Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **534,** 3305–3330 (2024) https://doi.org/10.1093/mnras/stae2249
Advance Access publication 2024 September 27

KiDS-1000 and DES-Y1 combined: cosmology from peak count statistics

Joachim Harnois-Déraps,[1★] Sven Heydenreich,[2] Benjamin Giblin,[3] Nicolas Martinet,[4] Tilman Tröster,[5] Marika Asgari,[1,6,7] Pierre Burger,[8,9,10] Tiago Castro,[11,12,13,14] Klaus Dolag,[15] Catherine Heymans,[3,16] Hendrik Hildebrandt,[16] Benjamin Joachimi[17] and Angus H. Wright[16]

[1] *School of Mathematics, Statistics and Physics, Newcastle University, Herschel Building, NE1 7RU Newcastle-upon-Tyne, UK*
[2] *Department of Astronomy and Astrophysics, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA*
[3] *Institute for Astronomy, University of Edinburgh, Blackford Hill, Edinburgh EH9 3HJ, UK*
[4] *Université d'Aix-Marseille, CNRS, CNES, LAM, F-13388 Marseille, France*
[5] *Institute for Particle Physics and Astrophysics, ETH Zürich, CH-8092 Zürich, Switzerland*
[6] *E.A Milne Centre, University of Hull, Cottingham Road, Hull HU6 7RX, UK*
[7] *Excellence for Data Science, AI, and Modelling (DAIM), University of Hull, Cottingham Road,Kingston-upon-Hull HU6 7RX, UK*
[8] *Waterloo Centre for Astrophysics, University of Waterloo, Waterloo, ON N2L 3G1, Canada*
[9] *Department of Physics and Astronomy, University of Waterloo, Waterloo, ON N2L 3G1, Canada*
[10] *Argelander-Institut für Astronomie, Auf dem Hügel 71, D-53121 Bonn, Germany*
[11] *INAF - Osservatorio Astronomico di Trieste, via Tiepolo 11, I-34131 Trieste, Italy*
[12] *INFN - Sezione di Trieste, Via Valerio 2, I-34100 Trieste, TS, Italy*
[13] *IFPU - Institute for Fundamental Physics of the Universe, via Beirut 2, I-34151 Trieste, Italy*
[14] *ICSC - Centro Nazionale di Ricerca in High Performance Computing, Big Data e Quantum Computing, Via Magnanelli 2, Bologna, I-40033, Italy*
[15] *Max-Planck-Institut fur Astrophysik, Karl-Schwarzschild Strasse 1, D-85748 Garching, Germany*
[16] *Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute (AIRUB), German Centre for Cosmological Lensing, D-44780 Bochum, Germany*
[17] *Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*

Accepted 2024 September 25. Received 2024 September 17; in original form 2024 May 27

ABSTRACT

We analyse the fourth data release of the Kilo Degree Survey (KiDS-1000) and extract cosmological parameter constraints based on the cosmic shear peak count statistics. Peaks are identified in aperture mass maps in which the filter is maximally sensitive to angular scales in the range 2–4 arcmin, probing deep into the non-linear regime of structure formation. We interpret our results with a simulation-based inference pipeline, sampling over a broad wCDM prior volume and marginalizing over uncertainties on shape calibration, photometric redshift distribution, intrinsic alignment, and baryonic feedback. Our measurements constrain the structure growth parameter and the amplitude of the non-linear intrinsic alignment model to $\Sigma_8 \equiv \sigma_8\,[\Omega_\mathrm{m}/0.3]^{0.60} = 0.765^{+0.030}_{-0.030}$ and $A_\mathrm{IA} = 0.71^{+0.42}_{-0.42}$, respectively, in agreement with previous KiDS-1000 results based on two-point shear statistics. These results are robust against modelling of the non-linear physics, different scale cuts, and selections of tomographic bins. The posterior is also consistent with that from the Dark Energy Survey Year-1 peak count analysis presented in Harnois-Déraps et al., and hence we jointly analyse both surveys with a common pipeline. We obtain $\Sigma_8^\mathrm{joint} \equiv \sigma_8\,[\Omega_\mathrm{m}/0.3]^{0.57} = 0.759^{+0.020}_{-0.017}$, in agreement with the *Planck*wCDM results. The shear-CMB tension on this parameter increases to 3.1σ when forcing $w = -1.0$, and to 4.1σ if comparing instead with $S_{8,\Lambda\mathrm{CDM}}^\mathrm{joint} = 0.736^{+0.016}_{-0.018}$, one of the tightest constraints to date on this quantity. Residual biases in the photometric redshifts of the DES-Y1 data and in the modelling of small scales physics could lower this tension, however it is robust against other systematics. Limits in the accuracy of our emulator prevent us from constraining Ω_m.

Key words: gravitational lensing: weak – methods: data analysis – methods: numerical – cosmological parameters – dark energy – dark matter.

1 INTRODUCTION

Cosmic shear cosmology has entered an era of high precision, with recent measurements from the Kilo Degree Survey[1] (KiDS), the Dark Energy Survey[2] (DES), and the Hyper Suprime-Cam Survey[3] (HSC) reaching a precision of a few per cent on parameters central to the standard model of cosmology (e.g. Asgari et al. 2021; Amon et al. 2022; van den Busch et al. 2022; Secco et al. 2022a; Dalal

★ E-mail: joachim.harnois-deraps@newcastle.ac.uk
[1] KiDS: kids.strw.leidenuniv.nl
[2] DES: www.darkenergysurvey.org.
[3] HSC: www.naoj.org/Projects/HSC.

et al. 2023; Li et al. 2023a, b). Based on the detection of weak correlations between the observed shapes of galaxies imparted by the foreground large-scale structure, cosmic shear is mostly sensitive to the structure growth parameter $S_8 \equiv \sigma_8\sqrt{\Omega_\mathrm{m}/0.3}$, a combination of the matter density parameter Ω_m and of the amplitude of the linear matter power spectrum smoothed on spheres of $8h^{-1}$Mpc, labelled as σ_8 (for lensing reviews, see e.g. Kilbinger 2015; Mandelbaum 2018). These Stage-III lensing surveys have been steadily improving the data quality and the analysis methods, in preparation for the next generation of cosmic shear experiments such as the Rubin observatory[4] (Ivezić et al. 2019), *Euclid*[5] (Laureijs et al. 2011), and the *Nancy Grace Roman space telescope*[6] (Akeson et al. 2019).

Despite the large effort that is being invested by international collaborations in constructing accurate lensing catalogues of hundreds of millions of galaxies, it is not entirely clear how to best analyse these vast data, striking an optimal compromise between accuracy and precision. To date the shear two-point (2pt) functions are still regarded as the baseline summary statistics, having been tested for over a decade and achieving an unmatched level of understanding and control in all aspects of the analysis, including measurements tools (e.g. TREECORR and NAMASTER, see Jarvis, Bernstein & Jain 2004; Alonso et al. 2019), theoretical predictions (e.g. Kilbinger et al. 2017), and the impact of systematics (see e.g. Mandelbaum 2018). The main drawback from these statistics is that they completely disregard the non-Gaussian information that is stored in the non-linear matter field, more precisely in the coupling between the phases of distinct Fourier modes, without which the cosmic web would look like a Gaussian random field. This is obviously sub-optimal, and this waste of information will be aggravated in the upcoming cosmic shear experiments. Accessing this non-Gaussian information is an active field of research: an array of novel weak lensing statistics are being developed specifically to utilize this complementary small-scale information. These new methods are reaching a level of maturity that makes them competitive at analysing existing cosmic shear data, carefully balancing the precision versus accuracy metric. Recent progress is largely due to the radically improved modelling of the signal, thanks to the increased accuracy of cosmological N-body codes and the availability of supercomputers (see Angulo & Hahn 2022, for a recent review on N-body codes). Recent examples of these 'beyond-2pt' cosmic shear data analyses include the three-point function (Fu et al. 2014; Secco et al. 2022b; Burger et al. 2024), peak count statistics (Kacprzak et al. 2016; Martinet et al. 2018; Shan et al. 2018; Harnois-Déraps et al. 2021; Zürcher et al. 2022; Liu et al. 2023; Marques et al. 2024; Gatti et al. 2024a), density split statistics (Brouwer et al. 2018; Gruen et al. 2018; Burger et al. 2022), shear clipping (Giblin et al. 2018), persistent homology (Heydenreich et al. 2022), moments of convergence maps (van Waerbeke et al. 2013; Gatti et al. 2020), cumulative distribution functions (Anbajagane et al. 2023), likelihood-free inference (Jeffrey, Alsing & Lanusse 2021; Lin et al. 2023; Gatti et al. 2024b), or convolutional neural network inference (Fluri et al. 2019, 2022).

At the moment, these alternative methods exhibit a constraining power that is similar to that of two-point functions, which is not surprising given the noise levels of current lensing data, which make difficult the extraction of information stored in the noisy higher order moments. The situation will change drastically with the upcoming surveys, where the cosmic web itself will be detectable with lensing, at which point the non-Gaussian information will take on a larger proportion of the signal.

All forecasts are clear about this: joint cosmic shear analyses that combine two-point functions and any complementary probe improve the constraints on cosmological parameters even in presence of systematic uncertainties (e.g. Li et al. 2019; Schneider et al. 2019; Zürcher et al. 2020; Pyne & Joachimi 2021; Harnois-Déraps, Martinet & Reischke 2022; *Euclid* Collaboration: Ajani et al. 2023; Giblin, Cai & Harnois-Déraps 2023). The main difficulty in many of these methods comes from their accrued dependence on numerical simulations, which adds a significant computational overhead to the data analysis compared to those for which an analytical model exists. Typically, simulations are needed for modelling the cosmological signal, for modelling some of the systematics such as baryonic feedback or intrinsic alignments of galaxies, and for the estimation of the covariance matrix (although this is not always necessary, as demonstrated by the recent likelihood-free inference analyses mentioned above).

This paper contributes an important step to this effort: we carry out a cosmological analysis based on lensing peak statistics measured from the fourth data release of the Kilo Degree Survey (KiDS-1000 hereafter). We use the exact same data as those used in the two-point function analyses of Asgari et al. (2021, A21 hereafter), while ignoring for now other non-lensing KiDS galaxy catalogues designed for galaxy clustering analyses (Bilicki et al. 2021; Vakili et al. 2023). Our method finds peaks in aperture mass maps with an aperture filter designed for the extraction of small-scale structure, with maximal sensitivity to scales of less than 4 arcmin, as in Martinet et al. (2018, hereafter M18) and Harnois-Déraps et al. (2021, HD21). This contrasts with the recent peak count analysis of Zürcher et al. (2022), in which peaks are extracted from convergence maps with pixel resolution of about 7 arcmin. Both methods have their advantages and downsides, ours strongly focuses on small, non-linear scales, which, as demonstrated in HD21 and Martinet et al. (2021a), have a higher potential for complementarity with two-point functions. Finding a posterior that is statistically consistent with that from HD21, we combine both likelihoods and carry out a joint KiDS-1000 + DES DR1 data (DES-Y1 hereafter) peak count analysis, finding the tightest constraints on S_8 to date from peaks alone.

After describing the data and simulations in Section 2.1, we detail our measurement techniques and analysis pipeline in Section 3, and we present our mitigation strategy for the key systematic uncertainties in Section 4. We finally show our results in Section 5 and discuss our findings afterwards. Supplementary material is provided in the Appendices, including a thorough discussion of B-modes (in Appendix A), supplementary pipeline validation tests (in Appendix B), and a detailed discussion on goodness-of-fit for noisy covariance matrices (in Appendix C).

2 DATA AND SIMULATIONS

We present in this section the survey data and the various simulation suites that are used for the cosmological analysis.

2.1 KiDS-1000 data

The Kilo Degree Survey (Kuijken et al. 2015) is a multiband photometric galaxy survey explicitly designed for weak lensing cosmology. Carried out at the European Southern Observatory by the

[4] LSST: www.lsst.org.

[5] *Euclid*: www.euclid-ec.org.

[6] *Roman Space Telescope*: roman.gsfc.nasa.gov

VST-OmegaCAM, we analyse here the public[7] fourth data release (Kuijken et al. 2019). The observation conditions are of exceptional quality, with a mean seeing of 0.7 arcsec in the r-band, used for shape measurements. The photometric redshifts are obtained from a combination of nine optical and infrared bands ($ugriZYJHK_s$, see Wright et al. 2020), thanks to the observations of the companion VIKING survey (VISTA Kilo-degree INfrared Galaxy; Edge et al. 2013). The galaxies selected in this analysis exactly match those used in the cosmic shear two-point function analyses of A21 and van den Busch et al. (2022), covering an effective area of 777.4 deg^2.

The KiDS DR4 data are reduced with the THELI (Erben et al. 2013) and Astro-WISE (Begeman et al. 2013) pipelines, following which the shear is inferred from lens*fit* (Miller et al. 2013; Fenech Conti et al. 2017). Shear additive and multiplicative biases (c- and m-corrections) are measured to a high accuracy (Giblin et al. 2021), where it is shown via a series of null tests that known residual systematics in the shear measurement could lead to no more than a 0.1σ shift in the structure growth parameter $S_8 \equiv \sigma_8\sqrt{\Omega_{\rm m}/0.3}$, the composite quantity that is best measured by cosmic shear. Note that strictly speaking, the results from the tests carried out in Giblin et al. (2021) are only shown to hold for two-point cosmic shear statistics.

Following A21, we split the full DR4 galaxies in five tomographic bins according to their individual best-fitting redshift $z_{\rm B}$ as measured by BPZ (Benítez 2000), with bin edges set to [0.1, 0.3, 0.5, 0.7, 0.9, and 1.2]. The tomographic redshift distributions, $n^a(z)$, are estimated via self-organizing maps (SOM, see Wright et al. 2020), which group galaxies based on their nine-band photometric properties and assign redshifts based on similar studies made on spectroscopic samples; galaxies for which no match is found are rejected. We further reject galaxies for which the SOM redshift catastrophically differs from the initial $z_{\rm B}$, resulting in the so-called 'Gold Sample' introduced in Hildebrandt et al. (2021) and used in the subsequent KiDS-1000 cosmic shear analyses mentioned above. As detailed in A21, the means and the error of the SOM redshift distributions are calibrated on KiDS-like mock data constructed from the MICE2 simulations (Fosalba et al. 2015; van den Busch et al. 2020) and accounted for during the inference stage of our analysis. The redshift accuracy is excellent due to the nine-band photometry, which helps breaking degeneracies in the galaxy spectral energy distributions: at worst, the difference between the mean redshift and that estimated from the matched spectroscopic sample is $z_{\rm est} - z_{\rm true} = 0.013 \pm 0.0118$, making this a subdominant source of uncertainty in our measurement. Note that Hildebrandt et al. (2021) further show that the SOM redshift distributions are fully consistent with independent estimates based on clustering cross-correlations with spectroscopic reference samples, providing extra robustness to the method. Fig. 1 shows the redshift distributions estimated in the five tomographic bins, along with the variations on these distributions allowed within our photometric uncertainty.

The SOM selection and the shear inference pipelines are both repeated on KiDS-like image simulations (Kannawadi et al. 2019), from which a relation between apparent size, magnitude, and the observed galaxy shape is used to calibrate the inferred lens*fit* shear.[8] Whereas previous cosmological analyses use a single m-calibration factor per tomographic bin, the aperture mass map statistics exploited in this paper are subject to local variations in the noise levels and seeing conditions, and we therefore use the above-mentioned relation to extract a shear calibration per object, m_a. This is not necessary, but allows us to capture possible correlations between the m-correction and the lens*fit* weights. These are inevitably noisier than the average over the full tomographic bins, but a large fraction of this noise cancels within our aperture mass map calculations as well, while providing optimal estimates of the local noise contribution (M18). Let us recall that this calibration corrects for known residual biases such as shape detection biases (Fenech Conti et al. 2017; Kannawadi et al. 2019) or blending of the images of galaxies (Hoekstra et al. 2015). While we apply the m-correction per object, the averaged multiplicative biases per redshift bin used in A21 enter our analysis at the inference level in the form of nuisance parameters over which we marginalize. Table 1 summarizes the survey properties relevant to our analysis.

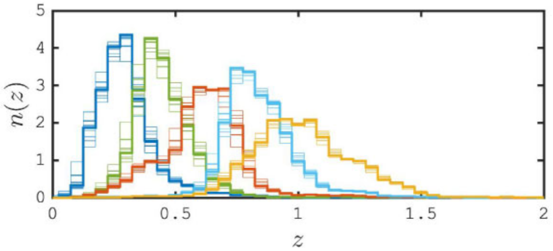

Figure 1. Tomographic redshift distribution of the KiDS-1000 data. The thinner lines represent the effect of photometric uncertainty on these distributions, characterized by $n^a(z) \rightarrow n^a(z + \Delta z_a)$, with Δz_a sampled 10 times from Gaussian distributions with widths listed in Table 1. All shifted $n^a(z)$ are then rebinned with the same z bins.

Table 1. Main properties of the KiDS-1000 data used in this work. The gold sample redshift selection based on $z_{\rm B}$ is identical to that presented in Hildebrandt et al. (2021). The effective number densities are listed in the second column, in gal arcmin^{-2}. The shape noise (per component) listed in the third column reflects the dispersion measured in the observed galaxy shapes, as documented in Giblin et al. (2021), while the fifth column shows the mean shape calibration coefficients. The redshift bias and errors listed in the fourth column are estimated from the SOM method in Hildebrandt et al. (2021), while the last column shows the additive $c_{1/2}$ terms, which has an uncertainty of 0.23×10^{-3} (Giblin et al. 2021).

tomo	$n_{\rm eff}$	σ_ϵ	$z_{\rm est} - z_{\rm true}$	m	$(c_1, c_2) \times 10^3$
bin1	0.62	0.27	0.000 ± 0.0106	-0.009 ± 0.019	(0.295, 0.156)
bin2	1.18	0.26	0.002 ± 0.0113	-0.011 ± 0.020	(0.004, 0.621)
bin3	1.85	0.27	0.013 ± 0.0118	-0.015 ± 0.017	(0.052, 0.728)
bin4	1.26	0.25	0.011 ± 0.0087	0.002 ± 0.012	(−0.360, 0.948)
bin5	1.31	0.27	-0.006 ± 0.0097	0.007 ± 0.010	(−1.363, 1.155)

2.2 DES-Y1 data

The DES-Y1 measurements is based on the public year-1 data release from the Dark Energy Survey Collaboration (Abbott et al. 2018), with source galaxy selections that exactly follow the main cosmic shear results described in Troxel et al. (2018). The lensing catalogue consists of 26 million galaxies covering a footprint of 1320 deg^2 with a galaxy density of 5.07 gal arcmin^{-2}. The per-galaxy shear signal is inferred with the METACALIBRATION method (Sheldon & Huff 2017). Every galaxy is assigned to one of the four tomographic bins based on the photometric redshift posteriors estimated from the the *griz* flux measurements, as detailed in Hoyle et al. (2018). Following Troxel

[7] KiDS-1000 data: http://kids.strw.leidenuniv.nl/DR4.

[8] A KiDS-1000 re-analysis has been presented in Li et al. (2023b) after correcting for an anisotropic error in the *lens*fit likelihood sampler. This error has not been corrected here, but their study shows the correction has a negligible impact on the inferred cosmology.

et al. (2018), the mean and uncertainty on the shear multiplicative calibration are given by $m_a = 0.012 \pm 0.023$.

Whereas the original DES-Y1 results estimated the tomographic $n^a(z)$ from a Bayesian photometric redshift analysis calibrated on the COSMOS2015 field (Laigle et al. 2016), the HD21 reanalysis instead opted for $n^a(z)$ estimates based on a direct reweighted calibration of matched spectroscopic data (Lima et al. 2008, DIR hereafter), following the DES-Y1 reanalyses of Joudaki et al. (2020) and Asgari et al. (2020). The uncertainty on the DIR mean redshift distributions is Δz_a = [0.008, 0.014, 0.011, and 0.009] for redshift bins $a = 1...4$, respectively. Both methods have their pros and cons. The calibration with COSMOS is by design based on a complete sample but suffers from imperfect redshifts and sampling variance (see e.g. Alarcon et al. 2021). In contrast, the spec-z samples used for the DIR method have (close to) perfect redshifts but are incomplete and not representative of the source sample, which is alleviated by the reweighting, but often cannot be fully eliminated (see Gruen & Brimioulle 2017). Importantly, the DIR$n^a(z)$ favours S_8 values that are smaller by $\Delta S_8 = 0.03$ compared to the COSMOS-calibrated $n(z)$, which is a 0.8σ shift (Joudaki et al. 2020).

2.3 Simulations

As mentioned in the introduction, the accuracy of simulation-based inference pipelines fully depends on the quality of the numerical simulations it is calibrated on. The same way 2pt analyses must carefully understand the scales, cosmologies, and redshifts that are well captured by their model, it is critical for our peak count analysis to identify the range of validity of our training simulations. The additional complexity here is that no simulation suite serves all purposes, and therefore we must carefully investigate, for all of them separately, the accuracy and limits of the measurements and how these impact the peak count statistics. The simulations used in this work are in many aspects identical to those presented in HD21, which we refer to for further details. Specifically:

(i) the cosmological dependence of the peak count statistics is calibrated on the wCDM *cosmo*-SLICS N-body simulations introduced in Harnois-Déraps, Giblin & Joachimi (2019). They sample a wide volume in S_8, $\Omega_{\rm m}$, w_0, and h with 25 points arranged in a Latin hypercube (plus one ΛCDM point), each evolved with a pair of N-body simulations designed to suppress sample variance in 2pt functions, then ray-traced in ten light cones of 100 deg^2 (10 000 deg^2 in total area). These form our *cosmology training set*, and resolve the non-linear physics to better than 2 per cent up to k-modes of 2.0 h^{-1}Mpc, when compared to the Cosmic Emulator (Heitmann et al. 2014). Smaller scales gradually lose precision, affecting mainly their ability to resolve substructure in most massive objects. The exact impact of this loss on weak lensing peak counts is investigated in HD21 with a separate set of simulations ran with a much higher force resolution, where it is found that this leads to at most a 1 per cent loss of the highest peaks, which is largely subdominant compared to both baryonic physics and statistical errors. We revisit this in Section 4 (see also point iv);

(ii) the covariance matrix that captures the sample variance is estimated from 124 fully independent SLICS N-body simulations described in Harnois-Déraps & van Waerbeke (2015). These are evolved from independent initial conditions at a fixed cosmology, and make our *covariance training set*. They resolve the same non-linear physics as the *cosmo*-SLICS, and are shown in Harnois-Déraps et al. (2019) and HD21 to produce marginalized errors on cosmological parameters that are fully consistent with those obtained with an analytical calculation, when analysing 2pt statistics. Burger et al. (2022) further show in the context of density-split statistics that a covariance matrix estimated from the SLICS or from a much larger number of log-normal FLASK mocks (Xavier, Abdalla & Joachimi 2016) produce fully consistent results, as expected for these mildly non-linear statistics. We further increase the effective number of covariance mocks by randomly rotating 10 times the shape noise components. This works particularly well given that the peak statistics is currently shape-noise dominated:[9] while the expectation value of standard 2pt statistics does not depend on the noise (only their variance does), shape noise affects both the signal and covariance of map-based statistics (see Appendix D of Heydenreich, Brück & Harnois-Déraps 2021);

(iii) for the KiDS-1000 analysis, the impact of galaxy intrinsic alignments is measured from the IA-infused lensing simulations described in Harnois-Déraps et al. (2022). These are also constructed from the *cosmo*-SLICS and therefore resolve the same physical scales. This *IA training set* assumes a linear coupling between the projected non-linear tidal field and the intrinsic ellipticity of every galaxy, and is therefore physically modelling the non-linear linear alignment model of Bridle & King (2007) without explicit redshift nor luminosity dependence. It is expected that this effective IA model does not fully capture the alignment signal, and that a more physical model such as the tidal alignment and torquing model (Blazek et al. 2019) or the halo-model of Fortuna et al. (2021) would provide a more accurate description, however current cosmic shear surveys do not have the statistical power to constrain parameters beyond the simpler NLA model (Secco et al. 2022a), which is therefore deemed sufficient for the current analysis. The IA infusion process has been shown in Harnois-Déraps et al. (2022) to accurately reproduce the NLA predictions for the 2pt correlation function down to scales of a few arcmin, beyond which the NLA is expected to fail in a manner that is undetectable in the current data. Burger et al. (2024) further show that these same simulations agree with the IA modelling of three-point shear statistics. The model fails at scales that correspond to high overdensities in our simulations, which contribute to lensing peaks that are excluded from our analysis. We infuse different levels of IA and marginalize over these choices in the end, as described in Section 4. for the DES-Y1 analysis, IA are included with a non-linear halo-based model, see HD21 for details;

(iv) limits in the force resolution of the *cosmo*-SLICS are bound to impact the weak lensing statistics in a manner that is not always predictable. We assess this with the SLICS-HR suite (Harnois-Déraps & van Waerbeke 2015), a high-resolution version of the SLICS light cones recently used in a combined lensing-clustering cosmological analysis (Duncan et al. 2022). The SLICS-HR consist of ten independent 10 × 10 deg^2 catalogues that are run at the same cosmology and with the same particle count and volume as the SLICS, but the N-body force accuracy has been increased such as to resolve k-modes up to 10 h^{-1}Mpc. We use these to validate the full inference pipeline in Section 4.7, acting as our *validation set*;

(v) the impact of baryon feedback is estimated with the *magneticum* hydrodynamical simulations,[10] forming our *baryons training set*. These have been shown to reproduce a number of key observations relevant to weak lensing studies (Castro et al. 2018), and notably the feedback on the matter distribution closely matches that

[9] The average shape noise contribution, computed from the scatter between the 10 noise realizations for a fixed underlying simulation, takes up about 90 per cent of the total error budget, 95 per cent for the auto-bins.

[10] *Magneticum* simulations: www.magneticum.org.

of the BAHAMAS (McCarthy et al. 2017), another suite of hydrosimulations with independent prescriptions for their subgrid physics. The training set consists of ten 10 × 10 deg^2*pseudo*-independent light cones extracted[11] from full hydrodynamical simulations, and another 10 light cones extracted from dark matter-only sister simulations, evolved from the same initial conditions (more details on the used simulations can be found in Martinet et al. 2021b). There is a large uncertainty on the exact impact of baryonic physics on the matter distribution (and therefore on our lensing statistics), which we account for by linearly scaling the relative baryonic bias with a nuisance parameter, $b_{\rm bary}$, which we marginalize over at the inference stage.[12]

(vi) different N-body codes and ray-tracing methods, even at fixed cosmology, will have a residual impact on the peaks statistics (Hilbert et al. 2020). We explore these numerical systematics with the public full-sky weak lensing simulations from Takahashi et al. (2017, T17 hereafter),[13] post-processed into KiDS-1000 mock data (North and South patches) as in Burger et al. (2024). The T17 simulations follow the non-linear evolution of 2048^3 particles in a series of nested cosmological volumes with side length starting at $L = 450h^{-1}$Mpc at low redshift, then increasing at higher redshifts. These result in 108 pseudo-independent full-sky lensing maps, seven of which are used in this work, with flat ΛCDM cosmological parameters set to ($\Omega_{\rm m}$, $\Omega_{\rm b}$, σ_8, h, $n_{\rm s}$, w_0) =(0.279, 0.046, 0.82, 0.7, 0.97, -1.0).

2.3.1 *Assembling mock surveys*

Most of these simulations have been introduced in HD21, in Heydenreich et al. (2022), in Burger et al. (2022), and in the references listed in the previous section; we encourage the interested reader to consult these for a more complete technical description. To summarize some of the key properties, all of the abovementioned simulations are organized in light cones of 100 deg^2 each, populated with galaxy samples that match the tomographic $n^a(z)$ distributions, number densities, and shape noise levels of the KiDS-1000 Gold Sample and DES-Y1 data. Except for the IA-infused simulations, the galaxy positions, the amplitude of their ellipticities $|\boldsymbol{\epsilon}_{\rm data}|$ and their multiplicative shear calibration factors m_a are exactly reproduced in each of the mock survey realizations (i.e. in the *cosmology, covariance, validation* and *baryon training sets*). To achieve this, the KiDS-1000 data are split into 18 tiles that each fit within 100 deg^2 regions, as depicted in Fig. 2, and the shear and convergence from every simulation is repeated across them, interpolated at the local galaxy positions. These tiles are analysed separately and combined only at the level of the summary statistics, ensuring that cross-tile correlations that exist in the data but not in the simulations are explicitly ignored. This effect is minor for localized non-Gaussian probes such as peak statistics, but is critical for e.g. shear 2pt functions. The shear and convergence are interpolated from the underlying simulations at the position of every galaxy, infused with the m_a from the data, then combined with the (randomly rotated) observed ellipticity following:

$$\boldsymbol{\epsilon}_{\rm mock} = \frac{\boldsymbol{\epsilon}^{\rm rand}_{\rm data} + \boldsymbol{g}}{1 + \boldsymbol{\epsilon}^{\rm rand}_{\rm data}\boldsymbol{g}^*} \, . \qquad (1)$$

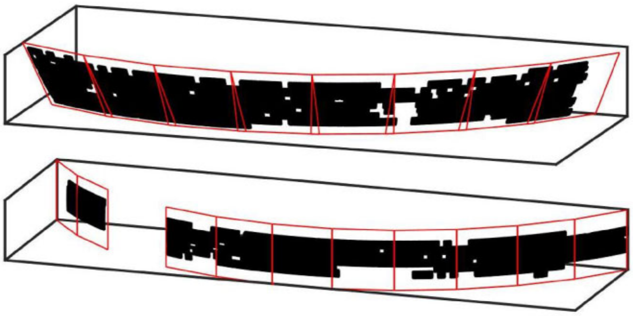

Figure 2. Tiling strategy adopted to pave the full KiDS-1000 data with flat-sky 10 × 10 deg^2 simulations (squares). Some of the tiles slightly overlap due to the sky curvature, in which case the data is split at the mean Dec in overlapping regions.

In the above expressions, bold-font symbols are spin-2 complex quantities and $\boldsymbol{g}$ is the m-biased simulated reduced shear. As described in HD21, this involves rotating each tile at the equator, which preserves the relative positions of galaxies but modifies their ellipticities, defined with respect to the North pole.

We repeat this construction for all light cones of the *cosmology training set*, the *covariance training set*, the *baryon training set*, and the *validation set*. Additionally, the uncertainty in the photometric redshifts is forward-modelled with a further 10 full survey realizations computed at the fiducial cosmology, in which the $n(z)$ is shifted by small amounts (details provided in Section 4). In total, this results in 414 simulated mosaic surveys that we analyse in preparation for the inference stage, with the majority (260) contributing to the *cosmology training set*. Each mock further contains 10 random rotations of $\boldsymbol{\epsilon}_{\rm data}$, to improve convergence of the signal.[14]

The *IA training set* are treated slightly differently, since for these the positions of the mock galaxies must be sampled from the simulated overdensity maps or halo catalogues, which do not correlate with the positions in the data (see Harnois-Déraps et al. 2022, for more details). The mosaic survey tiling is therefore not possible, so we use instead five light cones per IA model and explore four alignment strengths in KiDS, and one model in DES. Although these represent a lesser total area than the real data, their sole purpose is to capture the relative impact of IA on the signal, computed from ratios in which the sample variance cancels by design. We use 4 × 5 × 100 deg^2 of training data, which is enough to capture this.

The simulations are free of additive biases by construction, however Giblin et al. (2021) measures residual additive terms $c_{1/2}$ in the KiDS-1000 cosmic shear catalogues, caused by the shape measurement method itself. These are reported in Table 1 and subtracted from the observed ellipticities when analysing the real data. We follow Troxel et al. (2018) by not accounting for any low-level additive terms in the DES Y1 catalogue. The multiplicative biases are not easily removed from the data, hence we instead infuse the mocks with the m_a terms per object, and treat thereafter data and simulations on equal footings.

[11] The *magneticum* light cones were built with the public SLICER code: https://github.com/TiagoBsCastro/SLICER.

[12] Note that this parameter is not to be confused with $A_{\rm bary}$ used in A21 (see their table 2), which specifically relates to one of the free parameters entering their HMCODE halo model.

[13] T17: http://cosmo.phys.hirosaki-u.ac.jp/takahasi/allsky_raytracing/.

[14] For peak statistics, removing the shape noise from simulated data changes both the mean of the signal and the covariance, whereas for shear 2PCF, the mean of the signal is unchanged. For this reason, the best way to achieve convergence on the mean peaks signal is by averaging over multiple noise realizations.

3 METHODS

3.1 Aperture mass map statistics

There exists a number of methods for identifying and counting lensing peaks, including finding maxima on convergence maps (Li et al. 2019), on wavelet-transformed maps (Ajani et al. 2020) or on aperture mass maps (Schneider 1996). We here opted for the aperture mass maps for the following reasons: as argued in M18, this statistics is immune to masking-induced biases and strong B-mode leakage common to methods based on reconstruction of convergence maps, plus it benefits from a local estimation of the shape and Poisson noise, yielding more accurate signal-to-noise maps.

Specifically, we cover each of the 18 tiles with a 2D grid with a pixel size of 0.59 arcmin. We next reconstruct the mass inside an aperture filter Q centred on each pixel, at position $\boldsymbol{\theta}$ on the sky, from the sum of all tangential ellipticities $\epsilon_{a,\rm t}$ contained therein as:

$$M_{\rm ap}(\boldsymbol{\theta}) = \frac{1}{n_{\rm gal}(\boldsymbol{\theta})\sum_a w_a(1+m_a)} \sum_a w_a \epsilon_{a,{\rm t}}(\boldsymbol{\theta},\boldsymbol{\theta}_a) Q(|\boldsymbol{\theta}-\boldsymbol{\theta}_a|, \theta_{\rm ap}, x_c). \tag{2}$$

The tangential ellipticity about $\boldsymbol{\theta}$ is computed as $\epsilon_{a,{\rm t}}(\boldsymbol{\theta},\boldsymbol{\theta}_a) = -[\epsilon_1(\boldsymbol{\theta}_a)\ \cos(2\phi(\boldsymbol{\theta},\boldsymbol{\theta}_a)) + \epsilon_2(\boldsymbol{\theta}_a)\ \sin(2\phi(\boldsymbol{\theta},\boldsymbol{\theta}_a))]$, where $\boldsymbol{\theta}_a$ is the position of galaxy a and $\phi(\boldsymbol{\theta},\boldsymbol{\theta}_a)$ is the angle between both coordinates. The sum runs over all galaxies in the aperture, and $n_{\rm gal}(\boldsymbol{\theta})$ is the local galaxy density in the filter when centred at $\boldsymbol{\theta}$. As in M18 and HD21, our filter $Q(\theta, \theta_{\rm ap}, x_c)$, abridged to $Q(\theta)$, matches that of Schirmer et al. (2007), which is optimized for efficiently detecting NFW haloes:

$$Q(x) = \frac{\tanh(x/x_c)}{x/x_c}\left[1+\exp(6-150x)+\exp(-47+50x)\right]^{-1}. \tag{3}$$

In the above expression, we use the standard value of $x_c = 0.15$, while $x = \theta/\theta_{\rm ap}$, with θ the distance to the filter centre. We additionally use the same filter size, set to $\theta_{\rm ap} = 12.5$ arcmin, which is shown in M18 to better detect the cosmological signal over other filter sizes in KiDS data. We compute equation (2) at every pixel location to construct our signal map. The variance about this map is computed at every pixel location with:

$$\sigma^2_{\rm ap}(\boldsymbol{\theta}) = \frac{1}{2n^2_{\rm gal}(\boldsymbol{\theta})\left[\sum_a w_a\right]^2}\sum_a w_a^2|\boldsymbol{\epsilon}_a|^2 Q^2(|\boldsymbol{\theta}-\boldsymbol{\theta}_a|)\,, \tag{4}$$

where again the sum runs over all galaxies in the filter. The m-calibration estimated from the image simulations of Kannawadi et al. (2019) is meant to correct the inferred shear, not the ellipticity, which explains why it appears in the denominator of equation (2) but not in that of equation (4), which describes the noise map. Finally, we take the ratio between equation (2) and the square root of equation (4) at every pixel location to construct our signal-to-noise maps, $\mathcal{S}/\mathcal{N}(\boldsymbol{\theta}) \equiv M_{\rm ap}(\boldsymbol{\theta})/\sqrt{\sigma^2_{\rm ap}(\boldsymbol{\theta})}$, from which we identify peaks as local maxima and record their $\mathcal{S}/\mathcal{N}$-values. We repeat this process for the 10 realizations of random rotations and report the average, except for the *covariance training set*, for which we do not take the average; instead, each noise realization leads to an estimate of the covariance matrix, of which we take the average in the end.

As detailed in HD21, masking is dealt with naturally in aperture mass statistics, and no special treatment needs to be enforced as long as data and simulations are masked and analysed the same way. This is achieved by fixing galaxy positions in the simulations to that of the observed data, which ensures the impact of the mask is identical. In our case, we decided nevertheless to act upon masked pixels. These are identified from the galaxy catalogues as regions with an aperture galaxy density that is either critically low or null, then removed from the final $\mathcal{S}/\mathcal{N}(\boldsymbol{\theta})$ maps.

It has been shown that some additional information can be extracted by combining the peak count statistics measured from multiple filter sizes (e.g. Zürcher et al. 2022; Giblin et al. 2023), however M18 shows that this gain is mild for Stage III surveys. We therefore opted for a single-scale analysis here, but intend to revisit this in the future.

3.2 Tomography and selection

Tomographic decomposition of the lensing data allows us to probe the redshift evolution of the large-scale structures, which is largely driven by $\Omega_{\rm m}$ and w_0 via their impact on the growth of perturbations. A direct consequence of the improved sensitivity to these is a gain in precision in S_8, arising from degeneracy breaking. This decomposition is different for the KiDS and DES surveys, which we detail here.

3.2.1 *KiDS-1000*

From the five KiDS tomographic bins, we include both the auto- and the cross-redshift measurements, as first defined in Martinet et al. (2021a). To be specific, peaks are identified from the individual tomographic galaxy catalogues (the 'auto' redshift bins 1, 2, 3, 4, 5), from every possible combination of bin pairs (1∪2, 1∪3, 1∪4, 1∪5... 4∪5), triplets (1∪2∪3, 1∪2∪4, 1∪2∪5...), quadruplets (1∪2 ∪ 3∪4, 1∪2 ∪ 4∪5...), and quintets (i.e. no tomography). As shown in HD21, Zürcher et al. (2022) and Heydenreich et al. (2022), these 'cross-tomographic' catalogues contain a significant amount of additional information that is not contained within the 'auto' case. The tomographic peak function is presented in Fig. 3, showing in the different panels the 30 different redshift bin combinations. For each case we overlay the predictions from the *cosmology training set* in colour with the data measurements in black; the error bars are obtained from the *covariance training set*. A similar measurement is presented in Fig. B1, where the data are replaced by the mean over our *baryons training set*. In these figures, we have subtracted the peak function measured from pure shape noise fields, $N_{\rm peaks}^{\rm noise}$, to better highlight the cosmological dependence of the signal.

In all cases, we measure the peak function in $\mathcal{S}/\mathcal{N}$ bins of width 0.5 in the range [-2.5, 4.0], for a total of 13 bins per subpanel and 390 elements in total. The motivation behind this initial choice of range is driven by a number of requirements, notably that of having a large number of peaks per bin to ensure the data is Gaussian-distributed (with our selection, every bin has at least 200 objects, while bins outside this range have far fewer objects). Additionally, our analysis has strict requirements on the modelling precision and on the level of contamination by residual systematic effects, resulting in this bin selection being in fact 'aggressive'. We expand on this in Section 4, where we argue that instead the range $[-1.0 < \mathcal{S}/\mathcal{N} < 3.0]$ is a better choice with lower modelling errors, forming a 'clean' data vector of 7 × 30 = 210 elements in total that is used for the main cosmological analysis.

3.2.2 *DES-Y1*

Following HD21, our DES peak count analysis includes the auto- and cross-redshift measurements up to pairs of tomographic bins, for a total of 10 bin combinations. The peak function is measured in 12 $\mathcal{S}/\mathcal{N}$ bins in the range $[0.0 < \mathcal{S}/\mathcal{N} < 4.0]$, forming a data vector

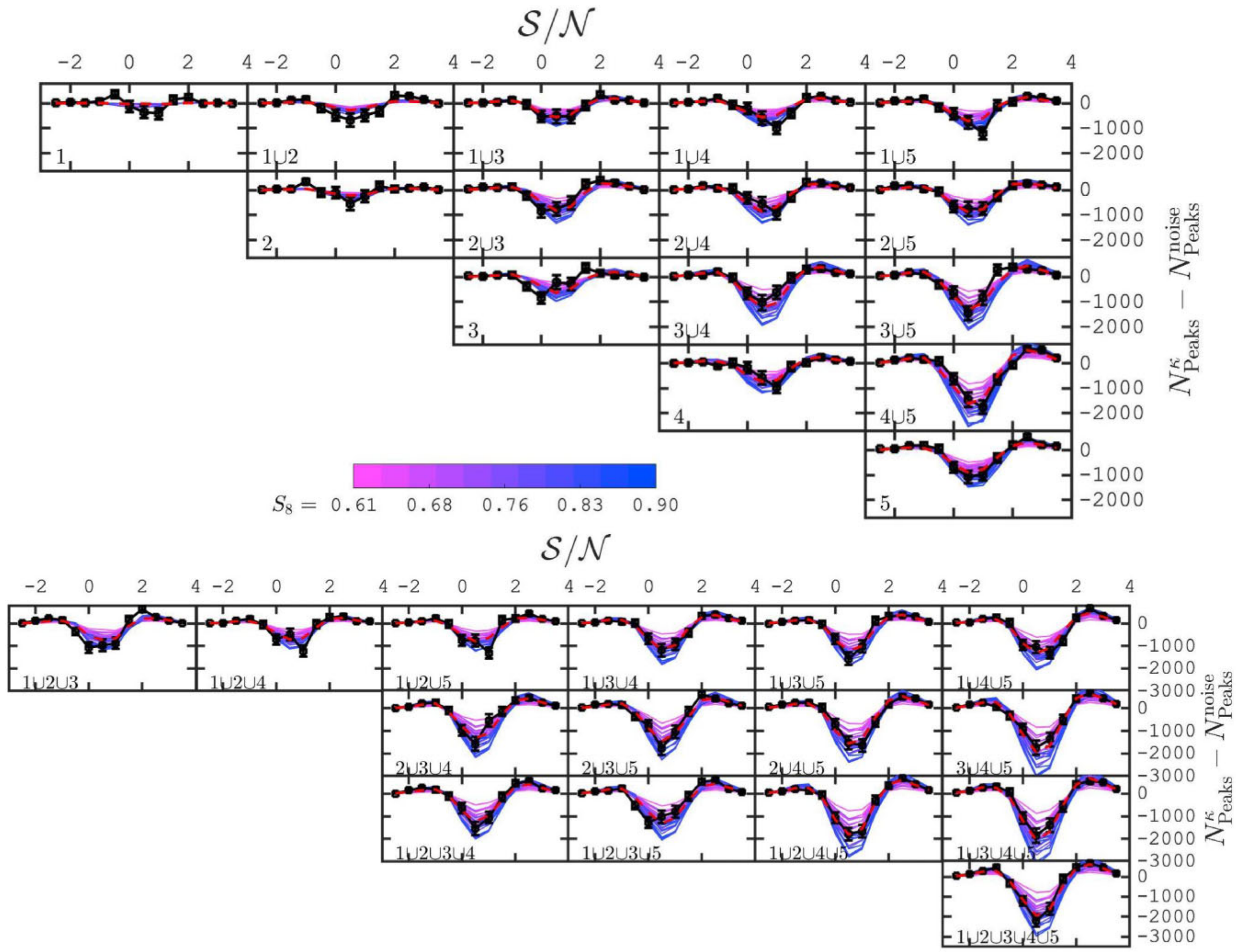

Figure 3. Tomographic weak lensing peak function $N^{\kappa}_{\rm peaks}(\mathcal{S}/\mathcal{N})$ measured in the KiDS-1000 data (black squares) and in the *cosmology training set* simulations, colour-coded by their S_8 value. The pure noise signal $N^{\rm noise}_{\rm peaks}$ has been removed to better highlight the variations with respect to cosmology. The panels show the results from different combinations of tomographic bins, in which the red dashed lines represents the best-fitting model inferred from our fiducial analysis, see Section 5.

with 120 elements. Although these details differ compared to the KiDS-1000 case described above, it is shown in HD21 to be accurate and competitive.

3.3 Analysis pipelines

Our cosmological inference pipeline heavily builds on the methods presented in HD21 and Heydenreich et al. (2022), which we briefly overview here. First, we model the peak function by training a Gaussian process regression[15] emulator (GPR) on the measurements from the *cosmology training set*, after averaging over the 10 noise realizations. The GPR can subsequently produce $N^{\kappa}_{\rm peaks}$ predictions within a fraction of a second everywhere inside the parameter volume covered by the *cosmo*-SLICS. This therefore determines the prior ranges over $\Omega_{\rm m}$, S_8, w_0, and h, which we report in Table 2.

Secondly, we must estimate the covariance matrix, which captures the correlation between the elements of our data vector, central to the error propagation. As mentioned before, the *covariance training set* consists of 124 full survey realizations, each duplicated 10 times with a distinct shape noise realization, producing 1240 *pseudo*-independent data vectors from which our covariance matrix C is extracted.[16] Since shape noise is added at the galaxy level, cross-redshift bins are correlated with the autobins. We show in Fig. 4 the cross-correlation coefficient matrix, defined as $C_{ij}/\sqrt{C_{ii}C_{jj}}$, which better highlights the correlations between the negative and positive peaks in each of the tomographic block. Also visible is the significant amount of correlation (and anticorrelation) present in the off-diagonal component. This matrix contains at most 390^2 elements and is thus invertible (since $390 < 1240$, see Hartlap, Simon & Schneider 2007), a criteria that is also naturally satisfied by the 'clean' KiDS-1000 data vector, which contains only 210 entries, and by the DES-Y1 data vector, which contains 120.

Having our model and covariance matrix, we are now in a position to evaluate the likelihood $\mathcal{L}$ of the model $\boldsymbol{x}(\boldsymbol{\pi})$ with parameters $\boldsymbol{\pi}$, given the data vector $\boldsymbol{d}$. We use the Sellentin & Heavens (2016) t-distribution likelihood, which is well suited for nearly Gaussian data

[15] We use the GPR toolkit provided by SCIKITLEARN (Pedregosa et al. 2011).

[16] In practice, we follow HD21 and estimate C from the average over 10 matrices, each computed from one of the noise realization.

Table 2. Priors used in the KiDS likelihood sampling. The ranges for the four cosmological parameters are determined by the cosmo-SLICS simulations, while the prescription from sampling the nuisance parameters describing the photometric redshifts Δz_a and intrinsic alignments $A_{\rm IA}$ are taken from Joachimi et al. (2021). In particular, the redshift parameters are correlated and drawn from a multivariate Gaussian distribution with means $\boldsymbol{\mu}$ taken from Table 1 (fourth column) and a covariance matrix $\mathbf{C_z}$ described in Section 4.3. The shear calibration parameters Δm_a are sampled from Gaussian priors centred on zero with a standard deviation (μ, σ) estimated in Giblin et al. (2021). The baryonic feedback parameter $b_{\rm bary}$ is used to scale the effect measured in the *baryon training set*.

Parameter	Range	Prior
Cosmology		
$\Omega_{\rm m}$	[0.1, 0.55]	Flat
S_8	[0.6, 0.9]	Flat
h	[0.6, 0.82]	Flat
w_0	[−2.0, −0.5]	Flat
Nuisance		
$\Delta z_a \times 10^2$	[−10, 10]	$\mathcal{G}(\boldsymbol{\mu}, \mathbf{C_z})$
$\Delta m_1 \times 10^2$	[−10, 10]	$\mathcal{G}(0.0, 1.9)$
$\Delta m_2 \times 10^2$	[−10, 10]	$\mathcal{G}(0.0, 2.0)$
$\Delta m_3 \times 10^2$	[−10, 10]	$\mathcal{G}(0.0, 1.7)$
$\Delta m_4 \times 10^2$	[−10, 10]	$\mathcal{G}(0.0, 1.2)$
$\Delta m_5 \times 10^2$	[−10, 10]	$\mathcal{G}(0.0, 1.0)$
Astrophysics		
$A_{\rm IA}$	[−5, 5]	Flat
$b_{\rm bary}$	[0, 2]	Flat

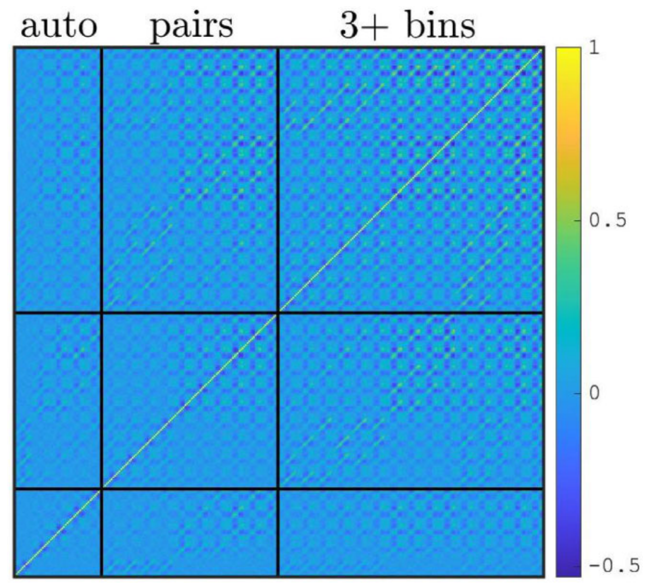

Figure 4. This figure highlights the correlations between the different elements of the KiDS-1000 data vector. From left to right, the 30 blocks show the correlation coefficients for the different redshift bin combinations, starting with singlets (i.e. autobins), pairs, triplets, quadruplet, and the no-tomographic case, with redshift increasing towards the right and the top of the figure.

vectors with simulation-based covariance matrices.[17] It is constructed as:

$$\ln\mathcal{L}(\boldsymbol{\pi}|\boldsymbol{d}) = \frac{N_{\rm sims}}{2}\ln\left[1+\chi^2/(N_{\rm sims}-1)\right] + {\rm const}, \qquad \text{with} \tag{5}$$

$$\chi^2 = [\boldsymbol{x}(\boldsymbol{\pi}) - \boldsymbol{d}]^{\rm T}{\rm C}^{-1}[\boldsymbol{x}(\boldsymbol{\pi}) - \boldsymbol{d}]. \tag{6}$$

In the above, $N_{\rm sims} = 1240$ is the number of realizations used to evaluate the covariance matrix C. The model depends on the four cosmological parameters $\Omega_{\rm m}$, S_8, w_0, and h, and on a set of 12 (9) astrophysical and nuisance parameters for KiDS (DES), which characterize the dependence of our signal on the systematic effects mentioned previously. This is an excellent approximation to the more general likelihood suggested by Percival et al. (2022) in our case. Finally, the posteriors are sampled both by the nested sampling algorithm MULTINEST (Feroz, Hobson & Bridges 2009) and by NAUTILUS (Lange 2023), implemented within COSMOSIS (Zuntz et al. 2015). While the latter sampler is more robust (Lange 2023), the former has been more widely used in the literature and is therefore useful to make fair comparisons with previous analyses. We report from these the mean and 68 per cent credible intervals computed from the 1D projected posteriors,[18] as well as the maximum a posteriori for some of our key results.

Since our likelihood function differs from the widely used multivariate Gaussian, the goodness-of-fit evaluation must be adapted accordingly. For Gaussian likelihoods, the $\chi^2_{\rm best-fit}$, estimated at the best-fitting parameters, is to a very good approximation sampling an underlying χ^2_ν probability distribution, which depends only on the number of degrees of freedom ν – this is only an approximation however, because of informative priors, non-linear modelling, and correlated error bars (see e.g. Joachimi et al. 2021). A good fit will have a $\chi^2_{\rm best-fit}$ close to the maximum of the χ^2_ν probability distribution, while a bad fit will land far in the tail, leading to a probability to exceed (PTE) that is smaller than our acceptance threshold, set to 0.01. For our Student-t distribution likelihood, we still assess the goodness-of-fit with PTE values, however the χ^2_ν curve needs to be modified (see Appendix C for details).

A few differences exist between the KiDS-1000 and DES-Y1 likelihoods which are worth highlighting here, as these influence the construction of our joint pipeline. First, the original DES-Y1 peak count analysis samples σ_8 instead of S_8; the latter is a better option as it exactly covers the training volume and is therefore adopted for our DES-Y1 re-analysis.

Secondly, the treatment of intrinsic alignments are simpler in the DES-Y1 analysis: the IA contribution is estimated from the alignments of dark matter haloes, which are assumed to fully correlate with the alignment of central galaxies. This non-linear prescription provides a single IA model that is then added to the predictions, without marginalization. As discussed in Section 5.2, not marginalizing over the IA in the KiDS analysis slightly underestimates the total error. This is likely less important in the DES-Y1 likelihood since the statistical error is larger.

This also connects with the third difference, which is that in the baseline DES-Y1 measurement, only the autotomographic redshift bins are included, in an attempt to avoid possible residual contamination from unmodelled IA in the cross-redshift bins. This turns out to be an overconservative data cut. Indeed, the recent DES-Y1 persistent homology cosmic shear analysis from Heydenreich et al. (2022) reveals that the constraints on S_8 are negligibly affected by these IA terms: they show that a full tomographic analysis including

[17] Using instead a multivariate Gaussian likelihood along with a Hartlap factor is less accurate, see Sellentin & Heavens (2016) for a full discussion.

[18] We refer to these intervals as 1σ regions, even though strictly speaking this notation should only apply to Gaussian posteriors.

Table 3. Priors used for sampling the nuisance parameters in the DES-Y1 peak statistics analysis. The sampling of photometric redshifts Δz_a and shear bias Δm_a nuisance parameters follows the original cosmic shear paper by Troxel et al. (2018). The baryonic feedback parameter $b_{\rm bary}$ is the same as in the KiDS-1000 likelihood, however there is no IA parameter here.

Parameter	Range	Prior
$\Delta z_1 \times 10^2$	$[-10, 10]$	$\mathcal{G}(0.1, 1.6)$
$\Delta z_2 \times 10^2$	$[-10, 10]$	$\mathcal{G}(1.9, 1.3)$
$\Delta z_3 \times 10^2$	$[-10, 10]$	$\mathcal{G}(0.9, 1.1)$
$\Delta z_4 \times 10^2$	$[-10, 10]$	$\mathcal{G}(1.8, 2.2)$
$\Delta m_a \times 10^2$	$[-10, 10]$	$\mathcal{G}(1.2, 2.3)$
Astrophysics		
$b_{\rm bary}$	[0, 2]	Flat

all cross-tomographic combinations shift the parameter by at most 0.3σ towards higher S_8 values, even when the inferred $A_{\rm IA}$ is as large as unity. Although their analysis is based on the different statistics (they use persistent homology instead of peak count), their results should hold here too, given that peaks are a subset of their data vectors. Therefore residual IA cannot play an important role in the DES-Y1 peak count analysis, justifying our choice to include the cross-redshift bins here (up-to-pairs, but not the triplets nor the quadruplets since these are not fully modelled yet for the DES-Y1 data).

A fourth difference in the likelihood concerns the treatment of the baryonic feedback: in HD21 the peak statistics are measured in the *magneticum* simulations to ensure that the selected elements from the data vectors are immune to unmodelled baryonic mechanisms, but no marginalization is included. This can potentially lead to a slightly overoptimistic precision on the DES likelihood compared to the KiDS-1000 likelihood, which includes marginalization over the $b_{\rm bary}$ parameter. We therefore decided to include in our joint analysis the same marginalization machinery for both the KiDS-1000 and the DES-Y1 pipelines. Moreover, we use a unique $b_{\rm bary}$ parameter to infuse baryonic feedback into both surveys, since these physical processes describe physics that affect the foreground matter distribution independently of survey-specific source selection. In total, the combined-survey analysis marginalizes over nine redshift bias parameters, nine shear bias parameters, one IA, and one baryon parameter. The sampling strategy of the DES-related parameters are listed in Table 3. Finally, given the absence of overlap between two survey footprints and the compatibility of the priors, the two likelihoods can be directly added at each evaluation point, without needing to consider cross-survey covariance.

Before running the analysis on the KiDS and DES data, we validated our pipelines on simulated data, as presented in the next section, and made no further modifications to thereafter. This method is not as strong as adopting a full blinding strategy at the catalogue level, however this avenue was not available anymore since many authors were already unblinded, having worked on previous cosmic shear analyses with the same data. In these conditions, our validation strategy is an excellent option to protect ourselves against confirmation bias.

4 SYSTEMATIC UNCERTAINTIES

As the amount of high quality lensing data keeps increasing, the statistical precision reaches unprecedented high levels, and consequently understanding and controlling the residual systematics in every segment of the data analyses has become one of the primary objectives and focus of development in the field of beyond-2pt statistics. We investigate here a number of such systematic effects that have been identified in the literature and mentioned earlier, including residual uncertainties related to interpolation in the modelling, shear calibration, photometric redshifts, astrophysics (intrinsic alignments and baryonic feedback), simulation-based covariance matrix, non-linear physics, source-lens coupling, and likelihood sampling strategies. Some of the systematics that are ignored in this work are those related to the effect of source blending, depth variations, PSF leakage, or the cosmology dependence of the IA signal. These will likely become important in the future, but can be safely omitted in current Stage-III lensing surveys (see HD21; Zürcher et al. 2022). Amongst those that we investigate here, many are shown to be subdominant or heavily suppressed by our range of $\mathcal{S}/\mathcal{N}$, while others are forward-modelled with nuisance parameters that are marginalized over in the likelihood analysis.

4.1 Modelling

The accuracy of the peak function modelling has two aspects to be considered: we must understand both how well the cosmology scaling is captured by the emulator, and whether any elements of our data vector are affected by either resolution limits of our simulations or the choice of gravity solver. Regarding the first aspect, the cosmo-SLICS have been shown to match in precision the commonly used COSMICEMU (Heitmann et al. 2014), and to even outperform the latter in terms of range, benefiting from more training nodes (Harnois-Déraps et al. 2019). The GPR interpolation uncertainty is fully propagated in the likelihood, and is quantified with a leave-one-out cross-validation test, in which the emulator is trained on all but one node, producing predictions at that node that are then compared with the measurement. As shown in Fig. 5, we cycle over all 26 nodes in this way and estimate an upper limit on the error (since the actual GPR has all the nodes). The accuracy degrades towards large positive or negative $\mathcal{S}/\mathcal{N}$ values, but most of the cross-validation lines lie well within the statistical precision on the data, shown with the dashed black lines. There are a handful of exceptions with poorer accuracy, attributed to removing extreme values of S_8 from the training set and therefore effectively demanding the GPR to extrapolate. With these edge nodes included, the full emulator has no such outliers. In fact, as argued in HD21, the most reliable estimate of the emulator's precision is evaluated by removing the fiducial cosmology and training on the others, which is shown as the thick black lines in the figure and always well within the statistical precision. For the KiDS modelling, the interpolation error is mostly at the 1 per cent error over the range $-1.0 < \mathcal{S}/\mathcal{N} < 3.0$ (our 'clean' range), and is otherwise always under 10 per cent. Similarly, the DES interpolation error is everywhere under 2 per cent (see HD21), the difference coming from the choice of $\mathcal{S}/\mathcal{N}$ cuts. This will likely be a limiting factor for future data analysis with sub-per cent accuracy requirement on the modelling, and will be addressed by increasing the number of nodes in the next generation of the *cosmology training set*. At the moment however, the interpolation error is low enough for our analysis. We nevertheless include it in our error budget by averaging over the square of the residuals (after the outliers have been removed):

$$\mathrm{Cov}_{\rm interp} = \mathrm{diag}\left\langle \left(N^{\rm GPR}_{\rm peaks} - N^{\rm sim}_{\rm peaks} \right)^2 \right\rangle , \tag{7}$$

and adding this to our statistical covariance. We could have instead used the errors directly provided by the Gaussian process emulator, however Heydenreich et al. (2021) have shown that the two meth-

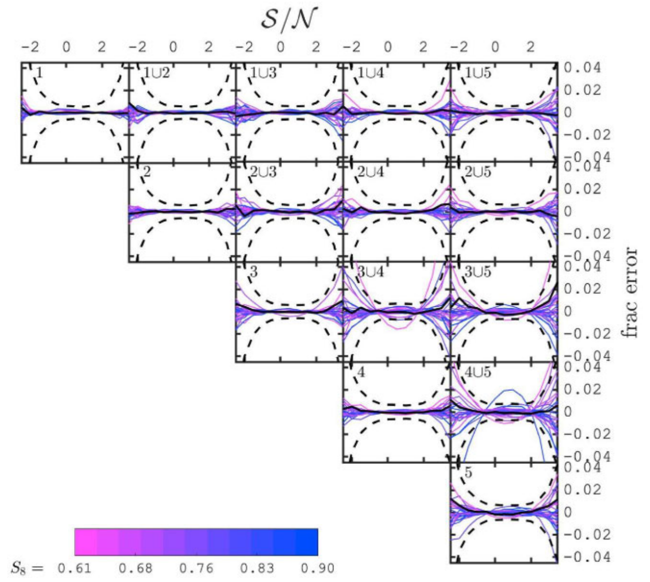

Figure 5. Accuracy of the KiDS-1000 GPR emulator, computed with a leave-one-out cross-validation test. The results are colour-coded with the S_8 value of the removed training point, and compared with the statistical precision on the measurement of the peak function (shown with the black dashed lines). The black solid line indicates the ΛCDM node, and the different panels show the auto- and cross-redshift (up to pairs) measurements; the other 15 tomographic combinations show a similar precision and hence are not shown. The outliers seen in a few panels are of extreme S_8 values and as such required the test emulator to extrapolate; this does not occur with the full emulator, and should therefore not be considered when estimating the interpolation error.

ods yield posteriors with negligible differences. This contribution, although small, helps with the goodness-of-fit in the data analysis.

The second aspect, concerning the training sets themselves, is discussed below in the section on N-body resolution.

4.2 Shape calibration

Table 1 shows the average multiplicative correction factors m_a that must be applied to the observed galaxy shapes in order to correct for a combination of residual PSF leakage, blending, and measurement noise, as assessed from Giblin et al. (2021). While in A21 the uncertainty on the shape calibration is absorbed directly in the analytical covariance matrix, our simulation-based method works instead at the level of the data vector, as for all other nuisance parameters. The $M_{\rm ap}$ estimator itself is unbiased (see equation 2), however we must propagate forward the uncertainty on the m_a calibration. The impact of potentially miscalibrated shape measurements is estimated by infusing a non-corrected global term $m_a \rightarrow m_a + \Delta m_a$ directly in the simulations and measuring the effect on the different elements of the peak function $N^{\kappa}_{\rm peaks}$. As we show later, this systematic effect is completely subdominant compared to the others due to the tight priors on Δm_a (reported in Table 2), and hence it is sufficient to model its impact with a reduced accuracy. In HD21 the estimation is based on a linear regression (i.e. $\partial N_{\rm peaks}/\partial \Delta m$ per data element) that is fit through 10 values of Δm_a. We use here only two points, at $\pm 1\sigma$, which is sufficient given the small values of Δm_a. The measured $\partial N_{\rm peaks}/\partial \Delta m$ is further discussed in Appendix B, and is used to modify the data vector for any value of Δm_a (see equation 8) sampled in the likelihood. For cross-redshift tomographic bins, we use the mean shift, e.g. $\Delta m_a^{1\cup 2} = (\Delta m_a^1 + \Delta m_a^2)/2$, which is consistent with what is currently done for all shear two-point function analyses. We could instead use an $n_{\rm gal}$-weighted mean to compute the Δm_a shift in cross-redshift tomographic bin, however this should have a negligible effect given the tight priors on these parameters, and we therefore leave this for the future.

4.3 Photometric redshifts

The KiDS-1000 uncertainty on the redshift distributions has been fully quantified in Hildebrandt et al. (2021), where it is shown that the mean of the $n(z)$ is captured to a high accuracy, varying by no more than 0.014 at the 1σ level.[19] The posteriors on the mean of the redshift distributions are used as priors on nuisance parameters in this work, summarized in Table 2. In this case however, the five redshift bias parameters Δz_a must be drawn from a correlated distribution. This is achieved in a two-step operation where we first draw five uncorrelated numbers from the priors, then rotate into the correlated space using a Cholesky decomposition of the redshift covariance matrix:

$$\mathbf{C_z} \times 10^5 = \begin{bmatrix} 11.20 & 2.600 & 1.562 & 0.056 & 0.622 \\ 2.600 & 12.78 & 4.081 & -1.692 & -0.2140 \\ 1.562 & 4.081 & 13.81 & -1.139 & 0.525 \\ 0.056 & -1.692 & -1.139 & 7.551 & 3.054 \\ 0.622 & -0.2140 & 0.525 & 3.054 & 9.496 \end{bmatrix},$$

which results in a correlated sampling of these five nuisance parameters (see A21, Hildebrandt et al. 2021, for more details). We produced a dedicated set of *redshift training set* simulations in which the $n(z)$ are shifted, but which are otherwise identical to the *cosmology training set* at the fiducial cosmology. Following HD21, we measure the peak function on full mock surveys with 10 shifts, each with a slightly different value of Δz_a sampled from the prior, then extract a linear fit per data element and estimate $\partial N_{\rm peaks}/\partial \Delta z_a$. This derivative is used to forward model redshift uncertainties on our data vector for arbitrary Δz_a values. Again, we use the mean shift when considering cross-redshift bins, and the DES-Y1 $\partial N_{\rm peaks}/\partial \Delta z_a$ measurements from HD21.

4.4 Astrophysics

Cosmic shear measurements are strongly affected by IA and baryon feedback. Using the *IA* and the *baryons training sets* described in Section 2.3, we estimate in a similar way $\partial N_{\rm peaks}/\partial A_{\rm IA}$ and $\partial N_{\rm peaks}/\partial b_{\rm bary}$, where $A_{\rm IA}$ and $b_{\rm bary}$ are free parameters that control the levels of IA and baryon contamination, respectively. The IA derivative is obtained by linear fitting the peak function's response to changes in $A_{\rm IA}$, measured from the *IA training set* infused with $A_{\rm IA} = 2.0$,1.0, 0.0, −1.0, and −2.0. Since IA is currently not well constrained and the NLA parametrization is an effective model, we adopt a wide top-hat prior over the range [−5.0; 5.0], as argued in Joachimi et al. (2021). This extrapolates our fit to larger $A_{\rm IA}$ values, which can in principle become inaccurate, however in the end the 3σ region of our posterior is fully contained within the training range (see Section 5). Similarly, the baryon derivative is measured from the *baryons training set*, which we use to infuse a baryonic correction whose strength is controlled by the parameter $b_{\rm bary}$. The case $b_{\rm bary} = 0.0$ corresponds to a dark matter-only universe, while $b_{\rm bary} = 1.0$ corresponds to the case where the feedback processes is exactly described by the *magneticum* physics. There is a large

[19] The full shape of the $n(z)$ is less accurate than its mean, and which consequences we leave for future work.

uncertainty on the amplitude of this baryon correction, hence we scale the measured baryonic correction with a free parameter $b_{\rm bary}$. Since the *magneticum* suites are already a strong model (see Martinet et al. 2021b, for a comparison with other hydrodynamical simulations), we sample the range $b_{\rm bary} \in [0.0, 2.0]$, thereby spanning a variety of realistic models (albeit imposing a fixed shape for the relative signal). As seen later, low $b_{\rm bary}$ values are not well constrained by the data while larger values are strongly disfavoured, hence we do not extend the prior limit beyond 2.0.

4.5 Implementation of forward-modelled systematics

Four sources of systematics are forward-modelled in our pipeline. Following Heydenreich et al. (2022), we construct systematics-infused data vector as:

$$\begin{aligned} N^{\rm syst}_{\rm peaks}&(\boldsymbol{\pi}, \Delta m_a, \Delta z_a, A_{\rm IA}, b_{\rm bary}) \\ &= N^{\rm GPR}_{\rm peaks}(\boldsymbol{\pi}) + \left[\partial N_{\rm peaks}/\partial \Delta m_a\right] \Delta m_a + \left[\partial N_{\rm peaks}/\partial \Delta z_a\right] \Delta z_a \ ... \\ &+ \left[\partial N_{\rm peaks}/\partial A_{\rm IA}\right] A_{\rm IA} + \left[\partial N_{\rm peaks}/\partial b_{\rm bary}\right] b_{\rm bary} \, , \end{aligned} \tag{8}$$

where the twelve parameters (Δm_a, Δz_a, $A_{\rm IA}$, $b_{\rm bary}$) are sampled from the priors described in Table 2. We marginalize over these nuisance parameters when inferring the values of the cosmological parameters. Equation (8) assumes that these different systematics are independent of cosmology and from each other, which we know is not entirely true. It has been shown that the cosmology dependence of the baryon feedback is a second order effect (McCarthy et al. 2017), supporting our simplified approach, however the intrinsic alignments couple to the tidal field that is in itself cosmology dependent. The shear calibration and redshift errors are independent of cosmology a priori, however the derivatives of the peak function with respect to Δm_a and Δz_a are not (see HD21), a secondary effect we neglect here. Moreover, it has been shown that the photometric and shape calibration errors are sometimes correlated (MacCrann et al. 2022). Although these approximation will become important in Stage-IV surveys, the current level of statistical precision allows us to relax the modelling of these effects without hurting our results. We illustrate this point in Section 5.2 by running inference MCMC chains in which the modelling of some or all of these systematic effects are switched off: the minor impact this has on the inference validates this approach. We also assume here that these systematic effects have a linear dependence on the nuisance parameter, which is probably not entirely true, but has been shown to be good enough for Stage-III lensing data in Heydenreich et al. (2022, see their fig. 7).

4.6 Other sources of systematics

In addition to the main systematic effects described in the last section, we consider here other known sources of errors that could potentially impact our results.

4.6.1 *N-body resolution*

Being completely simulation-based, our analysis relies on the quality of the underlying training samples. As mentioned already in Section 2.3, the *cosmology training set* has been shown to closely reproduce the non-linear clustering of the cosmic emulator (Heitmann et al. 2014), which is based on a completely independent N-body code. This agreement between different gravity solvers is key to assert the accuracy of the non-linear solution to structure formation (see e.g. *Euclid* Collaboration: Knabenhans et al. 2019, for a comparison between different N-body solvers), and the convergence of the solution must be assessed via a comparison with calculations carried out with a higher force/mass resolution simulations. As shown in HD21, known limits in the mass resolution of the *cosmo*-SLICS used for the peak function emulation mainly affect high peaks. More precisely, $N^{\kappa}_{\rm peaks}(\mathcal{S}/\mathcal{N} > 4.0)$ is systematically underpredicted by tens of per cent, while the $\mathcal{S}/\mathcal{N} = 4.0$ count is affected by no more than 5 per cent. This is in fact one of the main justification for our initial choice of upper $\mathcal{S}/\mathcal{N}$ limit.

The KiDS-1000 data are deeper than DES-Y1, and hence the sensitivity to such non-linear effects could be accrued here. We verify this by running our cosmological inference on the peak count statistics measured from the SLICS-HR, in which the increased force resolution results in a slightly larger number of large positive and negative peaks. Details are presented in Appendix B, but in short our data selection and marginalization scheme almost completely protects us against this, yielding no noticeable shifts on $\Omega_{\rm m}$ nor S_8. As in HD21, we nevertheless compute a multiplicative factor from the ratio between the SLICS-HR and the mean of the SLICS and optionally apply it on our model predictions during the likelihood sampling. The overall effect is smaller than the baryon and IA corrections, hence marginalizing over these latter two significantly washes out the impact of inaccurately modelled non-linear physics under question here. In the future we intend to look into multifidelity emulators as in Ho, Bird & Shelton (2022). The T17 and *magneticum* simulations were produced with a different N-body solver, and we show later that their reduced spatial resolution can affect quite significantly the peak count statistics. We treat this as a further uncertainty on the small scale physics and optionally add their scatter to the theoretical error in the covariance matrix, similar to equation (7). In two-point statistics analysis, this would be equivalent to including a theoretical error in the covariance matrix to account for difference between the $P(k)$ predictions provided by HALOFIT (Takahashi et al. 2012), HMCODE (Mead et al. 2016), or the BACCOEMU (Angulo et al. 2021), which can have a significant impact on the results (Aricò et al. 2023).

4.6.2 *Ray-tracing approximations*

Our ray-tracing method in itself contains approximations and algorithmic components that are bound to affect to some level the lensing statistics. Most importantly, the finite thickness of the mass sheets and the randomization process between them destroys correlations along the line of sight; in particular it can slice large galaxy clusters in two, and no structures larger than 257.5 h^{-1}Mpc can exist along the line-of-sight in our light cones (except for the T17 mocks, which we discuss below). This suppresses some of the large-scale power, as documented in Takahashi et al. (2017, see their Appendix B). However, smaller structures, such as those probed by the peak statistics, are left completely unaffected by this, which is why no forward modelling is needed here. This has been measured specifically for peak statistics in Zorrilla Matilla, Waterval & Haiman (2020) where it was found to play a subdominant role even for Stage-IV surveys. Of course, full on-the-fly light cones such as the ‘Onion Universe’ methods (Fosalba et al. 2008) avoid these problems, which we will consider for future analyses. The T17 simulations have thinner mass shells of 150.0 h^{-1}Mpc, but they are constructed such that the structures are preserved in groups of three shells, thus yielding a coherence length of 450.0 h^{-1}Mpc, further suppressing this residual systematic effect.

Another source of error comes from the fact that our simulations assume the Born approximation in the flat-sky limit, which introduces small inaccuracies at high-ℓ and low redshift, respectively (Hilbert et al. 2020). However, these are affecting the signal at a level much smaller than the statistical accuracy of our lensing data, and are not expected to matter here.

4.6.3 Covariance matrix

Estimation of the covariance matrix is one of the main computational challenges for non-Gaussian weak lensing probes, as it requires a large number of simulations with a resolution that is high enough to capture the non-linear physics being measured. Resorting to approximate methods such a FLASK (Xavier et al. 2016) and ICE-COLA (Izard, Fosalba & Crocce 2018) can significantly lower the computational cost of creating such mocks, but at the price of a reduced precision on the physics under investigation. We instead opted for mocks produced by a full N-body suite, our *covariance training set*, and are therefore only limited by the number of mocks and their box size. To test the convergence of our covariance matrix with respect to $N_{\rm sims}$, we run an inference analysis in which we increase the number of pseudo-independent realizations to 2120 (and adjusted the likelihood $N_{\rm sims}$ parameter accordingly), and find an excellent match to the posterior, with only the tail of the distribution being slightly modified. We could also have opted for a data compression such as in Zürcher et al. (2022) but that is not necessary given our results have converged, and our choice of likelihood accounts for the noise in the covariance matrix.

The simulation box size could also affect our results, however it has been shown in Harnois-Déraps et al. (2019) that the SLICS contains about 75 per cent of the ‘super-sample covariance’ term (SSC), when applied to 2pt statistics, yielding constraints on cosmological parameters that are highly accurate. Although this has not been demonstrated to date, peak count statistics are thought to be even less affected by the SSC, given that the covariance is close to being Poissonian, not Gaussian. As such, it scales with the number of peaks measured, which is independent of the survey window. In addition, as mentioned earlier, Burger et al. (2022) finds for the density-split statistics an excellent agreement between the SLICS covariance and that from full sky log-normal FLASK mocks (which contain an incomplete contribution from the trispectrum term but the full SSC), supporting our claim that the partly missing SSC must have a minimal influence on our error budget. This is also consistent with the recent findings from Linke et al. (2024) according to which the SSC term affects only the Fourier space estimators, whereas covariance matrix measured from intrasurvey real-space statistics such as the $M_{\rm ap}$ are unbiased.

4.6.4 Source-lens coupling and blending

An important difference between real and mock galaxies is that those in the data are clustered, which leads to a number of effects that are systematically absent from the calibration sample. For example, the quality of the shape measurements is lowered in regions of high density due to blending and obscuration. More importantly, the uncertainty in photometric redshifts is particularly severe in such areas, which often results in cluster members being wrongly assigned a higher redshift. This subsequently creates a small population of apparently high-redshift outliers that carry an unexpectedly weak shear component, thus diluting the overall lensing signal. Correcting for this can be partially achieved with ‘boost factors,’ however it was shown in HD21 and Zürcher et al. (2022) that even though the excess clustering around high peaks is indeed measured in the data, the impact this has on the inferred cosmology can be safely ignored. It was also shown in Gatti et al. (2024a) that source clustering had a minimal effect on the peak count statistics, supporting our choice to neglect this here.

4.6.5 Sampling the likelihood

Our likelihood sampling strategy, described in Section 3.3, assumes a flat prior for the four main cosmological parameters and the two astrophysical parameters ($A_{\rm IA}$ and $b_{\rm bary}$), and Gaussian priors for the parameters associated with photometric redshifts and shape calibration. This is not strictly speaking a non-informative approach, however the prior edges about the key measured parameters are sufficiently broad to have negligible impact on the posterior. Since it is found in Lemos et al. (2023) that MULTINEST tends to yield slightly overprecise constraints, we use the NAUTILUS sampler for our fiducial results, but report both.

We also note that our cosmology sampling strategy is different from the other KiDS-1000 cosmic shear analyses, mainly due to the volume where our emulator is valid. For example, A21 sample uniformly the parameters S_8, $\omega_{\rm c} \equiv \Omega_{\rm c}h^2$, $\omega_{\rm b} \equiv \Omega_{\rm b}h^2$, h, and $n_{\rm s}$. This choice is designed to avoid regions of parameter space that are strongly disfavoured by external data, and it was shown in Joachimi et al. (2021) that while it disfavoured high $\Omega_{\rm m}$ values already in prior space, the resulting S_8 prior space is highly uninformative. We could have taken a similar approach, however our emulator is much quicker, and hence it is more natural to sample the full training space, ensuring a wide sampling of $\Omega_{\rm m}$, S_8, and w_0.

Another aspect that currently limits our sampling strategy is the fact that we hold the value of many parameters fixed, notably $\Omega_{\rm b}$ and $n_{\rm s}$. In contrast, the DES-Y3 peak count analysis of Zürcher et al. (2022) use derivatives to marginalize over variation in these parameters, following the approach we adopt for IA and baryons. Neglecting to account for these has a small effect on current data sets (the DES-Y3 joint peaks + power spectrum analysis finds to be of about 0.13σ), which are thus ignored here.

4.6.6 $M_\times$ modes

The observed weak lensing signal can generally be decomposed into a combination of E- and B-modes, the latter of which can be estimated for any measurement by rotating all galaxies by 45 degrees; therefore, for the aperture mass map statistics, it is often referred to as $M_\times(\boldsymbol{\theta})$. The cosmic shear signal being a pure E-mode generator to first order, measurements of B-modes are therefore routinely used to assess the presence of residual systematics in lensing data (see e.g. Zürcher et al. 2022, for a recent application to peak statistics). Whereas the two-point function B-mode signal is zero in absence of systematics, the construction of aperture mass maps on a grid inevitably injects non-zero $M_\times$-modes due to the missing contribution from subpixel scales (Kilbinger, Schneider & Eifler 2006). This can be important: for a small-angle cut-off scale of 10 arcsec and an aperture of $\theta = 2.0$ arcmin, B-modes measured this way can reach about 10 per cent the size of the E-mode $M_{\rm ap}^2$ signal. This effect is accentuated for larger cut-off scales and smaller opening angles θ. Given our pixel scale of 35 arcsec, we do expect non-zero $M_\times$-modes to be introduced by our aperture map making, which we fully quantify in Appendix A. We show therein that the level of contamination is consistent with noise, that there is no evidence for residual systematics in the data from this measurement, hence that our cosmological analysis is clean of B-modes.

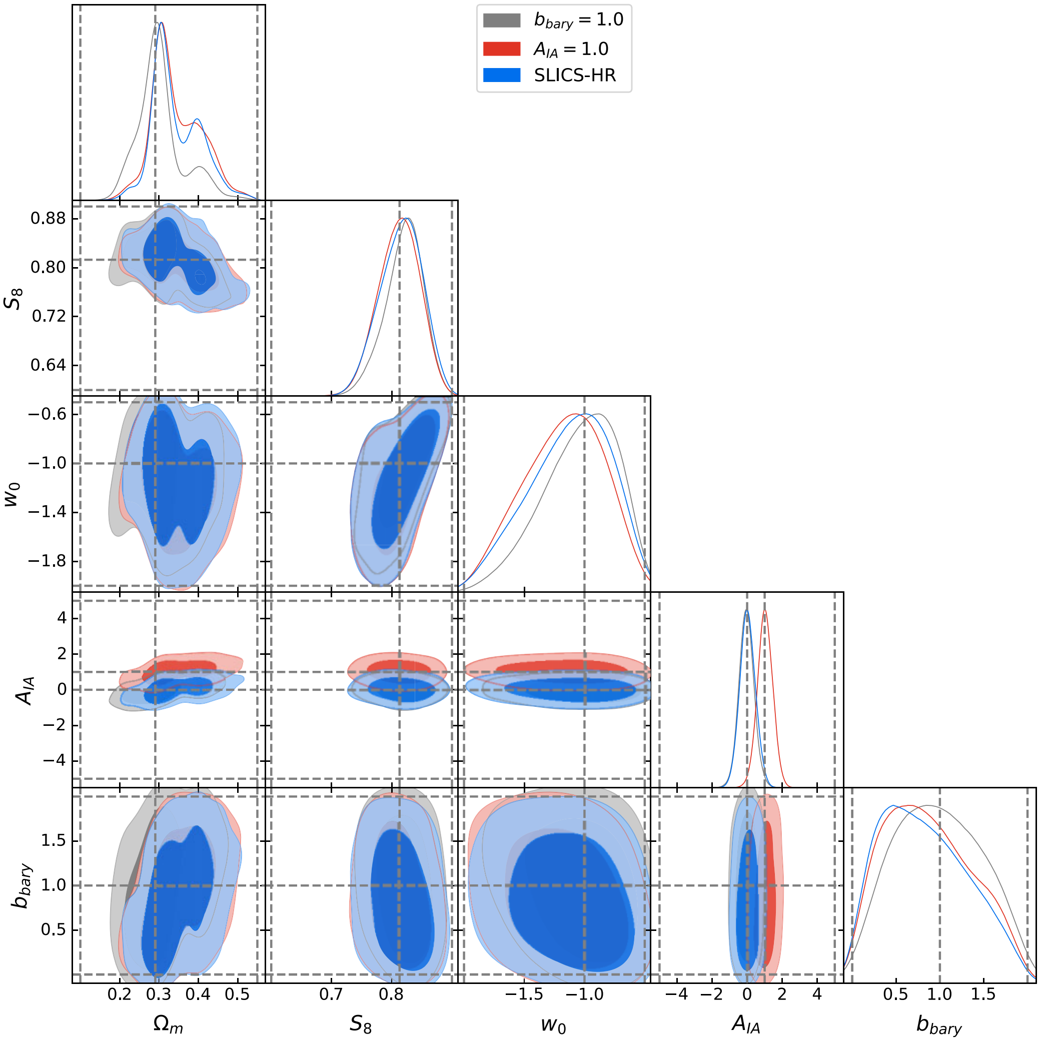

Figure 6. Full inference analysis on the *validation set* (SLICS-HR) peak count data (blue) with MULTINEST, optionally infused with intrinsic alignments (red) or baryon feedback (grey). We marginalize over these two effects plus shape calibration and photometric uncertainty. The priors are shown by the dashed grey lines at the edge of the panels, while the cross-hairs show the input truths.

4.7 Peak count to cosmology pipeline validation

We test our KiDS-1000 cosmology inference pipeline by analysing simulated data of known cosmology, infused with a controlled amount of residual systematics. In order to avoid confirmation bias, these tests are carried out with the *validation set*, which have not been used in the cosmology training nor for the covariance estimation, with an N-body force resolution that is higher than the other simulations used in this work.[20] In addition, we use the forward-modelling approach presented in Section 4 to infuse the simulated data vectors with either intrinsic alignments (assuming $A_{\rm IA} = 1.0$) or baryonic feedback (with $b_{\rm bary} = 1.0$). Fig. 6 shows the results for these three analysis cases. The maxima of the projected posterior distributions are all centred on the input truth, except for the $b_{\rm bary}$ parameter, which are away from zero even in the no-baryon cases. This is a projection effect similar to those discussed in Joachimi et al. (2021), Chintalapati, Gutierrez & Wang (2022), and Dark Energy Survey and Kilo-Degree Survey Collaboration (2023), and we have verified that reducing the lower prior limit to $b_{\rm bary} = -2.0$ pushed both the red and blue maxima towards the ground truth.

In this test, there is a secondary solution for $\Omega_{\rm m} \sim 0.4$ that is unexpected, and not observed in other peak count analyses (Martinet et al. 2018; Zürcher et al. 2022; Marques et al. 2024). As detailed in Appendix B, this feature persists when analysing data from the *cosmology training set* at the fiducial cosmology, from the *baryons training set* and from the T17 mocks, but can vanish at other cosmologies. This is caused first by the poor sensitivity of the current lensing data to $\Omega_{\rm m}$, as also seen in the large $\Omega_{\rm m}$ scatter reported in A21 between different two-point functions, but more importantly by limits in our GPR emulator, whose residual inaccuracy mostly affects this parameter. Inferring $\Omega_{\rm m}$ from single mock survey realizations (as opposed to a mean over several light cones or shape noise realizations) yields posteriors drawn from either one of these peaks, resulting in occasional strong biases on this cosmological parameter. We thoroughly verify that only $\Omega_{\rm m}$ is affected by this, and therefore do not report its value in our main analyses. See Appendix B for full details.

[20] We have further verified that the wCDM cosmology is correctly inferred when analysing data from the *cosmology training set* but these tests are easier to satisfy since the data is used for training the emulator. We discuss these in greater details in Appendix B.

However, the secondary $\Omega_{\rm m}$ solution corresponds to an S_8 posterior that is slightly lower than the main solution, which means that if a particular realization of the data prefers this region, it will on average have an S_8 value about 0.03 lower, which is of the size of our statistical precision. Conversely, realizations that prefer lower $\Omega_{\rm m}$ tend to have S_8 values that are 0.02 higher than the input truth. We further observe that this is not always the case: some individual mock survey realizations from the *covariance training set* have a best fit $\Omega_{\rm m} \sim 0.45$, yet their S_8 is unbiased compared to the input truth. Given that this $0.02 - 0.03$ shift is about a 1σ shift, this potentially dominates the systematic error budget on S_8, which we therefore must report as$\sigma(\mathrm{syst})={}^{-0.03}_{+0.02}$.

This additional systematics error take its roots from the tilt in the $[S_8 - \Omega_{\rm m}]$ posterior, which indicates residual correlation between these two parameters. We can suppress this tilt, and hence the additional error, by replacing S_8 with $\Sigma_8^\alpha \equiv \sigma_8[\Omega_{\rm m}/0.3]^\alpha$, where α is the parameter that best fits the $[\Omega_{\rm m} - \sigma_8]$ degeneracy. According to this metric, Σ_8^α is the most robustly measured quantity from peak statistics, with no need for a standalone σ(syst) term, in this case a significant advantage. With the validation data, we find $\Sigma_8^\alpha = 0.824^{+0.033}_{-0.033}$, with $\alpha = 0.582$, in excellent agreement with the input truth of 0.811 with the same α. We report the measurements of both S_8 and Σ_8^α in this paper, but while emphasize is on the former to better compare with previous measurements from the literature, the latter is more robust and has interesting properties which we highlight as well, notably on increasing the agreement with previous KiDS-1000 measurements and lowering the tension with external probes.

Back to Fig. 6, we observe that the posteriors on w_0 and $b_{\rm bary}$ are wide and significantly overlap with the prior limits, and we thus expect to be unable to place meaningful constraints on these parameters with our main KiDS-1000 analysis alone. We observe a degeneracy in the $[S_8 - w_0]$ plane here, however we show in Appendix B that it is not always seen when analysing other cosmologies, making it impossible to draw physically meaningful conclusions about this. Only the $[S_8 - A_{\rm IA}]$ plane is well constrained with the current KiDS-1000 peak count analysis: we achieve a 4.4 per cent precision measurement on S_8, with $S_8^{\rm SLICS-HR} = 0.816^{+0.039}_{-0.033}$ (truth is 0.813), and a precision of $\sigma_{A_{\rm IA}} = 0.45$ on $A_{\rm IA}$, sampling the likelihood with MULTINEST.

The DES-Y1 pipeline validation is presented in HD21, while that for the joint KiDS-DES is presented in Appendix B, showing again an excellent agreement between the inferred cosmology and the input truth.

Table 4. Summary of our cosmological inference analyses. Posteriors on $\Omega_{\rm m}$, h, and w_0 are prior-limited, so their constraints are not reported here. Unless explicitly specified in the first column, the KiDS-1000 measurements are based on the 'clean' data vector, i.e. $-1.0 < \mathcal{S}/\mathcal{N} \leq 3.0$. The last column presents the *maximum a posteriori* (MAP) values. Validation of the inference pipelines on mock data are presented in Appendix B.

	Peak count analysis of KiDS-1000				
	NAUTILUS		MULTINEST		MAP
	S_8	$A_{\rm IA}$	S_8	$A_{\rm IA}$	S_8
Fiducial	$0.733^{+0.032}_{-0.032}$	$0.71^{+0.49}_{-0.49}$	$0.733^{+0.021}_{-0.027}$	$0.74^{+0.43}_{-0.43}$	0.726
ΛCDM	$0.732^{+0.029}_{-0.029}$	$0.73^{+0.48}_{-0.48}$	$0.729^{+0.026}_{-0.026}$	$0.73^{+0.43}_{-0.43}$	0.718
Auto-only	$0.734^{+0.050}_{-0.050}$	$-0.2^{+1.0}_{-1.0}$	$0.732^{+0.039}_{-0.048}$	$-0.2^{+1.0}_{-1.0}$	0.725
Up to pairs	$0.750^{+0.036}_{-0.050}$	$0.10^{+0.69}_{-0.69}$	$0.742^{+0.032}_{-0.043}$	$0.11^{+0.63}_{-0.63}$	0.742
Up to triplets	$0.740^{+0.035}_{-0.035}$	$0.35^{+0.53}_{-0.53}$	$0.740^{+0.029}_{-0.029}$	$0.37^{+0.48}_{-0.48}$	0.738
No tomo	$0.695^{+0.033}_{-0.087}$	$-0.6^{+1.7}_{-2.2}$	$0.690^{+0.038}_{-0.068}$	$-0.7^{+1.5}_{-2.0}$	0.682
$-2.5 < \mathcal{S}/\mathcal{N} \leq 4.0$	$0.720^{+0.036}_{-0.026}$	$0.08^{+0.30}_{-0.30}$	$0.717^{+0.031}_{-0.022}$	$0.07^{+0.27}_{-0.27}$	0.728
$-2.5 < \mathcal{S}/\mathcal{N} \leq 3.0$	$0.717^{+0.031}_{-0.031}$	$0.14^{+0.31}_{-0.31}$	$0.713^{+0.023}_{-0.023}$	$0.22^{+0.27}_{-0.27}$	0.712
$0.0 < \mathcal{S}/\mathcal{N} \leq 4.0$	$0.739^{+0.031}_{-0.026}$	$0.77^{+0.47}_{-0.47}$	$0.744^{+0.023}_{-0.023}$	$0.80^{+0.38}_{-0.38}$	0.734
No IA	$0.726^{+0.024}_{-0.042}$	–	$0.720^{+0.021}_{-0.031}$	–	0.725
No baryons	$0.732^{+0.032}_{-0.032}$	$0.71^{+0.49}_{-0.49}$	$0.725^{+0.022}_{-0.027}$	$0.70^{+0.43}_{-0.43}$	0.725
No syst	$0.729^{+0.024}_{-0.056}$	–	$0.723^{+0.022}_{-0.048}$	– –	0.709
No GPR error	–	–	$0.732^{+0.019}_{-0.025}$	$0.71^{+0.40}_{-0.40}$	0.708
N-body error	–	–	$0.725^{+0.027}_{-0.027}$	$0.70^{+0.43}_{-0.43}$	0.717
No bin1	$0.734^{+0.043}_{-0.037}$	$0.10^{+0.74}_{-0.74}$	$0.735^{+0.035}_{-0.035}$	$0.10^{+0.70}_{-0.10}$	0.727
No bin2	$0.740^{+0.042}_{-0.048}$	$-0.78^{+0.72}_{-0.72}$	$0.740^{+0.040}_{-0.040}$	$-0.76^{+0.70}_{-0.70}$	0.738
No bin3	$0.775^{+0.049}_{-0.055}$	$0.97^{+0.63}_{-0.63}$	$0.777^{+0.048}_{-0.048}$	$0.95^{+0.60}_{-0.60}$	0.836
No bin4	$0.701^{+0.037}_{-0.037}$	$0.41^{+0.68}_{-0.68}$	$0.702^{+0.029}_{-0.034}$	$0.45^{+0.59}_{-0.59}$	0.659
No bin5	$0.720^{+0.036}_{-0.028}$	$0.53^{+0.61}_{-0.61}$	$0.723^{+0.030}_{-0.025}$	$0.53^{+0.57}_{-0.57}$	0.716
	Peak count analysis of DES-Y1				
DH21	–	–	$0.737^{+0.027}_{-0.031}$	–	–
This work	$0.743^{+0.036}_{-0.036}$	–	$0.742^{+0.030}_{-0.034}$	–	0.712
	Joint peak count analysis				
	NAUTILUS		MULTINEST		
Fiducial	$0.732^{+0.020}_{-0.020}$	$0.82^{+0.47}_{-0.47}$	$0.732^{+0.012}_{-0.010}$	$0.82^{+0.33}_{-0.33}$	0.745
ΛCDM	$0.736^{+0.016}_{-0.018}$	$0.81^{+0.46}_{-0.46}$	$0.736^{+0.012}_{-0.015}$	$0.79^{+0.40}_{-0.40}$	0.732
No baryons	$0.728^{+0.020}_{-0.016}$	$0.82^{+0.46}_{-0.46}$	$0.725^{+0.018}_{-0.014}$	$0.83^{+0.39}_{-0.39}$	0.726
No IA	$0.726^{+0.020}_{-0.016}$	–	$0.729^{+0.015}_{-0.015}$	–	0.721

5 RESULTS: KIDS-1000

We present in this section the results from our cosmological inference analyses, beginning with the fiducial KiDS-1000 pipeline, then reporting on the importance of various selection cuts and systematic effects. For reasons explained in Section 4.7, we report only the constraints on S_8 and $A_{\rm IA}$; results are summarized in Table 4 and further condensed in Fig. 10.

From our fiducial full tomographic KiDS-1000 analysis of the measurements presented in Fig. 3, we obtain:

$$S_8^{\rm KiDS} = 0.733^{+0.032+0.020\,({\rm syst})}_{-0.032-0.030\,({\rm syst})}\,, \qquad A_{\rm IA} = 0.71^{+0.49}_{-0.49}\,, \tag{9}$$

$$\Sigma_8^{\rm KiDS} = 0.765^{+0.030}_{-0.030}\,, \qquad \alpha = 0.600\,, \tag{10}$$

after marginalizing over three cosmological parameters ($\Omega_{\rm m}$, w_0, and h) and 11 nuisance parameters ($5 \times \Delta m_a$, $5 \times \Delta z_a$, $b_{\rm bary}$). Unless explicitly mentioned, all quoted parameter constraints correspond to the mean $\pm 1\sigma$ region of the marginalized posterior, not to be confused with the point of maximum likelihood in the higher dimensional space. This is therefore a 3.9 per cent measurement of the structure growth parameter Σ_8. The best-fitting model is shown with the red line in Fig. 3. The joint constraints on two of these parameters are shown in Fig. 7, along with results from different selections of tomographic bins. Importantly, the strong $S_8 - A_{\rm IA}$ degeneracy seen in the no-tomographic case (the tilted dark purple contour) is lifted by tomographic decomposition, which capture the different redshift dependence of the cosmological and IA signals. Indeed, in the no-tomographic case only, and under the NLA framework, large S_8 values can be hidden by large tidal alignments, both fitting equally well the same data. However, as seen by the coloured histograms in Fig. 3, the cosmological signal in all tomographic bins is affected by changes in S_8, while IA mostly modifies the parts of the data vector that include the lowest tomographic bins. This difference allows one to break the $[S_8 - A_{\rm IA}]$ degeneracy, an important verification of our

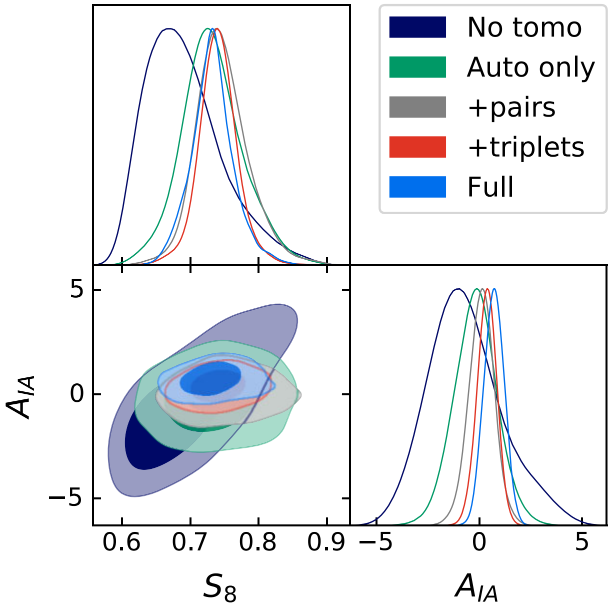

Figure 7. KiDS-1000 constraints on the two best-measured parameters from peaks count statistics, for different selection of redshift bins. Tomographic analyses all break the $S_8 - A_{\rm IA}$ degeneracy.

IA modelling. This result would be slightly different had we included redshift evolution of the IA signal, but this effect will be subdominant given the size of our statistical error bars. This will clearly need to be investigated with upcoming data sets.

Back to Fig. 7, all tomographic additions contribute to further tightening the constraints, once again demonstrating the power of using cross-redshift bins in non-Gaussian statistics. We also observe that all cases shown in Fig. 7 are consistent, providing statistical robustness to our measurement.

At our best-fitting parameters the measurement yields a χ^2 of 250, which reduces to $\chi^2_{\rm red} = 1.22$ after dividing by $\nu = (220 - 4.5) = 205.5$ degrees of freedom. Note that although we use six unconstrained parameters[21] in our likelihood evaluation (the four cosmological parameters plus $A_{\rm IA}$ and $b_{\rm bary}$), it was shown in Joachimi et al. (2021) that an effective number of $\nu = 4.5$ free parameters better describes the weak lensing data given the existing correlations and degeneracies, results which we have used here.[22] Our PTE for this measurement is 0.43, which is well above our threshold of 0.01, using the non-χ^2 distribution described in Appendix C. It is worth noting that the KiDS-1000 shear two-point correlation functions and band power analyses had a lower goodness-of-fit, with PTE = 0.034 and 0.013, respectively.

In many previous analyses, sampling and marginalization over w_0 is often excluded, being considered an extension to the vanilla ΛCDM scenario. In the present case, fixing w_0 to -1.0 when sampling the likelihood[23] results in minor changes to the reported (S_8, Σ_8, $A_{\rm IA}$) constraints, leading to $0.732^{+0.029}_{-0.029}$, $0.767^{+0.026}_{-0.026}$, and $0.73^{+0.48}_{-0.48}$. Interestingly, we find that the impact of opening up the w_0 dimension is far lower than for the two-point statistics, where Tröster et al. (2021) finds a degradation by a factor of a few on the S_8 constraints (compare their figs 1 and 6). Different degeneracy-breaking directions are likely causing this difference, which is promising for upcoming measurements of w_0 with alternative statistics (see Martinet et al. 2021b, for a Stage-IV lensing forecast on the dark energy parameter with peak statistics).

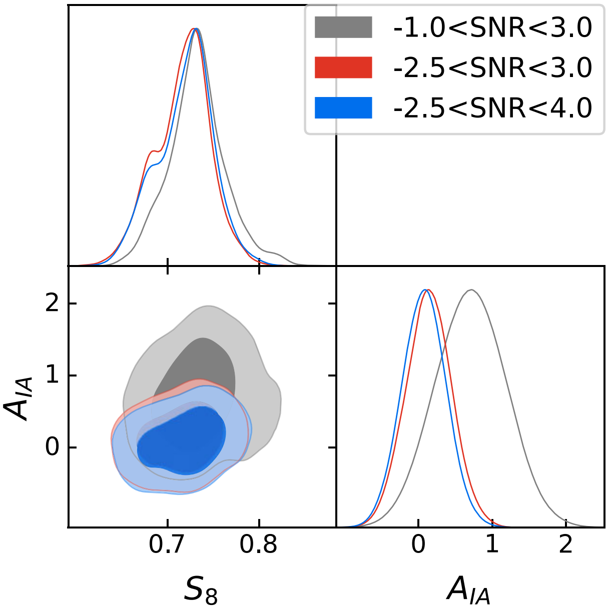

Figure 8. Effect of $\mathcal{S}/\mathcal{N}$ cuts on the KiDS-1000 constraints. Large peaks ($\mathcal{S}/\mathcal{N} > 3$) slightly increase the statistical precision on S_8, as seen by comparing the red and blue contours (see values in Table 4). Negative peaks ($\mathcal{S}/\mathcal{N} <$ -1), included in the red but excluded from the grey contours, help in breaking the [$S_8 - A_{\rm IA}$] degeneracy. The grey and blue contours correspond to the 'clean' and 'aggressive' cases, respectively.

One of the key questions to be explored by beyond-2pt statistics concerns the exact origin of the non-Gaussian cosmological information. Large peaks are often associated with massive galaxy clusters, which are known to be highly sensitive to the dark energy equation-of-state parameters for instance, however the wide projection effect and the fact that baryons, IA, and non-linear physics maximally affect these large $\mathcal{S}/\mathcal{N}$ peaks (Martinet et al. 2021b; Harnois-Déraps et al. 2022) complicate the picture. To (partly) answer this question, we investigate the constraining power contained in the highest ($\mathcal{S}/\mathcal{N} >$ 3) and lowest ($\mathcal{S}/\mathcal{N} <$ 0) bins by removing these sequentially from the 'aggressive' data vector ($-2.5 \leq \mathcal{S}/\mathcal{N} \leq 4.0$) in the likelihood. The results are shown in Fig. 8, where we observe that the negative $\mathcal{S}/\mathcal{N}$ peaks significantly help break the [$S_8 - A_{\rm IA}$] degeneracy, while the highest peaks help in tightening the S_8 constraints. In an analysis that ignored the role of IA, M20 found that the amount of information about S_8 that is contained in negative peaks is quite small, however here we find that they actually play a key role once IA are forward-modelled.

5.1 Internal consistency

It has been found in previous cosmic shear analyses (e.g. A21, Hamana et al. 2020; Amon et al. 2022) that internal consistency tests can help differentiate residual systematics from statistical fluc-

[21] We do not count as free parameters those nuisance parameters for which we impose a tight prior.

[22] It is not guaranteed that the exact same effective number of degrees of freedom applies here, given that the likelihood is not sampled over the same volume. We have checked that our goodness-of-fit is robust over choices for this quantity, with PTE varying between 0.48 and 0.37 over the range $2 < \nu < 7$.

[23] This still uses the same wCDM emulator, but only varying the other three cosmological parameters.

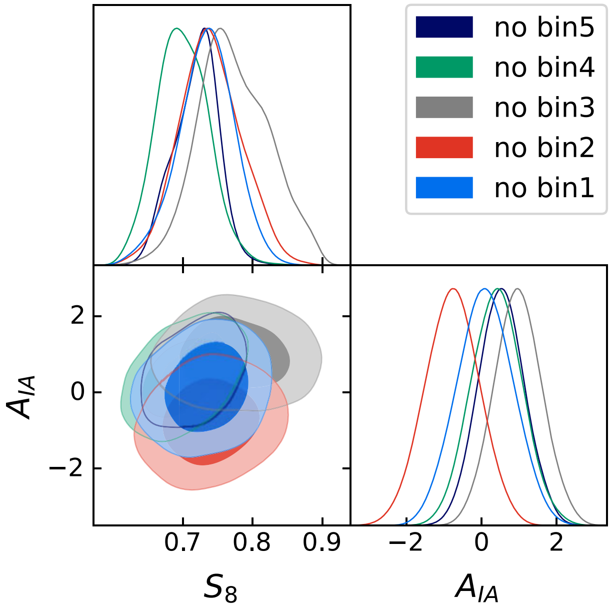

Figure 9. Internal consistency: effect of removing tomographic data from the KiDS-1000 analysis.

tuations. We therefore stress test our results by removing data from tomographic bins one at a time before proceeding to the inference. For example, we consider results obtained from an analysis where exactly no data from bin1 (i.e. 1, 1∪2...1∪2, 1∪2∪3... 1∪2 ∪ 3 ∪ 4∪5) is used, then no data from bin2, and so on. The results are shown in Fig. 9, where we observe that all cases are self-consistent, in agreement with the full selection. Note that the S_8 shifts per-bin are not expected to match exactly those measured with other lensing probes due to different responses of the cosmic shear estimators to noise in the data. For example, A21 found that removing the fifth tomographic bin maximally degrades the precision on S_8, confirming the large amount of information on this parameter carried by high redshift bins in shear two-point functions. In contrast, we find here that removing the third redshift bin has the worst impact on the precision. The third bin has the greatest number density of galaxies, hence better captures the information in peak statistics, whose mean value is affected by the noise level. The constraints on $A_{\rm IA}$ fluctuate about the fiducial results by less than 2σ, while those on S_8 agree within 1σ, as expected.

5.2 Impact of systematics

We present in this section additional variations with respect to the fiducial analysis, designed to better understand our results and assess their robustness to residual systematics. We first investigate the impact of IA on the uncertainty by fixing $A_{\rm IA}$ to 0.72, the best-fitting value in the fiducial analysis. Doing so, the error bars on S_8 shrink by less than 10 per cent, while the mean value is not affected, by construction. Setting instead the IA parameter to 0.0, we can estimate the bias on the inferred cosmology if IA are completely neglected. We measure in this case $S_8^{\rm no-IA} = 0.725^{+0.024}_{-0.042}$, a 0.22σ shift from the fiducial results. Intrinsic alignments are therefore a modest part of the error budget, suggesting that peak count analyses where IA are not modelled or held fixed (e.g. M18, HD21, Marques et al. 2024) likely yield both biased low and slightly optimistic constraints for S_8.

We next carry out a similar study this time removing the modelling of baryons, fixing the associated nuisance parameter to $b_{\rm bary} = 0.0$. As reported in Table 4, the measurements are mostly unchanged. As shown in M21, any non-zero residual feedback tends to lower the number of high $\mathcal{S}/\mathcal{N}$ peaks in all tomographic bins, which, when confronted to fixed data, must be compensated with an increased value of inferred S_8. Therefore, removing the baryon modelling goes the other way and reduces the inferred S_8. This is not clearly seen with the NAUTILUS chains, but the MULTINEST runs shows this shift with 0.2σ significance.

Then, removing modelling of all systematics (photo-z, shape calibration, IA, and baryons) results in S_8 values half way between the no-baryon and no-IA cases, but the error bars are the larger. This suggests that marginalization over these systematics helps in finding the true maximal likelihood, which is not at $b_{\rm bary} = A_{\rm IA} = 0$. Indeed, the error on S_8 becomes smaller than the fiducial case if $A_{\rm IA}$ and $b_{\rm bary}$ are fixed to their best-fitting value of 0.72 and 0.5, respectively, leading to $S_8^{\rm syst-fixed} = 0.728 \pm 0.030$.

The contribution to the error budget coming from the GPR interpolation uncertainty (equation 7) can be estimated from an MCMC run where the covariance matrix excludes this term, and we observe that the error on S_8 is reduced by just under 10 per cent. Similarly, adding an error on small scale non-linear physics estimated from the scatter between the cosmo-SLICS, *magneticum* dark matter-only and the T17 simulations (see Section 4.6.1), can degrade the error on S_8 by 12 per cent. This is an upper limit on the degradation, given that the *cosmology training set* has better resolution than these, hence the real uncertainty is certainly smaller. We do not include this latter error in the fiducial analysis here, because it is not accurately estimated, and instead report an upper bound on the effect.

Finally, we compared our fiducial NAUTILUS results with those from the MULTINEST nested sampler and recover negligible biases in the inferred parameters, but with smaller error bars ($S_8 = 0.733 \pm 0.032$ versus $0.733^{+0.021}_{-0.027}$ for MULTINEST). This is consistent with previous findings (Lemos et al. 2023) and justifies our choice of NAUTILUS as our main sampler. We nevertheless report results from both samplers to ease comparison with previous results.

5.3 Comparison with previous KiDS-1000 results

The S_8 measurement presented here is not the first carried out from KiDS-1000. Previous analyses include the measurements of A21 and van den Busch et al. (2022), the latter of which used an upgraded photometric calibration compared to the former, followed by that of Li et al. (2023b) based on upgraded shear measurements. Loureiro et al. (2022) carried out a *pseudo*-C_ℓ analysis, Fluri et al. (2022) used instead a convolutional neural network, while Longley et al. (2023) re-analysed the data within the LSST-DESC pipeline. We report these results as the purple symbols in Fig. 10, where we see that they all seem to prefer slightly higher values of S_8 compared to our own measurements, albeit not by a significant amount. Given the important differences in the analysis pipelines between these efforts, it is reassuring to recover $< 1\sigma$ agreements. The constraints from the wCDM band power analysis from Tröster et al. (2021) are reported in Fig. 11 and are broadly consistent with our peak statistics constraints, even though peaks are clearly more constraining on S_8 (0.732 ± 0.032 for peaks versus 0.742 ± 0.047 for band power), due to the reduced degeneracy in the $[S_8 - w_0]$ plane. It is worth noting that both statistics provide similar constraints on the $A_{\rm IA}$ parameter ($\sigma_{A_{\rm IA}} = 0.42$ for peaks, compared to $\sigma_{A_{\rm IA}} = 0.36$ for band power), which is reassuring given that both use the same NLA approach. This

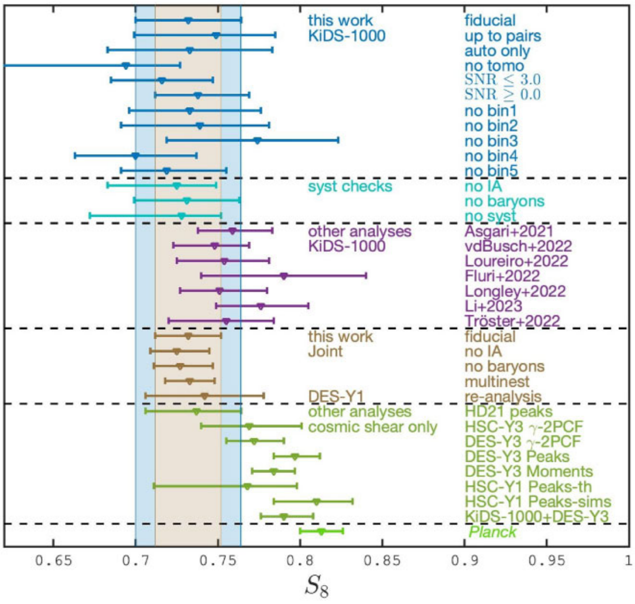

Figure 10. Summary of S_8 constraints from this work, from recent cosmic shear data analyses and from *Planck*. This figure shows the projected 1σ errors.

error is significantly reduced ($\sigma_{A_{\rm IA}} = 0.30$) when considering the more aggressive data selection ($-2.5 \leq \mathcal{S}/\mathcal{N} \leq 4.0$), but since the associated goodness-of-fit is poor, the results are not straightforward to interpret. We nevertheless expect tighter constraints on $A_{\rm IA}$ to be achievable coming from non-Gaussian probes.

We finally remark that our constraints on Σ_8^α aligns remarkably well with the band power measurements presented in A21 (they found $\Sigma_8^\alpha = 0.765^{+0.018}_{-0.024}$ with $\alpha = 0.58$, compared to our measurement of $\Sigma_8^\alpha = 0.765^{+0.030}_{-0.030}$ with $\alpha = 0.60$).

6 JOINT ANALYSIS WITH DES-Y1

The posterior obtained from the KiDS-1000 peak count analysis is fully consistent with that from the peak count analysis of the Dark Energy Year 1 (DES-Y1) presented in HD21. In particular, the latter finds $S_8^{\rm HD21} = 0.737^{+0.027}_{-0.031}$, which significantly overlaps with our $S_8^{\rm KiDS}$ 1σ results. Other parameters less well measured such as $\Omega_{\rm m}$ and w_0 are also largely overlapping at the 1σ level (see the lower part of Fig. 11), which means the intersection between the two likelihood hypervolumes must be large enough to safely combine the two data sets. Furthermore, both measurements are based on the similar analysis pipeline and, in particular, exploit the same simulations to model the cosmology dependence, thereby suppressing the risk of mis-interpreting the joint data due to non-uniform modelling of the signal.

6.1 Results: DES-Y1 re-analysis

As detailed in Section 3.3, there are differences between our DES-Y1 pipeline and that presented in HD21, including the S_8 sampling, the treatment of baryons, the inclusion of the emulator uncertainty on the covariance and the choice of sampler. The results from these re-analyses are presented in the lower panel of Fig. 11 (in green and black). The difference induced on these contours are small, but the goodness-of-fit improvement is important, with a PTE of 0.53 (using the same PTE estimator as HD21, we obtain 0.25, which is still a massive improvement compared to their PTE = 0.005).

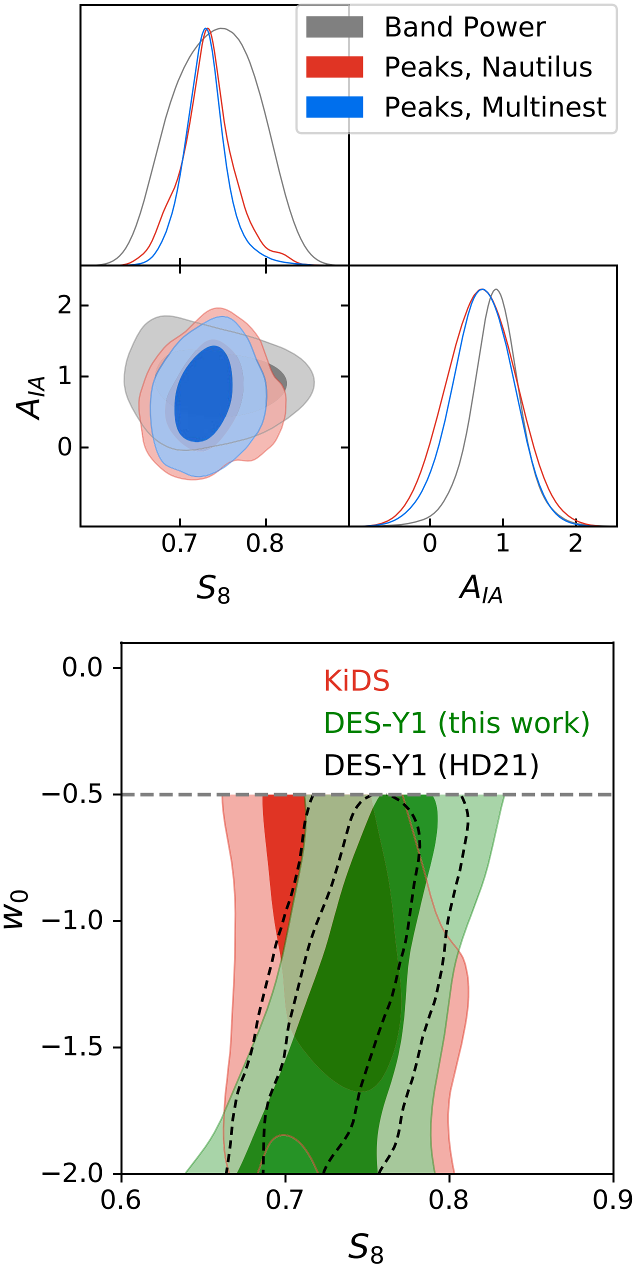

Figure 11. Comparison with the previous cosmic shear results. The top part shows a comparison with the KiDS-1000 band power analysis from Tröster et al. (2021, based on MULTINEST), while the bottom part compares the KiDS results with the DES-Y1 peak count analysis from HD21 (black) along with current re-analysis (green) and detailed in Section 6. Note that the posteriors obtained from the NAUTILUS sampler are typically wider and more accurate than those from MULTINEST used in HD21. The excellent agreement seen in this figure warrants the joint survey analysis.

We remark that our joint pipeline contains a slight inconsistency: we include IA with the NLA model in the KiDS-1000 data (with marginalization over $A_{\rm IA}$) and with the non-linear halo-based IA model for the DES-Y1 data (without marginalization, but with an on/off switch instead). We verify the impact of this feature by analysing the likelihood with the DES IA model turned on and off and report on the difference, which is subdominant ($\Delta S_8 = 0.002$). We also compare the results from turning off the modelling of baryons, and from replacing the wCDM by a ΛCDM analysis, finding in all cases results consistent with the fiducial analysis. The re-analysis presented in this work has slightly larger error bars compared to that of HD21, due to the marginalization over baryons, and to the fact that NAUTILUS yields constraints slightly larger compared to MULTINEST, as summarized in Table 4. Notably, we infer:

$$S_8^{\rm DES} = 0.743^{+0.036}_{-0.036}\,, \quad \Sigma_8^{\rm DES} = 0.762^{+0.036}_{-0.036}\,, \quad \text{with}\,\alpha = 0.559\,. \qquad (11)$$

6.2 Results: joint KiDS + DES

We present in this section the results from our joint KiDS-1000 + DES-Y1 peak count analysis. Sampling the joint likelihood with our fiducial setup, we achieve improved constraints on S_8 and Σ_8^{α} with:

$$S_8^{\rm joint, wCDM} = 0.732^{+0.020}_{-0.020}\,, \quad \Sigma_8^{\rm joint, wCDM} = 0.759^{+0.020}_{-0.017} \qquad (12)$$

and

$$S_8^{\rm joint, \Lambda CDM} = 0.735^{+0.016}_{-0.018}\,, \quad \Sigma_8^{\rm joint, \Lambda CDM} = 0.762^{+0.017}_{-0.017} \qquad (13)$$

computed with $\alpha = 0.572$ in both cases. These are the tightest results obtained from non-Gaussian cosmic shear statistics to date, comparable to the recent joint ΛCDM analysis of the KiDS-1000 and DES-Y3 data (Dark Energy Survey and Kilo-Degree Survey Collaboration 2023), which measured $S_8^{\rm NLA} = 0.792^{+0.016}_{-0.013}$. The 2D posterior is shown in Fig. 12 (in blue) and compared to the fiducial KiDS-1000 (red) and DES-Y1 (green) peak statistics constraints. Recall that the $A_{\rm IA}$ parameter affects only the KiDS likelihood since, as explained in the previous section, the DES likelihood assumes instead a fixed halo-based IA model with no free parameter. We should therefore use caution when interpreting this parameter. The reported value is close to the point of maximum likelihood ($S_8^{\rm ML} = 0.728$), and the size of the error bars on S_8 is consistent with our expectation: for example, we read from Table 4 that the ΛCDM KiDS-1000 analysis has a mean error of $\sigma_{S_8} = 0.029$. Scaling this precision by the square root of the area, we naively predict a joint survey error of around 0.018, and obtain 0.017. The error would be slightly larger had we included as well a marginalization of the IA in the DES-Y1 part of this analysis, possibly explaining this slight difference. At the joint best-fitting cosmology, the PTE values for the KiDS and DES pipelines are basically unchanged, while the joint analysis has a $\chi^2_{\rm red} = 1.15$ and a PTE of 0.96, all satisfying our goodness-of-fit criteria.

If we restrict the joint analysis to $w_0 = -1.0$, the S_8 values are minimally affected while the uncertainty is reduced, as expected from lowering the dimensionality of the likelihood. Alternatively, turning on the IA modelling in the DES likelihood only yields a 0.2σ downward shift, also expected whenever IA modelling is added. The smallness of this shift is once again showing that the intrinsic alignment do not significantly impact the peak count statistics as measured in the DES-Y1 data. In comparison, setting to zero the IA model in both KiDS and DES results in $S_8^{\rm joint, no-IA} = 0.725^{+0.020}_{-0.016}$. Holding fixed the baryonic feedback parameter to $b_{\rm bary} = 0.0$ has similar consequences on this joint analysis, shifting the best-fitting value to $S_8^{\rm joint, no-bary} = 0.727^{+0.020}_{-0.016}$, a 0.25σ shift compared to the fiducial case. All these values are summarized in Table 4 and in Fig. 10 (with the brown symbols).

The dark energy equation-of-state is constrained from this joint analysis, with

$$w_0^{\rm joint} = -1.12^{+0.42}_{-0.31}\,, \qquad (14)$$

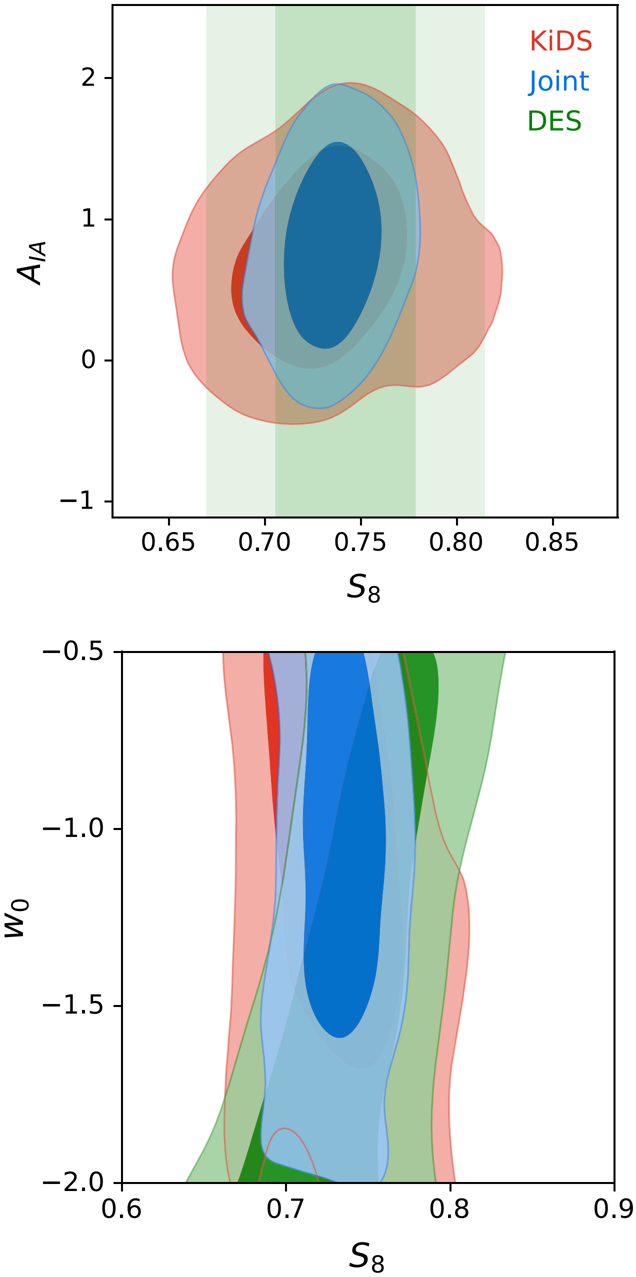

Figure 12. Joint peak count analysis of the KiDS-1000 and DES-Y1 data. In the upper panel, the vertical bands indicate the 1σ and 2σ confidence intervals from the DES-Y1 re-analysis presented in this paper. The $A_{\rm IA}$ parameter shown here describes only the IA signal in the KiDS likelihood, since the IA is fixed to a non-linear model in the DES likelihood (see the main text for details).

which is the first measurement of this quantity from peak statistics, and arguably one of the best from cosmic shear-only data analyses. The upper limit is close to the prior edge on w_0, which might lead to a slight underestimation of the error on this side. However, this measurement is robust against the choice of sampler ($w_0 = -1.09^{+0.29}_{-0.29}$ for MULTINEST), against baryon modelling ($w_0 = -1.05^{+0.51}_{-0.22}$ setting $b_{\rm bary} = 0.0$), IA ($w_0 = -1.13^{+0.44}_{-0.33}$ setting $A_{\rm IA} = 0.0$), and scale cuts ($w_0 = -0.958^{+0.45}_{-0.093}$ when including the aggressive $\mathcal{S}/\mathcal{N}$ cut in the KiDS-1000 data vector). As shown in Martinet et al. (2021a), aperture-mass maps statistics are highly sensitive to dark energy and these results seem to be showing exactly that. Previously, the shear two-point function measurement from Troxel et al. (2018) on DES-Y1 achieved $w_0 = -0.77^{+0.30}_{-0.37}$ when varying the baryonic feedback model, using the MULTINEST sampler. The GCNN analysis of Fluri et al. (2022) was also able to set constraints on dark energy, with $w_0 = -0.93^{+0.32}_{-0.29}$, although they recognize that their results are affected by the prior boundary on the low side, just like ours is on the high side.[24] Similarly, HD21 found $w_0 > -1.5$, also prior-dominated on one side. Other cosmic shear measurements of w_0 involve additional data (Tröster et al. 2021; Abbott et al. 2023), making this an unfair comparison.

It is worth mentioning that all peak count analyses based on the *cosmo*-SLICS yield S_8 constraints that are lower than the 2PCFs fiducial analyses. This could be pointing to limitations in the training set, but is quite speculative at this stage given that the Σ_8 values align well. Further investigations and novel simulation suites would be required to ascertain this, which we post-pone for future work.

6.3 Tension with *Planck*

The S_8 tension between recent CMB anisotropy and weak lensing data analyses is drawing a lot of attention, as it could point towards new physics or hidden systematics (see e.g. Abdalla et al. 2022, for a review). The *Planck* mission reports $S_8^{Planck} = 0.830 \pm 0.013$ (Planck Collaboration VI 2020), which is higher than many lensing results (see Amon et al. 2023, and references therein). The tension τ can be evaluated with a number of metrics, and we use here a relatively simple one used in A21, which compares the difference in the mean with the combined variances, $\text{var}[S_8]^{Planck}$, $\text{var}[S_8]^{\text{peaks}}$:

$$\tau = \frac{S_8^{Planck} - S_8^{\text{peaks}}}{\sqrt{\text{var}[S_8]^{Planck} + \text{var}[S_8]^{\text{peaks}}}} \qquad (15)$$

With the fiducial setup shown in Table 4, and using the above definition, our results from the KiDS-1000 peak count analysis are in $\tau = 2.0\sigma$ tension with the *Planck* nominal constraints on S_8 from their wCDM analysis. Tröster et al. (2021) finds a similarly low tension with the same KiDS-1000 data in a wCDM analysis, either using this simple tension metric or a more sophisticated method based on the full shape of the posteriors. Similarly, we evaluate our joint KiDS-DES analysis to be in $\tau = 2.7\sigma$ tension with *Planck*, an increase that is driven by the decrease in error bars. Using instead the MAP values (listed in Table 4) lowers the KiDS tension to 1.7σ, and the joint-survey to 1.65.

The tension reaches $\tau = 4.1\sigma$ in our joint-survey ΛCDM analysis, which could be pointing to a resolution of the S_8 tension that includes modification to the dark energy equation of state, as suggested by Tröster et al. (2021) and by the recent results from the Dark Energy Spectroscopic Instrument (DESI Collaboration 2024), however such statement cannot be definitive until better w measurements are obtained from cosmic shear data.

Note that this tension is not only seen in weak lensing, but also in other late-time probes, e.g. data involving galaxy clustering, as recently summarized in Alonso et al. (2023, see their fig. 7) and reviewed with greater details in Abdalla et al. (2022). The current work aligns with the existing trend, without providing an obvious solution. Again, large unaccounted contributions from IA and baryons could push the inferred S_8 value towards *Planck*, but our analysis prefers lower values: in particular, we measure $b_{\text{bary}} < 1.05$ at 95 per cent CL in the KiDS-1000 analysis, and < 0.82 in the joint analysis, excluding baryonic feedback models that are stronger than the *magneticum*. Also, the redshift estimation methods used to analyse the DES-Y1 data are suboptimal compared to recent developments (see e.g. Hildebrandt et al. 2021), potentially causing biases of up to 0.03 in the inferred S_8 value. The observed tension with *Planck* would be different if that bias was real.

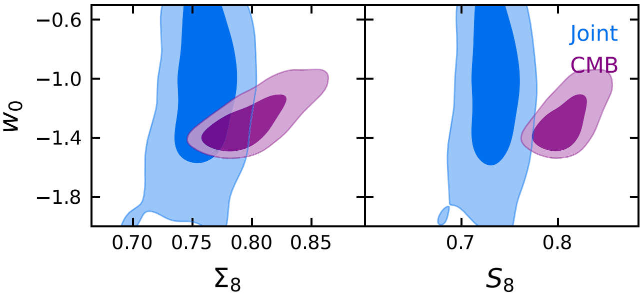

Figure 13. Joint constraint on S_8, w_0, and Σ_8^{α}, where $\alpha = 0.572$, comparing here the combined cosmic shear surveys with the CMB results. The tension is lower on Σ_8 than on S_8.

Interestingly, the tension on Σ_8 is reduced to $\tau = 0.72\sigma$ with the KiDS-1000 data, and to $\tau = 1.33\sigma$ with the joint data, in both cases evaluated at the α value preferred by the cosmic shear measurements.[25] This is in part caused by a degradation of the CMB constraints along this quantity, which completely relaxes the tension, and in part by a different projection angle of the high-dimensional posterior, as seen in Fig. 13. At the same time, when holding w fixed, the tension on Σ_8 is again increased reaching 2.3σ for KiDS, slightly lower than the 3σ reported in A21, and 3.1σ for our joint analysis. To summarize, the tension with *Planck* on S_8 is lowered in wCDM compared to ΛCDM, and is further lowered when considering the more robust Σ_8 parameter instead of S_8. This is discussed in more details in Appendix B.

A more in-depth analysis of this tension requires a robust determination of the inferred Ω_{m} parameter. We defer this improvement to future work, which will benefit from a denser training set, yielding a more accurate and robust emulator.

7 CONCLUSIONS

We report in this paper a 4.4 per cent measurement of S_8 from the tomographic peak count statistics measured from the KiDS-1000 data. Our simulation-based inference method exploits the non-linear features extracted from aperture mass statistics, sensitive to scales as small as 2.0 arcmin. We model the cosmological dependence with simulated wCDM weak lensing light cones, we estimate the covariance matrix numerically, and forward model the effect of intrinsic alignments, baryonic feedback, photometric redshift error, and galaxy shape miscalibrations. We find a value of $S_8^{\text{KiDS}} = 0.733^{+0.032}_{-0.032}$, which aligns well with previous KiDS-1000 measurements. We show that our results are robust to residual systematics and that, of these, intrinsic alignment of galaxies plays the most important

[24] We review the A21 definition that constraints are uninformed by the prior when the posterior drops below 0.135 of its maximum at the edges of a uniform prior volume. In the case of w_0 we find that the posterior is slightly above this threshold (0.156) at the upper edge. As this is at the borderline of the A21 criteria, we therefore caution that the error on this side might be slightly underestimated.

[25] The *Planck*wCDM measurements of Σ_8 are of $0.793^{+0.019}_{-0.031}$ and $0.797^{+0.018}_{-0.028}$ for the KiDS-1000 and joint-survey α parameters, respectively.

role, shifting the best-fitting S_8 value by 0.22σ if left unmodelled. We also show that the most robustly measured parameter in our analysis is $\Sigma_8^{\rm KiDS} \equiv \sigma_8\,(\Omega_{\rm m}/0.3)^\alpha = 0.765^{+0.030}_{-0.030}$, with $\alpha = 0.60$, in excellent agreement with previous KiDS-1000 analysis.

The inferred posterior distribution is consistent with the peak count measurement carried out on the DES-Y1 data using a similar analysis pipeline (HD21), allowing us to jointly analyse the two data sets, which yields $S_8^{\rm joint} = 0.732^{+0.020}_{-0.020}$, one of the tightest constraints on this parameter from lensing data alone. The combined data sets have enough statistical precision to allow the first measurement of the dark energy equation-of-state parameter from non-Gaussian statistics: $w_0^{\rm joint} = -1.12^{+0.42}_{-0.31}$, in agreement with the ΛCDM scenario, and robust to variations in the analysis choices.

Our best-fitting $S_8^{\rm joint}$ is also in statistical agreement with all previous KiDS-1000 analyses and with the HSC-Y3 and DES-Y3 γ-2PCF results, but lower than the DES-Y3 measurements from peaks and moments, and in 2.7σ tension with *Planck*. This joint-survey tension increases to 4σ in our ΛCDM analysis, but lowers to 2.3 and 3.1σ when considering instead the Σ_8 parameter, for the KiDS-only and joint survey analyses, respectively.

Our pipeline has been thoroughly tested, however we recognize it is incomplete. As detailed in Section 4, we hold fixed a number of cosmological parameters, which likely affect our results, including $\Omega_{\rm b}, n_{\rm s}$, and m_ν. We also consider a single baryonic feedback model (although we allow its amplitude to vary), knowing that other hydrodynamical simulations would provide slightly different responses. Furthermore, we model IA with the redshift-independent NLA model, which we know is an incomplete effective model, and we neglect source clustering, as it was shown to be completely subdominant. Finally, we have identified a systematic effect in our Gaussian process emulator that is caused by the relatively small number of training nodes, preventing us from extracting meaningful information about $\Omega_{\rm m}$, however all other parameters are unaffected by this. This limits our ability to further study in higher dimensional space the potential S_8 tension with the CMB. Addressing these will therefore be the object of future work. It will also be informative to compare our results to other non-Gaussian probes of cosmic shear, and possibly combine the methods to further reduce the uncertainty on S_8 and w_0.

ACKNOWLEDGEMENTS

We would like to thank Dr. Eric Tittley from the Royal Observatory of Edinburgh for his continuous help and support with the Cuillin cluster on which the bulk of the computations were carried on, as well as Elena Sellentin for discussions on the use of Hotelling's T^2 statistics.

JHD acknowledges support from an STFC Ernest Rutherford Fellowship (project reference ST/S004858/1). SH is supported by the U.D Department of Energy, Office of Science, Office of High Energy Physics under Award Number DE-SC0019301. BG acknowledges support from a UK STFC Consolidated Grant and from the Royal Society of Edinburgh through the Saltire Early Career Fellowship (ref. number 1914). NM acknowledges support from the French Agence Nationale de la Recherche (PISCO project/ANR-22-CE31-0004-01). TT acknowledges funding from the Swiss National Science Foundation under the Ambizione project PZ00P2_193352. KD is supported by the grant agreements ANR-21-CE31-0019/490702358 from the French Agence Nationale de la Recherche/Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for the LOCALIZATION project. TC is supported by the INFN INDARK PD51 grant, by the Fondazione ICSC National Recovery and Resilience Plan (PNRR) Project ID CN-00000013 'Italian Research Center on High-Performance Computing, Big Data and Quantum Computing' funded by MUR Missione 4 Componente 2 Investimento 1.4: 'Potenziamento strutture di ricerca e creazione di 'campioni nazionali' di R&S (M4C2-19)' – Next Generation EU (NGEU), and by the FARE MIUR grant 'ClustersXEuclid'. HH is supported by a DFG Heisenberg grant (Hi 1495/5-1), the DFG Collaborative Research Center SFB1491, as well as an ERC Consolidator Grant (No. 770935). BJ acknowledges support by the ERC-selected UKRI Frontier Research Grant EP/Y03015X/1 and by STFC Consolidated Grant ST/V000780/1. AHW is supported by the Deutsches Zentrum für Luft- und Raumfahrt (DLR), made possible by the Bundesministerium für Wirtschaft und Klimaschutz, and acknowledges funding from the German Science Foundation DFG, via the Collaborative Research Center SFB1491 'Cosmic Interacting Matters–From Source to Signal'.

The KiDS-1000 results in this paper are based on data products from observations made with ESO Telescopes at the La Silla Paranal Observatory under programme IDs 177.A-3016, 177.A-3017, and 177.A-3018, and on data products produced by Target/OmegaCEN, INAF-OACN, INAF-OAPD, and the KiDS production team, on behalf of the KiDS consortium. This project also used public archival data from the Dark Energy Survey (DES). 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 Michigan, the National Optical Astronomy Observatory, the University of Nottingham, The Ohio State University, the OzDES Membership Consortium, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, and Texas A&M University. Based in part on observations at Cerro Tololo Inter-American Observatory, National Optical Astronomy Observatory, which is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

The chains presented in this paper are plotted with GETDIST (Lewis 2019), available at `https://getdist.readthedocs.io`.

All authors contributed to the development and writing of this paper. The authorship list is given in three groups: the lead authors

(JHD, SH), followed by two alphabetical groups. The first alphabetical group includes those who are key contributors to both the scientific analysis and the data products (BG, NM, TT). The second group covers those who have either made a significant contribution to the data products, or to the scientific analysis.

DATA AVAILABILITY

The SLICS numerical simulations can be found at http://slics.roe.ac.uk/, while the SLICS-HR, the *cosmo*-SLICS, and the *magneticum* can be made available upon request. This work also uses public KiDS-1000 and DES-Y1 data, which can be found at `https://kids.strw.leidenuniv.nl/DR4/index.php` and `https://des.ncsa.illinois.edu/releases/y1a1`, respectively.

REFERENCES

Abbott T. M. C. et al., 2018, ApJS, 239, 18

Abbott T. M. C. et al., 2022, Phys. Rev. D, 105, 023520

Abbott T. M. C. et al., 2023, Phys. Rev. D, 107, 083504

Abdalla E. et al., 2022, J. High Energy Astrophys., 34, 49

Ajani V., Peel A., Pettorino V., Starck J.-L., Li Z., Liu J., 2020, Phys. Rev. D, 102, 103531

Akeson R. et al., 2019, preprint (arXiv:1902.05569)

Alarcon A. et al., 2021, MNRAS, 501, 6103

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

Alonso D., Fabbian G., Storey-Fisher K., Eilers A.-C., García-García C., Hogg D. W., Rix H.-W., 2023, JCAP, 2023, 043

Amon A. et al., 2022, Phys. Rev. D, 105, 023514

Amon A. et al., 2023, MNRAS, 518, 477

Anbajagane D. et al., 2023, MNRAS, 526, 5530

Angulo R. E., Hahn O., 2022, Living Rev. Comput. Astrophys., 8, 1

Angulo R. E., Zennaro M., Contreras S., Aricò G., Pellejero-Ibañez M., Stücker J., 2021, MNRAS, 507, 5869

Aricò G., Angulo R. E., Zennaro M., Contreras S., Chen A., Hernández-Monteagudo C., 2023, A&A, 678, A109

Asgari M. et al., 2020, A&A, 634, A127

Asgari M. et al., 2021, A&A, 645, A104

Begeman K., Belikov A. N., Boxhoorn D. R., Valentijn E. A., 2013, Exp. Astron., 35, 1

Benítez N., 2000, ApJ, 536, 571

Bilicki M. et al., 2021, A&A, 653, A82

Blazek J. A., MacCrann N., Troxel M. A., Fang X., 2019, Phys. Rev. D, 100, 103506

Bridle S., King L., 2007, J. Phys., 9, 444

Brouwer M. M. et al., 2018, MNRAS, 481, 5189

Burger P., Friedrich O., Harnois-Déraps J., Schneider P., 2022, A&A, 661, A137

Burger P. A. et al., 2024, A&A, 683, A103

Castro T., Quartin M., Giocoli C., Borgani S., Dolag K., 2018, MNRAS, 478, 1305

Chintalapati P. R. V., Gutierrez G., Wang M. H. L. S., 2022, Phys. Rev. D, 105, 043515

DESI Collaboration 2024, preprint (arXiv:2404.03002)

Dalal R. et al., 2023, Phys. Rev. D, 108, 123519

Dark Energy Survey and Kilo-Degree Survey Collaboration, 2023, Open J. Astrophys., 6, 36

Duncan C. A. J., Harnois-Déraps J., Miller L., Langedijk A., 2022, MNRAS, 515, 1130

Edge A., Sutherland W., Kuijken K., Driver S., McMahon R., Eales S., Emerson J. P., 2013, The Messenger, 154, 32

Erben T. et al., 2013, MNRAS, 433, 2545

Euclid Collaboration: Ajani V. et al., 2023, A&A, 675, A120

Euclid Collaboration: Knabenhans M. et al., 2019, MNRAS, 484, 5509

Fenech Conti I., Herbonnet R., Hoekstra H., Merten J., Miller L., Viola M., 2017, MNRAS, 467, 1627

Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601

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

Fortuna M. C., Hoekstra H., Joachimi B., Johnston H., Chisari N. E., Georgiou C., Mahony C., 2021, MNRAS, 501, 2983

Fosalba P., Gaztañaga E., Castander F. J., Manera M., 2008, MNRAS, 391, 435

Fosalba P., Gaztañaga E., Castander F. J., Crocce M., 2015, MNRAS, 447, 1319

Fu L. et al., 2014, MNRAS, 441, 2725

Gatti M. et al., 2020, MNRAS, 498, 4060

Gatti M. et al., 2024a, MNRAS, 527, L115

Gatti M. et al., 2024b, preprint (arXiv:2405.10881)

Giblin B. et al., 2018, MNRAS, 480, 5529

Giblin B. et al., 2021, A&A, 645, 105

Giblin B., Cai Y.-C., Harnois-Déraps J., 2023, MNRAS, 520, 1721

Gruen D., Brimioulle F., 2017, MNRAS, 468, 769

Gruen D. et al., 2018, Phys. Rev. D, 98, 023507

Gupta A., Nagar D., 1999, Matrix Variate Distributions, Monographs and Surveys in Pure and Applied Mathematics. Chapman and Hall/CRC, London

Hamana T. et al., 2020, PASJ, 72, 16

Harnois-Déraps J., van Waerbeke L., 2015, MNRAS, 450, 2857

Harnois-Déraps J., Giblin B., Joachimi B., 2019, A&A, 631, A160

Harnois-Déraps J., Martinet N., Castro T., Dolag K., Giblin B., Heymans C., Hildebrandt H., Xia Q., 2021, MNRAS, 506, 1623(HD21)

Harnois-Déraps J., Martinet N., Reischke R., 2022, MNRAS, 509, 3868

Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399

Heitmann K., Lawrence E., Kwan J., Habib S., Higdon D., 2014, ApJ, 780, 111

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

Hilbert S. et al., 2020, MNRAS, 493, 305

Hildebrandt H. et al., 2021, A&A, 647, A124

Ho M.-F., Bird S., Shelton C. R., 2022, MNRAS, 509, 2551

Hoekstra H., Herbonnet R., Muzzin A., Babul A., Mahdavi A., Viola M., Cacciato M., 2015, MNRAS, 449, 685

Hotelling H., 1931, Ann. Math. Stat., 2, 360

Hoyle B. et al., 2018, MNRAS, 478, 592

Ivezić Ž. et al., 2019, ApJ, 873, 111

Izard A., Fosalba P., Crocce M., 2018, MNRAS, 473, 3051

Jarvis M., Bernstein G., Jain B., 2004, MNRAS, 352, 338

Jeffrey N., Alsing J., Lanusse F., 2021, MNRAS, 501, 954

Joachimi B. et al., 2021, A&A, 646, A129

Joudaki S. et al., 2020, A&A, 638, L1

Kacprzak T. et al., 2016, MNRAS, 463, 3653

Kannawadi A. et al., 2019, A&A, 624, A92

Kilbinger M., 2015, Rep. Prog. Phys., 78, 086901

Kilbinger M., Schneider P., Eifler T., 2006, A&A, 457, 15

Kilbinger M. et al., 2017, MNRAS, 472, 2126

Kuijken K. et al., 2015, MNRAS, 454, 3500

Kuijken K. et al., 2019, A&A, 625, A2

Laigle C. et al., 2016, ApJS, 224, 24

Lange J. U., 2023, MNRAS, 525, 3181

Laureijs R. et al., 2011, preprint (arXiv:1110.3193)

Lemos P. et al., 2023, MNRAS, 521, 1184

Lewis A., 2019, preprint (arXiv:1910.13970)

Li Z., Liu J., Zorrilla Matilla J. M., Coulton W. R., 2019, Phys. Rev. D, 99, 063527

Li X. et al., 2023a, Phys. Rev. D, 108, 123518

Li S.-S. et al., 2023b, A&A, 679, A133

Lima M., Cunha C. E., Oyaizu H., Frieman J., Lin H., Sheldon E. S., 2008, MNRAS, 390, 118

Lin K., von wietersheim-Kramsta M., Joachimi B., Feeney S., 2023, MNRAS, 524, 6167
Linke L., Burger P. A., Heydenreich S., Porth L., Schneider P., 2024, A&A, 681, A33
Liu X., Yuan S., Pan C., Zhang T., Wang Q., Fan Z., 2023, MNRAS, 519, 594
Longley E. P. et al., 2023, MNRAS, 520, 5016
Loureiro A. et al., 2022, A&A, 665, A56
MacCrann N. et al., 2022, MNRAS, 509, 3371
Mandelbaum R., 2018, ARA&A, 56, 393
Marques G. A. et al., 2024, MNRAS, 528, 4513
Martinet N. et al., 2018, MNRAS, 474, 712
Martinet N., Harnois-Déraps J., Jullo E., Schneider P., 2021a, A&A, 646, A62
Martinet N., Castro T., Harnois-Déraps J., Jullo E., Giocoli C., Dolag K., 2021b, A&A, 648, A115
McCarthy I. G., Schaye J., Bird S., Le Brun A. M. C., 2017, MNRAS, 465, 2936
Mead A. J., Heymans C., Lombriser L., Peacock J. A., Steele O. I., Winther H. A., 2016, MNRAS, 459, 1468
Miller L. et al., 2013, MNRAS, 429, 2858
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Percival W. J., Friedrich O., Sellentin E., Heavens A., 2022, MNRAS, 510, 3207
Planck Collaboration VI, 2020, A&A, 641, A6
Pyne S., Joachimi B., 2021, MNRAS, 503, 2300
Schirmer M., Erben T., Hetterscheidt M., Schneider P., 2007, A&A, 462, 875
Schneider P., 1996, MNRAS, 283, 837
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
Sellentin E., Heavens A. F., 2016, MNRAS, 456, L132
Shan H. et al., 2018, MNRAS, 474, 1116
Sheldon E. S., Huff E. M., 2017, ApJ, 841, 24
Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, ApJ, 761, 152
Takahashi R., Hamana T., Shirasaki M., Namikawa T., Nishimichi T., Osato K., Shiroyama K., 2017, ApJ, 850, 24
Tröster T. et al., 2021, A&A, 649, A88
Troxel M. A. et al., 2018, Phys. Rev. D, 98, 043528
van Waerbeke L. et al., 2013, MNRAS, 433, 3373
van den Busch J. L. et al., 2020, A&A, 642, A200
van den Busch J. L. et al., 2022, A&A, 664, A170
Vakili M. et al., 2023, A&A, 675, A202
Wright A. H., Hildebrandt H., van den Busch J. L., Heymans C., Joachimi B., Kannawadi A., Kuijken K., 2020, A&A, 640, L14
Xavier H. S., Abdalla F. B., Joachimi B., 2016, FLASK: Full-sky Log-normal Astro-fields Simulation Kit, Astrophysics Source Code Library, record ascl:1606.015
Zorrilla Matilla J. M., Waterval S., Haiman Z., 2020, AJ, 159, 284
Zuntz J. et al., 2015, Astron. Comput., 12, 45
Zürcher D., Fluri J., Sgier R., Kacprzak T., Refregier A., 2020, JCAP, 1, 028
Zürcher D. et al., 2022, MNRAS, 511, 2075

APPENDIX A: *B*-MODES

To leading order, B-modes are not produced by gravitational lensing, hence their detection in cosmic shear data is generally regarded as an indication of residual systematics. As mentioned in Section 4.6.6, the aperture mass map statistics constructed on a grid inevitably induces B-modes from the missing subpixel contributions, resulting in a non-zero $M_\times$ signal. This section presents a careful investigation of the amplitude, origins, and consequences of these induced B-modes. In particular, and we find that finite sampling of the shear field itself is also a source of B-modes in aperture mass maps, on top of pixelization.

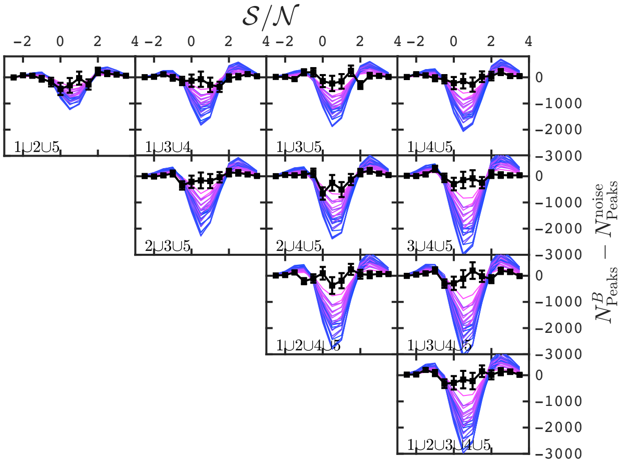

Figure A1. Noise-subtracted peak count statistics measured from B-modes data (black squares) for a representative subset of the 30 tomographic bins, compared with the E-modes cosmological predictions.

We first quantify here the strength of these effects by measuring the peak function from $M_\times(\boldsymbol{\theta})$ in our data, i.e. aperture mass maps in which the galaxies are rotated by 45 deg. The (noise-subtracted) signal $N^B_{\rm peaks}$ is shown in Fig. A1 for a representative subset of the tomographic bins. We observe that the residual signal is much flatter than what we would expect from a cosmological signal consistent with pure noise with a p-value of $p = 0.12$, above the threshold of $p = 0.01$ (the same threshold is used in the main text, and in the DES-Y3 results for this type of hypothesis testing, see Appendices G and D of Abbott et al. 2022; Zürcher et al. 2022, respectively). This agrees with A21, namely that there is no evidence of residual B-modes in the KiDS-1000 data. It is therefore safe to keep all data entries in our inference, but investigate further the source of the $M_\times(\boldsymbol{\theta})$ signal seen by eye in Fig. A1 to confirm it is not problematic.

We carried out peak count measurements of $M_\times(\boldsymbol{\theta})$ on 20 full survey realizations from the *covariance training set* (again, these are pure E-mode mocks rotated by 45 degrees for this exercise), expecting to find large p-values in all of these trials. Instead, this test revealed that p-values range from 10^{-10} to 0.1. Some of these trials seem to rule out completely the null hypothesis (that the B-modes are consistent with pure noise), even though no B-mode exists at the catalogue level. The observed $M_\times$ signal must therefore come from the aperture map method itself, and is consequently a poor test for residual observational systematics.

Interestingly, the measured $N^B_{\rm peaks}$ averaged over 20 noise realizations has a p-value of 1.0, namely $\langle N^B_{\rm peaks}\rangle = N^{\rm noise}_{\rm peaks}$, suggesting that these B-modes contain mostly noise, even though on a case-by-case some realizations see strong deviations. We hypothesize that this stems from E-modes leaking into B-modes due to an incomplete knowledge of the shear field: assuming a noiseless, pure E-mode shear field, the average cross-shear $\gamma_\times$ on a circle around every point in the field vanishes by definition, and thus $M_\times(\boldsymbol{\theta}) \equiv 0$ holds everywhere. However, that is no longer guaranteed once the shear field is only known at a discrete set of positions, as the average $\gamma_\times$ on a circle no longer necessarily vanishes. We investigate this by varying the number of source galaxies in our simulations. We achieve this by running our measurements on Stage-IV mocks created with a number density of 30.0 arcmin, introduced in Heydenreich et al. (2021), without any tomographic split. We measure on these the $M_\times$ signal from maps sampled (a) at every pixel location, (b) at all galaxy positions, and (c) at galaxy positions downsampled to match

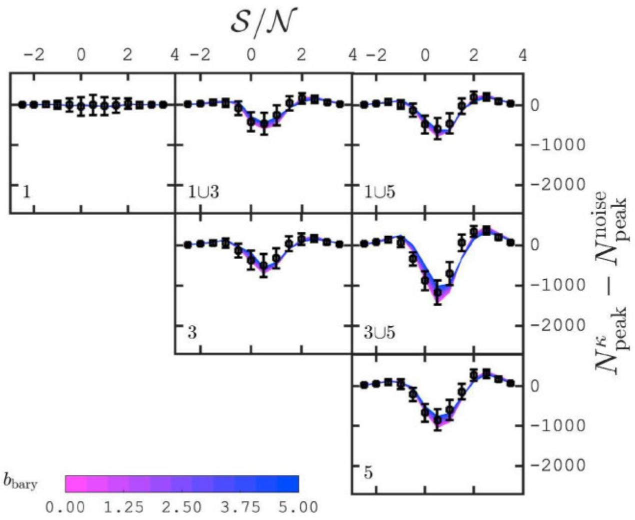

Figure B1. Tomographic weak lensing peak function in the *baryons training set*. The coloured lines are obtained by scaling the GPR predictions (at the *magneticum* cosmology) by the $b_{\rm bary}$ parameter, over the full prior range, demonstrating that peak statistics are fairly insensitive to changes in baryon feedback.

the KiDS-1000 number density. In the first case, we find that the B-mode field $M_{\times}(\boldsymbol{\theta})$ vanishes completely (up to numerical precision). The second case induces B-modes of approximately 0.5 per cent of the E-mode signal, whereas the third case (KiDS-1000-like number density) yields B-modes of approximately 4 per cent of the E-mode signal. We note that this is likely to be exacerbated by splitting the galaxies into different tomographic bins, which further decreases the number density per aperture. We further note that these tests were performed in the absence of shape noise to better isolate this effect.

More importantly, since these non-zero $N^{B}_{\rm peaks}$ are caused by finite sampling of the shear field, and that this sampling is exactly the same for the data and the *cosmology training set*, the same amount of leakage should occur on average. In particular, this should be fully converged in simulation-based model once averaged over the 50 mock survey × 10 noise realizations per cosmology (20 was shown to be enough in the discussion above). Therefore our inference must be immune to these by construction.

APPENDIX B: VALIDATION OF THE COSMOLOGY INFERENCE PIPELINE

In this section we present a series of validation tests we performed on our cosmology inference pipeline. First, we verified that the derivatives $\partial N_{\rm peaks}/\partial \Delta m_a$, $\partial N_{\rm peaks}/\partial \Delta z_a$, $\partial N_{\rm peaks}/\partial b_{\rm bary}$, and $\partial N_{\rm peaks}/\partial A_{\rm IA}$ are consistent with the results found in HD21 and Harnois-Déraps et al. (2022). The element-by-element values are different since these are survey-specific, but they agree qualitatively. We next verified that the impact of increased N-body force is of no consequence, consistent with HD21. This is achieved by carrying the inference with the *validation set* (high-resolution) instead of the mean of the *covariance training set*, as done in HD21.

We also verified that the peak function measured from the *baryons training set* is consistent with the *cosmology training set* in Fig. B1. This is an important test, as the baryon mocks based on a completely independent N-body code. This figures also shows the impact of varying $b_{\rm bary}$ on the data vector. Stronger feedback models (purple) tend to have fewer large peaks ($\mathcal{S}/\mathcal{N} > 2$), and more in the range $-1 < \mathcal{S}/\mathcal{N} < 1.5$. This is best seen in panel 3∪5 but a common feature to most panels. A few points lie outside the GPR predictions, which suggest that differences between N-body solvers/ray-tracing codes have a non-negligible impact on the data vector. We investigate this below, but first we examine in Fig. B2 the accuracy of our KiDS-1000 cosmology inference pipeline by showing the recovery of input parameters for two different cosmologies selected from the *cosmology training set*: the fiducial ΛCDM model as well as wCDM model 12. We find again an excellent agreement, except that the double peak solution in $\Omega_{\rm m}$ when analysing the former model. This was first seen with the *validation training set* in Fig. 6, but is absent from model 12; it therefore seems to be a cosmology-dependent feature, most likely associated with limits in our GPR emulation.

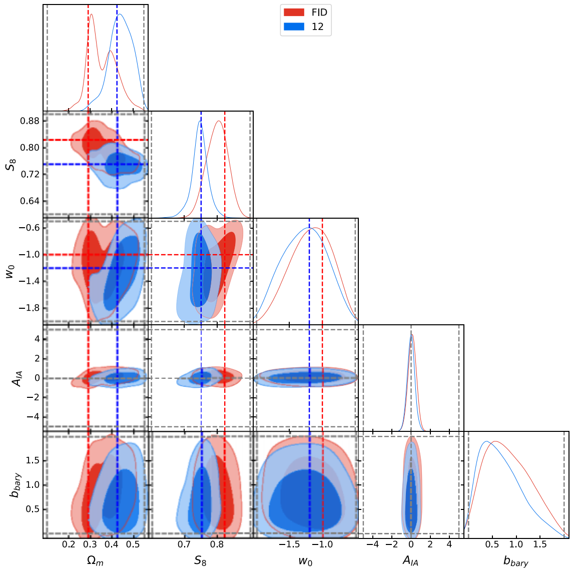

Figure B2. Cosmological constraints inferred from mock data vectors extracted from two of our *cosmology training set* models.

No parts of the data vector can easily explain the double peak solution in $\Omega_{\rm m}$. We have examined the posterior distributions resulting from our likelihood sampling and identified three tomographic bins (bins 1, 1∪2, and 1∪2∪3) where the data points were scattering outside the posteriors. Removing these from the analysis made minor differences, slightly broadening the contours; we therefore do not deem justified to remove them from the main analysis.

When inferring the cosmology from the *magneticum* directly or the T17 simulations, the results on all cosmological parameters are severely biased, as seen by the red contours in Fig. B3. As discussed earlier, this is likely caused by differences in the resolution of the N-body calculations and/or the ray-tracing algorithm being used in the creation of these mocks. In particular, the *magneticum* and T17 mocks both have lower mass resolution than the main *cosmo*-SLICS training set, which inevitably affects the accuracy of their measured peak statistics at scales as non-linear as those targeted by this analysis. One way to avoid these biases is to calibrate our emulator and explicitly enforce the desired data vector (*magneticum or T17*) at some point in parameter space. This can be achieved, for instance, by multiplying our theory by a calibration factor computed from the target data vector and the GPR prediction at the target cosmology. This is shown as the blue contours in Fig. B3, where the input cosmology is now correctly recovered, but the double $\Omega_{\rm m}$ solution persists, even if varying only that single parameter in the MCMC. We note that the *magneticum* and T17 mocks require

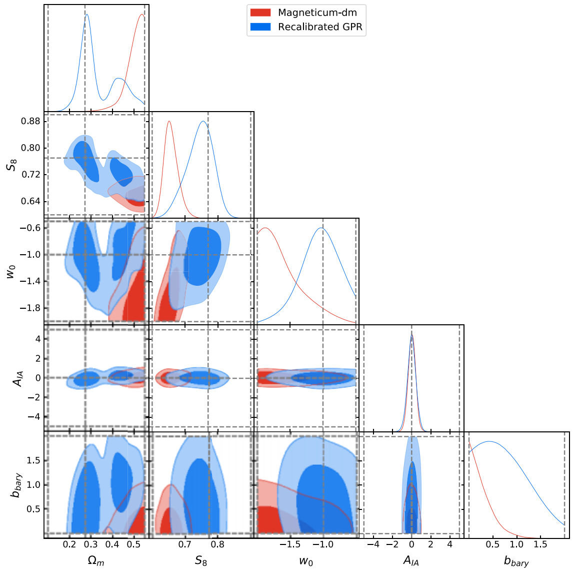

Figure B3. Cosmological constraints inferred from mock data vectors extracted from the *magneticum* dark matter-only model, with (blue) and without (red) recalibrating the emulator (see the main text). Similar results are obtained with the T17 simulations.

distinct calibrations, and that swapping them yields results almost as biased as the original case, due to their differences in small scales resolution. Because of this non-universality we do not calibrate the prediction in our main analysis, but optionally include the spread in these correction factors in the covariance matrix, accounting for added uncertainty about small scales physics.

The important conclusions drawn from these tests are that (1) small-scales structures that are not fully resolved or converged in N-body simulations greatly affect the non-Gaussian statistics we are investigating here, hence future analyses with increased precision will need to pay particular attention to such considerations, and (2) the sparsity of our cosmological training nodes impacts the GPR emulator mostly on the $\Omega_{\rm m}$ dimension, while all other cosmological parameters are well recovered. This means that the current analysis is robust in its measurements of S_8, w_0, $A_{\rm IA}$, and $b_{\rm bary}$, however our constraints on the matter density are unstable and hence we do not report on them. Since we use the same training nodes for the KiDS and DES analyses, this applies also to the joint-survey constraints.

We finally tested the joint-survey pipeline with the *validation set* defined for both KiDS-1000 and DES-Y1 analyses, and report our results in Fig. B4. We observe that it recovers very well the input truth: the best-fitting value is $S_8 = 0.818^{+0.030}_{-0.025}$, the maximum-likelihood value is 0.831, while the truth is 0.813. These results are obtained from wCDM pipeline assuming the 'clean' selection of $\mathcal{S}/\mathcal{N}$ bin, marginalizing over all nuisance parameters. All input parameters are accurately recovered, and we see here again the double $\Omega_{\rm m}$ solution, demonstrating that this parameter is subject to artificial degeneracies caused by our emulator, supporting our choice to not trust nor report its value in the main analysis. We also verified that the chain elements that fall in the secondary solution also tend to have a lower S_8 by about 0.03, which is larger than our statistical precision, however these are highly suppressed compared to the KiDS-1000 only pipeline, justifying our choice not to include this as a standalone systematic error. Again, the secondary solution yields unbiased Σ_8 inference, making the latter a more robust statistics.

We illustrate this last point in Fig. B5 where we present the projected posteriors on σ_8, S_8, and Σ_8 versus $\Omega_{\rm m}$ for our analyses of the KiDS-1000 data and of the *validation training set*. The KiDS-1000 inference prefers large $\Omega_{\rm m}$ values, consistent with being drawn from the secondary solution discussed earlier. If that is the case, the inferred value of S_8 might be biased low, but Σ_8 is robust. To illustrate this, we split the fiducial MCMC chain of the simulation analysis into low- and high-$\Omega_{\rm m}$ regions, and recover that both yield the similar Σ_8 posteriors, while their S_8 values differ by up to 0.03. We also show in this figure how the tension with *Planck* evolves under these change of variables.

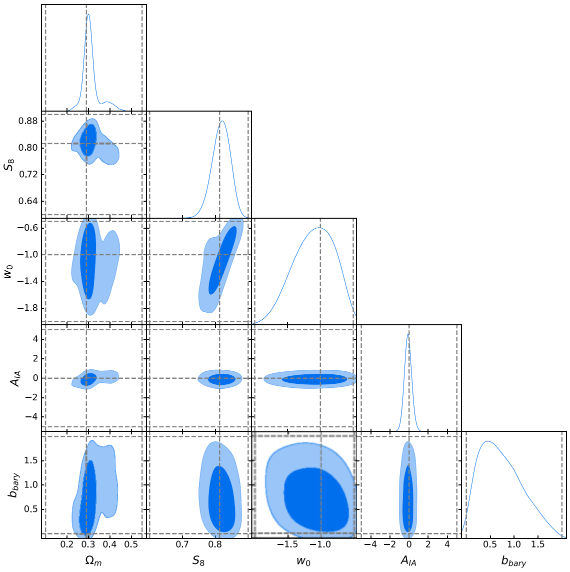

Figure B4. Joint-survey mock analysis of the *validation training set*.

APPENDIX C: GOODNESS-OF-FIT FOR STUDENT-*t* LIKELIHOODS

Noisy numerical covariance matrices need to be treated carefully in likelihood analyses to avoid biases incurred during the inversion. A commonly used approach is to debias the inverse matrix with the Hartlap-Anderson coefficient (Hartlap et al. 2007), however this often leads to overestimates in the contours. Instead, Sellentin & Heavens (2016) suggested to replace the Hartlap-corrected multi-variate Gaussian likelihood by a Student-*t* distribution, which better accounts for the noise present when estimating the matrix from $N_{\rm sims}$ realizations of the data.

Once the likelihood has been sampled and the best-fitting parameters found, one of the key subsequent steps is to estimate the goodness-of-fit. This is usually achieved by means of the p-value, which determines how likely it is that the difference between the best-fitting model and the measured data is due to a random noise fluctuation. Given the number of degrees of freedom ν, best-fitting χ^2 measurements from data that is well described by a multi-Gaussian likelihood will be sampled from a χ^2_ν distribution. Using this metric with noisy numerical covariance matrices will yield p-values that are biased towards low values if the inverse matrix is not Hartlap-corrected. Conversely, if corrected, the p-values are at risk to be on the high-side (Sellentin & Heavens 2016).

This is demonstrated by a toy model, which is created to follow our analysis: we generate a matrix A with 210^2 Gaussian random numbers (the same dimension as our KiDS-1000 analysis) and define

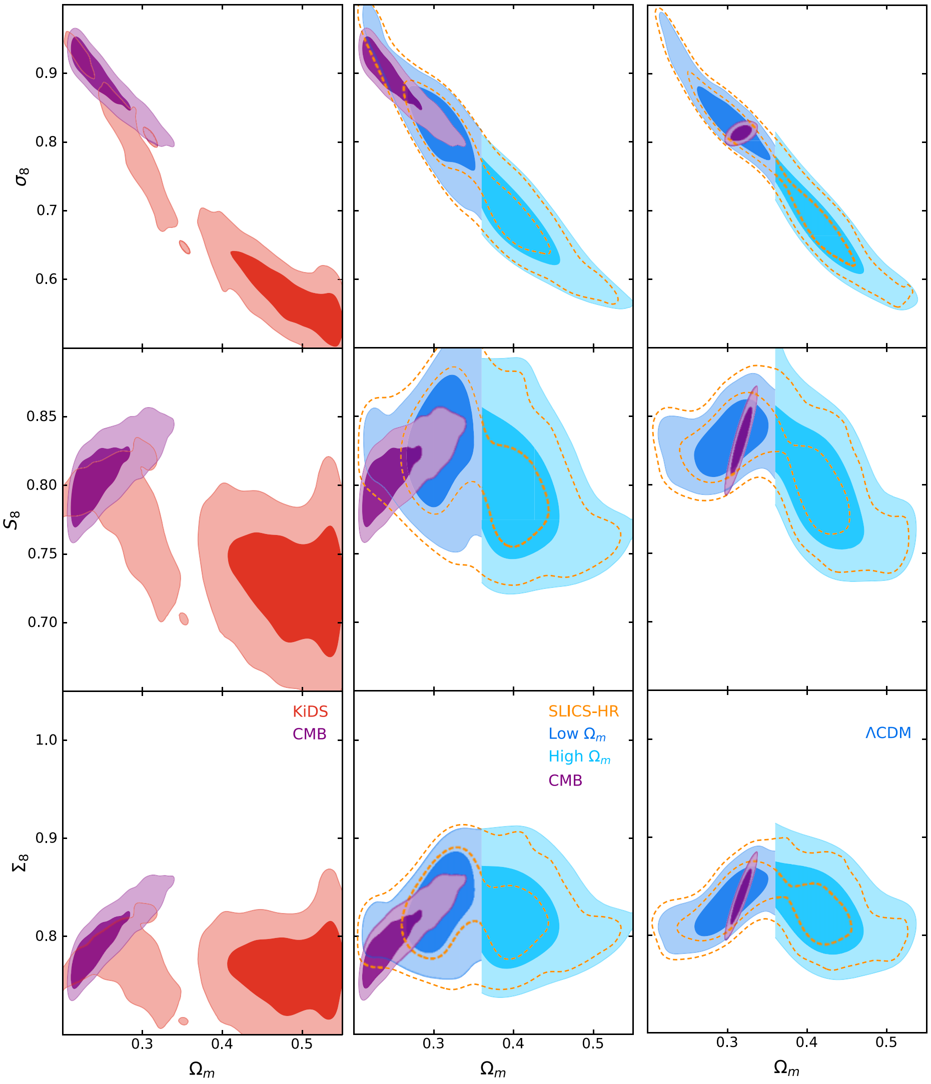

Figure B5. (*left panel:*) KiDS-1000 analysis: posterior distributions on the matter density and on the three clustering parameters (σ_8, S_8, Σ_8). The inferred value of $\Omega_{\rm m}$ from the data is higher than in previous KiDS-1000 analyses, consistent from being drawn from the biased secondary solution discussed in the main text, caused by our emulator. Overplotted in purple are the *Planck* results, which are in mild tension with KiDS for the first two clustering parameters, but in full agreement with Σ_8. (*middle:*) Mock analysis, here carried out on the *validation training set* (orange, dashed). Our sample is further split into low- and high-$\Omega_{\rm m}$ parts (dark and pale blue, respectively), highlighting the residual [$S_8 - \Omega_{\rm m}$] degeneracy, which vanishes for Σ_8. (*right:*) Same as middle panel, but here showing ΛCDM analyses.

a ‘true’ covariance matrix $\Sigma = A^T A$. We also define the ‘true’ data vector as the zero-vector.

Afterwards, the following procedure is repeated 10 000 times: we generate 1240 realizations of a multivariate normal distribution with mean 0 and covariance Σ, from which we estimate our sample covariance matrix $\mathbf{C}$, mimicking the *covariance training set*. We then also draw one additional realization $\mathcal{X}$ of the same multivariate normal distribution, which constitutes our measurement. Finally, we calculate the p-value given $\mathcal{X}$ and $\mathbf{C}$ and a chosen p-value test, assuming that the degrees of freedom equal the number of elements in the data vector (since our toy model contains no free parameters).

This procedure yields 10 000p-values which, if the chosen test is appropriate for our analysis, form a uniform distribution between 0 and 1. As can be seen in the lower panel of Fig. C1, the χ^2-based p-value tests are heavily biased towards 0, as is expected. The Hartlap-corrected p-values are more uniformly distributed, but still show slight biases towards 0 and 1, and consequently a reduced probability towards central values. Although this effect is relatively weak for our setup, it becomes more prominent if the degrees of freedom increase. Nevertheless, this means that a Hartlap-corrected p-value test is more likely to favour extreme values, but it appears to be relatively robust.

An unbiased solution to this problem can be achieved by deriving the sampling distribution of our quadratic statistics[26] defined in equation (6), specifically:

$$T^2_{\rm best-fit} = (\boldsymbol{d} - \boldsymbol{x}(\boldsymbol{\pi}_{\rm best-fit}))^T\, \mathbf{C}^{-1}\, (\boldsymbol{d} - \boldsymbol{x}(\boldsymbol{\pi}_{\rm best-fit}))\ , \quad \text{(C1)}$$

[26] The quadratic statistics described by equation (6) should not be labelled ‘χ^2’ unless it is sampling a χ^2_ν distribution. We used the χ^2 notation in the main text only to align with the notation in the weak lensing literature.

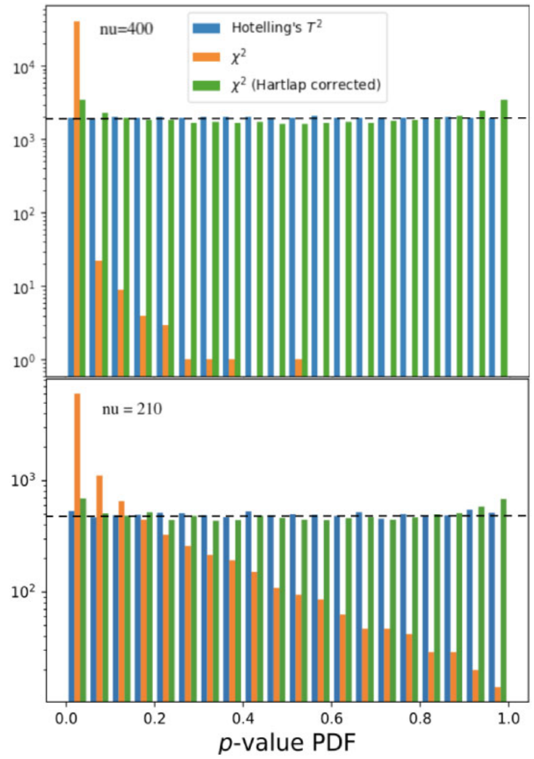

Figure C1. Distribution of p-values extracted from our toy examples based on three commonly used goodness-of-fit statistics, for $\nu = 210$ (lower) and $\nu = 400$ (upper). Given a noisy numerical covariance matrix, only the Hotelling's T^2 distribution returns p-values evenly sampling the range [0, 1]; the χ^2 distribution (orange) is heavily skewed towards low p-values, while the Hartlap-corrected χ^2 (green) is slightly skewed towards extrema p-values.

where the data $\boldsymbol{d}$ has dimension p and is drawn from a normal distribution $\boldsymbol{d} \sim N(\boldsymbol{\mu}, \Sigma)$, for unknown mean $\boldsymbol{\mu}$ and unknown covariance Σ, and the $\mathbf{C}$ covariance is drawn from a Wishart distribution with $N_{\rm sims} - 1$ degrees of freedom $(N_{\rm sims} - 1)\mathbf{C} \sim W_p(\Sigma, N_{\rm sims} - 1)$.

We now define $LL^T = \Sigma^{-1}$ and $\boldsymbol{w} = L\,(\boldsymbol{d} - \boldsymbol{x}(\boldsymbol{\pi}_{\rm best-fit}))$, such that $\mathrm{Cov}\,[\boldsymbol{w}, \boldsymbol{w}] = 1_p$, the $p \times p$ identity matrix. With this, we can express equation (C1) as

$$T^2_{\rm best-fit} = (N_{\rm sims} - 1)\boldsymbol{w}^T V^{-1} \boldsymbol{w}\ , \tag{C2}$$

where we have defined $V = (N_{\rm sims} - 1)LCL^T$ and note that $V \sim W_p(1_p, N_{\rm sims} - 1)$. We can now introduce an orthogonal matrix $M^T M = 1_p$ with the first row being $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$ and the others orthogonal to it, such that

$$M\boldsymbol{w} = \begin{pmatrix} \|\boldsymbol{w}\| \\ 0 \\ \vdots \\ 0 \end{pmatrix} . \tag{C3}$$

Conditional on M, we have that $Q = MVM^T \sim W_p(1_p, N_{\rm sims} - 1)$. Since this does not depend on M, $Q \sim W_p(1_p, N_{\rm sims} - 1)$ also holds in the unconditional case. With this transformation, equation (C2) can be written as

$$T^2_{\rm best-fit} = (N_{\rm sims} - 1)\|\boldsymbol{w}\|^2 \left(Q^{-1}\right)^2_{11}\ , \tag{C4}$$

with $\left(Q^{-1}\right)^2_{11}$ being the 1–1 entry of Q^{-1}. Writing out $\|\boldsymbol{w}\|^2$ as

$$\begin{aligned} \boldsymbol{w}^T \boldsymbol{w} &= (\boldsymbol{d} - \boldsymbol{x}(\boldsymbol{\pi}_{\rm best-fit}))^T\, L^T L\, (\boldsymbol{d} - \boldsymbol{x}(\boldsymbol{\pi}_{\rm best-fit})) \\ &= (\boldsymbol{d} - \boldsymbol{x}(\boldsymbol{\pi}_{\rm best-fit}))^T\, \Sigma^{-1}\, (\boldsymbol{d} - \boldsymbol{x}(\boldsymbol{\pi}_{\rm best-fit}))\ , \end{aligned} \tag{C5}$$

we recognize this as the usual χ^2 quantity where the true covariance Σ is assumed to be known. In other words, $\|\boldsymbol{w}\|^2 \sim \chi^2_\nu$, with $\nu = p - n_{\rm eff}$, where $n_{\rm eff}$ is the effective number of free parameters that are being varied when finding $\boldsymbol{\pi}_{\rm best-fit}$, which accounts for the fact that the model may contain both constrained and unconstrained parameters.

Returning to the last term in equation (C4), we have

$$\left(Q^{-1}\right)^{-1}_{11} = \frac{1}{\left(Q^{-1}\right)_{11}} = Q_{11} - Q_{12} Q_{22}^{-1} Q_{21}\ , \tag{C6}$$

where Q_{12} and Q_{22} are the $1 \times (p - 1)$ and $(p - 1) \times (p - 1)$ submatrices of Q. From this it follows (e.g. Gupta & Nagar 1999) that $\frac{1}{(Q^{-1})_{11}} \sim W_1(I_1, N_{\rm sims} - p) = \chi^2_{N_{\rm sims}-p}$. Putting things together, we therefore have that

$$T^2_{\rm best-fit} \sim (N_{\rm sims} - 1)\frac{\chi^2_{p-n_{\rm eff}}}{\chi^2_{N_{\rm sims}-p}} = \frac{(N_{\rm sims}-1)(p - n_{\rm eff})}{(N_{\rm sims}-p)} F_{p-n_{\rm eff}, N_{\rm sims}-p}, \tag{C7}$$

where $F_{p-n_{\rm eff}, N_{\rm sims}-p}$ is the F-distribution. For the case of no free parameters, $n_{\rm eff} = 0$, this reduces to Hotelling's T^2 distribution (Hotelling 1931).

To calculate a p-value, one just has to evaluate the cumulative distribution function of the F-statistics at $\frac{N_{\rm sims}-p}{(p-n_{\rm eff})(N_{\rm sims}-1)}T^2$, replacing the $\chi^2(p - n_{\rm eff})$ distribution. Clearly seen in Fig. C1, this constitutes an ideal solution for our toy model, so we use this statistics to estimate the p-values of our measurements.

When applied to our fiducial KiDS-1000 peak count data vector, along with the best-fitting model and our numerical covariance matrix, we obtain p-values of 0.43 with the (unbiased) Hotelling's statistics and for the (slightly biased) Hartlap-corrected χ^2 approach, and 0.02 for the (heavily biased) normal χ^2 statistics.

In this toy example, the difference between the Hotelling's and the Hartlap-corrected p-values distributions is quite small, however this is not always the case. The upper panel of Fig. C1 shows a second case where now the number of degrees of freedom has been increased to $\nu = 400$, overshooting our joint-survey setup, but close to typical sizes of data vectors used in 2pt statistics. Keeping $N_{\rm sims}$ unchanged, in this case the Hartlap-corrected distribution shows a clear excess towards low and high p-values. The Hotelling's distribution is still flat however, showcasing the advantage of the F-statistics.

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