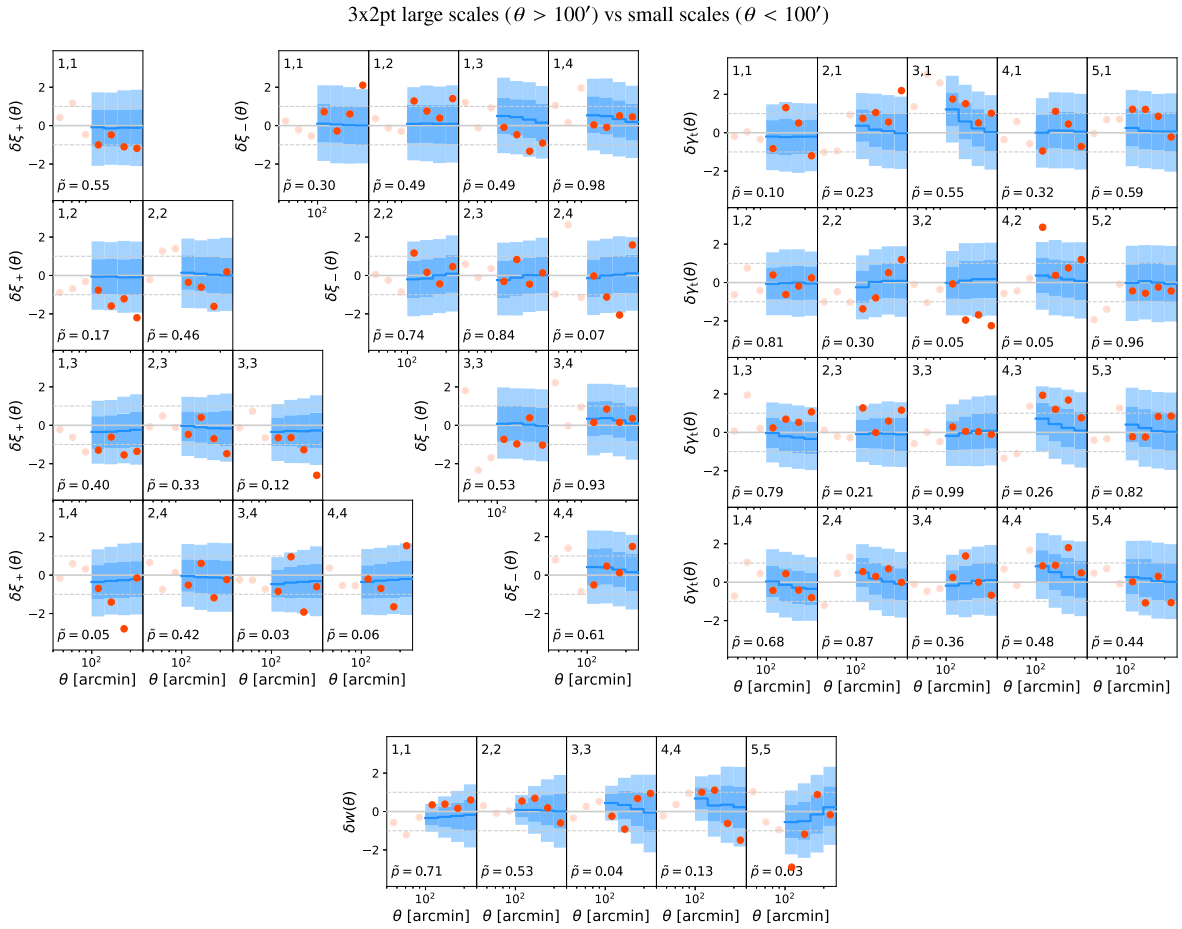

Figure 5. PPD for all 3 x 2 pt correlation functions at large scales ($\theta > 100$ arcmin) conditioned on observed correlation functions at small scales (*i.e.* data points at separation angle θ within scale cuts and $\theta < 100$ arcmin). See Fig. 2 for explanation of bands.

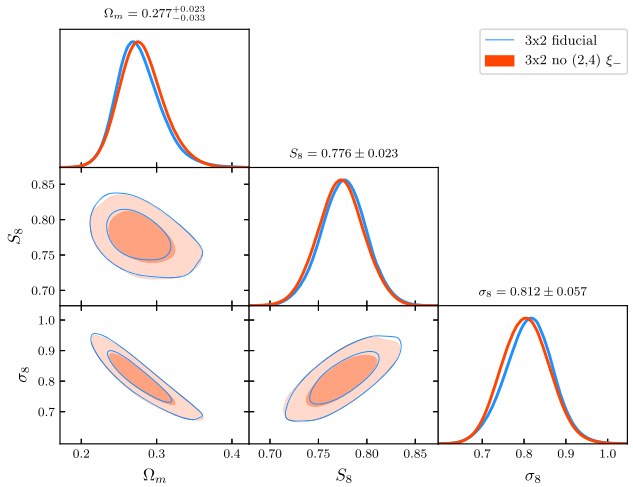

Figure 6. Impact on cosmological constraints of removing the $(2, 4)\,\xi_-$ bin from 3 x 2 pt.

yields a bad fit to ΛCDM, this measurement alone has little impact on the DES Y1 cosmological results. We also note that this bin combination makes up only about 1.5 per cent of the total DES Y1 data vector.

5 INTERNAL CONSISTENCY TESTS FOR DES YEAR 3

In addition to testing the internal consistency of the DES Y1 joint probes analysis, one of the goals of this paper is to lay out the set of consistency tests that will be applied to the DES Y3 joint probes analysis. We plan on performing the same tests as those that were applied here to DES Y1 (summarized in Table 1), supplemented by additional probe-specific tests, such as testing large *versus* small scales in cosmic shear alone. Importantly, we select the few most relevant tests to be performed on the unblinded data vector prior to any other comparison with our model. If these tests pass, we will allow comparing the data with the best-fitting prediction and move forward to the parameter-level unblinding stage (the blinding methodology is described in Muir et al. 2020). We keep the number of tests small in order to avoid redundancy and look-elsewhere effects (discussed in Section 3.3.2). We therefore decide to fix a threshold at $p = 0.01$ for each test used as part of the unblinding process and to select the following:

(i) Goodness-of-fit tests: (1) full 3 x 2 pt, (2) cosmic shear, (3) galaxy–galaxy lensing and clustering (referred to as 2 x 2 pt);
(ii) Consistency test: (4) cosmic shear *versus* galaxy–galaxy lensing and clustering (2 x 2 pt).

The three goodness-of-fit tests allow us to verify that the baseline model used in the analysis – including the cosmological model,

but also models for intrinsic alignments, photometric redshifts, multiplicative bias and galaxy bias – is a good fit to the full data set as well as each individual probe. The consistency test allows us to verify that cosmic shear measurements are compatible with galaxy–galaxy lensing and clustering measurements (2 x 2 pt) given the baseline model, indicating that it is sensible to combine them into the full 3 x 2 pt analysis. Because we a priori do not expect Y3 data to be sensitive enough to rule out both ΛCDM and wCDM, we have chosen to allow revisiting the potential for bugs or flaws with non-cosmological parts of the modelling pipeline if we find the data incompatible even in wCDM, before seeing the final parameter values in any model. Once the data have passed all unblinding criteria including those four tests, we will perform all other internal consistency tests, for which we will report calibrated $\tilde{p}$-values as well.

6 CONCLUSION

In the context of mild to severe tensions between cosmological constraints on the ΛCDM model reported by multiple experiments, it is crucial to assess the internal consistency of individual data sets. In this paper, we have performed a series of internal consistency tests of the DES Y1 3 x 2 pt data using the PPD. The PPD represents the distribution of possible (unobserved) data, conditioned on observed data, under a shared model. The PPD tests have the advantage of performing comparisons directly in data space and are not impacted by prior volume effects, making it a particularly useful consistency test. By comparing the PPD realizations to the true data, both with a χ^2 test statistic and graphically, we assess the consistency of the DES data. We perform two kinds of tests: *goodness-of-fit* tests to assess whether the model favoured by the data is actually a good fit to DES Y1 3 x 2 pt measurements, and *consistency* tests between disjoint subsets of the full data vector. In particular, we split the data vector into subvectors corresponding to different observables (cosmic shear, galaxy–galaxy lensing and clustering), measurements at small and large scales, and different redshift bins of cosmic shear data. The choice of χ^2 test statistic yields conservative measures of consistency and we propose a calibration method to overcome exceedingly conservative p-values that may occur when the data splits result in too different posterior distributions. This method is applied consistently to all tests and we report such calibrated $\tilde{p}$-values throughout this analysis.

In general, we find that the DES Y1 3 x 2 pt data are self-consistent, and have an acceptable fit to ΛCDM. A direct graphical comparison of the PPD realizations to the true data yields no obvious discrepancies. These results provides a strong validation of the DES Y1 measurements and cosmological constraints, as well as ΛCDM. However, there are a few peculiarities of the data. First, we find a somewhat low goodness-of-fit statistic for the full data set of $\tilde{p} = 0.065$ ($p = 0.046$ uncalibrated). Secondly, we find a low p-value for the consistency test comparing large-scale to small-scale data elements (with a split at separation angle $\theta = 100$ arcmin) which suggests a small tension close to the 2σ level. This indicates either insufficient accuracy of the modelling of small-scale measurements or some observational systematic effect likely to impact large-scale measurements, potentially explaining the overall low $\tilde{p}$-value. Finally, we find that the (2,4) bin combination of ξ_- consistently yields a low $\tilde{p}$-value. When this bin of ξ_- is excluded, the $\tilde{p}$-value for the full 3 x 2 pt data vector improves to $\tilde{p} = 0.13$. However, excluding this bin from the analysis has negligible impact on the DES Y1 cosmological constraints.

The methodology developed here will be applied to the forthcoming analysis of the 3 x 2 pt data vector measured from DES Y3 data. The improvements in statistical noise with Y3 data make such tests even more interesting. In particular, these tests will be essential to test the consistency with the cosmological model and look for unmodelled systematic effects. This would be even more relevant if the data were to show any sign of a real departure from the predictions of ΛCDM.

ACKNOWLEDGEMENTS

This paper has gone through internal review by the DES collaboration. The reviewers were Alex Alarcon, Andresa Compos, and Youngsoo Park.

The authors would like to thank Masahiro Takada for fruitful discussions at early stages of the project, and Vivian Miranda and Scott Dodelson for useful comments and discussions. The authors would like to kindly thank the anonymous referee for their comments which helped us improve this paper.

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 DES.

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

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

The DES data management system is supported by the National Science Foundation under grant numbers AST-1138766 and AST-1536171. The DES participants from Spanish institutions are partially supported by MICINN under grants ESP2017-89838, PGC2018-094773, PGC2018-102021, SEV-2016-0588, SEV-2016-0597, and MDM-2015-0509, some of which include ERDF funds from the European Union. IFAE is partially funded by the CERCA program of the Generalitat de Catalunya. Research leading to these results has received funding from the European Research

Council under the European Union's Seventh Framework Program (FP7/2007-2013) including ERC grant agreements 240672, 291329, and 306478. We acknowledge support from the Brazilian Instituto Nacional de Ciência e Tecnologia (INCT) do e-Universo (CNPq grant 465376/2014-2).

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

DATA AVAILABILITY

A general description of DES data releases is available on the survey website at https://www.darkenergysurvey.org/the-des-project/data-access/. DES Y1 cosmological data are available on the DES Data Management website hosted by the National Center for Supercomputing Applications at https://des.ncsa.illinois.edu/releases/y1a1. This includes the data vectors, redshift distributions, and some of the posterior samples used in this analysis. The `CosmoSIS` software (Zuntz et al. 2015) is available at https://bitbucket.org/joezuntz/cosmosis/wiki/Home.

REFERENCES

Aitken J. A. O., 2013, BMC Systems Biology, 7, 72
Asgari M. et al., 2020, A&A, 634, A127
Aylor K., Joy M., Knox L., Millea M., Raghunathan S., Kimmy Wu W. L., 2019, ApJ, 874, 4
Battye R. A., Charnock T., Moss A., 2015, Phys. Rev. D, 91, 103508
Bernal J. L., Verde L., Riess A. G., 2016, JCAP, 2016, 019
Chang C. et al., 2019, MNRAS, 482, 3696
DES Collaboration, 2018, Phys. Rev. D, 98, 043526
DES Collaboration, 2019a, Phys. Rev. D, 99, 123505
DES Collaboration, 2019b, Phys. Rev. D, 100, 023541
Dunn O. J., 1959, Ann. Math. Statist., 30, 192
Elvin-Poole J. et al., 2018, Phys. Rev. D, 98, 042006
Feeney S. M., Mortlock D. J., Dalmasso N., 2018, MNRAS, 476, 3861
Feeney S. M., Peiris H. V., Williamson A. R., Nissanke S. M., Mortlock D. J., Alsing J., Scolnic D., 2019, Phys. Rev. Lett., 122, 061105
Ferté A., Kirk D., Liddle A. R., Zuntz J., 2019, Phys. Rev. D, 99, 083512
Gelman A., 2013, Electron. J. Statist., 7, 2595
Gelman A., Carlin J. B., Stern H. S., Rubin D. B., 2004, Bayesian data analysis, 2nd ed. Chapman and Hall, Boca Raton, FL
Handley W., Lemos P., 2019, Phys. Rev., D100, 043504
Handley W. J., Hobson M. P., Lasenby A. N., 2015a, MNRAS, 450, L61
Handley W. J., Hobson M. P., Lasenby A. N., 2015b, MNRAS, 453, 4384
Heymans C. et al., 2021, A&A, 646, A140
Hildebrandt H. et al., 2017, MNRAS, 465, 1454
Joudaki S. et al., 2020, A&A, 638, L1
Köhlinger F., Joachimi B., Asgari M., Viola M., Joudaki S., Tröster T., 2019, MNRAS, 484, 3126
Krause E. et al., 2017, preprint (arXiv:1706.09359)
Leauthaud A. et al., 2017, MNRAS, 467, 3024
Lemos P., Köhlinger F., Handley W., Joachimi B., Whiteway L., Lahav O., 2020, MNRAS, 496, 4647
Louca A. J., Sellentin E., 2020, Open J. Astrophys., 3, 11
MacCrann N., Zuntz J., Bridle S., Jain B., Becker M. R., 2015, MNRAS, 451, 2877
Madan D. B., Seneta E., 1990, J. Bus., 63, 511
Marshall P., Rajguru N., Slosar A., 2006, Phys. Rev. D, 73, 067302
Mueller E.-M., Percival W., Linder E., Alam S., Zhao G.-B., Sánchez A. G., Beutler F., Brinkmann J., 2018, MNRAS, 475, 2122
Muir J. et al., 2020, MNRAS, 494, 4454
Neal R. M., 2000, preprint (arXiv:physics/0009028)
Park Y., Rozo E., 2020, MNRAS, 499, 4638
Planck CollaborationVI, 2020, A&A, 641, A6
Prat J. et al., 2018, Phys. Rev. D, 98, 042005
Raveri M., 2016, Phys. Rev. D, 93, 043522
Raveri M., Hu W., 2019, Phys. Rev. D, 99, 043506
Riess A. G., Casertano S., Yuan W., Macri L. M., Scolnic D., 2019, ApJ, 876, 85
Rozo E. et al., 2016, MNRAS, 461, 1431
Sellentin E., Heavens A. F., 2018, MNRAS, 473, 2355
Sheldon E. S., Huff E. M., 2017, ApJ, 841, 24
Simpson F. et al., 2013, MNRAS, 429, 2249
The Dark Energy Survey Collaboration, 2005, preprint (arXiv:astro–ph/0510346)
Troxel M. A. et al., 2018, Phys. Rev. D, 98, 043528
Zuntz J. et al., 2015, Astron. Comput., 12, 45
Zuntz J. et al., 2018, MNRAS, 481, 1149

APPENDIX A: GOODNESS-OF-FIT TESTS OF INDIVIDUAL OBSERVABLES

In this section, we present PPD goodness-of-fit tests for individual DES Y1 3 x 2 pt probes – cosmic shear in Fig. A1, galaxy–galaxy lensing in Fig. A2 and clustering in Fig. A3.

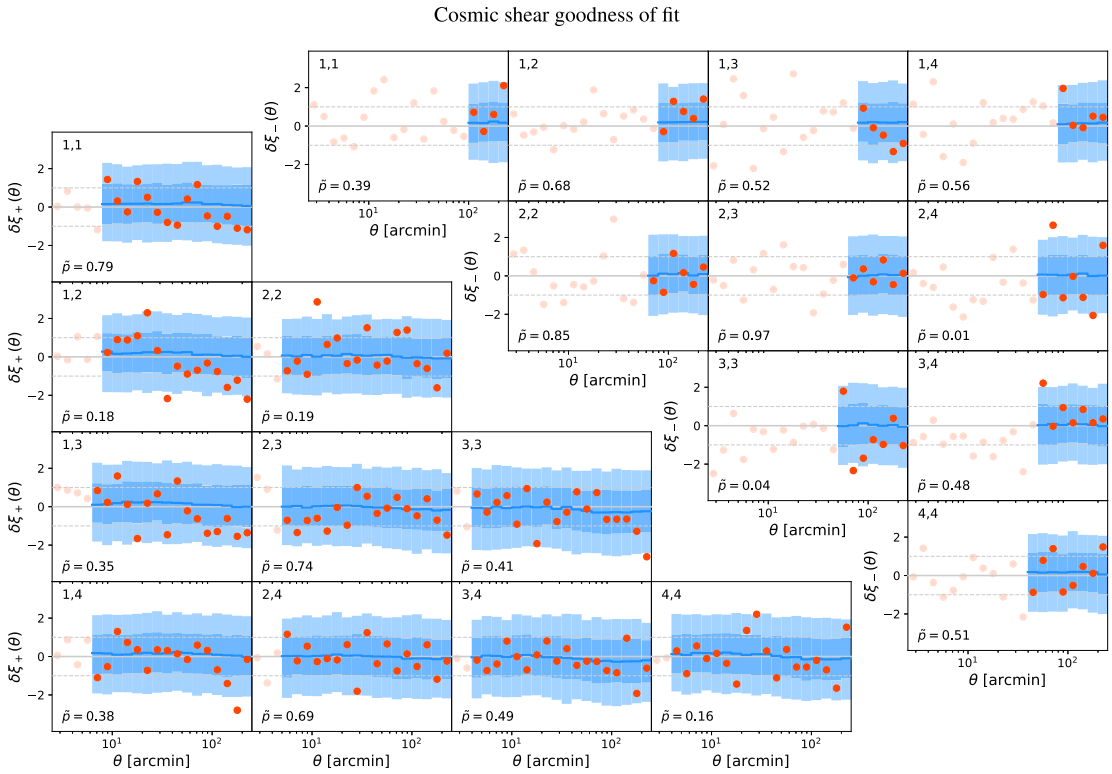

Figure A1. PPD goodness-of-fit test for cosmic shear alone. See Fig. 2 for explanation of bands.

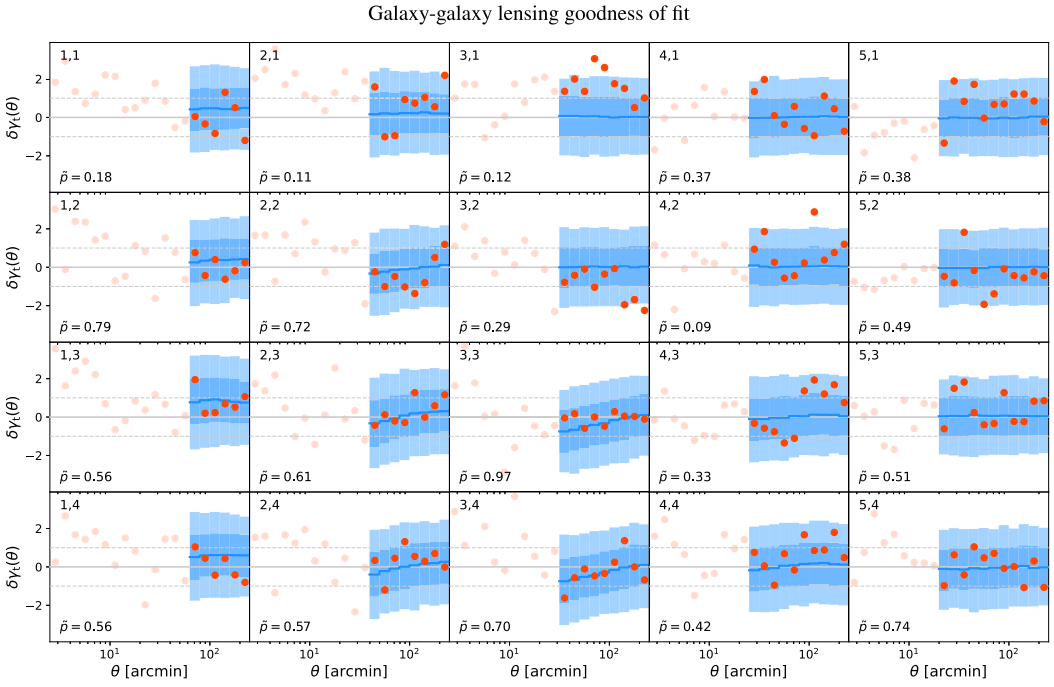

Figure A2. PPD goodness-of-fit test for galaxy–galaxy lensing alone. See Fig. 2 for explanation of bands.

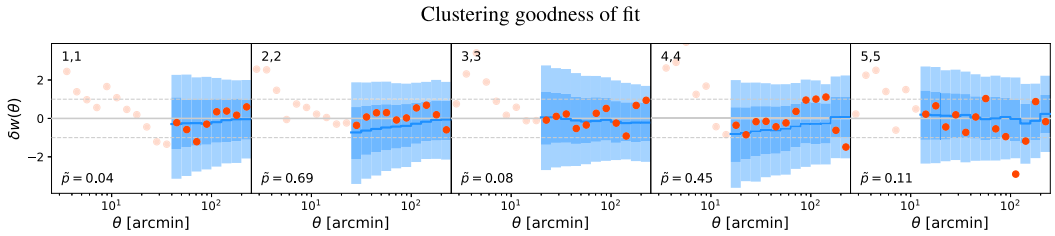

Figure A3. PPD goodness-of-fit test for clustering alone. See Fig. 2 for explanation of bands.

APPENDIX B: ADDITIONAL CONDITIONAL TESTS

In this section, we present additional results of PPD consistency tests. Figs B1 and B2 show, respectively, tests of galaxy–galaxy lensing and cosmic shear, conditioned on the two other probes. Fig. B3 shows the PPD test for cosmic shear ξ_- conditioned on ξ_+ measurements.

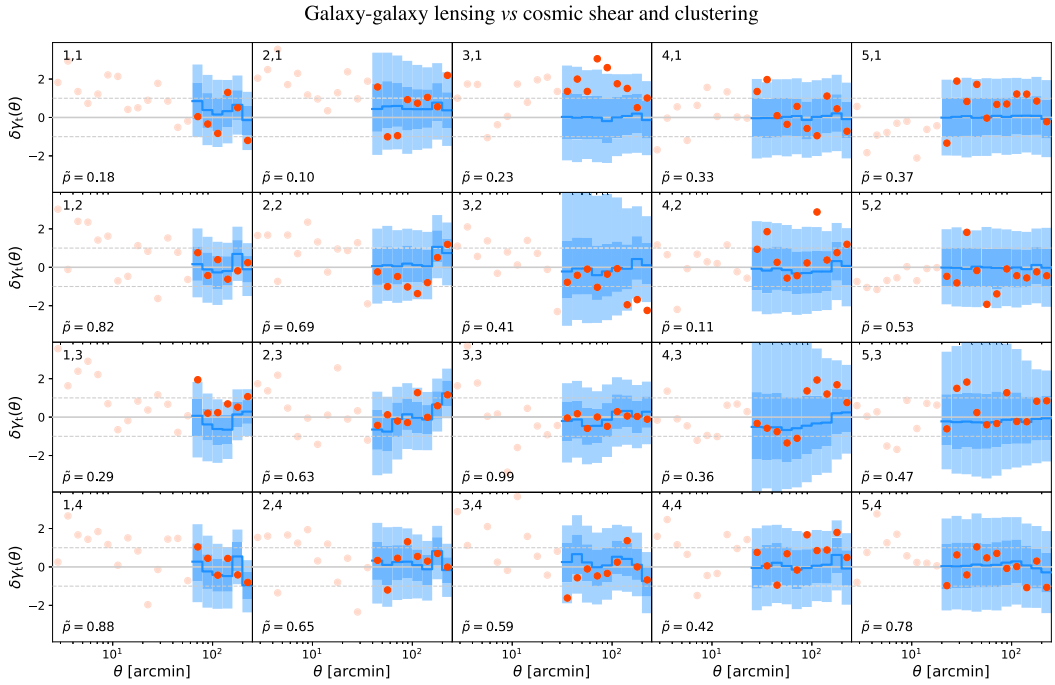

Figure B1. PPD for galaxy–galaxy lensing, conditioned on the posterior from cosmic shear and clustering. See Fig. 2 for explanation of bands.

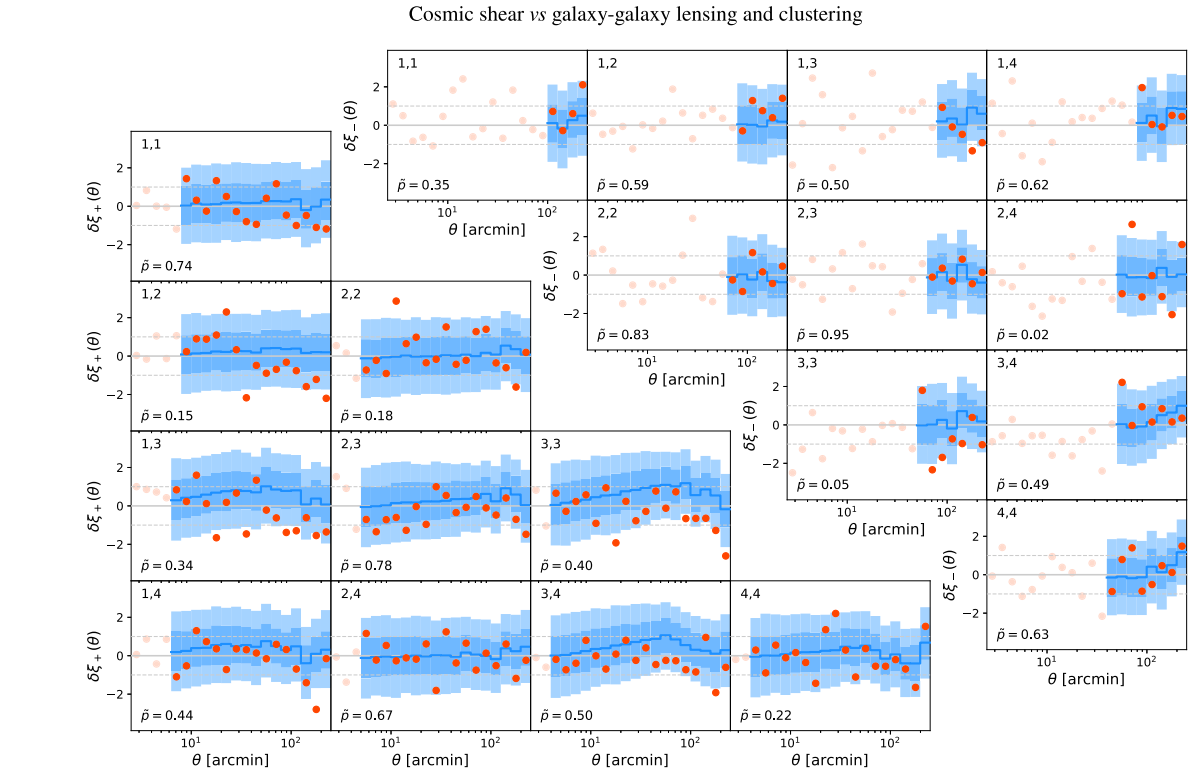

Figure B2. PPD for cosmic shear, conditioned on the posterior from galaxy–galaxy lensing and clustering. See Fig. 2 for explanation of bands.

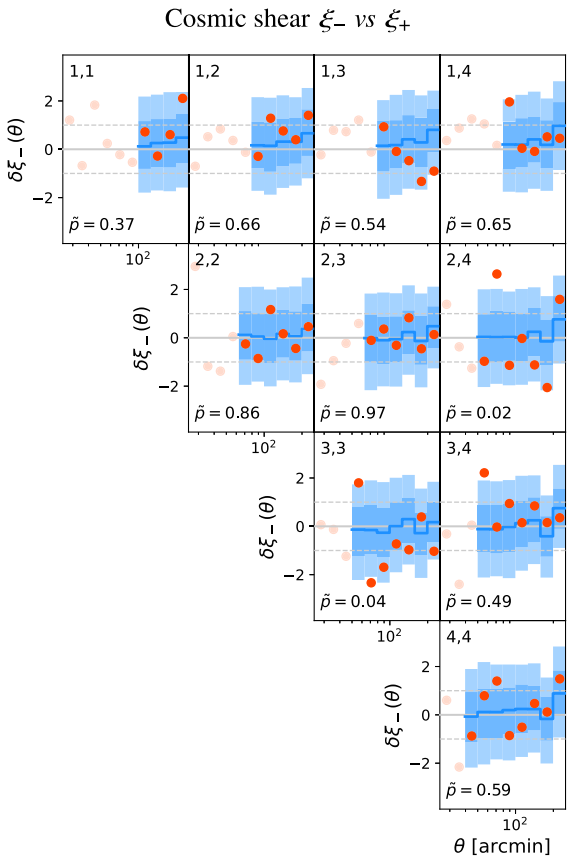

Figure B3. PPD for ξ_-, conditioned on the posterior from ξ_+. See Fig. 2 for explanation of bands.

APPENDIX C: CALIBRATION OF PPD TESTS

In this section, we present the distributions of uncalibrated p-values obtained for simulated data vectors generated at the 3 x 2 pt best-fitting cosmology, for five relevant PPD tests. Fig. C1 shows histograms these p-values – computed with importance sampling as explained Section 3.3.2 – for five cases: the 3 x 2 pt goodness-of-fit test (Section 4.1), the 3 x 2 pt large *versus* small scales consistency test (Section 4.5), and the three consistency tests between the two-point functions (Section 4.3). We compare these values to those obtained from the actual DES Y1 data (shown by the vertical red lines) and obtain *calibrated* $\tilde{p}$-values given by the fraction of simulated p-values below the observed ones. As expected, the goodness-of-fit test presents a distribution very close to uniform, while those for consistency tests depart from uniformity, with a concentration of simulated p-values at low values that depends on the constraining power of the data splits on one another.

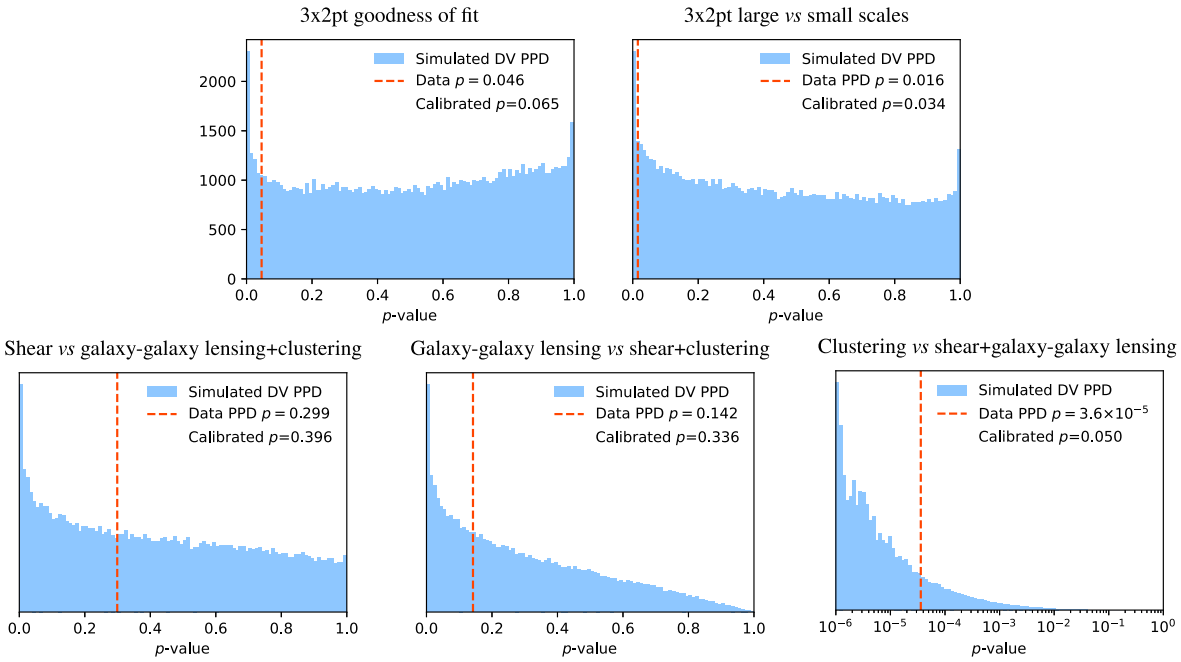

Figure C1. Histograms of uncalibrated p-values (blue histograms) for simulated data vectors (DV) at the 3 x 2 pt best-fitting cosmology, compared to observed uncalibrated p-values from DES Y1 data (in red). Note the last panel uses a logarithmic scale as uncalibrated p-values are found to be very small, which is expected for the comparison of clustering *versus* cosmic shear and galaxy–galaxy lensing.

[1]*Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA*
[2]*Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, HI 96822, USA*
[3]*Department of Physics and Astronomy, University College London, Gower Street, London, WC1E 6BT, UK*
[4]*Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL 60637, USA*
[5]*Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA*
[6]*Argonne National Laboratory, 9700 South Cass Avenue, Lemont, IL 60439, USA*
[7]*Kavli Institute for Particle Astrophysics & Cosmology, P. O. Box 2450, Stanford University, Stanford, CA 94305, USA*
[8]*Department of Physics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15312, USA*
[9]*Center for Cosmology and Astro-Particle Physics, The Ohio State University, Columbus, OH 43210, USA*
[10]*Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, E-08193 Bellaterra, Barcelona, Spain*
[11]*Department of Physics, Stanford University, 382 Via Pueblo Mall, Stanford, CA 94305, USA*
[12]*SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA*
[13]*Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[14]*Kavli Institute for the Physics and Mathematics of the Universe (WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan*
[15]*Department of Astronomy, University of California, 501 Campbell Hall, Berkeley, CA 94720, USA*
[16]*Santa Cruz Institute for Particle Physics, University of California, Santa Cruz, CA 95064, USA*
[17]*Département de Physique Théorique and Center for Astroparticle Physics, Université de Genève, 24 quai Ernest Ansermet, CH-1211 Geneva, Switzerland*
[18]*Faculty of Physics, Ludwig-Maximilians-Universität, Scheinerstr. 1, D-81679 Munich, Germany*
[19]*Max Planck Institute for Extraterrestrial Physics, Giessenbachstrasse, D-85748 Garching, Germany*
[20]*Universitäts-Sternwarte, Fakultät für Physik, Ludwig-Maximilians Universität München, Scheinerstr 1, D-81679 München, Germany*
[21]*Department of Physics, Duke University Durham, NC 27708, USA*
[22]*Institute for Astronomy, University of Edinburgh, Edinburgh EH9 3HJ, UK*
[23]*Cerro Tololo Inter-American Observatory, NSF's National Optical-Infrared Astronomy Research Laboratory, Casilla 603, La Serena, Chile*
[24]*Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, CP 66318, São Paulo, SP - 05314-970, Brazil*
[25]*Laboratório Interinstitucional de e-Astronomia - LIneA, Rua Gal. José Cristino 77, Rio de Janeiro, RJ - 20921-400, Brazil*
[26]*Fermi National Accelerator Laboratory, P. O. Box 500, Batavia, IL 60510, USA*
[27]*Instituto de Fisica Teorica UAM/CSIC, Universidad Autonoma de Madrid, E-28049 Madrid, Spain*
[28]*Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 3FX, UK*
[29]*CNRS, UMR 7095, Institut d'Astrophysique de Paris, F-75014 Paris, France*
[30]*Institut d'Astrophysique de Paris, Sorbonne Universités, UPMC Univ Paris 06, UMR 7095, F-75014 Paris, France*
[31]*Department of Physics and Astronomy, Pevensey Building, University of Sussex, Brighton BN1 9QH, UK*
[32]*Department of Astronomy, University of Illinois at Urbana-Champaign, 1002 W. Green Street, Urbana, IL 61801, USA*
[33]*National Center for Supercomputing Applications, 1205 West Clark St., Urbana, IL 61801, USA*
[34]*Physics Department, University of Wisconsin-Madison, 2320 Chamberlin Hall, 1150 University Avenue Madison, WI 53706-1390, USA*
[35]*INAF-Osservatorio Astronomico di Trieste, via G. B. Tiepolo 11, I-34143 Trieste, Italy*
[36]*Institute for Fundamental Physics of the Universe, Via Beirut 2, I-34014 Trieste, Italy*
[37]*Observatório Nacional, Rua Gal. José Cristino 77, Rio de Janeiro, RJ - 20921-400, Brazil*
[38]*Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA*
[39]*Department of Physics, IIT Hyderabad, Kandi, Telangana 502285, India*
[40]*Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029 Blindern, NO-0315 Oslo, Norway*
[41]*Institut d'Estudis Espacials de Catalunya (IEEC), E-08034 Barcelona, Spain*
[42]*Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, E-08193 Barcelona, Spain*
[43]*Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA*
[44]*Kavli Institute for Cosmology, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[45]*School of Mathematics and Physics, University of Queensland, Brisbane, QLD 4072, Australia*
[46]*Department of Physics, The Ohio State University, Columbus, OH 43210, USA*
[47]*Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA*
[48]*Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA*
[49]*Department of Astronomy/Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721-0065, USA*
[50]*Australian Astronomical Optics, Macquarie University, North Ryde, NSW 2113, Australia*
[51]*Lowell Observatory, 1400 Mars Hill Rd, Flagstaff, AZ 86001, USA*
[52]*Centre for Gravitational Astrophysics, College of Science, The Australian National University, ACT 2601, Australia*
[53]*The Research School of Astronomy and Astrophysics, Australian National University, ACT 2601, Australia*
[54]*Institució Catalana de Recerca i Estudis Avançats, E-08010 Barcelona, Spain*
[55]*Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton, NJ 08544, USA*
[56]*Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), E-28040 Madrid, Spain*
[57]*School of Physics and Astronomy, University of Southampton, Southampton SO17 1BJ, UK*
[58]*Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*

This paper has been typeset from a TEX/LATEX file prepared by the author.