

KiDS-1000 cosmic shear reanalysis using METACALIBRATION

Mijin Yoon,^{1,2*} Henk Hoekstra,¹ Shun-Sheng Li,^{1,2} Konrad Kuijken,¹ Lance Miller³, Hendrik Hildebrandt², Catherine Heymans^{2,5}, Benjamin Joachimi⁴, Angus H. Wright², Marika Asgari⁶, Jan Luca van den Busch², Robert Reischke^{2,7}, and Benjamin Stölzner²

¹ Leiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, the Netherlands

² Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute (AIRUB), German Centre for Cosmological Lensing, 44780 Bochum, Germany

³ Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK

⁴ Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

⁵ Institute for Astronomy, University of Edinburgh, Blackford Hill, Edinburgh, EH9 3HJ, UK

⁶ School of Mathematics, Statistics and Physics, Newcastle University, Herschel Building, NE1 7RU, Newcastle-upon-Tyne, UK

⁷ Argelander-Institut für Astronomie, Universität Bonn, Auf dem Hügel 71, D-53121 Bonn, Germany

October 2, 2025

ABSTRACT

A number of cosmic shear studies have reported results that are in mild tension with the Planck cosmic microwave measurement. To explore if this can be caused by biases in the shear estimation, we revisit the analysis of data from the Kilo-Degree Survey (KiDS) using an alternative shape measurement pipeline that is more robust to uncertainties in the calibration. To this end, we present an implementation of METACALIBRATION, and compare its performance to that of *lensfit*, which has been used in previous analyses of these data. We find that the multiplicative bias is reduced, especially for the most distant redshifts, as derived from multi-band image simulations designed to match the KiDS data (SURFS-based KiDS-Legacy-Like Simulations: SKILLS). For all tomographic bins we obtain a multiplicative bias $|m| < 0.017$, with negligible additive bias. Importantly, the calibration has a negligible sensitivity to key galaxy properties. The resulting robust shear estimates were used to obtain cosmological parameter constraints. We find that the parameter $S_8 \equiv \sigma_8 \sqrt{\Omega_m/0.3} = 0.789^{+0.020}_{-0.024}$ is consistent with the previous KiDS-1000 *lensfit* constraint of $S_8 = 0.776^{+0.029+0.002}_{-0.027-0.003}$ (statistical + systematic errors). Thanks to the higher effective source density, the constraining power is improved by about 28%. The difference in S_8 with the Planck value remains at a similar level, 1.8σ , implying that it is not caused by the shear measurements.

Key words. cosmology: observations – gravitational lensing

1. Introduction

The growth of the large-scale structure depends on the expansion history of the Universe and the theory of gravity. Comparison to models of structure formation can help shed light on the composition of the Universe or constrain modified gravity theories. Although galaxies can be used to trace the distribution of matter, to fully exploit their constraining power, we need to link their observable properties to the underlying dark matter distribution. Alternatively, it is possible to quantify the statistics of the total matter distribution directly: density inhomogeneities distort space-time, leading to coherent patterns in the observed shapes of distant galaxies, a phenomenon called weak gravitational lensing. The study of the cosmological weak lensing signal – or cosmic shear – is a major driver for the next generation of large imaging surveys, such as *Euclid* (Euclid Collaboration: Mellier et al. 2024), the Legacy Survey of Space and Time (LSST; Ivezić et al. 2019), the *Nancy Grace Roman* Space Telescope (Akeson et al. 2019), and the Chinese Space Station Telescope (CSST; Zhan 2011; Gong et al. 2019).

Much remains to be learned from the analysis of the precursors to these large projects, such as the impact of analysis choices (e.g. Abbott et al. 2023). A major motivation to scrutinise these data is that many lensing surveys prefer a less clumpy

distribution of matter in the low-redshift Universe compared to what is expected based on cosmic microwave (CMB) measurements by *Planck* (Planck Collaboration VI 2020). To leading order, the correlations between the shapes of galaxy pairs are sensitive to the variance in the projected matter distribution. This can be quantified by the parameter $S_8 \equiv \sigma_8 \sqrt{\Omega_m/0.3}$, which depends on σ_8 , the amplitude of linear density fluctuations within a sphere of radius $8 h^{-1}$ Mpc at the present day, and Ω_m , the mean matter density relative to the critical density. Although the significance varies (from 1σ to 3σ ; Asgari et al. 2021; Amon et al. 2022; Secco et al. 2022; Li et al. 2023c; Dalal et al. 2023), the tendency for a lower amplitude of density fluctuations is intriguing. If confirmed, this may suggest the need to modify the current baseline Λ cold dark matter (Λ CDM) model. Various alternatives have already been proposed to alleviate this discrepancy (see Di Valentino et al. 2025, for a summary).

Interestingly, the final analysis of the Kilo-Degree Survey (KiDS; Kuijken et al. 2015; Wright et al. 2024) is in agreement with the CMB measurements (Wright et al. 2025b). It may, however, be too soon to close the issue, because uncertainties in the modelling of the observed cosmic shear signal remain. First of all, intrinsic alignments of physically close galaxies also contribute to shape alignments (see e.g. Joachimi et al. 2015, for a review), while much the constraining power comes from relatively small scales. A major challenge for the modelling on these scales is that the matter distribution is affected by feed-

* E-mail: myoon@strw.leidenuniv.nl

back processes that suppress the predicted lensing signal (e.g. Semboloni et al. 2011). Uncertainties in the modelling have thus been put forward to explain the lower S_8 values (Amon & Efstathiou 2022), although we note that other ways to probe the low-redshift matter distribution find similarly low values (Dvornik et al. 2023; Miyatake et al. 2023).

Regardless of the myriad possibilities, it is worth recalling that accurate shapes and distances for the galaxies are a fundamental input. In this paper, we revisit the shape measurements using the 4th KiDS data release (DR4; Kuijken et al. 2019), which we refer to as KiDS-1000, while the performance is determined using SKILLS (SURFS-based KiDS-Legacy-Like Simulations; Li et al. 2023b), a set of realistic multi-band image simulations that were developed to calibrate the shear estimates for the final KiDS cosmic shear analysis (KiDS-Legacy; Wright et al. 2025b). Our work follows Li et al. (2023a) in that we do not change the calibration procedure of the photometric redshift distributions for the sources, another area of major activity, which has been studied extensively in van den Busch et al. (2022) for KiDS-1000 and developed further in Wright et al. (2025a) for KiDS-Legacy. We use the result of van den Busch et al. (2022) to account for the use of a different algorithm to measure shapes.

To date, all KiDS lensing analyses are based on shape measurements from the model-fitting algorithm *lensfit* (Miller et al. 2013; Fenech Conti et al. 2017). Performance evaluations using simulated data find that the average biases in the shear estimates are generally small and calibratable (Fenech Conti et al. 2017; Kannawadi et al. 2019; Li et al. 2023b). However, the simulations do not reproduce the actual observations perfectly (see Fig. 1), which may lead to residual biases in the calibration of *lensfit*. Although Li et al. (2023a) showed that the systematic uncertainty in S_8 is not limited by the fidelity of the simulations, the tests of the *lensfit* pipeline hinted at increased bias for the most distant redshift bins.

In this paper, we revisit the problem of shape estimation by considering an alternative to *lensfit*. We present a new pipeline for KiDS that uses METACALIBRATION (Huff & Mandelbaum 2017; Sheldon & Huff 2017) to determine the lensing shear. This approach uses the actual data to determine the response to a shear. Studies using simulated images suggest it is a promising way forward (Sheldon & Huff 2017; Hoekstra et al. 2021; Zhang et al. 2023; Sheldon et al. 2023).

The structure of the paper is as follows. We describe the KiDS-1000 data and simulated SKILLS data in Section 2. In Section 3, we describe our implementation of METACALIBRATION, and motivate some of the pipeline choices we made. The performance of our algorithm is examined using SKILLS in Section 4, while we test the robustness of the KiDS-1000 2-point statistics in Section 5. The inputs for the cosmological analysis are discussed in Section 5, and the cosmological/astrophysical constraints from the METACALIBRATION shear catalogue are presented in Section 6.

2. Data

Our ultimate aim is to use our METACALIBRATION pipeline as part of KiDS-Legacy, the final analysis of KiDS based on the 5th data release (DR5; Wright et al. 2024). To optimise the pipeline and to evaluate the performance, we follow Li et al. (2023a) and present results here that instead use DR4, which we summarise in Section 2.1. Moreover, the performance of the shear estimation is quantified using SKILLS (Li et al. 2023b), a set of simulated multi-band images that was developed for the calibration

of *lensfit* (Miller et al. 2013) for KiDS-Legacy. These simulated data are described in Section 2.2.

2.1. KiDS-1000 observation data

The Kilo-Degree Survey (KiDS) is a European Southern Observatory (ESO) public survey that was completed in 2019. It surveyed 1350 deg² in the *ugri* bands using the wide-field camera OmegaCAM mounted on the VLT Survey Telescope (VST). The KiDS observations were complemented by near-infrared (NIR) imaging in the *ZYJHK_s* bands from the VISTA Kilo-degree Infrared Galaxy (VIKING) survey. The resulting 9-band data enables high-fidelity photo-*z* estimation, ideal for cosmic shear tomography.

Here, we use data from KiDS DR4 (Kuijken et al. 2019), which covers a total of 1006 deg² of sky. Similar to previous lensing studies with these data (Asgari et al. 2021; van den Busch et al. 2022; Li et al. 2023a), we use the *r*-band for the shape measurements, because they are the deepest (5 σ limiting AB magnitude of 24.9) and have the best image quality (mean seeing of 0.7 arcsecond). We did not change the model for the point spread function, because the tests presented in Giblin et al. (2021) suggest it is sufficiently accurate.¹

As described in Section A, the selection of sources follows that of previous DR4 studies as much as possible. However, because of differences in the shape measurement algorithms, the final weighted results differ somewhat. As discussed in Section C, we found that some pointings showed significant PSF residuals, and we removed those from the analysis. The resulting effective area covers 771.9 deg² after applying the masks, slightly smaller than the previous KiDS analyses 777.4 deg² (Giblin et al. 2021; Li et al. 2023a).

Robust colours for the sources were measured using the Gaussian Aperture and PSF (GAaP) pipeline (Kuijken et al. 2015). The resulting matched photometry in the nine bands was used to determine photometric redshifts, z_B with BPZ (Benítez 2000). Here, the sample of sources is split based on z_B into 6 redshift bins for a tomographic lensing analysis (see Table 1 for the bin definitions).

As described in Wright et al. (2020); Hildebrandt et al. (2021), a self-organising map (SOM) was used to identify the subsamples of galaxies for which the underlying redshift distributions could be robustly calibrated using deep spectroscopic reference catalogues. This resulting ‘gold-sample’ selected based on ‘gold-flag,’ is used for the inference of cosmological parameters (see Section 5 for more details). Previous cosmological analyses, presented in Heymans et al. (2021); Asgari et al. (2020); Li et al. (2023a), were limited to $z_B \leq 1.2$ because of concerns about the fidelity of the $n(z)$ calibration of the 6th bin ($1.2 \leq z_B \leq 2.0$) due to the limited spectroscopic calibration catalogues. In this paper, we examine the possibility of including a 6th bin, but do not consider it for our cosmological parameter estimation in Section 6 to allow for a more direct comparison to previous work.

¹ We note that *lensfit* internally excludes chip images with fewer than 30 stars to ensure the accuracy of PSF model. This criterion excludes 756 individual chip images (out of a total of 159,136). As not all exposures in a pointing are affected, this mask only applies to about 30 pointings.

2.2. KiDS multi-band image simulation - SKiLLS

Simulated imaging data are essential to quantify the biases of shear measurement algorithms. The results, however, are only meaningful if the simulations capture the observing conditions and galaxy morphologies sufficiently well (Hoekstra et al. 2015, 2017). For the final analysis of KiDS data, Li et al. (2023b) developed SKiLLS, which features the first multi-band image simulations for KiDS that unify the shear and redshift calibration. We refer the interested reader to Li et al. (2023b) for details, in particular the improvements over previous setups (Fenech Conti et al. 2017; Kannawadi et al. 2019). Here, we summarize only its main features.

The input galaxy catalogue of SKiLLS is based on an N-body simulation, the Synthetic Universe For Surveys (SURFS, Elahi et al. 2018). We limit the sample to galaxies with $z < 2.5$ to allow for a robust assignment of galaxy properties, such as realistic spectral energy distributions (SEDs), a key ingredient for multi-band image simulations. The galaxy properties were computed using the open-source semi-analytic model named SHARK (Lagos et al. 2018). As described in Li et al. (2023b), the stellar mass-to-light ratios in the simulation needed to be adjusted to match the magnitude distributions of the actual data in all nine bands. Importantly, the simulations include realistic clustering of galaxies, which is important, because the ‘blending’ of sources along the line of sight is an important contributor to shear bias (e.g. Hoekstra et al. 2021; Li et al. 2023b).

Following Kannawadi et al. (2019), SKiLLS assumes that the surface brightness profiles of the galaxies can be described by a Sérsic profile (Sérsic 1963) with three parameters: the effective radius (galaxy size), Sérsic index (galaxy brightness concentration), and axis ratio (galaxy ellipticity). The values of these structural parameters are based on the catalogue of Griffith et al. (2012), which was derived from images taken by the Advanced Camera for Surveys (ACS) on the *Hubble* Space Telescope (HST). Specifically, we used the results from the COSMOS survey (Scoville et al. 2007) after selecting objects with a good shape fit and a reasonable size. The corresponding redshift information for the sources was extracted from the catalogue compiled by van den Busch et al. (2022), which is based on high-quality photometric and spectroscopic surveys in the COSMOS field.

However, where Kannawadi et al. (2019) aimed to emulate KiDS observations of the COSMOS field, the objective of SKiLLS is to simulate a larger sample of sources covering sufficient area in multiple bands. To this end, Li et al. (2023b) developed a vine-copula-based algorithm (see, e.g. Joe 2014; Czado 2019) that allowed galaxies to be generated from the SURFS-SHARK input, rather than repeatedly sampling from the measured values. The model learned the distributions of structural parameters and their mutual dependence and correlations with redshift and magnitude using the HST measurements as underlying truth. As the version of SURFS-SHARK that was used covers about 108 deg^2 , SKiLLS simulated a total 108 KiDS pointings. To reduce shape noise, we also created versions with the sources rotated by 90 degrees. To capture the impact of varying star density, stellar catalogues were generated using the TRILEGAL model (Girardi et al. 2005) at six different Galactic coordinates evenly spaced across the KiDS-1000 footprint (see Li et al. 2023b, for details).

Another important aspect of SKiLLS is that the multi-band images were generated with realistic seeing conditions drawn from the survey data in each band. Moreover, SKiLLS mimics realistic observational features such as 8×4 CCDs covering a single pointing ($\sim 1 \text{ deg}^2$) with corresponding CCD gaps and dither

patterns. For the r -band images, which were used for the galaxy shape measurements, SKiLLS used PSF models, represented as two-dimensional polynomial functions derived from KiDS-1000 data (Giblin et al. 2021). These were constructed for each exposure and each chip using the centres of the CCD images (for 32 images and 5 exposures). For the other optical bands, which were used for photometry, a Moffat profile for the PSF was adopted. The model parameters were determined from bright stars identified in the ASTRO-WISE² images.

Thanks to the multi-band images, SKiLLS allows us to determine photometric redshifts in a fully realistic fashion, and subsequently to split the source sample into tomographic redshift bins. For this purpose, we used the GAAP (Gaussian Aperture and PSF) pipeline, which measures accurate multi-band colours by accounting for PSF variations across filters and improves the signal-to-noise ratio by down-weighting the noise-dominated outskirts of galaxies (Kuijken et al. 2015, 2019). Photometric redshifts were then estimated with the public Bayesian Photometric Redshift (BPZ; Benítez 2000) code, using the recalibrated template set from Capak (2004) together with the Bayesian redshift prior from Raichoor et al. (2014). This enables a detailed comparison of key observables between the simulated and KiDS-1000 data. Figure 1 shows the distributions for the six tomographic redshift bins as a function of apparent r -band magnitude (MAG_AUTO; top panels), half-light radius (FLUX_RADIUS; middle panels) as measured by SourceExtractor (Bertin & Arnouts 1996), as well as galaxy shapes (using Eq. (1); see Section 3 for details). Despite significant effort to match the simulated output to the data, some deviations in galaxy properties remain. These differences are more pronounced after redshift binning, implying that correlations between morphology and galaxy SEDs need to be improved. Nonetheless, the magnitude distributions match well, except for the sixth bin, which indicates that SKiLLS lacks faint high-redshift sources. Moreover, the simulated data tend to lack small sources, while the distributions in galaxy shapes also do not quite match. The impact of these differences can be largely mitigated using an empirical reweighting scheme based on measured galaxy properties. Moreover, sensitivity analyses by Li et al. (2023a) demonstrated that uncertainties in the input do not significantly change the calibration, and thus have a negligible impact on the cosmological inference, although this may change as more survey data are included. As discussed below, reducing the sensitivity of the shear estimation algorithm to limitations in the simulated data is our main motivation to explore the prospects of METACALIBRATION in this paper.

3. Shear measurement pipeline

The observed galaxy images are modified by instrumental effects, such as detector imperfections or telescope optics, or by atmospheric turbulence in the case of ground-based telescopes (see Mandelbaum 2018, for a review). The challenge is to recover the true underlying correlations of galaxy ellipticities. One approach is to fit galaxy models to the data and find the best match. This is the approach used by *lensfit* (Miller et al. 2013). A benefit of fitting methods is that masked data or additional instrumental effects can be naturally accounted for. However, mismatches between the adopted model and the true morphologies can lead to model bias (e.g. Kacprzak et al. 2014).

Alternatively, one can try to infer galaxy ellipticities from the moments of the observed surface brightness distribution.

² ASTRO-WISE is a KiDS pipeline built for photometry measurement. (<http://www.astro-wise.org/>)

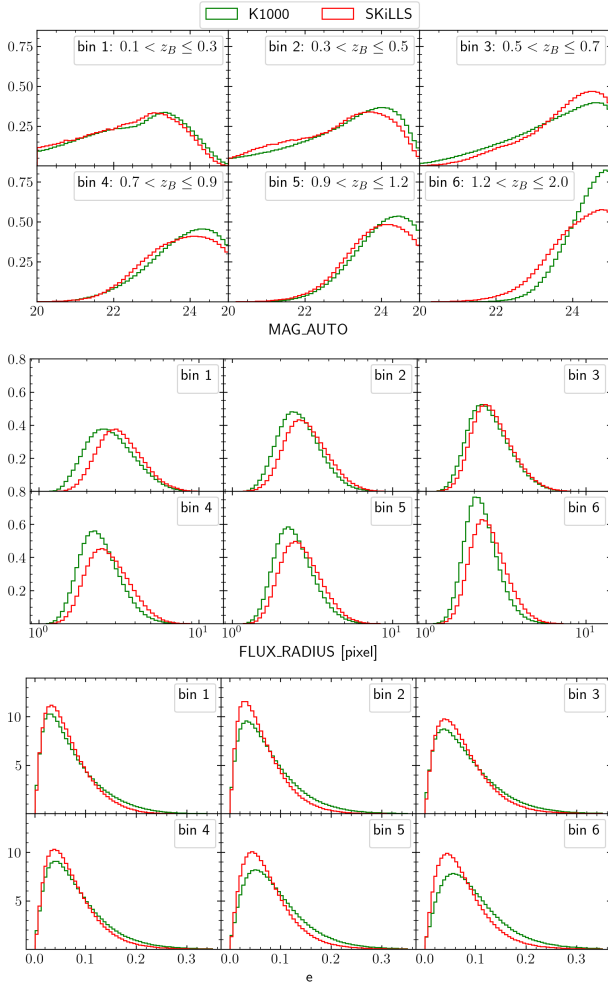


Fig. 1. Comparison of source features between KiDS-1000 and SKiLLS selected for the metacal catalogue for each tomo-bin (any weighting is not applied): r-band magnitude (top), size (middle), and ellipticity (bottom). The sixth bin shows the most deviation for r-band magnitude and ellipticity. Regarding size, the deviation gradually increases for high redshift bins.

Such moment-based approaches (e.g. Kaiser et al. 1995; Melchior et al. 2011) are computationally efficient. For instance, the complex polarisation (or third eccentricity), $e = e_1 + i e_2$, can be defined as

$$e = \frac{Q_{11} - Q_{22} + 2i Q_{12}}{Q_{11} + Q_{22}}, \quad (1)$$

where the (weighted) quadrupole moments, Q_{ij} , are defined as

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

where $I(x)$ is the observed surface brightness distribution of a source and $W(x)$ is a weight function.

Arguably, Eq. (1) is the simplest way to quantify the shape of an object, but the observed values are biased because of the convolution with the PSF. In the case of unweighted moments, this can be easily corrected for (Valdes et al. 1983). In practice, however, a compact weight function is needed to suppress the contribution of noise in the observed image, biasing the estimate for e even more. The bias in e depends on the higher-order moments of the surface brightness distribution. These can be used to

obtain a corrected estimate for the shear (e.g. Kaiser et al. 1995; Luppino & Kaiser 1997; Hoekstra et al. 1998), although biases remain because of ellipticity gradients and blending (Hoekstra et al. 2015). Moreover, as shown in Fenech Conti et al. (2017), and studied in more detail in Hoekstra et al. (2021), shear biases are already introduced when objects are detected. Hence, even a perfect shape measurement method results in a biased shear estimate for a sample of sources.

To evaluate the performance of shape measurement algorithms, simulated data can be used (e.g. Heymans et al. 2006a; Massey et al. 2007; Bridle et al. 2010). If the image simulations match the data perfectly,³ one could simply use the average observed e_i^{obs} for an ensemble of sources and calibrate any bias using the simulated data, for which the shear is known, using

$$\langle e_i^{\text{obs}} \rangle = (1 + m_i) \gamma_i^{\text{true}} + c_i, \quad (3)$$

where c_i quantifies the so-called additive bias, which corresponds to a preferred orientation, arising from an anisotropic PSF or detector effects. The multiplicative bias, m_i , quantifies the bias in the amplitude of the signal. Although both biases need to be accounted for, in the following we focus on the multiplicative bias, because it cannot be determined empirically; its value is determined using image simulations. Eq. (3) assumes a linear relation between the average observed shape $\langle e_i^{\text{obs}} \rangle$, and the shear, γ_i^{true} , which is a reasonable assumption for small shears. However, it is computationally efficient to use larger shears for the shear calibration simulations, while the shear can also be large around massive structures (e.g. Li et al. 2024). We therefore study the possible impact of non-linearity of the algorithm in Section D.

As demonstrated by Fig. 1, it is difficult to obtain a perfect match between simulated and observed data. However, the shear bias inferred from simulated data is only meaningful if they resemble the observations sufficiently well (Hoekstra et al. 2015). Kannawadi et al. (2019) and Li et al. (2023b) explored the implications of this in more detail, highlighting the trade-off between the realism of the simulated data and the sensitivity of the algorithm to deviations. As discussed in Kannawadi et al. (2019), the residual multiplicative shear bias (Δm) for a sample of galaxies depends on observable quantities, $\mathbf{D}$, such as signal-to-noise ratio (SNR) or size, as well as hidden features, $\mathbf{h}$, such as undetected blends. A mismatch of properties (q) between the actual and simulated galaxy populations, $\Delta q(\mathbf{D}, \mathbf{h}) := q_{\text{true}}(\mathbf{D}, \mathbf{h}) - q_{\text{sim}}(\mathbf{D}, \mathbf{h})$, implies a difference in multiplicative bias:

$$\Delta m := m_{\text{true}} - m_{\text{sim}} = b(\mathbf{D}, \mathbf{h}) \Delta q(\mathbf{D}, \mathbf{h}), \quad (4)$$

where $b(\mathbf{D}, \mathbf{h})$ quantifies the sensitivity of the multiplicative bias to a small change in the sample:

$$b(\mathbf{D}, \mathbf{h}) := \frac{\delta m[q]}{\delta q(\mathbf{D}, \mathbf{h})}. \quad (5)$$

Equation (4) shows that if the samples match perfectly, the inferred shear bias is correct, irrespective of the shape measurement method. Alternatively, a method that is insensitive to the parameters that influence the shear bias also yields a meaningful result. The sensitivity of a method can be reduced by mapping the bias as a function of $\mathbf{D}$, such as SNR and size (Fenech Conti et al. 2017; Kannawadi et al. 2019), but the sensitivity function cannot be determined perfectly as long as the hidden features, $\mathbf{h}$,

³ We assume implicitly that the PSF has been determined perfectly.

their correlation with the explicit features, $\mathbf{D}$, and their impact on the bias are uncertain.

An alternative, which aims to reduce the reliance on image simulations, is provided by METACALIBRATION (Huff & Mandelbaum 2017; Sheldon & Huff 2017). As discussed in more detail below, it uses the observed images themselves, so that $\Delta q \approx 0$ by construction. Given its promise, METACALIBRATION has been adopted as the main shear calibration process for the Dark Energy Survey Y1 (Zuntz et al. 2018) & Y3 analyses (Gatti et al. 2021), and the LSST (Ivezić et al. 2019). The main advantage of this approach is that the calibration is less sensitive to the choice of galaxy population properties that are set to be inputs of the image simulation for shear calibration. This has motivated us to consider METACALIBRATION as an alternative for KiDS as well.

METACALIBRATION infers the response to shear from the data themselves by manipulation of postage stamps of the observed galaxies. We only need to assume that it is possible to construct a sheared version of the true image (Huff & Mandelbaum 2017), which is the case for well-sampled data (but also see Kannawadi et al. 2021, for the case of undersampled data). We have verified that the KiDS PSF models that we use are sampled sufficiently well, so that the performance of METACALIBRATION is not degraded.

We will detail our specific implementation in Section 3.1, but review the key steps in the following (also see Fig. 2 for all the steps in the calibration pipeline). The postage stamp image is (i) deconvolved using a model for the PSF; (ii) sheared; and then (iii) reconvolved with a slightly dilated PSF to avoid the noise blowing up (see Sheldon & Huff 2017, for details). The shape estimate of the resulting sheared image is given by

$$\mathbf{e} \approx \mathbf{e}|_{\gamma=0} + \left. \frac{\partial \mathbf{e}}{\partial \gamma} \right|_{\gamma=0} \gamma + O(\gamma^2) \equiv \mathbf{e}|_{\gamma=0} + \mathbf{R}^{\gamma} \gamma + O(\gamma^2). \quad (6)$$

where $\mathbf{R}^{\gamma}$ is the 2×2 shear response tensor. We can estimate its components by measuring the shapes of the galaxies in the sheared images and computing

$$\mathbf{R}_{ij}^{\gamma} \equiv \frac{\partial e_i}{\partial \gamma_j} \approx \frac{e_i^{j+} - e_i^{j-}}{\Delta \gamma_j}, \quad (7)$$

where the subscripts indicate the two shear components, and the superscript the sign of the applied shear, so that ‘ $j+$ ’ means that the image was sheared by $+\gamma_j$, etc; hence, $\Delta \gamma_j = 2\gamma_j$. In practice, the off-diagonal elements vanish on average, while the diagonal elements tend to be equal. We therefore can capture the response by using $R^{\gamma} \equiv \text{Tr}(\mathbf{R}^{\gamma})/2$.

METACALIBRATION works well for isolated galaxies, but in the case of blended sources some residual bias remains (Hoekstra et al. 2021). Moreover, the selection of objects already introduces shear bias (Hirata & Seljak 2003; Fenech Conti et al. 2017), and METACALIBRATION cannot undo this. The choice of weight function for the shape measurement can further exacerbate any selection bias (e.g. Sheldon & Huff 2017). We minimise such additional selection bias by adopting an axisymmetric Gaussian with a fixed width for $W(\mathbf{x})$.

In principle, it is also possible to include the detection process, which is called METADETECTION (Sheldon et al. 2020; Zhang et al. 2023; Sheldon et al. 2023). However, implementing this would require a major overhaul of the whole shear calibration pipeline, because it needs to be applied to all bands in order to avoid selection biases during the photometric redshift estimation. Moreover, METADETECTION cannot capture the residual bias caused by blended sources at different redshift (MacCrann et al.

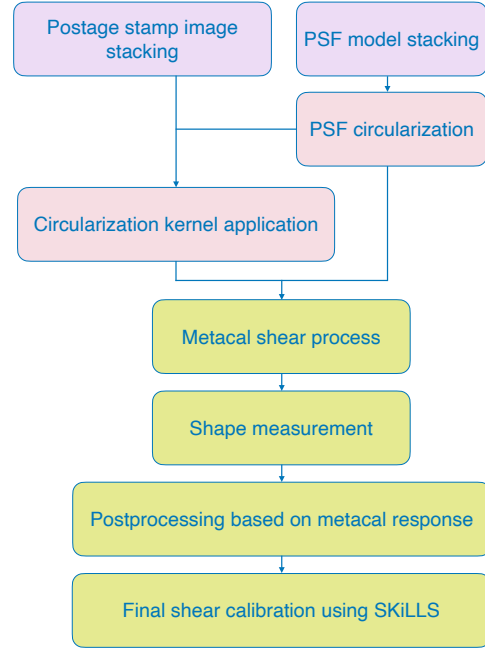


Fig. 2. Schematic work flow of the METACALIBRATION pipeline, which starts with the same input catalogue that was used in previous KiDS-1000 analyses.

2022; Li et al. 2023a). Given the realism of SKILLS (Li et al. 2023b,a) and the required level of accuracy, the implementation of METADETECTION is not warranted for our KiDS analysis.

PSF anisotropy is the main contributor to additive shear bias, and thus needs to be carefully accounted for. Different approaches can be used to do so. For instance, if the shape measurement itself does not remove the PSF anisotropy, the response can be determined with additional image manipulations in which the PSF is sheared (Sheldon & Huff 2017). Alternatively, Hoekstra (2021) proposed to correct the observed polarisations (Eq. (1)) using the KSB formalism (Kaiser et al. 1995; Hoekstra et al. 1998). However, to fully remove the PSF anisotropy, an empirically determined boost factor is required. As detailed in Section 3.1, we adopt another approach here: we used the PSF model to derive a convolution kernel that renders the PSF isotropic. In principle, it removes additive bias arising from the PSF, and we only need METACALIBRATION to determine the shear response R^{γ} .

In Fig. 2 we show the schematic workflow of our METACALIBRATION pipeline. It starts with the same input catalogue as used in previous KiDS-1000 analyses that used *lensfit*, and we apply the same selections where possible (see Section A for details). Moreover, we use the same PSF model, which was tested extensively in Giblin et al. (2021). The same work flow is used to analyse the SKILLS data and to determine the residual multiplicative bias after METACALIBRATION. Below, we describe our implementation in more detail.

3.1. Shape measurement setup

We exploit the fact that METACALIBRATION can be used in combination with any shape measurement. Naively, one might prefer an estimator with an intrinsically small bias. However, more important is that the results are robust against limitation of the image simulations that are used to quantify the biases. Here, we adopt Eq. (1) using weighted moments. Although simple, several choices remain to be made, such as the choice of weight func-

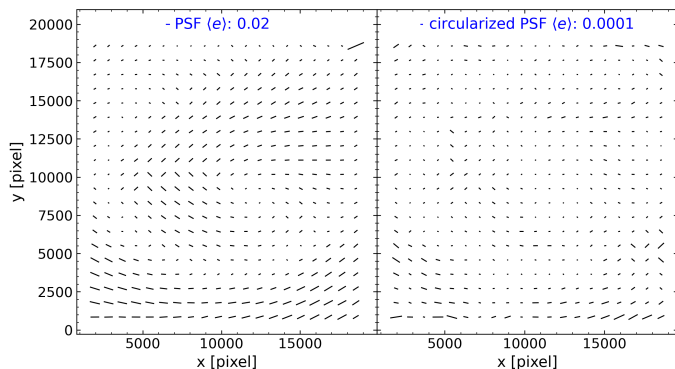


Fig. 3. Performance of the circularizing procedure applied to a representative KiDS-1000 pointing (at RA=9.4°, Dec =-32.1°) with the ellipticities shown as the length of bars (see the reference on top: blue bar) and the ratio of e_1/e_2 as their angle. The model PSFs, which were evaluated at the positions of the detected galaxies (whose average is shown on the grid), show some variation across the field of view, although the overall level of PSF anisotropy is small ($|e| \lesssim 0.02$; left panel). After circularising, the PSF shapes are consistent with zero ($|e| \lesssim 10^{-4}$; right panel). Please note that the ellipticity scale of the right panel is 100 times larger than the left panel.

tion, but also how to deal with the fact that the PSF is anisotropic and varies spatially, or how to combine the information from the (typically) five r -band exposures in KiDS observations.

As a first step, we create postage stamps with a fixed size of 48×48 pixels for each galaxy and exposure. We only retain complete stamps, thus removing any instances where a galaxy is located close to the chip boundaries in an exposure. To minimise the flux from neighbouring galaxies contaminating the shape measurement, we use the segmentation maps generated by SOURCEEXTRACTOR to mask those objects (replacing the pixel values of adjacent objects with zero). We found that this is sufficient to address the impact of neighbouring objects, as further filtering of recognized blends has a negligible impact on the shear bias (see Section A.2 for more details).

In principle, it is also possible to measure shapes on the individual images, but we decided to use stacked postage stamp images. A major advantage is that the stacking limits the METACALIBRATION step to only five image manipulations (as opposed to 25 in the case of processing all individual exposures). Moreover, the stacking procedure is fast, while the reduced noise in the images improves the shape measurement. We note that the galaxy images are well sampled, and hence the resampling does not introduce any bias or loss of information.

For each postage stamp, we generate the corresponding PSF model image using the location of the postage stamp. Here, we use the same PSF model that was used in previous KiDS-1000 analyses based on *lensfit*. We then stack the galaxy postage stamps, as well as the corresponding PSF images. In Section C, we test the performance of the PSF stacking procedure on the KiDS data.

In our pipeline, we account for PSF anisotropy by convolving the galaxy images with a kernel that circularises the (stacked) PSF. Following Kuijken (2006), we use two Gaussians fitted to PSF images, we estimate the PSF size, and find a kernel that yields a round PSF after convolution. We found that this performed better than a kernel based on a single Gaussian. To demonstrate the performance of the circularization step, we show the results for a representative pointing in Fig. 3. The original PSF anisotropy varies spatially, although the amplitude is

low. After the PSF circularisation kernel is applied, the resulting shapes are consistent with zero.

We apply the PSF circularisation kernels to the corresponding galaxy images, and use the result for the main METACALIBRATION step. For this, we use the functions that are available as part of the ngmix package (Sheldon & Huff 2017; Huff & Mandelbaum 2017). To determine the shear response, R_{ij}^{γ} , we adopted a shear of 0.01, so that $\Delta\gamma_i = 0.02$ (see Eq. 7) and use the MetacalGaussPSF option for the slightly dilated reconvolution kernel. Both the circularisation and the shearing introduce anisotropic correlations in the noise. In principle, this can be accounted for, but we chose not to (i.e., we use `fix_noise=False`), because it leads to noisier measurements. Importantly, the resulting bias is captured by the image simulations, and thus is naturally accounted for during the calibration using SKiLLS.

For the actual shape measurement of the METACALIBRATION outputs, we use the hsm module (Hirata & Seljak 2003; Mandelbaum et al. 2005) in GALSIM (Rowe et al. 2015). The choice of the weight function used to measure the galaxy shapes also matters. It can be matched to the source in size, shape or both, to maximise the SNR of the estimate. However, such optimisation introduces additional selection bias, because the weighting depends on the source ellipticity. As shown by Sheldon & Huff (2017), the total shear response consists of two terms: $\mathbf{R} = \mathbf{R}^{\gamma} + \mathbf{R}^S$, where the second term captures how the selection, or weighting scheme, depends on the shear. If the weight function is fixed, this sensitivity is minimized.

Blended sources bias $\mathbf{R}^S$ (Hoekstra et al. 2021), so that image simulations are needed to capture the total selection bias.⁴ Therefore, we do not compute $\mathbf{R}^S$, because it greatly simplifies the pipeline, while the residual bias can be captured well using image simulations (Hoekstra et al. 2021). These considerations also motivate our choice of a fixed circular Gaussian with a dispersion of 2.5 pixels. Although this is generally not the optimal choice, the distributions of FLUX_RADIUS in Fig. 1 show that it matches well the typical size of the galaxies in all tomographic redshift bins. This is sufficient, because we are interested in robust shear estimates, rather than precise shapes for individual galaxies. For larger, brighter galaxies the uncertainty in the shear estimate is determined by the intrinsic galaxy shape, irrespective of the weight function. As the tests in Section B show, the width of the weight function does not affect the SNR of the cosmic shear signal, but it can help to reduce the sensitivity to blended sources, in addition to the masking we already apply.

Figure 4 shows the shear response, R^{γ} , as a function of apparent magnitude for the six tomographic bins for SKiLLS (left) and KiDS-1000 (right). The shear response is a strong function of magnitude, but the results are quite similar between tomographic bins, especially for SKiLLS. Although the overall behaviour is similar between simulated and real data, we observe generally higher values for KiDS-1000. We attribute this to the fact that the morphologies in SKiLLS do not match the data perfectly, as alluded to in Fig. 1. In particular, for bright, resolved galaxies the use of a Sérsic profile may not be sufficient. We explore the implications of this in Section 4.4.

⁴ The detection bias can be captured using METADETECTION (Sheldon et al. 2020), although the impact of redshift-dependent blending still leaves a small residual bias that needs to be accounted for.

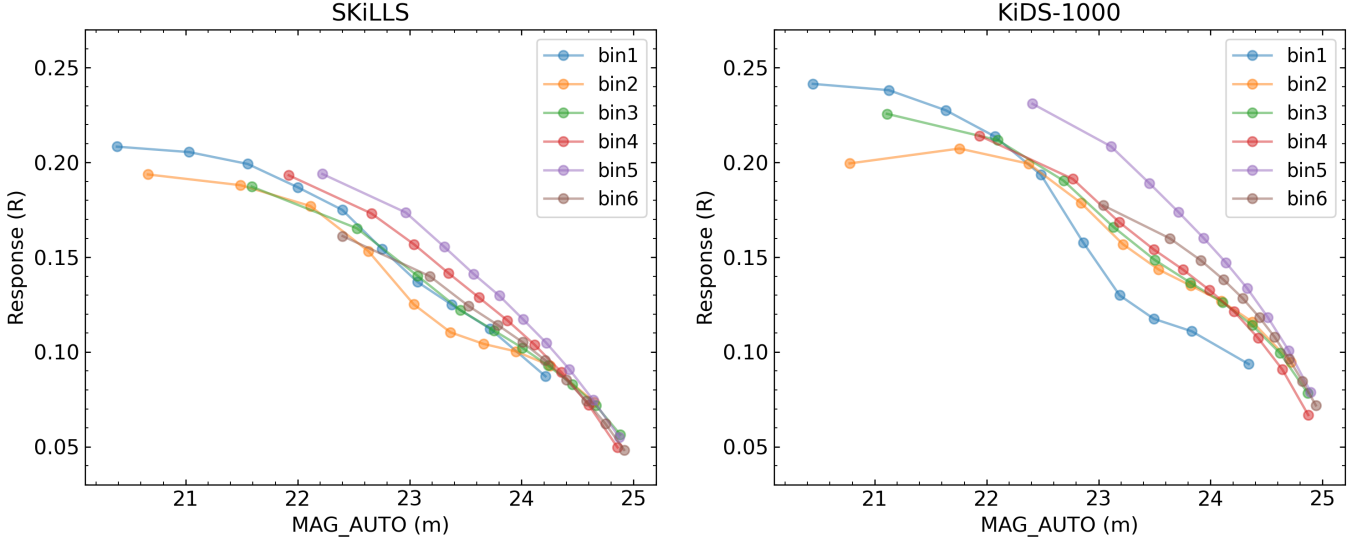


Fig. 4. The mean shear response, R^γ , as a function of apparent magnitude for the six tomographic bins of simulated SKILLS galaxies (left) and real KiDS-1000 galaxies (right). We used ten equi-populated subsamples for each tomographic bin to compute the averages.

3.2. Weighting scheme for the ensemble of galaxy shapes

In the previous section, we defined a simple estimate for the galaxy shapes, using Eq. (1) and a fixed weight function in Eq. (2). In the absence of noise and detection bias, an unbiased estimate for the shear γ_i is given by $\hat{e}_i/R_{ii}^\gamma$, where $\hat{e}_i$ is measured from the METACALIBRATION ‘no_shear’ image (this image is not sheared, but is convolved by the slightly dilated PSF used in the METACALIBRATION step). Unfortunately, the estimate for the shear response, R_{ii}^γ , is noisy for individual galaxies, and the ensemble average of the individual values for γ_i will therefore be biased.

Instead, following Sheldon & Huff (2017), we can consider the ensemble average of Eq. (6). Ideally, this should give us an unbiased estimate of the shear, but in practice blending and detection biases also contribute to multiplicative bias (e.g. Hoekstra et al. 2021). To explicitly account for this, we consider the following unbiased shear estimator:

$$\hat{\gamma}_i = \frac{\hat{e}_i}{\langle (1 + m_i) R_{ii}^\gamma \rangle} \approx \frac{(1 + m_i) R_{ii}^\gamma}{\langle (1 + m_i) R_{ii}^\gamma \rangle} \gamma_i + \frac{e_i^{\text{noise}}}{\langle (1 + m_i) R_{ii}^\gamma \rangle}, \quad (8)$$

where e_i^{noise} quantifies the uncertainty in the shear estimate. The intrinsic shape of the galaxy is the dominant contributor, but for fainter galaxies, the noise in the images can also become important.

The advantage of using Eq. (8) is that it yields an unbiased estimate, $\hat{\gamma}_i$, while avoiding taking ratios of noisy quantities for individual galaxies. However, the shears are now weighted by the product of $(1 + m_i)$ and R_{ii}^γ . This needs to be accounted for in the modelling of the lensing signal, as it modifies the effective source redshift distribution. In general, for narrow tomographic bins, we expect the impact to be negligible, because the shear bias tends to be small, while Table 1 shows that $\langle R^\gamma \rangle$ does not differ much between tomographic bins (also see Section 4.1).

As shown by Fig. 4, R^γ conveniently down-weights fainter galaxies, for which the shape measurements are noisier, but the precision of the shear signal can be improved further by using the inverse variance of the shear as an additional weight, w_j , for the j -th galaxy. Using Eq. (8) we find that the variance of the

shear estimate is given by

$$\sigma_{\hat{\gamma}_i}^2 \approx \frac{\langle \hat{e}_i^2 \rangle}{\langle (1 + m_i) R_{ii}^\gamma \rangle^2}, \quad (9)$$

where we assumed that the observed variance of $\hat{e}_i$ can be used as a proxy for e_i^{noise} . This ignores sub-percent broadening of the observed distribution due to cosmic shear.

Estimating the weight for individual galaxies is noisy, and tends to make the weight depend on ellipticity. Instead, we consider the variance for subsamples of galaxies. Specifically, we expect the measurement uncertainty to be mostly a function of the flux (or magnitude) and we therefore determine for each tomographic bin, t_z , the weight for ensembles binned by m_{AUTO} , the apparent magnitude determined by SOURCEEXTRACTOR. We thus compute the weight as

$$w_j(m_{\text{AUTO}}, t_z) = \sigma_{\hat{\gamma}}^{-2}(m_{\text{AUTO}}, t_z) = [\sigma_{\hat{e}}^{-2} \langle R^\gamma \rangle^2](m_{\text{AUTO}}, t_z), \quad (10)$$

where

$$\sigma_{\hat{e}}^2(m_{\text{AUTO}}, t_z) = \frac{1}{2} \left[\frac{\sum_j (\hat{e}_{1,j})^2 + (\hat{e}_{2,j})^2}{N_{\text{gal}}} \right]. \quad (11)$$

We ignored the multiplicative bias in the estimate of the weight, because it is small and does not vary much with magnitude (see Section 4.1). This weight is assigned to all the N_{gal} galaxies in each magnitude and tomographic bin. Our approach is similar to Gatti et al. (2021), who bin in S/N, but also resolution. We decided to ignore the resolution, because it tends to be noisier and does not improve the weighting much.

Our final estimator for the shear of an ensemble of sources is then

$$\langle \hat{\gamma}_i \rangle = \frac{\langle w \hat{e}_i \rangle}{(1 + \langle m_i \rangle) \langle w R_{ii}^\gamma \rangle}, \quad (12)$$

where we used that the value of m is determined for the ensemble by applying a linear regression of the input shear to $\langle w e_i \rangle / \langle w R_{ii}^\gamma \rangle$. In practice, we estimated $\langle m \rangle$ by taking the arithmetic mean

of $\langle m_1 \rangle$ and $\langle m_2 \rangle$. This is expected based on symmetry arguments, but is also supported by image simulations (Hoekstra et al. 2021).

Equation (8), or alternatively Eq. (12), can be used to compare the various approaches for shear estimation. In the end, each algorithm is defined by a shape estimate, a correction that is derived from the data, and a correction derived from image simulations. These appear in Eq. (8) as $\hat{e}_i$, R_{ii}^γ and m_i , respectively.⁵ The strength of METACALIBRATION is that R_{ii}^γ is determined directly from the observed images for any estimator, which results in small values for m (Sheldon & Huff 2017; Hoekstra et al. 2021). However, detection bias and blending ultimately limit the accuracy, and the residual bias needs to be determined using image simulations. However, these are relatively small corrections, and the sensitivity to the simulation input is reduced (Hoekstra et al. 2021).

A drawback of our simple shape estimator, Eq. (1), is the dependence of the shear response on size and magnitude, as this leads to an additional weighting of the shear. A better choice would be to replace Eq. (1) with an estimate of the shear, so that $R_i^\gamma \approx 1$ (or constant) for all galaxies. However, in doing so, one would need to ensure that the multiplicative bias remains small. Exploring this is beyond the scope of this paper.

Finally, it is worth reflecting on machine-learning approaches. These aim to learn the mapping between the shape estimate and the shear (Tewes et al. 2019). In the most extreme case, no additional information about the object is used, so that $R_i^\gamma = 1$, and the mapping is completely determined from the image simulations. Given the challenge to match the simulations and the data (Li et al. 2023b), this is undesirable, because it maximises the sensitivity to the limitations of the simulated data. In contrast, METACALIBRATION minimises the sensitivity to the image simulations. Only METADETECTION (Sheldon et al. 2020) can reduce this further, at the expense of a more complex data flow.

4. Metacalibration shear bias from SKiLLS

Here, we apply our METACALIBRATION pipeline to the SKiLLS data to determine the (residual) multiplicative and additive shear biases. Following, Li et al. (2023b), we do so in two steps. First, we use simulations with a constant shear applied to all the sources. Motivated by the findings of Hoekstra (2021), we calibrate the multiplicative and additive biases separately. The results for the multiplicative bias are discussed in Section 4.1. In Section 4.2 we quantify the additive bias, with a particular focus on the bias introduced by PSF anisotropy. The constant-shear simulations account for most of the biases, but as the shear applied does not depend on redshift, it cannot correctly capture the blending of sources at different redshifts (MacCrann et al. 2022; Li et al. 2023b). As described in Section 4.3, this second order correction to the multiplicative bias is determined using a second set of SKiLLS with variable, redshift-dependent input shears.

4.1. Multiplicative bias

For a given number of simulated sources, the uncertainty in the multiplicative bias, m , is determined by the shear applied; the larger the shear, the better m is determined. However, if the shear is too large, non-linearities in the shear measurement algorithm can bias the results (Kitching & Deshpande 2022; Jansen et al. 2024; Li et al. 2024).

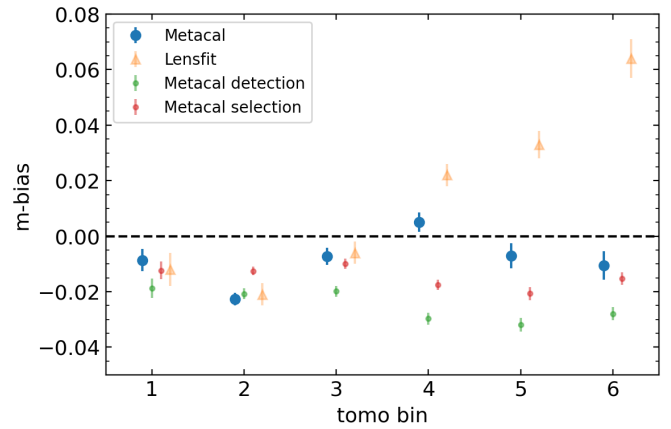


Fig. 5. Multiplicative bias as a function of tomographic bin estimated from fixed-shear simulations. We find that the biases for METACALIBRATION (blue points) are small ($|m| < 0.023$) and agree well with the level of bias expected for a perfect shape measurement method after accounting for selection bias (red points). The green points indicate the corresponding detection bias from SOURCEEXTRACTOR. For the first three bins, the multiplicative bias for *lensfit* (orange triangles) is similar, but it increases for more distant galaxies.

Following previous KiDS studies (Kannawadi et al. 2019; Li et al. 2023b), we applied a constant shear $\gamma_i = \pm 0.028$ (for a total shear $|\gamma| = 0.04$), resulting in four shear realisations. These values are higher than the average shear we expect in the data, but the results from Section D show that it yields accurate estimates for m , while reducing the simulation volume by more than an order of magnitude. To improve the precision further, we created pairs of images, where the sources were rotated by 90° , which dramatically reduces the shape noise contribution (Massey et al. 2007; Fenech Conti et al. 2017). We use the weighting scheme outlined in Section 3.2 to compute the mean shear for these pairs of images. Linear regression using Eq. (3) then yields estimates for m and c_i .

Table 1 lists the multiplicative bias values for the fixed shear, m_{fix} , for the six tomographic bins. We find that the average bias in each tomographic bin is small, with the largest value (bin 2) only $m = -0.0227 \pm 0.0023$, mainly due to contamination by stars (discussed in Section A.3). Figure 5 compares the METACALIBRATION results (blue points) to the biases determined for *lensfit* (orange triangles). For the first three bins, the multiplicative biases are comparable, but for the highest three redshift bins the values are considerably smaller for METACALIBRATION, highlighting the robustness of the approach.

The bias values reported here are the combination of detection bias and the bias subsequently introduced by the shape measurement algorithm. Following Fenech Conti et al. (2017) we estimate the detection bias using the input (sheared) ellipticities, assigning equal weights to all galaxies. The resulting detection bias is indicated by the green points in Fig. 5. As the actual shear estimates use a weighting scheme, a more appropriate comparison includes the weights as well (the red points in Fig. 5). The fair agreement between the red and the blue points demonstrates that METACALIBRATION behaves like a nearly perfect shape measurement algorithm. Some deviation is expected because of blending (Hoekstra et al. 2021), but the impact is small.

The values for m_{fix} give a fair estimate of the shear bias in KiDS-1000, but they require some further revision for the cosmological analysis. First of all, following Li et al. (2023b), we need to account for the blending of sources at different redshifts,

⁵ This interpretation changes the meaning of R^γ somewhat.

Table 1. METACALIBRATION summary statistics for the six tomographic bins.

Bin	Photo- z range	$\langle z \sigma_\gamma^{-2} \rangle$	$\langle z(1+m)R^\gamma \sigma_\gamma^{-2} \rangle$	$\langle R^\gamma \rangle$	n_{eff} [arcmin $^{-2}$]	$\sigma_{\gamma,i}$	m_{fix}	m_{final}	σ_m
1	$0.1 < z_B \leq 0.3$	0.338	0.292	0.21	1.01	0.330	-0.0086	-0.0082	0.0044
2	$0.3 < z_B \leq 0.5$	0.414	0.410	0.19	1.86	0.331	-0.0227	-0.0219	0.0027
3	$0.5 < z_B \leq 0.7$	0.580	0.577	0.20	2.77	0.351	-0.0072	-0.0037	0.0034
4	$0.7 < z_B \leq 0.9$	0.792	0.788	0.19	1.80	0.342	0.0051	0.0034	0.0038
5	$0.9 < z_B \leq 1.2$	1.046	1.039	0.20	1.84	0.364	-0.0071	-0.0024	0.0051
6	$1.2 < z_B \leq 2.0$	1.447	1.411	0.16	1.43	0.451	-0.0106	-0.0121	0.0057

Notes. Galaxies are assigned to tomographic bins based on their photometric redshift z_B . For each bin, we report the mean true redshift of the sources in SKILLS, weighted by the inverse shear variance, σ_γ^{-2} , as well as the results when the multiplicative bias and shear response are included in the weight. Other columns list the mean shear response, $\langle R^\gamma \rangle$, the effective source density n_{eff} and the uncertainty in the shear estimates of individual galaxies (each component), $\sigma_{\gamma,i}$. Finally, we report the multiplicative bias using a fixed shear, m_{fix} , which needs to be adjusted to account for the variable shear, as detailed in Section 4.3, to obtain the final shear bias, m_{final} and corresponding uncertainty, σ_m , both of which are used in the cosmological analysis.

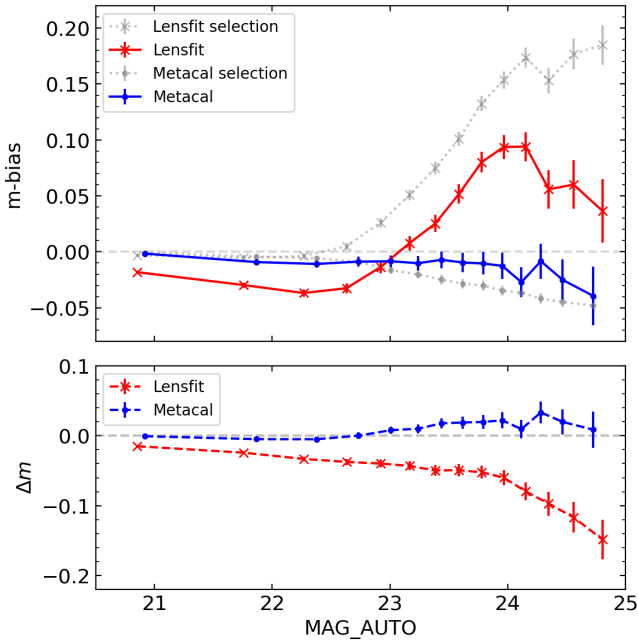


Fig. 6. *Top panel:* Multiplicative shear bias as function of apparent magnitude MAG_AUTO for METACALIBRATION (blue) and *lensfit* (red). For reference, the corresponding selection bias is also indicated (grey). *Bottom panel:* The difference, $\Delta m = m_{\text{fix}} - m_{\text{sel}}$, highlighting the excellent performance of METACALIBRATION.

which we investigate in Section 4.3. Moreover, additional selections are made. For instance, the scheme to correct for additive bias eliminates some sources, as does the redshift calibration. We discuss these below, and report m_{final} , the final value of the multiplicative bias, in Table 2.

A small average shear bias is clearly desirable, as it appears in the denominator for our estimator, Eq. (12). However, this does not necessarily mean the shear bias is small for any subset of sources. For instance, Fenech Conti et al. (2017) examined the dependence of m with galaxy size and SNR, which showed that m is still significant for faint, small galaxies. To explore this further, we show in the top panel of Fig. 6 the average multiplicative bias as a function of MAG_AUTO for METACALIBRATION (blue) and *lensfit* (red). For reference, we also indicate the corresponding selection bias as a function of apparent magnitude (grey). The difference, $\Delta m = m_{\text{fix}} - m_{\text{sel}}$, is shown in the bot-

tom panel. It shows that METACALIBRATION returns near-perfect shear estimates in the absence selection bias, whereas *lensfit* is increasingly biased for faint galaxies. As a consequence, the calibration of METACALIBRATION is more robust against mismatches in the magnitude distribution between data and simulations (see Fig. 1). We return to this in Section 4.4.

Generally, we expect the shear bias to also depend on galaxy size. In the case of METACALIBRATION, the observed size is strongly correlated with R^γ , which complicates such a comparison. However, to test the dependence on galaxy size, we can use the input properties instead. In Fig. 7 we show the difference $\Delta m = m_{\text{fix}} - m_{\text{sel}}$ as a function of the input r -band magnitude when we split the sample into quintiles by input size. The left panel of Fig. 7 shows the results for METACALIBRATION. The bias is generally small for all subsets, but only for the smallest, faintest galaxies do we observe a significant bias. For the other size bins, the value of Δm are consistent with one another, and thus do not depend on input size. The bump between $23 < m_r^{\text{input}} < 24$ is caused by blended sources (Hoekstra et al. 2021). The right panel in Fig. 7 shows the same, but for *lensfit*. We observe more spread between size bins, which indicates a slight dependence on input size. The biases for both methods are generally small, except for the smallest galaxies, where METACALIBRATION performs somewhat better (compare the grey and red results in the left panel of Fig. 7). In practice, however, these galaxies carry little weight in the analysis.

As discussed in Section 3.2, we weight the sources based on the measurement uncertainty using an inverse variance weight that depends on apparent magnitude for a given tomographic bin. The dashed lines in Fig. 8 show this for all bins combined for METACALIBRATION (blue) and *lensfit* (red). The behaviour for both methods is similar. We use these results to compare the statistical constraining power of our METACALIBRATION catalogue. To this end, we determined the uncertainty in the shear estimate from a single galaxy, $\sigma_{\gamma,i}$ and the effective number density, n_{eff} using the definition in Heymans et al. (2012). The results for each tomographic bin are reported in Table 1.

Fig. 9 shows the cumulative effective number density as a function of apparent magnitude, for METACALIBRATION and *lensfit* for both SKILLS (left) and the KiDS-1000 data (right). For both SKILLS and KiDS, METACALIBRATION yields $n_{\text{eff}} \approx 10.4$ arcmin $^{-2}$, while *lensfit* yields $n_{\text{eff}} \approx 9.3$ and 8.5 arcmin $^{-2}$, respectively. The 22% higher n_{eff} for METACALIBRATION when applied to KiDS-1000 is attributed to the higher fraction of sources for which shapes could be measured, which in turn is a conse-

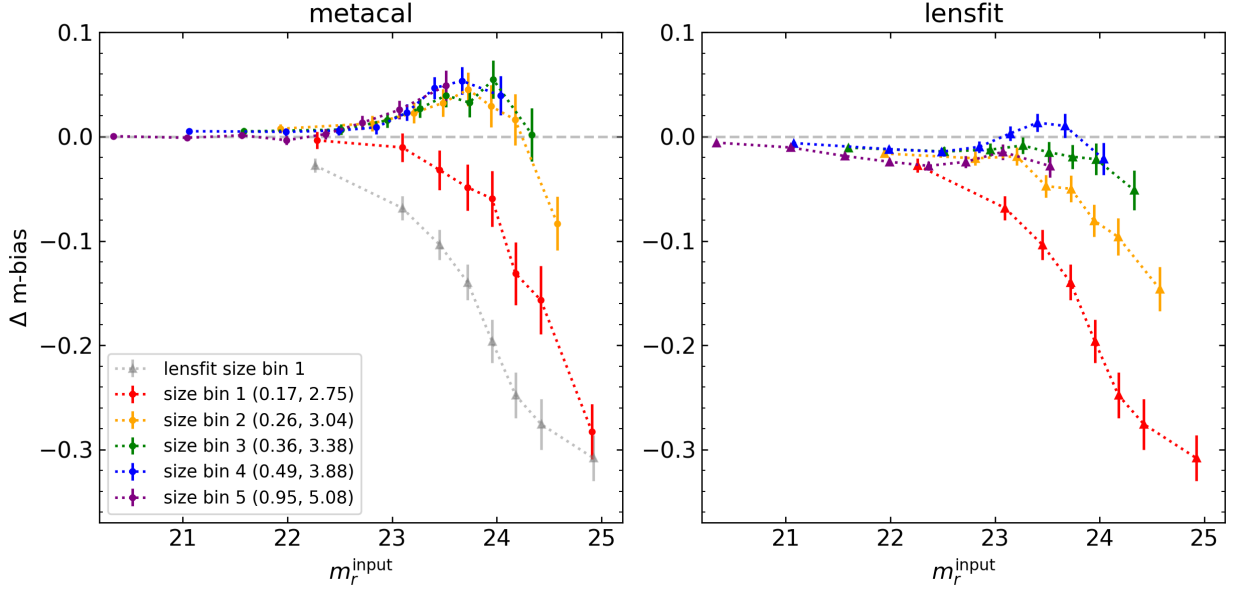


Fig. 7. The difference between $\Delta m = m_{\text{fix}} - m_{\text{sel}}$ as a function of m_r^{input} the input apparent magnitude when the sources are split into quantiles by input size. For reference, we also indicate the mean input effective radius in arcseconds, and the corresponding mean FLUX_RADIUS in pixels for each size bin. The left panel shows the results for METACALIBRATION, while the right panel shows the same for *lensfit*, using the sources in common between both catalogues. For reference, the grey points in the left panel correspond to the *lensfit* results for the smallest input galaxies, and should be compared to the red points for METACALIBRATION. The bump between $23 < m_r^{\text{input}} < 24$ is caused by blending (Hoekstra et al. 2021).

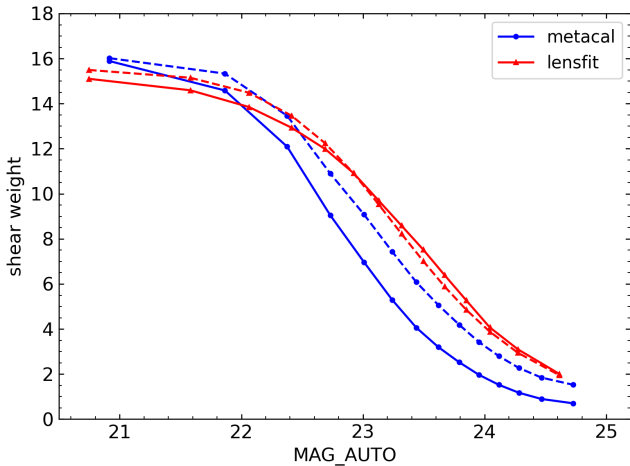


Fig. 8. Effective-shear-weight, $(1 + m)R^\gamma \sigma_\gamma^{-2}$ as a function of magnitude (solid blue line) compared to *lensfit* (solid red lines). For comparison, the dashed lines denote the corresponding inverse variance weights, $1/\sigma_\gamma^2$, which are used to compute the effective number density of sources. For *lensfit* the two weights are similar, but METACALIBRATION suppresses the contribution from faint sources, because of the extra factor R^γ . As discussed in the text, this slightly changes the effective source redshift distribution.

quence of our simple shape measurement and our estimate for the inverse variance.

The solid lines in Fig. 8 show how the effective shear weight⁶, $(1 + m)R^\gamma \sigma_\gamma^{-2}$, depends on apparent magnitude. For *lensfit* the weighting of the shears is similar to that of the shape

⁶ We call this the effective shear weight, which should be used to weight the redshift distributions used for the cosmological inference. This distinguishes it from the shear weight used to weight galaxies in the summary statistics.

estimates, so that $R^\gamma \equiv 1$, but our implementation of METACALIBRATION suppresses the contribution from faint sources, because of the extra factor R^γ . This changes the effective redshift distribution that is used to compute the predicted signal for a given sample. To stay close to the analysis of Li et al. (2023a) we ignore this here, and we used the inverse variance weights to compute the effective source redshift distribution. In Table 1 we report the corresponding mean weighted redshifts for SKiLLS for the tomographic bins.

For completeness, in Table 1 we also report the weighted mean redshifts when the additional weighting of the shears by $(1 + m)R^\gamma$ is included. For bins 2–5 the shifts are less than 0.01, comparable to the overall uncertainties in the mean redshifts. We do observe an unexpected large change for the first bin. When examining the mean shear response as a function of redshift in SKiLLS we found that it varied rapidly within the bin, which is not physical.⁷ Rather, we conclude this is a feature of SKiLLS, perhaps owing to the relatively small volume probed by the training set. This requires further study for stage IV projects, for which the observed changes in mean redshift exceed requirements (Euclid Collaboration: Mellier et al. 2024). For our purpose, we conclude that using the full shear weights or just the inverse variance weighting has no significant impact.

4.2. Correction for PSF anisotropy

We can expand the term describing the additive bias in Eq. (3) to distinguish between the various sources of additive bias. Specifically, we need to quantify the bias that arises from an incomplete correction for PSF anisotropy, also referred to as ‘PSF leakage’. To this end, we modify Eq. (3) to

$$\langle \epsilon_i^{\text{obs}} \rangle \equiv \frac{\langle \hat{\epsilon}_i \rangle}{\langle R^\gamma \rangle} = (1 + m_i) \gamma^{\text{true}} + \alpha_i e_i^{\text{PSF}} + \tilde{c}_i, \quad (13)$$

⁷ The shear response is mostly a function of galaxy size. We expect the mean galaxy size to be a smooth function of redshift.

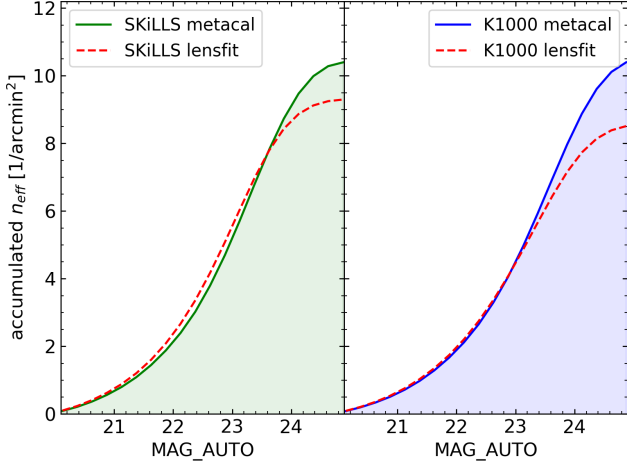


Fig. 9. Accumulated effective number density, n_{eff} , as a function of apparent magnitude for METACALIBRATION (solid line) and *lensfit* (dashed line). METACALIBRATION yields higher effective source density for both SKiLLS (left) and KiDS-1000 (right) for faint galaxies.

to express the result as a shear/ellipticity, and where α_i quantifies the leakage of the PSF anisotropy, e_i^{PSF} , into the shear estimate. We omitted the PSF modelling error term, $\beta_i \delta e_i^{\text{PSF}}$ here, because we used the input PSF from SKiLLS in our analysis. This is not a concern, because the impact of the PSF modelling itself was studied in Li et al. (2023a) who found it to be negligible for KiDS. Finally, any intrinsic additive bias and detector-related effects are captured by $\tilde{c}_i$.

In our setup of METACALIBRATION, we circularise the PSF before measurement. The results presented in Fig. 3 suggest that this works well, resulting in PSFs with negligible PSF anisotropy. In Fig. 10 we show the value for α for the tomographic bins. As anticipated, we find very small values. For reference, we also show the corresponding values for *lensfit* for the five bins used for KiDS-1000 (Li et al. 2023a). To allow for a fair comparison, we scaled the PSF ellipticities measured by the METACALIBRATION pipeline to match the *lensfit* values. This is needed, because of the circular weighting applied to the *lensfit* PSF measurements (Li et al. 2023a).

Although the recovered values are small, we observe a non-vanishing signal for the 4th tomographic bin. Moreover, the performance on actual data may differ. Therefore, to account for any residual PSF leakage, we follow Li et al. (2023a) and apply an empirical correction. To this end, we start by exploring how the PSF leakage varies with galaxy properties in SKiLLS. Here, we focus on signal-to-noise ratio and resolution, which are the dominant quantities for PSF-related shear bias. To quantify SNR, ν_{SN} , in our METACALIBRATION pipeline we use the following definition

$$\nu_{\text{SN}} \equiv \frac{\int d^2\mathbf{x} W(\mathbf{x}) I_{\text{obs}}(\mathbf{x}) / \sigma_{\text{noise}}^2(\mathbf{x})}{\sqrt{\int d^2\mathbf{x} W^2(\mathbf{x}) / \sigma_{\text{noise}}^2(\mathbf{x})}}, \quad (14)$$

where, $I_{\text{obs}}(\mathbf{x})$ is the observed image surface brightness, and $\sigma_{\text{noise}}(\mathbf{x})$ is the corresponding noise in the image. The optimal choice for the spatial weight function, $W(\mathbf{x})$, is the noiseless image, but here we use the same Gaussian weight function that was used in Eq. (2) to measure the quadrupole moments.

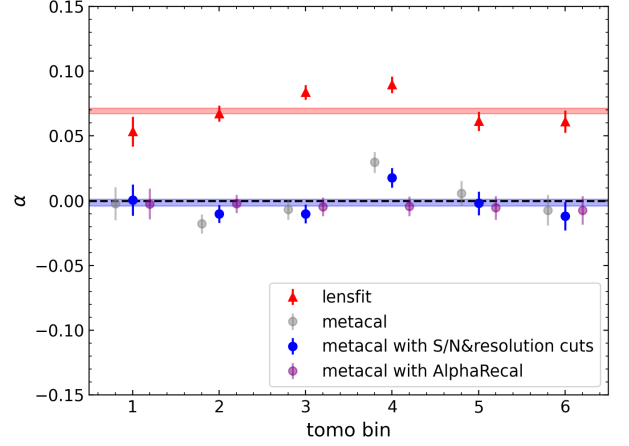


Fig. 10. Value of α for the different tomographic bins. The raw values from METACALIBRATION and *lensfit* are indicated by the grey and red points, respectively. The raw METACALIBRATION values are close to zero, and even more so after applying the cuts in signal-to-noise ratio and resolution (blue points; see text for details). After AlphaRecal, α is consistent with zero for all bins (purple points). To allow for a fair comparison, we scaled the PSF ellipticities measured by the METACALIBRATION pipeline to match the *lensfit* values.

We expect PSF leakage to be smaller for well-resolved objects. To capture this aspect, we define the resolution⁸

$$\mathcal{R} \equiv \frac{\sigma_{\text{PSF}}^2}{(\text{FLUX_RADIUS})^2}, \quad (15)$$

where the size of the galaxies is given by the half-light radius (FLUX_RADIUS) given by SOURCEEXTRACTOR. The PSF size is quantified by σ_{PSF} , which is computed using the stacked PSF model for individual sources, thus taking into account only the exposures that are used for METACALIBRATION. To determine σ_{PSF} , we fit an elliptical Gaussian to the PSF image using the `galsim.hsm.FindAdaptiveMom` routine in GalSim (Rowe et al. 2015), which returns the quadrupole moments. These are used to compute $\sigma_{\text{PSF}} \equiv |Q_{11}Q_{22} - Q_{12}Q_{21}|^{1/4}$.

Figure 11 shows one-dimensional projections of α as a function of resolution (top) and S/N (bottom). The ‘raw’ values are indicated by the orange points, which show that the amplitude of α quickly increases for small and noisy galaxies. To capture the dependence of the PSF leakage as a function of ν_{SN} and $\mathcal{R}$, we measured α on a 20×20 grid, and fitted the following function (Li et al. 2023a)

$$\alpha(\nu_{\text{SN}}, \mathcal{R}) = a_0 + a_1 \nu_{\text{SN}}^{-2} + a_2 \nu_{\text{SN}}^{-3} + b_1 \mathcal{R} + c_1 \mathcal{R} \nu_{\text{SN}}^{-2}. \quad (16)$$

The observed large values for α for small, faint galaxies, highlight the challenge of measuring their shapes robustly even after correcting the observed values for the dependence of α on $\mathcal{R}$ and ν_{SN} (referred to as ‘AlphaRecal’; see blue points in Fig. 11). We therefore cut our source sample, requiring $\nu_{\text{SN}} > 7$ and $\mathcal{R} < 0.6$. The applied cuts result in a 16% reduction in the number of sources. However, as Fig. 8 shows, those are down-weighted in practice. Hence, the impact on the effective number density of sources is negligible, while improving the robustness of the measurements. This cut reduces the mean α for the tomographic bins (blue points in Fig. 10) for the second and fourth

⁸ Our definition differs from that used in Li et al. (2023b), because they based it on parameters returned by *lensfit*. However, we expect that the rank-order of objects is largely similar in practice.

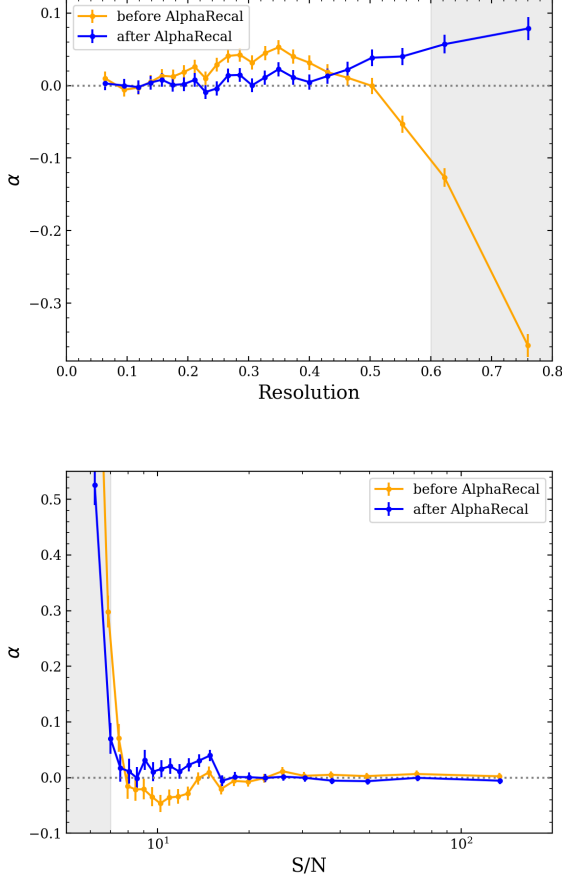


Fig. 11. 1-D projection of the alpha term measurement as functions of resolution and S/N, respectively. After applying the AlphaRecal process, alpha terms are consistent with zero in most bins (blue). We apply cuts based on remaining alpha terms for the low v_{SN} and resolution sources ($v_{\text{SN}} \leq 7$ and $R \geq 0.6$, grey shade).

tomographic bins. After accounting for AlphaRecal, the values are reduced further (see blue points in Fig. 11) and consistent with zero for all tomographic bins, as indicated by the purple points in Fig. 10.

4.3. Shear calibration parameters from SKILLS with variable shear

Up to now, we used image simulations with the same shear applied to all galaxies in the scene. In reality, the shear varies spatially and as a function of redshift. In principle, this is fine for METACALIBRATION, as its response to the applied shear is linear (see Section D), but only for isolated galaxies. However, the shear response is expected to be changed for blended sources at different redshifts (MacCrann et al. 2022), modifying the multiplicative bias when the redshift dependence of the input shear is included. To quantify this, we follow Li et al. (2023b) and re-simulate only sources that were identified as blended in the input catalogue, applying a realistic redshift-dependent shear.

We use the same simulations that were used by Li et al. (2023b) to calibrate the *lensfit* results, but adopt a different approach to calculate the change in the multiplicative bias. We take advantage of the robustness of our shape estimator, which allows us to measure the relevant quantities for individual sources. To determine how the multiplicative bias changes for the blended

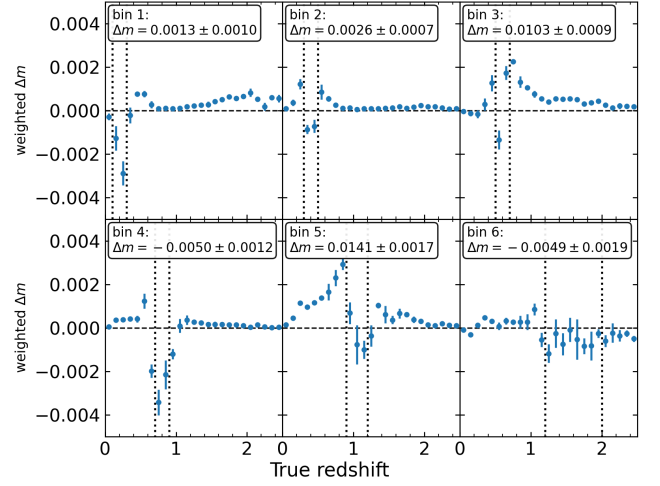


Fig. 12. Change in multiplicative bias, weighted by the normalised source redshift distribution, for blended-only sources with a realistic shear applied, as a function of redshift for each tomographic bin. The bin boundaries are indicated by the vertical dashed lines. $\Delta m = m_{\text{var}} - m_{\text{fix}}$ is computed using Eq. (19). The sum of all the data points in each tomographic bin yields the overall shift Δm for blended sources, which is reported in each panel.

sources using more realistic shears, γ_{var} , we compare the difference for the i^{th} source, $\Delta e_i = e_i^{\text{var}} - e_i^{\text{fix}}$. Using Eq. (8), this difference can be expressed in terms of the changes in m and R^γ for the ensemble of sources:

$$\Delta e_i = (1 + m + \Delta m)(R_i^\gamma + \Delta R_i^\gamma)(\gamma + \Delta \gamma_i) - (1 + m)R_i^\gamma \gamma, \quad (17)$$

where we dropped the subscript ‘fix’ for the quantities measured from the simulations with fixed shear. The change in multiplicative bias $\Delta m = m_{\text{var}} - m_{\text{fix}}$ is then given by

$$\Delta m = \frac{\langle \Delta \hat{e} \rangle}{\gamma \langle R^\gamma \rangle} - (1 + m + \Delta m) \left[\frac{\langle \Delta \gamma \rangle}{\gamma} \left(1 + \frac{\langle \Delta R^\gamma \rangle}{\langle R^\gamma \rangle} \right) + \frac{\langle \Delta R^\gamma \rangle}{\langle R^\gamma \rangle} \right]. \quad (18)$$

We assume that $|m| \ll 1$, while the changes in m and R^γ are small so that products of differences can be ignored. Hence, we can write

$$\Delta m \gamma \approx \frac{\langle \Delta \hat{e} \rangle}{\langle R^\gamma \rangle} - \left[\Delta \gamma + \frac{\langle \Delta R^\gamma \rangle}{\langle R^\gamma \rangle} \gamma \right], \quad (19)$$

so that Δm can be determined through linear regression of the right hand side against the input fixed shear, analogous to Eq. (3).

Figure 12, shows the values of Δm , multiplied by the normalised redshift distribution of the simulated blended sources, as a function of true redshift, to highlight which sources introduce biases. The sum of the points for each tomographic bin yields the total shift in multiplicative bias for the blended sources. We find that the largest deviations occur for bins 3 and 5, but the impact on the final bias estimates, m_{final} is small, as shown in Fig. 13, because the fraction of galaxies with a neighbour within $4''$ is only about one third. Compared to the fixed shear simulations (grey points), accounting for the redshift-dependent blending shifts the values by at most 1σ (blue points).

For reference, Fig. 13 also shows m_{final} for *lensfit* from Li et al. (2023b), who determined m_{var} for the blended sources. As those objects have small shears, the resulting uncertainty in m is large for a fixed number of sources. The advantage of the direct estimation of Δm is that the results are independent of the variable shear amplitude, resulting in much smaller uncertainties,

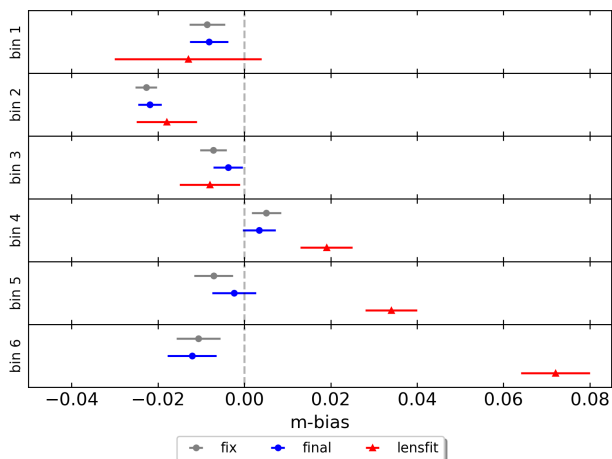


Fig. 13. Final estimates of the multiplicative bias, m_{final} , for each tomographic bin. The blue points include the redshift-dependent blending, which slightly shifts the values compared to the fixed-shear case (grey). For reference, we show the m_{final} for *lensfit* (red).

especially for lower redshift bins. Thanks to the improvement in precision, we required far fewer simulated sources: we reduced the numbers by a factor of eight compared to Li et al. (2023b).

4.4. Sensitivity tests

As shown in Fig. 1, SKILLS is able to reproduce the observed distributions of key galaxy properties well, but not perfectly. Equation (4) implies that this generally results in a change in the value of m . It is therefore useful to quantify the sensitivity of the multiplicative bias to such deviations. Figure 6 suggests that small mismatches in the distribution of apparent magnitudes will not have much impact, but Fig. 1 also indicates differences in the morphologies of the galaxies, such as their sizes and shapes. These also change the recovered biases (Hoekstra et al. 2017). We therefore follow Li et al. (2023a), and quantify the sensitivity of METACALIBRATION to rather extreme changes in the half-light radii, axis ratios and Sérsic indices of the input galaxies. We use the same input distributions that were used to test the sensitivity of *lensfit* in appendix C of Li et al. (2023a).

In Fig. 14, we show the resulting shifts in multiplicative bias, Δm , for the different extreme setups. From left to right, we shifted the half-light radii, axis ratios or Sérsic indices to higher (orange) or lower (green) values. The corresponding *lensfit* results from Li et al. (2023a) are shown by the light-coloured points. The sensitivity of both METACALIBRATION and *lensfit* to these extreme changes is well within requirements. Nonetheless, in the case of the source axis ratios, we find that the METACALIBRATION results are more robust, with *lensfit* showing an increased sensitivity for the higher redshift bins.

These tests suggest that our results are robust, but Euclid Collaboration: Csizi et al. (2024) recently showed that considering more complex morphologies may still impact the results. The changes are, however, sufficiently small, so that we can conclude that the galaxy morphologies in SKILLS are sufficiently realistic for KiDS cosmic shear measurements.

Changes in the observing conditions or the star density can also influence the shear bias (Hoekstra et al. 2015, 2017). In principle, SKILLS samples these variations, so that the final shear bias is representative for the survey as a whole. Provided the spatial variation in multiplicative bias is small, this is sufficient for cosmological parameter inference (Kitching et al. 2019).

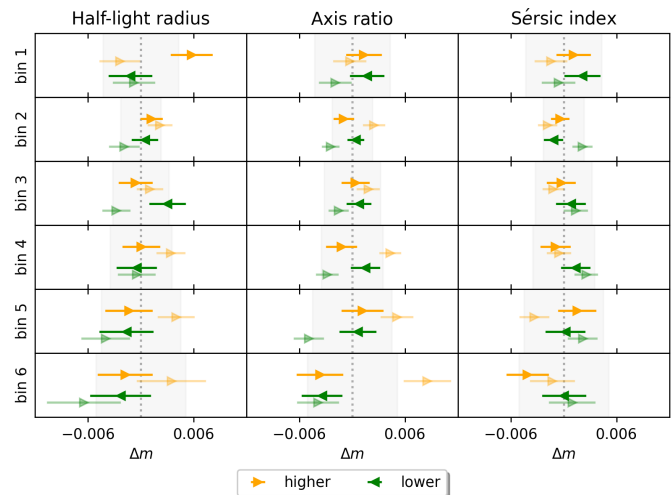


Fig. 14. Change in multiplicative bias to extreme changes in the input morphologies. The setup is the same as used by Li et al. (2023a). Their results for *lensfit* are shown as the fainter points with error bars. From left to right, we shifted the half-light radii, axis ratios or Sérsic indices to higher (orange) or lower values (green) values. The sensitivity of both METACALIBRATION and *lensfit* to these extreme changes is well within the estimated errors (grey shade).

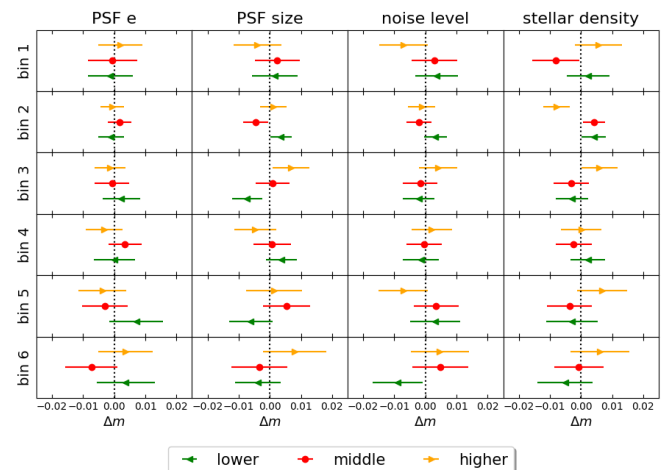


Fig. 15. Change in multiplicative bias when the data are split by, from left to right, PSF ellipticity, PSF size, background noise level, and stellar density. In each case, the simulations were split into three equally populated samples. The results for the lower (green), middle (red) and higher (orange) values are consistent with the average bias for the full sample. Only for the second tomographic bin do we observe a hint of a sensitivity to the star density.

Nonetheless, it is useful to explore the sensitivity of the results to such variation.

We follow Li et al. (2023b), and quantify the sensitivity to four key parameters: PSF ellipticity, PSF size, background noise level, and stellar density. To this end, we split the sources into three subsamples based on the parameter of interest and compute the change in multiplicative bias with respect to the full sample for each tomographic bin. We ensured that the variation between the three subgroups is large enough to capture the diversity in the data. For example, the star density in the mid-group/high-group is twice/four times higher than the low-group (3 stars/arcmin²). We do not observe significant changes in m , as shown by the results in Fig. 15. The largest deviation is seen for the high star

density case in the second tomographic bin. We explore this further in Section A.3, but conclude that our METACALIBRATION shear estimates are robust against the variations we expect across the survey.

5. KiDS data preparation for cosmological analysis

The main aim of this paper is to verify the methodology of METACALIBRATION and to investigate its impact on the cosmological inference. Therefore, we limit our cosmological analysis as close as possible to the KiDS-1000 cosmic shear analysis with *lensfit*, using the five tomographic bins. Although METACALIBRATION provides accurate shear estimates for the sixth bin, including these measurements would have required changing the redshift calibration methodology for the METACALIBRATION catalogue, which we leave for future work.

Nonetheless, the source selection differs because METACALIBRATION provides shape estimates for objects that are excluded by *lensfit*. We discuss the impact of these in more detail in Section A, reporting how each step affects the source sample. We apply the S/N and resolution cuts as described in Section 4.2. Another important selection in the source number relies on the reliability of the photometric redshift calibration. As discussed in Section 5.1, this step filters out the sources that lack sufficient counterparts with spectroscopic information.

Once the final sample is selected, we apply the AlphaRecal step to remove the residual PSF leakage (described in Section 4.2). There was no need to filter out objects with extreme shears after this correction. The same procedures were applied to the SKILLS data to derive the final values for m for each tomographic bin. We report these as m_{final} in Table 2.

5.1. Source redshift distributions

To assign galaxies to a tomographic bin, we use the photometric redshift, z_B , determined by BPZ (Benítez 2000). However, METACALIBRATION yields a different shape catalogue with modified shear weights. As a result, the corresponding source redshift distributions are different from those used by (Li et al. 2023a). To account for this, we use the same redshift calibration methodology as described in Hildebrandt et al. (2021); van den Busch et al. (2022). We use the METACALIBRATION weighting defined by Eq. (10), but ignore the re-weighting of the shears caused by the redshift dependence of R^γ and m . Fully accounting for the latter would require changing the redshift calibration pipeline, while the impact is negligible.⁹

The calibration of the bin-wise redshift distributions follows the method developed in Wright et al. (2020). A SOM is trained on a compilation of spectroscopic datasets using the r -band magnitude and colours derived from 36 combination of KiDS photometric bands. For the purpose of consistency with previous work, we adopt the SOM trained by van den Busch et al. (2022) on an extended compilation of spectroscopic data (see appendix B therein). We then parse the KiDS source sample into the SOM and identify regions of the colour space that are not sufficiently covered and represented by the calibration data (see van den Busch et al. 2022, section 3.2). We remove all galaxies in these parts of the colour space, per tomographic bin, sacrificing a small

⁹ For reference, we computed the impact using SKILLS and report the resulting mean redshifts in Table 1. This shows that we can safely ignore the additional redshift weighting of the shear for KiDS-1000. As discussed in Section 3 we attribute the change in the first bin to limitations of the input catalogue of SKILLS.

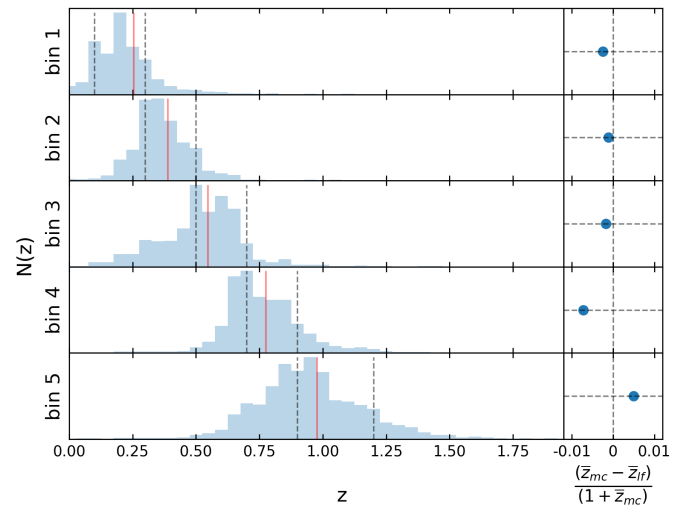


Fig. 16. Redshift distributions of KiDS-1000 METACALIBRATION (left panels) for each tomographic bin. Sources were selected based on their z_B , with the bin limits indicated by the dashed vertical lines. The red vertical line corresponds to the mean redshift of the bin. The change with respect to the *lensfit* catalogue is shown on the right. The mean redshifts are similar.

fraction of sources for a more robust redshift calibration of the remaining ‘gold-sample’ selected based on ‘gold-flag’.

After applying the gold selection, we obtain the final redshift estimate for each tomographic bin by computing the weighted redshift distribution of the spectroscopic calibration data. These so-called SOM weights are constructed by weighting the calibration data in each SOM cell such that their number density matches the METACALIBRATION. As shown in Fig. 16, the obtained redshift distributions of the five tomographic bins are very similar between METACALIBRATION and *lensfit*. The deviation of mean redshifts for each tomographic bin from the *lensfit* values are -0.003, -0.002, -0.003, -0.013, 0.010, respectively. Among them, the 4th tomographic bin deviates the most, with $dz/(1+z) = -0.007$. An important aspect of the redshift distribution estimation and calibration procedure within KiDS is the incorporation of informative priors on the bias of each tomographic bin’s $\langle z \rangle$. In Hildebrandt et al. (2021), these informative priors are derived for each tomographic bin using MICE2 simulations (Fosalba et al. 2015, 2014), and are sensitive to the make-up of the wide-field shear sample and the sample of calibrating sources used to estimate the $N(z)$. As we use a new wide-field sample here (i.e. the KiDS sample selected and weighted via metacalibration, rather than *lensfit*), the biases estimated in Hildebrandt et al. (2021) are not formally directly applicable to the source galaxies here. However, there is considerable overlap between the source samples from *lensfit* and metacalibration, and our estimated $N(z)$ are sufficiently similar to those from *lensfit* (as described above), that we opt to not recompute the $N(z)$ bias parameters using the simulations for this work. Instead, we adopt the bias estimates and prior ranges from Hildebrandt et al. (2021), shown as the last column in Table 2. In the future analysis of KiDS-Legacy with METACALIBRATION, this will be improved through the recomputation of bespoke bias parameters for the galaxy sample.

The effective number density n_{eff} , average shear response $\langle R^\gamma \rangle$, and shear variance $\sigma_{\gamma,i}$, are reported in Table 2 for each tomographic bin. We also present results for the joint METACALIBRATION sample, which yields $n_{\text{eff}} = 7.77 \text{ arcmin}^{-2}$, $\sim 16\%$ larger than *lensfit*. Compared to the results reported in Table 1, the im-

Table 2. METACALIBRATION summary statistics of five tomographic bins for the cosmological analysis.

Bin	Photo- z range	$\langle R^y \rangle$	n_{eff} (metacal)	n_{eff} (lensfit)	$\sigma_{\gamma,i}$	$m_{\text{SOM,gold}}$	m_{final}	$\sigma_{m,\text{final}}$	$\delta z = z_{\text{est}} - z_{\text{true}}$
1	$0.1 < z_B \leq 0.3$	0.21	0.74	0.68	0.308	-0.0061	-0.0056	0.0056	0.000 ± 0.0096
2	$0.3 < z_B \leq 0.5$	0.19	1.44	1.30	0.289	-0.0178	-0.0169	0.0027	0.002 ± 0.0114
3	$0.5 < z_B \leq 0.7$	0.20	2.28	1.97	0.313	-0.0093	-0.0054	0.0039	0.013 ± 0.0116
4	$0.7 < z_B \leq 0.9$	0.19	1.57	1.39	0.293	-0.0050	-0.0069	0.0042	0.011 ± 0.0084
5	$0.9 < z_B \leq 1.2$	0.20	1.74	1.35	0.330	-0.0091	-0.0036	0.0057	-0.006 ± 0.0097
1-5	$0.1 < z_B \leq 1.2$	0.19	7.77	6.69	0.306	-0.0104	-0.0089	0.0018	N/A

Notes. This is the summary statistics of KiDS-1000 metacal catalogue with the selections made for the cosmic shear cosmological analysis (comparable to the KiDS-1000-v2 catalogue (Li et al. 2023a)). The m-bias (gold) was recalculated based on the selections made during SOM photo- z calibration and AlphaRecal (the residual PSF correction). Through the extra selections, $\sim 25\%$ of effective number density is reduced (see Table 1). The total source number density (n_{eff}) is 7.77 arcmin^{-2} for the effective area 771.90 deg^2 .² The effective source numbers were calculated based on the dispersion of shear (σ_γ , Eq. (9)). Compared to the *lensfit* catalog, total n_{eff} increased by $\sim 16\%$.

provement in number density is somewhat reduced, because the SOM (Gold) selection preferentially removes faint sources. As shown in Wright et al. (2025a), this can be alleviated by using a weight rather than a binary selection for the faint sources.

5.2. Cosmic shear 2 point-statistics

The cosmic shear signal can be quantified using the two-point shear correlation functions, $\xi_{\pm}^{ij}(\theta)$, for samples of galaxies in tomographic bins i and j , separated on the sky by an angle θ . This can be estimated from the shape catalogue using

$$\xi_{\pm}^{ij}(\theta) = \frac{\sum_{ab} w_a w_b [e_i^j(\mathbf{x}_a) e_j^i(\mathbf{y}_b) \pm e_x^i(\mathbf{x}_a) e_x^j(\mathbf{y}_b)]}{\sum_{ab} w_a w_b R_a^\gamma R_b^\gamma}, \quad (20)$$

where the weight w is the shear weight given by Eq. (10). The tangential and cross ellipticities $\epsilon_{t,x}$ are computed with respect to the vector $\mathbf{x}_a - \mathbf{y}_b$ that connects the galaxy pair (a, b) .

For the cosmological inference, we use the complete orthogonal sets of E/B -integrals (COSEBIs; Schneider et al. 2010). These provide a convenient way to compress the information contained in $\xi_{\pm}(\theta)$, and cleanly decompose the signal into E - and B -modes, which correspond to curl-free gradient distortions and curl distortions, respectively. The COSEBIs are given by

$$E_n = \int_{\theta_{\min}}^{\theta_{\max}} d\theta \theta [T_{+n}(\theta) \xi_+(\theta) + T_{-n}(\theta) \xi_-(\theta)], \quad (21)$$

$$B_n = \int_{\theta_{\min}}^{\theta_{\max}} d\theta \theta [T_{+n}(\theta) \xi_+(\theta) - T_{-n}(\theta) \xi_-(\theta)]. \quad (22)$$

where $T_{\pm n}(\theta)$ are filters defined over a chosen angular range ($\theta_{\min}$, $\theta_{\max}$). Here, we use the range $2 - 300 \text{ arcmin}$, and the maximum COSEBI mode is chosen to be $n_{\max} = 5$, following the setup of Li et al. (2023a) for consistency. This choice of scale range reduces the impact of the uncertainty in modelling the baryonic feedback on scales smaller than 2 arcmin .

The measured COSEBIs using our METACALIBRATION catalogue are presented in Fig. 17 for the various tomographic bin combinations. The blue lines in the panels showing E_n indicate the best-fit model to the measurements, with a reduced $\chi^2 \approx 1.2$ (see below for more details). The uncertainties in the measurements are estimated with an analytical covariance, whose recipe is described in more detail in Asgari et al. (2020); Joachimi et al. (2021). Although the analysis of KiDS DR5 (Wright et al. 2024) employs a new covariance code by Reischke et al. (2025)¹⁰, we

continue to use the previous code as it is implemented in the KiDS-1000 pipeline. The two codes have been verified to be consistent with both the new implementation and simulation results. In the implementation used in this work, the main difference is the idealised treatment of the survey geometry in the mixed term and the model for non-linearities. These effects, however, were found to not affect the cosmological results. We therefore can safely use the old covariance code for the analysis here. For completeness, we will also carry out a single run with the updated covariance matrix in Section 6 (see Fig. 21).

The covariance consists of four distinct components: the super-sample covariance, which arises from modes outside the survey volume; the non-Gaussian covariance, which is due to the non-linearity of structure formation; the multiplicative shear bias uncertainty; and the disconnected (Gaussian) part. The first two components are calculated using the halo model (see Asgari et al. 2023, for a review). For the Gaussian covariance, we specifically account for the exact impact of the mask and weights (e.g. shape measurement weights) in the shape noise term by directly using the weighted pair counts. All other terms are projected from the harmonic space covariance, $C(\ell)$, to the COSEBIs through Eq. (25). For brevity, we direct readers to Reischke et al. (2025) for more detailed information on the implementation and formalism.

5.3. COSEBIs B -mode null test

To first order, weak lensing introduces correlations in the galaxy shapes that are curl-free. Some real-life effects, such as source redshift clustering, complicate this simple picture (e.g. Schneider et al. 2002). But we can safely ignore these for our analysis (Linke et al. 2025). Hence, we expect B_n to be consistent with zero, thus providing a non-trivial test for residual systematics.

The measurements for B_n are shown as the red points in Fig. 17, as well as the corresponding p -values of the null-test. We use the same range in angular scales ($2 - 300 \text{ arcmin}$) and maximum COSEBI mode ($n_{\max} = 5$) as used to compute E_n (black points). This differs from the setup used by Li et al. (2023a) for this test. Therefore, we also recalculated the B_n for *lensfit* for the same scale range and the number of COSEBI modes. The results are shown by the grey points. For our METACALIBRATION measurements, the smallest, $p = 0.02$ is found for the auto-correlation of the second tomographic bin (we consider the value, 0.02 , acceptable), where the contamination from stars is largest (check Section A.3 for the details of star contamination). Although the other auto-correlations also tend to have small p -values, the other bin

¹⁰ <https://github.com/rreischke/OneCovariance>

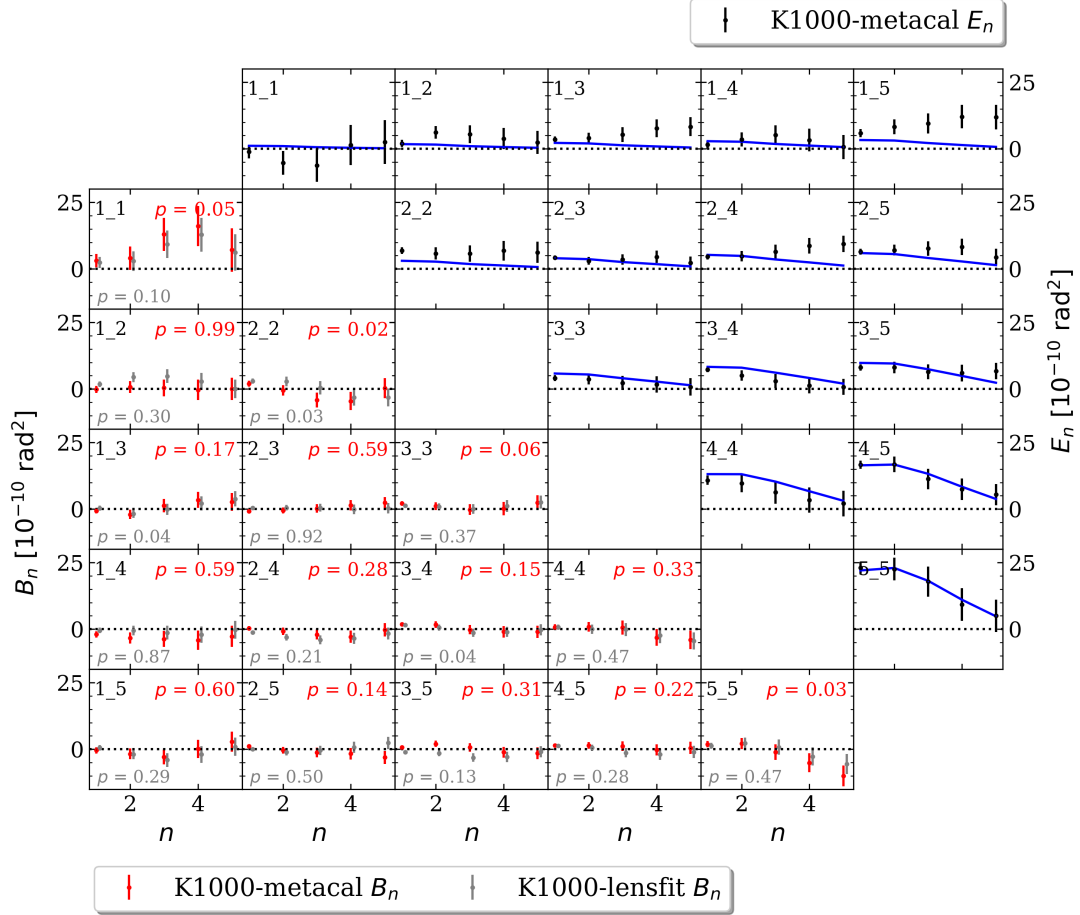


Fig. 17. Measurements for the first five COSEBI E_n (black) with the best-fit theory lines (blue) and B_n (red) measurements with their p-values (red) from the KiDS-1000 METACALIBRATION catalogue, converted from the real scale signals, $\xi_{\pm}(\theta)$ with the scale range, [2.0, 300.0] arcmin. As a comparison, the *lensfit* B_n measurements were denoted in grey with each corresponding p-value (grey). The theory with best-fit values from the fiducial cosmology analysis is fitted with the $\chi^2 \simeq 87.5$.

combinations show acceptable values. Moreover, the p -values are overall comparable to the *lensfit* results, while the estimated uncertainties are in fact smaller for METACALIBRATION. Hence, the robustness of the METACALIBRATION signal is at least comparable to that of *lensfit*. Following the previous conventional test ($n_{\max} = 20$), we also checked all the B-modes of the same scale range (2 – 300 arcmin) satisfy the minimum threshold, $p = 0.05$.

5.4. Accuracy requirement and validation of the first-order systematics model

The PSF anisotropy is the dominant source of additive bias. Although the measurements of B_n provide a useful test, any curl-free contributions would not be captured, or underestimated (see e.g. Hoekstra 2004). Therefore, we also directly examine the contribution from PSF systematics to ξ_+ . To do so, we follow Bacon et al. (2003) and define

$$\xi_+^{\text{sys}} = \frac{\langle \epsilon^{\text{obs}} \epsilon^{\text{PSF}} \rangle^2}{\langle \epsilon^{\text{PSF}} \epsilon^{\text{PSF}} \rangle}. \quad (23)$$

As discussed in detail in section 3.5 of Giblin et al. (2021), the main aim of ξ_+^{sys} is to quantify the impact PSF leakage and PSF modelling errors. In Fig. 18 we compare ξ_+^{sys} to the predicted cosmic shear signal for the tomographic bin combinations we use for the cosmological inference. We find that the amplitude

of ξ_+^{sys} is well within 10% of the uncertainty in the observed ξ_+ (indicated by the light blue regions) for most scales and tomographic bin combinations.

As is the case for previous KiDS analyses, we do observe an increase around 100 arcminutes, which coincides with the range where the denominator of Eq. (23) almost vanishes. The deviation is largest for the second tomographic bin, which may be impacted by our suboptimal selection of stars using the *lensfit* star flag (see Section A.3).¹¹

To explore this further, we examine possible contributions to the numerator. If the fraction of stars is f_{star} , we have

$$\begin{aligned} \langle e^{\text{obs}} e^{\text{PSF}} \rangle &= (1 - f_{\text{star}}) \alpha_{\text{gal}} \langle e^{\text{PSF}} e^{\text{PSF}} \rangle + f_{\text{star}} \alpha_{\text{star}} \langle e^{\text{PSF}} e^{\text{PSF}} \rangle \\ &\quad + \langle \delta e^{\text{PSF}} e^{\text{PSF}} \rangle \\ &= \alpha \langle e^{\text{PSF}} e^{\text{PSF}} \rangle + \langle \delta e^{\text{PSF}} e^{\text{PSF}} \rangle, \end{aligned} \quad (24)$$

where we explicitly distinguished between the PSF leakage for a galaxy and that of a star in the first step. This suggests that PSF leakage is naturally accounted for using our empirical ‘AlphaRecal’ procedure, as we expect $\alpha \approx 0$.

¹¹ Here, we focus on additive bias, but stars also lower the lensing signal, because they are not lensed. As SKILLS samples the distribution of stars across the survey, this is accounted for in our estimates of m .

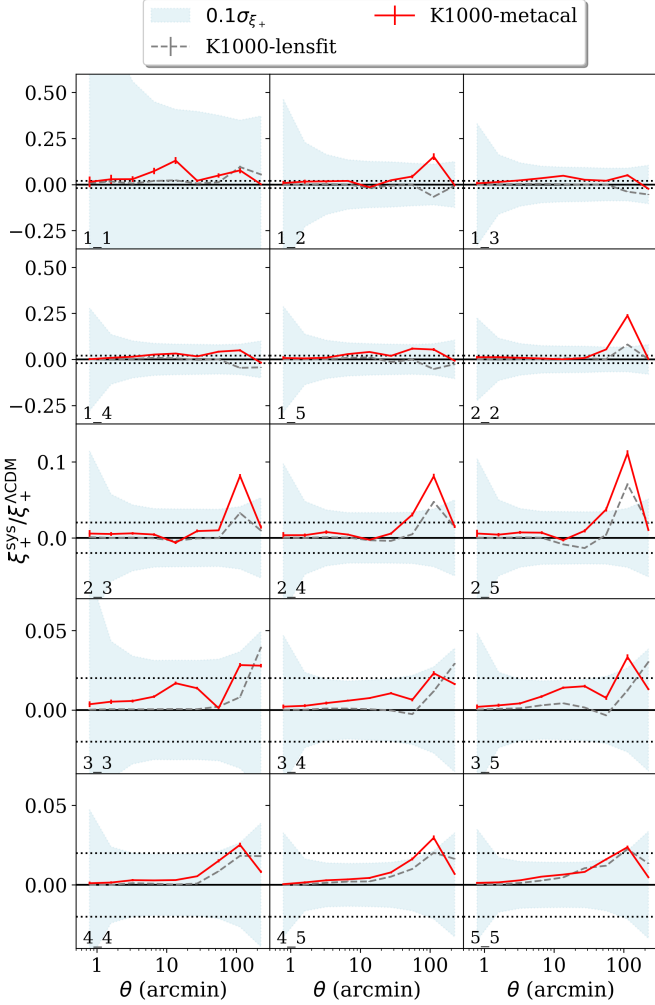


Fig. 18. Ratio of the PSF systematic contribution to the amplitude of the cosmic shear signal with the fiducial cosmology (same as figure 8 in Giblin et al. 2021). While most of the systematic signals are within the range of $0.1\sigma_{\xi_+}$.

Although PSF modelling residuals are small (Giblin et al. 2021), it is possible that residuals correlate with the PSF ellipticity, especially at the edge of the field of view. If $\langle \delta e^{\text{PSF}} e^{\text{PSF}} \rangle$ does not vanish, while $\langle e^{\text{PSF}} e^{\text{PSF}} \rangle$ does, we expect an increase in ξ_+^{sys} . Interestingly, we find that the amplitude at $100''$ is higher for pointings with low star densities. For such pointings, the PSF model itself is more noisy, thus supporting the notion that the bump in ξ_+^{sys} is caused by residual PSF modelling errors. However, given the negligible impact on the cosmological results, further exploration is beyond the scope of the paper.

6. Cosmological results

Our inference of cosmological parameters follows the previous KiDS-1000 analyses closely (for details, see Asgari et al. 2021). Consequently, we only summarise the most salient aspects here. The cosmological information is contained in the values for E_n , Eq. (21), presented in Fig. 17. These are related to the E -mode angular power spectrum, $C_{EE}(\ell)$, through

$$E_n = \int_0^\infty \frac{d\ell}{2\pi} C_{EE}(\ell) W_n(\ell), \quad (25)$$

where the weight functions $W_n(\ell)$ are Hankel transforms of $T_\pm(\theta)$ (Asgari et al. 2012).

$C_{EE}^{ij}(\ell)$ for the i th and j th tomographic bins is the sum of three terms that capture the contribution from gravitational lensing (G), intrinsic alignments of galaxies (I) and their cross-correlation:

$$C_{EE}^{ij}(\ell) = C_{GG}^{ij}(\ell) + C_{GI}^{ij}(\ell) + C_{II}^{ij}(\ell). \quad (26)$$

Using the Limber approximation (LoVerde & Afshordi 2008; Kilbinger et al. 2017), these angular power spectra can be computed using the matter power spectrum, $P_m(\ell, \chi)$, through

$$C_{XY}^{ij}(\ell) = \int_0^{\chi_H} d\chi \frac{q_X^i(\chi) q_Y^j(\chi)}{f_K^2(\chi)} P_m\left(\frac{\ell + 1/2}{f_K(\chi)}, \chi\right). \quad (27)$$

where X and Y stand for G or I, χ is the radial comoving distance that is integrated from the observer ($\chi = 0$) to the horizon ($\chi = \chi_H$), and $f_K(\chi)$ is the comoving distance. The kernels, q_X^i , depend on the source redshift distribution, $n_i(\chi)$. The lensing kernel is given by

$$q_G^i(\chi) = \frac{3H_0^2 \Omega_m}{2c^2} \int_\chi^{\chi_H} d\chi' n^i(\chi') \frac{f_K(\chi' - \chi)}{f_K(\chi')}, \quad (28)$$

and for the intrinsic alignment we choose a kernel that produces the NLA model Bridle & King (2007)

$$q_I^i(\chi) = -A_{IA} C_1 \rho_{\text{crit}} \frac{\Omega_m}{D_+(z)} \left(\frac{1+z}{1+z_0} \right)^\eta n^i(\chi), \quad (29)$$

with A_{IA} the intrinsic alignment parameter. The other parameters are $C_1 = 5 \times 10^{-14} h^{-1} \text{Mpc}^3$, ρ_{crit} is the critical density today, and $D_+(z)$ is the linear growth factor.

As our fiducial model, we adopt $\eta = 0$ and a pivot redshift $z_0 = 0.3$. Moreover, to ensure a direct comparison with Li et al. (2023a), we use the same prior choices for the twelve free parameters (cosmological and nuisance parameters), which are summarized in Table 3.

The uncertainties in m for each tomographic bin were included in the covariance matrix, while the two-point statistics were corrected using the value of m . The photo- z calibration parameters from the clustering redshift ($\delta_{z,i}$) were used as priors for the Monte Carlo Markov Chain (MCMC) runs. We used polychord (Handley et al. 2015a,b) as the fiducial sampling method for the MCMC runs, while the consistency tests in Section 6.2 used multinest (Feroz et al. 2009) to speed up the computations.

The main differences between the original KiDS-1000 analysis by Asgari et al. (2021) and the revision by Li et al. (2023a) pertain to changes in the baryonic feedback and intrinsic alignment models. Asgari et al. (2021) used the halo-model parametrization from Mead et al. (2015). Here, we follow Li et al. (2023a) and use an updated baryonic feedback model (HMCode-2020; Mead et al. 2021) based on the temperature of the AGN feedback (T_{AGN}) derived from a suite of large-volume hydrodynamic simulations, BARYONS and HALOES of MASSIVE SYSTEMS (BAHAMAS; McCarthy et al. 2017).

Asgari et al. (2021) used a wide prior $A_{IA} \in [-6.0, 6.0]$, while much of this prior range is excluded by models that aim to account for available observational constraints (e.g. Fortuna et al. 2021). Instead, Li et al. (2023a) adopted a prior $A_{IA} \in [-0.2, 1.1]$, which is considerably narrower. In general, it is preferable to reduce the prior ranges whenever possible, to avoid issues with sampling the high-dimensional parameter spaces. We investigate the impact of the different choices for these baryonic feedback models and intrinsic alignment priors in Section 6.2.

Table 3. METACALIBRATION fiducial cosmology priors

Parameter	Prior range
S_8	[0.1, 1.3]
ω_c	[0.051, 0.255]
ω_b	[0.019, 0.026]
h	[0.64, 0.82]
n_s	[0.84, 1.1]
$\log_{10}(T_{\text{AGN}}[\text{K}])$	[7.3, 8.0]
A_{IA}	[-0.2, 1.1]
δz_i	$\mathcal{N}(\mu_{z_i}; \mathbf{C}_{\delta z})$

Notes. The redshift shift (δz_i) for each tomographic bin, $\mathcal{N}(\mu_{z_i}; \mathbf{C}_{\delta z})$ refer correlated priors among tomographic bins with means μ_{z_i} and covariance $\mathbf{C}_{\delta z}$ reported in Table 2 and Hildebrandt et al. (2021).

6.1. Cosmological constraints from the METACALIBRATION pipeline

The blue lines in Fig. 17 show the best-fit model to the measurements to E_n . The model fit with the final best-fit parameters is fair with $\chi^2 = 87.5$. For reference, the best fits using the *lensfit* measurements correspond to $\chi^2 = 62.7$ (Li et al. 2023a) with the same degrees of freedom (DoF) $\gtrsim 75 - 4.5$. Due to the narrower A_{IA} prior that affected its constraint, the DoF is estimated slightly bigger than the previous KiDS studies, Asgari et al. (2021); van den Busch et al. (2022) with $\text{DoF} = 75 - 4.5$ that yielded $\chi^2 = 82.2$ and $\chi^2 = 63.2$, respectively.

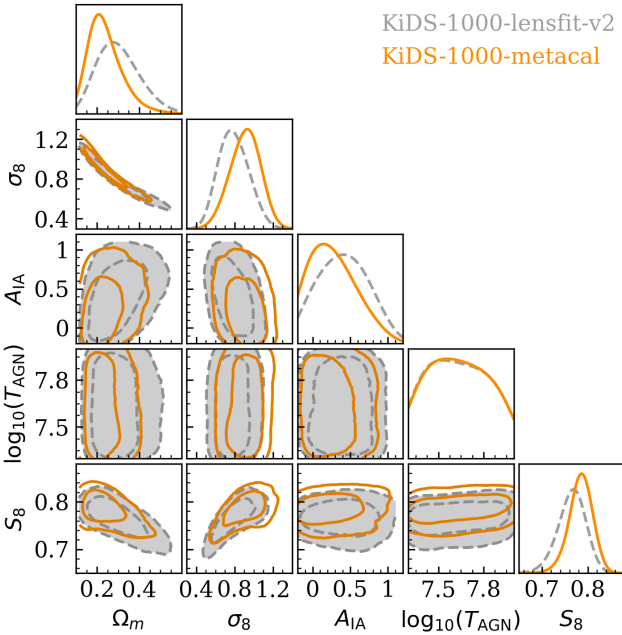


Fig. 19. METACALIBRATION cosmology compared to *lensfit*, both from the fiducial setting with polychord sampler and priors described in Table 3. The parameter constraints are summarized in Table 4 below. While the results from *lensfit* and METACALIBRATION are consistent, the constraint on S_8 from METACALIBRATION is 28% tighter than the one from *lensfit* and 0.5σ closer to the Planck value.

We compare the posterior distributions for the parameters listed in Table 4 for the METACALIBRATION and *lensfit* analyses in Fig. 19. The results agree well, with METACALIBRATION yielding slightly tighter constraints. The METACALIBRATION results prefer a

Table 4. METACALIBRATION fiducial parameter constraints

Parameter	MAP (fiducial)	Mean	Max
Cosmological Parameter			
S_8	$0.789^{+0.020}_{-0.024}$	$0.784^{+0.024}_{-0.021}$	$0.787^{+0.023}_{-0.025}$
Ω_m	$0.167^{+0.097}_{-0.041}$	$0.240^{+0.039}_{-0.090}$	$0.203^{+0.086}_{-0.058}$
σ_8	$1.057^{+0.155}_{-0.220}$	$0.908^{+0.154}_{-0.136}$	$0.928^{+0.142}_{-0.163}$
Σ_8	$0.777^{+0.018}_{-0.026}$	$0.778^{+0.022}_{-0.020}$	$0.781^{+0.020}_{-0.025}$
Astrophysical Parameter			
A_{IA}	$-0.105^{+0.499}_{-0.095}$	$0.263^{+0.188}_{-0.411}$	$0.092^{+0.396}_{-0.228}$

Notes. The second column lists the multivariable maximum posterior (MAP) and its 68% credible interval using its projected joint highest posterior density (PJ-HPD). The third and the last columns list the mean and maximum values from the marginal posterior of each parameter, respectively. Σ_8 is defined as $\sigma_8(\Omega_m/0.3)^{\alpha_{\text{best}}}$, similar to the definition of S_8 using the power of 0.5, with the best-fit value of $\alpha_{\text{best}} = 0.527$ in the plane of σ_8 and Ω_m along the degeneracy direction of posterior contour. For the A_{IA} constraint, only the upper bound of is trustworthy, because the lower bound is limited by the prior limit. The baryonic feedback parameter, $\log_{10}(T_{\text{AGN}})$, is not constrained, because of the narrow prior range.

smaller Ω_m , which pushes the S_8 constraint only slightly higher along a similar degeneracy direction. Although both analyses are based on the same multi-band imaging data, and we attempted to stay close to the choices made by Li et al. (2023a), we note that achieving this level of consistency is not trivial. For one, the METACALIBRATION catalogue includes more galaxies, especially for the high-redshift bins, while the shear calibration is quite different. Hence, the agreement is an important validation of the robustness of the calibration pipeline.

The inferred value for A_{IA} depends on the sample selection (Fortuna et al. 2021), but may also differ between shape measurement algorithms. For our METACALIBRATION measurements, we adopted a rather compact weight function, which minimises the sensitivity to the outer regions of galaxies, which may be more strongly aligned (Georgiou et al. 2019). Interestingly, we find a lower value for A_{IA} compared to Li et al. (2023a). However, a more quantitative exploration of the choice of weight function and the constraints on A_{IA} is beyond the scope of this paper.

Given the interest in the constraints on S_8 , we present the corresponding projected 2D posterior distribution for the parameters S_8 and Ω_m for the fiducial setup in Fig. 20. We find $S_8 = 0.789^{+0.020}_{-0.024}$, using the multivariable maximum posterior (MAP) and its projected joint highest posterior density (PJ-HPD) (see detailed discussion of the estimate in section 6.4 of Joachimi et al. 2021). We also show the posteriors from the *Planck* CMB measurements (Planck2018, TT,TE, EE + lowE + lensing; Planck Collaboration VI 2020), which yield $S_8 = 0.832 \pm 0.013$. We also measured the value of $\Sigma_8 \equiv \sigma_8(\Omega_m/0.3)^{\alpha_{\text{best}}} = 0.777^{+0.018}_{-0.026}$, where the best-fit value of $\alpha_{\text{best}} = 0.527$ was obtained from the orientation of the σ_8 and Ω_m posterior. The corresponding *Planck* value is $\Sigma_8 = 0.835 \pm 0.016$. Similar to previous analysis of KiDS-1000 data (Asgari et al. 2021; Li et al. 2023a) and other lensing studies (Amon et al. 2022; Secco et al. 2022; Li et al. 2023c; Dalal et al. 2023), we find a low value for S_8 . However, compared to the findings of Li et al. (2023a), the

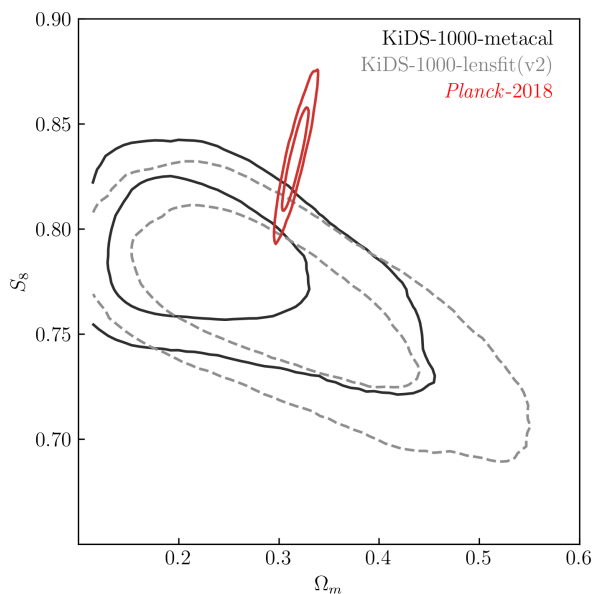


Fig. 20. Projected posteriors for $S_8 \equiv \sigma_8 \sqrt{\Omega_m/0.3}$ and Ω_m from the KiDS-1000 METACALIBRATION (black contours) and *lensfit* (grey dashed contours; Li et al. 2023a) catalogues. The constraints agree well, but thanks to the higher source density, the METACALIBRATION measurements improve the constraints by 28%. The constraints prefer a lower S_8 than *Planck* (red contours), although the difference in S_8 is reduced by 0.5σ to 1.8σ .

difference is reduced by 0.5σ to 1.8σ , while the uncertainties are smaller thanks to the higher source number density in the METACALIBRATION catalogue. The difference remains at the 2.1σ level when measured using Σ_8 . The difference is quantified using the Hellinger tension measure (see appendix G.1. of Heymans et al. 2021). Importantly, the good agreement between our results and those from Li et al. (2023a) suggests that the tension is not due to the shape measurements. Indeed, the most recent analysis of the KiDS-Legacy data Wright et al. (2025b) is based on *lensfit*, but finds good agreement with *Planck* thanks to improvements in the redshift calibration and the addition of more sky area.

6.2. Internal consistency tests

The fiducial setup used in the previous section reflects our current best knowledge of some of the nuisance parameters, and uses all the tomographic bins. However, it is useful to examine the robustness of the findings to some of these setup choices or specific subsets of the data. In this section we explore these, largely following what has been done in previous KiDS analyses (Asgari et al. 2021; Li et al. 2023a; Stölzner et al. 2025). To speed up the calculations, we switched to *multinest* (Feroz et al. 2009), but note that it is known to underestimate the errors by 10% compared to the fiducial run with *polychord* (Handley et al. 2015a,b). However, this is not a concern as we are mostly concerned with overall shifts in posterior distributions. We describe each of the tests below. The findings are summarized in Fig. 21, which shows that the results are robust with respect to the fiducial setup.

We start by exploring the constraining power of each of the tomographic bins. Asgari et al. (2021) found that the second tomographic bin yielded constraints that were somewhat inconsistent with the rest of the data (see their appendix B.2 and figure B.5), with that bin preferring a higher S_8 value. In contrast,

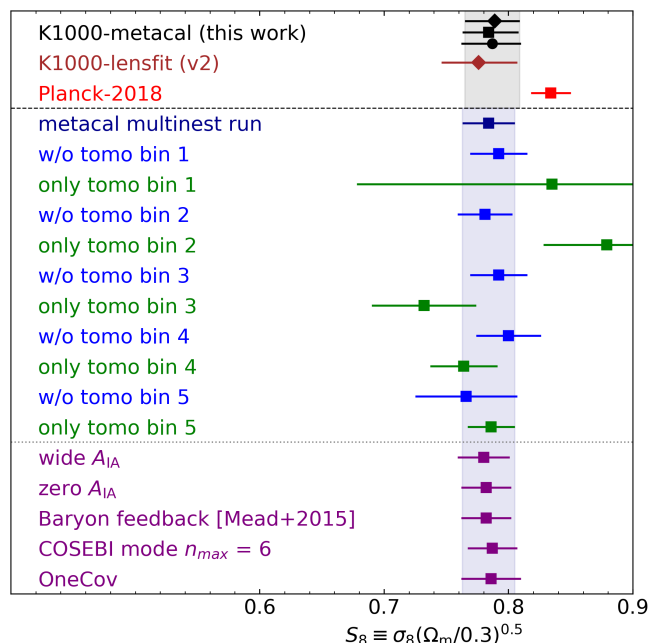


Fig. 21. Fiducial S_8 constraints from METACALIBRATION, *lensfit*, and *Planck* & METACALIBRATION internal consistency test results with different set-ups: tomo-bin splits, variation of intrinsic alignment priors, a baryonic feedback model from Mead et al. (2015), an extra COSEBI mode 6, and a new covariance estimation (Reischke et al. 2025). The consistency holds for all the varied settings, which validates the robustness of the cosmological analysis. Please note that for consistency tests, we used *multinest*, which yields slightly smaller errors than *polychord*. The diamond indicates the MAP + PJ-HPD (fiducial) estimate, the square denotes the mean of the marginal posterior, and the circle represents the maximum of the marginal posterior.

the recent KiDS-Legacy cosmic shear analysis (section 4.1.1 and figure 1 in Stölzner et al. 2025) finds consistency among all tomographic bins. We therefore conducted the same test using our METACALIBRATION catalogue. To do so, we ran the cosmological inference pipeline for each tomographic bin and its cross-correlations with the other bins to ensure sufficient constraining power. These are reported as ‘only tomo bin X’ in Fig. 21, while the joint posteriors for σ_8 and Ω_m are shown as green contours in Fig. 22. We also show the results when a particular bin and its cross-correlations are excluded by the blue points and contours. These results show excellent agreement between all the bins.

One change in our setup with respect to Asgari et al. (2021) is the prior for A_{IA} . If we use the same wide prior $A_{IA} \in [-6.0, 6.0]$, we obtain a consistent S_8 value with a constraint on $A_{IA} = 0.016 \pm 0.435$ (mean). Moreover, as shown in Fig. 23, for this choice of prior we find that the ‘only bin 2’ scenario prefers a higher value for S_8 that is in mild tension with the ‘without bin 2’ results. In addition to the wide prior, we also considered the impact of ignoring intrinsic alignments altogether, that is $A_{IA} = 0$. The results presented in Fig. 21 show that the inferred values for S_8 is robust to the adopted A_{IA} prior, while the S_8 constraint is tighter by ~ 5 percent. This demonstrates the benefit of measuring the A_{IA} parameter externally to increase the constraining power on S_8 , as was done by Wright et al. (2025b).

As shown in Fig. 19, the baryonic feedback-related parameter ($\log_{10}(T_{AGN})$) is not constrained within our adopted narrow prior range, which in turn was based on the validity of the underlying model. To examine the impact, we tested a different

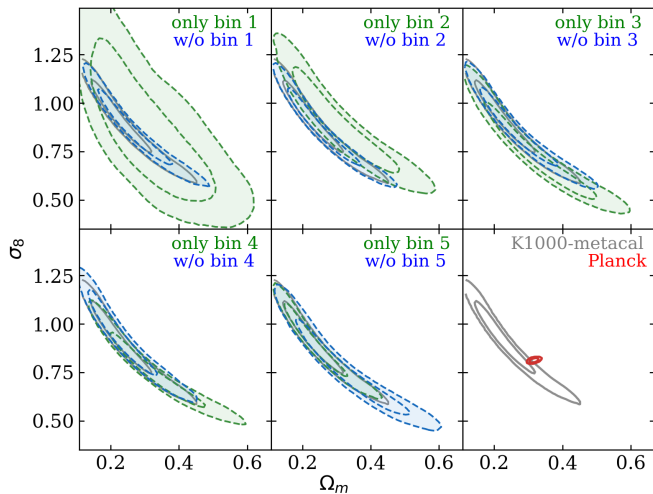


Fig. 22. Consistency between different tomographic bin combinations. For each test run, the independent cosmological constraints (in addition to the baryonic feedback parameter and intrinsic alignment parameter) we obtained using a single bin and its cross-correlations (green) or by excluding these (blue). The parameters shown (Ω_m and σ_8), as well as the values for S_8 , are consistent among all the test sets. Although not significant, the second bin shows the largest deviation, because it prefers a higher intrinsic alignment contribution.

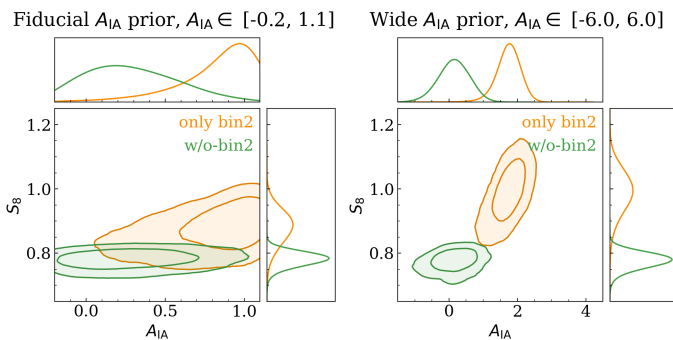


Fig. 23. Impact of the choice of IA prior for the internal consistency test. The left panel shows the posteriors considering only bin 2 (orange) or excluding this bin (green) for our fiducial setup. The right panel shows the same when the wide IA prior from [Asgari et al. \(2021\)](#) is used.

(older) baryonic feedback model from [Mead et al. \(2015\)](#), which is based on the Overwhelmingly Large Simulations (OWLS; [Schaye et al. 2010](#)). It has been used in a number of cosmological analyses ([Asgari et al. 2021](#); [Yoon & Jee 2021](#); [Yoon et al. 2019](#)). However, the differences between the models impact mostly small scales. Thanks to our conservative scale-cut at 2 arcmin, the results are insensitive to the adopted baryonic feedback model.

Finally, we examine the impact of settings that were adopted recently in the KiDS Legacy cosmic shear analysis ([Wright et al. 2025b](#)): including one more COSEBI mode $n_{\max} = 6$, and a new covariance estimation, called ‘OneCov’ ([Reischke et al. 2025](#)). Before running with the 6 COSEBI modes, we checked if this passes the B -mode test. All the tomographic bins passed, while the p -values increased for most bins. The S_8 constraint from the 6 COSEBI modes is consistent with the one from the fiducial run. While it decreases the errors on the S_8 constraint, the gain of constraining power is not significant. With the new covariance (OneCov), we also obtain a consistent S_8 constraint.

7. Conclusions

We revisited the cosmic shear analysis of data from the KiDS-1000 using an alternative shape measurement pipeline, based on METACALIBRATION, which is less sensitive to imperfections in the image simulations used to calibrate the remaining shear biases—in particular, the multiplicative bias. Although METACALIBRATION provides a simple framework to infer the shear response from the actual data, the choice of shape measurement method, weight function, or PSF treatment still impact the overall performance.

We adopted a simple shape measurement, which has the benefit of increasing the number of galaxies for which shapes can be measured. Our setup allowed for an efficient way to determine the bias for redshift-dependent blending, but it also down-weights fainter galaxies. Further optimisation may thus be possible. Nonetheless, we find that the effective number density is higher than that for the original KiDS-1000 analyses, based on *lensfit* ([Asgari et al. 2021](#); [Li et al. 2023a](#)), especially at high redshift. Moreover, the residual multiplicative shear biases are small for all tomographic bins, with the largest bias $m_{\text{final}} = -0.0169$ observed for the second bin (see [Table 2](#)). Compared to *lensfit*, the multiplicative biases are considerably smaller for the highest redshift bins, while the dependence on apparent magnitude is much reduced. Importantly, the calibration has negligible sensitivity to uncertainties in the properties of the galaxies used in the image simulations.

To account for PSF anisotropy we convolve the images with a kernel that circularises the PSF prior to the shape measurement. We find this works well, as the resulting additive bias is small. After we apply an empirical correction to capture any residual PSF leakage, the additive bias is negligible.

The METACALIBRATION measurements pass the tests that were also carried out in previous studies, and the constraint on $S_8 = 0.789^{+0.020}_{-0.024}$ is consistent with the results from [Li et al. \(2023a\)](#), but the constraints are improved by about 28% thanks to the increased effective source number density and the more robust shear calibration by METACALIBRATION. The difference with *Planck* is reduced by 0.5σ , but the difference seen in previous studies still remains at a similar level. This suggests that this is not caused by the shear measurement. Indeed, the recently completed KiDS-Legacy analysis ([Wright et al. 2025b](#)) agrees with *Planck*. The shift in these results is attributed to a combination of more data and improved redshift calibration.

In this study we aimed to stay as close as possible to the previous KiDS-1000 analysis by [Li et al. \(2023a\)](#). As a result, some of the calibration pipeline settings were suboptimal, such as star selection and clustering redshift calibration. We will revisit these when we use our METACALIBRATION pipeline to analyse the KiDS-Legacy data. We expect to improve upon [Wright et al. \(2025b\)](#), because of the improved performance for the highest redshift tomographic bins and the overall higher source number density.

Acknowledgements. MY, HHo, and SSL acknowledge support from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program with Grant agreement No. 101053992. SSL also acknowledges funding from the programme “Netzwerke 2021,” an initiative of the Ministry of Culture and Science of the State of Northrhine Westphalia. LM acknowledges support from STFC grant ST/W000903/1. MY, CH, and BS acknowledge support from the Max Planck Society and the Alexander von Humboldt Foundation in the framework of the Max Planck-Humboldt Research Award endowed by the Federal Ministry of Education and Research. CH also acknowledges support from the UK Science and Technology Facilities Council (STFC) under grant ST/V000594/1. H. Hildebrandt and AHW are supported by a DFG Heisenberg grant (Hi 1495/5-1), the DFG Collaborative Research Center SFB1491, an ERC Consolidator Grant (No. 770935), and the DLR project 50QE2305. BJ acknowledges support by the ERC-selected UKRI Frontier Re-

search Grant EP/Y03015X/1 and by STFC Consolidated Grant ST/V000780/1. MA acknowledges support by the UK Science and Technology Facilities Council (STFC) under grant number ST/Y002652/1 and the Royal Society under grant numbers RGS2222268 and ICAR1231094. RR is partially supported by an ERC Consolidator Grant (No. 770935).

References

- Abbott, T. M. C., Aguena, M., Alarcon, A., et al. 2023, *The Open Journal of Astrophysics*, 6, 36
- Akeson, R., Armus, L., Bachelet, E., et al. 2019, arXiv e-prints, arXiv:1902.05569
- Amon, A. & Efstathiou, G. 2022, *MNRAS*, 516, 5355
- Amon, A., Gruen, D., Troxel, M. A., et al. 2022, *Phys. Rev. D*, 105, 023514
- Asgari, M., Lin, C.-A., Joachimi, B., et al. 2021, *A&A*, 645, A104
- Asgari, M., Mead, A. J., & Heymans, C. 2023, *The Open Journal of Astrophysics*, 6, 39
- Asgari, M., Schneider, P., & Simon, P. 2012, *A&A*, 542, A122
- Asgari, M., Tröster, T., Heymans, C., et al. 2020, *A&A*, 634, A127
- Bagon, D. J., Massey, R. J., Refregier, A. R., & Ellis, R. S. 2003, *MNRAS*, 344, 673
- Benítez, N. 2000, *ApJ*, 536, 571
- Bertin, E. & Arnouts, S. 1996, *A&AS*, 117, 393
- Bridle, S., Balan, S. T., Bethge, M., et al. 2010, *MNRAS*, 405, 2044
- Bridle, S. & King, L. 2007, *New Journal of Physics*, 9, 444
- Capak, P. L. 2004, PhD thesis, UNIVERSITY OF HAWAII
- Czado, C. 2019, *Analyzing Dependent Data with Vine Copulas: A Practical Guide with R* (Cham: Springer International Publishing AG)
- Dalal, R., Li, X., Nicola, A., et al. 2023, arXiv e-prints, arXiv:2304.00701
- Di Valentino, E., Said, J. L., Riess, A., et al. 2025, *Physics of the Dark Universe*, 49, 101965
- Dvornik, A., Heymans, C., Asgari, M., et al. 2023, *A&A*, 675, A189
- Elahi, P. J., Welker, C., Power, C., et al. 2018, *MNRAS*, 475, 5338
- Euclid Collaboration: Csizi, B., Schrabback, T., Grandis, S., et al. 2024, arXiv e-prints, arXiv:2409.07528
- Euclid Collaboration: Martinet, N., Schrabback, T., Hoekstra, H., et al. 2019, *A&A*, 627, A59
- Euclid Collaboration: Mellier, Y., Abdurro'uf, Acevedo Barroso, J. A., et al. 2024, arXiv e-prints, arXiv:2405.13491
- Fenech Conti, I., Herbonnet, R., Hoekstra, H., et al. 2017, *MNRAS*, 467, 1627
- Feroz, F., Hobson, M. P., & Bridges, M. 2009, *MNRAS*, 398, 1601
- Fortuna, M. C., Hoekstra, H., Johnston, H., et al. 2021, *A&A*, 654, A76
- Fosalba, P., Crocce, M., Gaztañaga, E., & Castander, F. J. 2015, *Monthly Notices of the Royal Astronomical Society*, 448, 2987
- Fosalba, P., Gaztañaga, E., Castander, F. J., & Crocce, M. 2014, *Monthly Notices of the Royal Astronomical Society*, 447, 1319
- Gatti, M., Sheldon, E., Amon, A., et al. 2021, *MNRAS*, 504, 4312
- Georgiou, C., Chisari, N. E., Fortuna, M. C., et al. 2019, *A&A*, 628, A31
- Giblin, B., Heymans, C., Asgari, M., et al. 2021, *A&A*, 645, A105
- Girardi, L., Groenewegen, M. A. T., Hatziminaoglou, E., & da Costa, L. 2005, *A&A*, 436, 895
- Gong, Y., Liu, X., Cao, Y., et al. 2019, *ApJ*, 883, 203
- Griffith, R. L., Cooper, M. C., Newman, J. A., et al. 2012, *ApJS*, 200, 9
- Handley, W. J., Hobson, M. P., & Lasenby, A. N. 2015a, *MNRAS*, 450, L61
- Handley, W. J., Hobson, M. P., & Lasenby, A. N. 2015b, *MNRAS*, 453, 4384
- Heymans, C., Tröster, T., Asgari, M., et al. 2021, *A&A*, 646, A140
- Heymans, C., Van Waerbeke, L., Bacon, D., et al. 2006a, *MNRAS*, 368, 1323
- Heymans, C., Van Waerbeke, L., Bacon, D., et al. 2006b, *MNRAS*, 368, 1323
- Heymans, C., Van Waerbeke, L., Miller, L., et al. 2012, *MNRAS*, 427, 146
- Hildebrandt, H., van den Busch, J. L., Wright, A. H., et al. 2021, *A&A*, 647, A124
- Hirata, C. & Seljak, U. 2003, *MNRAS*, 343, 459
- Hoekstra, H. 2004, *Monthly Notices of the Royal Astronomical Society*, 347, 1337
- Hoekstra, H. 2021, *A&A*, 656, A135
- Hoekstra, H., Franx, M., Kuijken, K., & Squires, G. 1998, *ApJ*, 504, 636
- Hoekstra, H., Herbonnet, R., Muzzin, A., et al. 2015, *MNRAS*, 449, 685
- Hoekstra, H., Kannawadi, A., & Kitching, T. D. 2021, *A&A*, 646, A124
- Hoekstra, H., Viola, M., & Herbonnet, R. 2017, *MNRAS*, 468, 3295
- Huff, E. & Mandelbaum, R. 2017, arXiv e-prints, arXiv:1702.02600
- Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, *ApJ*, 873, 111
- Jansen, H., Tewes, M., Schrabback, T., et al. 2024, *A&A*, 683, A240
- Joachimi, B., Cacciato, M., Kitching, T. D., et al. 2015, *Space Sci. Rev.*, 193, 1
- Joachimi, B., Lin, C. A., Asgari, M., et al. 2021, *A&A*, 646, A129
- Joe, H. 2014, *Dependence Modeling with Copulas* (CRC press)
- Kacprzak, T., Bridle, S., Rowe, B., et al. 2014, *MNRAS*, 441, 2528
- Kaiser, N., Squires, G., & Broadhurst, T. 1995, *ApJ*, 449, 460
- Kannawadi, A., Hoekstra, H., Miller, L., et al. 2019, *A&A*, 624, A92
- Kannawadi, A., Rosenberg, E., & Hoekstra, H. 2021, *MNRAS*, 502, 4048
- Kilbinger, M., Heymans, C., Asgari, M., et al. 2017, *MNRAS*, 472, 2126
- Kitching, T. D. & Deshpande, A. C. 2022, *The Open Journal of Astrophysics*, 5, 6
- Kitching, T. D., Paykari, P., Hoekstra, H., & Cropper, M. 2019, *The Open Journal of Astrophysics*, 2, 5
- Kuijken, K. 2006, *A&A*, 456, 827
- Kuijken, K., Heymans, C., Dvornik, A., et al. 2019, *A&A*, 625, A2
- Kuijken, K., Heymans, C., Hildebrandt, H., et al. 2015, *MNRAS*, 454, 3500
- Lagos, C. d. P., Tobar, R. J., Robotham, A. S. G., et al. 2018, *MNRAS*, 481, 3573
- Li, S.-S., Hoekstra, H., Kuijken, K., et al. 2023a, *A&A*, 679, A133
- Li, S.-S., Hoekstra, H., Kuijken, K., Schaller, M., & Schaye, J. 2024, arXiv e-prints, arXiv:2411.03152
- Li, S.-S., Kuijken, K., Hoekstra, H., et al. 2023b, *A&A*, 670, A100
- Li, X., Zhang, T., Sugiyama, S., et al. 2023c, arXiv e-prints, arXiv:2304.00702
- Linke, L., Unruh, S., Wittje, A., et al. 2025, *A&A*, 693, A210
- LoVerde, M. & Afshordi, N. 2008, *Phys. Rev. D*, 78, 123506
- Luppino, G. A. & Kaiser, N. 1997, *ApJ*, 475, 20
- MacCrann, N., Becker, M. R., McCullough, J., et al. 2022, *MNRAS*, 509, 3371
- Mandelbaum, R. 2018, *ARA&A*, 56, 393
- Mandelbaum, R., Hirata, C. M., Seljak, U., et al. 2005, *MNRAS*, 361, 1287
- Massey, R., Heymans, C., Bergé, J., et al. 2007, *MNRAS*, 376, 13
- McCarthy, I. G., Schaye, J., Bird, S., & Le Brun, A. M. C. 2017, *MNRAS*, 465, 2936
- Mead, A. J., Brieden, S., Tröster, T., & Heymans, C. 2021, *MNRAS*, 502, 1401
- Mead, A. J., Peacock, J. A., Heymans, C., Joudaki, S., & Heavens, A. F. 2015, *MNRAS*, 454, 1958
- Melchior, P., Viola, M., Schäfer, B. M., & Bartelmann, M. 2011, *MNRAS*, 412, 1552
- Miller, L., Heymans, C., Kitching, T. D., et al. 2013, *MNRAS*, 429, 2858
- Miyatake, H., Sugiyama, S., Takada, M., et al. 2023, *Phys. Rev. D*, 108, 123517
- Planck Collaboration VI. 2020, *A&A*, 641, A6
- Raichoor, A., Mei, S., Erben, T., et al. 2014, *ApJ*, 797, 102
- Reischke, R., Unruh, S., Asgari, M., et al. 2025, *A&A*, 699, A124
- Rowe, B. T. P., Jarvis, M., Mandelbaum, R., et al. 2015, *Astronomy and Computing*, 10, 121
- Schaye, J., Dalla Vecchia, C., Booth, C. M., et al. 2010, *MNRAS*, 402, 1536
- Schneider, P., Eifler, T., & Krause, E. 2010, *A&A*, 520, A116
- Schneider, P., van Waerbeke, L., & Mellier, Y. 2002, *A&A*, 389, 729
- Scoville, N., Aussel, H., Brusa, M., et al. 2007, *ApJS*, 172, 1
- Secco, L. F., Samuroff, S., Krause, E., et al. 2022, *Phys. Rev. D*, 105, 023515
- Semboloni, E., Hoekstra, H., Schaye, J., van Daalen, M. P., & McCarthy, I. G. 2011, *MNRAS*, 417, 2020
- Sérsic, J. L. 1963, *Boletín de la Asociación Argentina de Astronomía La Plata Argentina*, 6, 41
- Sheldon, E. S., Becker, M. R., Jarvis, M., Armstrong, R., & LSST Dark Energy Science Collaboration. 2023, *The Open Journal of Astrophysics*, 6, 17
- Sheldon, E. S., Becker, M. R., MacCrann, N., & Jarvis, M. 2020, *ApJ*, 902, 138
- Sheldon, E. S. & Huff, E. M. 2017, *ApJ*, 841, 24
- Stölzner, B., Wright, A. H., Asgari, M., et al. 2025, arXiv e-prints, arXiv:2503.19442
- Tewes, M., Kuntzer, T., Nakajima, R., et al. 2019, *A&A*, 621, A36
- Valdes, F., Tyson, J. A., & Jarvis, J. F. 1983, *ApJ*, 271, 431
- van den Busch, J. L., Wright, A. H., Hildebrandt, H., et al. 2022, *A&A*, 664, A170
- Wright, A. H., Hildebrandt, H., van den Busch, J. L., et al. 2025a, arXiv e-prints, arXiv:2503.19440
- Wright, A. H., Hildebrandt, H., van den Busch, J. L., & Heymans, C. 2020, *A&A*, 637, A100
- Wright, A. H., Kuijken, K., Hildebrandt, H., et al. 2024, *A&A*, 686, A170
- Wright, A. H., Stölzner, B., Asgari, M., et al. 2025b, arXiv e-prints, arXiv:2503.19441
- Yoon, M. & Jee, M. J. 2021, *ApJ*, 908, 13
- Yoon, M., Jee, M. J., Tyson, J. A., et al. 2019, *ApJ*, 870, 111
- Zhan, H. 2011, *Scientia Sinica Physica, Mechanica & Astronomica*, 41, 1441
- Zhang, Z., Sheldon, E. S., & Becker, M. R. 2023, *The Open Journal of Astrophysics*, 6, 16
- Zuntz, J., Sheldon, E., Samuroff, S., et al. 2018, *MNRAS*, 481, 1149

Appendix A: Source selection for the KiDS META CALIBRATION catalogue

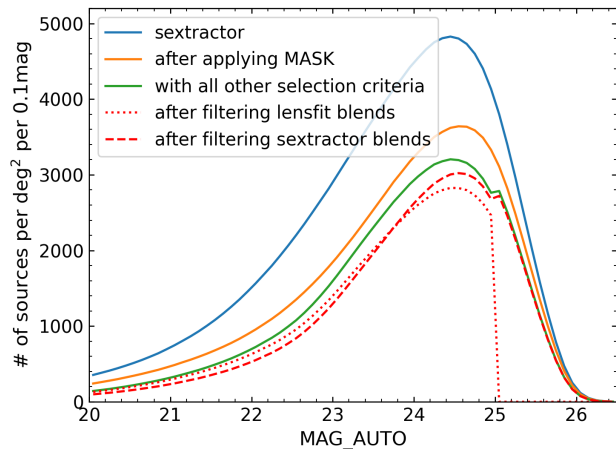


Fig. A.1. The impact of various selection criteria on the number density of sources as a function of magnitude. Removed objects in masked regions and flagged objects has the biggest impact. The two different selections of blended sources are discussed in Section A.2.

Our analysis is based on the same imaging data, released as part of KiDS-DR4 (Kuijken et al. 2019), that was used in previous KiDS-1000 shear studies. We therefore start from the same catalogue of sources detected by SOURCEEXTRACTOR. Moreover, to allow for a more meaningful comparison of the performance of *lensfit* and META CALIBRATION, where possible, we applied the same selections for various sources of contamination. In particular, we took advantage of the fact that *lensfit* classifies all objects in the input catalogue; the use of this *fitclass* enabled us to reproduce the star selection, without having to introduce new criteria.

For our cosmological analysis, we used the flagging of blended objection by *lensfit*, but we also explored an alternative approach based on the SOURCEEXTRACTOR flags, which is detailed in Section A.2. As detailed in Section A.1, we did have to modify the removal of unresolved binary stars, because the impact depends on the shape measurement algorithm.

The sample of galaxies used for the lensing analysis is based on the criteria listed in Table A.1, which also provides the number of sources affected by each selection. These results show that masking (to avoid bright stars, artifacts, etc.) and blended objects account for most of the reduction in sample size. As shown in Fig. A.1 those selections reduce the numbers over the full magnitude range. Once all criteria are combined, about 43% of the sources remain. This sample is reduced further after the photometric redshift calibration is considered, resulting in a total number of 30,401,986 sources. Although we closely follow the selection criteria used in previous KiDS-1000 analyses, the final META CALIBRATION sample is 30% larger than the *lensfit* v2 catalogue.

Appendix A.1: Unresolved binary stars

Unresolved binary stars show up as highly bright, compact, but highly elliptical sources. In previous analyses, the removal of these objects was based on *lensfit* outputs (see e.g. figure 30 in Wright et al. 2024). Although we expect many of these to be

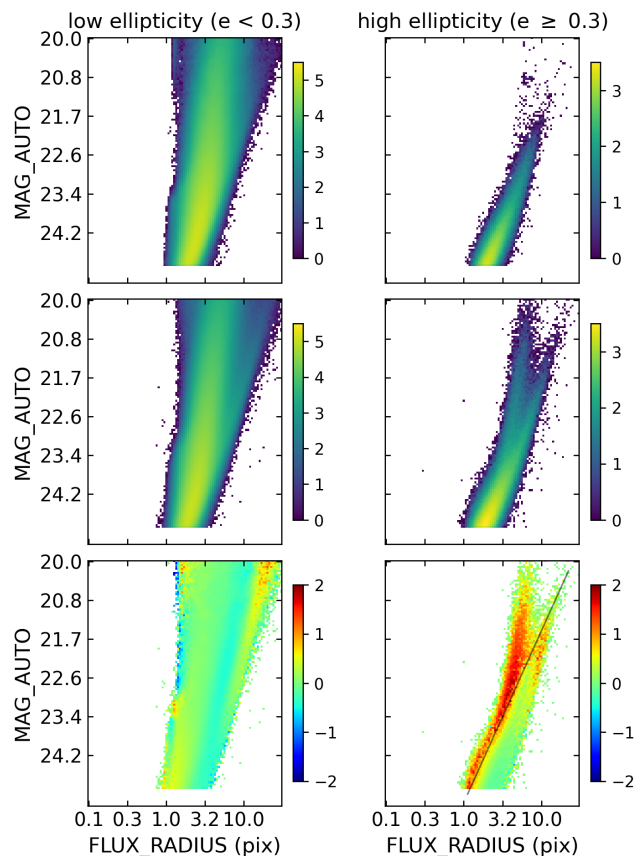


Fig. A.2. Logarithm of the number density of sources as a function of size (FLUX_RADIUS) and apparent magnitude (MAG_AUTO), split by the observed third eccentricity, $|e|$, as defined by Eq. (1). The top and middle rows show the results for SKiDS and KiDS, respectively. The bottom row shows the difference between the distributions (KiDS–SKiDS), highlighting a population of bright, compact, but highly elliptical sources in KiDS data that are lacking in SKiDS. These are binary stars that contaminate the source catalogue. We remove those objects that lie above the line drawn in the lower right panel, which is given by Eq. (A.1). The colour bars denote the logarithm of the number counts (base 10) in each size-magnitude bin, which are the same for all the subplots.

present in our catalogue, we decided to base our selection the shape estimates from Eq. (1). To do so, we compared the number density of sources as a function of size and magnitude in both SKiDS and KiDS, split by $|e|$. The results are presented in Fig. A.2. The bottom left panel shows that the distributions for objects with $|e| < 0.3$ agree well between SKiDS and KiDS-1000. In contrast, we observe an excess of bright, compact objects for elliptical sources ($|e| \geq 0.3$) in the actual KiDS data. Based on these results we defined the following empirical selection, and remove sources with $|e| \geq 0.3$ and (black line in Fig. A.2)

$$\text{MAG_AUTO} \leq -1.728 \times \ln(\text{FLUX_RADIUS}) + 25.56. \quad (\text{A.1})$$

As reported in Table A.1, this removes 0.83% of the sources. This fraction is larger than the 0.14% reported in Wright et al. (2024), but this may well reflect the ability of *lensfit* to reject some objects as bad fits at an earlier stage. In contrast, META CALIBRATION tends to return successful shape measurements for almost all bright sources.

Table A.1. Metacal catalogue source selection

selection criteria	parameters	# of filtered sources	fraction [%]
input KiDS-1000	–	112,335,001	100
masked area	(Mask & 27676)>0	32,644,554	29.06
9 band GAAP flags	FLAG_GAAP_anyband != 0	9,148,548	8.14
stars	fitclass = 1 or 2	8,255,613	7.35
asteroids	MAG_GAAP (g-r) > 1.5 and MAG_GAAP (i-r) > 1.5	4,339,934	3.86
potential blends	fitclass = -10 or contamination_radius ≤ 4.25	29,487,117	26.25
insufficient pixels on cutouts		1,666,036	1.48
binary stars (for e ≥ 0.3)	MAG_AUTO > - 1.728 × log(FLUX_RADIUS) + 25.56	933,517	0.83
magnitude cuts	20 < MAG_AUTO ≤ 25	19,113,615	17.01
all 6 selection criteria combined		63,679,324	56.68
remaining sources		48,655,677	43.31
tomographic binning (6 bins)	0.1 < z _B ≤ 2.0	47,080,020	–
tomographic binning (5 bins)	0.1 < z _B ≤ 1.2	40,412,178	100
+S/N and resolution cuts	v _{SN} > 7 and R < 0.6	33,906,706	83.90
+gold (photo-z SOM) selection	Gold_FLAG = 1	27,523,376	68.11
+AlphaRecal selection	calibrated e1, e2 < 1	27,523,376	68.11
North patch		12,933,729	46.99
South patch		14,589,647	53.01

Notes. The initial KiDS-1000 catalog, generated by the SOURCEEXTRACTOR detection runs, contains sources from 1000 tiles (see details in Appendix C), which are the tiles after excluding the ones with high PSF residuals. The fraction of filtered sources for each selection is calculated relative to the total detected sources. Once all criteria are considered, 56.68% of the original objects are filtered out. The final number of sources, 275,233,76, is 18% more than the one from the *lensfit* v2 catalogue, 23,401,764.

Appendix A.2: Selection of blended sources

Blended sources are a major source of shear bias, and the main reason for the difference between the multiplicative bias of METACALIBRATION and the detection bias from SOURCEEXTRACTOR (Hoekstra et al. 2021). A further complication is posed by blended sources at different redshift (MacCrann et al. 2022; Li et al. 2023b). Hence, removing blended sources may help reduce the overall bias and improve the robustness of the shear calibration. To this end, we use the segmentation map from SOURCEEXTRACTOR to mask pixels from neighbouring detected sources (see Section 3.1), but this cannot fully eliminate the impact of blended sources.

The image simulations provide a convenient test bed to explore the efficacy of blending selection criteria. In particular, the criteria could be optimised, and subsequently be applied to the data. However, it is important that the simulated data capture the clustering of galaxies adequately (Euclid Collaboration: Martinet et al. 2019). As shown in Li et al. (2023b), SKiLLS reproduces the observed distribution of magnitudes and separations between neighbouring galaxies well.

To explore the impact on the multiplicative bias, we considered two selection criteria to filter out the potential (recognized) blends:

- 1) select sources with a *lensfit* output flag `fitclass != -10` and a *lensfit* contamination radius ≤ 4.25 pixels;
- 2) select the sources with SOURCEEXTRACTOR output `FLAGS=0`.

The first is the selection used for the previous KiDS analyses, which are all based on catalogues produced with *lensfit* (Giblin et al. 2021). However, as one would like the METACALIBRATION pipeline to be independent of *lensfit*, it is useful to compare the performance to the second criterium, which corresponds to the one used by the DES pipeline (Gatti et al. 2021). As shown

in Fig. A.1, the selected sources have slightly different magnitude distributions, where the sharp drop for $m > 25$ arises from an internal selection in *lensfit*; most of them are assigned with `fitclass = -1` (flag for objects with insufficient data) and also with `contamination_radius = 0`.

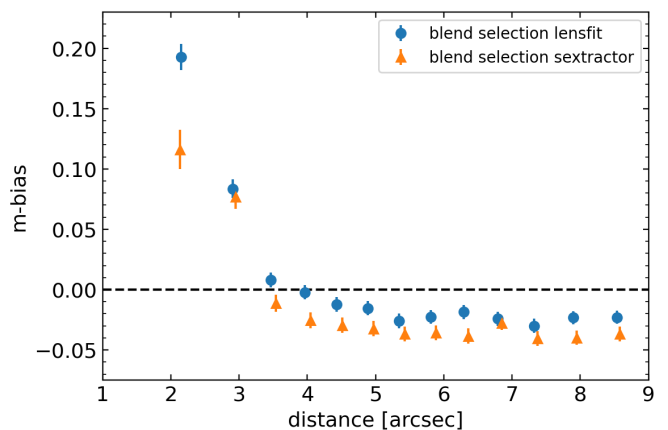


Fig. A.3. Multiplicative shear bias as a function of distance to the nearest detected neighbour, after removing blended sources using the two criteria described in the text. The blue points correspond to the baseline KiDS-1000 analysis (Giblin et al. 2021), whereas the orange points reflect the DES pipeline (Gatti et al. 2021).

Figure A.3 shows the multiplicative shear bias as a function of distance to the nearest detected source, after one of the two criteria was applied. For apparently isolated sources the bias does not vanish, because many blends are not detectable (Hoekstra et al. 2021). The smaller bias for *lensfit* (blue points) for large separations, suggests that it recognises blends a bit better than our SOURCEEXTRACTOR setup. Nonetheless, the difference

between both criteria is small, also for small separations, where the biases increase rapidly. Hence, we do not expect the final cosmological results to depend on the selection.

However, we adopted the selection based on the *lensfit* output, because it performs a bit better, but more importantly, because it allows for a more direct comparison with previous work. As Fig. A.3 shows, it might still be beneficial for future KiDS analyses to consider eliminating all sources with neighbours detected within $3''$, as the bias rises quickly, although they may still be of interest for galaxy-galaxy lensing studies.

Appendix A.3: Star contamination

Another important class of potential contaminants are stars that are not identified as such. At bright magnitudes, stars can generally be separated well based on their size, unless they are blended with another source. However, the size estimates are noisier for fainter stars, and overlap with the observed sizes of galaxies. As a result, the shear bias depends on the star density (Hoekstra et al. 2015, 2017). In Section 4.4 we explored the sensitivity of the multiplicative bias to variations in star density, and found that the impact appears negligible. However, for the second tomographic bin, Fig. 15 indicated a change in m for the highest star density sample. Therefore, we explore the contamination by stars in more detail here.

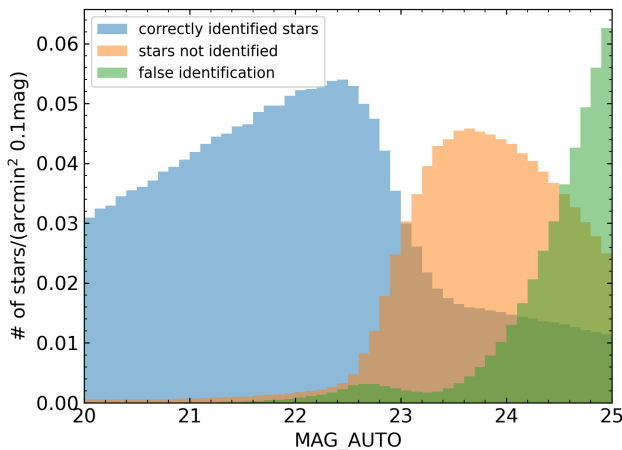


Fig. A.4. The number density of stars in the catalogue of sources detected by *sextractor* in SKiLLS. The sources were cross-matched with the input catalogues, resulting in a sample of correctly identified stars (*fitclass* = 1 or 2), missed stars, and incorrectly identified objects, denoted by the blue, orange, and green histograms, respectively. The *lensfit* flagging starts to fail for stars fainter than $\text{MAG_AUTO} > 22.5$.

As detailed in section 2.2 of Li et al. (2023b), SKiLLS includes a realistic population of stars, based on the population synthesis code *Trilegal* (Girardi et al. 2005). The star densities are sampled from the KiDS tiles to account for the variation with sky position.

To quantify the level of star contamination, we cross-matched the detected sources to the input catalogue of SKiLLS. This resulted in a sample of objects that were correctly flagged as stars by *lensfit* (*lensfit_class* = 1 or 2), indicated by the blue histogram in Fig. A.4, which works well for bright stars with $\text{MAG_AUTO} < 22.5$. At faint magnitudes, it is harder to distinguish stars from galaxies. The latter are more numerous and hence the rate of false positives increases, as indicated by the

green histogram in Fig. A.4. As these objects are removed from the galaxy catalogues, they do affect the shear bias.

The galaxy sample is contaminated by stars that are not identified as such, which is the case for most stars with $\text{MAG_AUTO} > 22.5$ (orange histogram in Fig. A.4). As a result, unidentified stars contaminate the tomographic bins by 2.9, 6.9, 4.3, 2.7, 0.3, and 0.3 percent, respectively. In the case of *lensfit*, the shear weights effectively reduce their impact to negligible levels, but as we opted for a weighting scheme that depends only on magnitude for *META*CALIBRATION, the contamination is reduced only slightly to 2.7, 4.4, 3.1, 1.9, 0.3, and 0.3 percent, respectively.

Although this level of contamination is not ideal, we note that the star counts match those in the KiDS data, and hence is accounted for in our estimates for m . Moreover, the results presented in Fig. 15 show that the relatively high fraction of stars in the second tomographic bin has only a modest impact on the shear bias, as $\Delta m \approx \pm 0.006$. As an extreme test, we also computed the change in m when all input stars are removed. In this case we find $\Delta m \approx \pm 0.0047$ for the second bin, consistent with our sensitivity test.

For a future analyses, that does not use the *lensfit* star selection, we plan to reduce the impact of stars by changing the weighting or selection schemes of stars. For instance, a simple cut based on the observed sizes (*FLUX_RADIUS*) seems more effective (especially for the sources fainter than magnitude 22.5). We leave this for future work, as we prefer to stay as closely as possible to the *lensfit* selection for a direct comparison.

Appendix B: Changing the size of weight function

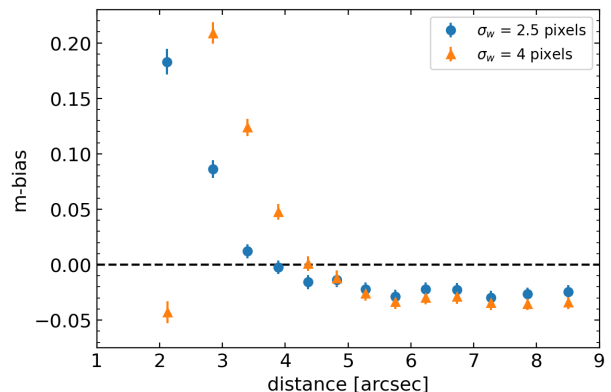


Fig. B.1. Multiplicative bias, m , measured as a function of distance to the nearest detected source ($\text{MAG_AUTO} < 24.5$) using a circular weight function with $\sigma_w = 2.5$ pixels (blue) and $\sigma_w = 4$ pixels (orange). The bias converges to similar values for isolated sources, but the effect of blending appears for smaller separations in the case of the more compact weight function.

As described in section 3.1, we measured the quadrupole moments of the source surface brightness distributions using a fixed weight function. By construction, this removes the sensitivity of the weighting scheme to the shear. Based on the observed distribution of galaxy sizes (see Fig. 1) we adopted a dispersion of $\sigma_w = 2.5$ pixels as baseline.

To quantify the impact of the width of the weight function, we repeated the *META*CALIBRATION shear calibration using a weight function with $\sigma_w = 4$ pixels. We compared the uncertainty in the inferred multiplicative bias, which is a measure for the overall precision with which the lensing signal can be measured. We found that the uncertainty in the shear measurement

increased by 16% when the width was increased from $\sigma_w = 2.5$ to $\sigma_w = 4$ pixels.

As the width of the weight function does not have a big impact on the precision of the shear estimate, other considerations can be taken into account to optimise the choice of weight function, such as the sensitivity to blending. A more compact weight function suppresses the light from close sources. To quantify the impact of blended sources, we determined the multiplicative bias as a function of distance to the nearest detected source for galaxies with $\text{MAG_AUTO} < 24.5$. The results for the two weight functions are presented in Figure B.1. The multiplicative bias converges to similar values for isolated sources, but the effect of blending appears for smaller separations in the case of the more compact weight function (blue points). This further motivates our choice for the fiducial weight function with $\sigma_w = 2.5$ pixels.

Appendix C: PSF residual check

Imperfect PSF modelling can introduce significant bias in shear measurement (Hoekstra 2004). We ignored this when quantifying the shear bias using SKiLLS, because we assumed the PSF model is known perfectly. Li et al. (2023b) explored the sensitivity of *lensfit* to PSF modelling errors using SKiLLS, and found that this can shift the multiplicative bias by about $\Delta m = 0.005$. Here, we do not revisit the actual PSF modelling, also because previous work has validated the fidelity of the PSF model (Giblin et al. 2021). Instead we examine a different aspect that is relevant for our analysis of KiDS data.

The PSF model was developed for the *lensfit* pipeline (see Miller et al. 2013; Kuijken et al. 2015, for details) and derived from stars in the individual exposures. This is used by *lensfit* to simultaneously fit a model of the galaxy surface brightness to the individual exposures (Miller et al. 2013). We use the same model, but our METACALIBRATION pipeline uses the stack of the postage stamp images. Therefore, we created a stacked PSF model for each galaxy.

During the step of image stacking, we excluded the chip images with fewer than 30 stars to ensure the accuracy of PSF model. Such low numbers might indicate problems with the images (e.g. a faulty video board) or because bright stars render large areas useless. This matches a selection that is conducted internally by *lensfit*, and excludes 756 individual chip images. Its necessity is demonstrated by a reduction in PSF residuals in the affected images.

In principle, the PSF of the stacked images should match the average PSF model for that location. A direct comparison of the two images thus tests if the stacking is done correctly. However, astrometric errors can introduce a bias between the two PSFs, thus impacting the accuracy of the METACALIBRATION measurements. The SKiLLS setup assumes perfect astrometry, and thus cannot be used to test this specific aspect. We therefore validated this step using the KiDS data.

We measured the ellipticities of the model PSFs that were reconstructed from the parametrized equations using the coefficients determined during the PSF modelling. At the locations of the stars, we measured ellipticities using `galsim.hsm.FindAdaptiveMom`, which uses adaptive elliptical weighting to estimate moment-based ellipticities. We repeated this measurement on the actual stacked KiDS *r*-band images, and show the difference ($de = e_i^{\text{mod}} - e_i^{\text{obs}}$) averaged in bins of 12×12 arcmin in Fig. C.1.

The top two panels show the results for KiDS-North, while the bottom two panels show the same for KiDS-South. The

residuals are generally consistent with zero, but we did identify five pointings that showed significant deviations.¹² These are marked by red squares in Fig. C.1. Upon further inspection, we found problems with the astrometry between the exposures, and excluded them from the analysis. We also omitted pointing ‘356p6_m29p2’, because of a corrupted file. Hence, the final selection results in 1000 pointings and an effective area of 771.90 deg^2 .

Appendix D: Linearity test of shear bias

To determine the shear biases we apply a constant shear of amplitude $|\gamma| = 0.04$ to the simulated galaxies, while we used more realistic, smaller values, to account for blended sources at different redshift in Section 4.3. However, if the value of m depends on the shear applied, our estimates may be biased (Heymans et al. 2006b). To validate the fidelity of the fixed input shear calibration, we quantify here the linearity of our METACALIBRATION implementation, using a wide range of shear inputs, $|\gamma| \in [0, 0.2]$.

The setup of the image simulations is similar to the baseline, but we made some simplifications to reduce computing time. We only modify the *r*-band images, and do not repeat the determination of photometric redshifts from the multi-band images. Also, we fixed the PSF model. Otherwise the analysis is unchanged, including the use of multiple exposures. Further details, as well as the corresponding results from the same simulations for *lensfit* can be found in Appendix A of Li et al. (2024).

The change in m with respect to the $|\gamma| = 0.04$ case is shown in Fig. D.1 as a function of the input shear. We assume that the change in m is the same for both shear components. The blue points indicate the results if we simply repeat our calibration, assuming a linear relation, but with a larger shear. Alternatively, we can consider all the simulations up to $|\gamma_{\text{max}}|$, indicated by the orange points. In both cases, we find that m does not depend on the input shear up to $|\gamma| = 0.075$; for larger shears, the relation becomes non-linear. However, even for the largest shears that we considered, the changes are well within the requirements for KiDS-1000. Comparing to the findings by Li et al. (2024) using *lensfit*, the level of non-linearity is comparable. Although Li et al. (2024) observed a small quadratic contribution, the dominant non-linearity depends on the cube of the input shear, which is expected based on symmetry arguments. Therefore, we also fit a third-order polynomial, and indicate the residuals by the green points in Fig. D.1.

These results show that our baseline setup provides an accurate estimate for m for the shears relevant for our cosmic shear analysis. In fact, given the linearity and the final calibration procedure, we could have used a somewhat larger shear for the fixed shear calibration, thus reducing the number of simulated images.

¹² These are the pointings ‘170p0_1p5’, ‘15p3_m32p1’, ‘28p8_m30p2’, ‘330p8_m27p2’, and ‘355p2_m33p1’.

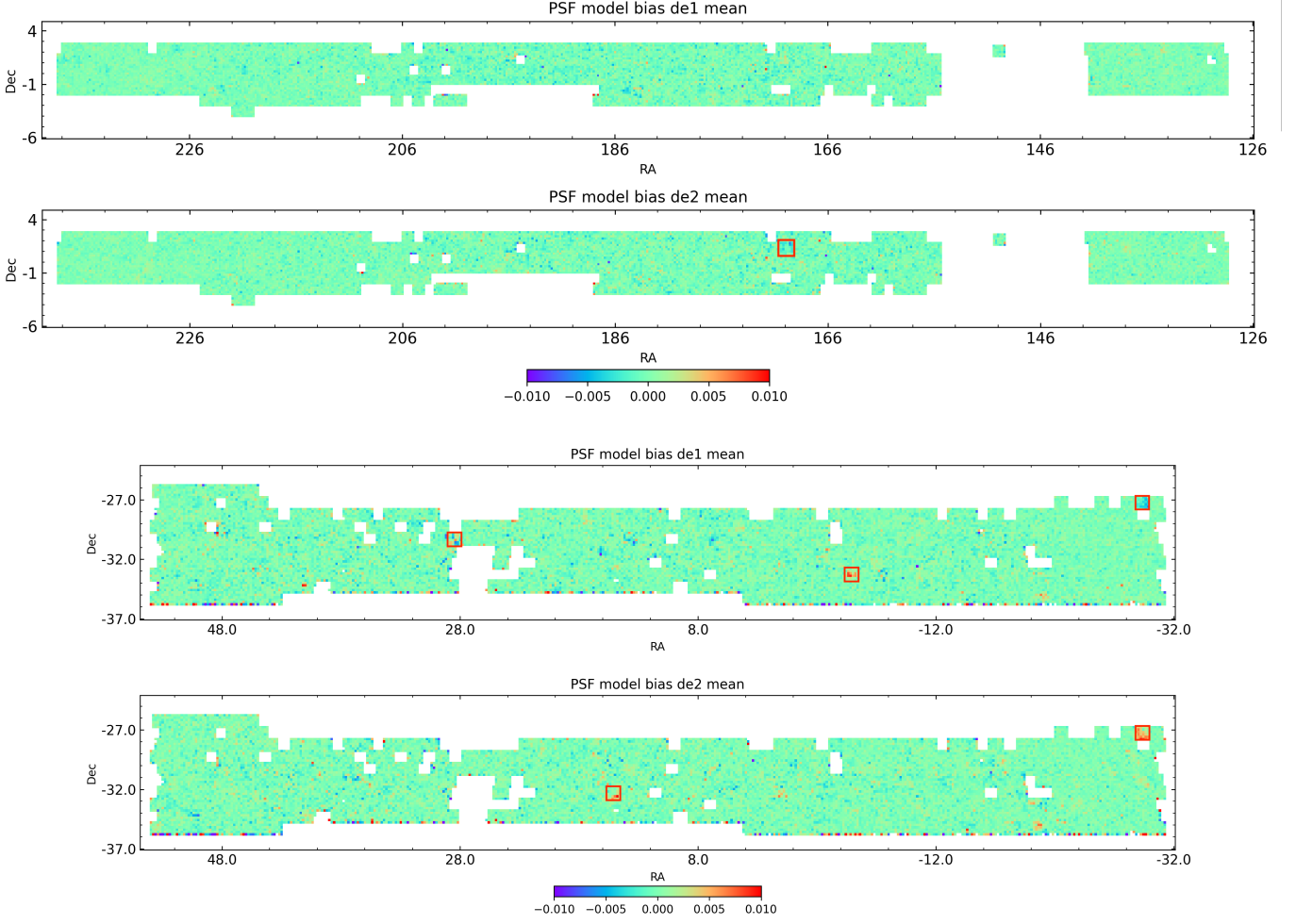


Fig. C.1. PSF model residual over the KiDS survey coverage. The residual is averaged over stars, which is defined as the deduction of the measured ellipticity of the stacked star images from the ellipticity of the PSF is modelled at the star location ($de = e_{\text{PSF}} - e_{\text{star}}$). (top: northern fields, de_1 , de_2 , bottom: southern fields, de_1 , de_2 , respectively) This residual figure demonstrates the accuracy of astrometry as well as the accuracy of PSF models.

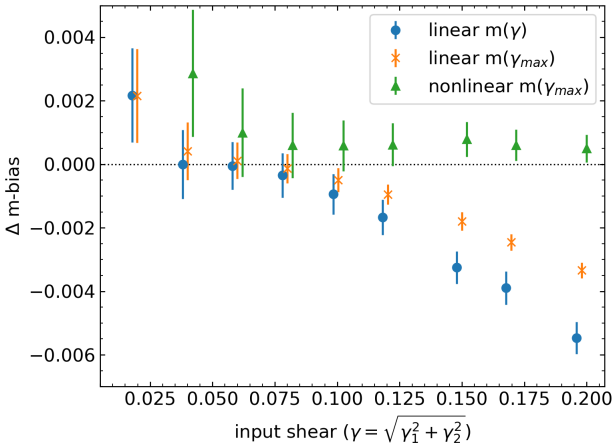


Fig. D.1. Change in m relative to the baseline input shear $|\gamma| = 0.04$ as a function of input shear. The blue points correspond to our linear calibration procedure with a single constant shear. The orange points use all shears up to $|\gamma_{\text{max}}|$, yielding similar but more precise results. Our META-CALIBRATION pipeline is well within the linear regime for our setup. The green points show the residuals if we consider a third-order polynomial instead.